Clustering Categorical Data using Silhouette Coefficient as a Relocating Measure

S.Aranganayagi[1*], K.Thangavel [2]

1. J.K.K.Nataraja College of Arts & Science, Komarapalayam- 638 183
*Department of Computer Science and Applications,
Gandhigram Rural University
Gandhigram – 624 302, Tamil Nadu, India.
E-mail : arangbas@gmail.com

2.Department of Computer Science, Periyar University
Salem – 636 011, Tamil Nadu, India..
E-mail : drktvelu@yahoo.com

Abstract

Cluster analysis is an unsupervised learning method that constitutes a cornerstone of an intelligent data analysis process. Clustering categorical data is an important research area data mining. In this paper we propose a novel algorithm to cluster categorical data. Based on the minimum dissimilarity value objects are grouped into cluster. In the merging process, the objects are relocated using silhouette coefficient. Experimental results show that the proposed method is efficient.

Keywords: *Data mining, clustering, categorical data, silhouette coefficient.*

1. Introduction.

Clustering has emerged as a popular technique for pattern recognition, image processing and data mining. In context of data mining clustering algorithms are increasingly used as a preprocessing tool for other data mining techniques. Data mining primarily works with large data bases. The attributes in the data set may be of numerical categorical or both numerical and categorical. Numerous algorithms exist to cluster the numerical data.

Clustering is the process of grouping the similar objects into groups. Basic principle of clustering hinges on a concept of distance metric or similarity metric. Most of the earlier work on clustering has been focused on numerical data, whose inherent geometric properties can be exploited to define the distance between two points.[1,2,7] Categorical data set are with values like {male, female}. It is not possible to find the distance between the male and female. K-Means algorithm is widely used to cluster numerical ones. Huang proposed an algorithm called K-modes which is an extension of K-means, instead of means he used the concept of modes. The simple matching measure is used to cluster the objects. The categorical clustering algorithm recently proposed are STIRR, CACTUS, COOLCAT, LIMBO, K-representative, Squeezer, K-histograms etc.[4, 8,10,11,12].

In this paper we propose an efficient categorical clustering algorithm using the concept of dissimilarity matrix and neighborhood concept. The proposed algorithm is an agglomerative clustering technique, where the clusters are merged based on the neighbor matrix. The placed objects are relocated using silhouette coefficient. The experimental results verify the efficiency of the algorithm.

In section 2 the related algorithm on categorical data is discussed. Section 3 briefs the definitions and notations used. Section 4 deals with the proposed algorithm. Section 5 deals with results and discussions and section 6 concludes the paper.

0-7695-3050-8/07 $25.00 © 2007 IEEE
DOI 10.1109/ICCIMA.2007.328

2. Related works.

In this section the past work on clustering categorical data are discussed. K-Prototypes algorithm is based on K-means but removes the numeric data limitation. It is applicable for both numeric and categorical data. The K-modes algorithm is used to cluster the data using modes. The Expectation Maximization(EM) algorithm is a popular iterative clustering technique[9]. Robust hierarchical Clustering with linKs (ROCK) is an adaptation of an agglomerative hierarchical clustering algorithm, which heuristically optimizes a criterion function defined in terms of the number of links between tuples. Informally the number of links between two tuples is the number of common neighbors they have in the dataset [2,7]. Clustering Categorical Data Using Summaries (CACTUS) attempts to split the database vertically and tries to cluster the set of projections of these tuples to only a pair of attributes [10]. The COOLCAT algorithm uses the entropy measure in clustering. The LIMBO algorithm clusters the categorical data using information bottle neck as a measure[11]. By varying the dissimilarity measure, fuzzy K-modes, K-representative and K-histogram is developed. In fuzzy K-modes, instead of hard centroid, soft centroid is used [3]. In K-representative[8] algorithm by Ohm mar sen et al, the measure relative frequency is used. Frequency of attribute value in the cluster divided by cluster length is used as a measure in K-representative. In K-histograms[11], the frequency of attribute value in the cluster is divided by frequency of attribute value in the data set is used as a measure.

3. Definitions and notations.

Definition 1: Dissimilarity Measure[13]

Let T = $\{O_1, O_2, \ldots O_n\}$ and O_i and O_j be the two categorical objects with m attributes. The dissimilarity between two distinct objects is defined as

$$d(O_i, Oj) = \sum_{k=1}^{m} \delta(x_k, y_k) \quad \text{where } 1 \le i \le n,\ 1 \le j \le n,\ i \neq j$$

and

$$\delta(x_k, y_k) = \begin{cases} 0 & \text{if } (x_k = y_k) \\ 1 & \text{if } (x_k \neq y_k) \end{cases}$$

where *d* is a dissimilarity matrix of size n x n and n is the number of objects.

