Host Based Intrusion Detection System Based on Fusion of Classifier using K-means Clustering

Manisha A.Rakate, BSCOER, Pune
Prof.R.H .Kulkarni BSCOER , Pune

Abstract—**Data mining techniques are applied on the distributed data for extracting the relevant data for decision making. Distributed systems such as sensor networks consist of large amount of data. These systems generate data over period of time and it is stored in the form of classifiers. Intrusion Detection Systems aim at detecting intruder for protection. Proposed novel technique of fusing classifiers to detect the coming request in Distributed intrusion detection application is anomaly or normal, with the help of k-nearest neighbor classifier. In this classifier object is classified by a majority vote of its neigh bourse, with the object being assigned to the class most common among its k nearest neighbors. The fusion of two (or more) classifiers is done by multiplying the hyper-distributions of the parameters and determines basic equations for that task. The fundamental playing point of this combination methodology is that the hyper-distributions are retained all through the combination process.**

Index Terms—**Data mining, Distributed systems, Intrusion Detection Systems, Classification rules, Fusion, Hyper distributions.**

I. INTRODUCTION

Machine learning is the study and research of algorithm that can be applied on different data [8]. The algorithms developed works by building structure on the data and is used to make decisions [9]. Machine learning is actually used for making decisions based on the algorithms used to get the required output. Machine learning is divided into following Categories as below-

A. Supervised Learning-

The Supervised Learning makes use of training data and determines the output or class labels based on how the test data is.

B. Un-supervised-

Learning-unsupervised learning tries to search the feature from non-labeled data. It groups the data according to the similar feature.

C. Reinforcement-Learning-

Reinforcement learning neither proper input nor proper output is provided. In this paper, the main idea is probabilistic classifier and generative classifier.

1) Probabilistic Classifier-

Probabilistic classifier gives o/p on basis of given data containing classes and probability.

2) Generative Classifier-

In Generative classifier, the output is generated randomly regardless of the input to the model.

The classifier which is used in the proposed system is combination of both probabilistic and generative classifier. This probabilistic generative classifier is usually based on Bayes theory.

$$p(c|x)= \frac{p(x|c) : p(c)}{p(x)}$$

The training data which is to be classified is clustered into different groups using clustering algorithm. The clustering algorithm used is K-means which is widely used data mining technique.

There are three different possibilities to fuse the classifiers

1) Firstly, the classifiers can be used as ensemble of classifiers i.e. combination of multiple classifiers which will increase the performance.
2) Secondly, the classifiers are combined at the parametric level of the classified rules.
3) Thirdly, by combining the outputs obtained from the classifiers and by discarding duplicate rules.

The proposed system is based on the second approach in which the fusion takes place by combining the classifiers at parametric level from the classified rules [13].In simple words, K- means algorithm to classify or to group the objects depending upon features in K groups. Where, K is a positive integer. The clusters are formed by reducing sum of squares of distance between the centroid and the data points. The Intrusion detection system is being proposed to check intrusion in the network using clustering and classification.

Intrusion Detection (ID) is a key procedure in Data Security assumes an imperative part locating diverse sorts of attacks and secures the system framework. Intrusion Detection is the procedure of observing and analyzing the

ISSN: 2278 – 1323 All Rights Reserved © 2015 IJARCET

emerging in a machine or system framework to distinguish all security issues. IDS give three imperative security capacities; monitor, detect and respond to unapproved activities. IDs are partitioned into two general classifications: host-based (HIDS) and system based (NIDS) [2]. A host-based Id obliges little projects (or specialists) to be introduced on singular frameworks to be administered. The details of HIDS and NIDS are mentioned in Section II.

The paper contain following contents: Literature Survey in Section II, Proposed System in Section III, Mathematical model in Section IV, Experimental results in section V followed by the Conclusion in Section VI.

II. LITERATURE SURVEY

The combination of different classifiers to create a joint has been turned out to be more helpful to analyze for the utilization of individual classifiers [1]. In information security, Intrusion Detection plays important role. It detects different types of attacks and secured fully the network systems. Intrusion Detection (ID) is a key procedure in Data Security assumes an imperative part locating diverse sorts of attacks and secures the system framework. Intrusion Detection is the procedure of observing and analyzing the emerging in a machine or system framework to distinguish all security issues. IDS give three commanding security capacities; monitor, detect and respond to unapproved activities. IDs are partitioned into two general classifications: host-based (HIDS) and system based (NIDS) [2]. A host-based IDs obliges little projects (or specialists) to be introduced on singular frameworks to be administered.

