Ensemble learning-based classification of microarray cancer data on tree-based features

Guesh Dagnew | B.H. Shekar

Department of Studies and Research in Computer Science, Mangalore University, Mangalore, Karnataka, India

Correspondence
Guesh Dagnew, Department of Studies and Research in Computer Science, Mangalore University, Mangalore, Karnataka, India. Email: guesh.nani@gmail.com

Abstract
Cancer is a group of related diseases with high mortality rate characterized by abnormal cell growth which attacks the body tissues. Microarray cancer data is a prominent research topic across many disciplines focused to address problems related to the higher curse of dimensionality, a small number of samples, noisy data and imbalance class. A random forest (RF) tree-based feature selection and ensemble learning based on hard voting and soft voting is proposed to classify microarray cancer data using six different base classifiers. The selected features due to RF tree are submitted to the base classifiers as the training set. Then, an ensemble learning method is applied to the base classifiers in which case each base classifier predicts class label individually. The final prediction is carried out hard and soft voting techniques that use majority voting and weighted probability on the test set. The proposed ensemble learning method is validated on eight different standard microarray cancer datasets, of which three of the datasets are binary class and the remaining five datasets are multi-class datasets. Experimental results of the proposed method show 1.00 classification accuracy on six of the datasets and 0.96 on two of the datasets.

1 | INTRODUCTION

Microarray cancer data analysis is one of the hot research areas across several disciplines such as Statistics, Machine Learning, Computational Biology, Pattern recognition and other related fields as it has significant contribution in the identification, diagnosis and treatment of cancer [1]. The researches in the area are aiming towards the enhancement of survival rates in cancer patients by improving the procedure and knowledge of screening and treatment. The main difficulty with microarray dataset classification arises from several problems such as noisy data, imbalanced class, lack of sufficient samples and the high curse of dimensionality of features that led to complex in analysis and have results inaccurate classification [2–4]. Many research works related to binary class microarray cancer data classification have been carried out. Classifying multi-class microarray data is still an open research topic due to a result of challenges in class imbalance whereby classes with a small number of samples are usually neglected because of the bias of most of the models towards classes having more number of elements [5–7].

The authors propose to explore the random forest (RF) tree-based feature selection method to get the relevant features and apply these features in the ensemble learning-based microarray cancer data classification. The RF tree is getting attention in many domains of machine learning tasks for feature selection as it provides optimal features by ranking according to their relevance. According to Breiman [8], RF tree is an improved version of bagged algorithms. It is designed to overcome the greedy nature of the bagged decision trees (DTs) by combining the predictions of multiple predictive models in ensemble techniques. The RF algorithm learns the predictions of the sub-trees in such a way that predictions from sub-trees are weakly correlated. Unlike the bagging algorithms that search across all features for the optimal split, RF tree considers a random sample of the features $m$ to search for the optimal split from the $p$ sample size. RF is an ensemble supervised learning algorithm used for classification and feature selection task. The RF tree-based feature selection framework is shown in Figure 1.

In the classification phase, we explored an ensemble learning-based microarray cancer data classification.
ensemble-based classifiers are meta-classifiers that are a combination of conceptually similar or different machine learning classifiers for classification purpose by employing hard voting that uses majority prediction and soft voting that uses averaging the class-probabilities of the predicted class level [9]. Combining predictions of multiple base classifiers in ensemble learning, works better when the sub-samples are weakly correlated. In our work, we use six heterogeneous classifiers namely logistic regression (LR), multilayer perceptron (MLP), support vector machine (SVM), RF simple DT and k-nearest neighbour (KNN) as base models. The prediction of the individual base classifiers is combined in two different methods of hard voting and soft voting. In the case of hard voting approach, the final class label of a given sample is computed on the basis of majority voting, whereas the soft voting method uses weighted probabilities for each class and the maximum probability is considered as the final class label.

We conduct comparative analysis of the proposed method with the state-of-the-art methods and the empirical results show that the proposed method is performing better when compared with many of the related works and is equally performing with some of them.

