Reduction of False Alarm Rate by using K-NN and Naive Bayes: A Review

Navita Datta
M.Tech (CSE)
DAVIET
Jalandhar

Rajeev Kumar
PhD
Asst.Prof, DAVIET
Jalandhar

Reeta Bhardwaj
M.Tech (IT)
Asst.Prof, DAVIET
Jalandhar

ABSTRACT
Interruption location is basic in orchestrate security. Most present framework interruption location structures (NIDSs) employ either misuse recognition or anomaly discovery. In any case, misuse recognition can't recognize darken interruptions, and anomaly location generally has high false positive rate. To overcome the imperatives of the two techniques, they intertwine both anomaly and misuse recognition into the NIDS. This paper presents a hybrid interruption recognition framework based on the combination of k-Means and two classifiers which are K-nearest neighbor and Naive Bayes. This paper includes picking features using an entropy based segment assurance computation that uses imperative properties and expels the irredundant qualities. The whole observation in this study is performed on KDD-99 Data set which is accepted at world level for surveying execution of various interruption recognition frameworks. The consequent stage is grouping stage using k-Means. The proposed framework can recognize all interruptions and categorize them into four segments: Denial of Service, User to Root, Remote to nearby and test. The main goal is to minimize the false ready rate of IDS.

General Terms
Intrusion detection system, NSL-KDD Dataset, Misuse detection, Anomaly Detection, Clustering, Classification, k-Means, Naïve Bayes, detection rate, false alarm rate, intrusion detection, KDD Cup 99Data set.

Keywords
KDD, NIDS, DoS, R2L, U2R, DR, FPR.

1. INTRODUCTION
Basically, compose based organizations and framework based attacks have grown continuously [1] [2]. The attacks based on framework can be assumed as an interruption which can be described as "any game plan of exercises which deals with the reliability, mystery or presence of a benefit". To control an interruption, interruption discovery frameworks are used. The three basic characteristics of interruption identification frameworks are precision, extensibility and adaptability. The strikes all things considered change their sorts; so the need to revive location rules to see the new attacks. A couple of strategies for instance, data mining, estimations, and innate figuring have been used for interruption location. Most starting late, the various data mining methods/techniques have been used to mine average plan from an audit data. Two data mining systems are used for anomaly discovery like association principles and repeat scenes. The association rules are used to find the connections between features and repeat scenes strategy is suitably used for perceiving occasions of back to back cases in a progression of events. Interruptions can be categorized into 2 parts: misuse and anomaly based. Certain plans of imprints are used in misuse that are taken from database and framework attempts to facilitate the moving toward attack with the ambush outlines set away in database and for any organization, the ambush is perceived. In anomalies, all movement that essentially gets sidetracked from regular lead is viewed as interruption which examines the malignant activities by standing out framework development from the commonplace utilize configuration picked up from the arrangement data. This methodology can perceive novel and covered interruptions, yet encounters a high rate of false alerts.

The rule inspiration driving interruption recognition is to distinguish the upcoming strikes which have incited incremental learning systems. An interruption recognition demonstrate can't change in accordance with the framework direct outline. So remembering the true objective to recognizing new ambushes and interminably change with the new framework lead, they display a hybrid interruption discovery framework that is made out of incremental misuse and anomaly recognition framework. This framework joins advantages of misuse moreover, anomaly recognition. The end isn't simply to get full recognition rate (DR) on poisonous activities yet also to diminish the False Positive Rate (FPR) on normal PC utilizes from an orchestrate action. Whatever is left of the paper is dealt with as takes after. Section 2 deals with the related work and portion 3 gives speculative establishment. The region 4 displayed the proposed work. The trial work is talked in zone 5 ultimately in portion 6 the paper conclusion is defined.

