Scientific Research Publishing

Open Journal of Statistics, 2017, 7, 173-181
http://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Clustering Categorical Data Based on Within-Cluster Relative Mean Difference

Jinxia Su, Chunjing Su

School of Mathematics and Statistics, Lanzhou University, Lanzhou, China
Email: jinxiasu@lzu.edu.cn

How to cite this paper: Su, J.X. and Su, C.J. (2017) Clustering Categorical Data Based on Within-Cluster Relative Mean Difference. *Open Journal of Statistics*, **7**, 173-181.
https://doi.org/10.4236/ojs.2017.72013

Received: January 17, 2017
Accepted: April 17, 2017
Published: April 20, 2017

Abstract

The clustering on categorical variables has received intensive attention. In dataset with categorical features, some features show the superior performance on clustering procedure. In this paper, we propose a simple method to find such distinctive features by comparing pooled within-cluster mean relative difference and then partition the data upon such features and give subspace of the subgroups. The applications on zoo data and soybean data illustrate the performance of the proposed method.

Keywords

Clustering, Categorical Variable, Distinctive Attribute, Pooled Within-Cluster Mean Relative Difference, Hamming Distance

1. Introduction

Data clustering is a technique for identifying groups with data instances in the same group which are more similar than the instances belonging to different groups. The issue of database clustering with categorical variables has received intensive attention ([1]-[8]) along with other publications in the same issue. The categorical data come from different areas of research, both social and nature sciences; this type of variable does not present a natural ordering for their possible values, results in the difficulty in clustering process.

There are various algorithms available for clustering categorical data, but no algorithm can achieve the best result for all the data sets. some new techniques have been developed recently, for example CACTUS (CAtegorical Clus Tering Using Summaries, see [9]), ROCK (RObust Clustering using linKs, see [10]), and neural networks based approaches (e.g. self-organizing map network) are used for this purpose. Arias-Castro and Xiao ([11]) proposed a sparse version of clustering method. Zhang *et al.* ([12]) proposed a novel statistical procedure

(HD Vector) for clustering categorical data based on frequency distributions of the hamming distance vector upon comparing with the uniform distributed sample. Unfortunately, their categorical sample space is still complex in computation since the comparison of the total number of possible positions in a categorical sample space is need.

This paper devotes to find the distinctive attributes among the categorical dataset using pooled relative within-cluster mean difference, then the data is clustered upon a single distinctive attribute. At each iteration, our algorithm recognizes one distinctive attribute and then identifies only one cluster with minimum of within-cluster mean relative difference, which will then be deleted from the dataset at the next iteration; this procedure repeats until there are no more significant clusters in the remained data.

The rest of the paper is organized as follows: A motivation example is illustrated in Section 2, and in Section 3, the methodologies are discussed. The performance of the proposed algorithm is explored through a real dataset in Section 4. Section 5 gives our conclusions.

2. Motivation

Considering the soybean disease dataset from the Machine Learning Depository at the University of California at Irvine, it comprises of 47 objects with 35 categorical attributes $\{v_k\}_{k=1}^{35}$ (There are 14 attributes have the identity observed value for all objects, so they can be treated as noninformationed attributes and be suppressed, thereafter only 21 attributes to be considered). As [12] pointed out, these data points come from four clusters: diaporthe stem canker, charcoal rot, rhizoctonia root rot, and phytophthora rot, with sample sizes 10, 10, 10, and 17 respectively. Summary in **Table 1** shows that some subgroups possess the attibutes with identical property (value) which can be defined as the subspace,

Table 1. Summary of the soybean data.

