Improved Filter Method for Feature Selection

Yongle Chen, Yiwen Zhong

College of Computer and Information, Fujian Agriculture and Forestry University, Fuzhou, 350200, PR China

Key Laboratory of Smart Agriculture and Forestry (Fujian Agriculture and Forestry University), Fujian Province University, Fuzhou, 350002, PR China

Corresponding author’s e-mail: yiwzhong@fafu.edu.cn

Abstract. To solve the problem of high data feature dimension in intrusion detection, a hybrid feature selection method has been proposed to reduce the feature dimension. This approach combines the filter and sequence floating forward search methods. Firstly, the original feature series is sorted by different filter methods, and the top ranked features are selected as the next original feature set. Based on this, the sequence floating forward search method is used to select the optimal feature subset with the Support Vector Machine (SVM) as the classifier. The method can avoid the selection and screening of the features based on the threshold and the evaluation value of single feature and categorical variable. Thereby the high evaluation characteristics obtained by the Filter method may be complementary to the low evaluation characteristics. The result shows that the proposed method can not only effectively reduce the number of features, but also can achieve a better classification performance.

1. Introduction

The large amount of noise and irrelevant features in network data may seriously interfere with and restrict the performance of the intrusion detection system. Feature selection is to select some feature set with the most ability to distinguish categories from the original feature set according to some criteria. How to optimize the feature selection method, select the optimal feature subsets that have the greatest impact on the performance of the classifier from the network data, reduce the dimension of the feature and computational complexity, improve the performance of the classifier and speed up the detection while ensuring the classification accuracy is the hot direction of current research.

Based on different evaluation functions, Feature selection can be divided into three different types: Filter method[1], Wrapper method[2] and Embedded method. The filter type feature subset evaluation criterion is to evaluate the relationship between a single feature variable and a categorical variable according to different metrics, and select a feature whose evaluation value is higher than the threshold value, and the evaluation function is established independently of the subsequent operation. Therefore, the algorithm is fast and requires less computation, effectively removing noise features. However, this method is usually easy to select redundant features, and the classification result is poor. Moreover, the filter method is based on the threshold and the evaluation value of single feature and the categorical variable to determine the selection and screening of features, so the feature with low evaluation value may be removed according to the evaluation result, but some features with low evaluation value may have a good combined classification effect, so the feature should not be selected only according to single evaluation value, but the sorted classification effect. The literature[3] uses information gain to filter noise and redundant features to achieve dimensionality reduction and ease of data processing.
The wrapper method uses the performance of the classification model as the evaluation function of the feature subset to gain a higher accuracy, but feature selection is slow due to high computational complexity. In this paper[4], a hybrid feature selection method combining ReliefF and (particle swarm optimization algorithm, PSO) is proposed. Firstly, ReliefF is used as a feature preselector to filter out the small correlation features, and then PSO is used as the feature selection search algorithm. The accuracy classification of the SVM is used as an evaluation function to obtain an optimal feature subset. The embedded method formalizes the restrictions on features into a mathematical formula and combines them with the objective function of the classification model, so both can be completed in the same optimization process and feature selection is automatically performed in the process of training the classification model. The literature[5] proposes a mixed fault diagnosis method based on wavelet analysis and random forests. The literature[9] combines the filter mode and the Wrapper mode to screen specific features in protein disordered region prediction and gene selection, improving detection speed and detection accuracy.

In view of the above situation, this paper proposes a combination of filter and sequence floating forward search method. Firstly, different filter methods are used to sort the original feature series, avoiding the selection and screening of features only based on threshold and the evaluation value of single feature and categorical variable. The comprehensive effect is selected as the evaluation basis and the top ranked features is selected as the next set of original features, thereby the high evaluation characteristics obtained by the Filter method may be complementary to the low evaluation characteristics. On this basis, the sequence floating forward search method is used to select the optimal feature subset with the SVM as the classifier.

