Volume 3, Issue 11, November 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Modified K- Means Clustering Algorithm for Data Mining in Education Domain

Anurag Bhardwaj*
Computer Sciences Corporation, India

Ashutosh Bhardwaj
Infosys Limited, India

Abstract— Data Mining is a process of extracting previously unknown, valid, potentially useful and hidden patterns from large data sets. Data mining is mainly used in commercial applications. Clustering is a multivariate analysis technique where individuals with similar characteristics are determined and classified (grouped) accordingly. In data mining, K-Means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. In this paper we concentrated on the application of data mining in an engineering education environment. The relationship between student's university internal examination results and their success was studied using cluster analysis and modified K-Means algorithm techniques. We proposed an improved K-Means algorithm using noise data filter. The algorithm developed density-based detection methods based on characteristics of noise data where the discovery and processing steps of the noise data are added to the original algorithm. By pre-processing the data to exclude these noise data before clustering data set the cluster cohesion of the clustering results is improved significantly and the impact of noise data on K-Means algorithm is decreased effectively and the clustering results are more accurate.

Keywords— Data Mining, Clustering, K-Means, Noise, Density.

I. INTRODUCTION

Data mining [1] is a technology used in different disciplines to search for significant relationships among variables in large data sets. The amount of data maintained in an electronic format has seen a dramatic increase in recent times. Data Mining is a process of extracting previously unknown, valid, potentially useful and hidden patterns from large data sets. These tools can include statistical models, mathematical algorithms, and machine learning methods [2] (algorithms that improve their performance automatically through experience, such as neural networks or decision trees). Consequently, data mining consists of more than collecting and managing data, it also includes analysis and prediction. Data mining can be performed on data represented in quantitative, textual, or multimedia forms. Data mining is a technology used in different disciplines to search for significant relationships among variables in large data sets. Clustering [3] is a multivariate analysis technique where individuals with similar characteristics are determined and classified (grouped) accordingly? Cluster analysis is one of the key technologies in the field of data mining and machine learning which has been applied in many areas: data mining and knowledge discovery, pattern recognition and pattern classification, data compression and vector quantization and plays an important role in biology, geology, geography, and marketing in addition.

The K-Means [4] clustering algorithm is proposed by Mac Queen in 1967 which is a partition-based cluster analysis method. It is used widely in cluster analysis for that the K-Means algorithm has higher efficiency and scalability and converges fast when dealing with large data sets. The K-Means algorithm [5] defines a random cluster centroid according to the initial parameters. Each consecutive case is added to the cluster according to the proximity between the mean value of the case and the cluster centroid. The clusters are then re-analysed to determine the new centroid point. This procedure is repeated for each data object. However it also has many deficiencies: the number of clusters K needs to be initialized, the initial cluster centers are arbitrarily selected, and the algorithm is influenced by the noise points.

The proposed algorithm developed density-based detection methods based on characteristics of noise data where the discovery and processing steps of the noise data are added to the original algorithm. By pre-processing the data to exclude these noise data before clustering data set the cluster cohesion of the clustering results is improved significantly and the impact of noise data on K-Means algorithm is decreased effectively and the clustering results are more accurate. The relationship between student's university [6] internal examination results and their success was studied using cluster analysis and modified K-Means algorithm techniques. The rest of the paper is divided as follows. Section 2 presents the related work. Section 3 covers the proposed algorithm. Section 4 present implementation and result. Section 5 concludes the paper.

II. RELATED WORK

[7] Describes extensions to the *K-Means* algorithm for clustering data sets. By adding suitable constraints into the mathematical program formulation, an approach is developed, which allows the use of the *K-Means* paradigm to efficiently cluster data sets with the number of objects in each cluster.

[8] presents a data clustering approach using modified K-Means algorithm based on the improvement of the sensitivity of initial center (seed point) of clusters. The proposed algorithm partitions the whole space into different segments and calculates the frequency of data point in each segment. The segment which shows maximum frequency of data point will have the maximum probability to contain the centroid of cluster. The number of cluster's centroid (k) will be provided by the user in the same manner like the traditional K-mean algorithm and the number of division will be k*k ('k' vertically as well as 'k' horizontally). If the highest frequency of data point is same in different segments and the upper bound of segment crosses the threshold 'k' then merging of different segments become mandatory and then take the highest k segment for calculating the initial centroid (seed point) of clusters. In this paper we also define a threshold distance for each cluster's centroid to compare the distance between data point and cluster's centroid with this threshold distance through which we can minimize the computational effort during calculation of distance between data point and cluster's centroid. It is shown that how the modified k-mean algorithm will decrease the complexity & the effort of numerical calculation, maintaining the easiness of implementing the K-Mean algorithm. It assigns the data point to their appropriate class or cluster more effectively.

