VOL. 11, NO. 9, MAY 2016 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

EFFECTIVE CLUSTERS CULLED OUT THROUGH ALGORITHMIC IMPLEMENTATIONS

Manalina and K. Mohana Prasad
Department of Computer Science, Sathyabama University, Chennai, India
manali625@gmail.com

ABSTRACT

Data mining is a technology that collects and search a bulk of data from database to discover relationship among data. It is an application that view data from different angles and group it into information that is useful in many perspectives. There are different types of clustering methods that used to grouping the generated data sets such as K-means etc. K-means algorithm is a centroid based technique and has input parameter as k. This technique has two restrictions such as k-means value selection and centroid selection i.e. the size of cluster is assigned by manually and the centroid value is selected by randomly. These two parameter impacts on the clustering performance massively. Another metric such as distance metric also have impact on choosing the cluster also presented. This paper presents powerful K-means (PKM). To show the performance of the proposed algorithm various set of dataset have been taken. That has been applied on traditional K-means and proposed algorithm. The experimental result shows proposed algorithm gives better result when compared to traditional k-means.

Keywords: data mining, clustering, K-Means, Euclidean distance, IKM.

1. INTRODUCTION

Data Mining is a promising field which discovers knowledge from large data set. It mines the earlier unfamiliar, actionable and valid information from huge databank. These data are recycled to make essential business decision. Data mining tools are very useful in predicting future trends and their performances. As it is recognized as knowledge discovery it consists of sequence of steps. First step contains cleaning of data to remove inconsistent and noisy data. Then data from different sources are joined and data which are appropriate for analysis are retrieved. Analysed data are transformed into form which are appropriate for data mining by applying aggregation operation. Data mining apply intelligent methods to extract useful data from huge database. After mining of the data, pattern which are Useful from the aspect of knowledge are evaluated. Finally mined knowledge is present to the user.

There are various steps are performed for mining data. Such as identify the objective to decide what is to be achieved after analysis process. After identifying, select all the data to accomplish the aim. Once the requisite data has been brought together, the usable formats of the attribute have to be decided and estimation of the building of all the data to decide the suitable tools are performed. The selection of the suitable data mining tools are data structure and business objective. The selection of suitable tools and business objective are combined with the data audit to decide the presentation of the solution. Then model are constructed and discussion of consequence of the study with the domain professional or business user are performed to authenticate the outcome. After the authentication the final outcome or report to the client or business unit are delivered and also share the outcome of findings to concern end-user.

1.1. Stage of clustering

Data clustering is a significant tool in various applications such as data mining [1], image segmentation [2] [3] [4] etc.

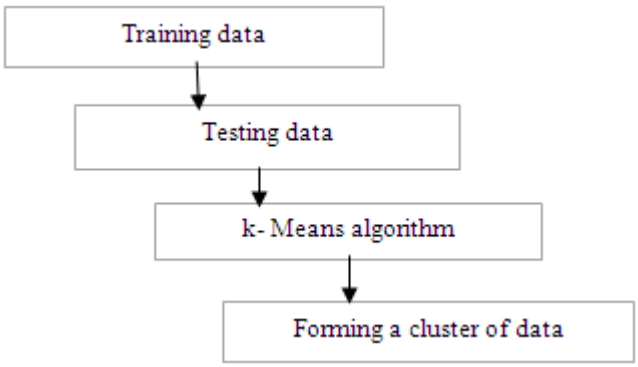

Figure-1. Various stage of clustering.

It is a process of building a collection of abstract or physical objects into a class of related objects. Data in one cluster can be defined as one group. The reason for using clustering in data mining is to provide scalability, ability to deal with any kind of data and also ability to deal with noisy data.

2. CLUSTERING METHOD

2.1 Partitioning method

Suppose a database of 'n' objects and the method make 'k' partition of the data. Partition characterize a cluster when k<=n. It means that each group must contains at least single object. Each object must be related to one group.

2.1.1 K-Means

K means clustering was first given by MacQueen in 1967 but the idea had recommended by Hugo steinhausin1957. The standard algorithm was first givenby staurt Lloyd [5], though it was not published outside of Bell Labs until 1982. In 1965 E.W. Forgy published the same method that's why it is also known as Lloyd–Forgy. It was proposed to resolve many clustering problem. It is based on partitioning clustering technique which goal to partition a group of n object in k cluster. The partition is done in a way that the intracluster similarity should high and the intercluster similarity become low. The similarity of cluster is determined by the mean value of the set of objects in a cluster which can be known as cluster's centroid.

