Optimized Intrusion Detection by CACC Discretization Via Naïve Bayes and K-Means Clustering

Vineet Richhariya, Nupur Sharma
[1]Lakshmi Narain College of Technology, Bhopal, India

Abstract
Network Intrusion Detection System (IDS), as the main security defending technique, is second guard for a network after firewall. Data mining technology is applied to the network intrusion detection, and Precision of the detection will be improved by the superiority of data mining. For IDS many machine learning approaches are ad-acute but they all work efficiently on basis of the training data accuracy. In this paper we used CACC Discretization algorithm to improve the training data representation and then used a crossbreed way by using Naïve Bayes and K-Means clustering. The database Discretization performs well in terms of detecting attacks faster and with reasonable false alarm rate.

Index Terms
Intrusion Detection System, Discretization, Crossbreed approach, Clustering, Classification.

1. Introduction

A significant challenge in providing an effective defense mechanism to a network perimeter is having the ability to detect intrusions and implement countermeasures. Components of the network perimeter defense capable of detecting intrusions are referred to as Intrusion Detection Systems (IDS). Intrusion detection techniques have been investigated since the mid 80's and, depending on the type and source of the information used to identify security breaches; they are classified as host–based or network–based. [1]

Host–based systems use local host information such as process behavior; file integrity and system logs to detect events. Network–based systems use network activity to perform the analysis. Combinations of these two types are also possible. Depending on how the intrusion is detected, an IDS is further classified as signature–based (also known as misuse system) or anomaly–based. Signature–based systems attempt to match observed activities against well defined patterns, also called signatures. Anomaly–based systems look for any evidence of activities that deviate from what is considered normal system use. These systems are capable of detecting attacks for which a well–defined pattern does not exist (such as a new attack or a variation of an existing attack). A hybrid IDS is capable of using signatures and detecting anomalies.[2]

The present paper introduces an adaptive approach for intrusion detection. The normal system activity is described using data mining techniques, namely Naïve Bayes and K-Means clustering. The intrusion behavior detection is optimized by CACC algorithm, which is for efficient database Discretization.

The rest of this paper is organized as follows. In section II, we discuss the related works; in section III we give our proposed architecture and experimental setup and evaluates explains in section IV. Finally, section V presents our conclusion, some discussion and future work.

2. Related Work

Data mining is the latest technology introduced in network security environment to find regularities and irregularities in large datasets [3].

ADAM (Audit Data Analysis and Mining) [4] is an intrusion detector built to detect intrusions using data mining techniques. It first absorbs training data known to be free of attacks. IDDM [5] and MADAM ID (Mining Audit Data for Automated Models for Intrusion Detection) [6] is one of the best known data mining projects in intrusion detection.

In [7] Authors discuss Different classifiers can be used to form a hybrid learning approaches such as combination of clustering and classification technique.

Researchers in [8] apply Clustering that is an anomaly-based detection method that is able to detect novel attack without any prior notice and is capable to find natural grouping of data based on similarities among the patterns.

In [9] Authors use K-Means and DB-Scan to efficiently identify a group of traffic behaviors that are similar to each other using cluster analysis.

Researchers in [10] has state that Naïve Bayes classifiers provide a very competitive result even this classifier having a simple structure on his experimental study. According to the author, Naïve Bayes are more efficient in classification task. Naïve Bayes classifier for anomaly-based network intrusion detection has proposed in [11]. He demonstrates that Naïve Bayes classifier more efficient in detecting network intrusion compare to neural network.

A comprehensive set of classifiers evaluated for detecting four type of attack category which are available on the KDD dataset [12]. The best classifier for each attack category has been chose and two appropriate

Manuscript received January 5, 2014
Manuscript revised January 20, 2014

classifier proposed for their selection models. Reference [13] proposed the best performed classifier for each category of attack by evaluates a comprehensive set of different classifier using the data collected from Knowledge Discovery Database (KDD).

For dataset Discretization CAIM (class-attribute interdependence maximization) algorithm is proposed by Authors [14]. A new Discretization algorithm based on difference-similitude set theory (DSST) is presented in [15]. It is different from the known algorithms because the reduction in the information system is prior to the data Discretization.

3. Proposed Architecture

Signature based learning approaches are able to detect attacks with high accuracy and to achieve high detection rates. In order to maintain the high accuracy and detection rate while at the same time to lower down the false alarm rate, we proposed a combination of two learning techniques.

For the first stage in the proposed hybrid learning approach, we grouped similar data instances based on their behaviors by utilizing a K-Means clustering as a pre-classification component. Next, using Naïve Bayes classifier we classified the resulting clusters into attack classes as a final classification task. We found that data that has been misclassified during the earlier stage may be correctly classified in the subsequent classification stage.

