Clustering Based Approach in Data Mining: A Review

Jyoti Sinha  
M. Tech Scholar  
Department of Computer Science & Engineering  
DIMAT, Raipur

Mr. Deepak Shrivastava  
Assistant Professor  
Department of Computer Science & Engineering  
DIMAT, Raipur

Abstract

The process of generating some useful information from a huge volume of data is called learning. There are two types of learning i.e. supervised and unsupervised. Clustering is a type of unsupervised learning. Clustering is the core of any data mining application. This paper presents a brief description of different clustering algorithm as well as the past work presented by different authors in clustering method. In this paper, a general working of the popular existing clustering algorithm is presented followed by some noteworthy contribution.

Keywords: Clustering, Supervised Learning, Unsupervised Learning, Hierarchical

I. INTRODUCTION

With the advancement of digital technology and introduction of internet, it has become possible to store and distribute the large amount of data. In most of the cases these data are in the form of some features of the particular entity. For example so many databases are available for biometric such as fingerprint, Iris, palm print etc. Database is generally prepared to analyze the data. One of the important tasks for these data is to group or classify them in to different categories or in other words in different cluster.

Clustering is an operation which is performed to group the similar object in similar categories. The main problem in clustering is to find group of similar objects in the database. In clustering similarity function is used for finding the similarity. Clustering is one of the vital tools for organizing and managing the documents for improving the document retrieval and browsing. Clustering is often misinterpreted as the classification but these are different from each other in some way. Classification operation, assign each object to any one of the predefined classes while in clustering classes are also required to be defined. Data clustering is also very useful because it improve the efficiency of the database system by minimizing the number of disk access. Clustering operation group the object of similar properties together in one class which enable the system to access the class in single attempt. Biomedical, insurance, marketing libraries, web content are some of the application areas of the data clustering [1][2].

Some of the essential requirement of clustering is as follows:-

1) Scalability- Scalability of the clustering algorithm refers to the number of data it can handle accurately. Most of the clustering based algorithm work very well for small number of data objects (less than 200) but if we increase the size of database then it start showing some inaccuracy.

2) Ability to handle different types of attributes-

3) Most of the clustering based methods are designed to deal with interval based data (Numerical data),but in some application it is required to cluster the data of different kind like binary, categorical(Nominal), ordinal data or mixture of both.

4) Ability to handle Arbitrary shape Cluster- Most of the cluster based approach obtain the cluster which is based on the Euclidean Distance or Manhattan Distance measure. Such kind of cluster based approach eventually find spherical shape cluster with similar size and density. It is very important for the cluster based approach that it must be able to detect the cluster of arbitrary shape.

5) Least requirement of input parameters: Most of the clustering algorithm requires some requirement of the input to be supplied from the user side before clustering (like number of desired cluster). The result of the clustering operation is very sensitive to the supplied input parameters.

6) Ability to deal with the noisy data: Most of the real word database contains some missing, erroneous and unknown data. Most of the clustering method are very sensitive to such noisy data and may lead to inaccurate cluster.

7) Insensitivities towards the order of input record- It was found that some of the clustering method are very sensitive to the order of the input and changing the order of the input may generate the different results which is undesirable. so a clustering algorithm must be insensitive toward the order of the input.

8) High dimensionality –Generally, A database or a data warehouse may contain more than one dimensions or attributes. Many clustering algorithms are good at dealing with low-dimensional data, i.e. only two to three dimensions. Human eyes are capable of judging the quality of clustering for up to three dimensions.

9) Constraint-based clustering- Real-Time applications may need to perform clustering under different kinds of constraints. Suppose that your job is to choose the locations for a given number of new automatic cash-dispensing machines (ATMs) in a
To decide upon this, we may cluster household while considering constraints such as the city rivers, highway networks and requirements of customer per region.

Interpretability and usability of the result: Users generally expect results of clustering to be interpretable, comprehensible, and usable. That is, clustering may require to be tied up with particular semantic interpretations and applications. It is important to know that how an applications aim may influence the selection of clustering methods [3].

