Intelligent Systems and Computer Technology
D.J. Hemanth et al. (Eds.)

doi:10.3233/APC200138

Effective Feature Subset Identification Using Adaptive Bee Colony Algorithm

Suja K [a,1] and Rengarajan A[b]

[a] *Research scholar, Dept of CSE, St Peter's University, Chennai, India*
[b] *Professor, Dept t of CSE, Vel Tech Multi Tech Engineering College, Chennai, India*

Abstract. The irrelevant features, along with redundant features, severely affect the accuracy of the learning machines. Feature subset selection as the process of identifying and removing many irrelevant and redundant features. The overall process of the optimal feature selection method is divided into two main steps, such as, i) preprocessing (ii) Optimal feature selection using clustering and tree generation. At first, preprocessing is done in the input micro array dataset. Then the Possibilistic fuzzy c- means clustering algorithm with optimal minimum spanning tree algorithm is applied on the high dimensional micro array dataset to select the important features. Here the proposed method is optimally select the features with the help of Adaptive artificial bee colony algorithm.

Keywords. micro array, possibilistic fuzzy c- means, minimum spanning tree, artificial bee colony

1. Introduction

Feature selection is an important topic in data mining, especially for high dimensional datasets [1]. With the aim of choosing a subset of good features with respect to the target concepts, feature subset selection is an effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility [2]. Feature selection techniques are often used in domains where there are many features and comparatively few samples [4]. Feature extraction creates new features from functions of the original features, whereas feature selection returns a subset of the features [5][19].

The feature subset selection can be done with the help of various algorithms such as best search, greedy forward selection algorithm, greedy backward elimination algorithm, genetic algorithm [7]. Many feature subset selection methods have been proposed and studied for machine learning applications. They are to be divided into four broad categories these are Embedded, Wrapper, Filter and Hybrid approaches [8]. In particular, we accept the minimum spanning tree based clustering algorithms, for the reason that they do not imagine that data points are clustered around centers or separated by means of a normal geometric curve and have been extensively used in tradition [9]. The embedded methods incorporate feature selection as a part of the training process and are usually specific to given learning algorithms, and therefore may be more efficient than the other three categories. Traditional machine learning stopping criterion is tested in each of iterations to determine whether or not the FS process should continue [3].

[1]K.Suja , Research scolar ,Department of CSE ,St Peter's University, Chennai,
Email-id: ksuja_venkat@yahoo.co.in

algorithms like decision trees or artificial neural networks are examples of embedded approaches [10]. In the wrapper approach the evaluation function calculates the suitability of a feature subset produced by the generation procedure and it also compares that with the previous best candidate, replacing it if found to be better. A They mainly focus on combining filter and wrapper methods to achieve the best possible performance with a particular learning algorithm with similar time complexity of the filter methods [11]. In cluster analysis, graph-theoretic methods are used in many applications. Sometime their outcomes have best agreement with human performance. The graph theoretic clustering is used to evaluate a neighborhood graph of instances [12]. We have clustered the features by graph-theoretic methods to select most representative feature related to target class. For this, we adopt MST in Fast clustering-based feature Selection algorithm (FAST). FAST algorithm performs in two steps. First of all, features are divided into various clusters. Then the most useful feature is selected from each cluster [13]. A feature in different clusters in relatively independent, the clustering based strategy of FAST has high probability of producing a subset of useful and independent features [14]. Features in different clusters are comparatively independent, the clustering based strategy of FAST features a high probability of producing a set of helpful and independent features [15].

2. Related works

Leonard K.M. Poon et al [16] have proposed a generalization of the Gaussian mixture models and demonstrate its ability to automatically identify natural facets of data and cluster data along each of those facets simultaneously. They have presented empirical results to show that facet determination usually leads to better clustering results than variable selection. Xinyue Liu and Menggang Li [17] have addressed that dimension selection; dimension weighting and data assignment were three essential tasks for high dimensional data clustering. They also pointed out that constraints were necessary to break the circular-dependency of the three essential tasks. Charles Bouveyron and Camille Brunet- Saumard [18] have therefore presented a review of existing solutions for model-based clustering of high- dimensional data. Model-based clustering was a popular tool which was renowned for its probabilistic foundations and its flexibility.

