International Journal of scientific research and management
(IJSRM) ||Volume||1||Issue||6 ||Pages|| 327-331 ||2013||
Website: www.ijsrm.in ISSN (e): 2321-3418

A FUZZY CLUSTERING ENSEMBLE APPROACH FOR CATEGORICAL DATA

K. Lakshmi priya[1], V.Kathiresan[2,] P.Sumathi[3]

[1]Research Scholar, Department of Computer Science, RVS college of Arts and Science
Coimbatore, Tamil Nadu 641402, India
priyakovaii@gmail.com

[2] Assistant Professor, Department of Computer Applications (MCA), RVS College of Arts and Science
Coimbatore, Tamil Nadu 641402, India
kathiresan.v@rvsgroup.com

[3]Assistant Professor, Department of Computer Science, Government Arts College
Coimbatore, Tamil Nadu 641402, India
sumathirajesh@yahoo.com

ABSTRACT

Data clustering is one of the essential tools for perceptive structure of a data set. It plays a crucial and initial role in machine learning, data mining and information retrieval. The intrinsic properties of the traditional algorithms intended for numerical data, can be employed to measure distance between feature vectors and cannot be directly applied for clustering of categorical data ,Wherever domain value are distinct haven't any ordering outlined. The final data partition generated by traditional algorithms, results in incomplete information and the core ensemble information matrix presents only cluster data point relations with many entries left unknown and disgrace the quality of the resulting cluster. In the proposed system, a new highly effective fuzzy cluster ensemble approach to categorical data clustering transforms the original categorical data matrix to an information-preserving numerical variation (QM), to which an effective hybrid graph partitioning technique can be directly applied. Using the fuzzy clustering algorithm, the quality matrix is determined efficiently and can be used to partition the categorical data under unsupervised circumstances

KEYWORDS

Clustering, Categorical Data, Cluster Ensembles, Link Based Similarity, Data Mining.

INTRODUCTION

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified.[12]

Data mining, *the extraction of hidden predictive information from large databases*, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses.[6] Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions.

Clustering is a mathematical tool that attempts to discover structures or certain patterns in a data set, where the objects inside each cluster show a certain degree of similarity. Clustering is a collection of data objects, similar to one another within the same cluster and are dissimilar to objects in the other clusters. Similar to classification, clustering is the organization of data in classes. However, unlike classification, in clustering, class labels are unknown and it is up to the clustering algorithm to discover acceptable classes.

In fuzzy clustering (also referred to as soft clustering), data elements can belong to more than one cluster, and associated with each element is a set of membership levels that indicate the strength of the association between that data element and a particular cluster. Fuzzy[7] clustering is a process of assigning membership levels, and using them to assign data elements to one or more clusters. The algorithm starts with random initial K cluster centers, and then at every iteration it finds the fuzzy membership of each data points to every cluster. Fuzzy clustering allows each feature vector to belong to more than one cluster with different membership degrees (between 0 and 1) and vague or fuzzy boundaries between clusters.

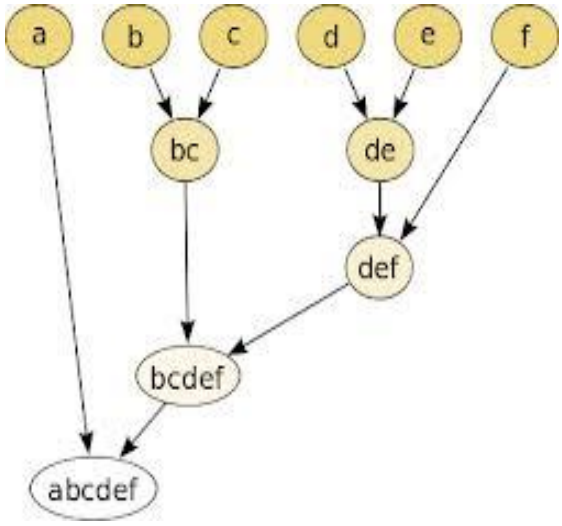

Figure 1: Cluster formation

CLUSTERING ALGORITHMS FOR CATEGORICAL DATA

- Complete-Linkage Clustering
- Average-Linkage Clustering
- K-medoids Clustering
- Fuzzy clustering

Complete Linkage Clustering

The complete-linkage (CL) hierarchical clustering Algorithm is also called the maximum method or the farthest neighbor method. It is obtained by defining the distance between two clusters to be the largest distance between a sample in one cluster and a sample in the other cluster.