The rest of the work is organized as follows. Section 2 presents literature that are considered to be state-of-the-art works in the domain of feature selection and classification of microarray cancer data. Section 3 introduces both the proposed feature selection and classification methodologies. In Section 4, we discuss the experimental setup of the proposed work. In Section 5 experimental results are presented. Section 6 provides discussion on experimental results. Conclusion and future work are covered in Section 7.

2 RELATED WORKS

In this section, we present related state-of-the-art works in the domain of feature selection and classification of microarray cancer data. Ebrahimpour et al. [10] and Das et al. [11] proposed an ensemble learning methods using a bi-objective genetic algorithm and hesitant fuzzy set approach respectively. An integrated Particle Swarm Optimization (PSO) algorithm with C4.5 classifiers called PSOC4.5 was proposed by Chen et al. [12] to address gene selection problem. They evaluate the method with average performance using a fivefold cross-validation method on various microarray cancer datasets. A modification of the analytic hierarchy process for gene selection was proposed in [13]. An adaptive rule-based classifier was proposed by Farid et al. [14] for big biological datasets. They use the DT and KNN to construct their model. A comparative research work is carried out by Nematzadeh et al. [15] aiming to classify cancer data using machine learning methods. It is a well-known fact that the feature selection plays a great role in minimizing complexity arising from the curse of dimensionality by selecting informative features and neglecting insignificant ones.

Liu et al. [16] proposed weighted Extreme Learning Machine (ELM) for classification of multi-class microarray cancer data. A hybrid method namely C-E MWELM was proposed to handle the class imbalance in multi-class microarray cancer data at both feature level and algorithmic level. The datasets namely Brain-Tumor1, Brain-Tumor2, 9-Tumors, 11-Tumors, Leukemia1, Leukemia2, Lung Cancer and SRBCT are used in their experimentation.

Lai et al. [17] suggested a hybrid filter and wrapper method named IG-ISSO for gene selection problem. Information gain (IG) is applied to select features, and Improved Simplified Swarm Optimization (ISSO) is proposed as a gene search engine to guide the search for an optimal gene subset. The SVM with a Linear Kernel is used as a classifier.

Lu et al. [18] introduced a hybrid feature selection method that combines Mutual Information Maximization and Adaptive Genetic Algorithm. Four classifiers are used to test the proposed model namely Backpropagation Neural Network, SVM, ELM and Regularized Extreme Learning Machine. Ramos-Gonzalez et al. [19] proposed the application of supervised machine learning for classification of several types of cancer via deep learning. The MLP neural network architecture is used for lung cancer classification.

Tabakh et al. [20] proposed an unsupervised gene selection method called MGSACO, which incorporates the Ant Colony Optimization (ACO) algorithm into the filter approach thereby minimizing the redundancy between genes and maximizing the relevance of genes. It works by an iterative improvement process where at each iteration a population of agents selects a subset of genes. Then, the performance of the proposed method that generates the subsets of genes is evaluated using a fitness function. Finally, the best subset of genes in all iterations is chosen as the final gene set. Three classifiers are used to test the selected genes for classification namely SVM, Naïve Bayes and DT. Motivated by the state-of-the-art methods, The authors propose to explore RF-based feature selection and device two ensemble learning classification approaches namely hard voting and soft voting.
3 | PROPOSED METHOD

In Sections 3.1 and 3.2, we are introducing the proposed RF-based feature selection and ensemble learning-based classification for microarray cancer data respectively.

3.1 | Random forest tree-based feature selection

Feature selection improves model prediction by removing features with negative influence on the general performance of the model. It simplifies interpretation, enhances training time as a result of handling the curse of dimensionality, improves generalization by overcoming overfitting and reduces variance.

Microarray cancer datasets are characterized by their high dimensional features, limited sample size and imbalanced number of samples across. Obtaining the best features is a challenging task in the high dimensional dataset such as in microarray cancer data. Hence, feature selection is a crucial research topic in the area of machine learning, which aims to improve decision-making by enhancing classification accuracy and computational complexity by obtaining optimal features [21].

