Journal of Intelligent Learning Systems and Applications, 2016, 8, 51-62
Published Online August 2016 in SciRes. http://www.scirp.org/journal/jilsa
http://dx.doi.org/10.4236/jilsa.2016.83005

Lymph Diseases Prediction Using Random Forest and Particle Swarm Optimization

Waheeda Almayyan

Department of Computer Science and Information Systems, College of Business Studies,
Public Authority for Applied Education and Training, Kuwait City, Kuwait
Email: wi.almayyan@paaet.edu.kw

Received 3 May 2016; accepted 30 July 2016; published 3 August 2016

Abstract

This research aims to develop a model to enhance lymphatic diseases diagnosis by the use of random forest ensemble machine-learning method trained with a simple sampling scheme. This study has been carried out in two major phases: feature selection and classification. In the first stage, a number of discriminative features out of 18 were selected using PSO and several feature selection techniques to reduce the features dimension. In the second stage, we applied the random forest ensemble classification scheme to diagnose lymphatic diseases. While making experiments with the selected features, we used original and resampled distributions of the dataset to train random forest classifier. Experimental results demonstrate that the proposed method achieves a remarkable improvement in classification accuracy rate.

Keywords

Classification, Random Forest Ensemble, PSO, Simple Random Sampling, Information Gain Ratio, Symmetrical Uncertainty

1. Introduction

Nowadays, Computer-Aided Diagnosis (CAD) applications have become one of the key research topics in medical biometrics diagnostic tasks. Medical diagnosis depends upon the experience of the physician beside the existing data. Consequently, a number of articles suggested several strategies to process the physician's analysis and judgment tasks about actual clinical assessments [1]. With reasonable success, machine-learning techniques have been applied in constructing the CAD applications due to its strong capability of extracting complex relationships in the medical data [2].

Raw medical data requires some effective classification techniques to support the computer-based analysis of

How to cite this paper: Almayyan, W. (2016) Lymph Diseases Prediction Using Random Forest and Particle Swarm Optimization. *Journal of Intelligent Learning Systems and Applications*, **8**, 51-62. http://dx.doi.org/10.4236/jilsa.2016.83005

such voluminous and heterogeneous data. Accuracy of clinically diagnosed cases is particularly important issue to be considered during classification. In most cases the size of medical datasets is usually great, which directly affects the complexity of the data mining procedure [3]. So, the large-scale medical data is considered a source of significant challenges in data mining applications, which involves extracting the most descriptive or discriminative features. Thus, feature reduction has a significant role in eliminating irrelevant features from medical datasets [4] [5]. Dimensionality reduction procedure aims to reduce computational complexity with the possible advantages of enhancing the overall classification performance. It includes eliminating insignificant features before model implementation, which makes screening tests faster, more practical and less costly and this is an important requirement in medical applications [6].

The lymphatic system is a vital part of the immune system in removing the interstitial fluid from tissues. It absorbs and transports fats and fat-soluble vitamins from the digestive system and delivers these nutrients to the cells of the body. It transports white blood cells to and from the lymph nodes into the bones. Moreover, it transports antigen-presenting cells to the lymph nodes where an immune response is stimulated.

Different medical imaging techniques have been used for the investigation of the lymphatic channels and lymph glands status [7]. The current state of lymph nodes with obtained data from lymphography technique can ascertain the classification of the investigated diagnosis [8]. The enlargement of lymph nodes can be an index to several conditions and extends to more significant conditions that threat life [9]. The study of the lymph nodes is important in diagnosis, prognosis, and treatment of cancer [10]. Therefore, the main contribution of this paper is to investigate the effectiveness of the suggested technique in diagnosing the lymph disease problem.

In this article, a CAD system based on random forest ensemble classifier is introduced to improve the efficiency of the classification accuracy for lymph disease diagnosis. The difference between this article and other articles that address the same topic is that a strong ensemble classifier scheme has been created by combining PSO feature selection and random forest decision tree methods, which yields more efficient results than any of the other methods tested in this paper.

Several approaches have been investigated using conventional and artificial intelligence techniques in order to evaluate the lymphography dataset. Karabulut *et al.* studied the effect of feature selection methods with Naïve-Bayes, Multilayer Perceptron (MLP), and J48 decision tree classifiers with fifteen real datasets including lymph disease dataset [11]. The best accuracy was 84.46% achieved using Chi-square FS and MLP. Derrac *et al.* proposed an evolutionary algorithm for data reduction enhanced by Rough set based feature selection. The best accuracy recorded was 82.65% with 5 neighbors [12]. Madden [13] proposed a comparative study between Naïve Bayes, Tree Augmented Naïve Bayes (TAN) and General Bayesian network (GBN) classifier, with K2 search and GBN with hill-climbing search in which they scored an accuracy of 82.16%, 81.07%, 77.46% and 75.06% respectively. De Falco [14] proposed a differential evolution technique to classify eight databases from the medical domain. The suggested technique scored an accuracy of 85.14% compared to 80.18% using Part classifier. Abellán and Masegosa designed Bagging credal decision trees using imprecise probabilities and uncertainty measures. The proposed decision tree model without pruning scored an accuracy of 79.69% and 77.51% with pruning [15].

