SIMILARITY MEASURES USING C-RANK CLUSTER APPROACH FOR CATEGORICAL DATA

M. Pavithra

Jansons Institute of Technology, India.

Email: smilepavi17@gmail.com

Abstract: Clustering is 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. The problem of clustering categorical data is to find a new partition in dataset. The underlying ensemble-information matrix presents only cluster-data point relations, with many entries being left unknown. This problem degrades the quality of the clustering result. 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. 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.

Keywords: Clustering, Data mining, Categorical data, Cluster Ensemble, link-based similarity, Refined matrix, C-Rank link based cluster.

Reference to this paper should be made as follows: M. Pavithra (2006) 'Similarity Measures Using C-Rank Cluster Approach For Categorical Data', *International Journal of Infinite Innovations in Engineering and Technology*, Vol. 3, No. 9, pp.01–05.

1 Introduction

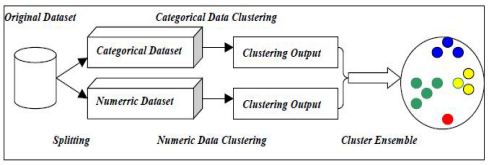
DATA clustering is one of the fundamental tools we have for understanding the structure of a data set. Clustering is a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters[14]. Clustering often is a first step in data analysis. Clustering is the process of discovering homogeneous groups or clusters according to a given similarity measure. Clustering maximizes intra-connectivity among patterns in the same cluster while minimizing inter-connectivity between patterns in different clusters[7]. Clustering is an important technique in discovering meaningful groups of data points. Clustering provides speed and reliability in grouping similar objects in very large datasets[3].Most previous clustering algorithms focus on numerical data whose inherent geometric properties can be exploited naturally to define distance functions between data points. However, much of the data existed in the databases is categorical, where attribute values can't be naturally ordered as numerical values[4]. An example of categorical attribute is shape whose values include circle, rectangle, ellipse, etc. Consensus clustering can provide benefits beyond what a single clustering algorithm can achieve. Consensus clustering algorithms often: generate better clustering's; find a combined clustering unattainable by any single clustering algorithm. The consensus clustering algorithm can be applied to the ensemble of all clustering's produced by discrete features of the data set[10]. Cluster ensemble (CE) is the method to combine several runs of different clustering algorithms to get a common partition of the original dataset, aiming for consolidation of results from a portfolio of individual clustering results[3]. 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[6]. 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[5].

2. Related Work

Clustering of categorical data has recently attracted the attention of many researchers[5][6]. The k-modes algorithm is an extension of k-means for categorical features. To update the modes during the clustering process, the authors used a new distance measure based on the number of mis-matches between two points. Squeezer is a categorical clustering algorithm that processes one point at the time. ROCK (Robust Clustering using links) is a hierarchical clustering algorithm for categorical data. It uses the Jaccard coefficient to compute the distance between points[7][8]. 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

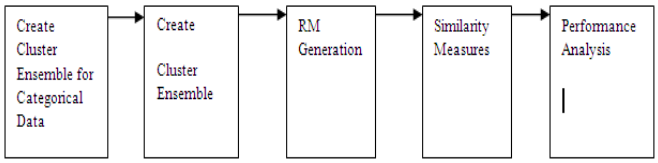
based on the idea that a cluster containing similar points has an entropy smaller than a cluster of dissimilar points. Thus, COOLCAT uses entropy to define the criterion for grouping similar objects. LIMBO is a hierarchical clustering algorithm that uses the Information Bottleneck (IB) framework to define 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.

Figure 1. Overview of Cluster Ensemble Algorithm Framework.



Proposed Work

Figure.2 shows the proposed rule based similarity using link based cluster approach . A brief explanation about its phases is followed.



Module Description

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.

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 = { x1 …. xn} be a set of N data points, A ={a1…. am} be a set of categorical attributes, and π = {π1……πM} be a set of M partitions. Each partition πi is generated for a specific categorical attribute ai € A. Clusters belonging to a partition πi = { Ci1,……,Ciki} correspond to different.

Type II (Full-space ensemble)

In this two ensemble types are created from base clustering results, each of which is obtained by applying a clustering algorithm to the categorical data set. In particular to a full-space ensemble, base clusterings are created from the original data, i.e., with all data attributes. To introduce an artificial instability to k-modes, the following two schemes are employed to select the number of clusters in each base clusterings: 1) Fixed-k, k= ⌐√N (where N is the number of data points), and 2) Random-k, k €{2,…….. ⌐√N¬ }[5].

