Applied Computing and Informatics (2016) 12, 117–127

Saudi Computer Society, King Saud University

Applied Computing and Informatics

(http://computer.org.sa)
www.ksu.edu.sa
www.sciencedirect.com

ORIGINAL ARTICLE

Multi Filtration Feature Selection (MFFS) to improve discriminatory ability in clinical data set

S. Sasikala [a,*], S. Appavu alias Balamurugan [b], S. Geetha [c]

[a] Anna University, Tamil Nadu, India
[b] K.L.N. College of Information Technology, Tamil Nadu, India
[c] Thiagarajar College of Engineering, Tamil Nadu, India

Received 13 June 2013; revised 21 March 2014; accepted 29 March 2014
Available online 5 April 2014

KEYWORDS

Medical data mining;
Biomedical classification;
Variance coverage factor;
Principal Component Analysis;
Multi Filtration Feature Selection

Abstract Selection of optimal features is an important area of research in medical data mining systems. In this paper we introduce an efficient four-stage procedure – feature extraction, feature subset selection, feature ranking and classification, called as Multi-Filtration Feature Selection (MFFS), for an investigation on the improvement of detection accuracy and optimal feature subset selection. The proposed method adjusts a parameter named "variance coverage" and builds the model with the value at which maximum classification accuracy is obtained. This facilitates the selection of a compact set of superior features, remarkably at a very low cost. An extensive experimental comparison of the proposed method and other methods using four different classifiers (Naïve Bayes (NB), Support Vector Machine (SVM), multi layer perceptron (MLP) and J48 decision tree) and 22 different medical data sets confirm that the proposed MFFS strategy yields promising results on feature selection and classification accuracy for medical data mining field of research.

1. Introduction

Data mining application in medicine has proved to be a successful strategy in the areas of medical services including prediction of usefulness of surgical procedures, clinical tests, medication procedures, and the discovery of associations among clinical and diagnosis data [37]. The applicability of data mining for healthcare applications is increasingly gaining importance. The availability of diverse-natured medical data for diagnosis and prognosis and of pervasive data mining techniques to process these data offers medical data mining a distinctive place to truly assist and impact patient care.

Due to proliferation of synergized information from enormous patient repositories, there is a paradigm shift in the insight of patients, clinicians and payers from qualitative analysis of clinical data to demanding a better quantitative visualization of information based on all supporting medical data.

* Corresponding author. Tel.: +91 9443831534.
E-mail addresses: nithilannsasikala@yahoo.co.in (S. Sasikala), app_s@yahoo.com (S. Appavu alias Balamurugan), sgeetha@tce.edu (S. Geetha).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.aci.2014.03.002
2210-8327

For instance, the physicians can evaluate the diagnostic information of many patients with identical conditions. In the same way, they can verify their findings too, with the conformity of peer physicians working on similar cases in other parts of the world. The patterns that are discovered denote valuable knowledge that helps medical discoveries, for example discovering that a certain combination of features may help in better, and more accurate diagnosis of a particular disease. Accurate diagnosis of diseases and subsequently, providing efficient treatment, form an important part of valuable medical services given for patients in a health-care system.

The unique characteristics of medical databases that pose challenges for data mining are the privacy-sensitive, heterogeneous, and voluminous data. These data may have valuable information which awaits extraction. The required knowledge is found to be encapsulated in/as various regularities and patterns that may not be apparent in the raw data. Extracting such knowledge has proved to be priceless for future medical decision making. Feature selection is crucial for analysing various dimensional bio-medical data. It is difficult for the biologists or doctors to examine the whole feature-space obtained through clinical laboratories at one time. In machine learning, all the computational algorithms recommend only few significant features for disease diagnosis. Then these recommended significant features may help doctors or experts to understand the biomedical mechanism better with a deeper knowledge about the cause of disease and provide the fastest diagnosis for recovering the infected patients as early as possible.

Feature selection methods [12] tend to identify the features most relevant for classification and can be broadly categorized as either subset selection methods or ranking methods. The former type returns a subset of the original set of features which are considered to be the most important for classification. Ranking methods sort the features according to their usefulness in the classification task. Most of the classifiers, irrespective of the application domain, uses the ranking strategy to select the final feature subset, in an ad hoc manner. Feature selection, as a pre-processing step to machine learning, is prominent and effective in dimensionality reduction, by removing irrelevant and redundant data, increasing learning accuracy, and improving result comprehensibility. Feature selection algorithms generally fall into two broad categories, the filter model and the wrapper model [37].The filter model depends on general characteristics of the training data to select some features without involving any learning algorithm. The filter model assesses the relevance of features from data alone, independent of classifiers, using measures like distance, information, dependency (correlation), and consistency. The filter method is further classified into Feature Subset Selection (FSS) and Feature Ranking (FR) methods. The wrapper model needs one predetermined learning algorithm in feature selection and uses its performance to evaluate and determine which features are selected. For each of the generated new subset of features, the wrapper model is supposed to learn the hypothesis of a classifier. It has a propensity to find features better suited to the predetermined learning algorithm resulting in superior learning performance, but it also tends to take more computation time and is economically more expensive than the filter model [37]. Whenever dealing with a large number of features, the filter model is usually chosen due to its high accuracy [9]. The hybrid model takes the advantages of the two previous models, and uses an independent measure to identify the best subsets for a given cardinality and applies a mining algorithm to select the best subset among all best subsets across different cardinalities. However, the ensemble of a filter based model with another filter based model, once for subset selection and again for ranking proves to be a promising approach, for medical data mining. The ensemble is brought about in a fashion so as to reduce the number of features and also to enhance the classification accuracy.

The objective of this research work is aimed at showing that the selection of more significant features from the available raw medical dataset helps the physician to arrive at an accurate diagnosis. The primary focus is on aggressive dimensionality reduction so as to end up with increase in the prediction accuracy. The features are subjected to a double filtration process, at the end of which, only the features that increase the accuracy, and form the subset with the lowest cardinality, with their corresponding rank, are obtained. The method employs an efficient strategy of ensemble feature correlation with ranking method. The empirical results show that the proposed Multi Filtration Feature Selection (MFFS) embedded classifier model achieves remarkable dimensionality reduction in the 22 medical datasets obtained from the UCI Machine Learning repository [10] and Kentridge repository [13].

2. Related work

Numerous works have been carried out in the field of dimensionality reduction for medical diagnosis. The following section presents the summary of those works, highlighting the strengths and weaknesses of each method.

