CLUSTER ENSEMBLE APPROACH FOR CATEGORICAL DATA CLUSTERING

M. Pavithra

Assistant. Professor in C.S.E

Jansons Institute of Technology, India.

Email: smilepavi17@gmail.com

ABSTRACT : Clustering, in data mining, is useful to discover distribution patterns in the underlying data[1]. Clustering algorithms usually employ a distance metric based (e.g., Euclidean) similarity measure in order to partition the database such that data points in the same partition are more similar than points in different partitions[3]. The problem of clustering becomes more challenging when the data is categorical, that is, when there is no inherent distance measure between data values. Various clustering algorithms are developed to cluster or categorize the datasets. Many algorithms are used to cluster the categorical data. Some algorithms cannot be directly applied for clustering of categorical data[5]. This project proposes an Algorithm called Average Weighted Quality (AWQ), which also uses k-means algorithm for basic clustering. Once the basic clustering is done by using consensus functions we can get cluster ensembles of categorical data[7]. This categorical data is converted to refined matrix. This project introduces a link-based approach to refining the aforementioned matrix, giving substantially less unknown entries.

Keywords : Data mining, Data Clustering, Categorical data, Cluster Ensemble Approach , Average Weighted Quality.

Reference: to this paper should be made as follows: M. Pavithra (2006) 'Cluster Ensemble Approach For Categorical Data Clustering', International Journal of Inventions in Computer Science and Engineering, Vol. 3, No. 9, pp.01–05.

INTRODUCTION

The cluster ensemble will differentiate various cluster outputs by using the clustering algorithms[4]. The main goal of ensembles has been to improve the accuracy and robustness of a given classification or regression task, and spectacular improvements have been obtained for a wide variety of data sets. Cluster ensemble methods are presented under three categories: Probabilistic approaches, Approaches based on co-association, and Direct and other heuristic methods. Categorical variables represent types of data which may be divided into groups. Examples of categorical variables are race, sex, age group, and educational level[8].

Categorical data is a statistical data type consisting of categorical values used for observed data whose value is one of a fixed number of nominal

categories, or for data that has been converted into that form[5]. Categorical data are always nominal whereas nominal data need not be categorical. Clustering the categorical data is remaining a challenging task in many techniques[9]. A critical problem in cluster ensemble research is how to combine multiple clustering's to yield a final superior clustering result. These problems are overcome by using different techniques. The link based similarity is used to improve the clustering result[12].

Related Work

Clustering Ensemble approaches are also referred to as the Consensus clustering approaches in which it mainly gains more and more consideration due to its diverse applications in the areas of data mining, machine learning, pattern recognition, bioinformatics, information retrieval, image processing and analyzing, statistical data analysis[1]. Clustering Ensemble techniques have the powerful ability to achieve the aggregation of the several partitions from different data sources and thus it improves the stability, compactness, and robustness of the traditional single clustering algorithms[2].

Cluster ensembles provide a solution to challenges inherent to clustering[4]. Cluster ensembles can find robust and stable solutions by leveraging the consensus across multiple clustering results[2]. The cluster ensemble combines various clustering outputs into single consolidated cluster. The cluster ensemble will differentiate various cluster outputs by using the clustering algorithms[11]. The main goal of ensembles has been to improve the accuracy and robustness of a given classification or regression task, and spectacular improvements have been obtained for a wide variety of data sets. Cluster ensemble methods are presented under three categories: Probabilistic approaches, Approaches based on co association, and Direct and other heuristic methods[4].

The categorical data is clustered and represented using the cluster ensembles[8]. Cluster ensembles are used as best alternative to the standard cluster analysis. The data set has been clustered by using any of the well known cluster algorithm and represented as a cluster ensemble. The cluster ensembles generate a final data partition based on incomplete information and the information is not prefect to make use of it[9].

PROPOSED WORK

Cluster Ensembles of Categorical Data

while averaging out emergent spurious structures that arise due to the various biases to which each participating algorithm is tuned. Conventional approach, the technique developed in acquires a cluster ensemble without actually implementing any base clustering on the examined data set[14]. In fact, each attribute is considered as a base clustering that provides a unique data partition. In particular, a cluster in such attribute-specific partition contains data points that share a specific attribute value (i.e., categorical label).

