Expert Systems with Applications 36 (2009) 5605–5612

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Data mining-based intrusion detectors

Su-Yun Wu [a], Ester Yen [b,*]

[a] Department of Information Management, Vanaung University, Taiwan
[b] Mathematical Sciences Research Institute, Berkeley, CA 94720-5070, USA

ARTICLE INFO

Keywords:
Classification tree
Support vector machine
Internet attack
Intrusion detection system (IDS)
.

ABSTRACT

With popularization of internet, internet attack cases are increasing, and attack methods differs each day, thus information safety problem has became a significant issue all over the world. Nowadays, it is an urgent need to detect, identify and hold up such attacks effectively. The research intends to compare efficiency of machine learning methods in intrusion detection system, including classification tree and support vector machine, with the hope of providing reference for establishing intrusion detection system in future.

Compared with other related works in data mining-based intrusion detectors, we proposed to calculate the mean value via sampling different ratios of normal data for each measurement, which lead us to reach a better accuracy rate for observation data in real world. We compared the accuracy, detection rate, false alarm rate for four attack types. More over, it shows better performance than KDD Winner, especially for U2R type and R2L type attacks.

1. Introduction

In recent years, as internet and personal computers are populated, utilization rate of internet keeps increasing. It is changing people's lives gradually, and the majorities of people study, recreate, communicate and buy through internet. Besides common people, enterprise structure and business mode also undergoes transformation due to internet, and large enterprise or government organizations, in order to achieve operation purpose and efficiency, develop many application and service items resting on internet; these are an irresistible tendency in the new era.

However, though internet brings about convenience and real-timeliness, consequently comes information safety problem; for example: servers are attacked and paralyzed, inner data and information are stolen, and so on. In the event of such cases, big losses may be caused in money and business credit. For example, in 2000, American Yahoo was subject to DDos attack, the servers were paralyzed for 3 hours approximately, 1 million users were affected, and the losses involved were too large to calculate. Other famous business internets, such as CNN, eBay, Amazon.com, Buy.com, and so on, also suffered such internet attacks.

Because of convenience of internet, it is easy to get access to attack knowledge and methods. At present, hackers are unnecessary to have a wide knowledge of specialized knowledge, and annual internet attack cases are increasing to a great extent. According to the statistics of American Computer Emergency Response Team/Coordination Center (CERT/CC) (http://www.cert.org/), annual network attack cases showed index growth, in recent years; according to the report of Information Security (http://www.isecutech.com.tw/), internet attacks have became new weapon of world war, and the report said that Chinese Military Hacker had drew up plan, with the view of attacking American Aircraft Carrier Battle Group to make it lose fighting capacity through internet. Such information reveals that it is an urgent need to effectively identify and hold up internet attacks nowadays.

Common enterprises adopt firewall as the first line of defense for internet safety, but the main function of firewall is to supervise accessing behaviors of internet, and it owns limited detection capacity for internet attacks. Therefore, Intrusion Detection System, IDS is always applied to detect internet *encapsulation*, to improve protective capacity of internet safety.

IDS appears like internet supervision and alarm device, to observe and analyze whether the internet attacks may occur, timely send alarm before risks are caused by attacks, execute corresponding response measures, and reduce occurrence of bigger losses. Moreover, some technologies are based on pattern check, with low mis-judgment rate, but the pattern-based should be upgraded on a regular basis, such technologies do not possess enough detection capacity for unknown and renewed attack manners. Recently, many researches applied the technology of data mining and machine learning, which can analysis bulk data, and such technologies own better detection capacity for unknown attacks. Though some research achievements have been scored, there is a lot of development potential.

* Corresponding author.
E-mail address: ester_yen@yahoo.com (E. Yen).

0957-4174/$ - see front matter Crown Copyright © 2008 Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.06.138

Under such circumstance with most same conditions, how is the efficiency of different machine learning methods applied in intrusion detection. Besides the said manners, what methods are there? Therefore, the research intends to compare the efficiency of different machine learning methods applied in intrusion detection, include classification tree, support vector machine, and so on, with the hope of providing possible suggestion for improvement, as the reference for building intrusion detection system.

The research process is shown in Fig. 1.

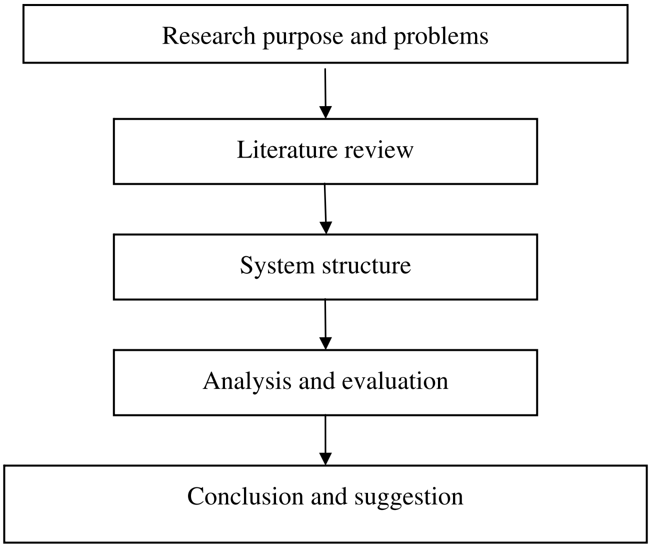


Fig. 1. Research flow.

