Application based, advantageous K-means Clustering Algorithm in Data Mining - A Review

Barkha Narang
Assistant Professor, JIMS, Delhi

Poonam Verma
Assistant Professor, JIMS, Delhi

Priya Kochar
Ex.Lecturer, GCW, Rohtak

Abstract : This paper has been written with the aim of giving a basic view on data mining. Various software's of data mining analyzes relationships and patterns in stored transaction data based on the user requirements. Several types of analytical software are available to mine useful data like statistical, machine learning, and neural networks. The four types of relationships sought using the analytical software's are classification, clustering, associations and finding patterns. In this paper we have discussed clustering and specifically K- means clustering technique. It is the most popular clustering technique with its own advantages and disadvantages. This paper focuses on the advantages in applications like market segmentation, in big data applications, real world problems like road detection, DNA gene expression and internet news group and elegant handling of continuous and noisy data. Different levels of analysis are available like genetic algorithms, artificial neural networks, decision trees, rule based induction methods and data visualization. K-means clustering has been integrated with these analytical tools as per the requirement of the application area.

Keywords: data mining, k-means clustering

I. INTRODUCTION

What is Data Mining?

Data mining, also called knowledge discovery or data discovery is the process of analyzing data from different perspectives and summarizing it into useful information. This information can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Data Mining Technology is going through continuous innovation. Companies have used powerful computers to shift through volumes of supermarket scanner data and analyze market research reports for years. However, continuous innovations in computer processing power, disk storage, and statistical software are dramatically increasing the accuracy of analysis while driving down the cost. For example, one Midwest grocery chain used the data mining capacity of Oracle software to analyze local buying patterns. They discovered that when men bought diapers on Thursdays and Saturdays, they also tended to buy beer. Further analysis showed that these shoppers typically did their weekly grocery shopping on Saturdays. On Thursdays, however, they only bought a few items. The retailer concluded that they purchased the beer to have it available for the upcoming weekend. The grocery chain could use this newly discovered information in various ways to increase revenue. For example, they could move the beer display closer to the diaper display. And, they could make sure beer and diapers were sold at full price on Thursdays.

Data Mining originates from the concepts of Data, Information, and Knowledge. The difference between them has lead to the discovery of data mining and data ware housing techniques. Also answering to the question on what can be done using data mining.The technological infrastructure required for data mining is based on two criterions that is size of the database and Query Complexity.With the advent of the social media, the data mining techniques are being used on every time high. These techniques are used for all size systems ranging from mainframes to hand held devices. There are two critical technological drivers:

- **Size of the database**: the more data being processed and maintained, the more powerful the system required.

- **Query complexity**: the more complex the queries and the greater the number of queries being processed, the more powerful the system required.

In conventional database technology, the indexing was sufficient. However with the larger amount of data and unstructured data , it becomes necessary to adopt some technologies that can handle such complications. Some vendors have added extensive indexing capabilities to improve query performance. Others use new hardware architectures such as Massively Parallel Processors (MPP) to achieve order-of-magnitude improvements in query time. For example, MPP systems from NCR link hundreds of high-speed Pentium processors to achieve performance levels exceeding those of the largest supercomputers.

II. LITERATURE REVIEW

Work has already been done by data classification method which integrates attribute oriented induction, relevance analysis, and the induction of decision trees. Such integration leads to efficient, high quality, multiple level classifications of large amounts of data, the relaxation of the requirement of perfect training sets, and the elegant handling of continuous and noisy data.

Historically as Wikipedia references, the terms of K-means clustering algorithm was first developed by J. Mac Queen (1967) and then the idea was followed by J. A. Hartigan and M.A. Wong around 1975. The standard algorithm as a technique for pulse-code modulation was proposed the by Stuart Lloyd in 1957, though it wasn't published until 1982. The consideration of K-means was demonstrated as early as 1956 by Steinhaus [15]. A simple local heuristic for problem was proposed in 1957 by Lloyds [16]. The method represents that first step choosing k arbitrarily point as facilities. In each stage, assign each point X into cluster with closest facility and then computes the centre of mass for each cluster. These centre's of mass become the new facilities for the next phase, and the process repeats until the solution stabilizes. [17, 18] In aspect of how the neighbours are computed for each centre there exists some different procedures for performing K-means: [22] • Lloyd's: Repeatedly applies Lloyd's algorithm with randomly sampled starting points. • Swap: A local search heuristic, which works by performing swaps between existing centres and a set of candidate centres. This algorithm iteratively changes centres by performing swaps. Each run consists of a given number (max swaps) executions of the swap heuristic. Elham Karoussi Data Mining, K-Clustering Problem 26 • Hybrid: A more complex hybrid of Lloyd's and Swap, which performs some number of swaps followed by some number of iterations of Lloyd's algorithm. To avoid getting trapped in local minima, an approach similar to simulated annealing is included as well. • EZ Hybrid: A simple hybrid algorithm, which does one swap followed by some number of iterations of Lloyd's.