A. Types of IDS

1) Host Based Intrusion Detection (HIDS):

Host based intrusion identification (HIDS) refers to intrusion identification that takes place on a single host framework. The information is gathered from an individual host framework. The HIDS specialist's screens exercises, for example, uprightness of framework, application activity, record changes, and host based system traffic, and framework logs.

2) Network Based Intrusion Detection (NIDS):

A system based intrusion location framework (NIDS) is utilized to monitor and investigate system activity to ensure a framework from system threats based on system where the information is movement over the system. A NIDS tries to catch malicious attacks such as denial of service (Dos) attacks, port scan and observing the system activity attacks. Intrusion Detection

Approaches There are right now several of methodologies being used to achieve the desirable components of an intrusion detection framework. There are two general methodologies to intrusion detection:

I. Anomaly Detection:

Anomaly IDS attempting to locate anomalies when any changes happen from the typical framework. Anomaly Detection is focused around analysis information assembled over a time of ordinary operation. Anomaly detection is a important tool for detection of fraud, network based intrusion, furthermore other irregular events that have incredible centrality however they are tough. The criticalness of anomaly detection is because of the way that anomalies in information mean vital noteworthy data in a huge extent of application domains [3].

II. Misuse Detection:

Misuse IDS attempting to locate Anomalous behavior by investigating the given traffic and run with a few rules focused around Analysis and correlation with the Rules the framework can identify any attacks, for example, matching sign. Misuse detection is similar here and there signature based detection wherein alarms are detected by particular signature of attack. This sort of attack signature includes specific traffic or action that is taking into account known intrusion activity [4].

B. Clustering for Intrusion Detection

Clustering is a serious issue with number of applications, and various different algorithms and techniques have risen throughout the years. The objective of clustering is to gathering information focus into similar clustering, where the homogeneity is generally measured by distance or similarity among information points. As of late, may applications oppose the need of clustering by information fusion? This is on account of that data contained in single information source is restricted by its particular perception, subsequently; consolidating various perceptions may encourage the far reaching Understanding of the issue. For example, with a specific end goal to examine memory persistent of microscopic organisms, a bacterium is seen at diverse trial conditions and evolutional times [8].
At that point the numerous perceptions are arranged by clustering algorithm. In scientometrics, a procedure has been proposed to join data mining information and bibliometrics information to examine the structure mapping of journal sets [9]. In bioinformatics, high throughput systems produce various genomic information. The test to supply clustering algorithm with the capacity to recover corresponded or correlative data about the basic practical parts of qualities and proteins has pulled in numerous investments [10]. Tragically, however the machine learning group has effectively centered on information combination for grouping [11] and novelty detection [12], the augmentation to unsupervised learning, for example, clustering, is still an uncertain furthermore progressing issue.

ISSN: 2278 – 1323 All Rights Reserved © 2015 IJARCET

C. Classification for Intrusion Detection

Classification is typical information mining strategy which is used to anticipate relationship for information examples. Classification strategies assess and characterize the information into known classes. Every information sample is checked with a known class label. Moreover this approach is utilized to take in a model utilizing the preparation train data samples. This model is utilized to arrange the data samples as peculiar conduct information or the typical conduct information [6]. By Sampath Deegalla and Henrik Bostrom ("Improving Combination of Dimentionality Reduction Methods for nearest neighbor Classification") this paper which is in the field of classification, explored in this paper "two novel strategies for intertwining peculiarities and classifiers in the conjunction with three dimensionality Reduction Strategies for Nearest Neighbor classifier in high measurements".[7]

III. PROPOSED SYSTEM

The planned system is to detect Intrusion while the data is distributed on different sites. Fig 1 shows the architecture diagram of the proposed system which consist of following steps:

1) The data is distributed at two sites viz. site1 and site2. The data given to the classifier is called as training data. The data set is stored at two sites mentioned above.
2) Before the data is classified the clusters are formed depending on the features. The algorithm used is K means which generates the k clusters depending upon the distance between the data to the centroid selected. The clustering is applied to improve the Performance of the system proposed and reduces the time to obtain the output here, rules.
3) The data in the clusters are now the input to the classifiers which would generate rules used to classify the unlabeled data. The classifier used is Probabilistic generative classifier which classifies the data in a probabilistic manner.
4) Before fusion the process is to identify which component of classifier1 belongs to classifier2. Hence, a similarity measure called "Hellinger distance" is used and the value greater than a threshold will be the components to fuse.
5) After generation of rules and finding the similarity measure the fusion process takes place based on parametric value and also type of the data. If data is categorical Dirichlet distribution is preferred and if data is continuous the normal-Wishart distribution. Here the rules from different sites are fused.
6) A test data i.e. the data with no label is given to the rules generated and the anomaly in the network is detected.

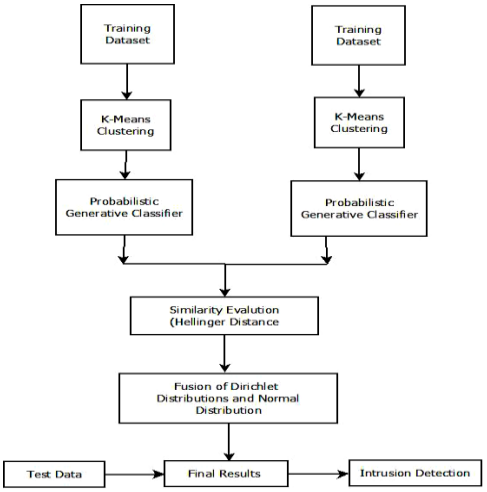


Fig.1: Architecture diagram of proposed system

Similarity measure:

Before fusing two classifiers it is essential to identify whether the component of the classifier1 belongs to the classifier2. To do so, similarity between two objects is to be computed to identify whether the component of the classifier1 belongs to the classifier2. Here "Hellinger distance" is used and is given by,

$$\frac{H(p(\text{component1}),(\text{component2}))}{\sqrt{(1 - BC(p(\text{component1});(\text{component2})))}}$$

BC is Bhattacharya coefficient and is given by,

1) For, continuous distribution

$$BC(p(\text{component1}),(\text{component2})) = \int (\sqrt{(\text{component1})(\text{component2})})$$

2) For, categorical distribution

$$BCp(p(\text{component1}),(\text{component2})) = \int \sqrt{N(x^{\text{cont}} |\mu j,\sum j)N(x^{\text{cont}} |\mu j,\sum j)}dx^{\text{cont}}$$

The fusion process is using Dirichlet distribution on categorical data and normal- Wishart distribution on continuous data. Algorithm for fusion and combination of classifier is given below:

Input: Two sets of data D1 and D2.

Output: Fused classifiers.

Algorithm1- for fusion and combination of classifiers given below:

1: Fused classifiers = fg
2: For each d1 in D1.
3: Search= false.
4: For each d2 in D2 do
5:Dist =Δ(d1, d2) ////Use Hellinger equation.
6: If dist< and class (d1) == class (d2) then
7: Fused classifiers. Add (fusion (d1, d2).
8: D2. Remove (d2).

ISSN: 2278 – 1323 All Rights Reserved © 2015 IJARCET

9: Search = True.
10: Break
11: End if.
12: End for.
13: If not found then
14: Fused classifiers .Add (d1).
15: End if
16: End for.
17: For each d2 in D2 do
18: Fused classifiers .Add (d2).
19: End for
20: Classifiers=
21: For each component in Fused classifiers do
22: Classifiers. Add (component. Point estimate)
23: End for.
24: Return Classifiers.