Name of group	Size of group	Subspace	Distinctive attributes
Diaporthe stem canker	10	$v_{2,17,18,19,20,21}(\equiv 0)$; $v_{4,11,15,16}(\equiv 1)$; $v_3(\equiv 2)$; $v_{13}(\equiv 3)$.	$v_{13}(\equiv 3)$
Charcoal rot	10	$v_{2,3,13,15,16,17,20,21}(\equiv 0)$; $v_{8,11,19}(\equiv 1)$; $v_{18}(\equiv 2)$ $v_{14}(\equiv 3)$.	$v_{3,13}(\equiv 0)$ $v_{18}(\equiv 2)$; $v_{19}(\equiv 1)$ $v_{14}(\equiv 3)$
Rhizoctonia root rot	10	$v_{4,11,15,18,19}(\equiv 0)$; $v_{7,13,14,16}(\equiv 1)$; $v_3(\equiv 2)$;	$v_{20}(\equiv 3)$
Phytophthora rot	17	$v_{15,17,18,19}(\equiv 0)$; $v_{11,21}(\equiv 1)$ $v_{14}(\equiv 2)$	$v_{14}(\equiv 2)$

also each subgroup has some attibutes with the value differently from other subgroup which called as the distinctive attributes, for example, all v_{18} equals to 2 in subgroup 2 whereas different in others subgroups. When this dataset is used to give a clustering partition upon v_{18}, the original subgroup 2 can be separated effectively from the database.

Also from Table 2, we can see that when one partition the data using different attibutes, the subgroups has different number of clusters and whin-cluster mean relative differences(Defined in Section 3), therefore results in different pooled whin-cluster mean relative difference $\left(\bar{W}\right)$. Since $v_{18}\left(v_{19}\right)$ gives the minimum of $\bar{W}$ among other attrbutes, so it can be considered as a distinctive attributes. Therefore in this paper we devote to a simple clustering method to partition the data along such distinctive attributes, that is, the clustering precedure of categorical dataset fully depends on these attributes.

3. Methodology

Suppose that we have the data set $X=\left(X_{ij}\right)$, where the element $X_{ij}\left(i=1,\cdots,n \text{ and } j=1,\cdots,p\right)$ denotes the j-th attribute of the i-th object. Notice that each categorical attributes v_k has a finite number of category levels $N\left(v_k\right)$.

3.1. Useful Measurements

While the Euclidean-based measure could yield satisfactory results for numeric attributes, it is not appropriate for data sets with categorical attributes. Therefore, some alternative measurements must be explored.

Hamming distance, named after Richard Hamming, is widely used to give the difference between two equal-length categorical vectors. The Hamming distance between the object x_i and x_j is defined as:

Table 2. The attribute-based clustering performance on soybean data.

Feature	N_r	W_r	$\bar{W}$	Feature	N_r	W_r	$\bar{W}$
		0.306, 0.262, 0.283		v_{10}	3	0.477, 0.493, 0.480	1.45
v_1	7	0.504, 0.491, 0.410	2.665	v_{11}	2	0.218, 0.509	0.727
		0.410		v_{12}	2	0.481, 0.513	0.994
v_2	2	0.464, 0.341	0.805	v_{13}	4	0.239, 0.334,	1.052
v_3	3	0.239, 0.301, 0.433	0.974			0.241, 0.238	
v_4	3	0.340, 0.468, 0.213	1.02	v_{14}	4	0.215, 0.389,	1.967
v_5	2	0.468, 0.520	0.988			0.285, 0.239	
v_6	4	0.520, 0.478,	1.985	v_{15}	2	0.4940.238	0.732
		0.505, 0.483		v_{16}	2	0.46, 0.434	0.896
v_7	4	0.213, 0.395	1.127	v_{17}	2	0.506, 0.267	0.773
		0.195, 0.324		$v_{18}\left(v_{19}\right)$	2	0.436, 0.239	0.675
v_8	2	0.505, 0.404	0.909	v_{20}	2	0.429, 0.361	0.790
v_9	2	0.490, 0.487	0.977	v_{21}	2	0.507, 0.296	0.803

$$d_{ij} = \sum_{k=1}^{p} 1_{\left(x_{ik} \neq x_{jk}\right)} \tag{3.1}$$

i.e., the hamming distance measures the number of attributes at which the corresponding objects are different.