It could be observed that the naive Sequential Forward Feature Selection (SFFS) (pure wrapper approach) [5] is impractical for feature subset selection from a large number of samples of high-dimensional features. Hence Gan et al. [4] proposed the Filter-Dominating Hybrid Sequential Forward Feature Selection (FDHSFFS) algorithm for high dimensional feature subset selection. This method proved to be fast but demanded huge computational complexity. Another variant of the SFFS method called improved F-score and Sequential Forward Search (IFSFS) was proposed by Xie and Wang [36] for feature selection to diagnose erythemato-squamous disease. This method was designed so as to improve the F-score and measured the discrimination between more than two sets of real numbers instead of measuring between only two sets of real numbers. The method's applicability to other medical data sets was not reported and hence it was a very specific system targeted at the diagnosis of erythemato-squamous disease only.

Another category of feature selection methods used Mutual Information score. Vinh et al. [32] proposed a novel feature selection method based on the normalization of this well-known mutual information measurement and utilized the information measurement to estimate the potential of the features. The method could not eclipse the strongly correlated features impact on the classification results. Correlated features may be accounted for redundancy and hence a single representative feature from that subset may be selected for further processing.

An incremental learning algorithm in which the most informative features are learnt at each step, is proposed by

Ruckstieb et al. [26] and is called as Sequential Online Feature Selection (SOFS). Another Scatter Search-based approach coupled with Decision Trees (SS + DT) is proposed by Lin and Chen [17]. The method acquired optimal parameter settings and selected the beneficial subset of features that resulted in better classification results. In [16] Koprinska empirically evaluated feature selection methods for classification of Brain–Computer Interface (BCI) data. A new feature selection method based on rough set theory has been proposed by Paul and Maji [23]. The proposed method identified discriminative and significant genes from high-dimensional microarray gene expression data sets.

Correlation Based Filter [3,18] is another strategy for feature selection. Ensemble methods have also been proposed. Raymer et al. [25] proposed a hybrid algorithm that coupled a genetic algorithm with k-nearest-neighbour classifier and applied it for protein–water binding from X-ray crystallographic protein structure data. MonirulKabi et al. [20] presented a new Hybrid Genetic Algorithm (HGA) for Feature Selection (FS), called as HGAFS. It employed a new local search operation that is devised and embedded in HGA to fine-tune the search in feature selection process. The search process is guided in such a way that the less correlated (distinct) features consisting of general and special characteristics of a given data set are generated in subsequent iterations.

A new approach called Redundancy Demoting (RD) has been proposed by Osl et al. [22]. It takes an arbitrary feature ranking as input, and performs improvement in ranking by identifying redundant features and demoting them to positions in the ranking in which they are not redundant. Hybrid schemes that combine wrapper-based and filter-based approaches are also in the literature [2,11,30] are such schemes where the features are ranked and then selected so as to offer superior classification accuracy. In the first stage, the filter model is used to rank the features by the relief algorithm and then the highest relevant features are chosen to the classes with the help of the threshold. In the second stage, they used shapely values to evaluate the contribution of features to the classification task in the ranked feature subset. Tanwani et al. [31] gave a study on comprehensive evaluation of a set of diverse machine learning schemes on a number of biomedical datasets. Sanchez-Monedero et al. [27] studied and proposed the suitability of Extreme Learning Machines (ELM) for resolving bio-informatics and biomedical classification problems.

After reviewing the works on feature selection for medical dataset [29] it is observed that most of the existing methods suffer from the following problems: (1) depending on the complexity of the search method, the iterations of evaluations are too large; (2) they rely on a univariate ranking that does not take into account interaction between the variables already included in the selected subset and the remaining ones. Moreover, a method that produces the best accuracy employs more number of features and hence more running time is involved in the construction of the respective classifiers. Contrarily, a method that outputs the fewest number of features produces inferior detection accuracy. A holistic and universal method that achieves the best classification accuracy with fewest features possible is still an open research problem. This paper makes an attempt to design such a feature selection sequence and it is called as "Multi Filtration Feature Selection (MFFS)".

This paper is organized as follows: Section 3 describes the proposed method with the suitable algorithm. Experimental results and discussions are presented in Section 4. The paper is concluded with a mention on the future scope of this work.

3. Proposed system

The proposed method involves four stages. The entire system flow of the proposed model is shown in Fig. 1. The individual stages are described in the following text.

3.1. System flow of the proposed method

3.1.1. Stage 1 – Relevant Feature Generation Phase (RFGP)

A representative of unsupervised dimensionality reduction method is Principal Component Analysis (PCA) [14,39] which aims at identifying a lower-dimensional space maximizing the variance among data [38]. PCA is a very effective approach of extracting features [6,21].

Let us denote the multi-dimensional dataset in the form of a matrix, A. The dimensions actually represent directions along which the data vary. The feature generation process, which removes the irrelevant features and redundant features, mainly finds an approximate "basis" to the set of directions. Only the crucial dimensions that serve as the corner stone upon which other dimensions are dependent are generated from the given dataset. The redundant-duplicate dimensions are finally eliminated with the sense that they can be reconstructed easily from the available set of basis dimensions. This is equivalent to finding the dimensions with maximal variance, since the points are found to be constant approximately along other dimensions. Variance factor is an important measure that denotes the degree of data spread in a multi-dimensional dataset. Thus dimensionality reduction is effectively contributed from this first step of filtration by choosing appropriate variance factor at which the system yields the minimum number of features with maximum accuracy.

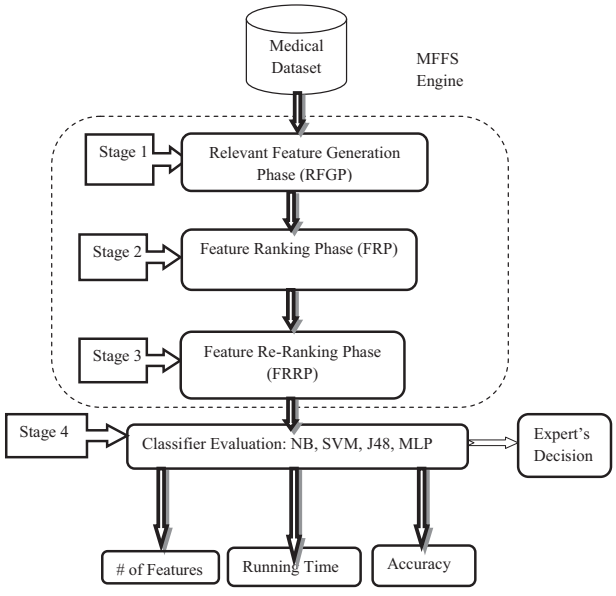

Fig. 1 System flow of the proposed MFFS model.

