Efficient Fusion of Cluster Ensembles Using Inherent Voting

Anandhi R J
Research Scholar, Dept. of CSE,
Dr MGR University, Chennai
Phone: 9845 705705(rjanandhi@hotmail.com)

Natarajan Subramanyam
Professor, Department of ISE, PESIT, Bangalore
Phone: 9945280225(snatarajan44@gmail.com)

***Abstract*—Discovering interesting, implicit knowledge and general relationships in geographic information databases is very important to understand and to use the spatial data. Spatial Clustering has been recognized as a primary data mining method for knowledge discovery in spatial databases. In this paper, we have analyzed an efficient method for the fusion of the outputs of the various clusterers, with less computing .We have discussed our proposed slice and dice cluster ensemble merging technique (SDEM) for spatial datasets and used it in our three-phase clustering combination technique in this paper. Voting procedure is normally used to assign labels for the clusters and resolving the correspondence problem, but we have eliminated by usage of Degree of Agreement Vector. Another common problem in any cluster ensembles is the computation of voting matrix which is in the order of n^2, where n is the number of data points, which is very expensive with respect to spatial datasets. In our method, as we travel down the layered merge, we calculate degree of agreement (DOA) factor, based on the count of agreed clusterers. Using the updated DOA at every layer, the movement of unresolved, unsettled data elements will be handled at much reduced the computational cost. Added advantage of this approach is the reuse of the gained knowledge in previous layers, thereby yielding better cluster accuracy and robustness**

Keywords- Data mining, Spatial data mining, Clustering ensembles, Consensus function, Degree of Agreement.

I. INTRODUCTION

With a variety of applications, large amounts of spatial and related non-spatial data are collected and stored in Geographic Information Databases. Spatial Data Mining, (i.e., discovering interesting, implicit knowledge and general relationships in large spatial databases) is an important task for the understanding and the usage of these spatial data. The importance of spatial data in our daily lives is rapidly increasing and so are the challenges and demands on the research and commercial communities to address the different facets of spatial data. [1] In these communities, spatial data have generated tremendous interest over the last decade.

With the rapid growth in size and number of available databases in commercial, industrial, administrative and other applications, it is necessary and interesting to examine how to extract knowledge automatically from huge amount of data. Knowledge discovery in databases, or Data Mining, is the effort to understand, analyze, and eventually make use of huge volume of data available. Through the extraction of knowledge in databases, large databases will serve as a rich, reliable source for knowledge generation and verification, the discovered knowledge can be applied to information management, query processing, decision-making, process control and many other applications. Therefore, data mining has been considered as one of the most important topics in databases by many database researchers.

Spatial data describes information related to the space occupied by objects. It consists of 2D or 3D points, polygons etc. or points in some d-dimensional feature space. It can be either discrete or continuous. Discrete spatial data might be a single point in multi-dimensional space while continuous spatial data spans a region of space. This data might consist of medical images or map regions and it can be managed through spatial databases [2, 3].

Clustering [3], one of the very important functionality of data mining, is to group analogous elements in a data set in accordance with its similarity such that elements in each cluster are similar, while elements from different clusters are dissimilar. It doesn't require the class label information about the data set because it is inherently a data-driven approach. So, the most interesting and well developed method of manipulating and cleaning spatial data in order to prepare it for spatial data mining analysis is by clustering that has been recognized as a primary data mining method for knowledge discovery in spatial database [4-7].

Clustering algorithms are used to partition unlabeled data into groups or clusters. Clustering data is often time consuming. This is especially true of iterative clustering algorithms such as the k-means family or EM. As larger unlabeled datasets become available, the scalability of clustering algorithms becomes more important. There are now unlabeled datasets which vastly exceed the size of a typical single memory [9].

