Comparative Study Between Classification Algorithms Based on Prediction Performance

A. G. Hari Narayanan[1,2], Megha Prabhakar[2(✉)], B. Lakshmi Priya[2(✉)], and J. Amar Pratap Singh[3(✉)]

[1] Department of Computer Application, Noorul Isalm Centre for Higher Education, Kumaracoil, Thucklay, Kanyakumari 629 180, Tamilnadu, India
hariag2002@gmail.com

[2] Department of Computer Science and IT, Amrita School of Arts and Sciences, Kochi Amrita Vishwa Vidyapeetham, Kochi, India
meghal2@live.com, lakshmipriyab007@gmail.com

[3] Department of Computer Science and Engineering, Noorul Isalm Centre for Higher Education, Kumaracoil, Thucklay, Kanyakumari 629 180, Tamilnadu, India
japsindia@yahoo.com

Abstract. In todays "data-centric" world, the prevalence of vast and immeasurable amount of data pertaining to various fields of study has led to the need for properly analyzing and apprehending this information to yield knowledge that becomes useful in decision making. Among the many procedures for handling this multitude of data, "classification" is the one that aids in making decisions based on categorization of data and "feature selection" is the process of picking out attributes relevant to the study. Keeping classification as the central idea of our study, we aim at presenting a comparative analysis of prediction accuracies obtained by two chosen classification algorithms, namely, SVM and RBFN. We proceed to introduce feature selection using both filter and wrapper methods along with SVM and RBFN to showcase a detailed analytical report on variations in performance when using classification algorithms alone, and with application of feature selection. The four approaches used for feature selection in our study are; Information Gain, Correlation, Particle Swarm Optimization (PSO) and Greedy method. Performance of the algorithms under study is evaluated based on time, accuracy of prediction and area under ROC curve. Although time and accuracy are effective parameters for comparison, we propose to consider ROC area as the criterion for performance evaluation. An optimal solution will have the area under ROC curve value approaching 1.

Keywords: Classification · Feature selection · SVM · RBFN
KNN · ROC curve

1 Introduction

We live in a data age. All around us and in every field of work like medicine, engineering, science, business etc., bulk amount of data is generated at an explosive rate. This growth can be attributed to computerization of our society and also the

J. K. Mandal and D. Sinha (Eds.): CSI 2017, CCIS 836, pp. 185–196, 2018.
https://doi.org/10.1007/978-981-13-1343-1_19

development of data management tools and storage mechanisms. It is common knowledge that businesses use this tremendous and widely available data to procreate meaningful information by understanding, analyzing and manipulating this data based on their requirements. This has essentially led to the development of "data mining", a hugely popular field of study today that deals with churning out useful information or knowledge from raw, unprocessed data using many data mining methods and techniques.

Following are the main functions in the process of mining data:

- Characterization: Summarization of general features of a target class
- Discrimination: comparing features of a target class against other conflicting classes
- Mining frequent patterns, associations and correlation: Extract usual and frequent patterns
- Classification: procedure of building a model to interpret and ascertain object classes
- Prediction: process of deducing class labels for objects whose labels are unknown
- Clustering: Congregating similar objects in groups
- Outlier Analysis: process of analyzing incompatible objects or outliers

The process of mining data is realized through a number of steps, were a combination of more than one these steps usually yields better results in terms of performance to the end user. These are:

- Cleaning and assimilating data from different sources
- Selecting data applicable to the analysis task
- Fine tune data to a form that supports the task in hand
- Extract desired results using mining methods
- Evaluate the result
- Deliver the result in a meaningful and useful manner

Realization of each of these steps or functions is possible through certain methods, techniques and algorithms available with various data management tools. A detailed discussion on all the above steps and functions is out of scope of our study as we focus mainly on classification, prediction and feature selection. The classification algorithms used here are Support Vector Machine (using SMO) and Radial Basis Function Network (RBFN). Along with classification, feature selection concept is also experimented to improve that performance of the algorithms under study. We chose two filter methods using Information Gain and Correlation, and two wrapper methods using PSO and Greedy method.

