

K-Means algorithm and modification using gain ratio

Ryan Dhika Priyatna, Tulus, Marwan Ramli

Computer Science, University of Sumatera Utara, Jl. Universitas Kampus USU,
Medan 20155

Abstract K-Means is one method in data mining that can be used to perform grouping clustering of data. Accurate data processing can be done by processing the data source. Each collection or data warehouse can provide important knowledge into valuable information, constraints on this method, if the cluster point is chosen randomly so that the resulting data may vary, if the value is not good, then the resulting grouping is less than optimal. Furthermore, failure to outliers in the process of grouping data include determining whether a data item is an outliers of a cluster of course and whether small amounts of data form a separate cluster. where the gain ratio is used to calculate the attribute's influence on the target of a data gain ratio is the development of the information gain, where the gain ratio eliminates the bias value of each attribute. The result of the research is to calculate the weights in each attribute by using the gain ratio and make the modeling and classification into the method of K-means.

1. Introduction

Data mining is a process of digging something meaningful by sifting through large data. In data mining there is a process that is clustering. Clustering is a method of analyzing data that has the purpose of grouping data that have the same characteristics. One method that uses clustering is K-means.

K-Means is a method of analyzing data that seeks to group data into groups. However K-means has a weakness in determining the starting point of randomly selected clusters (Random), which causes the value of membership will always change if re-calculated. This makes the researcher interested to do research on K-means method. In this research using two method / algorithm that is K-means and gain ratio from both methods produce performance comparison, So that performance can be measured in mean square error number of iterations performed.

Clustering is a grouping method based on the size of proximity (similarity). Clustering analyzes object data without notifying class labels. In many cases, class labeled data may be completely absent at the beginning. Clustering can be used to generate class labels for a group of data.

The low accuracy results of the K-means method in determining the starting point of the randomly selected cluster (Random), which causes the value of membership will always change if done re-calculation, In the algorithm is also not guaranteed the distance between each centroid is not spanned so that if there is two or more groups with adjacent mass center points then the result reduces the accuracy. The purpose of this study is to increase the accuracy of the K-means method by giving weight to each attribute by utilizing the gain ratio.



2. Methods

2.1 Clustering

Clustering is a grouping method based on the size of proximity. In many cases, class labeled data may be completely absent at the beginning. Clustering can be used to generate class labels for a group of data.

Data input for cluster analysis system is a set of data and similarity of size (or difference) between two data. While the output (output) of the cluster analysis is the number of groups that make up a partition or partition structure of the data set. One additional result of cluster analysis is the general description of each cluster and it is essential for a deeper analysis of the characteristics of the data set.

In Clustering are known four data types namely

1. Variable interval,
2. Binary variables,
3. The nominal, ordinal, and ratio variables,
4. Variable with other data types.

In figure 1 shows a simple example of the clustering process, where 3 clusters in the data can be easily identified. The similarity criteria used in this case is the distance (in this case the geometric distance). This process is called distance-based clustering. Another way to do clustering is conceptual clustering. In conceptual clustering, objects are grouped by their compatibility according to the descriptive concept.

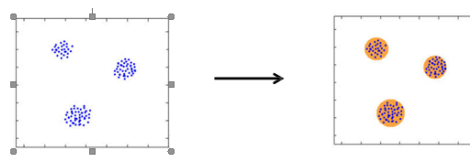


Figure 1. Process clustering

2.2 Data analysis

This study uses data sets obtained from UCI machine learning repository (absenteeism at work data set). The details of the data used can be seen in table.1.

Table 1. Data details used method

No	Data Set	Atribut	Tipe	Class	Amount Of Data
1	Absenteeism At Work Data Set	21	Classification, Clustering	3	741

This study will use the gain ratio as a measurement tool to see the correlation of the attributes in the data set, where the gain ratio will be used as a weighting basis for each attribute. It is expected that giving weight to each attribute can reduce the influence of irrelevant attributes, using the K-means method, so as to increase the accuracy of the process.

The general description of the stages of the method proposed in this study will be explained in figure 2.