The basic theoretical idea behind PCA is finding the principal components of the dataset that correspond to the components along which the variation is the most. This is achieved by finding the covariance matrix, i.e., we find the principal components of the data, which correspond to the components along which there is the most variation. This can be done using the covariance matrix, AA^T for our input matrix A, as follows.

Let the eigen values be represented as λ_i for the covariance matrix. Then, the corresponding diagonal matrix is given in Eq. (1) as:

$$L = \begin{bmatrix} \lambda_1 & 0 \dots 0 \\ 0 & \lambda_2 \dots 0 \\ & \vdots \\ 0 & 0 \dots \lambda_n \end{bmatrix} \tag{1}$$

The eigenvectors v_i of the matrix should satisfy AA^T as given in Eq. (2) as

$$AA^T v_i = \lambda_i v_i \tag{2}$$

On rewriting eigenvectors of the dataset as the rows of a matrix P, the system becomes

$$AA^T P = LP \tag{3}$$

It is apparent from Eq. (3) that the columns of this matrix P represents the principal components of the original matrix and hence confines to the directions of most variance [39,38].

PCA employs the entire features and it acquires a set of projection vectors to extract global feature from given training samples. The approach mainly consists of three primary processes such distinction process, binary session and pattern generation [29]. All these flavours make PCA [21] more suitable for applying on medical datasets, which typically have these characteristics. The variance coverage factor is playing a significant role in deciding the important features and hence this parameter is tuned so as to capture the classifier model with the best results.

3.1.2. Stage 2 – Feature Ranking Phase (FRP)

The correlation between each feature and the class and between two features can be measured and best-first search can be exploited in searching for a feature subset of maximum overall correlation to the class and minimum correlation among selected features. This is realized in the Correlation-based Feature Selection (CFS) method [7]. Correlation based Feature Selection is an algorithm that couples this evaluation formula with an appropriate correlation measure and a heuristic search strategy. CFS quickly identifies and screens irrelevant, redundant, and noisy features, and identifies relevant features as long as their relevance does not strongly depend on other features. CFS is a fully automatic algorithm—it does not require the user to specify any thresholds or the number of features to be selected, although both are simple to incorporate if desired. CFS operates on the original (albeit discretized) feature space, meaning that any knowledge induced by a learning algorithm, using features selected by CFS, can be interpreted in terms of the original features, not in terms of a transformed space. Most importantly, CFS is a filter, and, as such, does not incur the high computational cost associated with repeatedly invoking a learning algorithm.

The suggestion used by the CFS is on the basis that always features strongly correlated with the predicted class form the good feature subset than the features correlated with each other. The feature subset created by the CFS is computed by the merit of the feature subset 'S' containing 'k' features as in Eq. (4).

The following equation provides the merit of the feature subset 'S'.

$$\text{Merit}_s = \frac{k\overline{\omega_{cf}}}{\sqrt{k + k(k-1)\overline{\omega_{ff}}}} \tag{4}$$

where Merit_s is the evaluating hypothesis of a feature subset 'S' containing 'k' features, $\overline{\omega_{fc}}$ is the average value of feature–class correlation, and $\overline{\omega_{ff}}$ is the average value of feature–feature inter correlation.

The correlation between two entities 'i' and 'j', ω_{ij} is calculated as in Eq. (5)

$$\omega_{ij} = \frac{\sum(\mathrm{i} - \bar{i})(j - \bar{j})}{\sqrt{[\sum (i - \bar{i})^2][\sum (j - \bar{j})^2]}}. \tag{5}$$

where 'i' is the record's value of the independent variable and 'j' is record's value of the dependent variable which may be either feature or class label. $\bar{i}$ and $\bar{j}$ are the means of the values of the independent and dependent variables, respectively.

3.1.3. Stage 3 – Feature Re-Ranking Phase (FRRP)

In spite of feature extraction and selection, a problem is persistent namely the classifier may be biased towards the attributes with more values. Hence this biased nature has to be eliminated for which we employ Symmetrical Uncertainty (SU). It overcomes the problem of bias towards attributes with more values, by dividing information gain by the sum of the entropies of feature subsets S_i and S_j.

Symmetry is a desired property for a measure of correlations between features. However, information gain is biased in favour of features with more values. Furthermore, the values have to be normalized to ensure they are comparable and have the same influence. Therefore, we choose symmetrical uncertainty. It compensates for information gain's bias towards features with more values and normalizes its values to the range [0; 1] with value 1 indicating that knowledge of the value of either one completely predicts the value of the other and value 0 indicating that X and Y are independent. In addition, it still treats a pair of features symmetrically. Entropy-based measures require nominal features, but they can be applied to measure correlations between continuous features as well, if the values are discretized properly in advance. Therefore, we use symmetrical uncertainty in this work.

As CFS uses the best-first strategy search method to calculate the merit of the feature subset, however there is a necessity to fix the stopping criteria. Due to this strictly needed constrain correlation between features is computed using Symmetrical Uncertainty (SU) as specified in Eq. (6).

$$SU = 2.0 \times \left[\frac{H(S_j) + H(S_i) - H(S_i, S_j)}{H(S_j) + H(S_i)}\right] \tag{6}$$

where $H(S_j)$ and $H(S_i, S_j)$ are defined as in Eqs. (7) and (8) as:

Input :Training set with '*n*' features and '*m*' samples - $\{(x_{1,1},x_{1,2},...,x_{1,n},y_1).....(x_{m,1},x_{m,2},...,x_{m,n},y_m)\}$

Output : Best Feature Subset,MaxAccuracy,MinError,respectiveVariance Coverage Factor(δ)

//Initialization

1.δ ← 0.45,BestFeatureSubset ← {},S ← ∅

Phase I - Relevant Feature Generation Phase:

2.By forward feature selection strategy, populate features into S, where S is the feature subset.

3. For each FeatureSubset S,

4.ExtractedData = evaluatePCA(S,δ)

5.if (ExtractedData>BestFeatureSubset) then

6.BestFeatureSubset=ExtractedData

7.δ=δ+0.05

8.Repeat steps 2 to 7 until δ=0.95; Record the δ value and the respective BestFeatureSubset that gives maximum accuracy.