RF is a supervised learning algorithm containing ensembles of DTs to form the forest. The RF creates several DTs and merges them together to obtain a better result compared with the single DTs [22]. The RF method creates multiple trees using classification by searching for a random subset of input features at each splitting node and the tree grows fully without pruning [23].

The RF tree uses the Gini index impurity measure to decide the final class of features. It uses the Gini index to measure the impurity across the nodes of the tree. Given a dataset \( T \) with \( n \) classes, then the Gini index is computed as shown in Equation (1).

\[
\text{Gini}(T) = 1 - \sum_{j=1}^{n} (P_j)^2
\]

The Gini index for a given node \( t \) is computed based on Equation (2).

\[
\text{Gini}(T_t) = 1 - \sum_{j=1}^{n} (P_j|t)^2
\]

If the dataset \( T \) is split into \( k \) subsets \( T_1, T_2 \) and \( T_k \) with sizes \( N_1, N_2 \) and \( N_k \), the Gini index of the split data consists of samples from \( n \) classes, and the Gini index of the dataset \( T \) is computed based on Equation (3).

\[
\text{Gini}(T) = \frac{N_1}{N} \text{gini}(T_1) + \frac{N_2}{N} \text{gini}(T_2) + \frac{N_k}{N} \text{gini}(T_k)
\]

Based on the proposed feature selection method, we have obtained ranked set of features on eight different datasets. In selecting the candidate features, the RF uses Gini index. The selected features are shown in Figure 2a–h, respectively, to 2-class Leukaemia, 3-class Leukaemia, 5-class Lung, 3-class mixed-lineage leukemia (MLL), 2-class Ovarian, 2-class Prostate, 4-class SRBCT and 11-class Tumour datasets with their relevant ranking.

3.2 | Ensemble learning for microarray cancer data classification

We have proposed an ensemble learning method for microarray cancer data classification and details are presented in this section. Ensemble learning is the combination of the predictions of two or more or similar/dissimilar classification algorithms that are effective in handling class imbalance-related problems [16]. The ensemble classifier applies voting strategy to assign the final class label of a sample. The voting function combines the predicted results of several base classifiers that enhance the final prediction based on a certain strategy of combination of final class label [7].

The authors consider hard and soft voting techniques. In the case of hard voting, the class label of a sample is determined based on the majority voting approach, whereas in the case of soft voting the maximum average prediction of the classifiers is taken as the final class label. Initially, the base classifiers are initialized and trained independently on the training set. Then, a test set is provided to the voting function to obtain the final prediction of the voting function. The weight for each base classifier is provided manually in the case of soft voting. In predicting the final class label, the maximum average prediction of a given class is considered as the final class label, whereas majority voting is considered in the case of hard voting [9].

According to Cutler et al. [24,25], the ensemble classifier builds a function \( f \) in terms of the base classifiers \( h_1(x), h_2(x) \) and combines these base classifiers predictor function \( f(x) \), where \( f(x) \) is the most frequently predicted class or voted and is in Equation (4).

\[
f(x) = \text{majority voting}(h_1(x), h_2(x))
\]

The selected features of the microarray cancer data are loaded and split into training and test sets. The proposed method has a two-layer classification approach. In the first layer, six base classifiers with different behaviours, namely LR, MLP, SVM, RF, DT and KNN are used. In the case of SVM classifier, it does not support weighting probabilities since the default value for probabilities is set to false and hence must be turned on to true to obtain the probabilistic prediction when the soft voting is applied. The base classifiers are trained on the training data and the test data are used in the second layer to test the performance of the ensemble classifiers. In the second layer, the voting classifier is combined with hard and soft prediction as shown in Figure 3. The base classifiers are trained on the training set and test set is provided to the voting classifier which uses majority voting in the case of hard voting and weighted probabilities in the case of soft voting.

