International Journal of Contemporary Research in Computer Science and Technology (IJCRCST) ISSN: 2395-5325
Volume 3, Special Issue 3 (September '2017)
Proceeding of - International Conference on Recent Trends in Computational Life Science and Information Technology,
Conference held at Tiruppur Kumaran College for Women,Tirupur,Tamilnadu, India.

MBAT OPTIMIZATION TECHNIQUES TO IMPROVE CLUSTER PERFORMANCE AND REDUCE THE PROCESS TIME IN HEALTH CARE DATASET

V.Shanu,
M.Phil Scholar,
Department of Computer Science ,
Dr.SNS Rajalakshmi College of Arts and Science,
Coimbatore,Tamil Nadu, India.

S.Vydehi,
Professor & Head Of the Department,
Department of Computer Science,
Dr.SNS Rajalakshmi College of Arts and Science,
Coimbatore,Tamil Nadu, India.

Abstract: Medical data mining is an active research area in the present scenario. Medical data analysis and disease diagnosis have great impact on several medical systems. This includes heart disease prediction, diabetes detection, cancer and other type of health disorders. Data mining is the optimal choice to accomplish those processes in medical dataset. Optimal clustering in health care dataset is an important task due to its huge dimensionality. Medical data clustering emerges with numerous research challenges like clustering accuracy, delay, and minimizing intra cluster distance. In this paper, we propose a novel technique to perform optimal clustering on two different medical datasets heart disease and liver disease. To improve the cluster performance and accuracy, an optimization algorithm is used. The proposed system increases the cluster quality by deploying a hybrid technique which combines weighted fitness firefly (WFF) and Modified BAT (MBAT) Optimization Techniques. The modified BAT (MBAT) technique reduces the time utilization and WFF finds the optimal feature for cluster. Instead of random move of firefly, the optimal movements are identified and performed first. The MBAT is mainly used to reduce the multimodal optimization problems by applying the hybridization techniques. The results and experiments generated. And the proposed system shows the improvement on accuracy, specificity, and consistency etc.

Keywords: ***Healthcare data Clustering, Meta-heuristic Algorithm, Firefly algorithm, Optimization Techniques, Feature selection, BAT algorithm***.

I. INTRODUCTION

Clustering is a most important unsupervised machine learning technique widely used for all type of applications. Clustering algorithms have been applied to a wide range of problems such as data mining, pattern recognition, data compression, machine learning [1], etc. and clustering is applied for different types of data such as documents, health dataset, educational datasets and spatial dataset etc., Due to this different types of data availability, defining total number of cluster is inconvenient. When the number of clusters, K, is known a priori, clustering may be expressed as allocation of n objects in N dimensional space among K groups in such a way that objects in the same cluster are more similar in some aspects than the others in different clusters. This involves minimization of some optimization criterion. The K-means algorithm [2], starting with k random cluster centers then partitions a set of objects into k subsets. This method is a one the most popular and simple method that widely used in clustering. However, the k-means clustering has several drawbacks such as being attentive in local optima, as well as local maxima and being sensitive to initial cluster centers. One method to refined k-means algorithm is hybridizing it with efficient optimization method. There is different optimization algorithm like Particle Swarm Optimization (PSO) [3], Ant Colony Optimization (ACO) [4], Artificial Fish Swarm Algorithm (AFSA) [5] and Bee Colony [6]. The FireFly algorithm (FFA) was recently [7]. This swarm intelligence optimization technique is based on the assumption that solution of an optimization problem can be shown as a firefly which glows proportionally to its quality in a considered problem setting. Consequently, each brighter firefly attracts its partners, which makes the search space being explored efficiently. Yang used the FFA for nonlinear design problems [8] and multimodal optimization problems [9] and showed the efficiency of the FFA for finding global optima in two dimensional environments. In this paper, we use the firefly algorithm to find initial optimal cluster centroid and then initial k-means algorithm with optimized centroid to refined them and improve clustering accuracy. Proposed method experimental results compared with PSO [10], K-means, K-PSO method on standard datasets of Iris, WDBC, Sonar, Glass and Wine. The results show that the proposed algorithm has a higher efficacy than the other algorithms.

