A stable feature selection method based on relevancy and redundancy

Peng Shen, Xiaoming Ding, Wenjun Ren and Shu Liu*

College of Computer and Information Science, Southwest University, ChongQing, China

*Corresponding author e-mail: xmding@swu.edu.cn

Abstract. In this paper, the characteristics of software defect prediction are analyzed from the perspective of machine learning. To solve the problem of some redundant or uncorrelated features in defect data sets, a stable feature selection method based on relevancy and redundancy (RRSFS) is proposed. RRSFS combines the redundancy between features and the correlation between features and classes to select the optimal subset. RRSFS not only reduces the cost of data operation in the prediction model, but also enhances the stability of feature selection algorithm.

1. Introduction

A feature is an abstract representation of an attribute extracted from an entity. It can be composed of data or text. The characteristics of this document refer to the software defect feature metric. Feature Selection (FS) is a process of constructing the machine learning model M based on the original feature set A to select the optimal feature subset B. Its purpose is to reduce the feature scale. Does not affect the performance P of the model. Can be expressed as: \( P_M(B) \approx P_M(A), B \subseteq A \).

1.1. Feature selection method classification

According to the difference of feature subset evaluation criteria, the feature selection methods can be divided into the following three categories [1]:

1. Filter Feature Selection (FFS), FFS filters out irrelevant features before the machine learning model is established. It aims to enhance the correlation between features and classes, reduce the correlation between features and features, and model construction algorithms. Nothing.
2. Wrapper Feature Selection (WFS), WFS combines the classification results of the machine learning model, and uses the classification performance as the evaluation criterion of the degree of importance of the feature.
3. Embedded Feature Selection (EMS), EMS combines feature selection with machine learning training process, and automatically selects features based on feature importance indicators during model training.

1.2. Feature selection framework

There are many studies on feature selection frameworks, and the typical one is the feature selection framework proposed by Dash [2] and others. The framework mainly includes the following four parts:

1. Subset generation, subset generation requires feature selection through specific search strategies. Specific search strategies are generally divided into: global optimal search strategy, random search
strategy and heuristic optimal search strategy \[3\]. The global optimal search strategy, which covers all the feature subsets, its time complexity reaches \(O(2^n)\), and \(n\) is the feature number. This strategy is easy to be restricted in the case of many feature numbers \[4\]. The random search strategy is a random search method based on feature weights. By limiting the number of iterations, the time complexity can be less than \(O(2^n)\), but the random model has uncertainty and can not guarantee the optimal result; heuristic optimal The search strategy pays attention to the correlation between features and the rapid stability of computation, and at the cost of reducing global optimality, the pursuit of efficiency is improved, and its time complexity is \(O(2^n)\). Each of them has its own advantages and disadvantages. The specific choice of which strategy needs to be combined with the actual situation of the researcher's own research problems.

2. Related work

In recent years, there have been many researches based on feature selection. Most of them are based on feature correlation or redundancy analysis, and then the importance of feature is divided to construct feature subsets. Among them, Sun et al. \[5\] studied the selection method based on genetic feature subsets, first using PCA for dimension reduction, and then using genetic algorithm (GAS) to ignore some less important feature vectors, and selecting from low-dimensional feature tables. Feature subset. Based on the fuzzy rough set model, Guo \[6\] proposed a fast feature subset selection method, which can effectively identify related features and effectively identify the redundancy between all features. Zheng et al. \[7\] proposed an unsupervised feature selection method for autonomous normative learning. This method automatically selects a subset of samples containing the most important samples to establish an initial feature selection model, and then generalizes it by introducing other important samples. Ability until a robust generalized feature selection model is established. Zhu et al. \[8\] proposed an unsupervised feature selection method based on subspace learning regularization. By embedding subspace learning regularization, principal component analysis (PCA) was incorporated into the sparse feature selection framework, and the feature selection model was improved. It is interpretable and solves the proposed objective function through an optimization algorithm to achieve fast convergence and obtain stable optimal results. Das \[9\] et al. proposed a feature selection algorithm based on improved binary differential evolution algorithm, which simultaneously optimized the set approximation accuracy of rough set theory and the derived feature based on relational algebra. The most relevant feature subset is selected throughout the feature set. Bhuyan et al. \[10\] proposed a classification-based sub-feature selection model that uses correlation coefficients and fuzzy models to select features and sub-features of different categories. There are also many feature selection studies, the commonality is that one-sided attention to the relevance or redundancy of features, and that the
importance of the classification features is directly related to it, but the stability of feature selection is neglected.

