CHAPTER 1

AN OVERVIEW OF THE THESIS

1.1 INTRODUCTION

Remote sensing involves collection of data from a certain region using sensors located remotely, for example, as on a satellite. The purpose is to study the characteristics of the region and also classify it into different groups. Typically remote sensing data is collected by a set of sensors and in different frequency bands of light, because of the different reflective properties of the areas under study. Such data is called multispectral data whose resolution is dependent on various factors like the location of the sensors with respect to the region and the wavelength of the light used for collecting the multispectral data and also the signal processing performed on the received data.

1.2 OBJECTIVE

In this thesis we address some issues in the pattern recognition tasks of clustering, classification and prediction in remote sensing. Some of the objectives of remote sensing are:

i) To cluster the data into different groups using unsupervised learning methods.

ii) To arrive at a classification technique using data from known regions in order to classify unknown regions.

iii) To predict the trends following the temporal patterns of the remote sensing data.

The objective of clustering is to separate a set of regions into consistent clusters or groups so that members of any one group differ from another according to a chosen criterion (Anderberg M.R. 1973) such as, for example, the sum of squared error criterion. In general, clustering
procedures can be divided into two types, namely, partitional and hierarchical (Robert J. Schalkoff 1992). The partitional clustering technique aims at partitioning the patterns into a prespecified number of clusters where the desired number of clusters and initial cluster centres are specified apriori. The hierarchical clustering technique forms as many singleton clusters as the number of the pattern samples. A vital problem in cluster analysis (Dubes R. et al 1980) is to determine the correct number of clusters present in a data set. Each clustering algorithm imposes a particular structure on the data set such as hyperspherical, elliptical and concentric, rather than uncover the true structure present in it. Hence it is necessary to validate a clustering solution (Bailey T.A. et al 1982) and to identify the stable partition in a data set.

To identify similar groups from a remote sensing data, it is necessary to cluster a large number of patterns. An iterative scheme is normally adopted and the iterative nature of the clustering algorithm requires a significant amount of computation (Venkateswarlu N.B. et al 1992). The splitting and lumping of clusters (Kittler J. et al 1988) have to be verified at each stage to identify the correct number of regions. Thus determining the number of clusters and reducing computation for clustering are important issues in the pattern clustering task.

Pattern classification methods for remote sensing are mainly based on statistical methods like the minimum distance classifier, maximum likelihood classifier, etc., (Bendiktsson J.A. et al 1990). Here the classification is performed under Gaussian assumption and any pixel is assigned to the most likelihood class. In the above methods, the classification of a data sample requires the evaluation of decision function for each class being considered which makes the process relatively slow. In general, remote sensing data is subject to noise such as path radiance or back scattering effect (Bischof H. et al 1992). As a result, there can be an increase in the probability of error in classification. In spectral pattern recognition, spectral bands have varied levels of importance. Also the
information content of an important band can be watered down by the presence of insignificant data in other bands (King Sun Fu 1971). This problem is likely to become more complex when a large number of correlated spectral bands are used and as a result, the accuracy of the classifier begins to fall off (Marudhachalam M. et al 1987). Thus analyzing the cost of misclassification and reducing the probability of error in classification are the important issues in pattern classification.

Sutton R.S. (1988) has shown clearly that learning to predict is very important since it determines the outcome of the future event based on the current observations of the state of the environment. For example, through experience, one might learn to predict from particular cloud formations, whether there will be rain or from certain climatological conditions, whether there will be changes in oceanographic data and so on. Many of these problems involve temporal sequences of observations and predictions. For each prediction, the overall error measure has to match the expected value of the subsequent outcome and not the actual one. For making prediction for a fixed time later, such as the weekly prediction of Sea Surface Temperature (SST) from oceanographic data, both the temporal credit assignment and structural credit assignment issues have to be addressed. That is, it should be decided how each output of a temporal sequence of outputs needs to be changed and also which part of a network should be changed so as to influence the output in order to reduce the overall error.

1.3 MAIN CONTRIBUTIONS OF THE THESIS

The main contributions of this thesis are as follows:

(i) The partial sum approach is incorporated in the Self-Organizing Feature Map (SOFM) algorithm so that the amount of computation involved in clustering a large number of patterns which use Euclidean distance measure is reduced. Convex-combination method is used to improve the accuracy of the above algorithm and the weight update equations are derived in terms of adaption gain term.
(ii) The probability of error in classifying an unknown region in the study area is reduced by calculating the saliency factor expressed in terms of the sensitivity measure in Optimal Brain Damage (OBD). This is examined by computing the partial derivative of the criterion function with respect to each weight in the network using the chain rule.

(iii) To identify a stable partition and reduce the overlap between clusters in a data set, a regularity index, based on a generic bootstrap cluster stability statistic and the proposed performance indices, is defined.

(iv) Prediction at a fixed time is computed in the proposed Modified Temporal Difference (MTD) method by a multilayer connectionist network which is a non-linear function of each observation and weight vector.

(v) For the weekly prediction of SST, recursive equations consisting of overlapping sequences of interrelated predictions at different times are formed. The above factors describe the SST pattern characteristics of the summer and winter monsoon in the three oceans.

