Synopsis

We are living in an era dominated by Information Technology. The convergence of computers and communication networks have changed the way we live. Data generation and transmission on networks is happening in a seamless way. Newer concepts like E-commerce, E-governance, E-learning etc. have emerged as alternatives to traditional shopping, administration and learning. E-mail has become the fastest means of communication of information at highly affordable cost. All these changes are gradually transforming our traditional society into information rich digital multimedia society. Needless to say, in such a society every one has access to data, anytime, anywhere.

As a consequence, there is an exponential increase in the data transactions almost in all sectors. Storage and management of such large volume of data is therefore a major challenge in today's Information technology era. There has been a significant shift in the way the data are analysed in today's context. Earlier we have been using only mano valued data for varieties of decision making. But in today's world, the so called classical data is no longer sufficient for decision making but we are required to gather and analyze a more generalized form of data known as Symbolic data.

Symbolic data is a combination of quantitative and qualitative descriptions of objects. Thus symbolic data has interval type of data, lists, qualitative description of objects etc. It is very clear that what ever applications we see today, it represent symbolic data. For example, Medical data, Census data, University data, Credit card data etc. Such being the case, there is a need for methodologies which can take symbolic data and produce useful information from it for decision making. It is to be noted that conventional methods used with classical data no longer hold good for symbolic data.

Symbolic data always represents more realistic data and therefore possesses its own structure and internal variation. Symbolic objects are extensions of classical data types. Symbolic objects are defined by a logical conjunction of events linking values and variables in which the variables can take one or more values and all the objects need not be defined on the same variables. Symbolic data could be classified into either of type assertion, hoard or synthetic. A more detailed illustration of the different categories is given in Appendix-A.
Given a symbolic data set, the next step is to conduct statistical analysis as appropriate. Unlike classical data for which a number of methodologies exist, symbolic statistical analysis are new and available methodologies are only few in number.

Gowda and Diday (1990) first introduced a basic dissimilarity measure for Boolean symbolic objects. Later Ichino and Yaguchi (1994) using cartesian operators developed a generalized Minkowsky distance measure and computed clusters in symbolic data. Clustering of symbolic objects have been investigated by many researchers and most of the new methodologies have emerged from the works of Gowda and Diday in the early 90's.

It is seen that all the clustering procedures are based on the similarity or dissimilarity among the sample data. Some novel similarity measures for symbolic data were proposed by T.V.Ravi (1995), Srikanta prakash (1998) and Dinesh (2000). It is observed that they tried to improve the computation of similarity measures by different approaches. A chronology of different contributions for clustering of symbolic objects through different similarity dissimilarity measures is given below.

- Gowda and Diday (1991-a) have proposed a symbolic clustering methodology using a new similarity measure.

- Gowda and Diday (1991-b) have proposed a new dissimilarity measure for symbolic objects.

- Gowda and Diday (1992) have proposed a new similarity measure and a hierarchical agglomerative clustering algorithm.

- Gowda and Ravi (1995) have proposed a clustering procedure using both similarity and dissimilarity measures.

- Gowda and Dinesh (2000) have proposed a Fuzzy approach to compute similarity between samples of symbolic data.

A critical review of literature has opened up the following research issues.

1. Gowda and Diday (1992) distance measure has been extensively used for computing similarity indices $S_p$, $S_s$ and $S_c$ for both quantitative and qualitative type of data. Computation of $S_p$, $S_s$ and $S_c$ are based on some empirical formulae which are different for qualitative data and quantitative data. A critical analysis on the computation of span length
and the parameter $u_k$ indicate certain drawbacks with interval type of data and hence does not give correct similarity measures. Therefore we propose to introduce modifications to the Gowda—Diday Distance measure to overcome the drawbacks. This is discussed in detail in Chapter 2.

2. Clustering of Symbolic objects primarily requires the computation of similarity or dissimilarity measures from the given symbolic data set. This in turn depend upon the computation of $S_p$, $S_s$ and $S_c$ which employ different empirical formulae. We propose an alternative method that does not require the use of $S_p$, $S_s$ and $S_c$. This new method is based on the Binary representation of the symbolic data and use of Hamming distance to calculate the similarity indices. The details of this contribution are discussed in Chapter 3.

3. Clustering and classification are the two dominant applications of Neural Networks. There exists many such applications on classical data. However literature review indicate very few applications on symbolic data. Moreover the binary description of the symbolic objects proposed in the previous chapter fit well for training the neural networks. Thus the symbolic objects are viewed as vectors in the input space. We use Kohonens Self Organizing Feature Map (SOFM) to partition the input space into regions where each region represent the cluster. This method totally eliminates the conventional similarity measure computation and use of clustering algorithms. In addition this neural network generalizes incoming symbolic objects not seen in training and assigns them to right clusters. This work is discussed in detail in Chapter 4.

4. Symbolic data often has several features and hence usually has a higher dimension. Large number of such data are generated for every activity. It is therefore essential to classify such large dimension data into meaningful classes for effective decision making. Since Supervised Neural networks are well suited to perform classification, we propose to use a neural network that does not require its architecture pre defined and learn fast. The idea is to employ an incremental learning neural network that builds its architecture during learning and learn the given input data at a fast rate. We have implemented Platt’s Resource Allocating Neural Network (RAN) and present its learning and generalization characteristics for different symbolic data sets. This is discussed in Chapter 5.
5. All the above investigations are carried out on three benchmark data sets, Fat Oil, Microcomputer and Microprocessor data. We have also considered a practical university data set to verify the performance of the proposed methods. The clusters obtained by the proposed methods are compared with Gowda-Diday and Ichino methods. Further, the clusters obtained are also validated using Huebert $\Gamma$ statistics.