J Inf Process Syst, Vol.19, No.5, pp.652~662, October 2023
https://doi.org/10.3745/JIPS.04.0290

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Ensemble Gene Selection Method Based on Multiple Tree Models

Mingzhu Lou*

Abstract

Identifying highly discriminating genes is a critical step in tumor recognition tasks based on microarray gene expression profile data and machine learning. Gene selection based on tree models has been the subject of several studies. However, these methods are based on a single-tree model, often not robust to ultra-high-dimensional microarray datasets, resulting in the loss of useful information and unsatisfactory classification accuracy. Motivated by the limitations of single-tree-based gene selection, in this study, ensemble gene selection methods based on multiple-tree models were studied to improve the classification performance of tumor identification. Specifically, we selected the three most representative tree models: ID3, random forest, and gradient boosting decision tree. Each tree model selects top-n genes from the microarray dataset based on its intrinsic mechanism. Subsequently, three ensemble gene selection methods were investigated, namely multiple-tree model intersection, multiple-tree module union, and multiple-tree module cross-union, were investigated. Experimental results on five benchmark public microarray gene expression datasets proved that the multiple tree module union is significantly superior to gene selection based on a single tree model and other competitive gene selection methods in classification accuracy.

Keywords

Ensemble Tree Model, Gradient Boosting Decision Tree, Gene Selection, ID3, Random Forest

1. Introduction

The study of tumor genes based on microarray data has always been the focus of current cancer research microarray gene expression profile data extracted using biochip technology can be used to study the causes of tumors. By analyzing of microarray gene expression profile data, we can explore which tumor genes are related to cancer development and develop specific prevention mechanisms and treatment methods. However, the high-dimensional characteristics of gene expression profile data and many redundant and noisy genes pose significant challenges to gene data analysis. Therefore, selecting genes most associated with cancer, based on microarray gene expression data, is the key to improving tumor classification performance. Recently, many methods for gene selection for tumors have been proposed. Hybrid gene selection methods have attracted increasing attention recently.

Because the training process of tree modeling has the advantage of built-in feature selection processing, which can prevent overfitting, gene selection methods based on tree models have gained attention and

Manuscript received September 23, 2022; first revision March 15, 2023; accepted April 8, 2023.
* **Corresponding Author:** Mingzhu Lou (minzhulou@163.com)
School of Information and Engineering, Nanchang Institute of Technology, Nanchang, China (minzhulou@163.com)

research. For example, Rao et al. [1] proposed a hybrid feature-selection method based on bee colonies and gradient boosting trees. Horng et al. [2] proposed resampling to avoid overfitting and used decision rules in decision trees to select a set of reliable genes. Xiong and Wang [3] proposed a hybrid gene selection algorithm based on improved ant colony optimization and a random forest. Dagnew and Shekar [4] proposed a feature selection algorithm based on random forest trees and an integrated classification method based on hard and soft voting. Recently, Deng et al. [5] proposed a gene selection method based on extreme gradient boosting (XGBoost) [6] and a multi-objective genetic algorithm.

The afore-mentioned methods use a single-tree model for gene selection to improve tumor classification performance. However, gene selection based on a single-tree model often exhibits poor robustness for super-high-dimensional and small-sample microarray datasets, leading to the loss of useful information, and the final classification result is not ideal. The fundamental reason for this is the bias of the single-tree model learning algorithm. A good classification result of a tree model for one microarray dataset does not indicate good classification performance for other microarray datasets.

Motivated by the afore-mentioned limitations, in this study, ensemble gene selection methods based on multiple-tree models were investigated to improve the classification performance of tumor identification based on microarray gene data. We selected the three most representative tree models: ID3 [7], random forest [8], and gradient boosting decision tree (GBDT) [9]. Each tree model selects the top-n genes from the microarray dataset based on its intrinsic mechanism. The final subset of genes was selected using the intersection, union, and cross-union operations. The ensemble gene selection algorithms corresponding to the three different operations are respectively named multiple-tree model intersection (MTMI), multiple-tree module union (MTMU), and multiple-tree module cross-union (MTMCU). The effectiveness of the proposed gene selection method was evaluated using support vector machine (SVM) [10] and naïve Bayes [11] on five public microarray gene expression datasets. The experimental results demonstrated that MTMU is significantly superior to gene selection based on a single-tree model and other competitive gene selection algorithms in terms of classification accuracy.