3.1 Database Discretization

The task of extracting knowledge from databases is quite often performed by machine learning algorithms. The majority of these algorithms can be applied only to data described by discrete numerical or nominal attributes (features). In the case of continuous attributes, there is a need for a discretization algorithm that transforms continuous attributes into discrete ones. CACC is one of these algorithms.

CACC (class-attribute interdependence maximization) is inspired by the contingency coefficient. The main contribution of CACC is that it can generate a better discretization scheme and its discretization scheme can lead to the improvement of classifier accuracy. [16]

```
Input: Dataset with i continuous attribute, M examples and S target classes;
1. Began;
2. For each continuous attribute Ai
3. Find the maximum d_n and the minimum d_0 values of Ai;
4. Form a set of all distinct values of A in ascending order;
5. Initialize all possible interval boundaries B with the minimum and maximum.
6. Calculate the midpoints of all the adjacent pairs in the set;
7. Set the initial discretization scheme as D: {[d_0, d_n]}and Globalcacc = 0;
8. Initialize k = 1;
9. For each inner boundary B which is not already in scheme D,
10. Add it into D;
11. Calculate the corresponding cacc value;
12. Pick up the scheme D' with the highest cacc value;
13. If cacc > Globalcacc or k < S then
14. Replace D with D';
15. Globalcacc = cacc;
16. k = k + 1;
17. Goto Line 10;
18. Else
19. D' = D;
20. End If
21. Output the Discretization scheme D' with k intervals for continuous attribute Ai;
22. End
```

3.2 K-Means Clustering

Our approach first deploys the K-mean clustering algorithm in order to separate time intervals with normal and anomalous traffic in the training dataset. Figure 1 show the steps involved in K-Means clustering process. [17]

The main goal to utilize K-Means clustering approach is to split and to group data into normal and attack instances. K-Means clustering methods partition the input dataset into k- clusters according to an initial value known as the seed-points into each cluster's centroids or cluster centers. The mean value of numerical data contained within each cluster is called centroids. In our case, we choose k = 3 in order to cluster the data into three clusters (C1, C2, C3). Since *U2R* and R2L attack patterns are naturally quite similar with normal instances, one extra cluster is used to group U2R and R2L attacks.

The K-Means algorithm works as follows:

- Select initial centers of the K clusters. Repeat step 2 through 3 until the cluster membership stabilizes.
- Generate a new partition by assigning each data to its closest cluster centers.
- Compute new clusters as the centroids of the clusters.

A distance function is required in order to compute the distance (i.e. similarity) between two objects. The most commonly used distance function is the Euclidean [17] one which is defined as:

$$d(X,Y) = \sqrt{\sum_{m}^{i=1}(xi - yi)}^2 \quad (1)$$

In Eqn. (1); $X = (x_1 \ldots x_m)$ and $Y = (y_1 \ldots y_m)$ are two input vectors with m quantitative features. In the Euclidean distance function, all features contribute equally to the function value.

3.3 Naïve Bayes Classifier

Some behaviors in intrusion instances are similar to normal and other intrusion instances as well. In addition, a lot of algorithms including K-Means are unable to correctly distinguish intrusion instances and normal instances. In order to improve this shortcoming in classification, we combined K-Means technique with Naïve Bayes classifier.

Naïve Bayes has become one of the most efficient learning algorithms [18]. Naïve Bayes are based on a very strong independence assumption with fairly simple construction. It analyzes the relationship between independent variable and the dependent variable to derive a conditional probability for each relationship. Using Bayes Theorem we write:

$$P(H|X) = P(X|H)\, P(H) / P(X) \quad (2)$$

In Eqn. (2) X is the data record and H is some hypothesis represent data record X, which belongs to a specified class C. For classification, we would like to determine P(H|X), which is the probability that the hypothesis H holds, given an observed data record X. P(H|X) is the posterior probability of H conditioned on X. In contrast, P(H) is the prior probability. The posterior probability P(H|X), is based on more information such as background knowledge than the prior probability P(H), which is independent of X. Similarly, P(X|H) is posterior probability of X conditioned on H.

For intrusion detection Naïve Bayes is effective, we calculate prior probabilities and on that basis we calculate the posterior probability. Naïve Bayes algorithm specifies the class by taking maximum probabilities.

4. Experimental Evaluation

4.1 Dataset Description

Many Researchers use, the KDD Cup'99 benchmark dataset [19] for evaluation and comparison between the proposed approaches and the previous approaches. The entire KDD data set contains an approximately 500,000 instances with 41 features. The training dataset contains 24 types of attack, while the testing data contains more than 14 types of additional attack.

KDD dataset covered four major categories of attacks which is Probe, DoS, R2L and U2R. In order to demonstrate the abilities to detect different kinds of intrusions, the training and testing data covered all classes of intrusion categories as listed in the following as adopted from the [18].