How does data mining work?

Large-scale information technology has been evolving separate transaction systems and separate analytical systems. Data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. Generally, the following four types of relationships are sought using the analytical softwares:

- **Classes**: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.

- **Clusters**: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.

- **Associations**: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.

- **Sequential patterns**: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

Data mining consists of five major elements:

- Extract, transform, and load transaction data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

Different levels of analysis are available:

- **Artificial neural networks**: They are non-linear predictive models that learn through training. ANN resembles biological neural networks in structure.

- **Genetic algorithms**: GA s are optimization techniques. They use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.

- **Decision trees**: They are tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that one can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.

- **Nearest neighbor method**: A technique that classifies each record in a dataset based on a combination of the classes of the *k* record(s) most similar to it in a historical dataset (where *k* 1). Sometimes called the *k*-nearest neighbor technique.

- **Rule induction**: The extraction of useful if-then rules from data based on statistical significance.

- **Data visualization**: The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

What is Clustering?

Clustering is organizing data into such groups called clusters in which there is high intra-cluster similarity. Informally we can say that there is low inter-cluster similarity and also the aim of clustering is to find natural groupings among objects. In clustering, one is given a set of recorded historical transactions which tells various patterns on which customer bought what combinations of items. By using the right /appropriate clustering technique, one can tell the segmentation of ones customers. Clustering is actually called unsupervised learning. **I**f this question is asked to a data mining or machine learning person, they will use the term supervised learning and unsupervised learning to explain the difference between clustering and classification. So what is unsupervised learning? Let us suppose that one has a basket and it is full of fresh fruits. The task is to arrange the same type of fruits at one place. This time one doesnt know any thing about the fruits, one is seeing these fruits for the first time. So the task is how one will arrange the same type of fruits. One will first take on a fruit and one will select any physical character of that particular fruit. Example color. Then will arrange them based on the color, then the groups will be some thing like this. Red Color Group: apples & cherry fruits. Green Color Group: bananas & grapes. So now one will take another physical character as size, the groups will be some thing like this. Red Color and Big Size: apple. Red Color and Small Size: cherry fruits. Green Color and Big Size: bananas. Green Color and Small Size: grapes. Hence work done. In this case one didn't learn anything before, meaning there was no training data and no response variables available. This type of learning is known unsupervised learning. Clustering comes under unsupervised learning.

Clustering Definition

Cluster analysis or clustering is said to be a collection of objects. Clustering [1] tries to group a set of objects and find whether there is *some* relationship between the objects. In the context of machine learning clustering is *unsupervised learning*.

It is used in various applications in the real world. Such as data/text mining, voice mining, image processing, web mining and so on. It is important in real world in certain fields. How and why is important in the real world and how were the techniques implemented in several applications are presented.

Now the question arises that why should one want to do it and what are its real applications.

Why clustering?

- It helps in organizing huge, voluminous data into clusters which shows internal structure of the data. Example clustering of genes.
- Sometimes, the goal of clustering is partitioning of data. Example in Market segmentation.
- After clustering the data is ready to be used for other AI techniques. Example: Summarization of news wherein we group data into clusters and then find centroid.
- Techniques for clustering are useful in knowledge discovery in data. Example underlying rules, reoccurring patterns, topics, etc.

Existing Algorithms of Clustering

Various clustering algorithms exist like K-means, Clustering by Vertex Density in a Graph, Clustering by Ant Colony Optimization, A Dynamic Cluster Algorithm Based on L r Distances for Quantitative Data, The Last Step of a New Divisive Monothetic Clustering Method: the Gluing-Back Criterion [7}

K-means clustering

K-means clustering [10] also called Lloyd's algorithm in the computer science community, is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. The problem is computationally difficult (NP-hard); however, k-means clustering tends to find clusters of comparable spatial extent.