However, high-dimensional data were nowadays more and more frequent and, unfortunately, classical model-based clustering techniques show a disappointing behavior in high- dimensional spaces. This was mainly due to the fact that model-based clustering methods were dramatically over parameterized in this case. Haider Banka and Suresh Dara [19] have presented a Hamming distance based binary PSO algorithm for feature selection and classification in gene expression data. The experimental results validate that the proposed HDBPSO performs better using Hamming distance as proximity measure for this problem.

3. Proposed method

Feature subset selection is the process of identifying and removing many irrelevant features or data. The classification accuracy is affected by irrelevant features or data so that the proposed method is use optimal feature subset selection method. In our proposed method, initially the input dataset is preprocessed. In preprocessing we select numerical data and remove the non numerical data. Here the proposed method

uses microarray dataset as the input. Then we have to cluster the relevant features or data. For clustering, we use Possibilistic Fuzzy C-Means Clustering Algorithm (PFCM). To improve the efficiency of the proposed method we construct the minimum spanning tree after the clustering algorithm. Before the tree construction, each cluster is given as the input for optimization technique. The proposed method uses the adaptive artificial bee colony algorithm (AABC) for optimal feature or data selection. After getting the optimal result we generate the minimum spanning tree, here prim's algorithm is employed for tree generation. The detail procedure of the proposed work is illustrated in the further section and the overall block diagram of the proposed method is shown in below. The overall process of the proposed framework is divided into two steps, such as 1) Pre-processing, 2) Clustering and tree generation

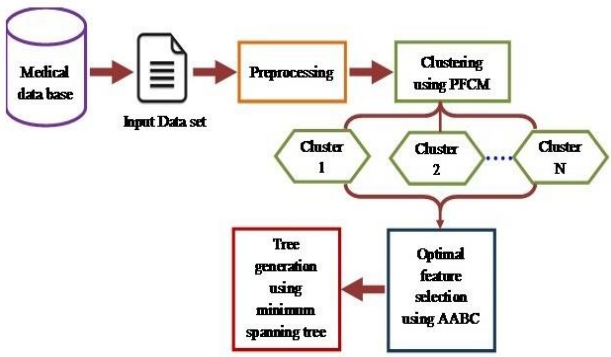


Figure 1. System Architecture

3.1 Preprocessings

Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running an analysis. If there is much irrelevant and redundant information or noisy and unreliable data present then knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time. In preprocessing stage the input dataset is given as the input. Here the input data has raw data. Raw data is highly susceptible to noise, missing values and inconsistency. The quality of data affects the results. In order to improve the quality of the data and consequently, of the results raw data is pre-processed so as to improve the efficiency and ease of mining process. Data pre-processing is one of the most critical steps in a data mining process which deals with the preparation and transformation of the initial dataset. In the paper, pre-processing is applied to the dataset for getting the numerical data from the non-numerical data. In the stage, the non-numerical data are removed and obtained the numerical dataset for proceeding further

3.2 Optimal Feature Selection

In order to select the optimal feature the proposed method use clustering and tree generation process. After preprocessing the input microarray dataset, we have to cluster the input data based on the clustering algorithm. Here the proposed method use Possibilistic fuzzy c means algorithm for clustering the data. The detailed explanation of PFCM algorithm described in below section

3.2.1 Possibilistic Fuzzy C-means clustering algorithm

The application of Possibilistic Fuzzy C-means clustering methods will improve the clustering process and accuracy of the data classification. The vital motive of the probabilistic fuzzy c-means (PFCM) cluster module is devoted to the specified set of data into clusters. By means of the clustering module, the training set is grouped into various subsets. The Probabilistic Fuzzy c-means constitutes a data clustering technique where each data point is a part of the cluster to a level indicated by the membership grade. In the clustering module, it is dependent on the reduction of the objective function which is illustrated in the following equation (1).

$$OB_{PFCM}(M,T,C;I)=\sum_{k=1}^{n}\sum_{i=1}^{c}(aM_{ik}^{m}+bT_{ik}^{\gamma})\times\left\|I_k-C_i\right\|_A^2+\sum_{i=1}^{c}\gamma_i\sum_{k=1}^{n}(1-T_{ik})^{\gamma} \quad (1)$$

Where,

M → Membership matrix

T → Typicality matrix

C →Resultant cluster centre

I →Set of all input data point

The constants a and b represent the comparative significance of fuzzy membership and typicality values in the objective function. The gradual procedure of PFCM is effectively elucidated below

