MAKING PREDICTIONS DIRECTLY FROM PAST EXPERIENCES

By

A. Julian Craddock
Bachelor of Cognitive Science Queen's University 1984
Master of Science (Computer Science) Queen's University 1986

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy
in
THE FACULTY OF GRADUATE STUDIES
(DEPARTMENT OF COMPUTER SCIENCE)

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA
1993

In presenting this thesis in partial fulfillment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

University of B.C.
Department of Computer Science
2075 WesBrook Mall
Vancouver, B.C.

Date: Dec 23, 1993

Abstract

This thesis considers the problem of making predictions about new experiences based upon past experiences. The problem is of interest to artificial intelligence because past experiences are a kind of domain knowledge that is readily available to computational agents, and are at least one form of knowledge that humans use to make predictions.

Instead of considering the problem in terms of first inducing a domain model from a set of past experiences, and then using some form of deduction to make predictions, this thesis develops a new technique called the reference class approach (RCA) that *directly* infers estimates of conditional probabilities from a knowledge base of past experiences. The resulting estimates can be readily used in a number of contexts such as non-monotonic reasoning, the characterisation of probability distribution functions, prediction and classification.

Given a knowledge base (KB) of descriptions of past experiences, a description of a new experience, and a proposition representing a query about the new experience, the RCA estimates the conditional probability of the proposition being true of the new experience. The RCA starts by identifying a subset of the KB called the reference class that contains all those past experiences in the KB whose descriptions cover everything that is known about the new experience in addition to providing a truth value for the proposition.

If there are no directly applicable past experiences, i.e., the reference class is empty, then the description of the new experience is modified until a non-empty reference class can be found. This thesis investigates two new approaches to modifying the description, namely syntactic generalisation and chaining. Previous research has proposed that logical implication can be used to semantically generalise an empty reference class to any non-empty reference class. This thesis shows that semantic generalisation does not work in the context of making predictions from a KB of past experiences. This thesis argues that we should syntactically generalise the description of the new experience. Chaining is a novel extension of syntactic generalisation that allows us to systematically increase what we know about a new experience by elaborating its description while generalising. Once a non-empty reference class has been identified the RCA estimates the conditional probability of the proposition being true by measuring the frequency with which the proposition is true in the reference class.

The RCA is an inductive technique in that it estimates probabilities directly from past experiences. One useful test of an inductive technique is to test whether or not it can be used to make accurate predictions from past experiences. This thesis argues that in order to implement the RCA we need a notion of irrelevance to pick the most appropriate generalised or chained reference class. This thesis shows that even with very simple notions of irrelevance, the RCA's estimates can be used to make predictions whose accuracy compares favourably with state of the art machine learning techniques on standard test data from the machine learning community.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Predicting the future from experience

This thesis considers the problem of making predictions about new experiences in the context of the following methodological assumption:

Assumption 1 *The only domain knowledge is a set of past experiences such that each past experience is described by a single ground sentence called a case.*

The thesis starts by describing the reference class approach (RCA) to estimating conditional probabilities from past experiences. Instead of considering the problem in terms of first inducing a domain model from the set of past experiences, and then using some form of deduction to make predictions, the RCA directly infers estimates of conditional probabilities from a knowledge base of past experiences. The resulting estimates are a form of domain knowledge that can be readily used in a number of contexts such as non-monotonic reasoning (e.g., [Bac90]; [Goo91]), prediction and classification (e.g., [Fis87]; [GS88]). The thesis concludes by demonstrating that a computational implementation called FRED[1] can use the RCA's estimates to make accurate predictions about a variety of domains considered bench marks in the machine learning, statistical and pattern matching literatures.

1.2 An overview of the issues

The RCA takes as input: 1. A new experience, such that the well formed sentence (wfs) β describes what is known to be true of the new experience, and the wfs α

[1] For Fred's relational experiential database.

describes what may or may not be true of the new experience, and 2. A set of cases called an episodic knowledge base (EKB) such that each case describes a single past experience. The RCA outputs an estimate of

> the conditional probability, $Prob(\alpha|\beta)$, that α is true of a new experience given that all we know about the new experience is that β is true.

Given the formulae α and β, the RCA estimates $Prob(\alpha|\beta)$ by incorporating β into a suitable *reference class*. The **intension** of the reference class is a pair (α, β) that specifies the **extension** of the reference class, i.e, the set of all past experiences in the EKB that are relevant with respect to estimating $Prob(\alpha|\beta)$. If the extension is empty, then the RCA identifies an alternative reference class, with a non-empty extension. If the extension of the reference class is not empty, then the estimate is obtained by measuring the frequency with which α is known to be true in the reference class extension.

Viewing the induction problem addressed in this thesis as a problem of finding a suitable reference class raises several issues. For example, it may be possible to incorporate β into many reference classes from which different estimates of $Prob(\alpha|\beta)$ can be obtained. In the context of estimating conditional probabilities this ambiguity has been referred to as the *reference class problem* (e.g., [Rei49] [Jr.83] [Lev80] [Pol83] [Pol84] [Bac90] [Goo91]). The RCA approach addresses the ambiguity by specifying the intension of a *single* reference class that is appropriate for making the desired prediction.

If a reference class with an empty extension is specified, then the RCA uses *syntactic generalisation*, and its novel extension *chaining*, to generate sets of possible alternatives. Syntactic generalisation generalises the reference class by generalising properties of the new experience that are known to be true, but which can be assumed irrelevant with respect to estimating $Prob(\alpha|\beta)$. For example,

Example 1 Syntactic generalisation might generalise 'x is rich and a lawyer' to 'x is rich' by assuming that 'x is a lawyer' is irrelevant and dropping the conjunct. However, syntactic generalisation can not generalise 'x is rich and a lawyer' by disjoining additional properties such as 'x is an elephant' to get "x is rich and a lawyer' or 'x is an elephant".

Chaining allows the RCA to assume that knowledge in addition to that specified by β is relevant. Although the addition of knowledge constrains the reference class even further, it has the desirable effect of increasing the number of possible generalisation. For example, if we know that 'x is a bird', and assume that 'x also has feathers', then we can generalise what we know about 'x' by generalising feathers or by generalising bird. Intuitively, the more ways in which the RCA can generalise the more likely it is to find a reasonable alternative to an empty reference class.

The cardinality of the sets of possible generalisations and 'chainings' of an empty extension may be very large. To apply the RCA to real world problems, this thesis considers the use of inductive biases that estimate *irrelevance* in order to identify a single most appropriate generalisation or chaining of an empty reference class. As discussed later in this chapter, this thesis assumes that probabilistic independence (e.g., [Pea88]) is an appropriate estimate of irrelevance.

The remainder of this section considers how the RCA addresses:

The relevant reference class problem: How do we specify the intension of a *reference class* of epistemologically relevant past experiences?

The adequate reference class problem: How do we determine whether or not the reference class extension is adequate with respect to estimating a conditional probability?

The inadequate reference class problem: How do we make estimates when the reference class extension is inadequate?

1.2.1 Relevant reference classes

Estimates of $Prob(\alpha|\beta)$ can be interpreted as reflecting the propensity of α to be true in a domain whenever β is true (e.g., [Bar82]). According to the *frequency* interpretation of probability theory such estimates can be obtained directly from a reference class of a random sample of past experiences (e.g., [Rei49] [Jr.83] [Bac90] [Goo91]) by measuring the frequency with which α is true whenever β is true.

The frequency interpretation is only appropriate with respect to obtaining an estimate of $Prob(\alpha|\beta)$ if two conditions are satisfied:

Condition 1: Given a set of past experiences, we must always know the truth value of α whenever we know that β is true, and

Condition 2: The past experiences that can be described by β must be a random sample of all the domain states that can be described by β.

I argue that we can not assume that either Condition 1 or 2 will be satisfied in the context of Assumption 1.

This section starts by revising the frequency interpretation of probability theory so that it is appropriate with respect to estimating conditional probabilities when Condition 1 fails. The section concludes by arguing that the revision is also appropriate if Condition 2 fails.

A revised frequency interpretation

The frequency interpretation of probability theory defines a conditional probability $Prob(\alpha|\beta)$ as follows:

Definition 1 (frequency interpretation [Bar82]) *The frequency interpretation of a conditional probability* $Prob(\alpha|\beta)$ *is*

$$\lim_{n\to\infty} \frac{\frac{T_{\alpha\wedge\beta}}{n}}{\frac{T_\beta}{n}} = \lim_{n\to\infty} \frac{T_{\alpha\wedge\beta}}{T_\beta} = Prob(\alpha|\beta)$$

where n is the total number of randomly sampled cases available, and T_γ is the number of cases for which γ is true.

Estimates of $Prob(\alpha|\beta)$ are obtained by calculating $T_{\alpha\wedge\beta}/T_\beta$. Intuitively, the estimate is obtained using a reference class whose intension is β and whose extension is the set of all past experiences for which β is true.

Consider a situation in which we wish to estimate $Prob(\alpha|\beta)$ but Condition 1 fails.

Example 2 Suppose we wish to estimate the conditional probability that some one is called Fred if they are tall, i.e., we wish to estimate $Prob(Fred|tall)$. Suppose our past experiences consist of three observations of tall men called Fred, five observations of tall men who are not called Fred, and 100 observations of tall men whose names we do not know.

In the previous example, the frequency interpretation does not provide us with a single number that estimates $Prob(Fred|Tall)$. All the frequency interpretation can tell us is that the estimate falls somewhere in the interval between $\frac{3}{108}$ and $\frac{103}{108}$. The reason the estimate is so imprecise is that the frequency interpretation includes past experiences in its estimate that are irrelevant, i.e., all those past experiences with tall men whose names are not known.

Chapter 3 of this thesis considers several revisions of the frequency interpretation that are appropriate when Condition 1 fails. Chapter 3 concludes that the most appropriate revision is

Definition 2 (revised frequency interpretation) *The frequency interpretation of a conditional probability $Prob(\alpha|\beta)$ is*

$$\lim_{n\to\infty} \frac{\frac{K_{\alpha\wedge\beta}}{n}}{\frac{K_{\alpha\wedge\beta}}{n} + \frac{K_{\neg\alpha\wedge\beta}}{n}} = \lim_{n\to\infty} \frac{K_{\alpha\wedge\beta}}{K_{\alpha\wedge\beta} + K_{\neg\alpha\wedge\beta}} = Prob(\alpha|\beta)$$

such that K_γ is the number of past experiences in the KB that can be described by γ.

Estimates of $Prob(\alpha|\beta)$ are obtained by calculating the frequency with which $\alpha \wedge \beta$ is *known* to be true in a reference class in which either $\alpha \wedge \beta$ or $\neg\alpha \wedge \beta$ is known to be true. Intuitively, the estimate is obtained using a reference class whose intension is the pair (α, β) and whose extension is the set of all past experiences for which α is known be true or known to be false, and β is known to be true.

The revised frequency interpretation of probability theory is appropriate when Condition 1 fails simply because the frequencies used to estimate conditional probabilities are calculated with respect to what we know, not with respect to what is true. For example,

Example 3 Consider the problem of estimating $Prob(Fred|Tall)$ in the previous example. According to the revised interpretation $K_{Fred \wedge Tall} = 3$, and $K_{\neg Fred \wedge Tall} = 5$, so the estimate is $\frac{3}{8}$, i.e., the proportion of tall men whose names are known to be Fred among all tall men whose names are known.

From the perspective of this thesis an agent is unlikely to interact randomly with its domain. As a consequence Condition 2 is unlikely to hold. For example, if an agent collects experiences as it interacts with its domain, then the experiences will reflect the non randomness of the agent's interactions. Accordingly, the agent's experiences will reflect the agent's propensity to observe certain domain properties. I now argue that if Condition 2 fails, but the following assumption is justified

Assumption 2 *The propensity that α can be used to describe a new experience whenever β is known to be true of the new experience, is the same for a new experience as it is among all the past experiences in the reference class extension.*

then the revised frequency interpretation is still appropriate. In the remainder of this section I argue that Assumption 2 is a reasonable assumption.

Consider the following situation in which Condition 2 does not hold:

Example 4 Suppose an autonomous agent, fresh from the factory, is switched on and left, immobile, in the middle of the corridor of the U.B.C. computer science department. The agent sees only two individuals, 'David' and 'Alan'. The agent observes 'Alan' two hundred times, and on each occasion observes that 'Alan' wears glasses. The agent observes 'David' twenty times, and on each occasion observes that 'David' does not wear glasses.

In the situation in the previous example, the agent's past experiences reflect the agent's propensity to observe 'David' and 'Alan'. Clearly the sample is not random. Instead, it is biased by the manner in which the agent experiences the world. Based on its past experience the agent should estimate that the probability that the next person it sees wears glasses is high. The agent's prediction is based upon the fact that, according to Assumption 2, it is ten times more likely to see Alan than David. However, because the agent's past experiences do not necessarily reflect the propensity of individuals in the domain to be old, the estimate will be unreasonable if Assumption 2 no longer holds, i.e., the estimate will be unreasonable if 'Alan' goes on leave.

Calculating frequencies

In order to estimate $Prob(\alpha|\beta)$, we need to be able to to measure an agent's propensity to observe specific domain properties. This requirement has an impact on the way in which past experiences are described in the EKB. Consider again the last example in the previous section. In order for the agent to calculate the frequency with which it observes individuals wearing glasses, the agent must be able to count different observations of the same individual separately. That is, it must be able to count that there have been two hundred occurrences of 'Alan', and twenty occurrences of 'David', in order to determine that it is ten times more likely that the next person it will see will be 'Alan' rather than 'David'. If the agent had simply represented all of its past experiences by the facts "There is an individual called Alan who wears glasses", and "There is an individual called David who does not wear glasses", then the agent would not be able to make this calculation. This issue is addressed in detail in Chapter 3.

1.2.2 Adequate reference classes

The previous section discussed the problem of specifying the reference class of cases in the EKB that an agent should use to estimate a conditional probability. This section discusses the problem of determining whether or not the reference class extension of available cases is *adequate* with respect to estimating the conditional probability.

Typically, the adequacy of a reference class is judged in terms of its *statistical adequacy* (e.g., Kyburg [Jr.88a]; [Jr.88b]). Intuitively, a reference class is statistically adequate if its extension contains sufficient items to make reliable estimates, i.e., estimates that are reasonable and not subject to change. While the mechanics of judging the *statistical* adequacy of a reference class are well understood, the problem of selecting an adequate, but epistemologically relevant reference class of cases remains problematic, and with the exception of Kyburg's work, largely ignored in the artificial intelligence literature. The reason the issue is problematic is that statistical adequacy and epistemological relevance are often incompatible. For example, the cardinality of a set of cases that is judged epistemologically relevant with respect to obtaining an estimate may be too small to statistically guarantee a reliable prediction. One way of side stepping the issue is to adopt the hypothesis that if epistemologically relevant statistics are available, then they should be assumed to be statistically adequate (e.g., [Bac90]; [Goo91]).

Although statistical adequacy is an important measure of adequacy, a more natural measure in the context of this thesis is whether or not the reference class can be used to obtain estimates that result in reasonable predictions about a domain, i.e., predictions that are correct. In this thesis, I argue that a reference class is adequate if it yields estimates of conditional probabilities that result in correctly predicting that α is true if β is known to be true. This thesis assumes that

Assumption 3 *Any non-empty reference class extension is adequate, and any empty reference class extension is inadequate, with respect to estimating Prob*$(\alpha|\beta)$.

This thesis justifies Assumption 3 in terms of:

1. Psychological evidence presented in the Chapter 2 that humans can use small reference classes to make reasonable predictions,

2. Experimental results presented in Chapter 5 demonstrating that adopting Assumption 3 allows us to make reasonable predictions about a variety of domains.

1.2.3 Inadequate reference classes

In the context of the revised frequency interpretation of probability theory and Assumption 3, an empty reference class is obviously inadequate with respect to estimating a conditional probability *Prob*$(\alpha|\beta)$ because

$$\frac{K_{\alpha\wedge\beta}}{K_{\alpha\wedge\beta} + K_{\neg\alpha\wedge\beta}}$$

is undefined as $K_{\alpha\wedge\beta} + K_{\neg\alpha\wedge\beta} = 0$.

If the reference class for estimating *Prob*$(\alpha|\beta)$ is inadequate, then the RCA must identify an adequate alternative. Unfortunately, the number of *possible* alternatives may be large. In order to reduce the number of alternatives that need to be considered, and to avoid considering unreasonable alternatives, this thesis assumes that

Assumption 4 *A reasonable estimate of Prob*$(\alpha|\beta)$ [2] *can be obtained by generalising any properties of the new experience that are epistemologically irrelevant with respect to estimating the probability of* α. *Moreover, probabilistic independence is a reasonable measure of epistemological relevance and can be estimated by applying statistics to the available past experiences.*

[2] The reasonableness of an estimate depends upon the context in which it is used. For example, a reasonable estimate might be an estimate that can be used to make reasonable predictions.

This thesis demonstrates that if Assumption 4 holds, then the RCA can identify reasonable alternatives to an inadequate reference class by generalising over anything that is known about a new experience that is irrelevant with respect to estimating $Prob(\alpha|\beta)$. Chapter 5 demonstrates that estimates obtained in this manner can be used to make reasonable predictions about a variety of domains. The remainder of this section considers a particular form of generalisation called *syntactic generalisation* and its novel extension *chaining.*

Syntactic generalisation

Intuitively, a reference class is a generalisation of another reference class if the former extension is a superset of the latter extension (e.g., [Rei49]), or if the former intension is logically implied by the latter intension (e.g., [Bac90]; [Goo91]) [3]. The problem with this intuitive notion of generalisation is that there may be a large number of generalisations of an inadequate reference class from which inconsistent estimates of the desired conditional probability can be obtained. For example,

Example 5 Suppose the reference class extension of the probability 'What is the probability of a lawyer named Fred being rich?' is empty. We might generalise the reference class to include the financial status of dead republicans and dwarf elephants because, the fact that Fred is a lawyer implies that Fred is either a 'lawyer', 'a dead republican', or 'a dwarf elephant'. However, it is unlikely that the probability of rich dead republicans or rich dwarf elephants bears any relevance to estimating the probability that Fred the lawyer is rich.

The problem occurs because notions of generalisation based upon set inclusion and logical implication are under-constrained. That is, any non empty reference class contains the empty reference class and must therefore be considered as a possible

[3] I refer to this sort of generalisation as being 'semantic' because it can be defined purely in terms of subset containment and logical implication and not in terms of the syntax of the wfss α and β in $Prob(\alpha|\beta)$.

alternative. In Chapter 4 I argue that we should constrain generalisation by only generalising what we know. I call this constrained form of generalisation syntactic generalisation because it depends upon the syntax of the wfss β used to describe a new experience.

Example 6 The reference class in the previous example can be generalised by considering all lawyers, regardless of their names, or by considering all Freds regardless of their profession. It can not be generalised by disjoining additional domain properties to 'is a lawyer named Fred'.

However, even if we only generalise what we know, we might still obtain different estimates of $Prob(\alpha|\beta)$ depending on how much of what we know is generalised. Following Reichenbach [Rei49] and Bacchus [Bac90], I argue that we should generalise as little of what we know as possible. I call an adequate alternative obtained by generalising as little as possible a most specific syntactic generalisation. Intuitively, the less we generalise to find an alternative, the more likely the alternative is to lead to a reasonable prediction [Bac90].

Chapter 4 demonstrates that there may be several most specific syntactic generalisations of an empty reference class, each resulting in a different estimate of the desired probability. In Chapter 4 I argue that inductive *biases* (e.g., [Lai88]; [Des92]; [Sch91]) should be used to identify a set of most reasonable most specific generalisation of an empty reference class. In Chapter 4, inductive biases make use of a notion of *probabilistic independence* (e.g., [Pea88]) to make assumptions about the relative *relevance* of different syntactic generalisations. The inductive biases can be used to select a single most relevant syntactic generalisation of an empty reference class.

Chaining reference classes

The number of different ways in which we can generalise what we know is limited by how much we know to start with. That is, the number of alternatives to an inadequate reference class that can be obtained by syntactic generalisation is a function of how much is known about the new experience of interest. Intuitively, given an inadequate reference class, the more we know the greater the number of possible alternatives, and the higher the likelihood of finding an alternative from which a reasonable estimate can be obtained.

The difficulty with finding a relevant alternative to a reference class by generalising what we know is that we might not know very much. As a result, we may be unable to find a reasonable syntactic generalisation of an inadequate reference class. Chapter 4 describes an extension of syntactic generalisation called *chaining* that allows the consideration of knowledge in addition to what is known about the situation of interest. For example,

Example 7 Suppose we wish to estimate the 'probability that an emu has feathers'. If the reference class is empty and we generalise on what we know we might approximate the desired probability using an estimate of 'the probability of anything having feathers'. However, if we know that emus are also birds, then we can take this information into account and estimate the 'probability that an emu has feathers' by the 'probability that an emu bird has feathers. If we now generalise on what we know we can approximate the probability using an estimate of 'the probability that a bird has feathers' which seems more likely to satisfy Assumption 4.

1.3 Relationship to existing work

The problem of making predictions from past experiences is a subject of research in a number of different AI paradigms. For example, the inductive problem of making predictions about a domain state from a set of cases is one of the primary paradigms of research in machine learning, statistics, case based reasoning, and neural nets. The deductive problem of making predictions about a domain state from a logical representation of experiential knowledge is one of the primary paradigms of research in non-monotonic reasoning.

As we shall see in the next chapter the the RCA extends techniques used by inductive and deductive approaches to making predictions from past experiences. The RCA is most closely related to non-parametric statistical techniques such as kernel estimation and k-nearest neighbours (e.g., [Eub88]; [Han82]; [Han81]), instance based machine learning (e.g., [AKA91]; [Sal90]; [Sal91]; [CS93]), memory based reasoning (e.g., [Dav90]; [SW86]), and case based reasoning (e.g., [SN91]; [Agh90]; [Kot89]) that use local averaging techniques to predict α when β is all that is known to be true about a new experience.

In contrast to Assumption 4, many existing local averaging techniques (e.g., [AKA91]; [Sal90]; [Sal91]; [CS93]; [Dav90]; [SW86]; [SN91]; [Agh90]; [Kot89]; [Eub88]; [Han82]; [Han81]) implicitly assume that a strong correlation exists between experiences whose known properties are *similar* [Eub88]. That is, if a property α is true whenever β is known to be true, then α will be known when wfss *similar* to β are known to be true. The specification of an appropriate similarity metric has been shown critical with respect to obtaining reasonable estimates of $Prob(\alpha|\beta)$ (e.g., [Han81], [Han82], [Eub88]). Unfortunately, the specification is often problematic, particularly when experiences are described in terms of categorical features, i.e., features with a finite number of unordered values [Han81] [Han82] [Eub88] [CS93].

This thesis demonstrates that by making Assumption 4, the RCA can obtain estimates that can be used to make predictions that are as reasonable as, or better than, many existing techniques that use similarity to obtain an estimate of $Prob(\alpha|\beta)$. In Chapter 5 the computational implementation FRED uses estimates obtained by the RCA to make predictions about a variety of domains. In general, Chapter 5 demonstrates that the RCA's estimates allow FRED to make predictions that are as reasonable as other inductive techniques.

1.4 Discussion and Contributions

The RCA described in this thesis is a novel framework for addressing the reference class problem in the context of making predictions about a domain from past experiences. This thesis makes the following contributions:

1. It provides a solution to the reference class problem in the context of Assumption 1, i.e., when the only domain knowledge is a set of past experiences.

2. It describes a new type of generalisation and its novel extension called chaining for identifying an adequate reference class.

3. It demonstrates that Assumption 4 is a reasonable assumption in the context of estimating probabilities directly from past experiences.

4. It demonstrates that probabilistic knowledge obtained by the RCA can be readily obtained from an EKB of cases and used to make reasonable predictions.

Although the reference class problem has been addressed in the context of nonmonotonic reasoning, this thesis shows that existing solutions fail to work when past experiences are the only source of domain knowledge. While the machine learning community has addressed the problem of making predictions from experiences it has not explicitly addressed the problem of identifying the reference class of past experiences relevant to making a particular prediction in the context of Assumption 4.

1.4.1 Outline of the thesis document

An outline of the chapters in the thesis follows:

1. Chapter 2 reviews the psychological, deductive, and inductive literature relevant to the RCA.

2. Chapter 3 describes a propositional language appropriate for describing experiences. Using the language, the revised interpretation of conditional probabilities is described in detail.

3. Chapter 4 describes the use of inductive bias as a technique for identifying an alternative to an inadequate reference class.

4. Chapter 5 describes three experiments that demonstrate that the RCA's estimates can be used to make reasonable predictions.

5. Chapter 6 discusses the strengths and weaknesses of the RCA in the context of Assumptions 1 through 4, and discusses implications for future work.

Chapter 2

Review

2.1 Introduction

This review considers existing techniques in the psychological, AI, and statistical literatures that address the problem of making predictions from past experiences. From the numerous research papers in the area I have selected a sample in order to highlight the issues of identifying a relevant, adequate reference class of past experiences. The intention of this chapter is to motivate techniques for addressing the issues in the context of the RCA, not to document or classify the extensive research in this area.

Figure 2.1 provides an overview of the techniques discussed in this Chapter. The Figure distinguishes between two general approaches to the problem of making predictions from past experiences:

1. Deductive Techniques, i.e.,

    - Default interpretations of direct inference (e.g., [Rei49]; [Bac90]; [Goo91]; [Jr.83]; [Pol84]; [Lev80]; [Pol83]).

2. Inductive techniques, i.e.,

    - Classification algorithms (e.g., [AKA91]; [GLF89]; [Fis87]; [Leb86]; [Qui86]; [FS84]; [Mic80]),
    - Case based reasoning algorithms (e.g., [SN91]; [Sla91]; [Agh90]; [SA77]).

As seen in Figure 2.1 the deductive techniques deduce predictions from some intermediate representation of past experiences. In contrast, the inductive techniques are

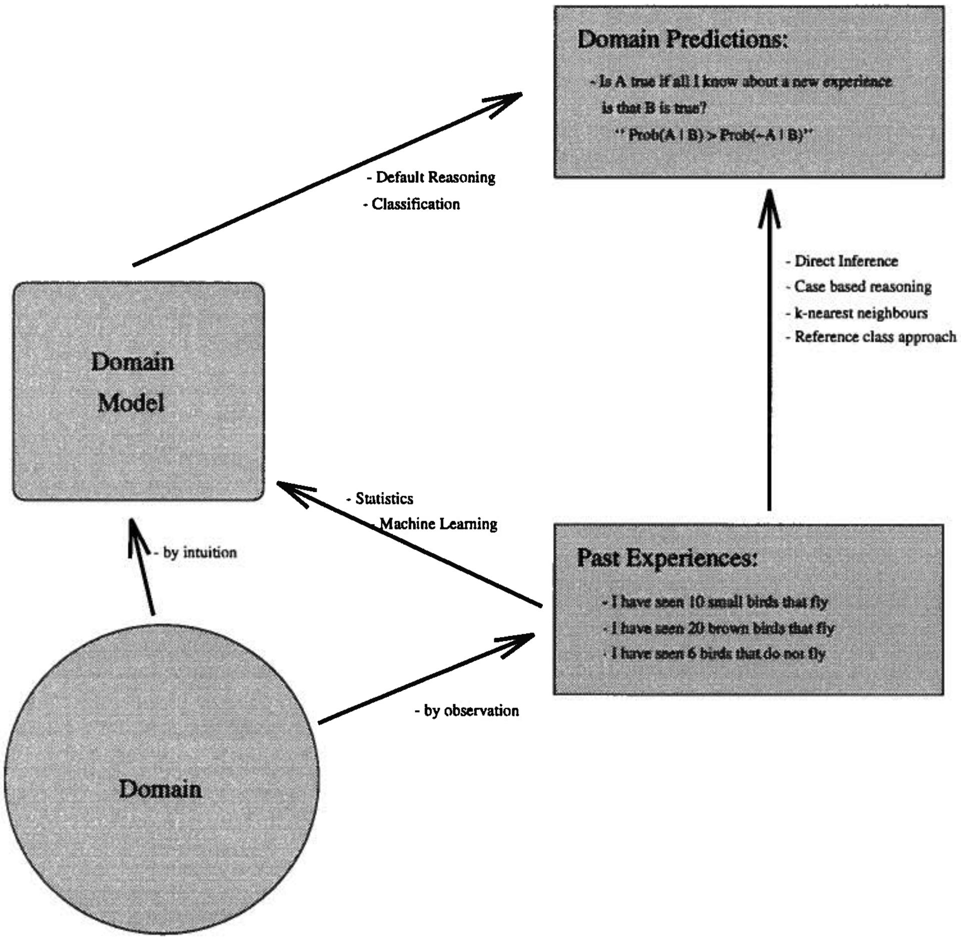
Domain Predictions:
- Is A true if all I know about a new experience is that B is true?
" Prob(A | B) > Prob(~A | B)"
- Default Reasoning
- Classification
- Direct Inference
- Case based reasoning
- k-nearest neighbours
- Reference class approach
Domain Model
- Statistics
- Machine Learning
- by intuition
Past Experiences:
- I have seen 10 small birds that fly
- I have seen 20 brown birds that fly
- I have seen 6 birds that do not fly
- by observation
Domain

Figure 2.1: Existing deductive and inductive approaches to the problem of deriving predictions from past experiences.

concerned with the problem of either: 1. First using induction to derive the intermediate representation, and then using some form of deduction to make predictions, or 2. Making predictions directly from past experiences. This review argues that deductive techniques for deducing predictions can be be extended to the problem of making predictions directly from past experiences, thus drawing a useful connection between inductive and deductive techniques. The review concludes by drawing support for some of the assumptions made in Chapter 1 from the psychological literature on episodic models of human memory (e.g., [FT78]; [Tul72]; [Tul76]; [Tul83]; [Tul83]; [Tul83]; [Tul85]; [TT73]).

2.2 Non-monotonic reasoning and direct inference

Non-monotonic reasoning is a deductive technique for making useful predictions from sparse domain descriptions. For example, the Yale shooting problem, the Nixon diamond problem and other canonical default reasoning problems all involve the use of small numbers of axioms to describe a domain. In contrast, direct inference is a local averaging paradigm that allows us to estimate conditional probabilities from statistical knowledge, knowledge of the form "The frequency with which α is true when β is true is x".

This section examines recent attempts to integrate direct inference (e.g., [Bac90]; [Goo91]) with a consistency based form of non-monotonic reasoning (e.g., [Rei80]) in which predictions are made if they are not contradicted by what is already known or assumed be known. In this context, the remainder of this section discusses the problem of identifying relevant, adequate, reference classes. The section concludes by arguing that techniques appropriate to the solution of the problem in the context of non-monotonic reasoning can be extended to address the same problem in the context of Assumption 1.

2.2.1 Default theories

Consistency based forms of non-monotonic reasoning assume the existence of a domain model called a default theory. A default theory contains no explicit knowledge about past experiences. Nor does it say how the knowledge in a default theory theory is derived from a set of past experiences. Instead, a default theory contains a series of statements that provide a static description of what the domain will be like in the future. This section describes a default theory containing statistical assertions.

Informally, the default theory considered in this section is a pair (D, W), where W is a set of closed well formed formulae (wffs) in a first order logic and D is a set of default assertions [1]. In this section the set D is assumed to consist of statistical assertions written in Bacchus' logic LP [Bac90] as

$$[\alpha(\vec{X})|\beta(\vec{X})]_{\vec{X}}$$

such that $\vec{X}$ is the set of vectors of domain objects that satisfy $\alpha(\vec{X})$ given that they satisfy $\beta(\vec{X})$. Each statistical assertion denotes the frequency with which a proposition α is true given that a proposition β is true. Thus, the statistical assertions can be interpreted, using the frequency interpretation of probability theory, as estimates of conditional probabilities. However, in general the statistical assertions are taken by [Bac90] to be "general scientific knowledge relating properties" as suggested by [Jr.88a].

2.2.2 Making a prediction

The maximally consistent sets that can follow from a default theory (D, W) are called extensions. Intuitively, the extensions of (D, W) can be thought of as filling in the gaps of what we do not know. For example, applying direct inference, we

[1] The reader interested in more detail is strongly advised to read Bacchus [Bac90] and Reiter [Rei80].

might estimate that $Prob(\alpha|\beta) = p$ if the statistical assertion

$$[\alpha(\vec{X})|\beta(\vec{X})]_{\vec{X}} = p$$

is true in at least one of (D, W)'s extensions. The remainder of this section discusses two potential problems with estimating $Prob(\alpha|\beta)$ in this manner:

1. The relevant reference class problem: (D, W) may have several extensions, allowing the derivation of conflicting estimates.

2. The inadequate reference class problem: (D, W) may not contain the statistical assertions necessary for estimating every possible conditional probability $Prob(\alpha|\beta)$.

The remainder of this section discusses existing techniques that address these two problems.

A relevant adequate reference class

In the direct inference paradigm the relevant reference class problem is a problem of choosing an adequate relevant reference class for estimating a conditional probability [Jr.83]. For example,

> If we are asked to find the probability holding for an individual future event, we must first incorporate the case in a suitable reference class. An individual thing or event may be incorporated in many reference classes from which different probabilities will result. This ambiguity has been called the problem of the reference class [Rei49, pg. 375].

Although a number of existing techniques have addressed the problem of choosing a most appropriate reference class (e.g., [Jr.74] [Lev80] [Jr.83] [Pol83] [Pol84] [Bac90] [Goo91]), this section only considers those techniques that address the problem in the context of direct inference and consistency based forms of non-monotonic reasoning.

When estimating the probability that a property is true of an individual from a default theory (D, W), a reasoner might start with the assumption that all the wffs W, and all the statistical knowledge D, is relevant. The difficulty with this assumption is that D may contain a large amount of statistical information that is not applicable to the situation of interest. For example, if we are interested in estimating the probability of leopards having spots we do not want to have to consider irrelevant statistical knowledge about the frequency of spotty children in our neighbourhood.

Given a situation of interest, Reichenbach [Rei49, pg. 203] suggests that the smallest reference class of related statistical assertions is the most appropriate. However, Reichenbach defines the smallest reference class to be the one whose members are 'included' in all other related, adequate reference classes. Unfortunately, defining set inclusion over empty sets is problematic [Jr.88a]. Bacchus' 34th lemma [Bac90] offers an alternative to set inclusion that allows the reasoner to condition upon the entire set of statistical knowledge to obtain the statistical knowledge that is "related" to the current situation of interest. For example,

Definition 3 (Bacchus' [Bac90] direct inference principle) *If $B(a)$ is true, and $[F(x) \mid B(x)]_x = p$ and that is all we know about a, then the probability associated with $F(a)$ is p. If we also know $C(a)$ and that $[F(x) \mid C(x)]_x = q$, and $\forall(x)\ B(x) \rightarrow C(x)$, then the probability of $F(a)$ is to be p rather than q as $[F(x) \mid B(x)]_x$ is more specific than $[F(x) \mid C(x)]_x$.*

Bacchus' interpretation of direct inference assumes that the properties of objects are determined by the properties of *similar* objects. Bacchus' interpretation of direct inference is non-monotonic and the notion of a relevant adequate reference class is determined solely by what statistics are not known. For example,

> Sanctioning the use of a wider reference class over a narrower one when there are no adequate statistics available for the narrower class is equivalent to non-monotonically assuming that the statistics over the narrower class do not differ from the statistics over the wider class. [Bac90, pg. 143]

Bacchus' direct inference principle defines the smallest adequate reference class appropriate to making predictions about an object. That is, if we wish to make predictions about large red birds and we only have statistics about red birds and birds, then we should use the statistics about red birds because they are more specific. Intuitively, the smallest or most specific reference class is preferred because considering a larger one "throws out information" [Rei49] [Bac90].

Bacchus' direct inference principle can be used to choose amongst statistical assertions to find a most reasonable alternative to an inadequate reference class. Indeed, the principle mirrors the use of specificity to impose preference orderings on conflicting defaults in default logics (e.g., [Eth87]; [AM91]; [Bou92]; [Poo91]). For example,

Example 8 Suppose we wish to estimate

$$Prob(Studies\ AI|Fred \wedge Graduate \wedge Large)$$

and the default theory only contains the statistical assertions

$$[Studies\ AI(X)|Graduate(X) \wedge Procrastinates(X)]_X$$

and

$$[Studies\ AI(X)|Graduate(X)]_X$$

According to Bacchus' direct inference principle the statistical assertion

$$[Studies\ AI(X)|Graduate(X) \wedge Procrastinates(X)]_X$$

is epistemologically relevant to making predictions about Fred, but the statistical assertion

$$[Studies\ AI(X)|Graduate(X)]_X$$

is not as $Graduate(X) \wedge Procrastinates(X) \rightarrow Graduate(X)$.