3. a stable feature selection method based on relevancy and redundancy

3.1. Relevancy Analysis

The feature correlation analysis aims to find out how much correlation between the feature data of the defect data and the category prediction, sort according to the degree of relevance, and then select the feature with the greatest correlation degree to construct the candidate feature subset. Feature correlations generally fall into three categories: strong correlation, weak correlation, and irrelevance [11]. It is assumed that the feature set of the defect prediction data set \( S_{DD} \) is \( F_X \), and \( X_i \) is the i-th feature. When \( X_i \) is an unrelated feature, the candidate feature subset \( F_{CX} = F_X - \{X_i\} \). The strong correlation feature directly affects the judgment of the category. The weak correlation feature has a limited influence on the judgment of the category, while the unrelated feature has no effect on the judgment of the category.

1) Mutual information, taking defect prediction as an example, the defect prediction data set \( S_{DD} = \{X_m \times n\} \), whose feature matrix is \( X_m \times n = [x_1, x_2, \ldots, x_m] \), assumes that the random feature variable \( S \) is The value is taken on \( X_m \times n \), and the corresponding probability is \( X_m \times n \). Then the uncertainty measure of \( S \) is defined as follows:

\[
H(S) = -\sum_{i=1}^{n} p(x_i) \cdot \log_2 p(x_i)
\]  

(1)

The class matrix in \( S_{DD} \) is \( Y_m \times 1 = [y_1, y_2, \ldots, y_m]^T \), and \( V \) is a random variable with a value on \( Y_m \times 1 \). The conditional entropy between \( S \) and \( V \) is:

\[
H(S|V) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \cdot \log_2 p(x|y)
\]  

(2)

Then the reduction of \( S \) uncertainty is called the mutual information between \( S \) and \( V \). The expression is as follows:

\[
I(S; V) = I(V: S) = H(S) - H(S|V) = H(V) - H(V|S)
\]  

(3)

The greater the mutual information between \( S \) and \( V \), the greater the correlation between the two random variables, and the method is suitable for correlation analysis.

2) Information gain, the information gain is to calculate the magnitude of the decrease of the entropy value of the characteristic variable \( S \) under the influence of the class variable \( V \), thereby evaluating the correlation between the feature and the class. The calculation formula is as follows:

\[
IG(S|Y) = H(S) - H(S|V)
\]  

(4)

3) Recursive feature deletion algorithm (RFE), the algorithm is a typical WFS model, which trains all features \( X_m \times n = [x_1, x_2, \ldots, x_m] \) in combination with the machine learning model \( M \). When the training achieves the expected classification effect or the training function is optimal, then the algorithm The weight minimum feature \( x_i \) will be selected and deleted; the second round of training will be entered. At this time, the feature number is \( n-1 \), and the next iteration is continued, and the feature number limit value is stopped. This algorithm is time consuming, but the accuracy is often high.

3.2. Redundancy Analysis

Feature redundancy analysis is based on correlation analysis. It aims to discover the degree of redundancy between features and features from the subset of candidate features under correlation analysis, and to remove redundant features with less relevance, so that not only the subsequent learning Model construction has less impact and reduces the amount of calculations. Based on the redundancy analysis, the representative feature selection algorithms are: the fast filtering feature selection algorithm (FCBF) proposed by Yu[12] and the maximum correlation minimum redundancy feature selection algorithm (mRMR) proposed by Peng [13] et al. The FCBF algorithm is based on symmetric uncertainty. It takes a feature that has a greater correlation with the label \( Y \) for a pair of highly correlated features, and removes the feature with less correlation with the label \( Y \), achieving reasonable de-redundancy; mRMR algorithm The aim is to maximize the correlation between features
and category labels Y, and to minimize the redundancy between features and features, so as to achieve reasonable redundancy.

