Learning in the Real World Environment: A classification Model Based on Sensitivity to Within-Dimension and Between-Category Variation of Feature Frequencies

by

Newman M. K. Lam
B. Comm., University of British Columbia, 1978
M.P.A., University of Victoria, 1982

A Dissertation Submitted in Partial Fulfillment of the Requirement for the Degree of

DOCTOR OF PHILOSOPHY

in the School of Public Administration

ACCEPTED
FACULTY OF GRADUATE STUDIES
DATE 1991-05-15 DEAN

We accept this thesis as conforming to the required standard

Dr. J. MacGregor, Supervisor (School of Public Administration)

Dr. B. Cunningham, Department Member (School of Public Administration)

Dr. M. Masson, Outside Member (Department of Psychology)

Dr. B. Johnson, Outside Member (Department of Mathematics)

Dr. B. Schaefer, Outside Member (Acquired Intelligence Inc.)

Dr. D. Russell, External Examiner (Clinical Neuropsychologist)

© Newman M. K. Lam, 1991

University of Victoria

All rights reserved. Thesis may not be reproduced in whole or in part, by mimeograph or other means, without the permission of the author.

Supervisor: Dr. James MacGregor

Abstract

Research on machine learning has taken numerous different directions. The present study focussed on the micro-structural characteristics of learning systems. It was postulated that learning systems consist of a macro-structure which controls the flow of information, and a micro-structure which manipulates information for decision making. A review of the literature suggested that the basic function of the micro-structure of learning systems was to make a choice among a set of alternatives. This decision function was then equated with the task of making classification decisions. On the basis of the requirements for practical learning systems, the feature frequency approach was chosen for model development. An analysis of the feature frequency approach indicated that an effective model must be sensitive to both within-dimension and between-category variations in frequencies. A model was then developed to provide for such sensitivities. The model was based on the Bayes' Theorem with an assumption of uniform prior probability of occurrence for the categories. This model was tested using data collected for neuropsychological diagnosis of children. Results of the tests showed that the model was capable of learning and provided a satisfactory level of performance. The performance of the model was compared with that of other models designed for the same purpose. The other models included NEXSYS, a rule-based system specially design for this type of diagnosis, discriminant analysis, which is a statistical technique widely used for pattern recognition, and neural networks, which attempt to simulate the neural activities of the brain. Results of the tests showed that the model's performance was comparable to that of the other models. Further analysis indicated that the model has

certain advantages in that it has a simple structure, is capable of explaining its decisions, and is more efficient than the other models.

Examiners:

Dr. J. MacGregor, Supervisor (School of Public Administration)

Dr. B. Cunningham, Department Member (School of Public Administration)

Dr. M. Masson, Outside Member (Department of Psychology)

Dr. B. Johnson, Outside Member (Department of Mathematics)

Dr. B. Schaefer, Outside Member (Acquired Intelligence Inc.)

Dr. D. Russell, External Examiner (Clinical Neuropsychologist)

Table of Contents

List of Tables

Acknowledgements

I would like to thank Dr. Alex Bavelas, who supervised me until his retirement, for the inspiration and encouragement he had given me without which I would not have initiated this study. I must thank Dr. James MacGregor, supervisor of my dissertation, for providing me the insight and direction necessary for completing this study. I am indebted to Dr. Brian Schaefer and Dr. Diane Russell whose study provided the data for testing the model I have developed.

Introduction

The pursuit of artificial intelligence (AI) has been divided into two main streams. In one stream, the emphasis has been on modelling human expertize through knowledge engineering. The knowledge engineering approach involves translating the decision-making knowledge of human experts into mechanical rules. The resulting products are generally referred to as expert systems. In the other stream, research emphasis has focused on the acquisition of knowledge through learning. There are some important differences between these two streams. The knowledge engineering approach produces machines that imitate human reasoning. The resulting AI systems consist of heuristics typified by "if-then" rules. The machine learning approach, as the other stream is often referred to, is based on learning algorithms. Learning algorithms do not necessarily imitate human reasoning, despite the fact that many of them may have implications for human learning.

The knowledge engineering approach has had some successes in practical applications. However, it is time-consuming, costly and has made relatively limited theoretical contributions. The systems it produces tend to be complicated in structure and lack flexibility for modification. The machine learning approach, conversely,

has a strong theoretical base and its results may lend insight into the constitution of intelligence. The weakness of this approach, however, is its limited success in practical application (Carbonell, Michalski and Mitchell, 1983).

Theory and practice often complement each other. Theory provides the basis for application development. The results of application development, in turn, provide insight for furthering theoretical exploration. There is a limit to application development, however, if there is no continuous theoretical contribution. In this paper, a theoretical approach is taken for developing learning models. The literature on machine learning provides a large variety of structures for developing learning systems. Some structures are suitable for complex problems which can only be solved by a chain of decisions, with the conditions for each decision affected by its preceding decisions. These structures are usually represented by a complex configuration of events which chart the decision paths available for solving the problems. On the other hand, there are structures designed for simple, basic problems which can be solved with a single decision. For example, the classification task is a typical example of simple problems. Simple problems generally require structures that

are relatively simple. In this paper, it is postulated that a complex problem can be broken down into a sequence of related but simpler problems. Similarly, a complex structure can also be broken down into a number of simpler structures, with the complex structure representing the organization of its simpler components. The complex structure, in this case, is referred to as the macro-structure of a system while its simpler components are called the micro-structures. In biology and physics, it has been shown that micro-structures of different systems, if not identical, may be highly similar. The same is assumed to be true for learning systems. Similar to the cells in biological systems, the micro-structure is considered to be the basic component for building learning systems.

This paper is divided into two major parts. The first part provides the literature research leading to the development of a learning model. In the second part, the framework, assumptions and pre-requisites of the proposed learning model are explained and tests of the proposed model reported, where its performance was compared with those of models designed for the same purpose.

The literature research is reported in Sections 1, 2 and 3 of this paper. In Section 1, a historical review of

learning machines is presented. This section allows us to gain an understanding of the types of research that have been conducted in machine learning. In Section 2, the basic issues of machine learning are examined. These basic issues include (i) methods for knowledge representation, (ii) functions of learning systems, and (iii) selection of knowledge for acquisition. The analysis of these issues leads to an understanding of how knowledge can be acquired through learning. With this understanding, the basic function of learning systems is defined. This basic function, which is typified by classification tasks, is assumed to be the function of the micro-structure in learning systems. In Section 3, literature concerning classification tasks is examined. This includes literature on concept learning, pattern recognition and parallel distributed processing.

Part II of this paper is divided into four sections beginning with Section 4 of the paper. In Section 4, a framework for constructing learning systems is defined with reference to the literature review. In Section 5, the decision rules enabled by the framework are explored. This leads to the proposed model which is also explained in Section 5. In Section 6, the system requirements of the proposed model are explained in detail. In Section 7, tests

of the model are reported. The paper concludes with Section 8, where findings of the paper are re-examined and ideas for further exploration discussed.

Part I: Literature Research

1. Historical Perspective

Scientific research in machine learning can be divided roughly into four periods with each period dominated by one school of thought. Initial research interest for machine learning stemmed from findings in biophysics, followed by periods dominated in turn by pattern recognition, concept learning, and production rules. The development reflects a movement from imitating the biophysical structure of the brain to imitating human behaviour. The following discussion provides an abstract of research on machine learning. Models that are relevant to the present study are explored in more details in Section 3 of this paper.

1.1 Biophysical Approach

Research interest for machine learning arose from findings on the machine-like activities of the neuron. Fascinated by the machine-like characteristics of the neuron, scientists began to design systems to imitate its activities (McCulloch and Pitts, 1943; Rashevsky, 1948). These systems were generally referred to as "neural nets" or "self-organizing systems". A distinguishing feature of such systems is that they begin with little or no task-oriented knowledge, and with random or partially random initial structures.

Learning in these systems is characterized by incremental changes in the likelihood of neuron-like elements to transmit a signal (Rosenblatt, 1958).

The majority of research in neural nets was conducted during the period from the early 1940's to the mid 1960's. Due to the primitive nature of computer technology at the time, most of this research was either theoretical (Rashevsky, 1948) or involved specially constructed hardware (Rosenblatt, 1958; Selfridge, 1959; Widrow, 1962). Notable research can also be found in the work of Ashby (1960), Block (1961), Yovitz (1962), Culberson (1963), Kazemierczak (1963) and Minsky and Papert (1969). Related research can be found in the work of Friedberg (1958, 1959) and Holland (1980) on the simulation of evolutionary processes.

There were high expectations for this line of research in the beginning. However, these expectations were not met and research on neural nets began to decline. Beginning from the mid 1980's, however, there was a resurgence of interest in neural nets, largely due to research on parallel distributed processing (McClelland, Rumelhart & Hinton, 1986) and back-propagation of errors (Rumelhart, Hinton & Williams, 1986a, 1986b). Incidentally, these models are no longer referred to as neural nets but as neural networks,

connectionist systems or parallel distributed processing systems.

1.2 Pattern Recognition

In the 1960's, developments in digital computer technology allowed complex patterns to be mechanically sensed, decoded, stored, retrieved and mathematically manipulated. This stimulated research on pattern recognition, which examined processes by which sensory signals are converted into meaningful perceptual experience. In pattern recognition, emphasis was on acquiring discriminant functions which separate patterns of different classes (Nilsson, 1965; Koford, 1966; Highleyman, 1967). In some models, statistical decision theories were used to derive discriminant functions (Watanabe, 1960; Sebestyen, 1962; Fu, 1968; Arkadev, 1971; Fukanagan, 1972; Duda & Hart, 1973; Kanal, 1974). A number of models were capable of adjusting their parameters to maintain stable performance in the presence of disturbance (Truxal, 1955; Davis, 1970; Mendel, 1970; Tsypkin, 1968, 1971, 1973; Fu, 1971; Fu & Tou, 1974).

1.3 Concept Learning

Shortly after pattern recognition entered the research scene, another discipline, referred to as concept learning,

began to emerge and gain attention. In concept learning, emphasis is on the human ability to formulate concepts from experience. Unlike pattern recognition which is based on numerical or statistical methods, models in concept learning utilized logic, graph, and symbolic descriptions for problem solving. Models in concept learning have been divided into four major categories: (1) hypothesis testing, (2) feature frequency, (3) prototype and (4) exemplar (Smith and Medin, 1981). Hypothesis testing models assume that concept can be learned by sampling, testing and eliminating hypothesis (Millward & Spoehr, 1973; Bower and Trabasso, 1966). Feature frequency models suggest that feature frequency provides the necessary information for concept learning (Kellogg, 1980a, 1980b, 1980c, 1981). Prototype models propose that concepts are represented by abstract prototypes which may not exist in reality (Beach, 1964; Reed, 1972, 1973; Rosch, 1975, 1978). Exemplar models suggest that concepts are represented by specific cases which are stored intact in memory (Medin and Schaffer, 1978; Medin, 1989).

1.4 Production Rules

In the 1980's, with the advancement of high-level programming technology, artificial intelligence became increasingly rule-based, with knowledge represented by decision rules rather than by numerical or symbolic

patterns. In these systems, learning was equated with the acquisition of task-specific decision rules. Learning models were no longer characterized by a single algorithm, but by systems of intricate rules. Unlike the earlier models which required little or no initial knowledge, these AI systems required a large amount of domain-specific knowledge to begin with. The initial knowledge had to be gained from human experts. Since human knowledge is often fuzzy rather than clearly defined, knowledge acquisition has been the "bottleneck" in the development of such systems (Feigenbaum, 1979; Johnson, 1983). Computer systems have been developed to facilitate knowledge acquisition by inductive inferences and to provide automatic explanation of the run-time reasoning (e.g. Muggleton, 1986). These systems perform the task of generalizing, specializing or grouping expert knowledge into a form that is more meaningful and manageable. There has also been research interest in expanding the initial knowledge base through a variety of learning techniques including inductive reasoning (Michalski, 1983), conceptual clustering (Michalski and Stepp, 1983) and analogical reasoning (Carbonell, 1983). Some models have facilitated learning through continuous interaction between the mechanical system and a human instructor (Rychener, 1983).

2. Basic Issues in Learning

Learning is a word very loosely defined in the English language. Simon (1983) provided a definition for "learning" which states that learning denotes changes in a system "that are adaptive in the sense that they enable the system to do the same task or tasks drawn from the same population more efficiently and more effectively the next time". Carbonell, Michalski and Mitchell (1983) suggested that learning can occur in two basic forms: knowledge acquisition and skill refinement. Knowledge acquisition is characterized by the conscious process through which skills are acquired. Skill refinement refers to the sub-conscious process through which the acquired skills are refined through repeated practice. The word "skill" refers to both physical as well as mental skills. The skill learned by a swimmer, for example, is more physical in nature. The skill for decision making, on the other hand, relies more on mental capacity. Whether physical or mental, learning begins with the acquisition of knowledge. With the exception of genetic knowledge, knowledge is generally gained either through interaction with the environment or through introspection. The knowledge gained from learning will be lost if it cannot be stored. The method a system uses to store knowledge is likely to reflect the type of knowledge it is designed to

gain. By examining the internal representation of knowledge in different types of systems, the fundamental characteristics of learning may be revealed.

2.1 Knowledge Representations

There are a large variety of methods for knowledge representation. In the simplest form, knowledge is stored as a set of numerical values. In the most complicated form, knowledge is represented by encoded programs. The list below shows the commonly used techniques for knowledge representation.

(1) Frequency matrices or graphs - Knowledge is represented by a set of numerical values indicating the frequency of occurrence of certain features. The feature frequencies, as these numerical values are called, are either stored in the form of a matrix or represented by a graph. This method of knowledge representation is commonly used in the feature frequency models of concept learning (Bourne et. al. 1976).

(2) Samples - Knowledge is represented by samples of a category. This method of knowledge representation is used in the prototype and exemplar approaches

to concept learning (see Sections 3.1.3 and 3.1.4). In the prototype approach, there is only one sample per category represented by the prototype of the category (Rosch, 1975). In the exemplar approach, samples are represented by all unique cases in a category (Medin, 1975, 1983).

(3) Algebraic expressions - Knowledge is represented by a set of algebraic expressions of a fixed functional form. This method of knowledge representation is frequently used in models for pattern recognition (see Section 3.2). The algebraic expression represents a plane which separate cases of two categories when cases of both categories are plotted in a multi-dimensional space.

(4) Hierarchical trees - Knowledge is represented as a hierarchical structure of concepts. Concepts can be the names of physical objects (e.g. apple), events (e.g. wedding), abstract terms (e.g. mathematics), or the category names for groups of objects (e.g. fruit), events (e.g. ceremony) or abstract terms (e.g. knowledge). Generally, the hierarchical structure assumes the form of an

inverted tree (Sacerdoti, 1974; Rosch, 1975). The hierarchical tree indicates the ordinal relationships among a set of concepts.

(5) Networks - This method of representation is very similar to that of hierarchical trees. Knowledge is represented by the relationship among a set of concepts. Instead of relating concepts by their ordinal structure, concepts are related by their spatial distance in a multi-dimensional network (Young, 1968; Quillian, 1968; Findler, 1979).

(6) Formal logic-based expressions and related formalisms - Knowledge is represented in the form of propositions, arbitrary predicates, finite-valued variables, statements restricting ranges of variables, or embedded logical expressions (Newell, Shaw & Simon, 1963; Nilsson, 1971). This form of knowledge representation is used for deductive inference. The logic-based expressions are modified until a targeted form is reached while preserving a set of premises which is assumed to be true.

(7) Formal grammar-based expressions - This form of knowledge representation bears certain similarities to logic-based expressions. The difference is that the transformation of grammar-based expressions is governed by grammatical rather than by logical rules. Knowledge is represented in the form of regular expressions, finite-state automata, context-free grammar rules, or transformation rules (Hopcroft & Ullman, 1969).

(8) Decision trees - This form of knowledge representation bears certain structural similarities to hierarchical trees. Its function however is quite different. This form of representation is based on decision theories. Knowledge is represented as a tree structure branching out at places where there are either multiple decision alternatives or event possibilities. The decision tree, therefore, does not indicate the ordinal relationship among concepts. Instead, it represents the consequential relationship among a set of pre-conceived decision alternatives and probable events. The decision tree is characterized by decision nodes at which branches of alternate

actions diverge, event nodes from where probable events spread out, event probabilities which indicate the chance of each event to occur, and outcome values which indicate the final consequences of different combinations of actions and events (Raiffa, 1970; Thompson & Thompson, 1986).

(9) Production rules - A production rule is characterized by a condition action pair (C => A), where C is a set of conditions and A is a sequence of actions. If all the conditions in a production rule are satisfied, the sequence of actions is then executed. Due to their simplicity and ease of interpretation, production rules are widely used in expert systems (Davis & King, 1977; McDermott & Forgy, 1978; Young, 1979).

(10) Frames and schemas - Frames and schemas provide a larger unit of representation than production rules. There are cases that condition-action pairs cannot be treated independently but have to be manipulated as a unified plan. Frames and schemas are designed to make such manipulation possible. Frames and schemas are characterized by

a set of relational slots with each slot having a prescribed role in the representation (Kuipers, 1975; Anderson, 1979). This form of representation is useful for the mapping and generalization of action plans. The internal structure of frames and schemas can be arbitrarily complex and consequently difficult to maneuver.

(11) Encoded programs - Knowledge is represented as encoded computer programs. This is the least flexible of all representations. There are program development systems which create encoded programs according to a set of input-output specifications. These systems are made for production rather than for learning, and exhibit no learning ability. Conceptually, it is possible for a program to rewrite its own codes but the author has not seen such a program yet.

2.2 Basic System Functions

Based on the design of knowledge representations, the basic functions of learning systems can be divided into the following categories:

(1) Classification systems - The basic function of these systems is to determine the category membership of new or unknown cases. Category membership is determined by strength of association or degree of similarity. For example, if case "X" has a stronger association with category "A" than with category "B", then "X" is deemed to be more likely a member of "A" than a member of "B". Frequency matrices, graphs, samples, and algebraic expressions are forms of representation that provide the means for assessing category membership. These forms of representation are very simple in structure and have few prerequisites. The methods for deriving these representations vary, however, in principle and in complexity.

Frequency matrices, graphs and algebraic expressions are derived through statistical

manipulation of input data. Frequency matrices and graphs are based on the assumption that the differences between categories are reflected by the presence or absence of some distinct features. Category membership can, therefore, be determined by some form of frequency analysis. Algebraic expressions, such as discriminant functions, require the analysis of variance for their derivation and are more sophisticated in form and in the use of statistical tools. Representing categories by their samples is a technique significantly different in principle. This approach assumes that the characteristics of a category are best described by a set of samples. Category membership should, therefore, be determined by some form of analysis involving the samples.

(2) Search systems - The basic function of these systems is to search for information in either a hierarchical structure or a multi-dimensional space. Hierarchical trees and networks are the respective forms of knowledge representation. The search begins with some input data which specify the purpose of the search. For example, the

statement "A robin is a bird" could be the input data. In a hierarchical tree, the search usually begins from the top and spreads downward until the target object "robin" is found. The system then back tracks to determine whether "robin" is a member of the category "bird".

Searching in a hierarchical structure can be seen as a sequence of classification tasks, with each step moving from categories that are more general in nature to categories that are more specific. For example, the search for "robin" may begin with decisions between very general categories, such as "living things" versus "non-living things". The task therefore is to determine which category "robin" belongs to. If "living things" is chosen and "plants" and "animals" are its sub-categories, another classification is required to determine whether robins are plants or animals. The classification continues and back-tracks if necessary until either the object "robin" is found or a conclusion is reached that "robin" is not available in the structure.

In a network structure, the difference between two objects is reflected by the spatial distance between them. For example, if the statement "Oranges are sweeter than lemons" is to be evaluated, the spatial distance between "orange" and "sweetness" has to be compared with the distance between "lemon" and "sweetness". To begin with, the criterion "sweetness" has to be found first. From this point, the search spreads out in the network until both "orange" and "lemon" are located. Once all three objects are located, their spatial relationship can be evaluated.

