Boosting Algorithm and Its Application in Secondary Chemistry Data Mining

Yingshuang Wu[1(✉)], Shijie Zhang[1], Yingshuang Wu[1], and Rong Zhang[2]

[1] College of Chemistry and Materials Science, Jiangsu Normal University, Xuzhou 221116, Jiangsu, China
wuletttetu@163.com
[2] Xingtai University, Xingtai 054000, Hebei, China

Abstract. In order to fundamentally improve the accuracy of data mining algorithm, efficiency and quality of the algorithm, the introduction of condition mode and depth priority strategy optimization data mining algorithm, please improve the data mining algorithm in various fields, by deleting invalid or using low frequency method to accelerate the resource positioning work, and improve the performance of the system resource positioning. This paper mainly combines the practical application of Boosting algorithm in middle school chemical data mining for in-depth analysis and discussion for reference.

Keywords: Boosting algorithm · Chemical data mining · Application

1 Introduction

In data mining, classification and prediction is a basic task. The traditional practice usually uses pattern recognition and statistical methods to obtain prediction rules. However, the proposal of Boosting algorithm can effectively solve the problems of blocked predictor accuracy.

2 Basic Overview of the Boosting Algorithm

The basic goal of Boosting algorithm is to use circular iteration to transform weak classifier science into classifier with low classification error rate, and then produce multiple rough estimates with good accuracy and low difficulty. Finally, these rules are processed comprehensively processed to form high-precision estimation, which belongs to the category of learning model. Boosting is a loop-class algorithm that can retain the weights of the samples in the training set to match it and reflect its own importance, and correct the data information according to the actual situation, to ensure that the learner concentrates in the differentiated samples, thus producing classifiers of different types. In addition, the algorithm can combine the classification errors of the generated single classification model in each cycle stage, adjust and change the weights of the training samples in time,

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
Y. Zhang and N. Shah (Eds.): BigIoT-EDU 2023, LNICST 580, pp. 35–40, 2024.
https://doi.org/10.1007/978-3-031-63130-6_5

and then focus on the classification activities of the samples with high classification difficulty in the next classification learning stage. In the Boosting, By modifying the data information by applying the weights to the training observations, The more the weight, the stool represents the greater effect of the observation on the classifier, Initial classes of samples were also given the same weights, To implement the first step of training the classification, Classifier construction by implementing a classification algorithm on samples in each cycle stage, And using the weights to measure the error of the classifier, Increase the weights of the samples misclassified by the classifier, Reducing the weights of the samples correctly classified by the classifier, The modification of the weights also highlights the corrected data to some extent, Promote the next study to focus on difficult to classify, The final classifier was obtained by weight comparison [1].

For the characteristics of Boosting, it highlights the advantages of simplicity, high efficiency and easy programming. In addition to the iteration number T, it can reduce the adjustment of parameters and other links, without the prior knowledge of weak learners. Therefore, diversified methods can be used to explore weak assumptions according to the actual situation. At the same time, each training set of Boosting is not in an independent state, and its own choice is greatly correlated with the previous learning results. The prediction functions of Boosting have high weights and often can only be generated in the specified order. In addition, the Boosting method requires a small change in the classifier without stability, namely, in the training set, which can cause significant changes to the classification model, such as support vector machines, neural networks and decision trees. Although Boosting presents many advantages, but there are also many shortcomings. The actual performance of Boosting at some level and in specific problems often over-relies on weak learners and data content, which is consistent with theoretical knowledge. In insufficient data sets, Boosting cannot play its own real power on the too weak or too complex level, and is too sensitive to the noise generated by the data set.

3 Basic Content of the Boosting Algorithm

3.1 Main Algorithms

The Ada Boost algorithm is a typical algorithm of the Boosting family. Its basic goal is to find out the hypothesis H (x) according to the provided example x and the prediction tag y by using boosting. Its pseudo-code is reflected as the algorithm input: the training set S = (x 1, y 1), (x 2, y 2),…(xm, ym)|. In this case, each member is a labeled training example, x 1 X, X represents a domain or an instance space, y 1 a label set Y, learners can accept (x 1, y 1) is randomly selected from X Y distributed to P, assuming two types of problems, Y = −1, +1, the basis of a diversification problem, Ada Boost repeatedly call a given weak or basic learning algorithm, the main idea is usually the training set of weight distribution, distribution weights on t can be recorded as Di. In the initial case, the weight of all examples is set to be equal, but the instance and weight of each misscore will continue to increase, thus reducing the difficulty that the learner is forced to focus in the training set. In general, the fundamental task of weak learning is embodied by combining the distribution Dt to find the matching weak hypothesis h 1 [2].

