COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 113 (2014) 465–473

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

A random forest classifier for lymph diseases

Ahmad Taher Azar[a,*], *Hanaa Ismail Elshazly*[b,c], *Aboul Ella Hassanien*[b,c], *Abeer Mohamed Elkorany*[b]

[a] *Faculty of Computers and Information, Benha University, Egypt*
[b] *Faculty of Computers and Information, Cairo University, Egypt*
[c] *Scientific Research Group in Egypt (SRGE), Egypt*

ARTICLE INFO

Article history:
Received 28 June 2013
Received in revised form
3 November 2013
Accepted 6 November 2013

Keywords:
Machine learning (ML)
Feature selection (FS)
Genetic algorithm (GA)
Random forest classifier (RFC)
Lymph diseases

ABSTRACT

Machine learning-based classification techniques provide support for the decision-making process in many areas of health care, including diagnosis, prognosis, screening, etc. Feature selection (FS) is expected to improve classification performance, particularly in situations characterized by the high data dimensionality problem caused by relatively few training examples compared to a large number of measured features. In this paper, a random forest classifier (RFC) approach is proposed to diagnose lymph diseases. Focusing on feature selection, the first stage of the proposed system aims at constructing diverse feature selection algorithms such as genetic algorithm (GA), Principal Component Analysis (PCA), Relief-F, Fisher, Sequential Forward Floating Search (SFFS) and the Sequential Backward Floating Search (SBFS) for reducing the dimension of lymph diseases dataset. Switching from feature selection to model construction, in the second stage, the obtained feature subsets are fed into the RFC for efficient classification. It was observed that GA-RFC achieved the highest classification accuracy of 92.2%. The dimension of input feature space is reduced from eighteen to six features by using GA.

1. Introduction

Computer aided diagnosis (CAD) systems have been used for many years. Medical diagnosis is known to be subjective and depends not only on the available data but also on the experience of the physician. It has been suggested that computer translation may hold part of the solution for processing the physician's interpretation [1,2]. Machine learning techniques are increasingly introduced to construct the CAD systems owing to its strong capability of extracting complex relationships in the biomedical data [3,4]. The high volume of medical data requires some helpful classification approaches to support the analysis of this data. Accuracy of classification algorithms used in disease diagnosing is certainly an important issue to be considered. Most medical data has the characteristic of high dimensionality datasets [5,6]. High dimensional data, in general, requires the extraction of most descriptive or discriminative features to be selected and hence the dimension of dataset is reduced [7]. In this context, dimension reduction plays an important role in diagnosing systems to remove irrelevant features from a data set [8,9]. Dimension reduction procedure is useful to decrease dataset complexity with the possible advantage of increased classification performance. Removing the number of irrelevant features for model implementation makes screening tests faster, more convenient and less costly. The current research work is focused on the determination of an optimal feature subset for

* *Corresponding author.*
E-mail addresses: ahmad.azar@fci.bu.edu.eg, ahmad_t_azar@ieee.org (A.T. Azar), hanosoma3002@yahoo.com (H.I. Elshazly), aboitcairo@gmail.com (A.E. Hassanien), a.korani@fci-cu.edu.eg (A.M. Elkorany).
0169-2607/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cmpb.2013.11.004

lymphography dataset in order to improve the diagnosis accuracy. The choice is a trade-off between computational time and quality of the generated feature subset solutions.

The lymphatic system aids the immune system in removing and destroying waste, debris, dead blood cells, pathogens, toxins, and cancer cells. It absorbs fats and fat-soluble vitamins from the digestive system and delivers these nutrients to the cells of the body where they are used by the cells. Also, it removes excess fluid, and waste products from the interstitial spaces between the cells. The lymphatic system consists of thin-walled lymphatic vessels, lymph nodes, and two collecting ducts [10]. Lymph vessels are closely associated with the circulatory system vessels. Larger lymph vessels are similar to veins. Lymph capillaries are scatted throughout the body. Contraction of skeletal muscle causes movement of the lymph fluid through valves. Lymph nodes are round or kidney-shaped, and range in size from very tiny to 1 in. in diameter. They are usually found in groups in different places throughout the body, including the neck, armpit, chest, abdomen, pelvis, and groin. Lymph nodes are garrisons of B, T and other immune cells. About two thirds of all lymph nodes and lymphatic tissue are within or near the gastrointestinal tract. The role of these nodes to filter the lymph before it can be returned to the circulatory system. Although these nodes can increase or decrease in size throughout life, any nodes that has been damaged or destroyed, does not regenerate.

