World Journal of Advanced Research and Reviews

eISSN: 2581-9615 CODEN (USA): WJARAI

Cross Ref DOI: 10.30574/wjarr

Journal homepage: https://wjarr.com/

(Review Article)

Voting-based efficient cluster ensemble fusion

Urvashi Soni [1, *] and Sunita Dwivedi [2]

[1] *Department of Computer Application, SAGE University, Indore,Madhya Pradesh,India.*

[2] *Makhanlal Chaturvedi Rashtriya Patrakarita Avam Sanchar Vishwavidyalaya university in Bhopal, Madhya Pradesh, India.*

World Journal of Advanced Research and Reviews, 2022, 16(03), 308–314

Publication history: Received on 22 October 2022; revised on 02 December 2022; accepted on 05 December 2022

Article DOI: https://doi.org/10.30574/wjarr.2022.16.3.1280

Abstract

Voting-based consensus clustering is a subset of consensus techniques that makes clear the cluster label mismatch issue. Finding the best relabeling for a given partition in relation to a reference partition is known as the voting problem. As a weighted bipartite matching problem, it is frequently formulated. We propose a more generic formulation of the voting problem as a regression problem with various and multiple-input variables in this work. We demonstrate how a recently developed cumulative voting system is an exception that corresponds to a linear regression technique. We employ a randomised ensemble creation method in which an excess of clusters are randomly chosen for each ensemble partition. In order to extract the consensus clustering from the combined ensemble representation and to calculate the number of clusters, we use an information-theoretic approach. Together with bipartite matching and cumulative voting, we use it. We provide empirical data demonstrating significant enhancements in clustering stability, estimation of the real number of clusters, and accuracy of clustering based on cumulative voting. The gains are made in comparison to recent consensus algorithms as well as bipartite matching-based consensus algorithms, which struggle with the selected ensemble generation technique.

Keywords: Voting; Consensus Clustering; Ensemble Generation; Co-Occurrence Matric~~es~~

1. Introduction

In data analysis, where the main goal is to find natural groups for unlabelled data objects, data clustering is a crucial problem. In the fields of pattern recognition, machine learning, applied statistics, communications, and information theory, the issue has been researched for many years. Data mining, text mining, bio-informatics, image analysis and segmentation, data compression, and data classification are just a few of the fields of application where it is used.

Data clustering is a well-known unsupervised learning task that is notoriously challenging. There has been interest in employing consensus clustering techniques to solve the clustering problem during the past few years. Recent advancements in the field of merging numerous classifiers, where the concept of voting can be easily implemented, are the main driving force behind the creation of cluster ensemble approaches. Due to the absence of generally agreed-upon cluster labels, the consensus of data partitions presents a more difficult problem than the case of classifier ensembles. Additionally, it is known that choosing the ideal number of clusters is a challenging challenge in general.

Finding a consensus partition that best summarises the ensemble is the aim of reconciling an ensemble of clustering solutions in order to produce a clustering solution that is more accurate and stable than the ensemble's individual components. Naturally, the effectiveness of the consensus method in combining the generated ensemble and the ensemble generation technique have a significant impact on the quality of a consensus clustering.

[*] Corresponding author: Urvashi Soni
Department of Computer Application, SAGE University, Indore,Madhya Pradesh,India.

A consensus clustering approach, also known as a consensus function, typically builds an aggregate representation of the ensemble and uses it as the foundation for obtaining a consensus partition. The cluster label correspondence problem is typically avoided by the construction of ensemble representations [12–14]. These representations include category feature spaces [14], pairwise co-association or co-occurrence matric~~es~~ [13], hyper-graphs, and meta-graphs [12]. Voting-based approaches, on the other hand, look for the best relabeling of the ensemble partitions in an effort to draw direct comparisons with consensus methods for multiple classifiers [15–20]. Relabeling involves matching the symbolic cluster labels on the various ensemble divisions. Relabeling makes it possible to aggregate the cluster-label assignments of the objects into an ensemble representation made up of a central (or median) aggregated partition.