Average Linkage Clustering

The hierarchical average-linkage (AL) clustering algorithm, also known as the unweighted pair-group method using arithmetic averages (UPGMA), is one of the most widely used hierarchical clustering algorithms. The average-linkage algorithm is obtained by defining the distance between two clusters to be the average distance between a point in one cluster and a point in the other cluster.

K-Medoids Clustering

Partitioning around medoids(PAM), conjointly known as K-medoids clustering, may be a variation of K-means with the target to attenuate the inside cluster variance. The concept of PAM is to pick out K representative points, or medoids, in X and assign the remainder of the information points to the cluster known by the closest medoid. Initial set of K medoids[5] square measure chosen arbitrarily. Subsequently, all the points in X square measure assigned to the closest medoid. In every iteration, a replacement medoid is decided for every cluster by the cluster. After that, all the data points with minimum total distance to all or any different points of the cluster. After that, all the points in X square measure reassigned to their clusters in accordance with the new set of medoids.

.

Fuzzy Clustering

In fuzzy clustering (also referred to as soft clustering), data elements can belong to more than one cluster, and associated with each element is a set of membership levels that indicate the strength of the association between that data element and a particular cluster. Fuzzy clustering is a process of assigning membership levels, and using them to assign data elements to one or more clusters. The algorithm starts with random initial K cluster centers, and then at every iteration it finds the fuzzy membership of each data points to every cluster. Fuzzy clustering allows each feature vector to belong to more than one cluster with different membership degrees (between 0 and 1) and vague or fuzzy boundaries between clusters.

PROBLEM DEFINITION

This paper presents an analysis that suggests this problem degrades the quality of the clustering result, and it presents a new link-based approach,

Table: Measure of LCE

which improves the conventional matrix by detecting unknown entries through similarity between clusters in an ensemble. In particular, an efficient link-based algorithm is proposed for the underlying similarity assessment.

While a large number of cluster ensemble techniques for numerical data have been put forward in the previous decade, there are only a few studies that apply such a methodology to categorical data clustering. The method introduced in [10] creates an ensemble by applying a conventional clustering algorithm (e.g., k-modes and COOLCAT) to different

data partitions, each of which is constituted by a unique

Rows	seppellength	seppewidth	petallength	petalwidth
1	0.279	0.279	0.315	0.279
2	0.244	0.253	0.315	0.244
3	0.307	0.251	0.244	0.307
4	0.244	0.258	0.3	0.244
5	0.334	0.315	0.315	0.334
6	0.244	0.379	0.235	0.244
7	0.244	0.258	0.315	0.244
8	0.334	0.258	0.3	0.334
9	0.315	0.334	0.315	0.315
10	0.244	0.258	0.3	0.244

subset of data attributes.

A cluster in attribute-specific partition contains data points that share a specific attribute value (i.e., categorical label). [2]Thus, the ensemble size is set by amount of categorical label s, across al l data attributes. The final clustering result is generated using the graph-based consensus techniques.

A cluster ensemble can also be achieved by generating base clustering's from different subsets of initial data. It is intuitively assumed that every clustering algorithm will provide different levels of performance for different partitions of a data set.

Heterogeneous ensembles is a number of different clustering algorithms are used together to generate base clustering's. Mixed heuristics was an addition to using one of the aforementioned methods, any combination of them can be applied as well as.

EXISTING SYSTEM

In existing system, the Novel link based approach has been established to discover the unknown values of the data partition, but do not produce good precision when computing the real large datasets. The final data partition generated by conventional algorithms, results in incomplete information and the core ensemble information matrix presents only cluster data point relations with many entries left unknown and disgrace the eminence of the resulting cluster. [3]The Weighted Triple-Quality algorithm is proposed, for the approximation of the similarity between clusters in a link network and simply counts the number of triples and computationally expensive for large datasets.

DISADVANTAGES

It combines the power of two novel techniques, key phrase discovery and orthogonal clustering, to generate clusters which are both reasonable and readable. Moreover, SHOC can work for multiple languages: not only English but also oriental languages like Chinese. The main contribution of this paper includes the following. (1) The benefits of using key phrases as Web document features are discussed. A key phrase discovery algorithm based on suffix array is presented. This algorithm is highly effective and efficient no matter how large the language's alphabet is. The concept of orthogonal clustering is proposed for general clustering problems. The reason why matrix Singular Value Decomposition (SVD) can provide solution to orthogonal clustering is strictly proved.

PROPOSED SYSTEM

A new highly effective fuzzy cluster ensemble approach to categorical data clustering transforms the original categorical data matrix to an information-preserving numerical variation (QM), to which an effective hybrid graph partitioning technique can be directly applied. Using the fuzzy clustering algorithm[1], the quality matrix is determined efficiently and can be used to partition the categorical data under unsupervised circumstances.

