ENSEMBLE LEARNING APPROACHES FOR CLASSIFICATION WITH HIGH-DIMENSIONAL DATA

Cao Truong Tran[1*]

DOI: 10.56651/lqdtu.jst.v12.n1.659.ict

Abstract

Classification with high-dimensional data is a significant challenge in machine learning because the abundance of features in high-dimensional data makes it difficult to identify meaningful patterns, which leads to overfitting and reduced classification performance. Moreover, the computational cost of processing high-dimensional data is often prohibitively expensive, requiring specialized hardware or optimized algorithms. Ensemble learning is a powerful machine learning technique that combines multiple models to improve classification accuracy. By aggregating the predictions of multiple models, ensemble learning can reduce overfitting, increase robustness, and improve performance on a wide range of real-world classification problems. Ensemble learning is effective for classification with high-dimensional data because it can combine multiple models to mitigate the effects of the curse of dimensionality, reduce overfitting, and enhance generalization performance. By using different learning algorithms or subsets of features, ensemble learning can improve the diversity of the models, leading to better overall performance on high-dimensional data. This paper proposes two hybrid ensemble machine learning approaches that integrate random subspace ensemble with bagging and boosting to enhance classification performance with high-dimensional data. Experimental results demonstrate that these methods significantly improve classification accuracy with high-dimensional data.

Index terms

Classification, ensemble learning, high-dimensional data, RSE, bagging, boosting.

1. Introduction

Classification is a machine-learning technique that involves grouping data into different categories or classes based on their similarities and differences. It is a supervised learning approach that requires labeled training data to build a predictive model that can classify unseen instances [1]. The main objective of classification is to accurately predict the class label of a new instance based on the information available in the training data. To achieve this, various classification algorithms such as decision trees,

[1]Institute of Information and Communication Technology, Le Quy Don Technical University
[*]Corresponding author email: truongct@lqdtu.edu.vn

support vector machines, and artificial neural networks have been developed over the years. The choice of algorithm depends on the nature of the problem and the properties of the data. Classification is widely used in a variety of applications, such as image recognition, spam filtering, and sentiment analysis [1], [2].

Classification of high-dimensional data is a complex task in machine learning due to various challenges that arise when dealing with large feature spaces [3]. One of the main difficulties is the curse of dimensionality, which refers to the exponential increase in the number of possible configurations of data points as the dimensionality grows. This makes it challenging to accurately estimate the underlying probability distribution of the data and may lead to overfitting [4], [5]. In addition, high-dimensional data often suffer from sparsity, where most of the features have a value of zero, and noise, which can obscure the underlying patterns in the data. Another issue is the presence of irrelevant and redundant features, which can negatively impact the performance of classification models by introducing noise and increasing the computational complexity [6].

Several solutions have been proposed to address these challenges of classification with high-dimensional [6]. One solution is feature selection, which aims to identify the most relevant features while discarding irrelevant and redundant ones [3], [5]. Another solution is feature construction, which transforms the high-dimensional data into a lower-dimensional space while preserving the most informative features [4]. Ensemble learning is another solution that combines multiple base classifiers to improve the overall accuracy and robustness of the classification model [7]. The choice of the appropriate solution depends on the specific characteristics of the data and the requirements of the classification task [6].

Ensemble learning is a powerful technique that aims to improve the accuracy and robustness of machine learning models by combining the predictions of multiple base models. It is particularly effective when the individual models have different strengths and weaknesses and can complement each other [8]–[10]. Ensemble methods can be classified into two main categories: bagging and boosting [11], [12]. Bagging generates multiple subsets of the training data and trains a base model on each subset. The final prediction is obtained by averaging the predictions of the base models [11]. Boosting, on the other hand, trains a series of base models iteratively, giving more weight to the instances that are misclassified in the previous iteration [12]. Ensemble learning has been shown to significantly improve the performance of machine learning models in various domains, including image classification, natural language processing, and bioinformatics [8]–[10].

