Nehaet al. International Journal of Institutional & Industrial Research ISSN2456-1274, Vol.3, Issue2, May-August 2018, pp13-17

An Optimized hybrid algorithm for clustering of data using BFO and PSO

Neha*, Sudeepta Thukral**
MTech Research Scholar*, Computer Science and Engineering Department, MIETCollege,Mohri Kurukshetra, Haryana, India
Assistant Professor**, Computer Science and Engineering Department, MIETCollege,Mohri Kurukshetra, Haryana, India

Abstract: **In data processing, K-means bunch is renowned for its potency in bunch massive information sets. the most aim in grouping information points into clusters is to lump similar things along within the same cluster specified objects lying in one cluster ought to be as shut as potential to every alternative (homogeneity) and objects lying in numerous clusters area unit any except one another. However, there exist some flaws in classical K-means bunch algorithmic program. First, the algorithmic program is sensitive in choosing initial centroids and may be simply cornered at an area minimum with regards to the activity (the total of square errors). Secondly, the KM downside in terms of finding a worldwide stripped total of the square errors is NP-hard even once the quantity of the clusters is capable a pair of or the quantity of attributes for information is a pair of, thus finding the best bunch is believed to be computationally balky. In this paper, KM clustering downside is resolved by optimized KM. The projected rule is called as BPSO within which the difficulty of a way to derive Associate in Nursing optimisation model for the minimum add of square errors for a given information set is taken into account. 2 biological process optimisation rules BFO and PSO square measure combined to optimize kilometer rule to ensure that the results of cluster is a lot of correct than cluster by basic kilometer algorithm. F-measure is employed to try to comparison of each basic K-means and BPSO rule.**

General Terms: **Bacterial foraging optimization, Particle Swarm Optimization, F-measure, K-means, Data Mining.**

Keywords: **PSO (Particle Swarm Optimization), BFO (Bacterial Foraging Optimization), KDD (Knowledge Discovery in Databases), BPSO (Bacterial Particle Swarm Optimization), KM (K-Means).**

1. Introduction

Generally, data mining is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cuts costs, or both or for future analysis. 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. For example, a supermarket can use the data mining capacity to analyze local buying patterns. For example, it is discovered that when men buy noodles on Thursdays and Saturdays, they tend to buy tomato-ketchups on the same days. If in more analysis it is discovered that the shoppers typically do their grocery shopping on Sundays. On Thursdays, however, they only buy 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.

2. Objective

Data mining plays a very important role in the analysis of diseases, buying patterns and many more, and clustering approach makes it easier to classify the similar data collected in respective groups as close as possible. Now a days, this is being done at a very large scale and has been named as big data analysis in which data size is of many terabytes. In the proposed work although such a large amount of data would not be used, yet work is performed to improve the clustering approach by using evolutionary optimization technique
.Following will be the key points:

data clustering using BFO and PSO is used with KM clustering approach.

Medical data will be used for analysis purpose. Data will be collected from the link: ftp://ftp.ics.uci.edu./pub/machine-learning-databas

F-measure would be the comparison parameter.

3. Literature Review

Nikhil Kushwaha et.al[1] proposed a GA based BF algorithm for function optimization. The performance studies on mutation, crossover, variation of step sizes, chemotactic steps, and the lifetime of the bacteria were done in the proposed method using test functions.

Tarun Kumar Sharma et.al[2] proposed ISBC, an improved ABC algorithm by embedding PSO into it. The results of experiments on a set of 6 datasets demonstrated good performance of ISBC to solve complex optimization problems when compared with two ABC-based algorithms.

Vipul Sharma et.al[3] presented an application of BFOA variants that would be useful for researchers to explore its use in other research problems.

Yang Yong[4] proposed a method based on KM clustering and genetic algorithm to effectively enhance classified performance of the minority kind of data in the imbalanced

© 2018 IJIR All Rights Reserved

data set. KM algorithm is used to cluster and to group the minority kind of sample. Finally, KNN and SVM sorter are used to prove the validity in the simulation experiment.

Youguo Li et.al[5] combined the largest minimum distance algorithm and the traditional KM algorithm to improve KM clustering algorithm. The shortcoming of determining initial focal point of traditional algorithm can be removed by improved algorithm.

Sunita Sarkar et.al[6] described a brief survey on PSO application in data clustering. PSO has emerged as a replacement to more conventional clustering techniques. PSO is a stochastic algorithm that mimics the capability of swarm that is both cognitive and social behavior. Data clustering with PSO algorithms showed good results in a wide variety of real-world data.

Gautam Mahapatra et.al[7] presented an alternative form of the recursive equation of Runge-Kutta method. The work showed how the bio-inspired BFOA, which is a mimicry of common type of bacteria like E. Coli, can be used to solve system of equation with rank less than or equal to n.

Hlaudi et.al[8] contributed in solving novel practical problems by the use of a hybrid system such as BFOA. The research result showed prediction accuracy of 99%.