A. Classification

Data objects may be associated with a class or a category and the objects may fall into one or more classes. In data mining terms, "classification" refers to the process of sculpting a model that identifies and distinguishes the classes among objects. Classification process is associated with a "training set" which refers to the original data that is used to generate the model using a classification algorithm. Once a successful model is created, a "test set" can be run on the model to predict the class labels of unclassified objects in the test set. For example, a data collection describing features of emails

received in an inbox can have them fall into any of the two categories; spam or not-spam. A model can be created using an available training data and the generated model can be used to determine whether a new email in the inbox is a spam or not.

Support Vector Machines

SVM is a highly accurate algorithm that categorizes objects by trying to compute a maximum margin hyper plane. This hyper plane would enhance the training data to a high dimension by separating objects that belong to different classes. SVMs can assort classes to objects that are linearly or non-linearly separable, making it even more acceptable. The separating hyper plane is calculated using the following formulae;

$$W.X + b = 0$$

where W is a weight vector and W = {w_1, w_2, w_3, w_4,….., w_n} and b or "bias" is a scalar value.

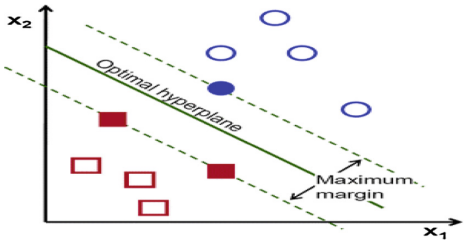

Fig. 1. Support vector machine (Color figure online)

There may be in most conditions infinite number of hyper planes available and SVM algorithm has to find out the best one out of them. The best hyper plane is estimated as the one having the largest margin so that it can precisely categorize future objects than the hyper planes with smaller margins (as shown in Fig. 1). The colored objects in Fig. 1 represents the "support vectors" and are equally close to the optimal dividing hyper plane. We chose to use the SMO package in WEKA to implement SVM based classification.

Radial Basis Function Network

RBFN is a neural networks based algorithm that uses radial functions to realize its calculations. It is associated basically with three layers, namely, the input layer, hidden layer and the output layer. In terms of classification process, the input layer represents all the attribute input units of the task. The hidden layer consists of nodes were each node represents a radial function. Each input node is connected to each of the nodes in the hidden layer. The output layer gives you the "class" of an object and is a weighted sum of all the outputs from the hidden layer. A general representation of the RBFN algorithm is depicted in Fig. 2.

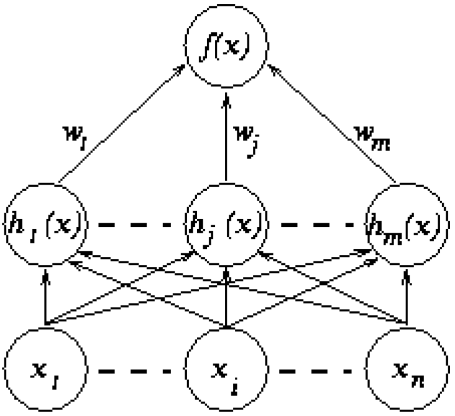

Fig. 2. Radial basis function network

Here, x represents the n input attributes or nodes that are dispatched forward to each of the hidden layer nodes. The hidden layer nodes are represented as the function h(x) and the output as f(x). The radial function h(x) at each hidden node is represented as,

$$\mathrm{h(x)} = \exp\left(-\frac{(x-c)^2}{r^2}\right)$$

where r is the radius or standard deviation, c is the center and can be calculated using any clustering method like the K-Means. The output function f(x) takes the following form,

$$\mathrm{f(x)} = \sum_{j=1}^{m} w_j h_j(x)$$

The initial step is the calculation of the centers c and the number of centers will represent the total number of nodes in the hidden layer. The second step attaches a weight w to the output of the hidden layer nodes in order to calculate the sum and give out the final output. Although neural networks are generally disregarded for their relatively larger time consumption for training and poor interpretability, they are well received and acknowledged for their ability to tolerate noisy data and classify objects into categories they are unaware of. In WEKA, RBFNetwork package is used to perform a radial basis function network based classification.

