2009 International Conference on Machine Learning and Computing
IPCSIT vol.3 (2011) © (2011) IACSIT Press, Singapore

Euclidean-based Feature Selection for Network Intrusion Detection

Anirut Suebsing, Nualsawat Hiransakolwong

Department of Mathematics and Computer Science
King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand

Abstract. Nowadays, data mining has been playing an important role in the various disciplines of sciences and technologies. For computer security, data mining are introduced for helping intrusion detection System (IDS) to detect intruders correctly. However, one of the essential procedures of data mining is feature selection, which is the technique (commonly used in machine learning) for selecting a subset of relevant features for building robust learning models, due to the fact that feature selection can help enhance the efficiency of prediction rate. In the previous researches on feature selection, the criteria and way about how to select the features in the raw data are mostly difficult to implement. Therefore, this paper presents the easy and novel method, for feature selection, which can be used to separate correctly between normal and attack patterns of computer network connections. The goal in this paper is to effectively apply Euclidean Distance for selecting a subset of robust features using smaller storage space and getting higher Intrusion detection performance. During the evaluation phase, three different test data sets are used to evaluate the performance of proposed approach with C5.0 classifier. Experimental results show that the proposed approach based on the Euclidean Distance can improve the performance of a true positive intrusion detection rate especially for detecting known attack patterns.

Keywords: Intrusion Detection System (IDS), Feature Selection, Data Mining, Euclidean Distance, C5.0

1. Introduction

The internet and local area networks are growing larger in recent years. As a great variety of people all over the world are connecting to the Internet, they are unconsciously encountering the number of security threats such as viruses, worms and attacks from hackers [1, 2]. Now firewalls, anti-virus software, message encryption, secured network protocols, password protection and so on are not sufficient to assure the security in computer networks, which some intrusions take advantages of weaknesses in computer systems to threaten. Therefore, intrusion detection is becoming a more and more important technology which follows up network traffic and identifies network intrusion such as anomalous network behaviors, unauthorized network access, and malicious attacks to computer systems [3].

The techniques of intrusion detection can be categorized into two categories [4]: anomaly detection and misuse detection. Anomaly detection identifies deviations from normal network behaviors and alert for potential unknown attacks, and misuse detection (signature-based detection) detects intruders with known patterns.

In the last, data mining are introduced for helping IDS to detect intruders correctly [5, 2], and accordingly IDSs have shown to be successful in detecting known attacks. On the contrary, many unknown attacks IDSs still undergo from false positive (FP: detect a normal as an attack connection), also known as a false detection or false alarm. Though some intrusion experts believe that most novel attacks can be adequate to catch by using a signature of known attacks [6]. Although data mining can help IDS to detect correctly intruders, data mining relies on feature selection which is one of the important procedures of data mining.

[+] Corresponding author. Tel.: +6623267439.
E-mail address: s9062952@kmitl.ac.th

Feature selection is intended to suggest which features are more important for the prediction, to find out and get rid of irrelevant features that reduce classification accuracy, discover relations between features and throw out highly correlated features which are redundant for prediction.

The goal in this paper is to effectively apply Euclidean Distance to select better feature subsets with using smaller storage space and getting higher Intrusion detection performance.

The paper is organized as follows: In Section 2, a background of feature selection is addressed, following with the fields of intrusion detection, Euclidean Distance and C5.0 algorithm. In Section 3, the data set used in this paper is addressed. The proposed method is described in Section 4. In Section 5, the experimental results are reported, and the remarkable conclusions are addressed in the final Section.

2. Background

The rapid developments in computer science and engineering have led to expediency and efficiency in capturing huge accumulations of data. The new challenge is to transform the enormous of data into useful knowledge for practical applications.

An earlier general task in data mining is to extract outstanding features for the prediction. This function can be broken into two groups—feature extraction or feature transformation, and feature selection [7]. Feature extraction (for example, principal component analysis, singular-value decomposition, manifold learning, and factor analysis) refers to the process of creating a new set of combined features (which are combinations of the original features).