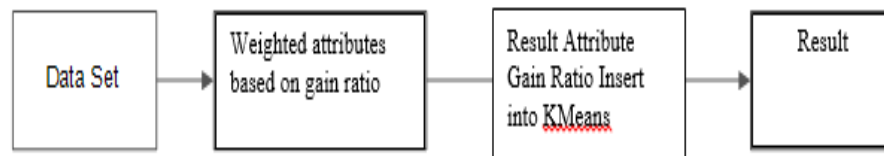


Figure 2. Attribute on gain ratio insert into K-means

2.3 General Architecture

The general architecture in this study can be seen in figure 3, this research is divided into several stages, namely weighting based on gain ratio, this study will use the gain ratio as a parameter to see the correlation between each attribute in the data, then determined by k-cluster wanted formed then calculate the distance of input data, and classify based on the smallest data.

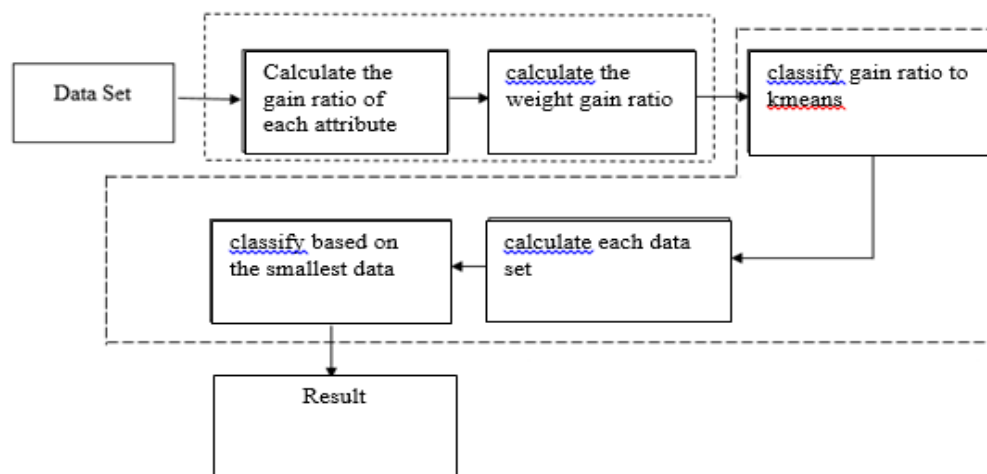


Figure 3. Classify gain ratio to K-means

3. Result and discussion

3.1 Data Set

A data set with a total of 10 records, in which the data has 3 attributes and 3 classes, all of which are used as training data, then entered a test data with the same number of characteristics with the training data where the test data will be classified. The details of the data can be seen in table 2.

Table 2. Data Set

ID	X1	X2	X3	Class	Information
1	33	90	172	Class 1	Data Sample 1
2	50	98	178	Class 3	Data Sample 2
3	38	89	170	Class 3	Data Sample 3
4	39	68	168	Class 3	Data Sample 4
5	28	90	172	Class 1	Data Sample 5
6	36	65	168	Class 2	Data Sample 6
7	28	84	182	Class 1	Data Sample 7
8	47	86	165	Class 3	Data Sample 8
9	36	65	168	Class 2	Data Sample 9
10	29	75	172	Class 1	Data Sample 10

3.2 Process With Kmeans

Kmeans clusters depend entirely on the selection of early centric groups. The data element K is selected as the starting center; Then the distance of all element data is calculated by euclidean distance formula. Data elements less than centroids distance are moved to the appropriate cluster. The process continues until no more changes occur in the k-1 group. Grouping this partition is the most popular and fundamental technique. Here are the steps of the K-means algorithm:

1. Determine the many k-clusters that you want to form.
2. Generate random values for the center of cluster center (centroid) as much as k-clusters.
3. Calculate the distance of each input data on each centroid using the euclidean distance formula until it finds the closest distance of each data with the centroid. Here is the euclidean distance equation.

$$d_{ij} = \sqrt{\sum_{k=1}^p (x_{ik} - x_{jk})^2 + (y_{ik} - y_{jk})^2}$$

With d_{ij} is the distance between the cluster x_{ik} with the cluster center x_{jk} i , and the distance between the cluster

y_{ik} with the cluster center y_{jk} to word to i.