The RSE approach is a popular ensemble learning technique that aims to address the challenges of high-dimensional data by training multiple base models on random subsets of the feature space [13]. This method can enhance the accuracy and robustness of machine learning models and is particularly useful for large feature spaces that have irrelevant or redundant features. By randomly selecting subsets of the feature space for each base model, this approach can effectively reduce the dimensionality of the problem and improve the model's generalization performance [14]. RSE has demonstrated its

effectiveness in various classification tasks such as image and text classification, and has found application in many real-world scenarios. However, it is essential to carefully select the model's hyperparameters, including the number of base models and the size of the feature subset, to achieve optimal performance [13], [14].

A hybrid ensemble method is a machine learning technique that combines multiple types of ensemble methods to leverage individual strengths and overcome their weaknesses. While different ensemble methods can be effective in addressing different sources of error, no single method is universally best for all problems. Hybrid ensemble methods can help to overcome this limitation by combining the strengths of different methods, resulting in a more comprehensive and accurate model. Studies in [15]–[17] have shown that bagging, when combined with RSE, provides better results than using bagging or RSE alone. When combined with RSE, studies in [18], [19] have also demonstrated that boosting produces superior outcomes compared to using boosting or RSE in isolation. However, the systematic study of combining RSE with bagging or boosting for classification with high-dimensional data has not been explored. Therefore, this paper proposes and evaluates the effectiveness of combining RSE with bagging and boosting for classification with high-dimensional data.

1.1. Goals

In this paper, our primary objective is to introduce hybrid ensemble machine learning techniques designed to classify high-dimensional data. Our focus is to explore and provide solutions to the following key questions:

1) In what ways can we effectively combine bagging with RSE to classify high-dimensional data?
2) In what ways can we effectively combine boosting with RSE to classify high-dimensional data?
3) Can the integration of RSE with bagging enhance the accuracy of high-dimensional data classification?
4) Can the integration of RSE with boosting enhance the accuracy of high-dimensional data classification?

1.2. Organisation

The paper is structured as follows: Section 2 provides an overview of related work, while Section 3 outlines the proposed method. Section 4 details the experimental design, the results and the discussion. Finally, Section 5 concludes the paper and highlights future work.

2. Related work

This section presents related work including ensemble learning, RSE, and classification with high-dimensional data.

2.1. Ensemble learning

Ensemble learning is a popular machine learning technique that involves combining multiple individual models to create a stronger and more accurate prediction model [8]. Ensemble learning works by combining the predictions of several different models, each trained on different subsets of the data or with different algorithms. The idea behind ensemble learning is that by combining the predictions of multiple models, the overall accuracy and robustness of the prediction model can be improved [9]. Ensemble learning can be used with a wide range of machine learning algorithms, including decision trees, neural networks, and support vector machines. One of the key benefits of ensemble learning is that it can help to reduce overfitting, which is a common problem in machine learning where a model performs well on the training data but poorly on new data [9], [10].

Bagging, or Bootstrap Aggregating, is a popular ensemble learning technique in machine learning that involves generating multiple models using subsets of the training data [11]. Bagging works by randomly selecting subsets of the training data with replacement, and training a separate model on each subset. The individual models are then combined using a simple averaging technique, known as the ensemble. The key idea behind bagging is that by creating multiple models that are diverse and have been trained on different subsets of the data, the overall variance and instability of the prediction model can be reduced [11]. Bagging can be used with a wide range of machine learning algorithms, including decision trees, random forests, and neural networks [20]. One of the key benefits of bagging is that it can help to reduce overfitting, a common problem in machine learning where a model performs well on the training data but poorly on new data [9]–[11].

Boosting is another popular ensemble learning technique in machine learning that involves combining multiple weak models to create a stronger and more accurate prediction model [12]. Unlike bagging, boosting works by iteratively adjusting the weights of the training examples to focus on the examples that were previously misclassified. In each iteration, a new weak model is trained on the updated weights of the training data, and added to the ensemble. The key idea behind boosting is that by focusing on the examples that are difficult to classify, the overall accuracy of the prediction model can be improved [21]. Boosting can be used with a wide range of machine learning algorithms, including decision trees, neural networks, and support vector machines. One of the key benefits of boosting is that it can help to reduce bias, a common problem in machine learning where a model is too simplistic and fails to capture the complexity of the data [9], [12].