3.3. Algorithm Description

The RRSFS combines the redundancy between features and features and the correlation between features and classes by two-stage multi-algorithm algorithm to select the optimal subset. The specific process of the algorithm is as follows:

Feature selection algorithm (RRSFS)

Input: SDD\(_n\) = \{X\(_m\times n\), Y\(_m\times 1\)\}, defect dataset; SDD\(_n\), the number of features n, label: Y

Output: Optimal feature subset \(F^*_\text{best}\)

1. begin
2. The data set is preprocessed, and the label \{Y, N\} is digitized as \{0, 1\}; then the feature data in SDD\(_n\) is normalized: \(X_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}\);
3. for k-fold crossValidation
4. Sampling the training samples to generate k sets of data;
5. Select three correlation feature selection algorithms IG, Gini, and RFE on the k-group data, and obtain the correlation feature subsets respectively \(F_{\text{IG}}, F_{\text{Gini}}, F_{\text{RFE}}\). Sort features according to feature importance in each subset, weighting features in each subset in order \(W_{f_i}^{\text{IG}}, W_{f_i}^{\text{Gini}}, W_{f_i}^{\text{RFE}}\);
6. Calculate the weighted sum of each feature in the three feature subsets: \(\text{sum}_{f_i} = w_{f_i}^{\text{IG}} + w_{f_i}^{\text{Gini}} + w_{f_i}^{\text{RFE}}\), sort(\(\text{sum}_{f_i}\)), Select the top N features to form a feature subset \(F_{\text{three}}\);
7. Find the correlation between each feature \(f_i, f_j\) in \(F_{\text{three}}\): \(\text{SU}(f_i, f_j) = \frac{2IG(f_i, f_j)}{H(f_i)+H(f_j)}\). Redundancy ordering is performed according to the correlation between features. A set of features \((f_i, f_j)\) with a higher degree of redundancy, retains the feature with the highest degree of relevance to the tag, and removes the feature with the least correlation, resulting in a feature subset \(F_{\text{new}}\);
8. Calculate the maximum correlation \(\max D_{F_{\text{new}}} = \frac{1}{|S|} \sum_{x_i \in S} F_{\text{new}}(x_i; Y)\), and the minimum redundancy \(\min R_C, R_C = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} C(x_i; x_j)\), obtain the feature subset \(F_{\text{mRMR}}\) based on redundancy, and fuse with \(F_{\text{best}}\) to obtain the optimal feature subset \(F^*_\text{best}\);
9. end for
10. return \(F^*_\text{best}\)

3.4. Experiment and Analysis

The experimental data in this chapter uses three data sets JM1, KC1, and MC2 in the defect data set MDP released by NASA's official website. Each data set is a function module of different NASA software systems collected by the bug tracking program. Static code metrics and their corresponding defect tag data. The defect rate of the data set is between 0.73% and 34.65%. In order to obtain objective data for analysis, this paper calculates the three major evaluation indicators Precision, Recall and F1-score in the defect prediction. The experimental results are shown in Table 1, Table 2, and Table 3.
Table 1. Comparison of experimental results of JM1 dataset feature selection method

| Method | Precision | Recall | F1-score |
|--------|-----------|--------|----------|
| *      | 0.81      | 0.82   | 0.74     |
| Gini   | 0.83      | 0.82   | 0.75     |
| RFE    | 0.68      | 0.82   | 0.74     |
| mRMR   | 0.81      | 0.82   | 0.74     |
| RRSFS  | 0.85      | 0.83   | 0.76     |

Table 2. Comparison of experimental results of KC1 dataset feature selection method

| Method | Precision | Recall | F1-score |
|--------|-----------|--------|----------|
| *      | 0.87      | 0.85   | 0.77     |
| Gini   | 0.86      | 0.75   | 0.78     |
| RFE    | 0.72      | 0.85   | 0.76     |
| mRMR   | 0.81      | 0.82   | 0.74     |
| RRSFS  | 0.87      | 0.85   | 0.78     |

