International Journal of Computer Engineering & Technology (IJCET)
Volume 7, Issue 6, November–December 2016, pp. 64–70, Article ID: IJCET_07_06_008
Available online at
http://www.iaeme.com/ijcet/issues.asp?JType=IJCET&VType=7&IType=6
Journal Impact Factor (2016): 9.3590(Calculated by GISI) www.jifactor.com
ISSN Print: 0976-6367 and ISSN Online: 0976–6375

WEIGHTED TRIPLE-QUALITY (WTQ) ALGORITHM FOR CATEGORICAL DATA

M. Pavithra
Assistant Professor, Department of Computer Science and Engineering,
Jansons Institute of Technology, India

ABSTRACT

Although attempts have been made to solve the problem of clustering categorical data via cluster ensembles, with the results being competitive to conventional algorithms, it is observed that these techniques unfortunately generate a final data partition based on incomplete information. The underlying ensemble-information matrix presents only cluster-data point relations, with many entries being left unknown. The paper presents an analysis that suggests this problem degrades the quality of the clustering result, and it presents a new link-based approach, which improves the conventional matrix by discovering unknown entries through similarity between clusters in an ensemble. In particular, an efficient link-based algorithm is proposed for the underlying similarity assessment. . A new link-based approach, which improves 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. C-Rank link-based algorithm is used to improve clustering quality and ranking clusters in weighted networks.

Key words: Clustering, Categorical Data, Cluster Ensembles, Link-Based Similarity and Data Mining.

Cite this Article: M. Pavithra, Weighted Triple-Quality (WTQ) Algorithm for Categorical Data. *International Journal of Computer Engineering and Technology,* 7(6), 2016, pp. 64–70.
http://www.iaeme.com/ijcet/issues.asp?JType=IJCET&VType=7&IType=6

1. INTRODUCTION

Data clustering is one of the fundamental tools we have for understanding the structure of a data set. It plays a crucial, foundational role in machine learning, data mining, information retrieval, and pattern recognition. Clustering aims to categorize data into groups or clusters such that the data in the same cluster are more similar to each other than to those in different clusters. Many well-established clustering algorithms, such as k-means [2] and PAM [3], have been designed for numerical data, whose inherent properties can be naturally employed to measure a distance (e.g., Euclidean) between feature vectors [4], [5]. However, these cannot be directly applied for clustering of categorical data, where domain values are discrete and have no ordering defined. An example of categorical attribute is sex = {male, female} or shape = {circle, rectangle; . . .}. As a result, many categorical data clustering algorithms have been introduced in recent years, with applications to interesting domains such as protein interaction data [6]. The initial method was developed in [7] by making use of Gower's similarity coefficient [8]. Following that, the k-modes algorithm in [9] extended the conventional k-means with a simple matching dissimilarity measure and a frequency-based method to update centroids (i.e., clusters' representative).

As a single-pass algorithm, Squeezer [10] makes use of a pre-specified similarity threshold to determine which of the existing clusters (or a new cluster) to which a data point under examination is assigned. LIMBO [11] is a hierarchical clustering algorithm that uses the Information Bottleneck (IB) framework to define a distance measure for categorical tuples. Clustering ensembles have emerged as a powerful method for improving both the robustness as well as the stability of unsupervised classification solutions. A cluster ensemble can be defined as the process of combining multiple partitions of the dataset into a single partition, with the objective of enhancing the consensus across multiple clustering results [12].

Therefore, it is difficult for users to decide which algorithm would be the proper alternative for a given set of data. Recently, cluster ensembles have emerged as an effective solution that is able to overcome these limitations, and improve the robustness as well as the quality of clustering results. The main objective of cluster ensembles is to combine different clustering decisions in such a way as to achieve accuracy superior to that of any individual clustering. Examples of well-known ensemble methods are:

- The feature-based approach that transforms the problem of cluster ensembles to clustering categorical data (i.e., cluster labels) [12],
- The direct approach that finds the final partition through relabeling the base clustering results.

This research uniquely bridges the gap between the task of data clustering and that of link analysis. It also enhances the capability of ensemble methodology for categorical data, which has not received much attention in the literature. In addition to the problem of clustering categorical data that is investigated herein, the proposed framework is generic such that it can also be effectively applied to other data types.

