Optimized Feature Selection Using IBAT

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

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

Abstract – Data mining is applied in almost all type of applications like ecommerce, business, education and health care. The most prominent area of data mining is handling high dimensional datasets. From the numerous health attributes, disease diagnosis, prediction and its risk analysis are performed by data mining techniques. The Intelligent health care systems are performing the data mining techniques for effective data analysis and management also. From the numerous health data, the proposed system handles different types of multi dimensional dataset with effective feature selection and optimal clustering. Even though there are several approaches in data mining is introduced to handle the dimensionality, still some issues arises while clustering those high dimensional datasets. So, the system proposes a new iterative approach, which concentrates on the effective feature selection and Clustering. The system concentrates on three main portions for accurate Clustering. One is Effective pre-processing, feature selection and Clustering. The second stage is the feature selection process, which performed using PCA (Improved Principle Component Analysis) and effective Clustering using IBAT (Improved BAT Algorithm). The system implements a new Improved BAT based algorithm with the use of effective weighted features from the IPCA. The system experimented with different types of high dimensional datasets from the UCI machine repository using IBAT technique. The system developed with the intension of high accuracy and less cluster overhead.

1. INTRODUCTION

Categorizing the data into a meaningful group is one of the fundamental ways of understanding and learning the valuable information. High-quality clustering methods are necessary for the valuable and efficient analysis of the increasing data. Clustering has a wide range of application in all the fields, namely pattern recognition, document categorization, bioinformatics applications, data compression, machine learning etc. Clustering is a renowned method in identifying fundamental structures and discovers useful information from a large amount of data. Data clustering is a tentative process for data analysis, where similar elements are identified and gathered into a set of elements called a cluster. Cluster analysis is an unsupervised classification technique with an objective to analyze the similarity of data and divide it into set natural clusters without any previous knowledge of the element [1]. Consider a clustering problem where a given dataset is partitioned into a certain number of natural and homogeneous subsets such that each subset is composed of elements similar to one another but different from those of any other subset.

1.1 Feature Selection:

Feature selection is one of the prominent preprocessing steps to machine learning. Feature selection is a process of selecting a subset of original feature set, so that the feature space is condensed according to a certain evaluation criterion. Feature selection has been a important part of research in data mining and it is very valuable and useful to handle the high dimensional inconsistent featured dataset. This improves the predictive accuracy brings better results. In present day applications such as genome projects, image retrieval, and customer relationship management, text categorization, the size of database has become exponentially large. This immensity may cause serious problems to many machine learning algorithms in terms of efficiency and learning performance. A high dimensional data can enclose high degree of redundant and irrelevant data which may greatly influence the performance of learning algorithms. So, while dealing with high dimensional data, feature selection becomes highly necessary. In the proposed system, the effective feature selection is developed to cluster high dimensional data's. Usually, the Feature selection algorithms can be divided into two broad categories, namely, the filter model and the wrapper model. The filter model relies on general characteristics of the training data to select some features without involving any learning algorithm. The wrapper model relies on a predetermined learning algorithm and uses its performance to evaluate and select the features. For each new subset of features, the wrapper model needs to learn a classifier. It tends to give superior performance as it finds tailor-made features which are better suited to the predetermined learning algorithm, but it also tends to be more computationally expensive, For the increased number of features, the filter model is usually a choice due to its computational efficiency. Different feature selection algorithms under filter model can be further classified into two groups, namely subset search algorithms and feature weighting algorithms. Feature weighting algorithms allocate weights to features individually and grade them based on their relevance to the objective. A feature will be selected subject to

a threshold value. If its weight of relevance is greater than a threshold value, the corresponding feature will be selected.