Definition2: Minimum dissimilarity measure

Minimum dissimilarity value for each object O_i , is defined as

$mrow(O_i) = \text{minimum}(d(O_i, O_j)) \quad 1 \le j \le n$

Definition 3: Neighbor Objects

The neighbor of objects 'neigh' is defined as,

$neigh(O_i) = \{ O_j / mrow(O_i) = d(O_i, Oj) \}$ where i = 1,2…n, j = 1,2…n

Definition 4: Clustering Criteria

Select an object O_i from the list of objects and let C_k be the cluster consisting of the objects O_j such that $mrow(O_j) \geq mrow(O_i)$ where $O_j \in neigh(O_i)$.

Definition 5: Cluster Merging Criteria

Let $C_j = \{O_{j1}, O_{j2}, O_{j3}, \ldots O_{jkj}\}$ be the cluster to be merged with suitable clusters: For each $O_{ji} \in C_j$, Neighbor Count (NC) of object O_{ji} is defined as, $NC(O_{ji}) = \{ count(C_l) / count(C_l) = | \{ neigh(O_{ji}) \in C_l \} | \}$ and the new cluster for object O_{ji} is defined as $newcluster(O_{ji}) = C_l$ such that $C_l = maximum\{ count(C_l) \}$

Definition 6: Silhouette Coefficient

The silhouette coefficient is a cluster validity measure.
Sil(i) = (b(i) – a(i)) / max(b(i) , a(i)) where a(i) – average dissimilarity between i and the other objects of the cluster to which I belongs. D(i,C) – For all other clusters C, average dissimilarity of i to all observations of C. b(i) - minimum of D(i,C) – dissimilarity between I and the neighbor cluster C. If the Sil(i) is nearer to one the object is placed in the correct cluster. If the Sil(i) is negative the object is placed in the wrong cluster. If it is around 0 the object is between the clusters.[9]

4. Proposed algorithm.

For the selected sample set, the dissimilarity matrix is constructed using Definition-1. Minimum value of each row is found using Definition-2. Neighbor of each object is generated using Definition-3. The proposed method contains 2 phases. First phase results in set of clusters using clustering criteria (Definition 4). Second Phase merges the clusters by relocating the objects using Definition-5. Quality of the clusters is evaluated using silhouette coefficient. Using this coefficient the objects in wrong clusters are placed in correct clusters. Thus the proposed algorithm improves the efficiency of the clusters. The proposed method is experimented with the datasets mushroom and zoo obtained from UCI data repository. The number of cluster resulted is more in number viz., 28 and 22. But all the clusters contain the objects related to same class. Thus the proposed method produces all pure clusters.

Phase I:

The steps involved in this phase are detailed below:
Step 1: Construct a dissimilarity matrix.
Step 2. Compute the threshold value, minimum dissimilarity of each object, mrow(O_j).
Step 3. Construct a neighbour matrix ‘neigh’.
Step 4.. Select the member of an object list O_1, form a new cluster with O_1 as a member. Group the objects[neigh(O_1)] in to clusters based on the criteria given in definition 4. Remove the clustered objects from the object list.
Step 5. Repeat the above step until the object list becomes empty.

Phase II

The steps involved in merging of clusters are detailed below:
Step 1: Select the cluster with least number of objects.
Step 2. The objects in the selected cluster are relocated based on the Cluster Merging Criteria, definition 6.
Step 3. The above steps are repeated until no more merging is possible.
Step 4: Using silhouette coefficient objects are relocated.

5. Results and Discussions.

The algorithm is experimented with real data sets mushroom and zoo data set obtained from UCI data repositories.[14]

5.1 Data Set

Mushroom: Each tuple records the physical characteristics of a single mushroom. The mushroom data set contains 23 attributes with 8124 tuples. Each is classified as either edible or poisonous. The number of edible and poisonous mushroom in the dataset are 4208 and 3916 respectively. We have selected 250 samples from the data set.
Zoo data set: Gives the description of features of animals in the zoo. Zoo data set consists of 101 tuples with 18 attributes, and is classified in to 7 categories.

5.2 Accuracy measure

The clustering accuracy r is defined as, $r = \sum_{i=1}^{k} a_i / n$ where n is number of instances in the data set, a_i is the number of instances occurring in both cluster i and its corresponding class, which has the maximal value. Thus the clustering error is defined as $e = 1 - r$. If a partition has a clustering accuracy of 100%, it means that it has only pure clusters. Large clustering accuracy implies better clustering [13, 15].

