A combined feature selection method based on clustering in intrusion detection

Ting Huang [1, a], Wenbo Chen[1, b] and Ruisheng Zhang[1, c]

[1] School of Information Science & Engineering, Lanzhou University, Lanzhou 730000, China.
[a]Corresponding author: huangt14@lzu.edu.cn,[b]chenwb@lzu.edu.cn,[c]zhangrs@lzu.edu.cn

Keywords: Feature selection, Relief algorism, k-means clustering, Relief+k-means.

Abstract. The rapid development of information technology generates high dimension and large scale data, which puts severer challenges to network security. Feature selection is proposed for the reduction of data dimension so that features of original data are utmostly retained and improving the effectiveness of data processing. This paper proposes a new method of feature extraction by combining two algorithms. Firstly, removes some noise and irrelevant features after researching for correlation between features and categories; Secondly, furtherly optimize subsets to select key features through feature selection algorithm, whose Evaluate Function is based on clustering algorithm. KDDCUP9910% datasets is used to testify the experiment, whose result shows the method guarantees effective detection rate and reducing the data dimension effectively at the same time , the effectiveness of data detection is improved.

1. Introduction

In the past decades, data information grows fastly as a result of the rapid development of network information technology, while the features of data samples becomes more and more. Especially in the aspect of network access data, whose data sets needs to be dealt with is very large and so is its object set attribute, so the analysis of the data processing is very time consuming, it's a deadly weakness to the development in network security. The performance of intrusion detection system largely depends on how to choose the features of the data samples. For high dimension and large scale data sets processed by intrusion detection system, it usually needs a dimension reduction to data with high-dimensional features, select the optimal feature subset to eliminate the influence of the redundant features, and retain the function of original data utmostly to improve the classification capacity of the intrusion detection system.

Feature selection is a common method of the dimension reduction, mainly used in the field of machine learning and pattern recognition, the role of this method is selecting the optimal feature subset of a data set to construct classification model ,which can make prediction precision close to or better than the original feature model, thus improve the generalization ability ,understandability and calculation efficiency of the model, reduce the frequency of "dimension disaster" at the same time [1]. In recent years, there are various different definitions of feature selection in literatures [1-4], although a bit different on the definition of feature selection, they all indicate the goal of feature selection is to find a minimum feature subset which can effectively identify the target. Therefore, the definition of feature selection mainly considers classification accuracy and class distribution. Literature [4] presents a basic framework of feature selection, which shown in figure 1.

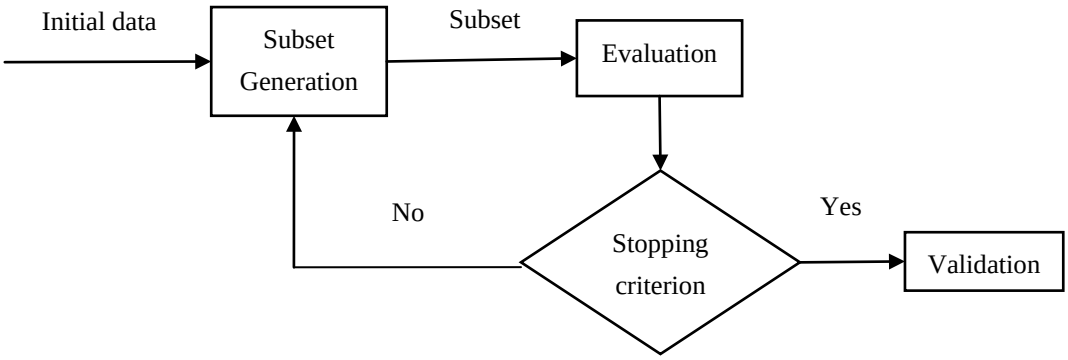

Fig. 1 Framework of feature selection

As shown above, the process of feature selection mainly includes four steps: subset generation, subset evaluation, stopping criterion, validation [5-6]. Subset generation refers to search feature subset and produce a subset, which provides feature subset for evaluation function; Subset evaluation mainly through evaluation function to evaluate the merits of the feature subset; Stopping criterion is related to evaluation function in the subset evaluation, it usually expressed as a threshold, the process of searching new feature subset may stop when the value of feature subset evaluation function reaches the threshold; In the process of the validation, feature subsets which selected should be verified validity on data set.