2. LITERATURE SURVEY

2.1 Metaheuristic Algorithms-General

The complete concepts of Evolutionary Calculation, Particle Swarm Optimization, Evolving Fuzzy Systems, Evolving Artificial Neural Networks and examples of recent applications of these methods have been discussed in literature. Authors in [2] have written a detailed note on various meta-heuristic methods along with the algorithms and their application in solving various combinatorial problems in their book. The book also highlights the beneficial features of the methods and the suitable problem areas so that the researchers find it useful for selecting the methods to solve the problems of their own. In [3], the author Mitchell has started with an overview of Genetic Algorithms (GA) and continued with the role of GA in problem solving and in building scientific model. The book also encompasses the theoretical foundations of GA. Integer programming has benefited from many innovations in models and methods. Glover [4] has suggested some of the promising directions for elaborating these innovations in the future from a framework that links the perspectives of artificial intelligence and operations research. To demonstrate this, four key areas have been examined: (1) controlled randomization, (2) learning strategies, (3) induced decomposition and (4) Tabu search. Each of these has been shown to have characteristics that appear usefully relevant to developments on the horizon. The authors have constructed an approach to cancer class prediction from gene expression profiling, based on a development of the simple nearest prototype (centroid) classifier [5]. The authors have shrunk the prototypes and hence obtained classifier that is often more accurate than competing methods. This method of "nearest shrunken centroids" has identified subsets of genes that best characterize each class.

2.2 Metaheuristic Methods in Feature Selection

Authors have proposed a novel classifier-independent feature selection method on the basis of the estimation of Bayes discrimination boundary. The experimental results on 12 real-world datasets have showed the fundamental effectiveness of the proposed method [6]. Authors have discussed the concepts of Feature Selection, its evaluation procedures and their applications in various fields. They have studied the problem of selecting an optimal feature-set for different dataset with the help of classification based on. It has been added and showed that the pooling features resulting from different texture models followed by a feature selection has resulted in a substantial improvement in the classification accuracy[7].

Sidelecki and Sklansky have devised and tested mathematical techniques for high speed detection and classification of multiple targets in [8]. They have developed an automated method for designing tree classifiers that classify targets with both near optimal accuracy and high speed, an algorithm for designing a time-varying classifier that is suitable for classifying multiple objects in cinematic sequences of images, devised efficient mapping algorithms for assigning computational tasks to hypercube-connected multiprocessors so as to minimize the task completion time and devised and tested genetic algorithms for automatic selection of small number of features for pattern classification. Practical pattern-categorizing and knowledge-discovery problems have required the selection of a separation of attributes or features to represent the patterns to be categorized. The author in the paper [9] has used a genetic algorithm to select such feature subsets to achieve multi-criteria optimization in terms of generalization accuracy and costs associated with the features.

Almuallim and Dietterich have presented five algorithms that identify a subset of features sufficient to construct a hypopaper consistent with the training examples [10]. They have proved that FOCUS-1 is a straightforward algorithm that returns a minimal and sufficient subset of features in quasi-polynomial time. An information-theoretic approach has been used to derive a new feature selection criterion capable of detecting features that are totally useless in [11] by Sheinvald et al. A feature subset selection algorithm based on branch and bound techniques has been developed by Narendra and Fukunaga to select the best subset of m features from an n-feature set [12]. The algorithm has been found to be very efficient and it has selected the best subset without exhaustive search. Computational aspects of the algorithm have been discussed here. Results of several experiments have demonstrated the very substantial computational savings realized.

3. PROPOSED SYSTEM AND RESEARCH METHODOLOGY

The system proposes a new iterative approach, which concentrates on the effective feature selection using IPCA (Improved Principle Component Analysis) and effective Clustering using IBAT (Improved BAT Algorithm). It is very difficult to find out a single feature selection method performing very good for any data set. Thus, rigorous validation is required to confirm the effectiveness of the feature selection method in a given context. Such validation will be expensive. So, to reduce the complexity of the search for optimal feature sub set, some regularity are to be found and use them as heuristics. In general, Feature Selection deals with data sets that contain large number of irrelevant and redundant attributes. Heuristic search is the convenient approach for selecting features explicitly. Heuristics are the search guides that facilitate decision making while selecting a feature into the subset. The essential criteria which can be served as best heuristics are finding Redundancy & Similarity and Learning & Prediction Efficiency. All heuristic search processes work around the above mentioned criteria.

