An Efficient Hybrid Genetic Fuzzy C-Means Algorithm on Data Mining Techniques

Priya Arjunan[1] Sasikala R[2]
[1,2]Department of Computer Science
[1,2]Sankara College of Science and Commerce

Abstract— The amount of data in world is growing day by day because the use of internet, smart phones and social networks. Big data is a collection of data sets which is very large in size as well as complex. Traditional database systems are not able to capture, store and analyze this large amount of data. In this research, proposed an algorithm for the clustering problem of big data using a combination of the genetic algorithm with the K-Means algorithm. The main idea behind this algorithm is to combines the advantage of Genetic algorithm and K-means to process large amount of data.

Key words: Data Mining, Clustering, K-Means, Genetic K-Means

I. Introduction

Data mining is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. It has been defined as the automated analysis of large or complex data sets in order to discover significant patterns or trends that would otherwise go unrecognized. The goal of data mining is to unearth relationships in data that may provide useful insights. Data mining tools can sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions, performance bottlenecks in a network system and identifying anomalous data that could represent data entry keying errors. The ultimate significance of these patterns will be assessed by a domain expert - a marketing manager or network supervisor - so the results must be presented in a way that human experts can understand. Data mining tools can also automate the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data quickly. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events. Data mining techniques can yield the benefits of automation on existing software and hardware platforms to enhance the value of existing information resources, and can be implemented on new products and systems as they are brought on-line. When implemented on high performance client/server or parallel processing systems, they can analyse massive databases to deliver answers to questions

II. Review of Literature

A. A Survey on Various Clustering Techniques with K-means Clustering Algorithm in Detail

Clustering is the division of data into groups of similar objects. In clustering, some details are disregarded in exchange for data simplification. Clustering can be viewed as a data modeling technique that provides for concise summaries of the data. Clustering is therefore related to many disciplines and plays an important role in a broad range of applications. The applications of clustering usually deal with large datasets and data with many attributes. Exploration of such data is a subject of data mining. This survey concentrates on clustering algorithms from a data mining perspective with K means Clustering algorithm.

B. Global K-Means (GKM) Clustering Algorithm: A Survey

K-means clustering is a popular clustering algorithm but is having some problems as initial conditions and it will fuse in local minima. A method was proposed to overcome this problem known as Global K-Means clustering algorithm (GKM). This algorithm has excellent skill to reduce the computational load without significantly affecting the solution quality. We studied GKM and its variants and presents a survey with critical analysis. We also proposed a new concept of Faster Global K-means algorithms for Streamed Data sets (FGKM-SD). FGKM-SD improves the efficiency of clustering and will take low time & storage space.

C. Survey on Various Enhanced K-Means Algorithms

Data Mining is defined as a technique used to extract and mine the invisible, meaningful information from mountain of data. Clustering is an important technique that has been introduced in the area of data mining. Clustering is defined as a method used to group similar data into a set of clusters based on some common characteristics. K-means is one of the popular partitional based clustering algorithms in the area of research. The impact factor of k-means is its simplicity, high efficiency and scalability. However, is also comprises of number of limitations: random selection of initial centroids, number of cluster K need to be initialized and influence by outliers. In view of these deficiencies, this paper presents a survey of improvements done to traditional k-means to handle such limitations.

D. A Survey on Clustering Algorithms and K-Means

The overall goal of data mining process is to extract information from a large data set and transfer it into an understandable form for future use .Clustering is important in data analysis and data mining applications. Clustering is a division of data into group of similar objects. Each group called a cluster consists of objects that are similar between themselves and dissimilar between compare to objects of other groups. This paper is aimed to study of all the clustering algorithms. In this paper we compare all types of clustering methods and gave a brief knowledge about k-means clustering.

E. *Survey on Different Enhanced K-Means Clustering Algorithm*

Data Mining is justify technique used to extract, meaning ful information from mountain of data and Clustering is an important task in Data Mining process which can be used for the purpose to make groups or clusters of the particular given data set which is based on the similarity between them. KMeans clustering is a clustering procedure in which the given data set is divided into K i.e number of clusters. The impact factor of k-means is its simplicity, high efficiency and scalability. However, is also comprises of number of limitations: random selection of initial centroids, number of cluster K need to be initialized and influence by outliers. In view of these deficiencies, this paper represents a survey of improvements done to traditional k-means to handle such limitations and we will compare K-means clustering algorithm with various clustering algorithm.