The remainder of this study is organized as follows. Section 2 provides preliminary information on the techniques used in our study. In Section 3, we present three methods for gene selection based on integrated-tree models. Section 4 presents the experimental results and an analysis of the proposed algorithms, and other related methods, on five public benchmark gene expression datasets. Finally, Section 5 presents the conclusions and future work.

2. Related Works

2.1 ID3 Decision Tree

A decision tree is a tree structure comprising multiple branch nodes and branches. The decision tree algorithm is an important algorithm in machine learning classification methods. ID3 [7], C4.5 [12], and the classification and regression tree (CRAT) [13] are the most commonly used decision tree algorithms. The ID3 algorithm is a classical decision tree algorithm. Many other decision tree algorithms can be generated based on the ID3 algorithm through extension and improvement.

The ID3 algorithm was proposed by Quinlan [7] in 1986, and its basic theory is information entropy. The ID3 algorithm tree judges the attribute-based information gain of each internal node, divides the

criteria according to judgment, and outputs the judgment results for each branch.

The core of the ID3 algorithm uses information entropy to determine the attributes used for node segmentation. Entropy is defined as follows:

$$E(X) = -\sum_{i=1}^{C} P(x_i) * log_2\left(P(x_i)\right), \tag{1}$$

where $P(x_i)$ is probability of the value of x_i for feature X, and C denotes the number of categories. ID3 is determined based on the information gain and the calculation method is as follows:

$$IG(X,Y) = E(X) - E(X|Y), \tag{2}$$

where the $E(X)$ is the entropy of feature X, and $E(X|Y)$ is the entropy of feature X under the condition of feature Y, which is defined as

$$E(X|Y) = \sum_{i=1}^{C_y} P(y_i) \sum_{j=1}^{c} P(x_j|y_i) log_2\left(P(x_j|y_i)\right). \tag{3}$$

2.2 Random Forest

Random forest [8] is a classical ensemble learning algorithm that processes sample data by combining multiple decision trees to form an ensemble classifier. In 2001, Breiman [8] proposed a random forest algorithm in *Machine Learning*, which uses the CART tree as the base learner to solve the overfitting problem that may occur during classification and regression. Fig. 1 shows the steps of random forest training sample data.

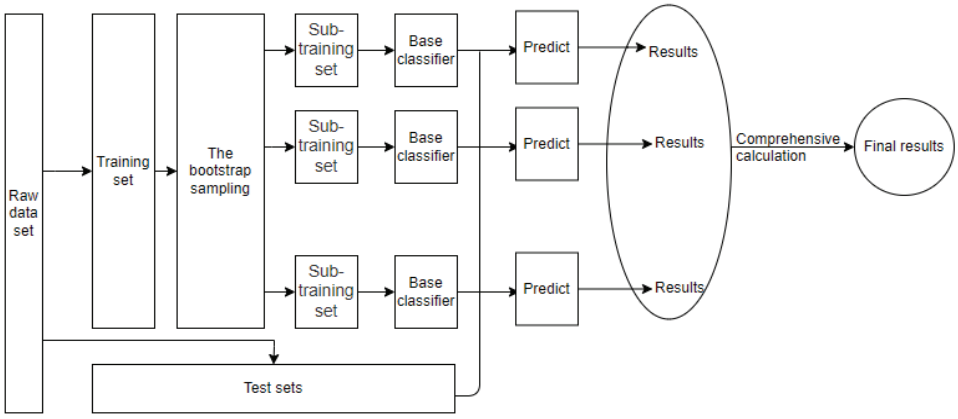


Fig. 1. Training process of random forest.

The random forest method scores the importance of each feature of the dataset and then selects the important features based on the score. The Gini coefficient was used as the evaluation index, calculated as follows:

$$Gini(D) = 1 - \sum_{i=1}^{n} p_i^2, \tag{4}$$

where D is the sample set and p_i is the probability of occurrence of each category.