B. Feature Selection

A realistic data set may contain huge amount of tuples represented using comparatively larger number of attributes or features. And all these attributes may not prove to be appropriate or relevant to the analysis task in hand and should be ignored based on some statistical conditions. Feature Selection or Attribute Subset Selection is a data preprocessing step that aims at discarding attributes that are irrelevant and not applicable to the analysis task in hand. This elimination of irrelevant features is done such that the results attained are proximately similar or better than those achieved by considering all the attributes. This process is intended at ignoring all the insignificant

features thus prompting better accuracy, performance and reduced time and complexity. The two approaches to feature selection are using filter and wrapper methods. Filter methods aspire to rank or score the attributes in a way independent to the classifier used. Thus, they provide you with a more generalized result than that of wrapper methods which tries to score attributes and generate a subset by exploring the feature set by means of a classification model. Each new subspace of attributes are tested against the model and accordingly scored based on performance. Wrapper approach of feature subset selection is thus considered better than filter methods as they furnish the optimal feature subset.

Feature Ranking Based on Information Gain

Information Gain is a measure used for choosing attributes that have the most information required to perform a task like classification. It removes ambiguity by recognizing only those features that holds the highest value among others. An attribute with the highest information gain is considered as the "splitting attribute" where the feature set gets split or divided. The expected average information required to classify an object is given by,

$$\text{Info}(\text{D}) = -\sum_{i=1}^{m} p_i log_2(p_i)$$

where p_i is the probability that a tuple belongs to class C_i and Info(D) is also called the entropy of D. Now, if the feature set is split at attribute A, then the amount of expected information required further to achieve an efficient classification is given as,

$$\text{Info}_{\text{A}}(\text{D}) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

Information Gain is given as,

$$\text{Gain}(\text{A}) = Info(D) - Info_A(\text{D})$$

The attributes with the highest gain value is chosen as the next splitting attribute so that the overall information required to classify an object or tuple is kept at minimum. The final output offered by this measure will be a ranked list of attributes of a feature set starting from the ones that hold highest value (information gain) or information. In WEKA, feature selection using Information Gain is realized using the InfoGainAttributeEval method

Feature Ranking Based on Correlation

Correlation among attributes in a feature set illustrates the dependency between attributes or how strongly on feature entails another. Correlation for a numeric data goes by the following formulae,

$$\text{r}_{\text{A,B}} = \frac{\sum_{i=1}^{n} (a_i b_i) - n\bar{A}\bar{B}}{n\sigma_A \sigma_B}$$

where r represents the correlation coefficient, n is the number of tuples, a_i and b_i are the values of A and B in tuple i, $\bar{A}$ and $\bar{B}$ are mean values of attributes A and B and σ_A and σ_B are the standard deviations of A and B. The principle followed for ranking and selecting features based on correlation is that the attributes remain highly correlated to the class and minimally correlated to each other. It is implemented in WEKA using CorrelationAttributeEval method.

Particle Swarm Optimization

PSO is a type of meta-heuristic optimization algorithm inspired from the social behavior of birds or fish. It was developed in the year 1995 with the objective of creating a model to express the social behavior of animals like a flock of birds or school of fish and later on came to be used as an optimization algorithm in various branches of science and engineering like machine learning, data mining, image processing etc. The main task of an optimization problem is to find the "best" solution and for this, it uses the concepts of communication and learning. PSO works similarly by enabling the swarm (search space) members to communicate learn about each other and achieve the global best solution. Each swarm member i have a position denoted by X_i, a velocity V_i that describes the movement of a particle and a personal best experience denoted by $Pbest_i$ as in Fig. 3.