Our proposed method is based on the pooled within-cluster mean difference of the clusters. Intuitively, when a p-dimension dataset is divided to some subgroups $C_1, C_2, \cdots, C_r$ according to the attribute v_r, this attribute has the same value in some specified subgroup, so it has no information in such subgroups, therefore the dimension d_r of the cluster becomes smaller and smaller. In order to give the dispersion corresponding to this phenomenon, a relative version of dispersion must be adopted.

Provided that we have partitioned the data into $N\left(v_k\right)$ clusters $C_1, C_2, \cdots, C_{N(v_k)}$ upon attribute v_k, denote n_r the number of objects in C_r and d_r the corresponding dimensions (after eliminate the identical attributers). Let

$$\bar{W}_r = \frac{1}{d_r} \frac{1}{n_r\left(n_r - 1\right)} \sum_{x_i, x_j \in C_r} d_{ij} \tag{3.2}$$

be the within-cluster mean relative difference (WCMRD) in cluster C_k, and

$$\bar{W}\left(v_k\right) = \sum_{r=1}^{N(v_k)} \bar{W}_r \tag{3.3}$$

be the pooled within-cluster mean relative difference (PWCMRD).

The idea of our method is to select the distinctive attributes sequentially, which results in the minimum pooled within-cluster mean relative difference comparing with the other attibutes, *i.e.*,

$$v_m = \arg\min_{v_k} \bar{W}\left(v_k\right), \tag{3.4}$$

thereafter, partition the dataset upon the finite characters of the selected attibutes and give the subspace of each subgroup at each iteration.

3.2. Clustering Procedure

Step 1 Initially the data set D is clustered according to the characters of $v_k\left(k = 1, 2, \cdots, p\right)$, *i.e.*, the objects are partitioned to $N\left(v_k\right)$ clusters such that the objects in each cluster have the same character on v_k;

Step 2 Find a distinctive attribute v_g satisfies

$$v_g = \arg\min_{k=1,2,\cdots,p} \bar{W}\left(v_k\right) \tag{3.5}$$

where $\bar{W}\left(v_k\right)$ be the pooled within-cluster mean relative difference of the clusters partitioned upon v_k.

Step 3 Partition the dataset based on v_g, and calculate the corresponding within-cluster mean relative difference $\bar{W}_r$ for each cluster $C_r\left(r = 1, 2, \cdots, N\left(v_g\right)\right)$.

Step 4 While $\bar{W}_r > W_T$ (where W_T is the threshold predefined to stop the procedure),

Update the data set D by C_r,
Repeat Step 1 and Step 2 until all $\bar{W}_r \le W_T$.
End.

3.3. The Stop Threshold W_T

The stop threshold W_T can be chosen arbitrarily. In fact, different W_T results are in different hierarchical clustering. In our paper, the threshold is adopted to be 0.35, means a different of 35% attributes in a cluster is accepted.

4. The Performance of the Proposed Method

4.1. Numerical Experiments

In the section, a simulated sample is deduced as reference [12]. Also the criterion of classification rate (CR) is adopted to give the accuracy of the assignment. The classification rate measures the accuracy of an algorithm to assign data points into correct clusters. With given K clusters,

$$CR(K) = \sum_{i=1}^{K} \frac{n_i}{n}$$

where n_i is the number of data points that have been correctly assigned by an algorithm, n is the total number of the data.

For the simulated sample, [12] obtains a mean CR 94.62%, with standard derivation 3.14%, we obtains a mean CR 96.02%, with standard derivation 2.57%, a litter better than their method.

4.2. Soybeansmall Data

The data set is derived from UCI Machine Learning Repository (archive.ics.uci.edu/ml/), it contains 47 objects, each has 35 categorical attributes. There are some attributes with exactly the same value, so after eliminate the attributes redundant, there only 21 attributes left in data set.

