ORIENTAL JOURNAL OF
COMPUTER SCIENCE & TECHNOLOGY
An International Open Free Access, Peer Reviewed Research Journal
Published By: **Oriental Scientific Publishing Co., India.**
www.computerscijournal.org

ISSN: 0974-6471
June 2017,
Vol. 10, No. (2):
Pgs. 474-479

Enhanced K-Means Clustering Algorithm Using Collaborative Filtering Approach

ANKUSH SAKLECHA and JAGDISH RAIKWAL

Institute of Engineering and Technology Devi Ahilya University,Indore, M.P, india.

http://dx.doi.org/10.13005/ojcst/10.02.31

(Received: May 10, 2017; Accepted: May 30, 2017)

ABSTRACT

Clustering is well-known unsupervised learning method. In clustering a set of essentials is separated into uniform groups.K-means is one of the most popular partition based clustering algorithms in the area of research. But in the original K-means the quality of the resulting clusters mostly depends on the selection of initial centroids, so number of iterations is increase and take more time because of that it is computationally expensive. There are so many methods have been proposed for improving accuracy, performance and efficiency of the k-means clustering algorithm. This paper proposed enhanced K-Means Clustering approach in addition to Collaborative filtering approach to recommend quality content to its users. This research would help those users who have to scroll through pages of results to find important content.

Keywords:Data Mining; Clustering; K-means Clustering; Collaborative filtering Centroids.

INTRODUCTION

Data mining is a technique that is used to extract and mine the important information from mass of data. Data mining is used as information Discovery in Database, information engineering. As we know that clustering is an unsupervised learning method well-known in the area of data mining. For creating subgroups that are handier than individual datum, clustering is used as a data reduction tool. Cluster analysis is one of the best data analysis methods. It mainly used for many practical applications in research area. The other term classification is a supervised learning method due to existence of predefined classes. The clustering method of high quality is used to achieve low inter-cluster similarity and high intra cluster similarity [3].Fig1 illustrate the clustering.

We can use different clustering algorithms for cluster the data. The well-known k-means algorithm is a clustering algorithm. It produces clusters for many applications. But for the larger data sets the computational complexity of the existing k-means algorithm is very high. Depending on the random choice of initial centroid this algorithm results in different types of clusters. For improving the efficiency of the k-means clustering algorithm many researcher publish number of papers and a good number of attempts were made

by researchers. Recommended systems are well known information filtering systems that suggest items to users. There are two main approaches for information filtering the first one is Collaborative filtering and the second one is content-based filtering.

1. Collaborative filtering: It choose items based on the similarities among the preferences of different users.
2. Content-based filtering: It choose items based on the similarities between the content description of an item and the user's preferences. This paper deals with the use of clustering techniques to develop a collaborative based recommendation system. It addresses the limitations of existing K-means used to implement recommendation systems, evaluation of experimental results, and conclusion.

K-Mean Clustering Algorithm

K-means is easiest and popular unsupervised learning algorithm which solves the popular clustering problem. It is famous partitioning technique in which objects are categorized as fitting in one of K groups. The principle thought is to characterize k centroids, one for every cluster. In each and every cluster there may appear centroid or a cluster agent. In this case we consider data based on real world and all objects inside a cluster which gives a suitable agent with the help of mean of the attribute vectors; we can utilize distinctive sorts of centroid in different cases. Let us take an example: Suppose we have a list of such keywords that appear in few minimum number of documents inside a cluster denotes cluster of documents. If the number of clusters are larger then the centroids then it will be again clustered which gives hierarchy within a dataset. To execute clustering of the data sets or samples, K-means is used. As we know that, he meaning of clustering is the separation of a dataset or samples into a dissimilar group such that related items fit in to the similar groups. In K-means

