A Comparative Study Of K-Means And Weighted K-Means For Clustering

Anand M. Baswade
M.Tech. Student of CSE Dept.
SGGSIE&T, Nanded, India

Kalpana D. Joshi
M.Tech. Student of CSE Dept.
SGGSIE&T, Nanded, India

Prakash S. Nalwade
Associate Professor
SGGSIE&T, Nanded, India

Abstract

k-mean [1] is one of the most important algorithm for Clustering. Major problem of using k-mean type algorithms in Data mining is selection of variables (attributes). k-meanalgorithm can not select variablesautomatically because they treat all variables equally in clustering process that result in poor clustering. To obtain good clustering results it is important to identify the subset of variables from all variables. So that subset of variables can be used for clustering. New k-mean type clustering algorithm called W-k-mean [2] that can automatically calculate variable weights. So the algorithm can be used as variable selection in data mining application where large and complex real data are often involved. This paper gives detail information about k-mean and W-k-mean algorithms and their performance.

Keywords: *Data mining, clustering, feature evaluation and selection.*

1. Introduction

Data clustering is a process of 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 define criteria. Means grouping the data into clusters according to their internal character, the element in each cluster should have as similar character as possible, the difference between clusters should be as big as possible. The k-means clustering algorithms [1],[3] are widely used in real world applications such as marketing research and data mining to cluster very large data sets due to their efficiency. Among clustering algorithms, k-mean clustering algorithm can be applied in many fields, including image and audio data compression, pre-process of system modelling with radial basis function networks, and task decomposition of heterogeneous neural network structure.

There are some weaknesses in k-mean algorithm like selection of number of cluster, selection of initially centroids etc. but in this paper we discussed about some other problem of k-mean and solution to that problem is W-k-mean that is also discussed in this paper.

A major problem of using the k-means 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. In practice, selection of variables for a clustering problem such as customer segmentation is often made based on understanding of the business problem and data to be used. Tens or hundreds of variables are usually extracted or derived from the database in the initial selection which forms a very high-dimensional space. It is well-known that an interesting clustering structure usually occurs in a subspace defined by subset of the initially selected variables. To find the clustering structure, it is important to identify the subset of variables [4],[5]. So the new algorithm which is proposed called W-k-mean helps to find out those variables which are most important and which are least important by calculating weightof variablesduring each iteration. So those variables which having least weight can be removed and clustering is performed on remaining variables to achieve good clustering results.

The variable weights produced by W-k-means measure the importance of variables in clustering. The small weights reduce or eliminate the effect of insignificant (or noisy) variables. The weights can be used in variable selection in data mining applications where large and complex real data are often involved.

In this paper we have studied the most important data mining algorithmk-means andextension of k-means by adding weights to variables that is W-k-mean for clustering.

2. k-Mean Algorithm

The traditional K-mean algorithm is based on decomposition, most widely used in data mining field. The concept is use K as a parameter, Divide n object into K clusters, to create relatively high similarity in the cluster, relatively low similarity between clusters. And minimize the total distance between the values in each cluster to the cluster center. The cluster center of each cluster is the mean value of the cluster. The calculation of similarity is done by mean value of the cluster objects. The distance between the objects is calculated by using Euclidean distance. The closer the distance, bigger the similarity of two objects, and vice versa.

Let $X_i = \{X_1, X_2, X_3, \dots \dots X_n\}$be the *n*objects.

$X_{i,m} = \{x_{i,1}, x_{i,2}, \dots \dots x_{i,m}\}$ be the *m* variables.

$Z_l = \{Z_1, Z_2, \dots \dots Z_k\}$ be the *k* centroids.

$u_{i,l} = \begin{cases} 1 \\ 0 \end{cases}$;$u_{i,l}$ is *n x k* matrix and it is 1 if objecti allocate in cluster *l.*

2.1 k-Mean Algorithm

Input: k, Data.

Output: n objects into k clusters.

Step 1. Choose k random positions in the input space.

Step 2. Assign the cluster centres Z_l to those positions.