In recent days continuous monitoring of different medical issues like liver disorder features, blood pressure, temperature, heart rate, respiratory rates, ECG, and Electromyography (EMG) can be done easily. These data can be stored for future analysis and diagnosis. But the large amount of data entails with huge storage space. Another aspect is that all data are not imperative except a few significant data, which possess a big challenge in disease analysis. To make the disease analysis in an effective manner, the data should be pre-processed, conditioned, clustered or classified. The study aims to improve the cluster performance and cluster quality through utilizing the FFA. And the results of FFA will be applied as a weighted feature

for the next iteration into the MBAT algorithm. This iterative result improves the clustering efficiency and reduces the time for feature selection.

II. LITERATURE REVIEW

The different data mining classification techniques were tested on variety of healthcare datasets such as, PIMA Indian Diabetes dataset, Wisconsin Breast Cancer dataset, BUPA Liver Disorder dataset, Stat Log Heart Disease dataset. The clustering was not more accurate based on the pure density based clustering algorithm. This method was appropriate only for limited dataset. Natures inspired algorithms are widely used to find best solutions to various optimization problems. Some of the examples of them are genetic algorithm [11], and ACO, simulated annealing, differential evolution, PSO and BCO. All these algorithms though work well but they suffer from some issues. Taking this into account a new category of evolutionary optimization algorithm has emerged. Firefly algorithm, Cuckoo search [12], Bat algorithm [13] and krill herd algorithm [14]. In this paper a new hybridization of meta-heuristic algorithms are used.

Table 1.0 meta-heuristic algorithm comparison table

Algorithm	Firefly	Bat algorithm
Year	2008	2010
Develop by	X.S Yang	X.S Yang
Based on	Flashing behavior of fire fly	Echo location behavior of micro bat
Objective function defined by	Brightness(light intensity) and attractiveness	Pulse rate emission and velocity
Features	High convergences rate, robust rate. Finds good optimum solutions in less number of iterations.	Accurate and efficient
Area of Application	Quadratic assignment problem, Travelling salesmen problem, digital image processing	Engineering design and classification

The existing clustering framework requires repeated re-clustering and cluster size specifications, when the data with an incrementally grows. This can be computationally demanding for large uncertain data sets. To address this problem, an effective feature selection and clustering method is proposed. An alternative way to lower the computational cost is to reduce the number of iterations by applying the effective feature selection process, which selects a set of points to as weighted features from large and uncertain dataset.

III. PROPOSED WORK

In order to reduce the feature selection time, clustering delay and the problem in processing huge medical data, the system introduces a three step implementation, which performs effective pre-processing, weighted Feature selection using GA algorithm, Weighted Fitness Firefly (WFF) and Modified BAT technique for increasing the cluster quality and to reduce the cluster delay. This uses an alternative way to lower the computational cost by reducing the number of iterations by performing the three phase works. The improved pre-processing technique calculates the Similarity & Dissimilarity probability Distribution of the cluster by applying a new Combined KL Divergence & Shannon Entropy Distance Measures. This performs the effective pre-processing steps in the given uncertain data sets.

- For accurate clustering process, a well known Feature selection algorithm is taken and modified as Forward GA algorithm. This significantly increases the accuracy of the proposed clustering process.
- Feature Selection and Optimization using Weighted Fitness Firefly (WFF) will be used and finally the improved and optimized clustering.

From the above contributions, the proposed system details are well studied. The further chapters and sections will discuss about the detailed analysis of the above stated methodologies.

The proposed system implements a new fusion based approach to improve the clustering efficiency and accuracy. So this includes 3 steps initially. The followings are the steps involved with the proposed system. One is pre-processing, feature selection and clustering process. The fast and effective clustering needs a fine grained dataset, and this need to be ranked and effective features and a best clustering algorithm. The proposed system has all the above features.

Fig1.0 overall process diagram of the proposed system

A. Data collection and pre-processing

Data collection and pre-processing steps are important tasks of machine learning process. Especially in high dimensional data environment, there is a necessity to avoid noisy and redundant data for better and accurate results. The collected data may contain incomplete, noisy and inconsistent datasets. The first phase collects data from UCI repository and performs the preprocessing steps. To be efficient in the ML process, feature selection requires the three procedures the current pre-processing step performs Dissimilarity between objects and inter-cluster distances. This is has been performed to improve the distribution of the clusters. So, it considered and applied the optimal pre-processing approach which is Combined KL Divergence & Shannon Entropy Distance Measures.

Similarity function provide guarantee about the desired clustering and make it possible for many unsupervised learning algorithms to increase their performance. For any pair of observations *xi*, *xj* in *XL*, the different types of similarity function that we can generate, are:

• **Similarity function (SF)**: involving *xi* and *xj*, specifies that they have the same label.

