PSO Based Fast K-means Algorithm for Feature Selection from High Dimensional Medical data set

Doreswamy
Department of Computer Science
Mangalore University
Mangalagangothri, Mangalore 574199
Email:doreswamyh@yahoo.com

Umme Salma M
Department of Computer Science
Mangalore University
Mangalagangothri, Mangalore 574199
Email:hmbssalma@gmail.com

***Abstract*—Features are the most important entity in any data mining and machine learning applications. They are the backbone of any model. Reliability, efficiency and accuracy of the model depends upon the choice of strong and relevant features. However, feature selection is always a time-consuming and challenging task. In this paper, we have proposed an approach where we combine a clustering technique and a stochastic technique to select effective features from the high dimensional breast cancer data set in quick time. In order to select strong and relevant features, we have used an improved version of K-means algorithm called fast K-means algorithm, which is much faster and more accurate than a general means algorithm. The fast K-means algorithm is embedded in Particle Swarm Optimization (PSO) algorithm to produce better results. The results were validated using various classification techniques and were evaluated on various performance evaluation measures. The results obtained were found to be highly supportive in nature. The feature subset generated using PSO based fast K-means algorithm on KDDcup 2008 data set produced an accuracy of 99.39% and its time complexity was found to be $O(log(k))$.**

Keywords*—*Data mining; Feature selection; PSO Algorithm; Fast K-means; Breast Cancer.

I. Introduction

Apart from female foeticide, one of the strong reasons for imbalanced sex ratio is the mortality rate of women dying because of diseases, especially deadly diseases like Tuberculosis, AIDS and Cancer. Among various types of cancers, breast cancer is the one which is killing millions of women around the world. It is the second most popular killer disease, killing one among every four women [2]. In 2009-2010, about 562,340 Americans died of cancer, more than 1,500 people a day [2]. In 2011-2012, approximately 230,480 new cases of invasive breast cancer and 39,520 breast cancer deaths were expected to occur among US women [7]. In 2013-14, the approximation of women getting diagnosed with breast cancer was predicted to be about 64,640 [26]. But the number of women diagnosed was found to be more. However, US has made the best of its effort to reduce the mortality rate of women dying because of breast cancer and it can be clearly observed in the declination of the statistics from 2009-2014. But, in the developing country like India, the count is increasing with an alarming rate [5]. A better way of treating the disease is to first analyze its patterns. Medical data mining is a new advancement which leads the experts in efficient decision-making and proper analysis of disease. In any data mining or machine learning process preprocessing is the first step and is of high importance. The productivity of the entire model depends upon the relevancy of the features selected. Thus, we have come up with an approach which is capable of selecting predominant features from the high dimensional data sets. Data mining, image processing and machine learning make use of various supervised and unsupervised learning techniques for pre-processing, classification and prediction of data. Among all unsupervised techniques, clustering is the most famous and widely used technique used for segmentation of data. Here are some of the works related to different clustering techniques used till date. The classic paper of clustering was introduced by [8] and the other classic papers on clustering are reviewed in [10].

The most simple and the most famous clustering algorithm is K-means clustering [9]. But it has its own drawbacks. First and the major drawback is the time needed by K-means algorithm i.e. as the data size increases K-means take more time to perform. And the second drawback is the problem of outlier identification. Identification of outliers can be dealt by using a density based clustering technique called DBSCAN [9], and the computation time can be reduced by using K-means++ algorithm, also called as, Advanced K-means [12], [3]. However, many improved versions of K-means such as, hierarchical K-means, [28],fuzzy K-means [16], fast K-means [17], [25], [23] etc were developed in order to address various conceptual problems and real-world applications. Even though clustering is famous for segmentation of data, it has also been used for feature selection [15], feature extraction [29], [33], classification [18] and prediction. A new approach for extraction of Darwinian features for face recognition using Kohonen clustering and K-means clustering was made to recognize faces with better accuracy [1].The customer classification model was built by using PCA and K-means clustering by analyzing the historical data from the exchange. The model produced better results and help to overcome the problems in suitability management [19]. Recently clustering has been used for the prediction of wireless sensor data, to predict the energy consumption [11]. K-means Clustering along with Apriori algorithm was used to predict the education trends in academic databases. This approach helps in profiling and grouping of students, based upon their academic records [24]. K-means clustering was used to predict the heart attack. Here the clustering was done using K-means and Maximal Frequent Item set Algorithm (MAFIA) was used to find the frequent patterns that can predict the occurrence of heart attack in future [4].