To search blindly in a network would be inefficient if not impractical. An efficient network system require some form of algorithm to guide the search process. The algorithm must provide the ability to decide, at every point in the search space, among the available paths, which path has the best chance of leading to the target object. The algorithm therefore must have the ability to assess the strength of association between each available path and the target object.

(3) Action systems - The basic function of these systems is to select a course of action based on some input information. Decision trees, production rules and frames and schemas are knowledge representations appropriate for these systems.

In decision trees, the outcomes of each combination of decisions and events are given. Each outcome is weighted compoundedly by the probabilities of the events leading to the outcome. A course of action is determined by selecting the combination of actions which provides the best expected return. The performance of a decision tree depends largely on the accuracy of the probabilities estimated for the given events. The probability for an event to occur is assessed with reference to the situation given by the input data. Therefore, the estimated probability for an event is, in essence, indicative of the strength of association between the given situation and the event. For example, if event "A" has an 80% chance of occurring while event "B" has only a 20% chance when situation "X" is given, it implies that event "A" is more

strongly associated with situation "X" than event "B".

Production rules, frames and schemas are rarely used as stand-alone techniques. They are used usually in conjunction with structural types of representation such as hierarchical trees or networks. Production rules, frames and schemas provide the means for representing the relationship between conditions and actions. Production rules are characterized by pair-wise relationship between a set of conditions and a sequence of actions. To select a production rule, a set of given conditions, either provided by the input data or derived from the system's operation, is compared with the conditions prescribed in a set of production rules. The rule with conditions most similar to the given ones will be chosen. In other words, the selection is made according to the strength of association between the given conditions and the conditions prescribed in the production rules.

Frames and schemas provide larger units of representation than production rules and are

widely used for representing action plans. In frames and schemas, conditions and actions are represented relationally in intricate structures. The process for selecting frames and schemas, however, is fundamentally similar to that of production rules. For example, in learning by analogy, a new action plan can be derived by comparing the structural, functional and conditional requirements of the new plan against those of existing plans. An existing plan with the most similarity to the new requirements is chosen and modified to formulate the new plan.

(4) Transformation systems - The basic function of these systems is to transform an expression from an original form to a target form. Logic-based expressions are transformed using a set of truth-preserving rules. Grammar-based expressions are transformed using grammatical rules. While transformation rules can be randomly applied to a given expression, the chance of it reaching the target form would be low. As with hierarchical trees and networks, some algorithm is required to guide the transformation process. The algorithm would involve comparing the current state of the

expression against its target form with respect to the available transformation rules.

2.3 Function of Micro-structure

In Section 2.1, it was shown that some structures for knowledge representation are higher level representations than others. The higher level representation usually indicate the organization of knowledge in a system. The lower level representations provide facilities for decision making. For example, in previous discussion, a search system is described as a hierarchical organization of classification tasks. At the higher or macro level, the system is represented by a hierarchical tree with knowledge stored in its nodes. At the lower or micro level, the knowledge stored in the nodes is represented by a structure suitable for classification tasks. The knowledge, for instance, can be represented by a frequency matrix or a graph. Problems are solved by moving around in the macro structure until a conclusion is reached. The movement is made by comparing contents of the nodes. The process begins at the macro level by choosing a starting point in the macro structure. Once a starting point is located, the focus is moved to the micro level where contents of the starting point and the nodes linking to it are revealed. Based on certain decision rules which compare contents of the nodes, a node will be chosen as the next step. From the chosen node, a further movement will be made by comparing contents

of the chosen nodes and the nodes linked to it. These steps will be repeated until the problem is solved or a conclusion is reached that the problem cannot be solved by the current structure.

In biological and physical systems, the microscopic elements of systems have structures that are either highly similar or identical. A simple example would be the cells in a living being or the atoms in a compound. In human organizations, the same is also true. Large human organizations are characterized by a hierarchical arrangement of departments, within each department a hierarchical arrangement of branches, and within each branch a hierarchical arrangement of personnel. It may be reasonable to assume that the same is also true for learning systems. In other words, the structures of micro-elements in learning systems may be essentially the same.

It would be difficult to understand a complex system without first understanding the function of its micro-elements. If we assume the micro-structures of learning systems to be the same, we may also assume their functions to be similar. If so, disregarding their designs and purposes, learning systems should show similar functionality at the micro level. In Section 2.2, it was indicated that the basic

function of learning systems involves making a choice among a set of alternatives. For example, in classification systems, a choice has to be made among a number of given categories. In search systems, a choice has to be made at every point of diversion to determine the next path to take in the searching process. In action systems, a combination of actions, a production rule or an action plan has to be chosen to derive a course of actions. In transformation systems, a transformation rule has to be selected at every point a transformation is required. For simple systems such as classification systems, a single choice represents the solution to a problem. In complex systems such as search systems, choices are continuously made until a conclusion is reached. Since the ability to choose among alternatives is central to all learning systems investigated in the present study, it is reasonable to assume that this ability is the function of micro-structures in learning systems.

3. Classification Models

In the previous discussion, the ability to choose among alternatives is assumed to be the function of micro-structures in learning systems. The act of choosing among alternatives is exemplified in classification tasks. In the following sub-sections, classification tasks are examined in the following fields: concept learning, pattern recognition and parallel distributed processing (neural networks).

3.1 Concept Learning

The bulk of literature on classification tasks was found in the area of concept learning. Concept learning examines how concepts are formulated through the observation of examples. A concept is considered to have been learned when the learner can successfully classify new cases of the concept correctly, or when the learner can correctly identify the concept rule. Literature in concept learning is concerned more with human learning than with machine learning. Research on concept learning usually involves human subjects sorting case descriptions (stimuli), into two or more pre-determined categories. A case description may be a string of letters, a stick figure, a sequence of numbers, a configuration of dots, and so on, depending on the needs of

the experiment. The case descriptions must contain some distinguishable features which indicate the differences between concepts. Models of concept learning have been divided into four major categories: hypothesis testing, feature frequency, prototype and exemplar models.

3.1.1 Hypothesis Testing Approach

The hypothesis testing approach, also known as the classical approach, is based on the assumption that concepts are distinguished by the presence or absence of some defining features. Concepts, therefore, can be learned through sampling, testing and elimination of hypotheses which postulates what the defining features are (Millward & Spoehr, 1973). Typically, this involves setting up a number of unique hypotheses each specifying one or a number of features as the defining features. A hypothesis is eliminated when at least one of the defining features it postulates is not found in all cases of the concept. This process of elimination continues until one or more hypotheses are found which correctly classify all cases of the concept.

The conventional approach in hypothesis testing rejects the notion that concepts are learned through inductive reasoning, despite the fact that some models suggest that hypotheses may be sampled on the basis of perceptual

salience. The hypothesis testing approach has been well supported by studies which used well-defined concepts. Well-defined concepts are those that can be defined either singly by one feature or jointly by a set of features. The strongest evidence for this approach is provided by Bower and Trabasso (1966). In this study, the rule for defining a concept was changed half way through an experiment by shifting the defining feature from one dimension to another. Subjects of the experiment were not informed of the change. Bower and Trabasso claimed that if deductive reasoning was used, subjects would detect the change very quickly. If inductive reasoning was used, performance would be impaired for a period of time until the change became evident. After the change, subjects' performance was hindered only temporarily. Bower and Trabasso therefore concluded that deductive rather than inductive reasoning was used in human learning. However, subjects in their study were instructed explicitly to look for defining features. The instruction given to the subjects, in this case, might have pre-determined the subjects' learning strategies. The reliability of their findings, therefore, is questionable.

Criticism of the hypothesis testing approach has been provided by Kellogg (1980a, 1980b, 1980c, 1981) and his associates (Kellogg, Bourne & Ekstrand, 1978; Kellogg,

Robbins & Bourne, 1978). Based on Rosch's (1973a, 1973b, 1978) proposition that natural categories are ill-defined, Kellogg argued that the hypothesis testing approach was inadequate to account for all situations in human learning. In Kellogg (1980b), it was shown that bi-conditional concepts can never be learned using hypothesis testing. McClosky and Glucksberg (1978) also provided evidence to support the ill-defined nature of natural categories (for further discussion on this topic, see Section 3.1.3). With increasing interest in natural categories, attention to the hypothesis testing approach has been declining in the last twenty years.

3.1.2 Feature Frequency Approach

In contrast to the hypothesis testing approach, feature frequency models are based on inductive learning. Feature frequency models assume that feature frequencies are calculated and stored in human learning. The idea behind this assumption is that features salient to a category will have a frequency distribution skewed in favor of that category (Barresi, Robbins & Dhain, 1975; Dougherty, 1978). Conversely, features that are irrelevant to any categories will have an even distribution of frequencies. Furthermore, features dissociated with a category will have a very low frequency of occurrence in the concept category. The

characteristics of a category can therefore be represented by its pattern of feature frequencies. The similarity between a category and a new case can be determined by summing the frequencies of features in common between the category and the new case (Neumann, 1974; Kellogg, Bourne and Ekstrand, 1978). The membership of a new case can then be determined by selecting the category which has the highest similarity with the new case.

Research on feature frequency models has been mainly empirical in nature. The research emphasis has been on demonstrating that feature frequencies are attended to in human learning. A number of techniques have been used to collect empirical evidence, including:

(1) Subject ratings of the typicality of items - Subjects are asked to rate the typicality of given cases to a concept category. The assumption is that cases with higher similarity with the category will be rated as more typical.

(2) Order in which category items are learned - The assumption is that cases with higher similarity with the category will be learned first.

(3) Verification time for category membership - The assumption is that the higher the similarity with

the category, the shorter will be the verification time.

(4) Probability of item output - Subjects are asked to list members of a category. The assumption is that those with higher similarity with the category will be listed more often.

(5) Estimation of feature frequency - Subjects are asked to recall how often some feature has occurred in a category.

Since the research interest has been empirical in nature, assumptions about how frequencies are accumulated, how they are interpreted and how they are used for decision making have not always been explicitly stated. However, some inferences can be made from the procedures used in these studies. In the existing literature, features have generally been treated as independent from each other, despite the fact that some studies recognize the contextual relationship among features (Reitman and Bower, 1973; Hayes-Roth and Hayes-Roth, 1977). In Neumann (1974) and Kellogg, Bourne and Ekstrand (1978), frequencies were reported as the sum of the actual number of occurrences. In Bourne et. al. (1976), frequencies were calculated as a proportion of the category total. Bourne et. al. also examined the implications of within-group and between-group frequency

differentials on concept learning. They found that subjects' performance was the worst when there was an even distribution of within-group frequencies. The effect on performance, however, was slight when between-group frequency distribution was varied.

There has been disagreement among studies over the definition of features. Kellogg (1981) maintained that only basic features were counted in human learning. However, it is difficult to define what constitutes a basic feature. Kellogg cited the nose, eyes and hair as basic features of a face. However, it can easily be argued that the shape of an eye is already a configurational feature involving the eyeball, eyelids and eyelashes. In some studies, conjunctive features (examples: Bourne et al., 1976; Toppino and Bucher, 1983) and relational features (Hayes-Roth and Hayes-Roth, 1977) have been used.

3.1.3 Prototype Approach

The prototype approach is strongly associated with the work of Rosch (1975, 1976, 1978) who proposed that natural categories are hierarchically structured, with a basic level of abstraction where members of one category are intrinsically separated from members of other categories (Rosch, Mervis, Gray, Johnson & Boyes-Braem, 1976). At this

level of abstraction, similarities between objects are maximized within-group and minimized between-group. At higher levels of abstraction, between-group similarity is not minimized because category features are too general. At lower levels of abstraction, within-group similarity is not maximized because category features are too specific. Prototypes exist at the basic level of abstraction. They are abstract representations which have the salient features of their respective categories at this basic level of abstraction. Rosch, Simpson and Miller (1976) identified three major principles for constructing prototypes:

(1) Gestalt Configuration: A prototype is represented by the configuration of a set of features. It is the configuration which is most similar to the feature configurations of its group members. This method of representation is most suitable for cases with features that are difficult to enumerate precisely in their entirety (Strange et al. 1970). Example: dot patterns.

(2) Family Resemblance: A prototype is represented by a set of relational items that has a high frequency of co-occurrence. This method is suitable for cases in which joint features are

more noticeable than individual features (Rosch & Mervis, 1975). Examples: stick figures.

(3) Average Values of Attributes: A prototype is represented by the average values of features. This method is most suitable for cases with easily identifiable and separable features. Example: letter strings.

Rosch's prototype models are supported by the work of Beach (1964), Reed (1972) and Murphy (1982). An alternative approach for constructing prototypes is provided by Shepard (1957, 1958, 1962), Kruskal (1964), and Posner (1967). This alternative approach is consistent with Collins and Quillian's (1969, 1970) proposition about memory organization. They proposed that concepts are organized in memory in the form of a multi-dimensional network. Assuming this form of memory organization, it has been proposed that a prototype is a node in the network which has the minimum average distance to all nodes belonging to the same category and the maximum average distance to nodes of other categories. This approach is referred to as the prototype distance approach.

There are obvious similarities between the prototype approach and the feature frequency approach. Both

approaches are based on inductive reasoning which implies that feature frequencies are attended to. In the feature frequency approach, the feature frequencies are used directly for decision making. In the prototype approach, the feature frequencies are used to generate the prototype which is then used for decision making. Kellogg (1980) suggested that prototypes consist of the most frequently occurring features of a category. For example, if apples are most often red, then red would become the prototype color for apples. According to this interpretation, the prototype approach is a modified form of the feature frequency approach.

3.1.4 Exemplar Approach

The exemplar approach, also referred as the context approach, opposes the probabilistic view of concept learning which is typified by feature frequency and prototype models. Medin and associates, proponents of the exemplar model, suggested that concepts are learned by storing examples intact in memory (see Medin and Schaffer, 1978; Medin and Smith, 1981; Medin, Altom, Edelson and Freko, 1982; Medin, 1983; Medin, Dewey and Murphy, 1983; Medin, Altom and Murphy, 1984; Medin and Shoben, 1988; Medin and Ortony, 1989; Medin and Ross, 1989).

The differences between the prototype and exemplar models are the subject of examination in a number of studies (Medin and Smith, 1984; Estes, 1986a, 1986b; Medin, 1986; Oden, 1987; Nosofsky, 1987, 1988). Medin (1989) cited a number of examples to support the exemplar approach to concept learning. Firstly, he suggested that human experts make decisions on the basis of examples. For example, if a patient is diagnosed as suicidal, it may not mean that the patient is similar to some prototype of a suicidal person, but rather that the patient reminds the clinician of a previous case who was suicidal. This view is consistent with Genero and Cantor's (1987) finding that prototypes served untrained diagnosticians well but that trained diagnosticians found exemplars to be more helpful.

The second reason cited by Medin is that the prototype of a category may change when placed in a different context. For example, in Roth and Shoben (1983), tea was judged to be a more typical beverage than milk in the context of secretaries taking a break, but this ordering reversed for the context of truck drivers having breakfast in the morning.

The third reason Medin used is that the typicality of combined concepts can not be predicted from the typicality

of the constituents. For example, in Medin and Shoben (1988), subjects rated small spoons as more typical of the category "spoons" than large spoons, and metal spoons as more typical than wooden spoons, but large wooden spoons as more typical than small wooden spoons or large metal spoons.

The above examples led Medin and associates to believe that if concepts are represented by prototypes, then more than one prototype will be required for each concept in order to account for the various contexts in which the concept may be explained. Based on this reasoning, Medin and associates proposed that a concept is represented by a set of examples. The model they proposed assumes that people store mental representations of all unique cases that they have encountered. A new case is classified by a sequential comparison of its similarity with all the stored cases. In Nosofsky (1984), the exemplar model was described as follows:

Let X and Y denote two categories;

x denote a stimulus in category X;

y denote a stimulus in category Y;

x_j denote the value of stimulus x on dimension j;

y_j denote the value of stimulus y on dimension j;

t denote a given test stimulus;

t_j denote the value of stimulus t on dimension j;

S(t,x) denote the similarity of stimulus t to stimulus x.

The probability of classifying "t" as a member of category "X" is given by

$$P(X|t) = \frac{\sum_{x \epsilon X} S(t,x)}{\sum_{x \epsilon X} S(t,x) + \sum_{y \epsilon Y} S(t,y)}$$

and the individual S(t,x), and similarly, S(t,y), are computed by

$$S(t,x) = \prod_{j=1}^{n} s_j$$

where $s_j = p_j$, $(0 \leq p_j \leq 1)$, if $t_j \neq x_j$;
and $s_j = 1$, if $t_j = x_j$.

The parameter, p_j represents an initial or minimum value assigned to dimension j. How this parameter should be derived however has not been clearly explained in the literature. Presumably, this parameter would vary according to the nature of the dimension j and the context of the classification task.

3.2 Statistical Pattern Recognition

Models in pattern recognition have been heavily influenced by the development of the digital computer. The digital computer allows complex patterns to be mechanically sensed, decoded, stored, retrieved and mathematically manipulated. This has stimulated interest in examining the biological and mental processes by which sensory signals are converted into meaningful perceptual experience. While the exact operation of these processes is not totally understood, the pattern recognition approach assumes that it involves the following steps (Fukanagan, 1972):

(1) Before recognition, a pattern must first be perceived by our sense organs.

(2) In order to recognize the pattern, the same pattern or a pattern in the same class must have previously been perceived.

(3) The past experience must be remembered (i.e. stored and retrievable) and some association between the past and the present perception must be established.

The above process is translated into the following mechanical steps:

(1) An input pattern is decoded into a vector of n-dimensional space with each dimension representing an aspect (or feature) of the pattern. This step corresponds to the perception of an input stimulus.

(2) A class of pattern is then represented by a particular distribution of the vector's values in the n-dimensional space. This step corresponds to the storage of a class of patterns in memory.

(3) Finally, statistical techniques are used to find one or more planes which maximize the variances between patterns of different classes and minimize the variances among patterns of the same class. These planes then serve as the mechanism for determining the class membership of new patterns.

The dividing planes are generally referred to as discriminant functions. They appear in the form of algebraic expressions with each variable in the expressions representing a dimension of the n-dimensional space. Since each dimension represents an aspect (or feature) of the patterns, the coefficients of variables indicate the relative importance of the features used in the classification task.

Some models in pattern recognition have recognized that classification performance can be improved by weighting features according to their relative effectiveness and/or by eliminating the less effective features (Fukanagan, 1972). A feature selection process is generally included in pattern recognition models to screen out the less effective features. This process requires that a criterion be established for evaluating feature effectiveness. The original measurement in the n-dimensional space is then mapped onto a lower dimensional space to either maximize or minimize the criterion.

3.3 Parallel Distributed Processing (Neural Networks)

Parallel distributed processing systems, also known as connectionist systems and more commonly referred to as neural networks, have their origin in the research on neural nets in the 1940's. Neural networks were originally proposed as models for brain organization and for psychological functions such as association, concept formation and word recognition. While research interest in this area had declined since the 1960's, there has been a renaissance of interest in the last few years.

There are several variants of neural networks, but all have a similar structure. The structure involves a very large number of simple processing elements arranged in parallel arrays. The processing elements can either be simple Boolean devices with two allowable states (for example, on and off) or they can have continuously graded activity. A typical neural network consists of three layer of elements, referred to as input units, hidden units and output units. Research on neural network has shown that systems with only input and output units are inadequate to solve certain types of problems. Rumelhart, Hinton and Williams (1986b) have shown that "exclusive-or" (XOR) problems can never be solved without hidden units. XOR concepts are characterized by

mutual presence or absence of some features. For example, in a two dimension-binary value situation, the concept of a "pair" would be represented by either (1,1) or (0,0) while any mixture of 1 and 0 would be the non-concept. Incidentally, XOR is also the type of problems that feature frequency models cannot not solve without using correlated features.