The state of the lymphatic system can be detected by lymphography medical imaging techniques [11]. Magnetic resonance lymphography holds much promise for the noninvasive evaluation of lymph nodes. The technique utilizes ultrasmall superparamagnetic particles of iron oxide and has been shown to be highly sensitive and specific in the diagnosis of malignant lymph nodes [12]. The current state of lymph nodes with extracted data from lymphography technique can ascertain the classification of the investigated finding [13]. The enlargement of lymph nodes can be an index to trivial conditions and extends to more significant conditions that threats life [14]. Additionally the status of the lymph nodes could also suggest the occurrence of cancer [9]. Therefore, the main contribution of this paper is to investigate the effectiveness of RFC in conducting the lymph disease diagnostic problem. Aiming at improving the efficiency and effectiveness of the classification accuracy for lymph disease diagnosis, a CAD system based on RFC is introduced. The difference between this study and other studies that address the same topic is that a strong classifier system has been created by combining GA feature selection and random forest decision tree methods, which has very important implications for dimension reduction and sound classification to discriminate between normal and abnormal cases. Furthermore, this method yields more efficient results than any of the other methods tested in this paper.

The structure of the paper is the following: Introduction and related research are briefly described in Sections 1 and 2. Section 3 explains theoretical approach of feature selection methods and RFC. The evaluation procedure is described in Section 4. The dataset and the experimental results are presented in Section 5. Finally, Conclusion and future directions are summarized in Section 6.

2. Related work

Detection of Lymph disease is a prevalent research topic in the literature. Polat and Gunes [15] proposed a novel hybrid classification system based on C4.5 decision tree classifier and one-against-all approach to classify the multi-class problems including dermatology, image segmentation, and lymphography datasets taken from UCI (University of California Irvine) machine learning database. In this work, C4.5 decision tree was initially executed for all the classes of datasets and they reported 84.48%, 88.79%, and 80.11% classification accuracy for dermatology, image segmentation, and lymphography datasets using 10-fold cross validation, respectively. The proposed method based on C4.5 decision tree classifier and one-against-all approach obtained 96.71%, 95.18%, and 87.95% for the above datasets, respectively. Iannello et al. [16] proposed decomposition methods named as One-per-Class (OpC), Error-Correcting Output Codes (ECOC), PairWise Coupling (PWC) for multiclass classification including lymph disease dataset. The comparison has been carried out by employing three different paradigms for the basic classifiers like Multi-Layer Perceptron (MLP) as a neural network, a Nearest Neighbor (NN) as a statistical classifier, and a Support Vector Machine (SVM) as a kernel machine. The experimental results for MLP achieved 82.90%, 79.32% and 75.84% using OpC, ECOC and PWC, respectively. For NN classifier, the results achieved 79.22%, 76.81% and 76.99% using OpC, ECOC and PWC, respectively. Finally, the performance results obtained by SVM using OpC, ECOC and PWC were 87.85%, 81.36%, and 79.44%, respectively. Some of the recent classification results obtained by other studies for Lymph disease dataset are presented in Table 1.

3. Genetic algorithm (GA) and random forest classifier: preliminaries

This section provides a brief explanation of the basic framework of genetic algorithm and random forest classifier, along with some of the key definitions.

3.1. *Genetic algorithm (GA)*

GA is a stochastic search method for solving optimal solutions within large and complicated search spaces. It's a popular type of evolutionary algorithm (EA) that has been successfully used for feature selection. The technique is based on ideas from Darwin's theory of natural selection and "survival of the fittest" [27]. Genetic algorithm operates on a set of individuals called population, where each individual is an encoding of the problem's input data and are called chromosomes. Each chromosome is composed of genes, each of them has a binary value that indicates the presence or not of a specific element of the set. The search for the best solution is guided by an objective function called fitness function. The selected solutions of higher fitness function are more ability to produce new solutions than the less of fitness value while those of weak fitness function will be eliminated gradually. Fitness function controls the selection of best solution and provides a criteria

Table 1 – Related work for lymph disease diagnosis studies.

Author	Method	Accuracy (%)
De Falco [17]	Differential evolution (DE)	80.18% compared to 85.14% using Part classifier.
Gutiérrez et al. [18]	Two-stage evolutionary algorithm	85.05%
Karabulut et al. [19]	Feature selection methods with NaïveBayes, Multilayer Perceptron (MLP), and J48 decision tree classifiers	Best accuracy was 84.46% achieved using Chi-square FS and MLP
Abellán and Masegosa [20]	Bagging Credal decision trees (B-CDT): B-C4.5/B-CDT without pruning (0% noise level) B-C4.5/B-CDT with pruning (0% noise level)	79.96%/76.24% 79.69%/77.51%
Derrac et al. [21]	Evolutionary instance selection enhanced by Rough set based feature selection (EIS-RFS)	Best accuracy was 82.65% with 5 neighbors
McSherry [22]	Conversational case-based reasoning (CCBR)	86.5%
Rodríguez et al. [23]	ENDF: Ensembles (i.e., Randomization) of nested dichotomies using a forest method as base classifier. FND: A forest method using nested dichotomies of decision trees as base classifier	82.57% (MultiBoost-W) 83.51% (MultiBoost-S)
Madden [24]	Naïve Bayes Tree Augmented Naïve Bayes (TAN) General Bayesian network (GBN) with K2 search: GBN-K2: GBN with hill-climbing search: GBN-HC	82.16% 81.07% 77.46% 75.06%
Li et al. [25]	Combination of radial basis function neural network (RBFNN) and co-operative co-evolutionary algorithm (Co-CEA): CO-RBFNN	85.27% (testing)
Polat and Gunes [26]	Fuzzy-Artificial immune recognition system (AIRS)	AIRS: 83.138% Fuzzy-AIRS: 90.00%

to evaluate the candidate individuals. During the genetic evolution of solutions, the selected solution is that classifies the maximum number of training samples. The fitness function in its simple form is formulated as [28]

$$\text{Fitness} = \frac{\text{number of correctly classified instances}}{\text{number of training samples}}$$

