An algorithm acceleration framework for correlation-based feature selection

Xuefeng Yan, *Yuqing* Zhang*, and *Arif Ali* Khan

College of Computer Science Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China

Abstract. Repeated calculations lead to a sharp increase in the time of correlation-based feature selection. Incremental iteration has been applied in some algorithms to improve the efficiency. However, the computational efficiency of correlation has generally be ignored. An algorithm acceleration framework for correlation-based feature selection (AFCFS) is proposed. In AFCFS, the criterion of the feature selection will be analyzed and reconstructed based on entropy granularity, and the algorithm structure will also be adjusted accordingly. Specifically, all repeated part of calculation will be saved in mapping tables and can be accessed in next time directly, so as to further reduce the calculation repetition rate and improve the efficiency. The experimental results show that AFCFS can greatly reduce the cost time of these algorithms, and keep the corresponding classification accuracy basically unchanged.

1 Introduction

Correction-based feature selection has been widely used in software defect prediction to construct the feature subset due to its simple principle and good stability. A good feature selection algorithm not only has higher classification accuracy, but also has less time complexity [1]. However, most existing methods focus on how to describe the correlations more accurately, and ignore the time cost. In most cases, there are often a large number of repeated calculations, and the time cost increases sharply with the expansion of the original datasets. Therefore, it is necessary to optimize the algorithm structure of this type of algorithm to improve their efficiencies without reducing its performance too much.

A few researchers have realized the expensive time cost of the feature selection based on correlation and optimized their algorithm structure accordingly, such as Feature selection with redundancy-complementariness dispersion (RCDFS)[2], Interaction Weight based Feature Selection algorithm (IWFS) [3], and fast greedy feature selection algorithm (FGS_KDE) [4].However, most of them only optimize the number of iterations by incremental iteration or weight update, and pay little attention on the calculation process of correlation itself. Several algorithms, such as FGS_KDE, have optimized the calculation of their basic correlation metric, but they are often only for a specific correlation metric, which is not universal and difficult to understand. Therefore, a simply and clear acceleration framework for this type of feature selection is proposed, we call it acceleration

* Corresponding author: 18361220609@163.com

framework for correlation-based feature selection (AFCFS). In this framework, all criteria based on correlation will be transformed into a combination of entropy or joint entropy, so that the repeated correction calculations can be find easily. All repeated calculations will be completed saved in mapping tables, the subsequent use will be transformed to the access to these tables rather than recalculation. In this way, the repeated calculations can be avoided to the greatest extent, so as to improve the operating efficiency without changing the performance of the algorithm itself.

2 Correlation-based feature selection

In the correlation-based feature selection, the core idea is to remove irrelevant features and redundant features. Irrelevant features refer to the features that cannot provide effective information for the classification task, while the redundant features refer to the features that cannot provide new information in addition to the selected features [3]. Therefore, the general way is maximizing the correlation between features and label but minimizing the correlation within features, which can be summarized into (1)

$$J(X_k) = corr(X_k, Y) - \sum_{X_j \in S} corr(X_j, X_k) \tag{1}$$

where $J(X_k)$ means the benefit to select the feature X_k, $corr(a,b)$ means the correlation between a and b, S is the set of the selected features, and X_j is the feature in S. Maybe some algorithms have only the first item or second item or some other forms, but the core idea is similar. Every time a new feature needs to be selected, the criterion value of all the unselected features needs to be calculated, and the largest one is added to S.

The correlation usually be measured by the information theory [5] metrics, such as mutual information, symmetric uncertainty and interactive mutual information. The most basic metrics in information theory are entropy and joint entropy. According to them, almost all other metrics can be derived, and their calculated as (2) and (3)

$$H(X) = -\sum_i^n p(x_i) \log p(x_i) \tag{2}$$

$$H(X,Y) = -\sum_i^n \sum_j^m p(x_i, y_j) \log p(x_i, y_j) \tag{3}$$

where X, Y is random variables, their possible values can be described as $X = \{x_1, x_2, \cdots, x_n\}$, $Y = \{y_1, y_2, \cdots, y_m\}$. $p(x_i)$ means the probability of that $X = x_i$, and $p(x_i, y_j)$ means the probability of that $X = x_i$ and $Y = y_j$ at the same time.