2. Literature review

2.1. Introduction on intrusion detection system

The concept of intrusion detection system was first suggested in a technical report by Anderson (1980); he considered that computer audit mechanism should be transformed and able to provide internal risks and threats for computer safety technicians, and suggested that statistics method should be applied to analyze users' behavior and detect those masqueraders who accessed system sources illegally. In 1987, Dorothy suggested a prototype of intrusion detection system: IDES (intrusion detection expert system), afterwards, the concept of intrusion detection system was known gradually, and his paper was also regarded as a significant landmark in intrusion detection area. Following this, intrusion detection system with various patterns was put forward, such as: Discovery, Haystack, MIDAS, NADIR, NSM, Wisdom and sense, DIDS, and so on (Bace, 2002).

Intrusion detection system is to supervise and control all cases happening to computer system or network system, analyze any signal arising from related safety problems, send alarms when safety problems occur, and inform related personnel or units to take relevant measures to reduce possible risks (Bace, 2002). Its framework includes three parts (Bace, 2002):

1. Information collection: Data collection: the source of these collected data can be separated into host, network and application, according to the position.
2. Analysis engine: Analysis engine is able to analyze whether or not there are symptom of any intrusion.
3. Response: Take actions after analysis, record analysis results, send real-time alarm, or adjust intrusion detection system, and so on.

2.1.1. Classification of intrusion detection system

Generally speaking, there are two kinds of classification methods for intrusion detection system:

1. According to different data sources, intrusion detection system includes host-based IDS and network-based IDS.
2. According to different analysis methods, intrusion detection system includes Misuse Detection and Anomaly Detection.

The following is to give a brief introduction on property, advantage and disadvantage of these intrusion detection systems.

(a) Classification based on different information source:

- (Host-based IDS) (Bace, 2002): Its data comes from the records of various activities of hosts, including audit record of operation system, system logs, application programs information, and so on. Taking Windows NT operation system as an example, its event logs mechanism searches and collects three patterns of system events: Operation system event, safety event and application event; and examples of application program information are as follows: Database system, WWW servers, and so on. Its advantage and disadvantage are stated as follows (Ertoz et al., 2004):
  - Advantage:
    1. It can judge whether or not the host is intruded more accurately: Because its data comes form system audit records and system logs of hosts, comparing with network-based intrusion detection system, it can more accurately judge network attacks or intrusion on hosts.
    2. It can detect attacks under encrypted network environment: Because the data comes from system files and transmitted encrypted data in network which are decrypted in hosts, thus the data is not affected.
    3. It does not need additional hardware: It just needs monitoring system installed in specified hosts, without additional hardware.
  - Disadvantage:
    1. Higher cost: Monitoring systems must be installed in each host; and because of different hosts, the audit files and log pattern are accordingly different, thus different intrusion detection systems are required in each host.
    2. It may affect system efficiency of monitored hosts: Intrusion detection system in monitoring state may occupy system sources of hosts.
- (Network-based IDS) (Bace, 2002): Its data is mainly collected network generic stream going through network segments, such as: Internet packets. And its advantage and disadvantage are stated as follows:
  - Advantage:
    1. Low cost: Only network-based IDS can detect all attacks in a LAN, and the cost is just for the device.
    2. It can detect attacks that cannot be done by host-based IDS, such as: Dos, DDos.
  - Disadvantage:
    1. The flux is large, and some packets may be lost, and it cannot detect all packets in network.
    2. In large-scale network, it requires more rapid CPU and more memory space, to analyze bulk data.
    3. It cannot deal with encrypted packets, and it may not receive attack information in encrypted packets accordingly.

(b) Classification based on different analysis method:
- Misuse Detection (Bace, 2002): It is also named signature-based detection, which can transform the information of attack symptom or policy disobeying into state transition-based signature or rule, and such information is stored in signature database. To judge whether or not it is attack, pre-treated case data should be first compared with the signature of signature database, and those conforming to attack signature data can be judged as attack. Its advantage is high detection rate and low false alarm rate for known attacks; however, its detection capacity is low for un-known detection methods, and attack database should be renewed on a regular basis.
- Anomaly Detection (Bace, 2002): It may establish a profiles for normal behavior of users, which comes from statistics data of users in the former period; when detection is performed, the profiles is compared with actual users' data, if the offset is below threshold value, user's behavior can be considered normal, and it has no intention of attacks; if the offset is above threshold value, user's behavior can be considered abnormal. Anomaly detection is based on an assumption that intruder's behavior is different from normal users' behavior. Detection rate of the method is high, and it is more likely to detect un-known attacks, but mis-judgment rate is also high.
- Hybrid: The advantage of misuse detection is low mis-judgment rate, as well as low detection capacity for unknown attacks; comparatively, anomaly detection owns the capacity of detecting unknown attacks, but with high mis-judgment rate. If these said two methods are combined for detection, they can supply disadvantage of each other, such as: MINDS (Ertoz et al., 2004), EMERALD, Prelude, and so on.