Fig. 1: Clustering

algorithm an iterative method is use to cluster the database. The numbers of preferred clusters and the first means used as input and in this case output is final means. The first and last means are the means of clusters, if an algorithm is required to create a cluster then K-means will be the initial and final tool. After completing this algorithm, each object of a dataset is a member of a cluster. The cluster is found out by looking throughout the means to find out the cluster with nearer mean to the object. The examined object belongs is a cluster with shortest distance mean. In the K-means clustering algorithm, it attempt to cluster the data samples in dataset into preferred number of clusters so to done this task K-mean algorithm perform few iterations until it meets few converges criteria. After every iteration the newly calculated means are somewhat nearer to the final means and they are updated at the end and then the algorithm converges and stops performing iterations. In other words k-means clustering is method of vector quantization. It is well known algorithm for analysis of cluster in data mining. The aim of K-means clustering algorithm is to divide n observations into k clusters and every observation belongs to the cluster with the closest mean, serving as a model of the cluster. Let us take into consideration the set of observations (x_1, x_2, x_3,…, x_n) and here each and every observation is a N-dimensional real vector. Now we apply k-means clustering to partition the n observations into k (k d" n) sets $S = \{S_1, S_2, S_3 \ldots, S_K\}$ in order to minimize the inter-cluster sum of squares (ICSS).The following is objective function:

K-Mean Clustering Alorithms's Step

K-Means clustering is used to classify data samples in K cluster and the value of K is determined as user-defined. In this algorithm first, the centroid of each cluster is selected for clustering and then according to the chosen centroid, the least distance data points from the specified cluster, is allocated to the particular cluster. For evaluating the distance of data point from the centroid Euclidean Distance is used. The Complexity of the algorithm is O(ABCD), where C = number of objects, B =number of clusters, D = dimension of each object, and A = number of iterations. Note: A, B, D << C.

Algorithm1: K- means Algorithm

1. Initialization: Firstly, we initialize the number

of clusters and the centroid that we have defined for every cluster.

2. Classification: In the second step we calculate the distance between the cluster centers and every data point. Now we give data point to the cluster center, which keeps the minimum distance from the cluster centers among all the cluster centers.
3. Recalculation: Now we use the given expression to recalculate the new cluster center:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

Where, 'c_i' = the number of data points in *ith*

cluster.

4. The following are Convergence Condition:

4.1 When given or defined number of iterations are done, Stop the algorithm.

4.2 When data points between the clusters are not exchanged then stop the algorithm.

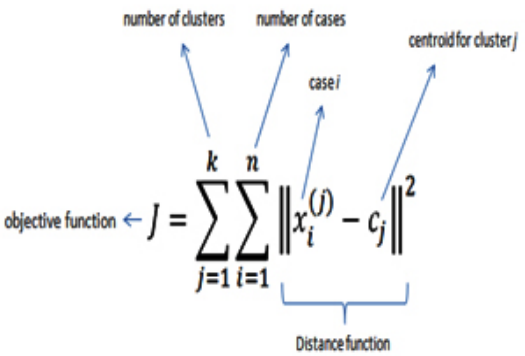
4.3 When we get a threshold value the algorithm is stopped.

5. If algorithm is not stopped then from step 2 the complete process is repeated again until and unless the given conditions are not satisfied[7].

Fig2 shows stepwise execution of K-Means Process.

Advantages of k-means

- Simple, robust and easy to understand.
- K-Means is computationally efficient and faster than hierarchical clustering provided large number of variables exists and k is kept small.
- More efficient algorithm than k-mediod.
- It gives tighter clusters than other clustering method.



Problem in the Original K-Means

K-Means cluster analysis is a useful clustering algorithm is also known as best machine learning method. In addition, it can provide great descriptive information about population sub-groups that share same patterns of response. However, k-means cluster analysis in general has some disadvantages.

First, we need to specify the number of clusters. But we don't know the true number of clusters. And figuring out the correct number clusters that represent the true number of clusters in the population is pretty subjective. On top of that, your results can change based on the location of observation, which are randomly chosen as primary centroids. K-means cluster analysis is not recommended if you have too many explicit variables. If you have a lot of clear variables, then you have to use a different clustering algorithm that can handle them better. K-means clustering that it believes that the underlying groups in the population are spherical, different, and are of approximately equal size. Consequently, there is a tendency to identify clusters with these characteristics; it won't work as well if clusters are elongated or not equal in size.

Related Work

Wang Shunye et al[3] inspired by the problem of random selection of initial centric and similarity measures. The researcher presented enhanced K-means clustering algorithm which is based on dissimilarity. This improved k-means clustering algorithm basically consists of 3 steps. (i) construction of the dissimilarity matrix.(ii)Huffman tree is created according to dissimilarity matrix using Huffman code algorithm. It gives the initial centroid as a output.(iii) The k-means algorithm is applies to initial centroids to get k cluster as output. As compared to traditional k-means the proposed algorithm gives better accuracy and results.

