Volume 5, Issue 6, June 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Modified K-Means Clustering using Decision Tree

Parvinder Singh, Sheenam Malhotra
C.S.E. Deptt, Sri Guru Granth Sahib World University,
Fatehgarh Sahib, Punjab, India

Abstract— Clustering is a process of putting similar data into groups. Clustering can be considered the most important unsupervised learning technique so as every other problem of this kind; it deals with finding a structure in a collection of unlabeled data. This paper proposedan enhanced K-Means algorithm using the decision tree approach.

Keywords— Clustering, K-Means, Enhanced K-Means, Decision Tree

I. INTRODUCTION

Clustering is a data mining technique of grouping set of data objects into multiple groups or clusters so that objects within the cluster have high similarity, but are very dissimilar to objects in the other clusters. Dissimilarities and similarities are assessed based on the attribute values describing the objects. Clustering algorithms are used to organize data, categorize data, for data compression and model construction, for detection of outliers etc.Common approach for all clustering techniques is to find clusters centre that will represent each cluster. Cluster centre will represent with input vector can tell which cluster this vector belong to by measuring a similarity metric between input vector and all cluster centre and determining which cluster is nearest or most similar one [1]. Cluster analysis can be used as a standalone data mining tool to gain insight into the data distribution, or as a preprocessing step for other data mining algorithms operating on the detected clusters. Many clustering algorithms have been developed and are categorized from several aspects such as partitioning methods, hierarchical methods, density-based methods, and grid-based methods.Further data set can be numeric or categorical. Inherent geometric properties of numeric data can be exploited to naturally define distance function between data points. Whereas categorical data can be derived from either quantitative or qualitative data where observations are directly observed from counts.

Data mining may be applied as per the steps given below:

1. Anomaly detection: This is the identification of the unusual records or data errors.
2. Association rule learning: It Searches the relationships between variables. This is sometimes referred to as market basket analysis.
3. Clustering: This is the process of finding groups and structures in the data that are in some way or another "similar", without using known structures in the data.
4. Classification: This is the task of generalizing known structure to apply to new data.
5. Regression: this process is used to search a function which modals the data with the least error.
6. Summarization: It provides a more compact representation of the data set, including visualization and report generation [10].

II. VARIOUS CLUSTERING TECHNIQUES

A. K-Means Clustering

It is a partition method technique which finds mutual exclusive clusters of spherical shape. It generates a specific number of disjoint, flat(non-hierarchical) clusters. Stastical method can be used to cluster to assign rank values to the cluster categorical data. Here categorical data have been converted into numeric by assigning rank value [2]. K-Means algorithm organizes objects into k – partitions where each partition represents a cluster. We start out with initial set of means and classify cases based on their distances to their centers. Next, we compute the cluster means again, using the cases that are assigned to the clusters; then, we reclassify all cases based on the new set of means. We keep repeating this step until cluster means don‟t change between successive steps. Finally, we calculate the means of cluster once again and assign the cases to their permanent clusters.

i. K-Means Algorithm Properties

- There are always K clusters.
- There is always at least one item in each cluster.
- The clusters are non-hierarchical and they do not overlap.
- Every member of a cluster is closer to its cluster than any other cluster because closeness does not always involve the 'center' of clusters.
- ii. K-Means Algorithm Process

- The dataset is partitioned into K clusters and the data points are randomly assigned to the clusters resulting in clusters that have roughly the same number of data points.
- For each data point:
- Calculate the distance from the data point to each cluster.

 If the data point is closest to its own cluster, leave it where it is. If the data point is not closest to its own cluster, move it into the closest cluster.

- Repeat the above step until a complete pass through all the data points results in no data point moving from one cluster to another. At this point the clusters are stable and the clustering process ends.
- The choice of initial partition can greatly affect the final clusters that result, in terms of inter-cluster and intracluster distances and cohesion [6].

B. Hierarchical Clustering

A hierarchical method creates a hierarchical decomposition of the given set of data objects. Here tree of clusters called as dendrograms is built. Every cluster node contains child clusters, sibling clusters partition the points covered by their common parent. In hierarchical clustering we assign each item to a cluster such that if we have N items then we have N clusters. Find closest pair of clusters and merge them into single cluster. Compute distance between new cluster and each of old clusters. We have to repeat these steps until all items are clustered into K no. of clusters. It is of two types:

i. Agglomerative (bottom up)-

Agglomerative hierarchical clustering is a bottom-up clustering method where clusters have sub-clusters, which in turn have sub-clusters, etc. It starts by letting each object form its own cluster and iteratively merges cluster into larger and larger clusters, until all the objects are in a single cluster or certain termination condition is satisfied. The single cluster becomes the hierarchy‟s root. For the merging step, it finds the two clusters that are closest to each other, and combines the two to form one cluster [5].

ii. Divisive (top down)-

A top-down clustering method and is less commonly used. It works in a similar way to agglomerative clustering but in the opposite direction. This method starts with a single cluster containing all objects, and then successively splits resulting clusters until only clusters of individual objects remain [4].

