An ensemble feature selection method for high-dimensional data based on sort aggregation

Jie Wang, Jing Xu, Chengan Zhao, Yan Peng and Hongpeng Wang

School of Management, Capital Normal University, Beijing, People's Republic of China

ABSTRACT

With the rapid development of the Internet, big data has been applied in a large amount of application. However, there are often redundant or irrelevant features in high dimensional data, so feature selection is particularly important. Because the feature subset obtained by a single feature selection method may be biased, an ensemble feature selection method named SA-EFS based on sort aggregation is proposed in this paper, and this method is oriented to classification tasks. For high-dimensional data sets, the results of three feature selection methods, chi-square test, maximum information coefficient and XGBoost, are aggregated by specific strategy. The integration effects of arithmetic mean and geometric mean aggregation strategy on this model are analyzed. In order to evaluate the classification and prediction performance of feature subset, three classifiers with excellent performance, KNN, Random Forest and XGBoost, are tested respectively, and the influence of threshold on classification performance is analyzed. The experimental results show that compared with the single feature selection method, the arithmetic mean aggregation ensemble feature selection can effectively improve the classification accuracy, and the threshold interval setting of 0.1 is a better choice.

1. Introduction

With the rapid development of the Internet and information technology, the scale of data that can be processed by various industries has been continuously developed, and problems such as ‘dimensional disasters’ have been brought about. Feature selection is a critical step in data preprocessing and important research content in data mining and machine learning tasks such as classification.

Feature selection is to effectively reduce feature dimension and improve classification accuracy and efficiency by deleting irrelevant and redundant features in data sets. It also has the function of denoising and preventing machine learning model from over-fitting (Chandrashekar & Sahin, 2014).

Feature selection is usually in the search space composed of all combinations of data features, through the feature subset search algorithm, to find a subset of features that are highly correlated with pattern recognition problems (such as classification learning problems), and based on the obtained optimal features. Subsets to improve the recognition performance of learning algorithms are distinguished according to the feature subset evaluation strategy. Feature selection algorithms can be divided into the following types: Filter, Wrapper, Embedded, Hybrid. Moreover, the ensemble (Yang, 2017) developed in recent years.

Ensemble learning is an effective method for machine learning. The core idea is to achieve better learning results by combining different learning models. Among many machine learning methods, ensemble learning methods are superior to single machine learning models. In recent years, the faster development of ensemble feature selection draws on the idea of ensemble learning. Different from several feature selection methods, only a single optimal feature subset is sought. The goal of combined feature selection is to obtain multiple optimal features. Set and aggregate learning results based on multiple optimal feature subsets. Due to the integration technology, the ensemble feature selection algorithm has better stability and robustness than other feature selection algorithms when dealing with high-dimensional data with multiple optimal feature subsets.

This paper proposes an ensemble feature selection based on sort aggregation (SA-EFS) method orient to the classification task. On the high-dimensional data set, the maximum information coefficient and chi-square are first used. The feature selection method such as the test method and XGBoost obtains a plurality of feature
subsets and sorts according to the feature importance degree. After the feature ranking result is normalized, the importance weight of the feature is obtained, and the candidate set of the optimal feature subset is obtained. On this basis, the learning results of multiple optimal feature subset candidate sets are aggregated according to different rules such as arithmetic mean and geometric mean, and the optimal feature subset is obtained. Finally, KNN, RF (Random Forest) and XGboost are used. The better-performing classification model was tested, and the area AUC index (Yang, 2017; Zomaya, 2013) under the receiver operating characteristic curve was used to evaluate the classification performance of different prediction models. The method based on sorting aggregation integrates the results of multiple feature selection methods. The feature importance is normalized and aggregated, which can effectively avoid the prediction model caused by one or several feature selection methods mis-selecting feature subsets – low-performance issues. Based on the experimental results from three UCI machine learning data sets, the SA-EFS method fully considers the effectiveness of different feature selection methods. The selected feature subsets after aggregation are more accurate and are available in KNN, RF, and XGboost. A higher predicted AUC value is obtained on the classification algorithm.