Subspace ensemble

Another alternative to generate diversity within an ensemble is to exploit a number of different data subsets. To this extent, the cluster

ensemble is established on various data subspaces, from which base clustering results are generated[8]. Similar to the study in, for a given N *d data set of N data points and d attributes, an N * q data subspace (where q < d) is generated by q = qmin + ⌞ α(qmax – qmin) ⌟ . where α € [0,1] is a uniform random variable, qmin and qmax are the lower and upper bounds of the generated subspace, respectively. In particular, qmin and qmax are set to 0:75d and 0:85d. An attribute is selected one by one from the pool of d attributes, until the collection of q is obtained[9]. The index of each randomly selected attribute is determined as h= ⌞1+βd⌟ , given that h denotes the hth attribute in the pool of d attributes and β € [0,1] is a uniform random variable[7]. Note that k-modes is exploited to create a cluster ensemble from the set of subspace attributes, using both Fixed-k and Random-k schemes for selecting the number of clusters.

Generating a Refined Matrix

The cluster ensemble and the corresponding BM, a large number of entries in the BM are unknown, each presented with “0”[2]. Such condition occurs when relations between different clusters of a base clustering are originally assumed to be nil. In fact, each data point can possibly associate to several clusters of any particular clustering. These hidden or unknown associations can be estimated from the similarity among clusters, discovered from a network of clusters.

RM(xi,cl) = {1, if cl=Ct(xi),

{ Sim(cl,Ct(xi)), otherwise

where Ct(xi) is a cluster label (corresponding to a particular cluster of the clustering πt) to which data

point xi belongs. In addition, sim(Cx,Cy) €[0,1] denotes the similarity between any two clusters Cx, Cy, which can be discovered using the following link-based algorithm. Note that, for any clustering πt € π, 1<=∑ RM(xi,C)<=kt. Unlike the measure of fuzzy membership, the typical constraint of ∑ RM(xi,C)= 1 is not appropriate for rescaling associations within the RM[4].

Performance Evaluation

This section presents the evaluation of the proposed link based method (LCE), using a variety of validity indices and real data sets[6]. The quality of data partitions generated by this technique is assessed against those created by different categorical data clustering algorithms and cluster ensemble techniques.

Investigated Data Sets

The experimental evaluation is conducted over two data sets. The "UCI Machine learning repository" data set is a subset of the well known attribute values collection— UCI Machine learning repository[6].

Data Normalization

A summary of the datasets taken from the UCI Machine learning repository is shown in Table 1. The datasets are selected in such a way that the problems chosen are with at least six classes and no missing values[6].

Table 1 Summary of Datasets

Data sets	No of Instances	No of Attributes	No of Classes	Missing Values	Area
Primary Tumour	699	10	10	NIL	Life

Illustration :

Primary Tumour Datasets

Accuracy:

The classification accuracy of standard methods (CO+SL,CO+AL and WTQ) and proposed method (C-Rank) based on number of samples is 200. If the number of cluster is 7 then type I,II,III cluster ensemble for proposed method (C-Rank) gets increased in their classification accuracy when compared to other standard methods(CO+SL,CO+AL,WTQ)are shown the

Table 1. The classification accuracy of standard methods (CO+SL,CO+AL and WTQ) and proposed method (C-Rank) based on number of samples is 300.

Number of Cluster	Ensemble Type	Classification Accuracy (%)			
		Co association with Single Link (CO+SL)	Co association with Average Link (CO+AL)	Weighted Triple Quality (WTQ)	C - Rank
4	Type I	32.52	40.17	44.53	53.69
		35.78	43.45		
	Type II	38.54	46.67	47.24	53.90
	Type III			47.83	54.22
5	Type I	32.88	40.55	44.80	54.40
		37.55	44.34		
	Type II	38.60	46.70	47.50	54.69
	Type III			47.77	55.10
6	Type I	32.98	40.78	44.93	54.88
		37.74	44.89		
	Type II	39.01	46.91	47.67	55.37
	Type III			47.89	55.94

7	Type I	33.10	41.45	45.39	56.44
		38.22	45.67		
	Type II	39.56	46.98	48.45	56.77
	Type III			49.56	57.20

Table 2: Comparison of Classification Accuracy of standard and proposed methods based on number of samples = 400

Number of Cluster	Ensemble Type	Classification Accuracy (%)			
		Co association with Single Link (CO+SL)	Co association with Average Link (CO+AL)	Weighted Triple Quality (WTQ)	C- Rank
5	Type I	38.56	43.03	46.67	56.59
	Type II	40.29	46.78	47.73	56.77
	Type III	42.02	47.88	48.60	57.30
6	Type I	38.88	43.49	46.72	56.69
	Type II	40.56	46.89	47.84	56.80
	Type III	42.34	47.90	48.68	57.60