An importance difference between Bacchus' principle and default logic specificity orderings is that the preference orderings over statistical assertions are an automatic consequence of the semantics of LP [Bac90], but are not possible propositionally in a default logic [Poo91].

It is natural to consider extending Bacchus' direct inference principle to deal with the problem of choosing among alternatives to an inadequate reference class in the context of Assumption 1. For example, we might choose the adequate alternative whose intension is logically implied by the intensions of all other adequate alternatives. Unfortunately, as discussed in Chapter 1, and as demonstrated in Chapter 4, using logical implication to impose a preference ordering over alternative reference classes of past experiences has undesirable consequences. In the next section I consider an extension of Bacchus' work that is more appropriate in the context of this thesis.

Assumptions of irrelevance

One property of Bacchus' direct inference principle is that the most specific reference class may not have any statistics. For example,

Example 9 Suppose we wish to estimate $Prob(flies|bird)$ and (D, W) has a single extension containing

$$[flies|large \wedge bird] = p$$

Using Bacchus' direct inference principle we can not estimate that $Prob(flies|bird)$ equals p because the statistical assertion in the default extension is too specific to apply.

The problem of having statistics that are too specific does not occur in the context of the RCA. That is, if we have statistics about large birds and small birds, then we must have statistics about birds because large birds are birds. However, the problem of having too specific statistics is of interest because it is the reciprocal of the problem of having too general statistics in the RCA. That is, if we have statistics about birds, then we may not have statistics about small or large birds. The problem of too general statistics occurs when we have partial knowledge about a new experience, i.e., we can know that x is a bird without knowing its size. I now demonstrate that a straightforward extension of Bacchus' solution to the too specific statistics problem is applicable to the too general statistics problem.

I start by considering Bacchus' solution to the too specific statistics problem. Bacchus' solution to the too specific statistics problem follows Kyburg's [Jr.69] solution and argues that in addition to a default theory (D, W)

> We need knowledge of relevant measure statements; we may ignore special characteristics of the object or event under consideration which are not known to be related to the property in question [Jr.69, pg. 185].

Bacchus includes knowledge of relevant measure statements by adding non-monotonic expectation independence assumptions to (D, W). These assumptions mirror similar assumptions found in conditional logics (e.g., [Bou91]), and default logics (e.g., [Del88]; [Sub90]).

Definition 4 (Bacchus' [Bac90] Expectation Indep. Assmp.) *The assertion*

$$E([Q(V)|P(V) \wedge R(V)]_V) = E([Q(V)|P(V)]_V)$$

is interpreted as 'knowing $R(V)$ is irrelevant to predicting $Q(V)$ when $P(V)$ is all that is known'.

Expectation independence assumptions logically minimise the default theory (D, W) by removing all facts and distinctions that are logically irrelevant with respect to making a *particular* prediction. Intuitively, the assumptions allow us to *specialise* from an inadequate to an adequate reference class. For example,

Example 10 Suppose we wish to predict the probability of Fred flying given that all we know is $bird(Fred)$. Suppose the default theory has a single extension containing the statistical assertion

$$[flies(x)|bird(x) \wedge yellow(x)]_x = p$$

If the expectation independence assumption

$$E([flies(x)|bird(x) \wedge yellow(x)]_x) = E([flies(x)|bird(x)]_x)$$

is true, then we can estimate that $Prob(flies(Fred)|bird(Fred))$ is p.

Because expectation independence assumptions are non-monotonic they introduce the possibility of deriving conflicting estimates of conditional probabilities. Bacchus addresses this problem by imposing a partial ordering on $(D, W)'s$ extensions that captures the direct inference principle's preference for inheriting statistical information from the most specific reference classes.

Bacchus' expectation independence assumptions only allow us to specialise an inadequate reference class. They do not allow us to generalise an inadequate reference class which is the situation of interest in this thesis. For example,

Example 11 Suppose the default theory (D, W) contains the assertion

$$bird(Tweety) \wedge yellow(Tweety) \wedge [fly(x)|bird(x)]_x = .75$$

and the expectation independence assumption

$$E([fly(x)|bird(x) \wedge yellow(x)]_x) = E([fly(x)|bird(x)]_x)$$

is true. With respect to the statistical assertion and the independence assumption, there is no viable theory in LP that allows us to estimate the probability of Tweety flying.

If we accept Bacchus' argument that we can specialise to an adequate reference class by ignoring irrelevant properties, then we should be able to argue that we can generalise to an adequate reference class in exactly the same way. Indeed, Goodwin [Goo91] provides an extension of Bacchus' logic LP that allows us to generalise by excluding irrelevant properties. Informally, Goodwin interprets the assertion

$$E([Q(V)|P(V) \wedge R(V)]_V) \;=\; E([Q(V)|P(V)]_V)$$

as both 'knowing $R(V)$ is irrelevant to predicting $Q(V)$ when $P(V)$ is all that is known', and as 'knowing $R(V)$ is irrelevant to predicting $Q(V)$ when $P(V) \wedge R(V)$ is all that is known'. Thus, in the context of the previous example we can estimate that the probability of Tweety flying is 0.75 because knowing that Tweety is yellow is irrelevant with respect to estimating the probability of Tweety flying.

2.2.3 Discussion

The deductive techniques reviewed in this section make various epistemological assumptions about the knowledge in a default theory in order to estimate conditional probabilities when the domain knowledge contained in the default theory is incomplete. The expectation independence assumptions discussed in this section are an attractive partial solution to the problem of estimating probabilities from past experiences. Goodwin's [Goo91] extension of Bacchus' [Bac90] non-monotonic expectation independence assumptions closely mirror Assumption 4 in that by assuming that something is irrelevant with respect to obtaining a particular estimate we can generalise the reference class.

In principle at least, the estimates of conditional probabilities obtained from a default theory by direct inference can be obtained directly from a reference class of past experiences. Unfortunately, the techniques reviewed in this section are not appropriate in the context of Assumptions 1 through 4 because:

1. The techniques assume the existence of a default theory and fail to address

the issue of obtaining the theory from a set of past experiences.

2. The expectation independence assumptions force the designer of the default theory to anticipate its every use in advance.

In order to address the first problem we would have to consider the problem of deriving a default theory from past experiences. This may or may not be an appropriate solution. However, even if were to derive a default theory from past experiences we would still be left with the second problem. In order to ensure that any estimate can be derived from a default theory (D, W), the designer of (D, W) would have to anticipate its every use. That is, the designer would have to know, in advance, which expectation independence assumptions were necessary. Unfortunately, there is no capacity within the techniques discussed in this section to automatically generate the assumptions as required.

In general, the techniques reviewed in this section rely too heavily on the intuition of the designer of the default theory and not enough on past experiences to address the issues discussed in Chapter 1. To address the problem in this thesis, the designer would have to be omniscient in order to cover every eventuality. This raises various epistemological concerns such as where the knowledge underlying the expectation independence assumptions comes from and how the techniques can obtain estimates about situations that have not been 'anticipated'.

2.3 Inductive classification algorithms

Inductive classification algorithms are the dominant paradigm in the artificial intelligence literature for making predictions about the properties of an object from a set of cases (e.g., machine learning [Qui87b], [Qui87a], [BFOS84], [CMM83] [BP89], [CN88], [GS88]; statistics [AKA91], [Das91], [Aha89] [HV74]); connectionism [MR81], [RJ86], [RHW86], [PG90], [RR89], [Koh90]). Classification algorithms take as input a set of cases and a set of hypotheses. Often, a case is a feature vector, i.e.,

$\langle f_1, \ldots, f_n \rangle$, such that each feature is a function mapping a single object to a single value. The values can be continuous or discrete. The set of all possible hypotheses, H, is a space of n-ary functions defined over the n features used to describe the objects in the cases. Intuitively, the n-ary functions in a hypothesis space H represents the set of all possible reference class intensions that can be considered in order to make a prediction.

The n features define an n-dimension feature space and each case a point in the space. If the classes are pre-specified by including class information with each case, then the algorithm is said to be *supervised*, otherwise the algorithm is *un-supervised.* If the classification algorithm is able to update the hypothesis each time a new case is presented without re-processing the entire set of cases received as input, then the algorithm is said to be *incremental* [GLF89]. The set of hypotheses form a space of n-ary functions defined over the n input features.

Classification algorithms select and output a *single* hypothesis $h \in H$ that is consistent with all the cases provided as input (e.g., [CMM83] [BP89]). The selected hypothesis serves as a: 1. Finite representation of the cases provided as input, and 2. Domain representation consisting of a finite set of adequate reference classes that have been selected on the basis of the cases used as input. The selected hypothesis divides the input cases into a set of classes. If the n features used to describe the domain object are thought of as the n dimensions of a feature space, then each class represents a different sub-region of the feature space. The shape and size of the regions is a function of the particular classification algorithm used and the agent's past experiences. For example, classes in EACH [Sal90] are represented as hyper-rectangular regions while decision tree classes [Qui83] are represented as hyper-cubic regions [FSK+93] in a feature space.

Given a new experience, classification algorithms use the selected hypothesis to perform one of two tasks: A *classification* task, or a *prediction* task. Classification tasks are usually statements of the form "Based upon its description, is object x

a member of class y?", e.g., "Is the winged, feathered, egg-laying object a member of the class of birds?" Prediction tasks are usually statements of the form "Is v a value of the feature f of the object x?". Prediction tasks are typically re-cast as classification tasks. For example, the prediction task 'Does the feathered, winged object fly?" can be treated as the classification task "Is the feathered, winged object a member of a class for which flying is true?"

2.3.1 The hypothesis space

Unlike the models of default reasoning discussed previously, a central motivation of classification algorithms is to obtain a representation of a set of cases that serves as a domain model. The particular representation chosen depends upon a number of factors. One of the most important is the set of hypotheses considered as possibilities by the classification algorithm. From the numerous research papers describing classification algorithms, I have selected a sample in order to highlight the techniques and problems associated with the inductive classification approach to the problem considered in this thesis. Once again, my intention is to provide a point of comparison with the RCA, not to document or classify the research in this area.

The selected classification algorithms differ considerably in the complexity of the hypotheses that are considered. Decision tree algorithms create decision trees. Concept learning algorithms create decision rules for classes. Nearest neighbours and kernel estimation algorithms, in the simplest instance, create a domain model consisting of a set of cases that are divided into subclasses in response to a particular classification or prediction task (e.g., [Han81]; [Han82]). Connectionist algorithms differ somewhat from other machine learning and statistical techniques in that they start with a network with a particular topology and change the weights on the connections.

Concept learning algorithms

The majority of machine learning research has focussed on the broad area of algorithms that learn concepts by clustering cases [Sal90] [GLF89]. The concept learning algorithms divide cases into clusters based upon the specified object properties. The general goal is to identify a set of concepts such that the specifications of the objects in the concepts will have a much greater intra-concept similarity than inter-concept similarity. A prediction is made about an object's properties by classifying the object into one of the existing concepts on the basis of its specification.

The decision tree algorithms (e.g., ID3 [Qui83]; C4.5 [Qui87b], [Qui87a]; CART [BFOS84]) are supervised, non-incremental, clustering algorithms that split a set of cases into subsets or classes according to a sequence of tests conducted on the values of their individual features. The algorithms are *divisive* [Sal90] in that they start by treating the entire data set as one big cluster that is gradually split into many small clusters each representing a single concept.

To choose a test, decision tree algorithms examine the information theoretic gain of the potential splits using functions such as entropy (e.g. C4.5) or the gini-function (e.g. CART). In the simplest case the test is based upon a single feature value. For example, a test might divide a set of cases into those describing objects with the colour red and those describing objects with some colour other than red. Generally, the root of the tree consists of the test with greatest information theoretic gain. Each leaf represents a single concept. The tests divide the feature space into hypercubic regions [FSK$^+$93] such that objects with geometrically close descriptions are allocated to the same concepts. A new case is classified into one of the leaves of the decision tree by performing the tests on its feature values as specified by the nodes of the decision tree. The properties of the object described by the case are assumed to be the same as those associated with the concept denoted by the leaf.

The induction rule algorithms (e.g., EBG [MD85]; CN2 [CN88]; ITRULE

[GS88]) are supervised, non-incremental, clustering algorithms that tackle the problem of learning concept definitions. Instead of a decision tree they produce a set of conditional decision rules for class membership in a set of pre-specified concepts.[2] For example, for each concept they start with a universal rule such as "If any condition, then object x is in current class". The conditions on the left hand side of the rule are generalised so that all instances of the class satisfy the membership requirements, and specialised so that all non instances of the class are excluded. While the rules are often expressed in a logic such as Prolog, they are often expressed in other forms such as schemata (e.g., EBG [MD85]). A new case is classified into the class whose decision rule it satisfies. As with decision trees each case is assumed to only satisfy the membership requirements of one concept.

The conceptual clustering algorithms (e.g., CLASSIT[GLF89]; COBWEB [Fis87]; UNIMEN [Leb86]; EPAM [FS84]; [Mic80]) are un-supervised, incremental, clustering algorithms that produce a classification scheme over a set of cases. Unlike the induction rule algorithms, the concepts are placed in a concept hierarchy that organises concepts in terms of their generality.

Due to their hierarchical nature, the conceptual clustering algorithms are very similar to decision tree algorithms. The main difference is that the tests performed to determine concept membership are more complicated as they often take into account contextual information about the properties of an object [GLF89]. Furthermore, each concept is associated with a number of necessary and probabilistic properties (e.g., COBWEB [Fis87]). For example, if an object is classified into the concept bird, then it may be possible to predict flies with a certain probability, feathers with another probability and so on.

[2] Decision trees can be shown to be equivalent to ordered lists of rules [FPSM92]

Pattern matching algorithms

Pattern matching algorithms are un-supervised algorithms that represent a domain by the set of cases provided as input. Unlike the clustering algorithms discussed in the previous section, no attempt is made to represent the cases more parsimoniously as a set of concepts. Each of the cases represents a point in a feature space. A prediction is made about an object described by a new case by selecting the nearest neighbours to that case in the feature space. Although some have argued (e.g. [Des92]) that pattern matching algorithms are inappropriate and unlikely to be useful, applications of various pattern matching algorithms have demonstrated otherwise (e.g., [FSK+93]; [Tur92]).

The k nearest neighbour algorithms (e.g. k-nearest neighbours [Das91]; IBL [Aha89]) choose the k-nearest neighbours (k-NN) to a new case using a *similarity metric*, usually based upon some notion of 'distance' in the feature space. Different k-NN algorithms differ on the similarity metric chosen to find the nearest neighbours. k-NN algorithms are particularly useful when applied to features with continuous numeric values [FPSM92]. A simple measure of similarity that is often used when features have discrete values is the inverse of the *Hamming distance d* between two cases [3].

Definition 5 *The Hamming distance between two cases I_1 and I_2 represented by the feature vectors $\langle f_1, \ldots, f_n \rangle$ and $\langle f_1', \ldots, f_n' \rangle$ respectively, is:*

$$d(I_1, I_2) = \frac{\sum_1^n |f_i - f_i'|}{n}$$

Techniques exist for applying k-NN to features with discrete values are discussed in detail in [Cre92], [Han81] and [Han82]. These authors argue that traditional techniques such as Hamming distance are not always appropriate. I return to this issue later in this section.

[3] Strictly speaking, the Hamming distance function requires the feature values to be integers or reals. Symbolic feature values must be normalised first.

The similarity metric divides the feature space into sub-regions in a manner analogous to the concepts defined by concept clustering algorithms. The difference is that the space is re-divided each time in response to the syntax of the case used to describe the new experience. Thus, the set of possible divisions of the feature space is not constrained to an a-priori defined set. Generally, an object x is classified into a class C by *voting*, i.e., If there are more instances of class C among the k nearest neighbours than any other class, then x is also classified as an instance of C [Das91]. The frequency of C among k can be used as an estimate of the conditional probability of C (e.g., [Han81]; [But93]). The parameter "learned" by the algorithms is k. k is learned by applying the k-NN algorithm to the same data set using different values of k [FSK$^+$93]. The value k that results in the highest predictive accuracy is chosen to classify all new cases.

The clustering algorithms (e.g., EACH [Sal90]; IBL2 [AKA91]) are extensions of k-NN algorithms that use clustering techniques similar to those used by concept learning algorithms to cluster individual cases into larger units. The primary difference is that clustering is *agglomerative* rather than divisive. The algorithms assume that to start with each case is a single cluster. Larger clusters are formed by combining smaller clusters together. For example, IBL2 and [Bra87] use an instance averaging technique based upon *median cluster analysis* to replace any two cases by the *average* of their feature values. The technique assumes that all feature values are numeric. The k-NN algorithm is then applied to make a prediction. EACH combines cases into hyper-rectangular shaped regions in feature space called exemplars whose necessary features are shared by every case in that region. A new case is classified into the *single* most similar exemplar and is assumed to share the necessary properties of that exemplar.

The clustering instance based algorithms are similar in many respects to the concept learning algorithms discussed in the previous section. They are motivated by two concerns: 1. Finding a more parsimonious representation of the domain than a set of cases, and 2. Improving predictive accuracy. The latter concern is motivated by the observation that clustering cases sometimes decreases the influence of inconsistent cases (cases that describe identical experiences differently) on predictive accuracy. By averaging the feature values (e.g., IBL2), or identifying the necessary feature values (e.g., EACH) of a set of cases the inconsistencies are factored out. It is interesting to note that a similar effect is often obtained by increasing the size of k in k-NN algorithms [Han81].

Connectionist algorithms

Connectionist algorithms (e.g. back-propagation networks [RHW86]; radial-bias function networks [PG90] [RR89]; Kohonen networks [Koh90]) are supervised algorithms that learn a function mapping an input space (the object features) to an output space (the desired predictions). The function consists of a network of a fixed topology such that the arcs are weighted and the nodes are divided into three sets: 1. A set of input nodes each representing a single feature value and a set of nodes are designated output nodes, 2. A set of output nodes each representing a single class, and 3. A set of hidden nodes. Given a case, a connectionist algorithm turns "on" the input nodes corresponding to features of the object and turns "off" all the other input nodes. The object is classified by observing which output node turns on as a result of turning on the input nodes.

Connectionist algorithms can be distinguished on the basis of the topology of the network and the method by which the weights on the arcs are updated. Back propagation repeatedly adjusts the weights of the connections in a neural network to minimise a measure of the squared differences between the actual output and the desired output of the algorithm. Internal "hidden nodes" that are not part of the input or output are used to represent important features of the domain by capturing regularities in the data. Radial bias functions differ from back propagation by the function that maps the nodes of the input to the nodes of the output. Kohonen networks learn a feature map between an input space and an output space.

2.3.2 An adequate relevant reference class

The predictive accuracy of any classification algorithm is dependent upon the ability of that algorithm to identify the intensions of a finite set of adequate reference classes that capture the domain knowledge that is epistemologically relevant to making the set of desired predictions. As seen in the previous section, different classification algorithms consider different sets of hypotheses and as a result are 'biased' to capturing different kinds of domain knowledge. The existence of biases is substantiated by comparative studies of classification algorithms (e.g., [FSK$^+$93] [FMM$^+$89] [RHW86] [Qui86]) indicating that there is no such thing as a universally appropriate classification algorithm.

One interesting conclusion of the comparative studies is that the best classification algorithm is not always the most sophisticated. For example, comparatively simple k-NN algorithms often outperform complex machine learning techniques (e.g., [FSK$^+$93]). Unfortunately, the conclusions of the comparative studies are often contradictory.

I now discuss two general techniques used by classification algorithms to simplify the problem of finding a set of relevant reference classes by: 1. Imposing constraints on the form of the cases used as input, 2. Biasing the process of selecting a hypothesis using domain specific heuristics. I briefly consider the two techniques and discuss their effect on the ability of the classification algorithms to find an epistemologically relevant domain model.

Constraining the input

All the classification algorithms considered in this section are concerned with the problem of selecting a single hypothesis that is consistent with a set of cases [4]. The problem of selecting a single hypothesis from a large set of possibilities is often simplified by imposing constraints upon the form of the cases accepted as input. Doing so reduces the size of the hypothesis space and simplifies the problem of selecting a single hypothesis just as Assumption 4, Chapter 1, allows the RCA to reduce the number of alternative reference classes that it considers. For example, the cases are often assumed to be noise free:

1. *Complete*, or transformable into a complete form, (e.g., [AKA91] [Qui83] [Qui89] [Tur92] [Des92] [SMT91]),
2. *Consistent*, (e.g., [Qui83] [AKA91] [CMM83] [BP89]), that is, identical experiences are described by identical cases, and/or
3. *Supervised*, (e.g., [AKA91] [Qui83] [Tur92] [Des92] [Sal90]), that is, each case is either specified as an example or counter example of a particular class.

The constraints are imposed upon the cases to simplify the problem of selecting a single hypothesis by reducing the size of the hypothesis space considered by the classification algorithm. For example, supervised cases tell the classification algorithm which predictions it will be asked to make. This constraint significantly reduces the set of possible classes that have to be considered. Further constraints can also be applied to the number of features used to describe each case and the number of concepts to be learned. For example, attempts to learn category rules are often restricted to severely constrained inputs in which the number of classes are small, i.e. only one or two (e.g. [Win75]).

[4] There are of course exceptions. For example, the variant space method [Val84] may consider a set of consistent hypotheses. A consequence of this approach is the familiar multiple extension problem in which the different predictions made from using the various consistent hypotheses must be combined.

The difficulty with reducing the size of the search space is that only simple hypotheses are considered. While existing concept learning algorithms are readily able to learn conjunctive category rules, learning more complex disjunctive category rules has proven difficult. Kearns' [Kea89] and Valiant's [Val84] analyses of the complexity of PAC (probably approximately correct) learning algorithms has shown that PAC learning certain classes of disjunctive concepts is NP-hard or that a prohibitively large number of cases is required for learning to take place.

Being able to consider more complicated hypotheses is important. Bundy, Silver and Plummer [BSP85] showed that certain sequences of cases cause inconsistencies to emerge and result in the failure of concept learning algorithms. This problem is called the disjunctive concept problem as such sequences of cases may indicate the existence of a disjunctive concept [Tho87]. Pattern matching algorithms handle the disjunctive concept problem very easily: A disjunctive concept is defined by the cases in its extension.

A consequence of imposing constraints on the cases that are acceptable as input is that the resulting classification algorithms can only be applied in certain circumstances. For example, Schaffer [Sch91] and Feng et. al. [FSK$^+$93] observe that if the cases available are sparse, the classification algorithms may be unable to find any meaningful regularities in the cases and thus be unable to find a predictive domain model. Pattern matching algorithms provide a solution to this problem in that they do not have to "learn" a domain representation but represent the domain by the cases themselves.

Inductive biases

Even if the hypothesis space is small there may be several hypotheses consistent with a particular set of cases. This problem is an interesting variant on the reference class problem. That is, there may be several hypothesis from which different predictions can be made. Most classification algorithms use *inductive bias* [Lai88] to search a hypothesis space for a *best* hypotheses. The term 'inductive bias' is used to describe the way in which hypothesis are selected for evaluation as possible representations of a set of cases [Lai88]. For example, Mitchell [Mit80] suggests biasing the process of hypothesis selection in favour of certain rules over others. Schaffer [Sch91] discusses the role of bias in decision tree pruning. Utgoff [Utg84] studies the problem of adapting the class of admissible hypothesis to the performance of the learning algorithm. Fisher [Fis87] discusses the problem of defining a metric of cluster goodness for identifying the best clustering of a set of cases. DesJardins [Des92] and Turney [Tur92] study the problem of using background domain knowledge in addition to the cases in order to choose the best hypothesis.

At first glance, pattern matching algorithms appear to avoid the problem of choosing a single hypothesis. After all, a KB consisting of the set of all cases provided as input is trivially consistent with the cases. However, pattern matching algorithms still partition the cases into classes using similarity metrics and the similarity metrics represent a form of inductive bias.

There is nothing inherently good or bad about any particular inductive bias. The value of each technique is conditional upon the domain in which it is employed [Sch91]:

> Suppose we are given a series of unfair coins and asked to guess, on the basis of experiments, whether each favours heads or tails. A basic strategy is to flip each coin a set number of times and then guess whichever face has appeared most often. Consider two variations of this strategy. The first calls for a guess of heads if heads is flipped in at least a third of the trial flips. This is clearly an example of bias and, as such, it has indeterminate effect on the performance of the strategy. Whether the bias is good or bad depends upon the problem distribution. The second strategy simply doubles the original number of trial flips. By contrast, this variation is a statistical improvement to the basic strategy. Regardless

of the mix of coins, it must increase expected performance [Sch91].

The difficulty is that it may be difficult to identify situations to which a particular algorithm is suited if the inductive bias is not obvious. For example, it is very difficult to examine the weights in a neural net to determine whether or not it will perform well or poorly on a particular data set.

Another problem with selecting an appropriate inductive bias is that the bias may change over time or indeed change depending on the prediction being made. For example, supervised algorithms adopt a bias that is particularly suitable to making predictions about a pre-identified set of object properties. If the set of desired predictions changes, then the algorithms must be re-applied to a new set of cases. The ability to shift bias is one of the strengths of incremental algorithms like the conceptual clustering techniques (e.g., COBWEB [Fis87]) and pattern matching algorithms (e.g., k-nearest neighbours [Das91]). However, it is interesting to note that most existing incremental techniques require the original set of cases in order to shift bias [Des92]. Even though some incremental techniques profess to learn a more parsimonious representation of a domain than a set of cases, they must retain the individual cases in order to shift bias.

2.3.3 Discussion

The RCA is an inductive technique by virtue of the fact that it is concerned with the problem of making predictions directly from past experiences. The RCA is particularly similar to un-supervised, incremental, inductive techniques such as pattern matching algorithms whose only representation of the domain is a set of cases such that each case describes a single past experience. This section discusses the relationship between the RCA and pattern matching algorithms in more detail.

The RCA and pattern matching algorithms such as k-NN are instances of non-parametric statistical smoothing techniques. Informally, smoothing techniques estimate conditional probabilities from inadequate reference classes by smoothing to, or interpolating from, adequate reference classes that are similar. The difference between the RCA and existing smoothing techniques lies in the nature of the smoothing. For example, k-NN algorithms smooth by interpolating from k most similar cases such that each case describes a single past experience. In contrast the RCA smoothes by generalising the syntax of a new case until an adequate reference class is obtained. In the context of the RCA, smoothing can be understood in terms of preference orderings that choose among 'possible smoothes', and expectation independence assumptions that determine what kind of smoothing is allowed.

Of particular interest to machine learning and statistics is the RCA's applicability in situations in which categorical variables, i.e., variables with a finite number of unordered values, need to be smoothed. Existing smoothing techniques have proven inappropriate in this context (e.g., [Han82]; [Han81]; [Eub88]). Some of the difficulties may be a consequence of using distance metrics to measure similarity when no concept of distance exists between the values of categorical variables (they are, after all, unordered). For example, techniques that translate categorical variables into a series of binary valued variables over which Hamming distance can be calculated are inappropriate as discussed by Eubank [Eub88]. In contrast, the RCA uses concepts that are ideally suited to smoothing categorical variables. For example, syntactic generalisation and chaining are both extensions of non-monotonic techniques that have been specifically designed to reason about categorical variables.

An obvious test of the RCA, as with all inductive strategies, is its performance on real data.

> ... all [inductive] algorithms must be subject to empirical verification. In particular, an [inductive] algorithm should be compared to other [inductive] algorithms by testing it on the same data set [Sal90]

Chapter 5 compares the performance of a particular implementation called FRED, that uses the RCA's estimates of conditional probabilities to make predictions, with the performance of a variety of existing inductive algorithms including implementations of k-NN. The results empirically verify the RCA in that they show that the RCA's estimates can be used to make reasonable predictions.

2.4 Case based reasoning algorithms

Case based reasoning algorithms (CBR) address the problem of retrieving solutions to past problems from an EKB in order to solve new problems. CBR start with an EKB of cases, each describing a problem and a solution to that problem. Given a case that describes a new problem with no solution, CBR select a *single* case in the EKB whose solution is applicable to the new problem. The retrieved solution is then modified, if required, and applied to the new problem. CBR share obvious similarities with pattern matching algorithms such as EACH that make a prediction by retrieving a single exemplar. It comes as no surprise that many of the issues surrounding the problem of choosing a most appropriate case reflect those found in pattern matching algorithms.

2.4.1 The KB design

The semantics of the cases considered by CBR are considerably more complicated than those considered by inductive classification algorithms. Generally, there are three major parts to each case in the EKB.

Definition 6 *(***Kolodner [Kol91, pg. 60]***) A case consists of:*

1. *The problem/situation description, the state of the world at the time the case was happening and .... what problem needed solving at that time.*

2. *The solution, the stated or derived solution to the problem specified in the problem description. Some case-based reasoners also store traces of how the problem was solved.*

3. *The outcome, the resulting state of the world when the solution was carried out.*

Not only do the cases used by CBR have more 'parts' than those used by classification algorithms [5], but the parts themselves often have a complicated structure. For example, in CASEY [Kot89], the first part of each case consists of a n-ary vector describing a patient with heart failure. The second part consists of a causal network modelling the underlying cause of the heart failure. In [SN91] the first part of each case consists of an influence graph that models the behaviours of a set of components for designing a particular fluid flow model. The second part of each case consists of the model itself.

2.4.2 Identifying the relevant knowledge

CBR take as input a new case consisting of a description of a particular problem/situation. CBR assume that a solution to the problem can be retrieved by:

1. Integrating the new case into the EKB, and retrieving all cases with similar state descriptions.

2. Evaluating the relevance of the retrieved cases to solving the problem described in the new case.

3. Transferring the solution of the best matched case to the new situation, adjusting it according to differences between the two cases [Kot89] [SN91].

For example, CASEY takes as input a description of a new patient with heart failure and returns the most appropriate causal network for modelling the cause of the heart failure. CHEF [Ham84] [Ham86] [Ham89] takes as input a description of a dinner menu and returns a plan of how to prepare the menu. Sycara and Navinchandra's

[5] The cases used by classification algorithms are only domain state descriptions.

[SN91] system takes as input a description of a fluid dynamic system in the form of an influence diagram and returns the most appropriate design for the system.

Given a new case consisting of a problem description, and no solution, CBR assume that the best solution to the problem is found by finding a *single* most appropriate case in an EKB, and returning its associated solution. CBR assume that solutions to a particular problem entail similar solutions for slight variations of the problem.

Making similarity judgements

In CBR similarity judgements can be as simple as measuring the similarity of vectors as in CASEY [Kot89] and pattern matching algorithms, or as complicated as matching influence graphs as in [SN91]. The similarity metric must be sensitive to small, yet crucial differences, in the description of problem states. For example, in CASEY very small differences between the description of the problem state of a new case and the retrieved case can dichotomise the causal explanations of the retrieved case and the new case to the point of incompatibility [Agh90]. Unfortunately, Aghassi [Agh90] reports that when CBR are applied to the domain of heart disease considered in CASEY truly similar cases are rare even in a large EKB (less than 7 per-cent of the new cases are similar to cases in the EKB). Aghassi's [Agh90] findings suggest that a large EKB may be required before the similarity metrics used by CBR will work.

If similar domain states recur in the domain, then CBR assume that as the number of cases increase the best matched case should approximate the new case more and more closely and the retrieved solution should be more and more appropriate. For example, Goodman[Goo86] reports a positive correlation between the accuracy of his system's solutions with the number of cases in the EKB. However, Aghassi [Agh90] reports that CASEY's performance is negatively correlated with the number of cases in the EKB. These incompatible results suggest that the problem of making similarity judgements in the context of CBR is not well understood.

Indexing cases

The choice of appropriate case features on which to base the similarity metric is crucial. If inappropriate features are chosen the right cases will not be retrieved at the right times. The choice of so-called "indexing" features is perhaps the biggest issue in CBR [Kol91]. The literature on CBR suggests that the indexes should be:

1. Predictive, e.g., similar indexes predict similar solutions,
2. Abstract enough to make a case useful in a variety of future situations, and
3. Concrete enough to be recognisable in future cases.

The indexes must allow similar cases representing similar problems with similar solutions to be identified. Unfortunately, choosing the most appropriate indexes is a non-trivial task and apparently domain dependent.

One popular indexing method is to organise the EKB into a generalisation hierarchy (see for example, [Kol88]; [Kot89]). However, in [Kol88] and [Kot89] the number of generalisation nodes required overwhelms the actual number of cases in the EKB as the size of the indexed EKB grows exponentially with the number of cases [Agh90]. Thus, even an O(n) brute force search through a simple O(n) list of cases is better than organising the cases in the generalisation hierarchy. The applicability of different indexing strategies is domain dependent. For example, a generalisation hierarchy only works if there is sufficient regularity in the domain to merit it. If the number of cases in the EKB is small or the domain complex, then finding a satisfactory indexing strategy is difficult.

Imposing an indexing strategy upon the EKB also increases the computational difficultly associated with integrating new cases into the EKB. For example, if the EKB is maintained as an unordered list of cases, then new cases can be integrated in constant time. If the EKB is maintained as a generalisation hierarchy, then each new case must be integrated into an exponential search space. As most CBR require that each new case be integrated with the existing cases in the EKB there is a clear tradeoff between the computational effort required to integrate a new case and the computational effort required to retrieve a single solution. This tradeoff does not appear to have been considered.

2.4.3 Discussion

CBR provide solutions to novel problems if the following assumptions are satisfied [Agh90]:

1. Similar cases recur,
2. Similar cases require similar solutions,
3. Similar solutions recur.

If similar cases do not recur then a search through the EKB for the relevant knowledge will fail. If similar cases do not require similar solutions then although the case retrieved from the EKB may provide a plausible solution to the new problem it may be more probable that there is a completely different solution. Finally, a rarity of "similar solutions precludes the successful transfer [of a solution to a new problem], almost entirely, given anything other than a large pool of cases [in the EKB]" [Agh90].

Aghassi's [Agh90] analysis of CASEY indicates that if any one of the three assumptions is unjustified the performance of CBR will be unacceptable. In particular, Aghassi concludes that "CBR [are] not a good idea if unforeseen circumstances routinely occur." Aghassi's conclusion indicates that unless the assumptions are justified, CBR are not a good approach for providing solutions to novel problems.

Aghassi's [Agh90] analysis of CBR indicates that there would be certain difficulties with applying CBR to the problem of making predictions about the properties of objects from a set of cases. Particularly problematic is the observation that CBR are only appropriate in the absence of unforeseen circumstances. This indicates that if current CBR were applied to the problem of making predictions they would only be good at making predictions about things that the inductive reasoner already knows. However, given sparse data about a complex domain, the inductive reasoner would be expected to be continually faced with unforeseen circumstances.

2.5 Episodic Memory

In Chapter 1 of this thesis:

Assumption 1: states that individual experiences are specified by cases in terms of domain properties.

Assumption 2: states that reasonable predictions can be made by aggregating over the members of a reference class extension.

Assumption 3: states that reasonable predictions can be made from small reference classes.

This section examines empirical support for these three statements from the psychological literature on human episodic memory (memory of specific domain states). First, some preliminary definitions are provided. Second, the effect of remembering the specific properties of experiences on the retrieval of knowledge from human memory are studied. Third, evidence for the hypothesis that memory categories are dynamic and formed in response to requests for specific knowledge is considered.

2.5.1 Preliminaries

Empirical experiments on human memory consist of a memory task, and a memory test:

Memory tasks consist of a domain object, usually a word, pair of words, or a list of words, and a set of instructions. In a typical psychological experiment a subject is given a set of instructions which are then followed by the presentation of one or more objects. The subject is told to apply certain *memory operations* specified in the instructions, such as reading the words, to the domain object. What is stored in the subject's memory as a result is a case referred to as an *episodic trace*: a specification of the memory task and the object. The episodic traces formed during a memory task are called *task-episodes.*

Memory tests consist of an object, commonly called the 'cue', a set of instructions, and, depending on the nature of the task, a second object called the 'target'. The subject's representation of the 'cue', instructions and 'target' form a *test-episode* which is matched against *task-episodes* already stored in memory. The effectiveness of the *test-episode* is measured by the amount of time that it takes the subject to make the desired response and/or the accuracy of the response.

It may be assumed (e.g., [HBP89]) that each task and test episode is a function of: 1. the properties of the 'cue' and/or 'target', 2. the instructions, 3. the memory processes, and 4. any other information present in the experimental setting. All information outside of the properties of the actual 'cue/target' is said to describe the 'context' of the episode.

Memory tests can be categorised on the basis of the type of instructions and the type of 'cue' and 'target'[HBP89]. For example,

1. *Recognition*: a subject recalls whether or not the cue word is remembered as having occurred in a particular context.

2. *Cued recall*: a subject attempts to recall one member of a study pair of words given the other member as a cue.

3. *Cued recall with a part-word cue*: a subject attempts recall of a 'target' word given a fragment of that target as a cue.