3 Algorithm acceleration framework

In this section, the algorithm acceleration framework proposed will be introduced, which mainly including the optimization on number of iterations and the optimization on correlation calculations.

3.1 Optimization on number of iterations

According to the formular (1), it can be seen a iteration on the selected feature set S , which represented as a summation symbol. Most papers given their criteria directly in this form. However, it should be noted that every time a new feature is selected, a "double loop" is actually carried out. The first loop represents all the candidate features, the criterion values of them is calculated. And the second loop represents the selected features. Therefore, the actual iteration number is $|S|(|F|-|S|)$, where $|F|$ is the total number of features and $|S|$ is the number of currently selected features. The incremental iteration method is used to eliminate the internal loop and realize the preliminary optimization.

The core idea of the incremental iteration is saving the summation symbol results of each feature in previous iteration to avoid recalculation. A "accumulation mapping table" will be designed, and each feature will have an item in this table. Every time a new feature is selected, the value corresponding to all unselected feature will be update unless it is the chosen one. In most cases, each unselected feature only needs to add its correlation with the new selected feature.

Through the above process, for each feature selected round, the number of iterations is $(|F|-|S|)$ actually, which becomes $1/|S|$ of the original. With the increase of the number of selected features, the efficiency improvement becomes more and more obvious.

3.2 Optimization on correlation calculations

By means of incremental iteration, the repeated traversal of selected features can be avoided. However, for some complex correlation measures, there are still many repeated calculations which need to be further optimized. For example, when calculating the mutual information between each feature and the classification label, it is necessary to calculate the entropy of the classification label itself. In fact, for a given data set, its value is uniquely determined, and the recalculations of the entropy of label are obviously unnecessary. Therefore, on the basis of incremental iterative optimization, it is necessary to optimize the correlation calculation with smaller granularity.

As the most basic metrics in information theory are entropy and joint entropy, we can take the calculation of entropy and joint entropy as the minimum "atomic computing" and make a new definition.

Definition 3.1 (atomic computing form) when a correlation expression contains only entropy calculation, joint entropy calculation and some arithmetic operation, without any other information theory measurement calculation, we call the expression "atomic computing form", in which entropy calculation and joint entropy calculation are called "atomic computing". All the correlation measurements based on information theory except these two are called "non atomic computing", such as conditional entropy, mutual information and symmetric uncertainty.

According to definition, we can transform the criteria of the correction-based feature selection into the "atomic computing form", and then a "calculation mapping table" can be designed by analyzing the repeated calculations on the "atomic computing". All of repeated calculations will be saved in this table. If subsequent calculation is needed, it can be directly taken from the table. Take mRMR [6] as an example, which is a typical correlation-based feature selection. According to the formular of mutual information as shown in (4), the criterion and corresponding "atomic computing form" of mRMR are (5) and (6). The $I(X;Y)$ represents the mutual information of X and Y , other symbols meanings are similar to those in (1) . Based on the (5), we can get that the entropy of each feature and label have been calculation many times and should be saved in the "calculation mapping table".

What's more, as each time a new feature is selected, the mutual information of each feature and label is also calculated repeatedly, it should be save in this table as well. That is to say, this "calculation mapping table" can not only save the "atomic computing", but also the "non atomic computing", as long as it has been calculated repeatedly. The purpose of transforming criteria into "atomic computing form" is to help us find smaller granularity repetitive computing units, so as to eliminate repeated calculations as much as possible.

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{4}$$

$$J(X_k) = I(X_k;Y) - \frac{1}{|S|}\sum_{X_j \in S} I(X_k;X_j) \tag{5}$$

$$J(X_k) = H(X_k) + H(Y) - H(X_k;Y) - \frac{1}{|S|}\sum_{X_j \in S} H(X_k) + H(X_j) - H(X_k;X_j) \tag{6}$$

4 Experiment and analysis

4.1 Experiment design

In this experiment, three typical correlation feature selection algorithms, mRMR[6], IWFS[3] and CFR[7], are used to verify the performance and time efficiency before and after the optimization of the proposed acceleration framework. Ten datasets from PROMISE Repository [8] are used, which are widely used in the field of software defect prediction.