Apart from this, meta-heuristic techniques were clubbed with clustering methods for solving many real-world problems. Diagnosis and detection of faults in motors were done by using Artificial ant based clustering [27]. Clustering of documents using K-means along with Particle Swarm Optimization (PSO) [6], Gravitational Search Algorithm (GSA) and Cuckoo Search [22] for clustering of web documents. Our main focus is on using Fast K-means algorithm embedded in PSO [3] for the selection of predominant features that can be later used for either classification of data or for the prediction purpose.

The paper is organized in the following way, after the introduction part the second section is preliminary view,followed by proposed model,results and discussions and finally the conclusion and future work.

II. Preliminaries

A. Fast K-means algorithm

Fast K-means algorithm is an improved version of K-means algorithm introduced by David Arthur and Sergei Vassilvitskii. The fast K-means algorithm differs from its parent form in the way they both choose the centroids. K-means algorithm tries a random approach where, at the beginning it randomly allocates a cluster center and then goes on searching for the proper center based upon its previous iteration values. Where as Fast K-means algorithm also termed as $K-means++$ algorithm uses a sophisticated procedure called $seeding\ procedure$ [3].

The main aim of any clustering is to minimize the intra class distance and to maximize the inter class distance. Thus the problem definition for both fast K-means and K-means can be defined as follows;

Consider a data set D of m rows and n columns,

$$D = x_1, x_2, ...x_k, ..., x_m \tag{1}$$

Here each row is a collection of n attributes also known as feature vectors and can be represented as

$$x_i = x_{i1}, x_{i2}, ...x_{in} \tag{2}$$

The objective is to minimize the distance between the centroid and the data points as much as possible.

The above problem can be addressed in two ways;

- by using $K-means$ algorithm
- by using $FastK-means(K-means++)$ algorithm

The pseudo code for K-means algorithm is given in Algorithm 1

The pseudo code for Fast K-means algorithm is given in Algorithm 2

Even though the objective of K-means and fast K-means algorithm is same. Researchers always prefer Fast K-means algorithm for handling high dimensional data sets because of the following reasons;

- Fast K-means follows $seeding\ procedure$ as the selection criteria for choosing the centroid where as general K-means algorithm follows $random\ procedure.$
- Fast K-means algorithm converges earlier than the general K-means algorithm.
- The performance of fast K-means is better than the general K-means.
- The most important is the complexity, the complexity of fast K-means is O(log(K)) [3]competitive and that of general K-means is O(K*m*n). Where K is a positive integer and indicates the number of clusters. m is the number of rows and n is the number of columns.

B. Particle Swarm Optimization Algorithm

Particle Swarm Optimization (PSO)is a nature inspired stochastic technique which mimics the social behavior of flock of birds. In 1995 Dr. Eberhart and Dr. Kennedy, came up with the concept of PSO which was very much similar to genetic algorithms but it does not require any operators for finding solution, instead the particles find the solution through optimization [13]. In PSO, particles are meant to move around the problem domain (search space) to find the optimal solution. In order to find the solution, the particles need to keep track of their local and global best solutions. The change in position of each particle depends upon its current velocity, current position, its distance from the particle's local best solution, and the distance from the global best solution [21]. The mathematical formulation of PSO are as shown in Equations (3) and (4).