Phase II - Feature Ranking Phase:

9.Perform Correlation-based heuristic evaluation on BestFeatureSubset using

$$\text{Merit}_S = \frac{k\overline{\omega_{cf}}}{\sqrt{k+k(k-1)\overline{\omega_{ff}}}}$$

10.Arrange the BestFeatureSubset in the decreasing order of the Merit score.

Phase III - Feature Re-Ranking Phase:

11.Rank[]=Create a Rank for BestFeatureSubset by usign Ranking-based heuristic of symmetrical uncertainty using

$$SU = 2.0 \times S\left[\frac{H(S_j)+H(S_i)-H(S_i,S_j)}{H(S_j)+H(S_i)}\right]$$

12.Return the re-ranked BestFeatureSubset

Phase IV - Classifier Evaluation Phase:

13.Run a 10-fold CrossValidation on the Original feature set and BestFeaturesubset using the Naive Bayes,SVM, J48,MLP classifier models.

14.Record the model that yields the maximum accuracy and minimum error.

15.Return BestFeatureSubset,MaxAccuracy,MinError,respectiveVariance Coverage Factor(δ)

Fig. 2 Algorithm of the proposed MFFS model.

$$H(S_j) = -\sum_{f s \varepsilon FS_j} p(S_j)\log_2(p(S_j)) \tag{7}$$

where a realistic model of a feature S_j can be formed by evaluating the training data, considering the individual's probability values of S_j. A new feature subset S_i can be worked out by partitioning the previously existing feature subset S_j, then the relationship between subsets S_i and S_j is given by:

$$H(S_i, S_j) = -\sum_{x \varepsilon X} P(S_i)\sum_{y \varepsilon Y} P(S_i/S_j)\log_2 P(S_i/S_j) \tag{8}$$

The algorithm is better explained by the Fig. 2.

3.1.4. Stage 4 – Classifier Evaluation

The proposed system is validated against standard successful classifier models [35]. Classifiers are constructed with the final subset of features obtained after subjecting the datasets to RFGP, FRP and FRRP steps sequentially. A detailed insight into various classifiers is presented in Section 4.4.

3.2. Algorithm for the proposed MFFS model

Traditional filter approaches usually select the top ranked features or eliminate the irrelevant features by using a threshold criterion. Since prediction is made after the single filtering phenomenon, they report feeble accuracy. Alternatively, when filtering is done, more than once, an improved accuracy may be obtained. Hence the proposed scheme is designed with Multi Filtration Feature Selection (MFFS) as the central logic.

It consists of the following steps:

4. Test results and discussion

4.1. Test scenario

Empirical study with the synthetic datasets has been executed, to investigate the performance of the proposed algorithm in the following perspectives:

1. Classification accuracy
2. Number of features selected
3. Average running time

Datasets, test set-up, procedure and objectives for the tests necessary for the evaluation of these goals are described below.

4.2. Datasets

The proposed approach has been evaluated by experiments on 22 biomedical datasets from the UCI machine learning repository [10] and Kentridge repository [13]. The 22 biomedical datasets used to test the proposed approach are summarized in Table 1. The last column in Table 1 indicates the "imbalance

ratio'' present in each dataset. It is the ratio between the cardinality of the class with the maximum instance to the cardinality of the class with the minimum instance.

4.3. Test set-up

The tests are carried out in a system with Intel i5, 8 GB RAM, DDR3, 500 GB hard drive on a Windows XP operating system. The proposed algorithm is implemented using Weka [34]. WEKA is acknowledged as a landmark system in the field of machine learning and data mining. It has attained widespread acceptance among the academia and industry spheres, and has become a widely used tool for data mining research. Another flavour that is highly encouraging is its ''Open Source'' nature. The free access given to the source code has enabled us to develop and customize the modules matching our work. The stepwise approach is as follows. The input to the system is given in the Attribute-Relation File Format (ARFF). The proposed algorithm is executed and the features in the ranked order are obtained as the output. A result is created in Weka using the name specified in \@relation''. The attributes specified under \@attribute'' and instances specified under \@data'' are retrieved from the ARFF file and then they are added to the created table. 10-fold cross validation is performed for all classifiers [8]. Fifty runs were done for each classification algorithm on each dataset with features selected by MFFS method. In each run, a dataset was split into training and testing set, randomly. The results obtained are shown in Tables 2–9.

4.4. Classification Models

4.4.1. Model M1 – Naïve Bayes (NB)

Naïve Bayesian Classifier is a simple probabilistic classifier [35] with an assumption of conditional independence among the features, i.e., the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. It only requires a small amount of training data to estimate the parameters necessary for classification. Many experiments have demonstrated that NB classifier has worked quite well in various complex real-world situations and outperforms many other classifiers. Kernel estimation has been used in cases of datasets with numerical attributes. Also supervised discretization is done for converting numerical attributes to nominal ones.

4.4.2. Model M2 – Support Vector Machine (SVM)

SVM [13,19] finds the hyper plane with maximum margin in between two classes. The Support Vector Machine (SVM) is actually based on learning with kernels some of which form the support vectors. A great advantage of this technique is that it can use large input data and feature sets. Thus, it is easy to test the influence of the number of features on classification accuracy. We implemented SVM classification [28] for two types of kernels: polynomial kernel and Gaussian kernel (Radial Basis Function – RBF). The SVM model with complexity parameter C as 1.0, epsilon as 1.0E−12, normalized training data, RBF kernel with gamma as 0.0.1, and tolerance parameters as 0.0010 produced the best results.

4.4.3. Model M3 – Decision Tree (DT)

A decision tree [1,33] is a predictive machine-learning model that decides the target value (dependent variable) of a new sample based on various attribute values of the available data. Decision tree's internal node represents different attributes; the branches between the nodes tell us the possible values that these attributes can have in the observed samples, while the end nodes are the target class labels. The J48 decision tree classifier [24] operates on the basis of constructing a tree and branching it based on the attribute with the highest information gain. The J48 tree with binary split allowed a confidence factor of 0.25 and reduced error pruning is employed.

4.4.4. Model M4 – Multilayer perceptron (MLP)

An MLP [15] can be viewed as a logistic regression, where the input is first transformed using a learnt non-linear transformation. The purpose of this transformation is to project the input data into a linearly separable space. This intermediate layer is referred to as a hidden layer. We have employed a back-propagation network with 0.3 as learning rate and 0.02 as momentum. The attributes are normalized in the range of (0.1, 0.9). The training was carried out for 500 epochs.