- The primary issue to be concentrated in determining the nature of the heuristic search process is, to determine the IS(Initial State) in the space, which in turn influences the order of search and the possible operators that can be used to generate the successor states. One might start with an empty set(0) and successively add features(Forward Selection). Or one might start with the set of all features and successively removes features(Backward Selection).
- Organization of search is directly connected with the computational complexity. It is not feasible to adopt exhaustive search as it has to search 2" feature subsets of 'n' features. Greedy approach is the more realistic approach for optimum feature subset selection. By keeping Greedy approach as the backbone, simple hill climbing or steepest ascent hill climbing can be followed. To achieve more sophistication, BFS (Best First Search) can also be used though it is expensive in some domains.
- The third key issue to be considered is concerned with the evaluation strategy that has to be applied to evaluate the alternate feature subsets. 'Concepts of Information Theory and strait forward evaluation of training set' are the better alternatives of this.
- Stopping criteria: For halting the search process, criteria have to be decided. Normally feature selection process has to be continued as long as there is no significant change in the classification performance.

Several Heuristic criteria includes, finding Best feature under the feature independence assumption- based on significance tests, Best step-wise feature selection, The best single-feature is picked first, Then next best feature with respect to the first, and so on, Step-wise feature elimination, Repeatedly eliminate the insignificant feature, Feature selection and elimination by Hybrid methods, Optimal branch and bound and performing feature elimination with backtracking. The followings are the main contributions of the proposed work.

- The system implements a new BAT based algorithm with the use of effective weighted features from the PCA. The system introduces a new High dimensional data clustering algorithm with IBAT technique.
- This also creates a new advanced Clustering for fast data clustering. The system developed with the intension of high accuracy and less training overhead.
- High dimensional data clustering and prediction of the class score using a renovation algorithm which is a combination of PCA (Improved Principle Component Analysis) and Improved BAT algorithms.
- IPCA for feature selection and dimensionality reduction
- And BAT algorithm for data Clustering and class prediction.

The optimized feature selection algorithm has been expanded with the new optimal Clustering algorithms, which can handle large dataset more rapidly, accurately and effectively, and keep the good scalability at the same time. The algorithm mainly aims at classified data, but it should disperse the value data in the dealing process. The main advantage of using IPCA is, it only requires minimum number of training dataset and it can effectively reduce dimensionality problems.

4. METHODOLOGY

Optimization algorithms are planned to improve the effectiveness or to reduce the cost of clustering algorithms. To reduce computational cost now-a-day's most of the pattern recognition problems use various optimization techniques. In this section, IBAT algorithm is used as an optimization technique to enhance the performance of an existing BAT model. This gives attentiveness to reduce the time complexity of the model that with the minimum number of iterations and also the accuracy has been enhanced as compared with traditional algorithms. The ultimate goal of any pattern recognition system is to achieve the best possible clustering performance for a given problem domain. Meta-heuristic algorithms like PSO and Simulated Annealing are powerful methods for solving many optimization problems. The fine adjustment of the parameters of the above techniques enhances the accuracy of the algorithms. BAT algorithm is based on the echolocation behavior of BATs. The proposed model has been compared with three types of clustering techniques on different datasets BAT and PSO, and Fitness Firefly. Like PSO, the position of a BAT is also updated by taking into consideration of velocity and frequency of BAT. In this section, a novel fusion approach has been proposed, where the model will decide the applicable method. Experimental results show that the proposed clustering technique with IBAT algorithm is superior and faster from others. The proposed Clustering is a semi-supervised approach where input data is partially trained and tested to predict class label of data element. The unseen data is used as input to the clustering algorithm which measures the performance of an algorithm. The objective is to boost the accuracy of clustering algorithms with minimum number of iterations. From the literature this found that the algorithm which gives good accuracy suffers from time complexity. To address the problem of accuracy with consideration of time complexity this section proposed a new algorithm known as IPCA_IBAT clustering algorithm. This section uses various gene expression datasets to evaluate the performance of the proposed algorithms.

The total working procedure of IPCA-IBAT algorithm is given below:

Step 1: Guarantee that the algorithm is going to execute number of iterations as per user input. Weight *wt* for IPCA is randomly assigned and gets updated after iterations.

Step 2: Variable *j* represents j^{th} feature of dataset *D*.

