Modified Mutual Information-based Feature Selection for Intrusion Detection Systems in Decision Tree Learning

Song, Jingping; Zhu, Zhiliang; Scully, Peter Matthew David; Price, Christopher

Published in:
Journal of Computers

DOI:
10.4304/jcp.9.7.1542-1546

Publication date:
2014

Citation for published version (APA):
Song, J., Zhu, Z., Scully, P. M. D., & Price, C. (2014). Modified Mutual Information-based Feature Selection for Intrusion Detection Systems in Decision Tree Learning. *Journal of Computers*, *9*(7), 1542-1546. https://doi.org/10.4304/jcp.9.7.1542-1546

Modified Mutual Information-based Feature Selection for Intrusion Detection Systems in Decision Tree Learning

Jingping Song[1,2], Zhiliang Zhu[1], Peter Scully[2], Chris Price[2]
[1]Software College, Northeastern University, Shenyang, China, 110819
Email: {songjp, zhuzl}@swc.neu.edu.cn
[2]Department of Computer Science, Aberystwyth University, Aberystwyth, United Kingdom, SY23 3DB
Email: {jis17,pds7,cjp}@aber.ac.uk

***Abstract*—As network-based technologies become omnipresent, intrusion detection and prevention for these systems become increasingly important. This paper proposed a modified mutual information-based feature selection algorithm (MMIFS) for intrusion detection on the KDD Cup 99 dataset. The C4.5 classification method was used with this feature selection method. In comparison with dynamic mutual information feature selection algorithm (DMIFS), we can see that most performance aspects are improved. Furthermore, this paper shows the relationship between performance, efficiency and the number of features selected. .**[€]

***Index Terms*—Feature selection, classification, C4.5, intrusion detection, mutual information**

I. Introduction

As network-based technology and applications develop rapidly, the threat of attackers, computer viruses and criminal enterprises has grown accordingly. So defence for computer security is necessary which is composed by intrusion detection, anti-virus software, firewalls, data encryption authentication and so on. In complex classification domains, features may contain false correlations and it is very difficult for learning algorithms to classify them. Moreover, some features may be irrelevant and others may be redundant because the information they have is contained by the other features. These extra features can increase computation time. That is the reason why these classification domains are suitable for applying feature selection methods [1].

There are three main models dealing with feature selection: wrapper methods, filter methods and embedded methods. Wrapper methods optimize a classifier as part of the selection process and choose those features with high prediction performance induced by specified learning algorithms [2]. Filter methods are independent of learning algorithms and they mainly identify a feature subset from the original space on the basis of given evaluation criterions. In the embedded model, feature selection is integrated into the process of training for given methods. Mutual information-based feature selection method was first proposed by Battiti in 1994 [3]. It was modified by Huawen Liu in 2009 and by Fatemeh in 2011 [4,5]. This paper proposes a modified mutual information feature selection method based on Battiti's work and compares the resulting performance with Huawen's work. After we calculate the selected features, we use the C4.5 classification methods to evaluate the performance. C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan, and is an extension of Quinlan's earlier ID3 algorithm [6]. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier. C4.5 uses the concept of information gain to make a tree of classificatory decisions with respect to a previously chosen target classification.

The rest of the paper is organized as follows. Section 2 introduces the KDD Cup 99 dataset and reviews the mutual information and the necessary of feature selection. Section 3 gives an overview of the proposed algorithm. Section 4 shows the results of employing the algorithm, and compares the results with the methods proposed by Huawen Liu. Computation time comparisons are showed in this part as well. Section 5 describes the conclusion and discusses potential future work.

II. Background

In this section, some concepts about mutual information are given and some feature selection results are presented. The KDD99 dataset is also introduced.