• **Dissimilarity function (DF)**: involving *xi* and *xj*, specifies that they have different labels.

SF and *DF* similarity function are then grouped into two defined subsets *_SF* and *_DF*, respectively. The similarity function is explained in the following sections.

While it was expected that different similarity sets would contribute more or less to improving accuracy of many similarity and dissimilarity calculation formulas and functions, it was found that some existing pre-processing techniques actually decrease the performance. It was observed that control over the dataset at the time of pre-processing can have unwell effects even when they are generated from the data labels that are used to evaluate accuracy, so this behavior is not caused by noise or errors in the implementation. Instead, it is a result of the similarity and as well as dissimilarity between a given set of constraints and the algorithm being used. So it is more important to know the similarity and dissimilarity functions increase clustering accuracy while others have no effect or even decrease accuracy. For this, the proposed system utilizes two important measures, KLSE -Kullback-Leibler & Shannon Entropy- that capture similar and dissimilar objects. These measures provide insight into the effect a given constraint set on a specific feature selection and clustering algorithms. In this paper, the pre-processing stage is performed with the above two functions.

B. Feature Selection Process:

After successful implementation of the KLSE the feature selection process is performed. Feature Selection is the process of extracting a subset of features from an original dataset. By utilizing a new fusion algorithm that combines the KLSE with GA algorithms. This fusion method significantly increases the accuracy of the clustering process in the next level. In this work, **GA and WFF** method is used for feature selection.

Steps of GA+KLSE

1.Choose initial object I from KLSE

2. Evaluate the fitness of each individual in the population using

$$E = \sum_{k=0}^{n} \binom{n}{k} xF^{k}$$

3. Repeat until termination: (time limit or sufficient fitness achieved)

a. Select optimum -weighted individuals to reproduce

b. Breed new generation through crossover and/or mutation (genetic operations) and give origin to offspring

c. Evaluate the individual fitness of the offspring

d. Replace non-weighted part of population with offspring

The feature selection process has begun with the output of the KLSE phase. With the similar and dissimilar object detection, the complexity of the feature selection process is reduced in this stage.

Mining optimal features is not an easy task when there is a huge set of medical objects. It has some challenge and algorithmic complexity. The number of objects grows exponentially with the number of items. But this complexity is tackled with some latest algorithms which can efficiently prune the search space. Secondly, the problem of finding results from rules, i.e. picking optimal results from set of outputs. In general the main motivation for using GAs in the feature selection is that they perform an optimal search and cope better with objects interaction than the existing algorithms often used in data mining. The use of KLSE in feature planning helps to predict optimal feature set based on the selected output of the KLSE. This section discusses several aspects of GAs for feature selection. The main areas of discussion include individual representation of feature points Genetic Operators involved and the choice of Fitness function. In KLSE_GA, an initial population consisting of a set of solution is chosen and then the solutions are evaluated. Relatively more effective solutions are selected to have more off springs which are, in some way, related to the original solutions. If the genetic operator is selected accurately then the final population will have better solutions so the genetic operator chooses an optimal gene from multiple iterations. GA improves the whole population. KLSE aims at producing one best solution. For the KLSE, we require several good initial solutions to ensure the required number of good initial solution.

In specific, a Firefly Algorithm (FA) is a recent nature inspired optimization algorithm, which simulates the flash pattern and characteristics of fireflies. In this paper, the **Weighted Fitness Firefly** is used for clustering process. WFF has been used for solving nonlinear optimization problems. Usually, the firefly the intensity decrease as their mutual distance increases.

Weighted Fitness Firefly (WFF)

Objective function
Get the initial population of fireflies fi(x=1, 2…n) from genetic function
Formulate intensity I and it belongs to
Define absorption coefficient
While (t < MaxGeneration)
For i = 1 : n (all n fireflies)
For j = 1 : n (n fireflies)
If (Ij>Ii),
Vary attractiveness with distance r via exp(- γr);
Move firefly i towards j;
Evaluate new solutions and update intensity;
End if
End for j
End for i
Grade fireflies according to the fitness and find the current best;
Cluster C.
End while
End

After implementing the weighted fitness firefly, the Modified bat algorithm, (MBAT) is used. This algorithm is based on BAT, which is developed on echo location behavior of micro bats. It is based on three important rules. For sensing distance, Mbat uses its‟ echolocation capacity. It also uses echolocation to differentiate between food and prey and background barriers even in the darkness. Bats

used to fly randomly with some characteristics like a velocity, fixed frequency and loudness to search for a prey. But in the MBAT, it fly based on the weighted feature generated from the WFF. It also features the variations in the loudness from a large loudness to minimum loudness. Bats find the prey using varying wavelength and loudness while their frequency, position and velocity remains fixed. They can adjust their frequencies according to pulse emitted and pulse rate.