4.5. Test procedure

For realizing dimension reduction via PCA, the orthogonal base components are obtained out of the datasets through linear transformation. For the evaluation the PCA variance coverage parameter δ is varied in the range of (0.45, 0.95). In preliminary test δ values outside this range did not lead to useful results for the analysis. So we tested in this range. The proposed MFFS calculates the correlations of feature-class and feature–feature using CFS and the feature subset space is searched. The subset with the highest merit is subjected to analysis. Then the resulting subset is re-ranked using symmetrical uncertainty principle. For performance analysis, the models M1-M4, generated according to the earlier discussion are applied to the optimal feature subset, returned after MFFS steps. The best accuracy in percentage along with the respective variance coverage and the number of features involved to achieve it are returned.

4.6. Test objectives

From the goals stated above, the following objectives are established:

O1 – enhancing the detection accuracy for the classifier model, which is measured using the detection accuracy metric expressed in percentage.
O2 – reducing the number of features so as to achieve the best accuracy in each classifier model.
O3 – reducing the running time for model generation which is measured in seconds.
O4 – determining the influence of the variance coverage of the PCA feature extraction model on the MFFS performance.

Test objectives O1 and O2 are the obvious test goals within the focus of this work. High detection accuracy with the least

number of features, which are shown in Tables 2–9 are proving the usefulness of applying the proposed MFFS scheme to feature selection.

Test objectives O3 is aimed at determining the overall quality of our feature selection approach. Test objective O4 is formulated to address the impact of the variance coverage of the PCA on the MFFS performance over each model.

To facilitate a logical sequence for the presentation of our research results, the test objectives framed based on the goals are ordered in a way to glide from the most specific to a more general case. Tables 2–9 summarize the evaluation of test objectives O1–O4.

4.7. Test results and discussion

In Table 2, the average classification accuracy of the chosen algorithms over unprocessed datasets is provided. Tables 3–8 show the best average classification accuracy with the four classifiers on each dataset and the best accuracy in each case is highlighted in bold typeface. Table 9 shows the average running time in seconds taken by the proposed system. Figs 3–11 show the performance of the proposed system.

It can be seen from Tables 2–8, that the classification accuracy based on the selected subsets by the proposed MFFS scheme is better than that based on the original feature set. This indicates that the selected feature subsets are representative and informative and, thus, can be used instead of the complete data for pattern classification. The list of such selected features is shown in Table 7.

From the empirical results obtained so far, it is worth noting that each method has its strengths and limitations. In particular, CFS obtains good classification accuracy in the least amount of running time but at the expense of selecting many more features; PCA selects the least number of features but suffers in terms of classification accuracy and also requires more running time than others; MFFS attains the best accuracy and robustness in a reasonable time with lowest number of features. Considering all these factors, the proposed MFFS scheme shows overall better performance than other methods. Tables 10 and 11 summarize and compare characteristics of

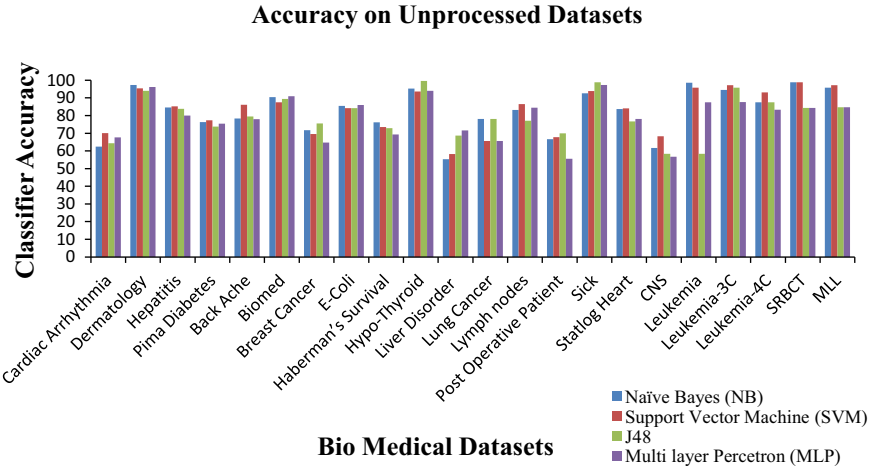

Fig. 3 Performance comparison of different classifiers on unprocessed biomedical data sets.

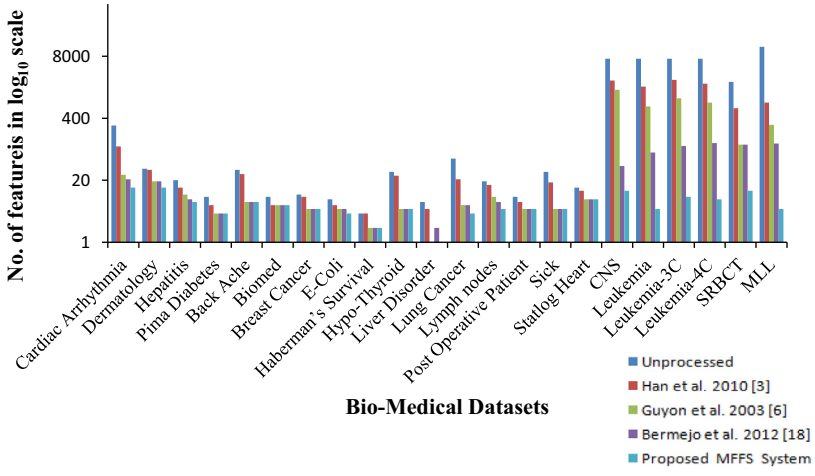

Fig. 4 Number of features selected by the existing systems and proposed MFFS (in $\log_{10}$ scale for uniform scaling) with Naïve Bayes classifier.

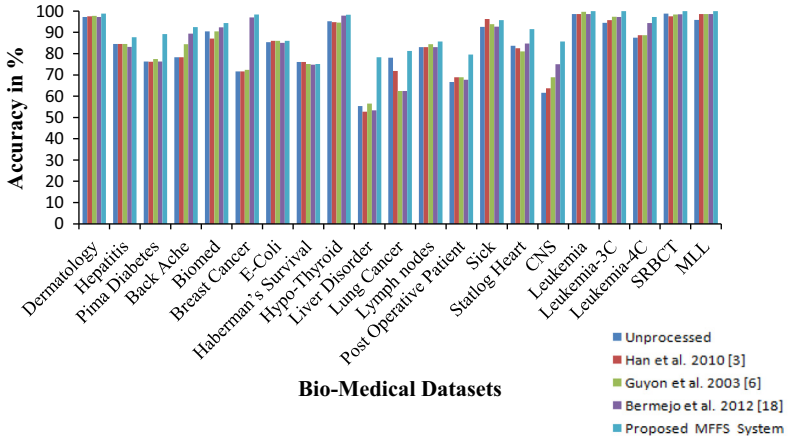

Fig. 5 Classification accuracy obtained for existing and the proposed MFFS by Naïve Bayes classifier.