For classification, the fitness function may include other factors such as maximization of prediction accuracy, minimization of error rate, etc. [29]. In genetic algorithms terminology, each iteration of the search is called a generation. From each generation the fittest individuals are selected and pooled out to form a base for a new population with better characteristics. Genetic algorithm is characterized by attributes such as objective function, encoding of the input data, crossover, mutation, and population size [28]. The point of GAs is to search a proper combination of multiple parameters to achieve the greatest level of satisfaction, either minimum or maximum, depending on the requirement of the problem [30].

3.2. *Random forest classifier (RFC)*

Random forests classifier (RFC) is one of the most successful ensemble learning techniques which have been proven to be very popular and powerful techniques in the pattern recognition and machine learning for high-dimensional classification and skewed problems [31]. A drawback associated with tree classifiers is their high variance. In practice it is not uncommon for a small change in the training data set to result in a very different tree. The reason for this lies in the hierarchical nature of the tree classifiers. An error that occurs in a node close to the root of the tree propagates all the way to the leaves. In order to make tree classification more stable, a decision forest methodology has been invented. The methodology was initially proposed by Ho [32], Amit and Geman [33] and later by Breiman [31], in an integrated form (as "random forest"). A decision forest is an ensemble of decision trees. It can be seen as one classifier which contains several classification methods or one method but various parameters of work. Consider a learning set $L = ((M_1, N_1), \ldots, (M_n, N_n))$ made of n vectors, $M \in X$ where X a set of numerical or symbolic observations and $N \in Y$ where Y is a set of class labels. For classification problems, a classifier is a mapping $X \rightarrow Y$. A new input vector is classified by each individual tree of the forest. Each tree yields a certain classification result. The principle of random forests is to build binary sub-trees using the training bootstrap samples coming from the learning sample L and selecting randomly at each node a subset of X. The decision forest chooses the classification which has the most votes over all the trees in the forest. The random forest methodology contains Breiman's "bagging" idea and Ho's "random selection features". Bagging, which stands for "bootstrap aggregation", is a type of ensemble learning introduced by Breiman [34], in order to improve the accuracy of a weak classifier by creating a set of classifiers. If the number of instances in a dataset is N, almost 2/3 of the original size is randomly selected through bootstrapping manner for N times. The remain instances have been used as an out-of-bag set to be evaluated. The set of out-of-bag are those observations that are not used to build the sub-trees. They have been used for evaluating the error prediction. At each node, a random feature selection is invoked for constructing a decision node. For m as a number of features, the size of feature selected considered at each split is typically equal to $\sqrt{m}$ or $\sqrt{m}/2$ [35]. The sub-trees are all maximal trees since no pruning process is performed.

Random forest training is accomplished for each decision tree. In this method, each classifier's training set is generated by randomly drawing N examples, with replacement, with N the size of the original training set. The learning system generates a classifier from the sample and aggregates all the classifiers generated from the different trial to form the final classifier. To classify an instance, every classifier records a vote for the class to which it belongs and the instance is labeled