The process of feature selection methods according to the search feature subspace can be divided into complete search, heuristic search and random search. Complete search method mainly has the breadth-first search, branch and bound search, beam search and best first search, etc. Heuristic search method mainly includes sequential forward selection, sequential backward selection, bidirectional search, plus-L minus-R selection, sequential floating search, etc. Random method includes random generation plus sequential selection, simulated annealing algorithm, tabu search algorithm, genetic algorithm, etc. [7]. According to the working principle of evaluation function, feature selection method also is divided into two kinds: filter and wrapper [8]. Filter method is usually independent of the subsequent learning algorithm, the algorithm directly using the training set to evaluate, get the best characteristic by its statistical performance, it can fast obtain the optimal subset, but the disadvantage is that performance of the assessment has the large deviation with subsequent learning algorithm; Wrapper method is using the training accuracy of subsequent learning algorithms to evaluate feature subset, its advantage lies in the small deviation, but large amount calculation of the algorithm is not suitable for large data sets.

Combining with the traits of filter and wrapper method, this paper proposes a combined form of feature selection methods, firstly estimate features and classes by filtering algorithm, preliminary eliminate some noise and irrelevant features; Then, using the method of heuristic search combine with clustering algorithm to optimize feature subset and select key features.

2. Related Work

2.1 Relief algorithm

Relief algorithm determines feature subset with the weight of data feature. The algorithm gives different weights to feature according to the correlation between features and categories. Optimal subset is selected by eliminating features, whose weights are less than a certain threshold. Relief algorithm could only process problems of two-category, so Kononenko improved it and proposed ReliefF algorithm in 1994, which could process multi-class problems and complemented a solution to data missing.

The algorithm works as follows: Extract a data sample Ri from training set, search k neighbor samples respectively from sample sets both of similar-class and different-class to Ri, calculate the feature weight and update it[10]. Iterate the procedure and determine the correlation weight between feature and category, then obtain correlation between every feature and category in each sample example .Sort features according to their weight, determine the validity by the set threshold. Construct feature subset with features have weight above the threshold, or a certain amount of features with higher weight, eliminate the others and eliminate invalid features .The formula w[A] to update weights is shown as follows (1) and (2):

$$w[A] = w[A] - \sum_{j=1}^{k} diff(A, R_i, H_j)/mk + D(c) \quad (1)$$

$$\text{where, } D(c) = \sum_{C \neq class(R_i)} \left[\frac{P(C)}{1-P(class(R_i))} \sum_{j=1}^{k} diff(A, R_i, M_j(C))\right]/mk \quad (2)$$

in which, m indicates sampling frequency of data sample,Mj(C) indicates the number j neighbor sample from different-class-set C,P(C) indicates the proportion of target samples to the total sample population in a certain class. class(Ri) indicates the class of Ri , diff(A, Ri, Rj) is to calculate the distance of two sample examples about feature A[10].

ReliefF algorithm is a relatively better filtering evaluation algorithm for its expandability, effectiveness, stability and evaluation efficiency, as well as available to all kinds of data types. It can well eliminate irrelevant features because its evaluation process to features mainly considers the correlation between features and categories. But the demerit of ReliefF algorithm is that the correlation between features was ignored when designed, and fail to eliminate redundant feature. The algorithm gives higher weight to features that more related to classes, regardless of redundancy with other features[10,11].

2.2 K-means feature selection algorithm

A n-dimension data set has 2N-1 subsets. It will be hard to find the optimal feature subset if N gets larger. For this reason, algorithms above focus mainly on calculating the effection of single feature to classification in each feature subset.

This algorithm uses a random searching feature selection method. Unlike calculating the effection single feature to category, this algorithm research the effection that correlation between features to category. In this algorithm, define a N-length array S to show the selection state in a feature subset：

$S=<S1,S2,\ldots,Si>, \quad Si \in \{0,1\}, \quad 1 \leqslant i \leqslant N.$

In which, Si=0 means the feature i is not selected, otherwise, Si=1, it's selected. For instance, if all features are selected in a feature subset, it's feature array $S=<1,1,\ldots,1>$.

The key factor of a feature selection method is the cost function used to evaluate the selected feature subset, which means the effectiveness of algorithm depends largely on the definition of cost function. In wrapper method, categorizer is usually used as evaluation criteria. Taking categorizer as evaluation criteria causes problems of effectiveness and over-fitting of some special categorizer, the cost function adopted in this method independent to the final categorizer. The selected feature subset has better performance, so has the effectiveness. This algorithm cluster the training data by k-means clustering algorithm and use the clustering result as cost function to evaluate a subset[12].