Step 1: Calculation of distance matrix

At first, the number of cluster is furnished by the user, which is identical in respect of every segment. When the number of cluster is determined, the evaluation of the distance between the centroids and data point for each segment is carried out. In this paper, Euclidian distance function as illustrated in equation (2) shown below is elegantly employed to evaluate the distance between centroids and data points with the ultimate motive of evaluating the distance matrix. Further, the distance matrix is determined in respect of each and every cluster

$$D(I_k,C_i)=\sqrt{\sum_{i,k=1}^{i=n,k=n}(I_k-C_i)^2} \quad (2)$$

Where, I_k represent the data point

C_i represent the centroid value

Step 2: Calculation of typicality matrix

After the estimation of the distance matrix, the typicality matrix is evaluated. The typicality matrix, in turn, is obtained from PCM [3]. The probability value of each data point in relation to every centroid is completed. Subsequently, typicality matrix is created

$$T_{ik} = \frac{1}{1 + \left[\frac{D^2 (I_k, C_i)}{\gamma_i} \right]^{1/(m-1)}} \tag{3}$$

Where, T_{ik} represent the typicality matrix.
$\gamma > 0$ is a user defined constant

Step 3: Calculation of membership matrix

The evaluation of the membership matrix M_{ik} is performed by means of assessing the membership value of data point which is gathered from the FCM. As illustrated in the help of the following equation (4), the membership value of each data point in relation to each centroid is completed

$$M_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{\| I_k - C_i \|}{\| I_k - C_j \|} \right)^{\frac{2}{m-1}}} \tag{4}$$

Step 4: Updation of centroid

After the generation of the clusters, the modernization of the centroids is performed in accordance with equation (5) shown below

$$C_i = \frac{\sum_{k=i}^{n} \left(M_{ik}^m + T_{ik}^\gamma \right) I_k}{\sum_{k=1}^{n} \left(M_{ik}^m + T_{ik}^\gamma \right)}, \qquad 1 \le i \le c \tag{5}$$

Subsequent to the modernization of the centroids in respect of each and every cluster, the task of evaluating the distance with the lately modernized centroids is started and continued till the evaluation of the modernization of the centroids. The relative procedure is performed again and again till the modernized centroids of each and every cluster becomes identical similar in successive iterations. To improve the efficiency of the proposed method we construct the minimum spanning tree after the clustering algorithm

3.2.2 Minimum spanning tree generation

A minimum spanning tree is a spanning tree of a connected, undirected graph. It connects all the vertices together with the minimal total weighting for its edges. In our proposed method minimum spanning tree is implemented using prims algorithm. For generating the minimum spanning tree we have to select the optimal features, so that the proposed method use adaptive artificial bee colony algorithm. The step by step procedure of adaptive artificial bee colony algorithm is described in further section

3.2.2.1 Adaptive Artificial Bee Colony Algorithm (AABC)

ABC algorithm is a swarm based meta-heuristic algorithm which was enthused by the sharp foraging behavior of the honey bees. It consists of three components namely, employed bees, onlooker bees and scout bees

Employed bees: The employed bees are coupled with the food sources in the region of the hive and they transfer the data to the onlookers about the nectar quality of the food sources they are exploiting

Onlooker bees: Onlooker bees are looking the dance of the employed bees inside the hive to pick one food source to exploit according to the data provided by the employed bees

Scout bees: The employed bees whose food source is abandoned become Scout and seeking new food source arbitrarily

The number of food sources denotes the location of probable solutions of optimization problem and the nectar amount of a food source denotes the quality of the solution.

A. Initial Phase

First the population of the food sources G_i (i = 1, 2, ….N) are generated arbitrarily. N denotes the size of the population. This generation process is called as initialization process. To evaluate the best food source, the fitness value of the generated food sources is calculated using equation. (6)

$$Fitness\ function = Max\ Accuracy \tag{6}$$

After the calculation of fitness value, the iteration is set to 1. After that, the phase of employed bee is carried out

B. Employed Bee Phase

In the employed bee phase, new population parameters are generated using the below equation,

$$G_{i,j}^{New} = G_{i,j} + \alpha_{ij}\left(G_{i,j} - G_{k,j}\right) \tag{7}$$

Where, k and j is a random selected index, α is randomly produced number in the range [-1, 1] and $G_{i,j}^{New}$ is the new value of the j^{th} position. Then the fitness value is computed for every new generated population parameters of food sources. From the computed fitness value of the population, best population parameter is selected, which has the highest fitness value by applying greedy selection process. After selecting the best population parameter, probability of the selected parameter is computed using the equation (8)

$$\Pr obability = \frac{Fitness}{\sum_{j=1}^{d} Fitness_j} \tag{8}$$

C. Onlooker Bee Phase

After computing the probability of the selected parameter, number of onlooker bees is estimated. Following, generate new solutions ($G_{i,j}^{New}$) for the onlooker bees from the solutions ($G_{i,j}$) based on the probability value. Then the fitness function is calculated for the new solution. Subsequently apply the greedy selection process in order to select the best parameter