In this article, a two-stage algorithm is investigated to enhance classification of lymph disease diagnosis. In the first stage, a number of discriminative features out of 18 were selected using PSO and several feature selection to reduce the dimension. In the second stage, we used a random forest ensemble classification scheme to diagnose lymphography types. While making experiments with the selected features, we used original and resampled distributions of the dataset to train random forest algorithm. We noticed a promising improvement in classification performance of the algorithm with resampling strategy.

The article commences with the suggested feature selection techniques and the random forest ensemble classifier. Section 4 briefly introduces simple random sampling strategy. Section 5 focuses on the applied performance measures. Section 6 describes the experiment steps and the involved dataset and shows the result of the experiments. The article concludes with conclusion and further research.

2. Feature Selection

The main objectives of the proposed approach are to improve the performance of classification accuracy and obtain the most important features. Essentially, the feature space is searched to reduce the feature space and prepare the conditions for the classification step. This task is carried out using different state-of-the-art dimension reduction techniques, namely Particle Swarm Optimization, Information Gain Ratio attribute evaluation and Symmetric Uncertainty correlation-based measure.

2.1. Particle Swarm Optimization for Feature Selection

The particle swarm optimization (PSO) technique is a population-based stochastic optimization technique first introduced in 1995 by Kennedy and Eberhart [16]. In PSO, a possible candidate solution is encoded as a finite-length string called a particle p_i in the search space. All of the particles make use of its own memory and knowledge gained by the swarm as a whole to find the best solution. With the purpose of discovering the optimal solution, each particle adjusts its searching direction according to two features, its own best previous experience (p_{best}) and the best experience of its companions flying experience (g_{best}). Each particle is moving around the n-dimensional search space *S* with objective function $f : S \subseteq \Re^n \rightarrow \Re$. Each particle has a position $x_{i,t}$ (*t* represents the iteration counter), a fitness function $f\left(x_{i,t}\right)$ and “flies” through the problem space with a velocity $v_{i,t}$. A new position $z_1 \in S$ is called better than $z_2 \in S$ iff $f\left(z_1\right) < f\left(z_2\right)$ [17].

Particles evolve simultaneously based on knowledge shared with neighbouring particles; they make use of their own memory and knowledge gained by the swarm as a whole to find the best solution. The best search space position particle *i* has visited until iteration *t* is its previous experience p_{best}. To each particle, a subset of all particles is assigned as its neighbourhood. The best previous experience of all neighbours of particle *i* is called g_{best}. Each particle additionally keeps a fraction of its old velocity. The particle updates its velocity and position with the following equation in continuous PSO [17]:

$$v_{pd}^{new} = \omega * v_{pd}^{old} + C_1 * rand_1\left(\right) * \left(pbest_{pd} - x_{pd}^{old}\right) + C_2 * rand_2\left(\right) * \left(gbest_{d_d} - x_{pd}^{old}\right) \tag{1}$$

$$x_{pd}^{new} = x_{pd}^{old} + v_{pd}^{new} \tag{2}$$

The first part in Equation (1) represents the previous flying velocity of the particle. While the second part represents the “*cognition*” part, which is the private thinking of the particle itself, where C_1 is the individual factor. The third part of the equation is the “*social*” part, which represents the collaboration amongst the particles, where C_2 is the societal factor. The acceleration coefficients (C_1) and (C_2) are constants represent the weighting of the stochastic acceleration terms that pull each particle toward the p_{best} and g_{best} positions. Particles’ velocities are restricted to a maximum velocity, $V_{\max}$. If $V_{\max}$ is too small, particles in this case could become trapped in local optima. In contrast, if $V_{\max}$ is too high particles might fly past fine solutions. According to Equation (1), the particle’s new velocity is calculated according to its previous velocity and the distances of its current position from its own best experience and the group’s best experience. Afterwards, the particle flies toward a new position according to Equation (2). The performance of each particle is measured according to a pre-defined fitness function (**Figure 1**).

2.2. Information Gain Ratio Attribute Evaluation

Information Gain Ratio attribute evaluation (IGR) measure was generally developed by Quinlan (Quinlan, 1993) within the C4.5 algorithm and based on the Shannon entropy to select the test attribute at each node of the decision tree [18]. It represents how precisely the attributes predict the classes of the test dataset in order to use the “best” attribute as the root of the decision tree.