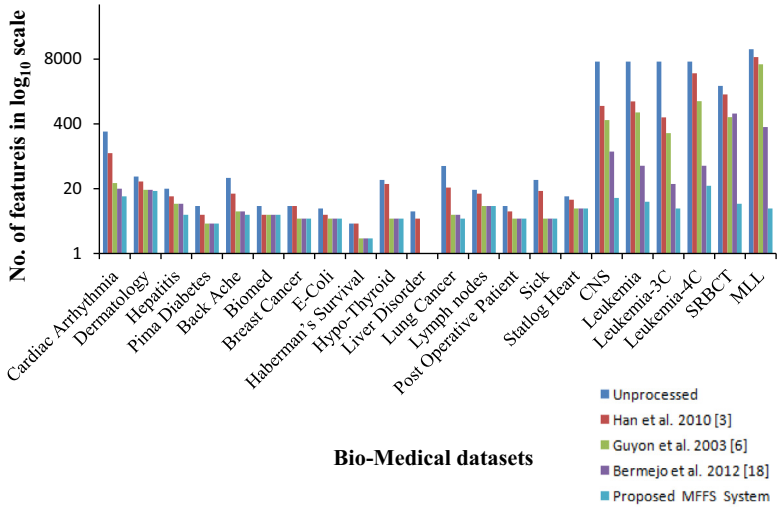

Fig. 6 Number of features selected by the existing systems and the proposed MFFS (in $\log_{10}$ scale for uniform scaling) with SVM classifier.

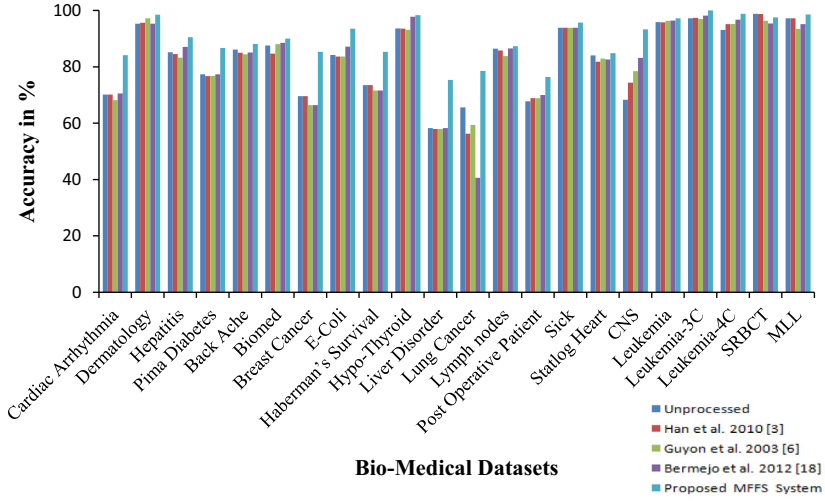

Fig. 7 Classification accuracy obtained for the existing systems and the proposed MFFS by SVM classifier.

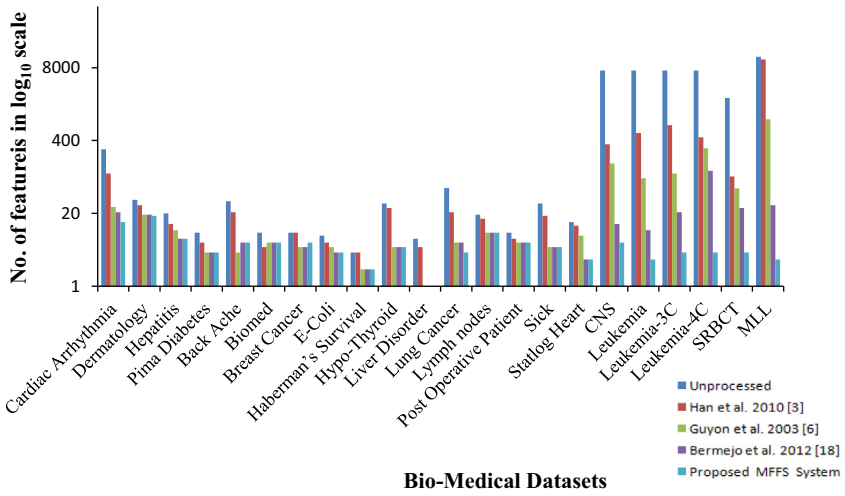

Fig. 8 Number of features selected by the existing systems and the proposed MFFS (in $\log_{10}$ scale for uniform scaling) with J48 classifier.

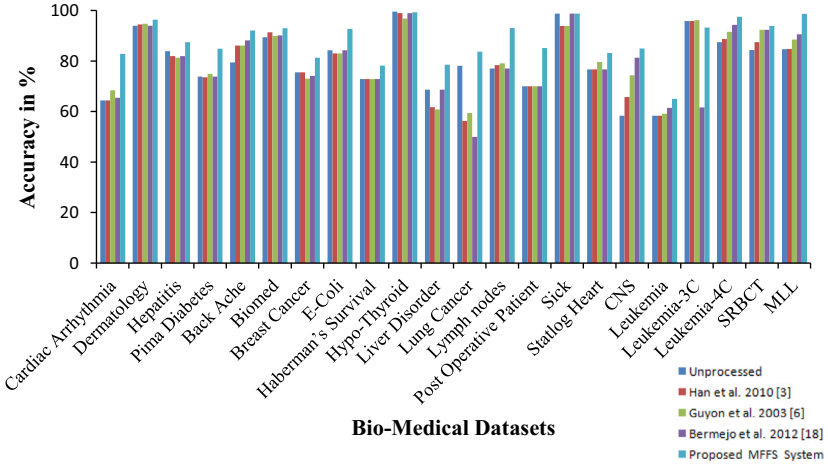

Fig. 9 Classification accuracy obtained for the existing systems and the proposed MFFS by J48 classifier.