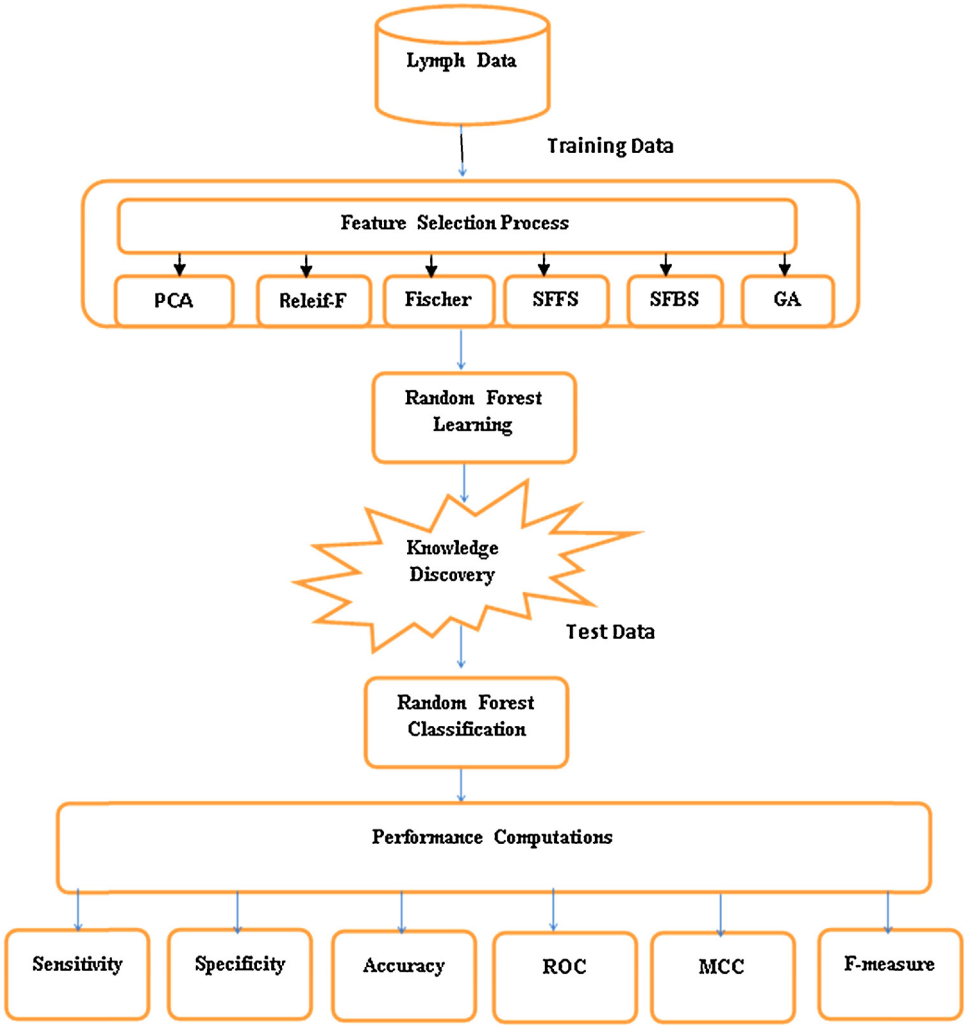


Fig. 1 – Random forest classifier combined with feature selection algorithms for lymph disease diagnosis.

as a member of the class with the most votes. In case that more than one class jointly receives the maximum number of votes, then the winner is selected at random. Every tree in the ensemble is grown on an independently drawn bootstrap replica of input data. Observations not included in this replica are "out-of-bag" for this tree [34]. The prediction error of the bagged ensemble is estimated by computing predictions for each tree on its out-of-bag observations; averaging these predictions over the entire ensemble for each observation and then comparing the predicted out-of-bag response with the true value at this observation. Bagging works by reducing variance of an unbiased base learner, such as a decision tree. This technique tends to improve the predictive power of the ensemble, as the random selection of features reduces the correlation between trees in the ensemble.

4. The proposed approach

The proposed system for diagnosis of lymph disease used in this study is given in Fig. 1. The main objectives of the proposed approach are (1) to improve the performance of classification accuracy and (2) obtain the important features as an indicator to the presence of lymph disease class. In this study, a genetic algorithm was applied to find an optimal feature set for the diagnosis of lymph diseases. The hybrid proposed system includes two main phases namely the feature selection phase and the classification phase. The feature selection phase consists of the GA technique which is used to reach an optimal solution within large and complicated dataset. Evolutionary computing is characterized by its simplicity, ability to reach good solutions with minimal knowledge and its efficiency in combinatorial and non-linear problems [36,37]. While for the classification phase the Random Forest technique is used in order to benefit of the prediction over the entire ensemble which improves the predictive power and reduces variance. A chromosome consisted of 15 genes, each representing an input variable. The encoding of a gene was binary, meaning that a particular variable was considered as an input variable (represented by '1') or not (represented by '0'). The assessment of the fitness of an input variable subset was based on the subset evaluator function with 10-fold cross-validation. The value of a subset of attributes was evaluated by considering the individual predictive ability of each feature along with the degree of redundancy between them. The initial population consisted of 60 chromosomes that evolved through

Table 2 – Lymphographic dataset description of attributes [36].

Attribute number	Attribute description	Possible values of attributes	Assigned values
1	Lymphatic	Normal, arched, deformed, displaced	1–4
2	Block of afferent	No, Yes	1–2
3	Block of lymph c (superior and inferior flaps)	No, Yes	1–2
4	Block of lymph s (lazy incision)	No, Yes	1–2
5	By pass	No, Yes	1–2
6	Extravasates (force out of lymph)	No, Yes	1–2
7	Regeneration	No, Yes	1–2
8	Early uptake	No, Yes	1–2
9	Lymph nodes diminish	0–3	0–3
10	Lymph nodes enlarge	1–4	1–4
11	Changes in lymph	Bean, oval, round	1–3
12	Defect in node	No, lacunar, lacunar marginal, lacunar central	1–4
13	Changes in node	No, lacunar, lacunar marginal, lacunar central	1–4
14	Changes in structure	no, grainy, drop-like, coarse, diluted, reticular, stripped, faint	1–8
15	Special forms	No, Chalices, vesicles	1–3
16	Dislocation	No, Yes	1–2
17	Exclusion of node	No, Yes	1–2
18	Number of nodes	0–80	1–8
19	Target Class	Normal, metastases, malign lymph, fibrosis	1–4

a maximum of 60 generations. The mutation and crossover probability was set at default values of 0.05 and 0.9, respectively.

5. Results and discussion

In this section, the performance of RFC and FS algorithms are presented. The simulations were performed by using an Intel (R) Core (TM) i3 CPU 530–2.93 GHz personal computer and a Microsoft Windows 7 64-bit operating system.