4. Classify each data based on its proximity to the centroid (smallest distance).
5. Updating the centroid value, centroid value obtained from the average cluster in question using the formula:

$$C_k = \frac{1}{n_k} \sum d_i$$

Where :

C_k = cluster value

n_k = amount of data in the cluster

d_i = the sum of the value of the incoming distance into the cluster

6. Doing looping from steps 2 to 5 until the members of each cluster nothing has changed.
7. If step 6 has been fulfilled, then the mean value of the cluster center (μ_j) in the last iteration will be used as a parameter to determine the classification of the data.

3.3 Process With Gain Ratio

Gain Ratio (GR) is a modification of information gain which reduces its bias. gain ratio improves the information gain by taking intrinsic information from each attribute . the steps in determining the gain ratio are as follows :

- a. Calculate the value of entropy on each attribute, by equation

$$\text{Entropy}(S) = \sum_{i=1}^n -p_i * \log_2 p_i$$

Where:

S = Set of cases

n = Number of partitions S

p_i = Proportion of S_i to S

- b. Calculate the value of the information gain in each attribute by the equation.

$$\text{Information Gain}(S, A) = \text{Entropy}(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} \times \text{Entropy}(S_i)$$

Where :

S = Overall dataset

A = Subset attribute

N = Number of attribute partition A

$|S_i|$ = Subset size of dataset owned attribute on A i-partition

$|S|$ = Total number of cases in dataset

- c. Calculate the value of split information for each attribute with the equation below

$$\text{SplitInfo}_A(D) = - \sum_{j=1}^v \frac{|D_j|}{|D|} \times \log_2 \left(\frac{|D_j|}{|D|} \right)$$

Where :

D = Overall dataset

A = Subset attribute

v = Number of partition attributes A

$|D_j|$ = Subset size of dataset owned attribute on A partition

$|D|$ = Total number of cases in dataset

- d. Calculate the gain ratio of each attribute with the equation

$$\text{Gain Ratio}(A) = \frac{\text{Gain}(A)}{\text{SplitInfo}(A)}$$

where the gain ratio is used to calculate the attribute effect on the target of a gain ratio data is the development of the information gain, where the gain ratio eliminates the bias value of each attribute.

4. Conclusion

1. The result of this research is to classify data set then give weighting using gain ratio and insert into method of kmeans so result obtained in this research increase accuracy of gain ratio.
2. To measure the level of accuracy used cross validation which in this study using the gain ratio by entering the method of kmeans.

5. References

- [1] Abdillah, Adli. 2017. Attribute Weighting At K-Nearest Neighbor Using Gain Ratio.Thesis. University of Nort Sumatera.
- [2] Fauziah ,Zarlis, M. 2017. Implementation of K-Means Algorithm At New Vocational High School Students For Clustering Majors. Infotekjar. 1 (2) .100-105.
- [3] Han, J, Kamber, M. 2012. Data Mining Concepts and Techniques. Elsevier: Amsterdam.
- [4] Handoko, K. 2014. Implementation of Data Mining in Improving the Quality of Learning in Higher Education Institutions Using K-Means Clustering Method (Case Study In Tkj Program Academy Community Solok Selatan). Teknosi. Ppt- 31-40.
- [5] Jannah. 2017. Application of Clustering Method With K-Means Algorithm For Students Graduation Prediction Department Of Informatics In University.
- [6] Larose, D. T. 2005. Discovering Knowledge In Data. John Wiley & Sons, Inc: New Jersey.
- [7] Martiano. 2016. Comparison Analysis of Performance of K-Means Method With Modification of K-Means Method. Thesis. University of North Sumatera.
- [8] Mega, Dhuhita, 2015. Clustering Using K-Means Method To Determine Underfive Nutritional Status. Journal of Informatics. 12: 160-174. Journal of Informatics, Vol.15, No.2, December 2015
- [9] Priyadarsini, R.P., Valarmathi, M.L., Sivakumari, S. 2011. Gain Ratio Based Feature Selection Method For Privacy Preservation. ICTACT Journal on Soft Computing 1 (4): 201-205.
- [10] Rizka, Ade. 2018. Particle Swarm Optimization Algorithm Analysis And Rank-Based Selection Method Fitness Assignment On Optimization Of Product Preparation In Warehouse Thesis. University of North Sumatera.