Chapter 7

Conclusions

This chapter concludes the thesis by summarizing the findings and the contributions of this thesis and presents several directions for future research.

7.1 Summary of Contributions

The wealth of information embedded in large centralized or distributed databases have generated a large amount of machine-readable data in the form of files and databases. To analyze the huge amount of collected data, the interdisciplinary field of Knowledge Discovery in Databases (KDD) has emerged. KDD comprises of many steps namely, data selection, data preprocessing, data transformation, data mining and data interpretation and evaluation. Data Mining forms a core activity in KDD. Data mining applies efficient algorithms to extract interesting patterns and regularities from large amounts of data where the data can be stored in information repositories. Clustering in data mining is a useful technique for discovering interesting data distribution and pattern in the underlying data. The limitations of the traditional hard computing paradigm in solving many real-life problems dealing with uncertainty or imprecise data, present a great opportunity for the application of Soft
7.1. Summary of Contributions

Computing (SC) techniques to data clustering. Therefore, it is necessary to investigate the principle of clustering from a soft computing perspective to device efficient methods that meet the specific requirement in the context of data mining. This thesis contributes to the field of data clustering by introducing an efficient clustering algorithm and a cluster validity assessment index. A summary of these contributions of the thesis are summarized as follows:

Chapter 2 presents a brief introduction to the subject of soft computing methodologies, viz., fuzzy set theory, ANN and GAs algorithms, with a special emphasis on some of the most common and successful hybrid soft computing techniques along with the related research in the field.

In Chapter 3, the Knowledge Discovery in Databases (KDD) process is introduced and described. Furthermore, a brief introduction to Data Mining and its key tasks is presented. The Clustering techniques and approaches are discussed in detail. Since, clustering is an unsupervised method and there is no a priori indication of the actual number of clusters present in the data set, there is a need for some kind of clustering results validation. A survey of most popular validity measures available in the literature is presented in this chapter as these are the primary focus of the thesis.

Chapter 4 provides the performance analysis of different clustering techniques from soft computing perspective. In particular, results of the experimental evaluation of the performance of the clustering algorithm based on hard and soft computing techniques on different benchmark numerical data sets is presented. It is concluded from the experimental results obtained, that the major weakness of the GA based FCM algorithm is the CPU time it requires. The dominant cost is that of calculating the fitness value for each population member for each generation. The GA based FCM approach is about 20 times slower than the \( k \)-means and KSOM algorithms and almost 18 times slower than the FCM algorithm. Experimental results show that the best validation result among all the tested indices is obtained by the PBM and the
7.1. Summary of Contributions

*PBMF* validity indices for different data sets. It is noted that the *PBMF* and *XBF* validity indices give the best cluster validation when used with the FCM and GA based FCM algorithms respectively with different data sets. Further, it is noted that in spite of the sensitivity of the soft computing based clustering algorithms to the selection of the various parameters values to suit the applications, great potentials are offered by soft computing approaches in solving the clustering problems and finding global optimal solutions. Experimental results indicate that clustering problems can be solved effectively by using soft computing methodologies in combination and there is ample space for improvement for efficiency of clustering algorithms using hybrid techniques.

The FCM algorithm and its derivatives are iterative algorithms. In addition, their calculation often involves a huge number of membership matrices and candidate cluster matrices. It is a computationally intensive method. Further, the performance of the FCM depends on the selection of the initial cluster centers. If a good set of initial cluster centers is chosen, the algorithm may take less iteration to find the actual cluster centers. An efficient *psFuzzy c*-means clustering algorithm is proposed in Chapter 5. The proposed algorithm is a modified *psFCM* called the *pshFCM* that finds better cluster centers for a given data sets as compared to the cluster centers obtained by the *psFCM* and the FCM algorithms. The proposed clustering algorithm consists of three phases. In Phase I, the *pshFCM* algorithm divides the data set into several unit blocks. The centroid of unit blocks replaces the input patterns and forms a new data set. By doing so, a given data set is drastically reduced to a simplified data set containing the centroids of the original patterns. In Phase II, random selection is applied to find the candidate set of cluster centers of the simplified data set. The *PBMF* cluster validity index is used to evaluate the fitness of the randomly selected candidate set of centers. The overall computation time is reduced since the number of calculations of the norm (distance metric) is reduced. In Phase III, the FCM algorithm is applied to find the cluster centers in the original data set with the optimal set of centers obtained in Phase II. The number of calculations of the norm is further reduced. The number of iterations needed to converge in Phase III of the
pshFCM is also less than the number of convergence iterations of the FCM algorithm for all the data sets used in the experiments. To illustrate the applicability and the efficiency of the proposed algorithm it is applied to the benchmark data sets and the experimental results are presented.