[9] Present a divisive hierarchical method for the analysis and segmentation of visual images. The proposed method is based on the use of the K-Means method embedded in a recursive algorithm to obtain a clustering at each node of the hierarchy. The recursive algorithm determines automatically at each node a good estimate of the parameter k (the number of clusters in the K-Means algorithm) based on relevant statistics. The author made several experiments with different kinds of images obtaining encouraging results showing that the method can be used effectively not only for automatic image segmentation but also for image analysis and, even more, data mining.

[10]The traditional *K-Means* algorithm is, however, inefficient while working on large numbers of data sets and improving the algorithm efficiency remains a problem. This paper focuses on the efficiency issues of cluster algorithms. A refined initial cluster centers method is designed to reduce the number of iterative procedures in the algorithm. A parallel *K-Means* algorithm is also studied for the problem of the operation limitation of a single processor machine when given huge data sets. The analytical results demonstrate that these improvements can greatly enhance the efficiency of the *K-Means* algorithm, i.e., allow the grouping of a large number of data sets more accurately and more quickly. The analysis has theoretical and practical importance for work on the improvement and parallelism of cluster algorithms.

[4]A major drawback of K-Means algorithm is that it is difficult to determine the parameter k to represent natural cluster, and it is only suitable for concave spherical clusters. The paper presents an efficient clustering algorithm which combines the hierarchical approach with the grid partition. The hierarchical approach is applied to find the genuine clusters by repeatedly combining together these blocks. Hilbert curve is a continuous path which passes through every point in a space between the coordinates of the points and the one dimensional sequence numbers of the points on the curve. The goal of using Hilbert curve is to preserve the distance of that points which are close in space and represent similar data should be stored close together in the linear order. The simulation shows that the clustering algorithm can have shorter execution time than K-Means algorithms for the large databases. Moreover, the algorithm can deal with clusters with arbitrary shapes in which the K-Means algorithm can not discover.

[5] So far, the K–means algorithm is the most widely used method for discovering clusters in data, and it has been used extensively in the commercial field, such as customer analysis. However, the efficiency of the algorithm needs to be improved when faced with large amounts of data. The improved algorithm avoids unnecessary calculations by using the triangle inequality. Applies the improved algorithm for customer classification. Experiments show that the optimizated algorithm take lower time overhead than the standard K–means algorithm and the superiority of proposed method is more remarkable as the number of clusters increases.

[11] Paper is a cluster analysis algorithm research carried out based on the existing data mining, which focuses on the current popular and commonly used K-Means algorithm, and presents an improved K-harmonic means clustering algorithm through using a new distance measure. Through the regulation of distance metric parameters can achieve better clustering effects than the traditional K-harmonic means, and has an advantage both in run time and number of iterations.

In [12] Cluster analysis is explained which a multivariate analysis technique is where individuals with similar characteristics are determined and classified (grouped) accordingly. Through cluster analysis, dense and sparse region can be determined in the distribution, and different distribution patterns may be achieved. The concepts of similarities and differences are used in cluster analysis. Different measures may be used in determining similarities and differences. This study utilizes the Euclidian distance measure.

The Proposed model [7] makes prediction about fail and pass ratio of students based on class performance as well as system inform the students about the ratio of class attendance. The proposed model also deals with entrance ratio of students in a particular department and exit ratio after successful completion of degree. Model was developed using DMX queries available in visual studio 2005. The proposed model identifies the weak students before final exam in order to save them from serious harm.

III. PROPOSED WORK

The K-Means algorithm [5] defines a random cluster centroid according to the initial parameters. Each consecutive case is added to the cluster according to the proximity between the mean value of the case and the cluster centroid. The clusters are then re-analysed to determine the new centroid point. This procedure is repeated for each data object.

