Clustering in Big Data: A Review

Anju
M.Tech Student
Department of Computer
Science and Applications
Maharishi Dayanand University

Preeti Gulia
Assistant Professor
Department of Computer
Science and Applications
Maharishi Dayanand University

ABSTRACT
BIG DATA[1] is a term for data sets that are so large or complex that traditional data processing[4] applications are inadequate. Accuracy in big data may lead to more confident decision making, and better decisions can result in greater operational efficiency, cost reduction and reduced risk. Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method are used for knowledge discovery from databases. Cluster is a group of objects that belongs to the same class. In other words, similar objects are grouped in one cluster and dissimilar objects are group in another cluster. Clustering methods can be classified into Partitioning Method, Hierarchical Method, Density-based Method. Clustering analysis is used in several applications like market research, pattern recognition, data analysis. K-means clustering is well known partitioning method. But this method has problem of empty cluster. The problems with existing system[6] were analysis, capture, search, sharing, storage, transfer, visualization, querying-updating. These problems can be reduced by using proposed algorithm. In this paper clustering and proposed algorithm is discussed.

Keywords
Clustering, K-Mean, Data mining, Big data

1. INTRODUCTION
Analysis of data[7] sets can find new correlations to spot business trends, prevent diseases, combat crime and so on. Scientists, business executives, practitioners of medicine, advertising and governments alike regularly meet difficulties with large data sets in areas including Internet search, finance and business informatics. Scientists encounter limitations in e-Science work, including meteorology, genomics, connectomics, complex physics simulations, biology and environmental research. Data sets are growing rapidly in part because they are increasingly gathered by cheap and numerous information-sensing mobile devices, aerial (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers and wireless sensor networks. The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s as of 2012, every day 2.5 exabytes (2.5×10^{18}) of data is created. One question for large enterprises is determining who should own big data initiatives that affect the entire organization.

Relational database[10] management systems and desktop statistics and visualization packages often have difficulty handling big data. The work instead requires "massively parallel software running on tens, hundreds, or even thousands of servers".[10] What is considered "big data" varies depending on the capabilities of the users and their tools, and expanding capabilities make big data a moving target. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration.

2. CLUSTERING
Cluster is a group of objects that belongs to the same class. In other words, similar objects are grouped in one cluster and dissimilar objects[11] are grouped in another cluster.

Clustering is the process of making a group of abstract objects into classes of similar objects.

1. A cluster of data objects can be treated as one group.
2. While doing cluster analysis, the first partition the set of data into groups based on data similarity and then assign the labels to the groups.
3. The main advantage of clustering over classification is that, it is adaptable to changes and helps single out useful features that distinguish different groups.

2.1 Applications of Cluster Analysis[6]
1. Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing.
2. Clustering can also help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns.
3. In the field of biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionalities and gain insight into structures inherent to populations.
4. Clustering also helps in identification of areas of similar land use in an earth observation database. It also helps in the identification of groups of houses in a city according to house type, value, and geographic location.
5. Clustering also helps in classifying[6] documents on the web for information discovery.
6. Clustering is also used in outlier detection applications such as detection of credit card fraud.
7. As a data mining function, cluster analysis serves as a tool to gain insight into the distribution of data to observe characteristics of each cluster.

2.2 Requirements of Clustering in Data Mining
The following points throw light on why clustering is required in data mining

- **Scalability** − Need to highly scalable clustering algorithms to deal with large databases.
- **Ability to deal with different kinds of attributes** − Algorithms should be capable to be applied on any kind of data such as interval-based (numerical) data, categorical, and binary data.
- **Discovery of clusters with attribute shape** − The clustering algorithm[8] should be capable of detecting clusters of arbitrary shape. They should not be bounded to only distance measures that tend to find spherical cluster of small sizes.
- **High dimensionality** − The clustering algorithm should not only be able to handle low-dimensional data but also the high dimensional space.
- **Ability to deal with noisy data** − Databases contain noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters.
- **Interpretability** − The clustering results should be interpretable, comprehensible, and usable.

