Comparative Study of K-means Type Algorithms

[1]Barkha H.Desai, [2]Nisha Shah, [3]Hetal Bharat Bhavsar

[1]*M.E. Computer Science and Engineering, Parul Institute of Engg. & Tech, Waghodia,India*
[2]*Computer Science and Engineering, Parul Institute. of Engg. & Tech, Waghodia,India*
[3] *Information Technology Dept. Saradar Vallabhbhai Patel Inst of Tech, Vasad*

***Abstract*---K-mean clustering is a partitioning method which contain k cluster and n object. It partition a set of n object into k cluster so resulting intracluster similarity is high but intercluster similarity is low. K-means uses Euclidean distance for measure similarity in objects. It has a problem when clusters are differing size, densities, and non-global shapes. It can not handle outlier . Another problem with k-means is selection of variables. The k-means type algorithms cannot select variables automatically because they treat all variables equally in the clustering process. To overcome this problem several algorithm has been proposed which improve the clustering result based on variable selection. In this paper we compared the K-means type clustering algorithm like k-means, WK-means, GW-K means algorithm. The comparision shows that the GW-K-means provides better accuracy than other two algorithm.**

Keywords-- K-Means, similarity measures, clusters, weighted K-means, variable weight

I. INTRODUCTION

K-means clustering is partitioning 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. The k-means type clustering algorithms [1], [2] are widely used in real world applications such as marketing research [3] and data mining to cluster very large data sets due to their efficiency and ability to handle numeric and categorical variables that are ubiquitous in real databases. A major problem of using the k-means type algorithms in data mining is selection of variables. The k-means type algorithms cannot select variables automatically because they treat all variables equally in the clustering process. To overcome this problem several algorithm has been proposed one of this algorithm is WK-means. This algorithm is proposed by[7]. This algorithm can update automatically weight variables based on the importance of the variables in clustering. WK-means adds a new step to the basic K-means algorithm to update the variable weights based on the current partition of data. The variable weights produced by WK-means measure the importance of variables in clustering. The small weights reduce or eliminate the effect of noisy variables. The weights can be used in variable selection in data mining applications where large and complex real data are often involved. GW-K-means proposed in[8] which simultaneously weight variable groups and individual variables in clustering where high dimensional and complex data are involved. In this paper we compare the K-means, WK-means and GW-K-means algorithm in terms of accuracy. The rest of this paper is organized as follows: Section 2 is introduction of K-means .Section3 covers WK-means. GW-K-means algorithm explain in Section4.Comparison and conclusion cover in Section5 and Section6 respectively.

II. K-MEANS

K-means[10][7] is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroids. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycentre of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroids. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more.[6]

Let X={X; X2; . . .; Xn} be a set of n objects. Object Xi = (xi,1; xi,2; . . . ; xi, m) is characterized by a set of m variables (attributes). The k-means type algorithms [4], [5] search for a partition of X into k clusters that minimizes the objective function P with unknown variables U and Z as follows:

$$\boldsymbol{P(U,Z) = \sum_{l=1}^{k} \sum_{i=1}^{n} \sum_{j=1}^{m} u_{i,l}\, d(x_{i,j},\ z_{l,j})}$$

(1)
subject to

$$\sum_{l=1}^{k} u_{i,l} = 1 \quad 1 \le i \le n,$$

(2)

Where,

- U is an n * k partition matrix, ui;l is a binary variable, and uij =1 indicates that object i is allocated to cluster l;
- Z ={Z1, Z2, . . . , Zk} is a set of k vectors representing the centroids of the k clusters;
- D(xi;j; zl,j) is a distance or dissimilarity measure between object i and the centroid of cluster l on the jth variable. If the variable is numeric, then

$$d(x_{i,j}, z_{l,j}) = (x_{i,j} - z_{l,j})^2 \qquad (3)$$

If the variable is categorical, the

$$d(x_{i,j}, z_{l,j}) = \begin{cases} 0 & (x_{i,j} = z_{l,j}) \\ 1 & (x_{i,j} \neq z_{l,j}) \end{cases} \qquad (4)$$

The algorithm is called k-modes if all variables in the data are categorical or k-prototypes if the data contains both numeric and categorical variables [1].

