Improvement of K-means using Ruleset

[1]Mary Ambika Babu, [2]Neeba E A
[1,2]Dept. of IT, Rajagiri School of Engg. & Technology, APJ Abdul Kalam Technological University, India

Abstract
The development of clustering analysis technology, there have been many application-based clustering algorithms, such as text clustering. K-means is a method of clustering observations into a specific number of disjoint clusters. The "K" refers to the number of clusters specified. Various distance measures exist to determine which observation is to be appended to which cluster. The algorithm aims at minimizing the measure between the centroid of the cluster and the given observation by iteratively appending an observation to any cluster and terminate when the lowest distance measure is achieved. The two big limitations that the K-Means algorithm has, number of cluster, K, must be determined beforehand and random selection of initial cluster centre, proposed K-Means improved algorithm based on the minimum rule set. This method proposed the concept of the minimum rule covering set. In order to solve the two big limitations of K-Means algorithm effectively. Performance analysis between traditional k-means and improved k-means is evaluated using Chi-Square method, Entropy and F-Measure method.

Keywords
Clustering, Text Clustering, Association Rules, K-means Algorithm, Chi-square Method, F-measure, Entropy

I. Introduction

The objective of cluster analysis is to assign observations to groups (clusters) so that observations within each group are similar to one another with respect to variables or attributes of interest, and the groups themselves stand apart from one another. In other words, the objective is to divide the observations into homogeneous and distinct groups. In contrast to the classification problem where each observation is known to belong to one of a number of groups and the objective is to predict the group to which a new observation belongs, cluster analysis seeks to discover the number and composition of the groups. Cluster analysis is an important research topic in the data mining field. Clustering analysis methods commonly used are: method based on classification, method based on hierarchical, method based on density, method based on grid and so on.

During data analysis many a times we want to group similar looking or behaving data points together. For example, it can be important for a marketing campaign organizer to identify different groups of customers and their characteristics so that he can roll out different marketing campaigns customized to those groups or it can be important for an educational institute to identify the groups of students so that they can launch the teaching plans accordingly. Classification and clustering are two fundamental tasks which are there in data mining for long, Classification is used in supervised learning (Where we have a dependent variable) while clustering is used in un-supervised learning where we don't have any knowledge about dependent variable. Clustering helps to group similar data points together while these groups are significantly different from each other.

There are multiple ways to cluster the data but K-Means algorithm is the most used algorithm. Which tries to improve the inter group similarity while keeping the groups as far as possible from each other. Basically K-Means runs on distance calculations, which again uses "Euclidean Distance" for this purpose. Euclidean distance calculates the distance between two given points using the following formula:

Euclidean Distance = $\sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2}$

Above formula captures the distance in 2-Dimensional space but the same is applicable in multi-dimensional space as well with increase in number of terms getting added. "K" in K-Means represents the number of clusters in which we want our data to divide into. The basic restriction for K-Means algorithm is that your data should be continuous in nature. It won't work if data is categorical in nature.

The K-means clustering algorithm (K-means clustering) [1] is one kind of classical clustering algorithm that is proposed by Mac Queen. The algorithm is simple and the complexity is low. The K-means algorithm main steps are as follows:

Input: Data set D, the K-means clustering counts K
Output: Clustering results C
Step 1: Selecting k points as initial central point
Repeat
Step 2: Assigning each point to the nearest the center, forming a k cluster. Recalculate the center of each cluster.
Until Center point does not change
Step 3: Returning clustering results C

