A New Link Based Approach for Categorical Data Clustering

Chiranth B.O[1], Panduranga Rao M.V[2], Basavaraj Patil S[3]

[1]Dept. of Computer science and Engineering,
B.T.L Institute of Technology, Bangalore, India
bo.chiranth@gmail.com

[2]Dept. of Computer Science and Engineering,
B.T.L Institute of Technology, Bangalore, India
raomvp@yahoo.com

[3]Dept of Computer Science and Engineering,
B.T.L Institute of Technology, Bangalore, India
csehodbtlit@gmail.com

Abstract: ***The data generated by conventional categorical data clustering is incomplete because the information provided is also incomplete. This project presents a new link-based approach, which improves the categorical clustering by discovering unknown entries through similarity between clusters in an ensemble. A graph partitioning technique is applied to a weighted bipartite graph to obtain the final clustering result. So the link-based approach outperforms both conventional clustering algorithms for categorical data and well-known cluster ensemble technique.Data clustering is one of the fundamental tools we have for understanding the structure of a data set. It plays a crucial, foundation role in machine learning, data mining, information retrieval and pattern recognition. The experimental results on multiple real data sets suggest that the proposed link-based method almost always outperforms both conventional clustering algorithms for categorical data and well-known cluster ensemble technique. This paper proposes an Algorithm called Weighted Triple-Quality (WTQ), which also uses k-means algorithm for basic clustering. Once using does the basic clustering consensus functions we can get cluster ensembles of categorical data. This categorical data is converted to refined matrix.***

Keywords: Clustering, categorical data, cluster ensembles, link-based similarity, data mining

1. Introduction

Dataclustering 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 [1] and PAM [2], have been designed for numerical data, whose inherent properties can be naturally employed to measure a distance (e.g., Euclidean) between feature vectors [3], [4]. 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={female, male} 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 [5]. The initial method was developedin [6] by making use of Gower's similarity coefficient [7]. Following that, the k-modes algorithm in [8] 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 [9] makes use of a prespecified similarity threshold to determine which of the existing clusters (or a new cluster) to which a data point under examination is assigned. LIMBO [10] is a hierarchical clustering algorithm that uses the Information Bottleneck (IB) framework to define a distance measure for Categorical tuples. The concepts of evolutionary computing and genetic algorithm have also been adopted by a partitioning method for categorical data, i.e., GAClust [11]. Cobweb [12] is a model-based method primarily exploited for categorical data sets. Different graph models have also been investigated by the STIRR [13], ROCK [14], and CLICK [15] techniques. In addition, several density-based algorithms have also been devised for such purpose, for instance, CACTUS [16], COOLCAT [17], and CLOPE [18].

Although, a large number of algorithms have been introduced for clustering categorical data, the No Free Lunch theorem [19] suggeststhere is no single clustering algorithm that performs best for all data sets [20] and can discover all types of cluster shapes and structures presented in data. Each algorithm has its own strengths and weaknesses. For a particular data set, different algorithms, or even the same algorithm with different parameters, usually provide distinct solutions. 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 ableto 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:

1. The feature-based approach that transforms the problem of cluster ensembles to clustering categorical data (i.e., cluster labels) [11],

2. The direct approach that finds the final partition through relabeling the base clustering results [15], [16],

3. Graph-based algorithms that employ graph partitioning methodology [17], [18], [19], and

4. The pairwise-similarity approach that makes use of co-occurrence relations between data points [20], [11], [12].

2. Cluster Ensemble Methods

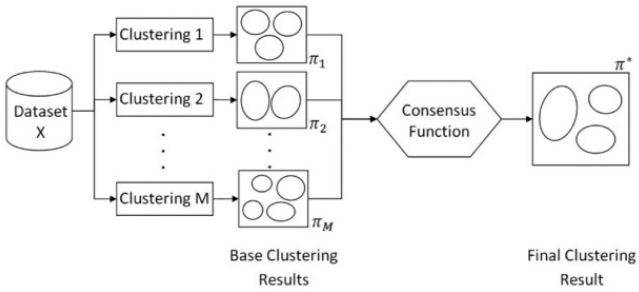

Figure 1. The Basic Process of Cluster Ensembles

It has been shown that ensembles are most effective when constructed from a set of predictors whose errors are dissimilar [17]. To a great extent, diversity among ensemble members is introduced to enhance the result of an ensemble [18]. Particularly for data clustering, the results obtained with any single algorithm over much iteration are usually very similar. In such a circumstance where all ensemble members agree on how a data set should be partitioned, aggregating the base clustering results will show no improvement over any of the constituent members. As a result, several heuristics have been proposed to introduce artificial instabilities in clustering algorithms, giving diversity within a cluster ensemble. The following ensemble generation methods yield different clustering of the same data, by exploiting different cluster models and different data partitions.

1. Homogeneous ensembles. Base clustering are created using repeated runs of a single clustering algorithm, with several sets of parameter initializations, such as cluster centers of the k-means clustering technique [11], [20], [19].