2. Background

2.1. **Information gain**

Information Gain[6] is an effective filter feature selection algorithm based on sample information, which can be used to measure the correlation between a feature and a category. The larger the information gain value of a feature, the higher the correlation between it and the category, and the greater its class distinguishing ability. If the information gain of feature A is defined as the difference between the original information demand and the new demand, and is represented by IG(A), so:

\[
IG(A) = \text{Info}(D) - \text{Info}_A(D) = -\sum_{i=1}^{m} p_i \text{lb}(p_i) - \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \text{Info}(D_j)
\]  

(2-1)

Where \( P_i = |D^i|/|D| \) represents the probability that any sample belongs to class \( C_i \), \( |D^i| \) is the number of samples belonging to class \( C_i \), \( |D| \) is the total number of samples, \( m \) is the number of sample categories, and \( v \) is the number of divided subsets \( D_j \).

2.2. **Fisher Score**

Fisher Score[7] is a feature selection method based on distance evaluation. The ratio of features is calculated according to Fisher's criterion. If the Fisher Score of the selected feature is larger, it means that the distance between different categories of the feature is greater, the distance between the same categories is smaller, and the classification ability of the feature is better. The Score formula of Fisher Score of the i-th feature \( S_i \) is as follows:

\[
S_i = \frac{\sum_{j=1}^{K} (u_{ij} - u_i)^2}{\sum_{j=1}^{K} p_{ij}^2}
\]  

(2-2)

Where \( K \) represents the number of categories; \( u_{ij} \) represents the average value of the i-th feature of the j-th sample point; \( u_i \) represents the average value of the i-th feature of all samples; \( p_{ij}^2 \) represents the variance of the i-th feature of the j-th sample, \((u_{ij} - u_i)^2\) represents the variance between classes, and \( p_{ij}^2 \) represents the variance within the class. The larger the \( S_i \), the stronger the classification ability of the feature.
2.3. SVM
The SVM[8] method is based on the VC dimension theory of statistical learning theory and the principle of structural risk minimization. It seeks the best compromise between model complexity and learning ability based on limited sample information, in order to obtain the best promotion ability. It is better to solve practical problems. Such as small samples, high latitudes, local minimum points and nonlinearity.

The principle of the SVM algorithm[10] is to map the features contained in the network packet to the high-latitude space, and find a hyperplane that can classify these features and achieve the best effect. It can correctly classify the data. And the spacing is the largest, so that the linearly inseparable data of the low latitude space is divided in the high latitude space.

2.4. Sequence floating search algorithm
The sequence floating search algorithms typically include sequential floating backward search and sequence floating forward search. The sequence floating backward search generally starts from the feature set and calculates and compares the evaluation value of the function \( J(\mathbf{X}_k^*\mathbf{x}) \) for each feature \( \mathbf{x} \) in the sequence. The feature that contributes the least to the evaluation function is deleted, so that the evaluation value of the remaining feature subset is optimized. However, the sequence floating forward search starts from the empty set, calculates and compares the function evaluation value for each feature to be added in the sequence, selects the feature with the largest evaluation value, and adds it to the empty set.

3. Feature selection algorithm
Although the Filter feature selection method is efficient and easy to calculate, the filter method evaluates the relationship between a single feature variable and a categorical variable according to different metrics, and removes the feature whose evaluation value is below than the threshold. But some features with low evaluation value may have a good combined classification effect.

This paper proposes a new hybrid feature selection method, in which Fisher Score and Information gain were used to sort the original feature set, respectively. and the top ranked features are selected as the next original feature set. Based on this, the sequence forward search method is used to select the optimal feature subset with SVM as the classifier.

