Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 07, July-2016

Dissimilarity Calculation Approach for Categorical Data Set

Mr. Yuvraj S. Sase [1]
ME Second Year (Computer Engineering)
Vishwbharati Academy College of Engineering,
Ahmednagar, Pune University

Mr. M. C. Kshirsagar [2]
Assistant Professor,
Vishwbharati Academy College of Engineering,
Ahmednagar, Pune University

***Abstract*— K-Means algorithm, which is probably most popular technique that solve the well-known clustering problem. The procedure follows the process of partitioning a group of data points into small number of clusters. The main limitation of algorithm is input required as number of clusters. It is very hard for user who is unknown to data set to calculate K value. There is no thumb rule for calculating K value, its complete trial and error method and very inefficient for any kind of user. An Improved K-Means algorithm [1] removes this problem of K-Means algorithm as it does not take any kind of input from user and decides K value in process automatically. But still an improved K-Means algorithm has problem with categorical data set. An improved K-Means algorithm fails with categorical data set. A proposed algorithm is addition to improved K-Means algorithm to solve problem of An improved K-Means algorithm. An proposed algorithm use dependant attributes to calculate dissimilarity between categorical attribute which further used with numerical data set final clustering result.**

***Keywords*— Clustering; Objects; Cluster Mean; Outliers; Dependent Attributes**

I. INTRODUCTION

Today every sector is moving to digital world. Thus they are generating huge amount of data. Knowledge discovery from this data is necessary for improvement in respective sector. Clustering is one of method to gain useful patterns from data set. Clustering is a data mining technique which helps in grouping or making clusters of data by considering their similarity and dissimilarity. There are many clustering algorithms proposed yet. The most popular one is K-Means algorithm, because of its simplicity and effectiveness against large data sets. K-Means algorithm uses Euclidean distance formula to decide dissimilarity between data objects. But K-Means algorithm has many problems like difficulty in finding K value, initial cluster selection etc.

An improved K-Means [1] algorithm is solution for all problems arises in K-Means algorithm. An improved K-Means algorithm removes the most basic problem of K-Means algorithm i.e. dependency on K value. This algorithm does not take K as input from user. K value is mainly calculated dynamically. This algorithm use outliers to find K value. Outliers are mainly decided on objective function. If any object satisfies conditions of objective function then that object is not outlier. Otherwise it is outlier for that cluster. This algorithm finds all outlier and then assigns new cluster to them. Next, all points are adjusted in the new formed clusters according to minimum distance and all process repeated till change happens in clusters. An improved K-Means algorithm solve dependency problem of K-Means algorithm. But improved K-Means algorithm does not works on categorical data set. The data type which can be divided into categories is called as categorical data. Such kind of categorical data has no numerical values instead they have groups. There is no way of calculating dissimilarity between these groups which leads to infertile environment for clustering.

In this paper, the proposed algorithm can find dissimilarity between categorical attributes. This algorithm uses distance equations to find out category attribute value. The proposed algorithm has two kinds input data sets, first is data set including categorical attributes which needed to be grouped in clusters and second one is data set containing dependant attributes on categorical attribute.

II. LITARATURE SURVEY AND RELATED WORK

A. Categorical Data

The data type which can be divided into categories is called as categorical data. For example,

- Age Group
- Colours
- Gender

Categorical data attributes are really very common in real life situations. Every sector is generating huge amount of categorical data. Categorical data is also data that is collected in an either/or, yes/no fashion. Categorical data is also divided into three types,

- Dichotomous Data: Categorical attribute which has only two values are called as Dichotomous data. Dichotomous variables are designed to give you an either/or response.
- Ordinal Data: Categorical attribute which has two or more categories in order is called as Ordinal data.
- Nominal Data: Categorical attribute which has two or more categories but not in order is called as Nominal Data.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 07, July-2016

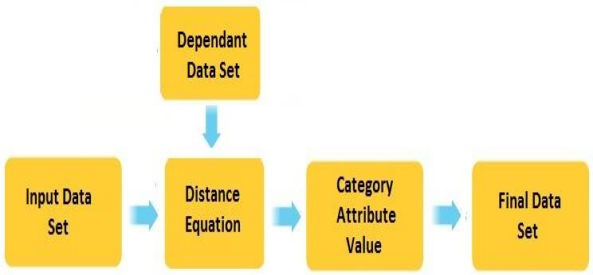


Fig. 1. Skeleton Diagram of Proposed System

B. Dependant and Independant Attribute

The dependent variable is simply that; a variable that is dependent on an independent variable(s) for example, student performance. Student performance depends on study time of student and natural intelligence of student. That means if change in these two variables will change student performance. In proposed algorithm dependant attribute has very crucial part in calculating dissimilarity of category attribute.