2. Related works

Feature selection is an important step in the data mining preprocessing stage. Many domestic and foreign scholars have carried out extensive research on Filter, Wrapper, Embedded, Hybrid, and Ensemble Feature Selection.

At present, the main way to obtain multiple optimal feature subsets is from the perspective of sample and feature (Yang, 2017). From the sample point of view, the integrated feature selection firstly sampled and partitioned the training data in some form, and then obtained multiple feature subsets based on the partitioned samples. From the feature point of view, the ensemble feature selection uses a certain feature subset search algorithm to obtain multiple optimal feature subsets (Zomaya, 2013). One commonly used ensemble feature selection method is to integrate base classifiers trained based on multiple feature subsets. For example, Kim et al. proposed EA-based Ensemble Classifier (2008), which aims to optimize the generalization performance of the classification, and uses the genetic algorithm to optimize the combination of multiple filter feature selections and multiple base classifiers and integrate them into strong classification. Device. Bolon-Canedo and other multi-optimal feature subsets selected by various filtering methods are used to construct classifiers and integrate them. Two kinds of Ensemble of Filters for gene microarray data are proposed (2014). Another type of ensemble feature selection is to select the best features from multiple feature subsets and integrate them to form an optimal feature set. Song, Ni, and Wang (2013) and other proposed feature selection algorithm based on cluster analysis (FAST FS), clustering the minimum spanning tree that is sufficient for the feature and category relationship and selecting the most relevant to each category from each tree cluster.

The features constitute the optimal feature subset. Jiang, Jiang, and Yu (2018) proposed a feature selection method based on sorting integration. The method first obtains the feature sorting sequence by using three feature selection methods such as Relieff, and then sums the weights of each feature obtained by the three methods as the total weight of the feature and selects the final feature set. Shi et al. (2018) proposed an ensemble learning algorithm based on Bagging method based on feature selection accuracy and difference and constructed a feature selector ensemble by the heterogeneous learner to achieve the impact of theft. Effective selection of prime factors. Du et al. (2019) introduced the idea of ensemble learning into feature selection and used three kinds of different feature selection methods for ensemble learning. Each method will get a set sequence sorted by feature importance, and different results will be aggregated. Get a final set of sequences.

3. Theoretical basis

3.1. Chi-square test feature selection

Chi-square test is a hypothetical test method for a wide range of count data. It belongs to the category of non-parametric tests, mainly comparing two or more sample rates (composed ratios) and the correlation analysis of two categorical variables. The basic idea is to compare the degree of agreement between the theoretical frequency and the actual frequency or the goodness of fit. The formula is as follows:

$$\chi^2 = \sum_{i=1}^{k} \frac{(A_i - E_i)^2}{E_i}$$

where $A$ is the actual value and $E$ is the theoretical value.

3.2. Feature screening of maximum information coefficient

The maximum information coefficient is developed on the basis of mutual information, which can quickly detect various types of associations and evaluate functional relationships. The maximum information coefficient is suitable for exploring the potential relationship between
pairs of variables in big data sets, which is fair and extensive.

Given a finite data set $D$ and ordered pairs $<X, Y>$, the sample size is $m$, and the $x$, $y$ plane is divided into several small meshes by $x$ value and $y$ value. This division is called $x \times y$ grid. $G$, so that all variables in $D$ fall into $G$ (Reshef et al., 2011), defined as:

$$I^p(X, Y, D, i, j) = \max(X, Y, D|G, i, j)$$

Equation (3) represents the maximum mutual information between $X$ and $Y$ when $G$ is divided into $i \times j$ grids.