Voting-based approaches represent a distinctive class of consensus procedures when viewed from the perspective of relabeling. A common solution to the computationally challenging problem of relabeling and aggregating an ensemble simultaneously is to pairwise relabeling each ensemble partition with respect to a reference partition [15,21]. The vote difficulty refers to this pairwise relabeling. The voting problem is frequently phrased as a weighted bipartite matching problem in the cluster ensemble literature [15–19]. In this study, we generalise the voting problem to be a multi-response regression issue. We demonstrate that this more generic formulation is a special case of bipartite matching and a more modern cumulative voting system presented in [20]. The former is a more confined least squares problem, whereas the later relates to fitting a linear model by least squares estimation (a.k.a. linear sum assignment problem LSAP). The research specifically explains how the ensemble generation technique has a significant impact on the effectiveness of these various unique circumstances.

The aggregated representation determined by voting is a soft partition. In general, the term "soft" is used in the literature to describe either a partition obtained using a statistical model-based clustering algorithm that maximises the likelihood function [22,23], where ujq reflects the uncertainty about the associated classification of each data object [24,25], or to describe a partition obtained using a clustering algorithm that optimises a fuzzy objective function [24,25]. In this study, the aggregated partition is conceptualised as a statistically soft partition. It is specifically obtained by averaging the probabilities of cluster-label assignment.

The number of clusters for each partition may vary and is typically more than the number of actual or desired clusters when using the ensemble generation strategies outlined in [13,26]. This method is seen as a distributional (statistical) representation of the ensemble because it results in an aggregated ensemble partition that is created by relabeling and averaging cluster-label assignments. According to this perspective, the issue of finding the best compression for this statistical distribution so that the most amount of information is saved becomes the issue of obtaining the best consensus clustering. Here, each of the bipartite matching and cumulative voting systems is taken into account when we extract the consensus clustering using the information-bottleneck based methodology presented in [20]. When cumulative voting is combined with the utilised ensemble generation technique, experimental results show significantly more accurate consensus solutions and better estimations of the number of clusters.

The voting procedures described here can, in theory, be used with either hard or soft ensembles. Bipartite matching schemes have been used with both hard [17–19] and soft ensembles [15, 16], and a concise explanation of the fundamental adjustments needed to use cumulative voting with soft ensembles can be found in [20]. However, we assume hard ensembles as input in order to keep things simple and focus on the analysis. The purpose of this article does not extend to analysis using soft ensembles.

In order to produce reliable results, clustering fusion integrates the output from multiple clustering algorithms. The possibility of clustering fusion techniques to enhance clustering performance is drawing more and more attention. The combination of cluster outputs is currently the typical operational technique for clustering fusion. A cluster ensemble is one technique for this combination or consolidation of findings from a portfolio of independent clustering outcomes [2].

Each data subset can be grouped, and then the clustering solution for all subsets can be combined. This makes it possible to cluster very huge datasets. Data may occasionally be geographically dispersed, making centralised data pooling impractical for reasons of cost and privacy. As a result, it's necessary to combine clustering solutions from several locations. Additionally, iterative clustering methods yield various partitions for the same data with various initializations since they are sensitive to initialization. A reliable and stable solution might be obtained by combining various partitions. It was demonstrated to be helpful in several contexts, including "Quality and Robustness" [8], "Knowledge Reuse" [9, 10], and "Distributed Computing" [11].

The remaining of the paper is structured as follows. Section 2 contains the pertinent work. Section 3 contains the suggested ensembles clustering merge approach. Section 4 of the paper includes a discussion of the experimental test platform and its findings. In part 5, we wrap up with a summary and some suggestions for further research.

2. Related work in clustering ensembles

To obtain the best possible partition of the initial dataset, a clustering ensemble combines numerous iterations of various clustering techniques. An ensemble of clustering solutions, P = P_l,P_2,..P_r, where r is the ensemble size, or the total number of clustering's in the ensemble, is known as a cluster ensemble when dataset X = xl x2,..,x_n is given. The Clustering-Ensemble approach first obtains the output of M clusters, then configures a common comprehension function to fuse each vector, and finally obtains the labelled vector. The objective of a cluster ensemble is to integrate the results of various clustering methods in order to provide more reliable and high-quality clustering. Despite the development of numerous clustering methods, the data mining and machine learning communities rarely use cluster ensembles. Co-association matrices were utilised by Fred and Jan in 2002 to create the final partition. The co-association matrix was subjected to a hierarchical (single link) clustering [12]. An adaptive Meta clustering approach was proposed by Zeng, Tang, Garcia-Frias, and GAO in 2002 for merging several clustering results using a distance matrix [13].

