THE UNIVERSITY OF CALGARY

Geological Mapping from Multi-source Data using Neural Networks

by

Grace Yen-Chew Yang

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF GEOMATICS ENGINEERING

CALGARY, ALBERTA

SEPTEMBER, 1995

THE UNIVERSITY OF CALGARY

FACULTY OF GRADUATE STUDIES

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies for acceptance, a thesis entitled "Geological Mapping from Multi-source Data using Neural Networks" submitted by Grace Yang in partial fulfillment of the requirements for the degree of Master of Science.

Supervisor, Dr. M.J. Collins
Department of Geomatics Engineering

CO-Supervisor, Dr. P. Gong
Department of Geomatics Engineering

Dr. M. A. Chapman
Department of Geomatics Engineering

Dr. R. R. Stewart
Department of Geology and Geophysics

Date: September 12, 1995

ABSTRACT

This study explores and analyzes the detailed operations of lithologic classification from remote sensing images and geophysical data using feedforward neural networks. A set of experiments was designed and performed to test the dependence of classification accuracy on various parameters. The variables used in the experiments are various combinations of input channels, the number of output classes, the number of hidden nodes, the training sample sizes, and the training coefficients (i.e. the momentum factor and the learning rate). The input channels consist of different types of images generated from gravity, magnetic, gamma ray spectrometry data and remote sensing images such as Thematic Mapper, radar and SPOT.

Through the analysis of classification accuracy with increased number of iterations, we demonstrated that the optimal choice of input channels is the most critical factor in achieving better accuracy result. The classification accuracy may be maximized by choosing an optimal combination of input data layers. When training the network, the size of individual training samples is more important than the total number of training samples in obtaining a satisfactory classification. The classification accuracy is inversely proportional to the number of output classes in this geological mapping. Generally speaking, the overall average accuracy of classification gets better by increasing the number of iterations to a certain degree, however, at the expense of some individual classification accuracy.

The variance in the individual classification accuracy were found to be significant which has lead to some criterion on the selection of the parameters. For lithologic mapping, the network should be structured in accordance with the importance of each individual class.

ACKNOWLEDGMENTS

Both my supervisors, Dr. Michael Collins and Dr. Peng Gong have helped me at different stages of my program. Dr. Peng Gong of the Department of Environmental Science, Policy, and Management, University of California at Berkeley, Berkeley, California, provided the software and the computing environment. Without them this series of experiments would not be possible. Dr. Michael Collins has provided guidance, encouragement and discussions that have stimulated my scientific thinking and interest in research. Dr. Collins' critical comments have made writing a challenging and enjoyable experience. The data were provided by Dr. C.F. Chung from the Geological Survey of Canada. Mr. Ruiliang Pu of the Department of Environmental Science, Policy, and Management, University of California at Berkeley, Berkeley, California, provided needed assistance and support during the most difficult stage of this study. Joseph, my husband, was always there with his support and unshakable faith in my intellectual capacity. Pearl, my daughter, is always sweet, lovely and diverting when I was in the pursuit of answers.

This thesis is dedicated to the memory of my father.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

This chapter introduces the framework of this study which has three distinct segments: the lithologic mapping, the multi-source data, and the back-propagation neural network. It begins by reviewing the relevant aspects of remote sensing's application in geology for reconnaissance lithologic mapping. Second, it presents the up-to-date developments in the integration of multi-source data, the nature of multi-source data, its benefit and difficulty of integration of multi-source data. Third, it introduces the neural network technology and its application in classification. Fourth, it lays out the hypotheses of this study. It is our intention to better understand the functioning of the neural network through analysis of the performance of the classification process. The final section of Chapter 1 summarizes the review.

1.1 Introduction to Geological Remote Sensing

For decades, geologists have successfully used remote sensing techniques to solve a wide variety of specific geologic problems that are difficult to solve by conventional methods alone, including mineral and energy resource exploration, nuclear siting and waste disposal, and the charting of glacier and shallow seas (Goetz, 1981).

Geological remote sensing can be defined as the study of the earth using electromagnetic radiation (EMR) with wavelengths ranging from ultraviolet (0.3 micrometer) to microwave (3 meters). In contrast, geophysical remote sensing can be

defined as the study of the earth using electromagnetic radiation of wavelengths shorter than ultraviolet (X ray, gamma rays, etc.) and longer than microwave (radio) (LeRoy, et al, 1977) together with other methods which do not detect electromagnetic radiation, such as magnetic, gravity, sonic and seismic techniques.

The most fundamental application of remote sensing in geology is in the augmentation of conventional methods for compiling and interpreting geologic maps of large regions. Regional geologic maps are important because they present compositional, structural, and chronological information essential for reconstructing the geologic evolution.

Geological maps, usually based on ground traverses, are categorical abstractions of the large amount of data available on the ground. They are limited by rock exposure, accessibility, and manpower resources. However, remote sensing techniques provide certain structural and lithologic information more efficiently than can be acquired on the ground.

Geologists are interested in mapping the regional linear features on images. These linear features, called lineaments, appear as linear or curvilinear geomorphic features on imagery. Lineaments are often associated with geological structures (faults, fractures or shear zones) or lithologic contacts. Lineaments are particularly important in mineral resource studies, because many, though not all, ore deposits are localized along fracture zones.

In areas where no geologic map is available but with good rock exposure, geologic reconnaissance maps can be compiled from Landsat images because many of the major structural and lithologic elements are well displayed. Remote sensing data also contribute to a better understanding of areas where considerable detailed geologic mapping is already

available.

1.1.1 Remote sensing and lithology classification

The spectral reflectance of earth materials is often the most useful and diagnostic criterion for lithologic discrimination. The process of "reflection" occurs within one-half wavelength of a material's surface, in the molecular structure of the material, and results in the instantaneous radiation of EMR. Spectral reflectance is a measure of the distribution of electromagnetric energy reflected by a material, and is expressed in images by photographic tone or color. Reflectance is a consequence of the chemistry and structure of the material modified by environmental factors and the physical condition of the material.

Remote sensing data represents the reflectance from the upper few micrometers or millimeters of the surface because of the high opacity and scattering characteristics of natural materials. The opacity in the visible and infrared portions of the spectrum is created by high absorption coefficients due to a variety of electronic and vibrational processes (Hunt, 1977). The principal constituents of igneous rock-forming minerals and, hence, all rocks have neither electronic nor vibrational transitions. The spectral information that appears as bands or wings of bands is due to the presence of other minor components that are present as impurities (Siegal, 1980).

In many areas of the world, vegetation cover obscures rock and soil materials, and remote sensing techniques must use variations in vegetation patterns as indicators of the underlying geologic materials. In some areas, plant species are very selective in associating with particular rock material, and geological mapping can be almost completed without actually observing the rock material themselves (Taranik, 1983). When rock types are not particularly well exposed, or when outcrops do not produce characteristic weathering patterns remote sensing technique alone can not predict the lithology of the subsurface

material.

The property related to the changes in surface temperature that are induced by diurnal solar heating may reveal some information concerning body properties as opposed to surface properties. This property, called thermal inertia, is defined as "k ρ c", where k is the thermal conductivity, ρ is the density, and c is the specific heat, can be analyzed to allow measurement up to a depth of about 10 cm or less (Gillespie, et al. , 1977).

The mid-infrared region beyond 8 μm is especially important for geologic mapping because spectral emittance variations provide a basis for distinguishing between silicate and non silicate rocks and for discriminating among silicate rocks. Near 10 μm, the manifestation of the fundamental S_i-O stretching vibration are diagnostic of the major types of silicates (Lyon, 1965). These two techniques of mid-infrared and thermal inertia can be used together to provide considerable fundamental lithologic information.

The majority of discernible features in the spectra of igneous rocks occurs as a result of the presence of iron, its oxidation state, and water. The same is true for sedimentary and metamorphic rocks, with the exception of carbonates, which display strong absorptions caused by vibrational processes due to the CO_3^{-2} ion and Al-O-H deformation in clay materials (Siegal, et al, 1980).

The electronic transitions in the visible and near infrared that occur in the transition elements, and the charge transfer bands in the ultraviolet, influence the spectral reflectance of minerals. It produces a distinguished spectral feature in the visible toward the ultraviolet and an absorption band between 0.85 and 0.92 μm associated with the Fe^{3+} electronic transition. These spectral features are characteristic of iron oxides and hydrous iron oxides, collectively referred to as limonite (Siegal, et al, 1980). The limonite in the altered zones is an important surface indicator for precious and base metals exploration.

Lithology encompasses several different factors such as mineralogy, grain size and degree of lithification. Lithology of a formation means that a certain rock type is present in a stratigraphic sub-division (Lorenzetti, 1992). Lithological classification maps are useful for geological mapping which are often used together with geophysical data for defining subsurface configurations. The subsurface configurations are important for all practical applications, because the conditions of formation of rock units and structural features influence the occurrence of ore and petroleum deposits. The spectral radiance, the spatial distribution of landforms and the brightness (average reflectance in the visible and near-infrared regions) on the imageries are all diagnostic properties of rocks.

The technique used for structural analysis is based on integration of Landsat data with other geological data sets such as gravity, magnetic, subsurface, and production data. Geologic mapping with the aid of remote sensing data entails the description of structure, lithologic units and geobotanical relationships.

1.2 The Nature of Multi-source Data

The rapid development of data acquisition technology has resulted in a large volume and diversity of modern multi-source spatial data. The integrated analysis of spatial data from multi-sources has become increasingly important in the communities of remote sensing and Geographical Information System (GIS), particularly in the development of geomatics. Geomatics deals with a wide range of data sets such as remote sensing images, geologic or topographic maps, gravity or magnetic maps, Global Positioning Systems (GPS), and various thematic maps, etc.

Spatial data are qualitative or quantitative observations of phenomena in space. In a spatial database, the element of spatial information can be defined as $T=\{x, y, z, h_1, h_2,..., h_n\}$, giving the values of a set of n spatial variables at location (x, y, z). The variable

h_i can be of any data type: nominal, binary, ordinal, interval or ratio. For example, in a geologic lithology map, h is in nominal type (e.g., rock type); for an image, each pixel records the spectral reflectance (e.g., h is a numerical value). Integration of these two different data types could be difficult, because there is a problem to find a mathematical model f (value, rock type) which combines a spectral value (e.g., '20') and a rock type (e.g. 'A) . The observations provide a set of data, which constitute an image of the phenomenon. The analysis of this image, and its synthesis, constructs our understanding of the phenomenon, transforming the data into information.

Different data sets acquired by different methods are recorded at different spatial scales. For example, geophysical data, such as gravity or magnetic, are usually point measurements or along flight lines for airborne geophysical surveys are measured in nominal scale (i.e. values in certain units). However, remote sensing imagery obtained by different sensors operating in different spectral regions represent the radiance of surface targets in a ratio scale. Both radar and TM data represent high-frequency surface information. The low-frequency information are provided by the geophysical data such as gravity and magnetic.

Each data set has its own possibilities and limitations in different applications. Different data sets also have different accuracy and resolution, or they differ in quality and in the degree of detail of the properties they measures. Data density (e.g. the number of measurements per unit area; or the average distance between the measurements), data coverage (e.g. the area where data is available), and data distribution (e.g. in terms of its regularity, i.e. is the data evenly distributed or not) also varied since detailed survey only covers some limited areas and whereas in areas with less interest, the data density is usually low.

1.2.1 The benefits of integration of multi-source data

The integration of multi-source data of the same area or object offers the potential for great improvement in all forms of image exploitation and especially in improvement of classification accuracy because by integration of multi-source data, one can make optimum use of the unique characteristics of multi-sources and backgrounds. Integration or incorporation of multi-source data can supply supporting identification criteria from different data sources to classify the pixels with mixed spectral/ physical features.

Classification is the following decision problem: given an input vector x decide in which of several known classes the input x belongs. The classes are assumed to be mutually exclusive and exhaustive. Useful characterizations of the classes are assumed to be either unknown or unavailable and must be estimated from a given collection of labeled training samples (i.e. input vector corresponding to each class). The absence of a priori class characterizations is the major difficulty in classification. Classification can be defined as a generalization or abstraction of a certain phenomena in order to improve our understanding (Gong, 1994). Classification reduces the level of complexity in spatial data and eases the process of decision making.

Data integration is a process for making different data sets compatible with each other. Data incompatibility can include the use of different geographical referencing systems, different spatial or temporal coverage, different scale, and different degrees of generalization. Integration of multi-source data may not only help to determine the quantitative relationships between the various data sets, but also complement the existing data sets, reduce the redundancy in order to extract the needed information and to achieve the ultimate goal of management, planning, and policy making.

1.2.2 The difficulties of integrating multi-source data

Many geological and geophysical data sets often have only partial coverage because it is very expensive to acquire geophysical or geological data in all the areas by geological field work or by using conventional geophysical surveys. In almost all cases, the data have very different spatial resolution and they are often incomplete and have an unbalanced spatial distribution.

Whenever multiple data sets are available for a region, data compatibility is often a problem because the multi-sensor data may have been initially acquired with different survey objectives and, thus, may have different temporal and spatial coverage. The flightline orientation, length of flightlines, and flightline spacing generally tend to be quite different. In addition, due to the great variety in data collection, processing, and representation, it results in differences in the reliability, uncertainty and completeness. These cause serious difficulties in integration (Gong, 1994).

1.2.3 Techniques for combining multi-source data

In the past, integration of multi-source data was considered as an overlap of multi-layer maps (Burrough, 1986). The concern was in data handling, and it was more geometric and topographic in nature than a true integration of multiple data types. More recent development in the integration of multi-source data has been directed toward the analysis and interpretation of multi-source data for the purpose of management decision making.

There are many approaches to deal with data integration. The most commonly used approach is the quantitative approach. The quantitative approaches include the rule-based expert system and several types of statistical and mathematical techniques such as Bayesian theory (based on statistical and subjective probabilities), Dempster-Shaefer theory of

evidence reasoning, and fuzzy set theory (based on fuzzy logic), etc. Evidential theory includes the Bayesian theory as a special case. The theory of evidence is a theory of probable reasoning because it deals with weights of evidence, and with numerical degrees of support based on evidence.

The partial belief function approach is examined by Moon (1990) as a means to integrate both geophysical and geological data sets for deposits of iron ore and base metal (such as gold and phyrhotite). Theoretically, the Dempster-Shafer methods appears to be a suitable method, but in practice the partial belief function is area (or site) dependent. When using evidential theory in data integration, one must first be able to reason over possibilities and also about the interrelation between the several sets of available data.

One problem in the use of rule-based expert systems is the knowledge acquisition and representation. Most human experts have difficulties in describing their knowledge explicitly and completely. Some human knowledge is inexpressible in terms of rules and sometimes may not be understandable even though it can be expressed in the if-then type of production rules (Hoffman, 1987).

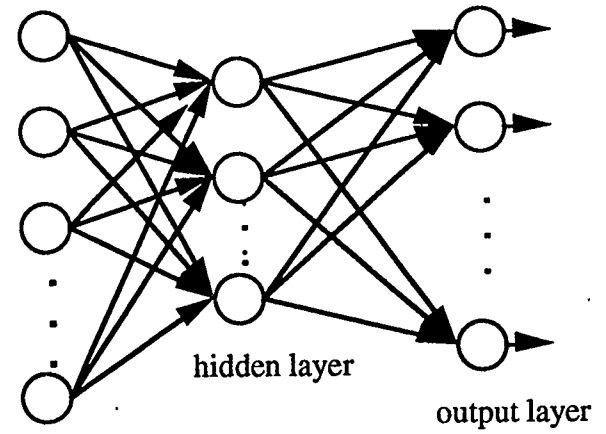
All these quantitative techniques are highly selective in the type of information they can handle. They usually assume the data are either in a discrete (nominal or ordinal) or a continuous (interval or ratio) format and thematic classification data can not be used. If the data are incomplete or noisy, it is difficult to apply these quantitative techniques. Furthermore, some techniques make certain statistical assumptions, such as Gaussian distribution.

The other trend of integrated analysis of heterogeneous spatial data is the development of neural computing technology which offers an alternative to the traditional statistical methods and rule-based expert systems. Neural networks are receiving a great

deal of attention from various fields. Due to its capability of automated learning from empirical data with little or no a priori information about the application and its tolerance with noisy or incomplete data, neural computing technology has gained enormous interdisciplinary popularity in recent years.

1.3 Neural networks

Neural networks were originally developed as computational systems that attempt to simulate the functionality and decision making process of the human brain. These computational systems, implemented in either hardware or software are made up of neurons, also called nodes or neurodes, which are inter-connected by weighted links or synapses. A neurode is a simple computation unit which maintains only one piece of dynamic information (its current level of activation).



the circle represents the neurode

Fig. 1.1 The structure of a feedforword back-propagation neural network

A neural network is similar to a computer in that both are general purpose systems that can be programmed to perform specific tasks. Programming or training a network to perform a specific task requires a training algorithm. Unlike traditional computer programs

incorporating algorithms to solve a particular problem, neural networks utilize a learning technique to develop an appropriate solution.

The neural network is "trained" by repeatedly presenting examples of the inputs and desired outputs of the problem to be solved. As each example is entered into the network, the difference between the actual output of the network and the desired output is used to modify the weights for each interconnection. Training of the neural network (or, equivalently, changing the values of the interconnection weights) continues until the actual output of all the training examples matches the desired outputs to within some specified tolerance. When this is achieved, the neural network is said to be "trained" and is ready to accept new inputs to predict the outputs.

A neural network performs "computations" by propagating changes in activation (i.e. level of stimulation) between the processors. The propagation of activation, and thus the nature of the "computation" performed by the network, is strongly affected by the weights (or strengths) of the numerous connections (or synapses) between the processors. Hence, the "program" or definition of the computation, is embodied within :

- the topology of the network (number of layers, nodes and how the nodes of each layers are connected);

- the connection strengths (or weights) of the network; and

- the mechanism of activation propagation, i.e. the algorithm for computing the activation value of a neurode as a function of its net input.

There are two main stages in the operation of an entire network: learning and recalling. Learning (training) is the process of adapting or modifying the connection weights in response to stimuli being presented at the input layer and optionally the output

layer. How do we choose the connection weights so that the network can do a specific task? We will encounter some examples where we can choose the weights *a priori* if we are a little clever. This embeds some information into the network by design. But such problems are the exception rather than the rule. In other cases we can often "teach" the network to perform the desired computation by iterative adjustments of the strengths of the weights. This may be done in two main ways:

- **Supervised learning.** The learning is done on the basis of direct comparison of the output of the network with known correct answers (or target patterns). This is sometimes called learning with a teacher. The network compares its output to the target and adapts itself according to the learning rules, that is the network is taught to classify input into one of several *a priori* categories.

- **Unsupervised learning.** A learning procedure in which the network is presented a set of input patterns. The network adapts itself according to the statistical associations in the input patterns. The only available information is in the correlation of the input data or signals. For example, Hebbian learning, self-organization or competitive learning are three of many unsupervised learning schemes.

Instead of having to specify every detail of a calculation (like the programmed instruction sequence), we simply have to compile a training set of representative examples. This means that we can hope to treat problems where appropriate rules are very hard to know in advance, as in expert systems and robotics.

Recalling refers to how the network processes a stimulus presented at its input layer and gives answers based on what it has learned during the training stage.

Rumelhart, et al., (Chs. 1-3 in PDP, 1986) provided an excellent description of the basic anatomy of all neural networks which they divide into seven basic aspects:

(1) a set of processing units.

(2) the state of activation of a processing unit.

(3) the function used to compute output of a processing unit.

(4) the pattern of connectivity among the processing unit.

(5) the rule of propagation employed.

(6) the activation function employed.

(7) the rule of learning employed.

These seven basic aspects will be described in more detail in chapter three.

1.3.1 Back-propagation neural network

The network topology (i.e. the number of layers, the number of nodes and their inter-connectivity) and the rules of learning and propagation and functions of output and activation are all variables in a neural network and lead to a wide variety of network types. There are basically two broad classes of networks:

1. Feed forward networks, in which the network nodes in the network are grouped into layers and communication is restricted to occur only between layers and in a forward direction, no lateral, self- or back-connections are allowed. Examples are: the Perceptron, the ADALINE (ADAptive LINear Element) and MADALINE networks, feed-forward back-propagation, the Boltzmann machine, and the Cachy Machine.

2. Other networks in which the links can form arbitrary topologies - any network, if feedbackward or recurrent (self- connections) are involved. Examples are: Adaptive Resonance Theory (ART) networks, Bi-directional Associative Memory Networks (BAM),

recurrent back-propagation network, Hopfield network, etc. From this basic binary classification, other subclasses arise due to variances in the activation functions and learning rules used.

Considerable research has been done in the field of multilayer neural networks with respect to feedforward network abilities. The Perceptron and the ADALINE and MADALINE networks are of substantial historical interest, and have paved the way for the development of other neural networks. The feedforward back-propagation neural network (BNN) is probably the most widely used neural networks today. It was developed by Werbos in 1974, rediscovered by Parker in 1982 (Parker, 1985), and, again, rediscovered by Rumelhart et al. in 1986 (Rumelhart et al., 1986).