Micheal Pazzani[12] discusses about recommending data sources for news articles or

web sites after learning the taste of the user by learning his profile. Various types of information have been mentioned in this paper which can be considered to learn the user's profile. Depending on the ratings given by a user for different sites, the ratings that other users have given to those sites and suggest demographic information about users. This paper explains how the above information can be added to provide recommendations for users.

Pallavi Purohit and Ritesh Joshi et al[2] introduce an best approach for K-means clustering algorithm due to its certain limitations. Due to the random selection of initial centroids, it provides poor performance and accuracy. The researcher's new research provides an algorithm that deals with this problem and improves efficiency and performance and cluster quality of the old algorithm. The proposed algorithm selects the initial centroid in a regular manner instead of randomly selecting. The new algorithm gives precise results and also minimizes the mean square distance. The new algorithm works better for dense dataset rather than sparse.

Navjot Kaur, Navneet Kaur[7]enhanced the old k-means by introducing new Ranking approach. The author introduces the ranking method to overcome the lack of execution time taken by old K-means. Ranking method is a way to find out the occurrence of similar data and improve search effectiveness. The tool is used to implement a better algorithm using Visual Studio and C #. Benefits of K-means have also been analyzed in this paper. Authors get K-means as fast, strong and easy-to-understand algorithm. Researcher also discussed that clusters are non-hierarchical and non overlapping in nature. In the process used in the algorithm, the students marks takes as data sets and then the initial centeroid is selected after then we calculate Euclidean distance for each data object and then for each data set, the threshold the value is set. After this, we apply the ranking method through which the cluster is formed on the basis of the minimum distance between the data point and the centroid. Author says that the scope of this paper in future is use of Query Redirection to create cluster.

Robert M Bell and Yehuda Koren[11] Koren "Scalable Collaborative Filtering with Jointly Derived Neighborhood Interpolation Weights". Neighbors are calculated based on the previous user-item relationship, which makes prediction easier. The weight of all the neighbors is calculated differently and for many interactions to provide a solution to the problem, they are interconnected between each other simultaneously. The proposed method has been asked to provide recommendation in 0.2 milliseconds. Training takes a lot of time in large-scale applications. The proposed method was tested on Netflix data, which contained 2.8 million queries that were processed in 10 minutes.

Jia Zhou and Tiejan Luo , has published a paper on Collaborative Filtering applications. This paper gives information about collaborative filtering techniques that were currently used in that generation. It has been said that collaborative filtering techniques used in that generation can be divided into an estimated-based method and model-based approach. Paper discusses the limitations of

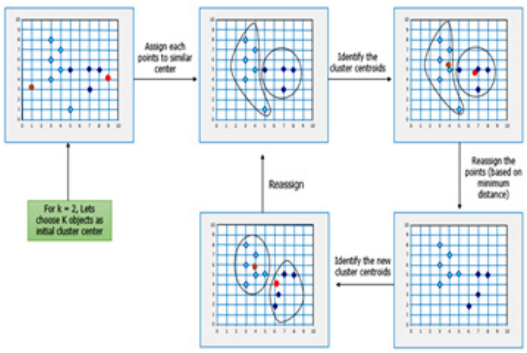


Fig. 2: K-Means Process

Cluster Number	Old K-Means	Enhanced K-Means
Cluster 1	3 to 3 years	3 to 9 years
Cluster 2	4 to 4 years	10 to 15 years
Cluster 3	5 to 5 years	16 to 20 years
Cluster 4	6 to 6 years	21 to 25 years
Cluster 5	7 to 7 years	26 to 30 years
Cluster 6	8 to 8 years	31 to 35 years
Cluster 7	9 to 9 years	36 to 40 years
Cluster 8	10 to 10 years	41 to 45 years
Cluster 9	11 to 11 years	46 to 49 years
Cluster 10	12 to 13 years	50 to 53 years
Cluster 11	14 to 17 years	54 to 58 years
Cluster 12	18 to 23 years	59 to 63 years
Cluster 13	24 to 31 years	64 to 68 years
Cluster 14	32 to 40 years	69 to 73 years
Cluster 15	41 to 50 years	74 to 78 years
Cluster 16	51 to 62 years	79 to 82 years
Cluster 17	63 to 77 years	83 to 88 years
Cluster 18	78 to 95 years	89 to 95 years

Fig. 3: Comparison on the basis of cluster creation

collaborative filtering techniques in that generation and suggests some improvements to increase system's recommended capabilities.