By treating each cluster in a separate clustering technique as a hyper edge, Strehl and Ghosh [9] suggested a hyper graph-partitioned approach to combining diverse clustering results. The cluster ensemble problem was resolved using three effective strategies that were developed. Each algorithm starts by converting the collection of clustering's into a hyper graph representation. The Cluster-based Similarity Partitioning Algorithm (CSPA) establishes a pairwise similarity metric using relationships between objects in the same cluster. The objects are then re-clustering using this induced similarity measure, producing a composite clustering. In the Hyper Graph Partitioning Algorithm (HGPA), a constrained minimum cut objective is used to approach the greatest mutual information objective. The cluster ensemble problem is essentially given as a problem of partitioning a suitably defined hyper graph, where hyper edges denote clusters. The goal of integration is seen as a cluster correspondence problem in their Meta CLustering Algorithm (MCLA). In essence, meta-clusters (groups of clusters) need to be found and consolidated.

In addition to formulating the process of cooperation between component clusters, Kai Kang, Hua-Xiang Zhang, and Ying Fan [19] also suggested a unique cluster ensemble learning technique based on dynamic cooperating clusters (DCEA). The strategy largely focused on how the component clusters worked together seamlessly during the training phase. By measuring the similarity between cluster centroids by computing their distances, this method first aligns the cluster centroids found by various component clusters. The next step then adjusts the aligned cluster centroids by a dynamic momentum term, and so on until the termination rule is satisfied.

A soft feature selection method (named LAC) was suggested by Muna AI-Razgan and Carlotta Domeniconi [20] that gives features weights based on the local correlations of data along each dimension. Distances along weakly connected dimensions are lengthened as a result of their low weight, which is applied to all dimensions. Strongly associated features receive a lot of weight, which has the effect of reducing distances along that dimension. As a result, the learnt weights execute a directional local reshaping of distances that enables improved cluster separation and, as a result, the finding of various patterns in various input space subspaces. The number of clusters k to be found in the data and the h factor, which regulates the strength of the incentive to cluster on additional features, both affect how LAC clusters the data.

3. Newly proposed ensemble merge algorithm

3.1. Definition

- The term "matching groups set, MG refers to sets of clusters from various clusters that have the largest cardinality in the intersection set, $MG_{[ij]}$ denotes matching pairs of the i^{th} cluster j^{th} cluster. For instance, MG $_{[1][I]}$ indicates that merging groups is achieved by combining the first and second cluster members.
- Degree of Agreement (DOA) Factor: The proportion of the level index to the total number of clusters.
- User-assigned value, often set to 50% of the number of clusters, for DOA_{Th}.

3.2. The Problem Specification

A consensus function Fx is described as a function that maps a set of clustering's to an integrated clustering when there are r groupings and the q-th grouping, x (q), has k (q) clusters. The ideal combined clustering has to have the greatest degree of informational overlap with the initial clustering's. Mutual information, a symmetric measure to quantify the statistical information shared between two distributions, is typically used to measure this shared information between clustering's [3, 18].

3.3. The Proposed voting solution

We go over our suggested Newly Proposed Ensemble Merge algorithm (EM: Ensemble Merge) for datasets in this section. B heterogeneous ensembles are applied to the same data set at the first level to get results for partitioning. Each ensemble's individual partitions are created one at a time.

These clustering findings are joined in pairs at the second level, known as matching groups set, or MG_{mk}, using the cardinality of similarity set between the clusters central components. The label naming problems are very simply and beautifully solved by using similarity between core points. As a result, significant computational costs are avoided.