Because the result of feature selection is in the form of feature subset rather than a certain value, we can use the classifier to train the feature subset obtained by the feature selection algorithm, and judge whether the result of feature selection is consistent by the classification accuracy of the final classifier. In order to avoid the influence of classifiers, two typical classifier algorithms, K-Nearest Neighbor (KNN) and Naive Bayes (NB) classifiers, are used. Since the time of feature selection is related to the number of features selected, we set the number of features selected as 1,2,3,⋯,20. In order to further reduce the time error, each parameter value of feature selection number corresponds to 10 times of ten-fold cross-validation.

4.1 Experiment result and analysis

In order to show the experimental results more intuitively, this paper describes the changes of classification accuracy and time cost with the change of the number of selected features, in the form of line graph. Different node shapes with the same color represent the same feature selection algorithm before and after optimization. Red, blue and green correspond to mRMR, IWFS and CFR respectively. Hollow circle corresponds to before optimization, and cross node corresponds to optimized. In order to facilitate the distinction, we record the optimized algorithm as FastMRMR, FastIWFS and FastCFR.

As the number of selected features increases, the classification accuracy of the selected feature subsets before and after optimization of the three algorithms is shown in Fig.1. Among them, the first 10 subgraphs correspond to the result on the 10 datasets based on the KNN classifier, and the last 10 subgraphs correspond to that on the NB classifier. It can be observed that, on the whole, whether for KNN classifier or NB classifier, lines with different colors are quite different, and lines of the same color can be roughly overlapped.

The former means that the difference between feature subsets obtained by different algorithms can be reflected by the corresponding classification accuracy rate, while the latter means that for the same feature selection algorithm, after the optimization of the acceleration framework proposed in this paper, the accelerated algorithm will not have too much impact on the results of feature selection algorithm. That is because that, both the optimization on iteration and the optimization of correlation calculation only transform the criteria into the access operation of tables. Therefore, it will not affect the results of the algorithm.