The proposed ensemble learning-based classification that uses hard and soft voting methods is presented in Figure 4. The hard voting method predicts final class label on the basis of majority voting, and the soft voting method uses weighted probability approach.
In the soft voting, the class labels are predicted based on the probabilities $p$ for a classifier $C_i$ as shown in Equation (5).

$$\hat{y} = \arg\max \sum_{i=1}^{m} (y = h_i(x))$$

The weighted probability vote is computing by associating a weight $w_i$ with classifier $C_i$ as shown in Equation (6), where $X_A$ is the characteristic function $C_i(x) = i \in A$ and $A$ is the set of unique class labels.

$$\hat{y} = \arg\max_i \sum_{j=1}^{m} w_j X_A(C_j(x) = i)$$

4 | EXPERIMENTAL DETAILS

The experimental setup of our work is described in terms of dataset description and performance measures in the following sections.
4.1 Dataset description

In this section, the description of the datasets used in this work is presented. Eight publicly available microarray cancer datasets downloaded from Shenzhen University data repositories [26], GEMS [27] and Elvira [28] were used in this work and their descriptions are shown in Table 1. The datasets are a mix of both binary and multiclass datasets. Three of the datasets namely 2-class Leukaemia, 3-class Leukaemia and 3-class MLL are from Leukaemia cancer family which comprises 7129 features in the case of 2-class Leukaemia and 3-class Leukaemia, and 12,582 features in the case of MLL dataset. The Ovarian and Prostate cancer datasets are a binary class containing 253 samples with 15,154 features and 102 samples with 12,600 features, respectively. The small round blue cell tumours (SRBCT) is a 4-class cancer dataset which contains four various childhood tumours namely Ewing’s sarcoma, Burkitt’s lymphoma, neuroblastoma, and rhabdomyosarcoma tumours. It has 83 samples and 2308 features. The 5-class lung cancer dataset has 203 samples and 12,600 features that include four types of cancers namely adenocarcinomas, small-cell lung carcinomas, squamous cell carcinomas, and carcinoids and normal lung. The 11-class tumour contains various tumour types, such as ovary, bladder, breast, colorectal, gastroesophageal, kidney, liver, prostate, pancreas, lung adenocarcinoma and lung squamous [27]. Moreover, the 2-class prostate cancer dataset contains 102 samples and 12,533 features that are a mix of normal and tumour cells [28].
### 4.2 Performance measures

To evaluate the performance of the proposed method, the following evaluation methods were used. The classification accuracy (CA), precision, recall, F-score, average precision (AP), confusion matrix and receiver operating characteristic (ROC) curve are used. The CA, precision, recall and F-score are based on the true positive where true positives are samples that are correctly classified as positives by the classifier, false positives which are samples the classifier has incorrectly classified as positives, true negative whereby the classifier has correctly classified as negatives and false negative whereby those samples that are classifier has incorrectly classified as negatives.

The classification accuracy is the ability of a model which shows how well a given classifier predicts the value of a class level for a new sample in a test set. As defined in Equation (7), the CA is the ratio of correctly classified samples to all samples in the test set.

\[
CA = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}
\]

The recall, which is widely known as sensitivity in medicine-related research, deals with the probability of a diagnostic test showing the proportion of patients with the disease who have a positive test result. In microarray cancer analysis, the recall measures the proportion of patients who have cancer and is detected by the algorithm as cancerous. Equation (8) shows the recall which is used to compute the ratio of correctly classified true positive cases to the sum of true positives and false negatives cases in the test data.

\[
Recall = \frac{TP}{TP + FN} \tag{8}
\]

In microarray cancer analysis, precision measures to what extent the patients that are diagnosed cancerous who have cancer. Precision computes the portion of positive classes which are correctly predicted. It is the ratio of true positive predictions to the sum of true positives and false positives as shown in Equation (9).

\[
Precision = \frac{TP}{TP + FP} \tag{9}
\]

The F-measure which is the harmonic mean of precision and recall is defined in Equation (10). It is essential in normalizing the bias of precision and recall metrics.

\[
F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{10}
\]

\[
E = \frac{FP + FN}{TP + TN + FP + FN} \tag{11}
\]

The AP shown in Equation (12) is the summary of a precision–recall curve as the weighted mean of precisions achieved at each threshold, with the increase in recall from the previous threshold used as the weight.

\[
AP = \sum_n (R_n - R_{n-1}) \times P_n \tag{12}
\]
where $R_n$ and $P_n$ are the recall and precision at $n$th threshold. The log loss is a loss function that computes the uncertainty level of the model’s prediction, which penalizes overconfidence and under the confidence of the prediction as shown in Equation (13).

$$E = \frac{1}{N} \sum_{i=1}^{N} (y_i \log(\hat{y}_i)) + (1 - y_i)\log(1 - \hat{y}_i)$$  \hspace{1cm} (13)

where $E$ stands to the error, $y_i$ is the actual class level (ground truth) and $\hat{y}_i$ stands to the predicted class level.