In data set $D$ containing two node variables $X$ and $Y$, the characteristic matrix of $X$ and $Y$ is defined as (Reshef et al., 2011):

$$M(X, Y|D)_{ij} = \frac{I^p(X, Y, D, i, j)}{\log \text{min}(i, j)}$$

Assuming that the mesh size of the partition is smaller than $B(m)$ and $B(m) = n^{0.6}$, the maximum information coefficient of the node variables $X$ and $Y$ is defined as (Reshef et al., 2011):

$$\text{MIC}(X, Y|D) = \max_{i \times j < B(m)} \{M(X, Y|D)_{ij}\}$$

where $i \times j < B(m)$ represents the division dimension limit of the grid $G$. Obviously, $0 \leq \text{MIC} \leq 1$. Since the mutual information between random variables is symmetrical, the maximum information coefficient also has symmetry, namely:

$$\text{MIC}(X, Y|D) = \text{MIC}(Y, X|D)$$

### 3.3. XGBoost feature selection and classification principle

In the training process of feature selection, in order to improve the efficiency of generating new trees, XGBoost gives the importance score of each feature in each iteration, which shows the importance of each feature to model training, and provides a basis for establishing a new tree with gradient direction in the next iteration. The statistical significance of features can be directly used as the basis for feature selection.

In this paper, we first classify XGBoost based on all features, then get the importance of feature variables (FI) and sort them in descending order based on the information in the generated model process. Finally, we input the filtered features into the classifier to construct the prediction model.

As a classifier, XGBoost achieves accurate classification effect by iteration calculation of weak classifier (Torlai, Perrone-Bertolotti, Thomas, & Baciu, 2017). It supports weak classification algorithm and weak regression model, and is suitable for establishing regression model. Because of its fast calculation speed, good model performance, excellent performance and efficiency in application practice, it has been widely praised in the academic circles.

The model representation is shown in Formula 6.

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F$$

where $K$ denotes the number of trees, $F$ denotes all possible regression trees, and $f$ denotes a specific regression tree. The model consists of $K$ regression tree. Object function: The goal of XGBoost is to establish $K$ regression tree so that the predicted value of tree group can be as close as possible to the true value (accuracy) and have the greatest generalization ability. The objective function is defined as in formula 7:

$$\text{obj}(\theta) = \frac{1}{n} \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Function 5 describes $y_i$. The objective function consists of two parts. The first part is loss function and the second part is general term. Training XGBoost, training objective function can be defined as formula 8:

$$\text{obj}(\theta)^{(t)} = \frac{1}{n} \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \Omega(f_k)$$

The first part is the training error, the second part is the complexity of each tree. Then the training objective function can be written as formula 9:

$$\text{obj}(\theta)^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)}) + f_t(x_i) + \Omega(f_k) + \sum_{i=1}^{t-1} \Omega(f_i)$$

### 3.4. Forecast model evaluation indicators

When we study the two-class problem in the data set prediction, the prediction may have four different results, as shown in the Table 1. Column P indicates that the prediction is a positive class sample, $N$ column indicates that the prediction is a negative class sample, $p$ row indicates that the sample is actually a positive class sample, and $n$ rows indicates that the prediction is actually a negative class sample. A good prediction should be that the value on the main diagonal of the matrix is large, not the value on the main diagonal.

|       | P   | N   |
|-------|-----|-----|
| p     | TP  | FN  |
| n     | FP  | TN  |

**Table 1.** Four prediction results.
The true positive rate (TPR) is the number of instances that are correctly predicted to be positive, divided by the total number of instances in which the actual class is positive.

\[
TPR = \frac{TP}{TP + FN} \tag{10}
\]

The false positive rate (FPR) is the total number of instances that are incorrectly predicted to be positive, divided by the total number of instances in which the actual category is negative.

\[
FPR = \frac{FP}{FP + TN} \tag{11}
\]

The area under the receiver operating characteristic curve is the AUC value (Huang & Ling, 2005), which represents the relationship between the true positive rate and the false positive rate of the predictive model, as shown in Figure 1. The x-axis is the false positive rate, the y-axis. For the true positive rate, each point on the curve corresponds to a threshold. For a classification method, there is a TPR and FPR under each threshold. The ROC curve can be obtained by adjusting the threshold of the prediction model. The better predictive model has a lower false positive rate and a higher true positive rate. Therefore, the receiver operating characteristic curve of the better predictive model is closer to the (0,1) point, that is, near the upper left corner. The main diagonal represents a predictive model of random prediction.

