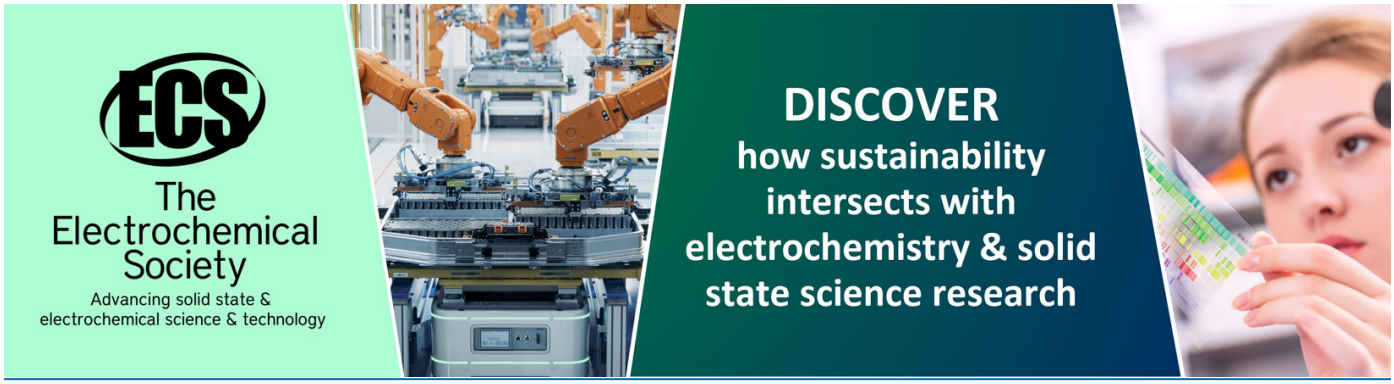
The Expansion of Initial Point Algorithm for K-Modes Algorithm



IOP Conf. Series: Journal of Physics: Conf. Series **930** (2017) 012027 doi:10.1088/1742-6596/930/1/012027

The Expansion of Initial Point Algorithm for K-Modes Algorithm

Juliandri[1], Zarlis M[2], and Situmorang Z[3]

[1] Department of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia.
[2] Department of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia.
[3] Department of Computer Science, Universitas Katolik Santo Thomas, Medan, Indonesia.

me@juliandri.com[1]

Abstract. The determination of the starting point in the k-modes algorithm is taken by random. Of course, such a thing can lead to an iteration of unpredictable numbers and accuracy. Therefore, it is necessary to develop different algorithms that are used to determine the starting point with hierarchical agglomerative clustering approach instead of randomly selecting the starting point in the initial iteration. At the end of this research is expected clustering process can produce more efficient iteration. The result of determining the value generated in this algorithm is the incorporation of a number of cluster central points on the variables based on the calculation of the approach which has the average linkage algorithm. Followed by calculating the difference of objective function on each iteration, of course, finished clustering process on k-modes. Iteration will stop after the difference of objective function is smaller than the specified limit.

1. Introduction

Dividing a group of objects into several groups and inserting them according to their homogeneity level is the basic operation in data mining. The initial step is the exploration of data in terms of forming a model used to recognize data patterns. Instinctively, it becomes the daily habit of each person to sort by the same value and classify or classify the object into different groups. Objects can be anything people, products, and services, objects and territories.

In relation to the object clustering method, data clustering problems are more widely studied in data mining and machine learning literature, because so diverse applications can be implemented, among others, exploring data information, learning systems, segmentation and others [1]. In addition, this method is also widely used to solve research problems from various disciplines such as medicine, data security and so on [9].

The problems within each study are almost always different and the use of various techniques in the clustering method is done in the hope of finding the right technique for each data type and output to be generated. In the classification of a proper technique used, the clustering method is divided into two types of methods, namely hierarchical clustering and non-hierarchical clustering [11]. The best-known technique in non-hierarchical clustering is the k-means algorithm.

Characteristics of clustering method operation in the most prominent data mining is not a problem if the existing data set contains a numeric value or a category type. The capability of the algorithm in the

clustering method is expected to solve the problem with the level of complexity that has various types of numerical attributes or category-shaped [10]. To that end (Huang, 1997) in his research describes k-modes algorithm to handle boundaries on category-specific datasets and develop new algorithms adapted from k-means algorithms in the hope that the process has the same efficiency level with the algorithm.