5.1. *Datasets*

The lymphography database was obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia [38]. There are 148 instances in total and there are no missing attributes. There are 18 numeric valued attributes and four classes, namely normal, metastases, malign lymph and fibrosis as shown in Table 2, described with the majority class prevalence and the entropy of classes. The latter two numbers measure the difficulty of the classification task; higher entropy and lower prevalence of majority class indicate the more difficult problems.

5.2. *Performance analysis*

The performance RFC was evaluated by using performance indices such as accuracy, sensitivity, specificity, precision, and *F*-Measure. Some of the main formulations are as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \times 100\% \quad (1)$$

$$\text{Senstivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\% \quad (2)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} \times 100\% \quad (3)$$

Precision or positive predictive value is the proportion of positive test results that are true positives (such as correct diagnoses). It is a critical measure of the performance of a diagnostic method, as it reflects the probability that a positive test reflects the underlying condition being tested for.

$$\text{PPR} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\% \quad (4)$$

In Eqs. (1)–(4), TP is the number of true positives; FN, the number of false negatives; TN, the number of true negatives; and FP, the number of false positives. They are defined as a confusion matrix. Considering imbalanced positive and negative samples in the datasets, another appropriate quantity for evaluating the classification accuracy of imbalanced positive and negative samples is the Matthews Correlation Coefficient MCC, which is given as follows [39]:

$$\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FN} \cdot \text{FP}}{\sqrt{(\text{TP} + \text{FN})(\text{TP} + \text{FP})(\text{TN} + \text{FN})(\text{TN} + \text{FP})}} \quad (5)$$

Obviously, the scope of the MCC is within the range of [−1,1]. The larger the MCC value, the better the classifier performance.

Receiver operating characteristic, ROC, curves are also used to evaluate the performance of a diagnostic test [40]. This method consists of a lot of information for comprehensibility and improving classifiers performance. The ROC curve plots

the true positive rate as a function of the false positive rate. It is parameterized by the probability threshold values. The true positive rate represents the fraction of positive cases that are correctly classified by the model. The false positive rate represents the fraction of negative cases that are incorrectly classified as positive. Therefore, it provides a trade-off between sensitivity and specificity. The advantages of ROC analysis are the robust description of the network's predictive ability and an easy way to change the existence network based on differential cost of misclassification and varying prior probabilities of class occurrences. However, it requires visual inspection because the best classifiers are hard to recognize when the curves are mixed.

5.3. *Feature selection and classification Accuracy*

Feature selection plays an important role in building classification systems. It cannot only reduce the dimension of data, but also lower the computation costs and gain a good classification performance. Six feature selection algorithms are employed to select features before passing the data sets to the RFC. They are; PCA [41], Relief-F [42], Fisher [43], Sequential Forward Floating Search (SFFS) and the Sequential Backward Floating Search (SBFS) [44,45] and GA. Threshold value is selected as 0.1 to select the features. The optimal FS of these algorithms on are summarized in Table 3. As noted from Table 3, the dimensionality of lymph disease data set is reduced and hence less storage space is required for the execution of the algorithms.

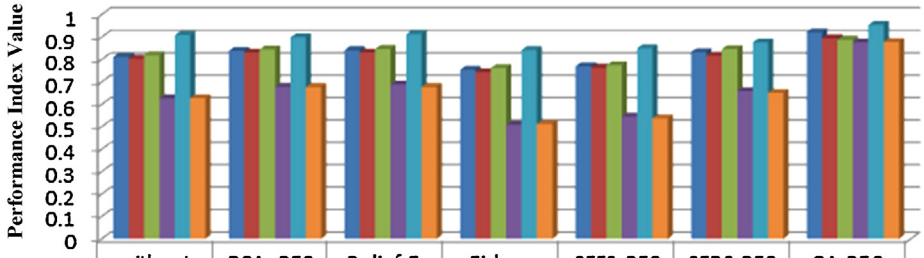
In order to get reliable estimates for classification accuracy on each classification task, every experiment has been performed using 10-fold cross-validation. Any result shown is always the average of the 10-folds. In this work, RFC has been run for all the classes of dataset used. Random forest models provided greater predictive accuracy than single-tree models, but they have the disadvantage that cannot be visualized; decision tree forest models are more of a "black box".

Table 3 – Selected features of lymph disease data set.

FS method	No of selected attributes	Selected attributes
PCA	15	{1–15}
Relief-F	7	{1, 2, 9, 14, 15, 17, 18}
Fisher	14	{2, 4, 5, 7, 8, 9, 10, 11, 13, 14, 15, 16, 17, 18}
SFFS	4	{2, 8, 10, 13}
SFBS	3	{2, 13, 15}
GA	6	{13, 2, 15, 14, 10, 11}

Generally, the larger a decision tree forest is, the more accurate the prediction. There are two types of size controls available (1) the number of trees in the forest and (2) the size of each individual tree. Specify the maximum number of levels (depth) that each tree in the forest may be grown to. Some research indicates that it is best to grow very large trees, so the maximum levels should be set large and the minimum node size control would limit the size of the trees. Therefore, maximum tree levels were adjusted at 50 trees. Surrogate splitters were used to handle missing values to compute the association between the primary splitter selected for a node and all other predictors including predictors not considered as candidates for the split. If the value of the primary predictor variable is missing for a row, the software will use the best surrogate splitter whose value is known for the row.