4. Method

4.1. Overall framework

This paper proposes an Ensemble feature selection based on sort aggregation (SA-EFS) method, which firstly uses different feature selection methods to obtain candidate sets of multiple optimal feature subsets. Then, according to different rules, the learning results of multiple optimal feature subset candidate sets are aggregated to obtain the optimal feature subsets. Finally, three classification algorithms with good performance are used to verify the proposed algorithm.

The overall framework for SA-EFS feature selection is shown in Figure 2:

The specific description of the method is as follows: (1) Using chi-square method, maximum information coefficient method and XGBoost method to select features, and sorting features according to importance, to obtain multiple sorted optimal feature subsets $FS_1, FS_2 \ldots FS_t$.
5. Experimental evaluation

5.1. Experimental data sets

The experimental data selected in this paper are all from the international UCI machine learning data set. They are the three data sets of sonar, hcc-survival and musk. The details are shown in Table 3. The sample data classification number is all 2, the number of samples is 208 ~ 776, and the feature number of the sample is 49–167 dimensions.

| Table 2. The pseudo code of SA-EFS algorithm. |
|------------------------------------------------|
| **Algorithm** | **IFL-Stacking** |
| **Input:** | Training dataset \( D = \{x_1, x_2, \ldots, x_{n-1}, y_n\} \), \( \alpha \), feature selection algorithms sets \{MIC, CHI, XGBoost\} classifier sets \( CS = \{KNNN, RF, XGBoost\} \) |
| **Output:** | Classify result of feature in set \( D \) |
| **Step 1:** | Get feature sets by different feature selection algorithms |
| 1: for \( A_i \) | 2: Using dataset \( D \) do feature selection by algorithm \( A \) |
| 3: Sorted features by the result of \( A_i \) | 4: Return the sorted feature subset \( F_S_i \) |
| **Step 2:** | Get weight sequence of feature selection method \( F_S_i \) |
| 5: For \( F_S_i \) in \{\( F_S_1, F_S_2, \ldots, F_S_t \)\} | 6: for \( f_j \) in \( F_S_i \) |
| 7: \( w_j^t = (n - j)/n \) | 8: Return the weight sequence \( \{W_1, W_2, \ldots, W_t\} \) of \{\( F_S_1, F_S_2, \ldots, F_S_t \)\} |
| **Step 3:** | Get the aggregation feature weight sequence \( W \) |
| 9: if Aggregation strategy = arithmetic mean | 10: for \( f_j \) in original feature set \( F_S \) |
| 11: \( w_j = (w_j^1 + w_j^2 + \ldots + w_j^t)/t \) | 12: else if Aggregation strategy = Geometric mean |
| 13: for \( f_j \) in original feature set \( F_S \) | 14: \( w_j^t = \sqrt[t]{w_j^1 \cdot w_j^2 \cdot \ldots \cdot w_j^t} \) |
| 15: Return the Aggregation feature weight sequence \( W \) | 16: Sorted FS according to \( W \) |
| **Step 4:** | Get best feature subsequence according to \( \alpha \) |
| 17: Put the first \( \alpha \) features in \( F_S \) to \( F_S_{best} \) (\( F_S \to F_S_{best} \)) | 18: Return \( F_S_{best} \) |
| **Step 5:** | Do classify by classifier in \( CS = \{RF, KNN, XGBoost\} \) |
| 19: for classifiers \( C_i \) in \( CS \) do | 20: Select the \( \alpha \)% features from \( F_S_{best} \) |
| 21: learn \( C_i \) based on \( D \) | 22: return the classify result |

(2) For each feature \( j \) in the \( F_S_i \), normalize its importance with \( (n - j)/n \) (there are \( n \) features in total), and obtain the feature weight set of the \( i \)th feature selection method \( W_i = \{w_1^i, w_2^i, \ldots, w_n^i\} \). (3) According to a certain aggregation strategy (such as arithmetic average or geometric mean), calculate the total weight of each feature in set \( F_S_1, F_S_2, \ldots, F_S_t \), sort the \( n \) features according to the total weight, and obtain the sorted feature sequence \( W \) (4) Based on the threshold \( \alpha \). The top \( \alpha \)% features are selected from the sorted feature sequences to form the optimal feature subset (5) Based on the optimal feature subset, the performance of the algorithm SA-EFS is verified by using a variety of good classifier algorithms.