Research has also shown that there is no necessity to have more than one layer of hidden units, since multiple layers of hidden units can always be collapsed into a single layer without losing effectiveness. The input units are usually Boolean elements, whereas the hidden and output units have continuously graded activity. Generally, input units are joined to hidden units and then through hidden units joined to output units. Theoretically, however, any units can be joined with any other units.

There are three distinct aspects to the functioning of a neural network. Firstly, the strength of connection, referred to as a weight, must be specified for every interconnection between two units. One way this can be done is by first assigning random numbers to the weights and then having the network learn a set of training cases according to a learning rule. Secondly, each hidden and output unit

must be given a threshold value which determines the activity of the unit. Similar to the weights, random numbers can be assigned to the thresholds which can then be modified through experience with training cases. Finally, when the connections between units are formed, the network must be provoked to perform the desired computation. An input pattern of activity has to be provided to start the system, then the initial pattern, through the interconnections, modifies itself and affects the activity of other connected elements to form a new pattern. This process may repeat for a number of steps until the network stabilizes, or it may run continuously.

Neural networks learn through a process called error propagation. This process involves comparing the output pattern of the network to a targeted pattern and systematically changing the weights and threshold values according to a learning algorithm. A commonly used algorithm for this purpose is the delta rule. Rumelhart, Hinton and Williams (1986b) provided a mathematical proof which indicates that the differences between output and targeted patterns are systematically reduced by the delta rule.

There are two basic problems which may make neural networks undesirable as a strategy for building practical problem solving systems. The first problem is that the reasoning behind a decision cannot be easily inferred by the activity of the elements in a neural network. The second problem is that there is no clear rule for determining the number of hidden units. As shown by the XOR problem in Rumelhart, Hinton and Williams (1986b), more hidden units do not necessary provide better performance. In that XOR problem, the problem was consistently solved by a system with one hidden unit. When two hidden units were used, the system was twice trapped in a local minimum and was unable to solve the problem. Research has shown that neural networks tend to specialize their functions when there are a large number of hidden units and generalize when the number of hidden units is small.

Part II: Model Development and Testing

4. Framework for Learning Systems

A criterion of the present study is that the model to be proposed must have the ability to solve real-world problems. Such a model cannot be development without first defining what real-world problems are.

4.1 Characteristics of Real World Problems?

The difference between real world and artificial problems may just be a matter of complexity. An artificial problem is like an experiment conducted in a laboratory. The nature of the problem is well-defined. Extraneous factors affecting the problem are eliminated. Relevant factors can be controlled and tested separately so that interaction effects can be eliminated. In a real-world problem, however, our understanding of the problem can be very vague. In some cases, we may not even be able to state the problem in any clear way. Granted that we can explain the problem, we may not be able to distinguish the relevant factors from the irrelevant ones. The number of factors involved are usually numerous, which renders a complete examination of their interaction effects impossible. In short, a real world problem has the following characteristics:

(1) the criteria for solving the problem are either unknown or ambiguous;

(2) very rarely can the problem be solved using simple criteria;

(3) information about the problem is incomplete due to missing values, insufficient sampling, and so on.

(4) relevant factors are neither clearly nor totally identified;

(5) potential factors are too numerous in number to be examined in detail;

(6) due to the large number of factors, it is impractical to examine the interaction effects between factors exhaustively.

There are a number of implications that stem from the above characteristics. Firstly, the enormity of data favors models that use summary information rather than case-specific information. Models that require analysis of every observed case would be too time-consuming to be of practical use. Secondly, the large number of factors favor models that do not require an exhaustive analysis of interaction effects. Models that do would be impractical. Finally, the ambiguous nature of real-world problems favors models that can handle natural categories. The above implications preclude hypothesis testing models which are appropriate

only for well-defined problems, exemplar models which require individual analysis of observed cases, and many types of discriminant functions which examine correlations between variables.

4.2 Requirements of Practical Problem Solving Systems

For a problem solving system to be of practical use, it must satisfy the following requirements:

(1) The system must be efficient enough to solve problems within a relatively short period of time.
(2) The system must show a level of performance comparable to its human counterpart.
(3) The system must have the ability to describe the deductive or the inductive reasoning behind its decisions. This requirement is a general requirement of expert systems (Feigenbaum, 1979).

The first requirement would preclude many types of discriminant functions and neural networks. The third requirement would preclude neural networks because there is not yet an easy way for extracting decision reasoning from neural networks.

4.3 Selected Framework

By elimination, there are only two types of model left for consideration: feature frequency models and prototype models. These two types of model represent the probabilistic approach to concept learning, and both require the calculation of feature frequencies. Both types of model are efficient because only summary data are required to be stored and manipulated. In addition, b h types of model assume that features can be treated as independent of each other. The models therefore are not required to calculate the correlations among features, which can be very time-consuming if there are a large number of features. Furthermore, both types of model can provide inductive reasoning for their decisions. By referring to either the prototype or the feature frequencies, these models can identify the features that distinguish the differences between categories and the features that determine the category membership of an item.

The above analysis indicates that a form of prototype or feature frequency model will be appropriate for constructing the micro-structure of learning systems. Prototype models could be regarded as a form of feature frequency model. In

general, feature frequency models operate within the following framework:

(1) Knowledge is represented either by frequency matrices or graphs. In the present study, frequency matrices were preferred because numeric values are easier to manipulate than graphs. The term "matrix" implies that all columns have the same size. In reality, this may not be the case. For this reason, the frequency matrices are referred to as frequency tables which could have variable column size.

(2) Values in the frequency tables are treated as independent of each other.

(3) Knowledge is equated to updating values in the frequency tables based on input data and feedback.

(4) Classification is made according to some method of interpreting the input data with respect to the frequency matrices.

5. Decision Rules for Problem Solving

In Section 2.3, the analysis of learning systems suggested that the microscopic function of learning systems involves making a choice among a set of given alternatives. In Section 3, this function was equated with classification tasks in which a choice must be made among a number of given categories. Classification tasks have been the subject of examination in concept learning, pattern recognition and neural networks. Among learning models proposed in these fields, feature frequency models were deemed most appropriate for the current study (see Section 4.3). Research on feature frequency models has been concerned with human learning rather than with machine learning. The emphasis has been on demonstrating that feature frequencies are calculated in concept learning (Kellogg 1980a, 1980b, 1980c; Kellogg, 1981). Exactly how feature frequencies are used for decision making was not explicitly stated in these studies. The literature, however, did suggest that category membership was determined on the basis of the similarities between a new case and the available categories. In this section, the methods for manipulating feature frequencies for decision making are explored.

5.1 Mathematical Representation

For the purposes of the current study, features will be coded numerically by the combination of a feature dimension number and a feature value. This method of representation is similar to the approach used in Bourne et al. (1976). In order to illustrate how this method works, the following example will be used. The example involves two categories: apples and peppers. Items in the two categories are described by their colours and shapes. In this case, colour and shape represent the dimensions of the features. Colours are divided into three groups: red, green and other. The "other" group represents all colours except red and green. Shapes are divided into two groups: round and other. Again, "other" represents all shapes other than the round shape. "Red", "green" and "other" are referred to as the values of the colour dimension. "Round" and "other" are the values of the shape dimension. To represent the above numerically, the following notations will be used:

k = category number ϵ $\{0,1,\ldots,n\}$;

i = feature value ϵ $\{0,1,\ldots,m\}$;

j = feature dimension number ϵ $\{0,1,\ldots,r\}$;

Since there are two categories, three dimensions and a maximum of three values per dimension,

n = 1 (the maximum value for k is 1),

m = 2 (the maximum value for i is 2),

r = 2 (the maximum value for j is 2).

It should be noted that the shape dimension does not have three values. In this case, the shape dimension has a null value which is not labelled and will not be used. To make the above notations meaningful, let us assume that:

"apples" is category 0 (when k = 0),
"peppers" is category 1 (when k = 1),
"shape" is dimension 0 (when j = 0),
"colour" is dimension 1 (when j = 1),
"other" is value 0 in the shape dimension
(when i = 0, j = 0),
"round" is value 1 in the shape dimension
(when i = 1, j = 0),
"other" is value 0 in the colour dimension
(when i = 0, j = 1)
"green" is value 1 in the colour dimension
(when i = 1, j = 1)

"red" is value 2 in the colour dimension
(when i = 2, j = 1).

As indicated above, a feature is represented by the combination of a feature value and a feature dimension number. For example, the feature red color is represented by (2 , 1) which indicates it is the third value (red) of the second dimension (colour). The above also indicates that (2 , 0) is a null combination since no feature is assigned to this combination.

Feature frequency models require that the number of times each feature has occurred be counted by category. The order that the features occur, however, is unimportant. The frequency of a feature is denoted by $F_k(i,j)$.

$F_k(i,j)$ = the number of times value i in dimension j
has occurred in category k

In the present example, let us assume that 110 items have been presented for classification. After the classification of each item, feedback is given indicating whether the classification is correct or not. If an item has been misclassified, it is taken out of the wrong category and placed in the correct category after feedback is received.

Since each item is eventually placed in the correct category, previous success or failure would not affect the model's performance at the present time. Among the 110 items classified, 100 of them are apples and 10 of them are peppers. Among the apples, 60 of them are red and 40 of them are green. All the apples are round in shape. In other words,

$F_0(0,0) = 0$ since no apple has the "other" shape,

$F_0(1,0) = 100$ since all apples are round,

$F_0(0,1) = 0$ since no apple has the "other" colour,

$F_0(1,1) = 40$ since 40 apples have the green colour,

$F_0(2,1) = 60$ since 60 apples have the red colour.

The above frequencies can be presented in a table form as follows:

Distribution of Apples: Frequency F_0

	shape 0	colour 1
0	other 0	other 0
1	round 100	green 40
2		red 60

The above table is denoted by F_0 which indicates that it is the frequency table for category 0. It should be noted that

the symbol "F" is used in two contexts. When it is shown as F_k, it represents the frequency table for category k. When it is shown as $F_k(i,j)$, it indicates the frequency (number of occurrences) of value i in dimension j for category k. Using this method of representation, the frequency table for the category "peppers", F_1, is created as follows:

		shape 0		colour 1	
Frequency	0	other	1	other	0
Distribution = F_1 =	1	round	9	green	7
of Peppers	2			red	3

Let us assume that a new item is currently presented for classification. Let us also assume that this new item is red in colour and round in shape. The new item is represented by the following table:

		shape 0		colour 1	
	0	other	0	other	0
The New Case = Y =	1	round	1	green	0
	2			red	1

The above table is denoted by the symbol "Y". Since the above table is the system's internal representation of the new item, the new item, for convenience, is also referred to as "Y". The elements in the table are denoted by Y(i,j), where

$$Y(i,j) = \begin{cases} 0 & \text{if value i is absent in dimension j of case Y} \\ 1 & \text{if value i is present in dimension j of case Y} \end{cases}$$

The table "Y" functions as a feature selector. It turns on only those features that are present in the new item by giving them the value of 1. It should be noted that a feature is either fully on or fully off. (While a feature could be on partially, for example, an apple that is 80% red, such complication is advised against unless there is an important reason for doing so. Further discussion on this topic is provided in Section 6.)

The classification of a new item is made by finding the category with which it has the highest similarity. The similarity between a new item Y and a category k is denoted by S(Y,k).

S(Y,k) = measurement of similarity between Y and category k.

In the following sub-sections, the methods for manipulating $F_k(i,j)$ to assess S(Y,k) will be examined.

5.2 Major Assumptions

The feature frequency approach to concept learning is based on a number of implicit assumptions (see Section 3.1.2):

(1) a category is represented by a set of frequency counts, each indicating the number of times a particular feature has occurred;

(2) features are treated as independent of each other despite the fact that they may be correlated.

(3) the strength of association between an item and a category is indicated by the degree of similarity between them.

The first assumption implies that individual items of a category need not be stored to represent the category. Since only the eventual frequency counts are stored, it also implies that the order that the items occurred is unimportant. The second assumption allows the frequency counts to be manipulated using simple mathematics such as

summation. The third assumption implies that the membership of a new item can be determined by finding the category with which it has the highest similarity.

In the present study, all of the above assumptions are adopted, despite the fact that there are arguments against some of them (see Medin, 1989). A major argument against the feature frequency approach is that the features in a category could be correlated and are not independent of each other as indicated by the second assumption. In contrast to this argument, Kellogg (1981) provided evidence that correlations between features are not attended to in human learning. However, it is an undeniable fact that the "exclusive-or" type of problems (see Rumelhart, Hinton and Williams, 1986b) could never be solved without attention to the correlations among features. Intuitively, it is reasonable to assume that correlations among features are attended to in concept learning. However, this does not mean that all possible correlations will be observed. If there is a small number of features, it would be possible to examine all possible correlations of the features. However, real-world problems usually have a large number of features. In this case, it would be impractical to examine all possible correlations. It is more likely that attention is

limited to a selective set of correlations which are most relevant to the classification task.

The existing literature on concept learning does not provide clear indications on how relevant correlations could be detected. In the present study, it is assumed that if a classification task necessitates the use of certain correlated features, these correlated features will be identified by the user of the learning system. For example, if the joint occurrence of the red colour and the round shape is essential for distinguishing apples from peppers, a new dimension called "red and round" will be created. This new dimension will have the Boolean values of "yes" and "no". This method of accounting for correlated features is consistent with the models proposed by Reitman and Bower (1973), and Hayes-Roth and Hayes-Roth (1977). In numerous situations, it may be unnecessary for a learning system to attend to all relevant correlated features. A correlated feature should be attended to only if

(1) the effect created by the correlated feature is essential to performing the required classification task;
(2) the effect will not be accounted for if the features are treated as independent.

For example, even though shapes and colours are often correlated in natural objects, the shape dimension alone is sufficient to distinguish apples from bananas. In this case, the correlations between shapes and colours (e.g. red and round, yellow and slender) are not essential to the classification task. In general, a system's efficiency would be hampered by the inclusion of unnecessary features. Correlated features, therefore, should not be included unless there is a valid reason for doing so.

5.3 Summed Feature Frequency

There are indications in some studies that similarity can be measured on the basis of summed feature frequency (Neumann, 1974; Kellogg, Bourne & Ekstrand, 1978). This method assumes that the similarity between a new item Y and a category k is measured by adding the feature frequencies $F_k(i,j)$ for only those features that are present in Y.

$$S(Y,k) = \sum_{i=0}^{m} \sum_{j=0}^{r} Y(i,j)F_k(i,j) \qquad (1.01)$$

In equation (1.01), the product of Y(i,j) and $F_k(i,j)$ will be equal to 0 if Y(i,j) is equal to 0 (i.e. when the feature represented by Y(i,j) is absent in dimension j). It will be equal to $F_k(i,j)$ if Y(i,j) is equal to 1. In effect, this equation sums the frequencies of only those features that are present in the new item Y.

To illustrate how equation (1.01) should be applied, the example in Section 5.1 will be used. This example involves two categories: apples and peppers. The feature frequencies of these two categories are shown in the following tables:

		shape 0	colour 1
Frequency	0	other 0	other 0
Distribution = F_0 =	1	round 100	green 40
of Apples	2		red 60

		shape 0	colour 1
Frequency	0	other 1	other 0
Distribution = F_1 =	1	round 9	green 7
of Peppers	2		red 3

Let us assume that a new item Y, which is round and red, is currently presented for classification.

		shape 0	colour 1
	0	other 0	other 0
The New Item = Y =	1	round 1	green 0
	2		red 1

Applying equation (1.01) will yield the following similarity scores:

For apples:

$$S(Y,0) = (0x\ 0) + (0x\ 0) + (1x100) + (0x40) + (1x60) = 160$$

For Peppers:

$$S(Y,1) = (0x1) + (0x0) + (1x9) + (0x7) + (1x3) = 12$$

The above similarity scores indicate that the new item Y is more similar to apples than peppers. Since apples are always round and more often red than peppers, our intuitive judgement would also suggest "apple" as the answer. In some situations, however, the results of equation (1.01) may not coincide with our intuitive judgement. This will occur, for example, if the new item is a green and slender object. Since apples are never slender and peppers are more often green, our intuitive judgement would say the object is a pepper. Given that a slender and green object is represented by:

			shape		colour	
			0		1	
		0	other	1	other	0
The New Item =	Y =	1	round	0	green	1
		2			red	0

applying equation (1.01) will provide these scores:

$$S(Y,0) = (1x\ 0) + (0x\ 0) + (0x100) + (1x40) + (0x60)$$

$$= 40$$

$$S(Y,1) = (1x1) + (0x0) + (0x9) + (1x7) + (0x3)$$

$$= 8$$

The above similarity scores suggest that the new item is more likely an apple than pepper. This decision would be inconsistent with our intuitive judgement. The inconsistency occurs because feature frequencies are expressed as the actual number of their occurrences. This method of calculating frequencies has the effect of favoring large categories. Given a sufficiently large difference in size between categories, an infrequent feature in the larger

category may have a higher frequency value than a frequent feature in the smaller category. In the above example, 40% of 100 apples is more than all the peppers added together. Therefore, as long as the new item has one of the valid features of apples (i.e. round, red, or green), it will always be classified as an apple. The only situation for a new item to be classified as a pepper is when it has none of the apple features. Consider, for example, that the new item is a slender and yellow object:

			shape		colour	
			0		1	
		0	other	1	other	1
The New Item =	Y =	1	round	0	green	0
		2			red	0

Applying equation (1.01) will provide the following scores:

$$S(Y,0) = (1x0) + (1x\ 0) + (0x0) + (0x40) + (0x60)$$

$$= 0$$

$$S(Y,1) = (1x1) + (1x0) + (0x9) + (0x7) + (0x3) = 1$$

In the above situation, the new item will be classified as a pepper because it has none of the features found in apples. The bias in equation (1.01) will become less obvious if category sizes are approximately the same. However, natural categories very rarely have equal population sizes. For this reason, equation (1.01) is inadequate as a decision making model for real-world problems.

5.4 Within-Dimension Variability

The problem with equation (1.01) is that it ignores the within- dimension variations of feature frequencies. A quick glance at the frequency distribution in the colour dimension would reveal that red is the salient colour of apples while green characterizes peppers. The importance of within-dimension variability has been examined by Bourne and associates (Bourne et. al., 1976). Their studies showed that a problem could be made more or less difficult by artificially varying the distribution of value frequencies within a dimension. One way to emphasize the within-dimension variations is to convert the value frequencies into a form relative to their dimensional total. By doing so, the effects of unequal category sizes will be eliminated. The following equations illustrate how equation (1.01) can be modified to take within-dimension variations into account.

Let $P_k(i,j)$ = proportion of items in category k that have value i in dimension j;

N_k = total number of items in category k;

$N_k(j)$ = total number of items in category k with non-missing value in dimension j.

The $N_k(j)$ above can be enumerated as follows:

$$N_k(j) = \sum_{i=0}^{m} F_k(i,j)$$

If there are no missing values in category k, $N_k(j)$ will have the same value as N_k and $P_k(i,j)$ can be calculated as:

$$P_k(i,j) = \frac{F_k(i,j)}{N_k}$$

In reality, however, there is often a good chance of having some values missing. Therefore, it is more appropriate to express $P_k(i,j)$ as:

$$P_k(i,j) = \frac{F_k(i,j)}{N_k(j)}$$

Substituting $P_k(i,j)$ for $F_k(i,j)$ in equation (1.01), we obtain:

$$S(Y,k) = \sum_{i=0}^{m} \sum_{j=0}^{r} Y(i,j)P_k(i,j) \qquad (1.02)$$

The following tables show the conversion of frequency values into a form relative to their dimensional total:

Frequency

		shape 0	colour 1
Frequency	0	other 0	other 0
Distribution	F_0 = 1	round 100	green 40
of Apples	2		red 60
	Total	100	100

Relative Frequency

		shape 0	colour 1
	0	other 0	other 0
	P_0 = 1	round 1.00	green .40
	2		red .60
		1.00	1.00

Frequency

		shape 0	colour 1
Frequency	0	other 1	other 0
Distribution	F_1 = 1	round 9	green 7
of Peppers	2		red 3
	Total	10	10

Relative Frequency

		shape 0	colour 1
	0	other .10	other 0
	P_1 = 1	round .90	green .70
	2		red .30
		1.00	1.00

The effects of the above conversion will be examined in the following examples. In the first example, the task is to classify a red and round object:

The New Item = Y =		shape 0		colour 1	
	0	other	0	other	0
	1	round	1	green	0
	2			red	1
	Total		1		1

Applying equation (1.02), the following similarity scores are obtained:

$$S(Y,0) = (0x\ \ 0) + (0x\ \ 0) + (1x1.00) + (0x.40) + (1x.60) = 1.6$$

$$S(Y,1) = (0x.10) + (0x\ \ 0) + (1x.90) + (0x.70) + (1x.30) = 1.2$$

The above scores indicate that the new item will be classified as an apple. This is consistent with the result provided by equation (1.01).