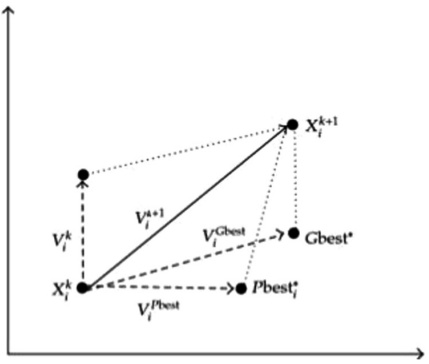

Fig. 3. PSO working

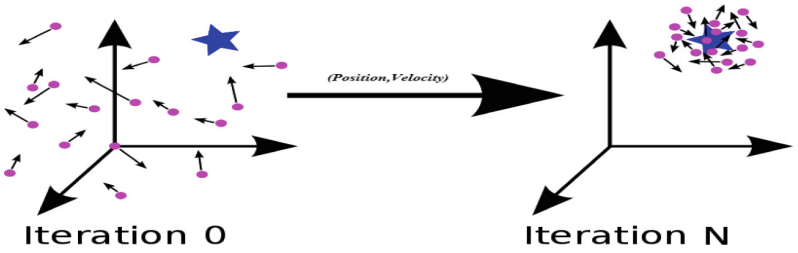

Fig. 4. Swarm movement

In advance to position, velocity and personal best, there is a global best value common to all swarm members and is not specific to any swarm member denoted by Gbest as in Fig. 4. With each iteration, the swarm members co-operate to achieve this global best solution by updating their position and velocity accordingly in each iteration. In Fig. 3, the particle X_i^k has velocity V_i^k and personal best value Pbest_i. In the next iteration k + 1, the particle moves parallel to its velocity vector, personal best vector and the global best vector to achieve the new position X_i^{k+1} which could be a better position for the particle i. Hence in every iteration, particles move closer to the global best solution in the swarm. The formulae for updating particle position in the swarm is given by,

$$X_i^{k+1} = X_i^k + V_i^{k+1}$$

$$V_i^{k+1} = wV_i^k + C_1\left(V_i^{Pbest}\right) + C_2\left(V_i^{Gbest}\right)$$

where w, c1 and c_2 are real valued coefficients, $V_i^{Pbest} = \left(Pbest_i^k - X_i^k\right)$ and $V_i^{Gbest} = \left(Gbest_i^k - X_i^k\right)$. This optimization capability of PSO makes it highly efficient and suitable for feature subset selection applications. It is implemented in WEKA using the PSOSearch algorithm.

Greedy Method

Greedy methods can be employed in algorithms to look for an optimum solution in every step. In WEKA, it is implemented using the GreedyStepwise algorithm along with an evaluator. The algorithm can start from any position in the search space with no/all attributes under consideration. It then moves forward/backwards by adding/removing attributes at every step. The evaluator algorithm verifies the performance of the subspace and if a decrease in evaluation if encountered, the algorithm terminates.

2 Related Works

To support our study, we have analyzed various research works done in the same context. Iain Brown and Christophe Mues in their study [1] have proposed a comparative study between traditional classification algorithms like logistic regression, neural networks and decisions trees while also exploring the possibilities of least square support vector machines, gradient boosting and random forests. Performance was measured based on the Area under the curve (AUC) value. Their study was carried out on a credit score feature set prone to imbalanced data. The implementation and analysis of the problem offered a result that favors random forests and gradient boosting classifiers which were found to have better performance than the other classifiers as they were found to have better tolerance to imbalanced data.

Yet another analysis study [2] offers a collative study between the Naïve Bayesian and Decision Tree algorithm. These were applied on a collection of crime information to predict the type of crime in different states of The United States of America. The

analysis was implemented in WEKA tool and evaluated based on accuracy value. The authors found that Decision Tree algorithm for classification surpasses Naïve Bayesian classifier by offering an accuracy of 83.9519%.

