CATCLUS – A Proposed Algorithm for Clustering Categorical Data

Srikanta Kolay
Electrical & Automation
SMS India Pvt. Ltd.
RDB Boulevard, 5th Floor, Unit-D, Plot No.-K1, Block-EP&GP, Sector-V, Salt Lake, Kolkata-700091, India

Kumar S. Ray
Electronics and Communication Science Unit
Indian Statistical Institute
203, B.T Road, Kolkata-700108, India

ABSTRACT
Classification of categorical data always involves more complexities compared to the numerical data. Because, a firm outline cannot be drawn in case of categorical data. Different types of assumptions are followed by various researchers to treat such kind of data. Again, dissimilarity measures applied in case of numerical data cannot be applied directly in this case. In this paper, a new clustering algorithm for categorical data is proposed. The algorithm is using a newly devised dissimilarity measure. This paper only includes the theoretical description of the proposed algorithm with appropriate example.

Keywords
Categorical Data, Clustering, Dissimilarity Measure, Algorithm.

1. INTRODUCTION

1.1 Clustering
Clustering is a popular approach to implementing the partitioning operation. Clustering methods partition a set of objects into clusters such that objects in the same cluster are more similar to each other than objects in different clusters according to some defined criteria. There are various categories of clustering algorithms such as partition based, hierarchical, density based, grid based etc. These algorithms are mainly based on the approach followed by them. Again, some algorithms are developed targeting a particular type of data such as categorical, numerical, Boolean, exponential etc. For example the k-means [1] algorithm, which is well known for its efficiency in clustering large data sets. However, working only on numeric values prohibits it from being used to cluster real world data containing categorical values. However an extension of the said algorithm is found in [2], which is capable of work with categorical attributes. Few example of categorical data clustering algorithms are ROCK [3], CACTUS [4], Squeezer [5]. A fuzzy set based approached can be found in [6].

1.2 Categorical Data
Data that consist of only small number of values, each corresponding to a specific category value or label. A categorical variable is a generalization of a binary variable in that it can take on more than two states.

There are a variety of categorical variables:

- Whether the mother was employed (yes, no);
- The mother's marital status (single, married, divorced, widowed).
- Map colour (red, yellow, green, pink, blue)

Some classes of attributes may be:

- An attribute is continuous if, between any two values of the attribute, there exists an infinite number of values. E.g. Temperature, colour, or sound intensity.
- An attribute is discrete if the elements of its domain can be put into a one-one correspondence with a finite subset of positive integers. E.g. number of children in a family or the serial numbers of books.
- The class of binary attributes consists of attributes with domains of exactly two discrete values. Eg. Yes/No responses to a poll or the Male/Female gender entries of a database of employees.

1.3 Dissimilarity between Categorical Attributes
Dissimilarity can be measured in many ways and one of them is distance, used for numeric attributes. For categorical attributes, however, distances between two objects is a poor estimate of the similarity. The dissimilarity between two objects i and j can be computed based on the ratio of mismatches:

$$d(i, j)=(p-m)/m,$$

where m is the number of matches (i. e. the number of variables for which i and j are in same state), and p is the total number of variables.

Suppose that let us have the sample data of the table given below, where Grade is categorical.

Table 1

Roll No	Grade
1	A
2	B
3	C
4	D

Since let us have only one categorical variable, Grade, let us set p=1 so that d(i, j) evaluates to 0 if objects i and j match and 1 if the objects differ.

0
d(2,1) 0
d(3,1) d(3,2) 0
d(4,1) d(4,2) d(4,3) 0

Thus let us get,

0			
1	0		
1	1	0	
0	1	1	0

2. CATCLUS: PROPOSED METHODOLOGY

In the new algorithm, each cluster will be represented by cluster representative based on the notion of means in numerical setting.

- Let us assume a cluster C={ X_1, X_2,……., X_n} of n categorical objects, each object having m categorical attributes X_i=($x_{i,1}$, $x_{i,2}$,……., $x_{i,m}$).
- Each X_i can take on values from dataset D_i, called the domain of X_i. Thus D=(D_1, D_2,……, D_n).

  Then the representative of cluster C may be defined as R=(r_1, r_2,……, r_m), where r_i={(c_i, rf_{ci})| c_i€D_i}

  with f_{ci} being the relative frequency of c_i within C , i.e. , rf_{ci} =n_{ci} /n, where n_{ci} is the number of objects in C having category c_i in attribute A_i.