F. *An efficient k-means clustering algorithm: analysis and implementation*

In k-means clustering, we are given a set of n data points in d-dimensional space R/sup d/ and an integer k and the problem is to determine a set of k points in Rd, called centers, so as to minimize the mean squared distance from each data point to its nearest center. A popular heuristic for k-means clustering is Lloyd's (1982) algorithm. We present a simple and efficient implementation of Lloyd's k-means clustering algorithm, which we call the filtering algorithm. This algorithm is easy to implement, requiring a kd-tree as the only major data structure. We establish the practical efficiency of the filtering algorithm in two ways. First, we present a data-sensitive analysis of the algorithm's running time, which shows that the algorithm runs faster as the separation between clusters increases. Second, we present a number of empirical studies both on synthetically generated data and on real data sets from applications in color quantization, data compression, and image segmentation.

III. METHODOLOGY

A. *Genetic*

Following are the steps of Genetic Algorithm for clustering of data:

- Input: k: the no of clusters
  d: the data set containing n objects
  p: population size Tmax: Maximum number of iterations
- Output: A set of K clusters

1) Initialize every chromosome to have k random centroids selected from the set of data.
2) For T=1 to Tmax (i) For every chromosome i
   - Allocate the object data to the cluster with the closest centroid.
   - Recomputed k cluster centroids of chromosome i as the mean of their data objects.
   - Compute the chromosome i fitness.
   - Generate the new group of chromosomes using GA selection, crossover and mutation.

B. *K-Means Algorithm*

1) *Input:*
- ky - the no. of clusters
- dy- data set which contains n number of objects

2) *Output:*
- Input the value of ky and set of data.
- If ky= = 1 then Exit
- Else
- Select k no. of objects from d randomly as the initial centers of cluster.
- For each data point in the cluster j reprint and state each object into the cluster of similar types objects, based on the mean value of object in the cluster.
- Update cluster means values and after that for every cluster compute the mean value of objects.
- Repeat from step iv until no data point was allocated, stop otherwise.

The sufficient criteria can be either no. of iteration or centroid’s change of position in consecutive iterations.

C. *Genetic K-Means Algorithm*

This algorithm combines the advantage of Genetic algorithm and K-means. Genetic Algorithm based clustering algorithm is expected to provide an optimal clustering, more superior to that of K-Means approach, but with a little more time complexity. Following are the steps of the algorithm of GK-means:

- Set the population.
- Compute fitness of every individual by following equation.
  Fitness (i) =2. (pi - 1)/Q-1 i=individual, p=position, Q=total individuals
- If satisfied with the fitness condition, then assign solution, Else
- Calculate sub population and migrate
- Counting the ith individual depends on the rate si, which is relative to its level of fitness that is Si = fitness (i) / summation (fitness (i));
- Translate population and assets individual wellness.
- Perform crossover and mutation on each sub population
- If termination condition satisfies, stop; else go to step 5.

D. *The Hybrid Genetic Fuzzy C-Means*

Hybrid approach which integrate Fuzzy C-Means (FCM) algorithms and Genetic Algorithms (GAs) to design an optimal classifier for the specific clustering problem. This integration allows automatic generation of a classifier system, with an optimized subset of features, from a database of examples. The generated clustering strongly outperform the classic FCM algorithm. A reasoned implementation of the hybrid algorithm, we called GFCM, is given along with a comparative study and performance evaluation results on several public benchmark databases. Results obtained show the efficiency of GFCM algorithm.

The most widely used clustering algorithm implementing the fuzzy philosophy is Fuzzy C-Means (FCM), initially developed by Dunn and later generalized by Bezdek, who proposed a generalization by means of a family of objective functions. Despite this algorithm proved to be less accurate than others, its fuzzy nature and the ease of implementation made it very attractive for a lot of researchers that proposed various improvements and applications. Usually FCM is applied to unsupervised clustering problems. In this research, however, show how it can be applied successfully to unsupervised clustering problems. Thanks to the integration with a Genetic Algorithm (GA), and hybrid Genetic Fuzzy C-Means (GFCM) is built up to concurrent solve the unsupervised and the feature selection problems. Several works have been proposed in the literature which make use of Evolutionary Algorithms (EAs) for fuzzy clustering, some of them are devoted to improve the performance of FCM-type algorithms using the GA to optimize parameters of these algorithms, and others are designed to create directly a fuzzy partition of data.