Step 3. For each $X_i \epsilon$Data

-compute distance $d(x_{i,j}, z_{l,j})$ for each Z_l

$$d(x_{i,j}, z_{l,j}) = \sqrt{\sum_{j=1}^{m} |x_{i,j} - z_{l,j}|^2}$$

-Assign X_i to the cluster with the minimum distance.

$$u_{i,l}=1 \text{ if } \sum_{j=1}^{m} d(x_{i,j}, z_{l,j}) \leq \sum_{j=1}^{m} d(x_{i,j}, z_{t,j}) \quad \text{for } 1 \leq t \leq k$$

$$u_{i,t} = 0 \text{ for } t \neq l$$

Step 4. For each Z_lMove the position of Z_l to the mean of the points in that cluster:

$$Z_{l,j} = \frac{\sum_{i=1}^{n} u_{i,l}\, x_{i,j}}{\sum_{i=1}^{n} u_{i,l}} \quad \text{for } 1 \leq l \leq k \text{ and } 1 \leq j \leq m.$$

Step 5. Stopping criteria

-No (or minimum) re-assignments of data points to different clusters.i.e. $u_{i,l}$ Remain unchanged.

OR

- No (or minimum) change of centroids.
i.e. Z_l Remain unchanged.

2.2 Performance analysis

2.2.1 Advantages:

1) K-means is classical algorithm to resolveclustering problems, simple and quickly and it is easy to implement and Understand.

2) Complexity of k-mean algorithm is O(ntk).where n is number of objects, t is number of iteration and k is number of cluster.

2.2.2 Disadvantages:

K-means only can be used under the situation that the average value has been defined. This may not suit some applications, such as mobile objects clustering, data concerned about classified attributes.

1) In k-mean algorithm user need to specify the number of cluster that is k.

2) It's sensitive to the initial centroids and change in initial centroids can lead to different clustering results with different initial value.

3) k-means is not fit to non-convex cluster, or big difference on size. Besides, it's sensitive to noisy data and isolated points data, a little data like this can make huge effects on average values. In the other way we can say k-mean algorithm is unable to handle noisy data and outliers.

The above listed disadvantages are discussed in many research papers but there is one more major problem in k-mean that is the k-means type algorithms treat all variables equally in deciding the cluster memberships of objects. This is not desirable in many applications such as data mining where data often contains a large number of diverse variables. A cluster structure in a given data set is often confined to a subset of variables rather than the entire variable set. Inclusion of other variables can only obscure the discovery of the cluster structure by a clustering algorithm. To overcome this major problem a new k mean type algorithm is designed that is W-k-mean.

3. W-k-Mean Algorithm

New k-mean type clustering algorithm called W-k-mean [2] that can automatically calculate variable weights. The variable weights produced by the algorithm measures the importance of the variable in clustering. So the algorithm can be used as variable selection in data mining application where large and complex real data are often involved. Select higher weighted variables and remove lower weighted variables for good clustering results.

Let $W_j = \{W_1, W_2, \ldots\ldots W_m\}$ be the weights for m variables such that $\sum_{j=1}^{m} W_j = 1$.

Objective of W-k-Mean is to Minimize:

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

3.1W-k-Mean Algorithm

Step 1. Randomly choose an initial $Z^0 = \{Z_1, Z_2, \ldots\ldots Z_K\}$ *and randomly generate a set of initial weights* $W^0 = \{w_1{}^0, w_2{}^0, \ldots\ldots w_m{}^0\}$ $(\sum_{j=1}^{m} W_j = 1)$. *Determine* U^0 *such that* $P(U^0, Z^0, W^0)$ *is minimized. set t=0;*

Step 2. Let $\acute{Z}=Z^t$ *and* $\hat{W}=W^t$ *,solve problem* $P(U, \acute{Z}, \hat{W})$ *to obtain* U^{t+1}*.if* $P(U^{t+1}, \acute{Z}, \hat{W}) = P(U^t, \acute{Z}, \hat{W})$*, output* $(U, \acute{Z}, \hat{W})$ *and stop ; otherwise go to step 3;*

$$u_{i,l} = 1 \text{ If } \sum_{j=1}^{m} w^{\beta}{}_{j}\, d(x_{i,j}, z_{l,j}) \le \sum_{j=1}^{m} w^{\beta}{}_{j}\, d(x_{i,j}, z_{t,j}) \quad \text{for } 1 \le t \le k$$

$$u_{i,t} = 0 \text{ for } t \ne l$$

Step 3. Let $\hat{U} = U^{t+1}$ *and* $\hat{W}=W^t$ *,solve problem* $P(\hat{U}, Z, \hat{W})$ *, to obtain* Z^{t+1}*. If* $P(\hat{U}, Z^{t+1}, \hat{W}) = P(\hat{U}, Z^t, \hat{W})$*, output* $(U, Z^t, \hat{W})$ *and stop ; otherwise go to step 4;*

$$Z_{l,j} = \frac{\sum_{i=1}^{n} u_{i,l}\, x_{i,j}}{\sum_{i=1}^{n} u_{i,l}} \quad \text{for } 1 \le l \le k \text{ and } 1 \le j \le m$$

