An Optimized Feature Subset Selection Method for Network Flows based on Machine Learning

Xiaoyan Zhang[1, a], Jingbo Xia[1,b], and Ruixin Li[2,c]

[1]Tan KahKee College, Xiamen University, Xiamen, Fujian, China

[2] Institute of Information and Navigation, Air Force Engineering University, Xi'an, Shaanxi, China

[a]zxyxjwxxj@163.com, [b]jingbo@sina.com, [c]sunflower@163.coml

Keywords: Network flows classification; Feature subset selection; Machine learning; Dynamic block.

Abstract. As the foundation of network cognition, management and optimizing, the classification of network traffic is making a significant difference in resource scheduling, safety analysis and future tendency prediction. Feature subset selection (FSS) based on machine learning plays an important role in classification problems, especially dealing with high-dimensional data like network traffic flows. To realize accurate traffic classification at lower price of evaluations, a hybrid feature subset selection method is proposed on the base of dynamic block, the size of which is flexible according to the classification performance. The performances are examined a few experiments. Our theoretical analysis and experimental observations reveal that the proposed method consumes fewer evaluations with similar or even better classification performance.

Introduction

Traffic classification is currently a significant challenge for network monitoring and management. Feature selection is an effective method to realize dimension reduction and decrease redundant information. Most of the input traffic features contain irrelevant or redundant information which may degrade the classification accuracy and increase the computational complexity. The purpose of FSS method is to extract the exact feature subset that can represent the whole feature dataset with fewer variables.

Filters and wrappers are the two major directions of FSS approach [1]. The filters evaluate the features by their intrinsic properties. As filters are independent from the specific classifier, they have the advantages of general and highly computational efficient. However, the relationships between the features are hardly to be discovered, which may lead to imperfect classification result. Wrappers employ the information of the classifier to find the best feature subset with much more classification accuracy. They can explore the relationship among features, but at the price of expensive computation on the evaluation of feature space. Moreover, the classification performance of the feature subset selected by wrappers is strictly related to the specific classifier, and it is hard to get similar performance with other classifiers.

In this paper, we aim to improve the efficiency and accuracy of the existing FSS methods. We put forward a FSS algorithm with the proposed concept of dynamic block to save evaluations and CUP time. Our proposal is to dispose the features at the level of block whose size is flexible according to the classification performance of every round of selection. We combine filter-based ranking measure with wrapper-based FSS method to take the advantages of them both. The features that are selected into the subset may interchange with the following selected ones.

The rest of the paper is organized as follows. The following section presents a set of FSS methods in detail. In section 3 the proposed concept of dynamic block and the feature selection algorithm are introduced. The 4th section contains the experiment and the corresponding analysis. In section 5 we provide a summary of the paper and future work.

Previous Work on Feature Subset Selection

Filter-based Feature Subset Selection

Filter-based feature ranking techniques use statistical measures to assign a score to each feature and rank the resulting features according to the value of the scores [2]. As common sense, the classification performance may be enhanced with the number growing of the selected features. However, the truth shows the opposite side.

Filter-based feature subset selection uses measurements as evaluation criteria to evaluate the quality of feature subsets. Filter-based FSS methods select subsets of variables as a pre-processing step, which is independent of the chosen predictor. Correlation-based feature selection (CFS) method is widely used in previous literatures. CFS algorithm tries to find a subset of features not only to reduce the dimensions of the dataset but also to improve the classification accuracy. It defines a merit for each selected subset of features. The merit is based on the hypothesis that a promising subset involves those features which are uncorrelated or less correlated to each other but correlated to the class label. The merit is mathematically defined as formula (1).

$$M = \frac{kr_{cf}}{\sqrt{k + k(k-1)r_{ff}}} \tag{1}$$

where M is the heuristic merit of a feature subset containing k features, r_{cf} is the mean feature-class correlation and r_{ff} is the average feature−feature correlation.

Wrapper-based Feature Subset Selection

Wrapper-based feature subset selection methods evaluate each feature subset with the classification performance, such as accuracy or F-measure. A learning machine is applied as a black-box to score subsets of features according to their predictive power. They consider all the subsets but evaluate their merits by building a classification model only with the selected features and considering the performance of model. Sequential forward selection (SFS) method is one of the wrapper algorithms that are widely used. It starts from an empty set and adds one feature at a time. For the stopping criterion, the procedure stops when the test result starts to get worse or the number of features reaches a predefined threshold [3] .

Conceptually, wrapper-based feature subset evaluation is very simple: the chosen feature subset is used as the basis for a classification model, and then the performance of this model is used as the score for that feature subset [4]. However, wrappers are often criticized because they seem to be brute force methods requiring massive amount of computation. The computational complexity is too high to be O(n2) even with efficient search strategies like Bestfirst and Greedy.

Hybrid Filter-wrapper Feature Subset Selection

Hybrid method is a more recent approach and a promising direction in the feature selection field. It uses the ranking information obtained using filter methods to guide the search in the optimization algorithms used by wrapper methods.