It is an unsupervised and iterative method of clustering. In this method each cluster is characterized by taking mean value of objects in a cluster. Since it is iterative in nature it follows an iterative technique to cluster a dataset or database. After the completion of k means each object of the dataset belongs to exactly one cluster. No two different clusters should contain the same data points. To determine which object will belong to which cluster is done on the basis of mean value. The object having nearest mean to the center of the cluster is included in that cluster.

Description

Given a set of objects Yi where i=1, 2, 3….n. K-mean algorithm aims to partition n-object into k (where k<=n) cluster c= {c1, c2,……..ck}. In order to decrease the within- cluster sum of squares (WCSS).The algorithm works as follows:
In k-means algorithm, it decide k of the objects randomly, each of them are initially a cluster mean or center.

- In the dataset, each data point is allocated to the closet cluster according to the Euclidean distance between every cluster center and every data point.
- Then the cluster center are recomputed.
- Repeat step 2 and 3 iteratively and stopped when there is no change in the clusters centers, identified as algorithm converges.

The Figure-2 shows the flow of K-means algorithm. The major advantage of using this algorithm is its simplicity and its speed which permit it to run on huge datasets. One drawback is that it does not give the same outcome in each run. One of the restrictions is that if the datasets are huge then there are chances that it will not behave properly. In the algorithm another basic problem encounter when the whole group of data point is massive, the convergence maybe lengthy to local minimum and ultimately the result after some repeats may not be the best response. To overcome these limitations, various cluster algorithms have been proposed recently.

3. DIFFERENT DISTANCE FUNCTION

There are various distance function. These distance functions have been used to measure the distance between two data points. Distance functions shows significant role in K -means clustering algorithm. Some of the distance functions are discussed as follows:

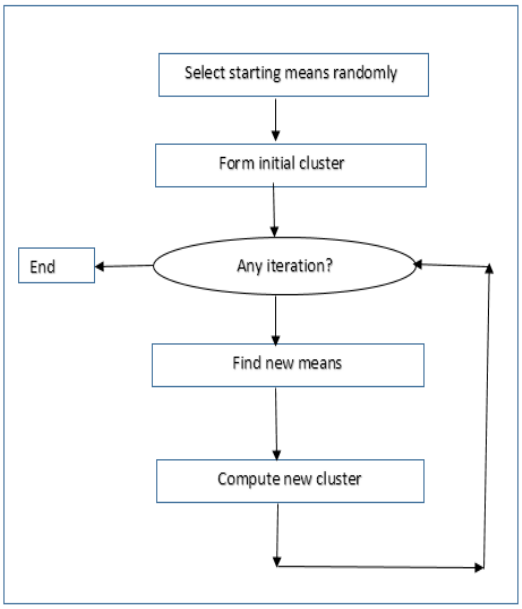

Figure-2. Flow chart: K-means algorithm.

3.1 Euclidean distance

Euclidean distance is a conventional distance connecting two objects. This distance function [13] is widely in used. This distance function is derived from Pythagorean formula. A formula has been derived to find Euclidean distance connecting two points (p, q) [12].
This formula is given by:

$$D(r, s) = \sqrt{(r1 - s1)^2 + (r2 - s2)^2 + \cdots + (rn - sn)^2} \quad (1)$$

Where D is the Euclidean distance. r and s are two data object. Where r = (r1, r2, r3…….rn). s= (s1, s2, s3…..sn).

This is equal to the Pythagorean formula [14]. One of the limitations of Euclidean distance is if the input element has a large scope, then it can override the other elements.

3.2 Manhattan distance

Like Euclidean distance Manhattan distance is one of the distance measures to find distance between different points. The distance is measured at right angles between two points along axes. It is summation of the absolute values of the difference of all points coordinates. In a plane z1 at (p1, q1) and z2 at (p2, q2), is | p1 – p2 | + | q1-q2 |. The formula is given as:

$$D = \sum_{k=1}^{n} | pi - qi | \quad (2)$$

Where D is Manhattan distance. p and q are two points.

These distance metric have great impact on the outcome of K-means algorithm. However in K-means we

use Euclidean distance. But when applying Manhattan distance instead of Euclidean distance on dataset the number of iterations and convergence time decrease in most of the cases. However in some cases both have same responsive time. Moreover Manhattan distance metric needs less calculation than Euclidean distance function also increases the computational time complexity of K-means.