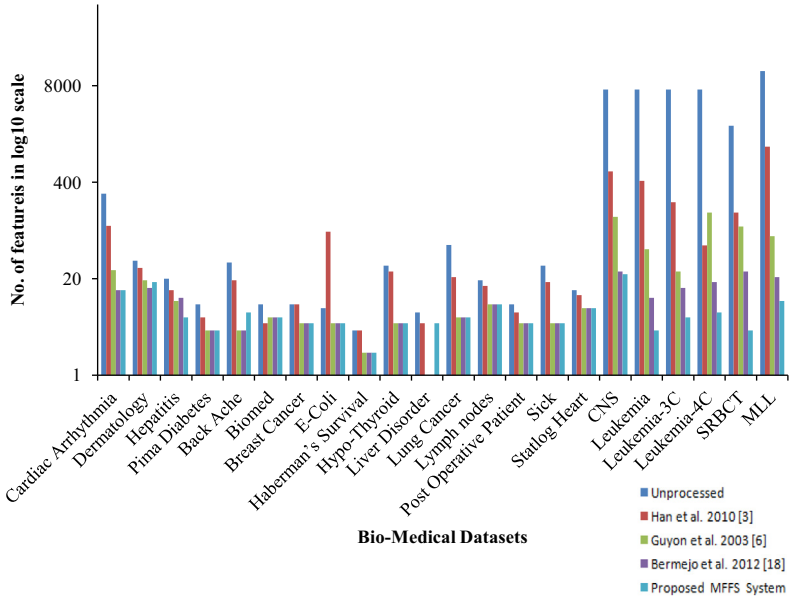

Fig. 10 Number of features selected by the existing systems and the proposed MFFS (in $\log_{10}$ scale for uniform scaling) with MLP classifier.

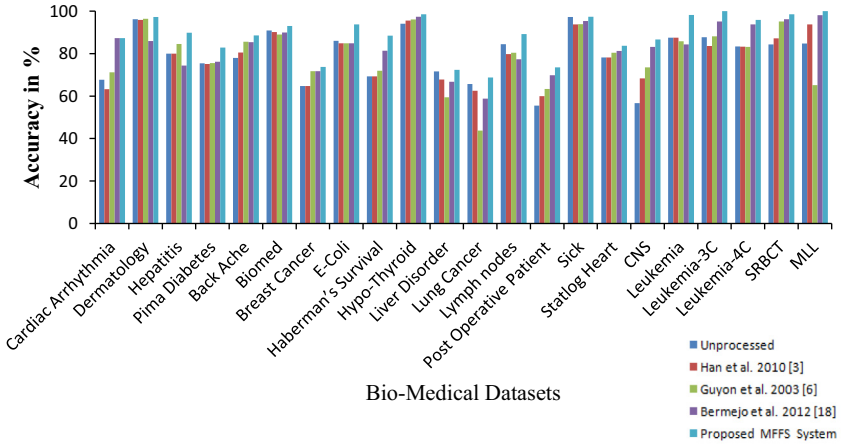

Fig. 11 Classification accuracy obtained for the existing systems and the proposed MFFS by MLP classifier.

our proposed method for selective 6 datasets with those of other previous works in the literature.

5. Conclusion

In this paper, we have proposed an efficient Multi Filtration Feature Selection (MFFS) method applicable to medical data mining. Empirical study on 6 synthetic medical datasets suggests that MFFS gives better over-all performance than the existing counterparts in terms of all three evaluation criteria, i.e., number of selected features, classification accuracy, and computational time. The comparison to other methods in the literature also suggests MFFS has competitive performance. MFFS is capable of eliminating irrelevant and redundant features based on both feature subset selection and ranking models effectively, thus providing a small set of reliable features for the physicians to prescribe further medications.

For simplicity, several key points are collected as follows.

(1) It seems that the classification performance is necessarily proportional to the removal of redundant features, heavily dependent on the inclusion of relevant features and the "Accuracy" metric is observed maximum with minimum number of features.
(2) The proposed MFFS algorithm operates invariably well on any type of classifier model. This shows the generalization ability and applicability of the proposed system.
(3) Our training and test database collects the popular and benchmark medical datasets. However, the proposed method can be tested and applied on real-world dataset too.
(4) The best accuracy rate achieved by our proposed system is superior to the existing schemes.

To make our system more practical, future work could include the following.

(a) Fitting the proposed system to classify any other real-world dataset.
(b) Applying the proposed method for a multi-label dataset, where a record may belong to many classes simultaneously.
(c) Ensemble with some optimization strategies like Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), and Genetic Algorithm (GA) etc.

Summarily, MFFS can be expected to serve as an excellent alternative for feature selection in the field of medical data mining.

Acknowledgment

This work is supported in part by the University Grant Commission (UGC), New Delhi, INDIA – Major Research Project under Grant No. F.No.:39-899/2010 (SR).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version,at http://dx.doi.org/10.1016/j.aci.2014.03.002.

References

[1] I.S.I. Abuhaiba, Efficient OCR using simple features and decision trees with backtracking, Arabian J. Sci. Eng. 31 (2B) (2006) 223–244.
[2] P. Bermejo, L.D.L. Ossa, J.A. Gamez, J.M. Puerta, Fast wrapper feature subset selection in high-dimensional datasets by means of filter re-ranking, Knowl.-Based Syst. 25 (1) (2012) 35–44.
[3] Y. Chen, S. Yu, Selection of effective features for ECG beat recognition based on nonlinear correlations, Artif. Intell. Med. 54 (1) (2012) 43–52.
[4] J.Q. Gan, B.A.S. Hasan, C.S.L. Tsui, A filter-dominating hybrid sequential forward floating search method for feature subset selection in high-dimensional space, Int. J. Mach. Learn. Cybern. 3 (4) (2012) 1–8.
[5] A.I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn Res. 3 (1) (2003) 1157–1182.
[6] H.U. Guz, A hybrid system based on information gain and principal component analysis for the classification of transcranial Doppler signals, Comput. Methods Programs Biomed. 107 (3) (2011) 598–609.
[7] M.A. Hall, Correlation Based Feature Selection for Machine Learning (PhD Dissertation), Dept. of Comp. Science, Univ. of Waikato, Hamilton, New Zealand, 1998.

