ABSTRACT

Spatial Data Mining i.e., discovery of interesting, implicit knowledge in spatial databases, is an important task for understanding and using spatial data as effective knowledge bases. It is widely recognized that combining multiple models typically provides superior results compared to using a single, well-tuned model. The idea of combining object partitions without accessing the original objects' features leads us to a general knowledge reuse framework termed as cluster ensembles. In this research work, we have brought out that by using a guided approach in combining the outputs of the various clusterers, we can reduce the intensive distance metrics computations and also generate robust clusters. We have proposed hybrid and layered cluster merging approach for fusion of spatial clusterings and used it in our three-phase clustering combination technique. The major challenge in fusion of ensembles is creation and manipulation of voting matrix or proximity matrix which is in the order of $n^2$, where $n$ is the number of data points. This is very expensive both in time and space factors, with respect to spatial data sets. Instead, in our method, we have eliminated this expensive voting matrix. We have computed a symmetric clusterer compatibility matrix of order $(m \times m)$, where ‘$m$’ is the number of clusterers and $m << n$, using the cumulative similarity between the clusters of the clusterers. This matrix is used for identifying two clusterers, which if considered for fusion initially, will provide increased information gain. In the subsequent layers, compatible clusterers are identified for the partially fused clusterers, so that this acquired knowledge will be used for further fusion. As we travel down the layered merge, for every layer, we calculate a factor called Degree of Agreement (DoA), based on the established clusterings. The apparent advantage is that we can prune the data sets after every $(m-1)/2$ layers, using the gained knowledge in previous layer. This helps in faster convergence compared to the existing cluster aggregation techniques. The destination of data points with less DoA (nebulous pool elements), will not be decided based on a sensitive parameter like ‘threshold’. The concluding clusters’ data points will provide the
guidance for the reallocation of these elements. This approach has increased the clustering accuracy by 5-10% when compared to our other models. The hybrid approach of combining the information theory, co-association and voting approaches of fusion has resulted in generation of compact clusters as indicated by the increase in inter cluster density as much as 30%. We have given a supplementary solution to predict the missing value problem, based on the ensemble outputs. We have also addressed the scalability issue of high dimensional data by using fusion of homogenous ensembles, derived from vertical slices. This has been achieved with no loss in accuracy rate in the aggregated clusters. The inputs for our work are the clusters derived from various clustering techniques, which have used the well established similarity metrics during their clustering process. We felt that there is no need of intensive similarity measures computation again and approximate measures are sufficient. Hence, we have proposed two similarity metrics for measuring the distance between the clusterings using minimum bounding circle and partition entropy. For data sets with comprehensible underlying data structure, we have shown that minimum bounding circle metrics perform better than the entropy based distances. Based on the nature, number and size of the clusterings, we have proposed four ensemble fusion models Inherent Voting Depth First Merge (IVDFM), Cyclic Merge (CycM), Slice and Dice Ensemble Merge (SDEM) and Nebulous pool Merge (NebM). The correctness and robustness of the proposed cluster ensemble algorithms are demonstrated by usage of various cluster validity metrics like accuracy, misclassification rate, cluster purity, Dunn indices, intra cluster density and inter cluster density. Synthetic data sets and real world data sets available in University of California Irvine’s data repository has been used for performing the experiments. Our proposed effective and efficient clusterer combining algorithms based on layered and hybrid approach, along with the results on synthetic as well as real data-sets show that our hybrid inherent voting spatial cluster ensembles can generate high quality, robust and compact clusters, apart from providing efficient tool for distributed clustering. We have also provided a solution based on homogenous ensembles to handle high dimensional data.
Keywords- Clustering ensembles, Spatial Data Mining, Degree of Agreement, Cluster Compatibility matrix. Cluster validity metrics.