Clustering of Categorical Data by Assigning Rank through Statistical Approach

Sovan Kumar Patnaik
Directorate of Horticulture, Odisha,Bhubaneswar

Soumya Sahoo
C.V Raman College of Engineering,Bhubaneswar

Dillip Kumar Swain
Indira Gandhi Institute of Technology,Saranga

ABSTRACT

Most of the earlier work on clustering has mainly been focused on numerical data whose inherent geometric properties can be exploited to naturally define distance functions between data points. Working only on numeric values prohibits it from being used to cluster real world data containing categorical values. Recently, the problem of clustering categorical data has started drawing interest. The k-means algorithm is well known for its efficiency in this respect. It is also well known for its efficiency in clustering large data sets. However, in this paper we use the k-means algorithm to categorical domains by assigning rank value to the attributes

Keywords

catagorical data; k-mean;Rank value

1. INTRODUCTION

Clustering is one of the primitive types of technique frequently used in the field of data mining[1]. Various Clustering techniques are used from Engineering field[2], web mining[3], medical science[2] to retail data mining. During the last decade, data-mining has emerged as a rapidly growing inter disciplinary field which merges together databases, statistics, machine learning and related areas in order to extract useful knowledge from data. Clustering can be defined as the process of organizing objects in a database into clusters/groups[5]. Clustering determine the objects within the same cluster have a high degree of similarity, while objects belonging to different clusters have a high degree of dissimilarity[6],

Traditionally, numerical clustering methods[7] have been viewed. Numerical techniques emphasize the determination of homogeneous clusters according to some similarity measures. Further it is easy to understand about the similarity between data through mathematical approach. Mathematical approach can be done by numerical operation. Most of the earlier work on clustering has been focused on numerical data whose inherent geometric properties can be exploited to naturally define distance functions between data points. But now a days it is seen that most of the data mining applications frequently involve many datasets that also consist of categorical attributes. So it is difficult for mathematical operation to determine cluster group. The distance functions cannot be naturally defined. Recently, clustering data with categorical attributes have drawn some attention.

2. CATAGORICAL DATA

In statistics, categorical data[8,9] is that part of an observed dataset that consists of categorical variables, or for data that has been converted into that form. Categorical variables represent types of data which may be divided into groups. More specifically, categorical data may derive from either or both of observations made of qualitative data, where the observations are summarised as counts or cross tabulations, or of quantitative data, where observations might be directly observed counts of events happening or they might counts of values that occur within given intervals.

A categorical variable is defined as one that can assume only a limited number of discrete values. The measurement scale for such a variable is unrestricted. Three type of categorical data we can observe .It can be nominal [10], which means that the observed levels are not ordered. It can be ordinal [10], which means that the observed levels are ordered in some way. Or it can be interval [10], which means that the observed levels are ordered and numeric and that any interval of one unit on the scale of measurement represents the same amount, regardless of its location on the scale.

2.1 Categorical Domains and Attributes

Let X be the data set of n instances having m attribute. Each attribute may have (k) number of values. ||Ω(X)||=n .Let A1, A2, …, Am be m attributes describing a space Ω and DOM(A1), DOM(A2), …, DOM(Am) the domains of the attributes. ||(Aj)||= m, DOM(Aj)|= a,b,c etc.A domain DOM(Aj) is defined as categorical if it is finite and unordered, e.g., for any a, b DOM(Aj), either a= b or a ≠ b. Aj is called a categorical attribute. Ω is a categorical space if all A1, A2, …, Am are Categorical. A special value, denoted by ϵ, is defined on all categorical domains and used to represent missing values.

2.2 Process of assigning rank value to categorical variables

A categorical variable is generalization of binary variables in that it can take on more than two states. For example map_colour is categorical variables that may have say, 5 state: red,yellow,green,pink,and blue. Let the number of state of categorical variables be M. The state can be representing by set of letter, symbol, or integers. It is better to assigning integer value to handle the data. In this paper we represent the categorical variable by assigning rank value. we adopt statistical method to assigning rank value. The process of assigning rank value is given in details in below.

Suppose A is an attribute having k categorical value. i.e.

DOM(A_k)= v1,v2,….vk

1. Calculate the number of occurrence of each domain value from the total sample
2. Determine the probability value of each domain.