A. KDD99 Dataset

The KDD Cup 99 dataset, originally developed by the DARPA 98 IDS evaluation program, has been the most widely used data set for the evaluation of anomaly detection methods. The whole DARPA dataset has almost 5 million input instances and each record represents a TCP/IP connection that is composed of 41 features. And the test dataset has about 2 million connection records. The dataset used in this work is a smaller subset, called 10 percent dataset, which contains 494021 instances and it was already used as the training dataset. For the test

[€] Manuscript received August 24, 2013; revised September 30, 2013; accepted October 25, 2013.

doi:10.4304/jcp.9.7.1542-1546

dataset, we used the original KDD Cup 99 dataset containing 311029 patterns [7].

A connection is a TCP data packet sequence from start to end in a certain time and data from source IP address to destination IP address in predefined protocol such as TCP or UDP. Each connection is labeled as either normal or attack. The attack type is divided into four categories of 39 types of attacks. The training and test dataset percentages for the four attack categories are shown in Table 1. Only 22 types of attacks are in the training dataset, and the other 17 unknown types only occur in the test dataset [8]. It is important to note that the test data is not from the same probability distribution as the training data, and it includes specific attack types not in the training data which make the task more realistic.

TABLE I.
PERCENTAGES OF NORMAL CONNECTIONS AND DIFFERENT KINDS OF ATTACKS IN KDD CUP 99

Categories	10% Training dataset (%)	Test dataset
Normal	19.69	19.48
Dos	79.24	73.90
Probe	0.83	1.34
R2L	0.23	5.21
U2R	0.01	0.07

The KDD Cup 99 data can be considered as a binary case or a multiple class case. The binary case regards all attack types as anomaly patterns and the other class is a normal pattern. A multiple class case deals with the classification based on different attacks. In ours work, we consider the KDD Cup 99 dataset as a binary case, we call the two patterns *normal* and *anomaly* data [9].

B. Mutual Information

Information theory was initially developed to measure the size of the amount of information in communicating data. And in this theory, entropy is an important measurement for information. It is capable of quantifying the uncertainty of random variables and scaling the amount of information shared by them effectively. In this paper, we only deal with finite random variables with discrete values [10].

Let X be a random variables with discrete values, its entropy is defined as

$$H(\mathrm{X}) = -\sum_{x \in X} p(\mathrm{x}) \log p(\mathrm{x}) \tag{1}$$

where H(·) is entropy, and p(x)=Pr(X=x) is the probability density function of X. Note that entropy depends on the probability distribution of the random variable.

Conditional entropy refers to the uncertainty reduction of one variable when the other is known. Assume that variable Y is given, the conditional entropy H(X|Y) of X with respect to Y is

$$H(\mathrm{X} \mid \mathrm{Y}) = -\sum_{y \in Y}\sum_{x \in X} p(\mathrm{x},\mathrm{y}) \log p(\mathrm{x} \mid \mathrm{y}) \tag{2}$$

where p(x,y) is the joint probability density function and p(x|y) is the posterior probabilities of X given Y.

Similarly, the joint entropy H(X,Y) of X and Y is

$$\begin{aligned} H(\mathrm{X},\mathrm{Y}) &= H(\mathrm{X}) + H(\mathrm{Y} \mid X) \\ &= H(Y) + H(X \mid Y) \\ &= -\sum_{y \in Y}\sum_{x \in X} p(\mathrm{x},\mathrm{y}) \log p(\mathrm{x},\mathrm{y}) \end{aligned} \tag{3}$$

To quantify how much information is shared by two variables X and Y, a concept termed mutual information I(X;Y) is defined as

$$\begin{aligned} \mathrm{I}(\mathrm{X};\mathrm{Y}) &= H(\mathrm{X}) - H(X \mid Y) \\ &= H(Y) - H(Y \mid X) \\ &= -\sum_{y \in Y}\sum_{x \in X} p(\mathrm{x},\mathrm{y}) \log \frac{p(\mathrm{x},\mathrm{y})}{p(\mathrm{x})\,\mathrm{p}(\mathrm{y})} \end{aligned} \tag{4}$$

If X and Y are closely related with each other, I(X;Y) will be very high. Otherwise, I(X;Y)=0 denotes that these two variables are totally unrelated. The mutual information could be applied for evaluating any arbitrary dependency between random variables. In this paper, we calculate the mutual information between two variables and measure the mutual dependence between them [11].