When spatial data are visualized, the attribute values defined numerically have to be classified into some class divisions. In this process, there exists the risk of leading us to miss-judgment or biased understanding, since much information of the original data may be lost, according to the classification method adopted. Therefore, the classification method of spatial data from the viewpoint of information-statistics was proposed as a new classification method based on minimization of information loss. This method is a sort of smoothing technique neglecting the characteristics of spatial data distribution. However, it is necessary to consider the spatial distribution of attributes, to adequately visualize data accompanied with information of "spatial distribution". [14].

978-1-4244-4711-4/09/$25.00 ©2009 IEEE

Cluster analysis encompasses a number of different algorithms and methods for grouping objects of similar kind into respective categories [15]. The attractiveness of cluster analysis is its ability to find categories or clusters directly from the given data. Many clustering approaches and algorithms have been developed and successfully applied to many applications. However, when a classical clustering technique, such as the k-means, is applied to geographically located data, without using the spatial information, the resulting partition has often a "chaotic" appearance over the geographic space, i.e., clusters look dispersed, and reflect only poorly any eventual underlying spatial structure. This is because classical clustering algorithms often make assumptions (e.g., independent, identical distributions) which violate Tobler's first law of geography: everything is related to everything else but nearby things are more related than distant things [16]. In other words, the values of attributes of nearby spatial objects tend to systematically affect each other. This is why, we decided that fusing the outputs of different clustering algorithms, especially for spatial data would produce robust clusters. And also we have used Tobler's first law in designing a simple yet efficient consensus function.

Clustering fusion is the integration of results from various clustering algorithms using a consensus function to yield stable results. Clustering fusion approaches are receiving increasing attention for their capability of improving clustering performance. At present, the usual operational mechanism for clustering fusion is the combining of clusterer outputs. One tool for such combining or consolidation of results from a portfolio of individual clustering results is a cluster ensemble [2].

Subsets of data can be clustered in such away that each data subset fits in memory and finally the clustering solution of all subsets can be merged. This enables extremely large datasets to be clustered. Sometimes, data is physically distributed and centrally pooling the data might not be feasible due to privacy issues and cost. Thus, merging clustering solutions from distributed sites is required. Moreover, iterative clustering algorithms are sensitive to initialization and produce different partitions for the same data with different initializations. Combining multiple partitions may provide a robust and stable solution. It was shown to be useful in a variety of contexts such as "Quality and Robustness" [8], "Knowledge Reuse" [9, 10], and "Distributed Computing" [11].

The rest of the paper is organized as follows. The related work is in section 2. The proposed slice and Dice ensembles clustering merge technique is in section 3. In section 4, we present experimental test platform and results with discussion. Finally, we conclude with a summary and some directions of future research in section 5.

II. Related Work in Clustering Ensembles

Clustering ensemble is the method to combine several runs of different clustering algorithms to get an optimal partition of the original dataset. Given dataset X = {x1 x2,.. ,xn}, a cluster ensemble is a set of clustering solutions, represented as P = P1,P2,..Pr., where r is the ensemble size, i.e. the number of clusterings in the ensemble. Clustering-Ensemble approach first gets the result of M clusterers, then set up a common understanding function to fuse each vector and get the labeled vector in the end. The goal of cluster ensemble is to combine the clustering results of multiple clustering algorithms to obtain better quality and robust clustering results. Even though many clustering algorithms have been developed, not much work is done in cluster ensemble in data mining and machine learning community. Fred and Jan 2002, used co-association matrix to form the final partition. They applied a hierarchical (single-link) clustering to the co-association matrix [12].Zeng, Tang, Garcia-Frias and GAO 2002, proposed an adaptive meta-clustering approach for combining different clustering results by using a distance matrix [13].