The standard (i.e. the feedforward) back-propagation neural network uses a Generalized Delta Rule (GDR) with a sigmoid function as its activation function. Back propagation is a learning rule for multi-layer feed forward networks in which weights are adjusted by backward propagation of the error signal from outputs to inputs. The distinguishing features of feedforward back-propagation networks are:

- multi-layered - the network has a minimum of one layer between the input layer and output layer, referred to as the hidden layer.

- fully-connected - in the restricted class of feed-forward networks under discussion, every unit (or input) feeds only and all the units in the next layer. There are no connections leading from a unit to units in previous layers, nor to other units in the same layer, nor to units more than one layer ahead.

- uses Generalized Delta Rule (GDR) for supervised learning or training - The learning algorithm is based on the minimization of the summed squared error (Least Mean Square or LMS) between the actual responses O_j and the associated desired responses d_j (the

delta) of the network over all training examples.

- uses a sigmoid function as its activation function (See Chapter 3 for a detailed explanation).

The training procedure of feedforward error back-propagation neural networks involves two stages. In the first stage, the presented training pattern generates a forward flow of activation from the input to the output layer. In the second stage, errors in the networks output generate a flow of information back to the input layer. It is this feature that gives the network its name. Back-propagation is actually a learning algorithm rather than a network design. It can be used in a variety of architectures.

In addition to the learning algorithms, there are several factors involved in neural network design. Network structure refers to the number of layers and the organization of the layers. The way the processing elements are connected; connections may use feedback or feed-forward; networks may have partial or full connectivity. Several networks can be cascaded together hierarchically or in some other fashion.

1.3.2 Application of BNN to geologic or lithology classification

The back-propagation neural network has many applications in geology. It has been applied to the detection of geological lineaments from TM imagery of the Canadian Shield by Parikh (1991) using edge images as input and digitized lineament maps as the desired output. In the study of Guo (1993), a BNN technique was successfully applied to lineament recognition and lithologic classification from aeromagnetic map. However, Guo clearly stated that it is difficult to classify lithology using only magnetic data because the magnetic field is affected only by magnetic geological bodies. In fact, lithologic classification from magnetic fields using BNN is actually a classification of magnetic anomalies which are related to different lithologies. In the work of Lorenzetti (1992), a

single hidden layer network was used to predict the lithology from P-wave and S-wave velocities (V_p and V_s) with 80 - 90 % accuracy. In his study, the network was trained using laboratory measurements of ultrasonic velocities on core samples. Inputs to the network are V_p, V_s, Poisson's ratio (σ), and depth (z).

With regard to the classification of multi-spectral remote sensing data, Bishchof et al. (1992) reported that an extension of the basic back-propagation network can incorporate textural information without explicit definition of a texture measure. Kanellopouloss et al. (1991, 1992) tested the possibility of discriminating a large number (20) of land cover classes using neural networks. The average accuracy of 84 % of their results outperformed traditional classifiers. Neural network also outperformed evidential reasoning in classification of alpine land cover and permafrost active layer depth (Peddle et al. 1994).

Previous studies on geological mapping from multi-source data using neural networks were carried out by An, et al. (1994) and Gong (in press). Gong used 13 input layers and one hidden layer to classify four lithologic units in the same area as used in this study, and compared the results to those generated by evidential reasoning. Neural network results were superior to the results from the evidential reasoning method. In the work of An et al. (1994), a total of sixteen input data layers were used with the exception of gravity data. Predictions of the same four lithologic units were examined with two training samples selected from two different base maps (bedrock map and the outcrop map).

1.4 Hypotheses

Geologic lithology mapping traditionally is compiled based on field survey data. It is time-consuming and requires a great deal of manpower. In many remote areas, the lithology maps have not been completed. Efforts have, therefore, been made to aid the geological field mapping in Arctic regions in Canada. Although a relatively high accuracy

of classification using neural network technique has been reported in some research work (An, 1994; Gong, 1995), the detailed operations of the neural network are not well understood in the classification application using multi-source data.

One phenomenon often noted in applying neural networks to classification is large variations in the classification accuracy of individual classes. Questions such as which parameter or parameters are primarily responsible for these differences in accuracy (for example: learning rate, momentum rate, the number of hidden units, the number of training iterations, network stability achieved, and the input data sets) remain unanswered in the application of neural networks to the classification of multi-source data (An, et al., 1994; Peddle et al. 1994). The goal of this study is to test the ability of a back-propagation neural network to predict the lithology from a multi-source data set and to try to find the answers to certain questions raised in An's and Peddle's paper. Also, given a fixed architecture of networks, the selection of optimal value of parameters are to be explored.

Our approach is to design experiments based on the hypotheses stated as follows:

Hypothesis 1 Classification accuracy may be maximized by choosing an optimal combination of input data layers.

An et al., (1994) considered the importance of a data layer to the lithology classification. An attempt was made to test the importance of a data layer to the classification. The method they used was to replace the value of one data layer with the mean value and then report the final results. In the application of neural networks on ecological land systems, Gong et al., (1994) also proposed a rationale for adding and dropping input variables in an effort to improve the convergence rate and reduce the mean square errors of the neural network.

This hypothesis states that a result of better accuracy after certain data sets are

included is an indication of higher contribution of those data sets and a lower accuracy indicates a negative contribution of those data layers. Conversely, if some data layers are important to the classification, the classification accuracy drops when those particular data sets were excluded from the classification. There would not be any significant changes in classification accuracy for data layers that have no importance to the classification with or without including them in the input data sets.

Hypothesis 2 The accuracy of classification is inversely proportional to the number of output classes, i.e. the classification accuracy decreases as the network tries to classify more classes.

Artificial neural networks have been widely used as a classifier in many complex classification tasks. In many other applications, such as handwritten zip code (le Cun, 1989), recognition of sonar targets (Gorman, 1988) or character recognition (Bebis, 1990). Their early experimental results show that as the number of classes involved in a classification task increases, the classification accuracy of these networks decreases. We anticipate that this will also be the case in lithologic classification.

This will be done by increasing the number of output classes in the classification.

Hypothesis 3 A larger number of training samples do not guarantee a better classification accuracy.

Training the network is critical to the classification. Training is often carried out with a training set consisting of a number of examples drawn from ground observations. The amount of data necessary for training a neural network has been explored by Zhuang et al. (1994); Heermann (1992); and Wann et al., (1990). Zhuang and Wann reported no significant improvement in classification with a higher percentage of training samples. However, Heermann suggested that the absolute size of the training set may be important.

Intuitively, we would think providing more examples to train the network meaning the network will have more knowledge to learn more about the relationship between the classes. Is there any relationship between classification accuracy and the size of the training samples? We will test this classification's dependency on training size in our study area with our data sets.

This will be done by selection of different training areas from an outcrop map.

Hypothesis 4 The optimal number of hidden nodes depends on the input data sets.

An important but difficult problem is to determine the optimal number of hidden units needed to provide us with the desired outputs. The difficulty is because an increase in the number of the hidden units lessens the output errors for the training examples, but increases the errors for novel examples. Murata et al., (1994) suggested using network information criterion to determine the number of hidden units. In our designed experiments we will test the dependency of the optimal number of hidden nodes on the input data sets.

Hypothesis 5 The dependence of classification accuracy on the training coefficients (i.e. learning rate and momentum factor) is less than the dependence on other parameters, such as the input data, the output classes, or the training sample sizes, etc.

A number of rules to select the training coefficients (learning rate and momentum factor) have been reported in different application (Kung & Hwang, 1988; Higashino et al., 1990). However, these rules have not been widely used. The selection of these coefficients has frequently been empirical (Rumelhart & McClelland, 1986; Watrous, 1986). Are these parameters equally important in producing higher classification accuracy ? Through our designed experiments we will gain more insight into it.

1.5 Summary

This chapter has briefly reviewed the integration of remote sensing and geophysical data in geological mapping using neural network and found a need for a more detailed study on the performance of the BNN.

The distribution of rocks or the lithologic information in any area may not be predicted from one single data set. The advantage of the neural network approach is that there is no prior knowledge about the statistical distributions of classes is needed.

Neural network is a form of automated pattern recognition in which a set of input patterns is related to an output by a transformation encoded in the network weights. They are particularly appropriate for applications in which the relationship is unknown (i.e., we know what the input is and we know what the output answer should be, but we cannot write the mathematical equations which will get us from the input to the output). Neural networks provide a practical and convenient pathway to a solution.

In a way, back-propagation networks operate similarly to linear regression in that a least-squares error criteria is used to determine the goodness of fit to the training data. Through the non-linear transfer function, back-propagation neural networks are able to determine complex non-linear mappings between a set of inputs and target outputs. It is this property which makes them especially well suited to the problem of determining lithology from various geophysical data and remote sensing images.

CHAPTER 2

STUDY AREA AND DATA SOURCES

This chapter provides a review of the geologic background of the study area, the sources and the characteristics of the data used in the study. It also summarizes the scientific knowledge of the geologic, geophysical, and the remote sensing technologies.

2.1 Study area

The study area covers approximate 13 km x 13 km in Melville Peninsula, N.W.T. in northeastern Canada, approximately between latitudes 68° 49' N and 68° 58' N and longitudes 82° 33' W and 82° 50' W (Fig. 2.1).

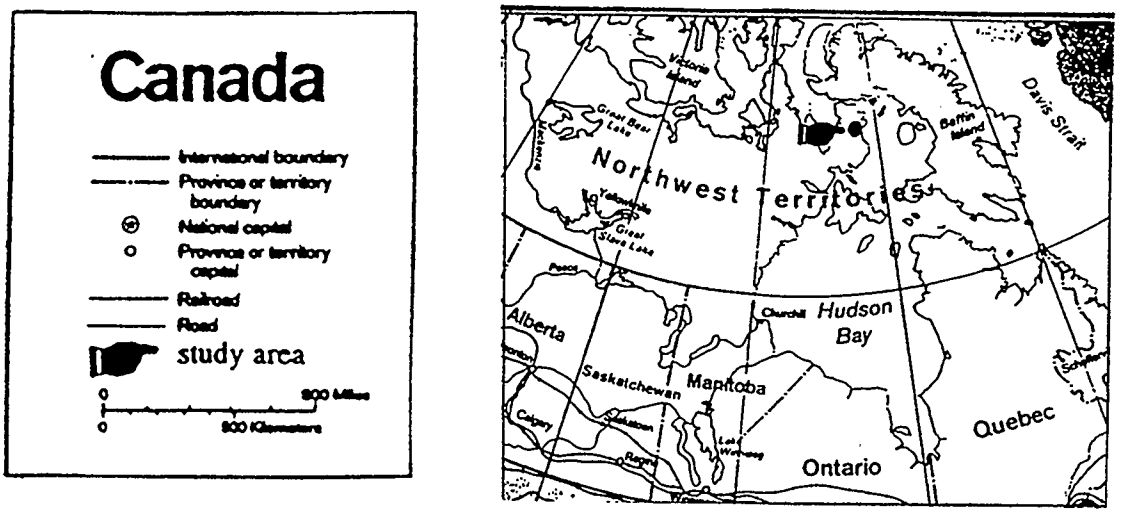


Fig. 2.1 Geographic location of the study area in Melville Peninsula, NorthWest Territories

This area was chosen for the study is because that it has little vegetation and relatively large areas of unweathered rock outcrop. It makes this area ideal for integration of remote sensing and geophysical data.

2.1.1 Geologic setting

The geology of Melville Peninsula consists of a highland with Precambrian basement gneiss rising above a lowland with flat lying Ordovician carbonate (Schau, 1993).

A number of lithologic units have been recognized in the area. The major lithologic units are:

class 1: Precambrian gneiss;

class 2: the Prince Albert Group (PAG)

class 3: the Hall Lake Plutonic

class 4: the Ordovician carbonate (Ship Point Formation).

Prince Alberta Group is a heterogeneous sequence containing many compositions such as acid and metaultramafic volcanogenic rocks, clastic metasediments, iron formation and intrusive sills and dikes. PAG is the main target to be identified because this formation associated with volcanogenic massive sulfide mineral deposits (An, et al. 1994). The plutonic complex is the most homogeneous - consisting of a single lithology - granodiorite, rich in potassic feldspars. In excess of 40% of calcium in the Ordovician carbonate is toxic to plants and inhibits the vegetation growth thus exposing the white carbonates directly (Schau, 1993).

2.2 Multi-source data

The data set used in the study is part of a Digital Data Atlas in the Hall Lake Area, Melville Peninsula, Northwest Territories (Schau et. al. 1993). The images contain 512 x 512 pixels. They were provided and processed by the Geological Survey of Canada (GSC) and recorded in a digital format. They include seven channels of Landsat Thematic Mapper (TM), one channel of synthetic aperture radar (SAR), one channel of SPOT and different types of images generated from gravity, magnetic, and gamma ray spectrometry. The geologic outcrop map (Fig. 2.2) was provided by GSC and used as base map for training the network.

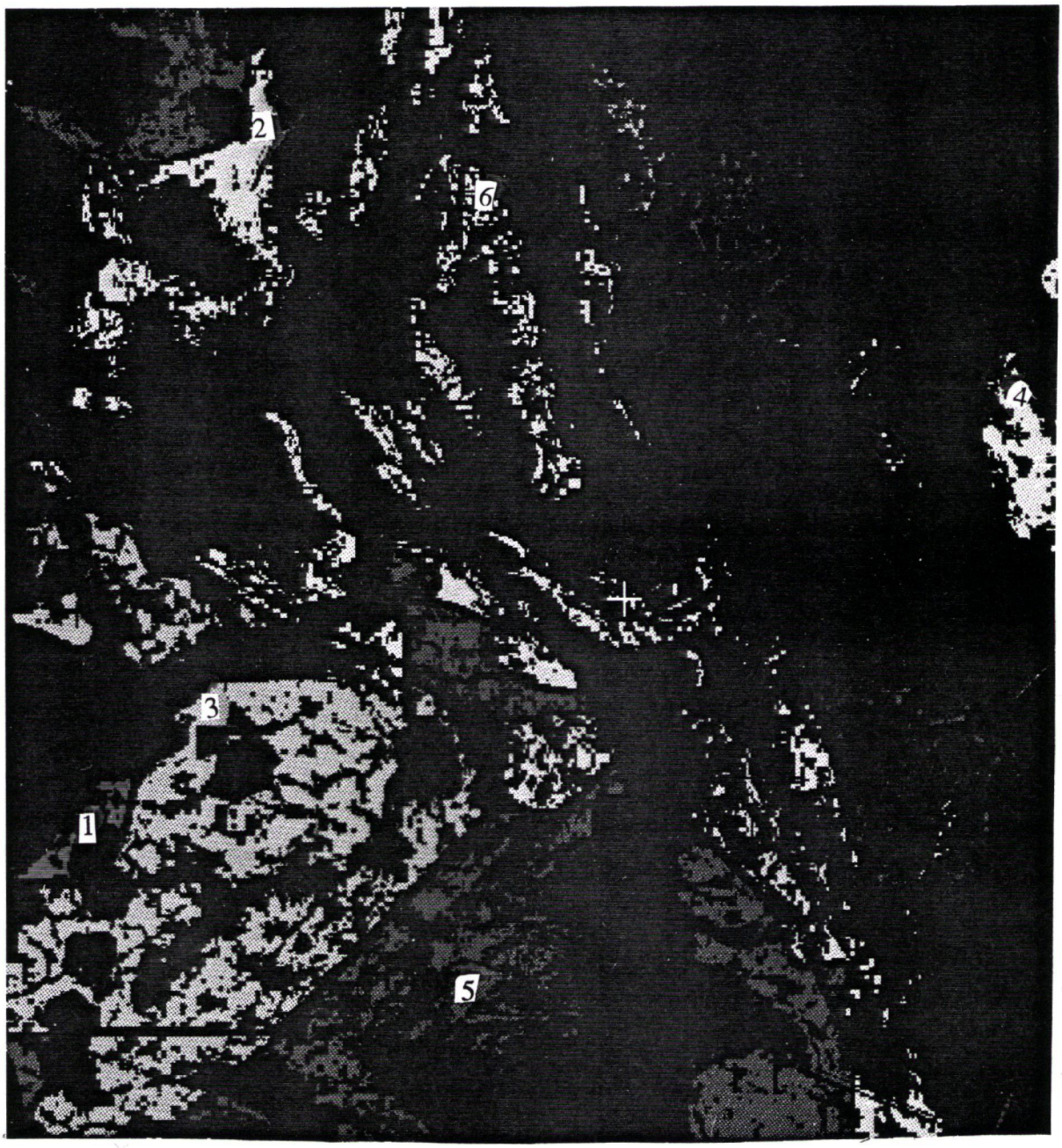
The outcrop map was generated from a surficial geology map which is a thematic map of surficial materials of marine, alluvial, glacial, outcrop and mixed origin, on which the outcrop was shown as one class. This distribution of surface materials were interpreted from aerial photographs and ground surveys.

TM data

Seven channels of Landsat TM image in the study area with 8 bit radiometric resolution were collected on July 14, 1988 and preprocessed and geometrically registered to Universal Transverse Mercator (UTM) coordinate base by Geological Survey of Canada (GSC). The TM image is characterized by northwest trending white streaks. TM 7 image is shown in Fig. 2.3. Table 2.1 is a summary of the source and basic characteristics of each data type available for this study.

Radar Data image

A C-band (5.66 cm wavelength) synthetic aperture radar (SAR) data, transmitted and received vertically (VV) polarized, were collected by the ERS-1 satellite on August 3,



unit 1 - Precambrian basement gneiss
unit 2 - volcanogenic sequence of Prince Albert Group
unit 3 - the grandodiorite of the Hall Lake plutonic
unit 4 - the Ship-Point formatin of Ordovician carbonate

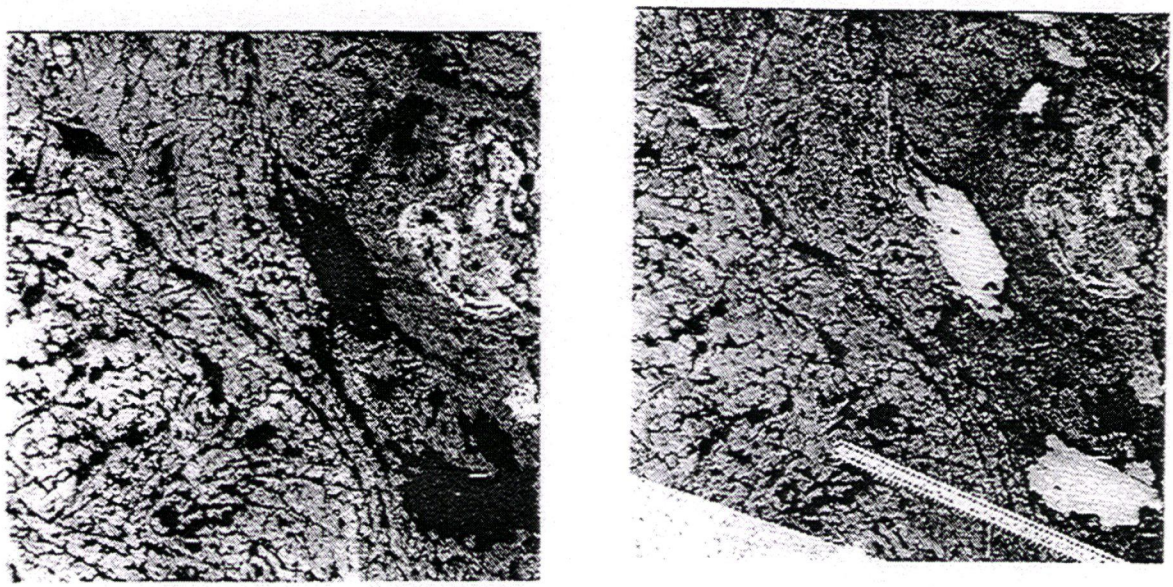
Fig. 2.2 Geological outcrop map of the study area

Data layers	Spectral resolution (μm)	Spatial resolution	resampled pixel size
Landsat TM 1	0.45 - 0.52	30 m x 30 m	
TM 2	0.52 - 0.60	"	
TM 3	0.63 - 0.69	"	
TM 4	0.76 - 0.90	"	
TM 5	1.55 - 1.75	"	
TM 7	2.08 - 2.35	"	
TM 6	10.40 - 12.5 "	120m x 120m	
SPOT	0.51 - 0.73	10 m x 10 m	
ERS 1 Radar- C band	5.66 cm wavelength	30 m x 30 m	
γ radiometric: flight line spacing 5 Km, Total exposure Potassium (K), Uranium (e U), Thorium (e Th), eU / eTh, eU / K, and eTh / K		5000 m x 130 m	30 m x 30m
Aeromagnetic: 800 m flight line spacing		800 m x 70 m	200 m x 200 m
Gravity:		10 km x 10 km	200 m x 200 m

Table 2.1 Data sets available for the study

1993. The original data was acquired from Radarsat International for GSC, and later was geometrically corrected and transformed to 30 m pixels. A low-pass 3 x 3 filtering operator was originally used to reduce speckle on the image. This radar image is shown in Fig. 2.3. Radar measured microwave backscatter which is expected to generate information of the dielectric properties of surface material as well as topography and the surface roughness. The variations in dielectric constants of rocks are too small to have a significant effect on backscatter and, therefore, the composition of rocks cannot be determined by a direct means using imaging radar. However, the roughness of the surface, manifested in the

weathering or jointing patterns, may be indicative of rock type in a given climatic environment (Daily et al., 1978).