On the other hand, feature selection is different from feature extraction because it does not produce new variables. Feature selection also known as variable selection, feature reduction, attribute selection or variable subset selection, is a widely used dimensionality reduction technique, which has been the focus of much research in machine learning and data mining and found applications in text classification, web mining, and so on. It allows for faster model building by reducing the number of features, and also helps remove irrelevant, redundant and noisy features. This allows for building simpler and more comprehensible classification models with classification performance. Hence, selecting relevant attributes are a critical issue for competitive classifiers and for data reduction. In the meantime, feature weighting is a variant of feature selection. It involves assigning a real-valued weight to catch feature. The weight associated with a feature measures its relevance or significance in the classification task [8]. Feature selection algorithms typically fall into two categories; Feature Ranking and Subset Selection. Feature Ranking ranks the features by a metric and eliminates all features that do not achieve an adequate score (selecting only important features). Subset selection searches the set of possible features for the optimal subset. Feature Ranking methods are based on statistics, information theory, or on some function of classifier's outputs [9]. In statistics, the most popular form of feature selection is stepwise regression. It is a greedy algorithm that adds the best feature (or deletes the worst feature) at each round. The main control issue is deciding when to stop the algorithm. In machine learning, this is typically done by cross validation [10].

In this paper, we adapt Euclidean Distance to select robust features which can bring to a successful conclusion of intrusion detection. Euclidean Distance is used to select features to build model for the detection of known and unknown attacks. And also, method of C5.0 is used to evaluation in this paper. Note that our proposed approach is categorized as the feature ranking selection. The following section is the introduction to intrusion detection.

2.1. Intrusion Detection

Network based and Host based IDSs are mainly two main types of IDS being used now. Individual packets going through networks are analyzed in a network-based system in NIDS. The malevolent packets which might be passed by a firewall filtering rules can be detected by the NIDS. In a Host based system, the IDS examines the activity on each individual computer or host [11]. The techniques of intrusion detection can be grouped into two groups [4]: anomaly detection and misuse detection.

Anomaly detection [4] tries to determine whether deviation from established normal usage patterns can be flagged as intrusions. Anomaly detection techniques are based on the assumption that misuse or intrusive behavior deviates from normal system procedure [12]. The advantage of anomaly detection is that it can detect attacks notwithstanding whether or not the attacks have been seen before. But the disadvantage of anomaly detection is ineffective in detecting insiders' attacks.

Misuse Detection or Signature-Based Intrusion Detection, traditional technique, [4] employs patterns of known attacks or weak spots of the system to match and identify attacks. This means that there are some ways to represent attacks in the form of a pattern or an attack signature so that even variations of the same attacks can be detected. The major drawback of misuse detection is that it cannot predict new and unknown attacks and has high false alarm rate.

In the view of the fact that Intrusion Detection System has some faults, especially misuse detection that cannot detect unknown attacks, intelligent computing techniques, such as statistical approaches, expert system, pattern matching, Artificial Neural Network, Support Vector Machines, Neuro-Fuzzy, Genetic Algorithm with above techniques and data mining, are being used to avoid above shortcomings of intrusion detection.

2.2. Euclidean Distance

Euclidean Distance is the most common use of distance [13, 14, 15]. In most cases when people said about distance, they will refer to Euclidean distance. Euclidean distance or simply 'distance' examines the root of square differences between coordinates of a pair of objects. In mathematics, the Euclidean distance or Euclidean metric is the "ordinary" distance between two points. The Euclidean distance between two points A = (x1, x2, x3, …, x_n) and B = (y1, y2, y3, …, y_n) is defined as:

$$d(A,B)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+(x_3-y_3)^2+\ldots+(x_n-y_n)^2}=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \tag{1}$$

2.3. C 5.0 Algorithm

Classification is an important technique in data mining, and the decision tree is the most efficient approach to classification problems—Friedman 1997 [16]. The input to a classifier is a training set of records, each of which is a tuple of attribute values tagged with a class label. A set of attribute values defines each record. A decision tree has the root and each internal node labeled with a question. The arcs emanating from each node represent each possible answer to the associated question. Each leaf node represents a predication of solution to the problem under consideration. C5.0, one of methods that be used to build a decision tree, is a commercial version of C4.5.

A C5.0 model is based on the information theory [17, 18]. Decision trees are built by calculating the information gain ratio. The algorithm C5.0 works by separating the sample into subsamples based on the result of a test on the value of a single feature. The specific test is selected by an information theoretic heuristic. This procedure is iterated on each of the new subsample and keeps on until a subsample cannot be separated or the partitioning tree has reached the threshold. The information gain ratio is defined as:

$$\text{Information Gain Ratio (D, S)} = \frac{Gain(D,S)}{H(\frac{|D_1|}{D},\ldots,\frac{|D_S|}{D})} \tag{2}$$

, where D is a database state, H $(\cdot)$ finds the amount of order in that state, when the state is separated into S new states S =$\{D_1,\ D_2,\ldots,\ D_S\}$.