$$\begin{aligned} v_i^{t+1} = wv_i^t + c_1 \times rand\left(particle\_best_i - x_i^t\right) \\ + c_2 \times rand\left(global\_best_i - x_i^t\right) \end{aligned} \tag{3}$$

$$x_i^{t+1} = x_i^t + v_i^{t+1} \tag{4}$$

Here, v_i^t indicates the velocity of i^{th} at the iteration t. w indicates the weight, c_1 and c_2 indicate the acceleration coefficients, and $rand$ is a function which generates a random number between 0 and 1. x_i^t is the current position of the particle i at iteration t, $particle\_best_i$ is the best solution found by the particle i at iteration t, and $global\_best_i$ is the global best solution. The working theme of basic PSO is outlined in Algorithm 3;

III. Proposed Work and Experiment

Our problem statement is to select top N strong and predominant features from high dimensional breast cancer data set that can aid in better decision making. The flowchart for the proposed method is given in Figure 1.

In order to carry out our work we have chosen a high dimensional numeric data set called KDD cup 2008 breast cancer data which contain 117 dependent attributes and one class attribute and 102924 rows [14]. The selected data sets are normalized in the range of $[-1, 1]$ and are subjected to hybrid $PSO\_FastK-means$ algorithm where we select the optimal feature subset that can provide us high accuracy. The pseudo code for hybrid $PSO\_FastK-means$ algorithm is as given in Algorithm 4

The normalized data set is subjected to $PSO\_fastK-means$ algorithm where it is evaluated to select the optimal features. There are two ways to evaluate the clustering

Algorithm 1: K-means Algorithm [?]

```
Data: $dataset\ of\ size\ n$
Result: $clusters$
1 initialization: k=2;
2 $Select\ k\ points\ at\ random\ as\ cluster\ centers.$
3 $Assign\ objects\ to\ their\ closest\ cluster\ center\ according\ to\ the\ Euclidean\ distance\ function.$
4 $Calculate\ the\ centroid\ or\ mean\ of\ all\ objects\ in\ each\ cluster.$
5 $Repeat\ steps\ 3\ and\ 4\ until\ the\ same\ points\ are\ assigned\ to\ each\ cluster\ in\ consecutive\ round$
```

Algorithm 2: Fast K-means Algorithm [3]

```
Data: $dataset\ of\ size\ n$
Result: $clusters$
1 initialization: k=2;
2 $For\ the\ first\ time\ choose\ one\ center\ c_1\ randomly\ from\ the\ data\ point\ X$
3 $Then\ take\ a\ new\ center\ c_1\ from\ X\ based\ upon\ the\ probability\ i.e.\ \frac{D(x)^2}{\sum x \in XD(x)^2}$
4 $Repeat\ 2\ and\ 3\ until\ all\ K\ centers\ have\ been\ taken.$
5 $Now\ follow\ steps\ 2\ to\ 5\ from\ K-means$
```

Algorithm 3: Pseudocode for PSO

```
Data: $Problem\_size, Population\_size$
Result: $Particle_{global\_best}$
1 initialization: Initialize the population randomly;
2 $Calculate\ the\ velocity\ of\ each\ particle$
3 $Update\ the\ position\ of\ each\ particle\ by\ considering\ its\ previous\ velocity\ and\ position$
4 $Repeat\ step\ 2\ and\ step\ 3\ till\ the\ convergence\ occurs$
```