Ibrahim M.El-Hasnony et.al[9] concerned on the most famous KM clustering approach which has some limitations such as local optimal convergence and initial point understanding. To overcome the problems of firefly, Hybrid KM with GA/PSO is proposed. The hybrid algorithm is compared with the standard GA and PSO approaches. Experimental results showed that the proposed method reduced the limitations and improved accuracy rate.

Khalid Raza et.al[10] clustered five different kinds of cancer datasets into different clusters with the help of four popularly used clustering algorithms and it was noticed that as per analysis there is no common learning algorithm that can give the best results in all types of different cancer datasets.

4. K-Means Algorithm

KM clustering means classification of data into groups of objects on the basis of attributes/features into K number of groups. KM formally described by Algorithm:

Algorithm: Basic K-Mean Clustering

1: Choose k points as initial centroid
2: **Repeat**
3: Assign each point to the closest cluster center.
4: Re compute the cluster centers of each cluster.
5: **Until** convergence criterion is met.

BFO

BFO is a nature-inspired optimization algorithm in which bacteria E. Coli searches for nutrients in order to maximize energy obtained per unit time. A step in which each bacterium searches food by taking small steps is called chemotactic. Flagella helps an *E. coli* bacterium to tumble or swim the two basic operations performed by a bacterium at the foraging time. Thus E. Coli searches for food can be categorized into four steps: Chemotactic, Swarming, Reproduction and Killing/ Dispersion.

PSO

PSO simulates the behavior of bird flocking, each single solution called “particle” is a "bird" in the search space. A fitness function is to be optimized and all particles have fitness values and have velocities which direct the flying of the particles. The particles fly through the problem space by following the current optimum particles. PSO is initialized with a group of random particles or solutions and then searches for optima. Every particle is updated by two "best" values, pbest which is the best solution or fitness it has achieved so far and gbest is the best value obtained so far by any particle in the population. After finding the two best values, the particle updates its velocity and position by (a) and
(b

v[ ] = v[ ] + c1 * rand( ) * (pbest[ ] – present[ ]) + c2 * rand( ) * (gbest[ ] – present[ ])……...(a)

present[ ] = present[ ] + v[ ]……..……(b)

v[ ] is the particle velocity, present[ ] is the current particle position. rand ( ) is a random number b/w 0 and 1. c1, c2 are learning factors.

5. Proposed Work

Problem Description

In this work a hybrid evolutionary optimization technique for data clustering is used which used KM objective function for data clustering. BFO and PSO are used in combination so that speed issue in global optimization

F-measure is a measure of a test's accuracy. Flow chart is shown below.

algorithms and iteration jump issue in PSO can be overcome.

Proposed Algorithm

In the proposed work, random direction of bacteria is controlled by PSO. The velocity of swarm particles in PSO is continuously updated considering fitness output function of BFO until swimming step. This updated velocity as per PSO algorithm will serve as the direction of bacteria. After that reproduction and killing/dispersion steps of bacteria follow. So the position of cluster heads amongst the data is optimized so that rest data settle down near to those cluster heads and get classified. Pseudo steps of the algorithm:

Load input clustering data
Find out the number of attributes in the data
For ii=1: number of attributes
Inputattribute=inputdata(attributes
) Choose number of cluster heads

Initialize all variables and steps in BFO and PSO optimization algorithm
For 1: killing/dispersion steps
For 1: reproduction steps For
1: swarming steps

© 2018 IJIR All Rights Reserved

For 1: chemotactic steps
Pass each initial bacteria positions to objective function Calculate the MSE using equation:
$MSE = \sum_{i=1 \text{ to } K} \sum_{x \in c_i} dist(C_i, X)^2$
Update the bacteria's positions as described in section 3.2 Call objective function again
Chemotactic steps end
Swarming steps end
Assign bacteria position as the current position in PSO
Take out the minimum value of the fitness function and that bacteria's position will serve as local best position of swarms Again index of minimum of fitness calculated in previous step serves the global best position for swarms
Update the velocity of swarm particles as in section 3.3.1 Direction of bacteria=velocity of swarms
Reproduce the bacteria
Reproduction steps end
Kill the bacteria if random value if greater than probability, otherwise disperse it
Elimination/dispersal steps end
Assign the data of attribute to corresponding cluster which will be nearest to them
Repeat all above steps for all attributes in the data Calculate the f-measure
End

6. Results And Discussion

In this work the data is classified by using optimized KM clustering approach. BFO and PSO are two Evolutionary algorithms that are combined to form BPSO algorithm. The results are compared in terms of F-measure by using MATLAB tool, six datasets described in Table 6.1 have been used which are taken from ftp://ftp.ics.uci.edu./pub/machine-learning databases/.