The method of C5.0 is very robust for handling missing data and in a large number of input fields [16]; therefore, C5.0 is used to evaluate our features in this paper.

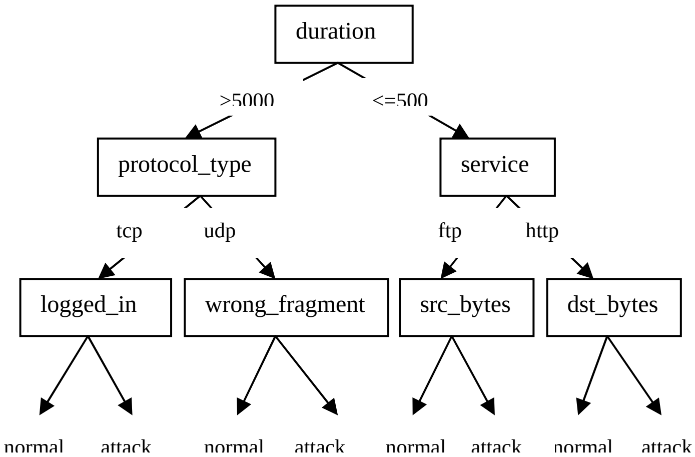

Fig. 1: Basis structure of C5.0

3. Intrusion Data set

In this paper, we choose the KDD Cup 1999 data set which was originally provided by MIT Lincoln Labs (The 1998 DARPA Intrusion Detection Evaluation Program) as the evaluation data set [2, 11, 12]. The data set was later prepared for KDD competition) (see "http://www.ics.uci.edu/˜kdd/databases/kddcup99/kddcup99.html" for more detail)

The data set is the real data which captured in the real network. It includes many kinds of attack data, also includes the normal data. The raw data was processed in to 39 attack types. These attacks are divided into four categories: probing (surveillance and other probing, e.g., port scanning), DoS (denial-of-service, e.g., SYN flood), U2R (unauthorized access from a user to root privilege, e.g., various "buffer overflow" attacks) and R2L (unauthorized access from remote to local machine, e.g., guessing password). For each TCP/IP connection, 41 input features plus one class label were extracted in the data set belonging to four kinds (9 basic Features, 13 Content Features, 9 Time-based Features and 10 Host-based Features) [12]. In Table 1, a total of 22 training known attack types, and additional 17 unknown types are summarized.

Table 1: Detail attack types [2]

Class	Known attack	Unknown attack
Probe	ipsweep, nmap, portsweep, satan	saint, mscan
DoS	back, land, Neptune, pod, smurf, teardrop	apache2, processtable, udpstorm, mailbomb
U2R	buffer_overflow, loadmodule, perl, rootkit	xterm, ps, sqlattack
R2L	ftp_write, guess_passwd, imap, multihop, phf, spy, warezclient, warezmaster	snmpgetattack, named, xlock, xsnoop, sendmail, httptunnel, worm, snmpguess

In this study, Training data set in the paper contained 49, 451 records, which were randomly generated from the KDD Cup 1999 for 10% training data set that consists of 9, 768 normal patterns, 39, 085 known DoS patterns, 435 known Probe patterns, 111 known R2L patterns and 52 known U2R patterns.

Test data set in the paper composed of three different test data sets, which were randomly selected from the KDD Cup 1999, 100% test data set. Table 2 gives the number of records on three different test data sets

Table 2: The number of records on three different test data sets

Dataset Name	Known attack	Unknown attack
Dataset-1	186, 745	19, 820
Dataset-2	49, 438	14, 781
Dataset-3	25, 419	10, 031

4. Proposed Approach

The proposed approach is used to select robust features to build model for the detection of known and unknown attacks. Note that the data which meets the demands of proposed methods must be numerical value. Therefore, the symbolic data should be transformed into numerical and make them under the same evaluation standard.

In proposed approach, the Euclidean Distance from equation (1) is used to compute ranking score between each attribute and class label by defining each attribute of KDD Cup 1999 training set, 41 attributes, as $A_1, A_2, A_3, \ldots, A_{41}$ respectively and class label as B; moreover, let x is a value in any attribute and y is a values in class label.