2.2. Random subspace ensemble

RSE is an ensemble learning technique that combines bagging and feature selection to improve the performance of machine learning models [13]. The main idea behind RSE is to randomly select subsets of features from the original dataset and use them to train individual models. The models are then combined using an averaging technique

to produce the final prediction. By using subsets of features, RSE can reduce the dimensionality of the data and improve the accuracy of the models, particularly for high-dimensional datasets [13], [22]. Algorithm 1 shows the main steps of RSM algorithm.

Algorithm 1: The training process of RSM

Input:
D, the original training data
m, the number of original features
n, the number of selected features
P, the number of learnt classifers
$\mathcal{B}$, a classifier learning algorithm
Output:
$\mathcal{H}$, a set of learnt classifiers

for $p \leftarrow 1$ *to* P **do**
  randomly select *n* features from *m* original features
  `/* project` D `on the selected features by removing not selected features on each sample */`
  $D^p \leftarrow$ project D on the p selected features
  `/* construct one classifier by using` D^p `as training data */`
  $classifier_p \leftarrow \mathcal{B}(D^p)$
  $\mathcal{H} \leftarrow \mathcal{H} \cup classifier_p$
end
return $\mathcal{H}$;

One of the key advantages of RSE is its ability to handle high-dimensional datasets [7], [14]. In these datasets, the number of features can be much larger than the number of samples, making it difficult for traditional machine learning models to make accurate predictions. By using subsets of features, RSE can reduce the dimensionality of the data and improve the performance of the models. RSE can also help to reduce overfitting, a common problem in machine learning, by creating multiple models that are trained on different subsets of the data [14].

RSE can be used with a wide range of machine learning algorithms, including decision trees, neural networks, and support vector machines [23], [24]. It has been shown to improve the performance of these algorithms, particularly for high-dimensional datasets. However, the performance of RSE can be sensitive to the choice of hyperparameters, such as the number of features and the number of models in the ensemble. Therefore, careful tuning of these hyperparameters is necessary to achieve optimal performance. Overall, RSE is a powerful technique for improving the performance of machine learning models, particularly for high-dimensional datasets [14], [23].

2.3. Classification with high-dimensional data

In recent years, high-dimensional data has become increasingly prevalent in many fields, including biology, finance, and social sciences [6]. High-dimensional data refers to data sets that have a large number of features or variables compared to the number of samples. In many cases, traditional machine learning methods may struggle to analyze such data due to the curse of dimensionality. Therefore, specialized techniques and algorithms have been developed to address the challenges of high-dimensional data in machine learning [3]–[5].

One approach for dealing with high-dimensional data in machine learning is feature selection or dimensionality reduction [3], [4]. Feature selection aims to select the most relevant features for the task at hand and discard the rest [3]. Dimensionality reduction, on the other hand, aims to transform the data into a lower-dimensional space while preserving important information [4]. Both methods can help to reduce computational complexity and improve the accuracy of machine learning algorithms [3], [4].

Another approach is to use ensemble learning, which combines multiple models to improve prediction accuracy [25], [26]. Ensemble learning is particularly effective in dealing with high-dimensional data because it can handle noisy or redundant features by averaging out their effects. Ensemble methods, such as random forest, have been shown to be highly effective for classification with high-dimensional data. By leveraging the strengths of multiple models, ensemble learning can improve classification performance and overcome the challenges of high-dimensional data [25], [26].

Overall, machine learning for classification with high-dimensional data is a challenging task that requires specialized techniques and algorithms. Feature selection, dimensionality reduction, and ensemble learning are among the most effective approaches for dealing with high-dimensional data in machine learning. By applying these techniques, researchers can better analyze and extract valuable insights from high-dimensional data sets in a wide range of fields [3], [4].

3. The proposed method

We propose two hybrid ensemble machine learning approaches for cancer classification from gene expression data. The first approach is to integrate bagging into RSE, and the second one is to integrate boosting into RSE. Algorithm 2 shows the first approach, and Algorithm 3 shows the second approach. The purpose of the integration is to promote the advantages of RSE, bagging and boosting in classifying cancer gene expression data.