But in the k-modes algorithm, the determination of the starting point is done randomly. This results in the coordinates of the center point not accurately representing every object contained in each cluster. The discussion of the percentage of the total improper object enters into each cluster is exposed by (Sun et al, 2002) and at the end of the calculation, iteratively k-modes substitutes the cluster center based on the number of frequencies of the object entering the cluster member [6]. In addition, this algorithm can also be trapped in local optima [3].

The problem that arises is how to determine the starting point for making the cluster central coordinates on this algorithm not determined by random. This is done because the selection of the cluster center is very important. The cluster center will directly impact the cluster information at the end of the iteration [2]. In his research, Cao also discussed the use of the MaxMin algorithm by developing the use of the average number of frequencies to replace random initialization of the MaxMin algorithm used in generating the cluster center that would be used to replace the cluster center on k-modes algorithms.

The same is also done by using Centroid Linkage algorithm which is a variant of Average Linkage algorithm to form the center point and Density-Based Multiscale Data Condensation algorithm to calculate the density at One particular area. The use of centroid linkage modification is also done by (Hadinata et al, 2017) to help determine more precisely the focal point of the fuzzy c-means algorithm. (He et al, 2011) in his research developed k-modes by changing distance calculations and incorporating weighted value attributes so as to produce a higher degree of accuracy whereas (Sangam, 2015) describes a concept of equality calculation using the coefficient basis of information entropy and successfully adds Value accuracy.

Therefore, in this research is presented about the formation of starting point by using the average linkage algorithm which is part of the linkage method and give weighting to the distance calculation to obtain higher accuracy. From the previously described method, it is expected to determine the value of the centroid more precisely without having to do the random process at the initialization stage and produce a more accurate distance accuracy calculation.

2. Basic Concept of Clustering Method

2.1. K-Modes

The k-modes algorithm is a simpler form of k-prototype described in the paper (Huang, 1997). In this algorithm there are three important modifications to the k-modes algorithm, including the use of differences in non-measurement, substituting k-means with k-modes and using frequency methods to update the mode values (Huang, 1997).

The process that occurs in the k-modes algorithm is $\{S_1, S_2, ..., S_k\}$ is the partition of X, where Si ¾ for lm1mk, and $\{Q_1, Q_2, ..., Q_k\}$ are values that often arise from $\{S_1, S_2, ..., S_k\}$. Thus, the total can be defined as follows:

$$E = \sum_{i=1}^{k} \sum_{i=1}^{n} y_{i,l} d(X_i, Q_i) \quad (1)$$

Where y_i, l is the element of the partition matrix $Y_{n\,x\,l}$ and d is like the equation below:

$$d(X, Y) = \sum_{j=1}^{m} \delta(x_j, y_j) \quad (2)$$

Where:

$$\delta(x_j, y_j) = \begin{cases} 0 & (x_j = y_j) \\ 1 & (x_j \neq y_j) \end{cases} \quad (3)$$

The function d (X, Y) provides an equation that is important for each category of each attribute. If calculating the frequency of the categories in the data set can be defined the following inequalities:

$$d_{\chi^2}(X, Y) = \sum_{j=1}^{m} \frac{(n_{x_j} + n_{y_j})}{n_{x_j} n_{y_j}} \delta(x_j, y_j) \quad (4)$$

Where n_{xj} n_{yj} is a numeric value that is an object within a dataset having categories x_j and y_j. Since, $d\chi^2$ (X, Y) is the equation for chi-square distance. Similar to the k-means algorithm, the purpose of clustering X is to find a set $\{Q_1, Q_2, ..., Q_k\}$ that can minimize the value of E.

The k-modes algorithm contains steps that can be followed as follows:

1. Select k initials modes, one for each cluster.
2. Allocate the object into the nearest cluster based on the value *d*. Update the mode values of each cluster once allocated by theory.
3. After each object is allocated into the cluster, re-test the inequality and compare it with the existing mode. If, an object is found to be closer to another cluster then move the object into the cluster. Update the mode values of each cluster.
4. Repeat step 3 until no object changes the cluster after the whole testing process is undertaken by the entire data set.