Let any $A_j = \{x_{1,j}, x_{2,j}, x_{3,j}, \ldots, x_{n,j}\}$ be a vector of attributes, where j ($1 \leq j \leq 41$) is an ordinal number of attributes of training set, and also n ($n \geq 0$) is the number of instances of training set.

Let $B = \{y_1, y_2, y_3, \ldots, y_n\}$ be a vector of class label, where n ($n \geq 0$) is the number of instances of training set.

Thus, the ranking score is $\{d_1(A_1, B), d_2(A_2, B), d_3(A_3, B), \ldots, d_{41}(A_{41}, B)\}$, where any

$$d_j(A_j, B) = \sqrt{\sum_{i=1}^{n}(x_{i,j} - y_i)^2} \tag{3}$$

where j ($1 \leq j \leq 41$) is an ordinal number of attributes of training set, and also n ($n \geq 0$) is the number of instances of training set.

After computing distance measure, the distance is score of known detection method of each attribute, $\{d(A_1, B), d(A_2, B), d(A_3, B), \ldots d(A_{41}, B)\}$. Then, sort scores of the ranking score from highest to lowest. Finally select features that have high scores to build model, which is used to detect accurately known and unknown attacks. The method of C5.0 is used to evaluate features that are taken from last step.

	Attributes					Class Label
Instances	**A_1**	**A_2**	**A_3**	**…**	**A_{41}**	**B**
	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	…	$x_{1,41}$	y_1
	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	…	$x_{2,41}$	y_2
	$x_{3,1}$	$x_{3,2}$	$x_{3,3}$	…	$x_{3,41}$	y_3
	⋮	⋮	⋮	⋮	⋮	⋮
	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	…	$x_{n,41}$	y_n

Fig. 2: Vectors of each attributes and a vector of class label

5. Experiments

In this section, an investigation on the performance of proposed feature selector based Euclidean is studied. The data sets (one train set and three test sets) described in Section 3 and C5.0 described in Section 2.4 are used to evaluate the proposed approach.

Note that the measurement in this paper of the experimental results is based on the standard metrics for evaluations of intrusion, Detection rate (TP) refers to the ratio between the number of correctly detected attacks and the total number of attacks while false alarm rate (FP: false positive) means the ratio between the number of normal connections that are incorrectly misclassified as attacks and the total number of normal connections.

From Table 3, the different between scores 441.72 and 360.65 is the highest value. Therefore, features that have scored more than four hundred scores are selected, getting 30 important features out of 41 features that show in Table 4. From Fig. 3, the proposed approach shows impressive detection rate of known attack for normal, DoS and Probe while the proposed approach does not demonstrates impressive detection rate for R2L and U2L. Maybe the number of records of R2L and U2L is 52 from 5 million records in the dataset. It is quite small. Moreover, the *warezclient* attack belonging to R2L is the majority patterns in the training set. However, in the test set, *guess_passwd* and *warezmaster* comprises most patterns of R2L. On the other hand, form Fig. 4, the proposed approach does not shows efficiency when it is used to detect unknown attack, but it can detect unknown attack for normal, Probe and U2L especially normal. It can detect quite excellent. Fig 5 shows the overall detection rate of the proposed approach on three different test sets. Fig 6 shows overall false positive rate of the proposed approach on three different test sets.

Table 5 (overall accuracy of C5.0 using the proposed approach to select features based Euclidean) shows capability of the proposed approach when it is used to detect known attack although it cannot show impressive used to detect unknown attack when comparing with detection rate of known attack. Furthermore, Table 6 shows overall false positive rate of C5.0 using the proposed approach to select features based Euclidean. Results from table 6 demonstrate that the proposed approach has drawback when used to detect unknown attack since percentage of overall false positive rate (FP) of unknown attack is quite high even if it is not more than 50 percent.