The Degree of Agreement (DOA) for each data point is determined during the merging process. The ratio of the cluster count to the merging level index is known as the DOA_{Th} factor. Additionally, the DOA value will accumulate until it hits the DOA_{Th} level. Any data point's DOA can be confirmed to belong to a specific cluster result once it has crossed the threshold. Thus, our solution completely avoids the traditional voting process that uses a large voting matrix to validate the majority. The data piece will not take part in any more computations once it has been confirmed to a cluster. As a result, the computational cost may also drastically decreased.

Because everything is related to everything else, but close-by things are more related than far-off things, this strategy will be quite helpful for handling data. The only difficulty will be the quantity of data points that oscillate between clusters. Early on in the iterations, all pertinent issues will be resolved, thereby decreasing the need for computational resources. Unsettled data objects with less than or equal to DOA_{Th} will be handled at the third level. When there are an even number of clusters, boundary elements will arise that can be resolved by merging the likelihood data with the final clusters. Data points that fall below the threshold will be labelled as noise or outliers.

The robust combined result will be produced by merging the final layer with the preceding combined clusters. This strategy requires less processing because we usually merge vote matrixes and eliminate them. The technique's three levels are implemented in order. They simply receive the outcomes from the prior tiers without interfering with one another. After completion, the algorithm ends without performing any feedback.

3.4. Pseudo Code

- Step 1: From Dataset D, create k clusters, each using m clusters.
- Step 2: DOA Increment Factor as 1/m
- Step 3: Identify merging groups For MG_1
- Step 4: for each pair in the merging group Sets, $MG_{ik'}$
- {Build dice vectors comprising the first data vector, and add it with a unit DOA vector. Then, modify the DOA vector using the second pair}
- Step 5: Place the element in $Final\_Kluster_i$ if (DOA of data elements> DOA_{Th}); else, place them in $Orphan_i$.
- Step 6: Using $Final\_Kluster_i$, $Orphan_i$, calculate MG_{i+1}.
- Step 7: Repeat steps 4 to 6 till all m slices are exhausted.
- Step 8: Group any leftover Orphan as anomaly .
- Step 9: Return the robust clusters & anomalies.

4. Test platform and results

Both homogeneous and heterogeneous data have been used in our test platform. As well as diverse groups. The latter scenario gives us ensemble clusters were made using K-means, PAM, FCM, and DBSCAN formulas. K-means is a quick and easy method for strong iterative method to divide a data collection into k Unconnected clusters. The DBSCAN approach works well when operates reasonably effectively with data and attributes data. The main benefit of partitioning around medoids (PAM) is Scalability, and consequently, value in mixed data. Therefore, the input for our merging strategy is formed by these four clustering techniques and various cluster sizes.

The majority of ensemble methods use sampling approaches to choose the data for the experimental platform; nevertheless, this strategy causes the loss of some naturally occurring data clusters, lowering the quality of clusters. We made an effort to avoid sampling and apply the Newly Proposed EM algorithm to the entire dataset. This is possible because during the cycle, only the matching pairs are chosen for merging. We measured an ensemble's accuracy as the degree of agreement between the ensemble partition and the "actual" partition using the Co-Association (CA) metric. For assessing the outcomes of clustering, the classification accuracy is frequently used. The percentage of accurately classified items in the data set is determined as CA for the partition in order to provide the best possible re-labelling of the clusters. Due to the combination of its cutting-edge technologies and functional range, Rapid Miner is one of the top open-source data mining solutions in the world. The datasets are processed by Rapid Miner, which serves as the standard for determining correctness. The test result using the Wine dataset encouraging and demonstrate greater cluster accuracy when compared to alternative non-ensembling approaches and homogenous cluster ensembles. On the basis of space complexity, our ensemble fusion method was contrasted with Alexander Strehl's method [3]. More details can be found in the graph below. When compared to alternative non-ensembling algorithms, the test results with Wine dataset (data repository) encouraging and exhibit greater cluster accuracy. Ensembles of homogeneous clusters are also used.