The expected IGR needed to classify a given sample *s* from a set of data samples *C* $IRG(s,C)$ is calculated as follow

$$IGR(s,C) = \frac{gain(s,C)}{split\_info(C)} \tag{3}$$

$$gain(s,C) = entropy(s,C) - entropy_p(s,C),$$

$$entropy(s,C) = -p(s|C)\log_2 p(s|C) - \left(1 - p(s|C)\right)\log_2\left(1 - p(s|C)\right)),$$

$$p(s,C) = freq(s|C)/|C|,$$

```
Begin
  Initialize Population with random positions and velocities
  WHILE (maximum iteration or minimum error criteria is not met)
        For each particle
        Calculate the fitness value;
        If the fitness value is better than the previous best fitness value (pbest) then
              update the current value as the new pbest;
              For each particle
        Choose the particle with the best fitness value of all particles as the gbest;
      Calculate the fitness value;
    Update particle’ s velocity and position according to equation (1) and (2);
  NEXT generation until stopping criterion;
End.
```

Figure 1. Pseudocode of PSO-based feature selection approach.

$$entropy_p(s,C)=\sum_i \frac{|C_i|}{|C|} entropy_p(s,C_i),$$

$$split\_info(C)=-\sum_i \frac{|C_i|}{|C|}\log\frac{|C_i|}{|C|},$$

where $freq(s,C)$, C_i and $|C_i|$ are the frequency of the sample s in C, the ith class of C and the number of samples in C_i, respectively.

2.3. Symmetrical Uncertainty

Symmetric uncertainty correlation-based measure (SU) can be used to evaluate the goodness of features by calculating between feature and the target class (Fayyad & Irani, 1993; Liu *et al.*, 2002) [19] [20]. The features having greater SU value get higher importance. SU is defined as

$$SU(X,Y)=\frac{2IG(X\,|\,Y)}{H(X)+H(Y)} \tag{4}$$

$$IG(X\,|\,Y)=\frac{H(X)}{H(X\,|\,Y)}$$

where $H(X)$, $H(Y)$, $H(X|Y)$, IG are the entropy of a of X, entropy of a of Y and the entropy of a of posterior probability X given Y and information gain, respectively.

3. Random Forest Ensemble Classification Algorithm

Ensemble learning methods which utilizes ensembles of classifiers such as neural networks ensembles, random forest, bagging and boosting have received an increasing interest because of their ability to deliver an accurate prediction and robust to noise and outliners than single classifiers [21] [22]. The basic idea behind ensembled classifiers is based upon the premise that a group of classifiers can perform better than an individual classifier. In 2001, Breiman proposed a new and promising tree-based ensemble classifier based on a combination tree of predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees and called it random forest. Random forest classifier consists of a combination of individual base classifiers where each tree is generated using a random vector sampled independently from the classification input vector to enable a much faster construction of trees. For classification, the classification single vote from all trees is combined using a rule based approach (such as, majority voting, product, sum, or Bayesian rule), or based on an iterative error minimization technique by reducing the weights for the correctly classified samples.

In Random Forest, the method to build an ensemble of classifiers can be summarized as follows:

- The Random Forest training algorithm starts with constructing multiple trees. In the literature, several methods were used such as random trees, CART, J48, C4.5, etc. In this article we are using the random trees in building the Random Forest classifier with no pruning, which makes it light, from a computational perspective.
- The next step is preparing the training set for each tree, which is formed by randomly sampling the training dataset using bootstrapping technique with replacement. This step is called the bagging step [23] and the selected samples are called the in-bag samples; the rest are set aside as out-of-bag samples. For each new training set that is generated, approximately one third of the data in the in-bag set is duplicated (sampling with replacement) and used for building the tree. Whereas, the remaining training samples, out-of-bag, are used to test the tree classification performance. **Figure 2** illustrates the data sampling procedure. Each tree is constructed using a different bootstrap sample.
- Random Forest increases the diversity of the trees by choosing and using a random number of features (in this work four features) to construct the nodes and leafs of a random tree classifier. According to Breiman [21] [23], this step minimizes the correlation among the features, decreases the sensitivity to noise in the data and, at the same time, increases the accuracy of classification.
- Building a random tree begins at the top of the tree with in-bag dataset. The first step involves selecting a feature at the root node and then splitting the training data into subsets for every possible value of the feature. This makes a branch for each possible value of the attribute. Tree design requires choosing a suitable attribute selection measure for splitting and the selection of the root node to maximize dissimilarity between classes. The information gain (*IG*) of splitting the training dataset (*Y*) into subsets (*Yi*) can be defined as:

$$IG = -\sum_i \frac{|Yi|}{Y} E(Yi) \tag{5}$$

- If the information gain is positive; the node is split else the node will become a leaf node that would provide a decision of the most common target class in the training subset.
- The partitioning procedure is repeated recursively at each branch node using the subset that reaches the branch and the remaining attributes continues until all attributes are selected. The highest information gain of the remaining attributes is selected as the next attribute. Eventually the most occurring target class in the training subset that reached that node is assigned as the classification decision.
- The procedure is repeated to build all trees.
- After building all trees, the out-of-bag dataset is used to test trees as well as the entire forest. The obtained average misclassification error can be used to adjust the weights of the vote of each tree. In this article, the implementation of the random forest classifier gives each tree the same weight.