4. *Free recall*: a subject reports the first word that comes to mind as having occurred in a particular context.

5. *Classification*: a subject classifies an episodic trace as belonging to a particular category.

Recognition, cued recall, free recall, and classification are predictive tasks. Each involves the presentation to a subject of a test-episode that specifies some properties of the experimental experience. From this specification the subject is asked to predict some other property. For example, a popular memory task is to ask a subject to memorise a list of word pairs. In cued recall the subject is given one word from each pair and is asked to predict the second.

2.5.2 Episodic Effects

It is well known that manipulations of the specific properties that can be used to describe a test or task episode modify performance on memory tests [WB88]. In particular, performance can be shown to depend on very specific properties associated with a very small number of task episodes. For example, Whittlesea and Brooks [WB88] showed that the correct prediction of a 'target' in a cued recall task is dependent upon the reinstatement of the experimental context as well as the presentation of the 'cue'. Performance disassociations due to the manipulation of the specific properties of small numbers of test and task episodes are referred to as *episodic effects.* This section reviews evidence suggesting that these effects are not minor, transient perturbations but due to: 1. The encoding of the specific properties of experiences [WB88], and 2. The use of these properties to retrieve reference classes containing a very small number of cases.

Experiments that study episodic effects manipulate properties of language units that are not expected to be true of all or most elements of a class of language units. For example, we might expect default knowledge such as 'a short word', or 'consisting of three letters "d", 'o", and "g"' to be true of all language units used to denote the concept 'dog' in English. We wouldn't expect knowledge such as 'bold type face', or 'written in red ink' to be part of the general knowledge about a language unit.

Schacter and Graf [SG89] find that performance on memory tests such as cued recall with a part word cue is facilitated for words that are studied and tested in the same sensory modality as opposed to words that are studied and tested in different sensory modalities. For example the word 'generation' might be presented visually during a memory task. If the word fragment 'gener' is presented visually during a memory test, the subject is more likely to correctly complete it as 'generation' than if the word fragment was written. A similar effect is reported when study and test episodes are presented in the same versus different symbolic fonts [WR87], and in the same versus different languages [RB87]. Whittlesea and Brooks [WB88] find that forcing subjects to encode only general properties during the memory task increases performance across changes in the memory task. However they also find that such general encodings result in less facilitation of performance on the memory task than if specific properties are encoded. This result suggests that specific properties specified in a memory test allow the retrieval of very specific experiences, i.e., a single experience in the memory task.

The results of experiments on manipulating the specific properties of experiences (e.g., [SG89], [WR87], [RB87]) suggest that: 1. The specific properties of experiences are described in memory as proposed by Tulving [FT78] [Tul72] [Tul76][Tul83] [Tul83] [Tul85] [TT73], and 2. The reinstatement of these properties in a test-episode results in a task-episode with the *same* matching specific properties being retrieved [HBP89], i.e., they allow us to retrieve a single past experience. This result supports

Assumption 1 in that the specific properties specified in a probability term can be used to specify the intension of the reference class of that probability term. The result also supports Assumption 3 by suggesting that given a task-episode human subjects will make a prediction by retrieving a single test-episode with the same specific properties. If subjects make predictions by ignoring or discounting specific properties such as modality, then performance on the cued recall tests should not depend upon modality.

While episodic effects are observed for words presented visually in the same font versus different fonts [JH87], little effect is observed when words are typed versus hand written [CM83], or when words are in upper versus lower case [SCS77]. Variations in format also appear to depend on the memory test. Graf and Ryan [GR90] report that format makes less difference when we are retrieving information about episodes than when we are accessing partially specified test-episodes. This suggests that some properties remain distinctive in memory [WW75]. These findings suggest that only some properties are used to describe episodes [6].

Transfer of appropriate processing (TAP) is a general proposal about the nature of human memory, that offers a general framework for theorising about specific episodic effects [GR90]. TAP assumes that performance on a memory test is expected to be facilitated to the extent that it specifies a *test-episode* that is similar to a preceding *task-episode.* The ease of memory processes is determined by the degree of overlap between the task and the test episodes [MBF77], i.e., the greater the number of specific properties of a task episode also specified in a test episode the more likely the subject is to retrieve the correct task episode. Tulving and others argue that each observation of a domain state is described separately in memory. Models in which descriptions of domain states are aggregated with existing memories are

[6] In the case of the RCA and the pattern matching algorithms all the cases are already described in terms of a set of properties so this result is not directly relevant. However, the result does suggest that care needs to be taken in defining the initial set of features that are used to describe domain objects as in CBR. This thesis does not address this important issue.

not able to separate domain states adequately, or to account for the prevalence of episodic effects [McC65].

2.5.3 Categories

The classification of objects into categories provides a tremendous amount of information about that object. For example, classifying an object as a 'bird' permits 'inferences' about how it moves, how fast it will go, ... Medin and Wattenmaker [MW87] suggest that "It is natural to categorise. Both our language and our experience lead us to treat non-identical stimuli in some way equivalent." Medin and Wattenmaker[MW87] further suggest that the categories that people normally create and use represent only a subset of the ways in which cases could be partitioned, i.e., there are lots of possible but useless categories. In this section I argue that certain categorical effects are consistent with aggregating over the members of a reference class extension defined in response to a specific situation of interest.

Franks and Bransford [MBF77] have suggested that categories such as 'bird' are defined in terms of a prototype that specific members resemble to a greater or lesser degree. Others have suggested that we classify on the basis of: 1. Family resemblance [Ros78], 2. *Exemplars* [MS78], 3. *Ideals*, or 4. Boundaries [AKA91]. Unfortunately, a consistent and comprehensive definition for categories has proven elusive (see for example, Armstrong, Gleitman and Gleitman[AGG83]). No one, for example, has found a rule for discriminating all games from all non-games, which persuaded Wittgenstein [Wit80] to advocate the use of *fuzzy categories* instead.

In fact, there has been increasing recognition that category representations are suffused with detailed knowledge about specific domain states that facilitates interactions with a highly specific context. For example, Kahneman, Slovic and Tversky [KST82] demonstrated that humans have considerable aptitude at "one shot learning", i.e., an ability to learn a concept from only one case or to remember exceptions to widely held generalisations. Such findings are in agreement with the TAP explanation for episodic effects discussed in the previous section. For example, when additional information is supplied about an object in a category, the properties associated with the category often change dramatically [HW90].

One possible explanation of such effects is that people construct categories that are relevant to making a particular prediction [Bar83] as in the RCA and the pattern matching algorithms. For example, answering the query 'If X is a purple bird, and in particular, X is a type of bird called a penguin, then does X fly?', might result in the formation of a category whose extension consists of all descriptions of purple birds that are penguins and from this category predicting flies.

Indeed, work by McCloskey and Glucksberg, [Bel84a, Bel84b, Bel84c], and Barsalou [Bar87] suggests that categories are much less stable than previously believed. For example, what is typical of a category varies widely as a function of its linguistic context [RS83] [HW90], i.e., tomato in the context ' ... fried green tomatoes ... ' evokes very different ideas of what is typical about tomatoes than tomato in the context ' ... tomato in the grocery store ...'. Similar effects are seen when using context to disambiguate word senses (see Simpson[SK89], Neill[Nei89], Simpson and Kellas[SK89], and Gorfein and Bubka[GB89]). Objects can also be *cross classified* into a large number of different categories. For example, a stump may be used as a chair or as a jack to support a car. While a great deal of work has addressed the ability to perform word sense disambiguation, little work has addressed the ability to form goal related categories and to perform *cross-classification.*

Exemplar theory [MS78] does not require humans to learn categories at all. Instead, every case gets stored in memory as in Tulving's [Tul72] model of episodic memory. However, it is improbable that a set of cases alone will be useful for making predictions. For example, vivid reasoning (discussed in Section 2.2.2) can only be used to make predictions from a set of cases about things that are already known. We might expect that the knowledge contained in a set of cases will have to be *generalised* [Lai88] in order to be useful. The RCA suggests that the cases are generalised by forming reference classes from which probabilities are estimated in response to the specific properties of a domain state about which a prediction is being made.

The findings of the apparent reliability of some predictions in classification may not be due to a reliance on the general properties of categories but rather due to the statistical stability granted by general access to large numbers of cases in a reference class [WB88]. For example, implementations of the RCA can use the properties of a set of objects to define the intension of a reference class. The cases in the extension of the reference class can be aggregated to produce a conditional probability. The predictions will be reliable if they are based upon a large reference class extension, where the effects of noise will be minimised. If the reference class extension is small, then the predictions will be unstable, that is, they will be strongly influenced by the addition of new cases and noise.

2.6 Discussion

Knowledge representations such as default theories, classification hierarchies, taxonomic hierarchies, and decision trees are founded on the premise that the cases used to describe past experiences should be *parsimoniously organised* on the basis of intuitively or statistically apparent structure. However, Bayesian arguments show that only external information about the likely mix of requests for estimates is relevant to this determination [Sch91], and this information is usually not available a priori. Furthermore, it is often unnecessary to impose a more parsimonious organisation on a set of cases if the cases are already sparse. Organising the cases parsimoniously results in already sparse domain knowledge being lost.

In the context of Assumption 1 there is no information about the likely mix of estimates that the RCA will be asked to provide. As a consequence it is necessary for the RCA to be able to respond to every eventuality as best as it can. I argue that by using syntactic generalisation and chaining to estimate probabilities directly, the RCA has this ability. In this thesis I demonstrate that these two techniques can be used to address the problem of making predictions directly from past experiences without depending upon the expertise and intervention of a knowledge base designer.

The reasonableness of the RCA's estimates are a function of the available past experiences. In particular, the RCA's will be reasonable only if Assumptions 3 and 4 hold. Assumption 3 states that any non-empty reference class is adequate with respect to making reasonable predictions. However, the larger the reference class, the more 'stable' the estimates and hence the less susceptible they will be to noise [Jr.88a] [Jr.88b] [Jr.88c] [Jr.88d] [Jr.91]. This thesis does not address the intriguing problem of balancing statistical stability with reasonableness although the problem can certainly be addressed within the RCA framework.

This thesis interprets Assumption 4 as stating that we should identify alternatives to an inadequate reference class by generalising irrelevant knowledge about a new experience. Bacchus [Bac90] assumes that the irrelevant knowledge has been identified by some external agent. However, in the context of Assumption 1 we must be able to identify the irrelevant knowledge using knowledge about the past experiences described in the KB. Of course, some knowledge may be more relevant than others with respect to a particular estimate. Thus, we we need to identify metrics that measure the relevance of knowledge with respect to specific estimates. In Chapter 5, I provide evidence suggesting that the reasonableness of the RCA's estimates is a function of these metrics as well as the past experiences that are available.

Although this thesis does not attempt to identify the 'best' metrics for measuring relevance it does provide, as an example, a candidate that performs well in a variety of domains. This thesis demonstrates that a simple implementation can use the RCA's estimates obtained using this metric to make more reasonable predictions than those obtained using k-NN and related techniques. This finding provides empirical support for the argument that Assumption 4 is a reasonable assumption.

In the following chapters I describe the RCA. The RCA provides a useful link between deductive models and inductive techniques:

1. Considerations as to the appropriate application of statistical knowledge in the RCA mirror considerations appropriate to the formalisation of non-monotonic logics as suggested by Kyburg [Jr.88a].

2. Inductive techniques used by classification algorithms and statistics can be used to manipulate cases in order to retrieve a most appropriate reference class of cases for making a prediction.

Chapter 3

A language for describing experiences

3.1 Introduction

This chapter describes a language, L, that is sufficiently expressive for talking about, and making predictions about, experiences. This chapter describes the properties of L that are used in this thesis for describing a KB of cases, and using the KB to estimate conditional probabilities. L has the following properties:

1. Individual ground sentences called *cases* describe specific experiences. An EKB contains a set of cases.

2. A distinguished set of ground terms called *labels* allow us to retrieve all cases describing past experiences that can be described by a case describing a new experience modulo the labels.

3. Probability terms, $Prob(\alpha|\beta)_{EKB}$, such that α and β are ground sentences and EKB is an EKB, are interpreted as *estimates* of the conditional probability of α given that β is all that is known to be true of a new experience with respect to EKB.

3.1.1 Chapter outline

The chapter is structured as follows:

1. Section 3.2 provides an overview of the language L.

2. Section 3.3 provides a definition of an EKB in terms of sentences of the language L.

3. Section 3.4 discusses two interpretations of probability terms in L.

4. Section 3.5 demonstrates, by means of example, that the choice of an interpretation of probability terms depends upon assumptions made about noise in the EKB.

5. Section 3.6 describes the properties that are true of the interpretation of probability terms used in Chapter 4 of this thesis.

3.2 An overview

This section discusses the: 1. Description of experiences using L, and 2. Prediction of conditional probabilities from a set of past experiences described in an EKB. The features described in this section are discussed in more detail in the remaining sections of this chapter. Appendix A contains a description of the formal properties of the language L.

3.2.1 Representing domain knowledge

In order to be useful, the language L must be expressive enough to describe experiences, but not so expressive that the retrieval of information about experiences from an EKB is computationally intractable. The following sections briefly address each of these issues.

Describing experiences

In this thesis an experience is described by a wfss in L. In L, experiences are described by disjoining, conjoining, and negating ground occurrences of a 3-ary relation R. A case is a well formed sentence (wfs) that specifies what is true of a particular experience.

Example 12 Let the feature 'colour' denote a variable with the possible values { 'red', 'green', 'blue', 'yellow', ... }. Suppose we wish to describe an experience in which a domain object is observed to be red in colour. In the language L, the fact that the object is red can be specified using the wfs

$$R(l_i, red, colour)$$

such that l_i is a label that denotes a particular observation of an object [1].

The relation R can describe objects that have more than one value for a feature.

Example 13 In the language L, the wfs

$$R(l_i, red, colour) \wedge R(l_i, yellow, colour)$$

describes an experience in which an object 'was observed to be red and yellow in colour'.

The relation R can also be used to describe n-ary domain relations. For example, consider a domain relation called a schedule that has a course number, a room number, time and an instructor, i.e.,

$$schedule(course, room, instructor, time)$$

An experience involving a particular schedule with course number 210, room number 312, instructor D. Poole, and time 2 p.m., can be described as

$$R(l_i, 210, course) \wedge R(l_i, 312, room) \wedge$$
$$R(l_i, Poole, instructor) \wedge R(l_i, 2pm, time)$$

such that l_i is a *label* denoting a particular observation of a 'schedule'.

An EKB contains a set of cases. For example,

[1] Labels are discussed in more detail later in the section.

Example 14 The cases

$$\begin{aligned}\{\ & R(l_1, 210, course) \wedge R(l_1, 312, room) \wedge R(l_1, Poole, instructor),\\ & R(l_2, 400, course) \wedge R(l_2, 210, room) \wedge R(l_i, Aha, instructor),\\ & R(l_3, 100, course) \wedge R(l_3, 109, room) \wedge R(l_i, Turney, instructor)\ \ \}\end{aligned}$$

describe three separate experiences. If the three cases were conjoined, then the EKB would contain a description of a single experience.

Retrieval from an EKB

First order logics such as *LP* [Bac90] contain features that are not needed to describe experiences. For example, there is no obvious way that a universally quantified variable can be observed in an experience. As a result universal quantification is excluded from L. Existential quantification is also excluded because all existentially quantified variables are effectively skolemized using labels. The exclusion of quantification from the language L means that the language is propositional and thus decidable. As a consequence we can be assured that it will be possible to retrieve all past experiences that can be described by a certain wfs modulo the labels.

However, even though L is propositional, the retrieval of cases from the EKB is NP-hard. Retrieval from an EKB can be efficient (i.e., polynomial) if we are willing to impose constraints on the syntax of the wffs that are used to describe the cases in the EKB. For example, we might restrict the wffs to conjuncts of positive literals as Levesque [Lev88] suggests in vivid reasoning. If we do, then an EKB effectively becomes a relational data base and retrieval is linear in the size of the EKB and sub-linear in the number of predicates used to describe the cases.

3.2.2 Estimating probabilities from an EKB

Estimates of conditional probabilities are represented in the language L as probability terms [2]:

$$Prob(\alpha|\beta)_{EKB}$$

Probability terms are used to estimate conditional probabilities. A probability term $Prob(\alpha|\beta)_{EKB}$ is interpreted as an estimate of the conditional probability of α when β is known to be true. The estimate is calculated by generalising over the *labels* in the wfss α and β.

Example 15 Suppose I have just observed a goat and a tiger and I want to estimate the probability that the tiger eats the goat. The estimate might be represented by the probability term

$$Prob(R(l_j, l_i, eats) \mid R(l_j, tiger, species) \wedge R(l_i, goat, species))_{EKB}$$

such that l_j denotes the particular observation of the tiger, and l_i the goat. The probability term is interpreted by generalising over the labels l_i and l_j to retrieve from the EKB all the descriptions of past experiences in which a tiger either eats or does not eat a goat.

Example 16 Suppose I have just observed two ants of different species and I want to estimate the probability of one ant eating the other. The estimate of the conditional probability could be represented by the probability term

$$Prob(R(l_j, l_i, eats) \mid R(l_j, l_k, species) \wedge R(l_i, l_l, species))_{EKB}$$

such that the labels l_j and l_i denote the two ants and l_k and l_l their species. The probability term is interpreted by generalising over the labels and retrieving from the EKB all observations of domain states in which an ant of one species eats an ant of another species.

[2] In the remainder of this thesis the subscript EKB in $Prob(\alpha|\beta)_{EKB}$ indicates that we are talking about an *estimate* of the conditional probability $Prob(\alpha|\beta)$.

$$\begin{array}{lc}
\{ & R(l_0, Craig, name) \wedge R(l_0, \top, glasses) \\
 & R(l_1, Craig, name) \wedge R(l_1, \top, glasses) \\
 & R(l_2, Craig, name) \wedge R(l_2, \top, glasses) \\
 & R(l_3, Craig, name) \wedge R(l_3, \top, glasses) \\
 & R(l_4, Craig, name) \wedge R(l_4, \top, glasses) \\
 & R(l_5, Alan, name) \wedge \neg R(l_5, \top, glasses) \\
 & R(l_6, Craig, name) \wedge R(l_6, \top, glasses) \\
 & R(l_7, Craig, name) \wedge R(l_7, \top, glasses) \\
 & R(l_8, Craig, name) \wedge R(l_8, \top, glasses) \\
 & R(l_9, Craig, name) \wedge R(l_9, \top, glasses) \\
 & R(l_{10}, Alan, name) \wedge R(l_{10}, \top, glasses) \\
 & R(l_{11}, Alan, name) \wedge \neg R(l_{11}, \top, glasses) \\
 & R(l_{12}, David, name) \wedge \neg R(l_{12}, \top, glasses) \quad \}
\end{array}$$

Figure 3.2: A set of cases describing 10 observations of Craig wearing glasses, two observations of Alan not wearing glasses, and one observation of David not wearing glasses.

Labels

As discussed in Chapter 1, a probability term $Prob(\alpha|\beta)_{EKB}$ can be interpreted by calculating the relative frequency with which $\alpha \wedge \beta$ is true with respect to some reference class. If Assumption 2 holds, then this frequency estimates the conditional probability of observing α given β.

In order to obtain the number of observations of domain properties, each distinct object in a case is denoted by an individual *label* - a distinct ground term - unique to that object and that case. Although the *same* object may be observed many different times, each observation of the object is denoted by a label unique to that object and to the case in which the object is described. In particular, two labels l_i and l_j denote distinct occurrences of objects if $i \neq j$. I now give an example showing how labels allow us to estimate a particular probability.

Example 17 Suppose I have seen Craig ten times, and each time I've seen Craig he has been wearing glasses. Suppose further that the only other people I have seen are David and Alan. I have seen David once, and he was not wearing glasses, and I have seen Alan twice, and he was not wearing glasses on either occasion. I might represent this knowledge using an EKB containing the set of cases in Figure 3.2. I now wish to estimate 'the probability that the next person I see will be wearing glasses'. If I have only kept track of the number of different people that I have seen wearing glasses, then I might estimate that the probability of the next person I see wearing glasses is one third as only one third of the people I have seen have been wearing glasses. A more accurate approximation of the probability would be $\frac{10}{13}$ as it is very likely, based upon previous experience, that the next person I will see will be Craig and every time I have seen Craig he was wearing glasses.

3.3 An EKB

Definition 7 *An EKB is a conjunct of all the axioms of L, together with a finite set of ground sentences written in L called cases. Each distinct label in a case is unique to that particular case.*

Examples in the remainder of this thesis only describe the cases in the EKB. The examples assume that the axioms consist of those described in Appendix A unless otherwise noted. For example,

Example 18

$$\begin{array}{ll}
\{R(l_0, red, colour) \ \wedge \ R(l_0, large, size), & \\
R(l_1, Ph.D., has\ degree) \ \vee \ R(l_1, MSc., has\ degree), & \\
R(l_2, l_3, colour) \ \wedge \ \neg(l_3 = red), & \\
R(l_4, l_5, father), & \\
\neg(R(l_6, red, colour) \ \wedge \ R(l_6, large, size)), & \\
\vdots & \vdots \\
R(l_{345}, Fred, name) & \}
\end{array}$$

might be the set of cases in a simple EKB.

Definition 8 *$EKB \vdash \alpha$ if there exists a case c_i in EKB such that the wfs α is implied by the conjunction of the axioms of the EKB and c_i.*

Example 19 Given the EKB in example 18, $EKB \vdash R(l_6, \mathit{red}, \mathit{colour})$ and $EKB \not\vdash R(l_8, MSc., \mathit{has\ degree})$.

If the cases in an EKB describe several experiences using the same syntax (modulo the labels of course), then the EKB can be expressed more parsimoniously as a set of pairs. This convention is sometimes adopted in this thesis to simplify the presentation of examples.

Definition 9 *$labels(\alpha)$ is the tuple $\langle l_1, \ldots, l_n\rangle$, such that $l_1 \ldots l_n$ are the distinct labels of α in order of first occurrence.*

Example 20 $labels(R(l_{17}, l_{23}, bigger) \wedge R(l_{17}, red, colour))$ is $\langle l_{17}, l_{23}\rangle$.

Definition 10 *$\alpha(X/Y)$, such that X is the n-ary tuple $labels(\alpha)$, and Y is the n-ary tuple $\langle l'_1, \ldots, l'_n\rangle$, is the result of substituting each occurrence in α of $l_i \in X$ by $l'_i \in Y$*

Definition 11 *$\langle n, \gamma_i\rangle$ is defined as follows: Let $\{\gamma_1, \ldots, \gamma_n\}$ be a subset of the cases in an EKB such that there is some m where X_i is the m-tuple $labels(\gamma_i)$. If $(\forall \gamma_i, \gamma_j)\ [\gamma_i = \gamma_j(X_i/X_j)]$, then $\{\gamma_1, \ldots, \gamma_n\}$ can be expressed as $\langle n, \gamma_i\rangle$*

Example 21 The EKB

$$\begin{array}{l}
\{(R(l_0, red, colour) \ \wedge \ R(l_0, large, size)),\\
(R(l_1, red, colour) \ \wedge \ R(l_1, large, size)),\\
(R(l_2, red, colour) \ \wedge \ R(l_2, large, size)),\\
(R(l_3, red, colour) \ \wedge \ R(l_3, large, size)),\\
(R(l_4, red, colour) \ \wedge \ R(l_4, large, size)),\\
(R(l_5, red, colour) \ \wedge \ R(l_5, large, size)),\\
(R(l_6, red, colour) \ \wedge \ R(l_6, large, size)),\\
(R(l_7, red, colour) \ \wedge \ R(l_7, large, size)),\\
(R(l_8, blue, colour) \ \wedge \ R(l_8, large, size)),\\
(R(l_9, blue, colour) \ \wedge \ R(l_9, large, size)), \quad \}
\end{array}$$

can be expressed as

$$\begin{array}{l}
\{ \quad \langle 8, R(l_0, red, colour) \ \wedge \ R(l_0, large, size)\rangle,\\
\{ \quad \langle 2, R(l_8, blue, colour) \ \wedge \ R(l_8, large, size)\rangle \quad \}
\end{array}$$

3.4 Interpreting probability terms

A probability term $Prob(\alpha|\beta)_{EKB}$ is interpreted with respect to the cases contained in the EKB by:

1. Describing the intension of the reference class of the probability term,

2. Retrieving all the entries in the EKB that are elements of the reference class extension,

3. Counting the number of elements in the reference class extension, and

4. Determining the proportion of elements in the reference class for which α is true.

Interpreting a probability term requires us to retrieve observations of specific properties from an EKB. I start by specifying intension of the reference class of a probability term $Prob(\alpha|\beta)_{EKB}$ in terms of the wfss α and β. I then show how a single wfs α is mapped to its extension. I then define the extension of a reference class in terms of its intension.

3.4.1 Defining the intension

The intension of a probability term $Prob(\alpha|\beta)_{EKB}$ specifies the domain knowledge that is relevant for predicting that α is true of a new experience given that all we know about the new experience is that β is true. I argue that as a minimum requirement the intension should take into account everything that we know.

Example 22 Suppose we wish to predict the propensity of a particular Ontario car to rust. We should take into account that we are interested in making a prediction about an Ontario car because the propensity of cars to rust may vary geographically. For example, the probability of cars in Ontario rusting may be very different from the probability of cars in California rusting.

An intuitive definition of the intension might specify that all past experiences that can be described by β (modulo the substitution of labels) are relevant with respect to interpreting $Prob(\alpha|\beta)_{EKB}$. For example, the intension of the reference class of $Prob(\alpha|\beta)_{EKB}$ might be simply defined as β. In Chapter 1, I suggested that β is an inappropriate intension because it does not take into account what we are trying to predict. For example,

Example 23 Suppose we are interested in estimating the probability of a red bird flying. It seems intuitive that our reference class should take into account the fact that we are interested in a red bird, thus the inclusion of β in the reference class intension. Equally intuitive, is the observation that our reference class should take into account the fact that we are interested in red birds that fly or do not fly. For example, considering objects that are *only* known to be red birds does not tell us anything about the propensity of red birds to fly.

I argue that it is necessary to take both α and β into account when defining the intension of a reference class.

Definition 12 *The pair* $\langle \alpha, \beta \rangle$ *is the intension of the probability term* $Prob(\alpha \mid \beta)_{EKB}$.

The next section discusses the problem of mapping from each wfs in the intension $\langle \alpha, \beta \rangle$ of a probability term $Prob(\alpha|\beta)_{EKB}$ to its extension.

3.4.2 Defining the extension

The extension of a single wfs γ is obtained by generalising over all the labels in γ. The extension contains all the objects described in an EKB that have the same properties as the objects generalised in γ. In this section a function h is defined that specifies the extension of any wfs in L with respect to a particular EKB.

h is defined as follows:

Definition 13 *Let* Ω *be the set of all possible EKBs. Let* Γ *be the set of all possible ground sentences in* L. *Let* R *be the set of all possible tuples of labels in* L. h *is the mapping* $h : \Omega \times \Gamma \rightarrow R$ *such that* $h(EKB, \alpha)$ *is:*

$$h(EKB, \alpha) = \{Y : EKB \vdash \alpha(X/Y)\}$$

Let $labels(\alpha)$ be an n-tuple. Informally, each tuple in $h(EKB, \alpha)$ denotes a past experience, described in the EKB, that can be described by α (modulo the substitution of labels). For example,

Example 24 Suppose, the EKB contains the set of cases,

$$\begin{aligned} \{ \quad & R(l_1, red, colour) \wedge R(l_2, red, colour), \\ & R(l_3, red, colour) \wedge R(l_4, red, colour) \quad \} \end{aligned}$$

$$h(EKB,\ R(l_{987}, red, colour)) = \{\langle l_1\rangle, \langle l_2\rangle, \langle l_3\rangle, \langle l_4\rangle, \}$$

such that each tuple corresponds to a past experience in which an object was observed to be red.

The function h addresses the problem of retrieving observations of objects from an EKB. If we wish to count cases that describe past experiences involving a *specific number* of domain objects, then we might wish to consider expanding the definition of h. For the sake of completeness, Appendix D provides the necessary expansions of the definition. The reader should note that it may, or may not, make sense to take the specific number of domain objects into account when interpreting a probability term.

The extension of a probability term $Prob(\alpha|\beta)_{EKB}$ whose intension only takes β into consideration is $h(EKB, \beta)$. In the next section I demonstrate that $h(EKB, \beta)$ may be an inappropriate extension. In the remainder of this thesis I use an extension that takes both α and β into consideration.

Definition 14 *The probability term $Prob(\alpha|\beta)_{EKB}$ with intension $\langle\alpha, \beta\rangle$ has extension*

$$h(EKB, \alpha \wedge \beta) \cup h(EKB, \neg\alpha \wedge \beta)$$

3.4.3 Interpreting probability terms to estimate $Prob(\alpha|\beta)$

Once the reference class of a probability term $Prob(\alpha|\beta)_{EKB}$ has been defined we can interpret the probability term. We start by counting the number of past experiences contained in the reference class extension by determining its cardinality.

Definition 15 *The cardinality $|\alpha|_{EKB}$ of a wfs alpha with respect to a particular EKB is $|h(EKB, \alpha)|$, the number of tuples in the extension $h(EKB, \alpha)$ of α.*

Example 25

$$|(R(l_{987}, red, colour) \wedge R(l_{987}, large, size))|_{EKB}$$

is the number of past experiences described in the EKB involving objects known to have the property 'colour red and size large'.

Definition 16 *The cardinality of the reference class extension*

$$h(EKB, \alpha \wedge \beta) \cup h(EKB, \neg\alpha \wedge \beta)$$

of the probability term $Prob(\alpha|\beta)_{EKB}$ is

$$|\alpha \wedge \beta|_{EKB} + |\neg\alpha \wedge \beta|_{EKB}$$

A probability term succeeds if its reference class extension has a cardinality greater than zero.

Definition 17 *If a reference class extension of a probability term is not empty, then the probability term* **succeeds***. Otherwise the probability term* **fails***.*

Definition 18 *The interpretation of a probability term, Prob, is*

$$Prob(\alpha|\beta)_{EKB} = \frac{|(\alpha \wedge \beta)|_{EKB}}{|(\alpha \wedge \beta)|_{EKB} + |(\neg\alpha \wedge \beta)|_{EKB}}$$

if it does not fail.

In the following section I demonstrate that the interpretation of a probability term depends upon the definition of the reference class. I argue that the interpretation presented in this thesis is a good one because it makes reasonable assumptions about the cases in the EKB.

3.5 Choosing among different interpretations

In this section I demonstrate that the interpretation of a probability term depends upon the definition of the reference class. I demonstrate, by means of example that some interpretations are more appropriate than others. In particular, this section shows that choosing the most reasonable interpretation depends upon making reasonable assumptions about the cases in the EKB. As an illustrative example, this section discusses the problem of choosing between two different interpretations when the cases in the EKB are incomplete. I argue, that if the cases in the EKB are incomplete, then the interpretation of probability terms given in Section 3.4.3 is a reasonable one.

3.5.1 Incomplete EKBs

I start by defining what I mean by an incomplete EKB.

Definition 19 *An EKB is incomplete with respect to a term $Prob(\alpha|\beta)_{EKB}$ if*

$$|\beta|_{EKB} > |\alpha \wedge \beta|_{EKB} + |\neg\alpha \wedge \beta|_{EKB}$$

Definition 20 *The number of observations, $N(\alpha, \beta)$, of domain properties described in the EKB for which β is known to be true but for which the truth of α is unknown is:*

$$N(\alpha, \beta) \; = \; |\beta|_{EKB} - (|\alpha \wedge \beta|_{EKB} + |\neg\alpha \wedge \beta|_{EKB})$$

Proposition 1 *If $N(\alpha, \beta) \neq 0$, then the EKB is incomplete.*

Definition 21 *p is the proportion of the $N(\alpha, \beta)$ past experiences in the extension of $\alpha \wedge \beta$, and $1 - p$ is the proportion of $N(\alpha, \beta)$ past experiences in the extension of $\neg\alpha \wedge \beta$.*

3.5.2 Two different interpretations of $Prob(\alpha|\beta)_{EKB}$

I now define two different interpretations of a probability term $Prob(\alpha|\beta)_{EKB}$ in which the proportion p of $N(\alpha, \beta)$ observations is unknown. Each interpretation may result in a different estimate of the conditional probability $Prob(\alpha|\beta)$.

Estimate I:

$$\begin{aligned} Prob^{I}(\alpha|\beta)_{EKB} &= \frac{|(\alpha \ \wedge \ \beta)|_{EKB}}{|\beta|_{EKB}} \\ &= \frac{|(\alpha \ \wedge \ \beta)|_{EKB}}{N(\alpha,\beta) + |\alpha \wedge \beta|_{EKB} + |\neg\alpha \wedge \beta|_{EKB}} \end{aligned}$$

Estimate II:

$$Prob^{II}(\alpha|\beta)_{EKB} = \frac{|(\alpha \ \wedge \ \beta)|_{EKB}}{|(\alpha \ \wedge \ \beta)|_{EKB} \ + \ |(\neg\alpha \ \wedge \ \beta)|_{EKB}}$$

Estimate II is the estimate given in Section 3.4.3.

If the EKB is incomplete, then the estimates I and II may provide different estimates of conditional probabilities. For example,

Example 26 Suppose we wish to interpret the probability term

$$Prob(R(l_i, \top, flies)|R(l_i, \top, wings))_{EKB}$$

with respect to the EKB

$$\begin{aligned} \{ \quad & R(l_1, \top, flies) \wedge R(l_1, \top, wings), \\ & \neg R(l_2, \top, flies) \wedge R(l_2, \top, wings), \\ & R(l_3, \top, flies) \wedge R(l_3, \top, wings), \\ & R(l_4, \top, yellow) \wedge R(l_4, \top, wings), \\ & R(l_5, \top, plane) \wedge R(l_5, \top, wings), \\ & R(l_6, \top, plane) \wedge R(l_6, \top, green) \quad \} \end{aligned}$$

The EKB is incomplete because $N(flies, wings) = 2$. The reference class extension, $h(EKB, R(l_i, \top, wings))$, of Interpretation I is

$$\{\langle l_1\rangle, \langle l_2\rangle, \langle l_3\rangle, \langle l_4\rangle, \langle l_5\rangle\}$$

and

$$Prob^{I}(R(l_i, \top, flies)|R(l_i, \top, wings))_{EKB} = \frac{2}{5}$$

The reference class extension of Interpretation II is $\{\langle l_1\rangle, \langle l_2\rangle, \langle l_3\rangle\}$ and

$$Prob^{II}(R(l_i, \top, flies)|R(l_i, \top, wings))_{EKB} = \frac{2}{3}$$

In the remainder of this section I compare the two different interpretations with the *Ideal* interpretation in which p is known.

Ideal estimate:

$$\begin{aligned} Prob^G(\alpha|\beta)_{EKB} &= \frac{|(\alpha \wedge \beta)|_{EKB} + pN(\alpha, \beta)}{|\beta|_{EKB}} \\ &= \frac{|(\alpha \wedge \beta)|_{EKB} + pN(\alpha, \beta)}{N(\alpha, \beta) + |\alpha \wedge \beta|_{EKB} + |\neg\alpha \wedge \beta|_{EKB}} \\ &= Prob^I(\alpha|\beta)_{EKB} + \frac{pN(\alpha, \beta)}{|\beta|_{EKB}} \end{aligned}$$

I assume that $Prob^G$ is the most reasonable estimate of $Prob(\alpha|\beta)$ if the EKB is incomplete because the probability p is known. I now show that estimates I and II make different a-priori assumptions about the value of p. I argue that the reasonableness of the estimates can be judged by comparing them to the Ideal estimate in which p is known.

3.5.3 Assume p = 0: Estimate I

Estimates I and II make implicit assumptions about the value of p. In this section I demonstrate that Estimate I assumes that $p = 0$. As a consequence Estimate I counts all $N(\alpha, \beta)$ past experiences as members of the extension of $\neg\alpha \wedge \beta$.

Example 27 Suppose we are trying to interpret the probability term

$$Prob(R(l_i, red, colour)|R(l_i, large, size))$$

Now suppose the EKB contains the following case:

$$R(l_6, large, size) \wedge (R(l_6, red, colour) \vee R(l_6, blue, colour))$$

$Prob^I$ assumes $R(l_6, large, size) \wedge \neg R(l_6, red, colour)$.

The difference between Estimate I and Estimate II when $p = 0$ can be clarified by comparing them to the Ideal interpretation. First, I examine the trivial situation in which the cases in the EKB are not incomplete.

Proposition 2 *If the cases are not incomplete, then* $N(\alpha, \beta) = 0$ *and*

$$Prob^{I} = Prob^{II} = Prob^{G}$$

Proof: The equality follows from the definitions of $Prob^{I}$, $Prob^{G}$, and $Prob^{II}$. □

In the absence of noise Estimates I and II are the same as the Ideal estimate of $Prob(\alpha|\beta)$.