2.1.2. Current analysis method

The following is to describe current analysis methods applied in intrusion detection system (Bace, 2002; Lu, Boedihardjo, & Manalwar, 2005; Patcha & Park, 2007; Verwoerd & Hunt, 2002):

- State transition analysis (Ilgun, Kemmerer, & Porras, 1995): State transition is applied to describe the relation of arising events, which is usually used for misuse detection.
- Statistical models: Statistics method is applied to construct normal behavior mode, including: threshold measures, mean and standard deviation, multivariate model, clustering and outlier detection (Jiang et al., 2006), which is usually used for anomaly detection.
- Neural network: Before detection, some time is needed for training, detection can be begun after constructing mode, and it includes: back-Propagation, SOM (self-organization map), and so on.
- Bayesian network (Ben Amor, Nahla, Benferhat Rue, Salem, & Elouedi, 2004): Graph method is applied to express the relation among variables; when performing detection, conditional probability is used to calculate proper detection value.
- Rule-based: Behavior or mode is expressed by rule method, and those conforming to rule can be judged to be attack behavior. It is commonly used for misuse detection.
- Data mining (Machine learning) methods: It concludes Markov process model (Kuo-Hua Yang, 2006), classification tree (Ben Amor et al., 2004; Yu-Shan Hsu, 2006), support vector machine, association rule, link analysis, sequence analysis, and so on.
- Other method: Other methods include immune system approaches (Aickelin, Greensmith, & Twycross, 2004), genetic algorithm, agent-based detection, and so on.

2.2. Machine learning

Machine learning (http://en.wikipedia.org/wiki/Machine_Learning) is a sub-area of artificial intelligence, which mainly develops some technologies that qualify computer to learn automatically. Its stress of research is how to use computer and statistics method to select useful information from bulk data. Therefore, machine learning and data mining are correlated closely to statistics method and theory of computer science.

Machine learning is widely applied in various areas currently, such as: Biological signature differentiation, search engine, medical diagnosis, bond market analysis, pronunciation and handwriting identification, computer vision, and so on. The following is to list common machine learning technologies (Alpaydin, 2004):

- Bayesian decision theory
- Multivariate methods
- Clustering
- Classification trees
- Linear discrimination
- Multilayer perceptions
- Local models
- Hidden Markov models
- Reinforcement learning

2.3. Classification tree

Classification tree is a prediction mode in machine learning, and it is also called Decision tree. It is tree pattern graph similar to flow chart structure; any internal node is a test property, each branch represents test result, and final nodes of leaves represent distribution situation of various types. The most fundamental and common algorithm used for classification tree is ID3 and C4.5 (Quinlan, 1993). There are two methods for tree construction. top-down tree construction and bottom-up pruning, and both ID3 and C4.5 belong to top-down tree construction; their algorithm is described as follows:

1. All paradigms of training data are placed into root of classification tree.
2. If the node does not contain any data, or the data at the node belongs to the same type, the node becomes empty leaf or all paradigms leaf of the same type; if the node contains more than one type of paradigms, it is necessary to assess all properties of data, according to certain assessment function, and a proper property is selected. According to the value of the property, paradigms at the node are divided into *N* parts, and each part is a new node connecting root node. The process is named splitting node.
3. After splitting of nodes, judge whether or not these now nodes are leaves; if not, new nodes are the root of sub-tree and used to construct new sub-tree.
4. The said steps proceed continuously with recursion method, until all new nodes are leaves. Decision tree got through the type of inductive method can totally classify paradigms of training data.

Assessment function in the second step usually applies impurity function for assessment; after calculation by impurity function, the property with greatest value will be selected, and impurity function includes:

Suppose that S is a set constituted by data sample s, and contains m C_i ($i = 1,\ldots,m$) with different labels, s_i is the sample of C_i type in S set, P_i is the probability of any sample possibly belonging to C_i, namely s_i/S; suppose again that A property contains v sub-set, S_j represents the set constituted by a_j samples belonging to A property in S set. When a property is selected as test property, it will contain samples of S_j set where v represents C_i type.

- Entropy function: $-\sum_{i=1}^{m} P_i \log_2 P_i$ (used for ID3 and C4.5).
- Information gain: $Gain(A) = I(T) - E(A)$.In the formula above: $I(T) = -\sum_{i=1}^{m} P_i \log_2 P_i$, $E(A) = \sum_{j=1}^{v} \frac{s_{1j}+\cdots+s_{mj}}{s} I(s_{1j},\ldots,s_{mj})$.
- Gain ratio: $GR(A) = \frac{Gain(A)}{I(A)}$, in the formula above: $I(A) = \sum_{j=1}^{v} \frac{S_j}{S} \log_2 \frac{S_j}{S}$.
- Gini function: $Gini(T) = 1 - \sum_i p_i^2$.After tree is constructed, to avoid overfit, constructed classification tree should be pruned, and there are usually two methods for this (Han & Kamber, 2006):
- Pre-pruning: In construction of tree, set a threshold, when *ramification* point condition is more than threshold, stop the construction of the tree from the ramification point.
- Post-pruning: After tree is finished, then prune it. Common methods include subtree replacement and subtree raising; principles for pruning assessment include error estimation, significance testing, and so on.

At last, after tree is finished and pruned, rules will be produced according the tree, shown in Fig. 2.

We use C4.5 type in the research.

2.4. Support vector machine

Support vector machine is put forward by Vapnik (1995); recently, it is widely applied in various areas, and it becomes a popular method in machine learning area.