C. DBSCAN Clustering

DBSCAN (Density Based Spatial Clustering of Application with Noise).It grows clusters according to the density of neighborhood objects. It is based on the concept of "density reachibility" and "density connectability", both of which depends upon input parameter- size of epsilon neighborhood e and minimum terms of local distribution of nearest neighbors. Here e parameter controls size of neighborhood and size of clusters. It starts with an arbitrary starting point that has not been visited [1]. The points e-neighbourhood is retreived, and if it contains sufficiently many points, a cluster is started. Otherwise the point is labelled as noise. The number of point parameter impacts detection of outliers. DBSCAN targeting low-dimensional spatial data used DENCLUE algorithm [4].

D. OPTICS

OPTICS (Ordering Points To Identify Clustering Structure) is a density based method that generates an augmented ordering of the data‟s clustering structure. It is a generalization of DBSCAN to multiple ranges, effectively replacing the e parametre with a maximum search radius that mostly affects performance. MinPts then essentially becomes the minimum cluster size to find. It is an algorithm for finding density based clusters in spatial data which addresses one of DBSCAN‟S major weaknesses i.e. of detecting meaningful clusters in data of varying density. It outputs cluster ordering which is a linear list of all objects under analysis and represents the density-based clustering structure of the data. Here parameter epsilon is not necessary and set to maximum value. OPTICS abstracts from DBSCAN by removing this each point is assigned as „core distance‟, which describes distance to its MinPts point. Both the core-distance and the reachability-distance are undefined if no sufficiently dense cluster w.r.t epsilon parameter is available [1].

E. STING

STING (STasticalINformation Grid) is a grid-based multi resolution clustering technique in which the embedded spatial area of input object is divided into rectangular cells. Statistical information regarding the attributes in each grid cell, such as the mean, maximum, and minimum values are stored as statistical parameters in these rectangular cells. The quality of STING clustering depends on the granularity of the lowest level of grid structure as it uses a multiresolution approach to cluster analysis. Moreover, STING doesnot consider the spatial relationship between the children and their neighbouring cells for construction of a parent cell. As a result, the shapes of the resulting clusters are isothetic, that is, all the cluster boundaries are either horizontal or vertical, and np diagonal boundary is detected. It approaches clustering result of DBSCAN if granularity approaches 0. Using count and cell size information, dense clusters can be identified approximately using STING [4].

III. LITERATURE REVIEW

Research in the various techniques in clustering is started in early 1990s. Now a days we have lots of clustering algorithms which are useful in different areas .we have different kind of clustering algorithms from which we can select

the best suited algorithm according to our requirement . K. A. Abdul Nazeer, M.P.Sebastianpresented an enhanced k-means algorithm which combines a systematic method for finding initial centroids and an efficient way for assigning data points to clusters. This method ensures the entire process of clustering in O(n2) time without sacrificing the accuracy of clusters. The previous improvements of the k-means algorithm compromise on either accuracy or efficiency. A limitation of the proposed algorithm is that the value of *k*, the number of desired clusters, is still required to be given as an input, regardless of the distribution of the data points. Evolving some statistical methods to compute the value of *k*, depending on the data distribution, is suggested for future research. Methods for refining the computation of initial centroids are worth investigating.

Malay K Pakhir proposed "modified algorithm" that maintained all important characteristic features of the basic k-means and at the same time eliminates the possibility of generation of empty clusters. It has been shown that the present algorithm is semantically equivalent to the serial k-means algorithm. Proposed clustering scheme was able to solve the empty cluster problem, to a great extent, without any significant performance degradation.

NehaAggarwal&KirtiAggarwalpresented the way to find the initial centres for the k-means so that every time the K-Means algorithm produces same result for the same dataset by using mid point based K-means algorithms and it also remove the limitation of k-means that the final cluster results heavily depends on the selection of initial centroids which causes it to converge at local optimum.

AhamedShafeeq&Hareesha [4] proposed "Dynamic clustering of data with modifief K-means algorithm". In which ,we can overcome the problem by finding the optimal number of clusters on the run. But the main drawback of the proposed approach was that it takes more computational time than the K-means for larger data sets.

PankajJadwal&Ruchi Dave proposed "An Improved and Customised I-K Means for Avoiding Similar Distance Problem", by which same distance problem can be solved and better result can be obtained. It depends on equal distribution of data in each cluster and quality factor.

Rajeev Kumar &RajeshwarPuran proposed " Enhanced K-Means Clustering Algorithm Using Red Black Tree and Min-Heap"**,** it saved the distances between data objects and clusters. It then dynamically changes them when required. However, the saving of the distances requires much space. Thus, although our algorithm is superior to the traditional k-means algorithm in terms of time complexity, it appears to be lagging behind in terms of space complexity. But the drawback of this algorithm is to sort its orientation.