The detailed process of the SA-EFS algorithm is as follows (Table 2):

5.2. Experimental design

In order to verify the effectiveness of the feature selection method based on sorting integration proposed in this paper, this method is implemented by programming. The experimental environment is Windows 10, 64-bit, 8 GB RAM, Intel Core i5-7200U (2.70 GHz).

Combining experiments were performed using five feature selection methods and three data sets, shown as Table 4.

Firstly, a data set is selected to select features by using chi-square, maximum information coefficient and XGBoost, and aggregate multiple feature selection results by using specific aggregation strategies (mean aggregation and geometric average aggregation). Then, a prediction model is constructed by XGBoost. The performance of the prediction model is evaluated by 5 fold cross-validation, and the performance evaluation index \( A \) of the prediction model is obtained. Finally, repetitive experiments are conducted on three data sets to compare the performance of the proposed method with that of single maximum information coefficient, XGBoost and chi-square verification. Finally, the thresholds of alpha interval are reduced by half, and KNN and RF classifiers are added to construct the prediction model. By comparing the results before and after the thresholds are reduced, the appropriate thresholds of alpha interval are searched.

5.3. Classification accuracy results

Since the number of features in the three data sets is small, the threshold \( \alpha \) is set between 10% and 100%, and the prediction accuracy of the model under different conditions is compared. The experimental results are shown in Table 5.

Tables 5–7 show the classification accuracy obtained by XGBoost for the feature subsets obtained by different feature selection methods under different thresholds \( \alpha \). The blackened font portion is the highest accuracy of the different feature selection methods at the same
Table 5. sonar Dataset classification accuracy.

| Select feature method | α  | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1   |
|-----------------------|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| MIC                   |    | 0.706 | 0.828 | 0.789 | 0.796 | 0.824 | 0.821 | 0.804 | 0.796 | 0.806 | 0.816 |
| DIS                   |    | 0.709 | 0.759 | 0.849 | 0.824 | 0.828 | 0.841 | 0.834 | 0.824 | 0.830 | 0.816 |
| CHI                   |    | 0.759 | 0.797 | 0.837 | 0.856 | 0.843 | 0.840 | 0.827 | 0.806 | 0.810 | 0.816 |
| FBEST                 |    | 0.787 | 0.839 | 0.873 | 0.847 | 0.809 | 0.819 | 0.833 | 0.823 | 0.821 | 0.816 |
| GBEST                 |    | 0.731 | 0.673 | 0.651 | 0.674 | 0.686 | 0.711 | 0.742 | 0.771 | 0.768 | 0.816 |

Table 6. hcc-survival Dataset classification accuracy.

| Select feature method | α  | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1   |
|-----------------------|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| MIC                   |    | 0.785 | 0.819 | 0.801 | 0.797 | 0.812 | 0.798 | 0.782 | 0.793 | 0.790 | 0.781 |
| DIS                   |    | 0.753 | 0.777 | 0.811 | 0.794 | 0.814 | 0.808 | 0.786 | 0.784 | 0.782 | 0.781 |
| CHI                   |    | 0.749 | 0.742 | 0.773 | 0.814 | 0.809 | 0.785 | 0.778 | 0.804 | 0.812 | 0.781 |
| FBEST                 |    | 0.739 | 0.775 | 0.840 | 0.822 | 0.781 | 0.784 | 0.787 | 0.790 | 0.786 | 0.781 |
| GBEST                 |    | 0.739 | 0.720 | 0.689 | 0.675 | 0.690 | 0.762 | 0.767 | 0.769 | 0.780 | 0.781 |

Table 7. musk Dataset classification accuracy.