Algorithm 4: Hybrid PSO fast K-means Algorithm [3]

```
Data: $Problem\_size, Population\_size$
Result: $Particle_{global\_best}$
initialization;
while number of iterations do
    Evaluate fitness of particle swarm by fastK-means()
    for $i = 1 \to Population\_size$ do
        $Particle_{velocity} \leftarrow Random\_Velocity()$
        $Particle_{position} \leftarrow Random\_Position()$
        $Particle_{cost} \leftarrow Cost(Particle_{position})$
        $Paricle_{best} \leftarrow Particle_{position}$
        if $(Particle_{cost} \leq Particle_{global\_best})$ then
            $Particle_{global\_best} \leftarrow Particle_{best}$
        end
    end
    while $Stop\_Condition$ do
        foreach $Particle \in Population$ do
            $Particle_{velocity} \leftarrow Update\_Velocity(Particle_{velocity}, Particle_{global\_best}, Particle_{best})$
            $Particle_{position} \leftarrow Random\_Position(Particle_{position}, Particle_{velocity})$
            $Particle_{cost} \leftarrow Cost(Particle_{position})$
            if $(Particle_{cost} \leq Particle_{global\_best})$ then
                $Particle_{global\_best} \leftarrow Particle_{best}$
            end
        end
    end
    $return\ Particle_{global\_best}$
end
```

methods,internal evaluation and external evaluation. In an internal evaluation method, the clustering algorithm is evaluated in terms of accuracy. And in an external evaluation, the clustering technique is evaluated using external objective

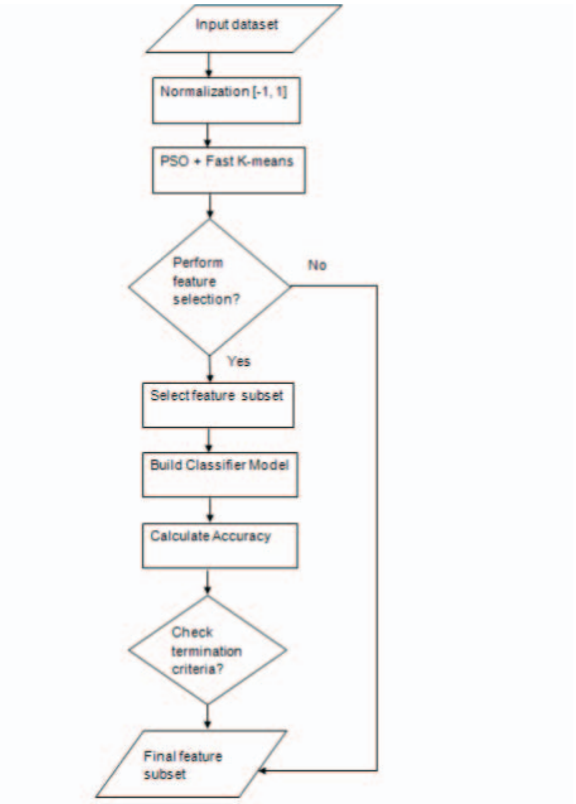

Fig. 1. Flowchart of the proposed model

functions such as distance function. The main target of any clustering algorithm is to minimize the intra-cluster difference and to maximize the inter-cluster difference. Here we follow the internal or intrinsic evaluation measure where the entire data set is subjected to PSO based fast K-means algorithm ($PSO\_fastK{-}means$) and accuracy of the clusters formed is calculated. In $PSO\_fastK{-}means$ algorithm, the population is randomly initialized and the particles are updated based upon two fitness values - local best ($Particle_{best}\ or\ Pbest$) and global best ($particle_{global\_best}\ or\ Gbest$) to find the optimal solution. It is to be noted that almost all particles gets converged near gbest fitness value after some iterations. The fitness of a particle is evaluated using fast K-means algorithm. The data points are clustered into two groups (benign and malignant). The indices obtained are subjected to accuracy check. The accuracy of each individual feature is calculated using Equation (5).

$$Accuracy = \frac{TP + TN}{FP + FN + TP + TN} \tag{5}$$

where TP: True positive is a malignant data correctly clustered as malignant.

TN: True negative is a benign data correctly clustered as benign.

FP: False positive is a malignant outcome clustered as benign and

FN: False negative is a benign outcome clustered as malignant.

The attribute with the highest accuracy get ranked first followed by next immediate one and so on. For our convenience we have chosen top 10 items for feature selection.

A. Experiment

The proposed work was built on Matlab 2014a platform installed on a desktop with 4GB RAM, Intel i3 processor and 500 GB hard disk. The high dimensional data set was subjected to fast K-means clustering algorithm for feature selection. The attributes selected using proposed method (M5) are as shown in Table I.