The proposed algorithm presented in Algorithm 2 aims to improve classification performance by integrating two ensemble learning techniques: RSE and Bagging. The first step in this approach involves using RSE to generate a subset of training data with a sub of features. By reducing the feature space before applying Bagging, the algorithm can build more accurate classifiers that are less prone to overfitting. Once

Algorithm 2: The training process of random subspace bagging ensemble (SubBag)

Input:
D, the original training data
m, the number of original features
n, the number of selected features
P, the number of random subspace classifers
K, the number of bagging classifers
$\mathcal{B}$, a base classifier learning algorithm
Output:
$\mathcal{H}$, a set of learnt classifiers

for $p \leftarrow 1$ *to* P **do**
  randomly select *n* features from *m* original features
  $X^p \leftarrow$ project *X* on the *p* selected features
  for $k \leftarrow 1$ *to* K **do**
    $X_k^p \leftarrow$ Bootstrap(X^p)
    $\mathcal{H}_k^p \leftarrow \mathcal{B}(X_k^p)$
    $\mathcal{H}_p \leftarrow \mathcal{H}_p \cup \mathcal{H}_k^p$
  end
  $\mathcal{H} \leftarrow \mathcal{H} \cup \mathcal{H}_p$
end
return $\mathcal{H}$;

the sub training data has been generated, Bagging is used to build a set of classifiers from this data. By using multiple models and averaging their predictions, Bagging can reduce variance and improve the robustness of the classifier. The integration of RSE and Bagging aims to capitalize on the strengths of both techniques. By reducing the feature space before applying Bagging, the algorithm can build more accurate classifiers that are less prone to overfitting. This is particularly important in high-dimensional data, where overfitting can be a significant problem due to the large number of features.

Algorithm 3 presents the second approach proposed for improving classification performance using ensemble learning. In this approach, RSE is used to generate sub training data with a subset of features, similar to the first approach. However, instead of using Bagging, Boosting is used to build a set of classifiers from this sub training data. By integrating RSE and Boosting, the proposed algorithm aims to leverage the strengths of both techniques to improve classification performance. By reducing the feature space with R, the algorithm can mitigate the effects of the curse of dimensionality, while Boosting can build a more accurate classifier that is robust to overfitting.

In summary, the first algorithm integrates RSE and Bagging, which involves generating a subset of training data with a sub of features using RSE, and then building a set of classifiers using Bagging. This approach aims to reduce overfitting and improve

Algorithm 3: The training process of random subspace bagging ensemble (SubBoost)

Input:
D, the original training data
m, the number of original features
n, the number of selected features
P, the number of random subspace classifers
K, the number of boosting classifers
$\mathcal{B}$, a base classifier learning algorithm
Output:
$\mathcal{H}$, a set of learnt classifiers

for $p \leftarrow 1$ *to* P **do**
  randomly select *n* features from *m* original features
  $X^p \leftarrow$ project X on the p selected features
  $X_1^p \leftarrow X^p$
  for $k \leftarrow 1$ *to* K **do**
    $\mathcal{H}_k^p \leftarrow \mathcal{B}(X_k^p)$
    $\mathcal{H}_p \leftarrow \mathcal{H}_p \cup \mathcal{H}_k^p$
    $X_{k+1}^p \leftarrow$ AdjustDis(X_k^p,$\mathcal{H}_p$)
  end
  $\mathcal{H} \leftarrow \mathcal{H} \cup \mathcal{H}_p$
end
return $\mathcal{H}$;

accuracy in high-dimensional data. The second algorithm integrates RSE and Boosting, which also involves generating a subset of training data with a sub of features using RSE, but building a set of classifiers using Boosting instead. This approach aims to mitigate the curse of dimensionality and build a more accurate and robust classifier.

