Analysis of Algorithm for classification of Data Using BPSO

Neetika, Gagan Kumar

Computer Science and Engineering Department, MIET College, Mohri, Kurukshetra , Haryana, India

Abstract—In data mining, K-means clustering is well known for its efficiency in clustering large data sets. The main aim in grouping data points into clusters is to lump similar items together in the same cluster such that objects lying in one cluster should be as close as possible to each other (homogeneity) and objects lying in different clusters are further apart from each other. However, there exist some flaws in classical K-means clustering algorithm. First, the algorithm is sensitive in selecting initial centroids and can be easily trapped at a local minimum with regards to the measurement (the sum of squared errors). Secondly, the KM problem in terms of finding a global minimal sum of the squared errors is NP-hard even when the number of the clusters is equal to 2 or the number of attributes for data point is 2, so finding the optimal clustering is believed to be computationally intractable. In this paper, KM clustering problem is solved by optimized KM. The proposed algorithm is named as BPSO in which the issue of how to derive an optimization model for the minimum sum of squared errors for a given data set is considered. Two evolutionary optimization algorithms BFO and PSO are combined to optimize KM algorithm to guarantee that the result of clustering is more accurate than clustering by basic KM algorithm. F-measure is used to do comparison of both basic K-means and BPSO algorithm.

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

I. 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, cut 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.

II. 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.

Taran 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 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 Mahaputra 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.

Poonam Sehrawat et al.[11] proposed a new efficient algorithm for mining interesting and understandable association rules in single scan without using the minimum support and the minimum confidence thresholds. The approach is implemented on Microsoft Visual Studio 4.0 to prove the practical significance of the proposed approach.

Sanjay Tiwari et al.[12] proposed algorithm by combining distance function and genetic algorithm. It was observed that when the distance weight was modified new rules in large numbers were found. This implied that when weight is solely determined through support and confidence values, there becomes a high chance of eliminating interesting rules.

P.Kalyani et al.[13] included three main practical issues: Generating huge amounts of heterogeneous data daily, Handling noisy and incomplete data, compute intensive tasks. The data mining techniques such as Fuzzy association rules and neural network techniques are proposed.

P.Ramachandran et al.[14] used data mining technology to identify potential cancer patients. The data was clustered using KM clustering algorithm. The research helped in preventing cost and time loss by detection of a person’s predisposition for cancer before going for clinical and lab tests.

Sandeep U.Mane et al.[15] presented an efficient hybrid evolutionary optimization algorithm PSO-ACO by combining PSO and ACO for optimally clustering N object into K clusters. The results showed that the proposed evolutionary optimization algorithm is robust and suitable for handing data clustering.

Amin Rostami et al.[16] proposed constraint KM Mode clustering algorithm to find the likelihood of diseases. The developed algorithm can handle both continuous and discrete data. The effectiveness was demonstrated by testing it for a real world patient data set.

Sundararajan S.et al.[17] tested five algorithms, namely a standard KM, PSO, KM + Genetic algorithm. Hybrid approach and the Hybrid Sequential clustering algorithm, in which clusters’ centers are found by swarms and further refined by KM algorithm. It was showed that the proposed clustering algorithm had better convergence to lower quantization errors.

Maheshwar et al.[18] proposed a firefly based genetic algorithm where the initial population is to be selected from a pool of population that is on the basis of fire-fly algorithm. In the proposed work FAG algorithm is applied to the publicly available datasets. The results obtained are very much satisfactory and competitive as compared to the basic genetic and firefly algorithm.

R.Jensi et al.[19] presented a hybrid data clustering algorithm FPAKM based on KM and Flower Pollination algorithm. The results are obtained by comparing F-measure values of the proposed algorithm with those of KM and flower pollination algorithm and from that optimal cluster centers were found for the proposed approach. In mere future, this algorithm can be applied to solve other optimization problems and in the same way other algorithms can be combined to provide desirable better results.

III. KM-MEANS CLUSTERING & EVOLUTIONARY ALGORITHMS USED

3.1 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 3.1:

Algorithm 3.1: 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.

3.2 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.

3.3 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).
\[ v[i] = v[i] + c1 \times \text{rand()} \times (pbest[i] - \text{present}[i]) + c2 \times \text{rand()} \times (gbest[i] - \text{present}[i]) \]…………(a)

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

\[ \text{present}[i] = \text{present}[i] + v[i] \]……………..........……..(b)

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

IV. PROPOSED WORK

4.1 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 algorithms and iteration jump issue in PSO can be overcome.