Table 3(a) gives the clustering results of the proposed method. It shows that only one objects are clustered incorrectly. This diagram is different from **Table 1** where we identified v_{11} (with PWCMRD 0.6548) instead of v_{14} (with PWCMRD 0.6584). **Table 3(b)** describes the detail of the proposed method on Soybean data, all except one object is assigned correctly. Also we can see the accuracy of the proposed method with $CR = 0.98$.

Figure 1 gives the details of partition for Soybean data. In each step, a distinct attribute is identified, and the data is separated along this attribute, and also the subgroup with largest W is chosen to be the target one to be separated in next step.

4.3. Zoo Data

The Zoo data set is available from UCI Machine Learning Repository (archive.ics.uci.edu/ml/), it contains 101 objects, each has 16 categorical attributes. There are some objects who posses exactly same value on all attributes, so

Table 3. Cluster result of soybean data by the proposed method.

(a)

True	C_1	C_2	C_3	C_4
diap.	10	0	0	0
char.	0	10	0	0
rhiz.	0	0	9	1
phyt.	0	0	0	17

(b)

Iterate	Distin. Var.	N	$\bar{W}$	Subspace
1	$v_{18}(=2)$	10	0.2511	$v_{2,3,13,15,16,17,20,21}(=0)$ $v_{8,11,19}(=1)$ $v_{18}(=2), v_{14}(=3)$
2	$v_{15}(=1)$	10	0.2632	$v_{2,17,18,19,20,21}(=0)$ $v_{4,11,15,16}(=1)$ $v_{3}(=2), v_{13}(=3)$
3	$v_{11}(=1)$	9	0.2546	$v_{4,11,15,18,19}(=0)$ $v_{7,13,14,16}(=1)$ $v_{3}(=2), v_{20}(=3)$
4	$v_{11}(=0)$	18	0.3638	$v_{15,18,19}(=0)$ $v_{11,21}(=1)$ $v_{20}(=3)$