TM7 radar

Fig. 2.3 TM 7 and radar image

The Gamma ray spectrometer data image

Different rock types have different characteristic concentrations of radioactivity; thus concentrations calculated from gamma ray spectrometry data can be used to help identify zones of consistent lithology and contacts between contrasting lithologies (Charbonneau et al., 1976). Gamma radiation in the energy range of interest, 1.46 to 2.62 MeV, is strongly attenuated by most materials, including water and overburden. In areas with thick overburden, attenuation theoretically limits the utility of gamma ray spectrometry data for geological mapping since radiation may be due to transported material unrelated to the bedrock. Variations in elemental ratios related to lithology are difficult to detect from individual radio-element images because the radio-element concentrations are strongly correlated.

Gamma ray spectrometry data were collected in this area by the GSC. The airborne gamma ray spectrometer surveys were flown in the north-south direction at a flight line spacing of approximately 5 km and recorded at 130 m intervals. Gamma ray spectrometry provides a method for estimating uranium (U), thorium (Th), and potassium (K) concentrations in material near the surface of the earth. Uranium and thorium concentrations are determined by measurement of gamma radiation from daughter elements in their respective decay series. To indicate that uranium and thorium isotope concentrations are assumed to be in equilibrium with their decay products, the calculated concentrations are called equivalent uranium (eU) and equivalent thorium (eTh). The uranium and thorium are preceded by the symbol "e" for equivalent.

The intensity of gamma rays in traversing matter decreases exponentially with distance. For practical geophysical purposes gamma radiation may be taken to be entirely absorbed by 1 to 2 meters of rock.

The seven radiometric data sets are: total exposure rate, concentration of potassium (K), equivalent uranium eU, and equivalent thorium eTh (as shown in Fig. 2.4), and the ratios eU / eTh, eU / K and eTh / K. The gamma ray spectrometer data represent measurements of gamma ray flux. The total exposure rate represents the total exposure rate of gamma ray photons produced by radiometric emitters which includes potassium, daughter products of uranium and thorium. The potassium content is determined from the gamma ray photons emitted by K^{40}, whereas uranium (eU) are measured indirectly from the gamma ray photons emitted by daughter product B_i^{214} and thorium (eTh) are measured from the gamma ray photons emitted by daughter product Tl^{208} (Schau et al., 1993).

The total exposure rates were recorded in micro Röentgens per hour. The Roentgen is the quantity of gamma radiation which produces 2.08×10^9 pairs of ions per cm^3 of air at Normal Temperature and Pressure (NTP). The concentration of potassium (K) is

generally expressed as percentage, whereas the concentrations of uranium (eU) and thorium (eTh) are expressed in equivalent parts per million (ppm).

The relative concentration of radioelements, corresponding to eU, eTh, and K concentration of radioelements, provide information on the surface cover material and often is related to lithological variations (Broome et al., 1987). Duval (1983) suggested that the ratios of concentration of radioelements suppress the variations due to soil moisture and those related to differences in altitude.

Minute traces of radioactive minerals are present in all igneous and sedimentary rocks. There is a large difference between the radioactivity of basalts and granites. Granites have a remarkably high content of K^{40}. This fact is of great consequence because granites are very common rocks and the gamma radiation from their potassium produces a radioactive background which may make it difficult to locate uranium and thorium ores. Sometimes the radioactivity of potassium feldspars in pegmatite dikes may be misinterpreted as being due to a concentration of uranium and thorium.

The original data were interpolated into a 500 m grid from flight line spacing of 5 km, and the pixel values were then linearly stretched between 0 and 255 and read into a PCI image file with 30 m x 30 m resampling (Schun, 1989), as shown in Figure. 2.4. All these data sets were compiled, gridded, registered, and geometrically corrected to a UTM topographic base map by GSC.

Aeromagnetic data image

Rocks are made up of basic building blocks known as minerals. From a macroscopic point of view, most rock-forming minerals are non-magnetic. Of rock-forming minerals, only a few minerals, such as magnetite (FeO_3), limonite ($FeTiO_3$), and pyrrhotite (FeS) have significant magnetic properties; magnetic rocks are those which

contain these magnetic minerals. The magnetism of a rock is made up of two components: the remnant magnetism (or permanent magnetism) and the induced magnetism. The magnetism induced in a rock is proportional to the ambient earth's magnetic field by a proportionality constant known as magnetic susceptibility. The total magnetism of a given rock is the vector sum of the remnant and induced magnetism components. Most sedimentary rocks are considered as non-magnetic since they contain negligible quantities of magnetic minerals. Most basic igneous rocks have high magnetic susceptibilities. Acid igneous rocks and metamorphic rocks can have highly variable susceptibilities which can range from negligible to extremely high. Magnetic anomalies, therefore, originated from either basement source or intrasedimentary volcanic sources. The magnetic susceptibility of a rock is proportional to the volume percent of magnetic minerals.

The aeromagnetic data were collected approximately 300 m above the mean terrain surface, along flight line spacing at about 800 m, with one sample approximately every 70 m. The data set was interpolated onto a 200 m grid and the pixel values were linearly stretched into 0 to 255 pixel values as shown in Fig. 2.4.

Gravity data image

Gravity instruments measure the earth's gravitational acceleration. After corrections are made to the measured gravity from the latitude and elevation at each gravity station, the corrected gravity differences (known as Bouguer gravity anomalies) are related to variations in subsurface rock densities. Gravity measurements can be used to direct ore reserve calculations for some massive sulfide ore bodies.

The gravity data were collected at stations approximately 10 km apart. The original unit was in milligals, later were linearly stretched into 0 to 255 pixel values. The sample spacing of the gravity data is very coarse compared to other data. It amounts to only four

true readings in whole area of study as shown in Figure 2.4.

Training patterns

The training patterns used in this study were extracted from the geological outcrop map. Since the vegetation has covered some of the outcrop in the area, this geologic outcrop map shows areas of rock outcrop subtracted the possible vegetated areas by using a Landsat TM vegetation index defined as (TM4 -TM3) / (TM4 + TM3) (Chung, 1993). The vegetation index map was classified into one class and was visually corrected on air photos.

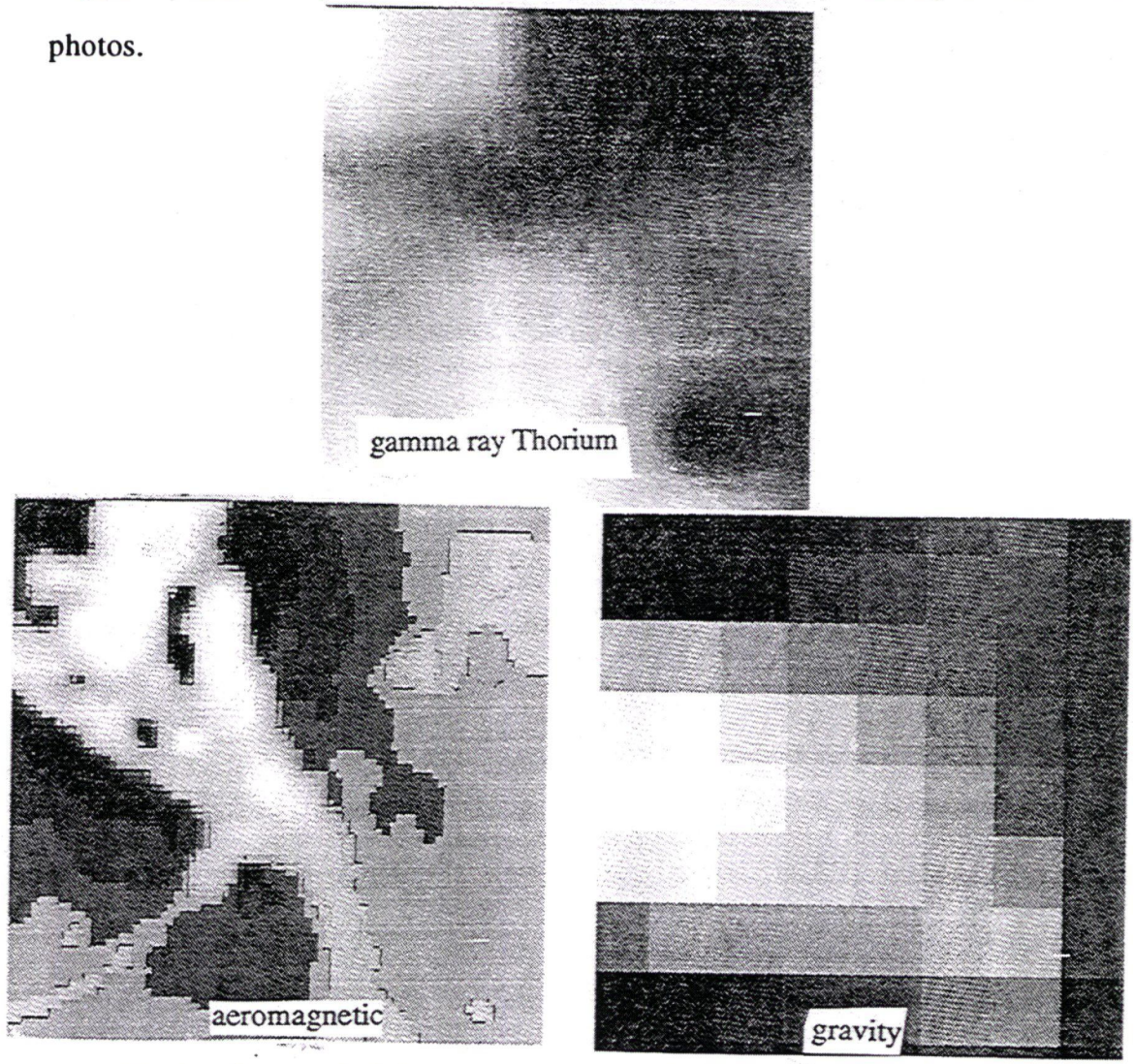


Fig. 2.4 Geophysical data sets: gamma ray thorium, aeromagnetic and gravity

CHAPTER 3

METHODOLOGY AND EXPERIMENTAL PROCEDURES

This chapter has two distinct segments. The first details the algorithm of the back-propagation neural network. The second outlines the experiments to be followed. The experimental design follows directly from the hypotheses given in Chapter one.

3.1 Fundamentals of neural networks

Set of processing units

All neural networks are composed of a set of simple processing units, called nodes or neurodes. All processing in a neural network is carried out by these nodes (i.e. there is no centralized control mechanism that computes values for nodes). Processing units may represent a specific concept or a piece of knowledge, such as features, letters, etc. (the idea is that one processing unit is equal to one concept); and an indescribable part of a larger concept (Rumelhart, et al., ch. 2; Hinton, et al., ch. 3., in PDP, 1986).

Neural networks are envisioned to be collections of individual processors, each capable of a few simple computations (multiplication, adding inputs, computing a new activation level, or comparing input to a threshold value) with control being completely via the passing of output between the processing units. The main tasks associated with a processing unit is to receive input from its neighbors (i.e. those processing units to which it is connected), compute an output, and send that output to its neighbors. Such a system is inherently parallel, because many processing units can be carrying out their computations at

the same time.

The processing units in a neural network can be classified as one of three types:

1. input units, which receive input from external sources, compute their activation level, compute their output as a function of activation level, and transmit this output to the rest of the network,

2. output units, which upon receipt of input from the rest of the network, compute and broadcast their output to external receivers and

3. hidden units, which only receive input from, and broadcast their computed output to, units within the network (i.e. no "outside" contact).

Processing units are usually organized into layers with full or random connections between successive layers. A hierarchical network structure is essential. The network must have a minimum of three layers: an input layer, to accept patterns from the outside world, an output layer to present the network's responses back to the outside world, and one middle (or hidden) layer.

State (level) of activation

After the net input calculated at each unit, each unit then has an activation level a which is most often represented as a continuous quantity between 0 and 1 (0 for inactive, 1 for active). The activation levels of all processing units (the input unit activation levels, the output unit activation levels and the activation levels of the hidden units) in a network represent the state of the network (Rumelhart, et. al., ch. 2 and ch. 8 in PDP, 1986).

Output function employed

Each processing unit transmits its output to its neighbors. This output is determined from the level of activation of the processing unit, and usually with a scalar value between 0 and 1. Associated with each processing unit is an output function f which defines how the output value for the processing unit is determined from its activation. Hence, for unit i of any layer, the relationship between the activation level (a_i) and output value O_i can be described mathematically as follows:

$$O_i = f[a_i] \tag{3.1}$$

In some neural network models, the output function, f, is unity. This means the output is the same as the activation function. In other neural network models, the output functions is a threshold function, i.e., a unit produces no output unless the activation exceeds some predefined level of activation (Rumelhart, et. al., 1986).

Pattern of connectivity among the processing units

Processing units are connected to other processing units and communicate with each other via these connections. There are no connections between nodes in the same layer. It is this pattern of connectivity and the strengths of the connections that mostly influences how a network will response (the various rules for activation, output, etc., will also affect how a network actually responds, but for a constant set of rules, the topology and connection strength will most affect what it computes).

The weight of a connection, w_{ik}, going from the unit in i^{th} layer to a unit in the k^{th} layer , can be positive, negative or zero. If $W_{i,k}$ is positive, the i^{th} unit encourages the activation in k^{th} unit (represented as an arrow), if $W_{i,k}$ is negative, the i^{th} unit discourages the activation of k^{th} unit and i^{th} unit has no effect on the activation in k^{th} unit if W_{ik} is

zero. The absolute value of the weight, $|W_{ik}|$, represents the strength with which the i^{th} unit excites or inhibits the k^{th} unit.

Memories are stored or represented in a neural network in the pattern of variable interconnection weights among the neurodes.

Rule of propagation employed

As stated previously, these processing units compare their output using their output function and then communicate that output to their neighbors. The rule of propagation describes how the inputs impinging on a unit (i.e. the outputs from other processing units) and the strengths of the connections are combined to compute the net input, S_i, to the processing unit. Most often, this rule is simply a weighted summation, described mathematically as follows for a set of N inputs:

$$S_i = \sum_{k=1}^{N} W_{i,k} O_k + b_i, \tag{3.2}$$

where $W_{i,k}$ is the weight of the interconnection from i^{th} unit to k^{th} unit and O_k is the output from k^{th} unit in the previous layer or the input in the current layer, b_i is a bias associated with i^{th} unit.

Activation function employed

The activation function, F, defines how the net input received by a unit i is combined with its current level of activation to compute a new level of activation. The net input is typically the sum of the weighted inputs of the neurode. The activation function which is continuous and bounded is usually expressed mathematically as follows:

$$a_{i\text{-new}} = F[a_{i\text{-old}}, S_i] \tag{3.3}$$

However, in simple network the activation function is equated to the value of the net input arriving at the processing unit i:

$$a_{i\text{-new}} = S_i \tag{3.4}$$

and is true for all feed-forward networks. Fig 3.1 shows a typical unit. The computation of a processing unit is split into two components. First, is a linear component that computes the weighted sum of the unit's values. Second, is a nonlinear component activation function, f, that transforms the weighted sum into the final value that serves as the unit's activation value, a_i. Usually, all units in a network use the same activation function.

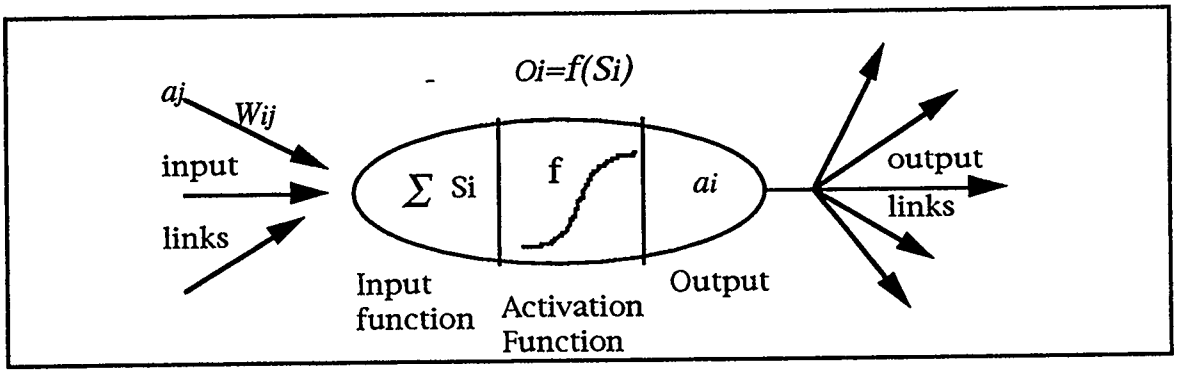


Fig. 3.1 A simple processing unit

The input, activation, and output functions of a neurode in a back-propagation network can usually be combined into one function, the transfer function. From eqns (3.1), (3.2), and (3.4) we know that the output is an explicit function of the input:

$$O_i = f(S_i) = f\left(\sum_{k=1}^{N} W_{i,k} O_k + b_i\right) \tag{3.5}$$

where O_k are all the input nodes (the outputs from previous layer) to node i. For the purpose of simplification, b_i is often set to 0.

Rule of learning employed

The learning rule defines how the network is modified in response to experience (i.e. training cases presented to the system). There are three ways that the network could be modified:

1) Modify the strength of the existing connections between processing units;

2) Develop new connections between the processing units; and

3) Remove existing connections between the processing units;

The second and the third types of network can be considered as special cases of the first. The most common forms of learning rules are Hebbian learning and the delta rule.

3.1.1 Back-propagation neural network algorithm

The back-propagation neural network (BNN) algorithm requires a pre-determined pattern for the output layer of the network. It also requires a single set of weights and biases that will satisfy all the input-output sets of combinations presented to it. The input is a set of observations and the outputs are the desired class membership values $d=\{d_1, d_2,..., d_k\}$. The process of obtaining the weights and bias is network learning, which is essentially the same as supervised training. During network training, the elements in a set of observations $X=\{x_1, x_2, ..., x_k)$ correspond to the nodes in the input layer. When the input (or the observation) X is fed into the network, the network will adjust the weights according to the designed learning rule such that the desired outputs can be obtained. Once this adjustment been achieved, another pair of input-output is presented and the network is to learn that association again.

The algorithm of BNN aims at minimizing the average sum-squared error between

the network output and the desired output. This algorithm allows the problem of representing information by path weights to be reformulated into a problem of minimizing an error function. Both the ADALINE and the Perceptron use a binary (1, -1) transfer function. However, BNN uses a sigmoid function as a transfer function. In the beginning of training, random initial values are assigned to all the connection weights. The perceptron adjusts its weights according the difference between the target and the actual output. The learning rule of the back-propagation networks is the Generalized Delta Rule, which is a generalized form of the LMS (Least Mean Square) rule. The distinguishing feature between Delta Rule and the Generalized Delta Rule is that the error is used to affect not just one set of weights (input-to-output, as with the ADALINE), but two sets of weights (input-to-hidden, and hidden-to-output).

The Generalized Delta Rule uses the chain rule from differential calculus to calculate the way in which these weights (and thresholds) depend on each other. This is typically done in two stages. First, starting at the output layer, the network adjusts the weights between the hidden and output layers. In the second stage, the connection weights between the input and the hidden layers are adjusted. The back-propagation method uses the adjustments and values to the hidden-to-output weights to help determine the changes made to the input-to-hidden weights.

To obtain an average sum-squared error, we sum the errors over all neurodes. The error E for each input pattern p is defined as the sum of the squared error at each of the outputs :

$$E_p = \frac{1}{2}\sum_z (d_{p,z} - O_{p,z})^2 \qquad (3.6)$$

and the total error E is the sum over all input patterns

$$E=\sum_{p} E_p=\sum_{p}\left(\frac{1}{2}\sum_{z}(d_{p,z}-O_{p,z})^2\right) \tag{3.6.1}$$

where

z is the index that ranges over all output nodes

p is the index that ranges over all input patterns

$d_{p,z}$ is the desired output for input pattern p at the i^{th} node

$O_{p,z}$ is the actual output for input pattern p at the i^{th} node.

