MNK Publication

International Journal of Latest Research in Science and Technology **ISSN (Online):2278-5299**
Volume 3, Issue 1: Page No.102-105 ,January-February 2014
https://www.mnkpublication.com/journal/ijlrst/index.php

MEDICAL DATA SET ANALYSIS – A ENCHANCED CLUSTERING APPROACH

P.Kalyani
Associate Professor, S.N.R.Sons College, Coimbatore.

Abstract- **This Paper 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.**

Keywords - Intelligent system; bacteria foraging optimization; hybrid intelligent system; swam based intelligent system; biological application

I. INTRODUCTION

In recent years DM has attracted great attention inthe information industry and in society as a whole. This is because, on the one hand, modern computers and other piece of equipment are able to produce and store virtually unlimited datasets characterizing a complex system [1-6]. In fact, database and information technology has been evolving systematically from primitive file processing systems to sophisticated and powerful database systems. On the other hand, however, there are no concise set of parameters that can fully describe the state of real-world complex systems studied nowadays by engineers, psychologists, economists, etc. These on the contrary inspire the development of advanced DM which may employ techniques such as Ant Colony Optimization algorithms, Bacteria Foraging Optimization algorithm and fuzzy logic etc.

II. INTELLIGENT DATA MINING

A common feature of all DM techniques is that they are all, to a certain extent, data analysis methods and can support/interact with each other. However, each discipline has its own distinct attributes that make it particularly useful for certain types of problems and situations. For example, the most fundamental difference between classical statistical applications and data mining may be suggested to be the size of the dataset. Statistical techniques alone may not be sufficient to address some of the more challenging issues in data mining, especially those arising from very large datasets. On the other hand, an Intelligent System (IS) is all about learning rules and patterns from the data. With the help of available computational power in IS tools, there is a great potential for significant advances in both theoretical and applied research in this DM area.

The term Intelligent Systems is used interchangeably with Soft Computing in this paper. It is a collection of methodologies that works Synergistically and provides, in one form or another, flexible information processing capability for handling real-life situations.

III. OPTIMIZATION TECHNIQUES

Soft computing has attracted many research scientists, decision makers and practicing researchers in recent years as powerful computational intelligent techniques, for solving unlimited number of complexreal-world problems particularly related to research area of optimization. Under the uncertain and turbulence environment, classical and traditional approaches are unable to obtain a complete solution with satisfaction for the real world problems on Optimization. Therefore, new global optimization methods are required to handle these issues seriously [7-8]. One such method is hybrid evolutionary computation, a generic, flexible, robust, and versatile framework for solving complex problems of global optimization and search in real world applications.

Formulation of a new non-linear membership function using fuzzy approach is to capture and describe vagueness in the technological coefficients of constraints in Medical data analysis problems. This non-linear membership function is flexible and convenience to the decision makers in their decision making process.

The hybrid optimization techniques are robust, less time-consuming, dependable, high quality solutions and an efficient productive tool for solving the non-linear real world problem in a Medical engineering environment. The hybrid line search with Evolutionary algorithms and hybrid line search with Fuzzy C Means clustering techniques developed in this study are user friendly, easy-to-use and can serve as a teaching and research tool, besides being useful for practicing scientist in the area of industrial engineering.

Publication History
Manuscript Received : 27 February 2014
Manuscript Accepted : 1 March 2014
Revision Received : 3 March 2014
Manuscript Published : 10 March 2014

In optimization of a design, the design objective could be simply to minimize the cost of production or to maximize the efficiency of production. An optimization algorithm is a procedure which is executed iteratively by comparing various solutions till an optimum or a satisfactory solution is found. With the advent of computers, optimization has become a part of computer-aided design activities. There are two distinct types of optimization algorithms widely used today. Several optimization techniques have been used in the literature in clustering technique. The bacteria foraging algorithm is proposed in this research work. For the first time bacteria foraging algorithm is applied in clustering technique problem for optimization.

IV .FUZZY C MEANS ALGORITHM

Fuzzy clustering is a powerful unsupervised method for the analysis of data and construction of models. In many situations, fuzzy clustering is more natural than hard clustering. Objects on the boundaries between several classes are not forced to fully belong to one of the classes, but rather are assigned membership degrees between 0 and 1 indicating their partial membership. Fuzzy c-means algorithm is most widely used. The FCM employs fuzzy partitioning such that a data point can belong to all groups with different membership grades between 0 and 1.

This algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point. More the data is near to the cluster center more is its membership towards the particular cluster center. Clearly, summation of
Membership of each data point should be equal to one. Each iteration membership and cluster centers are updated according to the formula.

Advantages

- Unsupervised
- Converges

Limitations:

- Long computational time
- Sensitivity to the initial guess (speed, local minima)
- Sensitivity to noise and One expects low (or even no) membership degree for outliers (noisy points).

V. BACTERIA FOR AGING OPTIMIZATION ALGORITHM (BFOA)

Bacteria Foraging Optimization Algorithm (BFOA) wasproposed by passino it is one of the bio- inspired optimizationalgorithm. Over the last five years, the optimization algorithm likes Evolutionary Programming (EP), Genetic Algorithm (GA) which really shows their performance in optimization problem. Recently natural bio inspired algorithm Ant Colony Optimization (ACO), and Particle Swarm Optimization (PSO) shows their best performance in the optimization problem. BFOA Algorithm based on the behaviour of biologically inspired E-coli bacteria, used to findoptimal solution [9-11]. E-coli bacteria search for rich nutrients in the search space by using their energy per unit time. The common characteristic bacteria's are grouped together. The bacterium can communicate with each other by sending signals.

The bacteria take swimming over the search space for rich nutrients is called Chemotaxis activity. The chemotatic movement of bacteria in two directions (i) bacteria can swim the constant direction (ii) tumble -bacteria can move in the random direction, the swimming decisions based on finding the nutrients in the search space.The Bacteria Foraging Optimization Algorithm can be used by most of the researchers recently. Initially it is applied in the field of Electrical engineering optimization problem. The researchers tried to hybridize BFOA with different other algorithm to find the local best and global best solution in the problem search space.

In the real E- coli bacteria, the movement can be achieved by set of tensile flagella. The flagella help E-coli bacterium to swim/ tumble, which are two basic swimming operations of bacterium at the time of chemotatic movement.
When the bacterium rotates in the clockwise direction, the flagellum pulls the cells. The bacterium can tumble in the random direction to find the rich nutrition gradient. The Flagella movement in the counter clockwise direction helps the bacterium to swim at a very fast rate. If the bacteria get the sufficient food, it will grow. This leads to reproduction of bacteria, (i.e) the bacteria divided in to two parts and spread over the search space. The sudden changes in the search space will affect the bacteria to find the optimum solution. So, the elimination Dispersal event can maintain the population in the environment.

VI. IMPLEMENTATION OF BFOA WITH FCM

The social foraging behavior of Escherichia coli bacteria has been used to solve optimization problems in medical data set. This chapter proposes a hybrid
approach involving Bacterial Foraging (BF) algorithms and Fuzzy C Means algorithms for medical data analysis. First illustrate the proposed method using the performance of the algorithm is studied with anemphasis chemotactic steps,
lifetime of the bacteria and the clusters .Simulation results clearly illustrate that the proposed approach is very efficient and could easily be extended for other global optimization problems.

After completing all the process the generated output is given to the FCM as input. The optimal value of BFOA through variousmedical data set is given as an input for FCM. The aim of FCM is to find cluster centers (centroids) that minimize dissimilarity function. The membership matrix (U) is randomly initialized as

$$\sum \text{Uij} = 1$$

where i is the number of cluster j is the image data point

The dissimilarity function can be calculated with this equation

$$Ci = \sum_{i=1}^{c} Ji = \sum_{i=1}^{c} \sum_{j=1}^{n} Ui\ j^{n}$$

where Uij is between 0 and 1

Ci is the centroid of cluster i

Dij is the Euclidean distance between ith and centriod (Ci) and jth data point

To calculate Euclidean distance (dij)

Euclidean distance (dij) = Cluster center value - current neuron

Dij = CCp – Cn

where CCp is the Cluster center value

Cn is the current neuron

i.e. Number of clusters is computed as

C = (N/2)1/2

N= no. of pixels in image

VII. PERFORMANCE EVALUATION

The main purpose of metric learning in a specific problem is to learn an appropriate distance/similarity function. Metric learning has become a popular issue in many learning tasks and can be applied in a wide variety of settings, since many learning problems involve a definite notion ofdistance or similarity.

