IG-C4.5:An Improved Feature Selection Method Based on Information Gain

Kai Luo
State Key Laboratory of Mathematical Engineering and Advanced Computing
Zhengzhou, Henan, China
e-mail: kakastudy@126.com

MeiJuan Yin
State Key Laboratory of Mathematical Engineering and Advanced Computing
Zhengzhou, Henan, China
e-mail: raindot_ymj@163.com

JunYong Luo
State Key Laboratory of Mathematical Engineering and Advanced Computing
Zhengzhou, Henan, China
e-mail: ljunyong@163.com

JianLin Li
Jiuquan Satellite Launch Center of China
Jiuquan, Gansu, China
e-mail: lijianlinstudy@gmail.com

Abstract—Feature selection is an important means to solve the problem of dimension reduction in anomaly network traffic detection. Focusing on the problem of traditional feature selection algorithm based on information gain neglect the redundancy between features, this paper proposes an improved feature selection method combining CFS and C4.5 algorithms—IG-C4.5. In the improved algorithm, the irrelevant features and the redundant features were removed by adding the judgments of redundancy between features, which effectively simplified the feature subset. The experimental results show that the proposed algorithm can effectively find the feature subsets with good separability, which results in the low-dimensional data and the good classification accuracy.

Keywords-Feature selection; Information gain; Redundant features; dimension reduction; Anomaly traffic detection

I. INTRODUCTION

In recent years, network traffic anomaly detection technology is developing rapidly with the network security issues become increasingly prominent. Selecting the appropriate feature subset is the key to ensure the proper accuracy, reliability and the ability of detection of the anomaly detection system.

The original high-dimensional feature space data objects often contain many redundant features and irrelevant features that may reduce the efficiency and accuracy of the algorithm and greatly increase the learning and training time and the space complexity. Therefore, the most critical issue facing researchers usually is using feature selection algorithm to find the feature subsets with good separability, so as to achieve dimensionality reduction and reduce the time and space complexity of machine learning.

II. RELATED WORK

Feature selection algorithms broadly fall into three categories: the filter model, the wrapper model, and the hybrid model.

In the filter model, a good feature set is selected as a result of pre-processing based on properties of the data itself and independent of the machine learning algorithm. Typically, an independent criterion is used in algorithms of the filter model. Some popular independent criteria are distance measures, information measures, dependency measures, and consistency measures [1]. For example, Relief [2] and its extended algorithm ReliefF [3] and IRelief [4] use Euclidean distance to measure the importance of feature subset. Dash and Liu [5] carry out a study of consistency measure at 2003. Hall [6] propose the Correlation-based Feature Selection (CFS) algorithm. Because of the information entropy does not require the distribution of data is known in advance and it can evaluate the uncertainty between features effectively. Feature selection algorithms based on information entropy is a research hotspot in recent years. Pen and Ding [7] study how to select good features according to the maximal statistical dependency criterion based on mutual information. Wang [8] use dynamic mutual information as evaluation criteria and eliminates irrelevance and redundancy features by approximate Markov Blanket. Zhang [9] propose a novel multi-label feature selection algorithm based on information entropy. As with many advantages, the algorithm proposed in this paper is also based on information entropy.

The wrapper model requires one predetermined mining algorithm and uses its performance as the evaluation criterion. Wrapper methods based on SVM have been widely studied in machine-learning community. SVM-RFE [10] uses a backward feature elimination scheme to recursively remove insignificant features from subsets of features. R-SVM [11] is a recursive support vector machine algorithm to analyze noisy high-throughput proteomics and microarray data. Liu [12] develops a novel similarity kernel and propose a novel method based on the new kernel that iteratively selects features that provides the maximum benefit for classification.

Both filter and wrapper models have advantages and drawbacks. Filter models are generally less-computationally intensive than wrapped models. However, they tend to miss complementary features that individually do not separate the data well. To take advantage of the above two models and avoid the pre-specification of a stopping criterion, the hybrid model is recently proposed to handle large data sets [13,14]. For example, A two phase procedure to select salient features for classification committees has been presented in [15]. Elimination of clearly redundant features in the filter approach-based first phase of the procedure speeds up the genetic search executed in the second, wrapper approach-based, phase of the designing process.

III. BASIS THEORY AND CONCEPTS

A. Information Gain(IG)

Entropy is a measure of the uncertainty of a random variable. The entropy of a variable X [16] is defined as

$$H(X) = -\sum_i p(x_i)\log_2(p(x_i)) \quad (1)$$

The entropy of X after observing values of another variable Y [16] is defined as

$$H(X|Y) = -\sum_j p(y_j)\sum_i p(x_i|y_j)\log_2(p(x_i|y_j)) \quad (2)$$

Where $p(x_i)$ is the prior probabilities for all values of X, and $p(x_i|y_j)$ is the posterior probabilities of X given the values of Y.