Since the output O_j is a function of net inputs S_j, Eqns. (3.1) and (3,4), and S_j is a function of the weights, therefore the error E_p is function of the weights. We would drop the subscript of p to indicate E as any one particular input pattern. The usual gradient descent algorithm suggests changing each W_{ij} by an amount ΔW_{ij} proportional to $-\partial E/\partial W_{ij}$, i.e.

$$\Delta W_{ij} = -\eta \partial E / \partial W_{ij} \tag{3.7}$$

By employing the gradient descent method, we can calculate the partial derivative of E with respect to each weight and then add up the partial derivative of E for each input sample. For each input sample, we want to find an efficient way to compute $\frac{\partial E}{\partial W_{ij}}$ (the partial derivative of E with respect to W_{ij}), where the weight, W_{ij}, is a weight connecting the i^{th} layer of nodes to the j^{th} layer of nodes. The effect of W_{ij} on error E, is expressed through the intermediate variable, O_j, the output of j^{th} node as

$$\frac{\partial E}{\partial W_{ij}}=\frac{\partial E}{\partial O_j}\frac{\partial O_j}{\partial W_{ij}}=\frac{\partial O_j}{\partial W_{ij}}\frac{\partial E}{\partial O_j}. \tag{3.8}$$

O_j is determined by adding up all inputs to node j and passing the result through a threshold function. Hence,

$$O_j = f(S_j) = f\left(\sum_i O_i W_{ij}\right)$$

therefore,

$$\frac{\partial O_j}{\partial W_{ij}} = \frac{df(S_j)}{dS_j}\frac{\partial S_j}{\partial W_{ij}} = \frac{df(S_j)}{dS_j} O_i \tag{3.9}$$

Substituting this result back into eqn (3.8), we get the following key equation

$$\frac{\partial E}{\partial W_{ij}} = O_i \frac{df(S_j)}{dS_j}\frac{\partial E}{\partial O_j} \tag{3.10}$$

Note that the partial derivative, $\frac{\partial E}{\partial O_j}$, can be expressed in terms of the partial derivative, $\frac{\partial E}{\partial O_k}$, in the next layer closer to the output layer, so that

$$\frac{\partial E}{\partial O_j} = \sum_k \frac{\partial E}{\partial O_k}\frac{\partial O_k}{\partial O_j} = \sum_k \frac{\partial O_k}{\partial O_j}\frac{\partial E}{\partial O_k} \tag{3.11}$$

where O_k is again determined by adding up all the input node k and passing the result through a threshold function. since $S_k = \sum_j O_j W_{jk}$, hence,

$$\frac{\partial O_k}{\partial O_j} = \frac{df(S_k)}{dS_k}\frac{\partial S_k}{\partial O_j} = W_{jk}\frac{df(S_k)}{dS_k} \tag{3.12}$$

Substituting this result back into the equation for $\frac{\partial E}{\partial O_j}$, yields the following, additional key equation:

$$\frac{\partial E}{\partial O_j} = \sum_k W_{jk}\frac{df(S_k)}{dS_k}\frac{\partial E}{\partial O_k} \tag{3.13}$$

In summary, the two key Eqns (3.10) and (3.13) have two important consequences: first, the partial derivative of system error E with respect to a weight depends on the partial derivative of E with respect to the following output; and second, the partial derivative of E with respect to the output depends on the partial derivatives of E with respect to the outputs in the next layer. From these results we can conclude that the partial derivative of E with respect to any weight in the i^{th} layer must be given in terms of computations already required on layer to the right in the j^{th} layer.

The partial derivative of E with respect to each output in the final layer is derived from (3.6):

$$\frac{\partial E}{\partial O_j} = -(d_z - O_z) \tag{3.14}$$

So far, the derivative of the threshold function f is not calculated. If we choose a sigmoid function for the threshold function:

$$f(S) = \frac{1}{1+e^{-S}} \tag{3.15}$$

the derivative $\frac{df(S_j)}{dS_j}$ is the slope of this function

$$\frac{df(S_j)}{dS_j} = \frac{d}{dS_j}\left(\frac{1}{1+e^{-S_j}}\right)$$

$$= (1+e^{-s_j})^{-2} e^{-s_j} = f(S_j)\,(1-f(S_j))$$

$$= O_j\,(1-O_j) \tag{3.16}$$

Usually, the derivative is expressed in terms of each node's output $O_j = f(S_j)$, rather than the sum of the inputs S_j.

Now Eqn. (3.7) can be written as follows:

$$\Delta W_{ij} = -\eta \partial E / \partial W_{ij} = -\eta O_i O_j (1-O_j) \frac{\partial E}{\partial O_j} \qquad (3.17)$$

for the nodes in the output layer: $\frac{\partial E}{\partial O_j} = -(d_z - O_z)$ (3.14)

for the nodes in the hidden layer $\frac{\partial E}{\partial O_j} = \sum_k W_{jk} O_k (1-O_k) \frac{\partial E}{\partial O_k}$ (3.18)

Once we compute the weight changes for each input sample combination, then we have to add up the weight changes suggested by individual input pattern combinations. Then we can make changes to the weights. The weight changes should depend on the learning rate, η , that should be as large as possible to encourage rapid learning, but not so large as to cause changes to output values that considerably overshoot the desired values.

The above back-propagation equations are incorporated into the following back-propagation procedures. To train a network by the back-propagation rule the procedures are:

1. select a learning rate η

2. initialize weights W_{ij} with random values. For each sample input, compute the result output using

   $O_j = f(S_j) = f(\sum_i O_i W_{ij})$,

   where O_i are all the input nodes to node j.

3. compute $\frac{\partial E}{\partial O_j}$ for nodes in the output layer using

$$\frac{\partial E}{\partial O_j} = -(d_z - O_z)$$

4. compute $\frac{\partial E}{\partial O_j}$ for all other nodes using

$$\frac{\partial E}{\partial O_j} = \sum_k W_{jk} O_k (1 - O_k) \frac{\partial E}{\partial O_k}$$

5. compute weight changes for all weights using

$$\Delta W_{ij} = -\eta O_i O_j (1 - O_j) \frac{\partial E}{\partial O_j}$$

6. add up the weight changes for all sample inputs, and change the weights.

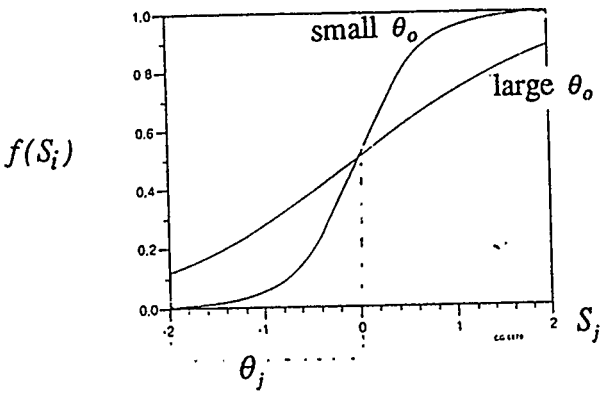
For back-propagation network, we must choose a sigmoid function for f. A more general form of the sigmoid function such as

$$f(S_i) = \frac{1}{1 + e^{-(\lambda S_i - \theta_j)/\theta_0}} \qquad (3.15.1)$$

which is continuous and differentiable (an important property for use in back-propagation learning), is used to maintain the value of activation for a processing unit within the bounds of 0 and 1. Here λ is a gain parameter, which is often set to 1. The parameter θ_j serves as a threshold or bias. The effect of a positive θ_j is to shift the activation function to the right along the horizontal axis. Some nodes may be easily activated to generate a high output value when θ_0 is small and θ_j is low. On the contrary, when θ_o is large, a node will have a slower response to the net input S_j. These effects are illustrated in Fig. 3.2.

The minimization technique stated above is called gradient search, an iterative technique which involves a random set of weight values, evaluating the outputs in feed forward manner and propagate the "errors" backward to previous layers. The development

of the back propagation algorithm provided a mathematical means for presenting the appropriate synaptic strength. Each input to the network is multiplied by the connection weights connecting it to processing elements in the hidden layer. A sum of these products is taken and passed through a non-linear transfer function such as a sigmoid function that compresses a wide domain of inputs to a limited range of outputs. The output of this sigmoid function is passed to the output layer where it is multiplied by the connection weights between the hidden layer and the output layer, and again a sum of products is taken to generate the output for the network.



$$f(S_i) = \frac{1}{1+e^{-(\lambda S_i - \theta_j)/\theta_0}}$$

Fig. 3.2 The sigmoidal activation function, with modification of bias and shape

Rumelhart, Hinton, and Williams (1986) suggest the addition of a momentum term to modify the adjustment of weights in order to keep the change in weight ΔW_{ij} at the $(n+1)^{th}$ iteration similar to the change undertaken at the n^{th} iteration. That is :

$$\Delta W_{ij}(n+1) = \Delta W_{ij} + \alpha \Delta W_{ij}(n) \qquad (3.19)$$

where α is a momentum factor that can take on values between 0 and 1. The parameter α can be called a smoothing factor since it smoothes the rapid changes of the weights. The

changes in weight, $\Delta W_{ij}(n+1)$, is used to indicate the weight change at the $(n+1)^{th}$ iteration, and $\Delta W_{ij}(n)$ indicates the weight change at n^{th} iteration.

3.1.2 Network operations

The neural network operations can be conceptualized as two separate operations: layer operations or the normalization and network operations which includes the training and recalling.

3.1.2.1 Normalization

For a back-propagation neural network, each input can take on any value between zero and one. That is, the input pixel values are continuous and normalized between the values of zero and one. The input nodes simply distribute the signal along multiple paths to the nodes in the hidden layers. The output of each input of each input layer is exactly equal to the normalized input and is in the range of 0 to 1.

The resolution used for each channel was eight bits (i.e. intensities in the range [0,255]). Then the input data normalized to the range of [0,1]. This normalization is achieved by finding the maximum and minimum in each channel and applying the following linear transformation to the original data:

$$\text{new data value} = \frac{\text{original data value} - \text{minimum value}}{\text{maximum value} - \text{minimum value}}$$

The BNN neural network algorithm was adapted from Pao (1989).

3.1.2.2 Software

This study employs neural network software implemented on the Sun Workstation in the remote sensing lab, at the Department of Geomatics Engineering. It was supplied by PCI Inc as part of their EASI/PACE image processing software. This software consists of

three programs: nncreat, nntrain and nnclass. These programs are used respectively for creating and building the synaptic network, training the network, and classification by the network. The programs can handle up to 16 input image channels.

3.2 Experiments to test the hypotheses

In geological mapping, each rock unit (or class) represents meaningful information about the depositional environment. Even with a small population one particular class may be more important than others if it is an indicator of certain mineral deposits. The population of classes is not a concern in lithology mapping. However, in the classification of land-cover or land-use, the small areas are usually neglected.

The experiments for this study follows from the hypotheses laid down in Chapter 1. The experiments are based on various sets of data input, architecture network and training coefficients. In this study, more emphasis will be placed on the classification accuracy of the individual classes since some classes may be more important than others. The process of classification (i.e. with increase of iteration number) is also what we will focus on.

These experiments were intended to reconstruct the lithologic map over the entire area from the available remote sensing data and airborne geophysical data. The training sample sets were extracted from the geologic outcrop map in the area. Six lithologic classes were identified in the area. The more important classes are:

class 1 - PreCambrian gneiss,

class 2 - PAG (Prince Albert Group),

class 3 - Hall Lake plutonic, and

class 4 - SPF (Ship Point Formation)

Among these classes, class 2 (PAG) is the most important class since it is associated with the massive sulfide mineral deposit in the area (An. et al., 1994).

3.2.1 Training the network

The performance of a feed forward BNN classifier and its convergence rate are affected by several variables. The following are a list and explanations of these input variables and training coefficients used in the experiments to test the hypotheses.

Input variables

- The number of layers in the network (or the number of hidden layers).
- The number of input nodes for classification: Usually the number of input nodes will equal the number of input data sets (or channels).
- The number of output nodes: The number of nodes in the output layer will equal the number of classified output classes .
- The number of nodes in the hidden layer: The hidden layer forms an n-dimensional space where n is the number of nodes in the hidden layer. The hidden layer allows the network to form its own internal representation of the data. The internal representation is the foundation on which the decision boundaries are formed (Rumelhart et al., , 1986). With too few hidden nodes, the network may not contain sufficient degrees of freedom to form a representation. The learning algorithm, since it is based on a heuristic, may also improperly train hidden nodes. With too many hidden nodes, the computation is expensive and may lose the generalization ability of the network. Therefore, a balance must be drawn.

- The size of the training samples: The amount of data used for training a neural network affects the accuracy and efficiency of the neural network classifier. Determining the number of training samples required for a satisfactory classification can be difficult. If the training samples is insufficient, the training set may not contain enough information for the neural network to find an optimal classification rule to correctly classify the unknown data. However, if the number of training samples is too large, it becomes too difficult to train the neural network.

- Momentum factor α: The momentum factor ranges from 0 to 1 and usually takes on values between 0.5 and 0.9. A large α speeds up the training.

- The learning rate η: The learning rate is very critical for a successful training, if it is too large, the networks will oscillate and will not converge. If η is too small, the network might learn very slowly or might not converge. In many examples in the literature, the reported range of η's that successfully produce rapid training is fairly small (Rumelhart & McClelland, 1986). Techniques that adaptively adjust η as the network trains have been developed to increase convergence speed (Jacobs, 1988 and others), but they have not found widespread use.

- The iteration number n: Each iteration consists of a forward pass, in which the nodes (processing units) compute their activation value, starting at the input layer, and propagate them forward to the next layer, and a backward pass, in which error signals are propagated backward through the network, starting at the output layer, and weights are changed to reduce the difference between the desired (target) output and the current output.

3.3 Experimental design

Several experiments are designed to gain insight into the network operation to

examine the multiple dependencies of the accuracy on training parameters, input channels and output classes. Classification commonly display significantly different accuracy levels among classes being differentiated. These differences may be due to variable strength of relationship between class memberships and spectral response characteristics, varying level of scene complexity. In all the experiments, only three layers (i.e. a single hidden layer) were used.

Each experiment is described separately in the following:

Experiment set I To test hypothesis 1 - classification accuracy may be maximized by choosing an optimal combination of input data sets (nodes). In order to examine whether the amount of information inherent in the available data have contributed to the lithologic classification, we will make the input data channels as the variable in the training and keep the other parameters constant. The number of nodes in the input layer equates to the number of input data channels.

The sets of data inputs selected are listed below:

1. 15 channels - TM 1 to TM 7, SPOT, radar, gravity, magnetic, and total exposure, K, eU, and eTh
2. 10 channels - TM 5, 6, 7, radar, gravity, magnetic, total exposure, K, eU, and eTh.
3. 7 channels - TM 7, gravity, magnetic, total exposure, K, eU, & eTh
4. 15 channels - TM 5, 6, 7, radar, SPOT, gravity, magnetic, TT, K, eU, eTh, eU / eTh, eTh / K, eU / K and mag / rad

The network parameters are kept constant as follows: momentum factor α =0.9,

learning rate $\eta = 0.01$, the number of hidden nodes is 30.

In these experiments, four output classes of the network are the four lithologic units: unit 1 - Precambrian gneiss; unit 2 - Prince Albert Group; unit 3 - Hall Lake Plutonic; and unit 4 - Ordovician carbonate. Since more diagnostic spectral information about the composition of minerals and rocks were provided in the short-wavelength infrared region than the visible and near-infrared regions (Goetz, 1981), we suspect the importance of TM 1, 2, 3, and 4 in aiding the detection of lithology. We showed two sets of the same number of input data sets (15 channels), one (network #1) with TM 1, 2, 3, and 4 as input data sets, another one (network # 4) without TM 1 to TM 4. We expect to obtain a better classification accuracy results with network #4 when comparing with the results from network # 1. We will also compare the results from another networks (# 3 and #2) : 10 and 7 channels (without TM 5, 6, and radar) of input data as stated above. We expect to obtain little difference in classification accuracies from these two networks.

Experiment set II To test hypothesis 2: the classification accuracy decreases as the network tries to classify more classes. Since we want to classify the lithology, the output classes will be the lithologic units from the outcrop map. The output nodes are determined by the number of lithology units to be classified. Two experiments are trained and tested with four and six lithological classes respectively. The additional classes are indicated as unit 5 and unit 6 in the geological outcrop map. We expect to obtain a less accurate classification result with six output classes than the four output classes.

Experiment set III To test hypothesis 3: a larger number of training samples do not guarantee a better classification accuracy. The selection of training samples from geologic outcrop map entails picking many small subsets of the image representing examples of desired classification. Each class contains multiple small regions from the overall image. The small regions were selected to incorporate variations in the data within a given class

such that each training pixel is unique and all pixels contain new information to the network to learn. The training samples available for each class reflect the intrinsic variability of the class.

The number of pixels and the percentage of training samples of each individual classes are shown in the Table 3.1.

	Class 1 gneiss	Class 2 PAG	Class 3 plutonic	Class 4 SPF	Total
Outcrop map	354 $(2\%;\frac{354}{17626})$	5837 (33.1%)	9203 (52.2%)	2232 (12.7%)	17626 (100%)
Training #1	107 $(30\%;\frac{107}{354})$	882 (15%)	1240 (13%)	805 (36%)	3034 $(17\%;\frac{3034}{17626})$
Training #2	39 $(11\%;\frac{39}{354})$	675 (12%)	571 (6%)	226 (10%)	1511 $(9\%;\frac{1511}{17626})$

Table 3.1 Number and percentage of training samples in the experiments

A neural network is trained separately with 9 and 17 percent of the image data, that is a total of 1511 pixels (approximately 9 % of 17626 outcrop pixels) and a total of 3034 pixels (approximately 17% of 17626 outcrop pixels) of samples from outcrop map were used in training the network. Training set #2 has more training samples for class 2 (675 samples) than class 3 (only 571 samples) although class 3 covers almost twice the area as

class 2 (9203 pixels vs 5837 pixels). We would not expect the classification accuracy from training # 1 is better than the result from training #2.

Experiment set IV The following experiments are designed to test hypothesis 4: the optimal number of hidden nodes depends on the input data sets. The variables are the number of hidden nodes and various number of input layers as shown in table 3.2.

	Number of input layers	Variable (parameter)
1	15	Number of hidden nodes :10, 20, 30, 40, 50
2	10	Number of hidden nodes :10, 20, 30, 40, 50
3	7	Number of hidden nodes :10, 20, 30, 40, 50

Table 3.2 Variables in the fourth set of experiments

where

15 channels includes: TM 5, 6, 7, radar, SPOT, gravity, magnetic, TT, K, eU, eTh, eU/eTh, eTh / K, eU / K and mag / rad,

10 channels includes: TM 5,6,7, radar, gravity, magnetic, total exposure, K, eU, and eTh,

7 channels includes: TM 7, gravity, magnetic, total exposure, K, eU, & eTh.

As stated in the Chapter 1, there is a phenomenon called "over-fitting" which occurred when the output errors for the training examples is lessened but the errors for novel examples increase. For instance, suppose that we have two networks, one which has

ten hidden nodes and another which has one hundred hidden nodes. After enough training with a thousand examples, the large network, which has a hundred hidden units, may produce better outputs for the training examples than the small one, but it may emit worse outputs for inexperienced inputs. Generating accurate outputs for known inputs competes against predicting accurate outputs for unknown inputs. The other disadvantage of more hidden nodes is longer training time for the network.

We expect that the optimal number of hidden nodes depends on the input data sets. We also expect the optimal number of hidden nodes is higher if using more input data layers.

Experiment set V To test the hypothesis 5: the dependence of classification accuracy on the training coefficients (learning rate and momentum factor) is less than the dependence on other parameters, such as the input data, the output classes, or the training sample sizes, etc.. We will compare the classification results when the only variable is momentum factor or the only variable is the learning rate as shown in Table 3.3. We expect that the variation in classification accuracy among various momentum factors is less than the accuracy variations in all previous experiments. We expect the same with various learning rates.

	Input channels	Variables
1	10	Momentum factors α : 0.9, 0.8, 0.7, 0.6, 0.5
2	10	Learning rate η: 0.01, 0.005, 0.001 , 0.0005

Table 3.3 Variables in the fifth set of experiments

Figure 3.3 is the schematic diagram of an back-propagation neural network used to predict lithology.

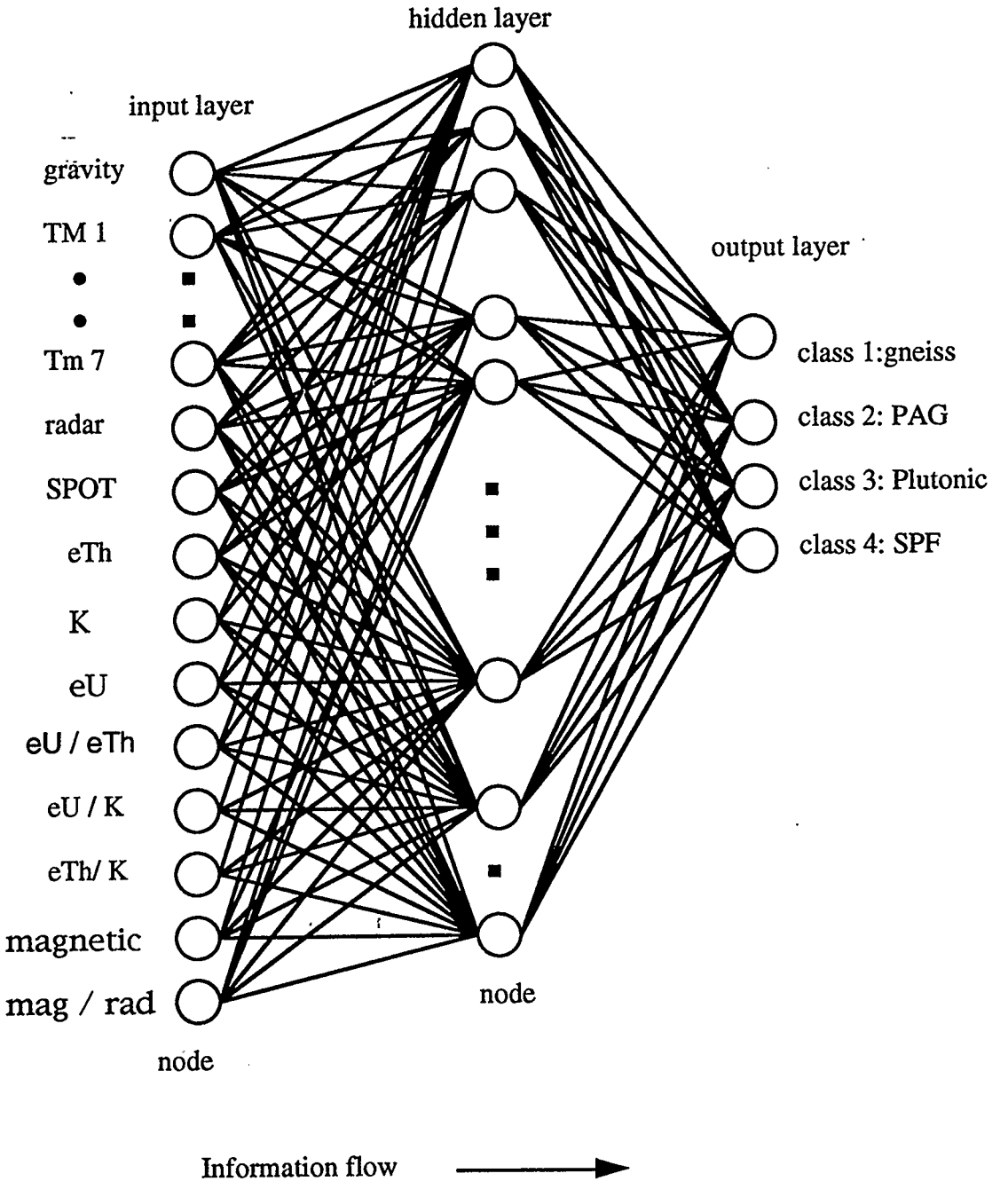


Figure 3.3. The schematic diagram of a back-propagation neural network used to predict lithology.