2.3 Gradient Boosting Decision Tree

The GBDT [9], also known as the multiple additive regression tree (MART), is an iterative machine learning algorithm [9,14]. The gradient boosting decision tree uses the residual gradient to optimize the regression tree and combines several weak learners into strong learner in a certain manner. The GBDT is a boosting model and its training process is shown in Fig. 2.

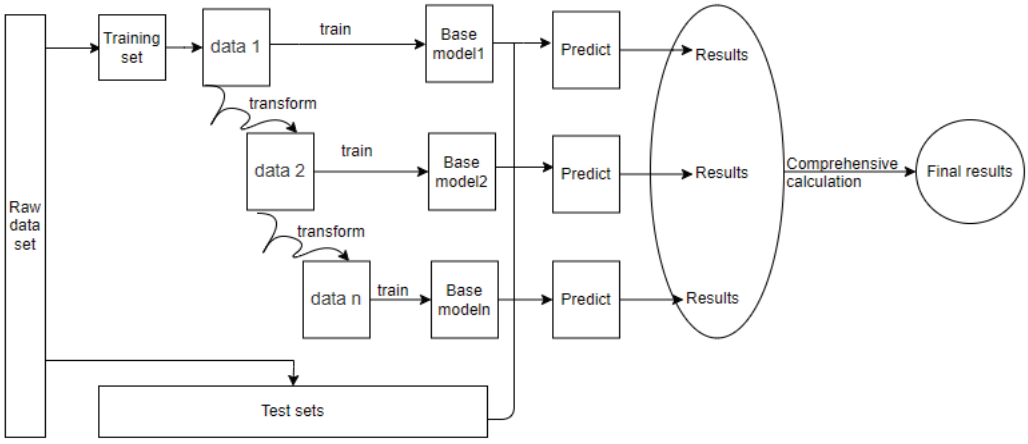


Fig. 2. Training process of GBDT.

The principle of GBDT is to optimize the residual so that the residual after each iteration is gradually reduced. After several iterations of GBDT, the output results are not different from the actual results, the GBD training model reaches the optimal value, and the loss function value reaches the minimum or remains unchanged. GBDT algorithm builds a model in the optimal descent direction based on the loss function, which is embodied as follows.

Assuming $\{x_i, y_i\}_{i=1}^{n}$ is the original data, softmax is the loss function of GDBT, $h(x)$ is the base learner, where $x_i = (x_{1i}, x_{2i,}, \ldots, x_{pi})$, p is the number of features, and y_i is the prediction label. For model β, the initialization estimate is

$$F_0(x) = arg\ \min_{\beta} \sum_{i=1}^{n} L(y_i, \beta). \tag{5}$$

Next, the gradient direction of residuals is calculated through iterations:

$$y_i^* = -[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}]_{F(x)-F_{m-1}(x)}, \tag{6}$$

where $i = \{1,2,\ldots N\}, m = 1{:}\,M$. The base learner was then used to train the sampled data and obtain the initialization model. According to the least-squares method, the parameter a_m of the model can be obtained as:

$$a_m = arg\ \min_{\alpha,\beta} \sum_{i=1}^{N} [y_i^* - \beta h(x_i; a)]^2. \tag{7}$$

To minimize the loss function, a new step size of the model was calculated using the formula:

$$\beta_m = arg\min_{\alpha,\beta} \sum_{i=1}^{N} L\big(y_i, F_{m-1}(x) + \beta h(x_i; a)\big). \tag{8}$$