3.1. The Filter stage
In Filter stage, Fisher Score and Information gain were used to sort the original feature set, respectively. Because the Information gain method and the Fisher score are different in the way of calculating the feature evaluation value, so it is necessary to introduce dimensionless to process the calculated value difference. The value of each feature should be at the same quantity level, so that the evaluation value of each feature is in the interval \([0, 1]\), and the negative influence of feature range on classification results should be eliminated. Taking the Information gain method as an example, the normalized evaluation of the i-th feature is expressed as:

\[
\text{std}_i^R = \frac{R_i - \min(SC_R)}{\max(SC_R) - \min(SC_R)} \quad (3-1)
\]

Where \( SC_R \) represents the score vector of the N-dimensional feature calculated by Information gain method, and \( R_i \) represents the Information gain value of the i-th feature.

In addition, the average weighting is used to calculate the composite ordering of features. The weighting formula is as follows:

\[
\text{std}_i = w_R \ast \text{std}_i^R + w_F \ast \text{std}_i^F \quad (3-2)
\]

Where \( \text{std}_i^R \) and \( \text{std}_i^F \) respectively represent the normalized evaluation value of Information gain method and fisher method, and \( w_R \) and \( w_F \) respectively represent the two evaluation weights, which are set as 0.5.
The basic idea of this stage is to use the Information gain method and the fisher method to calculate the feature evaluation value, after normalizing the evaluation values, the weighted evaluation value of each feature is calculated and the descending order is made according to the results.

3.2. The Wrapper stage
The top K features are selected based on the processing of the Filter stage as the original feature subset of the Wrapper stage. Those features are filtered by the strategy of sequential floating forward search. SVM is used as a classifier to construct a classification model to evaluate the pros and cons of feature subsets, and select the optimal feature subset set.

(1) Firstly, find a most important feature \( x^* \) from the set \( Y - X_K \) and add it to the current set \( X_K \).

(2) Then find one of the least important features \( X^- \) from the collection \( X_K \) each time. If the feature satisfies \( J(X_K - \{X^-\}) > J(X_{K-1}) \), then a subset with a size of \( k - 1 \) and better performance than the original subset is found. Similarly, a better subset of the size \( k - 2, k - 3, \ldots \) can be found.

(3) Keep repeating step 2 until the condition is not met, and then select the optimal feature subset.

4. Simulation experiment
This paper uses Python to build an experimental environment. The test uses two UCI data sets. The Ionosphere dataset contains 351 samples, 34 features, and 2 classifications; the Wisconsin Diagnostic Breast Cancer (WDBC) dataset contains 569 samples, 30 features, and 2 classifications.

4.1. Experimental result
This paper uses the All Feature, Fisher Score method, SFFS method and Filter-sffs method to select the features of the two UCI data sets. This paper compares performance probabilities of feature subsets selected by different search strategies through three performance metrics. The experimental results are shown in Table 4-1 to Table 4-3.

As can be seen from the second and fourth columns of each table, the performance of the classifier is improved after filtering irrelevant features through different feature selection methods. Meanwhile, the feature dimensions of the optimal feature subset screened by different feature selection methods are also listed in the third and fifth columns of the table. However, the filter method cannot effectively reduce the number of all features, which is because the filter method is a single feature selection method and cannot effectively remove redundant features. By comparing the performance of Accuracy, F1 and AUC, it can be seen that the Wrapper method can effectively improve the performance of the classifier. At the same time, after the feature selection method of SFFS, the number of features is greatly reduced.

| Feature selection method | Ionosphere | WDBC | Feature dimension | Feature dimension |
|--------------------------|------------|------|-------------------|-------------------|
| All Feature              | 93.78      | 97.34| 34                | 30                |
| Fisher score             | 95.29      | 96.92| 30                | 21                |
| SFFS                     | 97.17      | 97.87| 17                | 14                |
| Filter-SFFS              | 97.24      | 98.67| 15                | 13                |

| Feature selection method | Ionosphere | WDBC | Feature dimension |
|--------------------------|------------|------|-------------------|
| All Feature              | 95.21      | 96.77| 34                |
| Fisher score             | 95.79      | 96.92| 30                |
| SFFS                     | 97.59      | 97.12| 15                |
| Filter-SFFS              | 97.89      | 98.67| 16                |
Table 4-3  AUC

| Feature selection method | Ionosphere Feature dimension | WDBC Feature dimension |
|--------------------------|-----------------------------|------------------------|
| All Feature              | 97.87 34                    | 99.15 30               |
| Fisher score             | 98.12 25                    | 99.32 18               |
| SFFS                     | 98.67 16                    | 99.79 13               |
| Filter-SFFS              | 98.87 15                    | 99.79 15               |

Based on the comprehensive experimental results and the above analysis, the following conclusions can be drawn:

1. The SFFS method is superior to the pure Filter method in both the performance of the classifier and the number of selected features.