Kotsiantis has successfully and clearly explained some of the best supervised learning approaches [3]. This survey provides an in-depth analysis of some of the most popular algorithms under the categories of logic-based, perceptron-based, statistical and instance-based. Decision tree algorithm is chosen for analysis under logic-based approach, Single layered perceptron, multi layered perceptron and radial basis function networks are reviewed under the perceptron-based category, Naïve Bayesian and Bayesian networks are explored under the statistical category and K-nearest neighbor classifier is examined under the instance-based method. Further, the study proceeds to analyze Support Vector Machines which is one of the recent and popular classification algorithms used. The analytical survey offered by the author provides adequate prior knowledge about various supervised learning methods.

A comparative study [4] done on a feature set having information regarding a car manufacturer's product characteristics proposed the usage of classification algorithms like CHAID, C&R and QUEST belonging to tree category algorithms. Along with these, neural networks, Bayesian, logistic regression and SVM classifiers were also tested on the data for fault prediction. Support Vector Machines were found to have the overall better accuracy level compared to all the other classifiers chosen. The output of the study also suggests that even though the tree algorithms consume more time to build the model for classification, they offer good accuracy levels.

A survey report presented by Dash and Liu [5] illustrates in detail the use of feature selection in classification. Feature selection has become popular in the recent times as it reduces the time, complexity and is found to improve performance of a classifier. The authors in their study provides an extensive analysis on the types of feature selection methods and identifies four steps in feature selection namely, generation procedure, evaluation function, stopping criterion, and validation procedure. It is found that surprisingly, many feature selection methods do not perform or attempt the first two steps.

Some other notable study references are also discussed briefly. In [6], the author makes an analysis on comparative approach of study, proposes a recommended approach, and also offers a list of pitfalls to avoid while making a comparative study. [7] provides an insight into the evolutionary optimization algorithm known as Particle Swarm Optimization(PSO). PSO is applied on an SVM classifier to address prevalent classification problems and found that PSO-SVM hybrid approach gives better accuracy than other feature selection approaches. Instance-Based or Memory-Based Learning, Error back propagation, (k-NN) k-Nearest Neighbor algorithm and (MLP) multilayer perceptron algorithms are implemented in [8] on a data set using WEKA tool and the results favors kNN classifier that achieved 73.33% accuracy.

3 Experiment

In the experiment study, we chose to compare the predictive classification performances of two algorithms namely, SVM (using SMO) and RBFN taken individually and also with each of the feature selection methods. This gives us the following

different algorithms for comparison; SMO, InfoGainAttributeEval - SMO, CorrelationAttributeEval - SMO, PSO - SMO, GreedyStepwise - SMO, RBFN, InfoGain AttributeEval - RBFN, CorrelationAttributeEval - RBFN, PSO – RBFN and GreedyStepwise - RBFN. We chose to work in WEKA platform as it is a simple, easy-to-use and powerful tool that offers surplus amount of algorithms, methods and solutions to realize data mining tasks. A model is build using a chosen classifier from the above list on the training set. Using this model, a test data set is evaluated in order to study the performance of the algorithm and to analyze which feature selection approach provides better evaluation results for each classifier. The collative study is done based on three values obtained after prediction; time to predict, accuracy and ROC area. The experiment results are depicted in Table 1.

A. Time to Predict

This value corresponds to the time taken by a classifier to predict the class labels for a test set based on a predictive model. An efficient algorithm will always take the least amount of time for classification.

B. Accuracy

In WEKA, the classification result parameter "Correctly classified instances" is regarded as the accuracy for that classifier. It represents in number and percentage the amount of test tuples that were correctly classified against the prediction model. This value directly reflects the accuracy of the classifier used for the process.