Strethl and Ghosh [9], proposed a hypergraph-partitioned approach to combine different clustering results by treating each cluster in an individual clustering algorithm as a hyperedge. They introduced three efficient heuristics to solve the cluster ensemble problem. All algorithms approach the problem by first transforming the set of clusterings into a hypergraph representation. Cluster-based Similarity Partitioning Algorithm (CSPA) uses relationship between objects in the same cluster for establishing a measure of pair-wise similarity. This induced similarity measure is then used to re-cluster the objects, yielding a combined clustering. In Hyper Graph Partitioning Algorithm (HGPA) the maximum mutual information objective is approximated with a constrained minimum cut objective. Essentially, the cluster ensemble problem is posed as a partitioning problem of a suitably defined hypergraph where hyperedges represent clusters. In their Meta-CLustering Algorithm (MCLA), the objective of integration is viewed as a cluster correspondence problem. Essentially, groups of clusters (meta-clusters) have to be identified and consolidated.

Kai Kang, Hua-Xiang Zhang, Ying Fan [19] formulated the process of cooperation between component clusterers, and proposed a novel cluster ensemble learning technique based on dynamic cooperating (DCEA). The approach mainly concerned how the component clusterers fully cooperate in the process of training component clusterers. This method firstly aligns the cluster centroids discovered by different component clusterers which works by measuring the similarity between the cluster centroids through compute the distances, and then adjusts the aligned cluster centroids by a dynamic momentum term, and then the next iteration is going on until the termination rule is satisfied.

Muna Al-Razgan, Carlotta Domeniconi [20] proposed a soft feature selection procedure (called LAC) that assigns weights to features according to the local correlations of data along each dimension. Dimensions along which data are loosely correlated receive a small weight, which has the effect of elongating distances along that dimension. Features along which data are strongly correlated receive a large weight, which has the effect of constricting distances along that dimension. Thus the learned weights perform a directional local reshaping of distances which allows a better separation of clusters, and therefore the discovery of different patterns in different subspaces of the original input space. The clustering result of LAC depends on the number of clusters k to be

discovered in the data and h factor that controls the strength of the incentive to cluster on more features.

III. Proposed Slice and Dice Ensemble Merge Algorithm

A. *Defination*s

- Matching groups set, $MG_{[slice][Dice]}$: A set containing clusters from different clusterers with highest cardinality in intersection set ,i.e.,, $MG_{Slice[ij}$ refers to matching pairs of i^{th} clusterer's j^{th} cluster. For instance, $MGSlice_{[1]][1]}$ means merging groups is obtained by merging first slice clusterer and the dice is with second cluster elements.
- Degree Of Agreement Factor: Ratio of the index of the Slicing level to the total number of clusterers and is indicated as DOA.
- $DOA_{Th.:}$ User assigned value, normally will be set as 50% of the number of clusterers

B. *The Problem Specification*

Given r groupings with the q-th grouping x(q) having k(q) clusters, a consensus function Fx is defined as a function, mapping a set of clusterings to an integrated clustering. The optimal combined clustering should share the most information with the original clusterings. This shared information between clusterings is normally measured using mutual information, a symmetric measure to quantify the statistical information shared between two distributions [3,18].

C. *The Proposed Inherent voting solution*

In this section we discuss our proposed slice and Dice Ensemble Merge algorithm (SDEM: Slice and Dice Ensemble Merge) for spatial datasets. At the first level, B heterogeneous ensembles are run against the same spatial data set to generate partitioning results. Individual partitions in each ensemble are sequentially generated.

At the second level, these clustering results are combined in sliced pairs, called matching groups set, MGmk, using cardinality of similarity set between the core elements of the clusterers. The usage of similarity between core points resolves the label naming issues very easily and elegantly. This also prevents a lot of computational costs.