| Select feature method | α  | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1   |
|-----------------------|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| MIC                   |    | 0.843 | 0.839 | 0.846 | 0.843 | 0.839 | 0.837 | 0.838 | 0.836 | 0.835 | 0.833 |
| DIS                   |    | 0.830 | 0.837 | 0.846 | 0.849 | 0.849 | 0.845 | 0.842 | 0.842 | 0.839 | 0.833 |
| CHI                   |    | 0.851 | 0.859 | 0.859 | 0.845 | 0.842 | 0.849 | 0.840 | 0.844 | 0.830 | 0.833 |
| FBEST                 |    | 0.859 | 0.848 | 0.846 | 0.852 | 0.852 | 0.850 | 0.848 | 0.840 | 0.837 | 0.833 |
| GBEST                 |    | 0.821 | 0.826 | 0.817 | 0.827 | 0.821 | 0.828 | 0.823 | 0.832 | 0.832 | 0.833 |

threshold, and the box is the highest accuracy of the different feature selection methods at all thresholds on this data set. It can be seen from the experimental results that the feature subset obtained by the FBEST ensemble feature selection method has better prediction performance than the single method. Among the two polymerization methods used, the mean aggregation method shows the best performance in the musk dataset, the sonar dataset, and the hcc-survival dataset, while the geometric mean aggregation is not very effective on the three binary datasets. Obvious. Note that different aggregation methods should be chosen for different data sets. The threshold value of α is about 0.3, which can achieve a better classification accuracy. When the threshold value α is in the range of [0.4, 1], the classification accuracy will not increase but will decrease. The reason is that due to the certain correlation between different features, too many features will cause data redundancy and increase data noise, resulting in lower classification performance. How to reduce the redundancy of feature subsets is the direction of the next step.

5.4. Threshold alpha impact analysis

The change in threshold is reflected in the percentage of feature selection, which in turn affects the final classification result. In order to obtain more objective and accurate results, the impact of threshold α is analyzed. In the original interval [0.1, 1], the interval of α is set to 0.05, and whether the value of α will have a greater impact on the change of accuracy and increase the random forest classifier and KNN classifier for comparison experiments. The experimental results are shown in Figure 3. It shows the classification accuracy obtained by using the XGB-boost, random forest, and KNN as classifiers to select feature subsets with different feature selection methods at smaller partition intervals. It can be seen that the threshold values of the highest classification accuracy of the sonar, hcc-survival and musk datasets on XGB-boost are 0.3, 0.3, 0.1 (0.2, 0.3), respectively, and the highest classification accuracy threshold α is reached on the random forest. The values are 0.25, 1, 0.75, respectively. The highest classification accuracy threshold α on KNN is 0.1, 0.65, 0.7, which means that there are certain differences between different data sets. Different classifiers also have different effects on threshold α; the highest classification accuracy of random forests and KNN is lower than that of XGBoost, but it can also achieve a higher classification accuracy. Random forests can also obtain a single feature selection method in the mean integration feature selection method. Better classification performance, which means that the mean integration feature selection method has better adaptability to different classifiers. It is necessary to find a suitable threshold α for a particular...
Figure 3. Threshold alpha impact analysis.

data set, and setting a smaller threshold $\alpha$ does not cause a large change in the classification accuracy curve, which indicates that setting the threshold interval to 0.1 is a better choice.

6. Conclusion

This article uses the internationally accepted UCI machine learning data set. By data preprocessing, five feature selection methods and three data sets are combined to compare the performance difference between the proposed method and other methods. The results show that the feature selection method based on sorting integration has better feature selection. Models constructed based on this feature predict higher AUC values.

The main work and research results of this paper are summarized as follows:

(1) The three high-dimensional data sets used in this paper are real data provided by the international UCI machine learning platform, with large data volume and many characteristic variables. In order to improve the accuracy and efficiency of the model, we spend a large part of the time in data preprocessing.
(2) This paper introduces the idea of ensemble learning in feature selection and constructs an ensemble feature selection model. For the high-dimensional two-class dataset, the three feature selection methods of maximum information coefficient, XGBoost, and chi-square test are ensemble, and the integration effect of geometric and mean aggregation methods on this model is analyzed. Feature selection method, mean integration feature selection can effectively improve classification accuracy.