As shown in Table 4, RFC achieved 83.9%, 84.2%, and 75.5%, 77.1% and 83.4% classification accuracies for PCA, Relief-F, Fisher, SFFS and SFBS, respectively. The proposed method based on RFC and GA approach obtained 92.2%, 89.5%, and 88.9% for accuracy, sensitivity and specificity, respectively. Hence, these results verify efficiency of GA-RFC strategy. Value of AUC and MCC for GA-RFC achieved 0.954 and 0.877, respectively. Fig. 2 demonstrates the performance indices comparisons of RFC depending on feature selection strategies visually.



	without FS	PCA- RFC	Relief-F-RFC	Fisher-RFC	SFFS-RFC	SFBS-RFC	GA-RFC
Accuracy	0.812	0.839	0.842	0.755	0.771	0.834	0.922
Sensitivity	0.804	0.831	0.831	0.743	0.764	0.818	0.895
Specificity	0.819	0.845	0.848	0.763	0.775	0.847	0.889
MCC	0.627	0.678	0.688	0.508	0.546	0.66	0.877
ROC	0.909	0.899	0.913	0.843	0.851	0.877	0.954
KS	0.6277	0.6777	0.6779	0.5129	0.5381	0.6517	0.879

RFC combined with different FS methods

Fig. 2 – Performance indices comparisons for random forest classifiers combined with the feature selected algorithms.

Table 4 – Performance measures of RF classifier corresponding to the selected features.

Performance index	Without FS	PCA-RF	Relief-F-RF	Fisher-RF	SFFS-RF	SFBS-RF	GA-RF
Accuracy	0.812	0.839	0.842	0.755	0.771	0.834	0.922
Sensitivity	0.804	0.831	0.831	0.743	0.764	0.818	0.895
Specificity	0.819	0.845	0.848	0.763	0.775	0.847	0.889
Precision	0.807	0.827	0.831	0.743	0.742	0.801	0.874
MCC	0.627	0.678	0.688	0.508	0.546	0.660	0.877
F-Measure	0.803	0.828	0.829	0.743	0.752	0.808	0.879
AUC	0.909	0.899	0.913	0.843	0.851	0.877	0.954
KS	0.6277	0.6777	0.6779	0.5129	0.5381	0.6517	0.879

The performance of classifier, and hence the quality of selected features, will be evaluated also by means of Cohen's kappa coefficient or kappa statistic (KS) to measure the agreement between predicted and observed values of a dataset, while correcting the agreement that occurs by chance [46]. In classification performance comparisons, using the percentage of missing values as the single meter for accuracy can give misleading results. The cost of error must also be taken into account, while making such assessments. Kappa statistic, in this case, is a good index to investigate classifications that may be due to chance. Generally, KS takes values between (−1,1) and when its value calculated for classifiers approaches to '1', then the performance of the classifier is assumed to be more realistic. Therefore, in the performance analysis of classifiers, Kappa error is a recommended metric to consider for evaluation purposes and it is calculated by [47]:

$$KS = \frac{P_o - P_c}{1 - P_c} \tag{6}$$

where P_o is total agreement probability, and P_c is the hypothetical probability of chance agreement.

Table 4 gives also valuable information about the confidence of the RFC performances combined with different FS algorithm. If Kappa Statistic value for a classifier approaches to one, its performance is more valuable and less prone to be by chance. With this consideration, the classification reliability of GA-RFC is seen to be superior (KS = 0.879) and hence outperformed other methods.

As can be seen from above results, the proposed method based on GA-RFC has produced very promising results on the classification of multi-class dataset in classifying the possible lymph diseases patients. The proposed method was arrived to the highest classification accuracy among classifiers in Table 1. While Table 3 is examined for the selected features, it is seen that some features, i.e. Changes in node, Block of afferent, Special forms, Changes in structure, Lymph nodes enlarge and Changes in lymph are the optimum features selected with GA and achieved the highest accuracy. The results demonstrated that these features are enough to represent the dataset's class information. Random forest retains many benefits of decision trees while achieving better results through the usage of bagging on samples, random subsets of variables, and a majority voting scheme. It handles missing values, a variety of variables (continuous, binary, categorical), and is well suited to high-dimensional data modeling. Unlike classical decision trees, there is no need to prune trees in RFC since the ensemble and bootstrapping schemes help random forest overcome overfitting issues. Thus, it is believed that the RFC optimized system can be very helpful to the physicians for their final decision on their patients. This method can be used in many pattern recognition applications.