4. RELATED WORK TO VARIOUS CLUSTERING ALGORITHM

An efficient representation of a mixed prototype for categorical attribute and also the significance of various attributes to the clustering process are presented [7]. For this, notion of a distribution centroid to present the cluster center in the hard scenario for categorical attribute [18] [19] is presented. It records how many times a particular value occurred as a categorical attribute [20] in the cluster. It shows the distribution properties of categorical attribute in cluster while presenting the center cluster.

Huang's strategy is used for evaluating significance and in the proposed algorithm the above two ideas for clustering mixed data [17] is integrated. In the experiment result, the algorithm is used to cluster various datasets that occurred in real world. The datasets are Iris which is a pure numerical data, Soybean which is pure categorical data, Heart disease which is a mixed data [6] and the Credit approval which is a mixed data are used and the result gives higher clustering accuracy comparing to k-prototype [16], SBAC, KL-FCM-GM [17].

Fuzzy C-Means is a clustering algorithm in which portion of data can be assigned to more than one clusters and works by assigning belongings to each data point parallel to each cluster [9] center on the basis of distance between the cluster and the data object. It gives better result but has a limitation with trapping in local minimum and it is also sensitive when it comes to the of initial value selection and it has high convergence rate when dataset is large. To avoid this problem a hybrid of FCM and modified stem cell algorithm [21] has been proposed. To depict the efficiency of the proposed algorithm an experiment on well-known data set was performed. For this, a comparison has been made among the proposed algorithm with various algorithm like K-means, Ant Colony Optimization algorithm [9] [22], FCM and Artificial Bee Colony Algorithm [21]. Experiment result shows that the proposed algorithm has high convergence rate, low error rate, low cost and less number of iteration comparing to others algorithm. From the result it is clear that the proposed algorithm is better than above mentioned algorithm.

To advance the traditional FCM by implementing a new approach for determining the initial centers of cluster that is also one of the limitation of the classical FCM has been proposed [9]. This paper also presents the feature of WEKA an open source data mining platform [9]. It implements some popular data mining algorithm. Nonetheless the FCM algorithm [22] is not added into WEKA [23]. The proposed algorithm successfully added the FCM into WEKA to enhance the system functions of the open source platform. That enables the users to directly call FCM for fuzzy clustering analysis. A comparison is made on the basis of the strength and weakness of various algorithm depends on the simple metrics in the clustering field: number iterations, squared errors and some prior knowledge of metrology in addition. The experimental result shows that the proposed algorithm has smaller squared errors than the FCM algorithm while preserving the fast speed of convergence. Nonetheless this paper just improves in selecting the center of a cluster but does not improve the other limitation of the FCM algorithm.

A new hybrid procedure called Hybrid K-MICA [8] has been proposed for data clustering of m items into K cluster. A distinctive arrangement of two most important clustering algorithms first is K-means [11] and the second is Imperialist competitive algorithm [24]. The main contribution of this paper is to presenting a new ICA algorithm after modifying the existing one. Combine new ICA algorithm and CLS algorithm [4] to analyse the cluster. The imperialist competitive algorithm is a novel algorithm, which is capable of handling various optimization problems [21]. However it is still in its early stage and rigorous learning is essential for expanding its performance.

The proposed algorithm has been compared with various existing algorithm in term of standard deviation and also finding solution for best and worst. The convergence rate of proposed system is better than existing algorithm. The proposed work is relevant if the number of cluster can be find in priori.

A distinctive arrangement of two most important clustering algorithms Particle Swarm Optimization (PSO) [25] and K-means [11] has been proposed to attain improved clustering [28] outcome. In this paper first it discuss about K-means and PSO. K-means is the one of most effective algorithm in terms of execution time. But has the limitation of selecting the initial cluster center. Bad selection of initial cluster [27] may lead to a poor convergence rate or no convergence rate at all. To avoid this limitation meta-optimization has been proposed in which two algorithm combined to achieve the goal. In this proposed paper K-means is integrated with PSO [26]. The sequential combination of these two algorithms accomplishes firm data clustering and it avoids from being struck into a local minimum and gives ideal solution [10]. The experimental result shows that the sequential hybridization of both algorithms can result in best clustering outcome when compared with individual K-means and individual PSO algorithm.

5. PROPOSED ALGORITHM

The proposed system uses Manhattan distance to decrease the time to create a cluster in traditional k-mean. And the system uses Pearson coefficient to establish relationship between two data by using two metric we have introduced a novel algorithm improved k-means. That gives better result when compared with traditional k-means.