II. OVERVIEW OF CLUSTERING

Clustering basically divides the data in to groups of similar objects [3]. In broader sense, clustering algorithm can be categorized as:
1) Hierarchical Clustering
2) Spectral clustering
3) Grid-based Clustering
4) Density based clustering
5) Partition Clustering

Next section describes the above mentioned clustering algorithm in brief.

A. Hierarchical Clustering Algorithm

This algorithm divides the data object in such a way that it form the tree shaped structure. In a broader sense, it can be categorized as agglomerative hierarchical clustering and Divisive Hierarchical clustering. Agglomerative clustering which is also known as the bottom up approach, consider each data point to be a separate cluster which is merged on each iteration using some criteria. Single link, multiple link, centroid or wards method are used for merging operation. 

Divisive Hierarchical clustering is also known as the top down approach which consider all the data points as the single cluster and then these are divided in to a number of cluster based on some certain criterion

LEGCLUST [23], BRICH (Balance iterative Reducing and Clustering Using Hierarchies)[20], CURE(Cluster using Representative)[21] and Chemeleon[1] some example of this type of clustering.

B. Spectral Clustering Algorithm

A technique which relies on the eigen structure of the similarity matrix is known as the spectral clustering. This technique form the cluster by partitioning the data using similarity matrix. There are three main stages in any spectral clustering algorithm [24]. They are as follows-

1) Pre-processing- This stage construct the similarity Matrix.
2) Spectral Mapping- This stage constructs the eigen vector for similarity matrix.
3) Post Processing- This stage is responsible for grouping the data.

Some of the advantage of this algorithm is as-

1) Simple to implement
2) Objective function does not consider local optima.
3) Faster approach
4) Statistically much consistent
5) Assumption on cluster shape is not important.

One of the major down point of this method is that it tends to be complex for larger database. Algorithm proposed by Shi and Malik, algorithm proposed by Kannan Vempala and Vetta, Algorithm proposed by Ng, Jordan and Weiss[23] are some of the example of this algorithm.

C. Grid Based Clustering Algorithm.

In grid based algorithm, object space is quantized to finite number of cells which forms the grid structure[1]. All the operation is then done on this grid. Low processing time is the advantage of this method. Complexity of the cluster depends on the number of populated grid cells rather than on the number of objects in the dataset.

Major features of this algorithm are as follows-

1) Distance computation is not required.
2) Clustering operation is performed on summarized dataset.
3) Shapes are confined to union of grid-cells.
4) Algorithm complexity is generally given by O(Number of populated grid cells)

STING [1] is one of the examples of this kind of algorithm.

D. Density Based Clustering Algorithm

In density based clustering algorithm, cluster is grown as long as the density around the neighbourhood exceeds some threshold [1]. This algorithm is good for noisy dataset as it can handle the noise effectively.
Some of the important features of this algorithm.

1) It can handle the cluster of arbitrary shape.
2) It can handle noise effectively.
3) Only one scan of input dataset is required.
4) Only density parameter required to be initialized.

DBSCAN, DENCLUE and OPTICS [1] are some of the example of this algorithm.

E. Partition Clustering Algorithm

In this method, the data is split in to K-partition where each partition represents a cluster. The partition is done on the basis of some objective function.

One of the criterions is to minimize the square error criterion-

\[ E = \sum \sum ||p - m_i||^2 \]

Here p represents the points in cluster and m represents the mean of the cluster.

In this algorithm, two properties must be satisfied by the cluster-

1) Each group must have at least one object.
2) Each object must be the member of exactly one group.

One of the drawback of this algorithm[3] is that whenever the any point become closer to the centre of another cluster then it tend to give poor results due to overlapping of data-points.

III. NOTEWORTHY CONTRIBUTION

In the year 1975, Holland [8] first proposed genetic Algorithm in his paper “Adaptation in Natural and Artificial Systems”. Genetic algorithm is based upon the survival of the fittest rule proposed by the Charles Darwin. On the basis of the survival and then reproduction carried out by the fittest, genetic algorithm generate a encoded chromosome and then compute the fitness of these chromosomes. In the next phase of the algorithm, crossover and mutation operation is performed on each chromosome to generate next generation. This process of evolution carries out until the stopping criterion is achieved. Genetic algorithm is appropriate for large size, non-linear database in which the solution is unpredictable. Since this algorithm relies on the multi-point search algorithm, it always converges to the global optimal solution. Since Genetic algorithm is good for searching, it is generally used to solve the clustering based problems.