6. Conclusion

Classifying of multi-class classification problems such as lymphography database is important issue in pattern recognition applications. The idea of medical data mining is to extract hidden knowledge in medical field using data mining techniques. One of the positive aspects is to discover the important features. In this work, a hybrid method for diagnosis of lymph diseases based on GA and RFC is presented. So, GA is used for reducing the dimension of lymph diseases dataset and RFC is used for intelligent classification. The main aim of this system is to apply the unique features of RFC including better generalization performance, fast learning speed, simpler and without tedious and time consuming parameter tuning to perform lymph diseases diagnosis. Random forest is ensemble method that combines the predictions of many individual tree models (the base classifiers) to provide a prediction that tends to be more accurate than any of the individual classifiers predictions. The proposed GA-RFC system performance is compared with other feature selection algorithms combined with RFC such as PCA, Relief-F, Fisher, SFFS and SBFS. The performance of the proposed structure is evaluated in terms of sensitivity, specificity, accuracy and AUC. The results showed that RFC achieved 83.9%, 84.2%, and 75.5%, 77.1% and 83.4% classification accuracies for PCA, Relief-F, Fisher, SFFS and SFBS, respectively. The proposed method based on RFC and GA approach obtained 92.2%, 89.5%, and 88.9% for accuracy, sensitivity and specificity, respectively. Hence, these results verify efficiency of GA-RFC strategy. Value of AUC and MCC for GA-RFC achieved 0.954 and 0.877, respectively. The dimension of input feature space is reduced from 18 to 6 by using GA. Experimental results demonstrated also that the proposed system performed significantly well in lymph disease diagnosis. This research demonstrated that the GA can be used for reducing the dimension of feature vector and proposed GA-RFC model can be used to obtain efficient automatic diagnostic systems for other diseases. The future investigation will pay much attention to evaluate the proposed system in other medical diagnosis problems. Also, instead of Random forest, other classification algorithms can be used and tested with other optimization techniques.