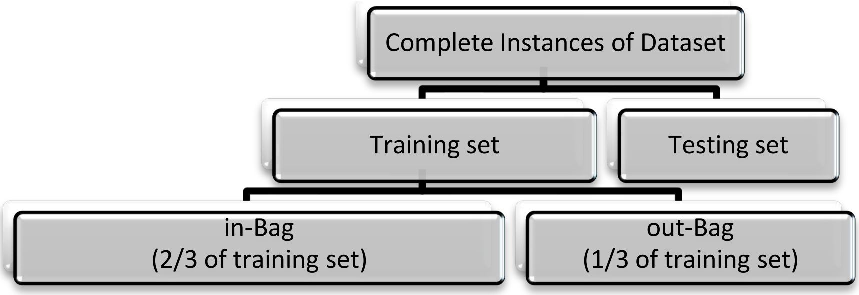

Figure 2. Data partition in constructing random forest trees.

4. Simple Random Sampling

Medical data usually experience class imbalance problems, due to the fact that one class is represented by a considerably larger number of instances than other classes. Subsequently, classification algorithms tend to ignore the minority classes. Simple random sampling has been advised as a good means of increasing the sensitivity of the classifier to the minority class by scaling the class distribution. An empirical study where the authors used twenty datasets from UCI repository has showed quantitatively that classifier accuracy might be increased with a progressive sampling algorithm [24]. Weiss and Provost deployed decision trees to evaluate classification performances with the use of a sampling strategy. Another important study used sampling to scale the class distribution and mainly focus on biomedical datasets [25]. The authors measure the effect of the suggested sampling strategy by the use of nearest neighbor and decision tree classifiers. In Simple random sampling, a sample is randomly selected from the population so that the obtained sample is representative of the population. Therefore, this technique provides an unbiased sample from the original data.

Regarding simple random sampling there are two approaches while making random selection, in the first approach the samples are selected with replacement where the sample can be selected more than once repeatedly with an equal selection chance. In the other approach the selection of samples is done without replacement where the sample can be selected only once, so that each sample in the data set has an equal chance of being selected and once selected it cannot be chosen again [26].

5. Performance Measures

When the data is inadequate, predicting classification performance of a machine learning method is difficult. Thus, Cross-validation is preferred when the scholar have a small amount of data [27]. When machine-learning methods explore data, decisions must be made on how to split dataset for training and testing. With the intention of estimating the performance of machine learning methods, the lymphography dataset is split into training and testing subsets, afterwards a 10-fold cross-validation, which is a commonly used technique for evaluation, is applied.

The performance of the suggested technique was evaluated by using four commonly used performance metrics, Precision, ROC, MCC and Cohen's kappa coefficient. The main formulations are defined in Equations (4)-(6), according to the confusion matrix. In the confusion matrix of a two-class problem, TP is the number of true positives that was classified correctly. FN is the number of false negatives that was classified incorrectly. TN is the number of true negatives that was classified as negatives. FP is the number of false positives that was classified as negatives. Accordingly, we can define Precision as:

$$\text{Precision} = \frac{\text{TN}}{\text{FP} + \text{TN}} \times 100\% \tag{6}$$

Receiver Operator Characteristic (ROC) curve is another commonly used measure to evaluate two-class decision problems in Machine Learning. The ROC curve is a standard tool for summarizing classifier performance over a range of tradeoffs between TP and FP error rates [28]. ROC usually takes values between 0.5 for random drawing and 1.0 for perfect classifier performance.

Considering class-imbalanced datasets, such as the case in this database, the Matthews Correlation Coefficient (MCC) is an appropriate measure that considered balanced. It can be used even if the classes are of very different in sizes, as it is a correlation coefficient between the observed and predicted classification decisions. The MCC measure falls within the range of [−1, 1]. The larger the MCC coefficient indicates better classifier prediction. The MCC measure can be calculated directly from the confusion matrix using the following formula:

$$\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FN} \cdot \text{FP}}{\sqrt{(\text{TP} + \text{FN})(\text{TP} + \text{FP})(\text{TN} + \text{FN})(\text{TN} + \text{FP})}} \tag{7}$$

Additionally, Kappa error or Cohen's kappa statistics is a recommended measure to compare the performances of different classifiers and hence the quality of selected features. Generally Kappa error value $\in [-1,1]$, so when Kappa error value calculated for classifiers approaches to 1, then the performance of classier is assumed to be more realistic [28]. The Kappa error measure can be calculated using the following formula:

$$\text{Kappa error} = \frac{P(A) - P(E)}{1 - P(E)} \tag{8}$$

where $P(A)$ is total agreement probability and $P(E)$ is the hypothetical probability of chance agreement.