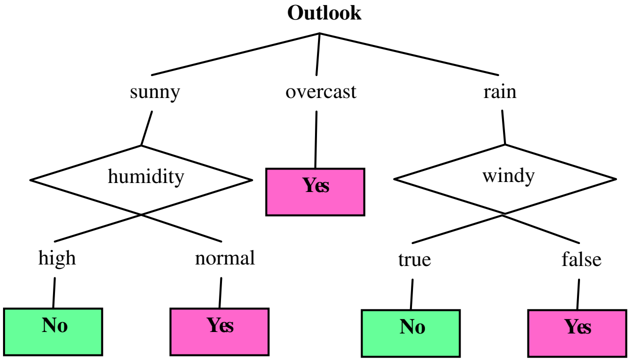
2.4.1. Optimize separate hyperplane

First, we take linearly separable as an example. Suppose that there is two types of classification problem now, class label r^t value applied by it is, its dataset is a vector of X, X^t. therefore, there is a set of data in the dataset, and support vector machine is to find out a hyperplane which will differentiate these datasets according to type. Taking the following graph as an example, It is to seek parameters w and b which is able to differentiate these two types of line $L2 = wx + b$. when performing classification later, we need know the value of $wx + b$, if it is above 0, it should be +1; and if is below 0, it should be -1 (see Fig. 3).

Support vector machine is to find out the hyperplane which is the most distant away from any data, this can minimize error rate.

The distance between data and hyperplane is shown in Fig. 4.

Support vector machine must meet the following conditions:



If outlook = sunny and humidity = high then play = No

If outlook = rain and windy = true then play = No

If outlook = overcast then play = Yes

Fig. 2. Classification tree mode and production rule.

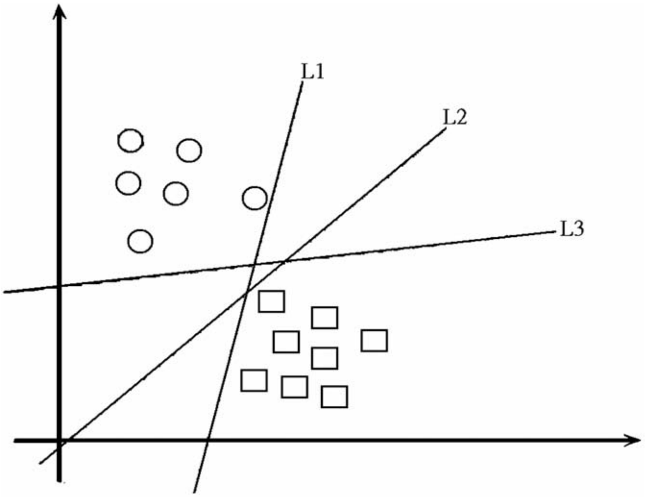


Fig. 3. Support vector machine.

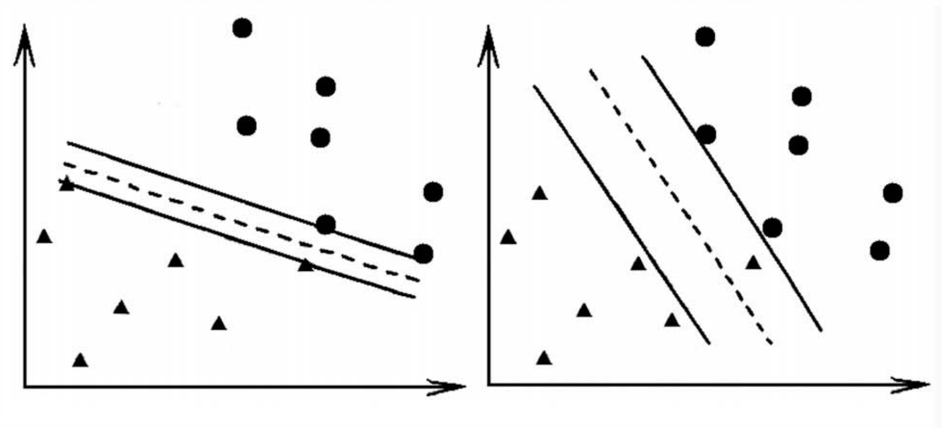


Fig. 4. Left: small distance between data and hyperplane and right: big distance between data and hyperplane.

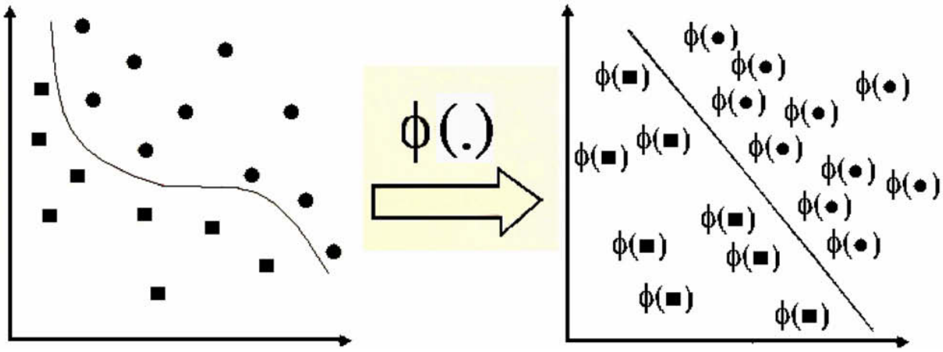


Fig. 5. Non-linear separable cases.

$$wX^t + b > +1,$$
$$wX^t + b < -1 \quad \forall t.$$

It also can be written into $r^t(wX^t + b) > + 1$.