When the merging happens, for each data point the degree of agreement (DOA) is calculated. This factor, DOA is the ratio of the index of the merging level to the total number of clusterers. And also the DOA value will be cumulative till it reaches the threshold level DOATh. . Once the DOA of any data point crosses the threshold, it can be affirmed to belong to a particular cluster result. Thus, the normal voting procedure with huge voting matrix, to confirm the majority does not arise at all in our method. Once the data element is confirmed to a cluster, it will not participate in further computations. Hence, the computational cost is also hugely reduced. This approach will be very useful when we are handling spatial data, because as per Tobler's First law of geography, everything is related to everything else but nearby things are more related than distant things. The number of data points which keep oscillating between the clusters, will be the only challenge. All the related points will be settled at the early stage of the iterations and thereby contributing a lot towards reducing computational costs. At the third level, the unsettled data objects i.e., data objects with less than or equal to DOATh will be handled. In case of even number of clusterers, we will have border elements which can be resolved by using likelihood merge with the final clusters. Data points below the threshold will be identified as Outliers/Noise.

This final layer merge with the earlier combined clusters will yield the robust combined result. This approach is not computationally intensive, as we tend to use the first law of geography in merging in slices along with elimination of voting matrix. The three levels of the technique are applied sequentially. They do not interfere with each other, but they just receive the results from the previous levels. No feedback process happens, and the algorithm terminates after the completion of all slices.

D. *Pseudo Code of SDEM algorithm*

Step1: Form k clusters each, from Dataset D using m clusterers.

Step2: Set DOA_Increment Factor ←1/ m

Step3:Identify merging groups for Slicei merge,MG_1

Step4: For every pair in the merging group Sets, $MG_{i\,k,}$
{
Construct Dice Vectors containing first data
vector and append it with unit DOA vector
Update the DOA vector using the second pair.
}

Step 5:if(DOA of data elements > $DOA_{Th),}$
place element in $Final\_Kluster_{Slicei}$
else place them in $Orphan_{Slicei}$

Step6: Compute MG_{i+1} using $Final\_Kluster_{Slicei}$ $Orphan_{Slicei}$

Step7: Repeat steps 4 thro 6 till all m slices are exhausted.

Step8:Classify any left over Orphans as noise.
Step 9:Return the robust clusters & Noise elements.

TABLE I. Pseudo Code of SDEM Algorithm

IV. Test Platform and Results

In our test platform, we have used both homogeneous as well as heterogeneous ensembles. In the later case, we have created the ensemble clusters using K-means, PAM, FCM and DBSCAN algorithms. K-means is a very simple and very powerful iterative technique to partition a data set into k disjoint clusters. DBSCAN method performs well with attribute data and performs fairly well with spatial data. Partitioning around medoids (PAM) is mostly preferred for its scalability and hence usefulness in Spatial data. We have added Fuzzy C means (FCM) as one of the clusterer, so that we get a robust partition in the end result. Hence these four clustering

techniques along with different cluster sizes form the input for our merge technique.

Most of the ensemble methods, have sampling techniques in selecting the data for experimental platform, but this heuristics results in losing some inherent data clusters, thereby reducing the quality of clusters. We have tried to avoid sampling and involve the whole dataset in SDEM algorithm. This is feasible because, only the matching pairs are taken for merging during the slice cycle. We used the Clustering accuracy (CA) to measure the accuracy of an ensemble as the agreement between the ensemble partition and the "true" partition. The classification accuracy is commonly used for evaluating clustering results. To guarantee the best re-labeling of the clusters, the proportion of correctly labeled objects in the data set is calculated as CA for the partition.

RapidMiner is one of the world-wide leading open-source data mining solution due to the combination of its leading-edge technologies and its functional range. The datasets are run through RapidMiner and is used as the benchmark for calculating accuracy. The test results with the IRIS dataset, Wine dataset, WDBC and Ionosphere dataset (Courtesy: UCI data repository) is promising and shows better cluster accuracy when compared to other non ensembling techniques, as well as homogenous cluster ensembles. Our ensemble fusion technique was compared with the approach of Alexander Strehl[3] on the grounds of space complexity. It was found that our technique was independent of the number of clusterers involved, whereas the approach of Alexander Strehl[3] had the space complexity increase exponentially when the number of clusterers increase. The graph shown below would provide more information. The test results with the IRIS dataset, Wine dataset, Half rings and Spiral dataset (Courtesy: UCI data repository) is promising and shows better cluster accuracy when compared to other non ensembling techniques. as well as homogenous cluster ensembles.