6. Experimental Study

This lymphography database was first obtained by the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia and then it was donated by the same contributors to UCI Machine Learning Repository [29] [30]. It comprised of 148 instances represented by 18 diagnostic features. The classification of the Lymph dataset will be with respect to condition of the subject as normal, metastases, malign lymph or fibrosis. The 18features along with description, mean and standard deviation are listed in **Table 1**. Each of the classes has the sample sizes of 2, 81, 61 and 4, respectively.

All the experiments were carried in Waikato Environment for Knowledge Analysis (WEKA) a popular suite of data mining algorithms written in Java as follows:

i) RF algorithm ensemble classifier is designed based on 150 trees and 10 random features to build each tree.

ii) The suggested algorithm is trained with lymphographic dataset using 10-fold cross validation strategy to evaluate the classification accuracy on the dataset.

As mentioned earlier, the suggested system for the purpose of enhancement of Lymph diseases diagnosis applied in this study is carried out in two major phases. In the first phase, the feature space is searched to reduce the feature space and prepare the conditions for the next step. This task is carried out using three feature selection techniques, PSO, IGR and SU. For PSO feature selection, population size is 20, number of iterations is 20, individual weight is 0.34 and inertia weight is 0.33. The optimal features of these techniques are summarized in **Table 2**. It is worth noting that the number of features has remarkably reduced, therefore less storage space is required for the execution of the classification algorithms. This step helped in reducing the size of dataset to only 9 to 13 attributes. **Figure 3** visualizes the feature selection techniques agreements. The Venn diagram shows the three feature selection techniques shares the lymphatic, block of afferent, regeneration, early uptake, lymph nodes diminish, changes in lymph, defect in node, changes in node and number of nodes attributes. It also indicates that the early uptake and changes in structure attributes were common between PSO and IGR techniques. Whereas, special forms characteristic was common between IGR and SU techniques.

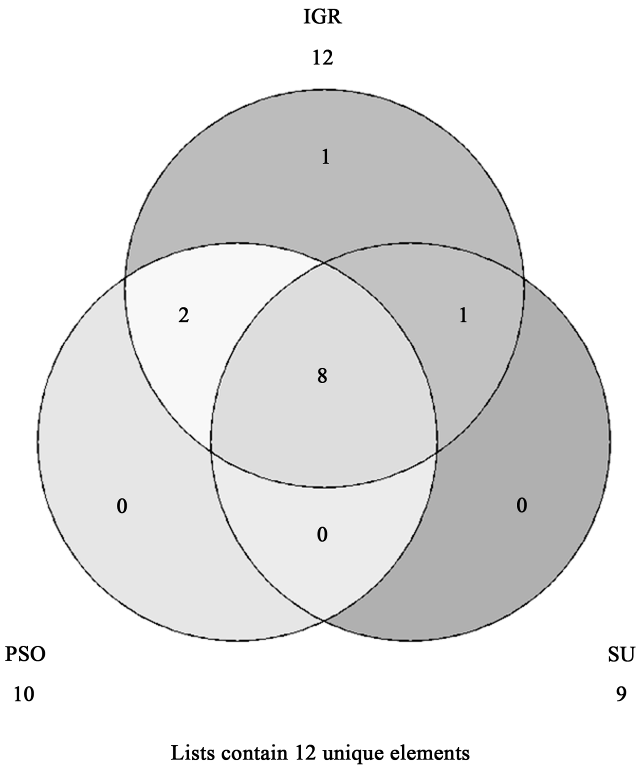

Figure 3. Feature selection techniques agreement.

Table 1. Lymphographic dataset description of attributes.

Attribute number	Attribute description	Possible values of attributes	Assigned values	Mean	S.D.
1	Lymphatic	Normal = 1, arched = 2, deformed = 3, displaced = 4	1 - 4	2.74	0.82
2	Block of afferent	No, Yes	1 - 2	1.55	0.50
3	Block of lymph c (superior and inferior flaps)	No, Yes	1 - 2	1.17	0.38
4	Block of lymph s (lazy incision)	No, Yes	1 - 2	1.04	0.21
5	By pass	No, Yes	1 - 2	1.24	0.43
6	Extravasates (force out of lymph)	No, Yes	1 - 2	1.51	0.50
7	Regeneration	No, Yes	1 - 2	1.07	0.25
8	Early uptake	No, Yes	1 - 2	1.7	0.46
9	Lymph nodes diminish	0 - 3	0 - 3	1.06	0.31
10	Lymph nodes enlarge	1 - 4	1 - 4	2.47	0.84
11	Changes in lymph	Bean = 1, oval = 2, round = 3	1 - 3	2.4	0.57
12	Defect in node	No = 1, lacunar = 2, lacunar marginal = 3, lacunar central = 4	1 - 4	2.97	0.87
13	Changes in node	No, lacunar, lacunar marginal, lacunar central	1 - 4	2.8	0.76
14	Changes in structure	No, grainy, drop-like, coarse, diluted, reticular, stripped, faint	1 - 8	5.22	2.17
15	Special forms	No, Chalices, vesicles	1 - 3	2.33	0.77
16	Dislocation	No, Yes	1 - 2	1.67	0.48
17	Exclusion of node	No, Yes	1 - 2	1.8	0.41
18	Number of nodes	0 - 80	1 - 8	2.6	1.91
19	Target Class	Normal = 1, metastases = 2, malign lymph = 3, fibrosis = 4			

Table 2. Selected features of lymph disease data set.