We should make the distance between data X^t and the hyperplane bigger than ρ, and make ρ maximum.

$$\frac{r^t(wX^t + b)}{\|W\|} \geqslant \rho \quad \forall t.$$

To get unique solution, make $\rho\|W\| = 1$; the said formula above can be transformed into quadratic optimization problem:

$$\min \frac{1}{2}\|W\|^2 \quad \text{subject to } r^t(wX^t + b) \geqslant +1 \quad \forall t.$$

Lagrange multipliers can be used for calculation.

Discrimination function can be got:

$$f(x) = \sum_i \alpha_i y_i x_i^T x - b^*$$

2.4.2. Non-linearly separable situation

When it is non-linearly separable situation, basic function $\phi(x)$ can be adopted, to make original non-linear feature space correspond to linear feature space, and then linear method can be applied for calculation, shown in Fig. 5.

After basis function is used, discrimination function may have inner product $\Psi(x^t)^T\Psi(x)$ of basis function, and another function $K(x^t,x)$ is applied to replace basis function, and the function is called kernel function.

3. System structure

3.1. System structure graph

The proceeding flow of the research is shown in Fig. 6.

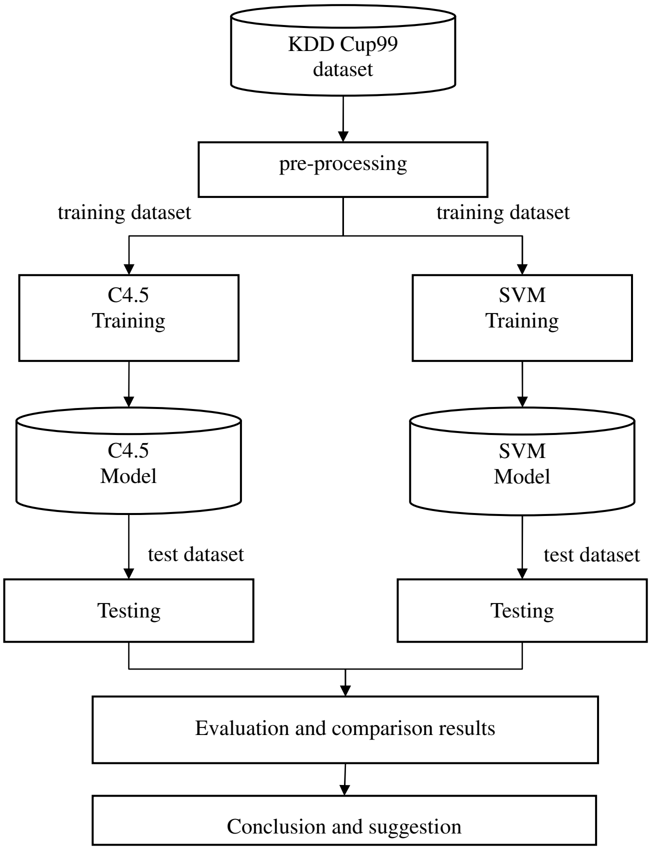


Fig. 6. System structure graph.

3.2. KDD Cup 99 dataset

The data applied in the research comes from KDD Cup 99 dataset, which was initially used for The Third International Knowledge Discovery and Data Mining Tools Competition. The dataset was got through selecting and arranging the data of DARPA of American Air Force by American Columbia University in MIT Lincoln in 1998, and it was intended to assess the efficiency of intrusion detection algorithm. Therefore, the research also applies the dataset.

There are approximately 4,940,000 kinds of data in training dataset, 10% of which is provided; there are 3,110,291 kinds of data in test dataset, and there are totally 47 types of network connection characteristic (characterized by continuous data and discrete data) in each kind of network connection record. And its property can be divided into three major types: Basis characteristic of network connection, characteristic of network connection content, network transmission characteristic; Data pattern include nominal, binary and numeric. Refer to Attachment I for detailed property. There are 23 types of attacks contained in training information, and 37 types of attacks contained in test information, 14 types of attacks more than training information, thus test information can be used to assess the detection capacity for unknown attacks. The attacks contained in test information can be separated into the following major types:

- Probe: Strictly speaking, it should not be regarded as true attacks but preparation step of attackers before launching attacks. Attackers usually apply probe to get information, to determine the targets and the type of operation system.

- Dos (Denial of service): Such attack may cause the stop of server operation, and the server cannot provide services. The attack usually occupies all system source of server, or occupies the band width and disables system resource and makes operation stop. Common attacks are SYN Flooding, Ping Flooding, and so on.
- U2R (User gain root): In the attack, users take advantage of system leak to get access to legal purview or administrator's purview, such as: Buffer Overflow is among them.
- R2L (Remote file access): The attack is to apply the advantage of server providing services, to get related safety setting or user's encrypted files, such as: Unicode leak, SQL Injection, and so on.

Table 1 lists attack pattern and type.

Table 2 lists the percentage of various data in training data and 10% kddcup.data_10_percent.gz.

