Volume 6, Issue 6, June 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Normalization Based K-means Data Analysis Algorithm

Navdeep Kaur
Research Scholar
Department of Computer Science, JCD Vidyapeeth
Affilated to Guru Jambheshwar University, Hisar, India

Krishan Kumar
Assistant Professor
Department of Computer Science, JCD Vidyapeeth
Affilated to Guru Jambheshwar University, Hisar, India

Abstract— Data mining is analysis step to knowledge discovery in the database process. It is the process of extraction knowledge from large databases. Clustering is a process of partitioning a set of data (or objects) into a set of meaningful sub-classes, called clusters. Used either as a stand-alone tool to get insight into data distribution or as a pre processing step for other algorithms. K-means is a good clustering technique. With the proposed algorithm, normalization of data prior to clustering. Then a efficient algorithm used for clustering which is better than simple k-means algorithm.

Keywords— Normalization, k-means, Clustering, Data mining, Algorithm.

I. INTRODUCTION

Data mining technology has created a new opportunity for exploiting the information from the databases. Patterns in the data, such as associations among similar items purchases, enables target marketing to focus on what things the customers are likely to purchase [1]. Data analysis is process of evaluating data using analytical and logical reasoning to examine each component of the data provided. This form of analysis is just one of the many steps that must be completed when conducting a research experiment. A clear understanding of data analysis and some of its baseline rules can help attest professionals perform their work and reach their objectives with greater efficiency and effectiveness. The use of data analysis services adds value in both external and internal audit engagements. Proper application of data analysis helps professionals who perform audit work by streamlining resources while maintaining an effective audit process. Simultaneously, auditors can address the risk associated with the specific audit areas as required by professional standards.Data from various sources is gathered, reviewed, and then analyzed to form some sort of finding or conclusion. Clustering is the process of partitioning a set of data into a set of meaningful sub-classes, called clusters. It helps users to understand the natural grouping or structure in a dataset. A good clustering method will produce high quality clusters in which the intra-class similarity is high and the inter-class similarity is low. The quality of clustering result also depends on both the similarity measure used by the method and its implementation. The quality of a clustering method is also measured by its ability to discover some or the entire hidden pattern [2]. K-Mean clustering is usually an extremely popular protocol to find the clusters inside a dataset through iterative calculations. It offers the luxury of simple execution as well as locating at the least nearby optimal clustering.

II. RELATED WORK

Frigui et al. (1996) proposed an algorithm named Competitive Agglomeration (CA), which combines the advantages of hierarchical and partitional clustering techniques. The CA algorithm starts by partitioning the data set into the large number of small clusters. As the algorithm progresses, adjacent clusters compete for data points and clusters that lose in the competition gradually become depleted and vanish. Thus, as the iteration proceeds, a sequence of partitions is obtained with a progressively diminishing number of clusters. The final partition is taken to have optimal number of clusters from the point of view of the objective function [3].

Khalid (1997) designed an approach that uses a weighted Mahalanobis distance as a distance metric to perform partitional clustering. This WMD prevents the generation of unusually large or unusually small clusters. The problem of this algorithm is in the estimation of optimal number of subgroup present in the data set. It achieves natural clustering with less iteration than the competing algorithm. Furthermore, by restricting the determinant of each cluster to unity, a monotonically decreasing criterion function is obtained which can be used to estimate the number of sub-structures in the data [4].

Abidi et al. (2000) describes a novice approach to perform the data mining tasks of Data Clustering in the general framework of Artificial Neural Network. In the ANN paradigm, typically, unsupervised learning based Self Organizing Map (SOM) is used for data clustering tasks. For well defined data, the SOM is able to cluster the data into visually distinct clusters which can be easily interpreted by data analysis. The general idea is that, given a high-level topological ordering by a SOM, and then feed this SOM-driven clustering information to a K-Means algorithm to refine the SOM's output by way of demarcating the output layer of the SOM into distinct clusters. The ensuing result of the experiment indicates that the post-processing of the SOM's output by the classical K-Mean algorithms results in a much better cluster structure of the data set [6].