5 | EXPERIMENTAL RESULTS

In this section, the experimental results of the proposed method on eight standard microarray cancer datasets namely 2-class Leukaemia, 3-class Leukaemia, 5-class Lung, 3-class MLL, 2-class Ovarian, 2-class Prostate, 4-class SRBCT and 11-Class Tumour that are both binary and multiclass are presented. The proposed ensemble learning method uses six base classifiers, namely LR, RF tree, SVM, MLP, DT and KNN. Both hard voting and soft voting techniques are applied to combine the final prediction label of the base classifiers into the final
ensemble learning model. As shown in Table 2, experimental results on all the datasets are presented in terms of classification accuracy, precision, recall, F-score and area under the curve (AUC). Accordingly, the proposed method achieves a classification accuracy of 100% in six of the datasets namely 2-class Leukaemia, 3-class Leukaemia, 3-class MLL, 2-class ovarian, 2-class prostate and 4-class SRBCT in the case of soft voting method. The model scored 96% in 5-class Lung and 11-class Tumour cancer datasets.

The proposed method scored 93% AP in the case of 5-class Lung and 11-class Tumour cancer datasets and 100% performance in the rest of the datasets. Similarly, 98% performance of AUC is achieved in the case of 5-class Lung and 11-class Tumour and AUC for the rest of the datasets is 100%.

The performance of the proposed method is presented in terms of visualized accuracy as shown in Figure 5a–h for 2-class Leukaemia, 3-class Leukaemia, 5-class Lung, 3-class MLL, 2-class Ovarian, 2-class Prostate, 4-class SRBCT and 11-class Tumour cancer datasets, respectively.

The misclassification rate $E$ is another evaluation metric that measures the number of misclassified samples in the test set. The misclassification error on the test data is computed at a threshold of every 10% of the training set and an average of the 10 errors registered is calculated to get the final error of the method for each data set as presented in Table 3. Even though the classification of many of the datasets is 100%, there is still misclassification error that indicates the penalty of over-confidence during correct and wrong predictions.

The errors due to the proposed method are computed using the misclassification error which is the difference between the proposed and the actual class label of each sample in the test set for all the datasets. Figure 6a–h presents the misclassification error rate tumour, 3-class Leukaemia, Lung, MLL, Prostate, Ovarian, SRBCT and 2-class Leukaemia, respectively.

The confusion matrix is also another evaluation technique applied here to measure the performance of the proposed method to exhibit correctly and wrongly classified samples in the test data considering true positive and true negative as correctly classified samples, and false positive and false negative categories as wrongly classified samples. Experimental results of the proposed method in terms of confusion matrices are presented in Figure 8a–h for datasets 11-class Tumour, 2-class Leukaemia, 3-class Leukaemia, 5-class Lung, 3-class MLL, 2-class Ovarian, 2-class Prostate and 4-class SRBCT cancer datasets, respectively.

The performance of the proposed method is also measured in terms of ROC curve. The ROC curve shows the probability that a randomly selected positive example got a higher rank by the ensemble-based classifier with respect to randomly selected negative example [29]. To plot the ROC curve, the false positive rate is shown on the x-axis and the true positive rate is presented on the y-axis. ROC is used as an evaluation technique as it has a high degree of tolerance in classifying data with a low-class imbalance. Moreover, the value of the AUC of the proposed method is reported showing all the positive test sets as true positives, showing 1.00 in five datasets namely SRBCT, 2-class Leukaemia, 3-class Leukaemia, MLL and Ovarian Cancer. The proposed method scores an AUC of 0.97 and 0.98 for the prostate and tumour cancers are shown in Figure 8b,c, respectively. Moreover, the AUC for SRBCT, 2-class Leukaemia, 3-class Leukaemia, Lung and MLL are shown in Figure 8a–h, respectively.