Other than the fast K-means algorithm we have tried to select the features from various widely used feature selection techniques and are tabulated in Table I.

TABLE I. FEATURES SELECTED FROM KDD2008 CUP BREAST CANCER DATA SET USING VARIOUS ALGORITHMS

Method	**Features selected from KDD data**
Relief (M1)	f27,f13,f29,f5,f96,f87,f88,f86,f6 and F97
Random Forest(M2)	f27,f112,f6,f77,f5 ,f29,f99,f79,f20 and f74
Information Gain(M3)	f29,f27,f73,f99,f68,f81,f33,f115,f15 and f80
Gain Ratio (M4)	f29,f27,f73,f99,f68,f81,f33,f115,f15 and f80
Proposed method(M5)	f75,f76,f77,f2,f55,f56,f57,f58,f3,f112

1) Comparison: In order to find out the potential of the selected attributes we subjected the features to both clustering and classification. For testing on clustering method we choose FastK-means algorithm to find out the accuracy of correctly clustered instances . The selected attribute subsets obtained from various method which are fed as input to the fast K-means algorithm are also evaluated by using various evaluation measures such as - precision, recall, F-measure, G-measure etc. Table II provides us the information regarding evaluating the performance of various feature subsets on KDD data. From Table II it is clear that the features selected from the proposed method are highly accurate than the features obtained from other feature selection methods. Even though the accuracy of both random forest (M2) and proposed method (M5)are same, but when we consider the time complexity the proposed method is far better than the other techniques. The time complexities of various feature selection methods for n training instances with f features and m iterations are tabulated in Table III.

TABLE II. EVALUATION OF FEATURES SELECTED FROM KDD2008 CUP BREAST CANCER DATA SET USING VARIOUS METHODS

Evaluation Measures	M1	M2	M3	M4	M5
Accuracy	66.74	99.39	99.37	99.37	99.39
Precision	69.19	98.55	98.35	98.39	98.55
Recall	83.84	99.2	99.0	99.1	99.2
AUC	69.49	99.087	99.085	99.085	99.087
G-Score	76.16	98.87	98.25	98.74	98.87
F-Score	75.81	98.87	98.25	98.74	98.87

TABLE III. TIME COMPLEXITIES OF VARIOUS FEATURE SELECTION METHODS

Feature selection method	**Time complexity**
Relief (M1) [31]	$O(m * n * f)$
Random Forest (M2) [30]	$O(t * Sqrt((logt) + 1) * n * logn)$
Information Gain(M3) [20]	$O(f * f)$
Gain Ratio (M4) [32]	$O(n * log(n))$
Proposed Method (M5)	$O(log(k))$

From Table III it is clear that the proposed method is computationally more efficient than other feature selection methods.

The attributes selected using PSO based fast K-means were also evaluated by subjecting them to various classifiers under random sampling scheme and their accuracies are tabulated in Table IV.

TABLE IV. ACCURACIES OF VARIOUS CLASSIFIERS ON FAST K-MEANS BASED FEATURE SUBSET

Method used	Accuracy of selected subset	Accuracy of Entire data set
ANN	67.13	65.01
MNN	68.55	66.23
KNN	63.3	60.30
SVM	62	61.55
Decission Tree	61.57	60.00
Naive Bayes	62.3	60.66
Bat	62.8	61.32
PSO	62.7	60.17

IV. RESULTS AND DISCUSSION

By analyzing the results obtained from Table II the accuracy of the feature subset selected using the proposed $PSObasedfastK - meansalgorithm$ is found to be 99.39 %, which is higher than accuracies of other models. Even though feature subset selected from random forest yields the same accuracy the precision, recall and F-measure of the feature subset generated from our proposed work is higher than remaining all subsets. Apart from accuracy the quick accomplishment of tasks, i.e. time complexity makes our model more productive and preferable. From Table IV it is clear that the feature subset (a subset of top 10 features) obtained by proposed method performs extremely well than the original data set (set of 117 features). Thus making our model quantitatively and qualitatively feasible.