CHAPTER 4

RESULTS AND DISCUSSION

This chapter presents the results and interpretation of the experiments. In order to evaluate the true classification accuracy, the testing area is the whole area excluding the training area, and the classification results are plotted against the iteration number at 100 iteration intervals.

Accuracy evaluation of the classification results are based on the total percent correct, both for all the classes and for individual class. There are two ways to calculate the accuracy for all classes: average accuracy and overall accuracy. The average accuracy is the mean value of all the individual accuracy and the overall accuracy is the average value of all individual accuracy times the number of total number of pixels of that individual class.

4.1 Experimental results and discussion

In all experiments, class 1 (gneiss) which accounts for 2 % of total pixels (354 of 17626) was always classified with 100 % accuracy. This can be explained by its gross radiometric signature: the granitic / gneissic units are moderately to highly radioactive.

Experiment I To test hypothesis 1 - classification accuracy may be maximized by choosing an optimal combination of input data channels.

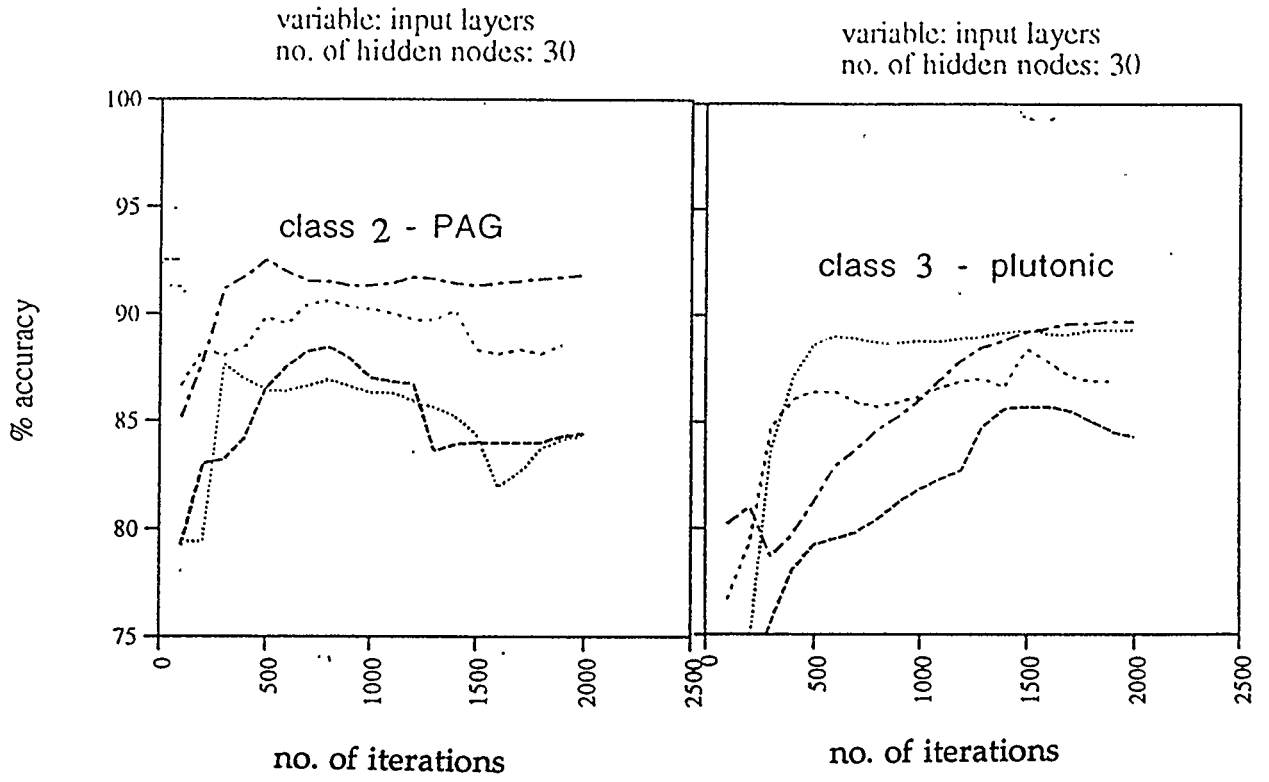
Results Four different combinations of input data sets (10 channels, 7 channels, and two different combinations of 15 channels) were used in the classification. The results

are shown in Fig. 4.1 (for class 2 and 3), Fig. 4.2 (for class 4) and Fig. 4.3. (for average and overall accuracy).

Discussion As we expected, the network # 4 (without TM 1, 2, 3, and 4 as input data sets) produced the highest average and overall classification accuracies, and highest accuracy for class 2 and 3, with exception for class 4. The network (#1) which includes all TM channels as its input data channels produced less accurate classification results than those results from networks (#2 and #3) with 10 or 7 input layers.

The exceptional high accuracy for class 4 (when using all TM channels as input sets) is because the high concentration of calcium in the Ship Point Formation inhibited the vegetation growth therefore they are easily detected from TM images. Figures 4.1, 4.2 and Fig. 4.3 clearly show that geophysical data sets contain more lithologic information the network needed in the classification. Because the spectral features of most rock types and derived soils are not very distinctive in the visible and near-infrared wavelength region, explains why the inclusion of TM 1 to TM 4 has little contribution to the classification accuracy. It also indicates that indiscriminately increasing the number of input data sets may produce less accurate results. The number of input data layer is not the factor that produces higher classification results, but more relevant input data would aid the improvement in classification accuracy. The input data sets should contain information relevant to the rock lithology to significantly affect the classification accuracy. The average and overall accuracies of network #3 and #2 are very close just as we expected.

Our experiments support hypothesis 1.



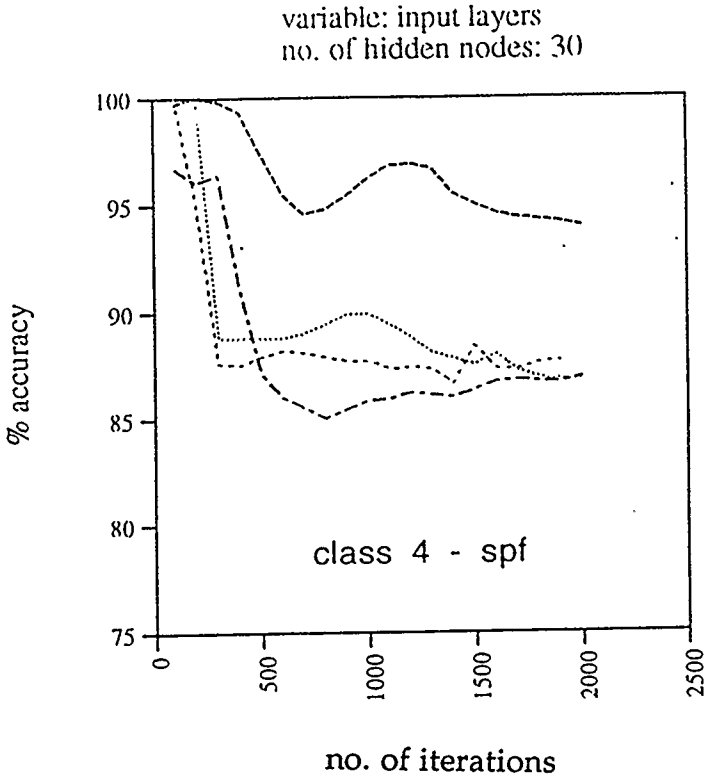
10 channels (TM 5,6,7, radar, gravity, magnetic and TT, eU, eTh, and K)

7 channels (TM 7, gravity, magnetic and TT, eU, eTh, and K)

15 channels(TM 1 to TM7, radar, SPOT, gravity, magnetic, TT, eU, eTh, & K)

15 channels (TM5,6,7, radar, SPOT, gravity, mag, mag/rad, TT, eU, eTh,K, eU/eTh, eTh/K, eU/K)

Fig. 4.1 Comparison of accuracy results for class 2 and class 3 with four sets of input data.



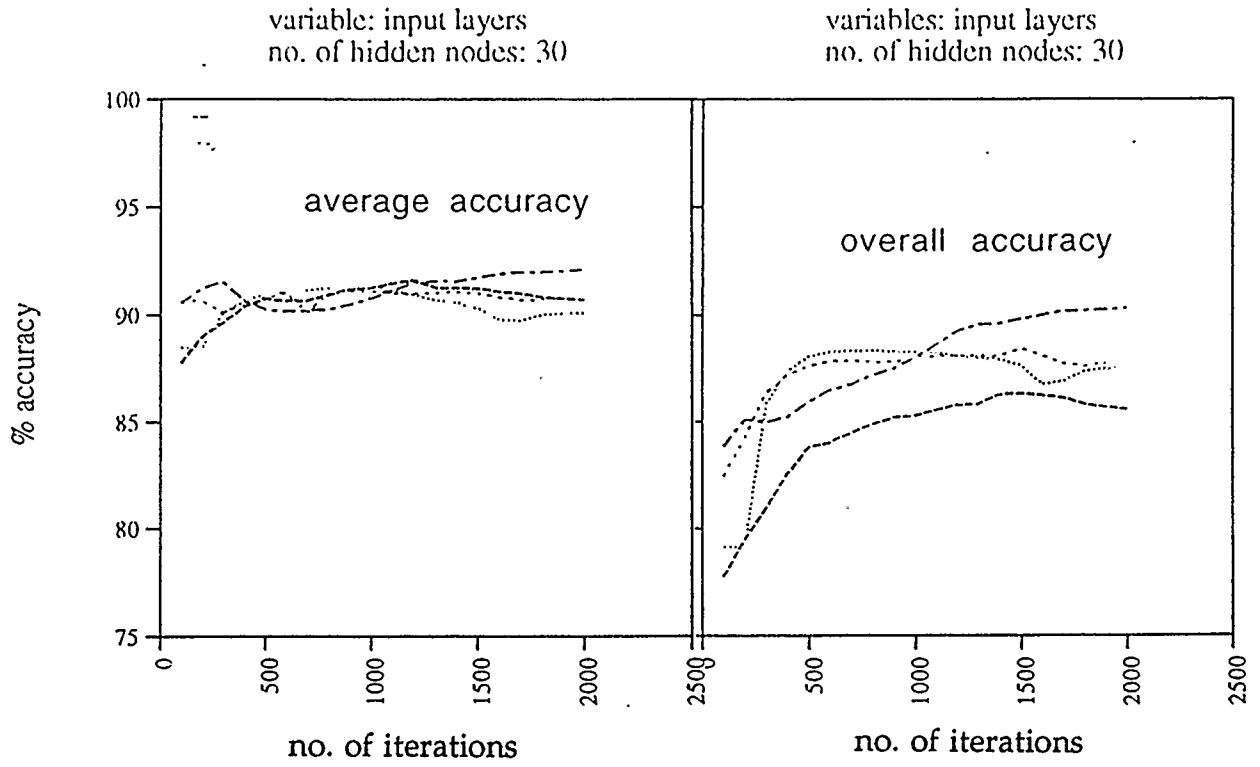
..... 10 channels (TM 5,6,7, radar, gravity, magnetic and TT, eU, eTh, and K)

---------- 7 channels (TM 7, gravity, magnetic, TT, eU, eTh, & K)

---------- 15 channels(TM 1 to TM7, radar, SPOT, gravity, magnetic, TT, eU, eTh, & K)

---------- 15 channels (TM5,6,7, radar, SPOT, gravity, mag, mag/rad, TT, eU, eTh,K, eU/eTh, eTh/K, eU/K)

Fig. 4.2. Comparison of accuracy results for class 4 with four sets of input data.



10 channels (TM 5,6,7, radar, gravity, magnetic and TT, eU, eTh, and K)

7 channels (TM 7, gravity, magnetic and TT, eU, eTh, and K)

15 channels(TM 1 to TM7, radar, SPOT, gravity, magnetic, TT, eU, eTh, & K)

15 channels (TM5,6,7, radar, SPOT, gravity, mag, mag/rad, TT, eU, eTh,K, eU/eTh, eTh/K, eU/K)

Fig. 4.3. Comparison of average and overall accuracy results with four sets of input data.

Experiment II To test hypothesis 2 - the accuracy of classification decreases as more classes being classified. The neural network of the first experiment classifies only four rock units, and the second neural network has six output rock units with four units in common.

Results The classification results of the four common classes for both experiments are plotted against the number of iterations as shown in Fig. 4.4 and Fig. 4.5.

Discussion As we expected, there is an overall decrease in accuracy for the second network (six output classes). The greatest decrease in accuracy is in class 2 (13 %), and a 9% decrease for class 4, with an exception of class 3 (an increase of 2 %). Generally speaking, the results show that the average and overall classification accuracy decrease as the number of output classes increases, but it is not always true for every individual class.

Experiment III To test hypothesis 3 - a larger number (or the size) of training samples does not guarantee a better classification accuracy. In our experiments we used two sets of training samples as summarized in the Table 3.1 in Chapter three.

Results When training with smaller samples (training #2, with training samples 9 % of total pixels), the classification accuracy of class 2 and class 4 unexpectedly increased 4 % and 11%, respectively. However, the classification accuracy for class 3 decreased almost 20 % (Fig. 4.6). About 2 % greater for the average accuracy and 9 % for overall accuracy when training with larger sample size (training #1) (Fig. 4.7).

Discussion From Table 3.1 we noticed that the population is unbalanced among the classes. Class 3 accounts for 52% of the total pixels (9203 of 17626), class 2 (5837 pixels) accounts for 33% and class 4 (2232 pixels) accounts for 13 % of the total population. Class 1 is very small, only 2 %, but the classification remains at 100%

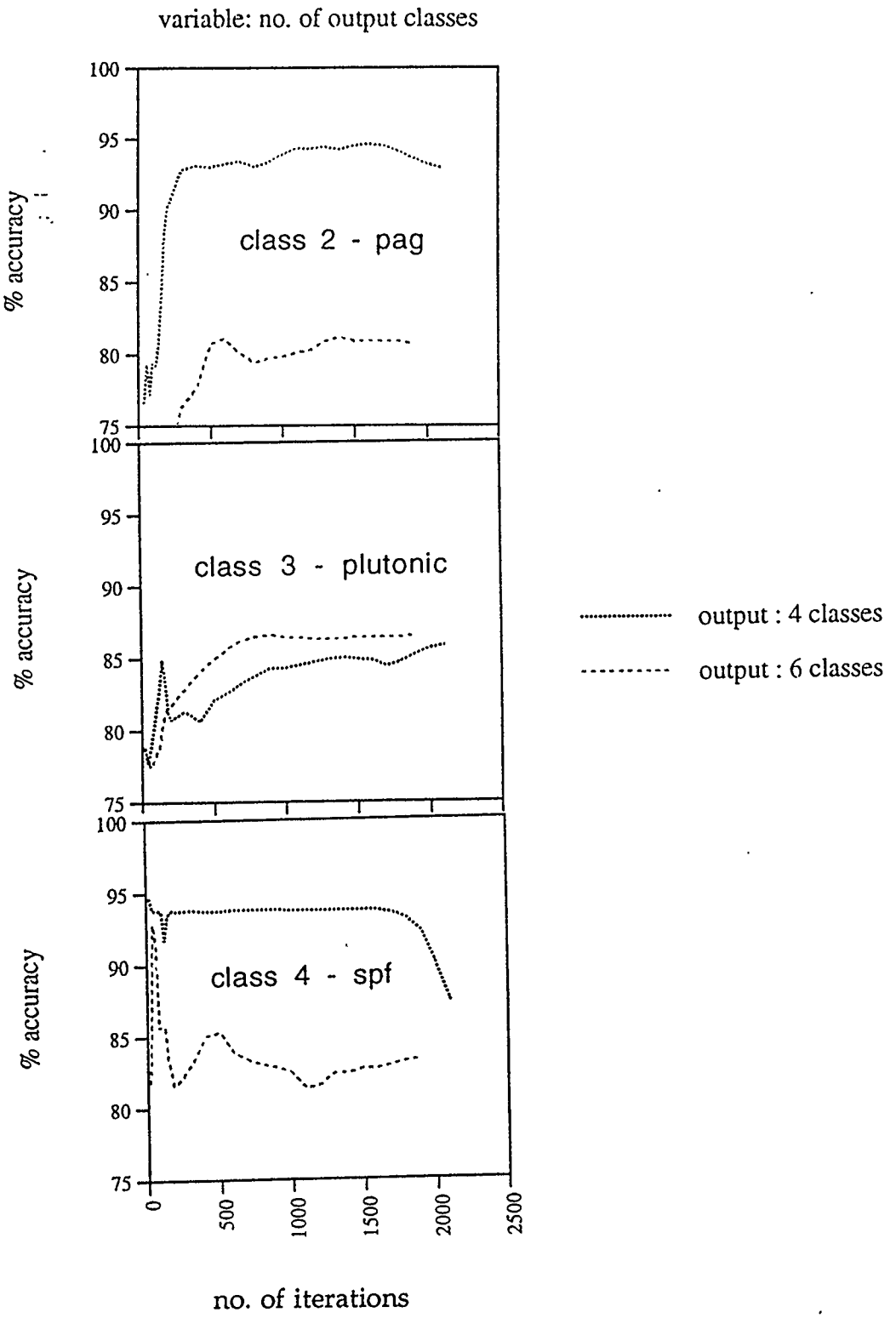


Fig. 4.4. Comparison of individual classification accuracy for different output classes.