4.2 Feature Selection

The number of features required is another major concern in processing the dataset as well.

- The primary benefit of feature selection is that the amount of data required to process is reduced, ideally without compromising the performance of the detector.
- In some cases, feature selection may improve the performance of the detector as it simplifies the complexity problem by reducing its dimensionality.

By many research studies we select 7 feature set for our experiment. These 7 features are best minimum features selected from 41 features of KDD dataset. These features are:

TABLE I. Selected Features for IDS

Features	Value type
Service	Cont.
dst_bytes	Cont.
logged_in	Disc.
count	Cont.
dst_host_count	Cont.
root_shell	Disc.
dst_host_rerror_rate	Cont.

4.3 Evaluation Measurement

An Intrusion Detection System (IDS) requires high accuracy and detection rate as well as low false alarm rate. In general, the performance of IDS is evaluated in term of accuracy, detection rate, and false alarm rate as in the following formula:

$$\text{Accuracy} = (TP+TN) / (TP+TN+FP+FN) \quad (3)$$

$$\text{Detection Rate} = (TP) / (TP+FP) \quad (4)$$

$$\text{False alarm} = (FP) / (FP+TN) \quad (5)$$

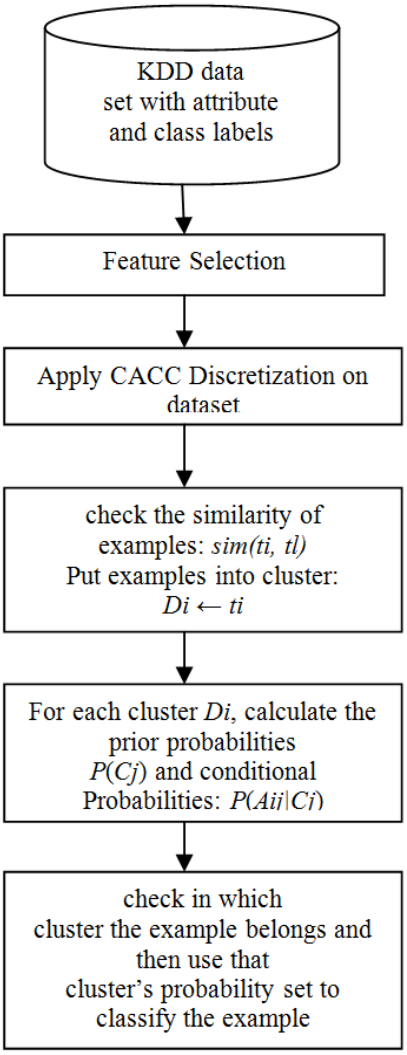

Fig. 1. NIDS process flow

Table II shows the categories of data behavior in intrusion detection for binary category classes (Normal and Attacks) in term of true negative, true positive, false positive and false negative. [16]

TABLE II.GENRAL BEHAVIOUR OF INTRUSION DETECTION DATA

	Predicted Normal	**Predicted Attack**
Normal	TN	FP
Attack	FN	TP

- True positive (TP) when attack data detected as attack
- True negative (TN) when normal data detected as normal
- False positive (FP) when normal data detected as attack
- False negative (FN) when attack data detected as normal

4.4 Results and discussion

Table III, IV and V represent the results of classification obtained from Naïve Bayes (NB) and proposed hybrid learning approach K-Means with Naïve Bayes (KM+NB) using the training and testing sets. KM+NB performed better than the single classifier NB in detecting Normal and attack instances. Since Normal, U2R, and R2L instances are similar to each other, KM+NB recorded a comparable result for R2L instances except for U2R instances.

TABLE III Confusion matrix on 7 features set with single Naïve Bayes

	Predicted Normal	**Predicted Attack**
Normal	7875	1852
Intrusion (attack)	6448	33227

TABLE IV Confusion matrix on 7 features set with KM+ Naïve Bayes

	Predicted Normal	**Predicted Attack**
Normal	8678	998
Intrusion (attack)	4089	35637

TABLE V Confusion matrix on 7 features set with discretized KM+ Naïve Bayes

	Predicted Normal	**Predicted Attack**
Normal	8998	821
Intrusion (attack)	3365	63218

Table VI shows the measurement in terms of Recall, detection rate, and false alarm using the training and testing sets of both single classifiers and hybrid learning approach. We can see that single classifier produced a slightly higher accuracy and detection rate but with high false alarm rates as well. Meanwhile, the hybrid approach recorded high accuracy and detection rate with low false alarm percentage. The clustering techniques used as a pre-classification component for grouping similar data into respective classes helped the proposed hybrid learning approach to produce better results as compared to single classifier. The hybrid approach also allows misclassified data during the first stage to be classified again, hence improving the accuracy and detection rate with acceptable false alarm. For instance, the hybrid learning approach

enhances the accuracy for single classifier especially for KM+NB combination.