2. RELATED WORK

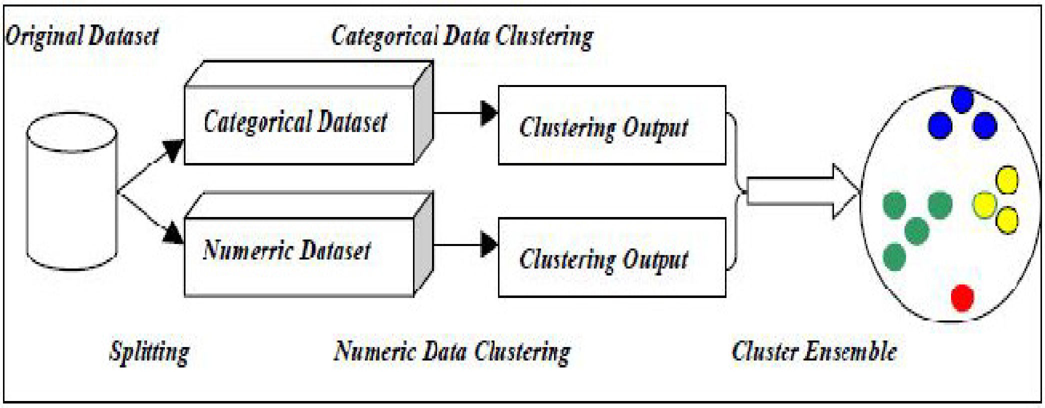
2.1. Link-Based Approach

The categorical data is clustered and represented using the cluster ensembles. 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. The novel link based approach is used to the cluster ensemble problem. The two consensus functions are generated from the RM: feature based partitioning and bipartite graph partitioning. In link based framework, it first creates a refined matrix using links between the cluster points and the two consensus methods are applied to generate the final ensemble cluster. The new link-based approach is used to improve quality of the conventional matrix. It achieves the result using the similarity between clusters and provides cluster view points of the cluster ensemble.

The COOLCAT algorithm is a scalable clustering algorithm that discovers clusters with minimal entropy in categorical data. COOLCAT uses categorical, rather than numerical attributes, enabling the mining of real-world datasets offered by fields such as psychology and statistics. The algorithm is based on the idea that a cluster containing similar points has entropy smaller than a cluster of dissimilar points. Thus, COOLCAT uses entropy to define the criterion for grouping similar objects. LIMBO is a hierarchical clustering algorithm that uses the Information Bottleneck (IB) framework to define a distance measure for categorical tuples [9][10].

The concepts of evolutionary computing and genetic algorithm have also been adopted by a partitioning method for categorical data, i.e., GAClust. Cobweb is a model-based method primarily exploited for categorical data sets. Different graph models have also been investigated by the STIRR, ROCK, and CLICK techniques. In addition, several density-based algorithms have also been devised for such purpose, for instance, CACTUS, COOLCAT, and CLOPE. The Cluster-based Similarity Partitioning Algorithm (CSPA) induces a graph from a co-association matrix and clusters it using the METIS algorithm

[12][13]. Hyper graph partitioning algorithm (HGPA) represents each cluster by a hyper edge in a graph where the nodes correspond to a given set of objects.



3. PROPOSED WORK

3.1. Module Description

3.1.1. Cluster Ensembles of Categorical Data

A cluster ensemble consists of different partitions. Such partitions can be obtained from multiple applications of any single algorithm with different initializations, or from the application of different algorithms to the same dataset. Cluster ensembles offer a solution to challenges inherent to clustering arising from its ill-posed nature: they can provide more robust and stable solutions by leveraging the consensus across multiple clustering results, while averaging out emergent spurious structures that arise due to the various biases to which each participating algorithm is tuned.

3.1.2. Creating a Cluster Ensemble Type I (Direct Ensemble)

The First type of cluster ensemble transforms the problem of categorical data clustering to cluster ensembles by considering each categorical attribute value (or label) as a cluster in an ensemble. Let X = { x_1 …. x_n} be a set of N data points, A ={a_1…. a_m} be a set of categorical attributes, and π = {π_1……π_M} be a set of M partitions. Each partition π_i is generated for a specific categorical attribute a_i € A.

3.1.3. Generating a Refined Matrix

Generating a refined cluster-association matrix (RM) using a link-based similarity algorithm. Cluster ensemble methods are based on the binary cluster-association matrix. Each entry in this matrix represents a association degree between data point. Refined cluster-association matrix is put forward as the enhanced variation of the original BM. Its aim is to approximate the value of unknown associations (“0”) from known ones (“1”), whose association degrees are preserved within the RM [12][14].