4. Experiment design and results

4.1. Experiment design

The approaches were assessed through experiments conducted on 12 high-dimensional gene expression datasets obtained from the benchmark study in [27]. Table 1 displays the key features of the datasets, including their names, tissue types, number of samples and genes, number of cancer types, and the distribution of samples across classes. The datasets used in the experiments have a relatively small number of samples and are often imbalanced, rendering traditional ten-fold cross-validation ineffective. To address this issue, leave-one-out cross-validation (LOOCV) is employed, whereby each dataset is partitioned into various training-test set pairs. Specifically, LOOCV utilizes a single sample as the test set and the remaining samples as the training dataset. Consequently,

Table 1. Datasets used in the experiments

Dataset	Tissue	#Sample	#Genes (#Features)	#Classes	Dist. Classes
alizadeh-v1	Blood	42	3687	2	21, 21
alizadeh-v2	Blood	62	3374	3	42, 9, 11
alizadeh-v3	Blood	62	3374	4	21, 21, 9, 11
bredel	Brain	50	20000	3	31, 14, 5
Bittner	Skin	38	2201	2	19, 19
Bredel	Brain	50	1739	3	31, 14, 5
Dyrskjot	Bladder	40	1203	3	9, 20, 11
Golub-v1	Bone marrow	72	1868	2	47, 25
Golub-v2	Bone marrow	72	1868	3	9, 20, 11
Liang	Brain	37	1411	3	28, 6, 3
Nutt-v3	Brain	22	1152	2	7, 15
Risinger	Endometrium	42	1771	4	13, 3, 19, 7

the number of training and test set pairs produced by LOOCV for each dataset equals the number of samples in the dataset.

Similar to ensemble learning algorithms, any decision tree algorithm can be used as a base classifier for the proposed method. The C4.5 algorithm is one of the most widely used decision tree algorithms; therefore, it has been chosen as the base classifier for the proposed method and the benchmark methods. The classification algorithm employed is C4.5, and all algorithms are implemented using WEKA [28]. The proposed methods are compared with popular ensemble machine learning techniques such as bagging, boosting, random forests, and RSE in the experiments. The parameters of C4.5 are set uniformly for the ensemble methods and default parameters are used in WEKA. Specifically, the confidence threshold for pruning is set to 0.25, the minimum number of instances per leaf is set to 2, and the number of folds for reduced error pruning is set to 3. To ensure fairness among ensemble methods, all ensemble methods are configured with the same number of base classifiers, which is 200. For the two proposed methods, SubBag and SubBoost, in order to achieve 200 base classifiers, the parameter P is set to 20, and the parameter K is set to 10.

Table 2. Classification error of different methods using C4.5 as classifier

Datasets	SubBag	Bag	SubBoost	Boost	Sub	RanFor	Single
alizadeh-v1	**2.38**	7.14	7.14	26.19	9.52	9.52	38.09
alizadeh-v2	**0.00**	1.61	1.61	11.29	3.22	3.22	11.29
alizadeh-v2	**1.61**	3.22	3.22	20.96	3.22	3.22	32,25
bredel	**14.00**	16.00	16.00	24.00	24.00	16.00	48.00
Bittner	15.78	15.78	**13.15**	34.21	18.42	18.42	57.89
Bredel	**14.00**	18.00	16.00	16.00	16.00	19.99	18.00
Dyrskjot	**7.49**	25.00	12.50	19.99	15.00	15.00	27.50
Golub-v1	**2.77**	6.94	**2.77**	12.50	**2.77**	**2.77**	18.05
Golub-v2	**4.16**	**4.16**	**4.16**	6.94	**4.16**	9.72	4.16
Liang	**5.40**	13.51	**5.40**	24.32	**5.40**	10.81	24.32
Nutt-v3	13.63	31.81	18.18	59.09	31.81	**9.09**	59.09
Risinger	**21.42**	28.57	23.80	42.85	23.80	28.57	61.90

4.2. Results and discussion

The classification error for the proposed methods and benchmark methods is displayed in Table 2, where abbreviations are utilized to represent different methods. Specifically, the proposed methods presented in Algorithm 2 and Algorithm 3 are denoted as “SubBag” and “SubBoost”, respectively. The other methods are referred to as “Bag”, “Boost”, “RanFor”, “Sub”, and “Single”, corresponding to Bagging, Boosting, RSE, Random Forest, and a single classifier, respectively.