A problem faced in clustering is to decide the optimal number of clusters that fits a data set. To obtain the optimal number of clusters \( c \), an evaluation methodology is required to evaluate the clustering results obtained from the clustering algorithms. Cluster validity indices have been used to evaluate the fitness of clusters produced by the clustering algorithms. In chapter 6, the limiting behavior of some of the existing clustering validity indices have been investigated and then a new \( v_{SBT} \) clustering validity assessment index and fuzzified \( v_{SBT} \) called \( v_{FSBT} \) has been proposed for eliminating the monotonically increasing tendency when the number of clusters becomes very large and approaches the number of data points. Limiting behavior of the proposed new \( v_{FSBT} \) index has been investigated. The proposed validity index determines the optimal number of clusters for fuzzy partitions with fuzzy \( c \)-means algorithm and is the most reliable of the indices tested in indicating the correct number of clusters. Tests of the performance of the proposed index on various data sets demonstrated its effectiveness.

As shown by both theoretical and experimental results, the proposed efficient pshFCM clustering algorithm and the proposed \( v_{SBT} \) validity index along with the fuzzified \( v_{SBT}, v_{FSBT} \), validity index give improved performance for different benchmark data sets. The research work presented in this thesis would therefore be useful to a large community of clustering practitioners working in different domains.

### 7.2 Scope for Future Research

- Although, the performance analysis of different clustering techniques based on soft computing techniques, *viz.*, fuzzy set theory, ANN and GA is made in Chapter 4. Other soft computing based optimization techniques like Particle Swarm Optimization (PSO), Ant Colony Optimization etc. along with the hybrid approaches can be investigated for data clustering, which may give
better clustering results.

- The proposed clustering algorithm \textit{pshFCM} in Chapter 5 uses the FCM clustering algorithm. Future research can investigate the use of other efficient clustering algorithms such as \textit{k}-harmonic means (KHM) and its variants. KHM has a soft membership function and a varying weight function. To illustrate the applicability and the efficiency of the proposed \textit{pshFCM} algorithm it is applied to the benchmark numerical data sets. To illustrate its wider applicability, the application of the proposed \textit{pshFCM} clustering algorithm to other application domains, e.g., bioinformatics and image segmentation needs to be carried out. Further, the \textit{k-d} tree method used to split the given data set takes the number of unit blocks (the number of splits) as an input from the user. Some heuristics needs to be developed to automatically split a given data set depending on the dimension of the feature vector and the statistical information of the patterns in the data sets. Furthermore, investigations needs to be carried out to understand the relationship between the number of split with the \textit{k-d} tree method and the convergence iterations (execution performance).

- The experimental results of the proposed \textit{vFSBT} clustering validity assessment index shows superior performance to other indices and is the most reliable of the validity indices tested. However, all the indices tested in the present work including the proposed index are dependent on the performance the underlying FCM clustering algorithm. Future research can investigate the use of other efficient clustering algorithms such as KHM as it is less sensitive to initial conditions. An extensive study needs to be undertaken from both theoretical and empirical point of view with regard to the proposed \textit{vFSBT} validity index. Further, experimentation is required to evaluate the validity of the new index in a noisy environment.