A metric or distance function is a function which defines a distance between elements of a set. A set with a metric is called a metric space. In many data retrieval and data mining applications, such as clustering, measuring similarity between objects has become an important part. Normally, the task isto define a function Sim(X,Y), where X and Y are two objects or sets of a certain class, and the value of the function represents the degree of "similarity" between the two.

Formally, a distance is a function D with nonnegative real values, defined on the Cartesian product X x X of a set X. It is called a metric on X if for every x,y,z εX:- D(x,y)=0 if x=y (the identity axiom);
- D(x,y) + D(y,z) ≥D(x,z) (the triangle inequality);
- D(x,y)=D(y,x) (the symmetry axiom).

A set X provided with a metric is called a metric space. Euclidean distance is the most commonuse of distance – it computes the root of square differences between coordinates of a pair of objects: (1) Manhattan distance or city block distance represents distance between points in a city road grid. It computes the absolute differences between coordinates of a pair of objects:

The purpose of the experimental part was to test the operation of the Fuzzy C -means algorithm by applying different metrics. Two metrics have been chosen: Euclidean distance and Manhattan distance. In the course of the experiments in order to determine cluster centres in k-means clustering algorithm sequentially all three metrics has been used. The results obtained have been analyzed and the clustering correctness has been tested.

Table 5.1: Bacteria Foraging Optimization Algorithmwith FCM results by applying different metrics

Distance	Euclidean Distance	Manhattan
Cluster centres	52.16 35.17 21.28 9.62 68.50 30.74 57.42 20.71 59.02 27.48 43.94 14.34	52 35 21 9 57 27 42 13 65 30 54 19
Cluster1 contains:	Records from cluster 1 – 50 Records from cluster 2 – 0 Records from cluster 3 – 0	Records from cluster 1 – 50 Records from cluster 2 – 0 Records from cluster 3 – 0
Cluster2 contains:	Records from cluster 1 – 0 Records from cluster 2 – 48 Records from cluster 3 – 2	Records from cluster 1 – 0 Records from cluster 2 – 42 Records from cluster 3 – 8
Cluster3 contains:	Records from cluster 1 – 0 Records from cluster 2 – 10 Records from cluster 3 – 40	Records from cluster 1 – 0 Records from cluster 2 – 3 Records from cluster 3 – 47
Correct-ness:	For cluster 1 – 100 % For cluster 2 - 96 % For cluster 3 - 80 %	For cluster 1 – 100 % For cluster 2 - 84 % For cluster 3 - 94 %

During the experiment the well-known UCI machine learning repository breast cancer detection data set was employed, containing the 50 data set are divided into two classes that are recurrent and non-recurrent. Each species has four attributes: age, menopause, tumor-size, inv nodes. However, it is uncommon to use this data set in cluster analysis, since the data set contains only two clusters with rather obvious separation. The experimental part has been carried out in Matlab
environment.

VI. COMPARISON OF TIME COMPLEXITY OF ACO-FCM AND BFOA-FCM

The time complexity of ACO-FCM is O(ndc2i) and time complexity of BFOA-FCM is O(ncdi) [12-14]. Now keeping no. of data points constant, lets assume n=130, d=2, i=5 and varying no. of clusters,we obtain the following table and graph. Where n=number of data point, c= numberof cluster, d= dimension, i=number of iteration.

Table 5.2: Ant Colony Optimization with FCM Clustering results by applying different metrics

Distance	Euclidean Distance	Manhattan
Cluster centres	50.06 34.28 14.62 2.46 68.50 30.74 57.42 20.71 59.02 27.48 43.94 14.34	50 34 15 2 57 27 42 13 65 30 54 19
Cluster1 contains:	Records from cluster 1 – 50 Records from cluster 2 – 0 Records from cluster 3 – 0	Records from cluster 1 – 50 Records from cluster 2 – 0 Records from cluster 3 – 0
Cluster2 contains:	Records from cluster 1 – 0 Records from cluster 2 – 47 Records from cluster 3 – 3	Records from cluster 1 – 0 Records from cluster 2 – 46 Records from cluster 3 – 4
Cluster3 contains:	Records from cluster 1 – 0 Records from cluster 2 – 12 Records from cluster 3 – 38	Records from cluster 1 – 0 Records from cluster 2 – 2 Records from cluster 3 – 48
Correct-ness:	For cluster 1 – 100 % For cluster 2 - 94 % For cluster 3 - 76 %	For cluster 1 - 100 % For cluster 2 - 92 % For cluster 3 - 96 %

The above tables show that all metrics correctly recognize cluster 1 records. Cluster 2 records are best recognized by Euclidean distance.