2. RELATED WORK
Hybrid interruption identification frameworks include misuse recognition and anomaly discovery frameworks that can perceive both known and cloud interruptions. A part of the interruption identification frameworks are said in turn off. Audit data analysis and mining (ADAM) [3] exploits the association guidelines for recognizing interruptions [1]; Next generate intrusion expert system(NIDES)[4] involves run based misuse recognition and anomaly discovery; Random Forest estimation [4] considered for interruption identification framework which considers social affair of portrayal tree for misuse location and use regions to get anomaly interruptions, for instance, ADAM [3]; Feedback learning intrusion prevention system (FLIPS) [5] uses hybrid approach for interruption balancing activity frameworks. The focal point of this study is an anomaly based classifier.

3. THEORETIC BACKGROUND
In this section, general methods and architecture are discussed which used to recognize interruption and their 2 basic classes of interruption based on misuse and anomaly are determined over here. The assorted blends of these frameworks can be named as hybrid frameworks are analyzed underneath.

3.1 Hybrid System Architecture

There are three different styles to deal with unite misuse and anomaly recognition. A few uses of anomaly at first glance to recognize threatening activities and subsequently the usage of stamp or misuse location to distinguish ambushes from vindictive activities. Affiliations which cope up the case of ambushes are set apart as strikes, those organizing to false alarms are set apart as ought not out of the ordinary and others are named as dark attacks. Hence, anomaly based part has been chosen to reduce the false positive rate.

A couple of employments use misuse and anomaly parallel. Both fragments make noxious actions autonomously. After that a few relationship portion is utilized to join the yield of both. The last third characterization uses misuse and a short time later anomaly based part to recognize assaults persistently.

3.2 System Profiling

As an amount of attacks are growing, IDS must be revived having signs to show new strikes. Framework profiling describes fresh stamps. Mastermind profiling have a lot of issues e.g. gathering strikes beginning from a framework in perspective of their sorts. These sorts of issues can be handled by frameworks, for instance, gathering and clustering.

4. SYSTEM ARCHITECTURE

The proposed framework uses K-implies gathering and KNN estimation [7]. Firstly, k-means is applying to figure out the offered dataset into regular gathering and weird bundles. Then the decision is taken on fixing the amount of bundles like five to k-means and then conversion of dataset records into normal one and atypical groups. The curious gatherings are U2R, R2L, PROBE, and DoS. The whole information is set apart with the pack documents.

Now the categorization is done of the enlightening record in two segments. One segment is meant for planning and the second is used for appraisal. In getting ready stage, use the named records to the KNN for planning reason. The K-NN classifier is set up with the checked records. At last, KNN is applied on unlabeled records. Then KNN categorizes the unlabelled record into common and strange gatherings. This work contains feature assurance, clustering and hybrid course of action. By then the exhibited figuring is inspected.

4.1 Module1: Feature Selection Algorithm

The entropy based segment decision procedure is used for picking up qualities and emptying the overabundance ones.

Estimation [8] involves 2 areas. The beginning section deals with the removal of insignificant features having poor desire ability to reach at the goal. It also registers common knowledge between the features and class. The figuring positions the features in diving solicitation of their degrees of relationship to the goal class. When this part is over, then knowledge measure counterparts to zero are emptied. The second section deals with the removal of dull features.

4.2 Module 2: Clustering

Packing is a categorization of information into social events of near sort of articles. Every social occasion or gathering contains objects that are practically identical among themselves but different to others.

The more vital complexity between social occasions, the better one is the grouping. Basically, clustering is an unsupervised learning in light of way that the class marks are unknown. A social event of estimations and recognitions are enhanced circumstances and the nearness of data in a collection. A few bundling computations are: k-Means [6], Agglomerative Hierarchical gathering &request and DBSCAN [7]. k-implies has been utilized in the proposed study.

4.3 Module3: Hybrid Classification

This section doles dominate imprints to articles. It arranges first with records near to class names in planning stage. The enlightening accumulations are secluded into look for range and new cases. It gathers a portrayal exhibit from chase space and picks class territory for every inquiry using one of the procedures - KNN[9], Naïve Bayes [6][9], Decision tree [6], and Support Vector Machine[5].