I now examine the more interesting situation in which the cases are incomplete, i.e., $N(\alpha, \beta) > 0$.

Proposition 3 *If* $N(\alpha, \beta) > 0$ *then*

$$Prob^{II} >= Prob^{I}$$

Estimate $Prob^{II}$ *always estimates a higher probability than* $Prob^{I}$*.*

Proof: The inequality follows from the definitions of $Prob^{I}$, and $Prob^{II}$. □

$Prob^{I}$ is a good or bad estimator of the conditional probability depending on what the value of p is when $N(\alpha, \beta) > 0$ as demonstrated by the following propositions. First, I show that when $p = 0$ $Prob^{I}$ is the same as the Ideal interpretation.

Proposition 4 *Let* $N(\alpha, \beta) > 0$ *and assume* $p = 0$*.*

1. $Prob^{I}(\alpha|\beta)_{EKB} = Prob^{G}(\alpha|\beta)_{EKB}$
2. $Prob^{II}(\alpha|\beta)_{EKB} > Prob^{G}(\alpha|\beta)_{EKB}$

Proof: Part 1 is proved as follows:

$$\begin{aligned} Prob^G(\alpha|\beta)_{EKB} &= \frac{|(\alpha\ \wedge\ \beta)|_{EKB} + pN(\alpha,\beta)}{|\beta)|_{EKB}} \\ &= \frac{|(\alpha\ \wedge\ \beta)|_{EKB} + pN(\alpha,\beta)}{N(\alpha,\beta) + |\alpha \wedge \beta|_{EKB} + |\neg\alpha \wedge \beta|_{EKB}} \\ &= Prob^I(\alpha,\beta)_{EKB} + \frac{pN(\alpha,\beta)}{|\beta|_{EKB}} \end{aligned}$$

If $p = 0$ then $Prob^G = Prob^I$. Part 2 follows from proposition 2. □

3.5.4 Assume $p = Prob^{II}$: Estimate II

From proposition 3 we can argue that $Prob^I$ is the most reasonable estimate of $Prob(\alpha|\beta)$ when $p = 0$. I argue in this section that it is more reasonable to assume $p = Prob^{II}(\alpha|\beta)_{EKB}$. I show that if $p = Prob^{II}(\alpha|\beta)_{EKB}$, then $Prob^{II}$ is a more reasonable estimate than $Prob^I$.

I argue that there is no information in an EKB that supports the assumption that $p = 0$. However, there is information in the EKB that supports the assumption that $p = Prob^{II}$. For example,

Example 28 Consider the probability term

$$Prob(R(l_i, red, colour)|R(l_i, large, size))_{EKB}$$

Suppose the cases in the EKB are incomplete with respect to the probability term because they contain

$$R(l_i, large, size) \wedge (R(l_i, red, colour) \vee R(l_i, blue, colour))$$

l_i is in the set N non-counted observations as its colour is not known to be red or not red. As before, $Prob^I$ treats l_i as an example of an observations of a large non-red object. $Prob^{II}$ says that maybe l_i is red, or maybe l_i is not red, so lets forget about l_i. When we say p=0 we say there is no way that l_i could be red which is not true because we already know that l_i is red or blue. It is more reasonable to say $p = Prob^{II}$ and assume that the chance that l_i is red is the same as it is for those observations of large objects where we know the colour is red or not red.

I now show that if $p = Prob^{II}(\alpha|\beta)_{EKB}$, then $Prob^{II}$ is more reasonable than $Prob^{I}$.

Theorem 1 *If* $p = Prob^{II}(\alpha|\beta)_{EKB}$ *and* $N(\alpha, \beta) > 0$*, then:*

1. $Prob^{G}(\alpha|\beta)_{EKB} = Prob^{II}(\alpha|\beta)_{EKB}$
2. $Prob^{G}(\alpha|\beta)_{EKB} > Prob^{I}(\alpha|\beta)_{EKB}$

Proof: The first part is proved as follows:

$$\begin{aligned} Prob^{G}(\alpha|\beta)_{EKB} &= \frac{|(\alpha \ \wedge \ \beta)|_{EKB} + pN(\alpha,\beta)}{|\beta)|_{EKB}} \\ &= \frac{|(\alpha \ \wedge \ \beta)|_{EKB} + pN(\alpha,\beta)}{N(\alpha,\beta) + |\alpha \wedge \beta|_{EKB} + |\neg\alpha \wedge \beta|_{EKB}} \end{aligned}$$

If $p = Prob^{II}$, then

$$\begin{aligned} Prob^{G}(\alpha|\beta)_{EKB} &= \frac{|(\alpha \ \wedge \ \beta)|_{EKB} + \frac{|\alpha\wedge\beta|}{|\alpha\wedge\beta|+|\neg\alpha\wedge\beta|}N(\alpha,\beta)}{N(\alpha,\beta) + |\alpha \wedge \beta|_{EKB} + |\neg\alpha \wedge \beta|_{EKB}} \\ &= Prob^{II}(\alpha|\beta)_{EKB} \end{aligned}$$

The second part follows from part 1 and proposition 2. □

3.5.5 $Prob^{I}$ versus $Prob^{II}$

The choice between $Prob^{I}$ and $Prob^{II}$ as an estimator of the conditional probability $Prob(\alpha|\beta)$ is a choice between assuming that $p = Prob^{II}$ or $p = 0$. That is, if $N(\alpha, \beta) > 0$, then

1. If $Prob^{II}(\alpha|\beta)_{EKB} = Prob^{G}(\alpha|\beta)_{EKB}$, then $p = Prob^{II}(\alpha|\beta)_{EKB}$.

2. If $Prob^{I}(\alpha|\beta)_{EKB} = Prob^{G}(\alpha|\beta)_{EKB}$, then $p = 0$.

The comparison of $Prob^{I}$ and $Prob^{II}$ to $Prob^{G}$ in this section supports the argument that the choice of reference class depends upon making assumptions about the noise in an EKB.

Any a-priori knowledge about the noise can be used to choose the most reasonable estimate. For example, if the cases in the EKB are incomplete and p is known to be 0, then $Prob^{I}$ is the most reasonable estimate. If on the other hand p is known to be 1, then the most reasonable estimate is

$$\begin{aligned} Prob^{G}(\alpha|\beta)_{EKB} &= \frac{|(\alpha \wedge \beta)|_{EKB} + pN(\alpha,\beta)}{|\beta|_{EKB}} \\ &= \frac{|(\alpha \wedge \beta)|_{EKB} + |\beta| - (|\alpha \wedge \beta| + |\neg\alpha \wedge \beta|)}{|\beta|_{EKB}} \\ &= \frac{|\beta| - |\neg\alpha \wedge \beta|}{|\beta|_{EKB}} \\ &= 1 - Prob^{I}(\neg\alpha|\beta)_{EKB} \end{aligned}$$

It is interesting to note that from Proposition 3 and the following proposition, $Prob^{II}$ is bounded by the Ideal estimate for $p = 0$ and $p = 1$.

Proposition 5 *Let $N(\alpha,\beta) > 0$ and assume $p = 1$.*

1. $Prob^{G}(\alpha|\beta)_{EKB} > Prob^{II}(\alpha|\beta)_{EKB}$

2. $Prob^{II}(\alpha|\beta)_{EKB} > Prob^{I}(\alpha|\beta)_{EKB}$

Proof: Part 1 is proved as follows:

$$\begin{aligned} Prob^{G}(\alpha|\beta)_{EKB} &= \frac{|(\alpha \wedge \beta)|_{EKB} + N(\alpha,\beta)}{|\beta|_{EKB}} \\ &> Prob^{II}(\alpha|\beta)_{EKB} \end{aligned}$$

Part 2 follows from proposition 2. □

From the previous proposition we see that even though $Prob^{I}$ and $Prob^{II}$ are both under estimators compared to $Prob^{G}$ when $p = 1$, $Prob^{II}$ is a better estimate. I argue that if we have no a-priori knowledge about p, as in the context of this thesis, then assuming $p = Prob^{II}$ is more reasonable that assuming that $p = 0$. Thus, $Prob^{II}$ is the estimate adopted in the remainder of this thesis.

3.6 Some properties of $Prob^{II}(\alpha|\beta)_{EKB}$

In this section I briefly discuss some useful properties of the estimate $Prob^{II}$. These properties are used in Chapter 4 when finding alternatives to the reference class of a failed probability term. In the following section I demonstrate that $Prob^{II}$ is not a conditional probability in the strict sense of the word because it only satisfies three of the four theorems of probability theory. However, I argue that its failure to satisfy one of the theorems is an advantage rather than a disadvantage in the context of Assumption 1.

The following theorems are true of the revised estimate of a probability term adopted in the remainder of this thesis.

Theorem 2

$$Prob^{II}(\alpha|\gamma)_{EKB} \; = \; Prob^{II}(\beta|\gamma)_{EKB} \; if \;\; \vdash \alpha \equiv \beta$$

Proof: The proof follows from the definition of L and $Prob^{II}$. □

Theorem 3

$$Prob^{II}(\top|\gamma)_{EKB} \; = \; 1$$

Proof: From the definition of $Prob^{II}$ we get

$$Prob^{II}(\top|\gamma)_{EKB} \; = \; \frac{|\top \wedge \gamma|}{|\top \wedge \gamma| \; + \; |\neg\top \wedge \gamma|}$$

From the semantics of L, $|\neg\top| = 0$ and

$$Prob^{II}(\top|\gamma)_{EKB} = \frac{|\top \wedge \gamma|}{|\top \wedge \gamma| + 0} = 1$$

□

Theorem 4

$$Prob^{II}(\neg\alpha|\gamma)_{EKB} = 1 - Prob^{II}(\alpha|\gamma)_{EKB}$$

if $Prob^{II}$ doesn't fail.

Proof: From the definition of $Prob^{II}$ we get

$$\begin{array}{lcl} Prob^{II}(\neg\alpha|\gamma)_{EKB} & = & 1 - Prob^{II}(\alpha|\gamma)_{EKB} \\ \frac{|\neg\alpha\wedge\gamma|}{|\neg\alpha\wedge\gamma| + |\alpha\wedge\gamma|} & = & 1 - \frac{|\alpha\wedge\gamma|}{|\alpha\wedge\gamma| + |\neg\alpha\wedge\gamma|} \\ \frac{|\alpha\wedge\gamma| + |\neg\alpha\wedge\gamma|}{|\alpha\wedge\gamma| + |\neg\alpha\wedge\gamma|} & = & 1 \\ 1 & = & 1 \end{array}$$

□

The theorem

$$Prob(\alpha \vee \beta|\gamma) = Prob(\alpha|\gamma) \vee Prob(\beta|\gamma) \ if \ \vdash \neg(\alpha \wedge \beta)$$

of probability theory is not a theorem of $Prob^{II}$ because it is possible to have a case c_i in the EKB such that:

$$[c_i \vdash (\alpha \vee \beta)] \wedge [c_i \not\vdash \alpha] \wedge [c_i \not\vdash \beta]$$

For example, we might observe an object to be red or blue but not know which. However, all this means is that if we know $\alpha \vee \beta$, then we should use the reference class of $Prob^{II}(\alpha \vee \beta|\gamma)_{EKB}$ rather than the combined reference class of $Prob^{II}(\alpha|\gamma)_{EKB}$ and $Prob^{II}(\beta|\gamma)_{EKB}$ because the former may be larger and the larger the reference class the better the statistics.

3.7 Discussion

This chapter presents a language L for describing experiences. Because quantification is not required to describe the experiences, L is propositional and thus decideable. As discussed in the introduction to this chapter, L can be restricted to allow computationally efficient retrieval from an EKB (i.e., retrieval that is possible in polynomial time). It is interesting to note that the restrictions required to make retrieval from an EKB efficient are the same restrictions that are required to make vivid reasoning [Lev88] efficient. This suggests, that considerations as to efficient retrieval from an EKB mirror considerations as to efficient retrieval from a VKB or a relational data base.

L is more expressive than the feature vectors often used to describe experiences in the psychological (e.g., [Tul86]), machine learning (e.g., [Fis87] [AKA91]), and pattern recognition (e.g., [Das91]) literatures. For example, L can be used to specify n-ary relations, disjuncts, conjuncts, and negations in a straight forward fashion.

This chapter presents a method for estimating conditional probabilities directly from a set of cases without the addition of any other domain knowledge. The chapter also demonstrates that different interpretations of probability terms can lead to different estimates. The chapter shows that the choice of a particular interpretation depends upon the assumptions that are made about any noise in an EKB. In particular, this chapter argues that $Prob^{II}$ is the most reasonable interpretation of a probability term and thus the most reasonable estimate of $Prob(\alpha|\beta)$ if the cases in the EKB are incomplete.

The next chapter addresses the problem of interpreting probability terms that **fail**.

Chapter 4

Generalisation and Chaining

4.1 Introduction

This chapter addresses the problem of deriving approximations to failed probability terms by *syntactic generalisation* and its novel extension *chaining*. *Syntactic generalisation* and *chaining* derive approximations to a failed probability terms [1] by:

1. Identifying a set, Θ, of intensions of adequate, yet epistemologically relevant, alternatives to the original reference class of a failed probability term $Prob^{II}(\alpha|\beta)$, by modifying the syntax of β, and

2. Deriving an approximation of $Prob^{II}(\alpha|\beta)$ by choosing a *single* item from the set
$$\{Prob^{II}(\alpha|\beta')|\langle\alpha,\beta'\rangle \in \Theta\}$$

4.1.1 Chapter outline

Section 4.2 distinguishes between the syntactic and semantic generalisations of a failed probability term. Both have the property that a reference class R' with intension $\langle\alpha,\beta'\rangle$ is an adequate generalisation of a reference class R with intension $\langle\alpha,\beta\rangle$, if $\beta \rightarrow \beta'$ and

$$h(EKB,\alpha \wedge \beta') \cup h(EKB,\neg\alpha \wedge \beta') \neq \emptyset$$

[1] This chapter is concerned only with estimates of conditional probabilities. Throughout this chapter the EKB subscript used to denote an estimate of a probability as opposed to the actual probability is dropped.

Intuitively, if $\beta \rightarrow \beta'$, then the domain knowledge specified by $\langle \alpha, \beta' \rangle$ is more general than the domain knowledge specified by $\langle \alpha, \beta \rangle$. For example,

Example 29 Consider the probability term[2].

$$Prob^{II}(flies \mid bird \wedge red \wedge dead)$$

Its reference class intension $\langle flies, bird \wedge red \wedge dead \rangle$ states that the relevant domain knowledge consists of all observations of dead, red, birds that fly or do not fly, i.e., the extension is

$$h(EKB, flies \wedge bird \wedge red \wedge dead) \cup h(EKB, \neg flies \wedge bird \wedge red \wedge dead)$$

If the probability term fails, then intension can be generalised to $\langle flies, bird \wedge dead \rangle$ because $(bird \wedge dead \wedge red) \rightarrow (bird \wedge red)$. The generalisation $\langle flies, bird \wedge dead \rangle$ states that the relevant domain knowledge consists of all observations of dead birds that fly or do not fly, i.e., the extension is

$$h(EKB, flies \wedge bird \wedge dead) \cup h(EKB, \neg flies \wedge bird \wedge dead)$$

The domain knowledge specified by the generalisation is more general than the domain knowledge specified by the original intension because all observations of dead red birds are also observations of dead birds.

Both syntactic and semantic generalisation have the additional property that they only consider the most-specific adequate generalisations of a failed probability term. Informally, an adequate reference class R_1 is a *most specific* adequate generalisation of R_2 if there does not exist an adequate reference class R_3 such that R_2 is a generalisation of R_3 and R_3 is a generalisation of R_1. Intuitively, the most-specific adequate generalisations are most likely to lead to reasonable approximations because they take into account as much information as possible about the situation of interest [Bac90]. For example,

[2] In the remainder of this chapter I sometimes write $flies$ instead of $R(l_i, flies, moves)$ and so on.

Example 30 Suppose we wish to estimate the probability that a red bird with large wings flies. We might represent the probability as the probability term

$$Prob^{II}(flies|bird \wedge wings \wedge red)$$

If the probability term fails, $Prob^{II}(flies|bird \wedge red)$ is a better approximation than $Prob^{II}(flies \mid bird)$ because its intension $\langle flies, bird \wedge red \rangle$ takes into account more information about the situation of interest, i.e., that we are interested in a red bird rather than a bird of any colour.

In Section 4.3 I argue that not all most-specific adequate generalisations of a failed probability term lead to reasonable approximations. For example,

Example 31 Suppose we wish to estimate the probability that a red bird with large wings flies. $Prob^{II}(flies|bird \wedge large\,wings)$ and $Prob^{II}(flies|bird \wedge red)$ are both equally specific generalisations but the former is a more reasonable approximation than the latter if knowing 'large wings' is more relevant to estimating the probability of flying than knowing 'red'.

I argue that semantic information obtained from an EKB should be used to select a single, most reasonable most specific generalisation of a failed probability term.

In Section 4.4 I demonstrate, by means of example, that the less that is known about the new experience, the less likely the most specific generalisation of a failed probability term $Prob^{II}(\alpha|\beta)$ is to be reasonable. Chaining is defined as a novel extension of generalisation in which we extend what is known about the new experience. For example, we might assume that if β is true, then γ is also true and approximate $Prob^{II}(\alpha|\beta)$ by generalising $Prob^{II}(\alpha|\beta \wedge \gamma)$. Chaining is based upon the assumption that the more that is known about the situation of interest the more likely it is that a single *relevant* most specific generalisation will be found. For example,

Example 32 Suppose we wish to estimate the probability of aspirin being prescribed as a treatment given that a patient appears flushed, i.e.,

$$Prob^{II}(aspirin \mid flushed)$$

All that is known about the situation of interest is that the patient is flushed. We might add to what is known about the situation of interest by assuming that $flushed$ patients are also $fevered$, to get

$$Prob^{II}(aspirin \mid flushed \wedge fevered)$$

(I argue later that this is a reasonable thing to do if we know that $Prob^{II}(fevered \mid flushed)$ is very high or very low), which, using generalisation, can be approximated by $Prob^{II}(aspirin \mid fevered)$.

In this chapter chaining is presented as a straightforward extension of syntactic generalisation that allows us to obtain more knowledge about a situation of interest before generalising.

4.2 Identifying Θ by generalising

This section starts by defining semantic generalisation. The section demonstrates by means of example that using semantic generalisation results in unreasonable approximations. The section concludes by defining syntactic generalisation as a constrained from of semantic generalisation that does not result in the consideration of unreasonable alternatives to a failed probability term.

4.2.1 Semantic generalisation

The reference class of a failed probability term $Prob^{II}(\alpha|\beta)$ has intension $\langle\alpha, \beta\rangle$ and extension $h(EKB, \alpha \wedge \beta) \cup h(EKB, \neg\alpha \wedge \beta)$. A natural way of generating adequate alternatives to the reference class is to consider the set of all reference classes whose extensions *include* the extension of the original reference class as a subset.

As suggested by Reichenbach [Rei49], a natural generality ordering upon the set of all reference classes whose extensions include

$$h(EKB, \alpha \wedge \beta) \cup h(EKB, \neg\alpha \wedge \beta)$$

is subset containment. For example,

Example 33 The sets of 'dead birds' and 'birds' are both generalisations of the set of 'dead red birds' as

$$h\left(EKB, \begin{array}{c} R(l_i, \top, dead) \wedge \\ R(l_i, bird, species) \end{array}\right) \supseteq h\left(EKB, \begin{array}{c} R(l_i, \top, dead) \wedge \\ R(l_i, bird, species) \wedge \\ R(l_i, red, colour) \end{array}\right)$$

and

$$h(EKB, R(l_i, bird, species)) \supseteq h\left(EKB, \begin{array}{c} R(l_i, \top, dead) \wedge \\ R(l_i, bird, species) \wedge \\ R(l_i, red, colour) \end{array}\right)$$

however, the set of 'dead birds' is a more specific generalisation than 'birds' as

$$h\left(EKB, \begin{array}{c} R(l_i, \top, dead) \wedge \\ R(l_i, bird, species) \end{array}\right) \subseteq h(EKB, R(l_i, bird, species))$$

Defining semantic generalisation in terms of set inclusion is problematic when we wish to generate the generalisations of a reference class with an empty extension, the case of interest in this thesis. As pointed out by Kyburg [Jr.88a] and Bacchus [Bac90], any reference class with a non-empty extension is a superset of a reference class with an empty extension. This means that an approximation to a failed probability term could be based upon any adequate reference class whose intension can be defined in the language L.

Example 34 Suppose the probability term

$$Prob^{II}(flies \mid red \wedge\ bird)$$

fails. Using subset containment we can approximate the probability term using any probability term with an adequate reference class that can be described in the language L. For example, if $Prob^{II}(flies \mid truck)$ and $Prob^{II}(flies \mid professor)$ have adequate reference classes, then both be used to approximate the probability of red birds flying.

Instead of ordering reference classes semantically by subset containment over their extensions Bacchus [Bac90] argues that we should order them semantically by implication over their intensions (e.g., [Bac90], [Lai88]).

Definition 22 *The reference class of the probability term* $Prob^{II}(\alpha|\beta)$ *is a semantic generalisation of the reference class of* $Prob^{II}(\alpha|\beta')$ *if* $\beta' \rightarrow \beta$.

By considering a generalisation ordering over the intensions of reference classes we avoid the difficulty of defining subset containment over empty reference classes.

Using logical implication, the set of all adequate semantic generalisations of a reference class can be defined as follows

Definition 23 *Let* Γ *be the set of all wfss in* L. *For any* $\beta' \in \Gamma$, *the set* $G(Prob^{II}(\alpha \mid \beta))$ *of intensions of all adequate semantic generalisations is*

$$\{\langle\alpha, \beta'\rangle \mid (\beta \rightarrow \beta') \wedge (\{h(EKB, \alpha \wedge \beta') \cup h(EKB, \neg\alpha \wedge \beta')\} \neq \emptyset)\}$$

Some elements of the set $G(Prob^{II}(\alpha|\beta))$ when substituted for $\langle\alpha, \beta\rangle$ in a failed probability term $Prob^{II}(\alpha|\beta)$ are more likely to result in the derivation of reasonable approximations than others. As discussed in Chapter 1, and in the introduction to this Chapter, this thesis assumes that these elements are among the most-specific generalisations.

The most-specific generalisations in the set $G(Prob^{II}(\alpha|\beta))$ can be identified by adopting a non-monotonic specificity assumption such as those discussed in Chapter 2 (e.g., [Bac90]; [Eth87]; [Bou92]; [Poo91]). A notion of specificity follows naturally from using implication to obtain the generalisations of a reference class.

$$\begin{array}{ll}\{ & \langle 50, \neg R(l_i, flies, moves) \wedge R(l_i, large, size) \wedge R(l_i, Scots, Race)\rangle, \\ & \langle 50, \neg R(l_j, flies, moves) \wedge R(l_j, Gaelic, lang.) \wedge R(l_j, Scots, Race)\rangle, \\ & \langle 1000, R(l_k, bird, species) \wedge R(l_k, flies, moves)\rangle, \\ & \langle 1, R(l_h, bird, species) \wedge \neg R(l_h, flies, moves)\rangle \qquad\qquad\qquad\qquad \}\end{array}$$

Figure 4.3: An EKB containing observations of Scotsmen and birds.

Definition 24 *The set $S(Prob^{II}(\alpha|\beta))$ of intensions of the most specific adequate semantic generalisations is:*

$$\left\{ \langle \alpha, \beta' \rangle \;\middle|\; \begin{array}{l} [\beta \to \beta']\; and\; [\langle \alpha, \beta' \rangle \in G(Prob^{II}(\alpha|\beta))] \\ and\, \forall(\langle \alpha, \gamma \rangle \in G(Prob^{II}(\alpha|\beta)))[(\beta \to \gamma) \to (\gamma \not\to \beta')] \end{array} \right\}$$

Unfortunately, there are two difficulties with defining the set Θ of all epistemologically relevant approximations to a failed probability term $Prob^{II}(\alpha|\beta)$ in terms of $S(Prob^{II}(\alpha|\beta))$. First, the set $S(Prob^{II}(\alpha|\beta))$ may be large. For example,

Example 35 According to our definition $S(Prob^{II}(\alpha|\beta))$ contains any $\langle \alpha, \beta' \rangle$ such that $\beta \to \beta'$. One way of generating a wfs β' is to simply disjoin a wfs to β. If the EKB is large, then a large number of these disjunctions may correspond to reference class extensions that are not empty.

Second, the elements of $S(Prob^{II}(\alpha|\beta))$ may be inappropriate in the context of approximating $Prob^{II}(\alpha \mid \beta)$. For example,

Example 36 Suppose the EKB is defined as in Figure 4.3, i.e., it consists of 50 observations of large Scotsmen that do not fly, 50 observations of Gaelic speaking Scotsmen who do not fly, 1000 birds that do fly, and 1 bird that does not. The probability term

$$Prob^{II}(flies|large \wedge Scots \wedge Gaelic)$$

fails as

$$\left\{ h\left(EKB, \begin{array}{c} flies \wedge large \wedge \\ scots \wedge Gaelic \end{array} \right) \bigcup h\left(EKB, \begin{array}{c} \neg flies \wedge large \wedge \\ scots \wedge Gaelic \end{array} \right) \right\} = \emptyset$$

As both large Scotsmen and Gaelic Scotsmen do not fly, it is reasonable to expect that large Scotsmen speaking Gaelic do not fly as well. However, using semantic generalisation, the set

$$S(Prob^{II}(flies|large \wedge Scots \wedge Gaelic))$$

contains the intension

$$\langle flies, (Gaelic \wedge Scots) \vee bird \rangle$$

which allows us to use the probability term

$$Prob^{II}(flies|bird \vee (Scots \wedge Gaelic))$$

to derive the counter intuitive approximation that the probability of large, Gaelic speaking Scotsmen flying is very high, i.e., the probability is $\frac{1000}{1001}$.

Example 37 The reference class of the probability term

$$Prob^{II}(rich|Queens\, graduate \wedge lawyer)$$

is generalisable by arbitrarily disjoining wfss to *Queens* $\wedge$ *lawyer* to include knowledge about the financial status of 'dwarf elephants', 'dead socialists', or 'U.B.C. graduates'. Considering knowledge about the financial status of 'U.B.C. graduates' when estimating the probability of 'Queens graduates' being 'rich' appears particularly suspect.

These two examples demonstrate that semantic generalisation, when defined by either subset containment over extensions or logical implication over intensions, is inappropriate when used to derive approximations of $Prob^{II}(\alpha|\beta)$.

4.2.2 Syntactic generalisation

In the previous section the set $S(Prob^{II}(\alpha|\beta))$ of most specific, adequate semantic generalisations of the reference class of $Prob^{II}(\alpha|\beta)$ can be generated by applying an operator to β that generates all those wfss logically implied by β. As seen in the previous section such an operator will generate inappropriate generalisations of the reference class of a failed probability term. In this section an alternative, syntactic operator, $\succeq$, is defined. Starting with a failed probability term, $Prob^{II}(\alpha|\beta)$, $\succeq$ generates a subset, $S^{\succeq}(Prob^{II}(\alpha|\beta))$, of $S(Prob^{II}(\alpha|\beta))$, that excludes, in particular, any intensions obtained by disjoining arbitrary wfss to β. Given a failed probability term $Prob^{II}(\alpha|\beta)$, the set $S^{\succeq}(Prob^{II}(\alpha|\beta))$ contains reference class intensions generated by 'generalising' the knowledge described by β.

Ignoring domain knowledge

Semantic generalisation is problematic because it allows us to generalise by considering knowledge that is not known to be true in the situation of interest. For example, we can generalise the reference class of "red birds" to "red birds or dead dogs" without knowing whether or not knowledge about "dead dogs" is appropriate to the situation of interest.

Syntactic generalisation only generalises what is known about the situation of interest, i.e., it only generalises knowledge described in β. For example, if "red" and "bird" are all that is known to be true, then the only generalisations allowed are those that can be obtained by generalising "red" and "bird". There are a number of ways of generalising β. For example,

Example 38 The wfs *red* might be generalised to similar colours to get "*red or orange or yellow*". In the context of making a prediction about a red object we might consider all objects that are red, yellow or orange as opposed to just *red*. The wfs *New York* might be generalised to "*New York or Big Apple*" as both names are often used to denote the same domain object.

The difficulty with generalising what is known about a new experience is that domain knowledge not readily available in an EKB is often required. For example,

Example 39 Generalising the wfs *red* to '*orange or red or yellow*' requires domain knowledge that says that *orange* and *yellow* are similar to *red* in some way and that this similarity is appropriate in the context of the particular situation of interest. For example, generalising *red* to '*orange* or *yellow*' may be inappropriate in the context of making predictions about whether or not a 'red sign' is a "stop sign". Similarly, generalising *New York* to '*New York* or the *Big Apple*' requires us to know something about the semantic equivalence of names.

This requirement clearly violates Assumption 1 of this thesis that the only available domain knowledge is that contained in an EKB.

I argue that a good way of generalising knowledge about a new experience is to *ignore* it. In the context of generalisation, knowledge is ignored by not incorporating it into the membership criteria of the generalised reference class. For example,

Example 40 Suppose we wish to obtain an approximation of the failed probability term

$$Prob^{II}(flies|red \wedge bird)$$

All that is known about the situation of interest is that it concerns a red bird. By ignoring what we know we can generalise. For example, by ignoring the fact that the object of interest is red, $red \wedge bird$ can be generalised to $bird$. The probability of $flies$ can now be approximated by $Prob^{II}(flies|bird)$ The reference class can be generalised even further by by ignoring the fact that the object of interest is a bird, obtaining the approximation $Prob^{II}(flies|\top)$

Generalising by "ignoring", is appropriate in the context of Assumption 1 of this thesis because it does not require any domain knowledge other than that readily found in the EKB.

Defining $\succeq$

The operator $\succeq$ is defined in terms of a minimal sub-ordering $\zeta \subseteq \succeq$ such that $\succeq$ is the reflexive, transitive closure ζ^*. In this case a necessary condition on ζ is that if $(\alpha_i, \alpha_j) \in \zeta$, then $h(EKB, \alpha_i) \supseteq h(EKB, \alpha_j)$. In this section I describe how ζ is used to generate the intensions of reference class generalisations in increasing order of generality.

In order to generalise the reference class of the probability term $Prob^{II}(\alpha|\beta)$ we generate the set

$$\{\langle\alpha, \beta'\rangle | (\beta', \beta) \in \zeta\}$$

of intensions of the most-specific generalisations. It is important to note that the set need not be the set of most-specific *adequate* generalisations. For example, every element of $\{Prob^{II}(\alpha|\beta') | (\beta', \beta) \in \zeta\}$ may fail.

$(\beta', \beta) \in \zeta$, if β' is obtained by *ignoring* a property predicate in β denoting the domain property, that is

Definition 25 *$\zeta(\beta', \beta)$ is defined as follows:*

$$\begin{array}{ll}
(\beta, \gamma_i \wedge \beta) \; and \; (\beta, \beta \wedge \gamma_i) \in \zeta & \\
(\alpha \wedge \beta', \alpha \wedge \beta) \in \zeta & if \;\; (\beta', \beta) \in \zeta \\
(\beta' \wedge \alpha, \beta \wedge \alpha) \in \zeta & if \;\; (\beta', \beta) \in \zeta \\
(\alpha \vee \beta', \alpha \vee \beta) \in \zeta & if \;\; (\beta', \beta) \in \zeta \\
(\beta' \vee \alpha, \beta \vee \alpha) \in \zeta & if \;\; (\beta', \beta) \in \zeta \\
(\top, \beta) \in \zeta & if \;\; (\neg\exists\beta')[\top \rightarrow \beta' \; and \; (\beta', \beta) \in \zeta]
\end{array}$$

Note, that a domain property can only be generalised by ignoring a property predicate if that property predicate is conjoined to others. The only generalisation of a domain property represented by a single property predicate is $\top$.

The reflexive transitive closure ζ^*, $\succeq$, can be specified succinctly in terms of the program written in pseudo Prolog in Figure 4.4. Given a wfs β, the first iteration

```
$gen((\alpha \wedge \beta), (\gamma \wedge \beta)) \leftarrow$
        $gen(\alpha, \gamma)$
$gen((\alpha \wedge \beta), (\alpha \wedge \gamma)) \leftarrow$
        $gen(\beta \wedge \gamma)$
$gen((\alpha \wedge \beta), \alpha) \leftarrow$
        $\neg gen(\beta, \delta)$
$gen((\alpha \wedge \beta), \beta) \leftarrow$
        $\neg gen(\alpha, \delta)$
$gen((\alpha \vee \beta), (\gamma \vee \beta)) \leftarrow$
        $gen(\alpha, \gamma)$
$gen((\alpha \vee \beta), (\alpha \vee \gamma)) \leftarrow$
        $gen(\beta, \gamma)$
```

Figure 4.4: A program in pseudo Prolog that partially defines the operator ζ for generating the intensions of the most specific syntactic generalisations of a reference class intension $\langle \alpha, \beta \rangle$.

of the program generates the set

$$\{\langle \alpha, \beta' \rangle \mid (\beta', \beta) \in \zeta\}$$

of intensions of the most specific generalisations of β. The second iteration generates the set of next most specific generalisations and so on. If the program can generate no generalisations of a wfs β, then the only generalisation is assumed to be $\top$.

Definition 26 *Let Γ be the set of all wfss in L. For any $\beta' \in \Gamma$, the set $G^{\succeq}(Prob^{II}(\alpha \mid \beta))$ of intensions of all adequate syntactic generalisations is*

$$\{\langle \alpha, \beta' \rangle \mid [(\beta', \beta) \in \succeq] \wedge (\{h(EKB, \alpha \wedge \beta') \cup h(EKB, \neg\alpha \wedge \beta')\} \neq \emptyset)\}$$

The set $S^{\succeq}(Prob^{II}(\alpha|\beta))$ of intensions of the most specific, adequate syntactic generalisations is defined as follows

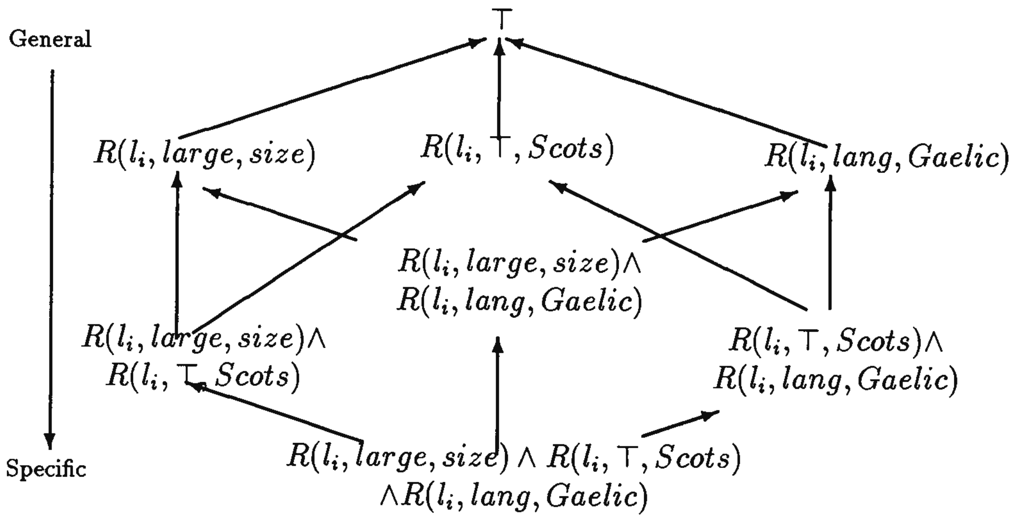
General
$\top$
$R(l_i, large, size)$
$R(l_i, \top, Scots)$
$R(l_i, lang, Gaelic)$
$R(l_i, large, size) \wedge$ $R(l_i, lang, Gaelic)$
$R(l_i, large, size) \wedge$ $R(l_i, \top, Scots)$
$R(l_i, \top, Scots) \wedge$ $R(l_i, lang, Gaelic)$
Specific
$R(l_i, large, size) \wedge R(l_i, \top, Scots)$ $\wedge R(l_i, lang, Gaelic)$

Figure 4.5: The wfss β' generated by applying $\succeq$ to *large*$\wedge$ *Scots*$\wedge$ *Gaelic.*

Definition 27 *The set* $S^{\succeq}(Prob^{II}(\alpha|\beta))$ *of intensions of the most specific adequate syntactic generalisations is:*

$$\left\{ \langle\alpha,\beta'\rangle \;\middle|\; \begin{array}{l} [\langle\alpha,\beta'\rangle \in G^{\succeq}(Prob^{II}(\alpha|\beta)] \textit{ and} \\ \forall\langle\alpha,\gamma\rangle \in G^{\succeq}(Prob^{II}(\alpha|\beta))[(\beta \rightarrow \gamma) \rightarrow (\gamma \not\rightarrow \beta')] \end{array} \right\}$$

Example 41 The wfss generated by applying $\succeq$ to $(large \wedge Scots \wedge Gaelic)$ are presented in Figure 4.5 organized in order of specificity. In the context of the probability term

$$Prob^{II}(flies|large \wedge Scots \wedge Gaelic)$$

and the EKB in Figure 4.3, the circled wfss in Figure 4.5 correspond to the wfss β' in the intensions $\langle flies, \beta'\rangle$ of the most specific adequate syntactic generalizations. As the β' are ordered by logical implication, every node that is logically implied by one of the circled nodes represents a less specific adequate generalization.