To validate the performance of the proposed method, we have made a comparative study with some of the state-of-the-art methods. The comparative analysis on all of the datasets in terms of classification accuracy is presented in Table 4. Results of the proposed method with the highest performance are presented in boldface. The soft voting method shows better performance, scoring 100% in six of the datasets and the case of the hard voting method, 100% performance is achieved in five of the datasets, and promising results are achieved on the remaining datasets. From the comparative results, we observe that the proposed method exhibits better performance when compared to many of the state-of-the-art methods.
6 | DISCUSSION AND ANALYSIS OF EXPERIMENTAL RESULTS

Timely and accurate models and timely decision-making by the domain experts are essential in cancer treatment at its early stage. Predominantly, the gene expression cancer datasets come with challenges such as imbalanced class label, the higher curse of dimensionality, small samples and noisy data. To tackle the higher dimensionality in microarray cancer datasets, we explored the RF tree-based feature selection on eight binary and multiclass microarray cancer datasets. The top-ranked features selected based on the RF tree algorithm are presented in Figure 2a–2h for 2-class Leukaemia, 3-class Leukaemia, MLL, Ovarian, Prostate, SRBCT, Tumour and Lung cancers, respectively. In the classification phase, we have proposed to explore an ensemble learning-based using hard voting and soft voting techniques. Six heterogeneous classification algorithms are
considered in the ensemble learning namely LR, MLP, SVM, RF, simple DT and KNN. Each of these classifiers are trained and predict individually at the first layer of the ensemble learning. The final class label is obtained by combing the prediction of each classifier using a certain combining function in the second layer of the ensemble learning method. Two combining functions namely hard voting and soft voting approaches are used in the second layer to collect and combine the prediction of each classifier. The hard voting technique uses the majority voting method in predicting the final class label of a given sample. The experimental results of the proposed ensemble learning-based cancer prediction show that it yields accurate and the standard predictions in many of the datasets. The effectiveness of the proposed method is validated using eight standard microarray cancer datasets where three of the datasets are binary class and the remaining five datasets are multi-class datasets. The effect of class imbalance in microarray cancer data is challenging when the classification accuracy is much larger than the AP and this is reflected on some of the datasets such as the lung and tumour datasets.

FIGURE 7 Confusion matrix the proposed ensemble learning approach on eight different datasets.
Cancer is a group of related diseases with high mortality rate characterized by abnormal cell growth which attacks the body tissues. Although there are several research outputs in the area across many disciplines, results so far are not up to the level of diagnosis, identification and treatment requirement. The authors proposed an ensemble learning method to classify microarray cancer data using RF tree-based feature selection. To validate the method, eight different standard microarray cancer datasets are used. The experimental results show that the proposed method is highly accurate compared to the state of the art methods. The accuracy of the proposed method is enhanced due to the RF tree-based feature selection followed by ensemble the prediction of six base classifiers to come up with the strong prediction. From the experimental results, we have observed that soft voting is outperforming or equally performing when compared with the hard voting approach. To confirm the validity of the method, we employ multiple evaluation methods such as classification accuracy, precision, recall,

FIGURE 8 Receiver operating characteristic curve and area under the curve of proposed ensemble learning approach on eight different datasets