Finally, the new model is updated according to Formula (9):

$$F_m(x) = F_{m-1}(x) + \beta_m h(x_i; a). \tag{9}$$

3. Integrated Gene Selection Method based on Tree Model

Feature selection is at the core of the tree model construction. Node generation and branch processing of the tree model preferentially select features that can better distinguish the training dataset. This improves the training efficiency of the decision tree, and accelerates its convergence. However, selecting the optimal features in the tree model construction process is an open problem. There is uncertainty regarding the selection of gene features, specifically for high-dimensional microarray datasets. Therefore, we adopted an integration strategy to eliminate the inaccuracy caused by a single-tree model in selecting the optimal gene by constructing multiple tree models. In contrast to the single-gene selection tree model, the integrated gene selection tree model based on an integration strategy can select the optimal feature subset with strong robustness. This study used the three most representative tree models for gene selection: decision tree, random forest, and GBDT. Fig. 3 demonstrates the gene selection based on the three tree models' intersection, union, and intersection-union operations. Each tree model calculates the importance score of each gene according to its criteria and ranks these genes in descending order. The top-n genes were then selected for each tree model, and three final gene subsets were obtained using intersection, union, and intersection-union operations. The ensemble gene selection algorithms corresponding to the three different operations are respectively, called MTMI, MTMU, and MTMCU. The formal definitions of the three methods are expressed by formulas (10)–(12), where M_{DT}^n, M_{RF}^n, and M_{GBDT}^n represent the sets of top-n genes selected by each method.

$$MTMI = M_{DT}^n \cap M_{RF}^n \cap M_{GBDT}^n, \tag{10}$$

$$MTMU = M_{DT}^n \cup M_{RF}^n \cup M_{GBDT}^n, \tag{11}$$

$$MTMIU = \{M_{DT}^n \cap M_{RF}^n\} \cup \{M_{DT}^n \cap M_{GBDT}^n\} \cup \{M_{RF}^n \cap M_{GBDT}^n\}. \tag{12}$$

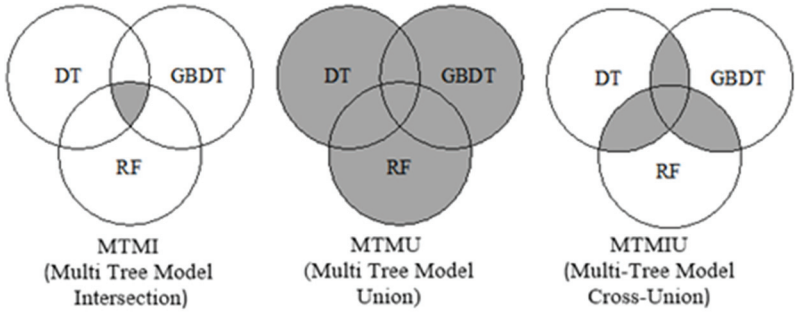


Fig. 3. Three types of ensemble gene selection based on multiple tree models.

4. Experimental Results and Analysis

4.1 Datasets

Five publicly available microarray gene-expression profile datasets were used to evaluate the performance of the proposed algorithm. The basic information on these datasets is listed in Table 1 [15-19]. From Table 1, ultra-high dimensionality and a small number of samples are common characteristics of these datasets.

Table 1. Microarray dataset

Dataset	Number of genes	Number of samples	Number of classes
HD [15]	22,283	31	2
Lymphoma1 [16]	4,026	66	2
Lymphoma_3 [17]	4,026	66	3
Lung_5 [18]	12,600	203	5
Ovarian [19]	15,154	253	2

4.2 Compared Algorithms

The experiments were conducted in two stages. First, three related tree models, the decision tree, random forest, and GBDT algorithms, were selected for comparison. Secondly, we selected some of the most advanced gene selection methods published recently, such as BGWOPSO [20], FCSVM-RFE [21], KDI [22], and MultiSURF [23], for experimental comparison.

To verify the effectiveness of the proposed integrated-tree model gene selection method, we compared it with three single-tree models: decision tree, random forest, and GBDT. These algorithms were used to select the top-n genes (5, 10, 20, 50, 100). Two standard classifiers, SVM and naïve Bayes, were used to evaluate the classification performance of selected gene subsets. To ensure the reliability of the experimental results, a 10-fold cross-validation method was used to evaluate the accuracy of each selected gene subset. The classification algorithms used in this study were implemented from the Python sklearn library.