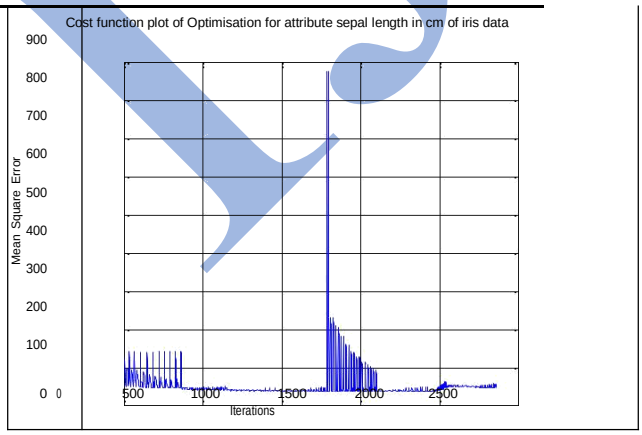
Table 6.1: Dataset Names

	Dataset	Attributes	Classes	Instances
1	Liver disorder	7	2	345
2	Iris	4	3	150
3	Wine	13	3	178
4	Glass	9	6	214
5	Thyroid	5	6	215
6	CMC	10	3	1473

Combined results of all attributes are used to calculate the min, max, std deviation and avg of F-measure. Example of iris data with 4 attributes is shown in Table 6.2.

Table 6.2: Fitness Value Function plot for Iris Attributes

1st



© 2018 IJIR All Rights Reserved

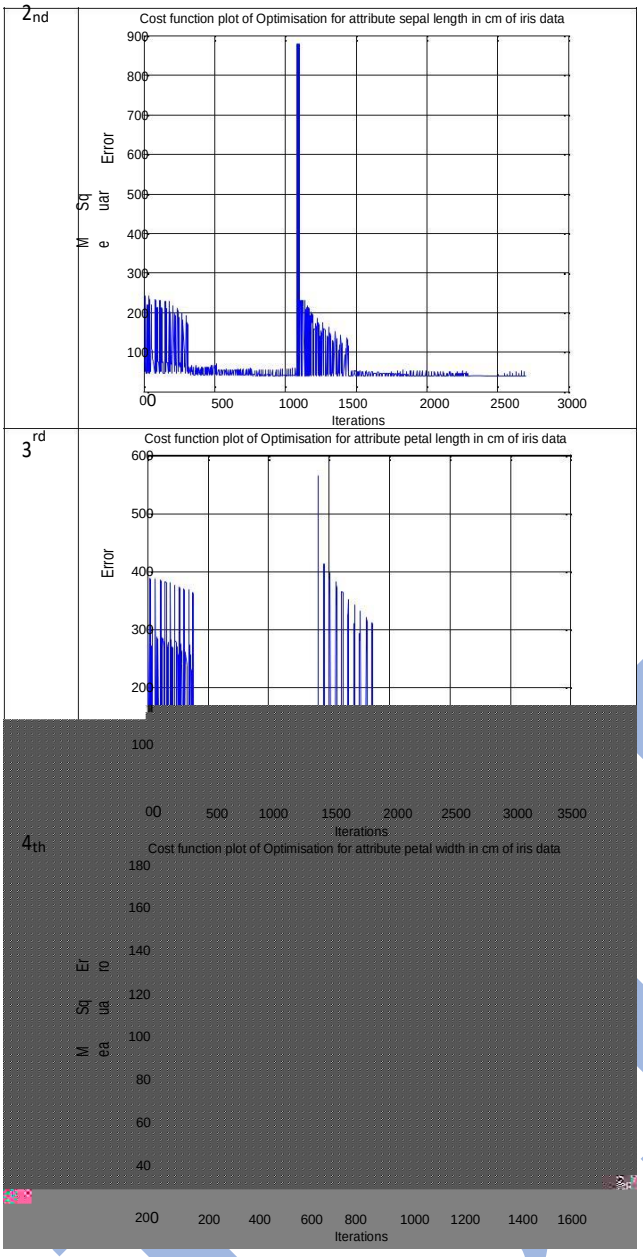


Table 6.3: F-measure Comparison

Data		F-meas.			
		Fmin	Fmax	Favg	Fstd
Liver	KM	.519459	.959984	.788855	.210787
	Pro.	.474188	.959936	.724632	**.217614**
Iris	KM	.566273	.959984	.798872	.189462
	Pro.	.565892	**.959984**	.775703	.170778
Wine	KM	.421926	.791354	.609064	.102975
	Pro.	**.436564**	.764669	.589900	.095845
Glass	KM	.292118	.698191	.389626	.114299
	Pro.	**.332840**	.689607	**.399367**	.104201
Thyr.	KM	.705597	.813951	.749209	.041964
	Pro.	.608734	**.834881**	.716553	**.087047**
CMC	KM	.395856	.503563	0.454815	.041383
	Pro.	**.395856**	**.503725**	**.456368**	.035027

© 2018 IJIR All Rights Reserved