3.2 Training Error Analysis

Schapire and Singer highlighted hypothesis training error boundaries during generalization of Freund and Schapice, If each weak hypothesis is somewhat better than the random conjecture, Constrained γ is the farthest from 0, Training errors will decrease exponentially, There are also some similarities between the Ada Boost and the Boosting algorithm, However, the previous adopted algorithms need to obtain a known lower bound γ before practical learning, There is relatively little information associated with such boundary knowledge in practical life, While Ada Boost can fundamentally adjust the error rate of weak assumptions, Therefore, highlight the advantages of adaptation.

3.3 Margin Theory

$\sqrt{\frac{TD}{m}}$ Freund and Schapire use VC dimension to further explore the generalization error of Boosting from the perspective of training error. VC dimension is the measure of learning algorithm complexity and learning ability, which can fundamentally derive the generalization error value. The main formula is Pr (H (x) = y) + O (). During this period, m mainly represents the sample number, d is VC dimension, T is the training rounds, P is the empirical probability of training set, this formula fully shows that if the training wheel value T is too large, Boosting will lead to overadaptation, but after a series of research can be seen that Boosting does not lead to overadaptation, in this case, Schapire pair training case margin for multi-level analysis, large positive boundary represents high credibility accuracy prediction, large negative boundary credibility error prediction. However, the smaller boundaries represent the predictions with low confidence. The interpretation of Schapire is mainly reflected that after the training error is reduced to 0, Boosting can still maintain the original state lifting boundary, further increase the minimum boundary, increase the reliability of classification, and reduce the overall error problem.

3.4 Combination and Classification Technology

Combined classifiers are a unified integration of classifiers, whose individual decisions are often combined through a pathway approach, and they expand the classification processing of new samples in this case. In the supervised learning link, the research field of high activity is reflected in the well-constructed classifier combination, according to the relevant experiments can show that the single classifier opinion difference during the combination of the accuracy is higher than the separate classifier, if the individual classifier finally obtained assumption error rate exceeds 0, then represents the final voting results and errors will increase. Therefore, the final assembled combination method uses individual classifiers with an error rate below zero, and these individual classifier errors are not correlated at some level. Combination classifier involved in special use and general due to two types, general technology combined with its own characteristics can be divided into change on the training case distribution, control the output target, introduce randomness, etc., for the combination of special application involves using targeted measures to implement training neural network B P algorithm or decision tree produce combination, special application need to analyze and explore the actual situation. The

implementation of combined classifier usually includes weighted stock, unauthorized stock and gate network, etc. Boosting uses weighted stock. For classification problems, the weights are often obtained by measuring the accuracy of each member classifier in the training set and other levels, and the weight and accuracy hold the proportional relationship. 5.5 The combustion combination method mainly adopts the naive Bayesian method to learn and master the weight-important points. The weight of the classifier is usually obtained by the assumed accuracy measured during the weighted training distribution [3].

4 The Application and Research Direction of Boosting in Data Mining

At present, the difficulties of machine learning and knowledge discovery in data mining are reflected in the large-scale and high-dimensional characteristics of data. In order to fundamentally meet the diversified requirements of massive data, many machine learning algorithms should be improved and improved according to the actual situation.

Learning in large-scale data. Three different decision tree algorithms can be extended to the large-scale data processing work. First, it is based on the intelligent sampling of training data during the growth of the decision tree, and the stage of selecting and testing for the tree often includes the analysis of training data and the prediction of output class characteristics. Among the many redundant dataset areas, it is possible to choose from data-based sampling. The second is based on developing smart data structures to prevent the storage of other training data in memory space. The SPRINT method can completely disperse the training data and integrate different examples or feature values into independent disk files. SLIQ usually consumes more memory than SPRINT, but the overall speed is relatively fast. The third is the use of decision tree integration, training data can be reasonably divided into multiple disjoint subsets, each subset can also independently generate independent decision tree, these trees can also be jointly voted to predict, although the accuracy of the individual tree is relatively low, but the overall combination of performance value is good, and can speed up the efficiency and quality [4].

Learn about high-dimensional data. The current C 4.5 and BP algorithms do not have good scale characteristics, However, at a statistical level, There are numerous examples of irrelevant or polynoised features providing the corresponding informative data, Therefore, we need to reasonably screen the characteristics according to the actual situation, There are three methods in this case: First, Provide a learning algorithm for performing an initial analysis of the training data and selecting a subset of features; second, Try to differentiate feature subsets based on learning algorithms, And estimate the algorithm performance supported by these different types of subsets, Ensure that well-performing subsets are available for subsequent use; third, The screening and weighting of features are all integrated into the learning algorithm. Although existing methods can be learned using scientific and reasonable computational time among many training examples, However, the existing prediction learning accuracy should not be adopted when the quality of the predictor cannot be greatly improved, The idea of a combination of learning came into being, It often gives matching results to the learner through multiple

approaches, As can be seen from the investigation and study, Although the learner was not significant in terms of learning performance alone, And in many cases, learning performance will be a significant leap, The combination is more suitable for solving large-scale data sets than the learners alone, This also, to some extent, represents some potential in data mining (Fig. 1).