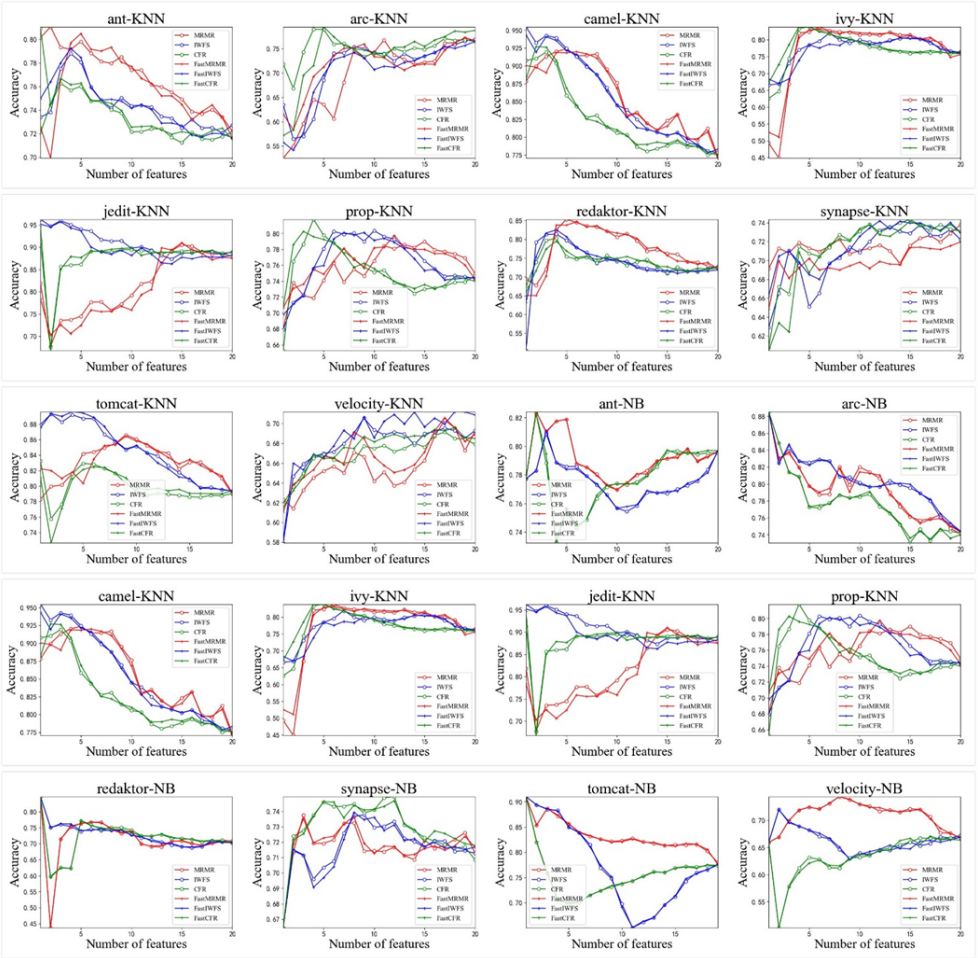

Fig. 1. Average classification accuracy versus different number of selected features.

As the number of selected features increases, the time cost of feature selection before and after optimization of the three algorithms is shown in Fig.2. Because we compare the time consumed by feature selection itself, which is independent of the classifier, for each data set, each algorithm does not need to distinguish classifiers. Instead, the time spent in selecting the same number of features on each dataset can be averaged. Finally, 10 subgraphs are obtained.

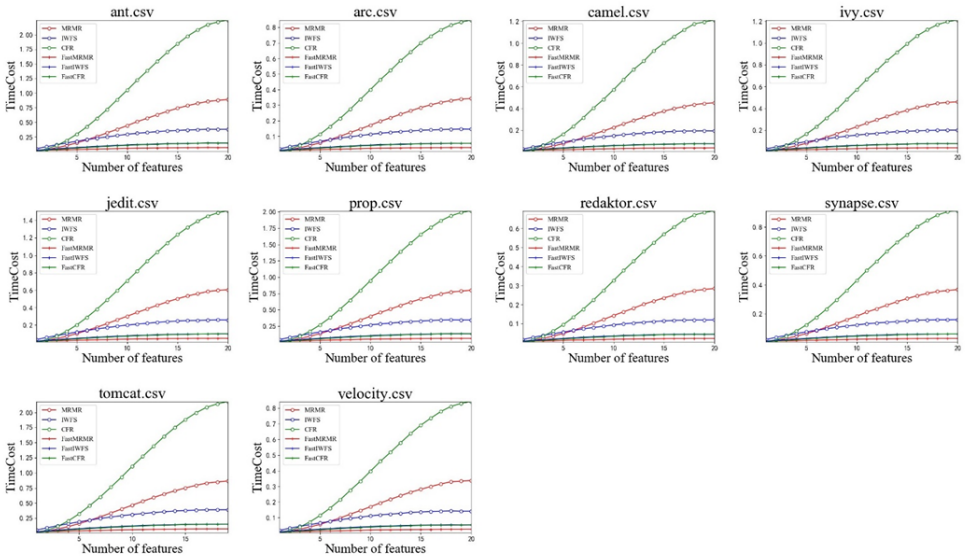

Fig. 2. Average time cost versus different number of selected features.

It can be found from Fig.2 that the time cost differences between the three original algorithms and the three optimized algorithms are basically the same on 10 datasets. Before optimization, the efficiency of CFR algorithm is the lowest, followed by mRMR, and IWFS is the best. After optimization, FastIWFS and FastCFR basically coincide, while FastMRMR is the best. After optimization, all the three algorithms can significantly reduce the time cost, and the optimization effect on CFR is the most obvious.