Proposed Work

The original K-means algorithm has been modified in the enhanced K-means clustering method and gives two phases of algorithm for determining the initial centroid and provides data points to nearby centroid to improve accuracy and efficiency of algorithm. The modified method is represent as Algorithm 2 is divided into two part. In part1 Using Equation 3 we are calculating middle point for each subset which will be initial centroids. For Calculating mean value for the given Dataset the Equation 1 is used. Derive K number of equal subsets from data set having initial no. of elements calculated through Floor Function given in Equation 4.The following Important Equations are used in algorithm for a given dataset

Equation 1: The mean value is
Dmean= (x1+x2+x3+.....+xn)/
Where n // total number of data points in data set d
Equation 2: The distance between to data points could be calculate
Dist = |x1-x2|
Equation3: The Centroid of dataset
Ci = Nearest (POINT) to Dmean
Equation 4: Floor = (No. of data points / Value of K)
Algorithm 2.Improved K-means Method
Input: D={x1,x2,x3,x4,.....xn}// set of n numbers of data points, K // The number of desire Clusters
Output: A set of k clusters
Steps:
Part1: Determine initial centroids
Step1.1: Using Equation 1 find the mean value for the given Dataset.
Step1.2: Find the distance for each data point from mean value.
Step1.3: Sort data points according to their distance from the mean value calculated in step1.2.
Step1.4: Derive K number of equal subsets from data set having initial no. of elements calculated through Equation 4.
Step1.5: Using Equation 3 calculates middle point for each subset which will be initial centroids.
Step1.6: Compute distance from each data point to initial centroids.
Part2: Assigning data points to nearest centroids
Step2.1: Calculate Distance from each data point to centroids and assign data points to its nearest centroid to form clusters and stored values for each data.
Step2.2: Calculate new centroids for these clusters.

Repeat

Step2.3: Calculate distance from all centroids to each data point for all data points.

Step2.3.1 **IF** The Distance stored previously is equal to or less then Distance stored in Step3.1 **Then**

Those Data points don't needs to move to other clusters.

Else

From the distance calculated assign data point to its nearest centroid by comparing distance from different centroids.

Step2.5: Calculate centroids for these new clusters again **UNTIL** convergence criterion met.

Number of Iterations

Old K-Means	Enhanced K-Means
7 iterations	3 iterations

Fig. 4: Comparison on the basis of number of iteration

Category	Cluster1: (3 to 9 years)	Cluster2: (10 to 15 years)	Cluster3: (16 to 20 years)	Cluster4: (21 to 25 years)	Cluster5: (26 to 30 years)	Cluster6: (31 to 35 years)	Cluster7: (36 to 40 years)	Cluster8: (41 to 45 years)	Cluster9: (46 to 49 years)	Cluster10: (50 to 53 years)
Books for Children	49.5%	30.8%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Literature and Fiction	0.0%	0.0%	4.8%	25.0%	13.3%	9.1%	0.0%	0.0%	0.0%	18.5%
Biography	3.0%	0.0%	0.0%	12.5%	6.7%	0.0%	0.0%	0.0%	0.0%	22.2%
Crime and Thriller	6.1%	0.0%	0.0%	0.0%	6.7%	9.1%	0.0%	0.0%	0.0%	0.0%
Spiritual	10.1%	0.0%	0.0%	0.0%	6.7%	0.0%	0.0%	0.0%	0.0%	3.7%
Education	2.0%	15.4%	4.8%	25.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Travel	7.1%	0.0%	0.0%	12.5%	0.0%	9.1%	0.0%	0.0%	0.0%	7.4%
Cookbooks	0.0%	0.0%	0.0%	0.0%	6.7%	9.1%	6.2%	26.3%	0.0%	0.0%
Teens and Young Adults	7.1%	46.2%	76.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Health and Fitness	0.0%	0.0%	4.8%	0.0%	6.7%	0.0%	0.0%	36.8%	15.0%	22.2%
Romance	13.1%	0.0%	0.0%	12.5%	6.7%	0.0%	43.8%	36.8%	20.0%	0.0%
Inspirational	1.0%	0.0%	0.0%	0.0%	13.3%	9.1%	0.0%	0.0%	15.0%	3.7%
Computer	0.0%	7.7%	4.8%	12.5%	6.7%	0.0%	0.0%	0.0%	0.0%	0.0%
History	0.0%	0.0%	0.0%	0.0%	6.7%	9.1%	0.0%	0.0%	0.0%	22.2%
Sci-fi and Fantasy	0.0%	0.0%	0.0%	0.0%	6.7%	0.0%	0.0%	0.0%	0.0%	0.0%
Business	0.0%	0.0%	0.0%	0.0%	0.0%	18.2%	50.0%	0.0%	15.0%	0.0%
Psychology	0.0%	0.0%	4.8%	0.0%	0.0%	9.1%	0.0%	0.0%	0.0%	0.0%
Language and Culture	0.0%	0.0%	0.0%	0.0%	0.0%	9.1%	0.0%	0.0%	20.0%	0.0%
Classics	1.0%	0.0%	0.0%	0.0%	6.7%	9.1%	0.0%	0.0%	0.0%	0.0%