IV. ISSUE AND CHALLENGES

Since there are various clustering algorithms available to automate or semi-automate the clustering procedure it is very difficult to choose the suitable algorithm for a particular dataset. As different algorithms applied on a dataset may produce different kind of results with different clusters. Each algorithm has its own run time, complexity, error frequency, resources used etc to complete the procedure of clustering. Another issue may be that the outcome of a clustering algorithm mainly depends on the type of dataset used. As the size and dimensions of dataset increases day by day this makes it difficult to handle for a particular clustering algorithm. Also the complexity of data set increases, which include data like audios, videos, pictures and other multimedia data which form very heavy database, this in turn create the time complexity of a clustering algorithm. Furthermore clustering algorithms do not concentrate on all of the requirements simultaneously and effectively which makes the result uncertain. Most of the clustering algorithms depends on the distance function used in the algorithm and if the given distance function do not perform efficiently then a new distance function may required which is difficult to formulate especially for multi-dimensional data this increases the tediousness of work. Also the output of a clustering algorithm can be interpreted in different ways which may create confusion for understanding the result by users. So we need an immense concern to choose a clustering algorithm for the dataset. The selection of a clustering algorithm may based on the type of dataset, time requirement, efficiency needed, accuracy required, error tolerance etc. so the main challenge is to choose the correct type of clustering algorithm for the data set which are based on user requirements among many known clustering algorithms so that user can get the desired results which helps in further research for data mining process.

V. PROPOSED WORK

In the proposed work the K-Means clustering algorithm is combined with Decision tree that provides a great help for the efficiency of K-Means.

Steps for Modified K-Means Clustering are as follows:

 The dataset is partitioned into K clusters and the datapoints are randomly assigned to the clusters resulting in clusters that have roughly the same number of data points.

- For each data point:
- Calculate the distance from the data point to each cluster along with apply Decision tree structure.

 If the data point is closest to its own cluster, leave it where it is. If the data point is not closest to its own cluster, move it into the closest cluster.

- Repeat the above step until a complete pass through all the data points results in no data point moving from one cluster to another. At this point the clusters are stable and the clustering process ends.
- The choice of initial partition can greatly affect the final clusters that result, in terms of inter-cluster and intracluster distances and cohesion

VI. CONCLUSION

This paper deals with study of modified clustering algorithm called Modified K-Means clustering using decision tree. It first defines the data mining process which is the method of finding predictive information from a huge amount of databases. Then it defines the clustering process which is the procedure of assemblage of the objects in groups whose members contain some kind of resemblance. After that a detailed study of clustering algorithms and their comparison in different perceptions are examined. This paper highlights the concerned issues and challenges which may be helpful for the upcoming researchers to carry on their work.

REFERENCES

[1] Manish Verma, MaulySrivastava, NehaChack, Atul Kumar Diswar, Nidhi Gupta," A Comparative Study of Various Clustering Algorithms in Data Mining*," International Journal of Engineering Reserch and Applications (IJERA),* Vol. 2, Issue 3, pp.1379-1384, 2012.

[2] Patnaik, Sovan Kumar, SoumyaSahoo, and Dillip Kumar Swain, "Clustering of Categorical Data by Assigning Rank through Statistical Approach," *International Journal of Computer Applications*,Vol*:* 43, Issue: 2, pp.1-3, 2012.

[3] Arockiam, L., S. S. Baskar, and L. Jeyasimman. 2012. Clustering Techniques in Data Mining.

[4] Han, J., Kamber, M. 2012. Data Mining: Concepts and Techniques, 3rd ed, pp. 443-491.

[5] Improved Outcome Software, Agglomerative Hierarchical Clustering Overview**.** Retrieved from: http://www.improvedoutcomes.com/docs/WebSiteDocs/Clustering/Agglomerative_Hierarchical_Clustering_Overview.htm [Accessed 22/02/2013].

[6] Improved Outcome Software, K-Means Clustering Overview. Retrieved from: http://www.improvedoutcomes.com/docs/WebSiteDocs/Clustering/K-Means_Clustering_Overview.htm [Accessed 22/02/2013].

[7] RenJingbiao and Yin Shaohong "Research and Improvement of Clustering Algorithm in Data Mining", 2010 2nd International Conference on Signal Processing Systems (ICSPS), ISSN: 978-1-4244-6893-5, 5-7 July 2010, V1-842 - V1-845,

[8] M. Srinivas and C. Krishna Mohan, "Efficient Clustering Approach using Incremental and Hierarchical Clustering Methods", ISSN: 1098-7576,18-23 July 2010, pp: 1-7

[9] Hemanta Kumar, KalitaDhruba Kumar and Bhattacharyya AvijitKar "A New Algorithm for Ordering of Points to Identify Clustering Structure Based On Perimeter of Triangle: OPTICS (BOPT)", 15th International Conference on Advanced Computing and Communications, ISBN: 0-7695-3059-1, 18-21 Dec. 2007, pp. 523 - 528

[10] H. G. Wilson, B. Boots, and A. A. Millward "A Comparison of Hierarchical and partitional Clustering Techniques for Multispectral Image Classification", ISBN: 0-7803-7536-X, vol.3, 24-28 June 2002, pp: 1624 - 1626