The second example involves a new item that is slender and yellow, indicated as follows:

		shape		colour	
		0		1	
	0	other	1	other	1
The New Item = Y =	1	round	0	green	0
	2			red	0

Applying equation (1.02), the following similarity scores are obtained:

$$S(Y,0) = (1x\ \ 0) + (1x\ \ 0) + (0x1.00) + (0x.40) + (0x.60) = 0$$

$$S(Y,1) = (1x.10) + (1x\ \ 0) + (0x.90) + (0x.70) + (0x.30) = .1$$

The similarity scores indicate that the new item is a pepper. This again is consistent with the result of equation (1.01).

The above two examples indicate that equation (1.02) is as effective as equation (1.01) in solving the obvious cases. In the following example, the difference between the two

equations will become apparent. This example involves a new item that is slender and green, indicated as follows:

The New Item = Y =		shape 0	shape 1	colour 0	colour 1
	0	other	1	other	0
	1	round	0	green	1
	2			red	0

Applying equation (1.02), the following similarity scores are obtained:

$$S(Y,0) = (1x\ \ 0) + (0x\ \ 0) + (0x1.00) + (1x.40) + (0x.60)$$

$$= 0.4$$

$$S(Y,1) = (1x.10) + (0x\ \ 0) + (0x.90) + (1x.70) + (0x.30)$$

$$= 0.8$$

As indicated by the similarity scores, the new item will be classified as a pepper. This decision would be consistent with our intuitive judgement but different from the decision arrived at by using equation (1.01). When equation (1.01)

was used, the ratio of similarity scores between apples and peppers was 5 to 1 in favor of apples. In the current case, the ratio is reduced by 10 times to 1/2. The decrease in the ratio is proportionally equal to the difference in category sizes (100 to 10) between apples and peppers. This indicates that the effect of unequal category sizes is eliminated when relative frequencies are used.

5.5 Prototype Modification of Relative Frequency

In close examination, the prototype approach to concept learning (Rosch & Mervis, 1975) also addresses the importance of within-dimension variations. This approach assumes that only the most frequent features are attended to in concept learning (Chumbley, Sala & Bourne, 1978; Goldman & Homa, 1977; Hayes-Roth & Hayes-Roth, 1977; Kellogg, 1980a; Neumann, 1977). Operationally, this is equivalent to emphasizing the pattern of within-dimension frequency distributions by assigning 1 to the mode of a distribution and 0 to all others.

$$S(Y,k) = \sum_{i=0}^{m} \sum_{j=0}^{r} Y(i,j)X_k(i,j) \qquad (1.03)$$

$$\text{where } X_k(i,j) = \begin{cases} 1 & \text{if } P_k(i,j) = \max\limits_{t} P_k(t,j) \\ 0 & \text{if } P_k(i,j) \neq \max\limits_{t} P_k(t,j) \end{cases}$$

In general, equation (1.03), which represents the prototype approach, would provide the same results as equation (1.02). By emphasizing the modal value of each dimension, equation (1.03) is sensitive to only one feature value in each dimension. This means that equation (1.03) will not be

sensitive to minor variations within a dimension and might cause difficulty in some situations, as indicated by the following example.

This example involves a new item which is slender and yellow.

The New Item = Y =		shape 0		colour 1	
	0	other	1	other	1
	1	round	0	green	0
	2			red	0

The following tables show the conversion of relative frequencies into the prototype form.

Frequency Distribution of Apples		Relative Frequency shape 0		colour 1		Prototype Form shape 0		colour 1	
	0	other	0	other	0	0	other	0	other	0
	P_0 = 1	round	1.00	green	.40	X_0 = 1	round	1	green	0
	2			red	.60	2			red	1
	Total		1.00		1.00					

			Relative Frequency			Prototype Form			
			shape		colour		shape		colour
			0		1		0		1
Frequency		0	other	.10	other 0	0	other 0	other	0
Distribution	P_1 =	1	round	.90	green .70	X_1 = 1	round 1	green	1
of Peppers		2			red .30	2		red	0
	Total			1.00	1.00				

Applying equation (1.03), the following similarity scores are obtained:

$$S(Y,0) = (1x0) + (1x0) + (0x1) + (0x0) + (0x1) = 0$$

$$S(Y,1) = (1x0) + (1x0) + (0x1) + (0x1) + (0x0) = 0$$

As shown in the above similarity scores, the prototype approach will have difficulty classifying this new item because the similar scores are identical for both categories. Using equation (1.01) or (1.02), however, the new item will be classified as a pepper. This is because the frequency distributions in the "shape" dimension are

slightly different between the two categories. The difference favors the pepper category if the "other" shape is present. The difference, however, would be masked when the prototype treatment is applied to the frequency data.

The above example brings up another issue, which is whether the variations in frequency distribution between categories should also be attended to. As shown above, the difference in frequency distribution between categories could be so small that it might be neglected in some situations.

5.6 Between Category Variability

The importance of between-category frequency variations was examined in Bourne et. al. (1976). Their findings suggested that this type of variations has little effect in human learning. Nevertheless, the following example indicates the potential importance of between-category frequency variations.

This example involves a classification task which distinguishes athletes who are soccer players from those who are not. The profiles of 200 athletes have been collected and fed into the system. Among these two hundred cases, half are soccer players and the other half are not. Among the 100 soccer players, 66 of them are from countries other than Canada and the United States. In addition, 68 of the 100 soccer players are under 50 years of age. Since playing soccer requires the use of both feet, none of them are wheel-chair athletes. In the non-soccer group, 34 of the cases are from countries outside North America, 68 of them are under 50 years of age and one of them is confined to a wheel chair.

In this example, the feature "confined to a wheel chair" is a very critical feature because the presence of this feature

alone is sufficient to conclude that an athlete is not a soccer player. The absence of this feature, however, is insufficient to determine whether the athlete is a soccer player. This feature, therefore, is not a defining feature of the non-soccer category because it is not present in all cases. This feature occurs so infrequently that it would not even stand out in the comparison of frequencies. As shown in the following calculations, a European wheel-chair athlete would be mistaken for a soccer player if only within-dimension frequency variations are attended to.

Consider a European wheel-chair athlete whose age is less than 50 years. His profile will be represented as follows:

		Racial Origin		Age		Wheel Chair	
		0		1		2	
New Case = Y =	0	other	1	50<	1	yes	1
	1	N.Amer	0	>=50	0	no	0
	Total		1.00		1.00		1.00

The relative frequencies will be tabled as follows:

		Racial Origin		Age		Wheel Chair	
		0		1		2	
Soccer Players = P_0 =	0	other	.66	50<	.68	yes	0
	1	N.Amer	.34	>=50	.32	no	1.00
	Total		1.00		1.00		1.00

		Racial Origin		Age		Wheel Chair	
		0		1		2	
Non-Soccer Players = P_1 =	0	other	.34	50<	.68	yes	.01
	1	N.Amer	.66	>=50	.32	no	.99
	Total		1.00		1.00		1.00

Applying equation (1.02), the relative frequency approach, the following similarity scores are obtained:

$$S(Y,0) = (1x.66) + (1x.68) + (1x\ \ 0) + (0x.34) + (0x.32) + (0x1.00) = 1.34$$

$$S(Y,1) = (1x.34) + (1x.68) + (1x.01) + (0x.66) + (0x.32) + (0x.99) = 1.03$$

As indicated in the above similarity scores, the new case will be incorrectly classified as a soccer player. The reason why this mistake occurs is because the feature "confined to wheel chair" has a very low relative frequency

of occurrence. Expressing frequencies as the actual number of occurrences (i.e. equation 1.01) would not solve this problem either, because only 1 out of the 200 cases is a wheel-chair athlete. The prototype approach, which exaggerates the pattern of within-dimension frequency distributions, would certainly not be able solve this problem.

The above problem could be solved by attending to the between-category distribution of frequencies. In this approach instead of measuring relative frequencies as a proportion of their dimensional total, relative frequencies are measured as a proportion of their total across categories. In other words, instead of expressing relative frequencies as

$$\frac{F_k(i,j)}{N_k(j)}$$

they are expressed as

$$\frac{F_k(i,j)}{\sum_{t=0}^{n} F_t(i,j)}$$

There are some mathematical implications embedded in the above two expressions. The first expression, which reflects the relative frequencies within a dimension, could be interpreted as the empirical conditional probability of feature i occurring in dimension j given category k. The second expression, which addresses the between-category frequency distribution, indicates the empirical conditional probability of a new case belonging to category k given that feature i is present in dimension j of the new case. Since the second expression could be derived by applying the Bayes' Theorem to the empirical data, it is referred to in the present study as the empirical Baysian probability. The use of Bayes' Theorem for classification tasks is consistent with Murphy's (1982) interpretation of prototype models, Nosofsky's (1984) interpretation of exemplar models, and Jajuga's (1986) interpretation of statistical pattern recognition. The following equations illustrate how the empirical Baysian probability is derived.

Let C_k = the event that the new case Y belongs to category k

E_{ij} = the event that feature value i is present in dimension j

$P(C_k|E_{ij})$ = the conditional probability of C_k given the event E_{ij}

The empirical prior probabilities of the C_k's are

$$P(C_k) = \frac{N_k(j)}{\sum_{t=0}^{n} N_t(j)} \qquad \text{for } k = 0,1,\ldots,n;$$

and the empirical conditional probabilities $P(E_{ij}|C_k)$ are given by

$$P(E_{ij}|C_k) = \frac{F_k(i,j)}{N_k(j)}$$

(It should be noted that $P(E_{ij}|C_k)$ is identical to $P_k(i,j)$ in equation 1.02)

According to Bayes' Theorem, the empirical posterior probabilities are

$$P(C_k|E_{ij}) = \frac{P(C_k)\ P(E_{ij}|C_k)}{\sum_{t=0}^{n} P(C_t)\ P(E_{ij}|C_t)}$$

$$= \frac{F_k(i,j)}{\sum_{t=0}^{n} F_t(i,j)}$$

Substituting $P(C_k|E_{ij})$ for $P_k(i,j)$ in equation (1.02), we obtain:

$$S(Y,k) = \sum_{i=0}^{m} \sum_{j=0}^{r} Y(i,j)\ P(C_k|E_{ij}) \qquad (1.04)$$

The following calculations show the conversion of frequency values into empirical Baysian probabilities:

Frequency		Racial Origin 0		Age 1		Wheel Chair 2	
Soccer Players = P_0 =	0	other	66	50<	68	yes	0
	1	N.Amer	34	>=50	32	no	100
	Total		100		100		100

Frequency		Racial Origin 0		Age 1		Wheel Chair 2	
Non-Soccer Players = P_1 =	0	other	34	50<	68	yes	1
	1	N.Amer	66	>=50	32	no	99
	Total		100		100		100

$$P(C_0|E_{00}) = 66/(66+34) = .66$$

$$P(C_1|E_{00}) = 34/(66+34) = .34$$

$$P(C_0|E_{10}) = 34/(34+66) = .34$$

$$P(C_1|E_{10}) = 66/(34+66) = .66$$

$P(C_0|E_{01}) = 68/(68+68) = .5$

$P(C_1|E_{01}) = 68/(68+68) = .5$

$P(C_0|E_{11}) = 32/(32+32) = .5$

$P(C_1|E_{11}) = 32/(32+32) = .5$

$P(C_0|E_{02}) = 0/(0+1) = 0$

$P(C_1|E_{02}) = 1/(0+1) = 1$

$P(C_0|E_{12}) = 100/(100+99) = .5$

$P(C_1|E_{12}) = 99/(100+99) = .5$

Based on the above calculations, the empirical Baysian probabilities are as follows:

Empirical Baysian Probability		Racial Origin 0	Age 1	Wheel Chair 2
Soccer Players = P_0 =	0	other .66	50< .5	yes 0
	1	N.Amer .34	>=50 .5	no .5

Empirical Baysian Probability		Racial Origin 0	Age 1	Wheel Chair 2
Non-Soccer Players = P_1 =	0	other .34	50< .5	yes 1.00
	1	N.Amer .66	>=50 .5	no .5

Given a new case Y:

		Racial Origin		Age		Wheel Chair	
		0		1		2	
New Case = Y =	0	other	1	50<	1	yes	1
	1	N.Amer	0	>=50	0	no	0
Total		1.00		1.00		1.00	

and applying equation (1.04), the following similarity scores are obtained:

$$S(Y,0) = (1x.66) + (1x.5) + (1x0) + (0x.34) + (0x.5) + (0x.5) = 1.16$$

$$S(Y,1) = (1x.34) + (1x.5) + (1x1) + (0x.66) + (0x.5) + (0x.5) = 1.84$$

As indicated by the similarity scores, the new case will be correctly classified as a non-soccer player. Among all the approaches examined to this point, the empirical Baysian approach is the only one sensitive to critical but infrequent features. This approach is also more sensitive to within-dimension frequency variations than the summed frequency approach (i.e. equation 1.01). For instance, a slender and green object, which was incorrectly classified by the summed frequency approach, will be correctly

classified as a pepper using the current approach. However, the current approach might not be sensitive enough to within-dimension frequency variations to handle cases that have close similarities with both categories. Consider for example a new case with the following features:

The New Case = Y =		shape 0	colour 1
	0	other 0	other 0
	1	round 1	green 1
	2		red 0

Since peppers are more often green than apples, our intuition might favor the new case being a pepper. However, when the empirical Baysian probabilities are used, the following probability tables result:

Frequency

Apples P_0 =		shape 0	colour 1
	0	other 0	other 0
	1	round 100	green 40
	2		red 60
	Total	100	100

Empirical Baysian Probability

P_0 =		shape 0	colour 1
	0	other 0	other 0
	1	round .92	green .85
	2		red .95

Frequency

Peppers		shape 0	colour 1
	0	other 1	other 0
	F_1 = 1	round 9	green 7
	2		red 3
	Total	10	10

Empirical Baysian Probability

		shape 0	colour 1
	0	other 1.00	other 0
	P_1 = 1	round .08	green .15
	2		red .05

These lead to the following similarity scores:

$$S(Y,0) = (0x0) + (0x0) + (1x.92) + (1x.85) + (0x.95) = 1.77$$

$$S(Y,1) = (0x1) + (0x0) + (1x.08) + (1x.15) + (0x.05) = 0.23$$

According to the above similarity scores, the new case will be classified as an apple. The lack of sufficient sensitivity to within-dimension frequency variations, in this case, might have led the empirical Baysian approach to a wrong answer.

5.7 Model with Dual Sensitivities

All the approaches discussed to this point have some deficiencies. This is because none of them are sufficiently sensitive to both within-dimension and between-category frequency variations. The following is the description of a proposed model which would provide for both types of sensitivities. The proposed model is based on Baysian probability with a uniform prior probability of occurrence for the categories.

We assume that the prior probability distribution of categories is the discrete uniform distribution.

$$P(C_k) = \frac{1}{n+1} \qquad \text{for } k = 0,1,\ldots,n.$$

Again, as in the previous approach, we assume that the conditional probabilities $P(E_{ij}|C_k)$ are given by the relative frequencies

$$P(E_{ij}|C_k) = \frac{F_k(i,j)}{N_k(j)}$$

then the Bayes' Theorem yields

$$P(C_k|E_{ij}) = \frac{P(C_k)\ P(E_{ij}|C_k)}{\sum_{t=0}^{n} P(C_t)\ P(E_{ij}|C_t)}$$

$$= \frac{\frac{F_k(i,j)}{N_k(j)}}{\sum_{t=0}^{n} \frac{F_t(i,j)}{N_t(j)}} \qquad (1.05)$$

When the $P(C_k|E_{ij})$ in equation (1.04) is enumerated using equation (1.05), the model combines the characteristics of the relative frequency approach with those of the empirical Baysian probability approach. It will be sensitive to both within-dimension and between-category frequency variations. As shown in the following examples, the model is unaffected by unequal category sizes and is sensitive to critical but infrequently occurring features.

In the first example, let us consider an object that is green and round which is to be classified between the categories of apples and peppers.

The New Case = Y =		shape 0	colour 1
	0	other 0	other 0
	1	round 1	green 1
	2		red 0

The conversion from relative frequencies into $P(C_k|E_{ij})$, as stipulated by equation (1.05), is as follows:

Frequency

Frequency Distribution of Apples F_0 =		shape 0	colour 1
	0	other 0	other 0
	1	round 100	green 40
	2		red 60
	Total	100	100

Relative Frequency

P_0 =		shape 0	colour 1
	0	other 0	other 0
	1	round 1.00	green .40
	2		red .60
		1.00	1.00

Frequency

Frequency Distribution of Peppers F_1 =		shape 0	colour 1
	0	other 1	other 0
	1	round 9	green 7
	2		red 3
	Total	10	10

Relative Frequency

P_1 =		shape 0	colour 1
	0	other .10	other 0
	1	round .90	green .70
	2		red .30
		1.00	1.00

Applying equation (1.05),

$$P(C_0|E_{00}) = 0/(0+0.1) = 0$$

$$P(C_1|E_{00}) = .1/(0+.1) = 1$$

$$P(C_0|E_{10}) = 1/(1+.9) = .53$$

$$P(C_1|E_{10}) = .9/(1+.9) = .47$$

$$P(C_0|E_{01}) = 0/(0+0) = 0$$

$$P(C_1|E_{01}) = 0/(0+0) = 0$$

$$P(C_0|E_{11}) = .4/(.4+.7) = .36$$

$$P(C_1|E_{11}) = .7/(.4+.7) = .64$$

$$P(C_0|E_{21}) = .6/(.6+.3) = .67$$

$$P(C_1|E_{21}) = .3/(.6+.3) = .33$$

the following probability tables result:

Apples

P_0 = 1	Baysian Probability shape 0	colour 1
0	other 0	other 0
1	round .53	green .36
2		red .67

Peppers

P_1 = 1	Baysian Probability shape 0	colour 1
0	other 1.00	other 0
1	round .47	green .64
2		red .33

Applying the above probabilities to equation (1.04), the following similarity scores are obtained:

$$
\begin{aligned}
S(Y,0) &= (0x \quad 0) + (0x \quad 0) \\
&\quad +(1x0.53) + (1x.36) \\
&\quad + (0x.67) \\
&= .89 \\
S(Y,1) &= (0x1.00) + (0x \quad 0) \\
&\quad +(1x.47) \quad + (1x.64) \\
&\quad + (0x.33) \\
&= 1.11
\end{aligned}
$$

As indicated by the similarity scores, the object will be classified as a pepper. Using the empirical Baysian probabilities, a green and round object will be classified as an apple. This example clearly shows that the proposed model is sensitive to within-dimension frequency variations and hence is not affected by the unequal category sizes between apples and peppers.

In the second example, the task is to classify a European wheel-chair athlete who is less than 50 years old.

		Racial Origin	Age	Wheel Chair
		0	1	2
New Case = Y =	0	other 1	50< 1	yes 1
	1	N.Amer 0	>=50 0	no 0
Total		1.00	1.00	1.00

Since the two categories, soccer players and non-soccer players, have an identical category size (100 cases each), they already have an equal prior probability of occurrence. The relative frequencies, therefore, are identical to the Baysian probabilities required by equation (1.05).

