2009 International Symposium on Computing, Communication, and Control (ISCCC 2009)
Proc .of CSIT vol.1 (2011) © (2011) IACSIT Press, Singapore

Efficient Consensus Function for Spatial Cluster Ensembles: An heuristic layered approach

Anandhi R J [1] and Natarajan Subramanyam [2+]

[1] Research Scholar, Dept. of Computer Science & Engg, Dr MGR University, Chennai, India
[2] Professor, Department of Information Science & Engg, PESIT, Bangalore, India

Abstract. 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 computations.We have discussed our proposed layered merging technique for spatial datasets and used it in our clustering combination technique in this paper. Voting procedure is normally used to assign labels for the clusters and resolving the correspondence problem when partitions are made from different clusters. We have eliminated the need for such voting using the matching groups. Based on the cardinality and the set intersections, most likely cluster groups across different clusters are grouped as matching pairs. When more than 50 percent of the clusterers agree upon the groupings, they are resolved into the final partition. 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.

1. Introduction

Spatial data mining i.e., discovering interesting, implicit knowledge and general relationships in large spatial databases is a demanding field since huge amounts of spatial and related non-spatial data are collected and stored in various applications, ranging from remote sensing to geographical information systems (GIS) and computer cartographies. The collected data far exceed people's ability to analyze it. Spatial Data Mining 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. The huge amount of spatial data and the complexity of spatial data types and spatial access methods make the efficiency of spatial data mining algorithms an important research challenge. This amount of data has surpassed the capacity of traditional data analysis methods. Standard GIS's are very slow in geographic analysis, especially on larger (national and state scale) data intensive business applications. Therefore, it is important to explore new methods for mining knowledge related to both non-spatial and spatial objects in large databases.

Statistical analysis [1] however, may not be a very efficient way to analyze large amounts of data that is spatially interrelated. Geographic data consist of spatial objects and non-spatial descriptions of these objects. Spatial data can be described using two properties: geometric and topological. For example, geometric properties include spatial location, area, perimeter etc., whereas topological properties include adjacency

[+] Corresponding author. Tel.: + (9945 280225,9845 705705); fax: +(91 80 2573 0551).
E-mail address: (rjanandhi@hotmail.com, snatarajan44@gmail.com).

(object A is a neighbour of object B) and inclusion (object A is inside object B). Spatial data are often organized by spatial indexing structures and retrieved by spatial access methods. These distinct features of spatial databases create problems when knowledge discovery methods for relational data are applied to spatial data. The data mining methods for spatial data should use spatial relationships and spatial access methods to effectively and efficiently discover knowledge. Therefore, the development of new algorithms should take into account spatial properties of the data and make use of spatial data structures and spatial query processing [2, 3].

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 [8,9]. Cluster analysis encompasses a number of different algorithms and methods for grouping objects of similar kind into respective categories [15-17]. Spatial Clustering [2], 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].

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 is more related than distant things [18]. 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. 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 [3].

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

The rest of the paper is organized as follows. The related work is in section 2. The proposed layered purging and merging 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.

2. Related Work in Clustering Ensembles

Cluster analysis encompasses a number of different algorithms and methods for grouping objects of similar kind into respective categories. 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. *Spatial clustering* is a process of grouping a set of spatial objects into clusters. For instance, spatial clustering is used to determine the “hot spots" in crime analysis and disease tracking. Hot spot analysis discovers unusually dense event clusters across time and space. Many criminal justice agencies are exploring the benefits provided by computer technologies to identify crime hot spots in order to take preventive strategies such as deploying saturation patrols in hot spot areas [2].

Clustering ensemble [3] 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 number of clusterings in the ensemble also called as the ensemble size. Clustering-Ensemble approach first gets the result of M clusterers, then sets up a common understanding function to fuse each vector and get the labelled 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.

Strethl and Ghosh [9-11], 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. Fred and Jan [10], used co-association matrix to form the final partition. They applied a hierarchical (single-link) clustering to the co-association matrix. Zeng, Tang, Garcia-Frias and GAO [15], proposed an adaptive meta-clustering approach for combining different clustering results by using a distance matrix. 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. Muna Al-Razgan, Carlotta Domeniconi [19] proposed a soft feature selection procedure (called LAC) that assigns weights to features according to the local correlations of data along each dimension.