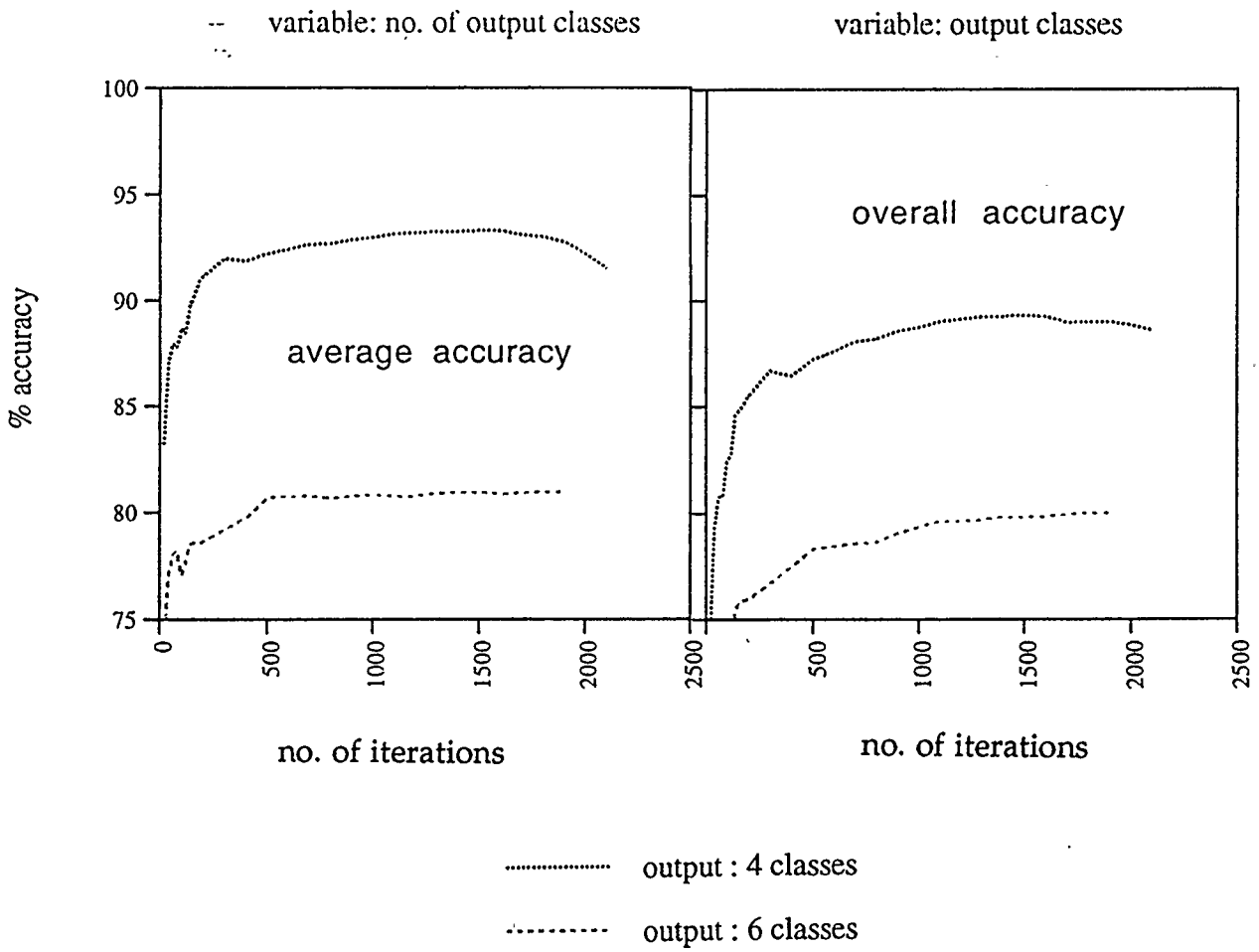


Fig. 4.5. Comparison of average and overall accuracy for different output classes

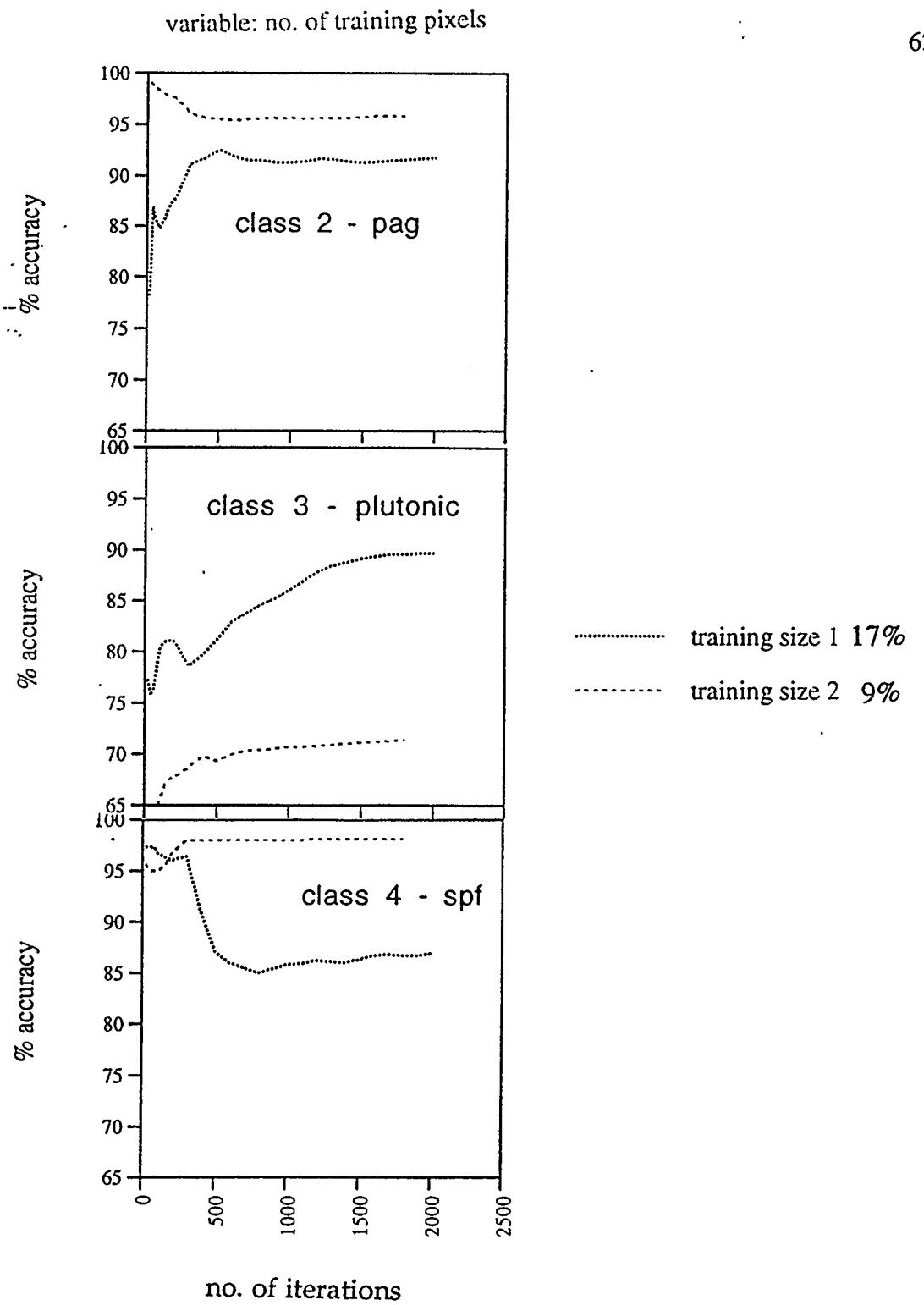


Fig. 4.6. Comparison of indivisual classification accuracy when trained with two different training samples.

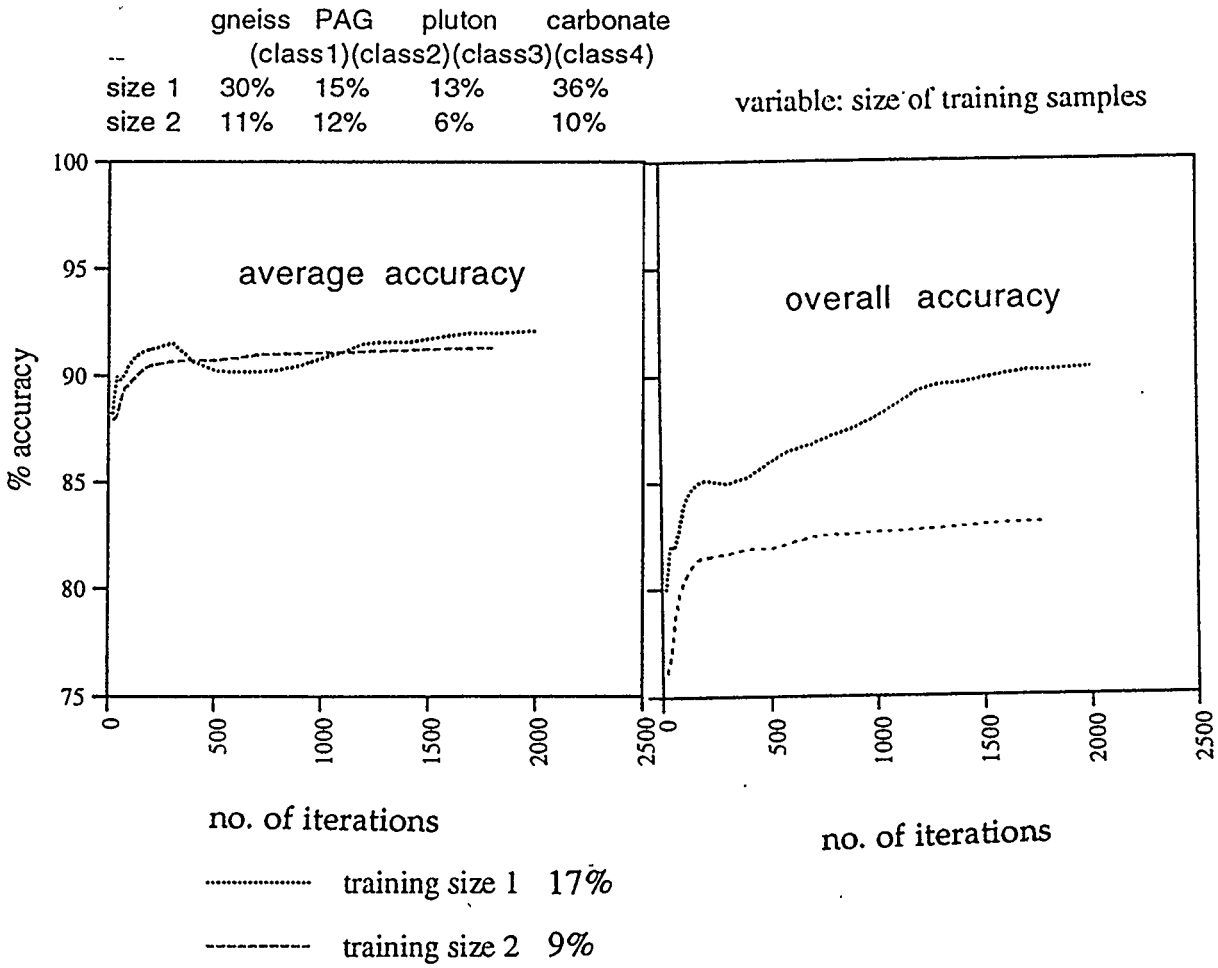


Fig. 4.7. Comparison of average and overall classification accuracy when trained with two different training samples. The percentage shown above indicates sample pixels among individual class.

accuracy. When comparing the classification accuracy among the classes, the average and overall accuracy tend to get better with larger training samples (training #1, 17% training samples). However, the accuracy of each individual class does not behave in the same fashion. If we count only the total number of training samples then the result of average and overall classification accuracy contradict the hypothesis. If we look at each individual class, hypothesis 3 is true, that is the larger number of training samples does not guarantee a better classification accuracy.

Because a complicated class might involve a wider range of input data, intuitively, a greater variety of training samples would needed for a complicated class. In this study, the second unit, PAG, contains rocks with very different spectral and geophysical properties and involves more complicated input pattern. While the third unit, Hall Lake Plutonic, has relatively consistent geophysical and spectral properties and thus involves relatively consistent input data. As a result, the training #2, which has 675 samples (12 %) of unit 2, 571 samples (6%) of unit 3 and 226 samples (10%) of unit 4, produced an improvement in classification accuracy for class 2. Although the third class - HLP covers an area about twice as large as the PAG, if we choose greater variety of training samples of unit 2 (675 samples) than unit 3 (571 samples) we would get a better classification accuracy for unit 2.

	Class 1 (354)	Class 2 (5837)	Class 3 (9203)	Class 4 (2232)	Total (17626)
Training #1	107 (0.6 %)	882 (5 %)	1240 (7 %)	805 (4.5 %)	3034 (17 %)
Training #2	39 (0.2 %)	675 (4 %)	571 (3.1 %)	226 (1.3 %)	1511 (8.6%)

Table 4.1 Percentage of training samples for each individual class.

Table 4.1 summarizes the percentage of each individual class in terms of total pixels (17626).

Training #1 has 7 % training samples of class 3 which covers more than half of the total area, however, training #2 only has 3 % training samples of class 3. This explains why the average and the overall classification accuracy for training # 1 is better - because the overall training samples is larger in training #1. It also explains why the class 2 classification accuracy of training #2 is better because the training samples for class 2 is larger in training #2.

It is the relative percentage of training samples between the classes which affects the individual classification accuracy not the number of the total training samples which influence the individual classification accuracy. From this experimental result we would conclude that if we want to increase the classification accuracy of certain class we should provide greater variety of training samples from that particular class.

Experiment IV To test hypothesis 4 - the optimal number of hidden nodes depends on the input data sets. As described in the previous chapter, four sets of experiments have been executed to test the hypothesis.

Results Fig. 4.8 to Fig. 4.11 present the classification result for two networks with different sets of input channels. The variable is the number of hidden nodes. The classification accuracy curve for different momentum factors are plotted in Figs. 4.8 and 4.9. The classification accuracy curve for different learning rates are plotted in Figs. 4.10 and 4.11.

Discussion As we expected, the optimal number of hidden nodes is dependent on the input data layers as shown in Figs. 4.8 to 4.11 and the classification accuracy has a larger

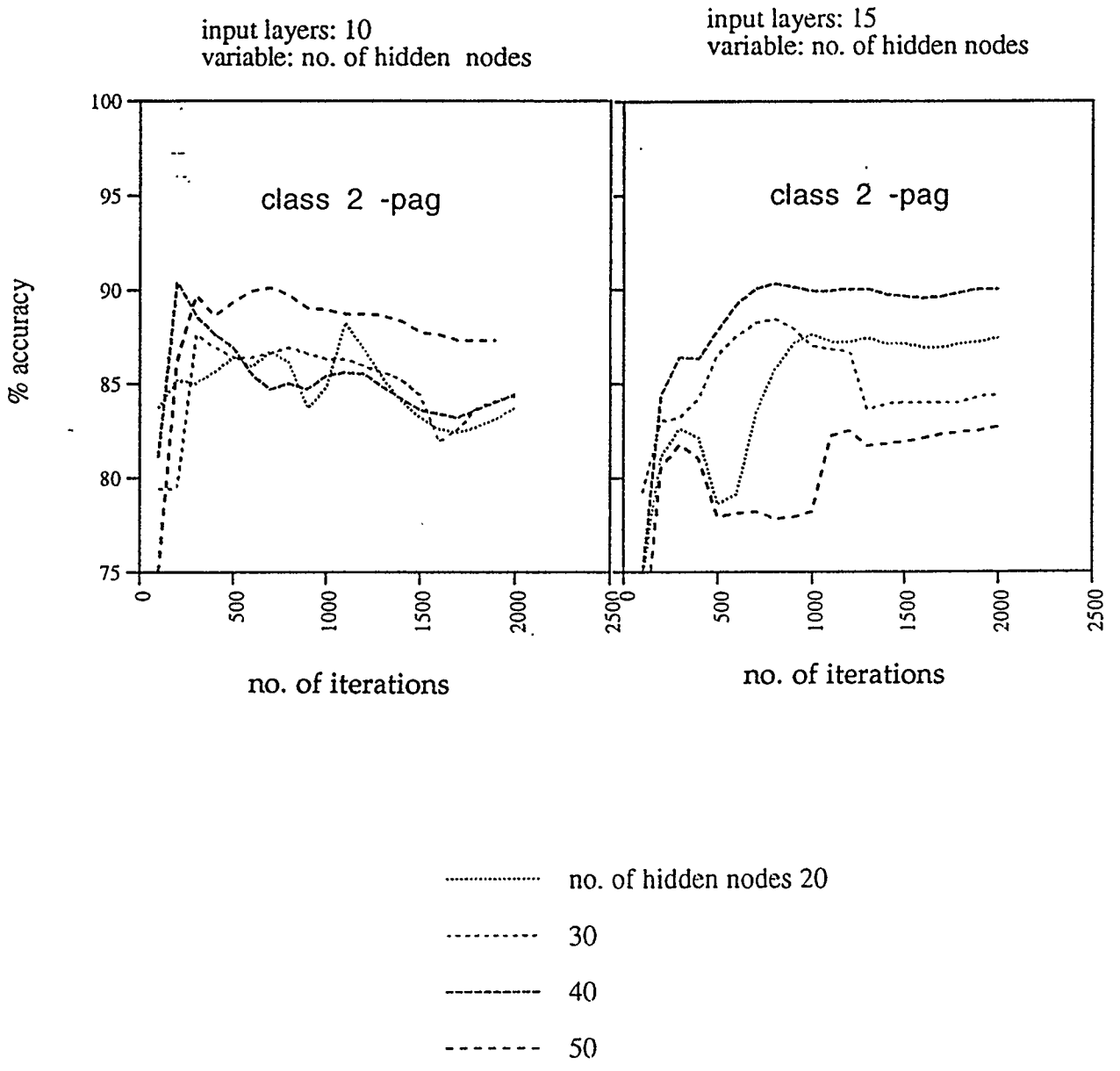


Fig. 4.8. Comparison of classification results for class 2 when using two sets of input data (10 and 15 channels) and the number of hidden nodes are varied.

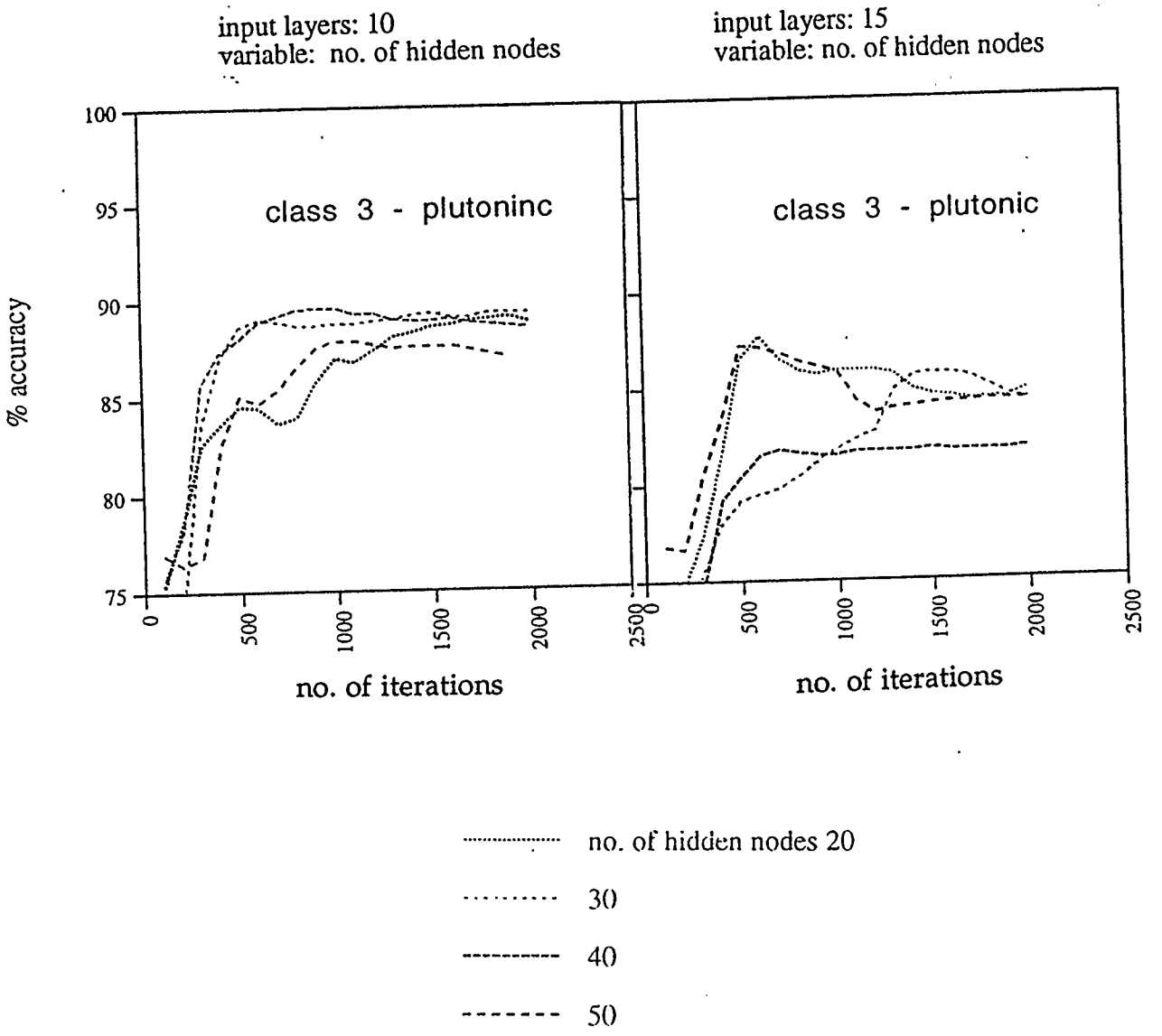


Fig. 4.9. Comparison of classification accuracy for class 3 when using two sets of input data (10 and 15 channels) and the number of hidden nodes are varied.