Just as with the k-means algorithm, the k-modes algorithm also produces a local optima solution tied to initialized initial mode values.

2.2. Clustering with the Hierarchy Approach

There are two types of hierarchical clustering, including agglomerative and divisive. Agglomerative determines the number of clusters based on the number of objects in the dataset and is followed by the search for the most similar pair. Then the similar pair is put together into a single cluster process continued until the desired number of clusters is obtained [1]**.** Therefore, the measure to determine the inequality of each cluster must be determined first [4]. This method has three types of variations including single linkage, complete linkage, and average linkage.

The divisive hierarchical clustering method has a process opposite to agglomerative clustering, starting with a large cluster and dividing the cluster into sections as desired [4].

2.3. Average Linkage

In the average linkage, the distance between two clusters is defined as the overall mean distance of each member of the object, especially the average linkage distance between C_i and C_j.

$$D_{AV}(C_i, C_j) = avg\ x \in C_i, y \in C_j(dist(x, y)) \quad (5)$$

Where, the function of dist is the function of choosing distance calculation [1].

2.4. Criteria Reject Measure

To evaluate if in the cluster studied has good quality or not until now still an issue. In the final stages, the evaluation process is always done by every researcher. Therefore, several evaluation criteria have been developed. These criteria are usually divided into two categories, namely internal and external.

In this study, the use of internal criteria as a benchmark for comparison between clusters using some similarity rules. This process usually measures the level of homogeneity of objects between clusters and measures the level of homogeneity of objects within a cluster. This process does not use external information other than the data contained in the dataset itself [9].

2.5. Data Set

The dataset in this study is Internet advertisements obtained from the UCI Learning Machine Repository or UCI Dataset. This dataset contains online advertisements contained on internet pages. This data set was donated by Nicholas Kushmerick in 1998 with the Learning to remove Internet advertisements published in 1999.

This study uses a C4.5 algorithm with an accuracy level of 97% in predicting the state of the image in an advertisement on the website page. The number of data sets consists of 2,821 nonads and 458 ads.

3. Methodology

3.1. Introduction

At this time practitioners and academics have been doing a lot of testing and research on new methodologies as well as mixing methodologies that exist in the soft computing and machine learning sections including combining both fields.

For that reason, the process in this methodology incorporates a detailed draft plan that is required to complete the research and also includes the basics of the method used. In the process of describing any limitations in research methods used then included in the research stage and work diagram.

3.2. Research Framework

There are some stages carried out in this Study as followed:

1. *Initial Research*. At this stage collected research materials from various sources of literature, such as books, journals (both print and online), proceedings, magazines, articles and other relevant sources. In addition, it also conducts consultations with supervisors in the preparation and thesis writing.
2. *Data Collection*. The dataset used in this study is Internet Advertisements obtained from the UCI Learning Machine Repository or UCI Data set.
3. *Early Point Determination*. The initial initialization process of the k-modes algorithm will be replaced by the determination of the starting point by using the average linkage approach and to evaluate the distance will be the addition of weighted values to the distance measurements to produce a more appropriate cluster center.
4. *Formation of New Algorithms*. The starting point has been obtained based on the algorithm adapted from the average linkage and its output is the coordinate center or centroid point which is more representative of the cluster center. Then, the cluster center is used as a cluster center at the initial initialization stage.
5. *Testing and Evaluation*. In this section will be discussed the value of homogeneity and heterogeneity of each cluster, by looking at and analyzing the cluster whether the cluster is ideal or not. These results are compared with k-modes that are not modified (native) methods. The results are the number of iterations, the average objective function and standard deviation and the homogeneity and heterogeneity of each cluster.
6. *Algorithm Creation*. After producing a new method that matches an ideal cluster value next is to write the work stages in the form of pseudocode.
7. *Conclusion*. From the output that has been produced, it can be concluded by writing it into report form, starting from the theoretical study, methodology procedure, testing, and suggestion that can be given for further research process.