2. The Filter-SFFS method outperforms the SFFS method in terms of classifier performance and is superior to the SFFS method in most cases in terms of the number of features selected.

In Table 4-4, a comparison of the runtime of the SFFS method with the Filter-SFFS method is listed. It can be seen that the run-time of the Filter-SFFS method is less than the time of the SFFS method. Because The feature selection method reduces the number of features in the data in the filter phase, thereby reducing the computational complexity of the Wrapper phase and improving search efficiency.

Table 4-4  Runtime

| Feature selection method | Ionosphere | WDBC |
|--------------------------|------------|------|
| SFFS                     | 1220s      | 960s |
| Filter-SFFS              | 870s       | 630s |

5. Conclusion

The data processed by the network intrusion detection system contains too much noise and irrelevant features, which makes the classification effect unsatisfactory. This paper proposes a hybrid feature selection method combining filter method and wrapper method in which the high evaluation characteristics obtained by the filter method may be complementary to the low evaluation characteristics, so as to improve the accuracy of the classifier. Firstly, the original feature series is sorted by different filter methods, and the top ranked features are selected as the next original feature set. Based on this, the sequence forward search method is used to select the optimal feature subset with SVM as the classifier. The result shows that the proposed method is better than the pure filter method and SFFS method.

Acknowledgement

This work was supported by Nature Science Foundation of Fujian Province of P. R. China (No. 2019J01401), and the special fund for scientific and technological innovation of Fujian Agriculture and Forestry University (No. CXZX2016026, No. CXZX2016031).

References

[1] Liu H, Yu L. (2005) Towards integrating feature selection algorithms for classification and clustering. IEEE Tran on Knowledge and Data Engineering, 17: 491-502.
[2] Yu L, Liu H. (2004) Efficient feature selection via analysis of relevance and redundancy. Journal of Machine Learning Research, 10: 1205-1224.
[3] Tesfahun, Bhaskari D L. (2013) Intrusion detection using random forests classifier with SMOTE and feature reduction[A]. In: Cloud & Ubiquitous Computing & Emerging Technologies, 2013 International Conference on[C]. Pune. pp.127-132.
[4] Xiao Yan, Jiang Qi Gang, et al. (2016) Object-oriented land use classification based on hybrid feature selection of ReliefF and PSO. Journal of Agricultural Engineering, 32: 211-216 (in Chinese).
[5] Cai Jin Ding, Yan Ren Wu, et al. (2011) Fault diagnosis of power electronic circuit based on wavelet analysis and random forests algorithm. Journal of Electric Power Science and Technology, 26: 647-657 (in Chinese).

[6] Wang Yan Hua, Tian Sheng Feng, et al. (2009) Feature weight vector machine[J]. Electronics and Informatics Newspaper, 31: 514-518 (in chinese).

[7] Dash M, Liu H. (1997) Feature selection for classification[J]. Intelligent Data Analysis, 11: 131-156.

[8] Cortes, Vapnik V. (1995) Support vector networks[J]. Machine Learning, 20: 273-297.

[9] Hsu H H, Hsieh C W, et al. (2011) Hybrid feature by combining filters and wrappers[J]. Expert Systems with Applications, 38: 8144-8150.

[10] Fugate M, Gattiker Jr. (2002) Anomaly detection enhanced classification in computer intrusion detection. LNCS 2388. In: Berlin. pp. 186-197.