V. CONCLUSION AND FUTURE WORK

From the available results and comparative analysis, we can strongly conclude that the proposed model - PSO based Fast K-means algorithm for feature selection and classification performs better than the other models used for comparison. It performs very well in selecting the attributes as well as in classification and is much faster than the basic K-means algorithm.

In future we are intended to carry two tasks;

1. Improving the classification accuracy and

2. Combining fast K-means algorithm with other strong meta-heuristic algorithms to carry out feature selection.

ACKNOWLEDGEMENT

The research scholar is financially supported by University Grant Commission(UGC),India under Maulana Azad National Fellowship Scheme 2013 dated on 31-01-2013.

We would like to thank KDD cup 2008 for Breast cancer data collection and sharing which was organized by Knowledge Discovery in Datamining (KDD) conference held in in Las Vegas, The United States of America during the year 2008. http://www.sigkdd.org/kdd-cup-2008-breast-cancer.

REFERENCES

[1] J. Adams, J. Shelton, L. Small, S. Neal, M. Venable, J. H. Kim, and G. Dozier, "Darwinian-based feature extraction using k-means and kohonen clustering," in *Midwest Artificial Intelligence and Cognitive Science Conference*. Citeseer, 2012, p. 83.

[2] American Society of Breast Cancer Research, "Cancer facts and figures," http://www.cancer.org/research/cancerfactsstatistics/cancerfactsfigures2009/index, accessed on 10.04.2015.

[3] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.

[4] N. Banu and B. Gomathy, "Disease forecasting system using data mining methods," in *Intelligent Computing Applications (ICICA), 2014 International Conference on*. IEEE, 2014, pp. 130–133.

[5] Breast Cancer India, "Trends of breast cancer in india," http://www.breastcancerindia.net/statistics/trends.html, accessed on 10.04.2015.

[6] X. Cui, T. E. Potok, and P. Palathingal, "Document clustering using particle swarm optimization," in *Swarm Intelligence Symposium, 2005. SIS 2005. Proceedings 2005 IEEE*. IEEE, 2005, pp. 185–191.

[7] C. DeSantis, R. Siegel, P. Bandi, and A. Jemal, "Breast cancer statistics, 2011," *CA: a cancer journal for clinicians*, vol. 61, no. 6, pp. 408–418, 2011.

[8] L. Engelman and J. A. Hartigan, "Percentage points of a test for clusters," *Journal of the American Statistical Association*, vol. 64, no. 328, pp. 1647–1648, 1969.

[9] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques*. Morgan kaufmann, 2006.

[10] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[11] H. Jiang, S. Jin, and C. Wang, "Prediction or not? an energy-efficient framework for clustering-based data collection in wireless sensor networks," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 22, no. 6, pp. 1064–1071, 2011.

[12] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 7, pp. 881–892, 2002.

[13] J. Kenndy and R. Eberhart, "Particle swarm optimization," in *Proceedings of IEEE International Conference on Neural Networks*, vol. 4, 1995, pp. 1942–1948.

[14] Knowledge Discovery in Data mining, "Kdd cup 2008," http://www.sigkdd.org/kdd-cup-2008-breast-cancer, accessed on 12.02.2014.

[15] M. H. Law, M. A. Figueiredo, and A. K. Jain, "Simultaneous feature selection and clustering using mixture models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 9, pp. 1154–1166, 2004.

[16] M. J. Li, M. K. Ng, Y.-m. Cheung, and J. Z. Huang, "Agglomerative fuzzy k-means clustering algorithm with selection of number of clusters," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 20, no. 11, pp. 1519–1534, 2008.