Step 4. Let $\hat{U} = U^{t+1}$ *and* $\acute{Z}=Z^{t+1}$ *,solve problem* $P(\hat{U}, \acute{Z}, W)$ *, to obtain* W^{t+1}*. if* $P(\hat{U}, \acute{Z}, W^{t+1}) = P(\hat{U}, \acute{Z}, W^t)$ *, output* $(\hat{U}, \acute{Z}, W^t)$ *and stop ; otherwise, set t=t+1 and go to step 2.*

$$w_j = \begin{cases} 0 & \text{if } D_j = 0 \\ \dfrac{1}{\sum_{t=1}^{h} \left[\frac{D_j}{D_t}\right]^{\frac{1}{\beta-1}}} & \text{if } D_j \ne 0 \end{cases}$$

Where

$$B > 1 \text{ and } \beta \le 0,\ D_j = \sum_{l=1}^{k}\sum_{i=1}^{n} u_{i,l}\, d(x_{i,j}, z_{l,j})$$

When $\beta = 1$, w_j is equal to 1 for the smallest value of D_j .The other weights are equal to 0. Although the objective function is minimized, the clustering is made by the selection of one variable. It may not be desirable for high-dimensional clustering problems.

When $0 < \beta < 1$, the largerD_j, the largerw_j, so does $w^{\beta}{}_{j}$.This is against the variable weighting principal, so we cannot choose $0 < \beta < 1$.
When $\beta > 1$, the larger D_j, the smaller w_j, and the smaller $w^{\beta}{}_{j}$. The effect of variable x_j with large D_j is reduced.
When $\beta < 0$, the larger D_j , the larger w_j However, $w^{\beta}{}_{j}$ becomes smaller and has less weighting to the variable in the distance calculation because of negative β.
From the above analysis, we can choose $\beta < 0$ or $\beta > 1$ in the W-k -means algorithm

Since the W-k -means algorithm is an extension to the k -means algorithm by adding a new step to calculate the variable weights in the iterative process, it does not seriously affect the scalability of the k-means type algorithms in clustering large data; therefore, it is suitable for data mining applications. the principal for variable weighting is to assign a larger weight to a variable that has a smaller sum of the within cluster distances and a smaller one to a variable that has a larger sum of the within cluster distances.
After getting weights of variables we can selects higher weighted variables and neglectlower weighted variables for better clustering results.MATLAB is used for implementing k-mean and W-k-mean algorithms and data sets obtained from UCI Machine Learning Repository [6] for comparing the results of k-mean and W-k-mean algorithm.

4. Conclusions

A new k-means type algorithm called W-k-Means that can automatically weight variables based on the importance of the variables in clustering. By knowing the weight of variables we can select those variables which are good for better clustering results. This capability is very useful in data mining application where large and complex dataset are often involved. The weights can be used to identify important variables for clustering. The variables which may contribute noise to the clustering process can be removed from data in future analysis.

5. References

[1] J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observation," Proc. Fifth Berkeley Symp.Math.Statistica and Probability, pp. 281-297, 1967.

[2] Joshua Zhexue Huang , Michael K. Ng , HongqiangRong , Zichen Li, "Automated Variable Weighting in k-Means Type Clustering," IEEE Transactions on Pattern Analysis and Machine Intelligence, v.27, n.5, pp.657-668, May 2005.

[3] 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.

[4] E. Fowlkes, R. Gnanadesikan, and J. Kettenring, "Variable Selection In Clustering," J. Classification, vol. 5, pp. 205-228, 1988.

[5]J.H. Friedman and J.J. Meulman, "Clustering Objects on Subsets ofAttributes,"J. Royal Statistical Soc. B., 2002.