TABLE VI Results in terms of various measurements

Methods	Naïve Bayes	KM + NB
Detection rate	80.840%	80.410%
False Alarm	21.03%	10.31%
Recall	83.74%	88.702%
F-Measure	89.220%	91.36%

5. Conclusion and Feature Work

In this paper, a hybrid learning approach through combination of K-Means clustering and Naïve Bayes classifier is proposed and efficiency of IDS is improved by discretized dataset. The approach was compared and evaluated using KDD Cup '99 benchmark dataset. The fundamental solution is to separate instances between the potential attacks and the normal instances during a preliminary stage into different clusters.

In the future, we recommend considering the Hybrid Intrusion Detection System which is better at detecting R2L and U2R attacks. The misuse detection approach better at detecting R2L and U2R attacks more efficiently as well as anomaly detection approach, which is better at detecting attacks at the absence of match signatures as provided in the misuse rule files.

REFERENCES

[1] W. Lee, J. S. Stolfo, and W. K. Mok, "A Data mining framework for adaptive intrusion detection," Proceedings of the 1999 lEEE Symposium on Security and Privacy, pp.120-132, 1999.

[2] Patcha and J-M Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends," Computer Network, 2007.

[3] Solahuddin, "Applying knowledge discovery in database techniques in modeling packet header anomaly intrusion detection systems," Journal of Software, 2008, 3(9): 68-76.

[4] D.Barbara, J.Couto, SJajodia, and N.Wu, "ADAM: A test bed for exploring the use of data mining in intrusion detection" , SIGMOD, vo130, no.4, pp 15-24,2001.

[5] Tomas Abraham, "IDDM: INTRUSION Detection using Data Mining Techniques", Technical report DSTO electronics and surveillance research laboratory, Salisbury, Australia, May2001.

[6] Wenke Lee and Salvatore J.Stolfo, "A Framework for constructing features and models for intrusion detection systems" ACM transactions on Information and system security (TiSSEC), vol.3, Issue 4, Nov 2000.

[7] C. F. Tsai, and C.Y Lin, "A triangle area-based nearest neighbors approach to intrusion detection," Pattern Recognition, 2010, 43(1):222-229.

[8] Y. Li and L. Guo, "An active learning based on TCM-KNN algorithm for supervised network intrusion," Computer and Securtiy, 2007, 26: 459-467.

[9] R. Luigi, T.E. Anderson, and N. McKeown, "Traffic classification using clustering algorithms. In Proceedings of the ACM SIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, Sept. 2011, Pisa, Italy, ACM Press, pp. 281-286.

[10] B.A. Nahla, B. Salem, and E. Zied, Naive bayes vs decision trees in intrusion detection systems. In Proceeding of the ACM Symposium on Applied Computing, Nicosia, Cyprus, 2004.

[11] P. Mrutyunjaya, and R. P. Manas, "Network intrusion detection Using naïve bayes," International Journal of Computer Science and Network Security, 2007, 7(12):258-263.

[12] N. H. Anh, and C. Deokjai, "Application of data mining to network intrusion detection: classifier selection model," Lecture Notes in Computer Science. 2008, 5297:399-408.

[13] G. Meera and S. K. Srivatsa, "Classification algorithms in comparing classifier categories to predict the accuracy of the network intrusion detection – a machine learning approach", Advances in Computational Sciences and Technology. 2010, (3):321–334.

[14] Kurgan, L.A, Cois K J, "CAIM Discretization Algorithm" Knowladge and Data Engg. Volume:2, Issue:6, 2004.

[15] Ming wau, Zayo chun Hang "Discretization algorithm based on Difference-similitude set theory" Machine Learning and cybernatics. 2005.

[16] Cheng-Jung Tsai, Chien-I. Lee, Wei-Pang Yang, "A discretization algorithm based on Class-Attribute Contingency Coefficient" Information Sciences 178, year-2008.

[17] Pradeep Rai,Shuba Singh. "A Survey of Clustering Techniques". International Journal of Computer Applications (0975-8887) Volume 7- No.12.October 2010.

[18] H. Zhang and J. Su., "Naive bayes for optimal ranking", Journal of Experimental and Theoretical Artificial Intelligence. 2008, 20: 79-93.

[19] KDD.(1999). Available at http://kdd.ics.uci.edu/databases/ - kddcup99/kddcup99.html.

[20] An Adaptive Hybrid Intrusion Detection System By Mahbod Tavallaee, THE UNIVERSITY OF NEW BRUNSWICK October, 2011.