[17] C.-H. Lin, C.-C. Chen, H.-L. Lee, and J.-R. Liao, "Fast k-means algorithm based on a level histogram for image retrieval," *Expert Systems with Applications*, vol. 41, no. 7, pp. 3276–3283, 2014.

[18] Y. Lin, F. Lv, S. Zhu, M. Yang, T. Cour, K. Yu, L. Cao, and T. Huang, "Large-scale image classification: fast feature extraction and svm training," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1689–1696.

[19] B. Liu, H. Qiu, and Y. Shen, "Establishment and implementation of securities company customer classification model based on clustering analysis and pca," in *Control Engineering and Communication Technology (ICCECT), 2012 International Conference on*. IEEE, 2012, pp. 325–329.

[20] H. Liu and H. Motoda, *Feature selection for knowledge discovery and data mining*. Springer Science & Business Media, 1998.

[21] S. Mirjalili, S. Z. M. Hashim, and H. M. Sardroudi, "Training feedforward neural networks using hybrid particle swarm optimization and gravitational search algorithm," *Applied Mathematics and Computation*, vol. 218, no. 22, pp. 11 125–11 137, 2012.

[22] A. J. Mohammed, Y. Yusof, and H. Husni, “A newtons universal gravitation inspired firefly algorithm for document clustering,” in *Advances in Computer Science and its Applications*. Springer, 2014, pp. 1259–1264.

[23] M. K. Pakhira, “A fast k-means algorithm using cluster shifting to produce compact and separate clusters (research note),” *International Journal of Engineering-Transactions A: Basics*, vol. 28, no. 1, p. 35, 2014.

[24] S. Parack, Z. Zahid, and F. Merchant, “Application of data mining in educational databases for predicting academic trends and patterns,” in *Technology Enhanced Education (ICTEE), 2012 IEEE International Conference on*. IEEE, 2012, pp. 1–4.

[25] M. Shindler, A. Wong, and A. W. Meyerson, “Fast and accurate k-means for large datasets,” in *Advances in neural information processing systems*, 2011, pp. 2375–2383.

[26] R. Siegel, D. Naishadham, and A. Jemal, “Cancer statistics, 2013,” *CA: a cancer journal for clinicians*, vol. 63, no. 1, pp. 11–30, 2013.

[27] A. Soualhi, G. Clerc, and H. Razik, “Detection and diagnosis of faults in induction motor using an improved artificial ant clustering technique,” *Industrial Electronics, IEEE Transactions on*, vol. 60, no. 9, pp. 4053–4062, 2013.

[28] S. N. Sulaiman and N. A. M. Isa, “Adaptive fuzzy-k-means clustering algorithm for image segmentation,” *Consumer Electronics, IEEE Transactions on*, vol. 56, no. 4, pp. 2661–2668, 2010.

[29] N. Van Huan, N. T. H. Binh, and H. Kim, “Eye feature extraction using k-means clustering for low illumination and iris color variety,” in *Control Automation Robotics & Vision (ICARCV), 2010 11th International Conference on*. IEEE, 2010, pp. 633–637.

[30] C. Vens and F. Costa, “Random forest based feature induction,” in *Data Mining (ICDM), 2011 IEEE 11th International Conference on*. IEEE, 2011, pp. 744–753.

[31] J. Wang, G. G. Yen, and M. M. Polycarpou, *Advances in Neural Networks-ISNN 2012: 9th International Symposium on Neural Networks, ISNN 2012, Shenyang, China, July 11-14, 2012: Proceedings*. Springer, 2012.

[32] S. Wang, G. Yu, and H. Lu, *Advances in Web-Age Information Management: Second International Conference, WAIM 2001, Xi’an, China, July 9-11, 2001. Proceedings*. Springer Science & Business Media, 2001, vol. 2.

[33] B. Zheng, S. W. Yoon, and S. S. Lam, “Breast cancer diagnosis based on feature extraction using a hybrid of k-means and support vector machine algorithms,” *Expert Systems with Applications*, vol. 41, no. 4, pp. 1476–1482, 2014.