4.3 Comparison with Single Tree Model-based Methods

Fig. 4 shows the accuracy comparison of MTMI, MTMU, and MTMCU, and the other three single-tree models using SVM and naïve Bayes classifiers for the five microarray datasets listed in Table 1. The original method is results obtained without gene selection. The left column in Fig. 4 shows the results of the SVM classification after selecting the top 5, 10, 20, 50, and 100 genes using all comparison algorithms. Accordingly, the right column shows the results of the naïve Bayes classification after selecting the top 5, 10, 20, 50, and 100 genes using all comparison algorithms.

As shown in Fig. 4, if the selected top number of genes was less than 100, there was no intersection of the genes selected by the decision tree, random forest, and GBDT in the HD dataset. This shows that the sorting results of the gene scores obtained by the three single-tree model algorithms differed significantly.

Table 2. Average accuracy of top-*n* genes on five datasets

	SVM					Naïve Bayes				
	Top-5	**Top-10**	**Top-20**	**Top-50**	**Top-100**	**Top-5**	**Top-10**	**Top-20**	**Top-50**	**Top-100**
Original	0.9598	0.9598	0.9598	0.9598	0.9598	0.9183	0.9183	0.9183	0.9183	0.9183
MTMI	0.8963	0.9147	0.8963	0.9409	0.9382	0.8809	0.8993	0.8963	0.9185	0.9382
MTMU	**0.9891**	**0.9871**	**0.9898**	**0.9921**	0.9772	**0.9842**	**0.9848**	**0.9898**	**0.9883**	0.9772
MTMCU	0.9492	0.9549	0.9404	0.9773	0.9565	0.9364	0.9354	0.9404	0.9658	0.9565
DT	0.9625	0.9233	0.9162	0.9241	0.8406	0.9431	0.9374	0.9162	0.8682	0.8406
GBDT	0.9783	0.9819	0.9751	0.9889	0.9609	0.9712	0.9792	0.9751	0.9762	0.9609
RF	0.9811	0.9834	0.9810	0.9882	**0.9879**	0.9685	0.9781	0.9810	0.9840	**0.9879**

The best results are highlighted in bold.

Table 2 lists the average classification accuracy of the top-*n* (5, 10, 20, 50, 100) genes selected by each method on five datasets. As observed from Table 2, for the SVM classifier, the MTMU algorithm is superior to other algorithms if top-5, top-10, top-20, and top-50 genes are selected respectively, and only slightly worse than RF if top-100 genes are selected. The hat MTMU algorithm achieves optimal average classification accuracy if top-50 genes are selected, which is higher than any average classification accuracy of other algorithms. Table 2 also shows that MTMI obtains the worst result because the Intersection cannot be formed when the number of genes selected by a single tree model is less than 100, thus reducing the average classification accuracy. In addition, similar comparison results can be obtained for naïve Bayes classifiers.

According to Fig. 4 and Table 2, the MTMU gene selection method was more effective than the single-tree model-based gene selection method.

4.4 Comparison with State-of-the-Art Methods

In this section, four of the most advanced gene selection methods published recently, namely BGWOPSO [20], FCSVM-RFE [21], KDI [22], and MultiSURF [23], were selected for experimental comparison. For a fair comparison, the size of the gene subsets selected by all compared algorithms was set to 50.

Tables 3 and 4 compares the classification accuracy between the proposed ensemble tree model gene selection methods and the four latest proposed methods using SVM and naïve Bayes classifiers, respectively.

Because no intersection of genes selected by DT, RF, and GBDT in the HD dataset when the top 50 genes were selected, the MTMI algorithm obtained no genes. As observed from Tables 3 and 4, the MTMU integrated tree model method achieved better results overall. The proportions of optimal values obtained by the MTMU on the SVM and naïve Bayes classifiers were 5/5 and 4/5, respectively.

The last columns in Tables 3 and 4 show each algorithm's mean classification accuracy for the five datasets. In the case of SVM, MTMU obtained a first mean classification accuracy of 0.9921 and MTMCU obtained a second classification accuracy of 0.9773. Similarly, using the naïve Bayes classifier, MTMU obtained the first mean classification accuracy of 0.9883, and MTMCU obtained the second mean classification accuracy of 0.9658.