Since each attribute of the cluster representative also contain the relative frequency of that attribute along with the categorical value, so the dissimilarity measure will also be dependent on the relative frequencies of the categorical attributes within the cluster along with the (dis)similarity measure between the categorical attributes.

2.1 Dissimilarity Measure

In the proposed algorithm, the dissimilarity measure is defined as

$$d(X,R) = \sum^{m}_{j=1} (1-rf_{xj}) \cdot \delta(x_j, r_j)$$
$$x_j \neq r_j$$

where,

$\delta(x_j,r_j)=0$, if $x_j=r_j$

=1, if $x_j \neq r_j$ (in case of binary attributes)

or in between 0 and 1, if $x_j \neq r_j$. (in case of ordinal attributes)

Here, X may or may not belong to the cluster of which R is the representative.

2.2 Cluster Representative Updation Process

To update the cluster representatives following steps to be carried out:

1.For every attribute of the objects in the cluster,

a. For each category of an attribute

i) count the relative frequency of each category

b.take the category with highest relative frequency

2. Choose the object with highest frequent category values in their attributes as new cluster representative .

3. Reassign the objects.

2.3 Algorithm

The steps in the proposed algorithm can be stated as follows:

1) Initially assume all elements in the dataset in one cluster.
2) Select k representatives R={ R_1, R_2, …., R_k }.
3) Then, k clusters are formed as per the k-modes algorithm.
4) After this, the new proposed dissimilarity measure is used to update representatives of each cluster.
5) The process is repeated until representative objects do not change.
6) After the formation of k clusters, next step will be validation phase. In this step, it is examined whether the user defined k number of clusters is appropriate or not. This step may detect outliers. Min, Max and Average intra cluster dissimilarity for each of the k-clusters is computed.
7) The average intra cluster dissimilarity is expected to be almost similar and preferably small for each of the k clusters.
8) If for any cluster, the average value is more, its *Max* and *Min* values are checked. If the Max value is high, then outlier object is detected, as it has the maximum distance from its cluster representative.
9) Now, taking this outlier as another new representative, the algorithm is repeated to assign objects to k+1 cluster.
10) The algorithm is repeated until no more new representatives are formed and existing representatives do not change.

3. EXPERIMENT (WITH SYNTHETIC EXAMPLE)

The proposed algorithm is explained below with the help of two synthetic datasets.

Example 1: Assume a new dataset as follows:

Record	Height	Sex	Use of specs	Record	Height	Sex	Use of specs
1	T	F	Y	11	M	M	N
2	T	M	Y	12	M	M	N
3	S	F	N	13	T	F	N
4	M	M	N	14	M	F	Y
5	T	F	N	15	T	M	Y
6	S	M	Y	16	M	F	N
7	S	F	N	17	S	F	N
8	S	M	N	18	S	M	N
9	M	F	Y	19	T	F	Y
10	S	M	Y	20	M	M	N

Fig 1: Sample Dataset 1

Where:

T=tall, M=Medium, S=Short

M=Male, F=Female

Y=Yes, N=No.

Let k=3.

Let us compute relative frequencies of all categories of all the attributes in the dataset.

Let us arrange the categories in descending order:

M>=S>T

M>=F

N>Y

Now, k representatives are selected so that most frequent categories are distributed in every representative.

Let us select, 5(T,F,N), 12(M,M,N) and 17(S,F,N).

Formation of initial k clusters as per the k-modes algorithm

3.1 5 (T,F,N)	3.2 12 (M,M,N)	3.3 17 (S,F,N)
3.4 13, 19, 16, 1, 2, 15	3.5 4, 11, 20, 16, 8, 18, 9, 14	3.6 3, 7, 16, 8, 18, 6, 10

Fig 2: Results on Sample Dataset 1

Red colored objects are objects common to all three clusters.

Using new dissimilarity measure, object number 16 is placed in cluster 1; 8 is placed in cluster 2 and 18 is placed also in cluster 2.

Now, representatives are updated. Only representative of cluster 1 changed; other two cluster representatives remain the same.

Now, it is again checked whether representatives change or not. In the 3rd step, representatives do not change and let us get the user defined k number of clusters.