In the year 1996, Murthy and Chowdhury in his paper [9] titled “In search of optimal clusters using genetic algorithms,” proposed a combination of Genetic algorithm and K-mean algorithm for solving the clustering based problems. In this paper, genetic algorithm is used for optimizing the specified objective function which is related to the clustering based problems. Several experiments have been carried out in synthetic and real life data set using proposed methods. Though K-means is one of the most popular method which is used for solving the cluster based problems. Incorporating genetic algorithm in K-means, improve the results significantly which is confirmed by the results published in this paper.


Though K-means clustering is very popular in general clustering, K-means clustering has three major disadvantages

1) Poor computational efficiency
2) Number of clusters is required to be supplied by the user.
3) Search may fall to local minima.

This paper addressed the above mentioned problems and proposed a solution for the first two problems and partial solution to the third problem.

In this paper a new algorithm has been presented which first search the cluster space location along with the number of clusters for optimizing the Bayesian Information criterion or Akaike information criterion. In this algorithm two different ways of class refinement method of k-means is used which was based on statistics and very efficient computationally. This gives a very fast and accurate number of classes and other parameters. Simulation results obtained in this paper shows remarkable improvement in selecting the true classes and speed of computation as compared to the simple K-means algorithm.

In the year 2003, Xi Xiao in his paper [11] titled “Gene clustering using self-organizing maps and particle swarm optimization,” proposed a method to cluster the genes in different groups by using hybrid method of genetic algorithm(GA) and Particle swarm optimization(PSO). Clustering of gene is basically a process of making a group of related genes. It is very useful for different genomic studies which are accomplished to analyze the function of genes.

Hybrid method in this paper is used to reduce the surface mount technology (SMT) set-up time. Simulation results of this method after applying on well-known dataset of Iris, Glass, Vowel and Wine shows that the performance of this method is better than the alone Genetic based approach or Particle swarm optimization(PSO) method. In this method, order clustering Schedule the orders that belong to the same cluster and hence reduce the production time as well as machine idle time.

In the year 2003, D.W. van der Merwe, A.P. Engelbrecht [12] in their paper titled “Data clustering using particle swarm optimization” presented two new methods of using PSO( Particle swarm Optimization) for data clustering. In this paper they showed that the Particle swarm optimization (PSO) can be used to find the centroids of a clusters specified by the user. K-means
Clustering is used in PSO for providing the seed for initial swarm. On the other hand, second method proposed by the authors basically refines the cluster formed by the K-means method.

In this paper, both the PSO based methods are tested on six different datasets and comparison is also carried out with pure k-means based method. Results clearly shows that Both the PSO (particle swarm optimization) outperform the the results obtained by the pure K-means based method.

In the year 2003, Xiao, X., et.al. in his paper[13] titled “Gene clustering using self-organizing maps and particle swarm optimization “ presented an approach to cluster the similar genes in the same group Gene clustering, the process of grouping related genes in the same cluster, is at the foundation of different genomic studies that aim at analyzing the function of genes. Microarray technologies have made it possible to measure gene expression levels for thousands of genes simultaneously. For knowledge to be extracted from the datasets generated by these technologies, the datasets have to be presented to a scientist in a meaningful way. Gene clustering methods serve this purpose. In this paper, a hybrid clustering approach that is based on self-organizing maps and particle swarm optimization is proposed. In the proposed algorithm, the rate of convergence is improved by adding a conscience factor to the self-organizing maps algorithm. The robustness of the result is measured by using a resampling technique. The algorithm is implemented on a cluster of workstations.