Table 2 indicates that the proposed hybrid ensemble methods can effectively decrease the classification error compared to solely utilizing Bagging or Boosting. Specifically, in 10 out of 12 cases, the “SubBag” algorithm outperforms “Bag”, with the two algorithms having the same accuracy in the remaining 2 cases. Therefore, the combination of Bagging and RSE yields superior outcomes than using solely Bagging. Similarly, in 11 out of 12 cases, the “SubBoost” algorithm surpasses “Boost”, and the two algorithms have the same accuracy in the remaining case. Consequently, the combination of Boosting and RSE also provides better results than utilizing only Boosting.

Table 2 also demonstrates that the proposed hybrid ensemble methods can effectively

decrease the classification error compared to solely utilizing RSE. Specifically, in 9 out of 12 cases, the "SubBag" algorithm outperforms "Sub", and the two algorithms have the same accuracy in the remaining 3 cases. Therefore, the combination of Bagging and RSE yields better results than utilizing solely RSE. Similarly, in 6 out of 12 cases, the "SubBoost" algorithm outperforms "Sub", and the two algorithms have the same accuracy in the remaining cases. Thus, the combination of Boosting and RSE also provides better outcomes than utilizing solely RSE.

Table 2 illustrates that the proposed methods outperform Random Forest. Specifically, in 10 out of 12 cases, the "SubBag" algorithm surpasses "RanFor", and in 8 out of 12 cases, the "SubBoost" algorithm outperforms "RanFor". Additionally, it is evident that all the ensemble methods significantly outperform utilizing only a single classifier.

In summary, the experimental results show that both proposed methods can reduce classification error compared to using only Bagging, Boosting, or RSE. The results also demonstrate that using ensemble methods is much better than using a single classifier.

5. Conclusions

The paper proposed two hybrid ensemble machine learning approaches for classifying high-dimensional data. Both methods leverage ensemble learning techniques to deal with high-dimensional data and improve classification performance. The first approach integrates RSE with bagging, which first generates sub-training data with a subset of features, and then builds a set of classifiers using bagging. The purpose of this integration is to use RSE to reduce feature space before using bagging to build more accurate classifiers. The second approach integrates RSE with boosting, which also generates sub-training data with a subset of features and builds a set of classifiers using boosting. By leveraging the strengths of both RSE and boosting, this approach can effectively deal with the challenges posed by high-dimensional data and improve classification performance. Experimental results have shown that both methods can improve classification accuracy for high-dimensional data. Overall, the proposed methods provide an effective approach for improving classification with high-dimensional data using ensemble learning.

References

[1] S. B. Kotsiantis, I. Zaharakis, P. Pintelas *et al.*, "Supervised machine learning: A review of classification techniques," *Emerging artificial intelligence applications in computer engineering*, vol. 160, no. 1, pp. 3–24, 2007.

[2] P. C. Sen, M. Hajra, and M. Ghosh, "Supervised classification algorithms in machine learning: A survey and review," in *Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018*. Springer, 2020. doi: 10.1007/978-981-13-7403-6_11 pp. 99–111.

[3] J. Hua, W. D. Tembe, and E. R. Dougherty, "Performance of feature-selection methods in the classification of high-dimension data," *Pattern Recognition*, vol. 42, no. 3, pp. 409–424, 2009. doi: https://doi.org/10.1016/j.patcog.2008.08.001

[4] B. Tran, B. Xue, and M. Zhang, "Genetic programming for feature construction and selection in classification on high-dimensional data," *Memetic Computing*, vol. 8, pp. 3–15, 2016. doi: https://doi.org/10.1007/s12293-015-0173-y

[5] B. Ghaddar and J. Naoum-Sawaya, "High dimensional data classification and feature selection using support vector machines," *European Journal of Operational Research*, vol. 265, no. 3, pp. 993–1004, 2018. doi: https://doi.org/10.1016/j.ejor.2017.08.040

[6] S. Ayesha, M. K. Hanif, and R. Talib, "Overview and comparative study of dimensionality reduction techniques for high dimensional data," *Information Fusion*, vol. 59, pp. 44–58, 2020. doi: https://doi.org/10.3390/info12010001