D. Scout Bee Phase

Determine the Abandoned parameters for the scout bees. If any abandoned parameter is present, then replace that with the new parameters discovered by scouts and evaluate the fitness value. Then memorize the best parameters achieved so far. Then the iteration is incremented and the process is continued until the stopping criterion is reached. This is the detailed description about the ABC algorithm and the working procedure. Here we are using the AABC algorithm for selecting the optimal features. The adaptiveness is attained by including the modified rate in the food source generation process. The working procedure is same as given in above process excepting the initial phase and the fitness function.

Initial Phase

The initial food sources are generated randomly within the range of the boundaries of the parameters using the equation given in below

$$G_{i,j}^{New} = G_j^{\min} + \alpha(0,1) \times \left(G_i^{\max} - G_j^{\min} \right) \quad (9)$$

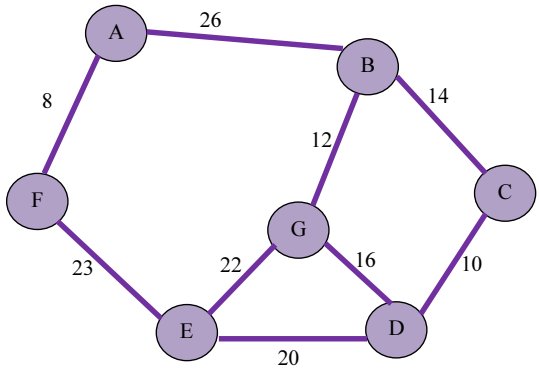
Where, $G_j^{\min}$ & $G_i^{\max}$ are the predetermined boundaries, i= 1,2 , .. N, and j = 1, 2, ..M. M- Number of optimization parameter. If the value of the parameter exceeds its predetermined boundary, then reset to its boundaries. After the initial phase, the same procedure is followed. The fitness function preferred here is eqn. (6). Based on the above procedures, we select the optimal features after that we generate the tree construction for our proposed work

Tree generation using Prims algorithm

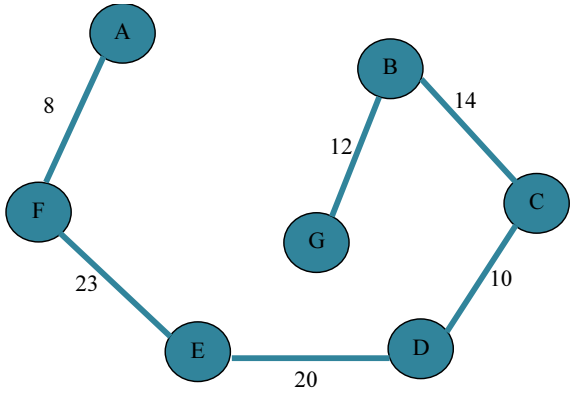
A Prim's algorithm is a greedy method which helps us to obtain minimum spanning tree. The Prim's algorithm uses the concept of sets. Instead of processing the graph by sorting order of edges, this algorithm processes the edges in the graph randomly by building up disjoint sets. The step by step procedure of prims algorithm is given below

Step by step procedure of Prims algorithm

Input: A non empty connected weighted graph composed of n vertexes and edges, possibly with null weights
Output: The minimal spanning tree in the final path
Step 1: Start from any arbitrary vertex.
Step 2: Note down all the edges emerging from this vertex
Step 3: Mark this edge as visited
Step 4: Select an edge with the minimum weight
Step 5: Traverse to the other end. Remove this edge from the list and insert it into the minimum spanning tree
Step 6: Repeat this process for the newly visited vertex
Step 7: Each time you visit a vertex, check if it was already visited, only then we do the process of adding its edges to the list and picking the minimum
Step 8: If not, then simply pick up the next minimum edge
Step 9: Repeat this process until all the nodes are visited
Example: To find a minimum spanning tree with the help of prims algorithm in the following weighted graph



The Prim's algorithm operates on two disjoint sets of edges in the graph. Prim's has a better running time if both the number of edges and the number of nodes are low. Prim's algorithm will proceed as follows. First we add edge {A, F} of weight 8. Next, we add edge {F, E} of weight 23. Next, we add edge {E, D} of weight 20. Next, we add edge {D, C} of weight 10. And then we add edge {C, B} of weight 14. Then finally we add the edge of {B, G} of weight 12. This produces a minimum spanning tree of weight 87. A minimum spanning tree is shown in below,



Finally the optimal minimum spanning tree is given to the classification purpose.