In this study, KDD99 compartment instructive gathering [10] [11] [12] is used for getting ready and testing [1][2]. DARPA interruption identification appraisal program was utilized to collect unrefined TCP/IP dump data [10],[12] for local area network in MIT Lincoln lab to take a gander at the execution of various interruption location procedures [1][2]. KDD-99 dataset, all records involve a course of action of features, out of them, some are discrete or few are persevering. The subjective regards are names without a demand which could be significant or numeric regards e.g. the estimation of feature tradition sort is one among the pictures {icmp, tcp, udp}. The numeric estimation of the component marked in is 0 or 1 to address whether the customer has viably marked in or not. For the quantitative qualities, the data are portrayed by numeric regards inside a constrained between time. Case can be the length. Since the part decision is material just to the discrete attributes, not to the relentless ones, the predictable features require being changed over to discrete one going before the component decision examination. Remembering the ultimate objective to evaluate the execution of this system they have used KDD99 enlightening file. In the first place, entropy based segment decision count is applied, and after that K-implies gathering figuring on the features picked. Starting their ahead, they classify the gotten data into Normal or Anomalous with the help of hybrid classifier.

In the study, 10 overlay cross endorsement appraisals on the instructive list are associated which are gathering precision, for instance, detection rate (DR) false positive rate (FPR), general request rate (CR) for evaluating the execution of the interruption discovery errand.

The significance of true positive, true negative, false positive, false negative are portrayed as follows.

True positive (TP): number of threatening records that are precisely named interruption.

True negative (TN): number of true blue records that are not named interruption.

False positive (FP): number of records that are incorrectly classified as assaults.

False negative (FN): number of records that are incorrectly appointed true blue activities.

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

$$FPR = \frac{FP}{TN + FN}$$

$$CR = \frac{TP + TN}{TP + TN + FP + FN}$$

Table1: Result for K-Means+ K-NN+ Naive bayes

Actual	Actual Normal	Predicted probe	Predicted DoS	Predicted U2R	Predicted R2L	Accuracy %
Normal	18954	106	11	216	19	96.03
Probe	3	1198	9	2	5	98.43
DoS	784	2678	89644	609	498	95.15
U2R	0	2	3	92	3	92
R2L	7	12	5	8	1342	97.67

The above table1 depicts the various types of attacks present and their corresponding prediction for the combination of the three methods as K-Means, K-NN and Naïve bayes.

Table 2: Method 1: K-Means clustering, Method 2: K-Means clustering and K-NN Method 3: K-Means clustering, K-NN and Naive Bayes Classifier.

Method Used	DR	FPR	Accuracy
1	0.935	0.018	0.97
2	0.958	0.013	0.98
3	0.981	0.008	0.99

The above table 2 portrays the detection rate, false positive rate and accuracy for the three unique techniques. In Method I, the detection rate is 99.35%, which have stretched out from 95.87%, the false ready rate reduces from 1.857% to 1.394%, and accuracy increments to 98.20%. In any case, in strategy 3, which is a mix of k-Means, k-NN and Naïve bayes classifier, the acknowledgment rate finishes 98.18% and the false positive rate has decreased from 1.394% to 0.830%. This demonstrates the displayed approach is superior to anything the traditional K-Means moreover, K-Means, K-NN.

5. CONCLUSION

This paper is based on a hybrid interruption location framework which merges advantages of anomaly and misuse identification. Anomaly location have high false alert rate. With a particular ultimate objective to reduce it they have associated the K-Means figuring for grouping took after by a hybrid classifier, uniting KNN and straightforward Bayes Classifier for recognizing interruptions. The burden of current method is the instructive list, everything considered, has for all intents and purposes nothing contrast among run of the mill and peculiar data. The differentiations are generally so little that course of action estimations misclassify them and a couple of records are misclassified. Hence, cushioned based estimation is one of data mining counts. The structure of ADAM has two phases: getting ready stage and on-line arrange. In the planning stage, the strike free getting ready data is supported to a module whose yield is an oversee based profile of common activities. Starting their forward, the conveyed profile is inputted to another module to play out a dynamic on-line estimation for association rules. The readiness data containing ambushes is supported into the module, and after that the module yields suspicious hot things. The suspicious hot things are set apart as false alarms or strikes. The stamped data is fed into classifier maker to set up the classifier. In the on-line organize, the test data is supported into the system. With the made profile, the anomaly recognition module can find suspicious hot things. These suspicious things are named false alerts, strikes and cloud ambushes by the readied classifier. The dark attacks are the suspicious things that can't be named false alerts or ambushes. The future scope of the above mentioned technique can be implemented using K-means Clustering and J48 classification in order to increase the accuracy and reduction of various anomalies.