Baysian Probability		Racial Origin	Age	Wheel Chair
		0	1	2
Soccer Players = P_0 =	0	other .66	50< .5	yes 0
	1	N.Amer .34	>=50 .5	no .5

Baysian Probability		Racial Origin	Age	Wheel Chair
		0	1	2
Non-Soccer Players = P_1 =	0	other .34	50< .5	yes 1.00
	1	N.Amer .66	>=50 .5	no .5

Applying equation (1.04) using these Baysian probabilities, we obtain the following similarity scores:

$$S(Y,0) = (1x.66) + (1x.5) + (1x0) + (0x.34) + (0x.5) + (0x.5) = 1.16$$

$$S(Y,1) = (1x.34) + (1x.5) + (1x1) + (0x.66) + (0x.5) + (0x.5) = 1.84$$

The similarity scores indicate that the athlete does not play soccer. This example illustrates that the proposed model is sensitive to between-category frequency variations and hence is sensitive to critical but infrequently occurring features.

5.8 Conclusion: Model Proposal

As shown in the above analysis, the sensitivities to both between-category and within-dimension frequency variations may be essential to a versatile problem-solving model. Without the sensitivity to within-dimension frequency variations, a model may be too sensitive to differences in category sizes. This might lead to incorrect decisions in some situations since category sizes could be arbitrarily different in the real world. Without sensitivity to between-category frequency variations, a model would be insensitive to critical but infrequently occurring features which might be the only cue to solving some real-world problems. These sensitivities therefore would improve the performance of classification models.

Sensitivity to critical but infrequently occurring features might have implications for expert systems as well as for human learning. One major problem in building expert systems is the difficulty in acquiring knowledge from human experts. When human experts are asked to explain the rationale behind a decision, rather than providing a theoretical explanation they often cite some previous examples. This phenomenon is one reason why Medin and associates believed that concepts were learned by storing

exemplars. It is also the justification for the rule-based approach to system development. The above analysis, however, suggests that this phenomenon could also be explained because of the existence of critical but infrequently occurring features.

In the above analysis, the examples involving a green and round pepper and a European wheel-chair athlete are the critical tests for determining the capability of different approaches. The following table indicates the decisions and the similarity ratios provided by these approaches:

	Green Round Pepper	European Wheel-Chair Athlete
Summed Frequency	8.75 to 1 favoring apple	1.3 to 1 favoring soccer
Relative Frequency	1.14 to 1 favoring pepper	1.3 to 1 favoring soccer
Prototype	2 to 1 favoring pepper	2 to 1 favoring soccer
Empirical Baysian	7.7 to 1 favoring apple	1.6 to 1 favoring non-soccer
Uniform Prior Baysian	1.25 to 1 favoring pepper	1.6 to 1 favoring non-soccer

The above table indicates that the Baysian approach which assumes the discrete uniform distribution of prior probabilities for the categories is the only approach versatile enough to correctly classify both examples. This is the approach proposed here to be the decision-making model for the micro-structure of learning systems. The model, once again, states that:

A decision is made by selecting the category (i.e. alternative) with the highest similarity score. Similarity scores are denoted by

$$S(Y,k) = \sum_{i=0}^{m} \sum_{j=0}^{r} Y(i,j)\ P(C_k|E_{ij})$$

where

$$P(C_k|E_{ij}) = \frac{\frac{F_k(i,j)}{N_k(j)}}{\sum_{t=0}^{n} \frac{F_t(i,j)}{N_t(j)}}$$

6. Requirements of the Model

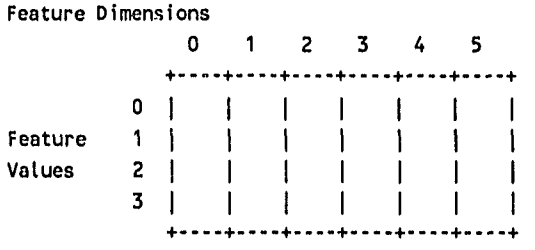
Despite the fact that the proposed model is very simple in structure and does not have many prerequisites, there are certain conditions that must be met before it can be applied. The following are the main requirements:

(1) the model requires data that have a finite set of clearly identifiable feature dimensions and a limited set of possible values for each dimension;

(2) the model must be given a finite set of alternatives on which its decisions are made;

(3) the model requires a learning phase during which it must be given feedback to indicate which of its decisions are correct.

There are numerous ways of storing data. In the present case, a two-dimensional array similar to the example below was used:



In the above structure, the feature dimensions are not required to have the same number of feature values (see examples in Section 5). For convenience, an array for storing input data is referred to as an input array. An array for storing frequency data is referred to as a frequency array.

An input array is required for each case to be processed by the system. Each element in an input array can have only two possible states: on or off. The value of 1 is assigned to the "on" state, and the value of 0 is assigned to the "off" state. When an element is in the "on" state, it means that the feature value in the feature dimension indicated by the element in the array is present in that case. Otherwise, it means the absence of the feature value.

The input array defines the type of data accepted by the model. Any data that fall beyond the structure of the input array will be ignored by the model. The user of the model

has the responsibility of identifying the feature dimensions and feature values for the model. The model treats the feature dimensions as if they are independent. The model does not compute the correlation between feature values in different dimensions. If there are significant correlated features, these features have to be identified by the user. Otherwise, the model would not take these features into consideration. Consider, for example, a pair of shoes and the task is to find out whether they are both black. In this case, there are two initial feature dimensions representing the colour of the left shoe and the colour of the right shoe. This problem cannot be solved by attending to these two feature dimensions independently. To solve this problem, a Boolean dimension for the joint feature "black for both shoes" has to be identified for the system.

The model treats feature values within a dimension as mutually exclusive. Each feature dimension, therefore, can have only one feature value in the on state. If a certain mixture of feature values is significant, a new feature value should be created to represent the mixed feature values. For example, if a colour of mixed blue and red is significant, instead of turning the colours of blue and red on, a new colour value called "mixed blue and red" should be

created and turned on while keeping all other colour values in the off state.

The following steps show the process through which the model acquires knowledge and makes decisions:

(1) The model requires a number of training cases.

(2) Based on the training cases, the model calculates the frequency of occurrence of each feature value in each feature dimension for each given alternative. To be consistent with the literature, the given alternatives are referred to as categories.

(3) Upon receiving a new case, the model, based on the frequencies calculated in (2), derives for each category a similarity score which indicates the similarity between the new case and the category. The similarity score is calculated by summing a series of Baysian probabilities, each indicating the probability of the new case belonging to the category on the evidence in one single feature dimension. The Baysian probability is calculated under the assumptions of:

(i) equal prior probability for all categories, and

(ii) given each category, the conditional probability of each feature value within each dimension is accurately specified by the corresponding empirical conditional probabilities based on the training cases.

(4) A decision is made by assigning the case to the category for which it has the highest similarity score.

For convenience, steps 1 and 2 are referred to as the training phase. Steps 3 and 4 are referred as the decision-making phase.

To train the model, a frequency array is required for each category. Since the input array and the frequency array have the same structure, the frequency array can be generated by adding the input arrays in the category which it measures.

Example

Input Case 1		Input Case 2		Input Case 3		Frequency Array
1 0	+	0 0	+	1 1	=	2 1
0 1		1 1		0 0		1 2

To make a decision, the values in the frequency array must be converted to frequencies relative to the total number of occurrences in each feature dimension.

Example

	Frequency Array		Relative Frequency
	2 1	=>	.67 .33
	1 2		.33 .67
Total	3 3		1.00 1.00

If there are no missing data, the number of occurrences should be identical for all feature dimensions in a category. If there are missing data, the missing data are ignored in the calculation. In this case, the frequency array may have variation in the totals for the dimensions.

In the decision-making process, frequencies in irrelevant feature dimensions should cancel each other out in the between-category comparison. However, random variation in

the irrelevant dimensions may still impede the model's performance. Besides, the model's efficiency will definitely be affected if there are too many irrelevant dimensions. Users therefore should be cautioned against creating too many correlated dimensions. Creating too many feature values will have a much more detrimental effect. Too many feature values in a dimension may dilute the relative frequency of the salient values. This may cause a significant change to the Baysian probability. In light of the potential damages to the model, users as a rule should not create correlated dimensions or new feature values unless there are reasons for doing so.

7. Tests of the Model

7.1 The Classification Problem Used in Testing

The problem used to test the proposed model pertained to neuropsychological interpretation of brain dysfunction in children and the data were collected from a battery of Halstead Reitan tests (Boll, 1981) used for such purpose. The data set consisted of scores of the Halstead Reitan tests for 154 cases and, for each case, the decisions (i.e. results of interpretation) of two neuropsychologists. For comparison purposes, the decisions of a neuropsychological expert system were also available.

Neuropsychological studies have shown that brain dysfunction is manifested in changes in behaviour. These changes can be detected by tests of IQ and various motor, sensory, speech and cognitive skills (Reitan and Davidson, 1974; Boll, 1981). There is evidence suggesting that children and adults show different behavioural effects to similar brain dysfunction (Boll, 1974). One of the Halstead Reitan test batteries is designed for children between the age of 9 and 14. The tests do not provide a conclusion but provide the vital information for performing further investigation. In the course of a neuropsychological interpretation,

neuropsychologists base their decision on the test results and other information derived from the case history and observations of the child's behaviour. The Halstead and Reitan tests, in particular, assist neuropsychologists in determining which part of the brain may be dysfunctional.

The training of medical specialists requires a considerable amount of time and resources. The service of medical doctors and psychologists is always in demand. For this reason, a considerable amount of research effort has been spent on devising AI systems that may assist the medical profession in performing their tasks. In particular, expert systems for medical interpretation and diagnosis have been an area of intensive study (Knights and Watson, 1968; Szolovits and Pauker, 1978; Weiss et al, 1978. Catanzarite and Greenburg, 1979; Reggia et al, 1980; Heaton et al, 1981; Wills, Teacher, Innocent and Bowley, 1982; Blois, 1983; Colbourn and McLeod, 1983; McMullin, 1983; Whitbeck and brooks, 1983; Yu, 1983; Wild, 1984; McLeod, 1985).

As pointed out by Whitbeck (1983), the function of medical diagnosis is not just to determine the nature of a disease but also to contribute to clinical reasoning. It is therefore necessary for computerized diagnostic systems to provide the reasoning behind their decisions in addition to

the decisions. Generally speaking, it is dangerous to entrust any important decisions totally to a machine. There has to be a way for people to examine the reasoning used by a machine on a case by case basis and judge its validity.

NEXSYS is an expert system developed by Schaefer and Russell (1985) for neuropsychological interpretation. This system was developed using the same data to be used for testing the present model. The NEXSYS system is rule-based and contains approximately 340 rules.

NEXSYS was developed using the technique of knowledge engineering. A neuropsychologist, referred to as DR, was consulted to develop the rules for the system. There are eight areas in the human brain. The first task in developing NEXSYS was to identify, for each of the 62 tests of the Halstead and Reitan battery, the functions of each brain area which the test purported to measure. This resulted in a list of tests associated with each brain area. The tests for each brain area were then sorted in groups according to the functions that they measure. For each group of tests, an exhaustive set of result patterns was prepared. Each pattern of results was then assigned to a conclusion qualified by the degree of dysfunction in a particular area of the brain.

Since there are no proven methods for confirming the results of neuropsychological interpretation, Schaefer and Russell measured the performance of NEXSYS in terms of the degree of agreement it has with human experts. Another neuropsychologist, referred to as RK, had already made decisions on the 154 cases. His decisions were used for comparisons of performance. The results of Schaefer and Russell's study are shown in Table 2.2 of Appendix 2. In Table 2.2, the results are shown by decision area and by type of comparison. There are eight decision areas represented by the eight areas of the brain. The eight decision areas are grouped into two groups of four, with each group representing a hemisphere of the brain. Three comparisons have been made. Between the two human experts, DR and RK, the percentage agreement varied from 79% in the left temporal (LT) lobe to 84% in the left and right frontal (LF and RF) lobes, with an overall agreement of 82%. Between NEXSYS and DR, the results varied from 77% in the LP lobe to 86% in the LO and RO lobes, with an overall agreement of 82%. Between NEXSYS and RK, there were larger variations in result, from 70% in the LT lobe to 84% in the RO lobe, with an overall agreement of 77%. This decrease in agreement and the increase in variation were expected since NEXSYS was based on the knowledge of DR instead of RK.

There are a number of reasons for using this problem for testing the proposed model. Firstly, it is a real-world problem. Secondly, the data for the problem meet all the requirements of the model. Data associated with natural problems are not always collected in such detail and in such structured form. Furthermore, the problem allowed a comparison to be made between the model's decisions and those of both human experts and a mechanical system which operates on different principles.

7.2 Structure of the Test Data

The test data consisted of the results of 62 tests performed on 154 cases. The coding forms for 4 of the cases were missing. This left 150 cases available for testing. For each of the 62 tests, the patient's score was coded as follows:

"0" indicated that the test score was less than 1 standard deviation below the mean;

"1" indicated that the test score was at least 1 standard deviation but less than 2 standard deviations below the mean;

"2" indicated that the test score was 2 or more standard deviations below the mean.

The 62 Halstead Reitan tests were treated as 62 separate feature dimensions. For each dimension, the feature values were represented by the above 3 codes. The input data was therefore represented by a 3x62 array, providing 186 elements.

For each input case, eight binary decisions, one for each area of the brain, were made by each of the experts, and by NEXSYS. These decisions indicated whether the respective part of the brain was dysfunctional or not. The binary decision was coded as follows:

"0" indicated that the respective part of the brain was judged to be normal;

"1" indicated that it was judged to be dysfunctional.

For convenience, the eight areas of the brain are referred to as decision areas. Since each expert, including NEXSYS, had to make a decision for each decision area, there were three decisions in each area.

In applying the proposed model, a pair of frequency arrays, one for each category, were created for each decision area. The frequency arrays provided the necessary data for the model's decision making process.

In the training phase, the model could be trained using the decisions of one or both human experts. The model could also be trained using the decisions of NEXSYS. This, however, was considered unnecessary since NEXSYS is an

"imitation" of DR, and it would be more direct to have the model trained by DR's decisions. The term "training" in the above context does not mean that the expert communicated with the model in person. It means that the model received the expert's decisions as feedback to its own decisions. In the case that the model was "trained" by both experts, the model received feedback from both experts, incrementing the frequency array twice, once for each expert.

7.3 Investigation and Results

There were a large number of issues about the proposed model that could be investigated. The single most important issue was whether the proposed model could learn and solve problems. Learning should be reflected by a learning curve that moves upward with the number of training cases the model has seen. The model's problem solving ability should be reflected by its overall performance, in terms of its degree of agreement with human experts. The performance of NEXSYS provided another standard for comparison. Also, existing learning models serving the same purposes were tested for comparison.

For ease of discussion, the tests conducted for this paper were grouped into the following categories:

(1) Learning ability of the proposed model - The purpose was to examine the learning ability of the proposed model. In addition, the assumptions made by the model were tested. A prototype modification of the model was also examined (see equation 1.03 in Section 5.5).

(2) Comparison between the proposed model, human experts and NEXSYS - The purpose was to examine the model's problem solving ability. The model's decisions were compared with those of the human experts. The degree of agreement achieved by the model were compared with those achieved by NEXSYS.

(3) The model's ability to classify new cases - The purpose was to examine the model's performance when cases unseen by the model before were presented. Comparison was made to the results in (2).

(4) Comparison of performance between the proposed model and discriminant analysis - The purpose was to confirm the model's problem solving ability by comparing its performance with the results of statistical functions generated using discriminant analysis.

(5) Comparison of performance between the proposed model and parallel distributed processing systems - The purpose again was to confirm the model's problem solving ability. The model's performance was compared with the results of neural networks

operating on the principle of parallel distributed processing.

7.3.1 Group 1: Learning Ability of Proposed Model

A number of tests were conducted to examine the model's learning ability. The first test examined the learning rate of the model. The model was based on the assumption that sensitivities to within-dimension and between-category frequency variations are essential to problem solving models (see Section 5). Subsequent tests examined whether this learning rate was maintained when this assumption was relaxed. In addition, a prototype modification of the model was also examined (see equation 1.03 in Section 5.5).

Test 1.1: The purpose of this test was to examine the learning rate of the proposed model as the number of training cases increased.

Method: Lee et al (1988) have shown that the order in which testing cases occur may affect the success rate of a model. In order to eliminate any potential ordering effect, 50 random sequences of the 150 cases were generated. The proposed model was then tested using all 50 random sequences. The results were averaged according to the order of appearance. Due to the enormous computing time required for testing 50 random sequences, the test was not conducted on all decision

areas. RK's decisions on the LF lobe of the brain were arbitrarily selected for training the model.

Results: Since the required decision was binary, the model would be expected to have a 50% agreement with the expert by chance alone. The test results are shown in Table 1.1 of Appendix 1. The results are divided into fifteen groups, according to the order of appearance. In each group, the scores for 50 random sequences of 10 cases were averaged. For the first 10 cases, the scores averaged to a 59% agreement with RK's decisions, which is slightly above the chance level. The percentage of agreement increased steadily with some minor fluctuations over the remaining fourteen groups. In the last group, where a minimum of 140 cases had already been shown, the average score rose to a 79.8% agreement with RK's decisions (see Appendix 7).

Conclusion: This test supported the learning ability of the model. The learning curve appeared to be still climbing in the last 10 cases. Whether the learning curve would have continued to rise is a question that could only be answered with more data. The model's performance in the last 10 cases was comparable to the performance of NEXSYS, which had a 75% agreement with

RK's decisions in the selected decision area. The two human experts had an 84% agreement for this set of decisions.

Test 1.2: The purpose of this test was to find out if the model would learn if it were made insensitive to within-dimension frequency variation.

Method: The method was identical to test 1 above. The model, however, was modified. The assumption of equal prior probability for all categories was relaxed. By relaxing this assumption, the model was no longer sensitive to within-dimension frequency variation (see equation 1.04 in Section 5.6).

Results: The modified model failed to show an upward learning curve (see Appendix 7). The results fluctuated between 52% and 66 % agreement (see Table 1.2 in Appendix 2). This was substantially lower than the 79.8% agreement achieved in Test 1.1.

Conclusion: The test indicated that for the model to learn, it must take into account within-dimension frequency variation.

Test 1.3: The purpose of this test was to examine the learning curve when the model is not sensitive to between-category frequency variation.

Method: The method was the same as in Test 1. Baysian probability, however, was no longer used in the decision making process. The frequency arrays were converted to an array of relative frequencies. The model would still be sensitive to within-dimension variation but would no longer be sensitive to between-category variation (see equation 1.02 in Section 5.4).

Results: The results were similar to test 2 (see Table 1.3 in Appendix 1). The modified model failed to show an upward learning curve (see Appendix 7). In general, performance was only slightly better than the chance level.

Conclusion: The test indicated that, for the model to learn, it must also be sensitive to between-category frequency variation.

Test 1.4: The purpose of this test was to examine the effectiveness of a prototype approach to modifying the frequency arrays.

Test 1.3: The purpose of this test was to examine the learning curve when the model is not sensitive to between-category frequency variation.

Method: The method was the same as in Test 1. Baysian probability, however, was no longer used in the decision making process. The frequency arrays were converted to an array of relative frequencies. The model would still be sensitive to within-dimension variation but would no longer be sensitive to between-category variation (see equation 1.02 in Section 5.4).

Results: The results were similar to test 2 (see Table 1.3 in Appendix 1). The modified model failed to show an upward learning curve (see Appendix 7). In general, performance was only slightly better than the chance level.

Conclusion: The test indicated that, for the model to learn, it must also be sensitive to between-category frequency variation.

Test 1.4: The purpose of this test was to examine the effectiveness of a prototype approach to modifying the frequency arrays.

Method: The method was identical to test 1. The frequency arrays were manipulated by assigning a value of 1 to the mode of each dimension, while all other elements were assigned a value of 0 (see equation 1.03 in Section 5.4).