4.2 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:

i. Load input clustering data
ii. Find out the number of attributes in the data For ii=1: number of attributes
iii. Inputattribute=inputdata(attributes) Choose number of cluster heads
iv. Initialize all variables and steps in BFO and PSO optimization algorithm
v. For 1: killing/dispersion steps For 1: reproduction steps For 1: swarming steps
vi. For 1: chemotactic steps
vii. Pass each initial bacteria positions to objective function Calculate the MSE using equation:

\[ \text{MSE} = \sum_{i=1}^{K} \sum_{x \in c_i} \text{dist}(C_i, X)^2 \]

viii. Update the bacteria’s positions as described in section 3.2 Call objective function again
ix. Chemotactic steps end Swarming steps end
x. Assign bacteria position as the current position in PSO
xi. 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
xii. Update the velocity of swarm particles as in section 3.3.1 Direction of bacteria=velocity of swarms
xiii. Reproduce the bacteria Reproduction steps end
xiv. Kill the bacteria if random value if greater than probability, otherwise disperse it
xv. Elimination/dispersal steps end
xvi. Assign the data of attribute to corresponding cluster which will be nearest to them
xvii. Repeat all above steps for all attributes in the data
xviii. Calculate the f-measure

V. 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 5.1 have been used which are taken from ftp://ftp.ics.uci.edu/pub/machine-learning databases.

<table>
<thead>
<tr>
<th>Table 5.1: Dataset Names</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td>Liver disorder</td>
</tr>
<tr>
<td>Iris</td>
</tr>
<tr>
<td>Wine</td>
</tr>
<tr>
<td>Glass</td>
</tr>
<tr>
<td>Thyroid</td>
</tr>
<tr>
<td>CMC</td>
</tr>
</tbody>
</table>

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 5.2.
Table 5.2: Fitness Value Function plot for Iris Attributes

Table 5.3: F-measure Comparison

<table>
<thead>
<tr>
<th>Data</th>
<th>F-meas.</th>
<th>Fmax</th>
<th>Favg</th>
<th>Fstd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liver</td>
<td>KM .519459</td>
<td>.959984</td>
<td>.788855</td>
<td>.210787</td>
</tr>
<tr>
<td></td>
<td>Pro .474188</td>
<td>.959936</td>
<td>.724632</td>
<td>.217614</td>
</tr>
<tr>
<td>Iris</td>
<td>KM .566273</td>
<td>.959984</td>
<td>.798872</td>
<td>.189462</td>
</tr>
<tr>
<td></td>
<td>Pro .565892</td>
<td>.959984</td>
<td>.775703</td>
<td>.170778</td>
</tr>
<tr>
<td>Wine</td>
<td>KM .421926</td>
<td>.791354</td>
<td>.609064</td>
<td>.102975</td>
</tr>
<tr>
<td></td>
<td>Pro .436564</td>
<td>.764669</td>
<td>.589900</td>
<td>.095845</td>
</tr>
<tr>
<td>Glass</td>
<td>KM .292118</td>
<td>.698191</td>
<td>.389626</td>
<td>.114299</td>
</tr>
<tr>
<td></td>
<td>Pro .332840</td>
<td>.689607</td>
<td>.399367</td>
<td>.104201</td>
</tr>
<tr>
<td>Thyx</td>
<td>KM .705597</td>
<td>.813951</td>
<td>.749209</td>
<td>.041964</td>
</tr>
<tr>
<td></td>
<td>Pro .608734</td>
<td>.834881</td>
<td>.716553</td>
<td>.087047</td>
</tr>
<tr>
<td>CMC</td>
<td>KM .395856</td>
<td>.503563</td>
<td>.454815</td>
<td>.041383</td>
</tr>
<tr>
<td></td>
<td>Pro .395856</td>
<td>.503725</td>
<td>.456368</td>
<td>.035027</td>
</tr>
</tbody>
</table>

Table 5.4: F-measure bar plots for each dataset
VI. CONCLUSION & DISCUSSION

6.1 Conclusion
In this dissertation, the problem of initializing centroids randomly by K-clustering has been solved by introducing an evolutionary clustering optimized KM technique. To optimize KM, an algorithm is proposed in which data clustering concept of K is utilized. Hybrid BFO and PSO are used. Cluster heads position optimized by BFO serves as input for particles’ positions in PSO and it updates the velocity of particles as per the velocity update formula of PSO. The position of centroids is initialized randomly only but later on the position changes as per tuning method of respective optimization technique. Comparison of results with basic KM algorithm is done in terms of F-measure. Standard deviation comparison as in Figure 5.2 shows improvement.

6.2 Future Scope
There are several different ways to extend the output results. First, the current model can deal with only a simple case of basic KM clustering in which the initial centroids are chosen randomly but there can be other ways to initialize the centroids, e.g., multiple runs, hierarchical clustering, etc. It could be considered a research study in itself to find a method of choosing these points. Second, the algorithms (BFO and PSO) combined in the proposed work to optimize the basic KM can be replaced by other combination of genetic algorithms and comparison among BFO, PSO and BPSO can also be done.

ACKNOWLEDGMENT
The author would like to thank her guide for her guidance and her family for their cooperation and moral support. This would not have been possible without their proper guidance. The author is thankful to all sources of information for their help in completing research work.

REFERENCES


[15]. Sandeep U. Mane, Pankaj G. Gaikwad,” Hybrid Particle Swarm Optimization (HPSO) for Data Clustering”;