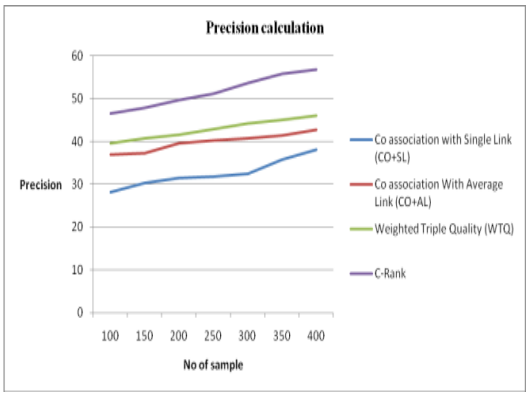

Fig.1 Graph for Performance of Precision based on number of samples

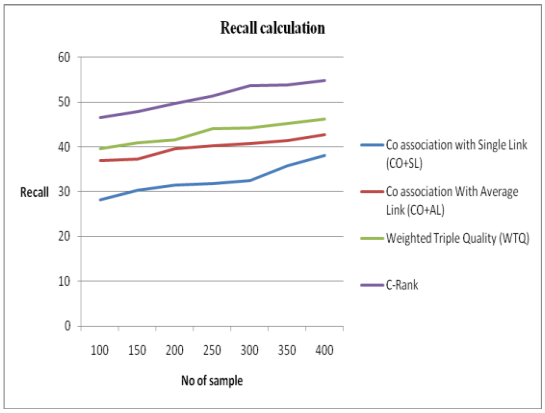

Fig.2 Graph for Performance of Recall based on number of samples

If the numbers of samples are more then precision value for proposed methods(C-Rank) has increased up to 58.79 % . The precision rate for standard methods(CO+SL,CO+AL,WTQ) are slightly less when compared to proposed methods which is shown in the Fig.1.

If the numbers of sample are more then recall value for proposed methods(C-Rank) has increased up to 54.60 % . The recall rate for standard methods(CO+SL,CO+AL,WTQ) are slightly less when compared to proposed methods which is shown in the Fig.2.

Conclusion

This paper presents a novel, highly effective link-based cluster ensemble approach (WTQ) to categorical data clustering. It transforms the original categorical data matrix to an information-preserving numerical variation (RM), to which an effective graph partitioning technique can be directly applied. The problem of constructing the RM is efficiently resolved by the similarity among categorical labels (or clusters), using the Weighted Triple-Quality similarity algorithm. The empirical study, with different ensemble types, validity measures, and data sets, suggests that the proposed

link-based method usually achieves superior clustering results compared to those of the traditional categorical data algorithms and benchmark cluster ensemble techniques. It also presents a Crank link based cluster approach for categorical data clustering.

Future Work

To improve clustering quality a new link-based approach the conventional matrix by discovering unknown entries through similarity between clusters in an ensemble and an efficient link-based algorithm is proposed for the underlying similarity assessment. To extend the work by analyzing the behaviour of other link-based similarity measures with this problem the quality of the clustering result.

References

1. N. Nguyen and R. Caruana, "Consensus Clusterings," Proc. IEEE Int'l Conf. Data Mining (ICDM), pp. 607-612, 2007.

2. A.P. Topchy, A.K. Jain, and W.F. Punch, "Clustering Ensembles: Models of Consensus and Weak Partitions," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 27, no. 12, pp. 1866-1881, Dec. 2005.

3. C. Boulis and M. Ostendorf, "Combining Multiple Clustering Systems," Proc. European Conf. Principles and Practice of Knowledge Discovery in Databases (PKDD), pp. 63-74, 2004.

4. 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.

5. Z. He, X. Xu, and S. Deng, "A Cluster Ensemble Method for Clustering Categorical Data," Information Fusion, vol. 6, no. 2, pp. 143-151, 2005.

6.A. Asuncion and D.J. Newman, "UCI Machine Learning Repository," School of Information and Computer Science, Univ. of California, http://www.ics.uci.edu/~mlearn/MLRepository.

html, 2007.

7.Y. Zhang, A. Fu, C. Cai, and P. Heng, "Clustering Categorical Data," Proc. Int'l Conf. Data Eng. (ICDE), p. 305, 2000.

8. M. Dutta, A.K. Mahanta, and A.K. Pujari, "QROCK: A Quick Version of the ROCK Algorithm for Clustering of Categorical Data," Pattern Recognition Letters, vol. 26, pp. 2364-2373, 2005.

9. E. Abdu and D. Salane, "A Spectral-Based Clustering Algorithm for Categorical Data Using Data Summaries," Proc. Workshop Data Mining using Matrices and Tensors, pp. 1-8, 2009.

10. B. Mirkin, "Reinterpreting the Category Utility Function," Machine Learning, vol. 45, pp. 219-228, 2001.

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. M. Law, A. Topchy, and A.K. Jain, "Multi objective Data Clustering," Proc. IEEE Conf. Computer Vision and Pattern Recognition, vol. 2, pp. 424-430, 2004.