Improved k-means

Input

Number of preferred clusters, k, and datasets d= {d1, d2, d3…., dn} having n objects.

Output

Clusters of dataset with decrease amount of time in execution when compared to traditional k-means.

Steps

Use centroid selection formula for selecting the initial centers of d datasets.

- Find distance for each object to the clusters center. Use Manhattan distance formula.
- Repeat step 1 and 2 until there is a change in cluster center.
- Use Pearson correlation coefficient formula to find relationship between two data.
- End process.

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} \quad (3)$$

Equation (3) is a Pearson correlation coefficient formula. Where r is the Pearson coefficient t and x and y are two data object to which we have to find relation

6. EXPERIMENTAL RESULTS

Data sets used in the experiment are taken from the UCI repository.

The detailed information of the data sets are given below:

Table-1. Data set used.

Data set	Size	Attribute	Class
Tissue	106	9	4
Wholesale customer data	441	8	1
Wiki4HE	700	10	1
Cardiotocographic data	800	17	5

Table-2. Traditional k-means algorithm.

Dataset	No. of class	No. of cluster	Execution time	No. of iterations
Tissue	4	3	0.04	5
Wholesale customer data	1	2	0.03	4
Wiki4HE	1	2	0.05	10
Cardiotocographic data	5	4	1.86	19

Table-3. Improved k-means algorithm.

Dataset	No. of class	No. of cluster	Execution time	No. of iterations
Tissue	4	3	0.02	5
Wholesale customer data	1	2	0.01	3
Wiki4HE	1	2	0.04	8
Cardiotocographic data	5	4	1.00	15

Experimental result shows that the IKM gives better result when compared with K-means algorithm in terms of execution time and the no. of iterations.

7. CONCLUSIONS

Data mining in latest years with the databank and artificial intelligence established a novel tool, its aim the huge volume of data from the mined useful knowledge, to attain the actual intake of data assets. Clustering plays a vital role in various application. The most commonly used effective clustering algorithm is K-means clustering. K-means is a significant issue of research currently these days in data mining. In K-means algorithm various kinds of distance metric can be used to determine the distance between objects. Additionally distance metric also have effect on size of the cluster. So distance metric should be taken wisely and rendering to the dataset. This paper have presented an analysis on various research effort done in this extent. Though k- means is standing at the stage of consideration and improvement. The survey comprises that numerous advances are essentially necessary on K-means to expand problematic of cluster initialization,

cluster excellency and effectiveness of algorithm. Experimental result shows that the IKM gives better result compared to K-means even if the dataset is large. For future enhancement it can be improved further for better result.

REFERENCES

[1] Busygin S., Prokopyev O., Pardalos P. M. 2008. Biclustering in data mining. Comput. Oper. Res. 35(9): 2964-2987.

[2] Aitkin M., Aitkin I. 2011. Theories of data analysis and Statistical inference. In Statistical Modeling of the National Assessment of Educational Progress. Springer, New York. pp. 1-21.

[3] Masmoudi Y., Chabchoub H., Hanafi S., Rebai A. 2010. A mathematical programming based procedure for Breast cancer classification. J. Math. Modeling Algorithms. 9(3): 247-55.

[4] Taherdangkoo M., Yazdi M., Rezvani M. H. 2010. Segmentation of MR brain image using FCM improved by artificial bee colony (ABC) algorithm. In: Proceedings of the International Conference on Information Technology and Applications in Biomedicine (ITAB). pp. 1-5.

[5] Malwinder singh, Meenakshi bansal. 2015. A Survey on Various K-Means algorithms for Clustering. IJCSNS International Journal of Computer Science and Network Security. 15(6).

[6] K. Mohana Prasad, Dr. R. Sabitha. Meta Physical Algorithmic Representation for Flawless Clustering. Journal Theoretical and Applied Information Technology (JATIT), ISSN: 1817-3195, 76(1): 82-87.

[7] Jinchao Ji , TianBai, Chunguang Zhou, ChaoMa, Zhe Wang. K. 2013. Prototypes clustering algorithm for mixed numeric and categorical data. ELSEVIER Neurocomputing. 120: 590-596.

[8] Taher Niknam, Elahe Taherian Fard, Narges Pourjafarian, Alireza Roust. 2011. An efficient Hybrid Algorithm based on modified imperialist Competitive Algorithm and K-Means for data clustering. Engineering Applications of Artificial Intelligence. 24: 306-317.