The fuzzy systems, which use GA to learn their structure from examples and to improve their performances, are called Genetic Fuzzy Systems (GFSs). The use of GAs to optimize the parameters of an FCM-type algorithm generates two different kinds of GFSs. Prototype-based algorithms encode the fuzzy cluster prototypes and evolve them by means of a GA guided by any centroid-type objective function , while fuzzy partition-based algorithms encode, and evolve, the fuzzy membership matrix .

A second possibility is to use the GA to define the distance norm of an FCM-type algorithm. The system considers an adaptive distance function and employs a GA to learn its parameters to obtain an optimal behavior of the FCM-type algorithm. A third group of genetic approaches are based on directly solving the fuzzy clustering problem without interaction with any FCM-type algorithm. These techniques, which are a recent trend in cluster analysis, have shown the potential to achieve high partitioning accuracy results. Previous approach employed Evolutionary Strategies, Evolutionary Programming, and recently Particle Swarm Optimization and Simulated Annealing. Details about other types of clustering are to be found in many search algorithms have been used for feature selection. Among these, EAs have proven to be an effective computational method, especially in situations where the search space is uncharacterized (mathematically), not fully understood, or/and highly dimensional. One particular application of these methods not only selects features but also assigns them weights according to their importance for the analysis to be performed.

Feature selection techniques do not normally offer the possibility of classifying the sets they analyze: they are, in fact, proposed often as filter techniques. There are, however, certain exceptions such as C4.5, which belongs to the supervised machine learning category. More recent work that integrates feature selection with classificatory analysis has been presented.

E. Data Sets

Dataset describes the contents of the heart-disease directory which is collected from UCI Repository.

IV. Results and discussion

The result of K-means, Genetic k-means, Fuzzy C Means and Hybrid Genetic Fuzzy C Means algorithm for clustering data. To group Heart disease with similar functionalities based on dataset.

S. No	Algorithm	Accuracy
1	K-Means	90 %
2	Genetic	93 %
3	Genetic K-Means	95 %
4	Fuzzy C-Means	97 %
5	Hybrid Genetic Fuzy C-Means	99 %

Table 1: Result of Clustering Algorithm

V. Conclusion future work

Hybrid Genetic Fuzzy C-Means (HGFCM), which proved to be a powerful extension of the famous Fuzzy C-Means algorithm. Essentially the GFCM algorithm is a learning algorithm which can mine databases to build an optimized and efficient classifier. In empirical tests on several well-known databases the classifiers build by HGFCM proved to be drastically better than other algorithm, obtaining an improvement up to 7 times in clustering accuracy. The technique has a number of qualities, despite the simplicity of the two integrated algorithms. It provides four different possibilities of analysis in a single algorithm and the comparisons with similar techniques, is that the HGFCM algorithm achieves an excellent accuracy when compared with other algorithm.

In future the research has been extended in the following direction of genetic algorithm and compared with weighting method using feature selection method on different data sets.

References

[1] "Anjan Goswami. Department of Computer Science and Engineering" Fast and Exact Out of-Core and Distributed K-Means Clustering 2001

[2] A. Ben-Dor, R. Shamir, and Z. Yakhini, "Clustering gene expression patterns" in *J Comput Biol* 6(3-4):281-97.

[3] Alizadeh A., Eisen M.B, Davis R.E, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 2000; 403(6769):503–511

[4] Arun. K. Pujari, "*Data Mining Techniques*", Universities press (India) Limited 2001, ISBN81-7371-3804.

[5] *Bagirov, A.M.[Adil M.]*, Modified global k-means algorithm for minimum sum-of-squares clustering problems, October 2008, pp. 3192-3199.

[6] Bingham E, Mannila H: Random projection in dimensionality reduction: applications to image and text data. Knowledge Discovery and Data Mining 2001:245-250

[7] Bloisi, D.D.[Domenico Daniele], Iocchi, L.[Luca], Rek-Means: A k-Means Based Clustering Algorithm.