(3) In order to verify the effectiveness of the feature selection method based on sorting integration proposed in this paper. First, we use XGBoost to construct the prediction model. Then, the performance of the prediction model is evaluated by 5-fold cross-validation, and the performance evaluation index AUC of the prediction model is obtained. The performance difference between the proposed method and the three methods is compared. Finally, the threshold of the interval between the values of alpha is reduced by half, and the prediction model is constructed by adding KNN and random forest classifier. The results before and after the threshold reduction are compared to find the appropriate interval between the thresholds.

(4) The model was evaluated using a comprehensive model evaluation index. The feature selection method based on sorting integration constructed in this paper achieves the best results on all three datasets with a threshold of 0.1, and the AUC is 0.873, 0.840, and 0.859, respectively, which are at a high level. It shows that the proposed feature selection method based on sorting integration has higher accuracy and higher practical value.

Due to time, the experiment was only tested on three data sets. The experiment has certain limitations. Therefore, the method needs to be applied to more high-dimensional data sets to further verify the validity of the model. Also, this study found that only a few features can bring useful information to the classification model. Too many features will result in redundancy of feature subsets and reduce the prediction accuracy of the classification model. Therefore, considering the correlation between different features, reducing the redundancy of feature subsets is also a problem that needs further study.

Acknowledgements
This work was supported by Social Science Project of Beijing Education Commission (grant number SM201910028017) and Capacity Building for Sci-Tech Innovation-Fundamental Scientific Research Funds of Beijing Education Commission (grant number 19530050142 &19530050187).

Disclosure statement
No potential conflict of interest was reported by the authors.

Funding
This work was supported by Social Science Project of Beijing Education Commission [grant number SM201910028017] and Capacity Building for Sci-Tech Innovation-Fundamental Scientific Research Funds of Beijing Education Commission [grant number 19530050142 &19530050187].

ORCID
Jie Wang http://orcid.org/0000-0002-1262-6719

References
Bolón-Canedo, V., Sánchez-Marono, N., & Alonso-Betanzos, A. (2014). Data classification using an ensemble of filters. Neurocomputing, 135, 13–20.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. Pergamon Press.

Du, C., Zhou, C., et al. (2019). Application of ensemble feature selection in gene expression data. Journal of Shandong University of Science and Technology(Natural Science), 38(1), 85–90.

Huang, J., & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. IEEE Transactions on Knowledge and Data Engineering (TKDE), 17(3), 299–310.

Jiang, L., Jiang, S., & Yu, Q. (2018). Feature selection method based on sorting integration in software defect prediction. Journal of Chinese Computer Systems, 39(7), 36–40.

Kim, K.-J., & Cho, S.-B. (2008). An evolutionary algorithm approach to optimal ensemble classifiers for DNA microarray data analysis. IEEE Transactions on Evolutionary Computation, 12(3), 377–388.

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., . . . Sabeti, P. C. (2011). Detecting novel associations in large data sets. Science, 334(6062), 1518–1524.

Shi, T., Jiang, W., Zhang, J., et al. (2018). Theft prediction method based on ensemble feature selection. Transaction of Beijing Institute of Technology, 38(9), 109–114.

Song, Q., Ni, J., & Wang, G. (2013). A fast clustering-based feature subset selection algorithm for high-dimensional data. IEEE Transactions on Knowledge and Data Engineering, 25(1), 1–14.

Torlay, L., Perrone-Bertolotti, M., Thomas, E., & Baciu, M. (2017). Machine learning–XGBoost analysis of language networks to classify patients with epilepsy. Brain Informatics, 4, 159–169. doi:10.1007/s40708-017-0065-7

Yang, J. (2017). Study on ensemble feature selection of biomics data.

Zomaya, A. Y. (2013). Stability of feature selection algorithms and ensemble feature selection methods in bioinformatics. Biological Knowledge Discovery Handbook: Preprocessing, Mining and Post Processing of Biological Data, 23, 333.