In proposed system, an unsupervised ensemble fuzzy clustering approach have been proposed that permit to dispose both of the flexibility of the fuzzy sets and the robustness of the ensemble methods. [1]

ADVANTAGES

While Web search engines can retrieve information on the Web for a specific topic, users have to step a long ordered list in order to locate the needed information, which is often tedious and less efficient. We propose a new link-based clustering approach to cluster search results returned from Web search engines by exploring both co-citation and coupling. Unlike document clustering algorithms in IR that are based on common words/phrases shared among documents, our approach is based on common links shared by pages. We also extend the standard clustering algorithm, K-means, to make it more natural to handle noise and apply it to Web search results. By filtering some irrelevant pages, our approach clusters high quality pages in Web search results into semantically meaningful groups to facilitate users accessing and browsing. Preliminary experiments and evaluations are conducted to investigate its effectiveness. The experimental results show that link-based clustering of Web search results is promising and beneficial.

The prominent future work includes an extensive study regarding the methods that can be used to define measures of similarity between categorical data. Possible measures include the shortest path between tipples and the commute distance between nodes on the database graph. Also the connection closeness between the expressive power of the category and the database graph can be analyzed and looked in. To add up an objective way of selecting the query language used for defining the database graph can also be better way for enhancement.

RESULT AND DISCUSSION

Following table shows the measures of LCE models being mostly higher than those of the corresponding baseline counterparts (Base), the quality of the RM appear to be significantly better than that of the original, binary variation.

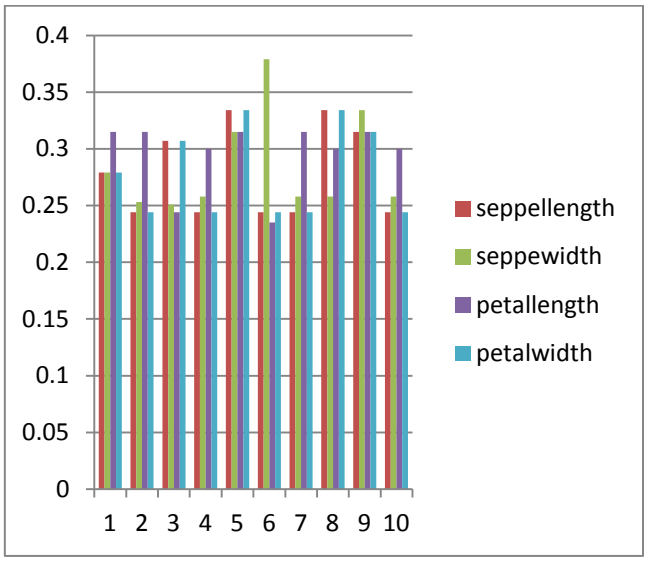

FIGURE 2: Measure of LCE

In order to further evaluate the quality of identified techniques, the number of times that one method is significantly better and worse (of 95 percent confidence level) than the others are assessed across experimented data sets To obtain a fair comparison, this pair wise assessment is conducted on the results with six data sets, where the clustering results can be obtained for all the clustering methods. Also note that CM consists of five clustering algorithms for categorical data and 35 different cluster ensemble models, each of which is a unique combination of ensemble type.

Despite its inefficiency, CSPA has the best performance among assessed ensemble methods. In addition, Cobweb is the most effective among five categorical data clustering algorithms included in this evaluation.

CONCLUSION

Cluster ensembles have emerged as an efficient answer that's able to overcome the limitation of grouping mixed information, and develop the strength yet because the quality of cluster results. The most objective of cluster ensembles is to affix completely different cluster selections in such some way on accomplish accuracy bigger to it of any person cluster. Cluster ensemble approach to categorical information cluster, transforms the first categorical information matrix to associate information-preserving graph partitioning technique may be directly applied to induce the ultimate information partition.

FUTURE WORK

The prominent future work includes an extensive study regarding the behavior of other link-based similarity measures within this problem context. Also, the new techniques are going to be applied to specific domains, together with tourism and medical data sets.

REFERENCES

[1] Natthakan Iam-On, Tossapon Boongoen, Simon Garrett, and Chris Price, "A Link-Based Cluster Ensemble Approach for Categorical Data Clustering," ieee transactions on kowledge and data engineering, vol .24, no .3, march 2012.

[2] Z. He, X. Xu, and S. Deng, "Squeezer: An Efficient Algorithm for Clustering Categorical Data," J. Computer Science and Technology, vol. 17, no. 5, pp. 611-624, 2002.