Recently the literatures have contained numerous references to the use of hybrid selection algorithms. Xie included a feature ranking in a sequential forward search method with the application of the F-score measure to rank the features [5] , while Peng added a random sampling method to choose features from the ranking . Zhang and Bonilla-Huerta proposed the similar methods including a Relieff estimation based ranking, which were also applied to compress the searching space.

Except for the good classification performance of hybrid methods on some aspects, the drawback is also obvious. The complexity of the wrapper search in hybrid methods is still intractable because of the large consumption of wrapper evaluations.

IWSS based on Dynamic Block

Development of IWSS

Incremental wrapper-based feature subset selection (IWSS) was presented by Ruiz R as a canonical method . The features that satisfy the relevance criterion would be selected into the feature subset. The advantage of this method is the low complexity of O(n) in the number of wrapper evaluations. However, the problem is that the features are kept into the subset once they are selected, even those which have poor performance of correlation with the latter selected ones. As a result, the size of selected feature subset cannot be restricted in a small range.

The algorithm of IWSS with replacement (IWSSr) was proposed to alleviate this problem. During the selection of a feature, the addition of the feature into the subset is not the only matter that is concerned, but also the interchange with one of the included features. As a result, the feature that has already been selected would be drove away from the subset when the performance of its combination with the latter selected features gets worse. We can find that the selected subset is more representative, but it is also obvious that all the features are disposed once at a time. Therefore, the process is inefficient to deal with high-dimensional problems.

Bermejo P put forward the method of IWSS with re-ranking (IWSSrR) by the introduced concept of block . The proposed algorithm can select features at the block level and improve the efficiency of the algorithm greatly. The size of block is initialized at the beginning and cannot be changed during the procedure of feature selecting. The size of block is also required to be large enough to give some freedom to the wrapper algorithm, which may increase extra evaluations.

Dynamic Block

We know that the ranking step sets the features with stronger classification ability to the front part of the line. As a result, the features ranked at the bottom are less efficient and few of them would be selected into the final feature subset. The bigger the size of block is, the more evaluations and computational resources are needed. It is reasonable to attenuate the size of block when the density of selected features goes down.

On the base of the analysis above, we propose the concept of dynamic block B. Dynamic block is defined as the number of features disposed at a time which is dynamically changing according to the result of feature selection. It is initiated at the beginning of the algorithm like IWSSrR. Under certain circumstances as follows, the size of dynamic block attenuates to the half.

1) The density of features selected in the blocks decreases.

$$D_n < D_{n-1}, D_n = \frac{|S_n| - |S_{n-1}|}{B_n} \tag{2}$$

where D_n is the density of features selected, B_n and S_n refer to the size of block and the number of all the selected features at the *n* round of feature selection.

2) The classification accuracy of the feature subset decreases.

From the description above, it can be seen that the attenuation of dynamic block is the penalty strategy to save evaluations when the density of features or classification performance goes down.

Algorithm Description

The main objective addressed in this work is to make a further step to reduce the evaluation consumption. The proposed algorithm IWSS based on dynamic block (IWSS-DB) composes of two stages, which are ranking and iterated feature selection. The stage of ranking requires O(n) filter evaluations. In this study, we use Information Gain (IG) which is one of the fastest attribute ranking methods to evaluate the predictive attributes. The stage of feature selection starts from the first block of features. The features selected from each round of selection by the wrapper method are added into the feature subset S. Then, the wrapper method runs again over the next block of features but taking into account the features already included in the feature subset S. That means the algorithm does not

only select new features into the subset, but also interchange with the ones that already included in S with redundant information. If the classification accuracy of the newly updated feature subset cannot exceed the former one, the feature subset will remain unchanged. After every round of selection, dynamic block will decide whether to attenuate according to the rules on section 3.2. Then, the block moves to the next part of the rank. This process keeps iterating until no more feature is added into the feature subset or dynamic block attenuates to zero.

Experiments

Experimental Setup

This paper applies traffic flow dataset of Moore to the experiment. The dataset contains 12 traffic categories. There are 24863 flows with 248 features in the dataset. We take one method out of each kind of approaches mentioned in Section 2 as comparisons, which are CFS, IG, SFS and IWSSrR. As for IG, we select the top 10 ranked features as the selected feature subset.

The experiment adopts Naivebayes as the classifier. The 10-fold cross-validation method is used to estimate the data. We take the average of the experiment which is conducted ten times as the result.

Results and Analysis

To test the advantages of IWSS-DB algorithm, we design the experiments to check the corresponding performance. In this part of experiment, we compare the performance of the proposed algorithm IWSS-DB with the other representative FSS methods. In this study, the initialized size of block is set to be 50.