Table 3: The ranking score computed by Euclidean (in Section 4)

Feature name	Scores
dst_host_srv_serror_rate	499.24
srv_serror_rate	499.23
serror_rate	499.04
dst_host_serror_rate	498.99
srv_rerror_rate	486.19
reerror_rate	486.07
dst_host_srv_rerror_rate	485.75
dst_host_rerror_rate	485.38
logged_in	484.72
root_shell	483.39
land	483.38
urgent	483.37
num_compromised	483.37
su_attempted	483.37
src_bytes	483.37
num_failed_logins	483.37
num_root	483.37
num_shells	483.36
num_file_creations	483.36
num_access_files	483.32
dst_bytes	483.30
is_guest_login	483.09
host	483.03
duration	483.01
wrong_fragment	482.46
dst_host_srv_diff_host_rate	481.21
srv_diff_host_rate	480.04
diff_srv_rate	476.68
dst_host_diff_srv_rate	474.17
protocol_type	441.72
service	360.65
dst_host_count	256.91
serv_count	225.08
dst_host_same_src_port_rate	224.81
same_srv_rate	223.86
dst_host_same_srv_rate	222.52
dst_host_srv_count	219.26
count	190.59
flag	152.15
num_outbound_cmds	NaN
is_host_login	NaN

Table 4: 30 features extracted by Euclidean

duration, protocol_type, logged_in, serror_rate, srv_serror_rate, reerror_rate, srv_rerror_rate, diff_srv_rate, srv_diff_host_rate, dst_host_diff_srv_rate, dst_host_srv_diff_host_rate, dst_host_serror_rate, dst_host_srv_serror_rate, dst_host_rerror_rate, dst_host_srv_rerror_rate, src_bytes, dst_bytes, land, wrong_fragment, urgent, host, num_failed_logins, num_compromised, root_shell, su_attempted, num_root, num_file_creations, num_shells, num_access_files, is_guest_login

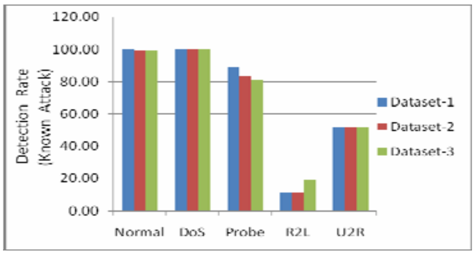

Fig. 3: Detection rate of known attack on three different test sets

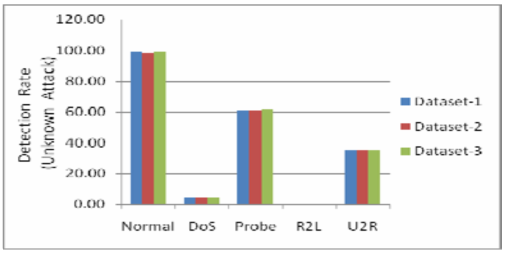

Fig. 4: Detection rate of unknown attack on three different test sets

Table 5: Overall detection rate of the proposed approach on three different test sets

Dataset Name	Known attack	Unknown Attack
Dataset-1	99.77%	56.52%
Dataset-2	99.32%	56.45%
Dataset-3	99.21%	56.59%

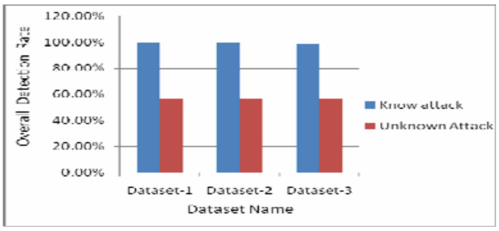

Fig. 5: Overall detection rate of the proposed approach on three different test sets

Table 6: Overall false positive rate of the proposed approach on three different test sets

Dataset Name	Known attack	Unknown Attack
Dataset-1	0.23%	43.48%
Dataset-2	0.68%	43.55%
Dataset-3	0.79%	43.41%

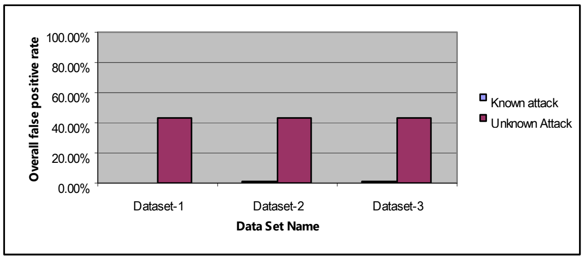

Fig. 6: Overall false positive rate of the proposed approach on three different test sets

6. The remarkable Conclusions

The proposed approach presented in this paper show that the feature selection method applied Euclidean Distance can extract the robust features to build model for the detection of known and unknown patterns, especially known patterns.

