CORRELATION-BASED FEATURE SELECTION FOR INTRUSION DETECTION DESIGN

Te-Shun Chou, Kang K. Yen, and Jun Luo
Department of Electrical and Computer Engineering
Florida International University
Miami, FL

and

Niki Pissinou and Kia Makki
Telecommunications and Information Technology Institute
Florida International University
Miami, FL

ABSTRACT

In a large amount of monitoring network traffic data, not every feature of the data is relevant to the intrusion detection task. In this paper, we aim to reduce the dimensionality of the original feature space by removing irrelevant and redundant features. A correlation-based feature selection algorithm is proposed for selecting a subset of most informative features. Six data sets retrieved from UCI databases and an intrusion detection benchmark data set, DARPA KDD99, are used to train and to test C4.5 and naive bayes machine learning algorithms. We compare our proposed approach with two correlation-based feature selection algorithms, CFS and FCBF and the results indicate that our approach achieves the highest averaged accuracies. Our feature selection algorithm could effectively reduce the size of data set.

I. INTRODUCTION

An intrusion detection system is a security management system for computers and networks. It examines activities from computer users and then identifies inappropriate, incorrect, or anomalous activities within computers or networks. During the past years, a variety of techniques have been proposed for intrusion detection systems. These techniques are mainly categorized into two groups: anomaly detection techniques and misuse detection techniques. Anomaly detection techniques use machine learning algorithm to search for intrusive activities by comparing network traffic to those acceptable normal patterns learned from training data [1], [2]. If the pattern of observed data is different from those learned normal ones, the data is classified as an attack. Misuse detection techniques build a set of predefined if-then-else rules to describe intrusive patterns [3], [4]. If the signature of observed network traffic is not matched with any of predefined rules, it is declared as an attack.

Generally, a data set that includes a large amount of network traffic is necessary to be collected in advance for designing an intrusion detection system. The size of data collected from the network is always large. It includes a great amount of traffic records with a number of various features such as the length of connection, the type of protocol, the network service and other information. Based on this set of data, misuse detection techniques specify well defined attack signatures and anomaly detection techniques construct acceptable user behaviors. Theoretically and ideally, the ability of discriminating attacks from normal behavior should be performed better if more features are added during the detection process. However, the answer is sometimes negative because some included features may be irrelevant with poor prediction ability to the class, and some features may be redundant since they are highly inter-correlated with one or more of the other features [5]. Therefore, analysis of these features is a very crisp step in the development of intrusion detection system. By keeping the most relevant features to the given classification task, the computation time can be reduced and the accuracy of detection may be enhanced. Finally the intrusion detection system can achieve the maximum overall performance.

1-4244-1513-06/07/$25.00 ©2007 IEEE

The algorithms of feature selection are mainly divided into two categories, filter and wrapper [6]. Filter method operates without engaging any information of induction algorithm. By using some prior knowledge such as feature should have strong correlation with the target class, or feature should be uncorrelated to each other, filter method selects the best subset of features [7], [8]. On the other hand, wrapper method employs a predetermined induction algorithm to find a subset of features with the highest evaluation by searching through the space of feature subsets and evaluating quality of selected features. The process of feature selection acts like "wrapped around" an induction algorithm. By including a specific induction algorithm such as ID3 [9] or C4.5 [10] to optimize feature selection, it often provides a better classification accuracy result than that of filter approach. However, wrapper method is more time consuming than filter method due to it is strongly coupled with an induction algorithm with repeatedly calling the algorithm to evaluate the performance of each subset of features. It thus becomes unpractical to apply a wrapper method to select features from a large data set that contains numerous features and instances [11]. Furthermore, wrapper method requires re-execute its induction algorithm to select features from data set while the algorithm is replaced with a dissimilar one. It is less independent of any induction algorithms than filter is.