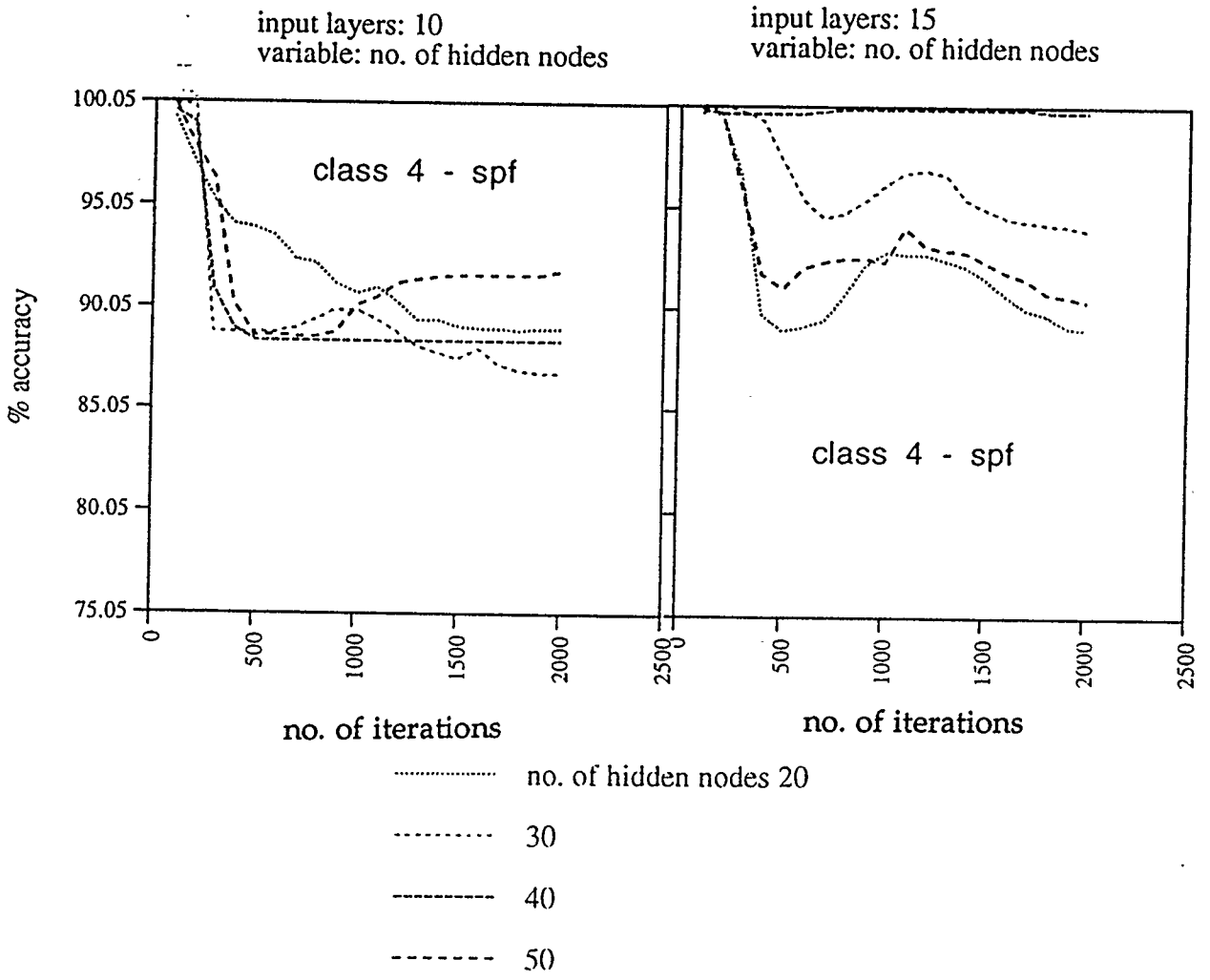


Fig. 4.10. Comparison of classification accuracy for class 4 when using two sets of input data (10 and 15 channels) and the number of hidden nodes are varied.

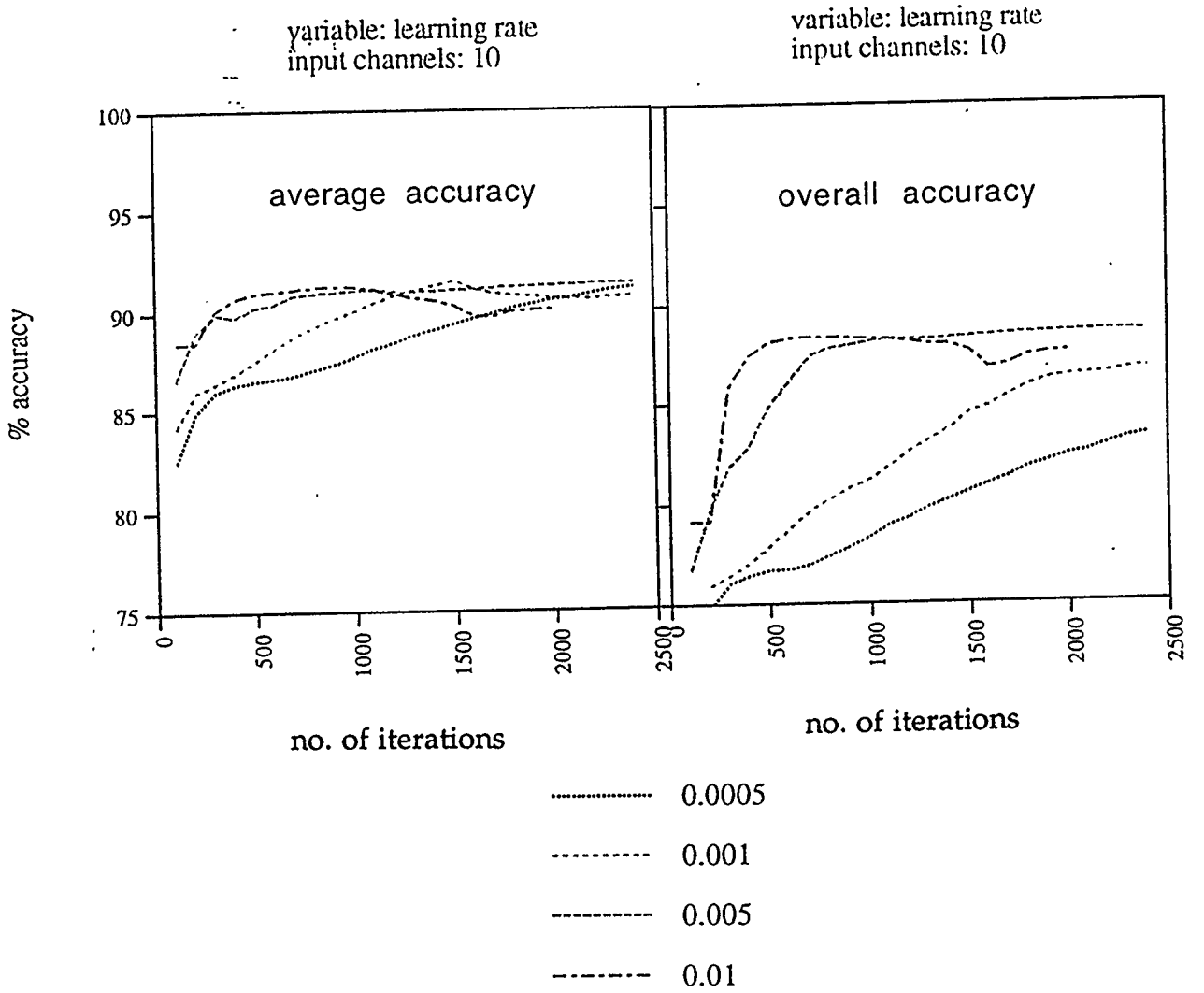


Fig. 4.11. Comparison of average and overall accuracies when training with different learning rates.

variation among various number of hidden nodes for the network with 15 input channels than the network with 10 channels (also true for 7 channels). For the network of 15 input layers, 40 hidden nodes produced the best accuracy, and with 10 input layers, 50 hidden nodes seems to be the optimal number of hidden nodes. Unfortunately, the rule of thumb of twice the number of input data sets being the optimal number of hidden nodes (Gong, in press) seems not suitable in this study. The optimal number of hidden nodes in this study is in the range of three or four times of the number of input data channels.

Experiment V The dependence of accuracy on the training coefficients (learning rate and momentum factor) is less than the dependence on other parameters, such as the input data, the output classes, or the training sample sizes, etc.

Results Fig. 4.12 shows the classification accuracy at various learning rates and Fig. 4.13 shows the classification accuracies at various momentum rates. The learning rate has similar performance as the momentum factor : the larger the learning rate, or the larger the momentum rate, the faster the network reaches its maximum classification accuracy. The momentum factor as a function of iteration number (N) at two different learning rates η (0.01 and 0.005) are plotted in Fig. 4.14. The relationship of learning rate as a function of iteration at two different momentum factors is shown in Fig. 4.15.

Discussion As we expected, the variation among the various momentum rates is about 1% for both average and overall accuracy. It has the least variation among all the experiments. The momentum rate affects the learning speed and the convergence rate. As the momentum factor increases, the number of iterations necessary to be less than the fixed error decreases. The relationship between the momentum factor and the iteration is approximately linear. The effect of learning rate η is to change the slope but not the intercept. Hence, a "slow" learning neural network will require more iterations to achieve the same error (was set at 0.035) than a "fast" learning neural network. In order to evaluate

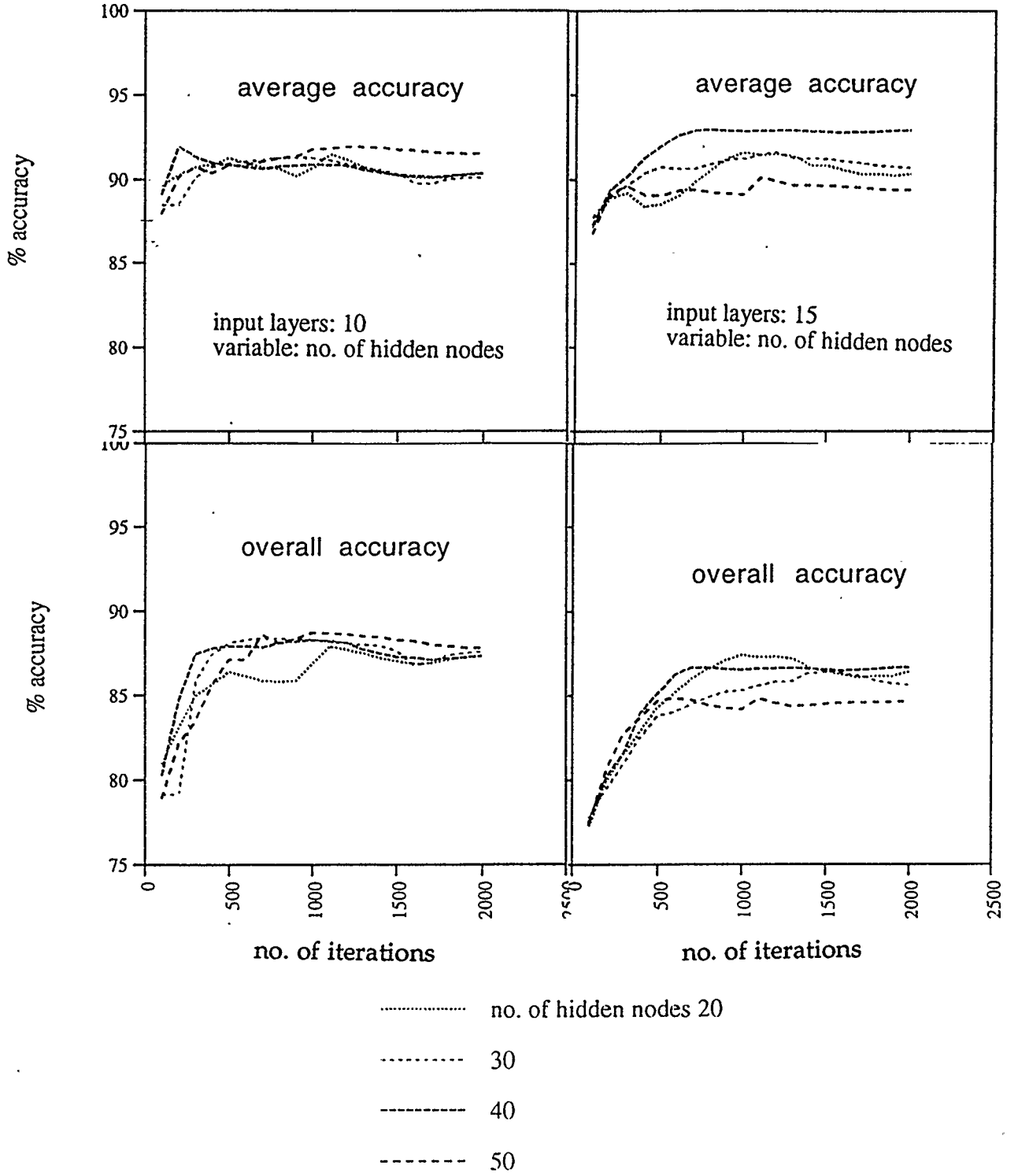


Fig. 4.12. Comparison of average and overall classification accuracies when using two sets of input data (10 and 15 channels) and the number of hidden nodes are varied.

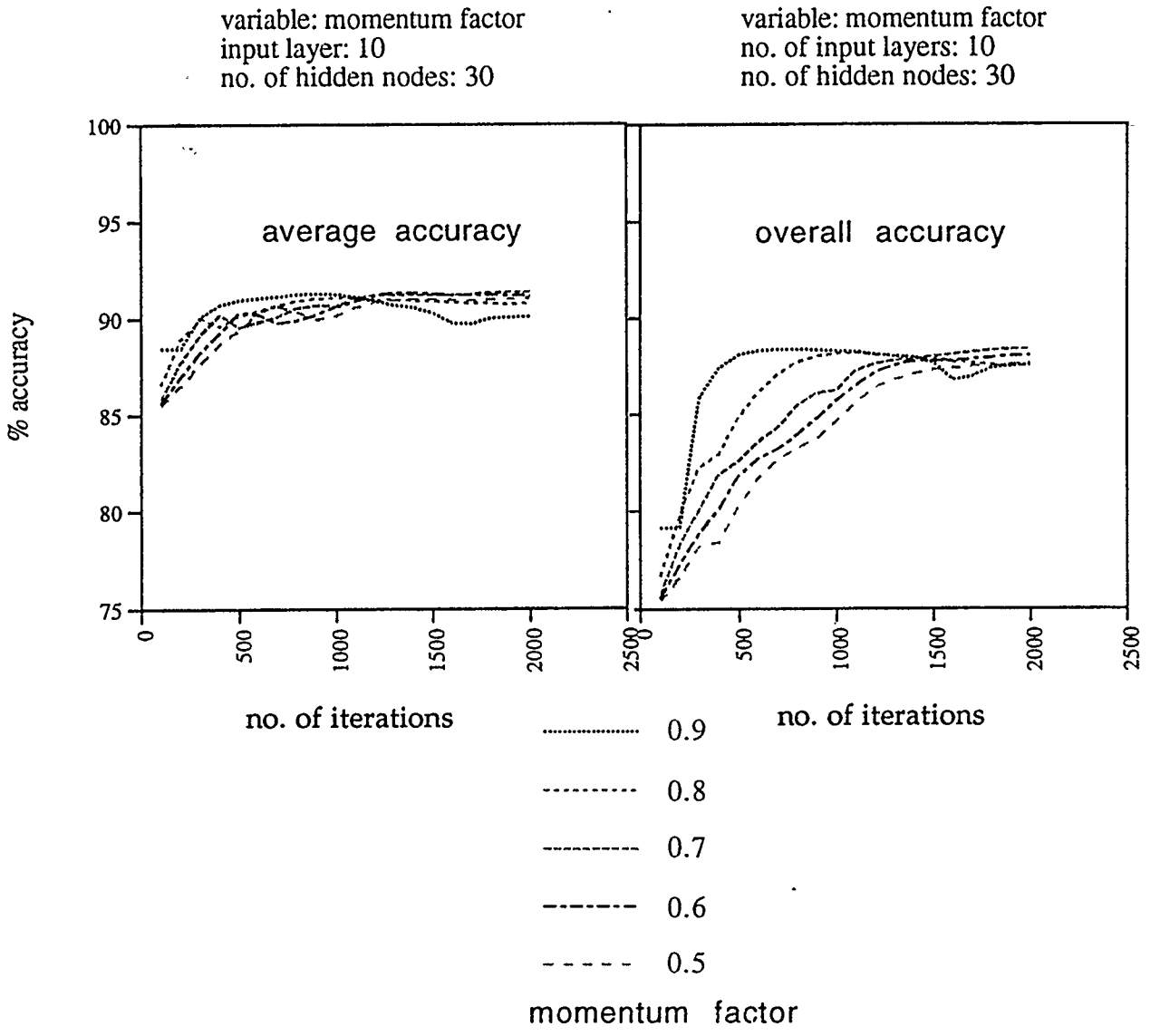


Fig. 4.13. Comparison of average and overall classification accuracies when training with various momentum factors.

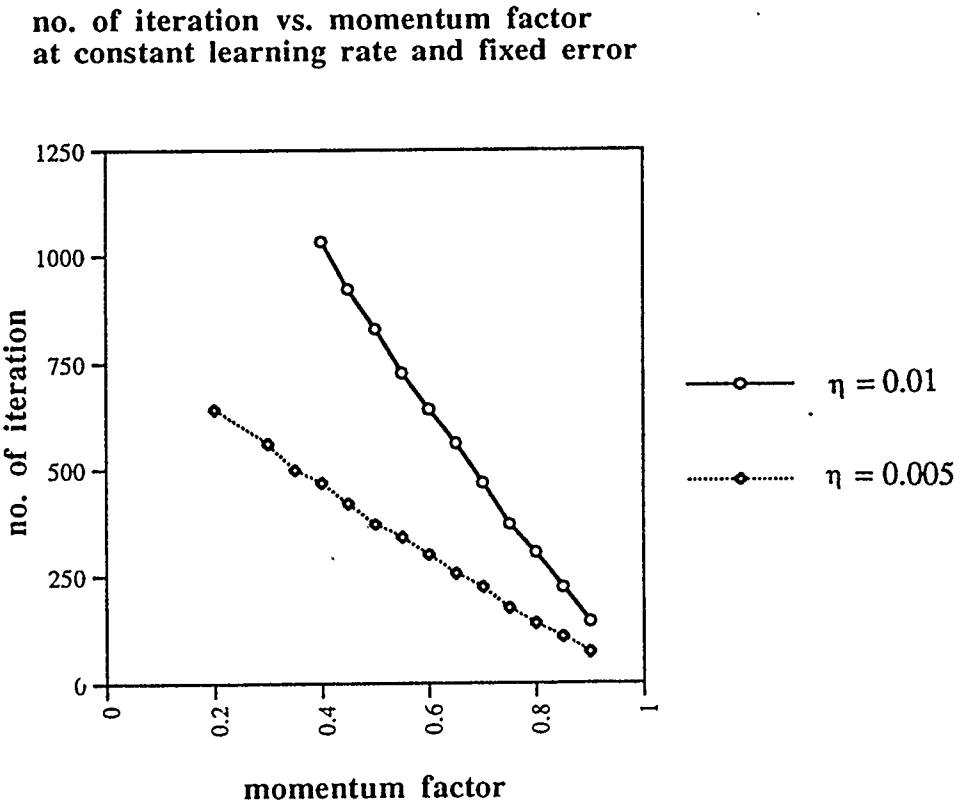


Fig. 4.14. Curve of momentum factor with constant learning rate and fixed error.

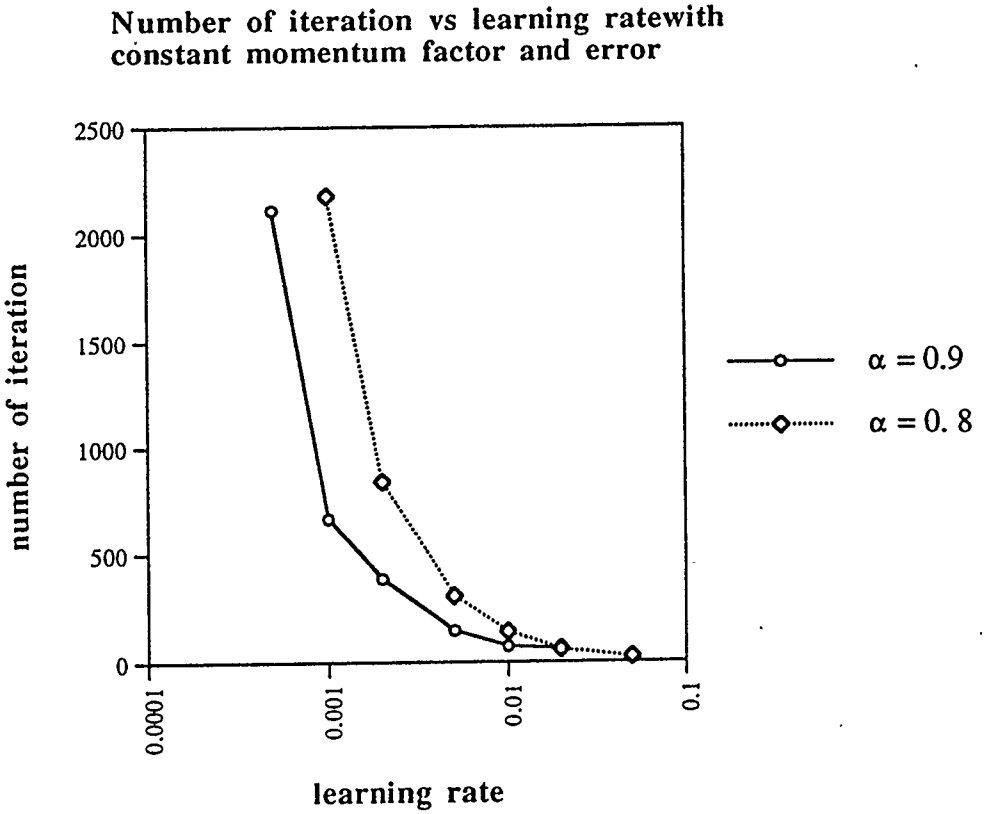


Fig. 4.15. Curve of learning rate with constant momentum factor and fixed error.

training sample set and with the testing sample vs. the iteration number, as shown in figure 4.16 .

