An Hybrid Approach for Classification of KDD Data

N.Raghavendra Sai
Research Scholar, Department of Computer Science,
Bharathiar University, Coimbatore, Tamil Nadu, India.
nallagatlaraghavendra@gmail.com
Dr.K.Satya Rajesh
HOD, Department of Computer Science & Engineering,
SRR & CVR Degree College, Vijayawada, AP, India.
ksatyarajesh@gmail.com

Abstract- The one-class order has been effectively connected in numerous correspondence, flag preparing, and machine learning undertakings. This issue, as characterized by the one class SVM approach, comprises in distinguishing a circle encasing all (or the most) of the information. The established technique to take care of the issue considers a synchronous estimation of both the inside and the span of the circle. In this paper, we think about the effect of isolating the estimation issue. For reasons unknown basic one-class grouping techniques can be effectively inferred, by thinking about a slightest squares plan. The proposed system enables us to infer some hypothetical outcomes, for example, an upper bound on the likelihood of false recognition. The significance of this work is shown on surely understood datasets.

Keywords- KDD, IDS, SVM, Data.

1. INTRODUCTION

The one-class classification machines has become a very active research domain in machine learning [1, 2] providing a detection rule based on recent advances in learning theory. In one-class classification, the problem consists in covering a single target class of samples, represented by a training set, and separate it from any novel sample not belonging to the same class, i.e., an outlier sample. It has been successfully applied in many novelty detection and classification tasks, including communication network performance [3], wireless sensor networks [4], forensic science [5], detection of handwritten digits [6] and objet recognition [7], only to name a few. Moreover, it has been extended naturally to binary and multiclass classification tasks, by applying a single one-class classifier to each class and subsequently combining the decision rules [8]. Since only a single class is identified, it is essentially a data domain description or a class density estimation problem, while it provides a novelty detection rule. Different methods to solve the one-class problem have been developed, initiated from the so-called one-class support vector machines (SVM) [9, 2]. The one-class classification task consists in identifying a sphere of minimum volume that englobes all (or most of) the training data, by estimating jointly its center and its radius. These methods exploit many features from conventional SVM [10], including a nonlinear extension thanks to the concept of reproducing kernels. They also inherit the robustness to outliers in the training set, by providing a sparse solution of the center. This sparse solution explores a small fraction of the training samples, called support vectors (SVs), and lying outside or on the sphere. In one-class SVM as defined in [9, 2], the resulting convex optimization problem is often solved using a quadratic programming technique. Several efforts have been made in order to derive one-class classification machines with low computational complexity [11]. In the same sense as least-squares SVM is derived from the classical SVM method [12, 13], some attempts have been made to derive from the one-class SVM a least-squares variant, such as in [14]. However, unlike the former, the latter do not have a decision function, thus inappropriate for novelty detection. In this paper, we propose to solve the one-class problem by decoupling the estimation of the center and the radius of the sphere englobing all (or most of) the training samples. In the same spirit as the classical one-class SVM machines, we consider a sparse solution with SVs lying outside or on the sphere. It turns our that the optimal sparse solution can be defined using a least-squares optimization problem, thus leading to a low computational complexity problem. This framework allows us to derive some theoretical results. We give an upper bound on the probability of false detection, i.e., probability that a new sample is outside the sphere defined by the sparse solution.

2. RELATED WORKS

Several metrics are used to evaluate and compare the performance of Intrusion Detection Systems (IDSs). The most basic metrics are the detection and false alarm rates. The detection rate is equal to the number of intrusions detected divided by the total number of intrusions in a data set, while the false alarm rate is equal to the number of normal instances detected as intrusions divided by the number of normal instances in a data set. False alarms are also referred to as false positives [7]. The diagnosis rate (or recall), meaning the number of correctly classified intrusions divided by the total number of intrusions, is also a relevant metric and we refer to it across this paper.

In the KDD Cup 1999 the criteria utilized for assessment of the member passages is the ACTE processed utilizing the disarray framework and a given cost lattice. The disarray lattice is gotten while characterizing the occurrences in the test dataset. Every segment of the perplexity framework speaks to the occasions in an anticipated class, while each line speaks to the cases in a real class. The cost network is given in Table 1.

	normal	Probe	DOS	U2R	R2L
normal	0	1	2	2	2
probe	1	0	2	2	2
DOS	2	1	0	2	2
U2R	3	2	2	0	2
R2L	4	2	2	2	0

Table 1 Cost matrix

3. DATASET

The KDD Cup 1999 utilizations an adaptation of the information on which the 1998 DARPA Intrusion Detection Evaluation Program was performed. The preparation dataset was gained in a seven week time allotment of checking the system and was handled into very nearly 5 million occurrences. The test dataset was procured amid a two week time allotment and contains 311029 occasions. Both preparing and test datasets are marked with the name of the assault write or as being ordinary movement [1]. There are 38 diverse assault writes in preparing and test information together and these assault composes fall into four principle classifications: test, disavowal of administration (DoS), remote to neighborhood (R2L) and client to root (U2R) [2].