Consequently, we addressed aspects of feature selection based on filter method. Our approach uses the concept of information theory to evaluate the worth of features and then eliminate both irrelevant and redundant features. The approach is closed to Fast Correlation-Based Filter (FCBF) [12], however we treat the correlation between features in a global perspective. We measure the total amount of information enclosing in a feature is the summation of inter-correlations to all of the rest of the features, but FCBF only considers on a feature of rest ones at a time. Therefore, FCBF may be tricked in situation where the dependence between pair of features is weak but the total inter-correlated strength of one feature to the others is strong. The result is that FCBF possibly keeps a feature that its information can be found in the remaining selected subset of features. In addition, FCBF requires adjusting a threshold for its feature selection procedure, while our algorithm does not.

For evaluating the performance of our proposed approach, two feature selection methods based on information entropy method are implemented, One is FCBF and the other is Correlation Based Feature Selection (CFS) [13]. Six small data sets retrieved from UCI databases [14] and a high-dimensional intrusion detection benchmark data set DARPA KDD99 [15] are used to train and to test C4.5 and naive bayes [16] machine learning algorithms. Experiments demonstrate that our approach outperforms CFS and FCBF feature selection methods. Results also show that the accuracies of our approach sometimes are even better than those of using full feature set.

This paper is organized as follows. Section 2 introduces the theoretical framework which forms the base of our proposed approach in measuring the goodness between pairs of features and between features and classes. Section 3 describes our proposed feature selection algorithm. We then demonstrate the experimental methodology, followed by a discussion of the experiment results. Finally, we present the conclusions and future work in the last section.

II. THEORETICAL FRAMEWORK

In information theory, *entropy* is a measure of the amount of uncertainty about a source of messages. Equations 1 and 2 define the entropy of Y before and after observing values of another variable X, respectively.

$$H(Y) = -\sum_i p(y_i)\log_2 p(y_i) \tag{1}$$

$$H(Y|X) = -\sum_j p(x_j)\sum_i p(y_i|x_j)\log_2 p(y_i|x_j) \tag{2}$$

where $p(y_i)$ is the prior probabilities for all values of random variable Y and $p(y_i|x_j)$ is the conditional probability of y_i given x_j. By treating Y as classes and X as features in a data set, the entropy is 0 without any uncertainty at all if all members of a feature belong to the same class. On the other hand, members in a feature are totally random to a class if the value of entropy is 1. The range of entropy is 0 to 1.

The amount by which the entropy of Y decreases reflects additional information about Y provided by X and is called *information gain* (or *mutual information*) as shown in Equation 3. It measures how well a given variable separates instances into another variable. It is symmetrical that the amount of information gained about Y after observing X is equal to the information gained about X after observing Y. It is defined as:

$$I(Y;X) = H(Y) - H(Y|X) \tag{3}$$

However, the range of information gain is not from 0 to 1 and it is biased if feature with more values. Therefore, we choose *symmetric uncertainty* [17] as our tool to find the strength of predictive from features to target classes and the strength of correlation between features themselves. It is a symmetric measure and normalizes its values from 0 to 1. Also, it averages the values of two uncertainty variables and hence it has no bias problem of pairs of features.

$$SU(Y;X) = 2\left[\frac{I(Y;X)}{H(X)+H(Y)}\right] \tag{4}$$

In our work, we apply symmetric uncertainty measure to find the correlation between features and target class. If a feature has a low symmetric uncertainty to the target class, it implies that feature has poor prediction ability to the class. On the other hand, the feature has strong prediction ability to the class if the symmetric uncertainty is high. Having all the symmetric uncertainties between features and the target class, features can be ranked in descending order according to their degrees of association to the target class Y such that $SU(Y; X_i) \geq SU(Y; X_j)$ where X_i and X_j are two features. Those features which have the lowest ranks are considered as irrelevant features and are filtered out.