From the above results, the integrated tree model gene selection method (MTMU) proposed in this study has an advantage in terms of classification accuracy compared to other methods. We believe that this is because the MTMU algorithm integrates multiple genes selected from different single-tree models, thus avoiding the bias of a single model and enhancing the robustness of gene selection.

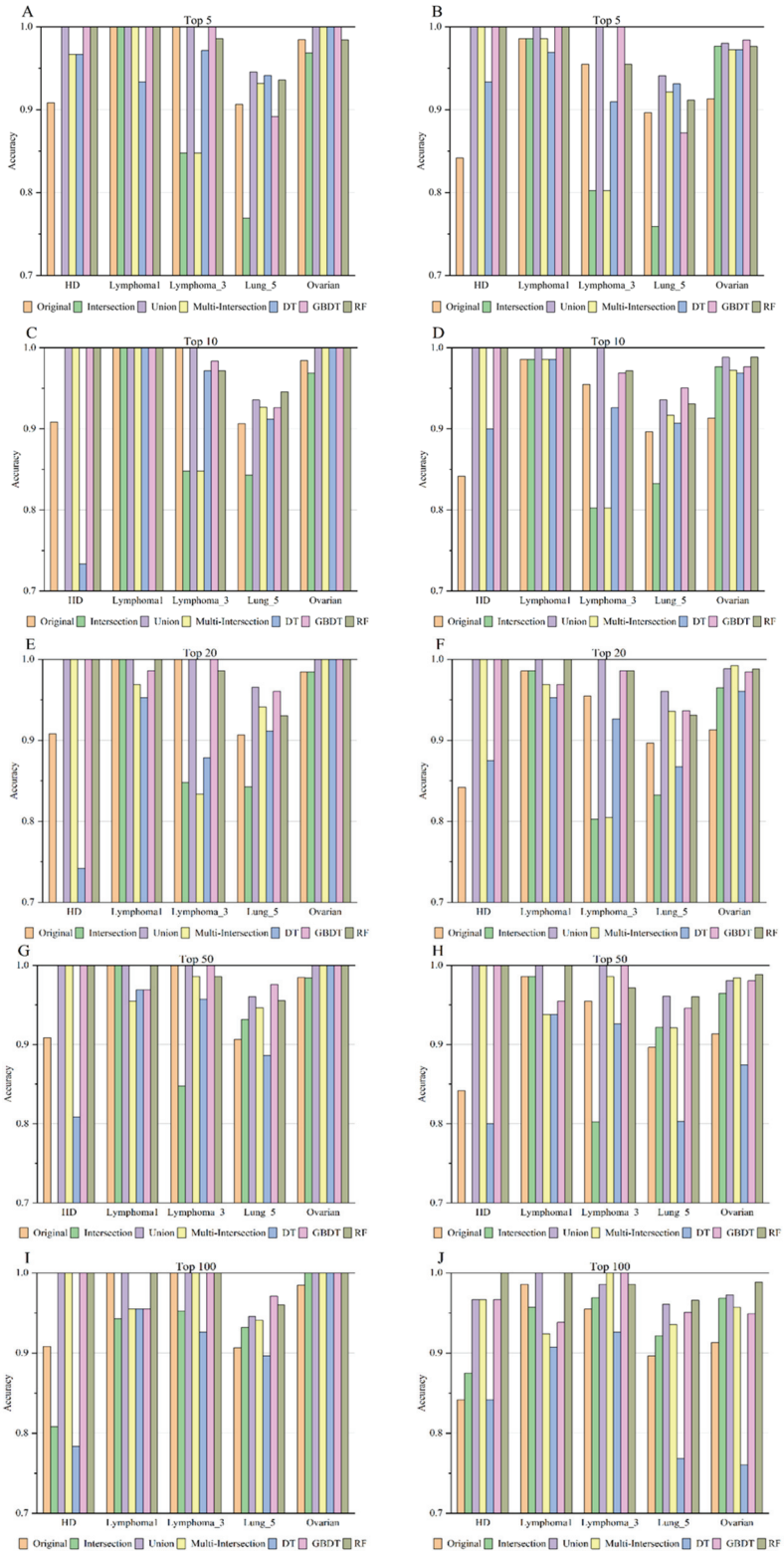


Fig. 4. Accuracy comparison based on selecting the top-*n* genes.