Type III: 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. 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.

3. Generating a Refined Matrix

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].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.

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].

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

| Lx U Ly |

ALGORITHM: WTQ(G,Cx,Cy)

G =(V,W), a weighted graph, where Cx,Cy € V ;
Nk ς V , a set of adjacent neighbors of Ck € V ;
Wk = ∑ Wtk;
Ct € Nk
WTQxy, the WTQ measure of Cx {and} Cy;
(1) WTQxy <---------- 0
(2) For each c € Nx
(3) If c € Ny
(4) WTQxy <-------WTQxy + (1/We)
(5) Return WTQxy

Following that, the similarity between clusters Cx and Cy can be estimated by,

$$Sim(Cx,Cy) = \frac{WTQxy}{WTQmax} * DC$$

5. Connector-based similarity measure

Connector-based similarity measure called C-Rank. C-Rank uses both in-links and out-links at the same time. A new link-based similarity measure called C-Rank, which uses both in-link and out-link by disregarding the direction of references. C-Rank, where L(p) denotes the set of undirected link neighbors of paper p. Similar to that the accuracy of Co-citation (Coupling) is improved by iterative SimRank (rvs-SimRank), Crank is defined iteratively. C-Rank unifies in-links and out-links into undirected links. C-Rank has the effect similar to increasing the weight of Co-citation (SimRank) when computing the score between old papers, increasing the weight of Coupling (rvs-SimRank) when computing the score between recent papers, and increasing the weight of a BP-based similarity measure when computing the score between old and recent papers.

1: Procedure unweighted CRank(G, L)
2: add G to L
3: if G is a clique or a singleton return
4: S: = sparsest vertex separator of G
5: A1, . . . ,Ak := connected components of G \ S
6: for i = 1 to k do
7: Gi: = sub-network of G induced on S [Ai
8: if Gi not already in L then
9: unweightedCRank(Gi , L)

Performance Evaluation

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.

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.

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
Primary Tumour	699	10	10	NIL	Life

Breast Cancer Datasets:

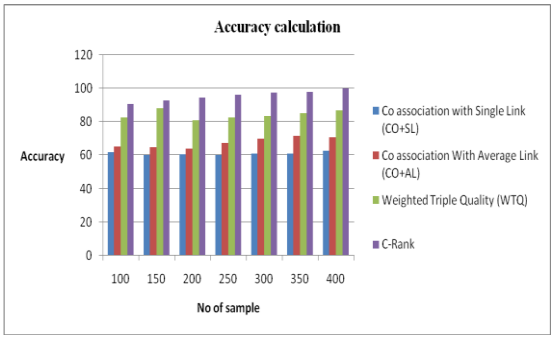


Fig.1 Graph for Performance of Accuracy based on number of samples

The above graph in the Fig.1 shows that the if number of sample is 100 then it shows the accuracy value for both proposed and standard methods in bar chart format. If number of sample is 400 then accuracy for proposed methods is 99.99 % but for standard method like WTQ has reached 90% , other standard methods are less when compared to proposed methods.

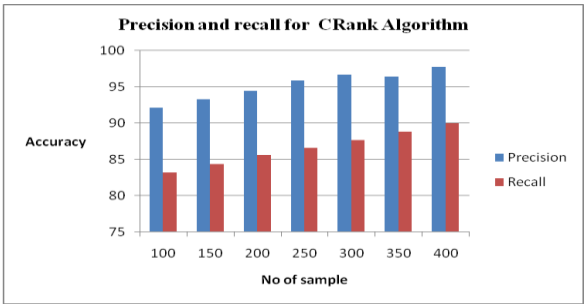


Fig.2 Graph for comparison of precision and recall for C-Rank algorithm

The above graph in the Fig.15 shows that comparison of precision and recall for C-Rank algorithm is that if number of sample are more then precision value is also gets increased when compared to recall value. If number of sample is 400 then precision value has reached 97% when compared to recall value by using C-Rank algorithm.