Similarly, we apply symmetric uncertainty measure to pairs of features. If the measure of mutual information between a pair of features is low, it represents these two features are independent to each other. For each pair of features, one feature only contains a little information about the other, i.e., knowing one feature cannot give any information about the other. On the contrary, the two

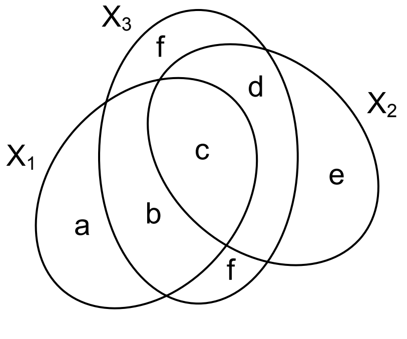

Figure 1. Illustration of Correlations of Three Features in Venn Diagram

features are highly inter-correlated with each other if they have a high mutual information measure. It means that one feature contains lots of information about the other and implies that knowing one feature can provide necessary information about the other. Under this circumstance, one of them can be considered as a redundant feature and can be discarded. For a better understanding of the idea of redundant feature, we use Venn diagram to illustrate correlations of multiple features X_1, X_2 and X_3. As shown in the figure, $SU(X_1; X_3) = b + c$, $SU(X_1; X_2) = c$ and $SU(X_2; X_3) = c + d$. Obviously, some redundant information will be included if we choose all three features since $SU(X_1; X_3) + SU(X_1; X_2) + SU(X_2; X_3) = b + 3c + d$ which is greater than $SU(X_1, X_2; X_3) = b + c + d$. Therefore, removing redundant features from the original feature space is necessary in order to discard needless information. The intrusion detection processing time can therefore be reduced by the use of a subset of the original feature space.

As shown in Figure 1, feature X_3 is highly inter-correlated with both features X_1 and X_2. By using Shannon's information-theoretic measure, we get the joint entropy:

$$\begin{aligned} H(X_1,X_2,X_3) &= (a+b+c+d+e)+f \\ &= H(X_1,X_2)+[H(X_3)+SU(X_1,X_2;X_3) \\ &\quad -SU(X_1;X_3)-SU(X_2;X_3)] \end{aligned} \quad (5)$$

The larger mutual information $SU(X_1; X_3)$ and $SU(X_2; X_3)$ are, the smaller area f will be. When f is very small, it represents that X_3 heavily depends on both X_1 and X_2. The measure of joint entropy $H(X_1, X_2, X_3)$ is approximately equal to $H(X_1, X_2)$. It implies that the total amount of information of X_1, X_2, and X_3 can be represented by the amount information of X_1 and X_2. Feature X_3 is considered as a redundant feature and can be removed with only losing a few information of the original feature space. Finally, we select the feature that is neither an irrelevant feature nor a redundant feature and call this type of feature as "significant feature".

III. FEATURE SELECTION ALGORITHM

The objective of feature selection is to select a subset from the original feature space which is more informative to target classes in executing machine learning tasks but to ignore the irrelevant and redundant features. In this paper, we develop a feature selection algorithm based on information-theoretical measures as described in Figure 2. Based on the entropy of a designated feature, symmetric uncertainty is obtained as a measurement of relevance on the given feature.

The algorithm mainly consists of two parts for achieving the goal of reducing dimensionality of the original feature space. In the first part (line 1-5), the algorithm removes irrelevant features with poor prediction ability to target class. In the second part (line 6-12), the algorithm eliminates redundant features that are inter-correlated with one of more other features. Finally, the remaining selected features are significant features that contain indispensable information about the original feature set.

Given a data set with a number of input features and a target class, the algorithm firstly calculates the mutual information between features and class. The algorithm then ranks the features in descending order according to their degrees of association to the target class. Once the importance of the input features are ranked, these terms whose information measure are greater than zero are kept; which means those removed features are totally irrelevant to target class and the remaining ones are predictive.

In the second part, it starts with calculating the inter-correlated strengths of each pair of features. The total amount of mutual information for each feature is acquired by adding all mutual information measures together that relate to that feature. For adjusting the discriminative power of mutual information performed on feature-to-feature and feature-to-class to the same level, we introduce factor w and its value is equal to the mean of summation of feature-to-class information divided by the mean of summation of feature-to-feature information. By multiplying w to each feature-to-class measure, both feature-to-class and feature-to-feature reach to the same important rank. Finally, the differences of them are computed and we only keep those features whose values are greater than zero, which means the selected features are the most "significant features" that restrain indispensable information of the original feature space.