Fig. 1 shows the classification performance of the feature subsets obtained through different FSS algorithms. From the comparison we can briefly come to the conclusion that filter based methods, which are CFS and IG in the experiment, perform worse than the wrapper and hybrid methods on all the aspects. The indexes of precision, recall and F-measure stand for the classification ability of the FSS method. It can be seen that SFS and IWSS-DB have better classification performance than IWSSrR. The situation goes the same to the indexes of the area under the curve of ROC and PRC, which are frequently used to quantitatively assess the identification model.

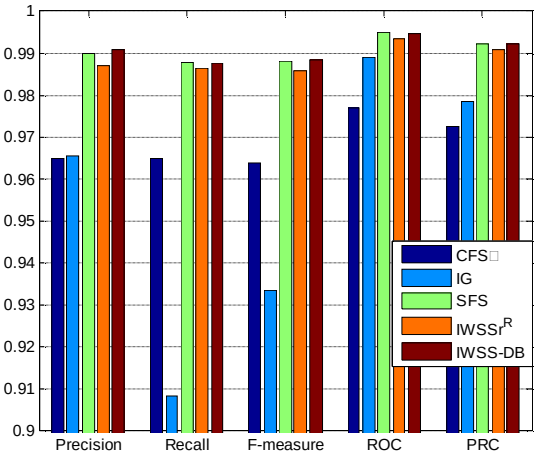

Fig.1. Performances of different algorithms

Table 1 shows the evaluations of different algorithms. It can be seen that filter-based methods consume fewer evaluations than the others, and that is the typical advantage of filters. The wrapper-based method SFS always selects the best feature subset with great classification performance, meanwhile it expends the most evaluations. Compared with SFS, the two hybrid methods IWSSr^R and IWSS-DB can decrease the evaluations to the half and maintain the classification performance on the same level. Furthermore, IWSS-DB consumes fewer evaluations than IWSSr^R with better performance, and sometimes even better than SFS on some aspects.

Table 1. Evaluations of all the different algorithms.

	CFS	IG	SFS	IWSSr^R	IWSS-DB
Evaluations	1479	248	6351	3925	3215

To observe the varying trend of classification performance along with the increasing of the evaluations, we record the result of every iteration of IWSSr^R and IWSS-DB. The comparison of performance is shown in Fig.2.

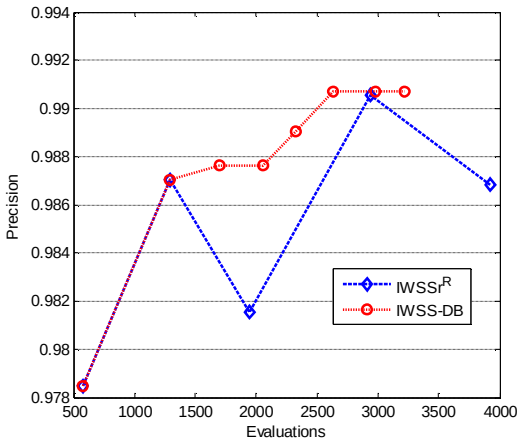

Fig.2. Varying trend of classification performance

It can be seen that the two methods' performance of the first two iterations are the same, that is because the dynamic block does not meet the condition of attenuation and the selected feature subset of the two selection rounds are the same. It is also clear that the classification performance of IWSS–DB keeps increasing, while that of IWSSr^R decreases at some iterations. Moreover, the number of evaluation of IWSS–DB stops increasing at the best performance, and IWSS–DB always has better performance at the same evaluations consumption.

Conclusions

To achieve efficient classification of network traffic flows, we have proposed a hybrid FSS method based on dynamic block whose size is flexible according to the classification accuracy. The experiment showed that the proposed IWSS-DB algorithm was able to select the feature subset of great classification performance with fewer evaluations which can lower computational complexity. From the description and analysis, we can see that the attenuation coefficient of dynamic block is also an important factor to the performance of the proposed method. As future work, our research will be focused on the research of its impact.

Acknowledgments

This work is supported by the Shaanxi Natural Science Foundation [2016JM6073].

References

[1] F.Amiri, M.Y. Rezaeiousefi, C. Lucas: Mutual information-based feature selection for intrusion detection systems. Journal of Network & Computer Applications Vol.34(4) (2011), p. 1184-1199
[2] S. M. Vieira, J.M.C.Sousa, U. Kaymak: Fuzzy criteria for feature selection. Fuzzy Sets & Systems Vol.189(1) (2012), p.1-18
[3] H.H. Hsu, C.W. Hsieh, M. D. Lu: Hybrid feature selection by combining filters and wrappers. Expert Systems with Applications Vol.38(7) (2011), p.8144-8150.
[4] R.Wald, T. M. Khoshgoftaar, A. Napolitano: Stability of Filter- and Wrapper-Based Feature Subset Selection. IEEE International Conference on TAI (2013), p.374-380

[5] J. Xie, C. Wang: Using support vector machines with a novel hybrid feature selection method for diagnosis of erythemato-squamous diseases. Expert Systems with Applications An International Journal. Vol.38(5) (2011), p.5809-5815