However it also has many deficiencies: the number of clusters K needs to be initialized, the initial cluster centers are arbitrarily selected, and the algorithm is influenced by the noise points. The proposed algorithm developed density-based detection methods based on characteristics of noise data where the discovery and processing steps of the noise data are added to the original algorithm. By pre-processing the data to exclude these noise data before clustering data set the

cluster cohesion of the clustering results is improved significantly and the impact of noise data on K-Means algorithm is decreased effectively and the clustering results are more accurate. The relationship between student's university internal examination results and their success was studied using cluster analysis and modified K-Means algorithm techniques.

The abnormal degree of each object in database is measured by local outlier factor LOF[13]. LOF first generates k-Neighbourhood and k-nearest neighbour distance of all objects, and then calculates the distance between each object and the objects which are in its k-Neighbourhood; at last LOF identifies local outlier according to the local outlier factor of each object [7]. The procedure of outlier

Detection is briefly described as follows:

I) Calculate k-nearest neighbourhood distance named

distance (p, i)(i E Nk(p» of each object p, distance (p, i) is defined as the direct connection distance between object p and

distance $\sqrt{(x^1 - y^1)^2 + (x^2 - y^2)^2 + \cdots (x^n + y^n)^2}$ is the dimension of the data set.

Calculate the density of each object p. The density of object p which reflects the distribution of the data near is defined by the reciprocal of k-nearest neighbour mean. It is described as follows:

$$lrd(p) = \frac{1}{\frac{1}{k}\sum_{i=1}^{k} dis\tan ce(p,i)}.$$

Calculate the local outlier factor of p

$$lof(p) = \frac{\sum_{i=1}^{k} \frac{lrd(i)}{lrd(p)}}{k},$$

Proposed Algorithm

Algorithm:

The improved clustering algorithm is described as follows:

Input: data set M {x), X2,.., xm} to be clustered, the number of clusters k;

Output: k clusters and the sum of dissimilarity between all objects and their nearest cluster centers is the smallest.

1) Traversal on data set M, calculate lof (p),

if lof (p) is much larger than 1
remove the isolated point p,
otherwise
leave p;
finally get the new data set N;

2) Calculate the mean of data set N as the first cluster center.

3) Find the next cluster center. Calculate the distance between the remaining points and the cluster center.
Calculate the distance between each object and the center of each cluster, and assign it to the nearest cluster.

4) Calculate the mean of objects in each cluster as a new cluster center.

5) Repeat 3) 4) until the criterion function E converged, return (m), m2 . . . mk).

According to the improved algorithm as isolated points are excluded from the data set with the application of density-based outlier detection so that the sample which is farthest from the cluster center can be selected as the next best initial cluster center, the sensitivity to outliers in the K-Means algorithm is decreased effectively [17], and the accuracy of clustering is improved. The algorithm is also suitable for large data sets and the time will increase when dealing with large data sets for an additional traversal.

IV. IMPLEMENTATION

We have evaluated our algorithm on several different real datasets and synthetic dataset. We have compared our results with that of *K-Means* algorithm in terms of the total execution time and quality of clusters. Our experimental results are reported on PC Intel 2.0 GHz CPU; 2 GB RAM 512 kB Cache.

We have used college dataset of examination department and Training and Placement departments. In our dataset we have used database of students, their attendance, sessional and practical marks and faculty.

We have used Orange tool for algorithm implementation and Python as a programming language. Orange is In this section we verify the accuracy, time, effectiveness and feasibility of the proposed algorithm. The college data set is used to test the clustering accuracy, time of our proposed algorithm. We can see that the improved algorithm proposed in this paper has a higher clustering accuracy than traditional K-Means algorithm. We applied college examination and training and placement data[14][15] set to prove our results. From fig 4.1 and fig 4.2 we find our results are more accurate as compared to traditional method.