IV. EXPERIMENTAL METHODOLOGY

In this section, we firstly describe the experimental data sets. We then introduce the discretization technique that is

```
// Remove irrelevant features
Input original data set $D$ that includes features $X$ and target class $Y$
For each feature $X_i$
    Calculate mutual information $SU(Y; X_i)$
Sort $SU(Y; X_i)$ in descending order
Put $X_j$ whose $SU(Y; X_i) > 0$ into relevant feature set $R_{XY}$
// Remove redundant features
Input relevant feature set $R_{XY}$
For each feature $X_j$
    Calculate pairwise mutual information $SU(X_j; X_k)\ \forall j \neq k$
$S_{XX} = \Sigma(SU(X_j; X_k))$
Calculate means $\mu_R$ and $\mu_S$ of $R_{XY}$ and $S_{XX}$, respectively. $w = \mu_S / \mu_R$
$R = w \cdot R_{XY} - S_{XX}$
Select $X_j$ whose $R > 0$ into final set $F$
```

Figure 2. Feature Selection Algorithm

used to transform continuous features to discrete ones. At last, we explain the experimental setup.

A. The Data Set

The DARPA KDD99 benchmark data set, also known as "DARPA Intrusion Detection Evaluation dataset", is chosen for analyzing the performance of our feature selection approach. During the entire course of work, a subset of 494,020 records is used. The data set describes each record in terms of 41 features plus a label of either normal or a type of attack. The content of these features is continuous and discrete with vary scales and ranges. Table 1 summarizes the distributions of the normal records and four main attack classes: *Denial of service* (*DoS*) *attacks*, *Remote to Local* (*R2L*) *attacks*, *User to Root* (*U2R*) *attacks*, and *Probe attacks*. In addition, all together six data sets are selected from UCI Repository of Machine Learning Databases. They have different numbers of records and features. Their characteristics are listed in Table 2.

B. Discretization of Features

In the six UCI data sets and KDD99 data set, each record is composed by a set of meaningful features. The type of features is either discrete or continuous, i.e., the former is a qualitative scale and the latter is quantitative. For qualitative scales, the values are simply labels without any order involved. They could be symbolic or numeric values which are distinct and separated. Also, it is a form of categorical data that has no "numeric" meaning. By using the features of KDD99 data set as an example, the value of feature *protocol_type* is one of the symbolic set {icmp, tcp, udp}. The numeric value of feature *logged_in* is 1 or 0 to represent the user successfully logged in the system or not. For quantitative scales, the data are characterized by numeric values within a finite interval. The distance between any two adjacent values is not necessary the same. Examples can be found in feature *duration* where it is given by numeric values to represent the lengths of record, and the values are within an interval [0, 58329].

Since symmetric uncertainty is calculated for discrete features only, all the continuous features in a given data set are required to be discretized prior to the feature selection analysis. Thus, we apply discretization method to transform continuous features to discrete ones prior to the analysis. For a numeric feature, cut points effectively decompose the range of continuous values into a number of intervals. These intervals can then be treated as categorical values of a discrete feature. In our work, *equal frequency binning* technique [13] is applied to each continuous feature individually. It is an unsupervised discretization method with no class information involved. It sorts the observed values of a continuous feature and then divides these values into a specified number of intervals. Each of the intervals has an approximate equal number of values. With the use of discretization of features, the complexity of every continuous feature is reduced as well.