3.3. Preprocess of data

The research intends to compare the efficiency of C4.5 and SVM under different circumstances, and KDD Cup 99 dataset is over large, various data is distributed unevenly, thus the research will sample training dataset (10% kddcup.data_10_percent.gz) and test dataset. Based on the normal proportion, select each 10,000 group of data where normal proportion is 10%, 20%, 30%, . . ., 90% in training dataset and test dataset; and make remaining data, namely attack data, even and sample them.

Besides sampling, C4.5 software applied in training stage is Weka 3.5.6, whose input form is arff; SVM software applied is Libsvm developed by a professor from Taiwan University Information Engineering Department, whose input form is also specified, thus the form of all data should be converted. Weka 3.5.6 also can convert information into SVM form, thus the research applies Weka software for conversion.

3.4. Training and testing

After pre-position treatment of data, training and test can be begun. C4.5 function in Weka is J48, some parameters need to be set, including reducedErrorPruning, confidenceFactor, minNumObj. Training stage of Libsvm also requires setting parameters, to provide python program seeking optimization parameter: grid.py, which is applied in the research to seek optimal parameter.

4. Analysis and evaluation

Detection and identification of attack and non-attack behaviors can be generalized as the following Table 3:

- True positive (TP): the amount of attack detected when it is actually attack.
- True negative (TN): the amount of normal detected when it is actually normal.
- False positive (FP): The amount of attack detected when it is actually normal, namely false alarm.

Table 2
Percentage of various data

Data type	Quantity	Percentage
Normal	97,277	19.69
Probe	4107	0.83
Dos	391,458	79.24
U2R	52	0.01
R2L	1126	0.22

Table 3
Confusion matrix

	Predicted attack	Predicated normal
Actual attack	True positive (TP)	False negative (FN)
Actual normal	False positive (FP)	True negative (TN)

Table 4
Accuracy comparison between C4.5 and SVM

Percentage of normal data (%)	C4.5 (%)	SVM (%)
10	40.67	35.35
20	43.46	41.19
30	51.22	46.93
40	57.27	54.85
50	64.53	63.24
60	72.24	69.01
70	78.70	76.36
80	84.45	85.14
90	91.90	92.21
Average	64.94	62.70

- False negative (FN): The amount of normal detected when it is actually attack, namely the attacks which can be detected by intrusion detection system.

Nowadays, intrusion detection system requires high detection rate and low false alarm rate, thus the research compares accuracy, detection rate and false alarm rate, and lists the comparison results of various attacks.

4.1. Comparison of accuracy

Accuracy refers to the proportion of data classified an accurate type in total data, namely the situation TP and TN, thus the accuracy is

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} * 100\%$$

Table 4 lists the results measured by original class label classification.

Fig. 7 is fold line figure.

There is not significant difference between accuracy of the two methods; however, C4.5 is better than SVM when the proportion of normal information is small; when the proportion of normal information is large (>70%), their accuracy is approximately equal, but

Table 1
KDD Cup 99 dataset attack pattern and classification

Attack type	Attack pattern
Probe	Ipsweep, nmap, portsweep, satan, mscan, saint
Dos	back, land, neptune, pod, smurf, teardrop, apache2, mailbomb, processtable, udpstorm
U2R	Buffer_overflow, loadmodule, perl, rootkit, ps, sqlattack, httptunnel, xterm
R2L	ftp_write, guess_passwd, imap, multihop, phf, spy, warezclient, warezmaster, snmpgetattack, named, xlook, xsnoop, snmpguess, worm, sendmail

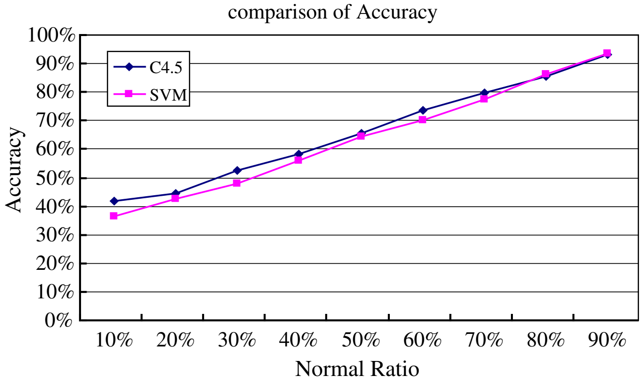


Fig. 7. Accuracy comparisons between C4.5 and SVM.

SVM is much better. According to the average, C4.5 is slightly better than SVM.

4.2. Comparison of detection rate

Accuracy refers to the proportion of attack detected among all attack data, namely, the situation of TP, thus detection rate is

$$\text{Detection rate} = \frac{\text{TP}}{\text{TP} + \text{FN}} * 100\%$$

Table 5 lists comparison results of detection rate between C4.5 and SVM.

Fig. 8 is its fold line figure.

In detection rate, C4.5 declines as the percentage of normal data rises, but SVM is not fixed. Integrally speaking, Curve of C4.5 is above that of SVM; obviously, its detection rate is better than that of SVM. According to the average value, C4.5 surpasses SVM by 12% approximately.

4.3. Comparison of false alarm rate

False alarm rate refers to the proportion that normal data is falsely detected as attack behavior, namely, the situation of FP, thus false alarm rate is

$$\text{False alarm rate} = \frac{\text{FP}}{\text{FP} + \text{TN}} * 100\%$$

Table 6 lists comparison results of false alarm rate between C4.5 and SVM.

