Amanjyot Kaur et al. International Journal of Recent Research Aspects ISSN: 2349-7688, Vol. 3, Issue 1, March 2016, pp. 140-142

A Review: Optimized BPSO Clustering Approach for Analysis of Data in Medical Field

Amanjyot Kaur, Gagan kumar

Department of Computer Science and Engineering, Modern Institute of engineering & technology, Mohri, Kurukshetra , Haryana, India

Abstract- In this paper, BPSO, a hybrid algorithm made up of BFO and PSO algorithms uses k means clustering approach for making clusters to obtain an optimal solution. Data mining plays a very important role in the analysis of diseases and clustering approach makes it easier to classify the data collected in respective groups. Medicine companies and medical appliance manufacturer are benefitted from these data analysis. Now a days, this is done at a very large scale and has been named as big data analysis in which data size is of many terabytes. Optimization forms an integral part of our day to day life. It can be defined as an art of selecting best alternative from a set of options. Several global optimization algorithms have been developed. PSO (Particle Swarm Optimization) is a powerful optimization technique. It consists of a population of solutions called as particles where the positions of particles are determined on the basis of position vector and velocity vector. The positions of particles get changed in search of optimal solution. The particles are distinguished as personal best and global best. Hybrid algorithms combine the desirable properties of different algorithms to mitigate weaknesses of individual algorithms and result in optimal solution. For example PSO combined with GA, DE and results in DE-PSO and GA-PSO which are better versions of PSO. Bacteria foraging optimization algorithm is another type of optimization algorithm which is based on the behavior of biologically inspired E-Coli bacteria. E-Coli bacteria search the search space for rich nutrients by using their energy per unit time. The common characteristic bacteria are grouped together. The bacterium communicates with each other by sending signals. The BFO is used by many researchers recently and they try to hybridize the BFO with different algorithms to find the local best and global best solution in the search space.

Keywords: Bacterial foraging optimization (BFO), Particle Swarm Optimization (PSO), Knowledge Discovery in Databases (KDD), Swarm Intelligence (SI) algorithms

I. INTRODUCTION

With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision-making. Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to as "the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable pattern in data". While data mining and knowledge discovery in databases (KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. Figure 1.1 shows data mining as a step in an iterative knowledge discovery process.

The task of the knowledge discovery and data mining process is to extract knowledge from data such that the resulting knowledge is useful in a given application. The Knowledge Discovery process in Databases comprises of a few steps leading from raw data collections to some form of retrieving new knowledge.

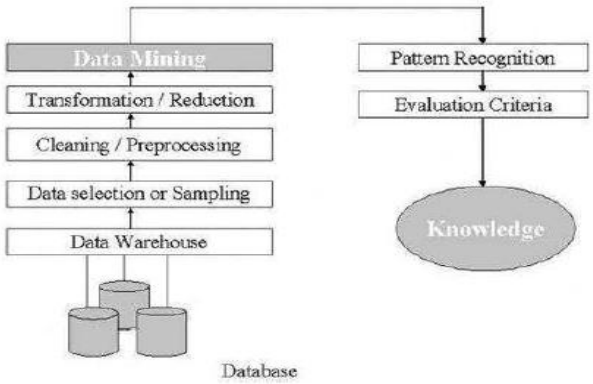

Figure 1: Complete Overview of Knowledge discovery from Databases

The iterative process consists of the following steps:

- Data cleaning: Also known as data cleansing, it is a phase in which noisy data and irrelevant data are removed from the collection.
- Data integration: At this stage, multiple data sources, often heterogeneous, may be combined in a common source.
- Data selection: At this step, the data relevant to the analysis is decided on and retrieved from the data collection.

- Data mining: It is the crucial step in which clever techniques are applied to extract data patterns potentially useful
- Pattern evaluation: In this step, strictly interesting patterns representing Knowledge is identified based on given measures.
- Knowledge representation: Is the final phase in which the discovered knowledge is visually represented to the user.

This essential step uses visualization techniques to help users understand and interpret the data mining results. It is common practice to combine some of steps together for specific application. For instance, data cleaning and data integration can be performed together as a pre-processing phase to generate a data warehouse (1.1). Data selection and data transformation can also be combined where the consolidation of the data is the result of the selection, or, as for the case of data warehouses, the selection is done on transformed data. The KDD is an iterative process. Once the discovered knowledge is presented to the user, the evaluation measures can be enhanced, the mining can be further refined, new data can be selected or further transformed, or new data sources can be integrated, in order to get different, more appropriate results.