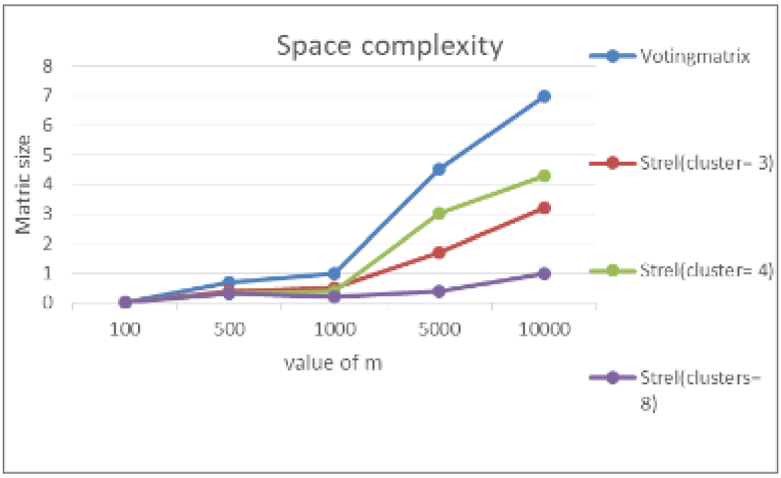

Figure 1 Space complexity comparison in wine data set

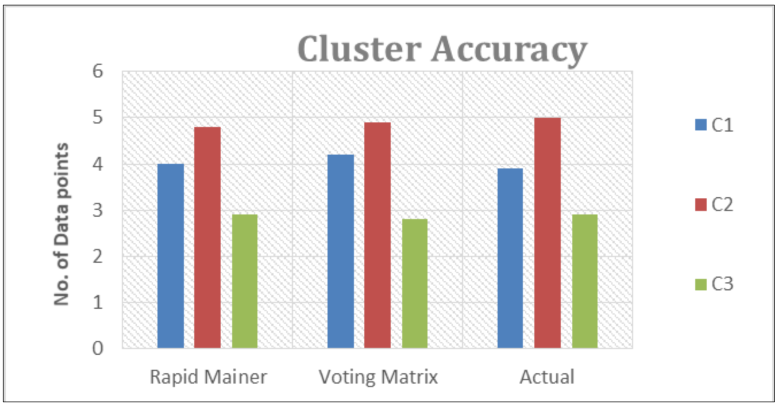

Figure 2 Cluster accuracy in wine data

When we compared our results with results from commercially available clustering software, our Newly Proposed EM technique was shown to produce accuracy that was on par with industry standards while also being more effective. The results we achieved by running our algorithm on the commercial clustering programme "Rapid Miner" were the same when we tested it on the "Wine" dataset (data repository). The majority of other datasets likewise attested to the fact that the ensembling method did not detect incorrect clusters. The two stages of the present methodologies were ensemble preparation and consensus function. For each cluster, the ensemble preparation stage calls for the construction of matrices of dimension m*(n+k), where m is the number of data objects, n is the number of attributes, and k is the number of clusters. As a result, the matrix dimension for the ensemble preparation stage will be in the order of mc*(n+k), where c is the number of clusters, and the Consensus function stage calls for the building of matrices m*(n+c) While our DOA_{Th} vector, has no dependency on c . A space complexity of the order m*(n+1) where m is the number of data objects and n is the number of attributes and is hence scalable.

5. Conclusion

In this study, we addressed the re-labelling issue that arises frequently in cluster ensembles and offered a practical approach to resolve it. The cluster ensemble is an extremely flexible architecture that supports a variety of uses. On mixed databases, we used the suggested layered cluster merging strategy. The cardinality of data points and the expanded dimensions are the key problems with mixed databases. The majority of the ensemble algorithms in use today must produce voting matrices of at least order n2. This restriction creates a significant bottleneck when n is very large and is also a prevalent factor in mixed datasets, making it difficult to build strong clusters quickly and accurately.

Our approach used merging based on ensemble to address the re-labelling. Our approach completely avoids the traditional voting process that uses a large voting matrix to determine the majority. The data piece will not take part in any more computations once it has been confined to a cluster. As a result, the computational cost is also drastically decreased. As different clustering solutions produce diverse outcomes for the same dataset, we apply ensemble approaches to improve cluster accuracy. Ensemble approaches produce more accurate clustering results by combining the results of several runs of the same or various clustering algorithms. Since client data cannot be shared in the modern world, ensemble methods are employed to classify the data before it is given to the analysts for usage in a variety of applications.