Results: The results were similar to Tests 2 and 3 (see Table 1.4 in Appendix 1). The modified model failed to show any learning ability (see Appendix 7). The results fluctuated around the chance level.

Conclusion: The results show that the prototype approach to frequency manipulation did not improve the model's performance. It actually reduced the model's performance to around the chance level.

7.3.2 Group 2: Comparison with Human Experts and NEXSYS

The purpose of the following tests was to provide a thorough evaluation of the model's performance compared with the two human experts and with NEXSYS.

The proposed model could be trained by one single expert or by both experts. Since the development of NEXSYS was based on the knowledge of DR, it was appropriate to train the proposed model using DR's decision. A separate test which had the model trained by both experts was also conducted. This test also examined the correlation between the differential of similarity scores and the difference in expert decisions. The details of this test, its results and analysis are presented in Appendix 6.

Test 2.1: The purpose of this test was to compare the proposed model's performance with the performance of NEXSYS.

Method: The model was trained by the decisions of DR. Since the development of NEXSYS used all 150 cases, the model was trained using all 150 cases. The model was then tested using the same 150 cases. The frequency arrays were held constant during the testing phase.

Results: The results are shown in Table 2.1 of Appendix 2. Between the model and DR, the levels of agreement fluctuated between 34% in the LF lobe and 87% in the RP and RO lobes, with an overall agreement of 86%. Between the model and RK, the fluctuation was between 73% in the LT lobe and 83% in the RP lobe, with an overall agreement of 78%. The levels of overall agreement were comparable with those of 82% and 77% achieved by NEXSYS (see Table 2.2 in Appendix 2). Between DR and RK, the overall agreement was 82%. As shown in Table I, the 86% and 78% overall agreements achieved by the model also averaged to 82%.

	% of Overall Agreement		
	DR	RK	AVERAGE
Model	86	78	82
NEXSYS	82	77	80
DR		82	82
RK	82		82

Table I: Overall levels of agreement between the model, NEXSYS, DR and Rk.

Conclusion: The results indicated that the model's performance was comparable to that of NEXSYS's. The model also achieved a level of agreement comparable to that between the human experts.

7.3.3 Group 3: Model's Ability to Learn New Cases

In the foregoing, the model was tested using the same cases that were used for training. The results therefore reflect only the model's ability to classify repeated cases. Performance is likely to decrease when the model was tested with new cases. The concern was whether the decrease would be substantial enough to render the model impractical.

Test 3.1: The purpose was to examine the model's ability to classify cases that it had not seen in the training phase.

Method: The method was similar to test 2.1. The model again was trained using DR's decisions. However, the model was trained using only 149 cases. All possible combinations of 149 cases out of the 150 were identified. The test was performed on the remaining case of each combination.

Results: In comparison to test 2.1, the results indicated a decrease in performance (see Table 3.1 in Appendix 3). The levels of agreement with DR's decisions fluctuated between 75% and 85% for the eight areas of the brain, with an overall agreement of 80%. The levels of

agreement with RK's decisions fluctuated between 73% and 83%, with an overall agreement of 76%. As shown in Table II, the average degree of agreement decreased from 82% in the previous test to 78% in the current test.

	% of Overall Agreement		
	DR	RK	AVERAGE
Model (Test 2.1)	86	78	82
Model (Test 3.1)	80	76	78
NEXSYS	82	77	80
DR		82	82
RK	82		82

Table II: Overall levels of agreements between Test 2.1, Test 3.1, NEXSYS, DR and RK.

Conclusion: As expected, the model showed a decrease in performance. However, the 78% average agreement achieved by the model can still be considered satisfactory.

7.3.4 Group 4: Comparison with Discriminant Analysis

Discriminant analysis is a commonly used statistical approach to pattern recognition (see Section 3.2 for discussion on statistical pat' ern recognition). Due to the system requirements imposed by the present study, this approach was precluded in the development of the present model. The performance of discriminant functions, however, could be used as a standard for judging the model's performance.

Test 4.1: The purpose was to compare the model's performance with that of discriminant functions.

Method: The 150 cases were divided into two halves. The first half consisted of the first 75 cases and the second half, the remaining 75 cases. The first 75 cases were used to generate a discriminant function for each combination of human expert and decision area. A total of 16 functions were therefore generated. The same 75 cases were also used to train the model. Two versions of the model were generated, one for each expert. The remaining 75 cases were used for testing. The discriminant functions were held constant during the test, and so were the frequency arrays of the

models. The discriminant functions were generated using the PC version of SPSS+ with the default options provided by the software.

Since unintended effects might be created by the way cases were divided into the two halves, another test was conducted as a reliability check. In this test, the 75 training cases were randomly selected, leaving the remaining cases for testing.

Results: Due to a large number of missing data in dimension 37 (aphasia dyscalculia), discriminant functions could not be generated until dimension 37 was excluded from the analysis. Furthermore, 25 training cases contained at least one missing value. These 25 cases were also excluded. The resulting discriminant functions showed a 99% agreement with the experts for the cases used in training. When cases with missing values were included, the degree of agreement with DR's decisions fluctuated between 80% and 91% for the eight areas of the brain, with an overall agreement of 84% (see Table 4.1 in Appendix 4). The degree of agreement with RK's decisions fluctuated between 81% and 89%, with an overall agreement of 86%. Using the same set of data, the model's overall agreement with DR's decisions and

RK's decisions were 87% and 85%, respectively (see Table 4.2 in Appendix 4). In the second test where training and testing cases were randomly selected, the results were similar (see Tables 4.5 and 4.6 in Appendix 4). As shown in Table III, the performance of the discriminant functions was comparable to that of the model as far as training cases were concerned.

	% of Overall Agreement (training case only)		
	DR	RK	AVERAGE
Test 4.1			
Fnct	84	86	85
Model	87	85	86
Test 4.2			
Fnct	83	84	83.5
Model	89	87	88

Table III: Average % of overall agreement between discriminant functions and the model for training cases only.

The performance of the discriminant functions decreased drastically in the testing phase when new cases were used (see Table 4.3 in Appendix 4). The degree of agreement fluctuated around the chance level for the eight areas of the brain, with an overall agreement of 58% with DR's decisions and 57% with RK's decisions. Similar results were obtained in the second test (see Tables 4.7 and 4.8 in Appendix 4). The results

indicated that the discriminant functions were not effective in classifying new cases. The model, on the other hand, showed an acceptable level of performance (see Tables 4.3, 4.4, 4.7 and 4.8 in Appendix 4). As shown in Table IV, the 81% and 77.5% of average agreement achieved by the model were better than that of the discriminant functions by a wide margin.

	% of Overall Agreement (testing case only)		
	DR	RK	AVERAGE
Test 4.1			
Fnct	55	59	57
Model	82	80	81
Test 4.2			
Fnct	58	57	57.5
Model	79	76	77.5

Table IV: Average % of overall agreement between discriminant functions and the model for testing cases only.

Conclusion: Since the results were similar in both tests, the test results could be considered reliable. In the above tests, one problem was that discriminant analysis could not process data with missing values. This could be a serious drawback of this approach. In the training phase, the discriminant functions as well as the model have shown impressive performance. However, when new cases were used, the performance of the

discriminant functions dropped drastically while the model held the same level of performance. This implies that the model was more flexible than the discriminant functions.

7.3.5 Group 5: Comparison with Neural Network

Neural networks have gained considerable research attention in the last few years (see Section 3.3 for discussion on neural networks). This type of model was excluded from our consideration because they did not meet the requirements imposed by the present study. Like discriminant functions, the results of neural networks could also be used as a standard for evaluating the present model.

Test 5.1: The purpose of this test was to compare the performance of the proposed model with that of a neural network.

Method: Similar to test 4.1, the 150 cases were divided into two halves of 75 cases each. The first half consisted of the first 75 cases and the second half, the remaining cases. The first half was used to train the neural network. The second half was used to provide new cases for testing. Two neural networks were developed, one for each expert, and two versions of the model were also created. Like all previous tests, The neural networks and the model were held constant during the testing phase.

The neural network used in this test was structurally consistent with those described by Rumelhart, Hinton and Williams (1986a, 1986b). The generalized delta rule was used for propagation. The networks consisted of 186 input units, 93 hidden units, and 8 output units. The 186 input units were analogous to the 3x62 data input array, while the 8 output units represented the eight decision areas. Since there is no analytical method for determining the number of hidden units to use, the number of hidden units was arbitrarily chosen. The initial values for the weights and the thresholds were randomly selected from the range -1 to 1. The learning rate was set at 0.25 and the momentum in weight space was set at 0.9. The training of the neural network was stopped after 10, 20, 40, 80, and 120 repeated presentations of all training cases, which corresponded to the totals of 750, 1500, 3000, 6000 and 9000 training presentations, respectively. At each stop, the neural network was tested using both the training and testing cases.

Results: The results are shown in Tables 5.1 and 5.2 in Appendix 5. Table 5.1 shows the results on the training cases for the eight areas of the brain at each stop. The performance of the neural network became

stable after 40 cycles. At this point, the network had a 99% overall agreement with DR's decisions and a 98% overall agreement with RK's decisions for the 75 training cases.

For the 75 new cases, the degree of overall agreement at the same stop (40 cycles) dropped from 99% to 81% with DR's decisions, and from 98% to 76% with RK's decisions (see Table 5.2 in Appendix 5). As shown in Table V, when the same data were used, the present model had an 87% overall agreement with DR's decisions and an 85% overall agreement with RK's decisions for the training cases (see also Table 4.2 in Appendix 4). For the new cases, the levels of overall agreement were 82% and 80%, respectively (see also Table 4.4 in Appendix 4).

	% of Overall Agreement (after 40 cycles)		
	DR	RK	AVERAGE
Training			
Network	99	98	98.5
Model	87	85	86
Testing			
Network	81	76	78.5
Model	82	80	81

Table V: Average % of overall agreement between Neural Networks and the model after 40 cycles.

Conclusion: The results indicated that the neural network had an extremely high level of performance in categorizing cases that it had repeatedly seen. However, with new cases its performance was comparable to that of the present model.

Test 5.2: The literature review indicated that performance may be affected when the number of hidden units is either increased or reduced. A second test was conducted to examine the behaviour of the neural network when the number of hidden units was changed.

Method: Two separate runs of the neural network were made. In the first run, the number of hidden units was increased to 186. In the second run, the number of hidden units was decreased to 45. For each run, the network was stopped after 3000 (i.e. 40 repetitions) and 4500 (i.e. 60 repetitions) presentations. Only DR's decisions were used to train the network. Other than the above changes, the method was identical to test 1.

Results: For both runs, the changes in performance were negligible between the two stops, especially where new

cases were concerned (see Tables 5.3 and 5.4 in Appendix 5). This again indicated that the network began to stabilize after 40 cycles. As shown in Table VI, when the number of hidden units was increased to 186, the performance of the network decreased drastically. The degree of overall agreement with DR's decisions dropped from 99% to 61% for the training cases, and from 81% to 58% for the new cases. When the number of hidden units was reduced to 45, the performance was restored to an acceptable level. The degree of agreement with DR's decisions was 93% for training cases and 80% for new cases.

% of Overall Agreement

	DR
Training Cases	
45 hidden units	93
93 hidden units	99
186 hidden units	61
Testing Cases	
45 hidden units	80
93 hidden units	81
186 hidden units	58

Table VI: % of overall agreement with DR after 40 cycles.

Conclusion: The above results indicated that performance of the neural network is dependent on the number of hidden units. Increasing the number of hidden units does not

necessarily improve performance. With an inappropriate number of hidden units, the performance in some decision areas may drop to as low as the chance level.

8. Discussion

8.1 Summary Conclusion

A number of conclusions may be drawn from results of the above tests. First of all, it appears that the model has the ability to learn and to solve a real world problem. The learning ability of the model was shown in the group 1 and group 3 tests. In particular, test 1.1 showed a steadily climbing learning curve for the model (see Appendix 7). At the end of the test, the model provided an average agreement of 79.8% with DR's decisions. A similar level of performance was also observed in test 3.1 where a full scale test was conducted on all decision areas. It was surprising to observe the poor performance shown by the model when its assumptions were relaxed one at a time, despite the fact that a decrease in performance was expected. The model was based on the assumptions that the sensitivities to (1) within-dimension and (2) between-category frequency variations are essential in performing classification tasks. In the previous discussion (see Section 5.8), it was postulated that these sensitivities would improve the performance of classification models. The findings in the group 1 tests suggested, however, that such sensitivities were not just beneficial but necessary for solving real-

world problems, since without these sensitivities the model failed to show any clear indication of learning.

Among the group 1 tests, the poorest performance was provided by the prototype model. This finding was not surprising. As shown in Section 5.5, the prototype approach would reduce a model's sensitivity to the difference in frequency patterns between categories.

The model's problem solving ability was also examined. In a number of tests, the emphasis was placed on the model's performance compared to other models serving the same purpose. In comparison to NEXSYS, a rule-based model specially designed for the problem, the model provided comparable performance (see group 2 tests). In comparison to discriminant analysis, a widely used statistical technique for pattern recognition, the model provided superior performance when new cases were used (see group 4 tests). In comparison to neural networks, models that simulate neural activities, the model provided comparable performance (see group 5 tests). When new cases were used, these tests indicated that the model performed as well as any other models tested in the present study. In addition, the model's performance was also comparable to the 82% agreement between the two human experts. Based on the above

evidence, it could be concluded that the model is capable of solving real-world problems similar to the one presented in this study. This does not imply that the model could solve all problems. As indicated in previous discussions, the model functions as the micro-structure of learning systems. The function of the micro-structure has been equated with the ability to perform classification tasks, or the ability to make a choice among a set of alternatives. Some problems might require a macro-structural organization. In this case, the model would only provide one part of a larger system.

An unexpected finding in the above tests was the poor performance shown by discriminant functions when new cases were used. A possible explanation to this poor performance is that the cases used for the tests were not randomly selected. The test data were selected for developing the NEXSYS system. Presumably, each of these cases might carry certain unique information for designing NEXSYS. These cases might not be a representative sample of the patient population. If the sample was biased, the same bias would be shown in the resulting discriminant functions. The bias might affect the functions' performance when new cases were presented. Since the model was not based on any sampling assumptions, its performance was not affected.

Another interesting finding was that neural networks performed poorly when the number of hidden units was increased. In Rumelhart, Hinton and Williams (1986b), it has been shown that interference caused by unnecessary hidden units could result in a local minimum. This might explain the decrease in performance.

Seidenberg and McClelland (1989) provided insight on the functions of hidden units. Their study indicated that when there were a large number of hidden units, some hidden units would specialize their functions to represent case-specific features. However, when the number of hidden was reduced, the hidden units would generalize their functions to represent general features of the cases. Presumably, the specialized hidden units would be activated only under specific conditions. Conversely, the generalized hidden units would be activated under general conditions. It is theoretically possible for a neural network to become over-specialized. An over-specialized neural network would become case-specific. The network would be effective when familiar cases were given. If novel cases were presented, the network might fail to perform because there might not be enough generalized hidden units to handle the new cases. Since the neural networks in the present study performed

poorly when the number of hidden unit was increased, it is possible that the neural networks had been over-specialized.

In comparison to neural network, the model is simpler in structure and can be made to provide inductive reasoning for its decisions. For example, the model can upon request provide a comparison of Baysian probabilities between categories and identify the dimensions which contribute most to the decisions. The reasoning provided by the model would be different from the reasoning provided by NEXSYS which is rule-based. One major drawback of neural networks is that they do not provide explanation for their decisions. Seidenberg and McClelland (1989) attempted to understand the decision rationale of a neural network by analyzing the input conditions under which certain hidden units were activated. The technique they used was time-consuming and would be impractical for complex problems.

In summary, the model chosen for the present study was supported by all the above findings and analyses. The model has a simple structure, is capable of explaining its decisions and provided a level of performance comparable to that of the human experts and the other models investigated in this study.

8.2 Implications for Human Learning

The model presented in this study was intended strictly for machine learning. However, any models that facilitate learning may have implications for human learning. The present model, if interpreted as a model of human learning, would suggest that only the within-dimension distributions of frequencies are attended to in concept learning. Even though the model is sensitive to both within-dimension and between-category frequency variations, the between-category sensitivity is derived using within-dimension frequencies (see Section 5.7). The within-dimension and between-category frequency variations are similar to the within-group and between-group frequency differentials mentioned in Bourne et. al (1976). In their study, it was shown that performance of human subjects was affected when the within-group frequency differential was reduced. However, a reduction in between-group frequency differential did not provide the same result. Their findings suggested that only within-group frequencies are attended to in human learning, which is consistent with the model.

8.3 Directions for Further Development

There are at least three directions for further research. The first direction is to find a way for the model to detect automatically correlated features relevant to a problem. In its present form, the model requires the user to specify relevant correlated features. The effects of certain correlated features (e.g. joint features) may already be sufficiently accounted for by the model. In this case, it may not be necessary to duplicate the effects by including the correlated features. However, there are certain types of correlated features with effects undetected by the model. There may be ways of detecting these features using the frequency data. For example, "exclusive or" features would show matching frequency distributions between the correlated dimensions. If these features could be detected, they could be automatically added to the model. An additional routine could be designed to eliminate features that do not improve the model's performance, so that the list of feature dimensions would not grow to an unmanageable extent. This feature elimination routine could apply to the original features as well as to the added features.

A second direction for improving the model is to examine whether the performance of the model can be improved by incorporating additional information available in the task

domain. The model, in its present form, uses a minimal amount of domain-specific information. Presumably, the model's performance could be improved by utilizing more prior information. For example, the model's performance might be improved by weighting the dimension according to certain prior knowledge of the task domain, or by weighting the feedback according to its probability of being true.

Another direction for further development would be to examine ways of incorporating the model ıto a macro-structure. Two types of macro-structures are widely used in artificial intelligence systems: the tree structure and the network structure. A more intriguing idea is to use the model for shaping a macro-structure. Since the model has the ability to determine the similarity between events, the model could theoretically be used to change the connections in a macro-structure so that the strength of connections between nodes could be maximized. Consider, for example, a network structure which has robins connected to living things and then connected to birds (robins-living things-birds). When their similarity scores are compared, it would be discovered that robins are more similar to birds than to living things. The connections could then be adjusted by connecting robins to birds first before birds are connected to living things (robins-birds-living). If the macro-

structure has a small number of nodes, the connections between nodes could be sequentially tested and adjusted. However, if there are a large number of nodes, a method would be required to detect connections that need adjustment. How this could be done would require further research and probably a different problem for testing.

Since the model has the ability to determine for a node its appropriate location in a macro-structure, the model could also be used for building macro-structures. Consider, for example, that a number of categories are given without their interconnections specified. A macro-structure could be built by first randomly selecting a category as the starting point. Using the model as a tool for determining the strength of association between categories, new categories could be added one by one to form an organized structure. Presumably, the structure would not be in its ideal form after it is constructed. The model could be applied to shape the structure until connections between nodes are sufficiently strengthened.

Bibliography

Anderson, J. R., Kline, P. J. & Beasley, C. M. A general learning theory and its application to schema abstraction. In G. H. Bower (Ed.), The psychology of learning and motivation 13. New York: Academic Press, 1979.

Ashby, W. Design of a Brain: The Origin of Adaptive Behaviour. New York: John Wiley & Sons, 1960.

Barresi, J., Robbins, D., & Dhain, K. Role of distinctive features in the abstraction of related concepts. Journal of Experimental Psychology: Human learning and Memory, 1975, 104, 360-368.

Beach, L. R. Cue Probabilism and inference behaviour. Psychological Monographs, 1964, 78(5), 1-20.

Block, H. D. The perceptron: A model of brain functioning, I. Review of Mathematical Physics, 1961, 34(1), 123-135.

Blois, M. S. Conceptual issues in computer-aided diagnosis and the hierarchical nature of medical knowledge. Journal of Medicine and Philosophy, 1983, 8, 29-50.