C. Empirical Setting

In order to evaluate the performance of our proposed feature selection algorithm on data sets, two representative feature selection algorithms, CFS and FCBF, built on the top of symmetric uncertainty are chosen. CFS method uses a correlation-based heuristic search algorithm to evaluate the worth of subsets of features. It considers good feature subsets contain features that are highly correlated with the class, yet uncorrelated with one another. The heuristic algorithm measures the merit of feature subsets from pairwise feature correlations and then the subset with the highest merit found during the search is reported. Rather than scoring the worth of subsets of features of CFS approach, FCBF method measures correlations between features and classes and correlations between pairs of features as well. It then selects features which are highly correlated with the class to predict but are less correlated to any feature already selected. In addition, we apply two machine learning algorithms, naive bayes and C4.5 algorithm, to evaluate the detection accuracy on selected features for each feature selection algorithm.

Table 1. KDD99 Data Set

Class	Number	Percentage
Normal	97277	19.69%
DoS	391458	79.24%
R2L	1126	0.23%
U2R	52	0.01%
Probe	4107	0.83%
Total	494020	100%

Table 2. UCI Data Sets

Name	Feature	Record	Class
Abalone	8	4177	3
Cmc	9	1473	3
Ionosphere	34	351	2
Pima	8	768	2
Wdbc	30	569	2
Wine	13	178	3

The experiments are performed on six UCI data sets and binary classification (normal/attack) of KDD99 data set. Four new sets of data are generated according to the normal class and four categories of attack (*DoS*, *Probe*, *U2R* and *R2L*). In each data set, records with the same attack category and all the normal records are included. For each data set, we run our proposed approach and the other two feature selection algorithms CFS and FCBF, and recorded these features selected by each algorithm. Throughout the entire experiments, the threshold of FCBF is set to 0. We then apply C4.5 and naive bayes machine learning algorithms on each original full data set as well as each newly obtained data set that includes only those selected features from feature selection algorithms. By applying 10-fold cross-validation evaluation on each data set, classification accuracies of six UCI data sets and standard measurements, such as the *detection rate* (*DR*) and *false positive rate (FPR*), for evaluating the performance of intrusion detection tasks are reported. The denotations of *True Positive* (*TP*), *True Negatives* (*TN*), *False Positive* (*FP*), and *False Negative* (*FN*) are defined as follows. Equations 6 and 7 describe detection rate and false positive rate, respectively. All the experiments are implemented by C++ and a data mining software TANAGRA [18].

- *True Positive* (*TP*): The number of malicious records that are correctly identified.
- *True Negatives* (*TN*): The number of legitimate records that are correctly classified.
- *False Positive* (*FP*): The number of records that were incorrectly identified as attacks however in fact they are legitimate activities.
- *False Negative* (*FN*): The number of records that were incorrectly classified as legitimate activities however in fact they are malicious.

$$DR = \frac{TP}{TP + FN} \tag{6}$$

$$FPR = \frac{FP}{TN + FP} \tag{7}$$

VI. RESULTS

Table 3 shows the number of features selected from our approach and those selected by CFS and FCBF algorithms. Table 4 summarizes the classification accuracies of six UCI data sets. Tables 5 and 6 summarize the percentages of detection rates and false positive rates performed on KDD99 data set with C4.5 and naive bayes learning algorithms, respectively. For an intrusion detection task, abnormal activities are expected to be correctly identified and normal activities are anticipated not to be misclassified. Therefore, a higher detection rate and a lower false positive rate are desired.

From Table 4, our approach shows higher averaged accuracies in comparison with the outcomes of CFS and FCBF feature selection algorithms. Especially in the abalone data set, we get the highest classification accuracy by using 2 out of 8 features performed on C4.5 learning algorithm, which is better than that of using full feature set. The averaged accuracies of Tables 5 and 6 also show that our approach outperform over CFS and FCBF feature selection algorithms. The averaged detection rates and the averaged false positive rates of our experimental results are better than those of using full feature set performed on C4.5 and naive bayes, respectively.