It can be explained by the principle of the acceleration framework and the structure of the three algorithms themselves. The core idea of the acceleration framework is to eliminate the "double loop" structure in the algorithm, and to eliminate the repeated calculation based on the "atomic computing" granularity in the correlation. As for IWFS, it has used the weight update rule to eliminate the "double loop", which is similar to the optimization of the number of iterations in the framework, so even the original IWFS, can give a higher efficiency than other original algorithms. It can be observed, after optimization, the time cost of the IWFS still have an obvious reduction, which can be attributed to the optimization on correction calculations. And for mRMR and CFR, which do not have any optimization themselves, performanced a very high time cost before optimization and had more obvious reductions in time cost. As the mainly correlation metric adopted in the mRMR is mutual information, which is simpler than the interactive mutual information in CFR and the complementary mutual information in IWFS, so it is faster than these two, no matter before or after optimization. What's more, we can also get that, if the complexity of correlation metric is similar, after optimization, the time cost will be similar too, just like IWFS and CFR, because almost all the repeated calculations have been eliminated.

5 Conclusion

Feature selection can effectively select the most representative features in the process of software defect prediction, and reduce the cost of data acquisition and model training. However, the existing feature selection algorithms based on correlation often have low efficiency due to the large number of repeated calculations. Therefore, it is necessary to propose a systematic framework to optimize this kind of algorithm.

An algorithm acceleration framework for correlation-based feature selection algorithm is proposed. It not only includes the optimization of the number of iterations, but also optimizes the repeated calculation in the correlation calculation. By analyzing and reconstructing the judgement criteria, the unnecessary iterative process and repeated correlation calculations are found and extracted. By designing the accumulation mapping table and the calculation mapping table, the original repeated calculation can be transformed into the access operation of the mapping table. Generally speaking, the access efficiency is much higher than that of these repeated calculations. Therefore, the proposed framework can effectively improve the efficiency of this kind of algorithm while maintaining the performance of the original algorithm unchanged. Three typical correlation-based feature selection algorithms mRMR, IWFS and CFR were verified by experiments. The experimental results show that, after the optimization, the classification performance of the corresponding feature subsets of the three algorithms are basically unchanged, and the time efficiency is greatly improved. What's more, the improvement effect becomes more and more significant with the increase of the number of selected features.

In the current framework, we have defined a complete scheme for the optimization of correlation-based feature selection, and researchers can adjust the structure of a specific algorithm to improve its efficiency. In the future, we will study how to transform this process into automatic process, that is, automatic generation of algorithm pseudo code.

This work is supported by the 13th Five-Year Planning Equipment Pre-Research Program under Grant Nos. 41401010401 and 41401010201.

References

1. Jie Cai, Jiawei Luo, Shulin Wang, Sheng Yang. Feature selection in machine learning: A new perspective[J]. NEUROCOMPUTING, 2018, 300(jul.26):70-79.
2. Chen Z , Wu C , Zhang Y , et al. Feature selection with redundancy-complementariness dispersion[J]. Knowledge-Based Systems, 2015, 89(NOV.):203-217.
3. Zeng Z , Zhang H , Zhang R , et al. A novel feature selection method considering feature interaction[J]. Pattern Recognition, 2015, 48(8):2656-2666.
4. Zhang Jinghong. Kernel density estimation entropy of mixed data and fast greedy feature selection algorithm [D]. Zhejiang University, 2017.
5. C.E. Shannon, A mathematical theory of communication, ACM SIGMOBILE Mobile Comput. Commun. Rev. 5 (1) (2001) 3 – 55.
6. H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (8) (2005) 1226 – 1238.
7. Gao W, Hu L, Zhang P, et al. Feature selection considering the composition of feature relevancy[J]. Pattern Recognition Letters, 2018, 112(SEP.1):70-74.
8. M. Jureczko and L. Madeyski, "Towards identifying software project clusters with regard to defect prediction," in Proc. 6th Int.Conf. Predictive Models Softw. Eng., 2010, Art. no. 9.