4. Performance Analysis

In this section we discuss the result obtained from the propose technique. For implementing the propose technique, we have used MATLAB. The proposed technique is done in windows machine having Intel Core i5 processor with speed 1.6 GHz and 4 GB RAM.

4.1 Dataset description

The proposed system is experimented with the widely applied datasets namely, orlraws10P, pixraw10P, warpAR10P, CLL_SUB_111, TOX_171, SMK_CAN_187 and GLA-BRA-180. The detailed explanation of dataset description is followed here

Table 1: The dataset description

Data Set	Number of Instances	Features	Classes
orlraws10P	100	10304	10
pixraw10P	100	10000	10
warpAR10P	130	2400	10
warpPIE10P	210	2420	10
CLL_SUB_111	111	11340	3
TOX_171	171	5748	4
SMK_CAN_187	187	19993	2
GLA-BRA-180	180	4915	4

4.2 Evaluation metrics

The evaluation of proposed medical data classification technique is carried out using the following metrics as suggested by below equations

Sensitivity: The sensitivity of the feature selection and the feature classification is determined by taking the ratio of number of true positives to the sum of true positive and false negative. This relation can be expressed as

$$S_t = \frac{TP}{TP + FN} \tag{10}$$

Specificity: The specificity of the feature selection and the feature classification can be evaluated by taking the relation of number of true negatives to the combined true negative and the false positive. The specificity can be expressed as

$$S_p = \frac{TN}{TN + FP} \tag{11}$$

Accuracy: The accuracy of feature selection and the feature classification can be calculated by taking the ratio of true values present in the population. The accuracy can be described by the following equation as

$$A = \frac{TP + TN}{TP + FP + TN + FN} \tag{12}$$

Where, TP represent true positive, TN represent true negative, FP represent false positive and FN represent false negative

4.3 Performance Analysis

The results of proposed work help to analyze the efficiency of the feature selection and classification process. The subsequent table.2 tabulates the results of eight micro array datasets. Table.2 is tabulated in the below section,

Table 2: Performance of the proposed method using various dataset

Data Set	Accuracy	Sensitivity	Specificity
CLL_SUB_111	98.2609	0.5	0.790476
GLA-BRA-180	94.8571	0.086957	0.978723
SMK_CAN_187	86.4473	0.86359	0.669717
TOX_171	88.3613	0.442581	0.948
orlraws10P	97.0213	0.12	0.983784
pixraw10P	97.8261	0.2	0.972222
warpAR10P	99.6078	0.4	1
warpPIE10P	93.4775	0.5	0.97931

From table.2, the evaluation metrics are analyzed for the eight different datasets, by which we can observe the efficiency of proposed feature selection and data classification system. The accuracy values of eight datasets are 98.2609%, 94.8571%, 98.8%, 86.4473%, 88.3613%, 97.0213%, 97.8261%, 99.6078% and 93.4775%. The sensitivity values for the eight datasets are 0.5%, 0.086%, 0.86% 0.442%, 0.12%, 0.2%, 0.4% and 0.5%. The specificity values for the eight datasets are 0.790%, 0.978%, 0.669%, 0.948%, 0.983%, 0.972%, 1% and 0.979%

5. Conclusion

In the section an effective techniques are used for medical data classification. Here optimal minimum spanning tree algorithm is applied on the high dimensional micro array dataset to select the optimal features. For selecting the optimal features the proposed method use Adaptive artificial bee colony algorithm. Before the tree construction, each cluster is given as the input for optimization technique. The proposed method uses the adaptive artificial bee colony algorithm (AABC) for optimal feature or data selection. After getting the optimal result we generate the minimum spanning tree, here prim's algorithm is employed for tree generation. A Test bed is implemented to find the optimal feature selection

Reference

[1] N.Magendiran and J.Jayaranjani .An Efficient Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data .International journal of innovative research in science, Vol.3, Issue.1, pp. 405-408, Feb.2014.