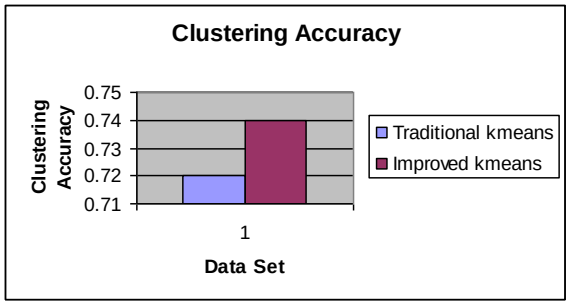

Fig 4.1

Fig 4.2

V. CONCLUSION

Data mining is a technology used in different disciplines to search for significant relationships among variables in large data sets. The K-Means clustering algorithm is a partition-based [16] cluster analysis method. It is used widely in cluster analysis for that the K-Means algorithm has higher efficiency and scalability and converges fast when dealing with large data sets. However it also has many deficiencies: the number of clusters K needs to be initialized, the initial cluster centers are arbitrarily selected, and the algorithm is influenced by the noise points. In view of the shortcomings of the traditional K-Means clustering algorithm, we proposed an improved K-Means algorithm using noise data filter. The algorithm developed density-based detection methods based on characteristics of noise data where the discovery and processing steps of the noise data are added to the original algorithm. By pre-processing the data to exclude these noise data before clustering data set the cluster cohesion of the clustering results is improved significantly and the impact of noise data on K-Means algorithm is decreased effectively and the clustering results are more accurate. Our proposed technique can be applied in the field of education. The relationship between student's university internal examination results [18] [19] and their success was studied using cluster analysis and modified K-Means algorithm techniques.

REFERENCES

[1] Han, J., Kamber, W., "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, USA, 5-10, 2001.
[2] Ren Jingbiao,Yin Shaohong, Research and Improvement of Clustering Algorithm in Data Mining, ICSPS, 2010, pg 842-845
[3] Behrouz.et.al, (2003) Predicting Student Performance: An Application of Data Mining Methods with the Educational Web-Based System Lon-CAPA © 2003 IEEE, Boulder, CO.
[4] Barkha H.Desai, Nisha Shah, Hetal Bharat Bhavsar, Comparative Study of K-Means Type Algorithms, UNIASCIT, 2011, pg 71-73
[5] Tian Jinan, Zhu Lin, Zhang Suqin, Improvement and Parallelism of *K-Means* Clustering Algorithm, IEEE 2005, pg 227-281
[6] Xiaoping Qin, Shijue Zheng , Tingting He□Ming Zou, Ying Huang, Optimizated K-Means algorithm and application in CRM system, International Symposium on Computer, Communication, 2010, pg 519-522
[7] Ran Vijay Singh, M.P.S Bhatia, Data Clustering with Modified K-Means Algorithm, IEEE-International Conference on Recent Trends in Information Technology, ICRTIT 2011, pg 717-721
[8] Jose, Javiar, Denail, A devisive Hierarchical K-Means based algorithm for image segmentation, IEEE 2010, pg 300-304.
[9] Michael K. Ng, A note on constrained *K-Means* algorithms, Elsevier Science, 2000, pg 515-520
[10] Hongbin Zhao, Qilong Han, Haiwei Pan, A Hierarchical Clustering Algorithm Based on Grid Partition, International Conference on Multimedia Communications, IEEE 2010, pg 187-190
[11] Juntao Wang, Xiaolong Su, An improved K-Means clustering algorithm, IEEE 2011, pg 44-46
[12] Yuqing, P., Xiangdan, H., Shang, L., "The K-Means Clustering Algorithm Based On Density and Ant Colony", IEEE Int. Conf. Neural Networks & Signal Processing Nanjing, China, 457-460, December 14-17, 2003.
[13] Erdogan and Timor (2005) A data mining application in a student database. Journal of Aeronautic and Space Technologies July 2005 Volume 2 Number 2 (53-57).
[14] Galit.et.al (2007) Examining online learning processes based on log files analysis: a case study. Research, Reflection and Innovations in Integrating ICT in Education.
[15] Kifaya(2009) Mining student evaluation using associative classification and clustering. Communications of the IBIMA vol. 11 IISN 1943-7765.
[16] Alaa el-Halees (2009) Mining Students Data to analyze e-Learning Behavior: A Case Study.
[17] Aher, B, Sunita and J.R.M.L., Lobo(2011) "Data mining in Educational System using WEKA".
[18] Bhardwaj, Kumar, Brijesh and Pal, Sourabh(2011) "Mining Educational Data to Analyze Student's Performance".
[19] Mustafa, Tasleem, Sattar, Ahsan, Raza, and Kha, Inayat(2010), M. "Data mining model for Higher Education System".