C. Feature Selection

Feature selection is the process of selecting a subset of relevant features for use in model construction. The central assumption when using a feature selection technique is that the data contains many redundant or irrelevant features. Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no useful information in any context [12,13]. In KDD99 dataset, some features may be irrelevant and others may be redundant since the information they add is contained in other features. These extra features can increase computation time for creating classifications, and can have an impact on the accuracy of the classifier built. For this reason, these classification domains seem to be suitable for the application of feature selection methods [14,15]. These methods are centered in obtaining a subset of features that adequately describe the problem at hand without degrading performance.

To verify that there are irrelevant and redundant

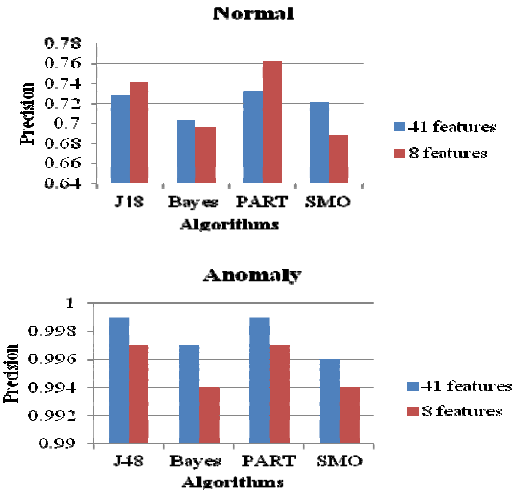

Figure 1. Precision comparison chart between all features and selected features

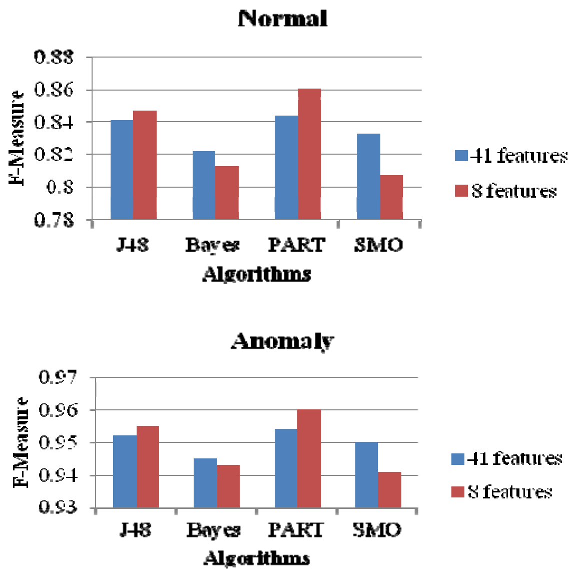

Figure 2. F-measure comparison chart between all features and selected features

features in KDD Cup 99 dataset, Correlation based Feature Selection (CFS) is used to select 8 features by Weka. Two performance measures (precision and F-measure) were calculated which will specifically be discussed in section 4 and we used four classification methods to calculate the two performances. Figure 1 shows the precision comparison between 41 features and 8 features by normal and anomaly types respectively. Similarity, figure 2 describes the other performance F-measure.

The two figures show for each classification method, that the two performances are quite close. For J48 and PART methods, the performances even get better. Another advantage of selecting features is the running time is shorter than using all features. We will show the computation time comparison in section 4.

III. Proposed Algorithms

First of all, the mutual information between each feature and class label in the KDD99 dataset is calculated. The results are shown in figure 3. Figure 3 shows that feature 5 has the largest mutual information value. This means that feature 5 and the class label have the largest correlation.