III. THE ALGORITHM

It is an iterative refinement technique. Due to its ubiquity it is often called the k-means algorithm.

Given an initial set of k means m_1 (1),…,m_k(1) , the algorithm proceeds by alternating between two steps:

Assignment step: Assign each observation to the cluster whose mean yields the least within-cluster sum of squares (WCSS). Since the sum of squares is the squared Euclidean distance, this is intuitively the "nearest" mean. (Mathematically, this means partitioning the observations according to the Voronoi diagram generated by the means), where each is assigned to exactly one, even if it could be assigned to two or more of them.

Update step: Calculate the new means to be the centroids of the observations in the new clusters.

Since the arithmetic mean is a least-squares estimator, this also minimizes the within-cluster sum of squares (WCSS) objective.

The algorithm has converged when the assignments no longer change. Since both steps optimize the WCSS objective, and there only exists a finite number of such partitioning's, the algorithm must converge to a (local) optimum. There is no guarantee that the global optimum is found using this algorithm.

The algorithm is often presented as assigning objects to the nearest cluster by distance. The standard algorithm aims at minimizing the WCSS objective, and thus assigns by "least sum of squares", which is exactly equivalent to assigning by the smallest Euclidean distance. Using a different distance function other than (squared) Euclidean distance, may stop the algorithm from converging. Various modifications of k-means such as spherical k-means and k-medoids have been proposed to allow using other distance measures.

IV. Applications of Clustering

The applications [2] of clustering usually deal with large datasets and data with many attributes. Data Mining explores such data. Following are the various applications where in we concentrate on advantages of k-means clustering algorithms in data mining:

1. Big Data applications place special requirements on clustering algorithms such as the ability to find clusters embedded in subspaces of maximum dimensional data, finding the answers to scalability of data, increasing end-user comprehensibility to the results, non-presumption of any canonical data distribution, and insensitivity to the order of input records. CLIQUE [3], is a clustering algorithm that satisfies all the above requirements. CLIQUE identifies dense clusters in high dimensional subspaces. It generates cluster descriptions in the form of expressions that are minimized for ease of understanding. It produces identical results irrespective of the order in which input records are presented and does not presume any specific mathematical form for data distribution. Experiments have shown that CLIQUE efficiently finds accurate cluster in large high dimensional datasets.

2. (a) A lot of work has been done by integrating[4]the self-organizing feature maps and ***K***-means algorithm for market segmentation. Cluster analysis is a common tool for market segmentation. Conventional research usually employs the multivariate analysis procedures. In recent years, due to the high performance of artificial neural networks in engineering, they are also being applied in the area of management. The two-stage method is a combination of the self-organizing feature maps and the K-means method. After using this method on the basis of Wilk's Lambda and discriminant analysis on the real world data and on the simulation data, results indicate that high rates of misclassification are decreased.
2. (b) As already discussed that the general idea of segmentation, or clustering, is to group items that are similar. A commonly used method is the multivariate analysis. These methods consist of hierarchical methods, like Ward's minimum variance method, and the non-hierarchical methods, such as the *K*-means method. Owing to increase in computer power and decrease in computer costs, artificial neural networks (ANNs), which are distributed and parallel information processing systems successfully applied in the area of engineering. Study has been done to integrate ANN and multivariate analysis to solve marketing problems. The method is a two-stage method, which first uses the self-organizing feature maps to determine the number of clusters and the starting point and then employs the *K*-means method to find the final solution. The above provides the marketing analysts a more sophisticated way to analyze the consumer behavior and determine the marking strategy.

2. (c).The Internet[5] is emerging as a new marketing channel, so understanding the characteristics of online customers' needs and expectations is considered a prerequisite for activating the consumer-oriented electronic commerce market. In this study, clustering algorithm based on genetic algorithms (GAs) are used to segment the online shopping market. Genetic Algorithms are believed to be effective on NP-complete global optimization problems, and they can provide good near-optimal solutions in reasonable time. Thus, clustering technique with GA provides a way of finding the relevant clusters more effectively. Research includes the initial optimization using genetic algorithms and then the K-means clustering which is called GAK-means, to a real-world online shopping market segmentation case. The results showed that GA K- means clustering improves segmentation performance in comparison to other typical clustering algorithms. Studies demonstrate the validity of the above method as a preprocessing tool for recommendation systems.