1.4 OUTLINE OF STUDIES AND RESULTS

This section discusses in detail the studies made so far and presents the results obtained. First it describes an approach to validate a clustering solution using bootstrap technique. Secondly, a learning rule using Self-Organizing Feature Map (SOFM) to reduce the computational requirements of a clustering algorithm is proposed. In the above two cases, experiments have been conducted on the 2-D data sets, real data and remote sensing data. Also an analysis of the cost of misclassification and classification accuracy of various patterns in the study area is made. Finally, a combination of backpropagation and TD methods for the 1°x1° weekly average prediction of SST from the oceanographic data is arrived at.
An approach to the problem of determining the number of clusters present in a data set is explained using bootstrap technique. Multiple fake data sets are generated by duplicating certain samples and deleting some of them from the original data set. Two performance indices are proposed by Meena K. et al (1992) for both partitional and hierarchical clustering techniques. For a partitional clustering technique, a K-cluster partition is generated using K-means algorithm. The value of the number of clusters which minimizes this performance index is taken as an estimate of the true structure of a data set. This index is good in identifying hyperspherical shaped clusters. However, when the clusters are elongated and concentric, the above index cannot be used for validating a clustering solution. That is because the compactness criterion is not suitable for these cluster shapes and they can be uncovered only by using the hierarchical clustering technique.

The performance of these indices is tested over different cluster shapes of synthetic, real (IRIS) (Fisher R.A. 1936) and remote sensing data. It has been observed that the smallest values of the performance indices are used to determine the number of clusters in a data set. The regularity index defined, based on the performance indices is applied to all the above data sets considered. As a result the overlap between clusters, resulting in intermixed clusters can be identified. An analysis of the regularity index on the artificial, real and remote sensing data is carried out. It has been shown that the values of this index based on the proposed performance indices identify the number of categories for the remote sensing data. The experimental results are compared with those obtained by using the Davies and Bouldin measure (Davies D.L. et al 1979). It has been observed that the proposed performance indices achieve the minimum for all values of the variance of a data set and can estimate its cluster tendency.

ISODATA algorithm is discussed as an adaptive clustering algorithm in which the splitting and lumping thresholds are calculated from the actual data. They identify the correct groups or clusters only if they are linearly separable in nature. But in reality, they are not always
linearly separable. A clustering algorithm called SOM_NET using SOFM (Meena K. et al 1995) significantly reduces computation for clustering. The partial sum approach is used in the above two algorithms to speed up the process of clustering. The speedup is derived in terms of the number of samples, groups, iterations and accuracy. Experiments have been conducted on the 2-D data sets, IRIS data and TM Landsat data BRAZIL. The remote sensing image BRAZIL is fed as a test input. The image obtained by using the feature map algorithm is compared with that obtained by using the histogram peak technique of cluster analysis (Meena K. et al 1993). The experimental results show that the clustering algorithm using feature map produces a more realistic and noiseless image than the histogram peak technique (Thomas M. Lillesand et al 1979).

In order to examine the neighbourhood relation between the pixels and decrease the probability of error, training areas are selected using SOFM and are classified using backpropagation algorithm (Meena K. et al 1995). The classification results for various number of nodes in the hidden layer are obtained. As remote sensing data includes noise or uncertainty of information in the spectral data, erroneous pixels are deleted and new training areas are made. By using the new ones, the network is retrained to obtain the final solution.

To enhance the accuracy of the classifier, it is necessary to eliminate redundant weights from the layers of a Multilayer Perceptron (Narendra K.S. et al 1991). This is achieved by employing the pruning technique (Ehud D. Karnin 1990) between the hidden layer and the output layer. Using this technique, a list of sensitivity numbers is generated and the weights with the smallest sensitivity are deleted to enhance the classification accuracy. From the data analysis, it has been observed that the backpropagation algorithm along with pruning produces better classification between training patterns than the maximum likelihood classifier.
A particular TD procedure is introduced by relating it to a classical supervised learning procedure, namely, the Widrow-Hoff rule (Widrow B. et al 1960). However, TD procedures cannot be used in making prediction for a fixed amount of time later, since each prediction is of a different event and there is no clear desired relationship between them. Backpropagation cannot be directly applied to the problem of prediction, since it needs to know the outcome for the backpropagation of error. But it is not available till a certain amount of time elapses resulting in large storage requirements.

A new scheme is designed wherein the advantages of backpropagation and TD methods have been combined in a single model which caters to fixed step problems. Conventionally the weightage to successive predictions in TD methods is achieved by an exponential decay function (Narendran R. et al 1993). However for a fixed step prediction, such as, for example, prediction of SST, the exponential recency is not appropriate, since it has a large negative slope and hence an alternative recency which leads to better prediction is proposed.

The stochastic, backpropagation, TD and MTD methods are used to predict SST values in the Arabian Sea, the Bay of Bengal and the Central Indian Ocean. One of the most important parameters in the climatic changes and oceanographic studies (Narendranath A. et al 1989) namely the SST in the three oceans is evaluated. The root mean squared error values are calculated between the actual SST values and those obtained by using the above methods. It is observed from the experimental results that the statistical methods suffer from their lack of adaptability and unsuitability while backpropagation gets entangled into local minima problems and hence MTD methods yield better results for the weekly prediction of SST.
1.5 SUMMARY

In this thesis, some issues in clustering, classification and prediction in remote sensing have been addressed. Some of the ANN techniques have been proposed to address them. It has been demonstrated that a learning rule using Self-Organizing Feature Map reduces the computation for clustering. Also the cost of misclassification and probability of error in classification is reduced by the backpropagation algorithm along with the pruning technique. It is found that the Modified Temporal Difference method yields better prediction for a fixed time later than other methods. The next chapter will give a survey of work done in the related areas and highlight some of the issues for study in this thesis.