III. Proposed work

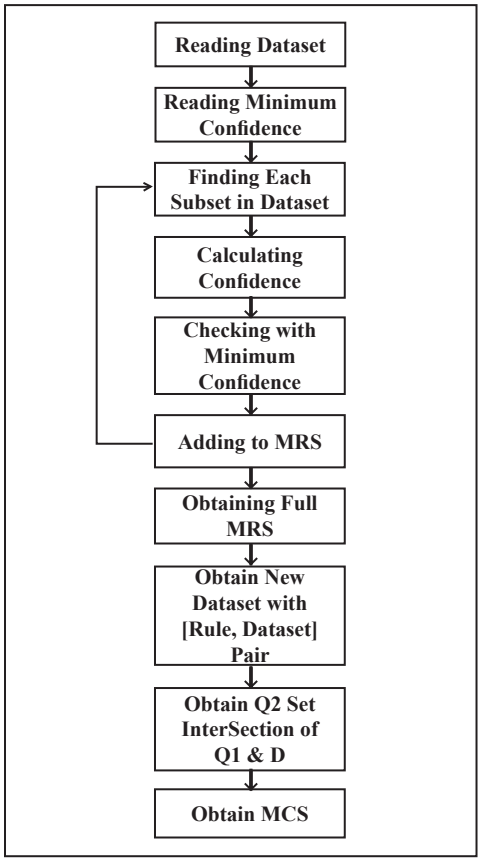

Fig. 1: Architecture

In fig. 1 the architecture read the dataset from frequent mining data repository fetching age, weight etc and find the minimum confidence. Calculate the confidence by support parameter. Then checking the minimum confidence generating the minimum rule set. The minimum rule set obtaining the new dataset with rule and dataset pair. Then dataset intersected by each rule and generate minimum covering set.

To solve the two limitations K-means algorithm has, we propose improved K-means algorithm based on minimum cover set .

Algorithm 1: Seeking the minimum rules set

Input: Frequent closed item sets FCI, Minimum confidence Minconf.

Output: minimum rules set MRS
For each item K in FCI
Finding all subset s in K;
For each subset s in K
If (s->K ¢ MRS)
  Confidence=support(IK)/SUPPORT(s);
If(Confidence>=minconf)
  Add s->K to MRS;

Algorithm 2: Seeking the minimum rules covering set

Input: Minimum rules set MRS

Output: Minimum rules covering set MCS

1. Obtain the corresponding data set D' according to the minimum rules set MRS. Form a new set Q and ensure that each item in the set consists of {rules, data set};
2. Obtain its subset Q1 according to Q;
3. Obtain subset Q2 which is the intersection set of Q1 with the primitive data set D;
4. Get subset Q3 which has least rules number in Q2;
5. If Q3 is not unique, then obtain the overlapped rate smallest subset, namely MCS.

For example, data set D is as follows Table 1. Set R consists of rules on this data set. Table 2 shows minimum rules set P getting from R and object N meeting rules. Then we can find out rules subset of P is r1 = {rules 1, rules 2, rules 3}, r2 = {rules 1, rules 3, rules 4}, and so on. Union of objects in these rules is D, but it is not the rules set covering the least rules number. There is rules subset r = {rules 3, rules 4}, which meets union of objects in r is D, and r contains the least rules number 2.Here has only one such set. Therefore, rules set r is the minimum rules covering set of data set D.

Table 1: Dataset

Object N	Items
1	ABCD
2	ABC
3	ACD
4	BCD

Table 2: Rules Set and Object Meeting Rules

Rules Set R	Objects
1 : BD=>C	1,4
2 : AB=>C	1,2
3 : AC=>D	1,3
4 : B=>C	1,2,4
5 : AB=>CD	1
6 : BD=>AC	1

To solve the limitation (requiring the user in advance to give the desired cluster number k) of the traditional K-means algorithm, we set K as N that is the number of elements in minimum rules covering set r. Meanwhile, according to the corresponding object that each of the rules has in the minimum covering set r and computing the average value of the object in each cluster, we can get K initial cluster. Then obtaining K initial cluster centers and leaving a good foundation for the later K-means clustering process.