3. CLUSTERING METHODS[11]

Clustering methods can be classified into the following categories −

- Partitioning Method
- Hierarchical Method
- Density-based Method
- Grid-Based Method
- Model-Based Method
- Constraint-based Method

• Partitioning Method

Suppose that are given a database of 'n' objects and the partitioning method constructs 'k' partition of data. Each partition will represent a cluster and $k \leq n$. It means that it will classify the data into k groups, which satisfy the following requirements −

- Each group contains at least one object.
- Each object must belong to exactly one group.

Points to remember −

- For a given number of partitions (say k), the partitioning method will create an initial partitioning.
- Then it uses the iterative relocation technique to improve the partitioning by moving objects from one group to other.

• Hierarchical Methods

This method creates a hierarchical decomposition of the given set of data objects. Many authors can classify hierarchical methods[5] on the basis of how the hierarchical decomposition is formed. There are two approaches here

- Agglomerative Approach
- Divisive Approach

• Agglomerative Approach

This approach is also known as the bottom-up approach. In this, it start with each object forming a separate group. It keeps on merging the objects or groups that are close to one another. It keep on doing so until all of the groups are merged into one or until the termination condition holds.

• Divisive Approach

This approach is also known as the top-down approach. In this, it has start with all of the objects in the same cluster. In the continuous iteration, a cluster is split up into smaller clusters. It is down until each object in one cluster or the termination condition holds. This method is rigid, i.e., once a merging or splitting is done, it can never be undone.

• Approaches to Improve Quality of Hierarchical Clustering

Here are the two approaches that are used to improve the quality of hierarchical clustering −

- Perform careful analysis of object linkages at each hierarchical partitioning.
- Integrate hierarchical agglomeration by first using a hierarchical agglomerative algorithm to group objects into micro-clusters, and then performing macro-clustering on the micro-clusters.

• Density-based Method

This method is based on the notion of density. The basic idea is to continue growing the given cluster as long as the density in the neighborhood exceeds some threshold, i.e., for each data point within a given cluster, the radius of a given cluster has to contain at least a minimum number of points.

• Grid-based Method

In this, the objects together form a grid. The object space is quantized into finite number of cells that form a grid structure.

Advantage

- The major advantage of this method is fast processing time.
- It is dependent only on the number of cells in each dimension in the quantized space.

• Model-based methods

In this method, a model is hypothesized for each cluster to find the best fit of data for a given model. This method locates the clusters by clustering[6] the density function. It reflects spatial distribution of the data points.

This method also provides a way to automatically determine the number of clusters based on standard statistics, taking outlier or noise into account. It therefore yields robust clustering methods.

• Constraint-based Method

In this method, the clustering is performed by the incorporation of user or application-oriented constraints. A constraint refers to the user expectation or the properties of desired clustering results. Constraints provide us with an interactive way of communication with the clustering process. Constraints can be specified by the user or the application requirement.

4. K-MEANS CLUSTERING ALGORITHM[14]

K-means clustering is a well known partitioning method. In this objects are classified as belonging to one of K-groups. The result of partitioning method is a set of K clusters, each object of data set belonging to one cluster. In each cluster there may be a centroid or a cluster representative. In case where it has consider real -valued data, arithmetic mean of attribute vectors for all objects within a cluster provides an appropriate representative; alternative types of centroid may be required within other cases.

4.1 Steps Of K-means Clustering Algorithm

K-Means Clustering algorithm is an idea, within which there is need to classify given data set into K clusters; value of K (Number of clusters) is defined by user which is fixed. In this first centroid of each cluster is selected for clustering & then according to chosen centriod, data points having minimum distance from given cluster, is assigned to that particular cluster[15]. Euclidean Distance is used for calculating distance of data point from particular centroid. This algorithm consists of four steps:

1. **Initialization:** In this first step data set, number of clusters & centroid that its defined for each cluster.
2. **Classification:** The distance is calculated for each data point from centroid & data point having minimum distance from centriod of a cluster is assigned to that particular cluster.
3. **Centroid Recalculation:** Clusters generated previously, centriod is again repeatly calculated means recalculation of centriod.
4. **Convergence Condition:** Some convergence conditions are given as below:

4.1 Stopping when reaching a given or defined number of iterations.

4.2 Stopping when there is no exchange of data points between clusters.

4.3 Stopping when a threshold value is achieved.

5. If all of above conditions are not satisfied, then go to step 2 & whole process repeat again, until given conditions are not satisfied.