Step 3: *Calculation of Frequency*: Let us denote the frequency of sound as *f* for a IBAT *B*. So, in order to find the frequency f_k of IBAT B_k can be computed using equation (1.1). Where c_1 is the *pulse rate* used to control the frequency f_k of IBAT B_k, and when it reaches near or far from the object, the value of c_1 is auto adjusted in iteration by equation (3.9).

$$f_k = c_1 x \sum_{i=1}^{m} (D_{ki}) / m \quad (1.1)$$

Step 4:*Calculation of distance:* Distance *S* of the object *z* from IBAT B_k is calculated by multiplying D_k with some random weight value multiplied by f_k for each object using equation (1.2); where, *z T*. *T* is the number of class labels. For example, in figure 1.3 (a) and (b) the value of *T*=3.

$$S_{objectz} = f_x x D_k x w \quad (1.2)$$

In equation (1.2) *wt* is a weight matrix of value in between -0.5 to 0.5 of size $m \times T$. Figure 1.4 depicts the complete working procedure of algorithm.

Step 5, 6 and 7: *Compute Error, Update the position of BAT and Update Weight*: Each object denotes a class. After calculating the error by equation (1.3), the BAT position can be changed using by equation

$$E_k = S_{objectz} - 1 \quad (1.3)$$

$$P_k = P_k + E_k \quad (1.4)$$

When BAT starts flying it assumes that the position is initialized to zero. Its position keeps on changing when it reaches nearer to the object. As closer it moves to the pray, error E_k and position P_k reduces to zero. As the BAT reaches nearer to its object the frequency starts reducing. That can be done by controlling the value of c_1 in equation (1.1) by using equation (1.5). c_2 is a constant treated as the *BAT learning parameter* and chosen nearer to 0.0011. Weight wt is computed and updated using equation (1.6).

$$c_1 = f_k + c_2 \times E_k^2 \times P_k \quad (1.5)$$

$$wt = wt + 2 \times \mu \times E_k \quad (1.6)$$

IBAT algorithm is an improved meta-heuristic algorithm that overcomes the optimization problem. After implementing the Improved PCA, the improved bat algorithm, (IBAT) is used. This algorithm is based on BAT, which is developed on echo location behaviour of micro bats. It is based on three important rules. For calculating and detecting the distance, IBAT uses its echo location capacity to get the optimal distance between objects. It also uses echolocation to differentiate between food and prey and background barriers even in the darkness. In general, the Bat flies at random with some features like a velocity fixed frequency and loudness to search for a prey. But in the IBAT, it fly based on the weighted feature generated from the IPCA. It also features the variations in the loudness from a large loudness to average 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 proposed system performs the prediction model based on the above IBAT algorithm. The proposed system successfully clusters with optimal feature selection the liver and Heart disease based on the given dataset. The system also predicts the score for the chance of Heart disease based on the boundary calculation. The proposed system implements an optimal clustering which does not depend on the features from IPCA completely. The system performs the statistical properties to evaluate the score of every attribute. The system finally provides the prediction accuracy over the given dataset.

5. EXPERIMENTAL EVALUATION

The experiment process is carried on the computer having Pentium processor with speed 2.6 GHz and 2 GB of RAM. There are different tools have used for the experiment. The existing techniques are verified in Weka and the proposed system is implemented using C#.net and Matlab.

Table 1.0 Performance comparison table

Type	BAT	IPCA_IBAT
Feature selection Time(s)	**8**	**1.4**
Clustering Time (s)	**5**	**2.6**

As in figure 1.0 the clustering delay of the proposed system is quite reduced than the existing algorithms. Due to the dimensionality reduction using IPCA and IBAT reduces clustering delay. For all the datasets used to evaluate the performance of IBAT algorithm there is 20 - 40 % improvement in every metric value when compared to the BAT algorithm. The improvement in the performance of the IBAT clustering algorithm is due to local search capability of FCM algorithm and so the global search property of BAT algorithm

is combined in the IBAT algorithm. From the experimental results, it observed that the performance of IBAT algorithm is better when compared with the original BAT algorithms.