Table 3. Results of all compared methods based on the SVM classifier

	HD	Lymphoma1	Lymphoma_3	Lung_5	Ovarian	Mean
MTMI	-	1.0000	0.8476	0.9317	0.9842	0.9409
MTMU	1.0000	1.0000	1.0000	0.9605	1.0000	0.9921
MTMCU	1.0000	0.9548	0.9857	0.9462	1.0000	0.9773
BGWOPSO	0.9083	1.0000	1.0000	0.9014	0.9845	0.9588
FCSVM-RFE	0.9333	1.0000	1.0000	0.9114	0.9843	0.9658
KDI	0.7750	0.9667	0.9857	0.8226	0.9685	0.9037
MultiSURF	1.0000	1.0000	0.9690	0.9064	1.0000	0.9751

Table 4. Results of all compared methods based on naïve Bayes classifier

	HD	Lymphoma1	Lymphoma_3	Lung_5	Ovarian	Mean
MTMI	-	0.9857	0.8024	0.9214	0.9646	0.9185
MTMU	1.0000	1.0000	1.0000	0.9610	0.9805	0.9883
MTMCU	1.0000	0.9381	0.9857	0.9210	0.9843	0.9658
BGWOPSO	0.8417	0.9857	0.9571	0.8764	0.9131	0.9148
FCSVM-RFE	0.9000	0.9833	0.9857	0.8867	0.9214	0.9354
KDI	0.7750	0.9357	0.9548	0.7733	0.9055	0.8689
MultiSURF	1.0000	1.0000	0.9548	0.8674	0.9842	0.9613

5. Conclusion

Gene selection based on tree models is both simple and efficient. This study analyzes the shortcomings of gene selection based on a single tree model and proposes ensemble gene selection methods based on multiple tree models, which solves the bias of the gene selection algorithm based on a single tree model and improves the robustness and stability of the gene selection algorithm. Specifically, we used three single tree models, ID3, random forest, and GBDT, to select the top-n genes. Subsequently, we conducted intersection, union, and cross-union operations to obtain three gene subsets. The ensemble gene selection algorithms corresponding to the three operations were named MTMI, MTMU, and MTMCU. Experimental results on five publicly available microarray gene expression datasets demonstrate that MTMU is significantly superior to other methods in classification accuracy. Further studies may involve the application of other tree models and more complex fusion mechanisms to improve the performance of the gene selection algorithm based on the integrated-tree model.

Acknowledgement

This work was partially supported by the funds from Jiangxi Education Department, PR China (No. GJJ211919), the grants from and National Nature Science Foundation of China (No. 62166028).

References

[1] H. Rao, X. Shi, A. K. Rodrigue, J. Feng, Y. Xia, M. Elhoseny, X. Yuan, and L. Gu, "Feature selection based on artificial bee colony and gradient boosting decision tree," *Applied Soft Computing*, vol. 74, pp. 634-642, 2019. https://doi.org/10.1016/j.asoc.2018.10.036

[2] J. T. Horng, L. C. Wu, B. J. Liu, J. L. Kuo, W. H. Kuo, and J. J. Zhang, "An expert system to classify microarray gene expression data using gene selection by decision tree," *Expert Systems with Applications*, vol. 36, no. 5, pp. 9072-9081, 2009. https://doi.org/10.1016/j.eswa.2008.12.037

[3] W, Xiong and C. Wang, "A hybrid improved ant colony optimization and random forests feature selection method for microarray data," in *Proceedings of 2009 5th International Joint Conference on INC, IMS and IDC*, Seoul, South Korea, 2019, pp. 559-563. https://doi.org/10.1109/NCM.2009.66

[4] G. Dagnew and B. H. Shekar, "Ensemble learning-based classification of microarray cancer data on tree-based features," *Cognitive Computation and Systems*, vol. 3, no. 1, pp. 48-60, 2021. https://doi.org/10.1049/ccs2.12003