C. ROC curve

Receiver Operating Characteristics curve plots the trade-off between true positive rates (TPR) and false positive rates (FPR). TPR corresponds to the magnitude of positive tuples that are correctly labeled and FPR represents the magnitude of negative tuples that are incorrectly labeled as positive. The area under the ROC curve exemplifies the accuracy of the classification algorithm used. A proficient and compelling classifier will have the average ROC curve value approaching 1 while any value less that 0.5 is equal to random guessing. ROC curve representations for the best two classifiers among SMO and RBFN are depicted in Figs. 5 and 6, where the X-axis represents true positives and the Y-axis depicts false positives.

D. Database

All the algorithms under consideration were run on a data set representing types of forests based on their spectral characteristics. There are 27 attributes that describe the spectral characteristics in wavelengths captured from satellite images. The feature set provides a training set having a total of 198 instances and a test set with 325 instances. The four class labels are described as s, h, o and d attributing to the four different forest types.

Table 1. Predictive analysis results for forest type's data

Prediction algorithm		Time to test (in seconds)	Accuracy (%)	Avg. ROC area
SMO	SMO	0.04	84	0.917
	InfoGain-SMO	0.04	84.3077	0.921
	Correlation-SMO	0.02	83.3846	0.913
	PSO-SMO	0.01	83.0769	0.914
	GreedyStewise-SMO	0.03	83.0769	0.912
RBFN	RBFN	0.03	13.5385	0.495
	InfoGain-RBFN	0.01	50.4615	0.744
	Correlation-RBFN	0.01	49.8462	0.738
	PSO-RBFN	0.01	48	0.749
	GreedyStewise-RBFN	0.01	53.5385	0.700

4 Experiment Results

From the results of the experiment study presented in Table 1, it is clearly specific that Support Vector Machines using SMO provides better prediction results than Radial Basis Function network classifier. SMO achieves better accuracy and average area under ROC curve that are indications of an efficient classifier. It is also found that feature selection in general has a positive effect on prediction results of both the classifiers. The finest performance is achieved by InfoGain-SMO classifier with an accuracy of 84.3077% and ROC area of 0.921, which is closer to 1. The least performance is displayed by the RBFN classifier without the use of any feature selection mechanism. Also, the results rank information gain and Particle Swarm Optimization as the best feature subset selection methods. Taking into consideration the time taken to

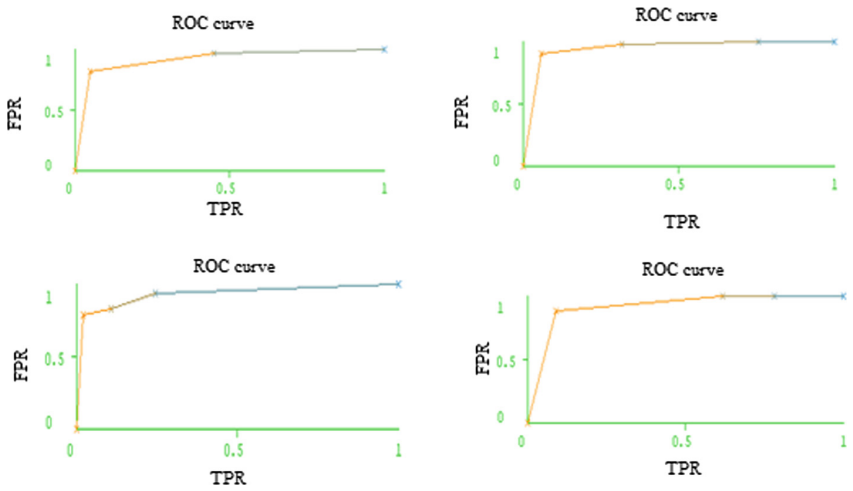

Fig. 5. ROC curve for classes d, h, o and s using SMO

predict the classes for tuples, it is found that RBFN holds a better stand when compared to SMO. This suggests that SMO, while being the better classifier than RBFN in terms of accuracy and ROC area, consumes more time to realize its output.