The pseudo code of Modified Bat Algorithm (MBAT) is shown below.

1. Define objective function F, attribute A.
2. Initialize the population of the bats
   Select initial population form the **WFF**
3. Select the parameters P from the WFF
4. While (Termination criterion not met)
   {
   Generate the new solutions from the WFF
   If (Pulse rate (A (WFF)) >current)
   Select a solution among the best solution
   Generate the local solution around the selected best ones.
   End if
   Generate a new solution by flying randomly
   If (loudness & pulse frequency (A (WFF))) < current)
   Accept the new solutions increase pulse rate and reduce loudness
   End if
   Rank the bats and find the current best
   }
5. Results and visualization

The algorithm starts with initialization of population of Mbats. Each Mbat is assigned a starting position which is a optimal feature generated form he WFF. The pulse rate and the loudness are defined from the feature selection process. Every Mbat will move from local solutions to global best solutions after every iteration based on the feature score. The values of pulse emission and loudness are updated if a bat finds a better solution after moving. This process is continued till the termination criteria are satisfied. The solution so achieved is the final best solution.

IV. EXPERIMENT AND RESULTS

The experiment is based on two datasets one is heart dataset and another one is liver dataset. With these dataset the performance of the work is analyzed. The entire data set has been taken for the experiment. The experimental results are depicted in Figures 2.0 and 3.0 and the data are tabulated in Table 1.0 for feature selection before optimization and feature selection after optimization. True Positive Rate is the proportion of datasets identified correctly by the WFF to the total datasets. False Positive Rate is the proportion of datasets not identified by the WFF to the total datasets.

The experiments are designed so that the different parts of the work could be evaluated. These include the evaluation of the features of the above dataset, the feature selection and also the feature creation methods. To this aim, first the features which were selected by the feature selection method named as WGA and their importance are discussed. Second, all the four possible combinations of the feature selection and creation methods are theoretically analyzed over the dataset. Finally WGA and genetic algorithms are implemented this proposed work was implemented using C#.net. The performance of this proposed work WGA Scheme was compared with the existing algorithms based on the following parameters.

- **Specificity** –measures the proportion of negatives that are correctly identified.
- **Sensitivity**- measures the proportion of positives that are correctly identified
- **Accuracy** – Determines the correctness
- **Precision** –Repeated process same result
- **Time taken** – Determines the processing time involved.

Sensitivity, specificity and accuracy are described in terms of TP, TN, FN and FP.

A. Specificity:

The sensitivity parameter measures the negatives that are correctly identified. In the given dataset d, the **specificity** is the ability of the test to correctly identify those without the disease (true negative rate) from the given dataset.

Specificity = TN/(TN + FP) = (Number of true negative assessment)/(Number of all negative assessment)

B. Sensitivity:

In health care analysis, sensitivity is the ability of a test to correctly identify those with the disease (true positive rate),

Sensitivity = TP/(TP + FN) = (Number of true positive assessment)/(Number of all positive assessment)

A. Accuracy:

The accuracy is the

Accuracy = (TN + TP)/(TN+TP+FN+FP) = (Number of correct assessments)/Number of all assessments)

Table 1.0 the comparative table with different metrics

Metrics	GA	WFF_MBAT
Specificity	0.89	0.92
Sensitivity	0.91	0.94
Accuracy	0.91	0.95
Precision	0.92	0.94

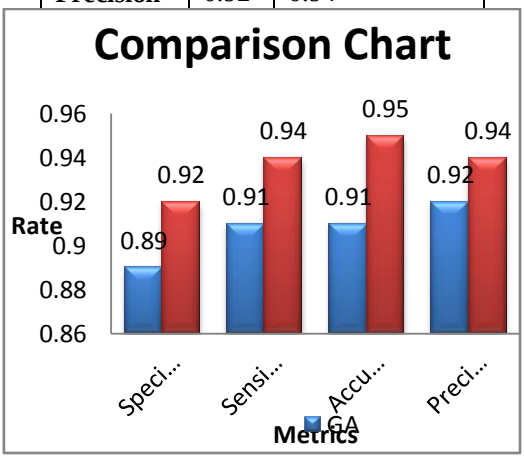

Figure 2.0 comparative chart

From the results shown in the graphs fig 2.0, it can be observed that the proposed WFF+MBAT, a hybrid based approaches provides better accuracy and increased true positive rate when it is analyzed with different type of datasets.The system finally performs the analysis to show the accuracy of the proposed system. The accuracy is calculated by true positive, true negative, false positive and false negative values.