Boll, T. J. Behavioral correlates of cerebral damage in children aged 9 through 14. In R. M. Reitan & L. A. Davison (Eds.) Clinical Neuropsychology: Current Status and Applications, London: Hemisphere Publishing, 1974, 91-120.

Boll, T. J. The Halstead-Reitan Neuropsychology Battery. In S. B. Filskov & T. J. Boll (Eds.) Handbook of Clinical Neuropsychology, New York: John Wiley & Sons, 1981, 577-607.

Bourne, L. E., Jr., Ekstrand, B. R., & Montgomery, B. Concept learning as a function of the conceptual rule and the availability of positive and negative instances. Journal of Experimental Psychology, 1969, 82, 538-544.

Bourne, L. E., Jr., Ekstrand, B. R., Lovallo, W. R., Kellogg, R. T., Hiew, C. C., & Yaroush, R. A. Frequency analysis of attribute identification. Journal of Experimental Psychology: General, 1976, 105, 294-312.

Bower, G. H., & Trabasso, T. Reversals prior to solution in concept identification. Journal of Experimental Psychology, 1963, 66, 409-418.

Carbonell, J. G. Learning by analogy: formulating and generalizing plans from past experience. T. M. In R. S. Michalski, J. G. Carbonell & T. M. Mitchell (Eds.) Machine Learning: An Artificial Intelligence Approach. Palo Alto, Calif: Tioga Publishing, 1983, 137-159.

Carbonell, J. G., Michalski, R. S. & Mitchell, T. M. An overview of machine learning. In R. S. Michalski, J. G. Carbonell & T. M. Mitchell (Eds.) Machine Learning: An Artificial Intelligence Approach. Palo Alto, Calif: Tioga Publishing, 1983, 3-24.

Catanzarite, V. A. & GreenBurg, A. G. Neurologist: a computer program for diagnosis in neurology. IEEE, 1979, 64-71.

Chumbley, J. I., Sala, L. S., & Bourne, L. E., Jr. Bases of acceptability ratings in quasinaturalistic concept tasks. Memory & Cognition, 1978, 6, 217-226.

Colbourn, M. J. & McLeod, J. The potential and feasibility of computer-guided educational diagnosis. In R. E. A. Mason (Ed.) Information Processing 83, North-Holland: Elsevier Science Publishers, 1983, 891-896.

Collins, A. M. & Quillian, M. R. Retrieval time from semantic memory. Journal of Verbal Learning and Verbal Behaviour, 1968, 8, 240-248.

Collins, A. M. & Quillian, M. R. Facilitating retrieval from semantic memory: the effective of repeating part of an influence. Acta Psychologica, 1970, 33, 304-314.

Culberson, J. T. The Minds of Robots. Urbana, Illinois: University of Illinois Press, 1963.

Davis, R. & King, J. J. An overview of production systems. In E. Elcock & D. Michie (Eds.) Machine Intelligence 8, Chichester, England: Ellis Horwood, 1977, 300-332

Davis, W. D. T. System Identification for Self-Adaptive Control. New York: Wiley and Sons, 1970

Dougherty, J. W. D. Salience and relativity in classification. American Ethnologist, 1978, 5, 6-80.

Estes, W. K. Structural aspects of associative models for memory. In C. N. Cofer (Ed.). The structure of human memory. New York: Freeman, 1976.

Estes, W. K. Array Models for Category Learning. Harvard University Press, 1986.

Feigenbaum, E. A. Themes and case studies of knowledge engineering. In D. Michie (Ed.) Expert Systems in the Micro-electronic Age. Edinburgh: Edinburgh University Press, 1979, 3-25.

Findler, N. V. (Ed) Associative Network: The Representation and Use of Knowledge by Computers. New York: Academic Press, 1979.

Friedberg, R. M. A learning machine: Part 1. IBM Journal of Research and Development, 1958, 2, 2-13.

Friedberg, R., Dunham, B., & North, T. A learning machine: Part 2. IBM Journal of Research and Development, 1959, 3(3), 282-287.

Fu, K. S. Sequential Methods in Pattern Recognition and Machine Learning. New York: Academic Press, 1968.

Fu, K. S. Pattern Recognition and Machine Learning. New York: Plenum Press, 1971.

Fu, K, S. & Tou, J. T. Learning Systems and Intelligent Robots. New York: Plenum Press, 1974.

Fukanagan, K. Introduction to Statistical Pattern Recognition. New York: Academic Press, 1972.

Genero, N. & Canter, N. Exemplar prototypes and clinical diagnosis: Toward a cognitive economy. Journal of Social and Clinical Psychology, 1987, 5, 59-78.

Goldman, D., & Homa, D. Integrative and meteric properties of abstracted information as a function of category discrimination, instance variability, and experience. Journal of Experimental Psychology: Human Learning and Memory, 1977, 3, 375-385.

Hayes-Roth, B., & Hayes-Roth. F. Concept learning and the recognition and classification of exemplars. Journal of Verbal Learning and Verbal Behavior, 1977, 16, 321-338.

Heaton, R. K., Grant, I., Anthony, W. Z. & Lehman, R. A. Conceptual Issues in Computer-aided Diagnosis and the Hierarchical Nature of Medical Knowledge. Journal of Medicine and Philosophy, 1983, 8, 29-50.

Highleyman, W. H. Linear decision functions, with applications to pattern recognition. Proceeding of IRE, 1967, 50, 1501-1504.

Holland, J. H. Adaptive algorithms for discovering and using general patterns in growing knowledge bases. Policy Analysis and Information Systems, 1980, 4(3).

Hopcroft, J. E. & Ullman, J. D. Formal Languages and Their Relation to Automata. Reading, Mass: Addison-Wesley, 1969.

Jajuga, K. Bayes classification rule for the general discrete case. Pattern Recognition, 1986, 19(5), 413-415.

Johnson, P. E. What kind of expert should a system be? Journal of Medicine and Philosophy, 1983, 8, 77-97.

Kanal, L. Patterns in pattern recognition: 1968-1974. IEEE Transactions on Information Theory, 1974, IT-20, 6, 697-722.

Kazemierczak, H. & Stienbuch, K. Adaptive systems in pattern recognition. IEEE Transactions of Electronic Computers, 1963, EC-12, 5, 822-835.

Kellogg, R. T. Simple feature frequency versus feature validity models of formation of prototypes. Perception and Motor Skills, 1980a, 51, 295-306.

Kellogg, R. T. Feature frequency and hypothesis testing in the acquisition of rule-governed concepts. Memory and Cognition, 1980b, 8, 297-303.

Kellogg, R. T. Is conscious attention necessary for long-term storage? Journal of Experimental Psychology: Human Learning and Memory, 1980c, 6, 379-390.

Kellogg, R. T. Feature frequency and concept learning: what is counted? Memory and Cognition, 1981, 9, 157-163.

Kellogg, R. T., Bourne, L. E., Jr., & Ekstrand, B. R. Feature frequency and the acquisition of natural concepts, American Journal of Psychology, 1978, 91, 211-222.

Kellogg, R. T., Robbins, D. W., & Bourne, L. E., Jr. Memory for intratrial events in feature identification. Journal of Experimental Psychology: Human Learning and Memory, 1978, 4, 256-265.

Knights, R. M. & Watson, P. The use of computerized test profiles in neuropsychological assessment. Journal of Learning Disabilities, 1968, 1(12), 696-709.

Kruskal, J. B. multidimensional scaling by optimizing goodness of fit to a non-metric hypothesis. Psychometrika, 1964, 29, 1-27.

Kuipers, B. A frame for frames: representing knowledge for recognition. In D. G. Bobrow & A. Collins (Eds.) Representation and Understanding: Studies in Cognitive Science. New York: Academic Press, 1975.

Lee, E.S., MacGregor, J. N., Bavelas, A., Lam, N., Mirlin, L. & Morrison, I. The effects of error transformations on classification performance. Journal of Experimental Psychology, Learning, Memory and Cognition, 1988.

McClelland, J. L., Rumelhart, D. E. & Hinton, G. E. The appeal of parallel distributed processing. In D. E. Rumelhart & J. L. McClelland (Eds.) Parallel distributed processing: Explorations in the Microstructure of Cognition, Vol. 1, London: MIT Press, 1986, 3-44.

McCloskey, M. E. & Glucksberg, S. Natural categories: Well defined or fuzzy sets? Memory and Cognition, 1978, 6, 562-572.

McCulloch, W. S. & Pitts, W. A logical calculus of ideas imminent in nervous activity. Bulletin of Mathematical Biophysics, 1943, 5, 115-133.

McDermott, J. & Forgy, C. Production system conflict resolution strategies. In D. A. Waterman & F. Hayes Roth (Eds.) Pattern Directed Inference Systems. New York: Academic Press, 1978, 177-199.

McLeod, J. & Jones, M. The computer as an aid for those with special needs. International Conference 85, Sheffield City Polytechnic, 1985, 69-78.

McMullin, E. Diagnosis by computer. Journal of Medicine and Philosophy, 1983, 8, 5-27.

Medin, D. L. A theory of context in discrimination learning. In G. H. Bower (Ed.). The Psychology of Learning and Motivation 9. New York: Academic Press, 1975.

Medin, D. L. Structural Principles in Categorization. In T. J. Tighe & B. E. Shepp (Eds.) Perception, Cognition and

Development: Interactional Analysis. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1983.

Medin, D. L. Commentary on "Memory storage and retrieval process in category learning. Journal of Experimental Psychology: General, 1986, 115(4), 373-381.

Medin, D. L. Concept and concept structure. American Psychologist, 1989, 44(12), 1469-1481.

Medin, D. L., Altom, M. W., Edelson, S. M., & Freko, D. Correlated symptoms and simulated medical classification. Journal of Experimental Psychology: Learning, Memory and Cognition, 1982.

Medin, D. L., Altom, M. W., & Murphy, T. D. Given versus induced category representations: use of prototype and exemplar information in classification. Journal of Experimental Psychology: Learning, Memory, and Cognition, 1984, 10, 333-352.

Medin, D. L., Dewey, G. I., & Murphy, T. D. Relationship between item and category learning: Evidence that abstraction is not automatic. Journal of Experimental Psychology: Learning, Memory, and Cognition, 1983, 9, 607-625.

Medin, D. L. & Ortony, A. Psychological essentialism. In S. Vosniadou & A. Ortony, (Eds), Similarity and Analogical Reasoning,New York: Cambridge University Press, 1989, 179-195.

Medin, D. L. & Ross, B. H. The specific character of abstract thought: Categorization, problem solving, and induction. In R. J. Sternberg (Ed), Advances in the Psychology of Human Intelligence, NJ: Erlbaum, 1989, 189-223.

Medin, D. L. & Shoben, E. J. Context and structure in conceptual combination. Cognitive Psychology, 1988, 20, 158-190.

Medin, D. L., & Schaffer, M. M. Context theory of classification learning. Psychology Review, 1978, 85, 207-238.

Medin, D. L., & Smith, E. E. Strategies and classification learning. Journal of Experimental Psychology: Human Learning and Memory, 1981, 7, 241-253.

Mendel, T. & Fu, K. S. Adaptive Learning and Pattern Recognition: Theory and Applications. New York: Spartan Books, 1970.

Michalski, R. S. A theory and methodology of inductive learning. In R. S. Michalski, J. G. Carbonell & T. M. Mitchell (Eds.) Machine Learning: An Artificial Intelligence Approach. Palo Alto, Calif: Tioga Publishing, 1983, 83-134.

Michalski, R. S. & Stepp, R, E. Learning from observation: conceptual clustering. In R. S. Michalski, J. G. Carbonell & T. M. Mitchell (Eds.) Machine Learning: An Artificial Intelligence Approach. Palo Alto, Calif: Tioga Publishing, 1983, 331-360.

Millward, R. B., & Spoehr, K. T. The direct measurement of hypothesis-sampling strategies. Cognitive Psychology, 1973, 4, 1-38.

Minsky, M & Papert, S. Perceptrons. Cambridge, Mass: MIT Press, 1969.

Muggleton, S. H. Inductive Acquisition of Expert System. University of Edinburgh, 1986, Doctoral Thesis.

Murphy, G. L. Cue validity and basic levels in categorization. Psychological Bulletin, 1982.

Neumann, P. G. An attribute frequency model for the abstraction of prototypes. Memory & Cognition, 1974, 2, 241-248.

Neumann, P. G. Visual prototype formation with discontinuous representation of dimensions of variability. Memory & Cognition, 1977, 5, 187-197.

Newell, A., Shaw, J. & Simon, H. A. Empirical exploration of the logic theory machine: a case study in heuristics. In E. Feigenbaum & J. Feldman (Eds.) Computers and Thought. New York: McGraw-Hill, 1963.

Nilsson, N. J. Learning machines. New York: McGraw-Hill, 1965.

Nilsson, N. J. Problem-Solving Method in Artificial Intelligence. New York: McGraw Hill, 1971.

Nosofsky, R. M. Choice, similarity, and the context theory of classification. Journal of Experimental Psychology: Learning, Memory, and Cognition, 1984, 10, 104-114.

Nosofsky, R. M. Attention and learning processes in the identification and categorization of integral stimuli. Journal of Experimental Psychology: Learning, Memory, and Cognition, 1987, 13, 87-108.

Nosofsky, R. M. Exemplar-based accounts of relations between classification, recognition, and typicality. Journal of Experimental Psychology: Learning, Memory, and Cognition, 1988, 17, 700-708.

Oden, G. C. Concept, knowledge, and thought. Annual Review of Psychology, 1987, 38, 203-227.

Posner, M. I., Goldsmith, R., & Welton, R. D. Perceived distance and the classification of distorted patterns. Journal of Experimental Psychology, 1967, 73, 28-38.

Quillian, M. R. Semantic memory. In M. Minsky (Ed.) Semantic Information Processing. Cambridge, Mass: M.I.T. Press, 1968, 227-270.

Raiffa, H. Decision Analysis: Introductory Lectures on Choices Under Uncertainty. Reading, Mass: Addison-Wesley, 1970.

Rashevsky, N. Mathematical Biophysics. Chicago: University of Chicago Press, 1948.

Reed, S, K. Pattern recognition and categorization. Cognitive Psychology, 1972, 3, 382-407.

Reed, S. K. Psychological Processes in Pattern Recognition. New York: Academic Press, 1973.

Reggia, J. A., Pula, T. P., Price, T. R. & Perricone, B. T. Towards an intelligent textbook of neurology. IEEE, 1980, 190-199.

Reitan R. M. & Davison, L. A. Description of psychological tests and experimental procedures. In R. M. Reitan & L. A. Davison (Eds.) Clinical Neuropsychology: Current Status and Applications, London: Hemisphere Publishing, 1974, 363-385.

Reitman, J. S., & Bower, G. H. Storage and later recognition of exemplars of concepts. Cognitive Psychology, 1973, 4, 194-206.

Rosch, E. H. Natural Categories. Cognitive Psychology, 1973a, 4, 328-350.

Rosch, E. H. On the internal structure of perceptual and semantic categories. In T. E. Moore (Ed.). Cognitive Development and the Acquisition of Language. New York: Academic Press, 1973b.

Rosch, E. H. Universals and cultural specifics in human categorization. In R. Breslin, S. Bochner, & W. Lonner (Eds.). Cross-cultural Perspectives on Learning. New York: halsted Press, 1975.

Rosch, E. H. Principles of categorization. In E. H. Rosch & B. B. Llyod (Eds.). Cognition and Categorization. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1978.

Rosch, E. H., & Mervis, C. B. Family resemblances: Studies in the structure of categories. Cognitive Psychology, 1975, 7, 573-605.

Rosch, E. H., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. Basic objects in natural categories. Cognitive Psychology, 1976, 8, 382-439.

Rosch, E. H., Simpson, S., & Miller, R. Structural bases of typicality effects. Journal of Experimental Psychology: Human Perception and Performance, 1976, 4, 491-502.

Rosenblatt, F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychological Review, 1958, 65, 386-407.

Roth, E. M. and Shoben, E. J. The effect of context on the structure of categories. Cognitive Psychology, 1983, 15, 346-378.

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. Nature, 1986a, 323(9), 533-536.

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.) Parallel distributed processing: Explorations in the Microstructure of Cognition, Vol. 1, London: MIT Press, 1986b, 319-362.

Rychener, M. D. The instructible production system: a retrospective analysis. T. M. In R. S. Michalski, J. G. Carbonell & T. M. Mitchell (Eds.) Machine Learning: An Artificial Intelligence Approach. Palo Alto, Calif: Tioga Publishing, 1983, 429-460.

Sacerdoti, E. D. Planning in a hierarchy of abstract spaces. Artificial Intelligence, 1974, 5, 115-135.

Schaefer, B. A. & Russell, D. L. NEXSYS: A knowledge-based system for the neuropsychological diagnosis of brain dysfunction in children. Cognitiva 85. Paris, 1985.

Sebestyen, G. S. Decision-making Processes in Pattern Recognition. New York: Macmillan, 1962.

Seidenberg, M.S. and McClelland, J. L. A distributed, developmental model of word recognition and naming. Psychological Review, 1989, 96(4), 523-568.

Selfridge, O. G. Pandemonium: A paradigm for learning. In D. Blake & A. Uttley (Eds.) Proceedings of the Symposium on Mechanization of Though Processes. London: HMSO, 511-529, 1959.

Shepard, R. N. Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. Psychometrika, 1957, 22, 325-345.

Shepard, R. N. Stimulus and response generalization: Tests of a model relating generalization to distance in psychological space. Journal of Experimental Psychology, 1958, 55, 509-523.

Shepard, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function. Psychometrika, 1962, 27, 125-140.

Simon, H. A. Why should machine learn? In R. S. Michalski, J. G. Carbonell & T. M. Mitchell (Eds.) Machine Learning: An Artificial Intelligence Approach. Palo Alto, Calif: Tioga Publishing, 1983, 25-38.

Smith, E. E., & Medin, D. L. Strategies and classification learning. Cambridge, Mass.: Harvard University Press, 1981.

Strange, W., Keeney, T., Kessel, F. S., & Jenkins, J. J. Abstraction over time of prototypes from distractions of random dot patterns: A replication. Journal of Experimental Psychology, 1970, 83, 508-510.

Szolovits, P. & Pauker, S. G. Categorical and probabilistic reasoning in medical diagnosis. Artificial Intelligence, 1978, 11, 115-144.

Thompson, B. and Thompson, W. Finding rules in data. Byte, 1986, Nov, 149-158.

Toppino, T. C., & Bucher, N. M. Acquiring conjunctive concepts: When and why does feature frequency affect feature identification? Memory and Cognition, 1983, 11(4), 407-414.

Truxal, T. G. Automatic Feedback Control System Synthesis. New York: McGraw Hill, 1955.

Tsypkin, Y. Z. Self-learning - What is it? IEEE Transaction on Automatic Control, 1968, AC-18 (2), 109-117

Tsypkin, Y. Z. Adaptation and Learning in Automatic Systems. New York: Academic Press, 1971.

Tsypkin, Y. Z. Foundations of the Theory of Learning Systems. New York: Academic Press, 1973.

Watanabe, S. Information-theoretic aspects of inductive and deductive inference. IBM Journal of research and Development, 1960, 4(2), 208-231.

Weiss, S. M. , Kulikowski, C. A., Amarel, S. & Safir, A. A model-based method for computer-aided medical decision-making. Artificial Intelligence, 1978, 11, 145-172.

Whitbeck, C. What is diagnosis? Some critical reflections. Metamedicine, 1981, 2, 319-329.

Whitbeck, C. & Brooks, R. Criteria for evaluating a computer aid to clinical reasoning. Journal of Medicine and Philosophy, 1983, 8, 51-65.

Widrow, B. Generalization and information storage in networks of adelaine 'neurons'. In M. C. Yovitz, G. T. Jacobi & G. D. Goldstein (Eds.) Self-Organizing Systems. Washington, D.C.: Spartan Books, 1962, 435-461.

Wild, U. W. SIC - A Program to Summarize, Interpret, and Compare Intellectual and Neuropsychological Test Performance, Department of Psychology, Simon Fraser University, (unpublished paper), 1985.