3.1.4. Weighted Triple-Quality (WTQ): A New Link-Based Similarity Algorithm

The Weighted Triple-Quality algorithm is efficient approximation of the similarity between clusters in a link network. WTQ aims to differentiate the significance of triples and hence their contributions toward the underlying similarity measure A cluster ensemble of a set of data points X, a weighted graph G =(V,M) can be constructed, where V is the set of vertices each representing a cluster and W is a set of weighted edges between clusters . The weight assigned to the edge that connects clusters is estimated by the proportion of their overlapping [9][11]. Members Shared neighbours have been widely recognized as the basic evidence to justify the similarity among vertices in a link network. For WTQ can be modified to discriminate the quality of shared triples between a pair of clusters in question [14]. The quality of each cluster is determined by the rarity of links connecting to other clusters in a network.

$$W_{xy} = \frac{|L_x \cap L_y|}{|L_x \cup L_y|}$$

```
ALGORITHM: WTQ ($G,C_x,C_y$)
G =(V,W), a weighted graph, where $C_x,C_y$ € V ;
$N_k$ ς V , a set of adjacent neighbors of $C_k$ € V ;
$W_k = \sum W_{tk}$;
        Ct € $N_k$
$WTQ_{xy}$, the WTQ measure of $C_x$ {and} $C_y$;
(1) $WTQ_{xy}$ <------------- 0
(2) For each c € $N_x$
(3) If c € $N_y$
(4) $WTQ_{xy}$ <-------$WTQ_{xy}$ + ($1/W_e$)
(5) Return $WTQ_{xy}$
```

Following that, the similarity between clusters C_x and C_y can be estimated by,

$$Sim(C_x,C_y) = \frac{WTQ_{xy}}{WTQ_{max}} * DC$$

4. EXPERIMENTAL RESULTS

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

4.1. Investigated Data Sets

The experimental evaluation is conducted over two data sets. The “UCI Machine learning repoistory” data set is a subset of the well known attribute values collection— UCI Machine learning repository.

4.2. 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.

Table 1 Summary of Datasets

Datasets	Number of Instances	Number of Attributes	Number of Classes	Missing Values	Area
Breast Cancer	1484	12	8	NIL	Life

4.3. Illustration: Breast Cancer Dataset: 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 in the Table 2. The classification accuracy of standard methods (CO+SL, CO+AL and WTQ) and proposed method (C-Rank) based on number of samples is 300.

Table 1 Comparison of Classification Accuracy of standard and proposed methods based on number of samples = 200.

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**
3	Type I	68.75	69.84	85.06	92.09
		73.93	75.84		
	Type II	79.66	80.20	86.85	92.64
	Type III			87.78	92.76
4	Type I	68.96	69.91	85.12	93.66
		73.94	74.85		
	Type II	79.65	84.66	85.60	94.45
	Type III			90.77	94.56
5	Type I	68.80	69.88	84.67	93.51
		74.21	76.65		
	Type II	77.23	83.78	86.39	95.30
	Type III			88.37	95.43
6	Type I	68.67	69.72	84.56	94.34
		74.77	79.72		
	Type II	82.56	88.45	87.89	95.48
	Type III			91.42	95.79

4.4. Precision

Precision is the degree of refinement in the performance of an operation (procedures and instrumentation) or in the statement of a result.

$$\text{Precision } (i,j)= n_{ij} / n_j \qquad (2)$$

where,

n_{ij} = number of member of class i in cluster j.

n_j = number of members of cluster j.

The Precision 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 Precision value when compared to other standard methods(CO+SL,CO+AL,WTQ)are shown in the Table 2.