Fig. 5: Cluster Analysis for Enhanced K-means algorithm

Result Analysis

We are applying the old and enhanced k-means algorithm for book recommendation system and the cluster created by enhanced k means algorithm is much better and accurate also the result shows that the total number of iterations are reduced hence complexity of algorithm is also reduced. Fig3. Illustrate the cluster created using old and new k means algorithm and also shows that number of iteration is less in enhanced k means algorithm. Fig4 illustrate the total number of iteration to create clusters using enhanced k means is less as compare to old k-means. Fig5 Illustrate Tabular form of Cluster created by the proposed algorithm and shows recommend books according to that.

CONCLUSION

In this research we propose a enhanced clustering algorithm which increase the efficiency of the algorithm because the number of iteration in enhanced k means is less than the old k means. Also result shows that cluster formation is better than the old k-means algorithm. This approach conquers the known defects of the k-mean algorithm, that is, dependency on the initial choice of cluster centroids. In addition, our approach guarantees high precision clustering. We measured the accuracy of our approach using different parameters like Recall, Accuracy and Precision. The proposed work represents an age-based clustering method that improves performance and accuracy of the K-means clustering algorithm in the area of users' recommendation of products like books. In this we have also analyzed the K- means clustering algorithm by implementing two data sets on both the existing K-means algorithm and the newly modified K-algorithm and using the graph also compared the performance of both methods. Experimental results confirmed that the advanced K-mean algorithm provides better results than existing algorithms. This paper has concluded that the increasing efficiency of the K-means algorithm and the users get better results compared to the purchase and execution time of their thoughts and books. The K-means algorithm is widely used for clustering of data on a large scale. But standard algorithms do not always guarantee good results because the accuracy and efficiency in the distribution environment decreases.

REFERENCES

1. Nikita Jain, Vishal Srivastava "DATA MINING TECHNIQUES: SURVEY PAPER"IJRET eISSN: 2319-1163 | pISSN: 2321-7308.
2. Pahlavi Purohit "A new Efficient Approach towards k-means Clustering Algorithm", International journal of Computer Applications,Vol 65-no 11,march 2013
3. Wang Shunye "An Improved K-means Clustering Algorithm Based on Dissimilarity" 2013 International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC)Dec 20-22,2013, Shenyang, China IEEE.
4. Juntao Wang & Xiaolong Su "An improved K -Means clustering algorithm" 2011 IEEE
5. Md. Sohrab Mahmud, Md. Mostafizer Rahman, and Md.Nasim Akhtar "Improvement of K-means Clustering algorithm with better initial centroids based on weighted average" 2012 7th International Conference on Electrical and Computer Engineering 20-22 December, 2012, Dhaka, Bangladesh, 2012 IEEE.
6. Shuhua Ren &Alin Fan "K-means Clustering Algorithm Based on Coefficient of Variation" 2011 4th International Congress on Image and Signal Processing2011 IEEE.
7. Navjot Kaur, Jaspreet Kaur Sahiwal, Navneet Kaur "EFFICIENT K-MEANSCLUSTERING ALGORITHM USING RANKING METHOD IN DATA MINING" ISSN: 2278–1323 International Journal of Advanced Research in Computer Engineering & Technology Volume 1, Issue 3, May2012.
8. SongJie Gong and Zhejiang "Joining User Clustering and Item Based Collaborative Filtering in Personalized
9. Friedrich Leisch1 and Bettina Gr un2, "Extending Standard Cluster Algorithms to Allow for Group Constraints", Compstat 2006, Proceeding in Computational Statistics, hysica verlag, Heidelberg, Germany,2006.