3. Clustering [11] also has an advantage in cases where information about the problem domain is available in addition to the data instances themselves. The k-means clustering algorithm can be modified to make use of this

information. Observations show improvement in clustering accuracy. We apply this method to the real-world problem of automatically detecting road lanes from GPS data and dramatic increase in performance has been found.

4. It has been found using K-means clustering via principal component analysis [6] that unsupervised dimension reduction is closely related to unsupervised learning.. DNA gene expression and Internet newsgroups are analyzed. Principal component analysis (PCA) is a widely used statistical technique for unsupervised dimension reduction. *K*-means clustering is a commonly used data clustering for performing unsupervised learning tasks. It has been proved that principal components are the continuous solutions to the discrete cluster membership indicators for *K*-means clustering. New lower bounds for *K*-means objective function are derived, which is the total variance minus the eigen values of the data covariance matrix. Several implications are discussed. On dimension reduction, the result provides new insights to the observed effectiveness of PCA-based data reductions, beyond the conventional noise-reduction explanation that PCA, via singular value decomposition, provides the best low-dimensional linear approximation of the data. On learning, the results suggest effective techniques for *K*-means data clustering. Experiments indicate that the new bounds are within 0.5-1.5% of the optimal values.

5. Generalization and decision tree induction, integrated with efficient clustering in data mining [8] helps to the relaxation of the requirement of perfect training sets, and leads to the elegant handling of continuous and noisy data. Several emerging applications in information-providing services, such as data warehousing and online services over the internet, call for various data mining techniques in order to understand user behavior, to improve the service provided and to increase business opportunities. To fulfill such a demand, the data mining techniques integrate attribute oriented induction, relevance analysis, and the induction of decision trees. Such integration leads to efficient, high quality clustering of large amounts of data.

V. CONCLUSIONS AND FUTURE DIRECTIONS

The conventional techniques of classification and clustering are little too outdated for the large amount of data being faced today. In today's time, where the data of 4.5 TB are being generated everyday, it becomes a necessity to find a method that can help in data mining embedded with a Artificial intelligence concept to ease the task with more accuracy.

A lot of work is directed towards classification of the available data mining techniques, comparative study of such techniques, etc. We are looking forward to take this review paper by bringing to notice the loop holes of K- means clustering algorithms and suggesting improvements in K-means clustering and other clustering algorithms.

REFERENCES

[1] http://stackoverflow.com/questions/5064928/difference-between-classification-and-clustering-in-data-mining

[2] Volume 27 Issue 2, June 1998 ,Pages 94-105 ACMNew York, NY, USA.(Springer, Chapter Grouping Multidimensional Data, pp 25-71).

[3] Rakesh, Johannes, Dimitrios, Prabhakar (June 1998) Volume 27 Issue 2, Pages 94-1058, Automatic subspace clustering of high dimensional data for data mining Applications . Proceedings of the 1998 ACM SIGMOD international conference on management of data , New York.

[4] R.J. Kuo, L.M. Ho , C.M. Hu (Sept 2002) Volume 29, Issue 11,Volume 29, Issue 11 Pages 1475–1493,Elsevier Computers & Operations Research, , September 2002, Integration of self-organizing feature map and *K*-means algorithm for market segmentation

[5] Kyoung-jae Kim , HyunchulAhn (February 2008) Volume 34, Issue 2, Pages 1200–1209 Expert Systems with Applications, A recommender system using GA *K*-means clustering in an online shopping market.

[6] Chris Ding, XiaofengHe, New York, NY, USA ©2004 *K*-means clustering via principal component analysis, ICML 04 Proceedings of 21st international conference on Machine Learning.

[7] Ming-Syan Chen, Dept. of Electr. Eng., Nat. Taiwan Univ., Taipei, Taiwan ;Jiawei Han. IEEE Data mining: an overview from a database perspective.

[8] Banks, D., House, L., Mc Morris, F.R., Arabie, P., Gaul, W. (Eds.) 15–18 July 2004, Classification, Clustering, and Data Mining

[9] M. Kamber, Burnaby, BC Canada ;L. Winstone ,Wan Gong , Shan Cheng. IEEE Generalization and decision tree induction: efficient classification in data mining .

[10] means clustering, from Wikipedia, the free encyclopedia

[11] KiriWagsta, Claire Cardie,Seth Rogers, Stefan Schroedl, page 577-584,Constrained K-means Clustering with Background Knowledge, Proceedings of the Eighteenth International Conference on Machine Learning, 2001,