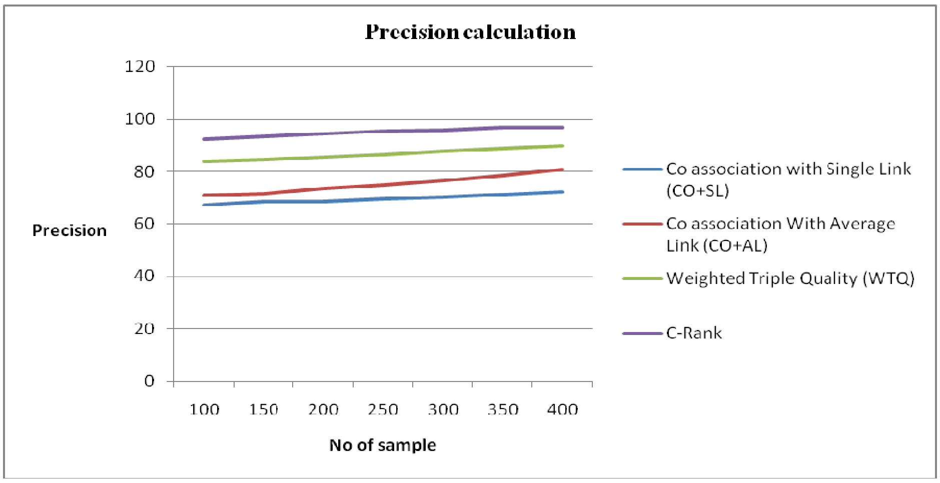


Figure 1 Graph for Performance of precision based on number of samples

The above graph in the Fig.1 shows that if number of sample is more then precision value for proposed methods (C-Rank) has increased up to 97.76% . The precision value for standard methods (CO+SL,CO+AL,WTQ) are slightly less when compared to proposed methods.

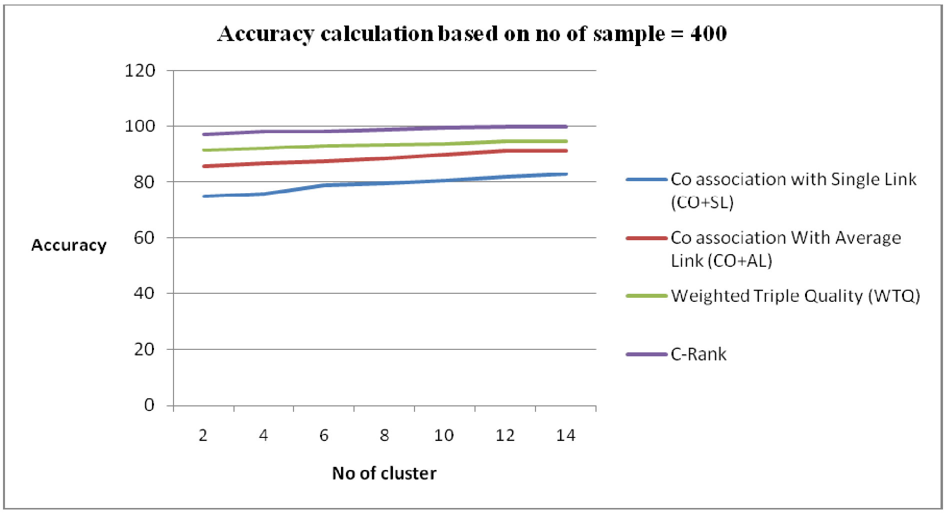


Figure 2 Graph for Performance of Accuracy based on number of sample = 400.

The above graph in the Fig.2 shows that if number of sample is more then accuracy value for proposed methods (C-Rank) has increased up to 99.99% . The accuracy value for standard methods (CO+SL,CO+AL,WTQ) are slightly less when compared to proposed methods.

5. 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.

6. 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. C-Rank link-based algorithm is used to improve clustering quality and ranking clusters in weighted networks. C-Rank consists of three major phases: (1) identification of candidate clusters; (2) ranking the candidates by integrated cohesion; and (3) elimination of non-maximal clusters. Finally apply this clustering result in graph partitioning technique is applied to a weighted bipartite graph that is formulated from the refined matrix.

REFERENCE

[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, "Multiobjective Data Clustering," Proc. IEEE Conf. Computer Vision and Pattern Recognition, vol. 2, pp. 424-430, 2004.

[13] M. Al-Razgan, C. Domeniconi, and D. Barbara, "Random Subspace Ensembles for Clustering Categorical Data," Supervised and Unsupervised Ensemble Methods and Their Applications, pp. 31-48,Springer, 2008.

[14] Natthakan Iam-On, Tossapon Boongoen, Simon Garrett, and Chris Price," A Link-Based Cluster Ensemble Approach for Categorical Data Clustering," IEEE Transactions On Knowledge And Data Engineering, Vol. 24, NO. 3, March 2012.

[15] Aswathy P Sreevatsan and Diya Thomas, A Survey on Weighted Clustering Techniques in MANETS, *International Journal of Computer Engineering and Technology (IJCET),* 5(12), 2014, pp. 15–22.

[16] Neeti Arora and Dr.Mahesh Motwani, A Distance Based Clustering Algorithm, *International Journal of Computer Engineering and Technology (IJCET)*, 5(5), 2014, pp. 109–119.