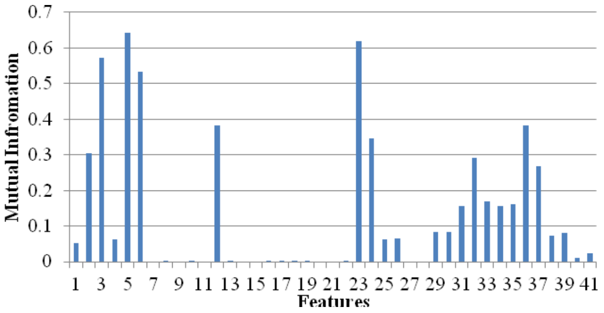

Figure 3. Mutual information of between each feature and class label in KDD99 dataset

We could rank the features by mutual information from Figure 3. But we could not select the features according to this way. Take the feature 5, 12 and 23 as an example, let C represent class label and mutual information between the three features and C are I(f5; C)= 0.6424, I(f12; C)= 0.3810, I(f23; C)= 0.6179. In descending order, the three is sorted as {f5, f23, f12}. But after f5 is selected, we should delete the correct instances induced by f5. Battiti proposed an evaluate function considering the mutual information between features, which is showed by formula (5). And the method called mutual information-based feature selection (MIFS). In this case, the mutual information between f5 and f23 is I(f5, f23)=1.4720 and the mutual information between f5 and f12 is I(f5, f12)=0.5436. According to Battiti's evaluate function, f12 will be selected, rather than f23. In 2009, Huawen Liu proposed a dynamic mutual information method called DMIFS. And DMIFS improved MIFS in respect to some performance.

$$I(f_i;C)-\beta\sum_{f_s\in S} I(f_i,f_s) \qquad (5)$$

In formula (5), f_i represents each feature in a set and f_s denotes a selected feature in a selected feature set S. There is a parameter β and Battiti suggested it should be between 0.5 and 1. But in our study, we think the parameter should be related to mutual information between each feature and class label, rather than a fixed value. So we put forward an improved algorithm named MMIFS as follows.

Input: A training dataset T=D(F,C).

Output: Selected features S.

(1) Initialize relative parameters: F←'initial set of all features', C ←'class labels', S=∅ .

(2) For each feature $f_i \in F$, compute the mutual information of the features with the class labels I(f_i; C).

(3) Selection of the first feature: find the f_i that maximizes the I(f_i,C), then S=S ∪ { f_i}, F =F\{f_i}.

(4) Repeat until the desired number of features is selected:

a. Computation of the mutual information between features: for all pair of features (f_i,f_s), where $f_i \in F$ and $f_s \in S$, compute I(f_i,f_s).

b. Selection of the next feature: choose the feature f_i as the one that maximizes $I(f_i;C)-\sum_{f_s\in S} I(f_i;C)*I(f_i,f_s)$

(5) Output the set containing the selected features: S.

The most important improvement of MMIFS is weighting all pairs of features (f_i,f_s) by the mutual information I(f_i; C). This study indicates that weighting the pair (f_i,f_s) by I(f_i; C) is better than a fixed value.

IV. Experiments and Results

A. Implemented System

C4.5 is used to classify the feature set that was selected by applying MMIFS. The classification is based on six measures: True Positive Rate (TPR), False Positive Rate (FPR), Precision, Total Accuracy, Recall, F-Measure. The six measures are calculated by True Positive (TP), False Positive (FP), True Negative (TN), False Negative (FN), as follows. True positive rate (TPR): TP/(TP+FN), also known as detection rate (DR) or sensitivity or recall. False positive rate (FPR): FP/(TN+FP) also known as the false alarm rate. Precision (P): TP/(TP+FP) is defined as the proportion of the true positives against all the positive results. Total Accuracy (TA): (TP+TN)/(TP+TN+FP+FN) is the proportion of true results (both true positives and true negatives) in the population. Recall (R): TP/(TP+FN) is defined as percentage of positive labeled instances that were predicted as positive. F-measure: 2PR/(P+R) is the harmonic mean of precision and recall.