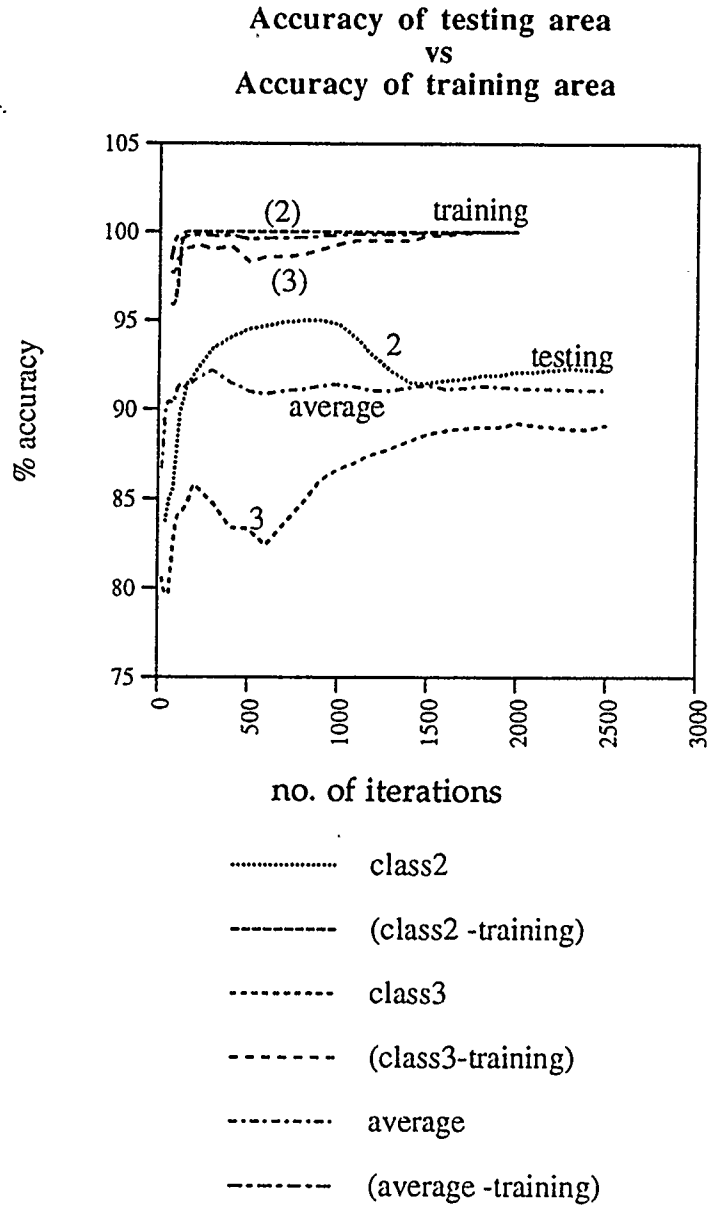


Fig.4.16. Comparison of classification accuracies of training set and testing set.

CHAPTER 5

SUMMARY AND CONCLUSIONS

5.1 Summary

This study has analyzed the performance of various neural networks in lithologic classification by increasing the number of iterations. Our approach gives us insights to better understand the various dependencies of classification accuracy on different training parameters. All classifiers considered in this study are of "supervised " type. We used the same training set (except in the second set of experiments) and assessed the performances of various networks on the same test set.

5.1.1 Summary of the findings

In order to use neural networks for pixel classification, the values of the spectral channels have to be mapped into a set of input neurons. This is the so called "data representation" or "input coding". Different coding techniques have been proposed in the literature (Bischof, 1992). We have adopted the simplest one: one input neuron per input channel. The following are the findings:

Finding 1 The selection of input data sets has a significant effect on classification accuracy. The classification results are sensitive to the choice of input channels.

Back-propagation is easily modified to accommodate more channels or to include spatial and temporal information. The input layer can simply be expanded to accept the

additional data. Expansion of the input greatly increases the computation (on the order of N^2), if the size of the input is doubled, the required computations will be four times as great. Therefore, new channels of information should not be added indiscriminately.

Finding 2 The objective of the network seems to aim at the improvement of the overall and the average accuracy rates.

After a certain number (1000 to 1500) of iterations, the network reaches its best classification results. More iterations do not always improve the accuracy rate. Generally speaking, the overall and the average accuracy of classification get better, however, at the expense of some individual classification accuracies.

Finding 3 An improvement of classification accuracy in the training set does not imply an improvement in classification accuracy of the testing set.

This is so called "over-fitting". A good network architecture should increase both the classification accuracies of the training set and the testing set as the learning proceeds. In our study, the classification accuracy achieved by the network classifiers were nearly 100% for the training set for all classes. The training error, which was defined as the number of misclassified samples in the training data, was reduced very quickly when the number of iterations was less than 300.

Finding 4 There is no general criteria for designing a suitable network architecture. Some prior knowledge about the input data sets and their relevance with the output classes would aid the choice of the input layers.

Finding 5 The variance in the individual classification accuracy was found to be significant which has lead to some criteria on the selection of the parameters. For lithologic mapping the network should be structured in accordance with the importance of each

individual classes. In this study area, the class 2 is of higher concern.

5.2 Conclusions

Lithologic units are rarely homogeneous, and are masked by various amounts of soil, vegetation, colluvium, organic debris and transported surficial materials, which mask and alter the spectral characteristics of the underlying rocks. Soil cover and other surficial deposits can greatly affect the spectral appearance of the ground and can mask structural features and other indicators useful for lithologic identification. By far the most significant problem in dealing with surficial materials is caused by the phenomenon of material transport. Water, ice , and wind are active transporting agents, capable of moving objects and materials of all sizes for great distances. Therefore, surficial deposits of this nature do not reflect the composition of the underlying material or bedrock. These factors make selection of representative training areas extremely difficult, especially in areas of highly dissected terrain. In this study area, these factors have been lessened because of its large non-vegetation area The geologic map produced showed a fairly good correspondence with existing geologic map.

All experiments carried out in these study are based on the concerned parameters, such as the network structure (the number of nodes in the input layer, the nodes in the hidden layer, and the nodes in the output layer), and the training coefficients (momentum factor, learning rate), the training condition (the training sample selection). The initial weights and the choice of activation also have influence on the final solution. For lithologic mapping, the network should be structured in accordance with the importance of each individual class.

Although our experiments were not designed to make recommendations about these parameters, we do gain some in-depth understanding of the neural network performance.

There is not a good way to determine a lower bound for its quantity. The network was done simply by trial and error. Once the decision of activation function is made, with the same network architecture, a number of selections would generate similar classification accuracy.

The neural networks approach is attractive because formal models do not have to be constructed to capture the complexity and variability of the objects to be detected. And the ability to learn is not dependent on assumptions about underlying statistical distributions. Neural network approach is well suited for multi-source image analysis and classification of lithology.

BIBLIOGRAPHY

Abrams, M. J., R. P. Ashley, L.C. Rowan, A.F.H. Goetz; and A.B. Kahle. 1977. Mapping of Hydrothermal Alteration in the Cuprite Mining District, Nevada, Using Aircraft Scanner Images for the Spectral Region 0.46 μm to 2.36 μm. Geology, Vol. 5, pp. 713-718.

An, P., W.M. Moon, and G.F. Bonham-Carter. 1992. On Knowledge-based Approach of Integrating Remote Sensing, Geophysical and Geological Information. IGARSS'92, Houston, Texas, May 1992, pp. 34-38.

An, P. and Chang-Jo. F. Chung. 1994. Neural Network Approach for Geological Mapping: Technical Background and Case Study. Canadian Journal of Remote Sensing, Vol. 20, No. 3. pp. 293-301.

Azimi - Sadjadi, M. R., S. Chaloum, and R. Zoughi. 1993. Terrain Classification in SAR Images using Principal Components Analysis and Neural Networks. IEEE Transactions on Geoscience and Remote Sensing, Vol. 31, No. 2, pp. 511-515.

Benediktsson, J.A., P.H Swain, and O.K. Esroy. 1993. Conjugate-gradient Neural Networks in Classification of Multi-source and Very-high-dimensional Remote Sensing Data. International Journal of Remote Sensing, Vol. 14, No. 15, pp. 2883-2903.

Bischof, H., W. Schneider, and A.J. Pinz. 1992. Multi-spectral Classification of Landsat-Images Using Neural Networks. IEEE Transactions on Geoscience and Remote Sensing, Vol. 30, No. 3, pp. 482-490.

Broome, H. J. 1990. Generation and Interpretation of Geophysical Images with Examples from the Rae Province, Northwestern Canada Shield. Geophysics. Vol. 55, No. 8, pp. 977-997.

Burrough, P. 1986. Principles of Geographical Information Systems for Land Resources Assessment. Oxford, University Press, Oxford, UK.

Charbonneau, B.W., P.G. Killeen, J.M. Carson, G.W. Cameron, and K.A. Richardson. 1976. Significance of Radioelement Concentration Measurements Made by Airborne Gamma Ray Spectrometry Over the Canadian Shield. In Exploration for Uranium Ore Deposits: International Atomic Energy Agency, Vienna, pp. 35-53.

Chen, J., P. Gong, J. Nie, and J.A.R. Blais. 1993. Application of Neural Networks in Forest Ecological Classification. Technical Papers of ASPRE/ACSM Annual Convention, New Orleans, Louisiana, Vol. 3, pp. 65-71.

Chung, C.F., P. Gong, A. N. Rencz, and M. Schau. 1993. Geological Mapping in Melville Peninsula, Northwest Territories, Canada Using Multi-source Remote Sensing and Geophysical Data. IGARSS'93, Tokyo, Japan, August 18-21, 1993, pp. 913-916.

Civco, D. L. 1993. Artificial Neural Networks for Land-Cover Classification and Mapping. International Journal Geographical Information Systems, Vol. 7, No. 2, pp. 173-186.

Crick, F. 1989. The Recent Excitement about Neural Networks, Vol. 337, pp. 129-132.

Daily, M., C. Elachi, T. Farr, and G. Schaber. 1978. Discrimination of Geologic Units in Death Valley using Dual Frequency and Polarization Imaging Radar Data. Geophysical. Research Letters, Vol. 5, pp. 889-892.

Dreyer, P. 1993. Classification of Land Cover Using Optimized Neural Nets on SPOT Data. Photogrammetric Engineering and Remote Sensing, Vol. 59, No. 5, pp. 617-621.

Duval, J.S. 1983. Composite Color Images of Aerial Gamma-Ray Spectrometric Data. Geophysics, Vol. 48, pp. 722-735.

Eliason, T.P., T.J. Donovan, and P.S. Chavez, Jr.. 1983. Integration of Geologic, Geochemical, and Geophysical Data of the Cement oil field, Oklahoma, Using Spatial Array Processing. Geophysics, Vol. 48, N. 10, pp. 1305-1317.

Freeman, J.A. and D.M. Skapura, 1991. Neural Networks: Algorithms, Applications and Programming Techniques, Addison - Wesley, New York.

Gillespie, A.R. and A.B. Kahle. 1977. Construction and Interpretation of A Digital Thermal Inertia Image. Photogrammetric Engineering and Remote Sensing, Vol. 43, pp. 983 - 1000.

Goetz, Alexander F.H. and L. Rowan. 1981. Geologic Remote Sensing. Science, Vol. 211, pp. 781-791.

Gong, P. 1993. Remote Sensing and Image Analysis. Lecture Notes, Department of Geomatics Engineering, The University of Calgary, 200p.

Gong, P., A. Zhang, J. Chen, R.J. Hall, and I.G.W. Corns. 1994. Ecological Land Systems Classification Using Multi-source Data and Neural Networks. GIS' 94 Symposium, Vancouver, B.C. February, 1994.

Gong, P. 1994. Integrated Analysis of Spatial Data from Multiple Sources- An overview. Canadian Journal of Remote Sensing, Vol. 20, No. 4, pp. 349-359.

Gong, P. Integrated Analysis of Spatial Data from Multiple Sources: using Evidential Reasoning and Artificial Neural Network Techniquesfor Geological Mapping. Photogrammetric Engineering and Remote Sensing (in press).

Graham, D.F. and Bonham-Carter G.F. 1993. Airborne Radiometric Data: A Tool for Reconnaissance Geological Mapping Using a GIS. Photogrammetric Engineering and Remote Sensing, Vol. 59, No. 8, pp. 1243-1249.

Guo, Yi, R.O. Hanse, and N. Harthill. 1992. Feature Recognition from Potential Fields Using Neural Networks. SEG (the Society of Exploration Geophysics) Expanded Abstracts. pp. 1 - 5.

Heermann, P.D. and N. Khazenie. 1992. Classification of Multi-spectral Remote Sensing Data Using a Back-Propagation Neural Network. IEEE Transaction on Geoscience and Remote Sensing, Vol. 30, N. 1, pp. 81-88.

Hepner, G.F., T. Logan, N. Ritter, and N. Bryant. 1990. Artificial Neural Network Classification Using Minimal Training Set. Photogrammetric Engineering and Remote Sensing, Vol. 56, No. 4, pp. 469-473.

Higashino, J., B.L. deGreef, and E.H.J. Persoon. 1990. Numerical Analysis and Adaptation Method for Learning Rate of Back-propagation. Proceedings of International Joint Conference on Neural Networks, Vol. 1. pp. 627-630.

Hoffman, R. 1987. The Problem of Extracting the Knowledge of Experts from the Perspective of Experimental Psychology. AI Magazine, Vol. 8, pp. 53-67.

Hunt, G.R. 1977. Spectral Signatures of Particular Minerals in the Visible and Near Infrared. Geophysics, Vol. 42, pp. 501 - 513.

Hutchinson, C.F. 1982. Techniques for Combining Landsat and Ancillary Data for Digital Classification Improvement. Photogrammetric Engineering and Remote Sensing, Vol. 48, No.1, pp.123 - 130.

Kanellopoulos, I. A. Varfis, G.G. Wilkinson and J. Megier. 1991. Classification of Remotely-Sensed Satellite Images Using Multi-layer Perceptron Networks. Artificial Neural Networks, T. Kohonen, K. Kakisara, O. Simula, and J. Kangas (Eds.). Vol. 2, pp. 1067-1074. Amsterdam, North-Holland.

Kanellopoulos, I. A. Varfis, G.G. Wilkinson and J. Megier. 1992. Land Cover Discrimination in SPOT HRV Imagery Using an Artificial Neural Network - A 20-Class Experiment. International Journal of Remote Sensing, Vol. 13, No. 5, pp. 917-924.

Kershaw, C.D. 1987. Discrimination Problems for Satellite Images. International Journal of Remote Sensing, Vol. 8, No. 9, pp. 1377-1383.

Kontoes, C., G.G. Wilkinson, A. Burrill, S. Goffredo, and J. Megier. 1993. An Experimental System for the Integration of GIS Data in Knowledge-based Image Analysis for Remote Sensing of Agriculture. International Journal Geographical Information System, Vol. 7, No. 3, pp. 247-262.

Kung, S.Y., and J. N. Hwang. 1988. An Algebraic Projection Analysis for Optimal Hidden Units Size and Learning Rates in Back-Propagation Learning. Proceedings of International Joint Conference and Neural Networks, Vol. 1, pp. 363-370.

le Cun, Y., W.R. Gardner, H.P. Graf., D. Henderson, R.E. Howard, and W. Hubbard. 1989. Neural Network Recognizer for Hand-Written Zip Code Digits. In Advances in Neural Information Processing Systems 1, D.S. Touretzky, (Eds.) pp. 321-331.

Morgan Kaufmann, San Mateo, California.

Legg, Christopher. 1992. Remote Sensing and Geographic Information Systems: Geological Mapping, Mineral Exploration and Mining. Ellis Horwood Limited, West Sussex, England. 166p.

LeRoy, L.W., D.O. LeRoy, and J.W. Raese, (Eds.), 1977. Subsurface Geology: Petroleum, Mining, Construction, 4th edition, pp. 767-789.

Liu, Z.K. and J.Y. Xiao. 1991. Classification of Remotely-sensed Image Data using Artificial Neural Networks. Int. J. Remote Sensing, Vol. 12, No. 11, pp. 2433 - 2438.

Lorenzetti, E.A. 1992. Predicting Lithology from V_p and V_s Using Neural Networks. SEG (the Society of Exploration Geophysics) Expanded Abstracts. pp. 14 - 17.

Lyon, R.J.P. 1965. Analysis of Rocks by Spectral Infrared Emission (8 to 25 microns). Economic Geology, Vol. 60. pp. 715 - 736.

Maren, A., C. Harston and R. Pap. 1990. Handbook of Neural Computing Applications. Academic Press Inc.

McCormack, M.D., David E. Zaucha, and Dennis W. Dushek. 1993. First-Break Refraction Event Picking and Seismic Data Trace Editing Using Neural Networks. Geophysics, Vol. 58. No. 1, 67-78.

Moon, Wooil M. 1990. Integration of Geophysical and Geological Data Using Evidential Belief Function. IEEE Transaction on Geoscience and Remote Sensing, Vol. 28, No. 4, pp. 711-720.

Murata, Noboru, Shuji Yoshizawa, and Shun-ichi Amari. 1994. Network Information

Criterion-Determining the Number of Hidden Units for an Artificial Neural Network Model. Transactions on Neural Networks, Vol. 5, No. 6, November 1994. pp. 865- 872.

Pao, Y. 1989. Adaptive Pattern Recognition and Neural Networks. Addison and Wesley: New York.

Parikh, J., J.S. DePonte, M. Damodaran, A. Karageorgiou; P. Podaras. 1991. Comparison of Back-propagation Neural Networks and Statistical Techniques for Analysis of Geological Features in Landsat Imagery, SPIE, Vol. 1469. pp. 526-538.

PCI, 1992. PACE C Programming's Manual Version 5.0a. Richmond Hill, Ontario, Canada

Peddle, D.R., G.M. Foody, A. Zhang, S.E. Franklin, E.F. LeDrew. 1994. Multi-source Image Classification II: An Empirical comparison of Evidential Reasoning and Neural Network Approaches. Canadian Journal. of Remote Sensing, Vol. 20, No. 4, pp. 396-407.

Richards, J.A., D.A. Landgrebe, and P.H. Swain. 1982. A Means for Utilizing Ancillary Information in Multispectral Classification. Remote Sensing of Environment, Vol. 12, pp. 463 - 477.

Rowan, L.C., A.G.H. Goetz, and R.P. Ashley. 1977. Discrimination of Hydrothermally Altered and Unaltered Rocks in Visible and Near-infrared Multi-spectral Images. Geophysics, Vol. 42. pp. 522-535.

Rumelhart, D.E., G.E. Hintom, and R.J. Williams. 1986. Learning Internal Representations by Error Propagation., In Parallel Distributed Processing -

Explorations in the Microstructure of Cognition, Vol. 1, Edited by D.E. Rumelhart, J.L. McClelland, The MIT Press: Massachusetts, pp. 318-362.

Salu, Y., and J. Tilton. 1993. Classification of Multi-spectral Image Data by Binary Diamond Neural Network and by Non-parametric, Pixel-by-Pixel Methods. IEEE Transaction on Geoscience and Remote Sensing, Vol. 31, No. 3, pp. 606-616.

Schau, M., A.Rencz, L. Dredge, C.F. Chung and L. Chorlton, 1993. Digital Data Atlas, Hall Lake Area, Melville Peninsula, Northwest Territories. Geological Survey of Canada, Open File 2786.

Schau, Mikkel, Dredge, L. Rencz, A.N., and C.F. Chung. 1993. Effect of Surficial Geolgy on Remotely Sensed Data of Bedrock in Northern Canada; IGARRS' 93, Tokyo, Vol. 4, pp. 2113-2115.

Shafer, G. A. 1976. Mathematical Theory of Evidence. Princeton Univ. Press, Princeton, NJ.

Siegal, B.S. and A.R. Gillespie (Eds.). 1980. Remote Sensing in Geology. John Wiley & Sons, Inc. New York.

Srinivasan, A. and J.A. Richards, 1990. Knowledge-based Techniques for Multi-source Classification. International Journal of Remote Sensing, Vol. 11, No. 3, pp. 505-525.

Sui, D.Z. 1994. Recent Applications of Neural Networks for Spatial Data Handling. Canadian Journal of Remote Sensing, Vol. 20, No. 4, pp. 368-378.

Swain, P.H. and S. M. Davis, 1978. Remote Sensing: the Quantitative Approach. McGraw Hill, New York, USA. 396p.

Taranik, J. V. 1983. Geological Remote Sensing: Landsat, Shuttle, and Beyond. Exploration and Economics of the Petroleum Industry. Matthew Bender & Co., pp. 185-197.

Werbos, P. 1974. Beyond Regression: New Tools for Prediction and Analysis in Behavioral Sciences. Ph.D. Dissertation, Harvard University, Cambridge, Massachusetts.

Yin, Yongyuan and Xiaomei Xu. 1991. Applying Neural Net Technology for Multi-objective Land Use Planning. J. of Environmental Management, Vol. 32, pp. 349-356.

Zhuang, X., B.A. Engel, D.FLozano-Garcia, R.N.Fernandez, and C.J. Johannsen. 1994. Optimization of Training Data Required for Neuro-Classification. International Journal of Remote Sensing, Vol. 15, No. 16, pp. 3271-3277.

Zhu, A-Xing and L. Band. 1994. A Knowledge-Based Approach to Data Integration for Soil Mapping. Canadian Journal of Remote Sensing, Vol. 20, No. 4, pp. 408-418.