Fusion using Dirichlet Distribution and normal- Wishart Distribution:

1) Fusion using Dirichlet distribution:This distribution is used for categorical inputs. For the fusion of two classifiers, the densities of two classifiers the densities of both are multiplied and then divide by the prior. The Dirichlet fusion is given by,

$$\frac{Dir_1(\partial \mid \alpha_1)Dir_2(\partial \mid \alpha_2)}{Dir_0(\partial \mid \alpha_0)}$$

$$\alpha \frac{\Pi_{(k=1)}^{K}(\theta_k)^{(\alpha 1,k-1)}\Pi_{(k=1)}^{K}(\theta_k)^{(\alpha 2,k-1)}}{\Pi_{(k=1)}^{K}(\theta_k)^{(\alpha 0,k-1)}}$$

$$= \Pi_{(k=1)}^{K}(k)^{((\alpha_(1,k)+\alpha_(2,k)-\alpha_(0,k))}-1)$$

Where, $Dir_1(\partial \mid \alpha_1)$ and $Dir_2(\partial|\alpha_2)$ are posteriors and $Dir_0(\partial|\alpha_0)$ be the prior. By using the result above the parameter ' can be derived as,

$$\alpha' = \alpha_1 + \alpha_2 - \alpha_0$$

2)Fusion using normal-Wishart distribution: This distribution is used for categorical inputs. For the fusion of two classifiers, the densities of two classifiers the densities of both are multiplied and then divide by the prior. The parameters for fused normal- Wishart distribution.

NW(x |, ,W,)is given by,

β′=β1+β2-β0
μ′=μ1+μ2-μ0
W′=W1+W2-W0

K means clustering:

The K means clustering algorithm is as follows Input:
K the number of clusters
D: a data set containing n objects
Output: A set of k clusters

Algorithm 2 - K means clustering:

1: Arbitrary choose k objects from D as in initial cluster centres.
2: Repeat
3: find similarity distance from centroids to documents.
4: Reassign each object to the most similar cluster based on the mean value of the objects in the cluster
5: Update the cluster means
6: Do 3 ,4, 5 Until no change.

In simple words, K- means algorithm to classify or to group the objects depending upon features in K groups. Where, K is a positive integer. The clusters are formed by reducing sum of squares of dist between the centroid and the data points.

IV. MATHEMATICAL MODEL

The mathematical model of the proposed system is represented using Deterministic finite automata as shown in Fig 4.1
. The DFA has five tuples:

{Q, ∑,δ,q0,F}

Q: Number of States{S1; S2; S3; S4; S5; S6; S7; S8} where S1-Data repository at site1, S2- K-means Clustering at site1, S3- PGC at site1,S4- Fusion Process, S5-Data repository at site2, S6-K-means Clustering at site2. S7-PGC at site1, S8- Test data repository
∑ input. Here, Training data set and Test Data. q0: initial state. Here, S1.
F: final state. Here, S4.

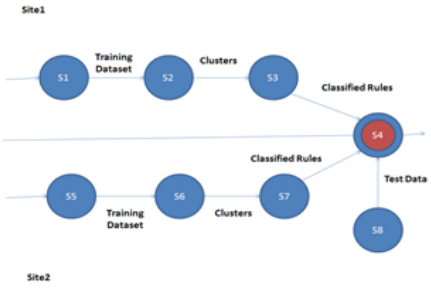
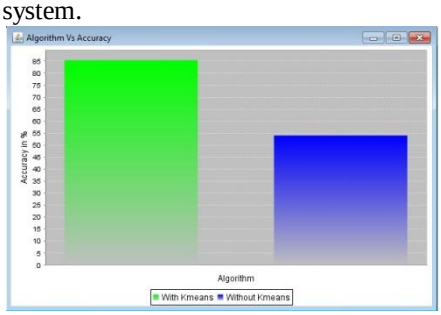


Fig.2: Deterministic Finite Automata for proposed System.

V. EXPERIMENTS AND RESULTS

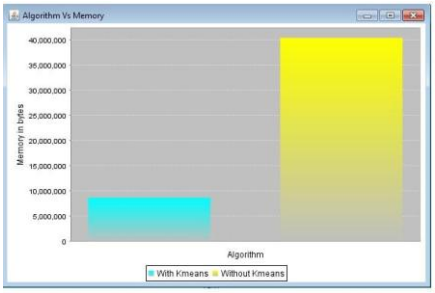
In the proposed system, classifier which is used is cluster based which increases the performance of the system. The results obtained are represented in form of the graphs first graph is in the form of accuracy of the proposed algorithm and the second graph is for memory required is for the proposed system.



Graph 1: Accuracy graph of the proposed system

The above graph shows the accuracy of the proposed system compare with the existing system. Form the graph it is concludes that the accuracy of the proposed system is more than the existing system. In the x-axis it shows the proposed and existing algorithm, and in y-axis it shows the accuracy in percentage.