4. Experimental Result and Discussion

4.1. Introduction

In this stage the authors will describe the results and discussion of the development of the starting point in the clustering method of the k-modes algorithm, the starting point which was originally formed by random, is now changed by using agglomerative clustering average linkage method.

4.2. Performance of K-Modes Algorithm

In the k-modes algorithm clustering method, the initial membership function is first initialized by means of a random way. The result is that all of the V vectors that become cluster centers become random at the beginning of the iteration. The size of the distance used for the data is categorical as in k-modes is by replacing the average with clusters with the modes or values that often appear.

This modification removes restrictions that only exist in k-means when using category values in the clustering process. For the measurement of the previously mentioned distance is simple mismatching or it can be defined as the number of unequal attributes between two objects as in equation 3.

4.3. K-Modes Pre-processing Data Set and Largest Frequency Strategy

Before performing the process of data classification, it is necessary to analyze the data to be used as test data. The dataset is obtained from a trusted source in data collection. In this research, data retrieval through UCI Machine Learning Data set site has been known as the curator of research data. Data used as the dataset is the data "Learning to Remove Internet Advertisements" which has been studied previously. Consisting of 3,279 rows and 1,558 columns.

In the dataset there are thousands of rows and columns which are divided according to the type of content: *url*image+button*, *origurl*labyrinth*, *ancurl*search+direct*, *alt*your*, and *caption*and*. Based on the classification of types of ads on the data then it can be done the category division, namely "ad" and "nonad". And as an additional display that affects the media appear, whether enlarged or not. It is characterized by a ratio of magnitude or aratio whose value is more than 1.

The ratio of 1 and 0 in k-modes is calculated based on the frequency of the number "1". Therefore, if all the data that have the type of url * image + button, origurl * labyrinth, ancurl * search + direct, alt * your and caption * and, but is the type of ad "ad" which has the most frequencies summed and so With "nonad" and which has a "ratio> 1" summed in cross.

After the data containing the 1 is summed then proceeded to form a percentage value based on the sum of data 1, divided by the total of all data in rows and columns counted cross-tabulated. Then, calculate the entire existing data to be the percentage value of the comparison between the amount of data that is worth 1 and the total data. Change of category data into percentage form is done, because the procedure of starting point determination is done by using agglomerative clustering average linkage algorithm, where the processing contained in the algorithms is not categorical data.

4.4. Testing K-Modes Algorithm

The testing stage is done by using the dataset which has been done percentage calculation. Testing process and calculation with percentage data, in this case, a k-modes algorithm using distance calculation Euclidean Distance as a benchmark of capture value similarity.

The testing process to be performed is to compare the quality of the cluster and the number of iterations between the two types of a k-modes algorithm that have changed the starting point using the average linkage algorithm with the k-modes algorithm whose starting point is still using the random way. The test also uses the same parameter, where the maximum iteration is 200 with the allowable error value is 0.05 and the cluster number = 3. Next, do the random value generation process to generate the random point of the center. Then, calculate the distance of each row of data to the cluster center by using the Euclidean Distance calculation.

Then, calculate the distance on the next lines until the overall distance of each cluster is calculated. After calculating the entire distance is done then the next step is to calculate the minimum distance to each cluster. The minimum distance or the smallest in each cluster is expressed as the cluster center on the data. If the minimum is in the first cluster the data is declared a member of that cluster. Furthermore, the results can be presented in a graphical form shown in the figure below.

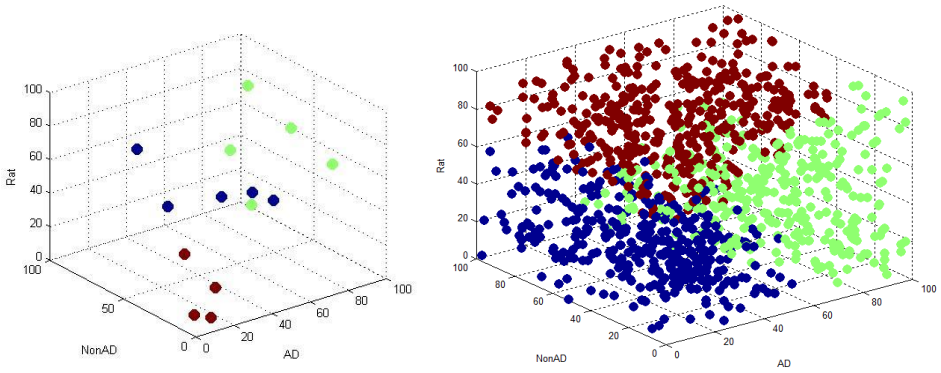


Fig. 1 Scattering 3D 15 lines of data.