K-means clustering algorithm divides n data samples into k cluster in accordance to the following criteria: Samples in the same cluster have high similarity, samples in different cluster have low similarity .This algorithm works as follows: select k data samples as original cluster center, calculate the distance of samples to the cluster centers. Clustering samples according to the distance. Recalculate the cluster center until members of every cluster unchanged.

In this paper, feature selection algorithm is mainly used to distinguish normal and abnormal data, so the number of cluster set as 2. The cost function for each data x:

$$f(x) = \begin{cases} 1, \text{if } l(x) \neq b(x) \\ 0, \quad \text{otherwise,} \end{cases} \tag{3}$$

where, l(x) indicates sample x belongs to the cluster, b(x) indicates sample x's actual catagory. The cost of a feature subset is calculated with the function:

$$C(s) = \sum_{i=1}^{n} f(xi) \tag{4}$$

where, n means the number of example in training set.

Thus, this feature selection method is concluded as follows: Calculate the cost C： S→q(0≤q≤n) of a given feature subset array, and alter a bit in the array, if the new cost Cn<C, process last step ,end it when Cn>C.

This algorithm can't ensure an optimal subset for the feature subset selected by a random method. A suboptimum subset can be selected by multiple iteration of the algorithm and sort all the M feature subsets.

2.3 Combination algorithm

The two algorithms have its own merits and demerits. Relief algorithm can well filter irrelevant feature but doesn’t take correlation between features into consideration, which fail to filter redundant feature; the feature selection algorithm based on k-means takes correlation between features into consideration but its randomness can’t guarantee a selected optimal subset, which needs more operations to verify an optimal subset, it’s definitely a complex work. Seen from the following experiment in section 3.3 ,the detection rate is relatively stable because of the basically consistent features selected by Relief algorithm; but features selected by k-means feature selection algorithm in section 3.4 is variable when detected and evaluated, so its unstable detection rate needs a further evaluation to select better feature subset.

The paper reaches the goal of dimension reduction of dataset by combining Relief algorithm and k-means feature selection. Select key feature subsets by Relief algorithm and then optimal by k-means feature selection. Simple combination of two algorithms doesn’t mean an ideal result. Compared with experiment in section 3.5, it can be seen that feature subset selected by Relief+k-means feature selection algorithm has certain changes because of the unstability caused by randomness of k-means feature selection, and its detection performance is no so stable. The paper will be revised accordingly.

The process of combined algorithm in the paper is as follows: calculate each feature weight by Relief algorithm and select high weight feature, useless feature is preliminarily excluded from the feature subset, then set the feature subset S1 as an initial one of k-means feature selection algorithm, change a certain bit of S1 and calculate its cost, select the optimal set and repeat the previous step until subset unchanged.

Algorithm pseudo-code is as follows:

（1） Using Relief algorithm get a feature subset S1 and set it as initial for the next step
（2） Calculate its cost C(S1)
（3） Set optimal subset Sbest=S1
（4） Find neighbor subsets of Sbest and make sure only one bit different with Sbest
（5） Calculate the cost of of all neighbor subsets
（6） Select the optimal Si
（7） If C(Sbest)≥C(Si)
（8） Sbest=Si
（9） else

（10）repeat (4) until Sbest unchanged
（11）end.

3. Experiment result

3.1 Experiment tools and data sets

The paper uses marked KDDCUP9910% training date set [13] as experimental date set, which includes 490,000 items of data, the scale of normal and abnormal data is 1:4.In the data set, data has 41 dimension features and 1 category tag. There are 9 discrete features in 41 dimension features of data, others are continuous features, which include character, numeric, and Boolean feature. So, the data usually needs a pre-processing when using it for experiment, firstly map character features into numerical values, then standardize and normalize the data sets to generate input data. Considering the large datasize in the paper, the processing data set for feature selection is actually proportional extracted 10% of original data; actually abnormal data is less than normal data, so the normal and abnormal data designed 4:1 to simulate this case, randomly extract training set data and test set data for 30,000 and 20,000 as testing data of feature selection result. Classifier adopts decision tree algorithm and R x64 3.1.1 software for programming processing of data.