Pavel et al. (1999) found the clustering in data mining techniques which presents an overview of pattern clustering methods from a statistical pattern recognition perspective. Some important applications of clustering algorithms are image segmentation, object recognition, and information retrieval. Clustering is an interesting, useful, and challenging problem in data mining technique. It has great potential in applications like object recognition, image segmentation, and information filtering and retrieval. However, it is possible to exploit this potential only after making several designs choice carefully [5].

Leeand Antonsson(2000) proposed an approach of Evolution Strategy (ES) implementing variable length genomes is developed to address the problem of dynamic partitional clustering. As opposed to static, dynamic partitional clustering does not require the prior specification of the number of clusters. The proposed ES implements variable length genomes that allow the algorithm to effectively search for both optimal clusters center position and cluster number. Cluster number is optimized during runtime, such clustering re referred to as dynamic. The proposed ES is developed as a general framework for dynamic partitioning; it would be interesting to observe its performance using different fitness criteria [7].

III. PROPOSED ALGORITHM

In the purposed algorithm , I use Gaussian (also called normal) normalization. It firstly computes the mean of Height and Weight for the data points. Then the items in the collection which is subtracting both Height and Weight by the means and then to the power two. The next step is to divide these two values by the number of items in the collection. Finally, the last loop populates the collection by adding new objects of type Data Point to it. The values used for the two properties are the normalized versions of the values in the collection. Now, most normalized values will be between -3and +3.

The purposed algorithm :

1. Load initial data set.
2. Find the maximum and minimum values of each feature from the dataset.
3. Normalize real scalar values of datasets with maximum and minimum values using equation :

$$v = \frac{v' - \min(e)}{\max(e) - \min(e)}$$

Where min(e) and max(e) are the minimum and the maximum values for attribute E.

4. Initialize the data points (n) and Number of Clusters (K)
5. Checkpoint Cluster Value (K)
6. If number K=1, then Exit
7. Else
8. Calculate Min(Data Point) and Max(Data Point)
9. Calculate Group Area Range (A_G) with Equation (Max(Data_Points) -Min(Data_Points))/Number of Clusters(K)
10. Data_Points Division in Number of Cluster (K) Group with Width $A_{G.}$
11. Frequency Calculation of Data_Points in Division Partitions.
12. Select highest Frequency Data_Points K Group.
13. Calculate Mean of Data-Point in group.
14. Initialize V=1
15. Analyze closest pair of Data_Points from collection of points and generate Data_Points set S_V and 1<V<=K having Data_Points and merge.
16. Analyze the closest Data_Points with Data_Points collection S_V and add to $S_{V,}$ then merge.
17. Repeat step 12 until the Data_Points in $S_V$is in Range 0.6<L<0.9 * (n/k)
18. If V<K, then V++, Search another pair of Data_Points.
19. Form Data_Points set S_V and Merge, move to Step 12.
20. Distance Calculation of each Data_Pointsdist$_i$, Set 1<=i<=n with centroids Cj, 1<=j<=K and d(dist$_i$,Cj)
21. Analyze the closest centroids Cj and assign it to cluster based on dist.
22. Set ClusterNum[i]=j and assign d(disti,Cj) as nearest distance ;//nearest cluster number
23. For each Recalculate centroids for each cluster j. Repeat steps 24 to 26
24. For each Data_Pointsdist$_i$
25. Distance computation from centroids with present closest cluster.
26. If dist<=nearestDistance, Data_Points stable in cluster and no move.
27. Else, for each centroids Cj, compute distance (disti,Cj), End Loop.
28. Assign Data_PointsDist$_i$ to Cluster with nearest centroids Cj.
29. Set ClusterNum[i]=j and assign d(disti,Cj) as nearest distance, End Loop.
30. Repeat until convergence with recalculation of centroids

First of all, no. of data points and no. of clusters are initialized. If no. of cluster is 1 then nothing will happen. Otherwise Maximum and Minimum data points are calculated. By using these data points, Group area range is calculated by using the above equation. Data points are divided in no. of clusters. Frequency calculation of data points is done. Highest frequency data points are selected and put into a group. Mean of data points in a group is calculated. Now closet pair of data points are analyzed and merge them. After this, distance is calculated for each data points. Cluster is assigned

to the closet centroid data points. This step is repeated until all data points assigned cluster. Main benefit of this algorithm is that after grouping the data points, the distance of a data point is calculated and compare with the group.