Now let us compute min(k), max(k) and avg(k) for all k.

$$\boldsymbol{Min(1) = 0, Max(1) = 1, Avg(1) = 0.57}$$

$$\boldsymbol{Min(2) = 0, Max(2) = 1, Avg(2) = 0.33}$$

$$\boldsymbol{Min(3) = 0, Max(3) = 1, Avg(3) = 0.5}$$

Almost in the same range.

Low Average value indicates high intra cluster similarity and vice versa. In this dataset, no outlier object is found and let us get the desired number of clusters. Object overlapping is removed (an advantage over k-modes algorithm).

Record	Height	Sex	Use of specs	education	Financial condition	habitat	Record	Height	Sex	Use of specs	Education	Financial condition	habitat
1	T	F	Y	G	P	V	11	M	M	M	G	P	V
2	T	M	Y	G	LM	V	12	M	M	N	G	P	C
3	S	F	N	M	LM	T	13	T	F	N	M	LM	T
4	M	M	N	M	UM	C	14	M	F	Y	G	P	V
5	T	F	N	PG	UM	M	15	T	M	Y	G	P	V
6	S	M	Y	M	LM	V	16	M	F	N	M	P	V
7	S	F	N	M	P	V	17	S	F	N	M	P	T
8	S	M	N	G	LM	T	18	S	M	N	M	P	V
9	M	F	Y	G	P	V	19	T	F	Y	PG	UM	M
10	S	M	Y	G	P	T	20	M	M	N	G	LM	C

Fig 3: Sample Dataset 2

Education: M=Matric, G=Graduate, PG=Post Graduate

Financial Condition: P=Poor, LM=Lower Middle, UM=Upper Middle, R=Rich

Habitat: V=Village, T=Town, C=City, M=Metropolitan

Using this dataset, taking k=3, if let us proceed let us get the final clusters as shown below:

1 (T,F,Y,G,P,V)	12(M,M,N,G,P,C)	17 (S,F,N,M,P,T)
2, 5, 9, 13, 14, 15, 19	4, 8, 11, 16, 18, 20	3, 6, 7, 10

Fig 4: Result on Sample Dataset 2

Now, let us compute Min(k), Max(k) and Avg(k) for all k.

Min(1)=0.5, Max(1)=3.16, Avg(1)=1.54

Min(2)=0.33, Max(2)=2.16, Avg(2)=1.19

Min(3)=0.33, Max(3)=2.66, Avg(3)=1.29

It is seen that *Avg* value for cluster 1 is high and its Max intra cluster dissimilarity is 3.16 (=d(1,5)). Next higher values for Max are 2.16(=d(1,13)) and 2.16(=d(1,19)).

Let us compute

.

d(5,13)=2

d(5,19)=1

d(13,19)=3

So now taking object 5 as $(k+1)^{th}$ cluster representative, 13 and 19 are assigned to the $(k+1)^{th}$ cluster.

When let us run the algorithm again, cluster representatives do not change and let us get the final (k+1) clusters as shown below

1 (T,F,Y,G,P,V)	5 (T,F,N,PG,UM,M)	12(M,M,N,G,P,C)	17 (S,F,N,M,P,T)
2, 9, 14, 15	13,19	4, 8, 11, 16, 18, 20	3, 6, 7, 10

Fig 5: Result on Sample Dataset 2

4. CONCLUSION

In this paper a new clustering algorithm is introduced. The algorithm is developed primarily for categorical data or attributes. Based on the requirement a new distance or dissimilarity measure is also formulated. The algorithm is well experimented with appropriate synthetic data and results are found as expected.

5. REFERENCES

[1] MCQUEEN, J. 1967. Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 281-297.

[2] Z. Huang Extensions to the k-means algorithm for clustering large data sets with categorical values Data Mining and Knowledge Discovery, 2 (3) (1998), pp. 283–304

[3] S. Guha, R. Rastogi, and K. Shim," ROCK: A Robust Clustering Algorithm for Categorical Attributes", 15th International Conference on Data Engineering, pp. 512-521, 2000.

[4] V., Ganti, J. Gehrke, R. Ramakrishnan, CACTUS – clustering categorical data using summaries, in: Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1999, pp. 73–83.

[5] Z. He, X. Xu, S. Deng, Squeezer: an efficient algorithm for clustering categorical data Journal of Computer Science & Technology, 17 (5) (2002), pp. 611–624

[6] D. Kim, K. Lee, D. Lee Fuzzy clustering of categorical data using fuzzy centroids Pattern Recognition Letters, 25 (11) (2004), pp. 1263–1271