Feature selection technique	Number of selected features	Selected features labels
1. PSO	10	1, 2, 7, 8, 9, 11, 12, 13, 14, 18
2. IGR	13	1, 2, 7, 8, 9, 10, 11, 12, 13, 14, 15, 18
3. SU	9	1, 2, 7, 9, 11, 12, 13, 15, 18

Afterwards, the selected features are used as the inputs to the classifiers. For the purpose of classification, three machine-learning classification paradigms, which are considered very robust in solving non-linear problems, are examined to estimate the lymph disease possibility. These methods include C4.5 as decision trees, k-NN as an instance based learner and feed-forward artificial neural network classifier (Multi-layered Perceptron MLP). The k-NN classifier is performed based on Euclidean distance measure for k = 1. While C4.5 classifier was applied with a confidence factor for pruning = 0.25 and a minimum number of instances per leaf of 2. And MLP classifier with a learning rate = 0.3 and momentum = 0.2. In **Table 3**, we depict the comparative results of the classification performance before and after applying the feature reduction phase that deploy PSO, IGR and SU algorithms to detect the most significant features. As **Table 3** is examined, it is seen that before the feature reduction step, the highest precision rate is associated with RF classifier was 84.3% with 18 features. The proposed method based on RF+PSO approach obtained 82.6%, 67.5%, 92.4% and 66.8% for Precision, Recall, MCC, ROC and Kappa, respectively with 10 features. RF+IGR and RF+SU techniques obtained an average

Table 3. Classification performances of lymphographic data—without sampling.

Classifier	Performance index	Without FS	PSO + FS	IGR + FS	SU + FS
RF	Precision	0.843	0.826	0.781	0.777
	MCC	0.712	0.675	0.591	0.568
	ROC	0.935	0.924	0.906	0.890
	Kappa error	0.7105	0.6676	0.5942	0.5811
k-NN	Precision	0.739	0.802	0.766	0.731
	MCC	0.505	0.629	0.557	0.494
	ROC	0.754	0.811	0.783	0.755
	Kappa error	0.5069	0.6322	0.5596	0.4893
MLP	Precision	0.813	0.795	0.760	0.801
	MCC	0.653	0.626	0.536	0.622
	ROC	0.914	0.893	0.866	0.869
	Kappa error	0.6626	0.6205	0.5282	0.6075
C4.5	Precision	0.774	0.726	0.738	0.754
	MCC	0.583	0.491	0.491	0.541
	ROC	0.785	0.740	0.756	0.817
	Kappa error	0.5874	0.4937	0.5014	0.5503

Precision rate of 78.1% and 77.7% with 12 and 9 features, respectively. Clearly we can observe that the PSO helped in reducing the dimension of features. Yet, this step did not improve the classification performance.

Table 4 describes the class distribution, which clearly shows that the lymphography dataset is imbalanced. A common problem with the imbalanced data is that the minority class contributes very little to the standard algorithms accuracy. This unbalanced distribution makes the lymphography dataset suitable to test the effect of simple random sampling strategy. We, therefore, used a simple random sampling approach with replacement to rescale class distribution of the dataset. The class distributions before and after simple random sampling are given in **Table 4**. The classification performance of this trained algorithm is tested with original distribution, *i.e.*, without resampling, of data using 10-fold cross validation scheme.

Table 5 shows the final classification results after applying the random sampling strategy on the reduced dataset to balance the number of instances in the minority classes. This step contributes to make a more diverse and balanced dataset. As it could be seen from results of **Table 5**, the highest precision rate before the feature reduction step is associated with k-NN classifier was 92.7% with 18 features. The proposed method based on RF and PSO approach obtained 94%, 89.8%, 98.3% and 92.3% for Precision, MCC, ROC and Kappa error, respectively with 10 features. While the proposed method based on RF and IGR approach obtained 95.4%, 92.5%, 98.4% and 92.3% for Precision, MCC, ROC and Kappa error, respectively with 12 features. In **Table 5**, it is also seen that the other performance indexes supports this improvement with increasing values compared to un-sampled classification strategy. We can observe that proposed RF+PSO model helped in improving the classification performance with a limited number of features. The results demonstrated that these features are fairly competent to represent the dataset’s class information. In terms of Precision, MCC, ROC and Cohen's kappa coefficient our proposed technique that deploys random sampling technique succeeded in significantly improving the classification accuracy of the minority while the classification accuracy of major class remains high. The outcomes from the suggested technique show better results compared to datasets which are un-sampled and also when these attribute selection techniques are used independently. As can be seen from above results, the proposed method based on RF+PSO has produced very promising results on the classification of the possible lymph diseases patients.