ISSN: 2278 – 1323 All Rights Reserved © 2015 IJARCET



Graph 1: Accuracy graph of the proposed system

The above graph shows the memory consumed by the proposed system compare with the existing system. Form the graph it is conclude that the memory required for the proposed system is less than the existing system. In the x-axis it shows the proposed and existing algorithm, and in y-axis it shows the required memory.

VI. CONCLUSION

The proposed system is successfully implemented which deals with the fusion of two classifiers at the parametric level of the classified rules. The results showed in form of graphs shows that the cluster based classifier shows better performance than classifying the rules using only the classifier alone. The system is successfully implemented the fusion of the classifier using parameter for intrusion detection. In future, a different similarity measure must be designed to identify the belongingness of the component of the classifiers to increase the quality of the induced rules.

REFERENCES

[1] Dominik Fisch , Edgar Kalkowski " Knowledge fusion for probabilistic generative classifier with data mining applications" , IEEE transaction on knowledge and data enginnering, Vol 26, no 3, March 2014
[2] N. C. Oza and K. Tumer, Classifier ensembles: Select real-world applications, Inf. Fusion, 9 (2008), pp. 4, 2012
[3] Defeng Wang, Yeung, D.S., and Tsang, E.C.,"Weighted Mahalanobis Distance Kernels for Support Vector Machines", IEEE Transactions on Neural Networks, Vol. 18, No. 5, Pp. 1453-1462, 2007.
[4] Glenn M. Fung and O. L. Mangasarian, "Multicategory Proximal Support Vector Machine Classifiers", Springer Science and Business Media, Machine Learning, 59, 77-97, 2005.
[5] Guang-Bin Huang, Dian Hui Wang and Yuan Lan, "Extreme learning machines: a survey", Published: 25 May 2011 Springer-Verlag, 2011.
[6] Hyeran Byun and Seong-Whan Lee, "Applications of Support Vector Machines for Pattern Recognition: A Survey", Springer-Verlag Berlin Heidelberg, 2002.
[7] Manish Joshi, "Classification, Clustering and Intrusion Detection System", International Journal of Engineering Research and Applications (IJERA), ISSN: 2248-9622, pp.961-964, Vol. 2, Issue 2, Mar-Apr 2012.
[8] Deegalla, S,"Improving Fusion of Dimensionality Reduction Methods for Nearest Neighbor Classification", Machine Learning and Applications, 2009. ICMLA'09. International Conference on.
[9] Ron Kovahi; Foster Provost (1998). "Glossary of terms". Machine Learning 30: 271274.
[10] C. M. Bishop (2006). Pattern Recognition and Machine Learning. Springer.ISBN 0-387-31073-
[11] Wernick, Yang, Brankov, Yourganov and Strother, Machine Learning in Medical Imaging, IEEE Signal Processing Magazine, vol. 27, no. 4, July 2010, pp. 25-38
[12] Mannila, Heikki (1996). "Data mining: machine learning, statistics, and databases". Int'l Conf. Scientific and Statistical Database Management. IEEE Computer Society.
[13] Friedman, Jerome H. (1998). "Data Mining and Statistics: What's the connection?".Computing Science and Statistics 29 (1): 39.
[14] K-Means Clustering Tutorial By Kardi Teknomo,PhD Preferable reference for this tutorial is Teknomo, Kardi. K-Means Clustering Tutorials. http: people.revoledu.com tutorial Last Update: July 2007
[15] S. Chen, B. Mulgrew, and P. M. Grant, "A clustering technique for digital communications channel equalization using radial basis function networks," *IEEE Trans. on Neural Networks*, vol. 4, pp. 570-578, July 1993.
[16] J. U. Duncombe, "Infrared navigation—Part I: An assessment of feasibility," *IEEE Trans. Electron Devices*, vol. ED-11, pp. 34-39, Jan. 1959.
[17] C. Y. Lin, M. Wu, J. A. Bloom, I. J. Cox, and M. Miller, "Rotation, scale, and translation resilient public watermarking for images," *IEEE Trans. Image Process.*, vol. 10, no. 5, pp. 767-782, May 2001.

ISSN: 2278 – 1323 All Rights Reserved © 2015 IJARCET