There are several data mining models, some of these are narrated below which are conceived to be important in the area of "Data Mining".

- Clustering: It segments a large set of data into subsets or clusters. Each cluster is a collection of data objects that are similar to one another with the same cluster but dissimilar to object in other clusters
- Classification: Decision trees, also known as classification trees, are a statistical tool that partitions a set of records into disjunctive classes. The records are given as tuples with several numerics and categorical attributes with one additional attribute being the class to predict. Decision trees algorithm differs in selection of variables to split and how they pick the splitting point.
- Association Mining: It uncovers interesting correlation patterns among a large set of data items by showing attribute value conditions that occur together frequently.

II. APPLICATION OF DATA MINING

Data mining has become an important area of research since last decade. Important area where Data mining can be effectively applied are as follows:

i. Health sector (Biology/Bioinformatics),
ii. Image Processing(Image segmentation),
iii. Ad-Hoc wireless Network(clustering of nodes),
iv. Intrusion detection system,
v. Finance sector , etc.

In this thesis focus has been given on clustering techniques and their application to machine learning and bioinformatics data.

III. RELATED SEARCH

In [1] author proposed a novel method for optimization of association rule mining. Our propped algorithm is combination of distance function and genetic algorithm. We have observed that when we modify the distance weight new rules in large numbers are found. This implies that when weight is solely determined through support and confidence, there is a high chance of eliminating interesting rules. With more rules emerging it implies there should be a mechanism for managing their large numbers. The large generated rule is optimized with genetic algorithm.

In [2] author proposed a new efficient algorithm for exploring high-class association rules by particle swarm optimization (PSO) algorithm. The proposed method mine interesting and understandable association rules without using the minimum support and the minimum confidence thresholds in only single scan. To prove the practical significance of the approach, this approach is implemented on Microsoft Visual Studio 4.0. Experimental evaluation shows the efficiency of proposed algorithm in terms of computation time.

In [3] author presents a hybrid data clustering algorithm (FPAKM) based on the K-Means and Flower Pollination algorithm. The results obtained by the proposed algorithm are compared with K-Means and flower pollination algorithm. It is revealed that the proposed algorithm finds optimal cluster centres, hence the F-measure value is increased. In mere future, this algorithm can be applied to solve other optimization problems.

In [4],author clustered five different kinds of cancer datasets into different clusters with the help of four popularly used clustering algorithms. As per our analysis there is no such common learning algorithm which can give the best results in all different types of cancer datasets which we are using. Every method predicts cluster on their own calculating equation. Selection of a particular clustering approach depends on the user that what kind of cluster they require to use for the dataset under study.

In [5] this research, author uses data mining technology such as classification, clustering and prediction to identify potential cancer patients. The gathered data is preprocessed, fed into the database and classified to yield significant patterns using decision tree algorithm. Then the data is clustered using K means clustering algorithm to separate cancer and non cancer patient data. Further the cancer cluster is subdivided into six clusters. Finally a prediction system is developed to analyze risk levels which help in prognosis. This research helps in detection of a person's predisposition for cancer before going for clinical and lab tests which is cost and time consuming.

In [6]author is concerned with the ideas behind design, implementation, testing and application of a novel swarm based intelligent system for Medical Data Set analysis. The unique contribution of this paper is in the implementation of a hybrid intelligent system Data Mining technique such as Bacteria Foraging Optimization Algorithm (BFOA) for solving novel practical problems, the detailed description of this technique, and the illustrations of several applications solved by this novel technique. This paper also aims to explore the possibilities of applying this hybrid Intelligent

System DM technique to environmental and biological applications. These two fields have attracted a lot of attention recently, which is not only because of the complexity of the problem, but also because of the massive quantities of the data that are available and increasing.

In [7] author included three main practical issues: Handling noisy and incomplete data, Generating almost daily huge amounts of heterogeneous data, processing compute intensive tasks. We suggest here in this study data mining techniques as Fuzzy association rules and neural network techniques. Knowledge management is providing the facility to find out these rules any time when need.