[9] Yinghua Lu, TinghuaiMa, Changhong Yin, Xiaoyu Xie, Wei Tian and ShuiMing Zhong. 2013. Implementation of the Fuzzy C-Means Clustering Algorithm in Meteorological Data International Journal of Database Theory and Application. 6(6): 1-18.

[10] Mohammad Taherdangkoo, Mohammad Hadi Bagheri. 2013. A powerful hybrid clustering method based on modified stem cells and Fuzzy C-means Algorithms. Engineering Applications of Artificial Intelligence ELSEVIER. 26: 1493-1502.

[11] Tsai C. Y. and Chiu C. C. 2008. Developing a feature Weight self-adjustment mechanism for a K-Means Clustering algorithm. Computational Statistics and Data Analysis. 52: 4658-4672.

[12] Gursharan Saini. 2014. Harpreet Kaur International Journal of Computer Science and Information Technologies. 5(4): 5978-5986.

[13] D. Randall Wilson and Tony R. Martinez. 1997. Improved Heterogeneous Distance Functions. Journal of Artificial Intelligence Research 6 (1997) 1-34 Submitted 5/96; published 1/97 © 1997 AI Access Foundation and Morgan Kaufmann Publishers. All Rights reserved.

[14] Antoni Moore. The case for approximate Distance Transforms. Presented at SIRC 2002-The 14th Annual Colloquium of the Spatial Information Research Centre University of Otago, Dunedin, New Zealand December 3-5th 2002.

[15] Richa Loohach and Kanwal Garg. 2012. Effect of Distance Functions on K-Means Clustering Algorithm. International Journal of Computer Applications (0975-8887) 49(6).

[16] Jinchao Ji, Wei Pang, Chunguang Zhou, Xiao Han, Zhe Wang. 2012. A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data. Knowledge-Based Systems ELSEVIER. 30: 129-135.

[17] Dharmendra K Roy and Lokesh K Sharma. 2010. Genetic k-Means Clustering Algorithm for Mixed Numeric and Categorical Data Sets. International journal of Artificial Intelligence and Applications (IJAIA). 1(2).

[18] S. Anitha Elavarasi1 and J. Akilandeswari. 2014. Survey on clustering algorithm and similarity measure for caterogical data. ICTACT journal on soft computing, January. 4(2).

[19] S. Anitha Elavarasi1 and J. Akilandeswari. 2014. Survey on clustering algorithm and similarity Measure for categorical data. ICTACT Journal of soft computing, January. 4(2).

[20] Yiu-ming Cheung, Hong Jia. 2013. Categorical- and numerical attribute data clustering based on a Unified similarity metric without knowing cluster Number Pattern Recognition ELSEVIER. 46: 2228-2238.

[21] Mohammad Taherdangkoo, Mahsa Paziresh, Mehran Yazdi. Mohammad Hadi Bagheri. 2013. An Efficient algorithm for function optimization: modified Stem cells algorithm. Research Gate, March, DOI: 10.2478/s13531-012-0047-8.

[22] Yun-Chia Liang, Alice E. Smith. 2004. AnAnt Colony Optimization Algorithm for the Redundancy Allocation Problem (RAP). IEEE Transactions on Reliability. 53(3).

[23] Anders Bergman. 2012. Analysis of Metrological Requirements for Electrical Measurement of HVDC Station Losses. IEEE Transactions on Instrumentation A Measurement. 61(10).

[24] Saeid ARISH, Ali AMIRI, Khadije NOORI. 2011. FICA: Fuzzy imperialist competitive algorithm, Sadhana. 36(3): 293-315._c Indian Academy of Sciences.

[25] Ai-Qin Mu, De-Xin Cao, Xiao-Hua Wang. 2009. A Modified Particle Swarm Optimization Algorithm. Natural Science. 1(2): 151-155.

[26] Pritesh Vora, Bhavesh Oza. 2013. A Survey on K-mean Clustering and Particle Swarm Optimization. International Journal of Science and Modern Engineering (IJISME) ISSN: 2319-6386. 1(3).

[27] Kalpana D. Joshi, P.S.Nalwade. 2013. Modified K-Means For Better Initial Cluster Centres, IJCSMC. 2(7): 219-223 ISSN 2320-088X.

[28] K. Mohana Prasad, Dr. R. Sabitha. Evolution of an Algorithm for Formulating Efficient Clusters to Eliminate Limitations. International Journal of Applied Engineering Research (Ijaer), ISSN: 0973-4562, 9(23): 20111-20118.