Chapter 9
Epilogue

9.0 Background

In this thesis, we emphasized on the fundamental research in devising new proximity measures and the associated clustering algorithms to cluster symbolic objects described by features of type interval and multivalued. In this chapter, we summarize the different models investigated during the course of this research work. In the sequel, we list out the major contributions accomplished through this research followed by the scope for future research.

9.1 Summary

We have made a successful attempt in exploring novel symbolic proximity measures useful for clustering symbolic objects, which are described by unconventional features of type interval and multivalued. In this direction, we have devised novel symbolic proximity measures (both symbolic similarity and dissimilarity measures) to estimate the degree of proximity between two symbolic objects. The proposed proximity measures approximate the proximity values by the use of interval type and multivalued type data and in addition, unlike conventional proximity measures, they are not necessarily symmetric. We suggested modifications to the conventional clustering techniques to make them work on the proposed unconventional proximity matrices. It has also been established theoretically that unlike conventional measures where similarity is a compliment of dissimilarity, the proposed similarity measure is not another aspect of the proposed dissimilarity measure. A more generalized
proximity measure called tendency measure is introduced by integrating both similarity and dissimilarity measures.

Further, several unsupervised learning schemes by the use of the proposed proximity measures have been developed to learn unlabelled symbolic objects. We presented learning schemes both on agglomerative and divisive techniques. As the agglomerative based techniques work on intermediate composite symbolic objects, it is not a reverse process of the divisive based learning schemes. The novel way of expressing the proximity values between two symbolic objects by the use of multivalued data type has been exploited for the purpose of dimensionality reduction through newly designed feature selection scheme. The designed dimensionality reduction technique has two stages, feature subset selection stage, where the features which are significant in discriminating the clusters present in a data set are selected and subset refinement stage where, the feature which are prominent in improving the cohesion with in clusters get added up into the selected feature subset. The importance of another property, non symmetricity, of the proposed symbolic proximity measures in simulating the reality during clustering symbolic objects has also been brought out. Two incremental clustering approaches namely pulling leader and attractive leader oriented approaches have been explored. The importance of the non symmetricity of the proposed proximity measures has also been corroborated by the use of the concept of mutually neibourhood value in clustering.

Nevertheless, two interesting applications of the proposed methodologies are also presented. All the proposed models have been theoretically characterized and experimentally validated thoroughly on different data sets taken from different domains. The proposed methodologies have been compared and contrasted for their performance superiorities with several other existing contemporary models.

In brief, the following section gives the list of the major contributions of the thesis.

9.2 Contributions

- Exploration of new symbolic similarity measure, which is not necessarily symmetric to estimate the degree of similarity between two symbolic objects described by features of type interval and multivalued.
• Exploration of new symbolic dissimilarity measure, which is not necessarily symmetric to estimate the degree of dissimilarity between two symbolic objects described by features of type interval and multivalued.

• Methods of approximating the proposed similarity and dissimilarity values by interval and multivalued data types.

• Modifications to make conventional clustering techniques work on the proposed unconventional proximity matrices, which are of type interval and multivalued in addition to being non symmetric.

• Introduction of a new proximity measure called tendency measure, which is a more generalized measure by integrating both similarity and dissimilarity measures.

• Theoretical establishment of the fact that the proposed similarity measure is not another aspect of the proposed dissimilarity measure.

• Development of unsupervised learning schemes to support unsupervised learning of unlabelled symbolic objects.

• Novel method of describing a composite symbolic object representing a cluster of symbolic objects.

• Exploitation of the proposed multivalued proximity measures for the purpose of dimensionality reduction through a method of feature selection, which works in proximity space instead of feature space.

• Introduction of two novel types of incremental clustering approaches called pulling leader oriented and attractive leader oriented approaches to bring out the importance of non symmetricity of a proximity measure in simulating the reality during clustering.

• Mutual neighborhood value based approaches for clustering by the use of non symmetric proximity measures.

• Application of the proposed methodologies in shape classification and numeral classification.
9.3 Scope for Future Research.

The methods proposed in this thesis have created many openings for the research community to develop various interesting approaches useful for clustering symbolic objects. Indeed, quite lot improvements could be done on even the proposed methodologies for their better performances.

We have presented in chapter 2 and chapter 3, the concepts of mutual average similarity and mutual average dissimilarity based on two parameters $\alpha$ and $\beta$. It shall be an interesting future research work to find the best optimal values for the parameters adaptively for a given set of symbolic objects while clustering. In case of MSV, MDV and MTV based clustering algorithms, we have taken into account only the magnitude of the vectors. However, devising a clustering algorithm by making use of even the direction of the vectors shall give an added dimension to symbolic clustering.

In chapter 4, while presenting unsupervised learning schemes based on the divisive clustering approach, we recommended removing all inconsistent edges, in case of existence of two or more inconsistent edges, at the same stage. This may miss out, some perhaps, important intermediate stages in clustering. Thus, looking for a method of ranking the existence of two or more inconsistent edges based on some criterion for their removal shall improve the performance of the proposed learning schemes by allowing them to stop at a better stage.

We have presented a novel way of describing a composite object representing a cluster of symbolic objects. However, we have not used the same way of describing the composite objects obtained at intermediate stages during clustering, as the proposed similarity and dissimilarity measures cannot be applied on such type of symbolic objects. The newly proposed way of describing a composite object, can even be applied while aggregating a large data set into a single symbolic object. Therefore, development of new proximity measures to work on such symbolic objects opens up a new avenue in the field of symbolic data analysis.

The multivalued proximity measure has been exploited for reducing the dimension of the feature space through a feature selection scheme, which works in the proximity
space instead of feature space. Analogously, one could think of a feature reduction technique through transformation, which works at proximity space instead of feature space.

Incorporation of a method of selecting best leaders would enhance the performance of the incremental clustering approaches presented in chapter 6. In case many leaders are selected at a time, then allowing them to grow parallelly perhaps simulates the reality in clustering in a more general way.

We have presented many clustering algorithms to cluster a set of symbolic objects. Devising an approach which combines all the proposed clustering methods together into single approach that produces more accurate clusters for a given data set by the use of the probability that two objects go together due to various clustering techniques.