The amount by which the entropy of X decreases reflects additional information about X provided by Y and is called information gain [16], defined as

$$IG(X|Y) = H(X) - H(X|Y) \quad (3)$$

B. Symmetrical Uncertainty(SU)

Symmetrical Uncertainty [16] is defined as

$$SU(X,Y) = 2[IG(X|Y)/(H(X)+H(X|Y))] \quad (4)$$

It compensates for information gain's bias toward features with more values and restricts its values to the range [0,1]. A value of 1 indicates that knowing the values of either feature completely predicts the values of the other; a value of 0 indicates that X and Y are independent.

C. Correlation Feature Selection(CFS)

Pearson's correlation [6] is often used for correlation-based feature selection, defined as

$$Merit_s = k\bar{r}_{cf} / (\sqrt{k + k(k-1)\bar{r}_{ff}}) \quad (5)$$

Where $Merit_s$ is the heuristic "merit" of a feature subset S containing k features, $\bar{r}_{cf}$ the average feature-class correlation, and $\bar{r}_{ff}$ the average feature-feature intercorrelation.

D. C4.5

C4.5 algorithm is an algorithm that is used to form a decision tree. This algorithm is a classification and prediction methods are very powerful and famous. Decision tree is useful to explore the data, find the hidden relationship between the number of candidate input variables to the target variables.

IV. IG-C4.5 ALGORITHM

A. The problem of traditional algorithm based on IG

Traditional selection algorithms based on information gain usually only focuse on searching for relevant features and neglect the redundancy between features, namely after a feature is selected, if another feature has relevance, this feature is often not necessary to be selected. This leads to the feature subset exists a lot of redundant features, which affect the performance of classifier.

B. Description of the algorithm

By improving the problem of traditional feature selection algorithm based on information gain (neglect the redundancy between features), this paper proposes an improved feature selection method combining CFS and C4.5 algorithms (IG-C4.5). Its main idea is: It uses the CFS algorithm to decide the best subsets for a subset selected by information gain and uses the decision tree algorithm C4.5 to select the final best subset among the best subsets across the accuracy of C4.5 algorithm.

C. Process of the algorithm

input: $S(F_1, F_2, ..., F_N, C)$ // a training data set
K // a predefined threshold
output: S_{best} // a selected subset

Step 1. For $i = 1$ *to* N, calculate IG_i for F_i; if $IG_i > K$, append F_i to $S_{IG-best}$.

Step 2. For each feature X in $S_{IG-best}$, calculate Merit. Where k is the number of $S_{IG-best}$, $\bar{r}_{cf}$ the average feature-class correlation (evaluated by IG) and $\bar{r}_{ff}$ the average feature-feature intercorrelation (evaluated by SU).

Step 3. If array *Merit* =NULL, end; otherwise, append the feature with the maximum *Merit* to S'.

Step 4. Evaluate S' by C4.5 across the accuracy V', if $V' > V_{best}$, then: $S_{best} = S'$, $V_{best} = V'$ and go to Step 2; otherwise remove it from *Merit* and go to Step 3. (The initial value of V_{best} is 0)

V. KDD CUP 99 DATA SET DESCRIPTION

Our experiment is based on the KDD CUP 99 dataset. The KDD CUP 99 dataset is a standard set of data collected through the 1998 DARPA intrusion detection evaluation program at the MIT Lincoln Labs. The data set includes a wide variety of intrusions simulated in a military network environment. The simulated attacks are grouped into four categories: Denial of Service(DoS), Probe, UserToRoot (U2R), RootToLocal (R2L).

The training dataset consisted of 494,021 records among which 97,277 (19.69%) were normal, 391,458 (79.24%) DOS, 4,107 (0.83%) Probe, 1,126 (0.23%) R2L and 52 (0.01%) U2R. In each connection are 41 attributes describing different features of the connection and a label assigned to each either as an attack type or as normal.

Due to the uneven distribution of the data set type, selecting a record randomly is possible to select only one or a few types of the dataset, those types of a small number may not be selected. It is disadvantageous to reflect the real network environment. In building the dataset, sample types should be as average as possible. Therefore, this paper uses a principle that the minimum number of samples to be selected first to build the datasets, namely: to ensure that the sample with a small number can be selected.

VI. EXPERIMENT RESULTS AND DISCUSSIONS

After some comparative experiments, the threshold value of information gain K is set to 0.2 in this paper.

From Table I , we can see that the IG algorithm chooses 23 features as the feature set, while IG-C4.5 selects 9 features. Table II details the relevant features.