The above optimization problem can be solved by iteratively solving the following two minimization problems:

1. PROBLEM P_1: FIX $Z = \hat{Z}$ AND SOLVE THE REDUCED PROBLEM $P(U, \hat{Z})$,
1. PROBLEM P_2: Fix $U = \hat{U}$ and solve the reduced PROBLEM $P(\hat{U}, Z)$.

Problem P1 is solved by

Algorithm

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

A major drawback of k-means type algorithms in data mining is selection of variables. The k-means type algorithms cannot select variables automatically because they treat all variables equally in the clustering process. Complexity of k-means algorithm is O(nkt).

III. WK-MEANS

The problem of k-means type algorithms in data mining is selection of variables. To avoid this problem K-means type new algorithm is proposed by[7] called WK-means. This algorithm is iteratively update the variable weight. WK-means adds a new step to the basic K-means algorithm to update the variable weights based on the current partition of data. The variable weights produced by WK-means measure the importance of variables in clustering. The small weights reduce or eliminate the effect of noisy variables.

Let W = {w1; w2; . . . ; wm} be the weights for m variables and β be a parameter for attribute weight w_j, we modify (1) as

$$P(U, Z, W) = \sum_{l=1}^{k} \sum_{i=1}^{n} \sum_{j=1}^{m} u_{i,l}\, \omega_j^{\beta}\, d(x_{i,j}, z_{l,j})$$

(5)

Subject to

$$\begin{cases} \sum_{i=0}^{n} u_{i,l} = 1, \ 1 \le i \le n \\ u_{i,l} \in \{0, 1\}, \ 1 \le i \le n, \quad 1 \le l \le k \\ \sum_{i=0}^{n} \omega_j = 1, \ 0 \le \omega_j \le 1. \end{cases}$$

(6)

The weights can be used in variable selection in data mining applications where large and complex real data are often involved but it is not always support for large data set. Complexity of WK-means algorithm is O(mnkt).

IV. GW-K-MEANS

WK-means, through it solves the problem of variable selection but it doesn't support for large dataset so another algorithm has been proposed by[8], named GW-K-means. This algorithm simultaneously update weight variable groups and individual variables in clustering high dimensional data. In this algorithm, the variables of the high dimensional data can be divided into several variable groups and a group weight is assigned to each variable group to identify the importance of the variable group. Simultaneously a variable weight is also assigned to each variable to identify the importance of the variable in the group. Both variable group weights and variable weights are used in the distance function to determine the clusters of objects. In this new algorithm, two additional steps are added to the iterative k-means clustering process to automatically compute the variable group weights and the variable weights.

Let X={X1,X2,…,Xn} be a set of *n* objects represented by the set *A* of *m* variables. Assume A is divided into *T* groups $\{G_t\}_{t=1}^{T}$ where $|G_t| = m_t$, $G_t \cap G_s = \emptyset$ for $s \neq t$, $U_{t=1}^{T} G_t = A$ and $\sum_{t=1}^{T} m_t = m$. Let W = $\{w_1, w_2, \ldots, w_T\}$ be the weights for $\{G_1, G_2, \ldots, G_T\}$ and V = $\{V_{t,i}\}$ be a set of *T* subsets of variable

weights where Vt = $\{v_1, v_2, \ldots, v_m\}$ is the subset of variable weights for variables in group Gt ($1 \le t \le T$). The clustering process to partition ***X*** into *k* clusters with weights for variable groups and individual variables in each group is modelled as minimization of the following objective function

$$P(U, Z, V, W) = \sum_{l=1}^{k} \sum_{i=1}^{n} \sum_{t=1}^{T} \sum_{j=1}^{m_t} u_{i,l} w_t v_{t,j} d(x_{i,j}, z_{l,j}) + \lambda \sum_{t=1}^{T} w_t \log(w_t) + \eta \sum_{t=1}^{T} \sum_{j=1}^{m_t} v_{t,j} \log(v_{t,j}) \quad (7)$$