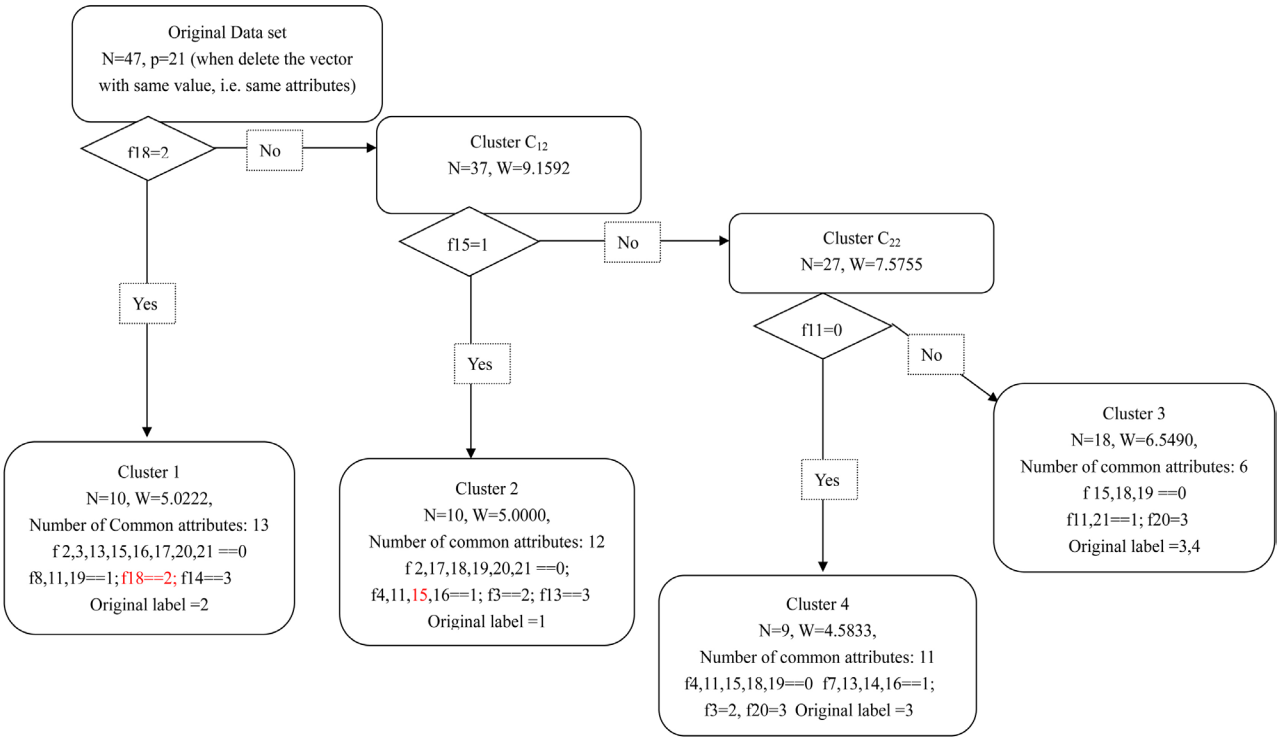

Figure 1. Performance on soybeansmall data.

it can be considered as the same ones, after eliminate the redundant objects, there only 59 objects left in data set.

Table 4 gives the clustering results of the proposed method, where "group" means the true category of the objects. It shows that the performance is poor on group 5, 6, 7, with 1 object in group 5 is clustered incorrectly into group 3, and the group 6 and 7 each has one object are considered as member of the new cluster. Since the objects in group 5 (frog, frog, newt, toad) and group 3 (pitviper, seasnake, slowworm, tortoise, tuatara) have more similarity than others (there are 11 attributes are the same among 16 attributes), so the two group can roughly be considered as one group. After combining the two subgroups, our proposed methods has a precision clustering with only one incorrect.

Table 5 describes the detail of the proposed method on zoo data.

5. Comparison with HD Vector Method

Zhang *et al.* ([12]) indicates that their method can archive a good results in both the zoo data and Soybean disease data, the comparison between HD vector method and *K*-modes as well as Autocluss algorithm shows the superiority of their method, the drawback of their method is the comparison of possible data with the number equals to the total number of possible positions in a categorical sample space, in our proposed this possible positions are not needed, therefore the algorithm can be faster than their method.

6. Conclusions

Categorical variables are widely explored in different fields to give a native

Table 4. Cluster result of Zoo data by the proposed method.

(a)

Clusters found	n_m	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8
Group 1	19	19	0	0	0	0	0	0	0
Group 2	12	0	12	0	0	0	0	0	0
Group 3	5	0	0	5	0	0	0	0	0
Group 4	5	0	0	0	5	0	0	0	0
Group 5	4	0	0	1	0	3	0	0	0
Group 6	6	0	0	0	0	0	5	0	1
Group 7	8	0	0	0	0	0	0	7	1

(b)

Clusters found	n_m	C_1	C_2	C_3	C_4	C_5	C_6
Group 1	19	19	0	0	0	0	0
Group 2	12	0	12	0	0	0	0
Group 3 + 5	9	0	0	9	0	0	0
Group 4	5	0	0	0	5	0	0
Group 6	6	0	0	0	0	5	1
Group 7	8	0	0	0	0	0	8

Table 5. Cluster result of Zoo data by the proposed method.