7 | CONCLUSION
| References                     | Leuk2 | Leuk3 | Lung | MLL | Ovarian | Prostate | SRBCT | Tumour |
|-------------------------------|-------|-------|------|-----|---------|----------|-------|--------|
| Mohapatra et al. [21]         | 0.99  | 0.74  | -    | -   | -       | -        | 0.99  | 0.71   |
| Jain et al. [30]              | 1.00  | 1.00  | -    | 1.00| 1.00    | -        | 1.00  | -      |
| Dash [31]                     | 0.99  | -     | -    | -   | -       | 0.98     | -     | -      |
| Tong et al. [32]              | 1.00  | -     | 0.95 | 1.00| -       | 0.93     | 1.00  | -      |
| Lv et al. [33]                | 1.00  | 1.00  | 0.96 | -   | -       | 0.96     | 0.95  | -      |
| Mollaee and Moattar [34]      | -     | 1.00  | -    | -   | -       | -        | 1.00  | -      |
| Chandra and Manish [35]       | 0.97  | -     | -    | 0.97| -       | 0.93     | -     | -      |
| Dash and Mista [36]           | 0.98  | -     | -    | -   | -       | 0.95     | -     | -      |
| Sharbaf et al. [37]           | 0.97  | -     | -    | 0.99| -       | 0.98     | -     | -      |
| Medjahed et al. [38]          | 0.96  | -     | -    | 0.96| -       | 0.98     | -     | -      |
| **Proposed (soft voting)**    | 1.00  | 1.00  | 0.96 | 1.00| 1.00    | 1.00     | 1.00  | 0.96   |
| **Proposed (hard voting)**    | 1.00  | 1.00  | 0.96 | 1.00| 1.00    | 1.00     | 1.00  | 0.86   |

f-measure, AP, ROC curve, misclassification error rate, and confusion matrix, and results are better than many of the state-of-the-art methods.

REFERENCES

1. Kumar, M., Rath, S.K.: Classification of microarray using MapReduce based proximal support vector machine classifier. Knowl-Based Syst. 89, 584–602 (2015)
2. Kar, S., Sharma, K.D., Maitra, M.: Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive K-nearest neighborhood technique. Expert Syst. Appl. 42(1), 612–627 (2015)
3. Lin, T.C., et al.: Pattern classification in DNA microarray data of multiple tumor types. Pattern Recogn. 39(12), 2426–2438 (2006)
4. Salem, H., Attiya, G., El-Fishawy, N.: Classification of human cancer diseases by gene expression profiles. Appl. Soft Comput. 50, 124–134 (2017)
5. Moayedkia, A., et al.: Feature selection for high dimensional imbalanced class data using harmony search. Eng. Appl. Artif. Intell. 57, 38–49 (2017)
6. Cerf, I., et al.: Parameter-free classification in multi-class imbalanced data sets. Data Knowl. Eng. 87, 109–129 (2013)
7. Piao, Y., Piao, M., Rya, K.H.: Multiclass cancer classification using a feature subset-based ensemble from microRNA expression profiles. Comput. Biol. Med. 80, 39–44 (2017)
8. Breiman, L.: Random forests. Mach. Learn. 45(1), 5–32 (2001)
9. Swamynathan, M.: Mastering Machine Learning with Python in Six Steps: A Practical Implementation Guide to Predictive Data Analytics Using Python. Apress (2019)
10. Ebrahimpour, M.K., Eftekhar, M.: Ensemble of feature selection methods: a hesitant fuzzy sets approach. Appl. Soft Comput. 50, 300–312 (2017)
11. Das, A.K., Das, S., Ghosh, A.: Ensemble feature selection using bi-objective genetic algorithm. Knowl-Based Syst. 123, 116–127 (2017 May 1)
12. Chen, K.H., Wang, KJ, Wang, KM, Angela, MA: Applying particle swarm optimization-based decision tree classifier for cancer classification on gene expression data. Appl. Soft Comput. 24, 773–780 (2014)
13. Nguyen, T., et al.: A novel aggregate gene selection method for microarray data classification. Pattern Recogn. Lett. 60, 16–23 (2015)
14. Farid, D.M., et al.: An adaptive rule-based classifier for mining biological data. Expert Syst. Appl. 64, 305–316 (2016)
15. Nematzadeh Z., Ibrahim R., Selamat A.: Comparative studies on breast cancer classifications with k-fold cross validations using machine learning techniques. In: 2015 10th Asian control conference (ASCC), pp. 1–6. IEEE (2015)
16. Liu, Z., et al.: A hybrid method based on ensemble WELM for handling multi class imbalance in cancer microarray data. Neurocomputing. 266, 641–650 (2017)
17. Lai, CM., Yeh, W.C., Chang, C.Y.: Gene selection using information gain and improved simplified swarm optimization. Neurocomputing. 218, 331–336 (2016)
18. Lu, H., et al.: A hybrid feature selection algorithm for gene expression data classification. Neurocomputing. 256, 56–62 (2017)
19. Ramos-González, J., et al.: A CBFR framework with gradient boosting based feature selection for lung cancer subtype classification. Comput. Biol. Med. 86, 98–106 (2017)
20. Tabakhi, S., et al.: Gene selection for microarray data classification using a novel ant colony optimization. Neurocomputing. 168, 1024–36 (2015)
21. Mohapatra, P., Chakravarty, S., Dash, P.K.: Microarray medical data classification using kernel ridge regression and modified cat swarm optimization based gene selection system. Swarm Evol. Comput. 28, 144–60 (2016)
22. Pedregosa, F., et al.: Scikit-learn: ‘Machine learning in Python’. J. Mach. Learn. Res. 12, 2825–2830 (2011)
23. Kawakubo, H., Yoshida, H.: Rapid feature selection based on random forests for high-dimensional data. Expert Syst. Appl. 40, 6241–6252 (2012)
24. Cutler, A.D., Richard, C., John, S.: Random forests. Ensemble machine learning, 157–175. Springer, Boston, MA (2012)
25. Weedon, M., Tsaptsinos, D., Denholm-Price, J.: Random forest explorations for URL classification. In: International Conference On Cyber Situational Awareness, Data Analytics and Assessment, pp. 1–4. Cyber SA, IEEE, 2017 (2017)
26. Zhu, Z., Ong, Y.S., Dash, M.: Markov blanket-embedded genetic algorithm for gene selection. Pattern Recogn. 40(11), 3236–48 (2007)
27. Statnikov I.T.A., Constantin, F: Aliferic Gene expression model selector. http://www-gems-system.org/
28. Cano S.M.A., Masegosa, A.: Elvira biomedical data set repository. http://lexugr.es/elvira/DBCRepository/
29. Arunkumar, C., Ramakrishnan, S.: Attribute selection using fuzzy rough set based customized similarity measure for lung cancer microarray gene expression data. Future Comput. Infor. J. 3(1), 131–142 (2018)