2. Homogeneous ensembles. Base clustering are created using repeated runs of a single clustering algorithm, with several sets of parameter initializations, such as cluster centers of the k-means clustering technique [11], [20], [19].

3. Data subspace/sampling. A cluster ensemble can also be achieved by generating base clustering from different subsets of initial data. It is intuitively assumed that each clustering algorithm will provide different levels of performance for different partitions of a data set [17].Data subspace/sampling. A cluster ensemble can also be achieved by generating base clustering from different subsets of initial data. It is intuitively assumed that each clustering algorithm will provide different levels of performance for different partitions of a data set [17].

4. Data subspace/sampling. A cluster ensemble can also be achieved by generating base clustering from different subsets of initial data. It is intuitively assumed that each clustering algorithm will provide different levels of performance for different partitions of a data set [17].

5. Mixed heuristics. In addition to using one of the aforementioned methods, any combination of them can be applied as well [17], [11], [3], [8], [20], [2], [19].

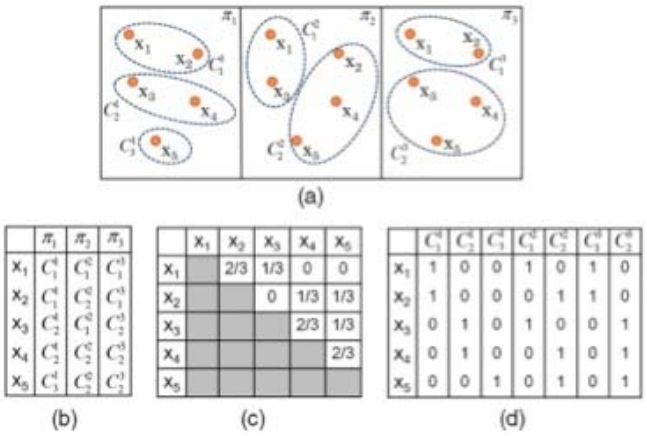

Figure 2. Examples of (a) cluster ensemble and the corresponding (b) label assignment matrix, (c) pair wise similarity matrix, and (d) binary cluster association matrix respectively.

3. A New Link Based Approach

Existing cluster ensemble methods to categorical data analysis rely on the typical pairwise-similarity and binary cluster-association matrices [8], [9], which summarize the underlying ensemble information at a rather coarse level. Many matrix entries are left “unknown” and simply recorded as “0.” Regardless of a consensus function, the quality of the final clustering result may be degraded. As a result, a link-based method has been established with the ability to discover unknown values and, hence, improve the accuracy of the ultimate data partition [13].

In spite of promising findings, this initial framework is based on data point- data point pairwise-similarity matrix, which is highly expensive to obtain. The link-based similarity technique, SimRank [2] that is employed to estimate the similarity among data points is inapplicable to a large data set. To overcome these problems, a new link-based cluster ensemble (LCE) approach is introduced herein. It is more efficient than the former model, where a BM-like matrix is used to represent the ensemble information.

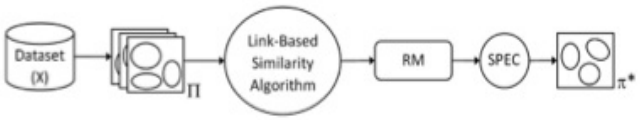

Figure 3. A Link Based Cluster Framework

4. Weighted Triple Quality (WTQ): A New Link Based Algorithm

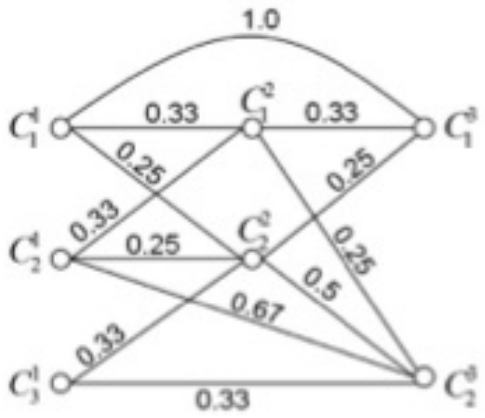

Figure 4. An Example of Cluster Network

Algorithm: WAQ (G, Cx, Cy):

$G = (V, W)$, a weighted graph, where $C_x, C_y \in V$;
$N_k \subset V$, a set of adjacent neighbors of $C_k \in V$;
$W_k = \sum_{\forall C_t \in N_k} w_{tk}$;
WTQ_{xy}, the WTQ measure of C_x {and} C_y;
(1) $WTQ_{xy} \leftarrow 0$
(2) **For each** $c \in N_x$
(3) **If** $c \in N_y$
(4) $WTQ_{xy} \leftarrow WTQ_{xy} + \frac{1}{W_c}$
(5) **Return** WTQ_{xy}

5. Experimental Design

The experiments set out to investigate the performance of LCE compared to a number of clustering algorithms, both specifically developed for categorical data analysis and those state-of-the-art cluster ensemble techniques found in literature. Baseline model is also included in the assessment, which simply applies SPEC, as a consensus function, to the conventional BM. For comparison, as in [8], each clustering method divides data points into a partition of K (the number of true classes for each data set) clusters, which is then evaluated against the corresponding true partition using the following set of label-based evaluation indices: Classification Accuracy (CA) [3], Normalized Mutual Information (NMI) [9] and Adjusted Rand (AR) Index [9]. Further details of these quality measures are provided in Section I of the online supplementaryNote that, true classes are known for all data sets but are explicitly not used by the cluster ensemble process. They are only used to evaluate the quality of the clustering results.as follows: Insert/Break/Continuous.