TABLE I. SELECTED FEATURES BY DIFFERENT METHODS

Method	The number of features	Feature NO
IG	23	2、3、4、5、6、10、12、23、24、25、27、28、29、30、32、33、34、35、36、38、39、40、41
IG-C4.5	9	3、4、5、23、24、30、33、35、40

TABLE II. NAME OF RELEVANT FEATURES

Feature NO	Feature Name
2	protocol_type
3	service
4	flag
5	src_bytes
6	dst_bytes
10	host
12	logged_in
23	count
24	srv_count
25	serror_rate
27	rerror_rate
28	srv_rerror_rate
29	same_srv_rate
30	diff_srv_rate
32	dst_host_count
33	dst_host_srv_count
34	dst_host_same_srv_rate
35	dst_host_diff_srv_rate
36	dst_host_same_src_port_rate
38	dst_host_serror_rate
39	dst_host_srv_serror_rate
40	dst_host_rerror_rate
41	dst_host_srv_rerror_rate

For these 3 feature set: the original feature set S with 41 attributes, $S_{IG-best}$ (selected by IG) with 23 attributes and S_{best} (selected by IG-C4.5) with 9 attributes, We use the C4.5 algorithm to evaluate and contrast. The experiment results are listed in Table III. The performance measures used are the error, true positive's rate and false positive's rate, defined as:

- TP shows the overall percentage of attacks detected.
- FP shows the false positive rate, that is the proportion of normal patterns erroneously classified as attacks.
- Error shows the overall percentage error rate for the two classes (Normal and Attack).

TABLE III. VALIDATION RESULTS USING DIFFERENT FEATURE SET

Feature Set	Error	TP	FP
S	13.1%	99.01%	1.67%
$S_{IG-Best}$	9.6%	99.51%	1.55%
S_{best}	1.3%	99.97%	0.69%

Table III shows the performance of IG-C4.5 algorithm has significant improvements compared to IG. For example, the false positive rate is reduced from 1.55% to 0.69% and the error rate is reduced from 9.6% to 1.3%. It shows the 9 attributes selected by IG-C4.5 algorithm is very effective for the C4.5 algorithm in anomaly detection.

VII. CONCLUSIONS AND FUTURE WORK

In this paper, we propose an improved feature selection method combining CFS and C4.5 algorithms which strives to reduce redundancy between features while maintaining information gain in selecting appropriate features. The experiment results show our method can efficiently achieve high degree of dimensionality reduction and enhance accuracy with selected features. The good results obtained using only 9 features implies that only part of features, rather than of the original 41, will be required in an anomaly detection system.

This work can be extended in various directions. We plan to explore a line of research that focuses on comparison of different classifiers and also of other methods of dimensionality reduction.

REFERENCES

[1] Liu H, Yu L. Toward integrating feature selection algorithms for classification and clustering[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(4):491-502.

[2] Kira K, Rendell L. A practical approach to feature selection [C] // Proc of the 9th International Conference on Machine Learning, 1992, 249-256.

[3] Kononenko I. Estimation attributes: analysis and extensions of RELIEF [C] // Proc of the European Conference on Machine Learning, Catania, Italy, 1994, 171-182.

[4] Sun Y. Iterative RELIEF for Feature Weighting: Algorithms, Theories, and Applications [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(6): 1035-1051.

[5] Dash M, Liu H. Consistency-based search in feature selection [J]. Artificial Intelligence, 2003, 151(1-2): 155-176.

[6] Mark A. Hall. Correlation-based Feature Subset Selection for Machine Learning[D]. Hamilton, NewZealand: University of Waikato, 1999.

[7] 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 and Machine Intelligence, 2005, 27(8): 1226-1238.

[8] Wang X, Yao X, Zhang Y, et al. Intelligent Science and Intelligent Data Engineering[M]. Springer Berlin Heidelberg, 2012:226-233.

[9] Zhang Z-H et al. Multi-Label Feature Selection Algorithm Based on Information Entropy [J]. Journal of Computer Research and Development, 2013, 50(6): 1177-1184.

[10] Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machines [J]. Machine Learning, 2002, 46(1-3): 389-422.

[11] Zhang X, Lu X, Shi Q, et al. Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data[J]. BMC Bioinformatics, 2006, 7.

[12] Liu J, Ranka S, Kahveci T. Classification and feature selection algorithms for multi-class CGH data [J]. Bioinformatics, 2008, 24: i86-i95.

[13] Sebban M, Nock R. A hybrid filter/wrapper approach of feature selection using information theory [J]. Pattern Recognition, 2002, 35(4): 835-846.

[14] Das S. Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection [C] // Proc of the 18th International Conference on Machine Learning, San Francisco, CA, USA, 2001, 74-81.

[15] Bacauskiene M, Verikas A et al. A feature selection technique for generation of classification committees and its application to categorization of laryngeal images [J]. Pattern Recognition, 2009, 42(5): 645-654.

[16] Liu H, Yu L. Efficient Feature Selection via Analysis of Relevance and Redundancy [J]. Journal of Machine Learning Research, 2004(5): 1205-1224