In our experiments, we need to determine the desired feature numbers which we expect to select in KDD Cup 99 dataset. Thus, we calculated total accuracy of different feature numbers which are obtained by MMIFS. The results are shown in figure 4.

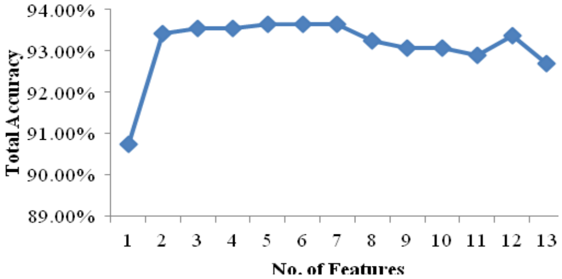

Figure 4. Total accuracy of different feature numbers

We can see from the figure that we tested 13 features obtained by MMIFS. The total accuracy does not increase as the numbers rise. The reason is because there are many noisy and redundant features in the dataset. Although the total accuracy levels are very close in figure 4, considering the number of instances in the KDD 99 dataset, slice improvement will result in large instances are correctly classified. The numbers between 2 and 13 could be used for comparison. But when we used DMIFS to get the features, we realised if the desired numbers are small, most of the features are the same as we got by MMIFS. Thus, we choose 10 features to compare the algorithms.

B. Results

In the following subsection, C4.5 is used to classify the dataset and compare the performance between DMIFS and MMIFS. The experiments were conducted a testing by using KDD 99 dataset and performed on a Windows machine having configuration and Intel (R) Core (TM) i5-2400 CPU@ 3.10GHz, 3.10 GHz, 4GB of RAM, the operating system is Microsoft Windows 7 Professional. We have used an open source machine learning framework Weka 3.5.0. We have used this tool for performance comparison of our algorithm with other classification algorithms. Table 2 shows the specific comparison.

Table 2 shows that most of the performances are improved by MMIFS compared to DMIFS, such as precision and F-measure. The total accuracies for these three methods are 92.65%, 92.94% and 93.02 respectively.

TABLE II.
COMPARISON RESULTS BETWEEN DMIFS AND MMIFS ALGORITHMS

Algorithm	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
C4.5	0.994	0.09	0.728	0.994	0.841	Normal
	0.91	0.006	0.999	0.91	0.952	Anomaly
C4.5 with DMIFS	0.993	0.086	0.736	0.993	0.846	normal
	0.914	0.007	0.998	0.914	0.954	anomaly
C4.5 with MMIFS	0.99	0.084	0.741	0.99	0.848	normal
	0.916	0.01	0.997	0.916	0.955	anomaly

Another advantage for applying feature selection methods on KDD 99 dataset is the saving in computation time. In C4.5 algorithm, we need to build a model from the KDD 99 training dataset first and then evaluate the model on the test dataset. Fig. 5 describes the time taken to build model comparison by the different feature numbers. The test model we used by C4.5 algorithm is 10-fold cross-validation and some computation time is spend for it. Fig. 6 illustrates the total time comparison by different feature numbers.

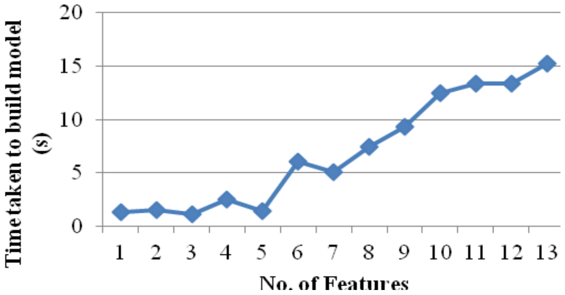

Figure 5. Time taken to build model comparison chart

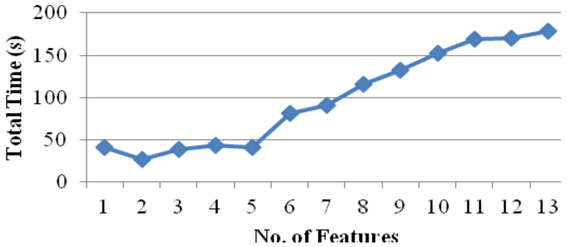

Figure 6. Total time comparison chart

We can see from fig.5 and fig.6 that as the number of features increases, the calculation time increases significantly. It indicates that the computation time is greatly affected by the numbers of features.