The dataset is extremely unbalanced; most instances are DoS traffic (79%), while the other three attack types together make less than 2% of the instances. Around 19% of the instances correspond to normal traffic. The test dataset has different distribution than the training dataset and contains several new attacks (17 new attacks out of 38 possible attacks). Figure 1 depicts the distribution of the full training dataset, 10% of the full training dataset and of the testing dataset. It can be noticed that the normal, probe and DoS connections keep their distribution across the three datasets while the same is not valid for U2R and R2L connections. For U2R connections a slight increase in number of instances in the test dataset versus the training dataset can be noticed. U2R instances represent 0.01% of the 10% training dataset and 0.2% of the test dataset. On th other hand, the proportion of the R2L connections dramatically increases in the test dataset (5.2%) comparing to the training one (0.2%). Furthermore, the R2L connections are spread in space posing real challenge for determining an accurate model for classification.

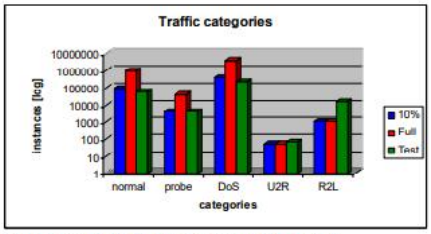

Figure 1 Traffic distribution in KDD Cup 1999 dataset

4. SVM

The machine learning method used in this paper is the support vector machines (SVMs) [11]. SVMs are a set of related supervised learning methods used for classification and regression. The examinations in this paper utilize direct SVM as actualized in Text Garden [10].

One-to-all, one-to-one and one-to-all-3categ IDSs

The one-to-all IDS utilize the 10% preparing dataset and pre-processes it as portrayed in Section 4.1. Subsequent to pre-processing, five preparing documents are made. In every one of the records, one assault write speaks to the positive class and the various assaults speak to the negative class. The SVM is trained on these five files and for each input file, it builds an output model that distinguishes between the positive class and all the other classes in the input, this is why the name one-to-all. Each connection in the test data is then fed to the models, each model decides if the connection belongs or not to a class with a certain degree of confidence. The connection is classified as belonging to the class that classified it with highest confidence. Figure 2 presents the workflow of the one-to-all IDS. The outcome of the voting is summarized into a confusion matrix and finally the average cost per text example is computed.

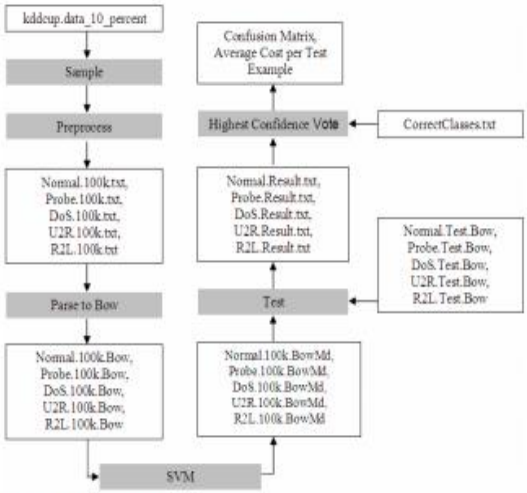

Figure 2 One-to-all IDS

The one-to-one IDS works similarly as the one-to-all IDS with two exceptions: the training files and the voting method. Each training files contains only two types of attacks: one represents the positive class and the other represents the negative class. This way 10 training files are prepared and 10 models are built. When a new connection has to be classified, each model decides for one of the two classes the connection belongs to. The connection is classified as belonging to the class to which the majority of the models assigned it to. Figure 3 presents the workflow of the one-to-one IDS.