[8] J.W. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2006.
[9] Y. Han, L. Yu, A variance reduction framework for stable feature selection, in: G.I. Webb, B. Liu, C. Zhang, D. Gunopulos, X. Wu (Eds.), Data Mining, IEEE Computer Society, Sydney, Australia, 2010, pp. 206–215.
[10] S. Hettich, C. Blake, C. Merz, UCI Repository of Machine Learning Databases, 1998 <http://www.ics.uci.edu/mlearn/MLRepository.html> (accessed 10 June 2013).
[11] B. Hualonga, X. Jing, Hybrid feature selection mechanism based high dimensional datasets reduction, Energy Procedia (2011) 30–38.
[12] M.M. Jazzar, G. Muhammad, Feature selection based verification/identification system using fingerprints and palm print, Arabian J. Sci. Eng. 38 (4) (2013) 849–857.
[13] L. Jinyan, L. Huiqing, Kentridge Bio-Medical Data Set Repository, 2002 <http://datam.i2r.a-star.edu.sg/datasets/krbd/> (accessed 10 June 2013).
[14] I.T. Jolliffe, Principal Component Analysis, Springer-Verlag, New York, NY, 1986.
[15] G. Kim, Y. Kim, H. Lim, H. Kim, An MLP-based feature subset selection for HIV-1 protease cleavage site analysis, J. Artif. Intell. Med. 48 (2) (2010) 83–89.
[16] I. Koprinska, Feature selection for brain–computer interfaces, in: T. Theeramunkong et al. (Eds.), Pacific Asia Conference on Knowledge Discovery and Data Mining PAKDD'09, Springer-Verlag, Berlin, Heidelberg, 2010, pp. 100–111.
[17] S. Lin, S. Chen, Parameter determination and feature selection for C4.5 algorithm using scatter search approach, Int. J. Soft Comput. 16 (1) (2011) 63–75.
[18] X. Lu, X. Peng, P. Liu, Y. Deng, B. Feng, B. Liao, A novel feature selection method based on CFS in cancer recognition, in: L. Chen, X. Zhang, L. Wu, Y. Wang (Eds.), Systems Biology, IEEE Computer Society, China, 2012, pp. 226–231.
[19] S.A. Mahmoud, S.M. Awaida, Recognition of off-line handwritten Arabic (Indian) Numerals using multi-scale features and support vector machines vs. hidden Markov models, Arabian J. Sci. Eng. 34 (2B) (2009) 429–444.
[20] M.d. MonirulKabi, M.d. Shahjahan, Murase Kazuyuki, A new local search based hybrid genetic algorithm for feature selection, Int. J. Neurocomput. 74 (17) (2011) 2914–2928.
[21] K. Moutselos, I. Maglogiannis, A. Chatziioannou, Integration of high-volume molecular and imaging data for composite biomarker discovery in the study of melanoma, Biomed. Res. Int. (2014), http://dx.doi.org/10.1155/2014/145243.
[22] M. Osl, S. Dreiseit, F. Cerqueira, M. Netzer, B. Pfeifer, C. Baumgartner, Demoting redundant features to improve the discriminatory ability in cancer data, J. Biomed. Inform. 42 (4) (2009) 721–725.
[23] S. Paul, P. Maji, Rough set based gene selection algorithm for microarray sample classification, in: Methods and Models in Computer Science, IEEE Computer Society, 2010, pp. 7–13.
[24] R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo, CA, USA, 1993.
[25] M.L. Raymer, T.E. Doom, L.A. Kuhn, W.F. Punch, Knowledge discovery in medical and biological datasets using a hybrid Bayes classifier/evolutionary algorithm, IEEE Trans. Syst. Man Cybern. 33 (5) (2003) 802–813.
[26] T. Ruckstieb, C. Osendorfer, P.V.D. Smagt, Minimizing data consumption with sequential online feature selection, Int. J. Mach. Learn. Cybern. 4 (3) (2012) 235–243.
[27] J. Sanchez-Monedero, M. Cruz-Ramrez, F. Fernandez Navarro, J.C. Fernandez, P.A. Gutierrez, C. Hervas-Martnez, On the suitability of extreme learning machine for gene classification using feature selection, in: A.E. Hassanien, A. Abraham, F. Marcelloni, H. Hagras, M. Antonelli, T. Hong (Eds.), Intelligent Systems Design and Applications, 2010, pp. 507–512.
[28] A. Shawkat, K.A. Smith Miles, A meta learning approach to automatic kernel selection for support vector machines, Neurocomputing 70 (1-3) (2006) 173–186.
[29] Q. Shen, R. Diao, P. Su, Feature selection ensemble, in: A. Voronkov (Ed.), Computing, Springer-Verlag, 2011, pp. 289–306.
[30] P. Smialowski, D. Frishman, S. Kramer, Pitfalls of supervised feature selection, Bioinformatics 26 (3) (2010) 440–443.
[31] A.K. Tanwani, J.M. Afridi, M.Z. Shafiq, M. Farooq, Guidelines to select machine learning scheme for classification of biomedical datasets, in: C. Pizzuti, M.D. Ritchie, M. Giacobini (Eds.), European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics, Springer-Verlag, Berlin Heidelberg, 2009, pp. 128–139.
[32] L.T. Vinh, S. Lee, Y. Park, B.J. Auriol, A novel feature selection method based on normalized mutual information, Int. J. Appl. Intell 37 (1) (2011) 100–120.
[33] L.M. Wang, S.M. Yuan, L. Li, H.J. Li, Improving the Performance of Decision Tree: A Hybrid Approach. Conceptual modeling. Lecture Notes in Computer Science, vol. 3288, Springer, 2004, pp. 327–335.
[34] Weka 3: Machine Learning Software in Java, 2013. The University of Waikato Software Documentation <http://www.cs.waikato.ac.nz/_ml/weka> (accessed 10 June).
[35] H.I. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, San Francisco, CA, USA, 2000.
[36] J. Xie, C. Wang, Using support vector machines with a novel hybrid feature selection method for diagnosis of erythemato-squamous diseases, Expert Syst. Appl. 38 (5) (2011) 5809–5815.
[37] E. Xing, M. Jordan, R. Karp, Feature selection for high-dimensional genomic microarray data, in: C.E. Brodely, A.P. Danyluk (Eds.), Machine Learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001, pp. 601–608.
[38] S. Yazdani, J. Shanbehzadeh, Mohammad Taghi Manzuri Shalmani, RPCA: a novel pre-processing method for PCA, Adv. Artif. Intell. 2012 (1) (2012) 1–7.
[39] M. Zahedi, A.G. Sorkhi, Improving text classification performance using PCA and recall-precision criteria, Arabian J. Sci. Eng. 38 (8) (2013) 2095–2102.