Main advantages:

1. K-means clustering is very Fast, robust & easily understandable. If data set is well separated from each other data set, then it gives best results.
2. The clusters do not having overlapping character & are also non-hierarchical within nature.

Main disadvantages:

1. In this algorithm, complexity is more as compared to others.

2. Need of predefined cluster centers.

3. Handling any of empty Clusters: One more problems with K-means clustering is that empty clusters are generated during execution, if within case no data points are allocated to a cluster under consideration during assignment phase.

***k*-means algorithm**

MSE=largenumber;

Select initial cluster centroids {*mj*}*j*

k=1;

Do

OldMSE=MSE;

*MSE*1=0;

For *j*=1 to *k*

mj=0; *nj*=0;

endfor

For *i*=1 to *n*

For *j*=1 to *k*

Compute squared Euclidean

distance *d*2(*xi*, *mj*);

endfor

Find closest centroid *mj* to *xi*;

mj=*mj*+*xi*; *nj*=*nj*+1;

*MSE*1=*MSE*1+*d*2(*xi*, *mj*);

endfor

For *j*=1 to *k*

nj=max(*nj*, 1); *mj*=*mj*/*nj*;

endfor

*MSE=MSE*1;

while (*MSE*<*OldMSE*)

Complexity

As discussed before, k-means algorithm[13] converges to local minimum. Before k-means converges, centroids computed number of times, & all points are assigned to their nearest centroids, i.e., complete redistribution of points according to new centroids, this takes O(nkl), where n is number of points, k is number of clusters & l is number of iterations. In existing enhanced k-means algorithm, to obtain initial clusters, this process requires O(nk).

Here, some points remain within its cluster, others move to another cluster. If point stays within its cluster this require O(1), otherwise require O(k). If suppose that half points move from their clusters, this requires O(nk/2), since algorithm converges to local minimum, number of points moved from their clusters decreases within each iteration.

5. PROPOSED WORK

The problems with existing system were analysis[6], capture, search, sharing, storage, transfer, visualization, querying-updating. One more problems with K-means clustering is that empty clusters are generated during assignment phase. The proposed work is to eliminate limitations of K-mean clustering algorithm.

1. **Initialization:** In this first step data set, number of clusters & centroid should be calculated automatically according to size of data.
2. **Classification:** The distance is calculated for each data point from centroid & data point having minimum distance from centriod[13] of a cluster is assigned to that particular cluster.

3. **Centroid Recalculation:** Clusters generated previously, centriod is again repeatly calculated means recalculation of centriod.

4. **Convergence Condition:** Some convergence conditions are given as below:

4.1 Stopping when reaching a given or defined number of iterations.

4.2 Stopping when there is no exchange of data points between clusters.

4.3 Stopping when a threshold value is achieved.

5. If all of above conditions are not satisfied, then go to step 2 & whole process repeat again, until given conditions are not satisfied.

6. ***Elimination of Empty Clusters: Clusters*** *generated previously are rechecked*

Clusters where no data points are allocated to a cluster under consideration during assignment phase are eliminated.