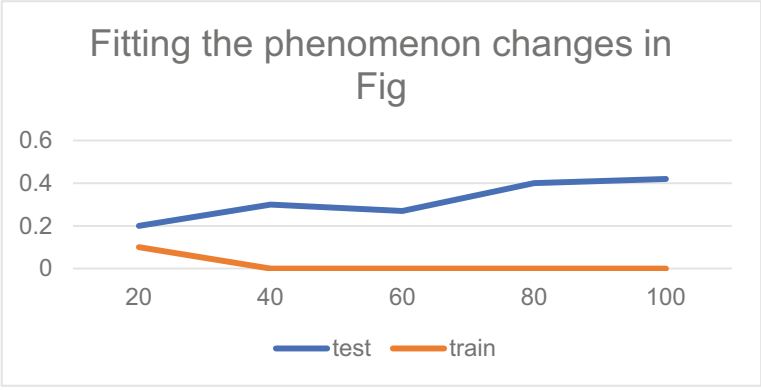

Fig. 1. Fitting the trend of phenomenon

From an overall perspective, the following Boosting combination learning methods have been applied in many fields:

4.1 Atlas Analysis

Integrating Boosting and Bayes optimization rules has developed a brand new algorithm that is used to handle the detection data classification of 215 patients with thyroid disease, saturated alcohol and ether mass spectrometry data classification, and mass spectrometry data investigation substructure classification. According to relevant investigations and studies, the two algorithms have been greatly improved to different degrees, and this method is used to carry out second-class prediction and multi-class prediction for many mass spectrometry data. Using Boosting combined classification regression tree as a classifier, can fundamentally on the green tea, wheat and cream, HIV carriers such as medical detection data retention index classification, found that part of the data classification effect is significant, but in the green tea data classification during overfitting, found Boosting results in chromatographic retention index classification is not ideal, shows that Boosting is a kind of excessively sensitive to data, noise, etc. (Fig. 2).

4.2 Prediction of Drug Efficacy

The Boosting method can be used to deal with the toxicity prediction such as nitrobenzene and lipooxidase inhibitors. After multiple linear regression and support vector regression comparison, the Boosting is slightly better in the prediction accuracy, which can provide more practical information and data for the diagnosis and treatment of tumor patients. At

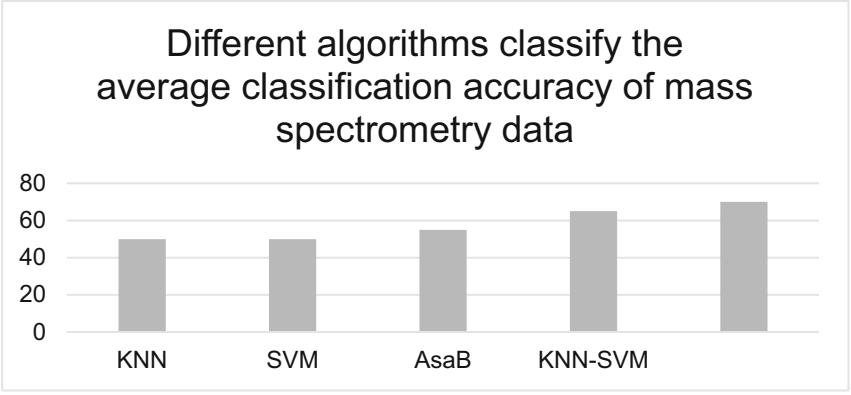

Fig. 2. Statistical graph of the average classification accuracy of mass spectrometry data classified by different algorithms

present, Boosting algorithm has been applied to the research of QSAR, but it focuses on solving the classification problem, and applying Boosting algorithm to QSAR research to solve the regression problem, which proves that this algorithm is widely used in drug molecular efficacy prediction.

5 Conclusion

To sum up, Boosting algorithm is the focus of attention in machine learning, and has received close attention from a wide range of industry insiders. Although this algorithm highlights some advantages, there are still some problems to be solved and studied. However, how to use Boosting algorithm to process large-scale chemical detection data is also a key problem to be solved in chemical data mining.

References

1. Dong, L., Cui, X.: Intelligent data mining algorithm for picking robot execution system optimization. Agric. Mech. Res. **45** (7), 224–227 + 237 (2023)
2. Qiu, D., Cai, X., Wu, L., Liang, S., Li, J.: Design of command ticket visual system based on data mining algorithm. Electron. Design. Eng. **30**(21), 61–65 (2022)
3. Li, J.: International road cost accounting method based on data mining algorithm Sankei, China **19**, 117–119 (2022)
4. Cai, L., Wei, X., Wang, J.: Analysis and application of data mining algorithm in big data. J. Honghe College **20**(05), 154–157 (2022)