V. CONCLUSION AND FUTURE WORK

This paper proposed a new feature selection method and the main improvement of this work is that it modifies the mutual information feature selection algorithm by changing the weighting parameter. We tested this method on the KDD 99 dataset and compared the results with the DMIFS algorithm. The results show that most of the performance indicators are improved. Future work will evaluate the algorithm against other datasets which have less noise and less redundant features. The value of the weighting parameter may not be optimum, and so further study will attempt to find values of the parameter that produce the best results. Finally, we will try to compare the method based on correlation coefficient of features with the method based on mutual information.

REFERENCES

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

[2] Z. Zhao, H. Liu, Spectral feature selection for supervised and unsupervised learning, in: Proceedings of the 24th international Conference on Machine Learning, Corvalis, Oregon, 2007, pp. 1151–1157.

[3] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Transactions on Neural Networks*, 1994, pp.537-550.

[4] H. W. Liu, J. G. Suna, L. Liu, H. J. Zhang, "Feature selection with dynamic mutual information," Pattern Recognition, 42 (2009) 1330 – 1339.

[5] Fatemeh Amiri, MohammadMahdi Rezaei Yousefi, Caro Lucas, Azadeh Shakery, Nasser Yazdani, "Mutual information-based feature selection for intrusion detection systems," *Journal of Network and Computer Applications*, 34 (2011), pp.1184–1199.

[6] Amuthan Prabakar Muniyandia, R. Rajeswarib, R. Rajaramc, "Network Anomaly Detection by Cascading K-Means Clustering and C4.5 Decision Tree algorithm," International Conference on Communication Technology and System Design, 2011.

[7] Kayacik, H. G., Zincir-Heywood, A. N., & Heywood, M. I., "Selecting features for intrusion detection: A feature relevance analysis on KDD 99 intrusion detection datasets," In Proceedings of *the third annual conference on privacy, security and trust*, 2005.

[8] Mahbod Tavallaee, Ebrahim Bagheri, W. Lu, and Ali A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in proceedings of *the Second IEEE Symposium on Computational Intelligence for Security and Defence Applications 2009*.

[9] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, Feature selection and classifi-cation in multiple class datasets: An application to KDD Cup 99 dataset, Expert Systems with Applications, 38(2011), pp.5947–5957.

[10] Estevez, P.A., Tesmer, M., Perez, C.A., Zurada, J.M., Normalized Mutual Information Feature Selection, *IEEE Transactions on Neural Networks*, 2009, pp.189-201.

[11] Jose, Martinez Sotoca, Filiberto Pla, Supervised feature selection by clustering using conditional mutual information-based distances, Pattern Recognition, 3 (2010) 2068–2081.

[12] Chebrolu, S., Abraham, A., & Thomas, J. P.. Feature deduction and ensemble design of intrusion detection systems, Journal of Computers & Security, 24(4), 2005, pp.295–307.

[13] Mukkamala, S., & Sung, A. H.. Feature ranking and selection for intrusion detection systems using support vector machines, In International conference on information and knowledge engineering (ICIKE) , 2002, pp.503–509.

[14] S. W. Lin, K. C. Ying, C. Y. Lee, Z. J. Lee, "An intelligent algorithm with feature selection and decision rules applied to anomaly intrusion detection," *Applied Soft Computing*, 12 (2012), pp. 3285–3290.

[15] Pedro Casas, Johan Mazel, Philippe Owezarski, Unsupervised Network Intrusion Detection Systems: Detecting the Unknown without Knowledge, Computer Communications, 37 (7), 2012, pp. 772–783.