Future prospective

Data mining's main objective is to automate the process of knowledge discovery. It is significant to highlight that in this study, it has been assumed that the user has a solid understanding of the data and the hierarchies employed in the mining process. The quality of the resulting clusters is still influenced by the key input of choosing the value of k. Using domain-specific knowledge as advice will help determine the value of k. We believe that using domain knowledge in semi-supervised clustering could enhance the quality of the mined clusters. Although we experimented with heterogeneous clusters, it might also be evaluated with additional novel clustering method combinations using different base clusters. This will ensure that more natural groupings are explored.

Compliance with ethical standards

Acknowledgments

The authors would like to thanks Faculty of Computer Application, SAGE University, Indore for supporting me. And to Dr. Sunita Dwivedi for his approval to carry out this research. Also, to all informants who agreed and actively helped in this research.

Disclosure of conflict of interest

There are no conflicts of interest with the study that the authors of this paper need to report.

References

[1] J.A. Hartigan, Clustering Algorithms, Wiley, New York, 1975.

[2] A.K. Jain, R.C. Dubes, Algorithms for Clustering Data, Prentice-Hall, Englewood Cliffs, NJ, 1988.

[3] L. Kaufman, P.J. Rousseeuw, Finding Groups in Data, Wiley, New York, 1990.

[4] A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering a review, ACM Computing Surveys 31 (3) (September 1999) 264–323.

[5] C. Fraley, A.E. Raftery, Model-based clustering, discriminant analysis, and density estimation, Technical Report, University of Washington, October 2000.

[6] J.M. Buhmann, Data clustering and learning, in M. Arbib (Ed.), Handbook of Brain Theory and Neural Networks, MIT Press, Cambridge MA, 2002.

[7] J. Kleinberg, An impossibility theorem for clustering, in Proceedings of Advances in Neural Information Processing Systems (NIPS), 2002.

[8] N. Tishby, F. Pereira, W. Bialek, The information bottleneck method, in Proceedings of the 37-th Annual Allerton Conference on Communication Control and Computing, , 1999, pp. 368–377.

[9] Elena D. Cristofor, Information-theoretic methods in clustering, Ph.D. Thesis, University of Massachusetts, 2002.

[10] D.M. Sima, Regularization techniques in model fitting and parameter estimation, Ph.D. Thesis, Faculty of Engineering, K.U. Leuven, Leuven, Belgium, 2006.

[11] J.H. Friedman, Regularized discriminant analysis, Journal of the American Statistical Association 84 (1989)165–175.

[12] A. Strehl, J. Ghosh, Cluster ensembles—a knowledge reuse framework for combining multiple partitions, Journal of Machine Learning Research 3 (2002) 583–617

[13] A. Fred, A.K. Jain, Combining multiple clusterings using evidence accumula- tion, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (6) (2005) 835–850.

[14] A. Topchy, A.K. Jain, W. Punch, Clustering ensembles models of consensus and weak partitions, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (12) (2005) 1866–1881.

[15] E. Dimitriadou, A. Weingessel, K. Hornik, A combination scheme for fuzzy clustering, International Journal of Pattern Recognition and Artificial Intelligence 16 (7) (2002) 901–912.

[16] 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..

[17] Su-lan Zhail,Bin Luol Yu-tang Guo "Fuzzy Clustering Ensemble Based on Dual Boosting" , Fourth International Conference on Fuzzy Systems and Knowledge Discovery 07

[18] 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.

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

[20] Matheus C.1., 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.

[21] 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

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

[23] 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.

[24] 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.

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

[26] 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.

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

[28] 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

[29] 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.

[30] Toshihiro Osaragi, "Spatial Clustering Method for Geographic Data", UCI Working Papers Series, paper41, Jan 2002.

[31] Jain, A.K, Murty, M.N., and Flynn P.1 Data clustering a review. ACM Computing Surveys,31, 3, 264-323

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

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

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

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