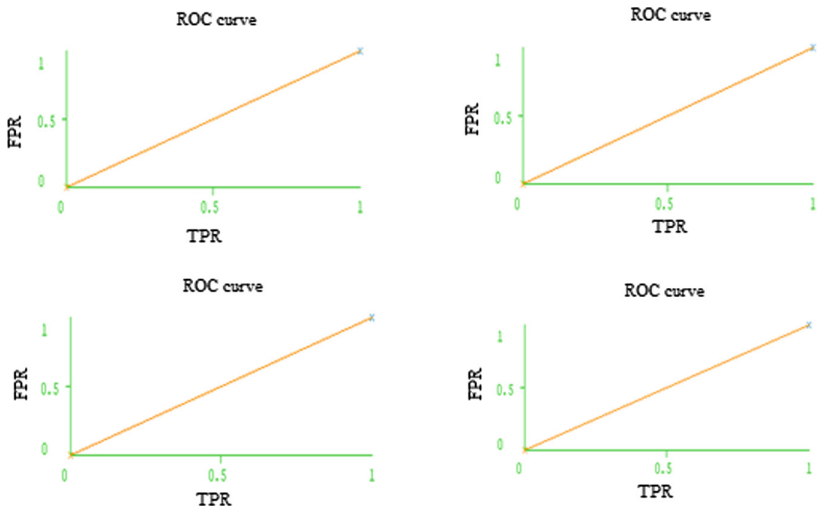

Fig. 6. ROC curve for classes d, h, o and s using RBFN

5 Conclusion

Classification is an important task in data mining that can provide efficient solutions to categorization problems in the real world. It is therefore of the utmost importance that classifiers of the highest efficiency in terms of performance are identified to reduce complexity, consumption of time and errors. The proposed study precisely aims at addressing this need by presenting a comparative study between two different classification algorithms and a combination of feature selection approaches along with them. It is found that Support Vector Machine classifier using SMO generally offers better prediction performance than RBFN classifier and feature selection used along with classification algorithms proves to provide better prediction results. These results are achieved on a numerical data set and may vary depending on the type of data used for the task.

For future works, more data sets or corpuses containing different types of data like numeric, nominal, mixed, etc. could be used for a similar kind of analytical experiment using different classification algorithms. Further, more evolutionary and genetic algorithms (GA) for feature selection besides PSO could be studied as they are highly efficient, popular and usually used along with classification algorithms to obtain better results.

References

1. Brown, I., Mues, C.: An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Syst. Appl. **39**(3), 3446–3453 (2012)
2. Iqbal, R., Murad, M.A.A., Mustapha, A., Panahy, P.H.S., Khanahmadliravi, N.: An experimental study of classification algorithms for crime prediction. Indian J. Sci. Technol. **6** (3), 4219–4225 (2013)
3. Kotsiantis, S.B.: Supervised machine learning: a review of classification techniques. Informatica **31**, 249–268 (2007)
4. Amooee, G., Minaei-Bidgoli, B., Bagheri-Dehnavi, M.: A comparison between data mining prediction algorithms for fault detection. IJCSI Int. J. Comput. Sci. **8**(6(3)) (2011)
5. Dash, M., Liu, H.: Feature selection for classification. Intell. Data Anal. **1**, 131–156 (1997)
6. SalzBerg, S.L.: On comparing classifiers: pitfalls to avoid and a recommended approach. Data Mining Knowl. Discov. **1**, 317–327 (1997)
7. Tu, C.-J., Chuang, L.-Y., Chang, J.-Y., Yang, C.-H.: Feature selection using PSO-SVM. IAENG Int. J. Comput. Sci. **33**(1), IJCS_33_1_18
8. Stylios, I.C., Vlachos, V., Androulidakis, I.: Performance comparison of machine learning algorithms for diagnosis of cardiotocograms with class inequality. In: International IEEE Conference on TELFOR 2014 (2014)