Table 3. Comparison of experimental results of MC2 dataset feature selection method

| Method | Precision | Recall | F1-score |
|--------|-----------|--------|----------|
| *      | 0.79      | 0.68   | 0.57     |
| Gini   | 0.78      | 0.66   | 0.61     |
| RFE    | 0.57      | 0.63   | 0.53     |
| mRMR   | 0.77      | 0.63   | 0.61     |
| RRSFS  | 0.80      | 0.71   | 0.69     |

Through the average of the evaluation indicators on the three data sets, it is found that the average values of Precision, Recall and F1-score output by the RRSFS algorithm are 0.84, 0.81, and 0.74, respectively, which are higher than the precision of the *, Precision, Recall, and F1-score: 0.82, 0.78, 0.70; the values of Precision, Recall, and F1-score higher than the Gini output: 0.82, 0.74, 0.69; the values of Precision, Recall, and F1-score higher than the RFE output: 0.66, 0.76, 0.68; higher than the mRMR output The average values of Precision, Recall, and F1-score are: 0.79, 0.76, and 0.70. The experimental output data objectively shows that the performance of the RRSFS algorithm is better than other algorithms. The results fully demonstrate the importance of the algorithm in analyzing the relevance and redundancy of defect data features in the feature selection of software defect data sets.

Figure 1. Feature selection algorithm evaluation index analysis.

The polyline of several methods, which fully validates the feature selection algorithm based on correlation and redundancy analysis proposed in this chapter. It can be clearly seen in Figure 1 that the RRSFS algorithm has three higher evaluation index fold lines in the three subgraphs. The RRSFS algorithm effectively solves the problem of feature selection in software defect data sets. The stability of the RRSFS algorithm is fully demonstrated by comparison experiments on three data sets. In
general, the RRSFS method proposed in this chapter can achieve better results in software defect prediction, which also shows the effectiveness of the RRSFS method in the feature selection direction.

4. Conclusion

In this paper, a stable feature selection method based on relevancy and redundancy (RRSFS) is proposed. The performance of the algorithm is verified by comparison experiments on JM1, KC1 and MC2 data sets. The results show that RRSFS has the feature selection algorithm. A certain advantage, this will be a strong pre-data processing work for the construction of the defect prediction model. RRSFS effectively filters the feature subsets suitable for model construction in the feature selection algorithm, which effectively reduces the computational complexity in model construction.

References

[1] Zheng L, Diao R, Shen Q. Efficient feature selection using a self-adjusting harmony search algorithm [C] / Computational Intelligence, 2013: 167-174.
[2] Dash M, Liu H. Feature selection for classification [J]. Intelligent data analysis, 1997, 1(1-4): 131-156.
[3] Sun Z, Bebis G, Yuan X, et al. Genetic feature subset selection for gender classification: A comparison study [C] / Sixth IEEE Workshop on Applications of Computer Vision, 2002. (WACV 2002). Proceedings. IEEE, 2002: 165-170.
[4] Guo C, Zheng X. Feature subset selection approach based on fuzzy rough set for high-dimensional data [C] / 2014 IEEE international conference on granular computing (GrC). IEEE, 2014: 72-75.
[5] Zheng W, Zhu X, Wen G, et al. Unsupervised feature selection by self-paced learning regularization [J]. Pattern Recognition Letters, 2018.
[6] Zhu Y, Zhang X, Wang R, et al. Self-representation and PCA embedding for unsupervised feature selection [J]. World Wide Web, 2017.
[7] Das S, Chang C C, Das A K, et al. Feature Selection Based on Bi-objective Differential Evolution [J]. Journal of Computing Science and Engineering, 2017, 11(4): 130-141.
[8] Bhuyan H K, Huque M D S. Sub-Feature Selection Based Classification [C] / 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI). IEEE, 2018: 210-216.
[9] Qiu Like, Guo Zhongwen, Liu Qing, et al. Feature Selection Algorithm Based on Redundancy Analysis [J]. Journal of Beijing University of Posts and Telecommunications, 2017(01): 40-45.
[10] Yu L, Liu H. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution [C] / Twentieth International Conference on International Conference on Machine Learning. AAAI Press, 2003.
[11] Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2005 (8): 1226-1238.