III. PROPOSED ALGORITHM

The proposed algorithm can find dissimilarity between categorical attributes. This algorithm uses distance equations to find out category attribute dissimilarity. The proposed algorithm has two kinds input data sets, first is data set including categorical attributes which needed to be grouped in clusters and second one is data set containing dependant attributes on categorical attribute. The proposed system is built to calculate values for each and every category tuple in each categorical attribute. This algorithm uses distance equations to find out category attribute value. Figure shows skeleton diagram of proposed system.

A. Algorithm

Steps of algorithm are as follow.

1: suppose c = c1, c2... cn be the single category feature and n is no of category tuple.

2: Select dependant attributes for categorical attributes.

3: Select any two category tuples.

4: Select 4 smallest values of dependent attributes for first category tuple as n1, n2, n3, n4 and two biggest values for second category tuple as n5.

5: Find distances between these values as,

$$d_1 = (n_1 - n_2)^2 \text{ represented as} N_1$$

$$d_2 = (n_3 - n_4)^2 \text{ represented as } N_2$$

$$d_3 = (n_1 - n_5)^2 + (c_1 - c_2)^2 \text{ represented as } N_3 + C_{12}$$

N3 has largest value among all.

6: Find distance equations as,

$$d_3 + d_1 = (N_1 + N_3) + C_{12}$$
$$= N_4 + C_{12} - - - - - - eq1$$
$$d_2 - d_3 = (N_2 - N_3) - C_{12}$$
$$= -N_5 - C_{12}$$
$$= -(N_5 + C_{12})$$

As distance is never negative,

$$= N_5 + C_{12} - - - - - - eq2$$

7: Adding eq1 and eq2 give final dissimilarity between selected categories that is C12.

8: Repeat from step 3 till get dissimilarity between all category tuples.

9: Apply improved k means algorithm on final dataset with categories values.

IV. IMPLEMENTATION ENVIRONMENT

The proposed system achieves the code separation using MVC design pattern. In MVC pattern M stands for Model, V stands for View and C stands for Controller. Model basically handle all business logic, in this case all algorithm steps like selecting categories, calculating distances, finding means, dividing data into clusters etc. Model is most essential part of system. Views are nothing but graphical interfaces which user going see and interact with system. View separates graphical interfaces from business logic, so graphical interface can be changed any time without changing business logic. Finally controller is connection between model and view. The selected implementation environments for proposed system are:

- PHP
- PHPDesktop
- Apache Server
- MySQL

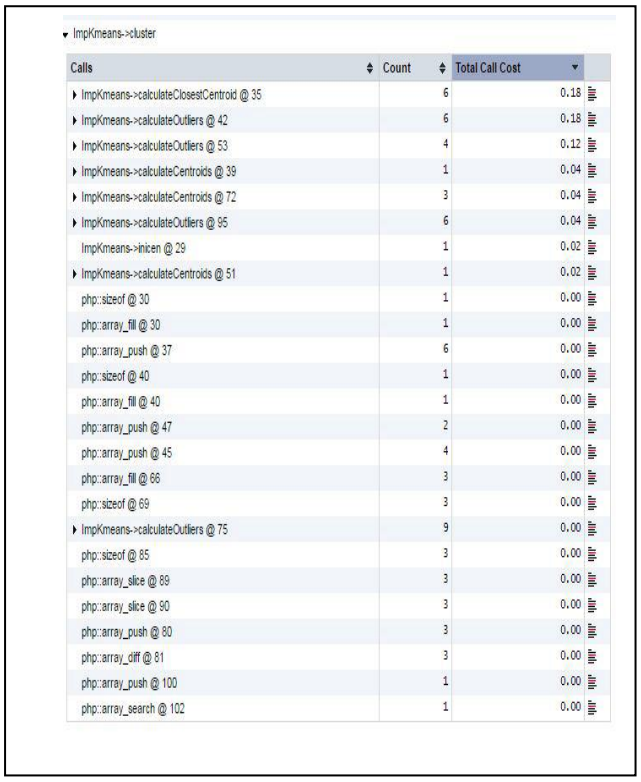
V. MATHEMATICAL MODELLING

A. Dissimilarity in Dichotomous Variable

Dichotomous variables are designed to give an either/or response. As it has only two values, variables either different or same. So they add same amount of dissimilarity percentage in cluster formulation. Let c1 and c2 be two dichotomous categorical tuples. Then distance d (c1, c2) is as shown in following equation

d (c1, c2) = 0/1

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 07, July-2016



ImpKmeans->cluster

Calls	Count	Total Call Cost
ImpKmeans->calculateClosestCentroid @ 35	6	0.18
ImpKmeans->calculateOutliers @ 42	6	0.18
ImpKmeans->calculateOutliers @ 53	4	0.12
ImpKmeans->calculateCentroids @ 39	1	0.04
ImpKmeans->calculateCentroids @ 72	3	0.04
ImpKmeans->calculateOutliers @ 95	6	0.04
ImpKmeans->inicen @ 29	1	0.02
ImpKmeans->calculateCentroids @ 51	1	0.02
php::sizeof @ 30	1	0.00
php::array_fill @ 30	1	0.00
php::array_push @ 37	6	0.00
php::sizeof @ 40	1	0.00
php::array_fill @ 40	1	0.00
php::array_push @ 47	2	0.00
php::array_push @ 45	4	0.00
php::array_fill @ 66	3	0.00
php::sizeof @ 69	3	0.00
ImpKmeans->calculateOutliers @ 75	9	0.00
php::sizeof @ 85	3	0.00
php::array_slice @ 89	3	0.00
php::array_slice @ 90	3	0.00
php::array_push @ 80	3	0.00
php::array_diff @ 81	3	0.00
php::array_push @ 100	1	0.00
php::array_search @ 102	1	0.00