Wills, K., Teacher, D., Innocent, P. & Du Bouley, G. H. An expert system for the medical diagnosis of brain tumours. International Journal of Man-Machine Studies, 1982, 16, 341-349.

Young, F. W. TORSCA-9, a Fortran IV program for nonmetric multidimensional scaling. Behavioral Science, 1968, 13(4), 343-344.

Young, R. M. Production systems for modelling human cognition. In D. Michie (Ed.) Expert System in the Micro Electronic Age. Edinburgh, Scotland: Edinburgh University Press, 1979, 34-45

Yovitz, M. C., Jacobi, G. T. & Goldstein, G. D. (Eds.) Self-Organizing Systems. Washington, D.C.: Spartan Books, 1962.

Yu, V. L. Conceptual obstacles in computerized medical diagnosis. Journal of Medicine and Philosophy, 1983, 8, 67-75.

Appendix 1: Results on Model's Learning Ability

Table 1.1: Results of Test 1.1
(% of Agreement between Proposed Model and RK for 50 Random Sequences)
(Scores Averaged by Order of Appearance for Groups of 10 Random Cases)

Cases in Random Order	1 to 10	11 to 20	21 to 30	31 to 40	41 to 50	51 to 60	61 to 70	71 to 80
Agreement	59.0	67.6	68.8	68.8	74.4	74.6	73.4	76.4

Cases in Random Order	81 to 90	91 to 100	101 to 110	111 to 120	121 to 130	131 to 140	141 to 150
Agreement	77.2	77.0	75.2	75.6	78.6	77.6	79.8

Table 1.2: Results of Test 1.2
(% of Agreement between RK and Model without Within-Dimension Frequency Sensitivity)
(for 50 Random Sequences)
(Scores Averaged by Order of Appearance for Groups of 10 Random Cases)

Cases in Random Order	1 to 10	11 to 20	21 to 30	31 to 40	41 to 50	51 to 60	61 to 70	71 to 80
Agreement	52.8	55.2	54.0	56.8	55.2	61.2	62.0	60.2

Cases in Random Order	81 to 90	91 to 100	101 to 110	111 to 120	121 to 130	131 to 140	141 to 150
Agreement	65.6	58.6	55.4	59.6	51.6	54.6	58.8

Table 1.3: Results of Test 1.3
(% of Agreement between RK and Model without Between-Category Frequency Sensitivity)
(for 50 Random Sequences)
(Scores Averaged by Order of Appearance for Groups of 10 Random Cases)

Cases in Random Order	1 to 10	11 to 20	21 to 30	31 to 40	41 to 50	51 to 60	61 to 70	71 to 80
Agreement	52.0	59.2	55.2	54.8	53.8	53.4	54.8	52.8

Cases in Random Order	81 to 90	91 to 100	101 to 110	111 to 120	121 to 130	131 to 140	141 to 150
Agreement	49.8	55.6	54.8	52.6	53.8	53.6	54.4

Table 1.4: Results of Test 1.4
(% of Agreement between RK and Model with Prototype Conversion of Frequencies)
(for 50 Random Sequences)
(Scores Averaged by Order of Appearance for Groups of 10 Random Cases)

Cases in Random Order	1 to 10	11 to 20	21 to 30	31 to 40	41 to 50	51 to 60	61 to 70	71 to 80
Agreement	52.2	54.4	56.2	50.6	51.8	50.4	45.4	52.2

Cases in Random Order	81 to 90	91 to 100	101 to 110	111 to 120	121 to 130	131 to 140	141 to 150
Agreement	53.4	45.2	53.2	51.4	51.8	47.6	46.8

Appendix 2: Results on Comparison between NEXSYS and Model

Table 2.1: Results of Test 2.1
(% of Agreement Between Human Expert and Model Trained by DR)

% of Agreement	Left Hemisphere				Right Hemisphere				
	LF	LP	LT	LO	RF	RP	RT	RO	OVERALL
Model vs. DR	84	85	85	85	85	87	85	87	86
Model vs. RK	81	78	73	76	74	83	80	80	78

Table 2.2: Results of Schaefer and Russell's Study
(% of Agreement Among Human Experts and NEXSYS)

% of Agreement	Left Hemisphere				Right Hemisphere				
	LF	LP	LT	LO	RF	RP	RT	RO	OVERALL
DR vs. RK	84	83	79	80	84	82	81	80	82
NEXSYS vs. DR	82	77	81	86	81	81	81	86	82
NEXSYS vs. RK	75	73	70	76	79	80	79	84	77

Appendix 3: Results on Model's Ability to Learn New Cases

Table 3.1: Results of Test 3.1
(Performance of Model with New cases)

% of Agreement	Left Hemisphere				Right Hemisphere				
	LF	LP	LT	LO	RF	RP	RT	RO	OVERALL
Model vs. DR	77	79	75	79	81	85	81	83	80
Model vs. RK	74	75	68	75	73	83	77	79	76

Appendix 4: Results on Comparison with Discriminant Analysis

Table 4.1: Results of Test 4.1 - Sequential Selection of Cases
(Performance of Discriminant Functions: Training Cases)
(including cases with missing data)

% of Agreement	Left Hemisphere				Right Hemisphere				
	LF	LP	LT	LO	RF	RP	RT	RO	OVERALL
Fnct vs. DR	91	84	81	80	81	84	84	85	84
Fnct vs. RK	87	88	83	89	88	83	85	81	86

Table 4.2: Results of Test 4.1 - Sequential Selection of Cases
(Performance of Model: Training Cases)
(including cases with missing data)

% of Agreement	Left Hemisphere				Right Hemisphere				
	LF	LP	LT	LO	RF	RP	RT	RO	OVERALL
Model vs. DR	87	84	85	85	91	88	87	89	87
Model vs. RK	81	80	84	[illegible]	88	91	88	88	85

Table 4.3: Results of Test 4.1 - Sequential Selection of Cases
(Performance of Discriminant Functions: Testing Cases)

% of Agreement	Left Hemisphere				Right Hemisphere				
	LF	LP	LT	LO	RF	RP	RT	RO	OVERALL
Fnct vs. DR	57	53	59	57	53	55	51	57	55
Fnct vs. RK	56	62	61	59	56	65	59	53	59

Table 4.4: Results of Test 4.1 - Sequential Selection of Cases
(Performance of Model: Testing Cases)

% of Agreement	Left Hemisphere				Right Hemisphere				
	LF	LP	LT	LO	RF	RP	RT	RO	OVERALL
Model vs. DR	81	81	88	80	75	83	81	87	82
Model vs. RK	76	77	76	75	77	85	84	83	80

Table 4.5: Results of Test 4.1 - Random Selection of Cases
(Performance of Discriminant Functions: Training Cases)
(including cases with missing data)

% of Agreement	Left Hemisphere				Right Hemisphere				
	LF	LP	LT	LO	RF	RP	RT	RO	OVERALL
Fnct vs. DR	77	87	89	77	85	85	81	79	83
Fnct vs. RK	87	85	80	76	85	88	85	84	84

Table 4.6: Results of Test 4.1 - Random Selection of Cases
(Performance of Model: Training Cases)
(including cases with missing data)

% of Agreement	Left Hemisphere				Right Hemisphere				
	LF	LP	LT	LO	RF	RP	RT	RO	OVERALL
Model vs. DR	89	91	85	89	85	91	91	91	89
Model vs. RK	83	85	87	81	88	93	93	89	87

Table 4.7: Results of Test 4.1 - Random Selection of Cases
(Performance of Discriminant Functions: Testing Cases)

% of Agreement	Left Hemisphere				Right Hemisphere				
	LF	LP	LT	LO	RF	RP	RT	RO	OVERALL
Fnct vs. DR	48	55	65	65	63	53	55	56	58
Fnct vs. RK	61	49	59	55	57	63	59	56	57

Table 4.8: Results of Test 4.1 - Random Selection of Cases
(Performance of Model: Testing Cases)

% of Agreement	Left Hemisphere				Right Hemisphere				
	LF	LP	LT	LO	RF	RP	RT	RC	OVERALL
Model vs. DR	75	73	71	80	83	83	81	87	79
Model vs. RK	73	77	73	76	73	83	76	75	76

Appendix 5: Results on Comparison with Neural Networks

Table 5.1: Results of Test 5.1
(Performance of Neural Network: - 93 Hidden Units)
(Training Cases Only)

% of Agreement	Left Hemisphere				Right Hemisphere				
	LF	LP	LT	LO	RF	RP	RT	RO	OVERALL
Fnct vs. DR									
10 cycles	83	43	91	79	79	93	93	91	82
20 cycles	63	75	95	79	87	93	91	97	85
40 cycles	100	93	100	100	100	100	100	99	99
80 cycles	100	100	100	100	100	100	100	99	100
120 cycles	100	100	100	100	100	100	100	99	100
Fnct vs. RK									
10 cycles	69	63	60	71	84	92	79	92	76
20 cycles	52	63	60	95	91	83	93	91	78
40 cycles	100	100	99	96	100	99	93	97	98
80 cycles	100	100	100	97	100	99	96	97	99
120 cycles	100	100	100	97	100	100	97	99	99

Table 5.2: Results of Test 5.1
(Performance of Neural Network: - 93 Hidden Units)
(New Cases Only)

% of Agreement	Left Hemisphere				Right Hemisphere				
	LF	LP	LT	LO	RF	RP	RT	RO	OVERALL
Fnct vs. DR									
10 cycles	71	52	75	76	80	84	83	83	75
20 cycles	49	71	81	76	79	81	85	85	75
40 cycles	79	81	85	83	80	81	80	80	81
80 cycles	76	80	87	87	77	81	81	81	82
120 cycles	77	80	85	87	77	81	81	81	81
Fnct vs. RK									
10 cycles	53	63	64	64	72	80	67	71	67
20 cycles	59	63	64	78	72	71	79	79	71
40 cycles	67	76	72	69	73	87	76	84	76
80 cycles	65	77	75	71	72	80	76	81	75
120 cycles	68	77	72	73	73	83	79	79	75

Table 5.3: Results of Test 5.2
(Performance of Neural Network: - 186 Units)
(Trained by DR's Decisions)

% of Agreement	Left Hemisphere				Right Hemisphere				
	LF	LP	LT	LO	RF	RP	RT	RO	OVERALL
Training Cases									
40 cycles	55	60	45	88	43	53	45	97	61
60 cycles	55	60	45	88	43	53	45	97	61
Testing Cases									
40 cycles	57	60	45	88	43	53	45	84	58
60 cycles	57	52	41	79	43	55	48	84	57

Table 5.4: Results of Test 5.2
(Performance of Neural Network: - 45 Units)
(Trained by DR's Decisions)

% of Agreement	Left Hemisphere				Right Hemisphere				
	LF	LP	LT	LO	RF	RP	RT	RO	OVERALL
Training Cases									
40 cycles	95	87	92	95	87	94	96	99	93
60 cycles	96	88	99	97	100	96	97	100	97
Testing Cases									
40 cycles	71	73	76	81	79	88	84	87	80
60 cycles	71	68	80	79	79	89	85	84	79

Appendix 6: Test of Model Trained by Both Experts

Purpose: The purpose of the following test was to examine the model's performance when trained by both human experts.

Method: The method was similar to test 2.1 in Section 7.3.2. The model was trained using all 150 cases and tested using the same 150 cases. The frequency arrays were held constant during the testing phase. In the training phase, the frequency arrays were updated first according to DR's decisions and then updated again according to RK's decisions. This was equivalent to presenting each case twice, the first time classified according to DR's decision and the second time according to RK's decision.

Results: The degree of agreement between the model and DR's decisions decreased from 86% in test 2.1 to 84% in this test (see Table 6.1). This decrease, however, was compensated by an increase in agreement with RK's decisions. The agreement with RK's decisions had increased from 78% to 80%. On average, the model still achieved the 82% average agreement with the two experts.

Table 6.1: Results of a Model Trained by Both Experts
(% of Agreement Between Model and Human Experts)

% of Agreement	Left Hemisphere				Right Hemisphere				
	LF	LP	LT	LO	RF	RP	RT	RO	OVERALL
Model vs. DR	83	83	82	80	85	87	85	88	84
Model vs. RK	81	81	75	80	77	86	81	82	78

Analysis: Generally speaking, a decision agreed on by more people is more reliable than a decision agreed on by a few. The quality of a decision, therefore, may be evaluated by the degree of consensus it has with others. In this test, the model's decisions could agree with one or both experts. An agreement with one expert was referred to as a 2-way agreement. An agreement with both experts was referred to as a 3-way agreement. Logically, a 3-way agreement is more preferable than a 2-way agreement. It is reasonable to expect that a model trained by both experts might result in more 3-way agreements. This expectation however was not supported by the results of the present study (see Table 6.2). In test 2.1, 72.75% of the model's decisions have a 3-way agreement. In the current test, 73.08% of the decisions have a 3-way agreement. The difference might be too small to be significant.

Table 6.2: % of Decisions That were 3-Way Agreements

	Left Hemisphere				Right Hemisphere				
	LF	LP	LT	LO	RF	RP	RT	RO	OVERALL
Test 2.1	74.67	72.67	68.67	70.67	71.33	76.67	73.33	74.00	72.75
Present Test	74.00	73.33	68.00	70.00	72.67	78.00	73.33	75.33	73.08

The model made its decision by choosing the category with the highest similarity score. If both categories have the same score, the decision would be difficult to make. Conversely, if there was a big difference in the scores, the decision could be considered as easy. It is therefore reasonable to assume that the differential of similarity scores between categories indicated how difficult a decision was. It is also reasonable to suspect that cases having an identical decision by both experts might be the easier ones. In the following analysis, a new variable, referred to as case type, was created by assigning the value of 1 to cases agreed by both experts and the value of 0 to the others. The correlation between case type and the differential of similarity scores were calculated (see Table 6.3).

Table 6.3: Results of a Model Trained by Both Experts
(Pearson Correlation between Case Type & the difference in Similarity Score)

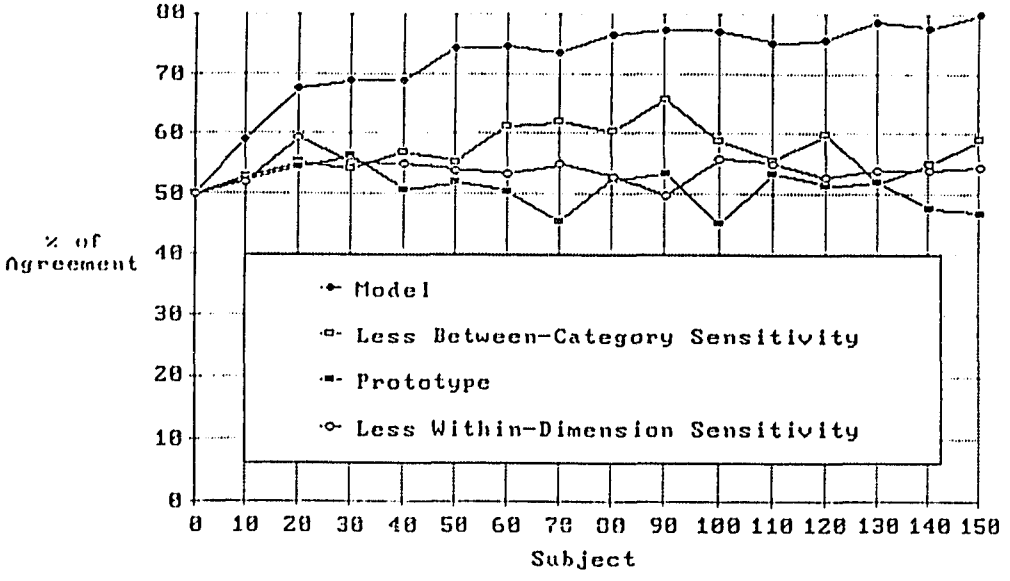
	Left Hemisphere				Right Hemisphere			
	LF	LP	LT	LO	RF	RP	RT	RO
Correlation	-.0165	-.1118	-.0917	+.2429	+.2406	+.2933	+.3455	+.3975
Significance Level	>.01	>.01	>.01	.01	.01	.001	.001	.001

As indicated in the above table, the correlations were not very strong. In some decision areas, however, the correlations were statistically significant.

Conclusion: The above results indicated that training the model using the decisions of both experts would neither improve the model's overall performance nor increase the number of 3-way agreements. Furthermore, it was found that those decisions agreed by both experts were not always easier decisions for the model to make.

Appendix 7: Learning Curve of the Model

The learning curve of the model is shown in the diagram below. As shown in this diagram, it is essential for the model to be sensitive to both within-dimension and between category variation in frequencies.



Appendix 8: An Example to Demonstrate the Model

The following example shows how the model can be applied to solve a simple classification problem. In this example, the operating procedures of the model are simplified to make them easier to understand. The classification problem involves two categories: A and B. The items in these categories are shown as follows:

Category A	Category B
1 1 1	0 0 0
1 1 0	1 0 0
1 0 1	0 1 0
	0 0 1
	0 1 1

Category A contains items that have at least two "1"s with a "1" in the first dimension, and category B has all the other items. Let us assume that 5 of the above items have already been classified. The following table shows the distribution of these 5 items in the two categories:

Category A	Category B
1 1 1	0 0 0
1 1 0	1 0 0
	0 1 0

The following table shows the internal representations of these 5 items, the resulting feature frequencies, relative frequencies and Baysian probabilities:

	Category A						Category B					
Dimensions	1		2		3		1		2		3	
Values	0	1	0	1	0	1	0	1	0	1	0	1
(1 1 1)	0	1	0	1	0	1						
(0 0 0)							1	0	1	0	1	0
(1 1 0)	0	1	0	1	1	0						
(1 0 0)							0	1	1	0	1	0
(0 1 0)							1	0	0	1	1	0
Frequencies	0	2	0	2	1	1	2	1	2	1	3	0
Rel. Freq .	0	1	0	1	.5	.5	.67	.33	.67	.33	1	0
Bayes Prob.	0	.75	0	.75	.33	1	1	.25	1	.25	.67	0

In the above table, the relative frequencies are calculated by dividing each feature frequency by the number of items in the category. This manipulation accounts for the within-dimension variation in frequencies. The Baysian probabilities are derived by (1) adding the relative frequencies across categories, and (2) dividing each relative frequency by its respective summed relative frequencies across categories. This manipulation accounts for the between-category variation in frequencies.

Let us suppose that an item X, which is shown as (1 0 1), is presented for classification. The similarity between item X

and category A is calculated by adding the Baysian probabilities in category A for the features shown in X.

$$S(X,A) = .75 + 0 + 1 = 1.75$$

Using the same procedure, the similarity between item X and category B is calculated as follows:

$$S(X,B) = .25 + 1 + 0 = 1.25$$

Since S(X,A) is greater than S(X,B), the model's decision is to put item X in category A. In this case, the feedback would indicate that the classification is correct.

With (1 0 1) placed in category A, the above table is revised as follows:

	Category A						Category B					
Dimensions	1		2		3		1		2		3	
Values	0	1	0	1	0	1	0	1	0	1	0	1
(1 1 1)	0	1	0	1	0	1						
(0 0 0)							1	0	1	0	1	0
(1 1 0)	0	1	0	1	1	0						
(1 0 0)							0	1	1	0	1	0
(0 1 0)							1	0	0	1	1	0
(1 0 1)	0	1	1	0	0	1						
Frequencies	0	3	1	2	1	2	2	1	2	1	3	0
Rel. Freq .	0	1	.33	.67	.33	.67	.67	.33	.67	.33	1	0
Bayes Prob.	0	.75	.33	.67	.25	1	1	.25	.67	.33	.75	0

If an item Y, represented by (0 0 1), is presented for classification, the similarity scores would be:

$$S(Y,A) = 0 + .33 + 1 = 1.33$$

$$S(Y,B) = 1 + .67 + 0 = 1.67$$

In this case, the model would classify Y as a member of B.