3.2 Evaluation criteria

(1)Accuracy A: Ratio of correctly determined category samples D to total sample number D

$$A = \frac{D}{Z} \times 100\%$$

(2)False Positive FP: Ratio of normal sample (which are misinterpreted as abnormal sample) number T to total normal sample number R.

$$FP = \frac{T}{R} \times 100\%$$

(3) False Negative FN: Ratio of abnormal sample (which are misinterpreted as normal sample) number F to total abnormal sample number Q.

$$FN = \frac{F}{Q} \times 100\%$$

3.3Relief algorithm preferences

It mainly to determine value m, k of Relief algorithm and weight threshold in the experiment, because algorithm in various data sets needs different preferences, it is no way to determine recognized value [14-16],thus conduct selection training of parameter through cross –validation, and select ideal parameter according to experiment result.

(1) m selection experiment

k is fixed to 1, m are 10,20,50,100,300 respectively, Accuracy ,False Positive and False Negative on table 1.m value of 10 is more appropriate seen from table 1.

Table 1 The m's selection of Relief algorithm

k=1	A	FP	FN
m=10	99.765	0.125	0.675
m=20	99.64	0.063	1.55
m=50	99.64	0.063	1.55
m=100	99.64	0.063	1.55
m=300	99.64	0.063	1.55

(2) k selection experiment

m is fixed to 10, k are 1, 10, 20, 50, 100 respectively, statistical Accuracy ,False Positive and False Negative on table 2.k value of 10 is more appropriate seen from table 2.

Table 2 The k's selection of Relief algorithm

m=10	A	FP	FN
k=1	99.64	0.063	1.55
k=5	99.64	0.063	1.55
k=10	99.71	0.063	1.2
k=50	99.64	0.063	1.55
k=100	99.64	0.063	1.55

(3) Weight threshold selection experiment

As can be seen from above experiments, m is fixed to 10, k value of 10,feature weight threshold are 0, 0.01, 0.02, 0.04, 0.05 respectively, statistical Accuracy, False Positive and False Negative on table 2.Weight threshold value of 0.02 is more appropriate seen from table 3.

Table 3 The threshold selection of Relief algorithm

threshold	A	FP	FN
0	99.64	0.063	1.55
0.01	99.64	0.063	1.55
0.02	99.79	0.044	0.725
0.04	99.79	0.044	0.8
0.05	99.68	0.119	1.125

(4)Relief algorithm result verification

Initial samples are selected randomly when run algorithm, so it leads to the selected features has certain randomness even if fixed value m and k. Run 10 times with fixed value m and k, evaluate the selected feature subsets, the detection condition shows in the following figure 2.It can be seen that key features remain unchanged though each feature subset differs slightly after running. The relatively stable detection rate demonstrates key features of algorithm selection are basically consistent. It needs a further feature eliminating due to possible redundant feature.

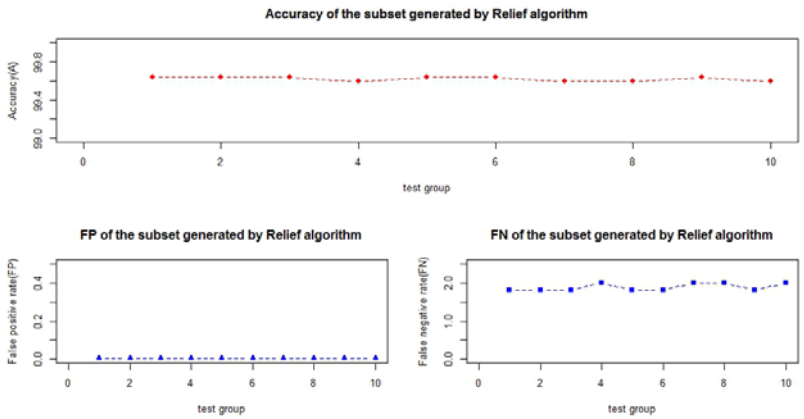

Fig. 2 Relief algorithm result verification

3.4 k-means feature selection algorithm experiment

For the feature subset by k-means feature selection algorithm is not the only one, it runs 10 time operations, detection result is shown in the following figure 3.We can see from the figure that the classification function has merits ,because of the randomness of feature selection of this algorithm. It needs more operations to compare to select better feature subset, but optimal subset can't be guaranteed.