[5] X. Deng, M. Li, S. Deng, and L. Wang, "Hybrid gene selection approach using XGBoost and multi-objective genetic algorithm for cancer classification," *Medical & Biological Engineering & Computing*, vol. 60, no. 3, pp. 663-681, 2022. https://doi.org/10.1007/s11517-021-02476-x

[6] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, 2016, pp. 785-794. https://doi.org/10.1145/2939672.2939785

[7] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, pp. 81-106, 1986. https://doi.org/10.1023/A:1022643204877

[8] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5-32, 2001. https://doi.org/10.1023/A:1010933404324

[9] J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics & Data Analysis*, vol. 38, no. 4, pp. 367-378, 2002. https://doi.org/10.1016/S0167-9473(01)00065-2

[10] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, article no. 27, 2011. https://doi.org/10.1145/1961189.1961199

[11] G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," in *Proceedings of the 11th Annual Conference on Uncertainty in Artificial Intelligence*, Montreal, Canada, 1995, pp. 338-345.

[12] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann Publishers, 2014.

[13] A. D. Gordon, "Review of Classification and Regression Trees by L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone, editors," *Biometrics*, vol. 40, no. 3, pp. 874-874, 1984. https://doi.org/10.2307/2530946

[14] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189-1232, 2001.

[15] F. Borovecki, L. Lovrecic, J. Zhou, H. Jeong, F. Then, H. D. Rosas, et al., "Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease," *Proceedings of the National Academy of Sciences*, vol. 102, no. 31, pp. 11023-11028, 2005. https://doi.org/10.1073/pnas.0504921102

[16] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, et al., "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, no. 6769, pp. 503-511, 2000. https://doi.org/10.1038/35000501

[17] T. Li, C. Zhang, and M. Ogihara, "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression," *Bioinformatics*, vol. 20, no. 15, pp. 2429-2437, 2004. https://doi.org/10.1093/bioinformatics/bth267

[18] A. Bhattacharjee, W. G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, et al., "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses," *Proceedings of the National Academy of Sciences*, vol. 98, no. 24, pp. 13790-13795, 2001. https://doi.org/10.1073/pnas.191502998

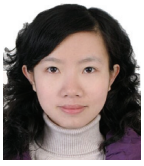
[19] Z. Zhu, Y. S. Ong, and M, Dash, "Markov blanket-embedded genetic algorithm for gene selection," *Pattern Recognition*, vol. 40, no. 11, pp. 3236-3248, 2007. https://doi.org/10.1016/j.patcog.2007.02.007

[20] Q. Al-Tashi, S. J. A. Kadir, H. M. Rais, S. Mirjalili, and H. Alhussian, "Binary optimization using hybrid grey wolf optimization for feature selection," *IEEE Access*, vol. 7, pp. 39496-39508, 2019. https://doi.org/10.1109/ACCESS.2019.2906757

[21] X. Huang, L. Zhang, B. Wang, F. Li, and Z. Zhang, "Feature clustering based support vector machine recursive feature elimination for gene selection," *Applied Intelligence*, vol. 48, pp. 594-607, 2018. https://doi.org/10.1007/s10489-017-0992-2

[22] Z. Hou and S. Y. Kung, "A kernel discriminant information approach to non-linear feature selection," in *Proceedings of 2019 International Joint Conference on Neural Networks (IJCNN)*, Budapest, Hungary, 2019, pp. 1-10. https://doi.org/10.1109/IJCNN.2019.8852186

[23] R. J. Urbanowicz, R. S. Olson, P. Schmitt, M. Meeker, and J. H. Moore, "Benchmarking relief-based feature selection methods for bioinformatics data mining," *Journal of Biomedical Informatics*, vol. 85, pp. 168-188, 2018. https://doi.org/10.1016/j.jbi.2018.07.015



Mingzhu Lou **https://orcid.org/0000-0001-9657-6300**

She graduated from Jiangxi Normal University majoring in Computer Science and Technology in 2003, and obtained a master's degree in computer applications technology in 2012 from Nanchang University. She is now a lecturer at the school of Information Engineering, Nanchang Institute of Technology, Nanchang, China. She presided over one science and technology project of Jiangxi Provincial Education Department and participated in four national and provincial scientific research projects.