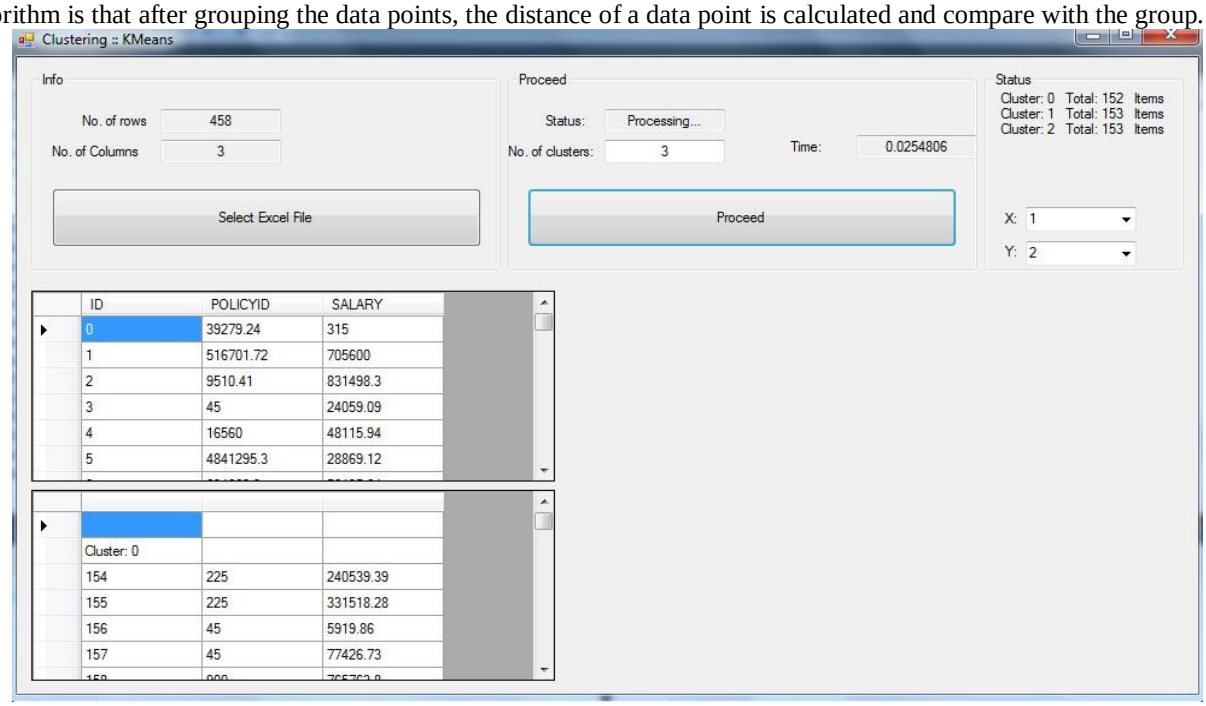

Fig 1: Normalized K-means clustering algorithm

IV. CONCLUSION

K-means is a good clustering algorithm. But it is always not provide efficient result. But with the normalization it provide the good result. This is an efficient data analysis algorithm which firstly normalize the collection of data. After that an efficient algorithm provide usually good result in clustering and this works very fast.

REFERENCES

[1] Clifton, C. and R. Steinheiser. 1998. **"Data Mining on Text",**Proceedings of the 22nd Annual IEEE International Computer Software and Applications Conference, COMPSAC98, pp. 630–635.

[2] Piatetsky-Shapiro G., Frawley W. (Eds.): **"Knowledge Discovery in Databases"**, AAAI Press, Menlo Park, CA, 1991.

[3] http://www.cs.waikato.ac.nz/~ml/weka/

[4] Frigui H. and Krishnapuram R. **"Competitive Fuzzy Clustering"**, IEEE 1996, Page No. 225-228.

[5] Ester Martin, Kriegel Hans-Peter introduced the Idea of **"Clustering for Mining in Large Spatial Database"**.

[6] BerkhinPavel introduced the Idea of **"Survey of Clustering Data Mining Techniques"**.

[7] Hasan A. Moh. , Chaoji V. , Salem S. and Zaki J. Moh. **"Robust Partitional Clustering by Outlier and Density Insensitive Seeding"**, ACM 2000.