3. Proposed layered merging approach

3.1. Definitions

- Layered set, LS[i,j]: A set containing maximum number of common cluster elements from different clusterers at layer i, where i -> 1 to m-1; j -> 1 to k. (m being the total number of clusterers each with k clusters)
- Degree Of Agreement Factor: Ratio of the index of the current merging level to the total number of clusterers and is indicated as DOA.[levelIndex]
- DOATh.: User assigned value, normally will be set as 50% of the number of clusterers or as per the requirement of the application.

3.2. The Proposed consensus function

In this section we discuss our proposed layered merge and purge algorithm for spatial datasets. Initially, B heterogeneous ensembles are run against the same spatial data set to generate partitioning results. Individual partitions in each ensemble are sequentially generated. This algorithm basically works with two

During the merge phase, layered sets LS[i,j] are formed by using maximum cardinality similarity set between the data elements of the clustering results. The usage of similarity between core points is a normally used approach, which is very sensitive to the presence of outliers. Instead we have found that the usage of cardinality of set intersection, serves as both as identifying the cluster partitions as well as resolves the label naming issues very easily and elegantly. When the merging happens in between the layered sets, 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.

During the purging phase, the confirmed data elements, i.e., their individual cumulative DOA factor value is equal to or more than DOATh, will not be permitted to participate in further computations. They will be considered as elements that have already got the majority, hence will be purged thereby increasing both the computational efficiency and space efficiency. 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. Once both the merge and purge phases are finished, the unsettled data objects i.e., data objects with less than or equal to DOA^{Th} will have to 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 layers along with elimination of voting matrix. The two phases 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 m-1 layers.

4. Test Platform and Results

In our test platform, we have used both homogeneous as well as heterogeneous ensembles (Fig 3). 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.

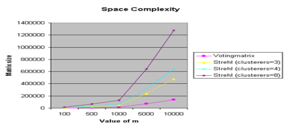

Fig. 1: Comparison of Space Complexity

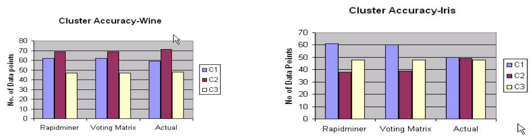

Fig. 2: Cluster accuracy

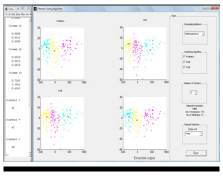

Fig. 3: View of heterogeneous ensemble

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 LMP algorithm. This is feasible because, only the matching pairs are taken for merging. 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. The clustered datasets available in UCI data repository is used as a standard bench mark data for testing the correctness and robustness of our technique. The ratio of correctly labelled objects to the total number of data points in the data set is calculated as clustering accuracy of the ensemble results. One more validation test that we have used is verification of our results with the results from Rapid Miner, one of the world-wide leading open-source data mining solutions. The datasets are run through RapidMiner and is used as the benchmark for calculating accuracy.(Fig:2) The test results with the IRIS dataset, Wine dataset, WDBC and Ionosphere dataset (Courtesy: UCI data repository) are 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 in fig 1, provides this 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 method has proved to give industry standard accuracy when compared with commercially available clustering software, yet 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.

5. Conclusion

In this paper we addressed the re-labelling 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 labelling 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. 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 clusterers. This will ensure exploring more natural clusters

6. References

[1] J. Zhang, A. Samal and L.-K. Soh, Polygon-Based Spatial Clustering, Proceedings in 8th International Conference on GeoComputation,Aug 05

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

[3] Su-lan Zhai,Bin, Guo: "Fuzzy Clustering Ensemble Based on Dual Boosting", Intl. Conf. on Fuzzy Systems and Knowledge Discovery 07

[4] Samet Hanan.:"Spatial Data Models and Query Processing",ACM Press:Modern Databases Systems:bject model, Interoperability, and Beyond.

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

[6] Matheus C.J., Chan P.K, Piatetsky-Shapiro G, "Systems for Knowledge Discovery in Databases", IEEE Transactions on Knowledge and Data Engineering 93.

[7] 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

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

[9] Ng R.T., and Han J., "Efficient and Effective Clustering Methods for Spatial Data Mining", Proc. 20th Int. Conf. on Very Large DataBases, 94.

[10] 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.

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

[12] 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.

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

[14] A.L.N. Fred and A.K. Jain, “Data Clustering using Evidence Accumulation”, In Proc. of International Conference on Pattern Recognition, 02.

[15] 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.

[16] Toshihiro Osaragi, “Spatial Clustering Method for Geographic Data”, UCl Workig Papers Series, paper41, Jan 2002.

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

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

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