Clustering Algorithm” proposed combination of genetic algorithm and k-means based clustering method which is faster than its predecessor and hence is known as the FGKA (Fast Genetic-K-means Algorithm). Basically FGKA is a faster form of GKA (Genetic-k-means algorithm). Some of the advantage of this method over GKA is efficient estimation of the object function value TWCV (Total within-cluster variation), efficient mutation operation . With the help of these improvements, FGKA has execution time 20 times faster than GKA. Though the performance of FGKA is much better than the GKA it has one of the major disadvantage is that for small mutation probability, cost of computing centroid and TWCV is very expensive.

In the year 2004, Lu et al in his paper [15] titled “Incremental genetic K-means algorithm and its application in gene expression data analysis” presented an algorithm to overcome the problem faced in FGKA (Fast genetic-K-means algorithm) this method is known as the incremental genetic k-means algorithm(IGKA). IGKA has all the advantages of the FGKA like global optimum convergence etc. this method (IGKA) outperform the previous method (FGKA) when the mutation probability is small. Main concept of IGKA (Incremental genetic k-means algorithm) is to compute the objective function value Total within cluster variation(TWCV) first and then to make a cluster centroids by incremental step. Though IGKA perform better than the FGKA in case of small mutation probability but fail to perform better in case of mutation probability is greater than the predefined threshold. This is one of the disadvantages of the IGKA that it is not useful for larger mutation probability and FGKA is not useful for smaller mutation probability.

In the year 2005, Cui, X et.al in his paper [16] titled “Document clustering using particle swarm optimization “ presented a PSO (Particle swarm optimization ) based document clustering algorithm. In any document clustering algorithm, quality and speed is of much importance as these parameters play important role in effective organization, navigation and summarization of the documents. Recent research revealed that partition clustering gives better results for clustering large database. But one of the most commonly used partition based clustering method k-means clustering tend to produce the local optimal solution which is undesirable. Therefore in this paper PSO based approach is adopted which tend to search globally rather than the locally as happen in the k-means clustering. In this paper PSO, K-means and hybrid method of clustering is applied to the four different text document dataset. Number of documents in the dataset is from 204 to more than 800 and the number of terms are from 5000 to over 7000. Obtained results revealed that the performance of the hybrid approach is better than rest of the two. It is able to generate more compact clustering result.

In the year 2006, M.G.H. Omran et.al [17] in his paper titled “Dynamic clustering using particle swarm optimization with application in image segmentation” presented a dynamic clustering approach based on particle swarm optimization (DCPSO). In this paper this algorithm is applied to classify the region of the image in unsupervised way. This approach automatically find out the optimum number of clusters and at the same time cluster the data in their respective cluster. This algorithm starts by dividing the datasets in to a large number of clusters to reduce the effect of number of cluster initialization. In this algorithm, binary based PSO (particle swarm optimization) is applied for selecting the best number of cluster. K-means clustering is used then to refine the centroid of the clusters. Simulation results reveal that the proposed approach is able to achieve the optimum number of clusters. Proposed method is applied to the natural images (MRI and satellite images) and the obtain result is compared with the result obtained by other unsupervised clustering method and it was found that the proposed method outperform the existing method in finding out the optimum number of clusters.

In the year 2009, H. Abolhassani, M. Mahdavi in their paper [18] titled “Harmony K-means algorithm for document clustering” presented harmony k-means based algorithm for document clustering. In this method harmony search optimization is used which earlier showed its usefulness in Markov chain theory. Being a Meta heuristic algorithm, it is able to achieve global optimum. Some of the advantages of this algorithm are

1) It requires few mathematical operations.
2) Stochastic random search is used in this algorithm.
3) Continuous variable can be used without any loss of precision.
4) It does not perform encoding and decoding in decision variable.
In the year 2010, Dharmendra K Roy and Lokesh K Sharma presented a paper [19] titled “Genetic K-Means clustering Algorithmfor mixed numeric and categorical data.” Which is the combination of both genetic and k-means algorithm.

In reality, most of the dataset contain both numeric and categorical type data. In such of type of dataset it is very challenging to divide this kind of mix dataset in to different clusters. In this paper modified genetic k-means algorithm is used to find the global optimal partition of the mixed dataset containing both numerical and categorical dataset.