3. P(A_{ij})=$_n$(A_{ij})/□ Ω n: number of A_{ij}
4. Assigning the rank value to the domain by adding the probability value of previous domain.
5. R(A_i)=P(A_{ij})+P(A_{ij-1}) R: Rank value of each attribute

Where P stands for Probability, Aij stands for no.of instance of each Attribute.

3. THE K-MEANS ALGORITHM

K-means is one of the oldest and widely used algorithm for clustering categorical data. K-means is a popular nonhierarchical clustering technique. In this case the k representative objects are called centroids. K-means is one of the simplest unsupervised learning algorithms that solve clustering problem. K-means algorithm [7] groups data vectors into a predefined number of clusters on the basis of the Euclidean distance. Euclidean distances among data are small for data within a cluster. Initially this algorithm was adopted by conjunction[14] of attribute method in categorical data types to determined the similarity between data attributes. But K-mean is effective in case of numeric data attribute.I n this paper we adopt k-mean numeric operation in data attributes having categorical values . Assume the set of data objects that is to be clustered is defined by a set of attributes A_1, A_2 , . . ., A_d . Each attribute A_j (1 ≤□j ≤□d) describes a domain of n_j categorical values denoted by *DOM(Aj*) =□ {a_{j1}, a_{j2} , . . ., $a_j n_j$ } . Assume that the data set is stored in a ***n*p*** matrix,
where ***n*** is the number of observations and ***p*** the number of categorical variables. An iterative method can be implemented in different ways. The K-mean algorithm[5] as follows:

1. Randomly initialize the k cluster centroid vectors

2. Repeat

(a) For each data vector, assign the vector to the cluster with the closest centroid vector, where the distance to the centroid is determined using

$$D(X_i, C_k) = \sum_{j=1}^{n} \left(\sqrt{x_{ij} - C_{kj}} \right)^2 \qquad (1)$$

where xi denotes the i-th data vector, j denotes the centroid C vector of cluster j, and d subscripts the number of features of each centroid vector. Where i=1 to n, j=1, 2, …, k.

(b) Recalculate the cluster centroid vectors using

$$C_k = \frac{1}{n_j} \sum_{x_i \in C_k} X_i \qquad (2)$$

where n_j is the number of data vector belongs to C_j.

3 Repeat until stopping criterion is satisfied.

4. Calculate the intra-cluster compactness of each k clusters using equn(4)

5. Calculate the inter class Separability using equn(5)

The K-means clustering process terminates when any one of the following criteria is satisfied: when the maximum number of iterations has been exceeded or when there will be little change in the centroid vectors over a number of iterations, or when the cluster membership doesn't change. For the purpose of this research the algorithm terminates when specified number of iterations has been exceeded.

Based on the new optimized cluster center the intra class compactness and interclass separability, have been calculated

For compactness [10] [11] Kim proposed:

$$v_u = \frac{1}{K} \sum_{K=1}^{K} \left(\frac{1}{|C_k|} \sum_{x_j \in C_k} \left\| x_j - v_k \right\|^2 \right) \text{……..(4)}$$

$$v_o = \frac{K}{\min_{j \neq i} \left\| v_i - v_j \right\|^2}$$
……….(5)

4. CLUSTER VALIDITY

Cluster validity measures the goodness of a cluster relative to other cluster. Cluster validation is important issue in cluster analysis because the result of clustering needs to be validated in most applications. In most clustering algorithms, the number of clusters is set as user parameter. There are a lot of approaches to find the best number of clusters. Different cluster validation methods include Silhouette method, Dunn's based index, Davies-Bouldin index and the C-index method [13]. We have used Dunn's based index method for the purpose of cluster validation. This technique [13] is based on the idea of identifying the cluster sets that are compact and well separated. For any partition of clusters, where ci represent the ith cluster of such partition, the Dunn's validation index, D, could be calculated with the following formula:

$$D = \min_{1 \le i \le n} \left\{ \min_{1 \le j \le n} \left\{ d\left(\min_{1 \le j \le n} (c_i, c_j) \right) \Big/ \max_{1 \le k \le n} \left\{ d'(c_k) \right\} \right\} \right\} \qquad (6)$$

where d(ci, cj) is the distance between clusters c_i, and cj (i.e. inter cluster distance), {d'(ck)} is the intra cluster distance of cluster ck, n is the number of clusters. The minimum is calculating for number of clusters defined by the mentioned partition. The main goal of this measure is to maximize the inter cluster distances and minimize the intra cluster distances.