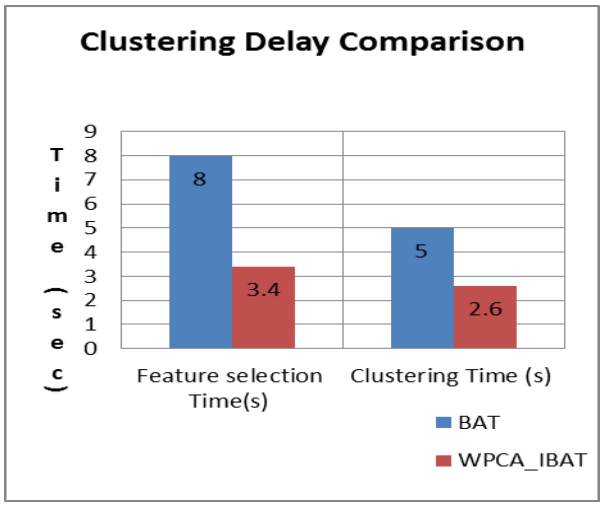

Figure 1.0 Clustering time analysis chart

6. CONCLUSION

In this paper, a new clustering scheme for multi dimensional data. The system studied the main two problems in the literature, which are feature selection, clustering accuracy and clustering delay. The study overcomes the above two problem by applying the effective enhanced improved component with IBAT algorithm. The PCA represents with the effective splitting criteria which has been verified by the IBAT algorithm. The system effectively identifies the cluster with reducing the intra cluster distance. The experimental results are evaluated using the two set of datasets. The experimental result shows that integrated extended improved ranked component with IBAT algorithm shows better quality assessment compared to traditional PCA and WFF techniques. From the experimental results, the execution time calculated for clustering object is almost reduced than the existing system. The proposed framework model can be used to analyze the existing work, identify gaps and provide scope for further works. The researchers may use the model to identify the existing area of research in the field of data mining in other dataset and use of other clustering algorithms.

REFERENCES

[1] Richards, John A. "Clustering and unsupervised classification." *Remote Sensing Digital Image Analysis*. Springer Berlin Heidelberg, 2013. 319-341.

[2] Hammouche, Kamal, Moussa Diaf, and Patrick Siarry. "A comparative study of various meta-heuristic techniques applied to the multilevel thresholding problem." *Engineering Applications of Artificial Intelligence* 21.5 (2010): 676-688.

[3] Mitchell, Melanie. *An introduction to genetic algorithms*. MIT press, 1998.

[4] Glover, Fred. "Artificial intelligence, heuristic frameworks and tabu search." *Managerial and Decision Economics* 11.5 (1990): 365-375.

[5] Golub, Todd R., et al. "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." *science* 286.5439 (1999): 531-537.

[6] Hua, Jianping, Waibhav D. Tembe, and Edward R. Dougherty. "Performance of feature-selection methods in the classification of high-dimension data." *Pattern Recognition* 42.3 (2009): 409-424.

[7] Fong, Simon, Yan Zhuang, Rui Tang, Xin-She Yang, and Suash Deb. "Selecting optimal feature set in high-dimensional data by swarm search." *Journal of Applied Mathematics* 2013 (2013).

[8] Siedlecki, Wojciech, and Jack Sklansky. "A note on genetic algorithms for large-scale feature selection." *Pattern recognition letters* 10.5 (1989): 335-347.

[9] Ghamisi, Pedram, and Jon Atli Benediktsson. "Feature selection based on hybridization of genetic algorithm and particle swarm optimization." *IEEE Geoscience and Remote Sensing Letters* 12.2 (2015): 309-313.

[10] Sangiovanni-Vincentelli, Alberto. "Constructive induction using a non-greedy strategy for feature selection." *Machine Learning Proceedings 1992: Proceedings of the Ninth International Workshop (ML92)*. Morgan Kaufmann, 2014.

[11] Song, Qinbao, Jingjie Ni, and Guangtao Wang. "A fast clustering-based feature subset selection algorithm for high-dimensional data." *IEEE transactions on knowledge and data engineering* 25.1 (2013): 1-14.

[12] Narendra, Patrenahalli M., and Keinosuke Fukunaga. "A branch and bound algorithm for feature subset selection." *IEEE Transactions on Computers* 9.C-26 (1977): 917-922.