In this algorithm, genetic algorithm is included in the k-means based clustering with enhance cost function for handling the categorical data. Simulation results obtained in this algorithm confirms that this method effectively recover the cluster structure from the categorical dataset.

In this algorithm, cluster center is represented in a modified way which is able to capture the characteristics very effectively because of having distribution pattern of all categorical data. Some of the special feature of this algorithms are as follows:

1) Efficient Computation of TWCVs.
2) Avoidance of illegal string elimination overhead.
3) Simplification of mutation operation.

In the year 2010, Feng et al in his paper [20] title “A fast divisive clustering algorithm using an improved discrete particle swarm optimizer” suggested a clustering algorithm by using particle swarm optimization. Being an important techniques, clustering is being used in many application areas like segmentation of the image, clustering of the documents and vector quantization. Divisive clustering, being a branch of Hierarchical clustering many application in clustering method because of being computation efficiency.

Two major principles which need to be taken account during divisive algorithm are as follows-

1) To decide which cluster need to be split.
2) To decide how to split the selected cluster.

One of the disadvantages of the divisive algorithm is that as compared to the partition algorithm its performance is less. Due to this it is very difficult to set the trade-off between computational time and clustering performance. To overcome this difficulty, in this paper a novel algorithm has been presented by incorporating the improved discrete particle swarm optimization techniques to a divisive clustering frame work. Exhaustive experiments have been carried out by taking synthetic as well as real world application. The results obtained are very promising and encouraging. For most of the database, the performance of this algorithm is better or comparable to the then existing algorithm. One of the high point of this algorithm is that it is much faster than the previous one and the then existing techniques.

In the year 2011 Senthilnath, J., Omkar, S. N., & Mani in their paper [21] titled “Clustering using firefly algorithm: performance study” suggested a clustering method which is based on the firefly algorithm (FA). Firefly is an optimization technique which is inspired by the nature. It basically copies the flash pattern and properties of the firefly. While clustering is basically a data analysis technique which groups the similar object in their respective cluster. In this paper, authors used FA(Firefly) for benchmark and compared the performance of this method with two more nature inspired techniques such as Artificial Bee Colony(ABC), Particle Swarm Optimization (PSO) and nine other different techniques. As many as thirteen different benchmark datasets from UCI machine learning repository has been taken to check the performance. Simulation result of this paper claimed that the performance of Firefly is sufficient enough to be used for clustering based problems.

In the year 2012, Yixian Yang et.al. In his paper [22] titled “Data clustering using bacterial foraging optimization “ proposed a new algorithm for clustering based problems. This method is based on the Bacterial Foraging (BF). This is also an optimization technique in which the group of bacteria forage to converge towards the certain solution by making the fitness function minimized. In order to test the algorithm, it is tested on some well-known benchmark datasets. The performance of this method is also compared with the performance of the other popular clustering method like K-means, Ant colony optimization (ACO) and Particle swarm optimization (PSO). Results shows that the performance of this method is also an effective method of clustering which can handle dataset of different cluster size, densities and dimensions.

In the year 2012, Mahendiran, in his paper [23] titled “Implementation of K-means clustering in cloud computing environment” suggested that k-means clustering can also be used in cloud computing environment. a huge volume of business data is stored in cloud data centers with low cost. In business organization, both data mining and cloud computing techniques are used to achieve maximum cost and minimum cost in several ways. In this paper k-means algorithm is used in google cloud by incorporating Google app engine and Cloud SQL. In this work, k-means algorithm is used for mining the large database and cloud computing is used to provide the solution of storing the large database with least cost.

IV. CONCLUSION

In this paper, a general method of clustering algorithm is discussed along with its advantages and disadvantages. Problem with k-means clustering is to find the optimal value for k and selection on initial centroid for each cluster. Some of the noteworthy contribution from different authors have also been presented and discussed.

REFERENCES

[1] Jiawei Han, Micheline Kamber, “Data Mining Concepts and Techniques” Elsevier Publication.