In [8] author describes that the most famous clustering approach is K-means which successfully has been utilized in numerous clustering problems, but this algorithm has some limitations such as local optimal convergence and initial point understanding. Clustering is the procedure of grouping objects into disjoint class is known as clusters. So, that objects within a class are extremely similar with one another and dissimilar with the objects in other classes. Firefly algorithm is mainly used for clustering problems, but it also has disadvantages. To overcome the problems in firefly this work used a proposed method of Hybrid K-Mean with GA/PSO. The hybrid method merges the standard velocity and modernizes rules of PSOs with the thoughts of selection from GAs. They compare the hybrid algorithm to the standard GA and PSO approaches. Experimental results show that the proposed method used to reduce the limitations and improve accuracy rate.

In [9] the proposed approach uses dynamic K-means algorithm is used for dynamic data clustering approaches. It can be applied to both known number of clusters as well as unknown number of clusters. Hence, the user can either fix the number of clusters or they can fix the minimum number of required clusters. If the number of clusters is static, it works like K-means algorithm. If the number of clusters is dynamic, then this algorithm determines the new cluster centers by adding one to the cluster counter in each iteration until the required cluster quality is achieved. The proposed method uses Modified Firefly algorithm to determine the centroid of the user specified number of clusters. This algorithm can be extended using dynamic k-means clustering to enhance centroids and clusters. Thus the proposed Dynamic clustering method increases the cluster quality and modified firefly algorithm increases optimality for the iris and wine datasets. Experimental results proved that the proposed methodology attains maximum cluster quality within a limited time and achieves better optimality.

IV. CONCLUSION

In data clustering class labels are not known in advance hence it is also known as unsupervised learning. The clusters so formed after clustering contain a set of objects that are similar within a cluster but far away from other cluster's objects. In past decades,non - linear nature inspired evolutionary algorithms were developed for solving most engineering design optimization problems because they take less amount of time to solve the real world problems. Nature inspired algorithms imitate the behavior of natural living objects hence known as Swarm Intelligence(SI) Algorithms. Dynamic K-means algorithm is used to attain maximum cluster quality. Bacterial foraging optimization [6] has already been implemented for solving novel practical problems In BPSO, Data clustering analysis using bacterial foraging optimization (BFO) and particle swarm optimization (PSO) is used with k means clustering approach to achieve better optimality .

V. REFERENCES

[1]. Sanjay Tiwari, Mahainder Kumar Rao,'' Optimization In Association Rule Mining Using Distance Weight Vector And Genetic Algorithm" International Journal of Advanced Technology & Engineering Research (IJATER), Volume 4, Issue 1, Jan. 2014.

[2]. PoonamSehrawat, Manju," Association Rule Mining Using Particle Swarm Optimization", International Journal of Innovations & Advancement in Computer Science, Volume 2, Issue 1 January 2014

[3]. R.Jensi and G.WiselinJiji,"Hybrid Data Clustering Approach Using K-Means And Flower Pollination Algorithm", Advanced Computational Intelligence: An International Journal (ACII), Vol.2, No.2, April 2015

[4]. Khalid Raza,"Clustering analysis of cancerous microarray data",Journal of Chemical and Pharmaceutical Research, 2014, 6(9)

[5]. P. Ramachandran, N.Girija,"Early Detection and Prevention of Cancer using Data Mining Techniques", International Journal of Computer Applications, Volume 97– No.13, July 2014.

[6]. Hlaudi Daniel Masethe, Mosima Anna Masethe,"Prediction of Heart Disease using Classification Algorithms", Proceedings of the World Congress on Engineering and Computer Science 2014 Vol II

[7]. P.Kalyani,"Medical Data Set Analysis Ñ A Enhanced Clustering Approach" International Journal of Latest Research in Science and Technology, Volume 3, Issue 1: Page No.102-105 ,January-February 2014.

[8]. Ibrahim M. El-Hasnony, Hazem M. El Bakry, Ahmed A. Saleh,"Data Mining Techniques for Medical Applications: A Survey", Mathematical Methods in Science and Mechanics,2014

[9]. Sundararajan S, Dr. KarthikeyanS,"An Hybrid Technique for Data Clustering Using Genetic Algorithm with Particle Swarm Optimization", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 12, December 2014.