REFERENCES

[1] K.J. Ciosa, G.W. Mooree, Uniqueness of medical data mining, Artif. Intell. Med. 26 (2002) 1–24.
[2] W. Ceusters, Medical natural language understanding as a supporting technology for data mining in healthcare, in: K.J. Cios (Ed.), Medical Data Mining and Knowledge Discovery, Springer, Heidelberg, 2000, pp. 32–60 (Chapter 3).
[3] I. Kononenko, Machine learning for medical diagnosis: history state of the art and perspective, Artif. Intell. Med. 23 (2001) 89–109.
[4] F. Calle-Alonso, C.J. Pérez, J.P. Arias-Nicolás, J. Martín, Computer-aided diagnosis system: a Bayesian hybrid classification method, Comput. Methods Programs Biomed. 112 (2013) 104–113.
[5] S.H. Huang, L.R. Wulsin, H. Li, J. Guo, Dimensionality reduction for knowledge discovery in medical claims database: application to antidepressant medication utilization study, Comput. Methods Programs Biomed. 93 (2009) 115–123.
[6] Z. Cselényi, Mapping the dimensionality density and topology of data: the growing adaptive neural gas, Comput. Methods Programs Biomed. 78 (2005) 141–156.
[7] I.A. Gheyas, L.S. Smith, Feature subset selection in large dimensionality domains, Pattern Recognit. 43 (2010) 5–13.
[8] H.H. Inbarani, A.T. Azar, G. Jothi, Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis, Comput. Methods Programs Biomed. (2013), Available online 16 October 2013, ISSN 0169-2607, http://dx.doi.org/10.1016/j.cmpb.2013.10.007
[9] M. Macaš, L. Lhotská, E. Bakstein, D. Novák, J. Wild, T. Sieger, P. Vostatek, R. Jech, Wrapper feature selection for small sample size data driven by complete error estimates, Comput. Methods Programs Biomed. 108 (2012) 138–150.
[10] Cancer Research UK. http://www.cancerresearchuk.org (accessed 03.11.13).
[11] A. Luciani, E. Itti, A. Rahmouni, Lymph node imaging: basic principles, Eur. J. Radiol. 58 (2006) 338–344.
[12] N. Jahan, P. Narayanan, A. Rockall, Magnetic resonance lymphography in gynaecological malignancies, Cancer Imaging 10 (2010) 85–96.
[13] R. Sharma, J.A. Wendt, J.C. Rasmussen, A.E. Adams, M.V. Marshall, E.M. Sevick-Muraca., New horizons for imaging lymphatic function, Ann. N. Y. Acad. Sci. 1131 (2008) 13–36.
[14] A. Guermazi, P. Brice, C. Hennequin, E. Sarfati, Lymphography: an old technique retains its usefulness, Radiographics 23 (2003) 1541–1558.
[15] K. Polat, S. Gunes, A novel hybrid intelligent method based on C4.5 decision tree classifier and one-against-all approach for multi-class classification problems, Expert Syst. Appl.: Int. J. 36 (2009) 1587–1592.
[16] G. Iannello, G. Percannella, C. Sansone, P. Soda, On the use of classification reliability for improving performance of the one-per-class decomposition method, Data Knowl. Eng. 68 (2009) 1398–1410.
[17] I. De Falco, Differential evolution for automatic rule extraction from medical databases, Appl. Soft Comput. 13 (2013) 1265–1283.
[18] P.A. Gutiérrez, C. Hervás-Martínez, F.J. Martínez-Estudillo, M.A. Carbonero, A two-stage evolutionary algorithm based on sensitivity and accuracy for multi-class problems, Inform. Sci. 197 (2012) 20–37.
[19] E.M. Karabulut, S.A. Özel, T. Ibrikçi., A comparative study on the effect of feature selection on classification accuracy, Proc. Technol. 1 (2012) 323–327.
[20] J. Abellán, A.R. Masegosa, Bagging schemes on the presence of class noise in classification, Expert Syst. Appl. 39 (2012) 6827–6837.
[21] J. Derrac, C. Cornelis, S. García, F. Herrera, Enhancing evolutionary instance selection algorithms by means of fuzzy rough set based feature selection, Inform. Sci. 186 (2012) 73–92.
[22] D. McSherry., Conversational case-based reasoning in medical decision making, Artif. Intell. Med. 52 (2011) 59–66.
[23] J.J. Rodríguez, C.G. Osorio, J. Maudes, Forests of nested dichotomies, Pattern Recognit. Lett. 31 (2010) 125–132.
[24] M.G. Madden, On the classification performance of TAN and general Bayesian networks, Knowl. Based Syst. 22 (2009) 489–495.
[25] M. Li, J. Tian, F. Chen, Improving multiclass pattern recognition with a co-evolutionary RBFNN, Pattern Recognit. Lett. 29 (2008) 392–406.
[26] K. Polat, S. Gunes, Automatic determination of diseases related to lymph system from lymphography data using principles component analysis (PCA), fuzzy weighting pre-processing and ANFIS, Expert Syst. Appl. 33 (2007) 636–641.
[27] D.E. Goldberg, Genetic Algorithm in Search, Optimisation, and Machine Learning, Addison-Wesley, Boston, 1989.
[28] K.M. Faraoun, A. Boukelif, A Genetic programming approach for multi-category pattern classification applied to network intrusion detection, Int. J. Comput. Intell. Appl. 6 (2006) 77–99.
[29] B.K. Sarkara, S.S. Sanab, K. Chaudhuric, A genetic algorithm-based rule extraction system, Appl. Soft Comput. 12 (2012) 238–254.
[30] A.A. Freitas, Evolutionary algorithms for data mining, in: O. Maimon, L. Rokach (Eds.), The Data Mining and Knowledge Discovery Handbook, Springer-Verlag, Berlin Heidelberg, 2005, pp. 435–467.
[31] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[32] T. Ho, Random decision forest, in: 3rd International Conf. on Document Analysis and Recognition, August 14–18, Montreal, Canada, 1995, pp. 278–282.
[33] Y. Amit, D. Geman, Shape quantization and recognition with randomized trees, Neural Comput. 9 (1997) 1545–1588.
[34] L. Breiman, Bagging predictors, Technical Report 421, Department of Statistics, University of California, Berkeley, USA, 1994.
[35] O. Okun, H. Priisalu, Random Forest for Gene Expression Based Cancer Classification: Overlooked Issues, Springer-Verlag, Berlin, Heidelberg, 2007, pp. 483–490.
[36] K. Cios, W. Pedrycz, R.W. Swiniarki, Data Mining Methods for Knowledge Discovery, Kluwer Academic Publishers, Norwell, MA, USA., 1998.
[37] T. Chen, T. Hsu, A GAs based approach for mining breast cancer pattern, Expert Syst. Appl. 30 (2006) 674–681.
[38] UCI. Machine Learning Repository. http://archive.ics.uci.edu/ml/index.html (accessed 03.11.13).
[39] B.W. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochim. Biophys. Acta 405 (1975) 442–451.
[40] J. Kerekes, Receiver operating characteristic curve confidence intervals and regions, IEEE Geosci. Remote Sens. Lett. 5 (2008) 251–255.
[41] T.N. Yang, S.D. Wang, Robust algorithms for principal component analysis, Pattern Recognit. Lett. 20 (1999) 927–933.
[42] K. Kira, L.A. Rendell, A practical approach to feature selection, in: Proceedings of the Ninth International Workshop on Machine Learning, Aberdeen, Scotland, United Kingdom, 1992, pp. 249–256.
[43] M. Robnik-Sikonja, I. Kononenko, Theoretical and empirical analysis of Relief and ReliefF, Mach. Learn. 53 (2003) 23–69.

[44] P. Pudil, J. Novovic¡ová, J. Kittler, Floating search methods in feature selection, Pattern Recognit. Lett. 15 (1994) 1119–1125.
[45] P. Somol, P. Pudil, J. Novovic¡ová, P. Paclík, Adaptive floating search methods in feature selection, Pattern Recognit. Lett. 20 (1999) 1157–1163.
[46] I.H. Witten, H. Ian, Data mining: practical machine learning tools and techniques, Morgan Kaufmann Ser. Data Manage. Syst. (2005) 153–168.
[47] A. David, Comparison of classification accuracy using Cohen's Weighted Kappa, Expert Syst. Appl. 34 (2008) 825–832.