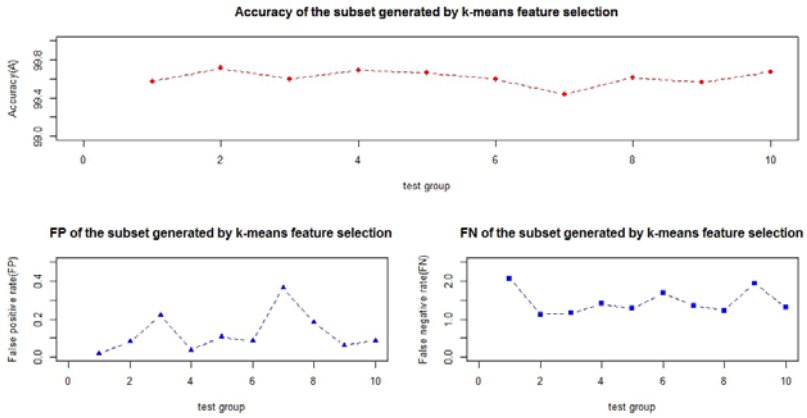

Fig. 3 the result of k-means feature selection algorithm

3.5 Algorithm comparison

If experiment on combined Relief algorithm and k-means feature selection algorithm, run iterative operation 10 times as well, because of its randomness. The experiment result shown in figure 4.

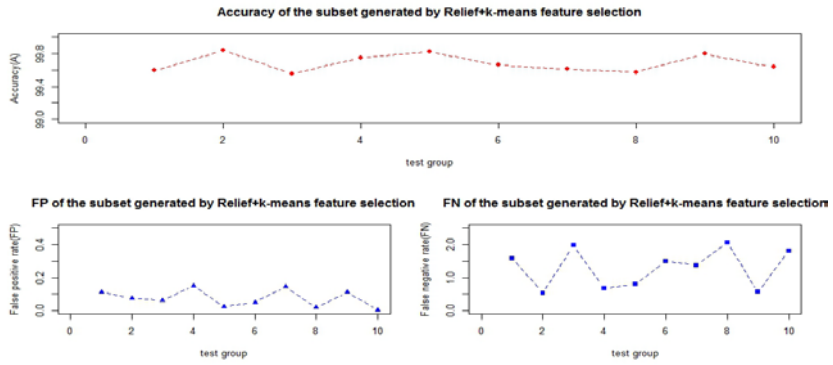

Fig. 4 the result of Relief+k-means feature selectionmethod

It can be seen that instability of classification function still exist, so the improved combined algorithm of the paper can solve this problem. Run operation 10 times with method of the paper, then get feature subsets and do classification verification, the result shows in figure 5.

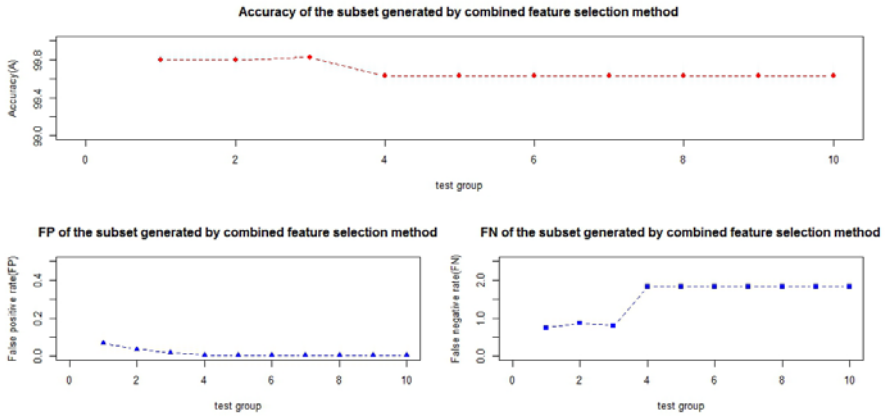

Fig. 5 the result of modified Relief+k-means feature selection method

We can see from the figure above that detection result of improved combining algorithm remains stable, while it has a high accuracy.

At last compare the sets evaluation results of original feature, Relief algorithm selection feature, Relief+k-means algorithm feature selection, and feature subsets of the method of the paper and other

related literatures, run 5-cross-validation, the detection result is shown in table 4, take optimal feature subset from feature subsets of Relief algorithm, k-means feature selection algorithm and Relief+k-means feature selection algorithm.

Table 4 The result of compare subsets

Feature selection method	Dimension	Modeling time(s)	Testing time(s)	A(%)	FP(%)	FN(%)
original feature	41	4.23	1.36	99.60	0.033	1.84
Relief	26	2.89	0.99	99.69	0.059	1.3
k-means	23	2.39	0.79	99.69	0.068	1.26
Relief+k-means	19	2.36	0.72	99.75	0.094	0.81
Paper method	12	1.35	0.43	99.77	0.084	0.84
Literature[17]	21	2.19	0.74	99.62	0.04	1.73
Literature[18]	13	1.38	0.45	99.68	0.097	1.23
Literature[19]	13	1.47	0.48	99.76	0.07	0.91

K-means feature selection algorithm can well guarantees high accuracy rate while obviously reduces dimension. Compared with simply combine Relief and k-means feature selection algorithm, the method of the paper reduces the randomness of the algorithm, and gets better feature subsets rather than simple usage of one algorithm. The method of the paper also has strengths in effectively reducing feature dimension of data sets in comparison with methods of other related literatures, while it has simply achieved algorithm and better classification result.