Fig. 2. Performance of Proposed System

B. Dissimilarity in Ordinal Variable

This kind of variable has multiple values but in predefined order. As it is in order difference between two simultaneous variables is always same. The difference between two ordinal variables is always multiplication of difference between two simultaneous variable and difference in their order.

Let c1 and c2 be two ordinal categorical tuples. d(s) be the difference between two simultaneous tuples and d(o) be the difference between order of c1 and c2. The distance d(c1; c2) is as shown in equation

$$d(c1, c2) = d(s) * d(o)$$

C. Dissimilarity in Nominal Variable

This kind of variable has multiple categories also not in order. So there is difficulty in calculating difference between nominal variables. In this case we use dependant variables on categorical attribute. Dependant attribute directly represent value of categorical data.

Let c1 and c2 be two nominal categorical tuples. tdep1 and tdep2 be the corresponding dependant variables. The distance d(c1, c2) is as shown in equation

$$d(c1, c2) = (tdep1 - tdep2)$$

VI. DISCUSSION ON PERFORMANCE

Performance testing is done using XDebug and Web grind tool. XDebug is a PHP extension which provides debugging and profiling capabilities. Web grind is an XDebug profiling web frontend in PHP5. Performance of proposed system is as shown in fig.

VII. CONCLUSION

This paper has seen An Improved K-Means Algorithm and its problems. This paper has seen the importance of categorical data as it is really common in real life situations. This paper has seen importance of clustering categorical data. The proposed algorithm calculated dissimilarity between categorical attributes which further used to formulate final clusters. This paper seen result generated by proposed system and its performance. But proposed system sill have several issues like accurate dependency calculation, performance issue etc. There are many issues could be seen in future.

ACKNOWLEDGMENT

I would like to sincere thanks to the peoples who support and help in this work. Firstly Mr M. C. Kshirsagar my guide, all other teaching staff, my friends and my family thank for great help from beginning to end.

REFERENCES

[1] Anupama Chadha and Suresh Kumar, " An Improved K-Means Clustering Algorithm: A Step Forward for Removal of Dependency on K" 2014 International Conference on Reliability, Optimization and Information Technology, Feb 6-8 2014

[2] Zhexue Huang and Michael K. Ng, "A Fuzzy k-Modes Algorithm for Clustering Categorical Data " IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 7, NO. 4, AUGUST

[3] I.H. Jarman, T.A. Etchells, P.J.G. Lisboa and C.M Beynon, "Clustering CategoricalData: A Stability Analysis Framework" IEEE , 2011.

[4] Dadong Yu, Dongbo Liu, Rui Luo, JianxinWangl, "Clustering Categorical Data Based on Maximal Frequent Itemsets," IEEE Sixth International Conference on Machine Learningand Applications, 2007.

[5] B.Suresh Kumar, H.Venkateswara Reddy, T.Ankamma Raju, Preethi Vennam, "Clustering Categorical Data Using Rough Membership Function," 2014 Sixth International Conference on Computational Intelligence and Communication Networks, 2014.

[6] ZHEXUE HUANG , "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values," Data Mining and Knowledge Discovery 2, 283304 (1998).

[7] Zhexue Huang , "Clustering Large Data Sets With Mixed Numeric And Categorical Values*," IJCA,

[8] Madhu Yedla, Srinivasa Rao Pathakota, T M Srinivasa, "Enhancing K-means Clustering Algorithm with Improved Initial Center," International Journal of Computer Science and Information Technologies, Vol. 1 (2) , 2010, 121-125.

[9] Ahamed Shafeeq B M and Hareesha K S , "Dynamic Clustering of Data with Modified K-Means Algorithm," 2012 International Conference on Information and ComputerNetworks (ICICN 2012)..

[10] Zengyou He, Xiaofe i Xu, Shengchun Deng, "Clustering Mixed Numeric and Categorical Data : A Cluster Ensemble Approach "High Technology Research and Development Program.

[11] Pramod Nutan Dhara, and Rishi Sayal, "A Comparative Analysis Beetween K-Means And Y-Means Algorithms In Fishers Iris Data Sets," International Journal of Engineering and Technology (IJET), Vol 5 No 1 Feb-Mar 2013 245.