Fig. 9 is its fold line figure.

In comparison of false alarm rate, SVM is inferior to C4.5 only when the proportion of normal information is 30%, 50% and 60%, but it is better than C4.5 otherwise. According to the average value, SVM is better C4.5 in false alarm rate.

4.4. Accuracy comparison between different attacks

Accuracy of various attacks refers to the proportion that the type of data is corrected classified, and there are four types compared in the research. They are Probe, Dos, U2R, R2L. Table 7 lists comparison results of accuracy of various attacks by C4.5 and SVM; two data in the table is accuracy information of C4.5 when it is above, and they are accuracy data of SVM when it is below.

From the table, we can know that

- For Probe attack: Accuracy of SVM is better than that of C4.5 when the proportion of normal information is 20%, 50% and 90%, but it is not more than 4%; however, in other circumstances, C4.5 surpass SVM by 10%. Obviously, the accuracy of C4.5 is better than that of SVM in such attack detection.

Table 5
Detection rate comparison between C4.5 and SVM

Percentage of normal data (%)	C4.5 (%)	SVM (%)
10	73.79	67.84
20	70.90	64.61
30	74.09	58.34
40	66.47	52.98
50	70.14	66.08
60	69.15	45.08
70	68.70	40.60
80	65.30	48.20
90	65.00	58.50
Average	70.62	58.68

Table 6
False alarm rate comparison between C4.5 and SVM

Percentage of normal data (%)	C4.5 (%)	SVM (%)
10	3.50	2.60
20	2.25	2.10
30	2.13	2.57
40	1.33	0.65
50	1.14	1.44
60	0.95	1.60
70	1.17	0.47
80	1.78	0.26
90	1.23	0.63
Average	1.44	1.00

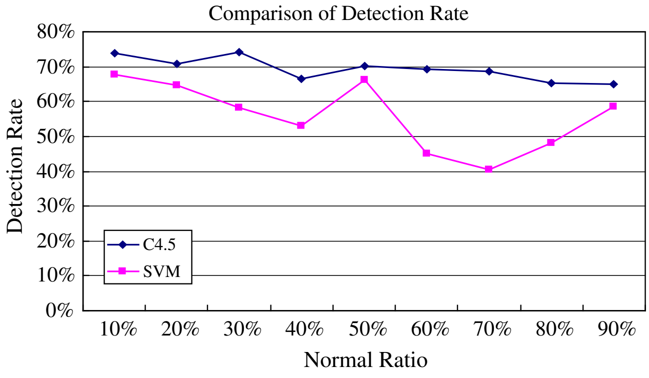


Fig. 8. Detection rate comparisons between C4.5 and SVM.

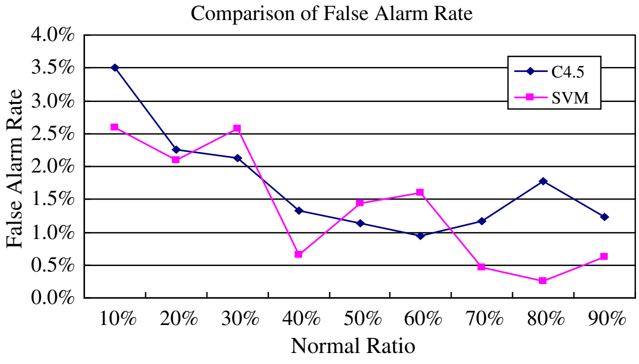


Fig. 9. False alarm rate comparisons between C4.5 and SVM.

Table 7
Accuracy comparison of four kinds of attacks by C4.5 and SVM

Percentage of normal data (%)	Probe (%)	Dos (%)	U2R (%)	R2L (%)
10	90.86	62.53	58.63	20.15
	86.01	62.01	43.81	22.59
20	80.35	62.87	47.82	19.15
	83.52	68.52	48.98	18.01
30	93.94	63.00	51.38	16.88
	83.84	66.16	34.64	19.37
40	83.63	61.50	53.93	17.45
	83.63	55.92	40.39	9.27
50	83.75	60.95	44.74	15.70
	87.07	59.22	50.23	24.29
60	84.42	71.51	50.51	14.44
	77.27	52.26	37.51	10.60
70	84.46	66.67	27.59	17.33
	70.41	39.37	17.08	14.69
80	88.85	54.55	44.42	12.97
	77.37	55.10	39.06	12.83
90	78.79	58.30	57.92	13.70
	81.82	59.78	40.27	15.74
Average	86.30	62.96	50.06	17.43
	82.70	60.04	41.05	17.46

Table 8
Detection rate comparison of various attacks through KDD Winner, C4.5 and SVM

	Probe (%)	Dos (%)	U2R (%)	R2L (%)
KDD winner	83.30	97.10	13.20	8.40
C4.5	86.30	62.96	50.06	17.43
SVM	82.70	60.04	41.05	17.46

- For Dos attack: When the proportion of normal data is low, SVM is better; however, if the proportion of normal data is above 40%, especially, it is 40%, 60% and 70%, C4.5 is better than SVM.
- For U2R attack: Integrally speaking, C4.5 is better than SVM.
- For R2L attack: According to the average value, these two methods are similar in accuracy. When the proportion of normal data is 10%, 30%, 50% and 90%, SVM is better, and C4.5 is better otherwise.
- According to the average value, except that these two methods are similar in accuracy in R2L attack, C4.5 is superior to SVM in accuracy otherwise.