Both the data sets results in pure clusters. Thus the error rate is 0%. The objects which are placed in the wrong clusters are relocated using silhouette coefficient. Experimental results show that the proposed method produces effective clusters. For all the objects the silhouette coefficient is more than 0, this is shown in figure 1 and 2.

6. Conclusion.

In many existing methods, the number of clusters 'K' is given as an input parameter. In the proposed algorithm the similarity measurement is used to cluster the categorical data and the objects are clustered without getting any input as 'K' or the threshold value. The objects merged, or relocated from one cluster to another cluster is validated using silhouette coefficient. Further we planned to extend this algorithm to very large data set.

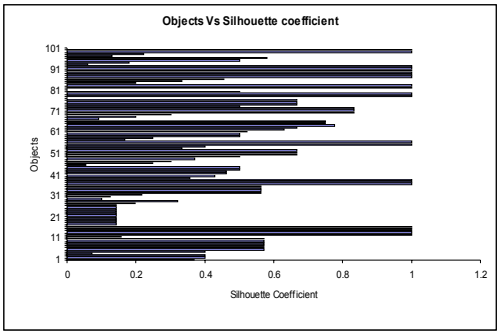

Figure -1 - Silhouette plot for Zoo data set

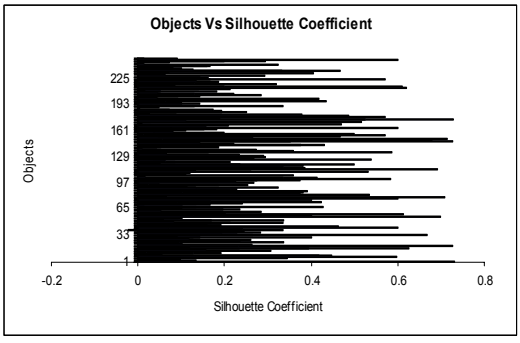

Figure -2 - Silhouette plot for Mushroom data set

References

[1]. P.Adriaans and D. Zantige , Data Mining, UK, Addision Wesley,1996

[2]. Arun.K.Pujari , Data Mining Techniques, , Universities Press,2001

[3]. S.Aranganayagi and K.Thangavel , A Novel Clustering Algorithm For Categorical Data, Computational Mathematics, Narosa Publishing House, New Delhi, India, 2005

[4] Daniel Barbara, Julia Couto, Yi Li, COOLCAT An entropy based algorithm for categorical clustering, Proceedings of the eleventh international conference on Information and knowledge management, 582 - 589 , 2002 [5]. George Karypis, Eui-Hong (Sam) Han, Vipinkumar CHAMELEON: A hierarchial clustering algorithm using dynamic modeling, IEEE Computer, 1999.

[6] J.Graham Williams and Zhexue Huang,(1997) Mining the Knowledge Mine: The hot spot methodology for mining Large Real World databases, Advanced Topics in Artificial Intelligence, Lecture notes in Artificial Intelligence, Vol. 1342,pp 340-348, Springer, Verlag,1997.

[7] Jiawei Han, Micheline Kamber, Data Mining Concepts And Techniques, Harcourt India Private Limited,2006

[8].Ohn Mar San, Van-Nam Huynh, Yoshiteru Nakamori, An Alternative Extension Of The *K*-Means algorithm For Clustering Categorical Data, J. Appl. Math. Comput. Sci Vol. 14, No. 2, 241–247,2004

[9]. Pavel Berkhin, Survey of Clustering Data Mining Techniques, Technical report,Accrue software,2002

[10]Venkatesh Ganti, Johannes Gehrke, Raghu Ramakrishnan. CACTUS –Clustering Categorical Data using summaries, In Proc. of ACM SIGKDD, International Conference on Knowledge Discovery & Data Mining, 1999, San Diego, CA USA.

[11] Zengyou He, Xiaofei Xu, Shengchun Deng, Bin Dong, K-Histograms: An Efficient Algorithm for Catgorical Data set, www.citebase.org.

[12]. Zhexue Huang , A Fast Clustering Algorithm to cluster Very Large Categorical Datasets in Data Mining, In Proc. SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, 1997.

[13]. Zhexue Huang, Extensions to the k-means algorithm for clustering Large Data sets with categorical value, Data Mining and Knowledge Discovery 2, 283-304, Kluwer Academic publishers, 1998.

[14]. www.ics.uci.edu/~mlearn/MLRepository.html