The third IDS tries to adapt to the nature of the training data. Given the unbalanced nature of the data, it attempts to build a better model for classifying minority classes. In order to achieve this, two sets of one-to-all training files are used. The first set is formed of two files in which the positive class is represented by normal and DoS connections respectively, and the negative class is represented by all other types of connections (one-to-all test files). The second set of training files contains only three types of connections: probe, R2L and U2R filtered from the full dataset, resulting in three one-to-all files (one-to-all 3categ files since the “all” stands for the other two minority categories). SVM is trained on all five files and a two level voting is applied to the new instances. The one-to-all IDS uses the 10% getting ready dataset and pre-processes it as depicted in Section 4.1.

Resulting to pre-processing, five getting ready reports are made. In each one of the records, one attack compose addresses the positive class and the different ambushes address the negative class.

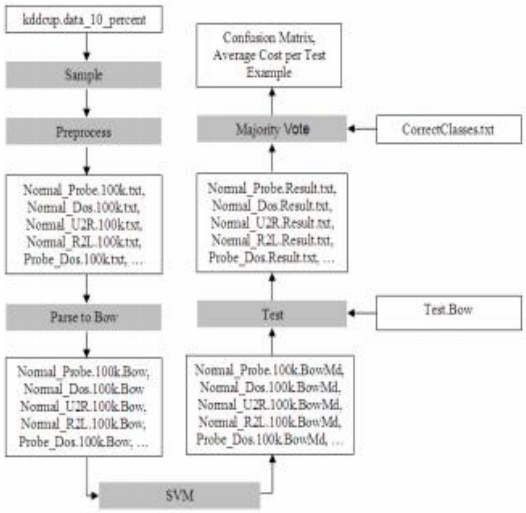

Figure 3 One-to-one IDS

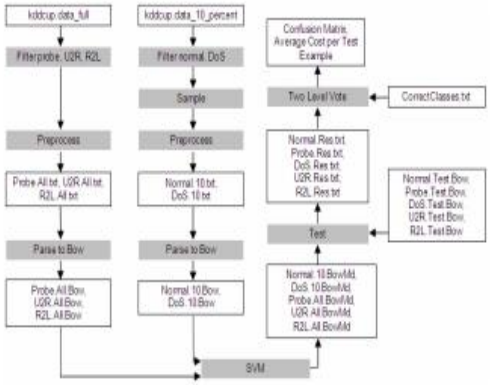

Figure 4 One-to-all-3categ IDS

5. Results

When dealing with such large and unbalanced datasets as the one provided for the KDD Cup 1999, an important step is to understand the data and find a suitable model for it. Our approach was to build models on a 100.000 instance dataset obtained as explained in Section 4.1 and classify the test dataset using the three IDSs described in Section 4.3. Table 2 presents the results obtained at this stage. The one-to-one IDS has the poorest ACTE, the one-to-all-3categ IDS has the best ACTE while the results for one-toall IDS are somewhere in between. The one-to-all IDS has a high detection rate, a good diagnosis rate but a very high false alarm rate meaning that it classifies most of the normal traffic as intrusion. This framework doesn't recognize test, R2L and U2R interruptions by any stretch of the imagination. All the movement is delegated DoS or typical, however it appears that it mistakes DoS for ordinary frequently. This might be due to the SVM cost parameters that are not optimized for this dataset or to the nature of the dataset. The one-to-one scenario has lower false alarm rate, but has poor diagnosis performance, meaning that it detects most of the alarms, but it doesn't classify them correctly. The high ACTE appears to originate from misclassifying DoS assaults (more than 220.000 occurrences out of 311.000) for R2L assaults. At last, the one-to-all3categ IDS gives the best outcomes: great ACTE, great location and determination rates and low false alert rate. In any case, this outcome may be additionally enhanced by parameter tuning or expanding the measure of the preparation dataset.

	One-to-all	One-to-one	One-to-all-3categ
ACTE	0.5306	1.6656	0.2641
Detection rate	99.2%	95.0%	90.3%
Diagnosis rate	91.3%	3.3%	90.1%
False alarm rate	99.6%	12.8%	1.6%

Table 2 Results for 100.000 instance training set

The next step in the approach was to tune SVM parameters in order to build more accurate models. The 10% training dataset (494021 instances) with 10 fold cross validation were used to build the models and the three resulting IDSs were then tested. The results are listed in Table 3.