At last, average results got in the research is compared with the results obtained through KDD Cup 99 winner, shown in Table 8.

We can see that the accuracy of KDD Winner is very high in Dos attack, but it is far worse than C4.5 and SVM in U2R and R2L.

5. Conclusions and suggestions

5.1. Conclusions

The research compares accuracy, detection rate, false alarm rate and accuracy of other attacks under different proportion of normal information. KDD Cup 99 dataset is current benchmark dataset in intrusion detection; however, its data is not distributed evenly, error may occur if only one set is used. Therefore, in comparison, the research applies different normal data proportion for training and test, finally get one average value, and hopes to obtain a more objective results.

For comparison results of C4.5 and SVM, we finds that C4.5 is superior to SVM in accuracy and detection; in accuracy for Probe, Dos and U2R attacks, C4.5 is also better than SVM; but in false alarm rate, SVM is better.

5.2. Future research suggestions

- Dataset KDD Cup 99 applied in the research is popularly used in current intrusion detection system; however, it is data of 1999, and network technology and attack methods changes greatly, it cannot reflect real network situation nowadays. Therefore, if newer information is got and tested and compared refresh, they can more accurately reflect current network situation.
- Through test and comparison, the accuracy and detection rate of C4.5 is higher than that of SVM, but false alarm rate of SVM is better; if we combine the two methods, overall accuracy can be increased greatly.
- In sampling, the research supposes that the distribution of attack data other than normal data is even, which cannot surely get optimal results, and this should be improved and validated in future.
- C4.5 parameters set in the research is not optimal, thus the future work should optimize the parameters according to C4.5 parameters and different training dataset.
- SVM applied in the research uses its built-in grid.py to optimize its parameters, and it needs approximately 2 hours to search parameters for 10,000 groups of data in the research; however, it is not suitable, for intrusion detection system requires real-timeliness. The future research should aim at the direction where the parameters can be optimized rapidly.

References

Aickelin, Uwe, Greensmith, Julie, & Twycross, Jamie (2004). *Immune system approaches to intrusion detection – A review*. Berlin, Heidelberg: Springer.

Alpaydin, Ethem (2004). *Introduction to machine learning*. MIT Press.

Anderson, James P. (1980). *Computer security threat monitoring and surveillance*, technical report, James P. Anderson Co., Fort Washington, Pennsylvania.

Bace, Rebecca G. (2002). NIST special publication on intrusion detection systems.

Ben Amor, Nahla, Benferhat Rue, Salem, Elouedi, Zied. (2004). Naïve Bayes. vs. decision trees. In: *Symposium on applied computing proceedings of the 2004 ACM symposium on applied computing*.

Confusion Matrix. <http://www2.cs.uregina.ca/~dbd/cs831/notes/confusion_matrix/confusion_matrix.html>.

Dorothy, Denning. 1987. An intrusion detection model. *IEEE Transaction on Software Engineering*.

Event Monitoring Enabling Responses to Anomalous Live Disturbances (EMERALD). http://www.sdl.sri.com/projects/emerald/.

Ertoz, L., Eilertson, E., Lazarevic, A., Tan, P., Srivastava, J., Kumar, V., et al. (2004). *The MINDS – minnesota intrusion detection system. Next generation data mining*. MIT Press.

Han, J., & Kamber, M. (2006). *Data mining: concepts and techniques* (2nd ed.). Morgan Kaufmann Publishers.

Ilgun, K., Kemmerer, R. A., & Porras, P. A. (1995). State transaction analysis: A rule-based intrusion detection approach. *IEEE Transaction on Software Engineering, 21*(3).

Jiang, Sheng Yi et al. (2006). A clustering-based method for unsupervised intrusion detections. *Pattern Recognition Letters*.

KDD Cup99 Winner. <http://www.cse.ucsd.edu/users/elkan/clresults.html>.

Kuo-Hua Yang (2006). Intrusion Detection Systems based on Hybrid Hidden Markov Models and Naïve Bayes Classifiers. National Taiwan University of Science and Technology. URN: etd-0728106-205122. Master thesis.

Libsvm. http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

Lu, C. -T., Boedihardjo, A. P., Manalwar, P. (2005). Exploiting efficient data mining techniques to enhance intrusion detection systems. in: *Information Reuse and Integration, Conf 2005, IRI – 2005 IEEE International Conference on*.

Patcha, A., & Park, J.-M. (2007). Network anomaly detection with incomplete audit data. *Elsevier Computer Networks, 51*(13).

Prelude. <http://www.prelude-ids.org/>.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. Morgan Kaufmann Publishers.

Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer-Verlag.

Verwoerd, Theuns, & Hunt, Ray (2002). Intrusion detection techniques and approaches. *Computer Communications, 25*(15), 1356–1365.

Weka. <http://www.cs.waikato.ac.nz/ml/weka>.

Yu-Shan Hsu (2006). A Hybrid IDS Framework via Decision Trees and SVMs. National Taiwan University of Science and Technology. Master thesis.