In the *Normal-DoS* data set, the difference in detection rates is very slight for all of the feature selection algorithms. With our approach, the detection rate is the same as that of using full feature set in C4.5 learning algorithm. In the *Normal-U2R* and *Normal-R2L* data sets, we have satisfactory detection rates and false positive rates. Though CFS and FCBF approaches achieve low false positive rates, they have very poor detection rates. In the *Normal-Probe* data set, both CFS and FCBF approaches fail to achieve an acceptable presentation on detection rates while using naive bayes leaning algorithm, whereas our approach gains very high detection rates performed on both leaning algorithms.

For any feature selection algorithm, false positive rates are mostly low because sufficient normal records present in

Table 3. Number of Selected Features

	Data Set	Full Set	Ours	CFS	FCBF
UCI	Abalone	8	2	4	1
	Cmc	9	2	2	2
	Ionosphere	34	8	2	2
	Pima	8	4	3	1
	Wdbc	30	15	5	1
	Wine	13	6	6	9
KDD99	Normal-DoS	41	12	4	4
	Normal-Probe	41	12	4	4
	Normal-U2R	41	5	1	2
	Normal-R2L	41	7	1	3

Table 4. Classification Accuracies of C4.5 and Naive Bayes on Full and Selected Feature Sets of UCI Data Sets

Data Set	C4.5				Naive Bayes			
	Full Set	Ours	CFS	FCBF	Full Set	Ours	CFS	FCBF
Abalone	51.90	56.00	51.90	51.90	63.23	53.60	51.90	51.90
Cmc	63.68	54.65	52.89	52.89	53.36	52.61	52.27	52.27
Ionosphere	74.93	74.93	74.93	74.93	99.15	97.72	94.02	94.02
Pima	65.10	65.10	65.10	65.10	89.97	87.50	85.03	77.34
Wdbc	62.74	62.74	62.74	62.74	99.30	99.30	99.65	94.02
Wine	94.94	94.94	94.94	94.94	98.88	97.75	97.75	98.88
Average	68.88	68.06	67.08	67.08	83.98	81.41	80.10	78.07

Table 5. Detection Rates of C4.5 and Naive Bayes on Full and Selected Feature Sets of KDD99 Data Sets

Data Set	C4.5				Naive Bayes			
	Full Set	Ours	CFS	FCBF	Full Set	Ours	CFS	FCBF
Normal-DoS	99.97	99.97	99.86	99.31	99.12	99.16	99.37	99.19
Normal-Probe	98.51	97.78	95.52	94.91	98.27	96.54	62.53	45.31
Normal-U2R	48.08	48.08	0	7.69	82.69	69.23	0	7.69
Normal-R2L	93.52	97.69	0	27.44	99.11	93.25	0	33.84
Average	85.02	85.88	48.85	57.34	94.80	89.55	40.48	46.51

Table 6. False Positive Rates of C4.5 and Naive Bayes on Full and Selected Feature Sets of KDD99 Data Sets

Data Set	C4.5				Naive Bayes			
	Full Set	Ours	CFS	FCBF	Full Set	Ours	CFS	FCBF
Normal-DoS	0.04	0.03	2.19	7.58	0.01	0.01	2.76	7.77
Normal-Probe	0.02	0.38	0.36	0.36	1.29	0.87	0.15	0.10
Normal-U2R	0	0	0	0	0.63	0.50	0	0
Normal-R2L	0.01	0.01	0	0.02	1.31	0.49	0	0.08
Average	0.02	0.11	0.64	1.99	0.81	0.47	0.73	1.99

any of four data sets. As for the number of misclassification attack records, our approach provides acceptable detection rates on *Normal-DoS*, *Normal-Probe* and *Normal-R2L* data sets using both C4.5 and naive bayes learning algorithms. It is not only because each of the above data set supplies sufficient attack records but also the attacks mostly have a same attack signature. For example, the *DoS* attack type includes near 400,000 data records distributed in 10 different attacks. Mostly of them are *netpune* and *smurf* attacks that account for around 99%. In the *Probe* attack category, 95% of attacks are *ipsweep*, *portsweep* and *satan* that are distributed in 4107 attacks. As for *R2L* attack class, more than 90% of attacks are *warezclient* attack while 8 different kinds of attacks present. In contrast, the classification presented on *Normal-U2R* data set is satisfactory neither on full feature set approach nor on one of three feature selection algorithms. The *Normal-U2R* data set includes 52 attack records which are insufficient for learning on a classification algorithm.