4. Conclusions

Network intrusion detection system faces the problems of large datasize, high feature dimension and so on, which bring difficulties to effectiveness and real-time of algorithms. This paper proposes a combining intrusion selection algorithm aiming at real-time needs of anomaly intrusion detection on network environment, focusing on invalid and redundant features filtering .Experimental result shows that the proposed algorithm combining the traits of two feature algorithms can filter irrelevant and redundant features effectively, with the help of computational results of decision tree classifier, it can reduce computing time on the basis of effectively reducing feature amount and guaranteeing classification precision.

5. References

[1] Kira K, Rendell L A. The feature selection problem:Traditional methods and a new algorithm[C]. Proc of the 9th National Conf on Artificial Intelligence. Menlo Park,1992: 129-134.

[2] John G H, Kohavi R, Pfleger K. Irrelevant features and the subset selection problem[C]. Proc of the 11th Int Conf on Machine Learning. New Brunswick, 1994: 121-129.

[3] Koller D, Sahami M. Toward optimal feature selection[C].Proc of Int Conf on Machine Learning. Bari, 1996: 284- 292.

[4] Manoranjan Dash, Huan Liu. Feature selection for classification[J]. Intelligent Data Analysis, 1997, 1(3):131-156.

[5] Liu H, Motoda H. Feature selection for knowledge discovery and data mining[M]. Boston: Kluwer Academic Publishers, 1998.

[6] Molina L C, Llu´ıs Belanche,`Angela Nebot. Feature selection algorithms: A survey and experimental evaluation[R]. Barcelona: Universitat Politecnicade Catalunya, 2002.

[7] Yao X, Wang X D, Zhang Y X, Quan W. Summary of feature selection algorithms [J]. Control and Decision, 2012, 11:161-166 +192.

[8] Langley P. Selection of relevant features in machine learning[C]. Proc of the AAAI Fall Symposium on Relevance. New Orleans, 1994: 1-5.
[9] Kononenko, I. Estimation attributes: analysis and extensions of Relief[C]. Proceedings of the 1994 European Conference on Machine Learning, 1994:171-182.
[10] Huang Y,McCullagh P J,Black N D. An optimization of ReliefF for classification in large datasets[J]. Data &Knowledge Engineering,2009,68(11) : 1348－1356.
[11] Gilad － Bachrach R,Navot A,Tishby N . Margin based feature selection-theory and algorithms[J].Proceedings of the twenty-first international conference on Machine learning. ACM,2004,1(1) : 43.
[12] Kang, S. H. and Kim, K. J. A feature selection approach to find optimal feature subsets for the network intrusion detection system[J]. Cluster Computing-the Journal of Networks Software Tools and Applications.2016,19(1): 325-333.
[13] KDD Cup 1999: Available on:http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html. (2007)
[14] Pang Z, Zhu D, Chen D, et.al. A Computer-Aided Diagnosis System for Dynamic Contrast-Enhanced MR Images Based on Level Set Segmentation and ReliefF Feature Selection[J]. Computational and Mathematical Methods in Medicine, 2015, (10):450531.
[15] Beretta L,Santaniello A. Implementing ReliefF filters to extract meaningful features from genetic lifetime datasets[J]. Journal of biomedical informatics, 2011, 44(2):361-369.
[16] Song Q,Ni J,Wang G. A fast clustering－based feature subset selection algorithm for high -dimensional data[J]. Knowledge and Data Engineering, IEEE Transactions on, 2013, 25(1):1-14.
[17] Xiao L Z, Liu Y X. Step feature selection algorithm for intrusion detection[J]. Computer Engineering and Application, 2010, 11:81-84 + 87.
[18] Wu X N, Peng X J, Yang Y Y, Fang K. Two-level feature selection method based on SVM for intrusion detection [J]. Journal on Communications, 2015, 2015:23 - 30.
[19] Li A, Jiang J H. A feature selection method for intrusion detection using the adaptive genetic algorithm [J]. Applied Science and Technology, 2016 01:53 + 49-59.