The most specific element in $G^{\succeq}(Prob^{II}(\alpha|\beta))$ is β, and the least specific element is $\top$. Given a failed probability term $Prob^{II}(\alpha|\beta)$, the approximation obtained by substituting the least specific generalization, $\top$, of β will succeed if the EKB contains any observations of the property α or its negation $\neg\alpha$. Thus, we can be assured, that if the EKB contains any knowledge about α, an approximation of $Prob^{II}(\alpha|\beta)$ can be obtained.

It is interesting to note that if β is a conjuct of property predicates such as $large \wedge Scots \wedge Gaelic$, then the result of applying $\succeq$ to β is a lattice with minimal element $\top$ and maximal element β as seen in Figure 4.5. The circled candidates in Figure 4.5 correspond to the minimal candidates in deKleer's model of fault diagnosis [dW83].

4.3 Selecting a best aproximation from Θ

The previous section demonstrated how semantic and syntactic generalization generate a set of intensions Θ of most-specific adequate generalizations of the reference class of a failed probability term $Prob^{II}(\alpha|\beta)$. In this section I consider the problem of approximating $Prob^{II}(\alpha|\beta)$ from the set

$$\{Prob^{II}(\alpha|\beta') \mid \langle\alpha, \beta'\rangle \in \Theta\}$$

Considering the most-specific adequate generalizations Θ may lead to conflicting approximations of a failed probability term. This occurs because Θ is generated without taking into consideration information about the relevance of the individual generalizations with respect to the prediction of interest [3]. I argue that this information can be readily extracted from an EKB and should be used to choose a single intension from Θ.

[3] The problem exists independently of whether semantic or syntactic generalization is used to generate Θ.

4.3.1 Conflicting approximations

If Θ contains the intensions of n adequate most specific generalizations, then in the worst case it is possible to derive n conflicting approximations. For example,

Example 42 Suppose we wish to interpret the probability term

$$Prob^{II}(R(l_i, flies, moves)|R(l_i, red, colour) \wedge R(l_i, bird, species))$$

with respect to the EKB

$$\begin{aligned}\{\quad &\langle 1, R(l_i, red, colour) \wedge R(l_i, \top, flies)\rangle,\\ &\langle 10, R(l_j, red, colour) \wedge \neg R(l_j, \top, flies)\rangle,\\ &\langle 10, R(l_k, bird, species) \wedge R(l_k, \top, flies)\rangle,\\ &\langle 1, R(l_l, bird, species) \wedge \neg R(l_l, \top, flies)\rangle \quad \}\end{aligned}$$

that is, an EKB containing 1 observation of a flying red object, 10 observations of non-flying red objects, 10 observations of flying birds, and 1 observation of a non-flying bird. The probability term fails as the EKB contains no observations of red birds. As

$$\zeta(R(l_i, red, colour) \wedge R(l_i, bird, species)) = R(l_i, red, colour),\ R(l_i, bird, species)\}$$

syntactic generalisation gives rise to two possible approximations of the failed probability term, i.e.,

$$\begin{aligned}&1.\ \ Prob^{II}(R(l_i, flies, moves)|R(l_i, red, colour)) &&= \tfrac{1}{11}\\ &2.\ \ Prob^{II}(R(l_i, flies, moves)|R(l_i, bird, species)) &&= \tfrac{10}{11}\end{aligned}$$

The first approximation suggests that the probability of a red bird flying is very low (0.09), while the second suggests that the probability of a red bird flying is very high (0.91).

If we have n conflicting approximations we might consider one of six possibilities:

1. Use a higher order probability to describe the n approximations [Jr.88d], or
2. Use the n approximations to place upper and lower bounds on the actual probability (e.g., [Jr.88d], [Goo91], [Bac90]), or
3. Average over the set of n approximations, or
4. Require all n approximations to be the same before deriving a conclusion (e.g., skeptical non-monotonic reasoning), or
5. Arbitrarily choose one of the n approximations (e.g., credulous non-monotonic reasoning), or
6. Use knowledge in the EKB to choose the most reasonable of the n approximations.

I use the following example to argue that the first five possibilities are inappropriate if everything that we know about the new experience is not equally relevant with respect to estimating a particular probability.

Example 43 Consider the problem of approximating

$$Prob^{II}(R(l_i, flies, moves)|R(l_i, red, colour) \wedge R(l_i, bird, species))$$

in the preceding example. Averaging over the two approximations obtained by syntactic generalisation we obtain 0.5 as an approximation of the probability of a red bird flying which is unsatisfactory in that it hides the divergent nature of the underlying probabilities. Arbitrarily choosing a single approximation from the candidates results in the wildly different approximations of $\frac{1}{11}$ or $\frac{10}{11}$, depending on which candidate is chosen. Requiring all the approximations to be the same leaves us in the same situation as we were with the failed probability term in the first place - ignorance. Finally, although bounding the approximation by the interval [0.09, 0.90] has some semantic merit, it does not tell us much more about whether or not red birds fly than the original probability term in which the probability was bounded by [0, 1].

I use the following example to argue that the sixth possibility is appropriate:

Example 44 Consider again the problem of approximating

$$Prob^{II}(R(l_i, flies, moves)|R(l_i, red, colour) \wedge R(l_i, bird, species))$$

Using information in the EKB we might decide that being a bird is much more relevant to predicting flying than being red. Subsequently we might approximate the failed probability term by

$$Prob^{II}(R(l_i, flies, moves)|R(l_i, bird, species))$$

obtaining a probability of $\frac{10}{11}$.

I now address the problem of choosing a single most relevant, most specific adequate generalisation.

4.3.2 Choosing a single approximation

In this section I argue that the problem of choosing a single most relevant, most specific adequate generalisation can be characterised as a problem of making the most reasonable assumption about the probabilistic independence of the knowledge that is generalised. I show that because probabilistic independence can not be directly measured in the context of a failed probability term, it is necessary to obtain an *estimate* of probabilistic independence using an appropriate inductive bias. The adoption of a particular bias, and thus the choice of a particular generalisation, is good only if the resulting probability term is an accurate approximation of the failed one.

In this section I discuss how inductive biases are applied to the problem of choosing a single most relevant, most specific generalisation. I do not argue for a particular inductive bias because I believe that the appropriateness of an inductive bias will vary depending upon the situation in which it is applied.

Probabilistic Independence

Given a failed probability term $Prob^{II}(\alpha|\beta)$, I argue that the appropriateness of a most-specific syntactic generalisation $Prob^{II}(\alpha|\beta')$ is a function of whether or not β and β' are independent with respect to predicting α.

Definition 28 *(***Independence, Pearl [Pea88]***) γ is independent of α given β, written $I(\alpha,\ \beta,\ \gamma)$, if*

$$Prob(\alpha|\beta \wedge \gamma) \;=\; Prob(\alpha|\beta)$$

If $I(\alpha, \beta, \gamma)$ is true, then the conditional probability $Prob(\alpha|\beta \wedge \gamma)$ is the same as the conditional probability $Prob(\alpha|\beta)$ and the generalisation $Prob(\alpha|\beta)$ is appropriate.

Example 45 If we know a-priori is that $R(l_i,\ red,\ colour)$ is independent of $R(l_i,\ flies,\ moves)$, then it is appropriate to generalise

$$Prob^{II}(R(l_i, flies, moves)|R(l_i, red, colour) \wedge R(l_i, bird, species))$$

by ignoring $R(l_i, red, colour)$ to approximate the failed probability term by

$$Prob^{II}(R(l_i, flies, moves)|R(l_i, bird, species))$$

The process of generating the set $S^{\succeq}(Prob^{II}(\alpha|\beta))$ of syntactic generalisations of the reference class of $Prob^{II}(\alpha|\beta)$ can be thought of as a process of making a series of independence assumptions. Whether or not an element of $S^{\succeq}(Prob^{II}(\alpha|\beta))$ will lead to a good approximation of $Prob^{II}(\alpha|\beta)$ depends on how reasonable the independence assumption was that was used to generate it. For example,

Example 46 The probability term

$$Prob(R(l_i, flies, moves)|R(l_i, Scots, Race) \wedge R(l_i, tongue, Gaelic))$$

is a reasonable approximation of

$$Prob\left(R(l_i, flies, moves)| \begin{array}{c} R(l_i, Scots, Race) \wedge R(l_i, tongue, Gaelic) \\ \wedge R(l_i, large, size) \end{array}\right)$$

only if it is reasonable to make the independence assumption

$$I\left(R(l_i, flies, moves), R(l_i, large, size), \begin{array}{c} R(l_i, Scots, race) \wedge \\ R(l_i, Gaelic, lang.) \end{array}\right)$$

I argue that a failed probability term should be approximated by choosing a single item from Θ obtained by making a most reasonable independence assumption.

Making reasonable independence assumptions

The problem with viewing syntactic generalisation as a process of making independence assumptions is that we do not have a-priori knowledge about independence assumptions. Nor, given a failed probability term, can we use our knowledge in the EKB to generate the assumptions. That is, if $Prob^{II}(\alpha|\beta)$ fails, there is no way of determining from the cases in the EKB whether or not $Prob^{II}(\alpha|\beta')$ is a reasonable generalisation because we can not tell whether or not $I(\alpha, \beta, \beta')$ is true. For example,

Example 47 Consider the failed probability term

$$Prob^{II}(R(l_i, flies, moves)|R(l_i, red, colour) \wedge R(l_i, bird, species))$$

As shown previously in Example 44, there are two possible approximations that can be obtained from the two most-specific adequate syntactic generalisations. However it is impossible to decide whether or not either of these have been obtained by ignoring an independent property because in order to see if either $R(l_i, red, colour)$ or $R(l_i, bird, species)$ are independent (according to Definition 28) we have to be able to calculate

$$Prob^{II}(R(l_i, flies, moves) | R(l_i, red, colour) \wedge R(l_i, bird, species))$$

which fails.

I argue that we can use inductive biases to measure the *reasonableness* of making independence assumptions of the form $I(\alpha, \beta, \gamma)$. These biases can be used to identify a single item in Θ that is the result of making the most reasonable independence assumption. For example, an inductive bias might state that it is more reasonable to assume

$$I(R(l_i, flies, moves), R(l_i, red, colour) \wedge R(l_i, bird, species), R(l_i, red, colour))$$

than

$$I(R(l_i, flies, moves), R(l_i, red, colour) \wedge R(l_i, bird, species), R(l_i, bird, species))$$

Although this thesis does not address the issue of finding the best inductive bias for measuring the reasonableness of independence assumptions, I suggest that statistical metrics that measure the *associativity* between two properties (e.g., [Edw76]) can be used to good effect. For example, given a feature f_i and a wfs α, the independence of the value of f_i with respect to predicting α might be measured by:

Context free dependence: (e.g., [Tur92] [MC90]) f_i has context free dependence for predicting α where there is a value v_j of f_j such that

$$Prob^{II}(\alpha | R(l_i, v_j, f_i)) \neq Prob^{II}(\alpha | \top)$$

Context free correlation: f_i has context free relevance for predicting α when the Pearson product moment correlation [Edw76] between the multi-valued feature f_i and property α is not 0:

$$r_{\langle f_i, \alpha \rangle} \neq 0$$

Context free clustering: A property γ has context free relevance for predicting α when the *clustering* metric (See Appendix C) between γ and α is not equal to 0.

In Chapter 5 I demonstrate that context free correlation and context free clustering both be used to make reasonable predictions in a machine learning domain.

Appendix B and Appendix C describe two different metrics of associativity that can be used to select a single most specific generalisation.

4.4 Extending syntactic generalisation

In the previous section syntactic generalisation is used to generate a single most relevant, most-specific, yet adequate, syntactic generalisation of the reference class of a failed probability term $Prob^{II}(\alpha|\beta)$. The difficulty with constraining semantic generalisation through the application of operators such as $\succeq$ is that there may not be much information about the situation of interest to generalise on. As a result, it may not be possible to generate a reasonable approximation of a failed probability term.

This section describes an extension of syntactic generalisation called chaining that allows us to consider additional information about the situation of interest during generalisation. The section starts by specifying chaining as a straightforward extension of syntactic generalisation. The section concludes by showing in certain circumstances that chaining a failed probability term $Prob^{II}(\alpha|\beta)$ is equivalent to generalising by disjoining a wfs to β. In Chapter 5 chaining is demonstrated to have practical applications when making predictions from incomplete EKBs.

4.4.1 Chaining

A failed probability term $Prob^{II}(\alpha|\beta)$ is *chained*, as opposed to syntactically generalised, in two steps. In step one additional knowledge about the domain state is conjoined to β. In step two the resulting expression is syntactically generalised.

Definition 29 *To chain $Prob^{II}(\alpha|\beta)$ on γ, $Prob^{II}(\alpha|\beta)$ is rewritten as*

$$\begin{aligned} Prob^{II}(\alpha|\beta) \approx\ & Prob^{II}(\alpha|\beta \wedge \gamma) \times Prob^{II}(\gamma|\beta) \\ & + Prob^{II}(\alpha|\beta \wedge \neg\gamma) \times Prob^{II}(\neg\gamma|\beta) \end{aligned}$$

followed by the syntactic generalisation of $Prob^{II}(\alpha|\beta \wedge \gamma)$ and $Prob^{II}(\alpha|\beta \wedge \neg\gamma)$.

It is important to note that the re-expression of a failed probability term $Prob^{II}(\alpha|\beta)$ in terms of γ by chaining is not a theorem of $Prob^{II}$ [4]. However, it is a theorem of probability theory. By defining chaining in this manner I explicitly assume that the re-expression will be a theorem of the underlying population from which the past experiences were obtained and will therefore be appropriate.

In the remainder of this section I consider a particular example of chaining in which the probability terms $Prob^{II}(\alpha|\gamma \wedge \beta)$ and $Prob^{II}(\alpha|\neg\alpha \wedge \beta)$ are syntactically generalised by ignoring β. That is, applying $\succeq$ to $Prob^{II}(\alpha|\beta \wedge \gamma)$ and $Prob^{II}(\alpha|\beta \wedge \neg\gamma)$ yields the generalisations $Prob^{II}(\alpha|\gamma)$ and $Prob^{II}(\alpha|\neg\gamma)$. $Prob^{II}(\alpha|\beta)$ can now be approximated by:

$$\begin{aligned} Prob^{II}(\alpha|\beta) \approx\ & Prob^{II}(\alpha|\gamma) \times Prob^{II}(\gamma|\beta) \\ & + Prob^{II}(\alpha|\neg\gamma) \times Prob^{II}(\neg\gamma|\beta) \end{aligned}$$

Generalisation by ignoring β makes intuitive sense in that if $Prob^{II}(\alpha|\beta)$ fails and there are no relevant generalisations of β, then β can be ignored. However, it is important to note that, in general, the two probability terms $Prob^{II}(\alpha|\beta \wedge \gamma)$ and $Prob^{II}(\alpha|\beta \wedge \neg\gamma)$ in the expression

$$Prob^{II}(\alpha|\beta) \approx Prob^{II}(\alpha|\beta \wedge \gamma) \times Prob^{II}(\gamma|\beta)$$

[4] If it were then there would be no need to chain or generalise.

$$
\begin{aligned}
\{ \quad & \langle 1, R(l_i, bird, species) \wedge \neg R(l_i, \top, feathers) \rangle \\
& \langle 1000, R(l_j, bird, species) \wedge R(l_j, \top, feathers) \rangle \\
& \langle 500, R(l_k, \top, feathers) \wedge R(l_k, \top, flies) \rangle \\
& \langle 50, R(l_l, \top, feathers) \wedge \neg R(l_l, \top, flies) \rangle \\
& \langle 10000, R(l_m, flying\,fish, species) \wedge \neg R(l_l, \top, flies) \rangle \quad \}
\end{aligned}
$$

Figure 4.6: An EKB of birds, flying fish and other flying feathered things.

$$+ Prob^{II}(\alpha|\beta \wedge \neg\gamma) \times Prob^{II}(\neg\gamma|\beta)$$

can be generalised by applying $\succeq$ to $\beta \wedge \gamma$ and $\beta \wedge \neg\gamma$.

Example 48 The probability term $Prob^{II}(flies|bird)$ fails with respect to the EKB in Figure 4.6. Syntactic generalisation approximates the desired probability by

$$Prob^{II}(flies|\top) = \frac{500}{10550}$$

allowing us to derive the counter-intuitive conclusion that the probability of birds flying is low. By chaining the failed probability term can be re-expressed as:

$$
\begin{aligned}
& Prob^{II}(flies|bird) \approx \\
& \quad Prob^{II}(flies|bird \wedge feathers) \times Prob^{II}(feathers|bird) \\
& \quad + Prob^{II}(flies|bird \wedge \neg feathers) \times Prob^{II}(\neg feathers|bird)
\end{aligned}
$$

As both $Prob^{II}(flies|bird \wedge feathers)$ and $Prob^{II}(flies|bird \wedge \neg feathers)$ fail they are generalised by making the independence assumption $I(flies, feathers, bird)$ to give:

$$
\begin{aligned}
& Prob^{II}(flies|bird) \approx \\
& \quad Prob^{II}(flies|feathers) \times Prob^{II}(feathers|bird)
\end{aligned}
$$

$$+\,Prob^{II}(flies|\neg feathers)\ \times\ Prob^{II}(\neg feathers|bird)$$

and we can conclude that the probability of the bird flying is 0.91.

Choosing a γ_i to chain on

The set $\{\gamma_1, \gamma_2, \ldots\}$ of wfs that can be conjoined with β to obtain an approximation

$$\begin{aligned} Prob^{II}(\alpha|\beta)\ \approx\ & Prob^{II}(\alpha|\gamma_i)\ \times\ Prob^{II}(\gamma_i|\beta) \\ & +\,Prob^{II}(\alpha|\neg\gamma_i)\ \times\ Prob^{II}(\neg\gamma_i|\beta) \end{aligned}$$

of a failed probability term is the set Γ of all wfss in the language L. Thus, chaining leads us to the consideration of large numbers of possibly irrelevant approximations, just as semantic generalisation does.

If every member γ_i of the set $\{\gamma_1, \gamma_2, \ldots\}$ satisfied the independence assumption

$$Prob^{II}(\alpha|\beta \wedge \gamma_i)\ =\ Prob^{II}(\alpha|\gamma_i)$$

then chaining by conjoining any member γ_i would result in the same approximation. That is, if β were truly independent of α given every γ_i, then any γ_i could be used to chain without worrying about the relevance of the approximation. Unfortunately, as discussed earlier we can only estimate whether or not the elements of $\{\gamma_1, \gamma_2, \ldots\}$ satisfy the independence assumption and these estimates are subject to error.

I argue that by requiring each element, γ_i, of $\{\gamma_1, \gamma_2, \ldots\}$ to satisfy certain constraints, the possibility of errors in estimating independence can be reduced. I start by arguing that in order for γ_i to be considered for chaining $Prob^{II}(\gamma_i|\beta)$ must not fail. If $Prob^{II}(\gamma_i|\beta)$ fails, then there is no way of determining whether or not γ_i is probable in the situation of interest. I argue further that not only should $Prob^{II}(\gamma_i|\beta)$ not fail but that

$$(Prob^{II}(\beta|\gamma_i)\ \approx\ 1)\ and\ (Prob^{II}(\gamma_i|\beta)\ \approx\ 1)$$

That is, the closer $Prob^{II}(\gamma_i|\beta)$ and $Prob^{II}(\beta|\gamma_i)$ are to 1 the more likely it is that γ_i is applicable to the situation of interest.

Consider, the requirement that $Prob^{II}(\gamma_i|\beta)$ be close to 1.

Example 49 Suppose the probability term $Prob^{II}(flies|bird)$ fails. If we consider the reference class of all birds we might find that with respect to our EKB most of them are known to also have feathers. That is,

$$Prob^{II}(feathers|bird) \approx 1$$

Using this information, we might conclude that if we knew more about the bird of interest in the failed probability term, then we would know that it has feathers. We might add this information to what we know to obtain the approximations

$$\begin{aligned} & Prob^{II}(flies|bird) \\ & \quad \approx Prob^{II}(flies|feathers) \times Prob^{II}(feathers|bird) \\ & \quad + Prob^{II}(flies|\neg feathers) \times Prob^{II}(\neg feathers|bird) \\ & \quad \approx Prob^{II}(flies|feathers) \times Prob^{II}(feathers|bird) \\ & \quad \approx Prob^{II}(flies|feathers) \end{aligned}$$

Intuitively, the approximation says that if all birds have feathers, then the probability of feathered objects flying can be used to approximate the probability of birds flying.

However, the requirement that $Prob^{II}(\gamma_i|\beta)$ be close to 1 is not enough. I argue that we also need $Prob^{II}(\beta|\gamma_i)$ to be close to 1. For example,

Example 50 Suppose the probability term $Prob^{II}(webbed\,feet|duck)$ fails. If we consider the reference class of all ducks we might find that with respect to our EKB

1. $Prob^{II}(bird|duck) \approx 1$ *and* $Prob^{II}(duck|bird) \approx 0.2$
2. $Prob^{II}(quack|duck) \approx 1$ *and* $Prob^{II}(duck|quack) \approx 1$
3. $Prob^{II}(webbed\,feet|bird) \approx 0.4$ *and* $Prob^{II}(webbed\,feet|quack) \approx 1$

Using this information we might obtain the approximations

$$\begin{aligned} & Prob^{II}(webbed\,feet|duck) \\ & \quad \approx Prob^{II}(webbed\,feet|bird) \times Prob^{II}(bird|duck) \\ & \quad + Prob^{II}(webbed\,feet|\neg bird) \times Prob^{II}(\neg bird|duck) \\ & \quad \approx Prob^{II}(webbed\,feet|bird) \times Prob^{II}(bird|duck) \\ & \quad \approx Prob^{II}(webbed\,feet|bird) = 0.4 \end{aligned}$$

and

$$\begin{aligned} & Prob^{II}(webbed\,feet|duck) \\ & \quad \approx Prob^{II}(webbed\,feet|quack) \times Prob^{II}(quack|duck) \\ & \quad + Prob^{II}(webbed\,feet|\neg quack) \times Prob^{II}(\neg quack|duck) \\ & \quad \approx Prob^{II}(webbed\,feet|quack) \times Prob^{II}(quack|duck) \\ & \quad \approx Prob^{II}(webbed\,feet|quack) \approx 1 \end{aligned}$$

Intuitively, the first approximation is unreasonable because we are using the reference class of birds to approximate the probability of ducks having webbed feet and most birds are not ducks. The second approximation is more reasonable because most birds that quack are ducks.

Chaining and semantic generalisation

If $Prob^{II}(\gamma_i|\beta) \approx 1$, then the chaining in the previous examples is equivalent to the semantic generalisation of $Prob^{II}(\alpha|\beta)$ by disjoining the wfs γ_i with β.

Theorem 5 *Suppose the probability term $Prob^{II}(\alpha|\beta)$ fails and that it is chained by γ and generalised by ignoring β, i.e.,*

$$Prob^{II}(\alpha|\beta) \approx$$

$$\begin{aligned} & Prob^{II}(\alpha|\gamma) \times Prob^{II}(\gamma|\beta) \\ & + Prob^{II}(\alpha|\neg\gamma) \times Prob^{II}(\neg\gamma|\beta) \end{aligned}$$

If $Prob^{II}(\gamma|\beta) \approx 1$, then by Theorem 4, Chapter 3, the previous expression can be rewritten as

$$Prob^{II}(\alpha|\beta) \approx Prob^{II}(\alpha|\gamma)$$

The reference class of $Prob^{II}(\alpha|\gamma)$ is now the same reference class as the reference class of the generalisation $Prob^{II}(\alpha|\beta \vee \gamma)$ obtained by disjoining β with γ.

Proof: As $Prob^{II}(\alpha|\beta)$ fails,

$$h(EKB, \alpha \wedge \beta) \bigcup h(EKB, \neg\alpha \wedge \beta) = \emptyset$$

It follows that

$$(h(EKB, \alpha \wedge \beta) = \emptyset) \wedge (h(EKB, \neg\alpha \wedge \beta) = \emptyset)$$

The reference class extension of the generalisation $Prob^{II}(\alpha|\beta \vee \gamma)$ obtained by disjoining γ with β is

$$\begin{aligned} &= h(EKB, \alpha \wedge (\gamma \vee \beta)) \bigcup h(EKB, \neg\alpha \wedge (\gamma \vee \beta)) \\ &= h(EKB, (\alpha \wedge \gamma)) \vee (\alpha \wedge \beta)) \bigcup h(EKB, (\neg\alpha \wedge \gamma) \vee (\neg\alpha \wedge \beta)) \\ &= h(EKB, (\alpha \wedge \gamma)) \bigcup h(EKB, (\alpha \wedge \beta)) \\ &\quad \bigcup h(EKB, (\neg\alpha \wedge \gamma)) \bigcup h(EKB, (\neg\alpha \wedge \beta)) \\ &= h(EKB, (\alpha \wedge \gamma)) \bigcup \emptyset \bigcup h(EKB, (\neg\alpha \wedge \gamma)) \bigcup \emptyset \\ &= h(EKB, (\alpha \wedge \gamma)) \bigcup h(EKB, (\neg\alpha \wedge \gamma)) \end{aligned}$$

which is the reference class extension of $Prob^{II}(\alpha|\gamma)$. □

As chaining can be shown, in certain circumstances, to be equivalent to semantic generalisation by disjunction, I suggest that only considering the most specific adequate *chainings* may lead to reasonable approximations, i.e., those obtained by disjoining the most specific γ_i. In Chapter 5 I show that choosing the most specific adequate chaining such that

$$\frac{Prob^{II}(\gamma_i|\beta) + Prob^{II}(\beta|\gamma_i)}{2}$$

is closest to 1 results in reasonable predictions.

4.5 Discussion

The problem of estimating a conditional probability $Prob(\alpha|\beta)$ is characterised in this thesis as a problem of identifying an adequate reference class. The problem is solved in three steps:

1. Specify an initial reference class.
2. Identify an appropriate alternative reference classes if the extension of the initial one is empty.
3. Aggregate over the members of the reference class extension to calculate an approximation of the desired probability.

Chapter 3 discusses steps 1 and 3. The current chapter discusses two techniques for identifying alternatives to the reference class of a *failed* probability term: Generalisation and Chaining.

In this chapter I argue that generalisation and chaining should be used to identify a single most-relevant, most-specific adequate syntactic generalisation of the reference class of a failed probability term. I argue that this generalisation is the most likely to result in a good approximation of a failed probability term. I also argue that generalisation must be constrained. In particular, I argue that semantic generalisation will result in unintuitive approximations. By means of example, I argue for a constrained syntactic form of generalisation that only generalises what is known about a situation of interest. I then show that in certain circumstances, chaining should be used to relax this constraint.

In Chapter 5 I support the arguments made in this chapter with empirical evidence resulting from applying the syntactic generalisation and chaining techniques to a prediction problem.

Chapter 5

Testing the RCA

In this chapter a computational implementation called FRED [1] uses the RCA's estimates of conditional probabilities to perform a predictive task. Given an EKB, and a case $\alpha \wedge \beta$, FRED calculates $\mathit{Pred}(EKB,\ \alpha \wedge \beta)$ such that

$$\mathit{Pred}(EKB, \alpha \wedge \beta) = \begin{cases} 1 & \mathit{if}\ \mathit{Prob}^{II}(\alpha|\beta) > \mathit{Prob}^{II}(\neg\alpha|\beta) \\ 0 & \mathit{if}\ \mathit{Prob}^{II}(\alpha|\beta) < \mathit{Prob}^{II}(\neg\alpha|\beta) \\ 0.5 & \mathit{if}\ \mathit{Prob}^{II}(\alpha|\beta) = \mathit{Prob}^{II}(\neg\alpha|\beta) \end{cases}$$

Intuitively, if $\mathit{Pred}(EKB,\ \alpha \wedge \beta) = 1$, then FRED's prediction is correct. If $\mathit{Pred}(EKB,\ \alpha \wedge \beta) = 0$, then FRED's prediction is incorrect.

The predictive task is a useful way of validating the RCA. This thesis assumes that if the RCA's estimates of conditional probabilities allow FRED to make correct predictions, then the techniques used by the RCA to obtain those estimates are reasonable. The particular predictive task used in this chapter has the additional advantage that the ability to make correct predictions is a common metric for comparing inductive techniques with otherwise distinct theoretical foundations. In this chapter, FRED is applied to the predictive task in three different experiments. It is important to note that what is measured as success in the experiments is not FRED's ability to make a single correct prediction but rather FRED's ability to make correct predictions in the long run.

Experiment 1 tests the hypothesis that the reasonableness of the estimates obtained by the RCA using syntactic generalisation are a function of the metric used to estimate the reasonableness of the independence assumptions. Experiment 1 tests

[1] For Fred's Relational Experiential Database.

the hypothesis by contrasting FRED's performance using estimates obtained by two different versions of syntactic generalisation with the performance of a k-nearest neighbours algorithm. The versions of syntactic generalisation differ with respect to the statistical metric of associativity that is used to estimate the reasonableness of the independence assumptions made in order to syntactically generalise.

Experiment 1 has three main results. First, performance on the predictive task is shown to be a function of the size of the EKB. Second, FRED's performance is shown to compare favourably with the performance of an implementation of k-nearest neighbours. Third, the reasonableness of the estimates obtained using syntactic generalization is shown to be a function of the metric of associativity used to estimate the reasonableness of independence assumptions.

Experiment 2 tests the hypothesis that, using syntactic generalisation, the RCA makes reasonable estimates of conditional probabilities. Using the version of syntactic generalization that resulted in the most accurate predictions in Experiment 1, FRED is applied to the predictive task in seven different data sets. Experiment 2's results are consistent with Experiment 1's. In particular, FRED's predictive performance is shown to compare favourably with a variety of other inductive techniques.

Experiment 3 tests the hypothesis that the RCA should chain rather than syntactically generalise given a case $\alpha \wedge \beta$ such that: 1. β specifies little about the situation of interest, and 2. The EKB is incomplete with respect to estimating $Prob^{II}(\alpha|\beta)$. Experiment 2 tests the hypothesis by contrasting the performance of FRED using syntactic generalisation with FRED using chaining. Experiment 2 demonstrates that chaining is more appropriate than syntactic generalisation if the only adequate syntactic generalisation of $Prob^{II}(\alpha|\beta)$ is $Prob^{II}(\alpha|\top)$.

5.0.1 An overview

This section provides an overview of the data and algorithms used in experiments 1, 2 and 3. The section is structured as follows:

1. The data sets used in Experiments 1, 2, and 3 are briefly described. A more complete description is provided in Appendix E.

2. The versions of syntactic generalisation and chaining used by FRED, and a k-nearest neighbours algorithm are described.

5.0.2 The data sets

In experiments 1, 2 and 3, FRED uses the RCA's estimates of conditional probabilities to make predictions about seven different data bases: 1. The *Soybean data base*, 2. The *Fisher soybean data base*, 3. The *Breast cancer data base*, 4. The *1984 congressional voting data base*, 5. The *modified 1984 congressional voting data base*, 6. The *mushrooms data base*, and 7. The *LED 7 digit data base.* All seven data bases were obtained from the machine learning data base repository at the University of California at Irvine and were not modified for use in this thesis.

Each case in the seven data sets describes a single domain object in terms of a set of exclusive features. In six of the seven data sets the features have discrete values, that is, a finit number of values. In the breast cancer data set, four of the features have real or continuous values. In five of the seven data sets the values of some of the features are unknown. In the terminology of this thesis, the five data sets may be incomplete with respect to making some predictions. Each case in the seven data sets is also categorized, apriori, into two or more classes. Finally, with the exception of the LED data set, the amount of noise is unknown. In the LED data set there is a 10 percent probability that the value of any of the seven binary valued features has been reversed.

An overview of the properties of the seven data sets used in the experiments is provided in Table 5.1. In the table, column 1 states the abbreviated name of the data base, column 2 states the number of cases in the data base, column 3 states the number of classes that the cases are divided into, column 4 states the number

Data base	*Size*	*Classes*	*Features*	*Real*	*Discr.*	*Missing Values*
Breast	286	2	9	4	5	Yes
Votes	435	2	17	0	17	Yes
Votes1	435	2	16	0	16	Yes
Mush	8124	2	22	0	22	Yes
Fisher	40	4	35	0	35	No
Soya	541	14	35	0	35	Yes
LED	3000	10	7	0	7	No

Table 5.1: An overview of the data sets used in experiments 1, 2 and 3.

of features used to describe each case, column 5 states the number of feaures that have real or continuous values, column 6 states the number of attributes that have discrete or nominal values, and column 7 states whether or not the data base is potentially incomplete.

5.0.3 The implementation

This section briefly describes the version of k-nearest neighbours used in experiment 1, the versions of syntactic generalisation used by FRED in experiments 1, 2 and 3, and the version of chaining used by FRED in experiment 3.

Syntactic generalisation

In Experiment 1 FRED uses two different versions of syntactic generalisation. The versions differ with respect to whether a correlation or a clustering statistic is used to estimate the reasonableness of the independence assumptions made when generalising. A description of the correlation statistic is provided in Appendix B. A description of the clustering statistic is provided in Appendix C. Experiments 2 and 3 use only the clustering version of syntactic generalisation.

To make a prediction FRED first attempts to use the RCA to to estimate

$Prob^{II}(\alpha \mid \beta)$ without generalisation. Note that because

$$Prob^{II}(\alpha|\beta) \;=\; 1 \;-\; Prob^{II}(\neg\alpha|\beta)$$

is a theorem, FRED only has to interpret $Prob^{II}(\alpha|\beta)$ to perform the predictive task. If the interpretation fails, then FRED generates the set $S^{\succeq}(Prob^{II}(\alpha|\beta))$ of intensions of the most specific adequate syntactic generalisations of $Prob(\alpha \mid \beta)$.

Following Chapter 4, FRED generates $S^{\succeq}$ by making a series of independence assumptions. Using a statistical metric of associativity, FRED assigns to each element of $S^{\succeq}(Prob^{II}(\alpha|\beta))$ an estimate of the reasonableness of that independence assumption. The estimate is simply the sum of the correlations or clusterings between α and each of the remaining ungeneralised features whose values are specified in β. FRED approximates $Prob^{II}(\alpha|\beta)$ by choosing the generalisation obtained by making the most reasonable independence assumption. If there is more than one most reasonable generalisation, then FRED averages over each of them.

Chaining

FRED implements chaining as an extension of syntactic generalisation with clustering. Briefly, if $Prob^{II}(\alpha \,|\beta)$ fails, then $Prob^{II}(\alpha \,|\beta)$ is chained on γ to obtain

$$Prob^{II}(\alpha|\beta) \;\approx\; Prob^{II}(\alpha|\gamma) \times Prob^{II}(\gamma|\beta) \;+\; Prob^{II}(\alpha|\neg\gamma) \times Prob^{II}(\neg\gamma|\beta)$$

such that $Prob^{II}(\gamma|\beta)$ and $Prob^{II}(\beta|\gamma)$ both succeed and

$$\frac{Prob^{II}(\gamma|\beta) + Prob^{II}(\beta|\gamma)}{2}$$

is closest to 1. If there are n equal possibilities $\Gamma = \{\gamma_1 \ldots \gamma_n\}$, then FRED averages the result of chaining over each of

$$\{\gamma_i \in \Gamma|(\forall\gamma_j \in \Gamma)(\exists Y)[\gamma_i(X/Y) \rightarrow \gamma_j]\}$$

i.e., FRED averages over the most specific γ_i.

k-nearest neighbours

The version of k-nearest neighbours used in Experiment 1 uses the Hamming distance similarity metric described in Chapter 2 and a value of k equals 1. Informal experimentation on my part showed the metric and the value of k to result in the most reasonable predictions. A more complete description of k-nearest neighbours techniques can be found in Hand [Han82] [Han81], and Dasarathy [Das91].