Our SDEM method has proved to give industry standard accuracy when we compared our results with commercially available clustering software, yet provides with better efficiency. When we tested our algorithm with the 'Wine' dataset (Courtesy :UCI data repository), we got same results obtained by running it on commercial clustering software 'RapidMiner'. Most other datasets also confirmed that the ensembling approach has not resulted in identifying wrong clusters. The current approaches consisted of two stages: Ensemble preparation and Consensus function. The ensemble preparation stage requires building up of matrices of dimension m*(n+k) for each clusterer, where m is the number of data objects, n is the number of attributes and k being the number of clusters, hence for the ensemble preparation stage the matrix dimension will be in the order of mc*(n+k) where c is the number of clusterers and the Consensus function stage requires building matrix of size m*(n+c).Whereas in our DOA vector has no dependency on 'c' and hence is scalable, and has the space complexity of the order m*(n+1) i.e. m*n, where m is the number of data objects, n is the number of attributes .

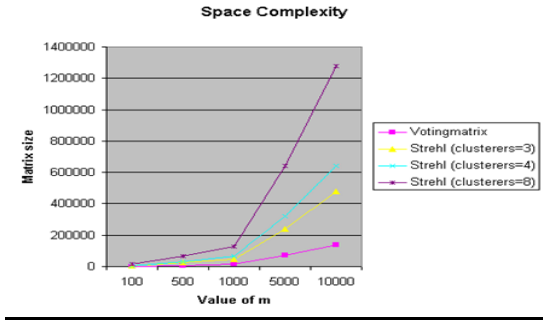

Figure 1. Space Complexity comparison in winedata

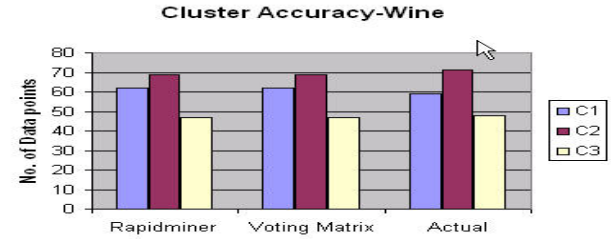

Figure 2. Cluster accuracy comparison in Wine data

V. CONCLUSION

In this paper we addressed the re-labeling problem found in general in most of cluster ensembles problem and provided an effective algorithm to solve it. The cluster ensemble is a very general framework that enables a wide range of applications. We applied the proposed layered cluster merging technique on spatial databases. The main issue in spatial databases is the cardinality of data points and also the increased dimensions. Most of the existing Ensemble algorithms have to generate voting matrix of at least an order of n2. When n is very huge and is also a common factor in spatial datasets, this restriction is a very big bottleneck in obtaining robust clusters in reasonable time and high accuracy. Our algorithm has resolved the re labeling using layered merging based on first law of geography. The normal voting procedure with huge voting matrix, to confirm the majority does not arise at all in our method. Once the data element is confirmed to a cluster, it will not participate in further computations. Hence, the computational cost is also hugely reduced. We use ensemble methods to get better cluster accuracy as different clustering results give different results for the same dataset. Ensemble methods use results of a number of runs of same or different clustering algorithms to give clustering results with better accuracy. Privacy is an important aspect in today's world, as customer data cannot be shared, ensemble methods are used to classify the data and is then provided to the analysts for various applications.