30. Jain, I., Jain, V.K., Jain, R.: Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification. Appl. Soft Comput. 62, 203–215 (2018)

31. Dash, R.: An adaptive harmony search approach for gene selection and classification of high dimensional medical data. J. King Saud Univ. Comput. Inform. Sci. (2018)

32. Tong, M., et al.: An ensemble of SVM classifiers based on gene pairs. Comput. Biol. Med. 43(6), 729–737 (2013)

33. Lv, J., et al.: A multi-objective heuristic algorithm for gene expression microarray data classification. Expert Syst. Appl. 59, 13–19 (2016)

34. Mollaee, M., Moattar, M.H.: A novel feature extraction approach based on ensemble feature selection and modified discriminant independent component analysis for microarray data classification. Biocybernet. Biomed. Eng. 36(3), 521–529 (2016)

35. Chandra, B., Gupta, M.: An efficient statistical feature selection approach for classification of gene expression data. J. Biomed. Informat. 44(4), 529–535 (2011)

36. Dash, R., Misra, B.: Gene selection and classification of microarray data: a Pareto DE approach. Intell. Decis. Technol. 11(1), 93–107 (2017)

37. Sharbat, F.V., Mosafer, S., Moattar, M.H.: A hybrid gene selection approach for microarray data classification using cellular learning automata and ant colony optimization. Genomics. 107(6), 231–238 (2016)

38. Medjahed, S.A., et al.: Kernel-based learning and feature selection analysis for cancer diagnosis. Appl. Soft Comput. 51, 39–48 (2017)

How to cite this article: Dagnew G, Shekar BH. Ensemble learning-based classification of microarray cancer data on tree-based features. Cogn. Comput. Syst. 2021;3:48–60. https://doi.org/10.1049/ccs2.12003