5. EXPERIMENTAL RESULT

The experiment have been carried out in categorical data. These data are collected from UCI repository. The Details about the data is shown in table-1.

Table 1: Information about datasets

Database	No of Samples	No of Attributes	Classes
Car Evaluation	1728	6	4
Balance Scale	625	4	3
Hayes-Roth	160	5	3

The K-means clustering algorithm has been described in Section 3 and the Ecludian distance have been calculated as per the equation (1).

Table 2 .Cluster Compactness

Data set	Compactness	
	Criterion	K-mean
Car Evaluation	Cluster1	17.0396
	Cluster2	18.5895
	Cluster3	24.8017
	Cluster4	23.953
Balance Scale	Cluster1	47.9155
	Cluster2	39.8128
	Cluster3	37.6207
Hayes-Roth	Cluster1	305.7694
	Cluster2	369.1610
	Cluster3	294.5871

Basing on this Distance the Cluster Compactness and Separability have been determined as per the equation(2)&(3).

Table 3. Cluster Separability

Data Set	K-mean
Car Evaluation	0.4821
Balance Scale	0.3129
Hayes-Roth	0.1936

Table-2&3 shows the cluster compactness and separability. Table-4 indicates the cluster validity

Table 4.Cluster Validation by Dunn's Index

Data Set	Index value
Car Evaluation	53.791
Balance Scale	50.152
Hayes-Roth	5.207

6. CONCLUSION

The biggest advantage of the k-means algorithm in data mining applications is its efficiency in clustering large data sets. However, its use is limited to numeric values. Due to filtering capacity of K-mean, this algorithm is only used in case of numeric data sets. The Agglomerative and Divisive Hierarchical Clustering algorithm was adopted the dataset of categorical nature initially. Due to complexity in both of the above algorithm, this paper has presented a new approach to assign rank value to each catagoric attribute for K-mean Clustering. The categorical data have been converted into numeric by assigning rank value. It is a that a categorical dataset can be made clustering as numeric datasets.. It is observed that implementation of this logic, k- mean yield same performance as used in numeric datasets. It is a initial experiment. In future more experiment in this effect will be taken to strengthening the algorithm.

7. REFERENCES

[1] Han, J., Kamber, M. 2010. Data Mining:Concept and Technique, 2nd ed, 383–464.

[2] Ng.H.P,Ong.S.H,Foong K.W.C.2006. Medical Image Segmentation Using K-Means Clustering And Improved Watershed Algorithm, 4244-0069- IEEE

[3] Hammouda.K.M 2001 "*Web Mining*: *Clustering* Web Documents. A Preliminary Review.

[4] Anderberg,M.R.1973.ClusterAnalysisforApplicationsNew wwork:Academic press.

[5] Agresti, A. 1984. Analysis of Ordinal Categorical Data, New York: John Wiley &Sons.

[6] Kaufman, L. and Rousseeuw, P. J. 1990 Finding groups in data: an introduction to cluster analysis, New York: John Wiley & Sons.

[7] Mishra, B.B., Sahoo, S., and Patnaik., S.K. 2011. KPSO:AnEvolutionary Approach for Data Clustering. In Proceeding of National conference on Future Trends in Information & Communication Technology , 137–142

[8] Andersen, Erling B. 1980. Discrete Statistical Models with Social Science Applications. North Holland,

[9] Mar San. O Huynh. V, Nakamori.Y An Alternative Extension Of The *K*-Means Algorithm For Clustering Categorical Data Int. J. Appl. Math. Comput. Sci., 2004, Vol. 14, No. 2, 241–247

[10] Alan Agresti 2nd edition 2007" An Introduction to Categorical Data Analysis" JohnWiley & Sons, Inc.,

[11] Kim, D.J., Park, Y.W., and Park, D.J. 2001.A novel validity index for determination of the optimal number of clusters, IEICE Trans. Inf. Syst., vol. E84D, 281–285.

[12] Mishra, B.B.,Patnaik, S.K., and Mohanty P. 2011 Simultaneous Learning for both Clustering and Classification in Multi-objective Framework.In Proceeding of National conference on Future Trends in Information & Communication Technology 137–142.

[13] Dunn, J.C. 1974.Well separated clusters and optimal fuzzy partitions, J. Cybern., vol. 4, 95-104.

[14] Huang.Z,1998 "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values" Data Mining and Knowledge Discovery 2, 283–304 (1998) °c 1998 Kluwer Academic Publishers. Manufactured in The Netherlands