IV. Result

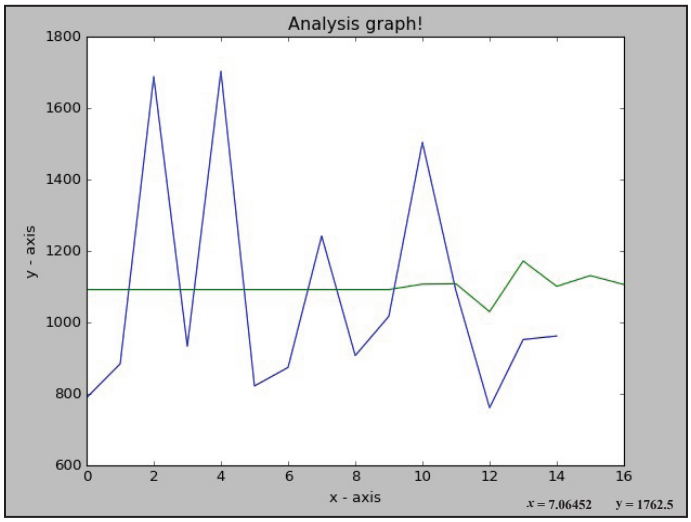

Fig. 2: Time Analysis Graph

Improved K-means algorithm is more effective than the traditional K-means algorithm. It is more time consuming compared to traditional K-means algorithm. Using improved K-means algorithm we get more accurate predictions in the projects related to the society rather than using traditional K-means algorithm.

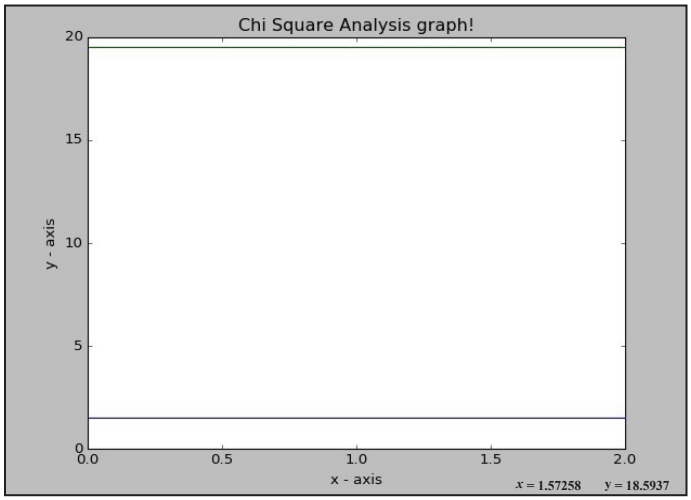

Fig. 2: Chi-Square Analysis Graph

In fig. 3 the performance analysis is evaluated by chi-square method better accuracy in K-means compared traditional K-means.

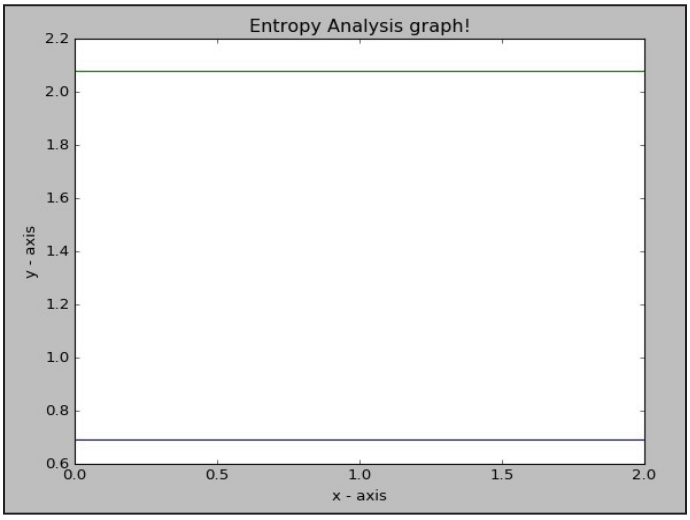

Fig. 4: Entropy Analysis Graph

In fig. 4 the performance analysis is evaluated by entropy method better improvement in K-means compared traditional K-means.