[2] Yogesh R. Shepal and Ashraf Shaikh .A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data .International Journal of Research Studies in Science, Vol. 1, Issue.7, pp. 1-6, Oct 2014.

[3] Shahla Nemati, Mohammad Ehsan Basiri, Nasser Ghasem-Aghaee and Mehdi Hosseinzadeh Aghdam .A novel ACO–GA hybrid algorithm for feature selection in protein function prediction IEEE international journal of expert systems with applications, Vol. 36, pp. 12086-12094, 2009.

[4] Pallavi U. Mudaliar, Tejaswini A. Patil, Smita S. Thete and Kavita P. Moholkar .A Fast Clustering Based Feature Subset Selection Algorithm For High Dimensional Data .International journal of emerging trend in engineering and basic science, Vol.2, Issue.1, pp. 494-499, Feb 2015.

[5] Sujatha Kamepalli and Radha Mothukuri .Implementation of Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data .International journal of emerging trends and technology in computre science, Vol.3, Issue.3, pp. 30-34, June 2014.

[6] J.K.Madhavi and G.Venkatesh Yadav .An Improved Fast Clustering method for Feature Subset Selection on High-Dimensional Data clustering . International journal of application or innovation in engineering and management, Vol.3, Issue.10, pp. 26-30, Oct 2014.

[7] Swapnil A. Sutar and Prof. Devendra P. Gadekar .Survey on Fast Clustering Based Feature Selection Algorithm for High Dimensional Data .International journal of science and research, Vol. 3, Issue.12, pp. 691-694, Dec 2014.
[8] Shilpa N.Dande and Prof. Parinta Chate .A Fast Clustering Based Feature Subset Selection Algorithm for High Dimensional Data . International journal of innovative research and studies, Vol.3, Issu.3, pp. 26- 37, March 2014.
[9] Avinash Godase and Poonam Gupta .Improvised Method Of FAST Clustering Based Feature Selection Technique Algorithm For High Dimensional Data . International journal of application or innovation in engineering and management, Vol.4, Issue.6, pp. 135-140, June 2015.
[10] M.Kalyan and Mohammed Ali Shaik .A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data . IOJETR Transactions on Data Ware House, Vol.18, pp. 1071-1080, Oct 2014.
[11] Karthikeyan.P, Saravanan.P and Vanitha.E .High Dimensional Data Clustering Using Fast Cluster Based Feature Selection .International journal of engineering research and applications, Vol.4, Issue.3, pp. 65-71, March 2014.
[12] D.KARTHIKA and S. DIVAKAR .Improving Improving Improving Improving Improving the Efficiency he Efficiency of Fast Using Semantic . International Journal of Scientific and Research Publication, Vol.4, Issue.1, pp. 1-5, Jan 2014.
[13] A.Gowri Durga and A.Gowri Priya .Feature Subset Selection Algorithm for High Dimensional Data using Fast Clustering Method .International Journal of Computing and Technology, Vol.1, Issue.2, pp.240-242, March 2014.
[14] K.Suman and S.Thirumagal .Feature Subset Selection with Fast Algorithm Implementation.International Journal of Computer Trends and Technology, Vol.6, no. 1, pp.1-5, Dec 2013.
[15] Kumaravel.V and Raja.k, "Feature Subset Selection Algorithm for High-Dimensional Data by using FAST Clustering Approach", journal of computer science and engineering, pp. 21-25,2013.
[16] Leonard K.M, Poon, Nevin L, Tengfei Liu, Zhang and April H. Liu .Model-based clustering of high-dimensional data: Variable selection versus facet determination . IEEE international journal of approximate reasoning, vol.54, pp. 196-215, 2013.
[17] Xinyue Liu and Menggang Li .Integrated constraint based clustering algorithm for high dimensional data .SI computational intelligence techniques for new product development of neuro computing, Vol. 142, pp. 478-485, Oct 2014.
[18] Charles Bouveyron and Camille Brunet-Saumard .Model-based clustering of high-dimensional data: A review .IEEE international journal of computational statistics and data analysis, Vol. 71, pp. 52-78, March 2014.
[19] V.D.Ambeth Kumar, M.Ramakrishnan, V.D.Ashok Kumar, S.Malathi .Performance Improvement using an Automation System for Recognition of Multiple Parametric Features based on Human Footprint" for the International Journal of kuwait journal of science & engineering, Vol 42, No 1, pp:109-132, 2015