4.5. Testing K-Modes Algorithm with Early Point Determination

The k-modes testing stage using the starting point determination begins by searching for variables that are always changed to be fixed in the next iteration.

In k-modes that have been shown previously iterative changes are done in the centroid. Where the section determines the position of the member, whether it is part of one cluster or another part. For that in this study, the centroid is searched by using agglomerative clustering average linkage algorithm.

4.6. Average Linkage Algorithm In Search Centroid Value

The calculation phase for the initial point search on this algorithm using agglomerative average linkage algorithm approach. The agglomerative average linkage algorithm in the cluster center search to determine the starting point begins by determining the number of clusters = number of rows or $c = n$. The agglomerative average linkage algorithm is expected to reduce the number of iterations contained in the k-modes algorithm by substituting an arbitrarily selected cluster center in the initial iteration.

In the average linkage selection algorithm done column by column and cluster center search begins by taking the first column as observation data, that is ad column. The value of the index for each cluster is used to unify the cluster that has the smallest overall value. Combine the C value with index i, j the smallest. Calculate the value of combinations i and j with an average of two values to fill in the new value. Next, calculate the smallest value between two-row groups whose numbers are not equal to the value itself. Having obtained the smallest value as a whole then combine the column that has the smallest value. In this explanation combine columns C and J.

Then, to fill the confused value with that column, calculate the value by using the mean between two values. Calculate the average value tangent to the combined column into one.

After the entire row is filled with an average of two values for the first column then the next step of the contents of the column is not tangent to the initial value or the contents back with the previous value. When this process is done then the first iteration of this algorithm is done. The next step repeats the process as in the previous process, which is to find the smallest value of the line and the smallest overall value to unify the columns.

4.7. Centroid Initialization On Calculation K-Modes Algorithm

The testing stage of k-modes calculation is done by counting all data in the same way and parameters. The parameters used are standard error value 0.05 with truth value 95%, the maximum iteration count is 200 iteration besides the same amount of data with previous and cluster number as much 3.

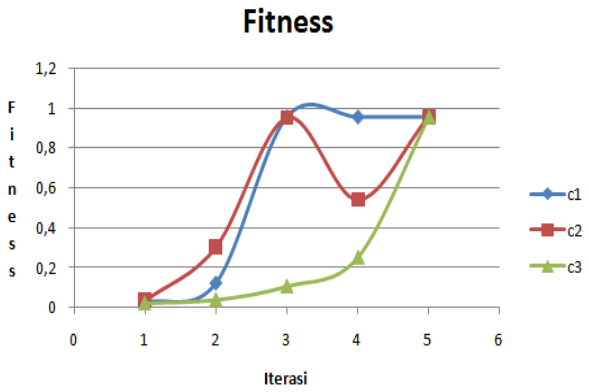
Next, calculate the distance of the center point to the observed data in the same way.

4.8. Comparison Between the Two Methods

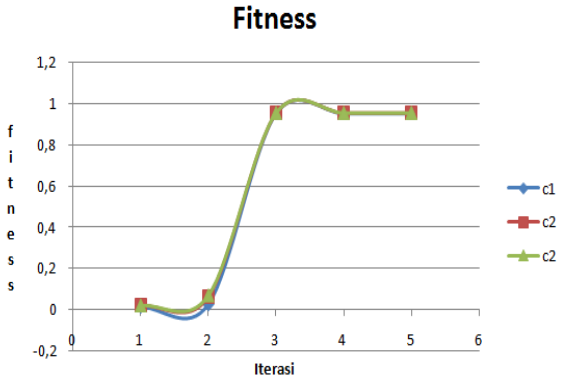
For comparison between k-modes using the determination of the center point of the average linkage and using only any number in determining the center of the cluster is shown in the graph.