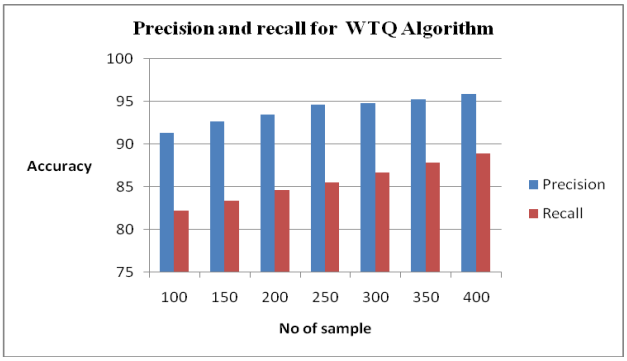


Fig.3 Graph for comparison of precision and recall for WTQ algorithm

If the number of sample are more then precision value is also gets increased when compared to recall value. If number of sample is 400 then precision value has reached 96% when compared to recall value by using WTQ algorithm which is in the Fig.3.

Primary Tumour Datasets:

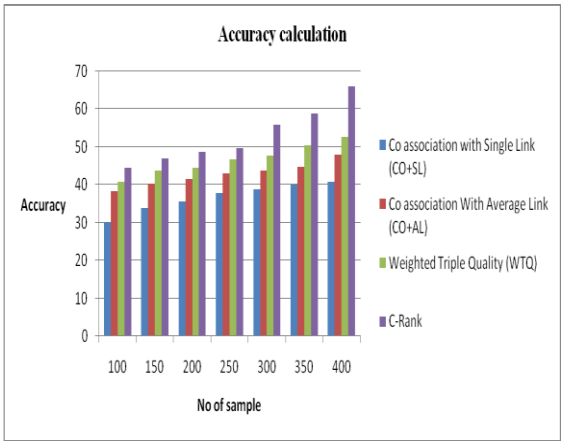


Fig.4 Graph for Performance of Accuracy based on number of samples

If the number of sample is 100 then it shows the accuracy value for both proposed and standard methods in bar chart format. If number of sample is 400 then accuracy for proposed methods is 68.66 % but for standard method like WTQ has reached 53.80 % , other standard methods are less when compared to proposed methods which is shown in the Fig.4.

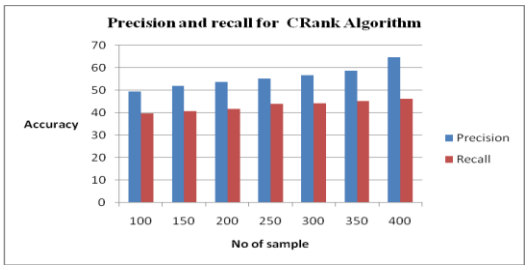


Fig.5 Graph for comparison of precision and recall for C-Rank algorithm

The comparison of precision and recall for C-Rank algorithm is that if number of sample are more then precision value is also gets increased when compared to recall value. If number of sample is 400 then precision value has reached 65 % when compared to recall value by using C-Rank algorithm which is shown in the Fig.5.

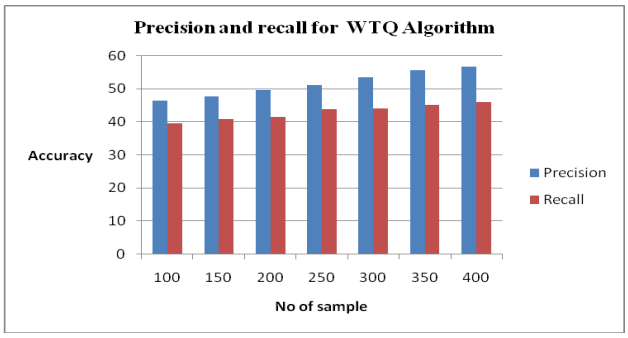


Fig.6 Graph for comparison of precision and recall for WTQ algorithm

The comparison of precision and recall for WTQ algorithm is that if number of sample are more then precision value is also gets increased when compared to recall value. If number of sample is 400 then precision value has reached 56% when compared to recall value by using WTQ algorithm which is shown in the Fig.6.

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. 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.

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, "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.A. Topchy, A. K. Jain, and W. Punch. Combining multiple weak clusterings. In Proceedings IEEE International Conference on Data Mining, pages 331–338, 2003.

16.C. L. Blake and C. J. Merz. UCI repository of machine learning databases, 1998. URL http://www.ics.uci.edu/~mlearn/MLRepository.html.