[7] G. Yu, G. Zhang, C. Domeniconi, Z. Yu, and J. You, "Semi-supervised classification based on random subspace dimensionality reduction," *Pattern Recognition*, vol. 45, no. 3, pp. 1119–1135, 2012. doi: https://doi.org/10.1016/j.patcog.2011.08.024

[8] R. Polikar, "Ensemble learning," in *Ensemble machine learning*. Springer, 2012, pp. 1–34. doi: DOI: 10.1007/978-1-4419-9326-7_1

[9] O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1249, 2018. doi: https://doi.org/10.1002/widm.1249

[10] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," *Frontiers of Computer Science*, vol. 14, no. 2, pp. 241–258, 2020. doi: https://doi.org/10.1007/s11704-019-8208-z

[11] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996. doi: https://doi.org/10.1007/BF00058655

[12] R. E. Schapire, "The boosting approach to machine learning: An overview," *Nonlinear estimation and classification*, pp. 149–171, 2003. doi: https://doi.org/10.1007/978-0-387-21579-2_9

[13] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE transactions on pattern analysis and machine intelligence*, vol. 20, no. 8, pp. 832–844, 1998. doi: 10.1109/34.709601

[14] Z. Yu, D. Wang, J. You, H.-S. Wong, S. Wu, J. Zhang, and G. Han, "Progressive subspace ensemble learning," *Pattern Recognition*, vol. 60, pp. 692–705, 2016. doi: https://doi.org/10.1016/j.patcog.2016.06.017

[15] P. Panov and S. Džeroski, "Combining bagging and random subspaces to create better ensembles," in *Advances in Intelligent Data Analysis VII: 7th International Symposium on Intelligent Data Analysis, IDA 2007, Ljubljana, Slovenia, September 6-8, 2007. Proceedings 7*. Springer, 2007. doi: https://doi.org/10.1007/978-3-540-74825-0_11 pp. 118–129.

[16] G. Wang and J. Ma, "A hybrid ensemble approach for enterprise credit risk assessment based on support vector machine," *Expert Systems with Applications*, vol. 39, no. 5, pp. 5325–5331, 2012. doi: https://doi.org/10.1016/j.eswa.2011.11.003

[17] W. Chen, H. Hong, S. Li, H. Shahabi, Y. Wang, X. Wang, and B. B. Ahmad, "Flood susceptibility modelling using novel hybrid approach of reduced-error pruning trees with bagging and random subspace ensembles," *Journal of Hydrology*, vol. 575, pp. 864–873, 2019. doi: https://doi.org/10.1016/j.jhydrol.2019.05.089

[18] N. García-Pedrajas and D. Ortiz-Boyer, "Boosting random subspace method," *Neural Networks*, vol. 21, no. 9, pp. 1344–1362, 2008. doi: https://doi.org/10.1016/j.neunet.2007.12.046

[19] G. Wang and J. Ma, "Study of corporate credit risk prediction based on integrating boosting and random subspace," *Expert Systems with Applications*, vol. 38, no. 11, pp. 13 871–13 878, 2011. doi: https://doi.org/10.1016/j.eswa.2011.04.191

[20] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001. doi: https://doi.org/10.1023/A:1010933404324

[21] R. E. Schapire, "Explaining adaboost," in *Empirical inference*. Springer, 2013, pp. 37–52. doi: https://doi.org/10.1007/978-3-642-41136-6_5

[22] Y. Tian and Y. Feng, "Rase: Random subspace ensemble classification," *The Journal of Machine Learning Research*, vol. 22, no. 1, pp. 2019–2111, 2021.

[23] R. Vyškovskỳ, D. Schwarz, and T. Kašpárek, "Brain morphometry methods for feature extraction in random subspace ensemble neural network classification of first-episode schizophrenia," *Neural computation*, vol. 31, no. 5, pp. 897–918, 2019.