V. CONCLUSIONS AND FUTURE WORK

Based on information-theoretical measure, a feature selection algorithm is introduced in this paper. Correlation analysis is employed between feature and feature and between feature and class for removing irrelevant and redundant features from both low and high dimensional feature spaces. Having gone through comprehensive analyses in C4.5 and naive bayes learning algorithms, our approach has quite different outcome from CFS and FCBF feature selection algorithms. The experimental results demonstrate that our algorithm has a superior performance in six UCI databases and KDD99 data set. The results show that our approach can be a practical feature selector to select informative features from data sets for classification. We will study more on correlations of feature-to-feature and feature-to-class in order to keep the most significant features for increasing accuracies of classification tasks. We will also perform more experiments on different feature selection techniques and machine learning algorithms in order to verify the practical feasibility in the real world classification problems.

REFERENCES

[1] K. M. Faraoun and A. Boukelif, "Neural Networks Learning Improvement using the K-Means Clustering Algorithm to Detect Network Intrusions," International Journal of Computational Intelligence, vol. 3 no. 2, pp. 161-168, 2006.

[2] J. T. Yao, S. L. Zhao, and L. V. Saxton, "A Study on Fuzzy Intrusion Detection," Proceedings of SPIE, Data Mining, Intrusion Detection, Information Assurance, And Data Networks Security, Orlando, Florida, USA, pp. 23-30, 2005.

[3] N. Bashah, I. B. Shanmugam, and A. M. Ahmed, "Hybrid Intelligent Intrusion Detection System," Transactions on Engineering, Computing and Technology, vol. 6, pp. 291-294, June 2005.

[4] M. Sabhnani and G. Serpen, "KDD Feature Set Compliant Heuristics Rules for R2L Attack Detection," International Conference in Computer Security and Management, Las Vegas, Nevada, pp. 310-316, June 2003.

[5] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," Journal of Machine Learning Research, vol. 3, pp. 1157-1182, 2003.

[6] G. John, R. Kohavi, and K. Pfleger, "Irrelevant Features and the Subset Selection Problem," Proceedings ML-94, pp. 121-129, Morgan Kaufmann, 1994.

[7] H. Almuallim and T. G. Dietterich, "Learning with Many Irrelevant Features," Proceedings AAAI-91, pp. 547-551, MIT Press, 1991.

[8] K. Kira and L. A. Rendell, "The Feature Selection Problem: Traditional Methods and a New Algorithm," Proceedings AAAI-92, pp. 129-134, MIT Press, 1992.

[9] J. R Quinlan, "Induction of decision trees," Machine Learning, vol. 1, pp. 81-106, 1986.

[10] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, 1993.

[11] J. Biesiada and W. Duch, "Feature Selection for High-Dimensional Data: A Kolmogorov-Smirnov Correlation-Based Filter Solution," Proceedings of the 4th International Conference on Computer Recognition Systems, 2005.

[12] L. Yu and H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution," Proceedings of The Twentieth International Conference on Machine Leaning, pp. 856-863, Washington, D.C., August, 2003.

[13] M. Hall, Correlation Based Feature Selection for Machine Learning, Doctoral Dissertation, The University of Waikato, Department of Computer Science, 1999.

[14] C. L. Blake and C. J. Merz, UCI Repository of Machine Learning Databases, 1998.

[15] KDD'99 archive: The Fifth International Conference on Knowledge Discovery and Data Mining. URL: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[16] T. M. Mitchell, *Machine Learning*, McGraw-Hill, New York, 1997.

[17] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge University Press, 1988.

[18] TANAGRA: http://eric.univ-lyon2.fr/~ricco/tanagra/