Table 4. Class distribution of the Lymphographic dataset before and after simple random sampling.

Index	Class	Class distribution	
		Before sampling	After sampling
1	Normal	2	1
2	Metastases	81	74
3	Malign	61	69
4	Fibrosis	4	4

Table 5. Classification performance of Lymphographic dataset—with sampling

Classifier	Performance index	Without FS	PSO + FS	IGR + FS	SU + FS
RF	Precision	0.907	0.940	0.954	0.886
	MCC	0.833	0.898	0.925	0.790
	ROC	0.946	0.983	0.984	0.964
	Kappa error	0.8328	0.8972	0.9229	0.7954
k-NN	Precision	0.927	0.940	0.926	0.880
	MCC	0.872	0.898	0.865	0.774
	ROC	0.936	0.947	0.947	0.917
	Kappa error	0.8713	0.8972	0.8594	0.7696
MLP	Precision	0.907	0.935	0.935	0.853
	MCC	0.833	0.883	0.883	0.721
	ROC	0.946	0.952	0.950	0.924
	Kappa error	0.8328	0.8847	0.8847	0.7185
C4.5	Precision	0.888	0.841	0.827	0.774
	MCC	0.792	0.700	0.673	0.557
	ROC	0.910	0.900	0.873	0.855
	Kappa error	0.795	0.7052	0.6798	0.5562

7. Conclusion

The main goal of medical data mining is to extract hidden information using data mining techniques. One of the positive aspects is to support the analysis of this data. Therefore, accuracy of classification algorithms used in disease diagnosing is certainly an essential issue to be considered. In this article, a random forest classifier approach has been investigated to improve the diagnosis of lymph diseases. The proposed RF + PSO model improved the accuracy performance and achieved promising results. The experiments have shown that the PSO feature selection technique helped in reducing the feature space, whereas adjusting the original data with simple random sampling helped in increasing the region area of the minority class in favor of handling the existing imbalanced data property. The future plan will take into consideration by applying the proposed technique in other medical diagnosis problems.

Acknowledgements

The Public Authority for Applied Education and Training in Kuwait supports this work (grant number BS-14-02). The author would like to kindly appreciate and gratefully acknowledge, UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/) for obtaining the lymphographic dataset.

References

[1] Ciosa, K.J. and Mooree, G.W. (2002) Uniqueness o Medical Data Mining. *Artificial Intelligence in Medicine*, **26**, l-24. http://dx.doi.org/10.1016/s0933-3657(02)00049-0

[2] Ceusters, W. (2000) Medical Natural Language Understanding as a Supporting Technology for Data Mining in Healthcare Medical Data Mining and Knowledge Discovery. Cios KJ Editor, Heidelberg: Springer, pp. 32-60,.

[3] Calle-Alonso, F., Pérez, C.J., Arias-Nicolás, J.P. and Martín, J. (2012) Computer-Aided Diagnosis System: A Bayesian Hybrid Classification Method. *Computer Methods and Programs in Biomedicine*, **112**, 104-113. http://dx.doi.org/10.1016/j.cmpb.2013.05.029

[4] Cselényi, Z. (2005) Mapping the Dimensionality Density and Topology of Data: The Growing Adaptive Neural Gas. *Computer Methods and Programs in Biomedicine*, **78**, 141-156. http://dx.doi.org/10.1016/j.cmpb.2005.02.001

[5] Huang, S.H., Wulsin, L.R., Li, H. and Guo, J. (2009) Dimensionality Reduction for Knowledge Discovery in Medical Claims Database: Application to Antidepressant Medication Utilization Study. *Computer Methods and Programs in Biomedicine*, **93**, 115-123. http://dx.doi.org/10.1016/j.cmpb.2008.08.002

[6] Luukka, P. (2011) Feature Selection Using Fuzzy Entropy Measures with Similarity Classifier. *Expert Systems with Applications*, **38**, 4600-4607. http://dx.doi.org/10.1016/j.eswa.2010.09.133

[7] Luciani, A., Itti, E., Rahmouni, A., Michel Meignan, M. and Clement, O. (2006) Lymph Node Imaging: Basic Principles. *European Journal of Radiology*, **58**, 338-344. http://dx.doi.org/10.1016/j.ejrad.2005.12.038