Benefits of proposed Implementation over traditional

1) No need of predefined cluster center

2) There will be no Empty clusters at end

6. CONCLUSION AND FUTURE SCOPE

Clustering is process of grouping objects that belongs to the same class. Similar objects are grouped in one cluster and dissimilar objects are grouped in another cluster. Clustering analysis is used in several applications like market research, pattern recognition, Data analysis. K-means clustering[6] is very Fast, robust & easily understandable. If data set is separated from one other data set, then it gives best results. The clusters do not having overlapping character & are also non-hierarchical within nature. One more problems with K-means clustering is that empty clusters are generated during execution, if within case no data points are allocated to a cluster under consideration during assignment phase. In this the proposed work is to remove the empty cluster and done automatic clustering SCOPE OF RESEARCH The advent of laptops, palmtops, cell phones, & wearable computers is making ubiquitous access to huge quantity of information possible. Advanced analysis of data for extracting useful knowledge is next natural step within world of ubiquitous computing. Accessing & analyzing[14] data from a ubiquitous computing device offer many challenges. Visualizing patterns such as classifiers, clusters, associations & others, within portable devices are usually difficult. The small display areas offer serious challenges to interactive data mining environments. Data management within a mobile environment is also a challenging issue. Moreover, sociological & psychological[11] aspects of integration between data mining technology & our lifestyle are yet to be explored.

7. REFERENCES

[1] Piatetsky-Shapiro, Gregory (1991), Discovery, analysis, and presentation of strong rules, in Piatetsky-Shapiro, Gregory; and Frawley, William J.; eds., Knowledge Discovery in Databases, AAAI/MIT Press, Cambridge, MA.

[2] Agrawal, R.; Imieliński, T.; Swami, A. (1993). "Mining association rules between sets of items in large databases". Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93. p. 207. doi:10.1145/170035.170072. ISBN 0897915925.

[3] Hahsler, Michael (2005). "Introduction to arules – A computational environment for mining association rules and frequent item sets" (PDF). Journal of Statistical Software.

[4] Michael Hahsler (2015). A Probabilistic Comparison of Commonly Used Interest Measures for Association Rules. http://michael.hahsler.net/research/association_rules/measures.html

[5] Hipp, J.; Güntzer, U.; Nakhaeizadeh, G. (2000). "Algorithms for association rule mining --- a general survey and comparison". ACM SIGKDD Explorations Newsletter **2**: 58. doi:10.1145/360402.360421.

[6] Tan, Pang-Ning; Michael, Steinbach; Kumar, Vipin (2005). "Chapter 6. Association Analysis: Basic Concepts and Algorithms" (PDF). Introduction to Data Mining. Addison-Wesley. ISBN 0-321-32136-7.

[7] Pei, Jian; Han, Jiawei; and Lakshmanan, Laks V. S.; Mining frequent itemsets with convertible constraints, in Proceedings of the 17th International Conference on Data Engineering, April 2–6, 2001, Heidelberg, Germany, 2001, pages 433-442

[8] Agrawal, Rakesh; and Srikant, Ramakrishnan; Fast algorithms for mining association rules in large databases, in Bocca, Jorge B.; Jarke, Matthias; and Zaniolo, Carlo; editors, Proceedings of the 20th International Conference on Very Large Data Bases (VLDB), Santiago, Chile, September 1994, pages 487-499

[9] Zaki, M. J. (2000). "Scalable algorithms for association mining". IEEE Transactions on Knowledge and Data Engineering **12** (3): 372–390. doi:10.1109/69.846291.

[10] Hájek, Petr; Havel, Ivan; Chytil, Metoděj; The GUHA method of automatic hypotheses determination, Computing 1 (1966) 293-308

[11] Hájek, Petr; Feglar, Tomas; Rauch, Jan; and Coufal, David; The GUHA method, data preprocessing and mining, Database Support for Data Mining Applications, Springer, 2004, ISBN 978-3-540-22479-2

[12] Omiecinski, Edward R.; Alternative interest measures for mining associations in databases, IEEE Transactions on Knowledge and Data Engineering, 15(1):57-69, Jan/Feb 2003

[13] Aggarwal, Charu C.; and Yu, Philip S.; A new framework for itemset generation, in PODS 98, Symposium on Principles of Database Systems, Seattle, WA, USA, 1998, pages 18-24

[14] Brin, Sergey; Motwani, Rajeev; Ullman, Jeffrey D.; and Tsur, Shalom; Dynamic itemset counting and implication rules for market basket data, in SIGMOD 1997, Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD 1997), Tucson, Arizona, USA, May 1997, pp. 255-264

[15] Piatetsky-Shapiro, Gregory; Discovery, analysis, and presentation of strong rules, Knowledge Discovery in Databases, 1991, pp. 229-248