V.CONCLUSION

Data clustering is the process of grouping objects into similar clusters. These clustering processes are usually called as unsupervised learning process. A perfect clustering method will make high quality clusters with high in intra class similarity and low inter class similarity. The quality of a clustering result depends on both the similarity and dissimilarity measure used by the method and its implementation. The quality of a clustering method is also measured by its ability to discover some or the entire hidden pattern, but the proposed system performs the pre-processing step by calculating both similarity and dissimilarity values Clustering large datasets has some important issues like excessive time, Computational complexity and so on. In the present work the capabilities of naturally inspired algorithms that are used efficiently to optimize the performance. The proposed system creates a new hybrid based approach which combines a set of algorithms for effective clustering. The system proposed a new Improved and Optimized Clustering technique with several data mining methods. The proposed system uses KLSE+GA+WFF+MBAT to bring the effective clustering results than the existing algorithms. Currently the implementation works on liver and heart disease dataset collected from UCI repository. The assumption that is made in proposing this algorithm is that the other type of clinical data can be applied with high dimensional attributes. The algorithm can be extended to incorporate qualitative/categorical dataset with huge dimensionality. The weighted feature selection process can be tested with the other types of evolutionary algorithms.

VI.REFERENCES

[1] Witten, Ian H., et al. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[2] Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. "Data mining cluster analysis: basic concepts and algorithms." *Introduction to data mining* (2013).

[3] Kennedy, J.; Eberhart, R. (1995). "Particle Swarm Optimization". Proceedings of IEEE International Conference on Neural Networks. IV. pp. 1942–1948

[4] D. Karaboga, B. Basturk, "A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm", Journal of Global Optimization 39 (2007) 459–471

[5] Zhang, Chao, et al. "Improved artificial fish swarm algorithm." *Industrial Electronics and Applications (ICIEA), 2014 IEEE 9th Conference on*. IEEE, 2014.

[6] Kefayat, M., A. Lashkar Ara, and SA Nabavi Niaki. "A hybrid of ant colony optimization and artificial bee colony algorithm for probabilistic optimal placement and sizing of distributed energy resources." *Energy Conversion and Management* 92 (2015): 149-161.

[7] Fister, Iztok, Xin-She Yang, and Janez Brest. "A comprehensive review of firefly algorithms." *Swarm and Evolutionary Computation* 13 (2013): 34-46.

[8] Yang, X.-S.: "Nature-Inspired Metaheuristic Algorithms ". Luniver Press, (2008)

[9] X.S. Yang, "Firefly algorithm, stochastic test functions and design optimization," International Journal of Bio-Inspired computation, vol. 2, no. 2, pp. 78-84, 2010.

[10] X.S. Yang, "Firefly algorithm, levy flights and global optimization," in Research and Development in intelligent systems XXVI. Springer, 2010,pp.209-218.

[11] Johnson, Eric G., et al. "Advantages of genetic algorithm optimization methods in diffractive optic design." *Critical Review Collection*. International Society for Optics and Photonics, 2017.

[12] X.-S. Yang, S. Deb, "Cuckoo search via L´evy flights", in: Proc. Of World Congress on Nature & Biologically Inspired Computing (NaBIC2009), December2009, India. IEEE Publications, USA, pp. 210-214(2009).

[13] P. Musikapun, P. Pongcharoen, " Solving Multi-Stage Multi- Machine Multi-product Scheduling Problem Using Bat Algorithm", 2nd international Conference on Management and Artificial Intelligence, IPEDR Vol.35, 2012

[14] Wang, Gai-Ge, et al. "A hybrid meta-heuristic method based on firefly algorithm and krill herd." *Handbook of Research on Advanced Computational Techniques for Simulation-Based Engineering*. IGI Global, 2016. 505-524.