[8] Sharma, R., Wendt, J.A., Rasmussen, J.C., Adams, A.E., Marshall, M.V. and Sevick-Muraca, E.M. (2008) New Horizons for Imaging Lymphatic Function. *Annals of the New York Academy of Sciences*, **1131**, 13-36. http://dx.doi.org/10.1196/annals.1413.002

[9] Guermazi, A., Brice, P., Hennequin, C. and Sarfati, E. (2003) Lymphography: An Old Technique Retains Its Usefulness. *RadioGraphics*, **23**, 1541-1558. http://dx.doi.org/10.1148/rg.236035704

[10] Cancer Research UK. http://www.cancerresearchuk.org

[11] Karabulut, E.M., Özel, S.A. and İbrikçi, T. (2012) A Comparative Study on the Effect of Feature Selection on Classification Accuracy. *Procedia Technology*, **1**, 323-327. http://dx.doi.org/10.1016/j.protcy.2012.02.068

[12] Derrac, J., Cornelis, C., García, S. and Herrera, F. (2012) Enhancing Evolutionary Instance Selection Algorithms by Means of Fuzzy Rough Set Based Feature Selection. *Information Sciences*, **186**, 73-92. http://dx.doi.org/10.1016/j.ins.2011.09.027

[13] Madden, M.G. (2009) On the Classification Performance of TAN and General Bayesian Networks. *Knowledge-Based Systems*, **22**, 489-495. http://dx.doi.org/10.1016/j.knosys.2008.10.006

[14] De Falco, I. (2013) Differential Evolution for Automatic Rule Extraction from Medical Databases. *Applied Soft Computing*, **13**, 1265-1283. http://dx.doi.org/10.1016/j.asoc.2012.10.022

[15] Abellán, J. and Masegosa, A.R. (2012) Bagging Schemes on the Presence of Class Noise in Classification. *Expert Systems with Applications*, **39**, 6827-6837. http://dx.doi.org/10.1016/j.eswa.2012.01.013

[16] Kennedy, J. and Eberhart, R.C. (2001) Swarm Intelligence. Morgan Kaufmann Publishers, Burlington.

[17] Kennedy, J. and Eberhart, R.C. (1997) A Discrete Binary Version of the Particle Swarm Algorithm. *Proceedings of the IEEE International Conference on Systems*, *Man*, *and Cybernetics*, **5**, 4104-4108. http://dx.doi.org/10.1109/icsmc.1997.637339

[18] Quinlan, J.R. (1993) C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, Burlington.

[19] Fayyad, U. and Irani, K. (1993) Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. *Proceedings of the* 13*th International Joint Conference on Artificial Intelligence*, Chambéry, 28 August-3 September 1993, 1022-1027.

[20] Liu, H., Hussain, F., Tan, C. and Dash, M. (2002) Discretization: An Enabling Technique. *Data Mining and Knowledge Discovery*, **6**, 393-423. http://dx.doi.org/10.1023/A:1016304305535

[21] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32.

[22] Dietterich, T.G. (2000) An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. *Machine Learning*, **40**, 139-157. http://dx.doi.org/10.1023/A:1007607513941

[23] Breiman, L. (1996) Bagging Predictors. *Machine Learning*, **24**, 123-140.

[24] Weiss, G. and Provost, F. (2003) Learning when Training Data Are Costly: The Effect of Class Distribution on Tree Induction. *Journal of Artificial Intelligence Research*, **19**, 315-354.

[25] Park, B.-H., Ostrouchov, G., Samatova, N.F. and Geist, A. (2004) Reservoir-Based Random Sampling with Replacement from Data Stream. *Proceedings of the* 2004 *SIAM International Conference on Data Mining*, 22-24 April 2004,

Lake Buena Vista, 492-496. http://dx.doi.org/10.1137/1.9781611972740.53

[26] Mitra, S.K. and Pathak, P.K. (1984) The Nature of Simple Random Sampling. *The Annals of Statistics*, **12**, 1536-1542. http://dx.doi.org/10.1214/aos/1176346810

[27] Schumacher, M., Hollander, N. and Sauerbrei, W. (1997) Resampling and Cross-Validation Techniques: A Tool to Reduce bias Caused by Model Building? *Statistics in Medicine*, **16**, 2813-2827. http://dx.doi.org/10.1002/(SICI)1097-0258(19971230)16:24<2813::AID-SIM701>3.0.CO;2-Z

[28] Ben-David, A. (2008) Comparison of Classification Accuracy Using Cohen's Weighted Kappa. *Expert Systems with Applications*, **34**, 825-832. http://dx.doi.org/10.1016/j.eswa.2006.10.022

[29] Cestnik, G., Konenenko, I. and Bratko, I. (1987) Assistant-86: A Knowledge-Elicitation Tool for Sophisticated Users. In: Bratko, I. and Lavrac, N., Eds., *Progress in Machine Learning*, Sigma Press, Wilmslow, 31-45.

[30] UCI (2016) Machine Learning Repository. http://archive.ics.uci.edu/ml/index.html