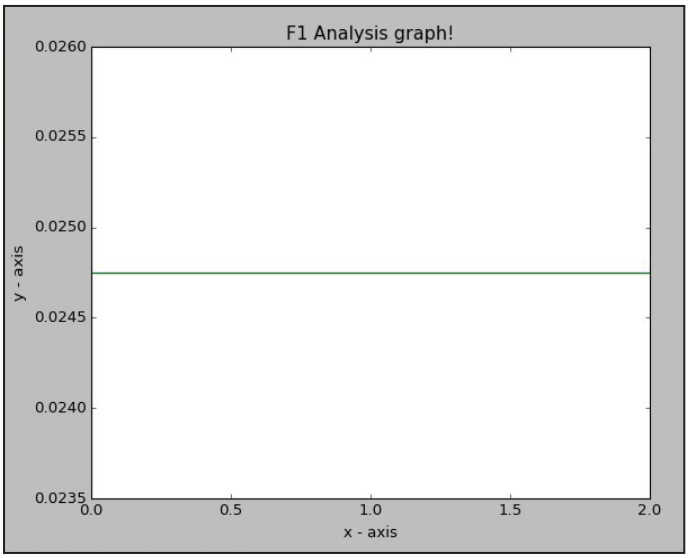

Fig. 5: F1-measure Analysis Graph

In fig. 5 the performance analysis is evaluated by F1-measure method better performance in K-means compared traditional K-means.

IV. Conclusion

We propose improved K-means algorithm based on the minimum rule set. In order to solve limitations of traditional K-means algorithm. Firstly number of cluster, K, must be determined beforehand and random selection of initial cluster centre. We can see that the K-means algorithm based on the minimum rule set is more effective than the traditional K-means algorithm. Performance analysis between traditional k-means and improved k-means is evaluated by chi-square method, Entropy and F-measure method.

References

[1] Gang Liu,Wray Buntine, Weiping Fu, Yudan Du,"An Association Rules Text Mining Algorithm Fusion with K-means Improvement", 4th International Conference on Computer Science and Network Technology (ICCSNT), pp. 781-785, 2015.

[2] Laith Mohammad Abualigah, Ahamad Tajudin Khader, Mohammed Azmi Al-Betar,"Multi-objectives-based text clustering technique using K-mean algorithm", 7th International Conference on Computer Science and Information Technology (CSIT), pp. 1 – 6, 2016.

[3] Guohua Wu, Hairong Lin, Ershuai Fu, Liuyang Wang,"An Improved K-means Algorithm for Document Clustering", International Conference on Computer Science and Mechanical Automation (CSMA), pp. 65 – 69, 2015.

[4] Caiquan Xiong, Zhen Hua, Ke Lv, XuanLi,"An Improved Kmeans Text Clustering Algorithm by Optimizing Initial Cluster Centers", 7th International Conference on Cloud Computing and Big Data, pp. 265 - 268, 2016.

[5] Hong Zhang, Hong Yu, Ying Li, Baofang Hu,"Improved K-means Algorithm on the Clustering Reliability Analysis", International Symposium on Computers & informatics, pp. 90-96, 2015.

[6] Huang Xiuchang,"An Improved K-means Clustering Algorithm, Journal of Networks, pp. 412-417, 2014.

[7] Liu Guoli, Wang Tingting, Yu, Li Yanping, Gao Jinqiao, "The improved research on k-means clustering algorithm in initial values", Proceedings International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC), pp. 2124 – 2127, 2013.

[8] Ghousia Usman, Usman Ahmad, Mudassar Ahmad,"Improved K-Means Clustering Algorithm by Getting Initial Cenroids", World Applied Sciences Journal, pp. 809-814, 2013.

[9] Xiaohui Cui, Thomas E. Potok, Paul Palathingal,"Document Clustering using Particle Swarm Optimization", Proceedings, IEEE Swarm Intelligence Symposium, pp. 185-191, 2005.

[10] Unnati R. Raval, Chaita Jani,"Implementing and Improvisation of K-mean Clustering", IJCSMC, pp. 72–76, 2015.

Mary Ambika Babu received his B.Tech. degree in Computer Science Engineering from Albertian institute of science and Engineering, kerala, India, in 2016, the M.Tech degree in Network Engineering,from Rajagiri school of engineering and technology, Kerala, India in 2018.

E. A. Neeba working as an assistant professor in the Department of Information Technology, Rajagiri School of Engineering & Technology, APJ Abdul kalam technological university, Kochi, Kerala. Also research scholar at Department of Computer Science and Engineering, School of Computing, Veltech Dr. RR & Dr. SR University, Avadi, Chennai-62, Tamil Nadu, India. Her research interest includes Data Mining and Big Data.