The average difference in both methods is that the k-modes method with any cluster center changes to the 4th iteration. While the mean on k-modes using cluster center determination at the start of initialization is stable in the 2nd iteration. It can be seen in the picture below.

The ratio of fitness values between the two k-modes is shown in the figure below.



Fitness k-modes with any cluster center.



Fitness k-modes by using cluster center determination at the beginning of the iteration.

Fig. 2 Fitness value.

5. Conclusion

Dividing a group of objects into groups and inserting them according to their homogeneity level is the basis of operations in data mining. The initial step is the exploration of data in terms of forming a model used to recognize data patterns.

The use of starting point determination in the classification process is the focus of this research. Especially in the k-modes algorithm, the use of any value in determining the center point produces a number of fluctuating iterations from several experiments in this study. On the other hand, the use of procedurally determined procedural center values in k-modes is a starting point in obtaining better cluster quality without generating a number of fluctuating iterations.

From the result of the research, the starting point determined by using agglomerative clustering average linkage is more appropriate to fill the center point of k-modes algorithm compared with any value. In addition, after passing the Sum of Square Error calculation, the determination of the center point using any value is much greater than the central point determined through the agglomerative clustering average linkage.

The use of more diverse data sets would be better so that it is more representative of different types of datasets. In addition, cluster quality analysis process will be better if reviewed not only using one point of view.

References

[1] Aggarwal, C. C. 2014. Data Clustering Algorithms and Applications, Data Mining and Knowledge Discovery Series. Chapman& Hall/CRC Book., CRC Press, New York.

[2] Cao, Fuyuan., Liang, Jiye & Bai, Liang. 2009. A New Initialization Method for Categorical data Clustering. An International Journal : Expert System with Application. Elsevier. Vol. 36, DOI:10.1016/j.eswa.2009.01.060, pp. 10223-10228.

[3] Chaturvedi, Anil., Foods, Kraft., Green, Paul.E & Carrol, J. Douglass. 2001. K-modes Clustering. Journal of Classification. Vol. 18, DOI: 10.1007/s00357-001-0004-3. Pp. 35-55.

[4] Everitt, B.S., Landau,S., Leese,M. & Stahl, D. 2011. Cluster Analysis 5th Edition. King's College, London. Willey Series in Probability and Statistics. Jhon Willey & Sons, Ltd. UK.

[5] Hadinata, Edrian., Sembiring, Rahmat.W., Kusumasari, Tien Fabrianti & Herawan, Tutut. 2017. The Algoritm Expantion for Starting Point Determination Using Clustering Algorithm Method with Fuzzy C-Means. Recent Advances on Soft Computing and Data Mining. Springer.
[6] He, Zengyou., Xu, Xiaofei & Deng, Shengchun. 2011. Attribute Value Weighting in k-modes Clustering. Elsevier, Expert System With Applications. Vol. 38, DOI:10.1016/j.eswa.2011.06.027, pp. 15365-15369.
[7] Huang, Zhexue. 1997. A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets In Data Mining. Cooperative Research Centre for Advanced Computational System(ACSys). Autralia.
[8] Huang, Zhexue. 1997. Clustering Large Datasets With Mixed Numeric and Categorical Values. Cooperative Research Centre for Advanced Computational System(ACSys). Autralia.
[9] Maimon, O., & Rokach, L(Editor). 2010. Data Mining and Knowledge Discovery Handbook Second Edition. Springer. DOI. 10.1007/978-0-387-09823-4.
[10] Michalski, R., Bratko, I., Kubat, M. 1998. Machine Learning and Data mining:Methods and Applications. Wiley, New York.
[11] Yim, O., & Ramdeen, Kayle T. 2015. Hierarchical Cluster Analysis : Comparison of Three Linkage Measures and Application to Psychological Data. The Quantitative Method for Psychology. TQMP. Vol.11, No.1, DOI: 10.20982/tqmp.11.1.p008, pp.8-22.
[12] Sun, Ying., Zhu, Quiming. & Chen, Zhenxin. 2002. An Iterative Initial Points Refinement Algorithm for Categorical Data Clustering. Pattern Recognition Letters. Elsevier Science.Vol. 23, pp. 875-884.