The key goal of spatial data mining is to automate knowledge discovery process. It is important to note that in this study, it has been assumed that, the user has a good knowledge of data and of the hierarchies used in the mining process. The crucial input of deciding the value of k, still affects the quality of the resultant clusters. Domain specific Apiori knowledge can

be used as guidance for deciding the value k. We feel that semi supervised clustering using the domain knowledge could improve the quality of the mined clusters. We have used heterogeneous clusterers for our testing but it can be tested with more new combinations of spatial clustering algorithms as base clusteres. This will ensure exploring more natural clusters

REFERENCES

[1] Han, J., Kamber, M., and Tung, A., 2001a, Spatial Clustering Methods in Data Mining: A Survey'',in Miller, H., and Han, J., eds., Geographic Data Mining and Knowledge Discovery. Taylor and Francis..

[2] Su-lan Zhai1,Bin Luo1 Yu-tang Guo : "Fuzzy Clustering Ensemble Based on Dual Boosting" , Fourth International Conference on Fuzzy Systems and Knowledge Discovery 07

[3] Samet, Hanan.: "Spatial Data Models and Query Processing". In Modern Databases Systems: The object model, Interoperability, and Beyond. Addison Wesley/ ACM Press, 1994,Reading, MA.

[4] Zhang, J. 2004. Polygon-based Spatial clustering and its application in watershed study. MS Thesis, University of Nebraska-Lincoln, December 2004.

[5] Matheus C.J., Chan P.K, and Piatetsky-Shapiro G, "Systems for Knowledge Discovery in Databases", IEEE Transactions on Knowledge and Data Engineering 5(6), pp. 903-913, 1993.

[6] M.Ester, H. Kriegel, J. Sander, X. Xu. Clustering for Mining in Large Spatial Databases. Special Issue on Data Mining, KI-Journal Tech Publishin, Vol.1, 98

[7] K.Koperski, J.Han, J. Adhikasy. Spatial Data Mining: Progress and Challenges. Survey Paper.

[8] Ng R.T., and Han J., "Efficient and Effective Clustering Methods for Spatial Data Mining", Proc. 20th Int. Conf. on Very Large DataBases, 144-155, Santiago, Chile, 1994.

[9] A.L.N. Fred and A.K. Jain, "Robust data clustering", in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR, USA, 2003.

[10] A.Strehl, J.Ghosh, "Cluster ensembles - a knowledge reuse framework for combining multiple partitions", Journal of Machine Learning Research, 3: 583-618, 2002.

[11] A.Strehl, J.Ghosh, "Cluster ensembles- a knowledge reuse framework for combining partitionings", in: Proc. Of 11th National Conference On Artificial Intelligence, NCAI, Edmonton, Alberta, Canada, pp.93-98, 2002.

[12] B.H. Park and H. Kargupta, "Distributed Data Mining", In The Handbook of Data Mining, Ed. Nong Ye, Lawrence Erlbaum Associates, 2003

[13] A.L.N. Fred and A.K. Jain, "Data Clustering using Evidence Accumulation", In Proc. of the 16th International Conference on Pattern Recognition, ICPR 2002, Quebec City

[14] Zeng, Y., Tang, J., Garcia-Frias, J. and Gao, G.R., "An Adaptive Meta-Clustering Approach: Combining The Information From Different Clustering Results", CSB2002 IEEE Computer Society Bioinformatics Conference Proceeding.

[15] Toshihiro Osaragi, "Spatial Clustering Method for Geographic Data", UCl Workig Papers Series, paper41, Jan 2002.

[16] Jain, A.K, Murty, M.N., and Flynn P.J:Data clustering: a review. ACM Computing Surveys,31, 3, 264-323

[17] Tobler, W.R. : Cellular Geography, Philosophy in Geography. Gale and Olsson, Eds.,Dordrecht, Reidel.

[18] Thomas M. Cover and Joy A. Thomas. Elements of Information Theory. Wiley, 1991.

[19] Kai Kang, Hua-Xiang Zhang, Ying Fan, "A Novel Clusterer Ensemble Algorithm Based on Dynamic Cooperation", IEEE Fifth International Conference on Fuzzy Systems and Knowledge Discovery 2008.

[20] Muna Al-Razgan, Carlotta Domeniconi, "Weighted Clustering Ensembles".