Subject to

$$\begin{cases} \sum_{l=1}^{k} u_{i,l} = 1, \ 1 \le i \le n \\ u_{i,l} \in \{0,1\}, 1 \le i \le n, 1 \le l \le k \\ \sum_{t=1}^{T} u_{i,l} = 1, \ 0 \le w_t \le 1 \\ \sum_{j=1}^{m_t} v_{t,j} = 1, \ 0 \le v_{t,j} \le 1, 1 \le t \le T \end{cases} \quad (8)$$

where

- *U* is an *n*k* partition matrix whose element $u_{i,l}$ is a binary variable and $u_{i,l} = 1$ indicates that object *I is allocated to cluster l.*
- *Z={Z1,Z2,…Zk}is a set of k vectors representing the centers of the k clusters;*
- $\lambda > 1, \eta > 1$, are two given parameters;
- $d(x_{i,j}, z_{i,j})$ is a distance or dissimilarity measure between object i and the center of the cluster l on the j-th feature. If the feature is numeric, then

GW-K-means algorithm is use for very large dataset for calculating automatic variable weighted for group and individual variable with more accurate result.

V. COMPARISON

In the following Table 1, we had presented comparison of different K-means type algorithm. The major difference between these three algorithm is in terms of variable selection method, the type of data handled and accuracy of generated clusters.

Characteristics	**K-Means**	**WK-Means**	**GW-K-Means**
Name	K-Means	Automated Weighted Variable	Weighted Variable Group
Data support	Limited volume of numerical dataset	Often support for large and complex data set	Very High dimenational data
Constraint	**D**ifficulty in produced the quality of cluster **N**ot support automatic variable selection	**N**ot scalable for the large Data set	Nil
Accuracy	Less	More accurate than k-means	More accurate than W-K-means

Table 1: Comparison of K-means, WK-means and GW-K means

VI. CONCLUSION

In this paper, we have presented comparison of three clustering algorithms: K-means, WK-means and GW-K-means. The WK-means and GW-K-means have been proposed to solve the problem of variable selection in K-means algorithm. From the comparison we conclude that the GW-K-means is better than K-means and WK-means, because it reduced the effect of noise variables and improve the accuracy of clustering process.

VII. REFERENCES

[1]Z. Huang, "Extensions to the k-Means Algorithms for Clustering Large Data Sets with Categorical Values," Data Ming and Knowledge Discovery, vol. 2, no. 3, pp. 283-304, 1998.
[2]J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observation," Proc. Fifth Berkeley Symp. Math. Statistica and Probability, pp. 281-297, 1967.
[3]P.E. Green, F.J. Carmone, and J. Kim, "A Preliminary Study of Optimal Variable Weighting in k-Means Clustering," J. Classification, vol. 7, pp. 271-285, 1990.
[4]J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observation," Proc. Fifth Berkeley Symp. Math. Statistica and Probability, pp. 281-297, 1967.
[5]M. Anderberg, Cluster Analysis for Applications. Academic Press,1973.
[6]http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html
[7] Joshua Zhexue Huang, Michael K. Ng, Hongqiang Rong, and Zichen Li,"Automated Variable Weighting in k-Means Type Clustering" IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 27, NO. 5,pp-658,2005.
[8] "GW-K-Means Clustering Algorithms Weighted Variable Groups" Vol. 5 No.1/ Jan. 2011,pp32-39
[9] http://www.cs.uiuc.edu/homes/hanj/bk2
[10] http://www.esnips.com/doc/bd2547cc-4904-4056-98f6-633de30ef690/Data-Mining-Concepts-and-Techniques---J.Han--M.Kamber