5.1 Experiment 1

Experiment 1 considers the following situation:

> Given a case $\alpha \wedge \beta$, what is the propensity to correctly predict that α is true given that β is true?

Experiment 1 compares the performance of FRED on the Fisher and Soybean data bases using the two versions of syntactic generalisation described previously with the performance of the k-nearest neighbours implementation. I start by describing how the EKB and a set of *test cases* are selected from the data base of interest. I then describe a procedure for measuring the performance of each algorithm. I conclude by describing the performance of FRED and k-nearest neighbours on the two data sets.

5.1.1 The EKB and test cases

The EKB is obtained by randomly selecting N percent of the cases in each of the classes described in the data sets. The remaining cases are assigned to the test set. Each case in the test set can be expressed as $\alpha \wedge \beta$ such that α specifies the value of the class.

5.1.2 The procedure

The k-nearest neighbours implementation, with $k = 1$, and FRED, using the correlation and clustering versions of syntactic generalisation, compute:

$$\frac{\Sigma_{i=1}^{n}(Pred(EKB, \alpha_i, \beta_i)}{n}$$

using the n cases $\alpha_i \wedge \beta_i$ in the test set thirty times for each of N equals 10, 20, 30, 40, 50, 60, 70, 80, and 90 per-cent. Intuitively, the computed value represents the implementations propensity to make correct predictions.

The results

By averaging the results of the thirty runs for each value of N we can obtain a stable estimate of predictive performance given N. The averaged estimates for FRED, using the correlation and clustering versions of syntactic generalisation, and for the k-nearest neighbours implementation, are plotted in Figures 5.7 and 5.8 as a function of N. Estimates for the COBWEB algorithm reported in [Fis87] are also included as a further, informal, point of comparison. Statistically significant differences in the performance of the four algorithms are plotted as bold points. An informal comparison of the results of Experiment 1 with several other techniques is provided in the discussion of this chapter.

5.2 Experiment 2

Experiment 2 considers the same situation as experiment 1. Experiment 2 measures the performance of the best version of syntactic generalisation identified in Experiment 1 on five additional data bases: The breast cancer data base, the mushrooms data base, the LED data base, the votes data base and the modified votes data base.

With the exception of the breast cancer data base, the EKB and test cases are obtained from each of the data bases using the procedure outlined in experiment 1.

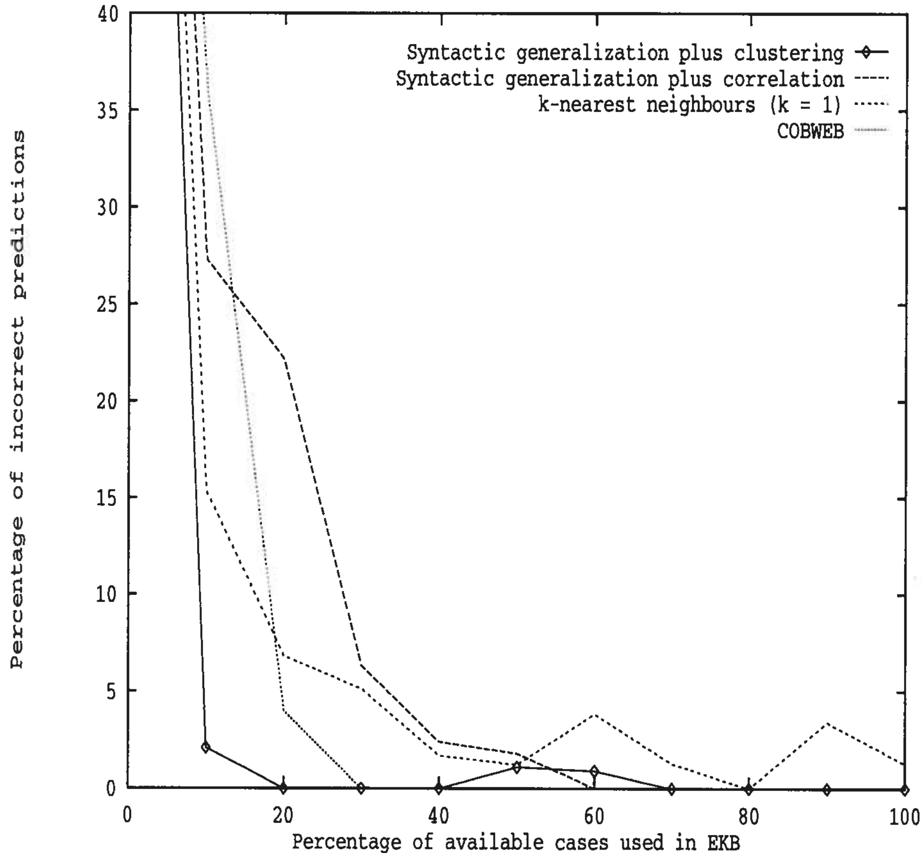

Figure 5.7: Percentage of incorrect predictions for the Fisher soybean data base as a function of N.

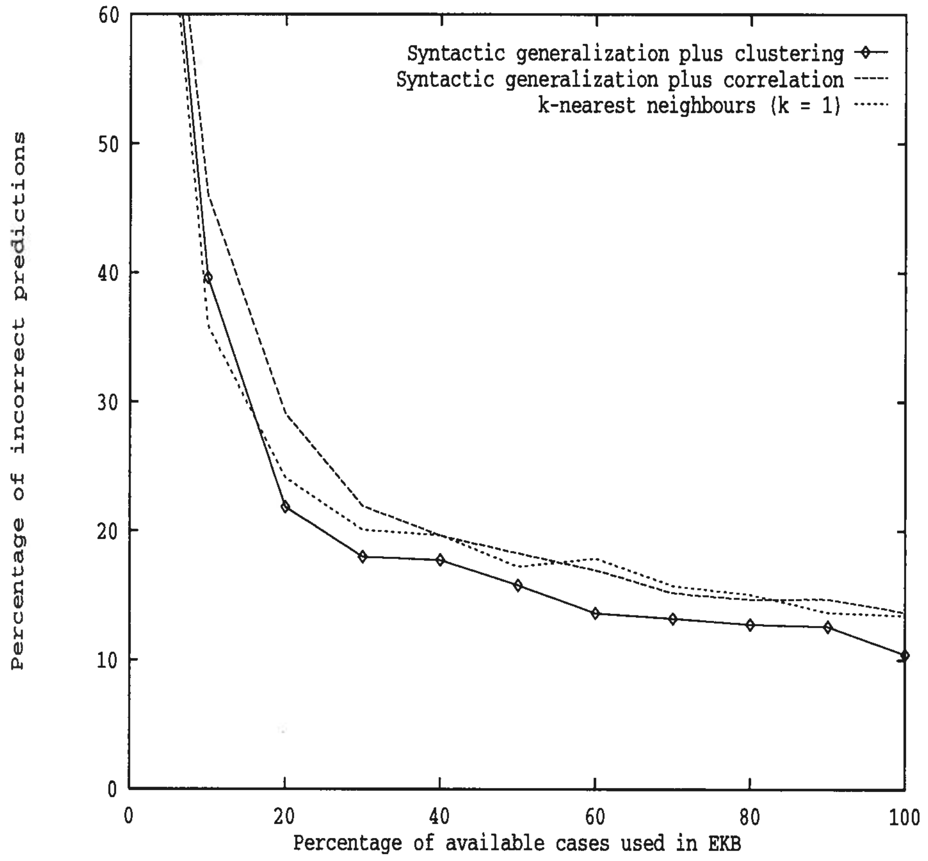

Figure 5.8: Percentage of incorrect predictions for the Soybean data base as a function of N.

Percentage Error	*Data base*	*Percentage Data in EKB*
27.4	Breast cancer	N = 70
5.3	Votes	N = 70
11.9	Votes (best attribute removed)	N = 70
0.0	Mushrooms	N < 2.5
0.0	Fisher soya bean	N = 10
30.7	LED (10 percent noise)	N < 10

Table 5.2: Percentage error using syntactic generalisation with clustering as a function of N for the five data sets examined in experiment 2.

The breast data base was pre-compiled by dividing the values of each of the real valued features into 5, equal lengthened, non-overlapping categories. No attempt was made to optimise the categorisation of the real valued features.

Figures 5.9, 5.12, 5.13, 5.10 and 5.11 plot the average of the thirty runs for each value of N as a function of the percentage error for each of the five data sets considered in Experiment 2. A brief summary of the results for each of the five data sets is presented in Table 5.2. An informal comparison of the results of Experiment 2 with other inductive techniques is provided in the discussion at the end of the chapter.

5.3 Experiment 3

In Chapter 4, I hypothesised that chaining might be more appropriate than syntactic generalisation given a probability term $Prob^{II}(\alpha|\beta)$ such that little is known about the new experience. Experiment 2 tests that hypothesis by applying syntactic generalisation and chaining in the following situation:

> Given a case $\alpha \wedge \beta$ and an EKB, what is the propensity to correctly predict $Prob^{II}(\alpha|\beta) > Prob^{II}(\neg\alpha|\beta)$ when the only adequate syntactic generalisation of $Prob^{II}(\alpha\ |\beta)$ is $Prob^{II}(\alpha\ |\top)$?

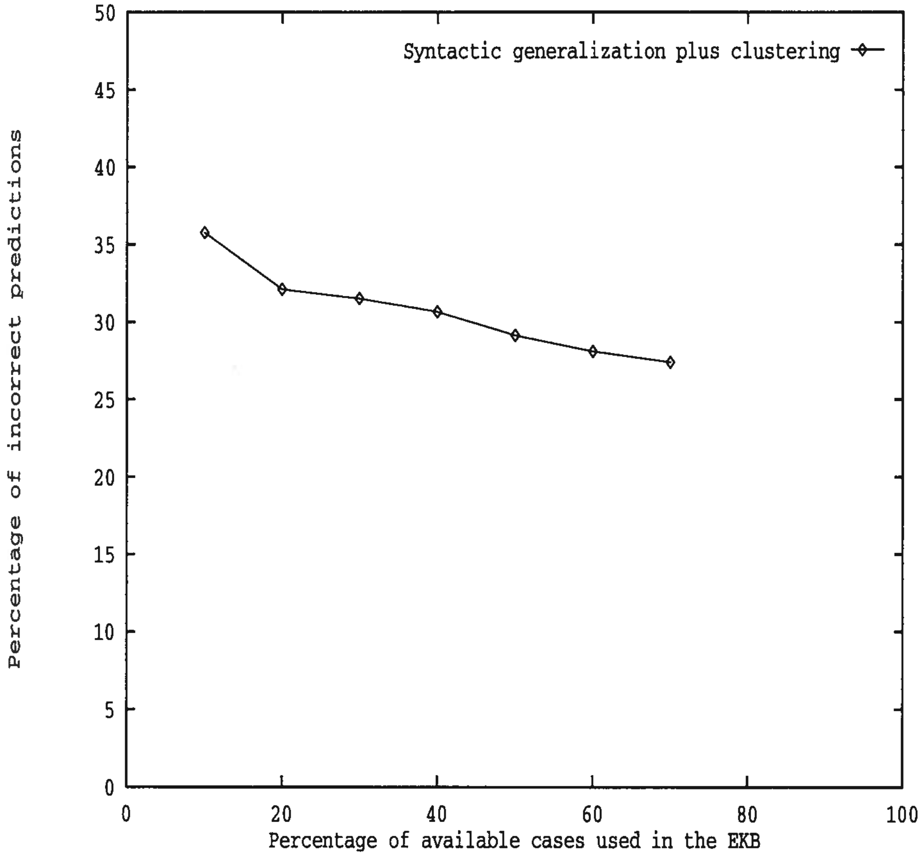

Figure 5.9: Percentage of incorrect predictions for the breast cancer data base as a function of N.

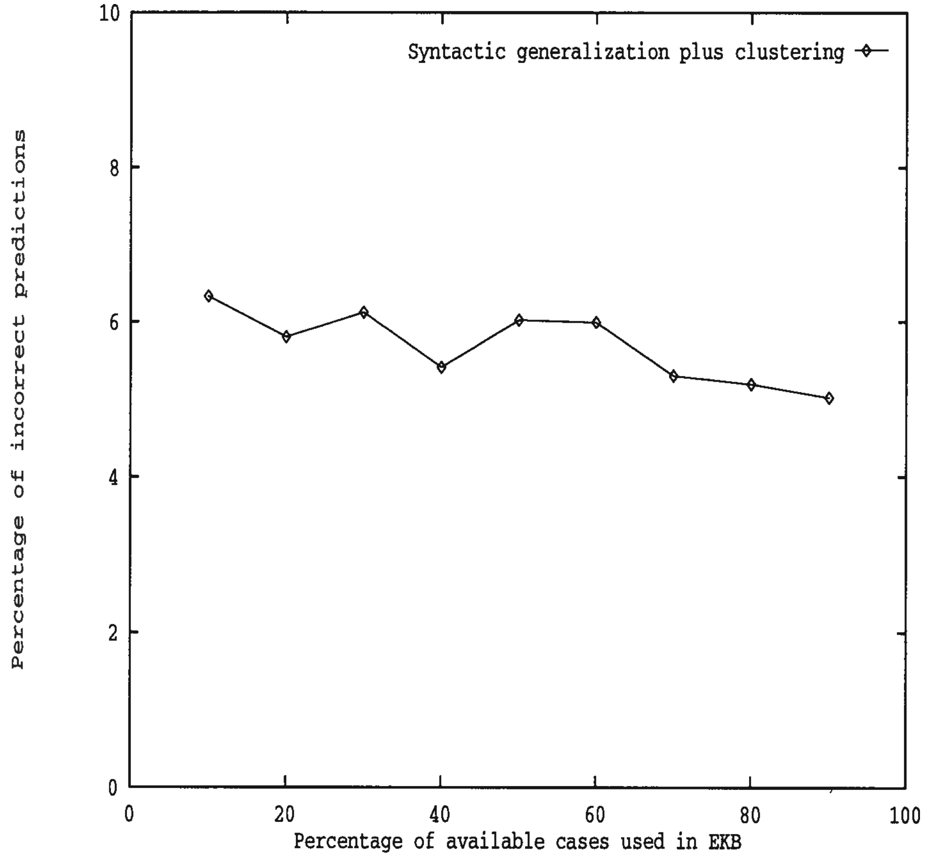

Figure 5.10: Percentage of incorrect predictions for the votes data base as a function of N.

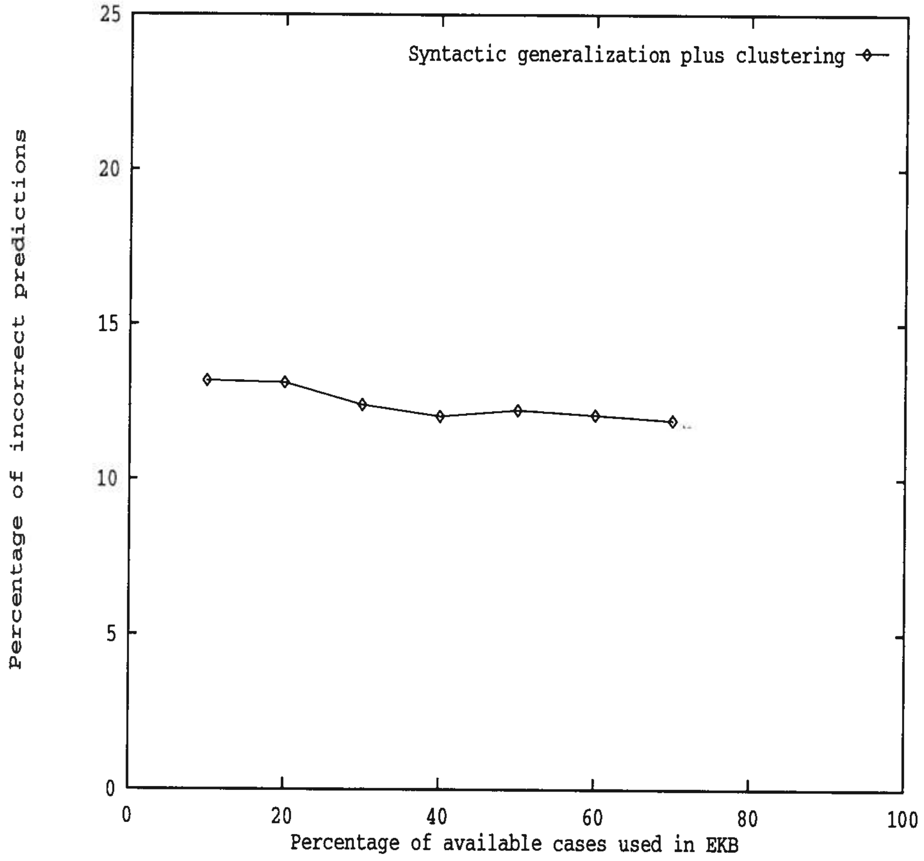

Figure 5.11: Percentage of incorrect predictions for the modified votes data base as a function of N.

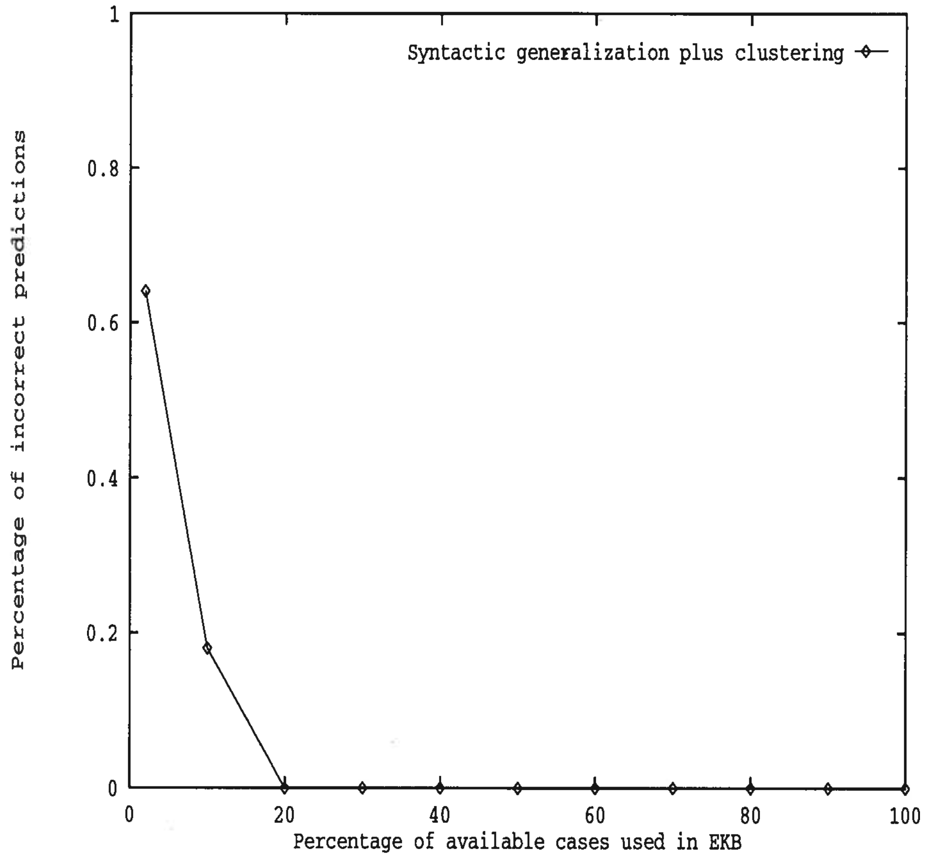

Figure 5.12: Percentage of incorrect predictions for the mushrooms data base as a function of N.

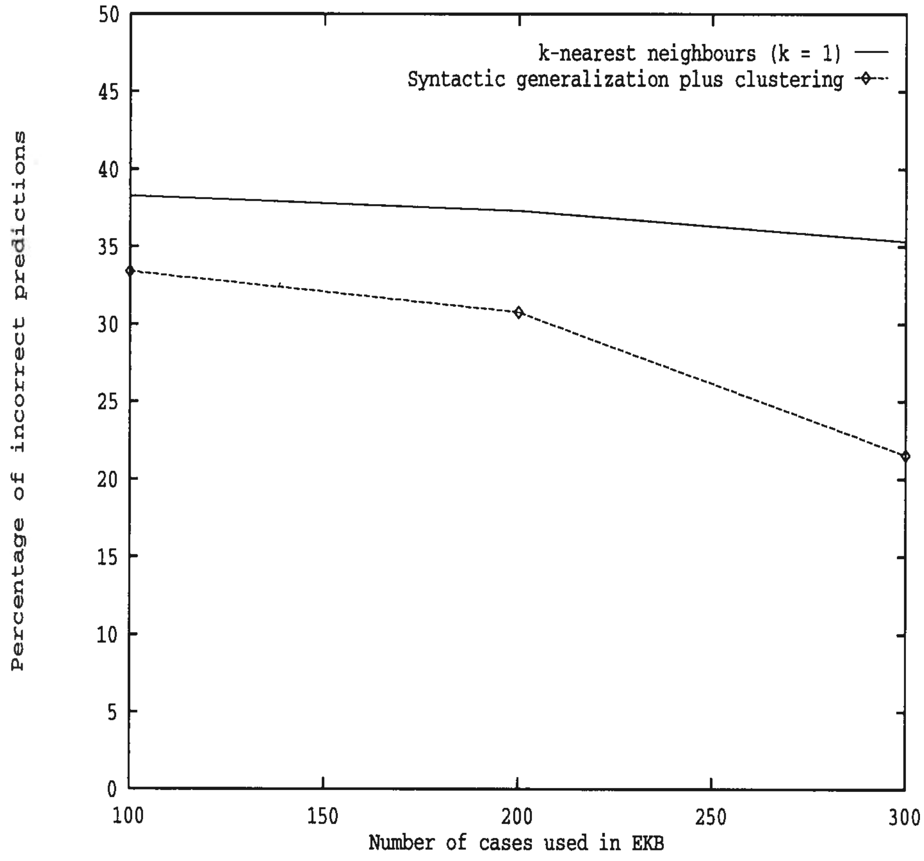

Figure 5.13: Percentage of incorrect predictions for the LED data base as a function of N.

I start by modifying the Fisher data base to construct an EKB and a test set in which the situation of interest can be examined. Using the EKB and the test set, I demonstrate that in situations in which the only adequate syntactic generalisation of a probability term $Prob^{II}(\alpha|\beta)$ is $Prob^{II}(\alpha|\top)$, chaining is more appropriate than syntactic generalisation.

5.3.1 The modified Fisher data base

In the Fisher data base each case $\beta \wedge S$ describes a soybean plant such that:

1. S specifies the values of 9 features that describe the stem of the diseased soya bean plant, and

2. β specifies the values of the 27 remaining features.

S can be written as a conjunct of nine property predicates

$$s_1 \wedge s_2 \wedge s_3 \wedge s_4 \wedge s_5 \wedge s_6 \wedge s_7 \wedge s_8 \wedge s_9$$

such that each property predicate s_i specifies the value of one of nine exclusive stem features: lodging, stem cankers, canker lesions, fruiting bodies, external decay, mycelium, discolouration, sclerotia, and fruit pods. By randomly dividing the nine stem features into three equal subsets, each case in the data base can be expressed as $\beta \wedge S^1 \wedge S^2 \wedge S^3$ such that

- S^1 specifies the values of the three features in the first subset.

- S^2 specifies the values of the three features in the second subset.

- S^3 specifies the values of the three features in the third subset.

I now use the Fisher data base to construct an incomplete EKB and a set of test cases such that for each case $\alpha \wedge \beta$ in the test case:

- If $Prob^{II}(\alpha|\beta)_{EKB}$ fails, then the only possible adequate syntactic generalisation is $Prob^{II}(\alpha|\top)$.

As in Experiment 1, N percent of the *Fisher* data base is placed in the EKB and the remainder in the test set. Every case in the EKB is replaced by three cases, each specifying the values of three of the nine stem features, i.e., each case $\beta \wedge S^1 \wedge S^2 \wedge S^3$ is replaced by

$$(\beta \wedge S^1),\ (\beta \wedge S^2)(X/Y^2),\ (\beta \wedge S^3)(X/Y^3)$$

such that Y^j is a tuple of labels unique to $\beta \wedge S^j$ in the EKB[2].

Each case in the EKB describes the stem of a diseased soybean plant in terms of 33 features. The only features common to all the cases in the EKB are those whose values are specified by β, e.g.,

Example 51 If a case in the EKB specifies the values of the features lodging, stem cankers, and canker lesions, then it does not specify the values of fruiting bodies, external decay, mycelium, discolouration, sclerotia, or fruit pods.

The procedure

In this section I measure the propensity of syntactic generalisation and chaining to correctly predict

$$Prob^{II}(\alpha|\beta) > Prob^{II}(\neg\alpha|\beta)$$

such that α specifies the value of one of the nine stem features, and β specifies the values of three of the eight remaining stem features.

Using *Syntactic generalisation plus the clustering metric*, and *Chaining with syntactic generalisation plus the clustering metric*, FRED computes

$$\frac{\Sigma_{k=1}^{n}(\Sigma_{i=1}^{6})(\Sigma_{j=1}^{3}(\,Pred(EKB, s_i, S^j \wedge \beta))))}{n \times 3 \times 6}$$

[2] The substitution is necessary if the set of new cases is to satisfy the definition of an EKB.

for all n cases $\beta \wedge S^1 \wedge S^2 \wedge S^3$ in the test set such that

$$\neg(S^j \rightarrow s_i) \wedge ((S^1 \wedge S^2 \wedge S^3) \rightarrow s_i)$$

The procedure is repeated ten times for each of N equals 20, 40, and 80 per cent, randomly selecting S^1, S^2 and S^3 each time.

Example 52 Consider the test case $\beta \wedge S^1 \wedge S^2 \wedge S^3$ such that

$$\begin{array}{lll} S^1 & is & R(l_i, \top, lodging) \wedge R(l_i, absent, cankers) \wedge R(l_i, tan, lesions), \\ S^2 & is & R(l_i, \top, fruiting) \wedge R(l_i, dry, decay) \wedge R(l_i, \top, mycelium), \ and \\ S^3 & is & R(l_i, black, colour) \wedge R(l_i, \top, sclerotia) \wedge R(l_i, diseased, pods). \end{array}$$

From the feature values specified in S^1, S^2 and S^3 we predict the values of the six other features. For example, one of the 18 possible predictions is

$$Prob^{II}\left(R(l_i, present, sclerotia) \middle| \begin{array}{c} R(l_i, \top, lodging) \wedge R(l_i, absent, cankers) \\ \wedge R(l_i, tan, lesions) \wedge \beta \end{array}\right)$$

For each $Prob^{II}(s_i|S^j \wedge \beta)$:

1. $S^j \wedge \beta$ only specifies the values of 3 out of a possible 36 features, i.e., it specifies little of the situation of interest.

2. $Prob^{II}(s_i|S^j \wedge \beta)$ and every syntactic generalization except for $Prob^{II}(s_i|\top)$ fails because there are no cases in the EKB that specify the value of the feature specified in s_i as well as the values of the features specified in S^j.

5.3.2 Results

The results of experiment 2 are plotted in Figure 5.14. As seen in the figure chaining offers a significant improvement over syntactic generalization. Of particular interest is the observation that as N increases the performance of syntactic generalization decreases and the performance of chaining increases.

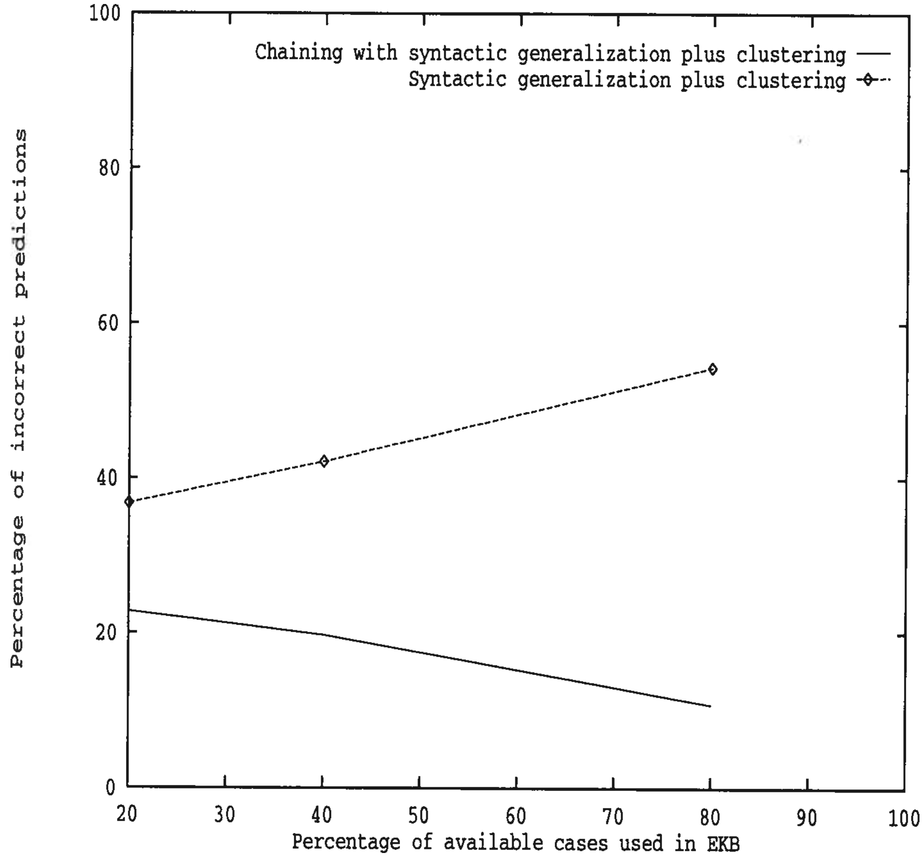

Figure 5.14: Percentage of incorrect predictions for the modified Fisher data base as a function of N for syntactic generalization and chaining.

5.4 Discussion

This chapter tests the hypothesis that the RCA described in Chapters 3 and 4 can be used to make reasonable predictions. In this section I discuss the results of three experiments that test that hypothesis. I conclude that:

1. The RCA can be used to make reasonable predictions.
2. The performance of syntactic generalization is a function of the measure of associativity used to estimate independence when generalizing.
3. Chaining results in more reasonable predictions than syntactic generalization in certain circumstances.

5.4.1 Experiment 1

Experiment 1 tests the hypothesis that the RCA can use syntactic generalization to obtain reasonable estimates of conditional probabilities. The hypothesis is tested by having FRED use the RCA's estimates to make predictions about two data sets. The propensity of FRED to make correct predictions using the RCA's estimates is compared with the propensity of an implementation of k-nearest neighours to make correct predictions.

In Figure 5.7 the propensity of FRED and k-nearest neighbours to make incorrect predictions about the Fisher data base declines as the percentage N of cases in the EKB increases. Indeed, both techniques perform very well as N approaches 100 percent. The result is consistent with the general observation that the performance of any reasonable algorithm approaches Bayes' optimum for a set of data as N increases (e.g. [FSK$^+$93]). This suggests that both techniques are reasonable with respect to performing the prediction task. When N is low there are statistically significant differences in the propensity of the four techniques to make incorrect predictions. I now discuss these differences.

The first noticeable difference is the difference between FRED's performance using the two different versions of syntactic generalization. As seen in Figure 5.7 syntactic generalization plus clustering performs significantly better than syntactic generalization plus correlation. The difference between the two versions of syntactic generalization demonstrates that not all estimates of relevance are equally good.

I suggest that the difference between the two versions of syntactic generalization is a consequence of the fact that correlation measures the relationship between features while clustering measures how well knowing the value of one feature predicts the value of another. The results plotted in Figure 5.7 for N equals 10, 20 and 30 suggest that the latter is more appropriate in the context of syntactic generalization than the former.

The second noticable feature of Figure 5.7 is the difference between syntactic generalization with clustering and k-nearest neighbours for N equals 10, 20 and 30. I hypothesize that the reason that the syntactic generalization algorithm with clustering does so well is that in the Fisher data base not all features are equally relevant with respect to predicting diagnostic category. Indeed, as noted in Appendix E knowing the values of very few features is sufficient for predicting the diagnostic category. The clustering statistic is very good at identifying these features so that the most relevant adequate syntactic generalization can be identified. In contrast, k-nearest neighbours treats all the features as equally relevant. As a result k-nearest neighbours would not be expected to be as good as syntactic generalization with clustering in the Fisher data base.

The third noticeable difference is the relatively poor performance of the COBWEB algorithm compared to k-nearest neighbours and syntactic generalization with clustering. The difference is a result of the fact that COBWEB is an unsupervised clustering algorithm whose primary function is to find good clusters. There is a tendency in the machine learning literature to make conclusions about inductive algorithms without considering their true nature. For example, even though the

RCA's estimates can be applied to the general problem of making predictions, it might be more appropriate to only compare the RCA only with inductive algorithms that also estimate conditional probabilities and use these estimates to make predictions.

With some exceptions, the results presented in Figure 5.8 replicate those in Figure 5.7. As in Figure 5.7, Figure 5.8 shows that the propensity of k-nearest neighours and FRED to make incorrect predictions decreases as N increases. However, unlike Figure 5.7, none of the differences between FRED using the clustering version of syntactic generalization and k-nearest neighbours are statistically significant.

The absence of a statistical difference between k-nearest neighbours and syntactic generalization is interesting when the results in Figure 5.7 are considered. One possible explanation for the difference in results is that the similarity metric used by k-nearest neighbours and the clustering and correlation statistics make different assumptions about the independence of feature values. For example, both the clustering and correlation statistics assume that feature values are independent. That is, the value of a feature f_i is independent of the value of another feature f_j. The k-nearest neighbours similarity metric used in this chapter does not assume that feature values are independent. An informal analysis of the Fisher data base indicates that the feature values necessary for predicting the diagnostic category are independent, while the values of the features in the Soybean data base tend to be dependent. As both the clustering and correlation statistics assume indendence they would be expected to only perform well when the assumption is justified. It might be possible to improve the performance of syntactic generalization on the soybean data base by using a statistical measures of association that does not assume that feature values are independent.

Of further interest is the fact that the performance of syntactic generalization with clustering is not significantly different from the performance of syntactic generalization with correlation. This result suggests that the choice of a particular statistical measure of association may not be critical with respect to making reasonable predictions. I tested this hypothesis by using random selection to choose the most reasonable, most specific generalization for $N = 100$. The propensity to make incorrect predictions averaged over 400 trials was 15 per-cent, compared to 10 per-cent for clustering, and 14 per-cent for correlation and k-nearest neighbours. The results of the informal study indicate that if the number of cases in the EKB is large, any metric for selecting a most reasonable, most specific generalization is effective.

5.4.2 Experiment 2

Experiment 2 has three main results. First, the experiment validates the hypothesis that the RCA's estimates obtained by syntactic generalization are reasonable. Second, the experiment demonstrates that FRED, using the clustering version of syntactic generalization, can make predictions that are reasonable when compared to existing techniques. Third, the experiment demonstrates that the RCA can obtain reasonable estimates in a variety of situations. For example, the results on the LED data base demonstrate that the estimates are reasonable when the data is noisy. The results on the breast cancer data base demonstrate that the estimates are reasonable when real valued features are used to describe experiences.

Tables 5.3, 5.4, 5.5, 5.6, 5.7, and 5.8 provide an informal comparsion of FRED's performance with existing techniques. The survey of results for the data bases using existing machine learning techniques is adapted from Holte [Hol93] and Clark and Niblett [BN92]. As noted in Holte, the comparison is necessarily informal because some of the results may have been obtained under slightly different experimental conditions. When the experimental condition is the same as that found in this thesis the results are presented in italics. All results obtained in this thesis are presented in bold.

Percentage Error	*Algorithm*
19.0	*IR* [Hol93]
2.5	*C4* (pruned) [Hol93]
1.3	**1 nearest neighbours**
0.0	COBWEB [Fis87]
0.0	**Syntactic Generalization with correlation**
0.0	**Syntactic Generalization with clustering**

Table 5.3: A survey of results for the "Fisher 1987 soya bean" data base.

Percentage Error	*Algorithm*
35.0	Bayes [CN87] [CN88]
34.7	Nearest neighbour [WK90]
32.0	AQ15 [Sal91]
31.7	*IR* [Hol93]
28.2	Bayes [WK90]
28.0	*C4 (pruned)* [Hol93]
27.7	*GINI decision tree* [BN92]
27.4	**Syntactic Generalization with clustering**
27	CN2 (unordered, laplace) [CB91]
26.7	ID3 (pruned) [Bun89]
23.9	Bayes [Bun89]
22.4	EACH with feature adjustment [Sal91]

Table 5.4: A survey of results for the breast cancer data base.