6. Conclusion

This paper presents a novel, highly effective link-based cluster ensemble approach 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. The prominent future work includes an extensive study regarding the behavior of other link-based similarity measures within this problem context. Also, the new method will be applied to specific domains, including tourism and medical data sets.

References

[1] D.S. Hochbaum and D.B. Shmoys, “A Best Possible Heuristic for the K-Center Problem,” Math. of Operational Research, vol. 10, no. 2, pp. 180-184, 1985. R. Caves, Multinational Enterprise and Economic Analysis, Cambridge University Press, Cambridge, 1982. (book style)
[2] D.S. Hochbaum and D.B. Shmoys, “A Best Possible Heuristic for the K-Center Problem,” Math. of Operational Research, vol. 10, no. 2, pp. 180-184, 1985.
[3] A.K. Jain and R.C. Dubes, Algorithms for Clustering. Prentice-Hall, 1998.
[4] P. Zhang, X. Wang, and P.X. Song, “Clustering Categorical Data Based on Distance Vectors,” The J. Am. Statistical Assoc., vol. 101, no. 473, pp. 355-367, 2006.
[5] J. Grambeier and A. Rudolph, “Techniques of Cluster Algo- rithms in Data Mining,” Data Mining and Knowledge Discovery, vol. 6, pp. 303-360, 2002.
[6] K.C. Gowda and E. Diday, “Symbolic Clustering Using a New Dissimilarity Measure,” Pattern Recognition, vol. 24, no. 6, pp. 567- 578, 1991.
[7] J.C. Gower, “A General Coefficient of Similarity and Some of Its Properties,” Biometrics, vol. 27, pp. 857-871, 1971.
[8] Z. Huang, “Extensions to the K-Means Algorithm for Clustering Large Data Sets with Categorical Values,” Data Mining and Knowledge Discovery, vol. 2, pp. 283-304, 1998.
[9] 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.
[10] P. Andritsos and V. Tzerpos, “Information-Theoretic Software Clustering,” IEEE Trans. Software Eng., vol. 31, no. 2, pp. 150-165, Feb. 2005.
[11] D. Cristofor and D. Simovici, “Finding Median Partitions Using Information-Theoretical-Based Genetic Algorithms,” J. Universal Computer Science, vol. 8, no. 2, pp. 153-172, 2002.
[12] D.H. Fisher, “Knowledge Acquisition via Incremental Conceptual Clustering,” Machine Learning, vol. 2, pp. 139-172, 1987.
[13] D. Gibson, J. Kleinberg, and P. Raghavan, “Clustering Catego- rical Data: An Approach Based on Dynamical Systems,” VLDB J., vol. 8, nos. 3-4, pp. 222-236, 2000.
[14] 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.
[15] M.J. Zaki and M. Peters, “Clicks: Mining Subspace Clusters in Categorical Data via Kpartite Maximal

Cliques," Proc. Int'l Conf. Data Eng. (ICDE), pp. 355-356, 2005.
[16] 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.
[17] 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.
[18] 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.
[19] D.H. Wolpert and W.G. Macready, "No Free Lunch Theorems for Search," Technical Report SFI-TR-95-02-010, Santa Fe Inst., 1995.
[20] L.I. Kuncheva and S.T. Hadjitodorov, "Using Diversity in Cluster Ensembles," Proc. IEEE Int'l Conf. Systems, Man and Cybernetics, p p. 1214-1219, 2004.

Chiranth B Oreceived the B.E. degree in Information Science and Engineering from Golden Valley Institute of Technology, KGF. At present pursuing the Master of Technology in Computer Science and Engineering Department at BTL institute of Technology, Bangalore.

M.V.Panduranga Rao is a research scholar at National Institute of Technology Karnataka, Mangalore, India. His research interests are in the field of Real time and Embedded systems on Linux platform and Security. He has published various research papers across India and in IEEE international conference in Okinawa, Japan. He has also authored two reference books on Linux Internals. He is the Life member of Indian Society for Technical Education and IAENG.

S. Basavaraj Patil Started career as Faculty Member in Vijayanagar Engineering College, Bellary (1993-1997). Then moved to Kuvempu University BDT College of Engineering as a Faculty member. During this period (1997-2004), carried out Ph.D. Research work on Neural Network based Techniques for Pattern Recognition in the area Computer Science & Engineering. Also consulted many software companies (ArisGlobal, Manthan Systems, etc.,) in the area of data mining and pattern recognition His areas of research interests are Neural Networks and Genetic Algorithms. He is presently mentoring two PhD and one MSc (Engg by Research) Students. He is presently working as professor at BTL institute of technology.