Iterate	Distin. Var.	N	$\bar{W}$	Subspace
1	$v_2 = 1$	12	0.1485	$v_{1,4,8,11,12}(=0)$ $v_{2,3,9,10,14}(=1)$ $v_{13}(=2)$
2	$v_4 = 1$	19	0.2498	$v_{2,11}(=0)$ $v_{4,9,10}(=1)$
3	$v_5 = 1$	5	0.1538	$v_{2,4,6,8,9,12,14,16}(=0)$ $v_{3,5,10}(=1)$
4	$v_{12} = 1$	5	0.1667	$v_{1,2,4,5,10,13}(=0)$ $v_{3,6,8,9,12,14}(=1)$
5	$v_7 = 1$	7	0.3	$v_{1,2,4,5,8,9,12,15}(=0)$ $v_7(=1)$
6	$v_7 = 0$	2	0.1	$v_{1,2,4,5,6,7,8,9,11,12,13,14,15,16}(=0)$ $v_{3,10}(=1)$
7	$v_{14} = 0$	3	0.133	$v_{1,2,4,5,10,12,13,14,15,16}(=0)$ $v_{3,6,8,9}(=1)$
8	$v_{14} = 1$	6	0.333	$v_{1,2,4,5,10,12,13}(=0)$ $v_{3,6,8,9,14}(=1)$

clustering algorithm to deal with such type data; a pooled-within-cluster-mean-different based method is proposed to select some distinctive attributes, and then the data are clustered upon such distinctive attributes; the subspaces are also investigated.

The applications on zoo data and soybean data (from UC Irvine Machine Learning Repository) illustrate the performance of the proposed method. The results show a high accuracy and simplicity in practical applications.

References

[1] He, Z., Xu, X. and Deng, S. (2008) k-ANMI: A Mutual Information Based Clustering Algorithm for Categorical Data. *Information Fusion*, **9**, 223-233.

[2] Andritsos, P. and Tsaparas, P. (2010) Categorical Data Clustering. In Sammut, C. and Webb, G., Eds., *Encyclopedia of Machine Learning*, Springer, Boston, 154-159.

[3] Bontemps, D. and Toussile, W. (2013) Clustering and Variable Selection for Categorical Multivariate Data. *Electronic Journal of Statistics*, **7**, 2344-2371. https://doi.org/10.1214/13-EJS844

[4] Anderlucci, L. and Hennig, C. (2014) The Clustering of Categorical Data: A Comparison of a Model-Based and a Distance-Based Approach. *Communication in Statistics—Theory and Methods*, **43**, 704-721. https://doi.org/10.1080/03610926.2013.806665

[5] Bouguessa, M. (2015) Clustering Categorical Data in Projected Spaces. *Data Mining and Knowledge Discovery*, **29**, 3-38. https://doi.org/10.1007/s10618-013-0336-8

[6] Silvestre Cardoso, C.M. and Figueiredo, M. (2015) Feature Selection for Clustering Categorical Data with an Embedded Modelling Approach. *Expert Systems*, **32**, 444-453. https://doi.org/10.1111/exsy.12082

[7] dos Santos, T.R.L. and Zarate, L.E. (2015) Categorical Data Clustering: What Similarity Measure to Recommend? *Expert Systems with Applications*, **42**, 1247-1260.

[8] Clarke, B.S., Amiri, S. and Clarke, J.L. (2016) EnsCat: Clustering of Categorical Data via Ensembling. *BMC Bioinformatics*, **17**, 380. https://doi.org/10.1186/s12859-016-1245-9

[9] Ganti, V., Gehrke, J. and Ramakrishnan, R. (1999) CACTUS—Clustering Categorical Data Using Summaries. *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM Press, San Diego, 73-83.

[10] Guha, S., Rastogi, R. and Shim, K. (2000) Rock: A Robust Clustering Algorithm for Categorical Attributes. *Information Systems*, **25**, 345-366.

[11] Arias-Castro, E. and Xiao, P. (2017) A Simple Approach to Sparse Clustering. *Computational Statistics & Data Analysis*, **105**, 217-228.

[12] Zhang, P., Wang, X. and Song, P.X.K. (2006) Clustering Categorical Data Based on Distance Vectors. *Journal of the American Statistical Association*, **101**, 355-367. https://doi.org/10.1198/016214505000000312