Table 5.3: Time Complexity when Number of cluster varying

S.No.	Number of Cluster	ACO-FCM Time Complexity	BFOA-FCM Time Complexity
1	1	2500	2400
2	2	4600	4000
3	3	7000	5500
4	4	9400	7000

Based on the time complexity with number of clusters varying, the BFOA-FCM is better than compare to ACO-FCM and other existing methods.

VII. CONCLUSIONS

In this review, the two major phases in medical data analysis such as i) Clustering Technique, ii) Swarm intelligence Algorithm are studied and appropriate methods and algorithms are selected for further proceedings and the reason for their adoption are discussed in brief. a novel approach to Medical data set analysis based on the Ant Colony Optimization (ACO) with FCM Algorithm have been used to find out the optimum label that minimizes the maximizing a posterior estimate the medical data set. A method based on ACO with use of FCM is implemented.

ACKNOWLEDGMENT

We like to express sincere appreciation and deep gratitude to all participants in this work.

REFERENCES

[1] P. Berkhin, "Survey clustering Data Mining Techniques", Technical Report, Accrue Software, San Jose, California, 2002.

[2] Ashish Ghosh, Anindya Halder, Megha Kothari and Susmita Ghosh, "Aggregation pheromone density based data clustering", Information Sciences, Vol. 178, Issue 13, 1 July 2008, pp. 2816-2831.

[3] B. Duran and P. Odell, "Cluster Analysis: A Survey", New York: Springer-Verlag, 1974.

[4] Diego Alejandro Ingaramo, Guillermo Leguizamon and Marcelo Errecalde, "Adaptive clustering with artificial ants", Journalof Computer Science and Technology, Science Press, Vol. 5, No. 4, 2005.

[5] C. Fernandes, A.M. Mora, J.J. Merelo, V. Ramos and J.L.J. Laredo, "KohonAnts: A Self-Organizing Ant Algorithm for Clustering and Pattern Classification", http://arxiv.org/abs/0803.2695v1, 2008.

[6] Haifang Li, Xia Wen and Hui Jin, "The Clustering Algorithm Research ofImage Emotional Characteristics Based on Ant Colony", Proceedings of the 2008 International Symposium on Intelligent Information Technology Application Workshops, 2008, pp. 455-458.T. Marler and J.S. Arora, "Survey of multi-objective optimization methods for engineering," Struct. Multidisc. Optim. 26, pp. 369–395, 2004.

[7] A. Abraham and V. Ramos, "Web usage mining using artificial ant colony clustering and linear genetic programming", Proc. Congress on Evolutionary Computation (IEEE Press), Australia, 2003, pp.1384-1391.

[8] W. Bin and S. Zhongzhi, " A clustering algorithm based on swarm intelligence", Proc. of the Int. Conf. on Info-tech. and Info-net, Beijing, China, 2001, pp. 58-66.

[9] Korani W, " Bacterial Foraging Oriented By ParticleSwarm Optimization Strategy for PID Tuning", GECOO 2008, 12-16 July, 2008, Atlanta, USA, pp. 1823-1826.

[10] S. Camazine, J.-L. Deneubourg, N.R. Franks, J.Sneydd, G. Theraulaz and E. Bonabeau, "Self-Organization in Biological Systems", Princeton University Press, 2001.

[11] J. Kennedy and R. Eberhart, "Swarm Intelligence", Morgan Kaufmann Publishers, Inc., San Francisco, CA, 2001.

[12] Fisher R.A. The use of multiple measurements in taxonomic problems.Ann.Eugenics, 1936,7(2), p.179-188.

[13] Prodip Hore, Lawrence O. Hall, and Dmitry B. Goldgof "Single Pass Fuzzy C Means", CSEEE, vol.28, 2000.

[14] M. Brej and M. Sonka, "Object localization and border detection criteria design in edge-based image segmentation automated learning from examples", IEEE Transactions on Medical imaging, vol. 19, pp. 973-985, 2000.