From the experimental results obtained, it is evident that the Euclidean-based feature selection is very promising over the known attack patterns. In addition, it produced smaller features showing other advantages because, in the real-world applications, the smaller features are always advantageous in terms of both data management and reduce the computing time. Therefore, the proposed approach can select a subset of robust features using smaller storage space and getting higher Intrusion detection performance, improving the performance of a true positive intrusion detection rate especially for detecting known attack patterns.

For the future work, the following directions are proposed: (1) setting the threshold by the automatic system and also (2) generating robust features to build model that can detect unknown patterns correctly.

7. Acknowledgements

We would like to esteem thanks Salvatore J. Stolfo, MIT Lincon Lab, and UCI KDD group for the KDDCUP 1999 data set.

8. References

[1] S. Hansman and R. Hunt. A Taxonomy of network and computer attacks. Computers & Security. 2005, 24, 31-43.

[2] C. H. Lee, S. W. Shin, and J. W. Chung. Network Intrusion Detection Through Genetic Feature Selection. *Proceeding of the Seventh ACIS International Conference on Software Engineering, Artificial Interlligence, Networking, and Parallel/Distributed Computing (SNPD'06).* 2006.

[3] R. H. Gong, M. Zulkernine, and P. Abolmaesumi. A Software Implementation of a Genetic Algorithm Based Approach to Network Intrusion Detection. *Proceedings of the Sixth International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing and First ACIS International Workshop on Self-Assembling Wireless Networks (SNPD/SAWN'05).* 2005.

[4] Y. Bai and H. Kobayashi. Intrusion Detection Systems: Technology and Development. *Proceeding of the 17th International Conference on Advanced Information Networking and Applications (AINA'03).* 2003.

[5] J. S. Han and B. Cho. Detecting intrusion with rule-based integration of multiple models. Computer & Security. 2003, 22, 613-623.

[6] S. Stolfo, W. Fan, W. Lee, A. Prodromidis, and P. K. Chan. Cost-based Modeling for Fraud and Intrusion Detection: Results from the JAM Project. *DARPA Information Survivability Conference*. 2000.

[7] S. Mukkamala and A. H. Sung. A comparative study of techniques for intrusion detection. *Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'03)*. 2003.

[8] G. John, R. Kohavi, and Pfleger. Irrelevant features and the subset selection problem. *Int. Conf. on Machine Learning, Morgan Kaufman, San Francisco*. 1994, 121-129.

[9] W. Duch, T. Winiarski, J. Biesiada, and A. Kachel. Feature Ranking Selection and Discretization. *Int. Conf. on Artificial Neural Networks (ICANN) and Int. Conf. on Neural Information Processing (ICONIP), Istanbul*. June 2003, pp, 251-254.

[10] K. Ron. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*. 1995, 2, 1137-1143.

[11] W. Hu, J. Li, and J. Shi. Optimal Evaluation of Feature Selection in Intrusion Detection Modeling. *Proceeding of the 6th world congress on Intelligent Control and Automation, Dalian, China*. June 21- 23 2006.

[12] W. Xuren, H. Famei, and X. Rongsheng. Modeling Intrusion Detection System by Discovering Association Rule in Rough Set Theory Framework. *International Conference on Computational Intelligence for Modelling Control and Automation, and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06).* 2006.

[13] http://www.itl.nist.gov/div897/sqg/dads/HTML/euclidndstnc.html

[14] http://people.revoledu.com/kardi/tutorial/similarity/EuclideanDistance.html

[15] A. Karnik, S. Goswami. and R. Guha. Detecting Obfuscated Viruses Using Cosine Similarity Analysis. *Proceedings of the First Asia International Conference on Modelling & Simulation (AMS'07)*. 2007.

[16] R. Yeh, C. Liu, B. Shla, Y. Cheng, and Y. Huwang. Imputing manufacturing material in data mining. *Springer Science+Business Media, LLC*. 2007.

[17] C. Chan, Y. Liu, and S. Luo. Investigation of Diabetic Microvascular Complications Using Data Mining Techniques. *International Joint Conference on Neural Networks (IJCNN 2008)*. 2008.

[18] J. Du and W. Guo. Data Mining on Patient Data. IEEE, 2005.

[19] A. Suebsing and N. Hiransakolwong. Feature Selection Using Euclidean Distance and Cosine Similarity for Intrusion Detection Model. *Asian Conference on Intelligent Information and Database Systems (ACIIDS 09)*. 2009.