[24] W. Cai, D. Yu, Z. Wu, X. Du, and T. Zhou, "A hybrid ensemble learning framework for basketball outcomes prediction," *Physica A: Statistical Mechanics and its Applications*, vol. 528, p. 121461, 2019. doi: https://doi.org/10.1016/j.physa.2019.121461

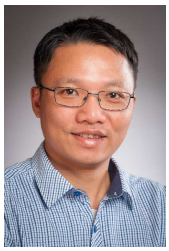
[25] H. Yu and J. Ni, "An improved ensemble learning method for classifying high-dimensional and imbalanced biomedicine data," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 11, no. 4, pp. 657–666, 2014. doi: 10.1109/TCBB.2014.2306838

[26] V. Kumar and S. Minz, "Multi-view ensemble learning: an optimal feature set partitioning for high-dimensional data classification," *Knowledge and Information Systems*, vol. 49, pp. 1–59, 2016. doi: https://doi.org/10.1007/s10115-015-0875-y

[27] M. C. de Souto, I. G. Costa, D. S. de Araujo, T. B. Ludermir, and A. Schliep, "Clustering cancer gene expression data: a comparative study," *BMC bioinformatics*, vol. 9, p. 497, 2008. doi: https://doi.org/10.1186/1471-2105-9-497

[28] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, pp. 10–18, 2009. doi: https://doi.org/10.1145/1656274.1656278

Manuscript received: 10-04-2023; Accepted: 12-06-2023.



Cao Truong Tran received the PhD degree in computer science from Victoria University of Wellington, New Zealand. He also did postdoc at Victoria University of Wellington. He is researching in the field of machine learning and evolutionary computation, specialized with evolutionary machine learning for data mining with missing data. He servers as a reviewer of international journals, including IEEE Transactions on Evolutionary Computation, IEEE Transactions on Cybernetics, Pattern Recognition, Knowledge-Based Systems, Applied Soft Computing and Engineering Application of Artificial Intelligence. He is also a PC member of international conferences, including IEEE Congress on Evolutionary Computation, IEEE Symposium Series on Computational Intelligence, the Australasian Joint Conference on Artificial Intelligence and the AAAI Conference on Artificial Intelligence.
Email: truongct@lqdtu.edu.vn

HỌC CỘNG ĐỒNG CHO PHÂN LỚP VỚI DỮ LIỆU NHIỀU CHIỀU

Trần Cao Trưởng

Tóm tắt

Phân loại với dữ liệu nhiều chiều thường gặp nhiều thách thức như hiện tượng quá khớp (overfitting), hiệu suất phân loại giảm, và yêu cầu tính toán cao. Để cải thiện hiệu suất phân loại, bài báo giới thiệu hai phương pháp học cộng đồng mới (ensemble learning) dựa trên việc tích hợp phương pháp học cộng đồng lựa chọn ngẫu nhiên các tập thuộc tính (RSE: Random subspace ensemble) với bagging và boosting.

Phương pháp đầu tiên tích hợp RSE với bagging, giúp giảm không gian đặc trưng bằng cách tạo ra các tập dữ liệu con với tập con các thuộc tính, sau đó sử dụng bagging để xây dựng một tập các bộ phân lớp. Phương pháp kết hợp này giúp tạo ra những bộ phân loại chính xác hơn.

Phương pháp thứ hai tích hợp RSE với boosting, tạo ra các tập dữ liệu con với tập con các thuộc tính và sử dụng boosting để xây dựng một tập các bộ phân loại. Phương pháp này tận dụng những ưu điểm của cả RSE và boosting để giải quyết các thách thức của dữ liệu nhiều chiều, cải thiện hiệu suất phân loại.

Cả hai phương pháp đều được kiểm chứng thông qua các thí nghiệm, kết quả cho thấy cả hai phương pháp đều cải thiện đáng kể độ chính xác phân loại với dữ liệu nhiều chiều. Tổng thể, các phương pháp học cộng đồng đề xuất trong bài báo cung cấp một cách tiếp cận hiệu quả để cải thiện việc phân loại dữ liệu nhiều chiều bằng học cộng đồng.

Từ khóa

Phân lớp, học cộng đồng, dữ liệu nhiều chiều, RSE, Bagging, Boosting.