Percentage Error	*Algorithm*
16.0	3-nearest neighbours [BMMZ92]
14.0	1-nearest neighbour [BMMZ92]
13.8	*K-nearest neighbour* [AK89]
11.8	*Marsh* [BN92]
8.1	NTgrowth [AK89]
8.0	AQ15 (TRUNC-SG) [BMMZ92]
6.4	CN2 (ordered, entropy) [CB91]
5.3	**Syntactic Generalization with clustering**
5.2	CN2 (unordered, laplace) [CB91]
4.8	*IR* [Hol93]
4.4	*C4* (pruned) [Hol93]

Table 5.5: A survey of results for the votes data base.

Percentage Error	*Algorithm*
16.6	*Marsh* [BN92]
16.2	*IR* [Hol93]
13.0	*Information Gain* [BN92]
12.8	*GINI* [BN92]
11.9	**Syntactic Generalization with clustering**
10.6	*C4 (pruned)* [Hol93]

Table 5.6: A survey of results for the modified votes data base.

Percentage Error	*Algorithm*
7.3	*Marsh* [BN92]
5.0	IR [Hol93]
3.4	*GINI* [BN92]
3.4	*Information Gain* [BN92]
0.9	Neural Network [Yeu91]
0.0	C4 (pruned) [Hol93]
0.0	**Syntactic Generalization with clustering**

Table 5.7: A survey of results for the mushrooms data base.

Error Rate	*Algorithm*
33.2	Marsh [BN92]
32.9	Information Gain [BN92]
32.8	GINI [BN92]
30.7	**Syntactic Generalization with clustering**
26.3	Bayes optimum [Bun89]

Table 5.8: A survey of results for the LED data set with 10 percent added noise and $N = 6.7$ percent.

5.4.3 Experiment 3

In Chapter 4 I demonstrated by example that chaining is more appropriate than syntactic generalization if little is known about the situation of interest. Experiment 2 tests the hypothesis that chaining is more appropriate than syntactic generalization given:

1. A probability term $Prob^{II}(\alpha|\beta)$ such that β tells us little about the situation of interest, and

2. An incomplete EKB, i.e., one in which

$$(\exists\alpha,\beta)(|\beta| > |\alpha \wedge \beta| + |\neg\alpha \wedge \beta|)$$

Experiment 2 tests the hypothesis by comparing the propensity of FRED using chaining to make correct predictions with the propensity of FRED using syntactic generalization plus clustering to make correct predictions.

In Figure 5.14 the differences between chaining and syntactic generalization plus clustering are statistically significant for N equals 20, 40, and 80. These findings support the hypothesis that chaining is more appropriate than syntactic generalization in the situation tested.

Of additional interest is the trend seen in Figure 5.8 for the performance of syntactic generalization to decrease as N increases while the performance of chaining continues to increase. I hypothesize that the reason that the performance of syntactic generalization decreases is that predictions are made by generalizing over every case in the EKB. Given a probability term $Prob^{II}(\alpha|\beta)$, as N increases the cardinality $|\top|$ of the reference class increases quickly relative to the cardinality $|\alpha|$.

Chapter 6

Conclusions

6.1 Introduction

This thesis addresses the problem of designing computational agents that make predictions about the properties of a well defined class of objects in the context of the following methodological assumption:

Assumption 1 The only domain knowledge is a set of past experiences such that each past experience is described by a single ground sentence called a case.

In particular, this thesis discusses the RCA to the following induction problem:

> Given an EKB, predict whether or not a property α will be true of a new experience, given that all we know about the new experience is that β is true.

In Chapter 1 I argued that in order to solve the induction problem the RCA must address three issues:

Relevant reference class problem: How do we identify the relevant cases in the EKB for predicting α given β?

Adequate reference class problem: How do we predict α given β when there are only a few relevant cases in the EKB?

Inadequate reference class problem: How do we predict α given β when there are no relevant cases in the EKB?

The main contributions of this thesis are:

- A solution to the three problems in the context of Assumption 1.
- A description of a new form of generalisation called syntactic generalisation.
- A description of a novel extension of generalisation called chaining.
- A demonstration that syntactic generalisation and chaining in the context of the RCA can be used to make reasonable predictions from a set of cases.

6.1.1 Chapter outline

The remainder of this chapter is structured as follows:

1. I summarise the results of the thesis in the context of the literature reviewed in Chapter 2.
2. I discuss extensions to the RCA in the context of relaxing the explicit assumptions made in Chapter 1.

6.2 Thesis Summary

In this section I briefly review the results of Chapters 3, 4 and 5 in the context of the literature reviewed in Chapter 2.

6.2.1 Chapter 3: Describing experiences

Chapter 3 describes a language L for talking about experiences and making predictions from an EKB. In this section I discuss the: 1. Expressiveness of L, 2. Efficiency of retrieving wfss of L from an EKB, and 3. The interpretation of probability terms in L.

Expressiveness

The language L shares functional similarities to vector languages often used by classification algorithms and models of human episodic memory to describe past experiences. For example, a feature vector can be represented using L as a conjunct of property predicates such that each property predicate denotes the value of an exclusive feature.

However, L is significantly more expressive than a vector language. That is, L allows us to describe experiences and make predictions about situations that can not be expressed using a vector language. For example, L allows us to describe arbitrary n-ary relations. That is, it allows us to describe objects as having more than one value for a feature, or objects that are related to other objects. L also allows us to describe experiences using disjunction and negation in addition to conjunction. I argue that L's added expressiveness makes it a “natural” language for talking about experiences.

Efficiency

As mentioned in Chapter 3, the retrieval of cases from an EKB is NP-complete. However, the situations under which retrieval is efficient (i.e, computable by a polynomial time algorithm) are well documented (e.g., Borgida and Etherington [BE89], Crawford and Kuipers [CK89], Davis [Dav90], Etherington et. al. [EBBK89] [EKP90], Levesque [Lev89]). Of particular interest is the observation that we can expect the retrieval of cases to be efficient under exactly the same conditions that vivid reasoning [Lev88] is efficient.

If a probability term does not fail, then the RCA estimates probabilities by ‘looking up’ in an EKB the members of reference class extension. ‘Look up’ forms the basis for *efficient* reasoning in several recent computational models (e.g., [Lev86] [Lev89] [EBBK89] [Fri87] [Dav90] [Dav87] [SW86]). For example, given a successful probability term $Prob(\alpha|\beta)_{EKB}$ such that α denotes the value of a single exclusive feature, and β and all the cases in the EKB are conjuncts of property predicates, each denoting the value of a single exclusive feature, retrieval of the reference class extension from an EKB is sub-linear in the size of the EKB and linear in the number of property predicates, i.e., retrieval is similar to look-up from a relational data base as in vivid reasoning [EBBK89].

If generalisation or chaining are necessary, then making a prediction is considerably less efficient than simple data base look-up. However, the fact that chaining and generalisation are inefficient is not unreasonable. An inductive reasoner should be expected to make quick, accurate predictions about what it knows and slower predictions about what it doesn't.

Estimating conditional probabilities

Chapter 3 discusses the problem of estimating a conditional probability such as $Prob(\alpha|\beta)$. In agreement with the existing literature, the chapter shows how the intension and extension of the reference class of cases can be specified in terms of the wfss α and β. Chapter 3 shows how labels are used to retrieve the cases in the reference class extension from an EKB.

Chapter 3 also demonstrates that the appropriateness of a particular interpretation of a probability term depends upon the 'form' of the cases contained in the EKB. For example, as discussed in the introduction and various texts on empirical probability theory (e.g., [SM82] [Bar82]), we can estimate a conditional probability $Prob(\alpha|\beta)$ by

$$\frac{T_{\alpha\wedge\beta}}{T_\beta}$$

such that $T_{\alpha\wedge\beta}$ is the number of cases in which $\alpha \wedge \beta$ is true. Implicit in this interpretation is the assumption that if we have a case, then we know if α and β are true or false.

This assumption does not apply in the context of this thesis. There is no requirement that an agent collecting experiences has to describe each experience in terms of the same set of properties. That is, the agent may not *know* whether or not α and β are true or false. In chapter 3 I argue that if the cases in the EKB are incomplete, i.e., $N(\alpha, \beta) \neq 0$, and nothing is known about the incomplete cases, then the most reasonable estimate of $Prob(\alpha|\beta)$ is

$$Prob^{II}(\alpha|\beta) = \frac{K_{\alpha\wedge\beta}}{K_{\alpha\wedge\beta} + K_{\neg\alpha\wedge\beta}}$$

such that $K_\gamma = |\gamma|$ as defined in Chapter 3.

6.2.2 Chapter 4: Generalisation and chaining

Chapter 4 addresses the empty reference class problem. The Chapter demonstrates by means of example, that semantic generalisation is inappropriate. The Chapter argues that using generalisation the most appropriate alternative to an empty reference class is the most specific, most reasonable syntactic generalisation. The Chapter demonstrates that if we do not know much about the situation of interest, then we should chain as well as generalise.

Syntactic versus semantic generalisation

At first, the finding that semantic generalisation is inappropriate appears inconsistent with Bacchus' [Bac90] and Goodwin's [Goo91] application of semantic generalisation to the apparently analogous task of discriminating among contradictory theories in *LP*.

The reason that semantic generalisation works in the context of *LP* is that the KB designer can carefully exclude erroneous independence assumptions. In contrast, the independence assumptions used in this thesis must be automatically generated without the benefit of a KB designer's intuitions. However, if the KB designer makes a mistake or can not anticipate all uses of the knowledge, then semantic generalisation leads to counter intuitive approximations in the context of LP just as it does in this thesis. For example,

Example 53 Suppose LP is provided with the following statistical assumptions

1. $E\left(\left[flies(x) \middle| \begin{array}{c} Arctic(x) \\ \wedge bird(x) \\ \wedge black(x) \end{array} \right]\right) = E\left(\left[flies(x) \middle| \begin{array}{c} Arctic(x) \\ \wedge bird(x) \end{array} \right]\right)$
2. $E\left(\left[flies(x) \middle| \begin{array}{c} Arctic(x) \wedge bird(x) \\ \wedge black(x) \end{array} \right]\right) = E([flies(x)|Arctic(x)])$
3. $E\left(\left[flies(x) \middle| \begin{array}{c} Arctic(x) \\ \wedge bird(x) \\ \wedge black(x) \end{array} \right]\right) = E\left(\left[flies(x) \middle| \begin{array}{c} \left(\begin{array}{c} Arctic(x) \\ \wedge black(x) \\ \wedge bird(x) \end{array} \right) \\ \vee lawyer(x) \end{array} \right]\right)$

Suppose $T3$ is the LP theory resulting from making the third expectation independence assumption and using the statistical knowledge

$$[flies(x)|(Arctic(x) \wedge bird(x) \wedge black(x)) \vee lawyer(x)]_x = .08$$

to approximate

$$[flies(x)|Arctic(x) \wedge bird(x) \wedge black(x)]$$

$T3$ is preferred to the theories $T1$ and $T2$ that result from making the first two expectation independence assumptions because the statistical knowledge

$$[flies(x)|(Arctic(x) \wedge bird(x) \wedge black(x)) \vee lawyer(x)]_x = .08$$

is the most specific. Thus, using semantic generalisation we derive the counterintuitive approximation that the frequency of black, arctic birds that fly is low because the frequency of flying lawyers is low.

Chaining

The set of all most specific alternatives to the reference class of a failed probability term $Prob^{II}(\alpha|\beta)$ is $S(Prob^{II}(\alpha|\beta))$. Chapter 4 argues that syntactic generalisation should be used to select a subset of S, the subset obtained by ignoring what we know. For example, the operator $\succeq$ can be used to identify a subset $S^{\succeq}(Prob^{II}(\alpha|\beta))$ that excludes any generalisations obtained by disjoining arbitrary properties to β. Unfortunately, $S^{\succeq}(Prob^{II}(\alpha|\beta))$ may exclude an appropriate alternative if β specifies little about the situation of interest.

Chapter 4 presents chaining as a novel means of making predictions in situations where syntactic generalisation fails to identify an appropriate alternative to a failed probability term. In particular, given a probability term $Prob^{II}(\alpha|\beta)$, chaining can be used to extend the set $S^{\succeq}(Prob^{II}(\alpha|\beta))$ by elaborating what we know about the situation of interest. As discussed in Chapter 4, Chaining, in certain situations, identifies a subset of those elements in $S(Prob^{II}(\alpha|\beta))$ obtained by semantic generalisation.

The difficulty with applying chaining lies in knowing what to chain on. Chapter 4 describes a particular heuristic for identifying the most relevant information to chain on. The heuristic represents a form of inductive bias and like all inductive biases it must be empirically tested. The results of Chapter 5 suggest that the heuristic can be used to make reasonable predictions.

6.2.3 Chapter 5: Experimental results

Chapter 5 describes three experiments. The experiments demonstrate that:

1. Syntactic generalisation and chaining can be used in the context of the RCA to make reasonable prediction.

2. The performance of syntactic generalisation depends upon the estimate of independence used when choosing a most reasonable, most specific syntactic

generalisation.

3. Chaining is more appropriate than syntactic generalisation when the only generalisation of a failed probability term $Prob^{II}(\alpha|\beta)$, is $Prob^{II}(\alpha|\top)$.

The results in Chapter 5 are of general interest to both the non-monotonic reasoning and the machine learning communities. This is a natural consequence of the observation made in Chapter 2 that both communities are addressing different aspects of the same problem.

The RCA and non-monotonic reasoning

The results described in Chapter 5 are of general interest to the non-monotonic reasoning community because they provide empirical validation of some of the non-monotonic and direct inference techniques reviewed in Chapter 2. For example, the results suggest that extra-logical preference assumptions (e.g., [Eth87], [AM91], [Bou92], [Poo91], [Jr.88a]) and irrelevance assumptions (e.g., [Bou91], [Pea88], [Bac90], [Sub90]) are a good basis on which to build techniques for selecting reasonable alternative reference classes. Previously, advocates of preference and irrelevance assumptions have relied on arguments with little or no empirical support.

Existing instantiations of preference and irrelevance assumptions have been semantic. The results of Chapter 5 reinforce the argument in Chapter 4 that preference and irrelevance assumptions need to be syntactically constrained.

The RCA and machine learning

The results in Chapter 5 are also of general interest to the machine learning community. The results suggest that techniques that have traditionally been considered the domain of deductive reasoning are applicable to the inductive problems considered the domain of machine learning.

In agreement with the machine learning literature (e.g., [FSK$^+$93], the experimental results in Chapter 5 suggest that there is no such thing as a 'best' inductive technique. In particular, the differences in performance between syntactic generalisation plus correlation and syntactic generalisation with clustering suggest that the appropriateness of inductive biases depend upon the information in a particular EKB.

The problem of selecting an appropriate inductive bias for estimating independence is not unlike the problem of "fine tuning" a machine learning algorithm. However, I argue that there are two important differences:

1. The fine tuning in a machine learning algorithm is not always obvious. In the context of the RCA, any fine tuning can be clearly identified with the measure of association used to estimate the reasonableness of independence assumptions.

2. Methods for fine tuning machine learning algorithms are often ad-hoc, frequently involving the programmer's intuitions. In contrast the statistical literature contains readily available and often well considered knowledge as the appropriateness of a particular metric of associativity.

6.3 Implications and Future Work

There are numerous directions - from theoretical extensions to practical applications of the RCA - to explore in the future. The RCA is a framework from which future research is to be hung. In this section I address several possible extensions of the RCA in terms of the four explicit assumptions stated in Chapter 1.

6.3.1 Assumption 1

> **Assumption 1** The only source of domain knowledge are past experiences such that each past experience is described by a single ground

sentence called a case.

If we relax Assumption 1, then large amounts of additional domain knowledge can be used to make predictions. For example, Appendix A demonstrates how knowledge about exclusive features can be added to the EKB. This knowledge is used in Chapter 5 to extend the number of cases that can be considered part of a reference class extension.

Recent research argues that the ability to incorporate domain knowledge in addition to past experiences is particularly important in machine learning (e.g., [Des92]; [Mic93]; [SBN93]; [Paz93]). Given the non-monotonic reasoning heritage of the generalisation techniques used by the RCA, it is easy to imagine how the consideration of deductive domain knowledge might take place. For example, it is is easy to see how statistical assertions such as those found in Bacchus' [Bac90] logic LP can be incorporated. The domain knowledge encoded by the statistical assertion

$$[Fly(X)|Bird(X)] \; = \; p$$

might be represented by adding the two tuples

$$\begin{array}{lll} \langle & n, \; R(l_i, flies, moves) \wedge R(l_i, bird, species) & \rangle \\ \langle & m, \; \neg R(l_i, flies, moves) \wedge R(l_i, bird, species) & \rangle \end{array}$$

to an EKB of past experiences such that $\frac{n}{m+n} \; = \; p$. The actual values n and m might indicate a degree of belief in the statistical assertion such that as n and m increase so does our belief that the statistical knowledge is reliable (this is a simple consequence of the concept of statistical adequacy discussed in Chapter 1).

In the remainder of this section I briefly discuss how some of this domain knowledge might be used by the RCA to extend generalisation and chaining.

Extending generalisation and chaining

In chapter 4, I argue that syntactic generalisation using $\succeq$ is appropriate because it does not violate Assumption 1. If we relax Assumption 1, then we can consider potentially more powerful alternatives to $\succeq$. For example, the wfs "red in colour" might be generalised by replacing $R(l_i, red, colour)$ with

$$R(l_i, red, colour) \vee R(l_i, orange, colour) \vee R(l_i, yellow, colour)$$

if red, yellow and orange were judged to be 'similar' colours with respect to making a particular prediction. Similarly, we might generalise the wfs "lives in Manhattan" to "lives in Manhattan or the Big Apple" if we knew that 'Manhattan' and the 'Big Apple' were the same place.

We might also extend syntactic generalisation by considering techniques that generalise α as well as β in a failed probability term $Prob^{II}(\alpha|\beta)$. For example,

Example 54 Suppose the probability term

$$Prob^{II}(flies|dead \wedge bird)$$

fails. In this thesis syntactic generalisation only considers generalisations of *dead* and *bird*. Suppose we had the additional domain knowledge that dead things do not move and flying is a form of moving. Using this knowledge we might generalise *flies* to *moves* and consider whether or not dead birds move.

Chaining can be extended if we have access to universal facts of the form "all birds have feathers" or "Manhattan is the same thing as the Big Apple". Knowledge of this sort can be used to decide what to chain on.

Example 55 Suppose the probability term

$$Prob^{II}(mugged|lives\,in\,Manhattan)$$

fails. If we know that

$$R(x, Manhattan, Lives) \rightarrow R(x, New\,York, Lives)$$

then we can chain on "Lives in New York" to get

$$Prob^{II}(mugged|lives\,in\,NewYork)$$

6.3.2 Assumption 2

> **Assumption 2:** The propensity that α can be used to describe a new experience, whenever β is known to be true of the new experience, is the same for a new experience as it is among all the past experiences in the reference class extension.

If we relax Assumption 2 and include domain knowledge about the propensity of observing a property α in a situation of interest, then we can use this knowledge to improve the efficiency and performance of the RCA.

Improving efficiency

If we know in advance the propensity to observe certain properties, then this knowledge can be used to improve the efficiency of the RCA by structuring the EKB. For example, if cases are simply appended the end of an EKB then algorithms for looking-up cases will spend a significant amount of time linearly searching through the EKB. If, on the other hand, cases are inserted intelligently so that cases that are used frequently are easier to look up, then the look-up algorithms will be more efficient. For example, we might store cases in a partial lattice ordered by logical implication. This would allow a look-up algorithm to use indexing techniques to find the relevant cases without having to recalculate measures of irrelevance for every new experience.

If we know in advance what it is that are going to predict we can improve the efficiency of the RCA by coalescing the cases in the EKB whose descriptions are the same with respect to the prediction of interest. For example, in the soya bean data set cases that are indistinguishable except for the value of the feature 'Date of observation', and we know that we will never want to predict the 'Date of observation' then we might coalesce those cases because 'Date of observation' is not predictive of any other feature value.

However, there is a tradeoff between the time taken to structure an EKB and the time taken to make a prediction. If our EKB must be restructured each time we obtain a new experience, then the advantages of structuring the EKB may be outweighed by the time required.

Improving predictive performance

A priori knowledge about the propensity to observe certain properties can also be used by the RCA to choose a more reasonable generalisations. For example, if we are provided with knowledge about *independence*, then this knowledge can be directly applied instead of using inductive biases to estimate the reasonableness of independence assumptions. We might also use knowledge about independence to choose the most appropriate estimator of independence.

6.3.3 Assumption 3

> **Assumption 3** Any non-empty reference class extension is adequate, and any empty reference class extension is inadequate, with respect to estimating $Prob(\alpha|\beta)$.

Assumption 3 contradicts the assumption of statistical adequacy adopted in existing work that addresses the reference class problem (e.g., [Jr.88a], [Bac90], [Goo91]). However, the most important property of a computational model for making predictions is its ability to make accurate predictions. I argue that the RCA's adoption of the criteria of psychological adequacy is supported by both the empirical results of Chapter 5 and the observations in Chapter 2 that humans may also use small reference classes.

An open problem is whether or not Assumption 3 will continue to applicable if the EKB gets very large or if the cases are very noisy. Current research is currently addressing this issue.

6.3.4 Assumption 4

> **Assumption 4** A reasonable estimate of $Prob(\alpha|\beta)$ can be obtained by generalising any properties of the new experience that are epistemologically irrelevant with respect to estimating the probability of α. Moreover, probabilistic independence is a reasonable measure of epistemological relevance and can be estimated by applying statistics to the available past experiences.

The reasonableness of Assumption 4 is supported by the examples in Chapter 4 as well as the experimental results in Chapter 5. However, this thesis does not begin to do justice to the problem of identifying a metric for determining which properties of a new experience are epistemologically relevant with respect to estimating a particular conditional probability.

The results of the three experiments in Chapter 5 were obtained using very simple measures of irrelevance. What is intriguing about the results in Chapter 5 is that the measures worked as well as they did. There is reason to believe that the results can be improved upon by adopting more sophisticated measures of irrelevance. For example, I observed in Chapter 5 that the measures of irrelevance used in this thesis assume that there are no inter-correlations between the properties that are being generalised with respect to the property being predicted. An important direction for future research lies in the problem of identifying which inductive biases should be used in conjunction with the RCA to measure the reasonableness of independence assumptions.

6.4 Discussion

The strength of the RCA lies in its simplicity and in its:

1. Ability of the RCA to use syntactic generalisation and chaining to estimate probabilities from readily available knowledge.

2. Capability to make reasonable estimates of conditional probabilities.

Induction algorithms can be classified on the basis of how much information beyond the cases is supplied as input [Win75]. The RCA as described in this thesis is purely inductive. No domain knowledge other than cases is used to make predictions. Most classification algorithms fall somewhere between the extreme of pure induction and learning by being told [Sal90] where the classification algorithm is given a complete description of the target concepts. For example, the classification algorithms EACH [Sal90] and IBL [AKA91] are provided cases augmented with information about which classes the cases will be used to predict instances of. Other algorithms, (e.g., [Des92] [DeJ81] [Mit83]) require considerable amounts of domain specific knowledge in order to make predictions. As a result the algorithms are of necessity domain dependent. Because the algorithms are domain dependent we can not be sure that they solve the problem of making predictions from experiences in general, or the problem of making predictions given a particular representation and a particular domain.

The RCA is a domain independent approach for making predictions about domains. The RCA does not convert cases into another representational form. Like pattern matching algorithms "it does not need a domain theory to explain which conversions are legal, or even what the representations mean" [Sal90, page 14]. A consequence of the RCA's domain independence is that the syntactic generalisation and chaining techniques that it uses are generally applicable to the problem of making predictions from experiences.

Appendix A

The language L

This appendix describes the formal properties, syntactic and semantic, of L in more detail. I start by describing the syntax of L. I then briefly describe the semantics of L, concentrating on those aspects which are unique to the language.

A.1 Language Syntax

In this section I define the syntax of a language for talking about experiences.

Alphabet = { 0, s, l, R, E, ′, (,), =, ∨, ∧, ¬, ⊥, }.

Term ::= Number | L-term.

Unary Function symbols ::= s | l.

Number ::= 0 | s(Number).

Constant ::= C(Number)

L-term ::= l(Number)

P-term :: $= Prob(Sentence \mid Sentence)_{EKB}$ such that α and β are sentences of L.

Predicate ::= Equality predicate | Property predicate.

Equality Predicate ::= (Term = Term)

Exclusive Predicate ::= $E(L - term)$.

Property Predicate ::= *R(L-term | constant, L-term | constant, L-term | constant).*

Sentence ::= Predicate | (Sentence $\vee$ Sentence) | (Sentence $\wedge$ Sentence) | $\neg$ (Sentence) | $\perp$

With the addition of the symbols { 0, 1, 2, ..., 9, 'red', 'yellow', 'colour', 'large', 'size', ... } to the alphabet of L I sometimes

1. Write lower case Greek letters such as α and β to represent individual sentences.
2. Use $\alpha \rightarrow \beta$ for $\neg\alpha \vee \beta$, and $\alpha \leftrightarrow \beta$ for $\alpha \rightarrow \beta \wedge \beta \rightarrow \alpha$.
3. Use the usual definitional extensions such as '$\top$' for '$\neg \perp$'.
4. Call the set of terms of the form s(Number)*numbers* and I will write 1 for s(0), 2 for s(s(0)), ... , where 's' stands for the 'successor' function.
5. Call the set of L-terms of the form l(Number) *labels* and I will sometimes write l_{Number} for l(Number).

I include the usual axioms of first order predicate calculus including those for quantification and equality [1]. In addition to the usual FOPL axioms I add the following:

Distinct labels The following axioms say that the function l maps distinct numbers to distinct values in the range of the functions. These axioms are as follows:

Axiom 1.1 $(\forall x_i)(\forall x_j)(\neg(x_i = x_j) \rightarrow \neg(l(x_i) = l(x_j)))$.

Axiom 1.2 $(\forall x_i)\neg(s(x_i) = 0)$

[1] See Johnstone [Joh87, pages 23-24] for a complete list of the FOPL axioms.

Axiom 1.3 $(\forall x_i)(\forall x_j)(\neg(x_i = x_j) \rightarrow \neg(s(x_i) = s(x_j)))$

From axioms 1.2 and 1.3 any two distinct numbers can be shown to be indeed distinct. With the addition of axioms 1.1 the individual labels can be proven to be distinct.

Example 56 If I have the labels $l(34)$ $l(62)$, then $l(34)$ and $l(62)$ are distinct, as $34 \neq 62$.

Exclusive labels are defined as follows:

Exclusive labels: If a label $l(i)$ is exclusive (written $E(l(i))$), then

Axiom 2.1

$$\begin{aligned}&(\forall x)(\forall y)(\forall y')(\forall z)\\&\quad R(l(x), l(y), l(z)) \wedge E(l(z)) \wedge R(l(x), l(y'), l(z)) \rightarrow \neg(y = y')\end{aligned}$$

Axiom 2.1 allows useful theorems about the properties of exclusive labels to be derived. For example, The counter positive of axiom 2.1,

$$\begin{aligned}&(\forall x)(\forall y)(\forall y')(\forall z)\\&\quad (R(l(x), l(y), l(z)) \wedge E(l(z)) \ \wedge \ \neg(y = y'))\\&\quad \rightarrow \neg(R(l(x), l(y'), l(z))))\end{aligned}$$

can be applied as follows:

Example 57 Let 'colour' be exclusive. Let

$$R(l(34), red, colour) \ \vee \ R(l(34), green, colour)$$

be a sentence in L. By axioms 2.1, 1.1, 1.2, 1.3 and the FOPL axioms for disjunction

$$\neg(R(l(34), yellow, colour))$$

is also true, assuming that 'red', 'green' and 'yellow' correspond to distinct values in the object language, say $l(10)$, $l(11)$, and $l(12)$.

A.2 Language Semantics

This section explores the intended interpretation of formulae written in the language L. Intuitively, the semantics provides a model for the formulae of L that specifies all the ways that the world could be given the information contained in a particular formula. I now define those aspects of the semantics of L of interest by defining an interpretation structure M[2].

A.2.1 An interpretation

Definition 30 *An interpretation M with respect to a language L is:*

$$M = \langle U, \varphi, \vartheta \rangle$$

The components of M are:

1. *U : The domain. There are two distinct types in the domain U:*
   - *$S = \{s_1, s_2, \ldots\}$, A countable set of objects (e.g. David's bicycle).*
   - *$N = \{0, 1, 2, \ldots\}$, The set of natural numbers.*

   The domain is the union of these two types:

   $$U = S \cup N$$

2. *ϑ: A mapping defined on the variables of L such that if x is an individual variable, then $\vartheta(x)$ is an individual in U.*
3. *φ: A mapping defined on the numbers, function symbols and predicate symbols of L such that:*
   (a) *Each individual number symbol in L is assigned a specific number in N.*
   (b) *Each individual function symbol l is assigned a 1 to 1 function from N to S.*

[2] See Johnstone [Joh87] for a complete review of the semantics of FOPL.

(c) *The function symbol Prob is assigned a function from* $\Gamma \times \Gamma \times \Omega$ *to* $[0,1]$*, such that* Γ *is the set of all wfss in* L*, and* Ω *is the set of all possible EKBs.*

(d) *The 3-place predicate symbol* R *in* L *is assigned a relation in* S^3*.*

(e) *The 1-place predicate symbol* E *in* L *is assigned a relation in* S*.*

(f) *The 2-place predicate symbol* $=$ *in* L *is assigned a binary relation in* U*.*

s, l and Prob are the only function symbols in the language. I interpret the function s as follows:

$$s: \ N \rightarrow N$$

such that $s(x) \ = \ x \ + \ 1$. s is the simple successor function used in defining the natural numbers. This is consistent with axioms 1.2 and 1.3.

The predicate symbols E, '=', and R are the only predicate symbols.

1. R is assigned a 3-ary relation in

$$S \times S \times S$$

2. The unary-predicate symbol 'E' is a unary relationship over the S. $E(l_l)$ is true in an interpretation M if and only if $\forall l_k$ and $l_i \neq l_j$, $\{ \langle l_k, l_i, l_l \rangle, \langle l_k, l_j, l_l \rangle \} \not\subseteq R$.

3. The binary predicate symbol '=' is a binary relationship over the domain U. In particular $l(i) = l(j)$ iff $i = j$.

The axioms of L can be used to prove that l is a $1-1$ mapping which is agreement with the semantics of L. The axioms of L ensure that distinct numbers are indeed distinct. In the semantics each distinct number indexes a distinct object, using the function l.

A.2.2 A valuation for L

The standard FOPL truth valuation can now be defined over the formulae of a language L using the interpretation $M = \langle U, \varphi, \vartheta \rangle$. M provides a truth value for the atomic sentences of the language and the rest of the sentences are assigned a truth value inductively. As the usual first order theory valuation of variables, functions, relations and logical symbols holds for L I will not reiterate it here. The interested reader is referred to Bell and Machover [BM77] for a detailed discussion of the valuation of a FOPL language.

Appendix B

Estimating independence by correlation

This appendix contains a detailed description of the *correlation* statistic. The statistic is used to order the most specific alternatives to the reference class of a failed probability term by estimating the reasonableness of the independence assumption $I(\alpha, \beta, \gamma)$ made in Chapter 4 in order to syntactically generalise a reference class.

The appendix starts by describing how the correlation between pairs of features is calculated. The correlation statistic is used to justify the independence assumptions made when generalising a reference class. The appendix concludes with a a detailed example showing how correlations are used to choose a particular generalisation of a failed probability term $Prob^{II}(\alpha|\beta)$ from the set of most specific alternatives $S^{\succeq}_{s}(Prob^{II}(\alpha|\beta))$.

B.1 The correlation coefficient

Following Edwards[Edw76], the correlation coefficient r may be defined as the covariance of two variables, divided by the product of the standard deviations, S_X and S_Y, of the variables [1]

$$r = \frac{C_{XY}}{S_X \times S_Y}$$

[1] There are three important special cases of the correlation coefficient: 1. the phi coefficient, 2. the point biserial coefficient, and 3. the rank order correlation coefficient. The formulae for each of these special cases are given in Edwards [Edw76, sections 7.2-7.5] and are equivalent to the formulae provided above. In the case in which both variables are dichotomous, the phi coefficient provides a more efficient calculation of correlation. The point biserial coefficient is applied when one variable is dichotomous and the other is continuous. Finally, the rank order correlation coefficient can be used when both the X and Y variables consist of a set of ranks.

	$R(l_i, flies, moves)$	$\neg R(l_i, flies, moves)$
1. $R(l_i, black, colour)$	4	2
2. $R(l_i, pink, colour)$	8	0
3. $R(l_i, blue, colour)$	100	1
4. $R(l_i, green, colour)$	3	0
5. $R(l_i, yellow, colour)$	0	0
6. $R(l_i, red, colour)$	6	0
7. $R(l_i, white, colour)$	50	0
8. $R(l_i, orange, colour)$	2	0

Table B.9: Frequency counts calculated from Figure B.1 for the feature *moves* and the feature *colour* with respect to the reference class of the probability term $Prob(R(l_i, flies, moves) \mid R(l_i, yellow, colour) \wedge R(l_i, bird, species))$.

Variable	*Inter-correlations*		
Colour	−.07		
Size	.16	−.5	
Moves	.42	.01	.15
	Species	Colour	Size

Table B.10: Inter-correlations of the features: *Species, Colour, Size, and moves* in Figure B.1.

From this equation we can derive the the following general equation for the correlation coefficient[2]:

$$r_{XY} = \frac{\sum(X-\bar{X})(Y-\bar{Y})/(n-1)}{\sqrt{\frac{\sum(X-\bar{X})^2}{n-1}}\sqrt{\frac{\sum(Y-\bar{Y})^2}{n-1}}} \quad \text{(B.1)}$$

$$= \frac{\sum xy}{\sqrt{\sum x^2}\sqrt{\sum y^2}} \quad \text{(B.2)}$$

[2] See An Introduction to Linear Regression and Correlation by A. Edwards for a detailed description of the derivation. $\bar{X}$ and $\bar{Y}$ are the averages of the X and Y variables and n is the number of instances.

$$
\begin{array}{lll}
\{ & \langle 4, & R(l(1), black, colour) \wedge R(l(1), flies, moves) \\
 & & \wedge R(l(1), bird, species) \wedge R(l(1), large, size)\rangle, \\
 & \langle 2, & R(l(2), black, colour) \wedge R(l(2), walks, moves) \\
 & & \wedge R(l(2), bird, species) R(l(2), small, size)\rangle, \\
 & \langle 1, & R(l(3), pink, colour) \wedge R(l(3), bird, species) \\
 & & \wedge R(l(3), small, size)\rangle, \\
 & \langle 8, & R(l(4), pink, colour) \wedge R(l(4), bird, species) \\
 & & \wedge R(l(4), flies, moves)\rangle, \\
 & \langle 100, & R(l(5), blue, colour) \wedge R(l(5), bird, species) \\
 & & \wedge R(l(5), flies, moves)\rangle, \\
 & \langle 1, & R(l(6), blue, colour) \wedge R(l(6), bird, species) \\
 & & \wedge - R(l(6), flies, moves)\rangle \\
 & \langle 3, & R(l(7), green, colour) \wedge R(l(7), bird, species) \\
 & & \wedge R(l(7), flies, moves)\rangle, \\
 & \langle 1, & R(l(8), purple, colour) \wedge R(l(8), bird, species) \\
 & & \wedge R(l(8), flies, moves)\rangle, \\
 & \langle 6, & R(l(9), red, colour) \wedge R(l(9), bird, species) \\
 & & \wedge R(l(9), flies, moves)\rangle, \\
 & \langle 50, & R(l(10), white, colour) \wedge R(l(10), bird, species) \\
 & & \wedge R(l(10), flies, moves)\rangle, \\
 & \langle 2, & R(l(11), orange, colour) \wedge R(l(11), bird, species) \\
 & & \wedge R(l(11), flies, moves)\rangle \qquad \}
\end{array}
$$

Figure B.15: An EKB containing cases that describe domain states in terms of values of the features colour, moves, species and size.