6. REFERENCES

[1] James P. Anderson, "Computer security threat monitoring and surveillance," Technical Report 98-17, James P. Anderson Co., Fort Washington, Pennsylvania, USA, April 1980.

[2] D. E. Denning, "An intrusion detection model," IEEE Transaction on Software Engineering, SE-13(2), 1987, pp. 222-232.

[3] Daniel Barbara, Julia Couto, Sushil Jajodia, Leonard Popyack and Ningning Wu, "ADAM: Detecting intrusion by data mining," IEEE Workshop on Information Assurance and Security, West Point, New York, June 5-6, pp. 11-16, 2001.

[4] Debra Anderson, Thane Frivold, and Alfonso Valdes, "NIDES Next-generation Intrusion Detection Expert System (NIDES)", A Summary, Computer Science Laboratory,SRI-CSL-95-07,May 1995

[5] Te-Shun Chou and Tsung-Nan Chou, "Hybrid Classified Systems for Intrusion Detection," Seventh Annual Communications Networks and Services Research Conference, pp. 286-291, 2009.

[6] N.B. Amor, S. Benferhat, and Z. Elouedi, "Naïve Bayes vs.decision trees in intrusion detection systems," Proc. of 2004 ACM Symposium on Applied Computing, 2004, pp. 420-424.

[7] Yihua Liao and V. Rao Vimuri, "Using K-nearest Neighbour Classifier for Intrusion Detection," Department Of Computer Scinece, University Of California.

[8] T. S. Chou, K. K. Yen, and J. Luo, Network Intrusion Detection Design Using Feature Selection of Soft Computing Paradigms," World Academic of Science, Engineering and Technology, 47, pp. 529-541, 2008.

[9] Z. Muda, W. Yassin, M.N. Sulaiman and N.I. Udzir, "A K-Means and Naive Bayes Learning Approach for Better Intrusion Detection," Information Technology Journal, 10, pp. 648-655, 2011.

[10] MIT linconin labs, 1999 ACM Conference on Knowledge Discovery and Data Mining (KDD) http://www.acm.org/sigs/sigkdd/kddcup/index.php?section=1999

[11] The KDD Archive. KDD99 cup dataset, 1999, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html .

[12] M. Tavlle, E. Bagheri, W. Lu, and A. A. Gorbani, "A detailed analysis of the KDD CUP 99 Data Set," Proc. of IEEE Symposium 1st Int'l Conf. on Recent Advances in Information Technology | RAIT-2012 |Computational Intelligence for Security and Defense Applications (CISDA'09), pp. 1-6, 2009.

[13] Mukkamala S., Janoski G., and Sung A.H., "Intrusion detection using neural networks and support vector

machines," In Proc. of the IEEE International Joint Conference on Neural Networks, 2002, pp.1702-1707.

[14] J. Zhang and M. Zulkernine, "A Hybrid Network Intrusion Detection Technique Using Random Forests," Proc. of IEEE First International Conference on Availability, Reliability and Security (ARES'06), p. 8, 2006.

[15] D. Md. Farid, N. Harbi, S. Ahmmed, Md. Z. Rahman, and C. M. Rahman, "Mining Network Data for Intrusion Detection through Naïve Bayesian with Clustering", World Academy of science, Engineering and Technology, 66, pp. 341-345, 2010.