[3] S. Guha, R. Rastogi, and K. Shim, "ROCK: A Robust Clustering Algorithm for Categorical Attributes," Information Systems, vol. 25, no. 5, pp. 345-366, 2000.

[4] D. Gibson, J. Kleinberg, and P. Raghavan, "Clustering Categorical Data: An Approach Based on Dynamical Systems," VLDB J., vol. 8, nos. 3-4, pp. 222-236, 2000.

[5] V. Ganti, J. Gehrke, and R. Ramakrishnan, "CACTUS: Clustering Categorical Data Using Summaries," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 73-83, 1999.

[6] D. Barbara, Y. Li, and J. Couto, "COOLCAT: An Entropy-Based Algorithm for Categorical Clustering," Proc. Int'l Conf. Information and Knowledge Management (CIKM), pp. 582-589, 2002.

[7] Y. Yang, S. Guan, and J. You, "CLOPE: A Fast and Effective Clustering Algorithm for Transactional Data," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 682-687, 2002.

[8] N. Nguyen and R. Caruana, "Consensus Clusterings," Proc. IEEEInt'l Conf. Data Mining (ICDM), pp. 607-612, 2007.

[9] X.Z. Fern and C.E. Brodley, "Solving Cluster Ensemble Problems by Bipartite Graph Partitioning," Proc. Int'l Conf. Machine Learning (ICML), pp. 36-43, 2004.

[10] A. Strehl and J. Ghosh, "Cluster Ensembles: A Knowledge Reuse Framework for Combining Multiple Partitions," J. Machine Learning Research, vol. 3, pp. 583-617, 2002.

[11] A.P. Topchy, A.K. Jain, and W.F. Punch, "A Mixture Model for Clustering Ensembles," Proc. SIAM Int'l Conf. Data Mining, pp. 379-390, 2004.

[12] Chandrika Satyavolu, Prof. T.Y. Lin "The Theory of Attributes from Data Mining Prospect" feb1 2001 .

[13] Cao, T.H., H.T. Do, D.T. Hong and T.T. Quan, 2008. Fuzzy named entity-based document clustering. Proceedings of IEEE International Conference on Fuzzy Systems, June 1-6, Hong Kong, pp: 2028-2034.

[14] Casillas, M.T., G. de Lena and R. Martınez, 2000. Document clustering into an unknown number of clusters using a genetic algorithm. Trans. Neural Networks, 11: 586-600.

[15] Chakrabarti, S., 2003. Mining the Web: Discovering Knowledge from Hypertext Data. Morgan Kaufmann Publishers, California.

[16] Cooley, R., B. Mobasher and J. Srivastava, 1999. Data preparation for mining world wide web browsing. J. Knowledge Inform. Syst., 1: 5-32.

[17] Cristofor, D. and D. Simovici, 2002. An information-theoretical approach to clustering categorical databases using genetic algorithms. Proceedings of 2nd Society for Industrial and Applied Mathematics Industrial Data Mining, Workshop on Clustering High Dimensional Data, (SAIM IDMWCHDD`02), USA., pp: 37-46.

Author Profile

K. Lakshmi priya received her BCA in 2010, from Bharathiar University, Coimbatore.Msc(IT) in 2012, from Bharathiar University, Coimbatore. At present she is pursuing her M.Phil in the area of Data Mining in RVS College of Arts and Science, Coimbatore.

Kathiresan.V received his B.Sc., in 2003 and MCA in2006 from Bharathiar University, Coimbatore. He obtained his M.Phil. in the area of Data mining from Periyar University, Salem in 2007. At present he is working as a Assistant Professor, Department of Computer Applications (MCA) in RVS College of Arts and Science, Coimbatore. His research interest lies in the area of Data mining. He got Faculty Excellence Award from RVS College of Arts & Science for the Academic year 2007-08, 2008-09, 2009-10, 2010-11, 2011-12 and 2012-13consecutively.

Dr P. Sumathi received her B.Sc Physics in 1994 and MCA in 1997 from Bharathiar University. She
Obtained her M.Phil in Mother Teresa University, Kodaikanal in the year 2004. She was awarded with her Ph.D from the Bharathiar University in 2009.She has an experience of 15 years in the teaching field. Her research interest lies in the area of Data mining .She worked as the Head of the Department, Computer Science in the PSG College of Arts & Science, Coimbatore .At present she is working as the Assistant Professor in the PG & Research Department of Computer Science, Government Arts College, Coimbatore.