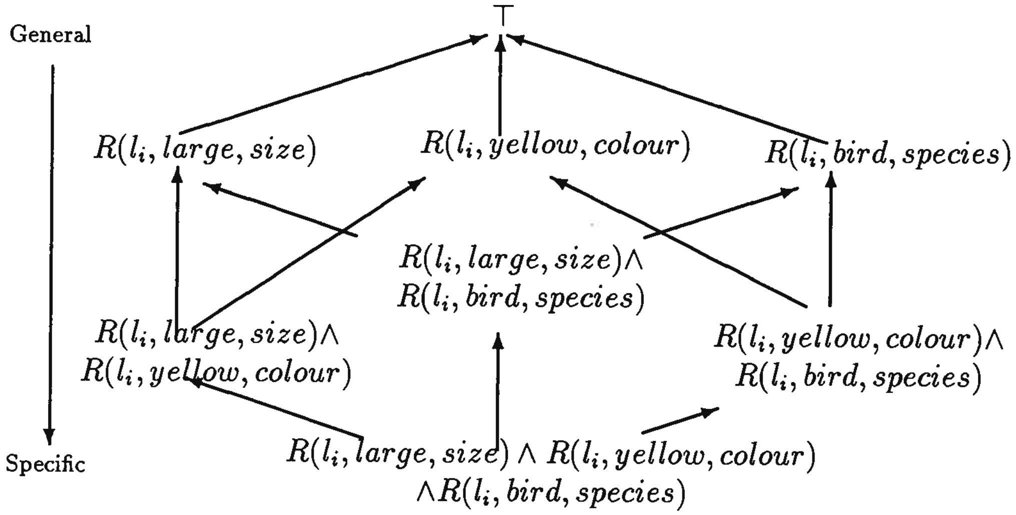
General
Specific
⊤
R(l_i, large, size)
R(l_i, yellow, colour)
R(l_i, bird, species)
R(l_i, large, size)∧ R(l_i, bird, species)
R(l_i, large, size)∧ R(l_i, yellow, colour)
R(l_i, yellow, colour)∧ R(l_i, bird, species)
R(l_i, large, size) ∧ R(l_i, yellow, colour) ∧R(l_i, bird, species)

Figure B.16: The lattice of generalisations of the reference class of $Prob^{II}(R(l_i, flies, moves)| R(l_i, yellow, colour) \wedge R(l_i, bird, species) \wedge R(l_i, large, size))_{EKB}$ obtained by syntactic generalisation.

B.2 An example

The lattice of generalizations of the failed probability term

$$Prob^{II}\left(R(l_i, flies, moves)\middle|\begin{array}{c}R(l_i, yellow, colour) \wedge R(l_i, bird, species)\\ \wedge R(l_i, large, size)\end{array}\right)_{EKB}$$

generated by applying $S^{\succeq}(Prob(\alpha|\beta)$ is presented in Figure B.16. With respect to the EKB in Figure B.15 there are three most-specific generalisations with adequate statistics. The intentions of these candidates are circled in Figure B.16.

Using the correlations in Table B.10 we select the single adequate syntactic generalisation

$$Prob^{II}(R(l_i, flies, moves)|R(l_i, bird, species) \wedge R(l_i, large, size))$$

because it is obtained by ignoring the value of the least relevant feature colour.

Appendix C

Estimating independence by clustering

This appendix describes the clustering statistic used in Experiments 1 and 2 in Chapter 5. The clustering statistic provides an estimate of the reasonableness of the independence assumption $I(\alpha, \gamma, \beta)$ by measuring how well β predicts α. In the context of Chapter 5 α and β are assumed to be single property predicates, each specifying the value of a mutually exclusive feature. The clustering statistic measures how predictive the values of the feature specified by β are of the values of the feature specified by α. Intuitively, the higher this measure the less likely α is to be independent of β.

C.1 Optimal predictability

Suppose we wish to know how well the values of a feature Y called the *source* predict the values of a feature called the *target* X with respect to an EKB. Suppose X can have m different values $x_1, \ldots, x_m$ in the EKB and suppose that Y can have n different values $y_1, \ldots, y_n$.

I argue that the ability to predict the values of X from Y depend upon the number m and n of values that X and Y have. I define the optimal error for predicting X from Y as follows:

Definition 31 *The optimal error for predicting the value of a feature X with m*

values from a feature Y with n values is:

$$\begin{array}{ll} 0 & if \;\; n >= m \\ \frac{1}{m} & if \;\; n = 1 \\ \frac{m-n}{m} & if \;\; (n > 1) \wedge (n < m) \end{array}$$

Example 58 Suppose we wish to measure the predictability of the values of colour from the values of size. In our EKB size can have 2 different values (small and large) and colour can have 4 different values (red, green, yellow and blue). At the very best, given a value for size we can only predict that the probability of colour having a particular value is 0.5. If $n >= m$, then at best there can be at least one value of Y for every value of X. If this were the case, then knowing the value of Y would tell us the value of X. If $n = 1$, then at best knowing the value of Y allows us to guess which value X has.

C.2 Actual predictability

I define the actual error associated with predicting the values of X from Y as follows:

Definition 32 *let X have the possible values $x_1, \ldots, x_m$ in the EKB and let Y have the possible values $y_1, \ldots, y_n$. The actual error for predicting a value of X given a value of Y is:*

$$\Sigma_{j=1}^{n} \left[\frac{\Sigma_{i=1}^{m}[Prob^{II}(y_j|x_i) \times (\Sigma_{l=1}^{m} Prob^{II}(y|x_l) - Prob^{II}(y_j|x_i)]}{\Sigma_{i=1}^{m} Prob^{II}(y_j|x_i) \times m} \right]$$

such that

$$Prob^{II}(y_j|x_i) \;=\; \frac{|y_j \wedge x_i|}{|x_i|}$$

C.3 Estimating independence

The estimate of independence between two features X and Y is a function of the difference between the actual and the optimal. In Chapter 5 I use:

$$1 \;-\; \sqrt{(optimal - actual)}$$

Appendix D

Expanding h

D.1 Expansion I

In this section I consider the problem of excluding tuples from h that represent observations of domain states that describe fewer objects than the number of objects of interest.

Suppose the EKB contains the case $R(l(1), red, colour)$ and that we wish to count the number of observations described in the EKB of two objects in the same domain state that are both red, i.e., we wish to know the cardinality of

$$h(EKB, R(l_i, red, colour) \land R(l_j, red, colour))$$

The set $h(EKB, R(l_i, red, colour) \land R(l_j, red, colour)$ contains the set of all 2-tuples that can be substituted for l_i and l_j. Among these 2-tuples is the 2-tuple $\langle l_1, l_1 \rangle$ which is counter intuitive as the case $R(l_1, red, colour)$ only describes an observation of a *single* object that is red, not two.

The definition of h can be extended to avoid counting the tuple $\langle l_1, l_1 \rangle$ by requiring that all labels in the tuples be distinct, i.e., by requiring that $h(EKB, \alpha)$ is defined as

$$\{\langle l'_1, \ldots, l'_n \ : \ (\forall l_i, lj) l_i \neq l_j \; and \; EKB \vdash \alpha(l_1/l'_1, \ldots, l_n/l'_n)\}$$

D.2 Expansion II

In this section I consider the problem of excluding tuples from h that represent observations of domain states that describe more objects than the number of objects of interest.

Suppose the EKB contains the case

$$R(Othello, l_1, parent) \wedge R(Tulving, l_1, parent) \wedge R(Dick, l_1, parent)$$

describing a parent with three children, Tulving, Othello and Dick. Suppose we wish to count the number of observations described in the EKB of a parent with two children, i.e., we wish to know the cardinality of

$$h(EKB, R(l_i, l_k, parent) \wedge R(l_j, l_k, parent))$$

Using the extended definition of h from the previous section, the set contains the subset

$$\{\langle Othello, l_1, Tulving\rangle, \langle Othello, l_1, Dick\rangle, \langle Tulving, l_1, Dick\rangle, \}$$

which is counter intuitive because the case

$$R(Othello, l_1, parent) \wedge R(Tulving, l_1, parent) \wedge R(Dick, l_1, parent)$$

describes a parent that has three children and not two.

h can be extended to avoid counting the tuples

$$\{\langle Othello, l_1, Tulving\rangle, \langle Othello, l_1, Dick\rangle, \langle Tulving, l_1, Dick\rangle, \}$$

by redefining h as follows:

$$\left\{ \langle l'_1, \ldots, l'_n\rangle \; : \; \begin{array}{l} EKB \vdash \alpha(l_1/l'_1, \ldots, l_n/l'_n) \; and \\ (if(\alpha(l_1/l'_1, \ldots, l_n/l'_n) \vdash R(i, v, f)) \; and \\ (\alpha(l_1/l'_1, \ldots, l_n/l'_n) \not\vdash R(i, v', f)) \; then \\ (EKB \not\vdash R(i, v', f))) \end{array} \right\}$$

D.3 Expansion III

In this section I consider the problem of excluding tuples from h that are syntactic variants.

Suppose the EKB contains the case

$$R(l_1, red, colour) \wedge R(l_2, large, size)$$

and that we wish to count the number of observations described in the EKB of a large object and a red object occurring in the same domain state. Using the extended definitions of h from the previous two sections, the set

$$h(EKB, R(l_i, red, colour) \wedge R(l_j, large, size))$$

contains the subset

$$\{\langle l_1, l_2\rangle, \langle l_2, l_1\rangle\}$$

which is counter intuitive in that the case

$$R(l_1, red, colour) \wedge R(l_2, large, size)$$

should only count as one observation of a red object and a large object in the same domain state.

In this case we can avoid counting the same observation more than once by considering the maximal sets of all subsets S of $h(EKB, \alpha)$ such that

$$(\forall X \in S)(\nexists Y \in S)\, such\, that\, (X \neq Y)\, and\, (l_i \in X)\, and\, (l_i \in Y)$$

By picking a single maximal set we exclude any tuples that are the result of counting the same observation described in an EKB more than once.

Appendix E

Machine Learning Data Sets

This appendix contains a brief description of the seven data sets used by Experiments 1, 2 and 3 in Chapter 5. A more complete description of these, and other machine learning data sets, can be obtained from the Machine learning data base repository at the University of California at Irvine.

Soybean The Soybean data base is divided into two parts: 1. The *Soybean training data base* containing 250 cases, and 2. The *Soybean testing data base* containing 296 cases. In each data base the cases are divided into fourteen classes such that each data base contains roughly the same number of cases in each class. Each case describes an observation of a single diseased soybean plant in terms of a set of 36 nominal valued exclusive features. Each case describes a single diseased soybean plant in terms of 36 features: Diagnostic category, data of observation, characteristics of the plant stand, local precipitation, local temperature, presence of hail, crop history, crop damage, severity of damage, seed treatment, per-cent germination, plant growth characteristics, ... Each feature is exclusive and the possible values of each feature are discrete.

Example 59 The case

$$R(l_i, low, precipitation) \wedge R(l_i, normal, temperature) \wedge \ldots$$
$$\ldots \wedge R(l_i, rotten, roots) \wedge R(l_i, charcoal - rot, disease)$$

describes a diseased soybean plant with charcoal rot disease that was exposed to low precipitation, normal temperature, ...

Fisher soybean The Fisher data base is a subset of the Soybean data base that contains only 47 contains divided into four diagnostic categories. Each case in the Fisher data base describes a soybean plant with: Diaporthe stem canker, Charcoal rot, Rhizoctonia rot, or Phytophtora rot. An informal analysis of the data base demonstrates that the knowing values of some subset of the nine features:

> lodging, stem cankers, canker lesions, fruiting bodies, external decay, mycelium, internal discolouration, sclerotia and fruit pods,

is sufficient for correctly predicting which of the four diagnostic categories the plant belongs to.

Example 60 If we know that a soybean plant has stem cankers above the second node, then the soybean plant has diaporthe stem canker, i.e.

$$Prob(R(l_i, \top, \mathit{Diaporthe}) \mid R(l_i, \mathit{second}, \mathit{stem\ cankers})) = 1$$

Breast The "Breast" data base contains 286 cases describing two hundred instances of women who have had breast cancer. The cases are divided into two classes: 85 instances of women who have had a re-occurrence of breast cancer and 201 instances of women who have not had a re-occurrence of breast cancer after an operation. There are nine attributes describing the original cancer nodes with multi-valued discrete and real values. The data set comes form the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. The prediction task is to predict whether or not a women will have a re-occurrence of breast cancer given a description of the cancer.

Votes The "Votes" data base contains 435 examples of the key votes of 267 democrats and 168 republicans during the 1984 U.S. congress. The congressmen voted on such issues as immigration and education spending. The votes have been simplified to yea, nay or abstained. The prediction task is to predict whether or not a voter is a democrat or a republican on the basis of his or her voting history.

Modified Votes The "Votes one" data set is derived from the "Votes" data set by deleting the most significant attribute *physician fee freeze* [BN92].

Mushrooms The "Mushrooms" data set consists of 8124 data. Each data records whether mushrooms from the Agaricus and Lepiota families are poisonous or edible. Each mushroom is described in terms of twenty two discrete attributes. The prediction task is to predict whether or not a mushroom is edible or inedible given values for each of the twenty two attributes.

LED The 7-digit "LED" data set is Breiman's [BFOS84] manufactured test data on the digit recognition problem. Each datum describes a single faulty LED display representing a digit from 0 to 9 in terms of seven binary valued attributes. The LED display is made faulty by adding 10 per-cent noise independently to each element. The prediction task is predict the digit in the LED display given values for each of the seven binary attributes. The prediction task has a theoretical minimum error of 27.3 percent [BN92].

Bibliography

[AGG83] S. Armstrong, L.R. Gleitman, and H. Gleitman. On what some concepts might not be. *Cognition*, 13:263–308, 1983.

[Agh90] D.S. Aghassi. Evaluating case-based reasoning for heart failure diagnosis. Technical Report MIT-LCS-TR-478, MIT, 1990.

[Aha89] D.W. Aha. Incremental learning of independent, overlapping, and graded concept descriptions with an instance-based process framework. Technical Report TR 89-10, University of California, Irvine, 1989.

[AK89] D. Aha and D. Kibler. Noise tolerant instance based learning algorithms. In *Proceedings of the 11th IJCAI*, pages 794–799. San Mateo, CA: Morgan Kaufmann, 1989.

[AKA91] D. Aha, D. Kibler, and M. Albert. Instance-based learning algorithms. *Machine Learning*, 6(1):37–66, 1991.

[AM91] N. Asher and M. Morreau. Commonsense entailment: A modal theory of nonmonotonic reasoning. In *Proceedings of 12th IJCAI*, pages 387–392. San Mateo, CA: Morgan Kaufmann, 1991.

[Bac90] F. Bacchus. *Representing and reasoning with probabilistic knowledge: A logical approach to probabilities.* MIT Press, 1990.

[Bar82] V. Barnett. *Comparative statistical inference: Second edition.* London: Wiley, 1982.

[Bar83] L.W. Barsalou. Ad hoc categories. *Memory and Cognition*, 11:211–227, 1983.

[Bar87] L.W. Barsalou. The instability of graded structure: Implications for the nature of concepts. In U. Neisser, editor, *Concepts and conceptual development: Ecological and intellectual factors in categorization*, pages 91–107. Cambridge: Cambridge University Press, 1987.

[BE89] A. Borgida and D. Etherington. Hierarchial knowledge bases and efficient disjunctive reasoning. In *Proceedings of Knowledge Representation (KR89)*, pages 33–43. San Mateo, CA: Morgan Kaufmann, 1989.

[Bel84a] F.S. Bellezza. Reliability of retrieval from semantic memory: Common categories. *Bulletin of the psychonomic society*, 22:324–326, 1984.

[Bel84b] F.S. Bellezza. Reliability of retrieval from semantic memory: Information about people. *Bulletin of the psychonomic society*, 22:511–513, 1984.

[Bel84c] F.S. Bellezza. Reliability of retrieval from semantic memory: Noun meaning. *Bulletin of the psychonomic society*, 22:377–380, 1984.

[BFOS84] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and regression trees.* Belmont: Wadsworth, 1984.

[BM77] J. Bell and Machover M. *A course in mathematical logic.* New York, NY: North-Holland, 1977.

[BMMZ92] F. Bergandano, S. Matwin, R. Michalski, and J. Zhang. Learning two-tiered descriptions of flexible concepts. *Machine Learning*, 8:5–44, 1992.

[BN92] W. Buntine and C. Niblett. A further comparison of splitting rules for decision tree induction. *Machine Learning*, 8:75–85, 1992.

[Bou91] C. Boutilier. Inaccessible worlds and irrelevance: Preliminary report. In *Proceedings of 12th IJCAI*, pages 413–418, 1991.

[Bou92] C. Boutilier. Conditional logics for default reasoning and belief revision. Technical Report 92-1, University of B.C., Vancouver B.C., 1992.

[BP89] R. Board and L. Pitt. On the necessity of occam algorithms. Technical Report UIUCDCS-R-89-1544, UIUC, 1989.

[Bra87] G. Bradshaw. Learning about speech sounds. In *Proceedings of the 4th International Workshop on Machine Learning.* San Mateo, CA: Morgan Kaufmann, 1987.

[BSP85] A. Bundy, B. Silver, and D. Plummer. An analytical comparison of some rule-learning algorithms. *IEEE Expert*, 2(3):137–181, 1985.

[Bun89] W. Buntine. Learning classification rules using bayes. In *Proceedings of the 6th international machine learning workshop.* New York: Morgan Kaufmann, 1989.

[But93] L.J. Buturovic. Improving k-nearest neighbours density and error estimates. *Pattern Recognition*, 26(4):611–616, 1993.

[CB91] P. Clark and R. Boswell. Rule induction with cn2: Some recent improvements. In *Machine Learning - EWSL-91*, pages 151–163. Springer-Verlag, 1991.

[CK89] J.M. Crawford and B. Kuipers. Towards a theory of access limited logic for knowledge representation. In *Proceedings of Knowledge Representation-89*, 1989.

[CM83] R. Clarke and J. Morton. Cross modality facilitation in tachistoscopic word recognition. *Quarterly journal of experimental psychology*, 35A:79–96, 1983.

[CMM83] J. Carbonell, R. Michalski, and T. Mitchell. An overview of machine learning. In *Machine learning, R. Michalski J. Carbonell and T. Mitchell Eds.*, pages 25–36. San Mateo, CA: Morgan Kaufman, 1983.

[CN87] P. Clark and T. Niblett. Induction in noisy domains. In I. Bratko and L. Lavrac, editors, *Progress in machine learning*, pages 11–30. Wilmslow, England: Sigma Press, 1987.

[CN88] P. Clark and T. Niblett. The cn2 induction algorithm. *Machine learning*, 3(4):261–283, 1988.

[Cre92] R. H. Creecy. Trading mips and memory for knowledge engineering. *Communications of the ACM*, 35(8):46–83, 1992.

[CS93] S. Cost and S. Salzberg. A weighted algorithm for learning with symbolic features. *Machine learning*, 10(1):57–78, 1993.

[Das91] B.V. Dasarathy. *Nearest neighbour pattern classification techniques.* Los Alamitos, CA: IEEE Press, 1991.

[Dav87] L. Davis. *Genetic algorithms and simulated annealing.* Los Angeles, CA: Morgan Kaufmann, 1987.

[Dav90] E. Davis. Partial information and vivid representations. In *Representation in mental models*, pages 123–145. Los Angeles, CA: Morgan Kaufmann, 1990.

[DeJ81] G. DeJong. Generalizations based on explanations. In *Proceedings of the 7th IJCAI*, pages 67–69. Morgan Kaufmann, 1981.

[Del88] J. Delgrande. An approach to default reasoning based on a first-order conditional logic: Revised report. *AI journal*, 36:63–90, 1988.

[Des92] M. DesJardins. Pagoda: A model for autonomous learning in probabilistic domains. Technical Report UCB/CSD 92/678, University of California, Berkeley, 1992.

[dW83] J. deKleer and B.C. Williams. Diagnosing multiple faults. *AI*, 32:372–388, 1983.

[EBBK89] D. Etherington, A. Borgida, R. Brachman, and H. Kautz. Vivid knowledge and tractable reasoning: Preliminary report. In *Proceedings of the 11th IJCAI*, pages 1146–1152. San Mateo, CA: Morgan Kaufmann, 1989.

[Edw76] A. L. Edwards. *An introduction to linear regression and correlation.* New York , NY: Freeman, 1976.

[EKP90] D. Etherington, S. Kraus, and D. Perlis. Nonmontonicity and the scope of reasoning. Technical Report CS-TR-2457, University of Toronto, April 1990.

[Eth87] D. Etherington. A semantics for default logic. In *Proceedings of 8th IJCAI*, pages 495–498. San Mateo, CA: Morgan Kaufmann, 1987.

[Eub88] R.L. Eubank. *Spline smoothing and non-parametric regression.* New York, NY: M. Decker, 1988.

[Fis87] D.H. Fisher. Knowledge acquistion via incremental conceptual clustering. *Machine Learning*, 2:139–172, 1987.

[FMM$^+$89] D. Fisher, K. McKusick, R. Mooney, J.W. Shavlik, and G. Towell. Processing issues in comparison of symbolic and connectionnist learning systems. In *Proceedings of the 6th international workshop on machine learning*, pages 169–173, Cornell University, Ithaca, New York, 1989.

[FPSM92] W.J. Frawley, G. Piatetsky-Shapiro, and C. Matheus. Knowledge discovery in databases: An overview. *AI magazine*, 14(3):57–70, 1992.

[Fri87] A. Frisch. Knowledge retrieval as specialised inference. Technical Report 214, University of Rochester, 1987.

[FS84] E.A. Feigenbaum and H. Simon. Epam-like models of recognition and learning. *Cognitive Science*, 8:305–336, 1984.

[FSK$^+$93] C. Feng, A. Sutherland, R. King, S. Muggleton, and R. Henery. Comparison of classification algorithms in machine learning, statistics and neural networks. In *To appear in Machine Learning*, 1993.

[FT78] A. Flexer and E. Tulving. Retrieval independence in recognition and recall. *Psychological Review*, 85:153–171, 1978.

[GB89] D.S. Gorfein and A. Bubka. A context sensitive frequency based theory of meaning achievement. In D.S. Gorfein, editor, *Resolving Semantic Ambiguity*. Springer Verlag, 1989.

[GLF89] J.H. Gennari, P. Langley, and D. Fisher. Models of incremental concept formation. *Artificial Intelligence*, 40:11–61, 1989.

[Goo86] M. Goodman. Case based reasoning in battle planning. In *Proceedings of the DARPA-sponsored case-based reasoning workshop*, pages 264–269, 1986.

[Goo91] S. Goodwin. Statistically motivated defaults. Technical Report CS-91-03, University of Regina, 1991.

[GR90] P. Graf and L. Ryan. Transfer appropriate processing for implicit and explicit memory. *Journal of experimental psychology: learning, memory and cognition*, 1990.

[GS88] R.M. Goodman and P. Smyth. The induction of probabilistic rule sets - the itrule algorithm. In *Proceedings of the sixth international workshop on machine learning. Cornell University, Ithaca, New York*, pages 129–132. San Mateo, CA: Morgan Kaufmann, 1988.

[Ham84] K. Hammond. Indexing and causality: The organization of plans and strategies in memory. Technical Report 351, Yale University, 1984.

[Ham86] K. Hammond. Case-based planning: An integrated theory of planning, learning, and memory. Ph.d. Dissertation, 1986.

[Ham89] K. Hammond. *Case-based planning: Viewing planning as a memory task*. New York, NY: Academic, 1989.

[Han81] D.J. Hand. *Discrimination and Classification*. London: J. Wiley and Sons, 1981.

[Han82] D.J. Hand. *Kernel discriminant analysis*. London: J. Wiley and Sons, 1982.

[HBP89] M. Humphreys, J. Bain, and R. Pike. Different ways to cue a coherent memory system: A theory for episodic, semantic, and procedural tasks. *Psychological Review*, 96(2):208–233, 1989.

[Hol93] R. C. Holte. Very simple classification rules perform well on most commonly used data sets. *Machine Learning*, 11:63–91, 1993.

[HV74] J. Hermans J.D. Habbema and A.T. VanDerBrught. Cases of doubt in allocation problems, k populations. *Bulletin of international statistics institute*, 45:523–529, 1974.

[HW90] S. Hollbach-Weber. Acquiring categorical aspects: A connectionnist account of figurative noun semantics. In *Pragmatics in AI: the 5th Rocky mountain conference on AI*, 1990.

[JH87] L. Jacoby and C. Hayman. Specific visual transfer in word identification. *Journal of experimental psychology: Learning, memory and cognition*, 13:456–463, 1987.

[Joh87] P.T. Johnstone. *Notes on logic and set theory.* Cambridge, UK: Cambridge university press, 1987.

[Jr.69] H.E. Kyburg Jr. *Probability theory.* Englewood Cliffs, N.J.: Prentice-Hall, 1969.

[Jr.74] H. E. Kyburg Jr. *The logical foundations of statistical inference.* Dordrecht, Netherlands: D. Reidel, 1974.

[Jr.83] H. E. Kyburg Jr. The reference class. *Philosophy of Science*, 50:374–397, 1983.

[Jr.88a] H. E. Kyburg Jr. Epistemological relevance and statistical knowledge. Technical Report 251, University of Rochester, 1988.

[Jr.88b] H. E. Kyburg Jr. Probabilistic inference and probabilistic reasoning. Technical Report 248, University of Rochester, 1988.

[Jr.88c] H. E. Kyburg Jr. Higher order probabilities and intervals. Technical Report 236, University of Rochester, November, 1988.

[Jr.88d] H. E.Kyburg Jr. In defense of intervals. Technical Report 268, University of Rochester, November, 1988.

[Jr.91] H. E. Kyburg Jr. Evidential probability. Technical Report 376, University of Rochester, March, 1991.

[Kea89] M.J. Kearns. *The computational complexity of machine learning.* Cambridge, MA: MIT press, 1989.

[Koh90] T. Kohonen. The self organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.

[Kol88] J. L. Kolodner. Retrieving events from case memory: A parallel implemenation. In *Proceedings of the DARPA-Sponsored Case-based reasoning workshop*, pages 233–249. San Mateo, CA: Morgan Kaufmann, 1988.

[Kol91] J.L. Kolodner. Improving human decision making through case-based decision analysis. *AI magazine*, Summer 1991:52–68, 1991.

[Kot89] P.A. Koton. Using experience in learning and problem solving. Technical Report MIT-LCS-TR-441, MIT, 1989.

[KST82] D. Kahneman, P. Slovic, and A. Tversky. *Judgement under uncertainty: Heuristics and Biases.* Cambridge: Cambridge university press, 1982.

[Lai88] P.D. Laird. *Learning from good and bad data.* Boston, MA: Kluwer Academic Publishers, 1988.

[Leb86] M. Lebowitz. Concept learning in a rich input domain: Generalization based memory. In R.S. Michalski J.G. Carbonell and T.M. Mitchell, editors, *Machine Learning: An artificial intelligence approach*, volume 2. Los Altos, CA: Morgan Kaufmann, 1986.

[Lev80] I. Levi. *The Enterprise of Knowledge.* Cambridge, MA: MIT Press, 1980.

[Lev86] H. Levesque. Making believers out of computers. *Artificial Intelligence*, 30:81–108, 1986.

[Lev88] H. Levesque. Logic and the complexity of reasoning. *Journal of Philosophical Logic*, pages 1–26, 1988.

[Lev89] H. Levesque. Logic and the complexity of reasoning. Technical Report KRR-TR-89-2, University of Toronto, 1989.

[MBF77] C. Morris, J. Bransford, and J. Franks. Levels of processing versus transfer appropriate processing. *Journal of verbal learning and verbal behaviour*, 16:519–533, 1977.

[MC90] A.J. Katz Gately M.T. and D.R. Collins. Robust classifiers without robust features. *Neural computation*, 2:472–479, 1990.

[McC65] J. McCullers. Type of associative interference as a factor in verbal paired-associate learning. *Journal of verbal learning and verbal behaviour*, 4:12–16, 1965.

[MD85] R. Mooney and G. DeJong. Learning schemata for natural language processing. In *Proceedings of IJCAI 1985*, pages 681–687. Los Angeles, CA: Morgan Kaufmann, 1985.

[Mic80] R.S. Michalski. Knowledge acquisition through conceptual clustering: A theoretical framework and algorithm for partitioning data into conjunctive concepts. *International journal of policy analysis and information systems*, 4:219–243, 1980.

[Mic93] R.S. Michalski. Inferential theory of learning as a conceptual basis for multistrategy learning. *Machine Learning*, 11:111–151, 1993.

[Mit80] T.M. Mitchell. The need for biases in learning generalizations. Technical Report CBM-TR-117, Rutgers University, 1980.

[Mit83] T. Mitchell. Learning and problem solving. computers and thought lecture. In *Proceedings of the 8th IJCAI*, pages 1139–1151. San Mateo, CA: Morgan Kaufmann, 1983.

[MR81] J. McClelland and D. Rumelhart. An interactive activation model of the effect of context in perception: Part 1 an account of basic findings. *Psychological Review*, 88:375–407, 1981.

[MS78] D. Medin and L. Schaffer. Context theory of classification learning. *Psychological review*, 85(3):207–238, 1978.

[MW87] D. Medin and W. Wattenmaker. Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology*, 19:242–279, 1987.

[Nei89] W.T. Neill. Lexical ambiguity and context: an activation-suppression model. In D.S. Gorfein, editor, *Resolving Semantic Ambiguity*, pages 21–72. New York, NY: Springer Verlag, 1989.

[Paz93] M. Pazzani. Learning causal patterns: Making a transition from data-driven to theory-driven learning. *Machine Learning*, 11:173–194, 1993.

[Pea88] J. Pearl. *Probabilistic reasoning in intelligent systems: Networks of plausible inference.* San Mateo, CA: Morgan Kaufmann, 1988.

[PG90] T. Poggio and F. Girosi. Networks for approximation and learning. *Proceedigs of the IEEE*, 78(9):1481–1497, 1990.

[Pol83] J. Pollock. A theory of direct inference. *Theory and Decision*, 15:29–96, 1983.

[Pol84] J. Pollock. Foundations for direct inference. *Theory and Decision*, 17:221–256, 1984.

[Poo91] D. Poole. The effect of knowledge on belief:conditioning, specificity and the lottery paradox in default resoning. *Artificial Intelligence*, 49:281–307, 1991.

[Qui83] R. Quinlan. Learning classification procedures and their application to chess end games. In *Machine Learning, R. Michalski and J. Carbonell and Mitchell, T. Eds.*, pages 463–482. Morgan Kaufmann, 1983.

[Qui86] J.R. Quinlan. Induction of decision trees. *Machine learning*, 1:81–106, 1986.

[Qui87a] J.R. Quinlan. Generating rules from decision trees. In *International joint conference on artificial intelligence*, pages 304–307, Milan, Italy, 1987. San Mateo, CA: Morgan Kaufmann.

[Qui87b] J.R. Quinlan. Simplifying decision trees. *International journal of man-machine studies*, 27(3):221–234, 1987.

[Qui89] J.R. Quinlan. Unknown attribute values in induction. In *The proceedings of the 11th IJCAI*, pages 164–168. Morgan Kaufmann, 1989.

[RB87] H. Roediger and T. Blaxton. Effects of varying modality, surface features, and retention intervals on priming in word-fragment completion. *Memory and Cognition*, 15:379–388, 1987.

[Rei49] H. Reichenbach. *Theory of probability.* Los Angeles, CA: University of California Press, 1949.

[Rei80] R. Reiter. A logic for default reasoning. *Artificial intelligence*, 13:72–105, 1980.

[RHW86] D.E. Rumelhart, G.E. Hinton, and R.J. Williams. Learning represenations by back-propogating errors. In D.E. Rumelhart, G.E. Hinton, and R.J. Williams, editors, *Neurocomputing: foundations of research*, pages 696–699. Cambridge, MA: MIT Press, 1986.

[RJ86] D. Rumelhart and McClelland J. *Parallel distributed Processing: Volume I.* Cambridge, MA: MIT press, 1986.

[Ros78] E.H. Rosch. Principles of categorization. In E. Rosch and B.B. Lloyd, editors, *Cognition and Categorization*, pages 12–78. Hillsdale, NJ: Erlbaum, 1978.

[RR89] S. Renals and R. Rohwer. Phoneme classification experiments using radial basis functions. In *Proceedings of the international joint conference on neural networks*, volume 1, pages 461–467. San Mateo, CA: Morgan Kaufmann, 1989.

[RS83] E.M. Roth and E.J. Shoben. The effect of context on the structure of categories. *Cognitive Psychology*, 15:346–378, 1983.

[SA77] R. Schank and R. Abelson. *Scripts, Plans, Goals, and Understanding.* Hillsdale, NJ: Lawrence Erlbaum, 1977.

[Sal90] S. L. Salzberg. *Learning with nested generalized exemplars.* Boston, MA: Kluwer academic publishers, 1990.

[Sal91] S.L. Salzberg. A nearest hyperrectangle learning method. *Machine Learning*, 6:251–276, 1991.

[SBN93] L. Saitta, M. Botta, and F. Neri. Multistrategy learning and theory revision. *Machine Learning*, 11:153–172, 1993.

[Sch91] C. Schaffer. Overfitting avoidance as bias. In *Proceedings of the IJCAI Workshop on Evaluating and Changing Representation in Machine Learning*, pages 24–30, Sydney, Australia, 1991. San Mateo, CA: Morgan Kaufmann.

[SCS77] D. Scarborough, C. Cortese, and H. Scarborough. Frequency and repition effect in lexical memory. *Journal of experimental psychology: Human perception and performance*, 3:1–17, 1977.

[SG89] D. Schacter and P. Graf. Modality specification of implicit memory for new association. *Journal of experimental psychology: Learning, memory and cognition*, 15:3–12, 1989.

[SK89] G.B. Simpson and G. Kellas. Dynamic contextual processes and lexical access. In D.S. Gorfein, editor, *Resolving Semantic Ambiguity.* New York, NY: Springer Verlag, 1989.

[Sla91] S. Slade. Case-based reasoning: A research paradigm. *AI magazine*, 12(1):42–55, 1991.

[SM82] R.L. Scheaffer and J.T. McClave. *Statistics for engineers*. Boston, MA: Duxbury Press, 1982.

[SMT91] J.W. Shavlik, R.J. Mooney, and G.G. Towell. Symbolic and neural learning algorithms: an experimental comparison. *Machine Learning*, 6(2):111–143, 1991.

[SN91] K.P. Sycara and D. Navinchandra. Index transformation techniques for facilitating creative use of multiple cases. In *Proceedings of 12th IJCAI*, pages 347–352. San Mateo, CA: Morgan Kaufmann, 1991.

[Sub90] D. Subramanian. A theory of justified reformulations. In D.P. Benjamin, editor, *Change of representation and inductive bias*, pages 140–161. MIT Press, 1990.

[SW86] C. Stanfill and D. Waltz. Toward memory-based reasoning. *Communciations of the ACM*, 29(12):1213–1228, 1986.

[Tho87] C. Thornton. Hypercuboid formation behaviour of two learning algorithms. In *Proceedings of IJCAI-87*, pages 301–303, Milan, Italy, 1987. San Mateo, CA: Morgan Kaufmann.

[TT73] E. Tulving and D. Thomson. Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, 80:352–373, 1973.

[Tul72] E. Tulving. Episodic and semantic memory. In E.Tulving and M. Donaldson, editors, *Organization of memory*, pages 23–89. New York, NY: Academic press, 1972.

[Tul76] E. Tulving. Ecphoric processes in recall and recognition. In J. Brown, editor, *Recall and recognition*, pages 1–21. London: Wiley, 1976.

[Tul83] E. Tulving. *Elements of episodic memory*. New York: Oxford, 1983.

[Tul85] E. Tulving. How many memory systems are there? *American psychologist*, 40:385–398, 1985.

[Tul86] E. Tulving. What kind of hypothesis is the distinction between episodic and semantic memory? *Journal of experimental psychology: Learning, memory and cognition*, 12:307–311, 1986.

[Tur92] P. Turney. Exploiting context when learning to classify, submitted to iea/aie 93. Personal correspondence, 1992.

[Utg84] P. Utgoff. *Shift of bias for inductive concept learning.* Boston, MA: Kluwer academic publishers, 1984.

[Val84] L.G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[WB88] B. Whittlesea and L. Brooks. Critical influence of particular experiences in the perception of letters, words, and phrases. *Memory and Cognition*, 16(5):387–399, 1988.

[Win75] P. Winston. Learning structural descriptions from examples. In P. Winston, editor, *The psychology of computer vision.* New York, NY: McGraw-Hill, 1975.

[Wit80] L. Wittgenstein. *Remarks on the philosophy of psychology.* Oxford: Basil Blackwell, 1980.

[WK90] S.M. Weiss and I. Kapouleas. An empirical comparison of pattern recognition, neural nets, and machine learning classification methods. In *Proceedings of 11th IJCAI*, pages 781–787. San Mateo, CA: Morgan Kaufmann, 1990.

[WR87] M. Weldon and H. Roediger. Altering retrieval demands reverses the picture superiority effect. *Memory and Cognition*, 15:269–280, 1987.

[WW75] M. Watkins and M. Watkins. Buildup of proactive inhibition as a cue-overload effect. *Journal of experimental psychology: Human learning and memory*, 1:442–452, 1975.

[Yeu91] D. Yeung. A neural network approach to constructive induction. In *Proceedings of the 8th international conference on machine learning.* San Mateo, CA: Morgan Kaufmann, 1991.