	One-to-all	One-to-one	One-to-all-3categ
ACTE	0.2625	0.2479	0.2653
Detection rate	90.2%	90.9%	90.3%
Diagnosis rate	90.1%	90.7%	90.1%
False alarm rate	1.6%	2.02%	1.6%

Table 3 Results for 10% training set

The one-to-all IDS improved the overall performance as well as the detection, diagnosis and false alarm rates. Both detection and diagnosis rates are quite good and false alarm rate is low, meaning the system detects and correctly determines the class of over 90% of connections and has a small false alarm rate (1.6%). The one-to-one IDS also improved: it has the smallest ACTE and good detection and diagnosis rate. The false alarm rate is slightly higher than for the one-to-all IDS. The most unexpected result comes from the one-to-all-3categ IDS: there is no improvement in the detection, diagnosis and false alarm rates. The ACTE slightly increases, due to more expensive (see the cost matrix) misclassifications.

We can go more into detail with the analysis of the performance of the three IDSs by comparing the output confusion matrices listed in Table 4, Table 5 and Table 6. Rows represent the labels of the connections and columns represent the class attributed by the IDS. The last row displays the rate of true positives (e.g. 71.0% of the connections classified as normal are normal) and the last column displays the accuracy (e.g. 98.3% of normal traffic was classified as normal).

	normal	probe	DOS	U2R	R2L	%
normal	59611	300	678	4	0	98.3
probe	1053	2922	191	0	0	70.1
DOS	7242	22	222589	0	0	96.8
U2R	54	0	0	11	5	15.7
R2L	15959	16	2	2	368	2.2
%	71.0	89.6	99.6	64.7	98.6	

Table 4 One-to-all confusion matrix (ACTE = 0.2625)

	normal	probe	DOS	U2R	R2L	%
normal	59367	211	818	12	185	97.9
probe	901	3002	148	0	115	72.0
DOS	7047	52	222754	0	0	96.9
U2R	32	0	0	32	6	45.7
R2L	14791	11	2	11	1532	9.3
%	72.2	91.6	99.5	58.1	83.3	

Table 5 One-to-one confusion matrix (ACTE = 0.2479)

	normal	probe	DOS	U2R	R2L	%
Normal	59593	313	672	5	10	98.3
probe	767	3120	181	6	92	74.8
DOS	7113	324	222406	0	10	96.7
U2R	60	0	0	5	5	7.1
R2L	16186	11	2	1	147	0.8
%	71.1	82.8	99.6	29.4	55.6	

Table 6 One-to-one-3categ confusion matrix (ACTE = 0.2653)

It can be seen in Table 4 that the one-to-all IDS performs well on normal and DoS connections, on probe it has a rather poor performance (70.1% diagnosis) and misclassifies most of U2R (15.7% diagnosis) and R2L (2.2% diagnosis) connections. Most of the misclassified probe, U2R and R2L connections are classified as normal. The models for normal and DoS traffic are fairly accurate since they had a large set of training instances to build on. The one-to-one IDS performs better than one-to-all IDS as can be seen in Table 5. This IDS performs significantly better than one-to-all IDS on classifying U2R and R2L connections: it classifies 45.7% of U2R connections and 9.3% of R2L connections. The R2L connections are spread in space so that linear SVM proves to be inefficient for building a good model for classifying these instances. We noticed a tradeoff: the more accurate the SVM model for classifying R2L connections, the poorest in classifying normal connections and the other way around. The one-to-all-3categ IDS performs worse than the other two IDSs in classifying R2L and U2R attacks, and performs slightly better on classifying probe attacks. It seems indeed that linear SVM is limited in building a good model for separating normal traffic from R2L due to the spread of these connections. Even though we introduced the one-to-all-3categ IDS in order to perform better at separating the three minority classes from the two major ones (normal and DoS), it seems like the model built using SVM is not accurate enough so that this voting system proves efficient. Most of the R2L connections do not pass the first level voting, being classified as normal. Comparing to relevant results in the literature, the IDSs studied in the paper are less accurate. The one-to-one IDS with 0.2479 ACTE would rank 8th in the KDD Cup 1999 contest. Higher accuracy can be obtained by increasing the complexity of the system. SVMs with different kernels can be used for building better models, but with this approach, classification speed would decrease [11], this is undesired in real time IDSs. Hybrid systems that combine several machine learning methods or that combine machine learning methods with the more classical ones based on signatures could be used.

6. Conclusion

In this paper we studied the performance of linear SVM in classifying normal and attack connections sniffed from a computer network. We proposed a two level voting IDS that proved to perform well on a small training set but performed relatively poor when the training dataset increased. In the context of intrusion detection in a computer network, attacks such as R2L and U2R that result in small number of traffic packets seem to pose a real challenge for detection and diagnosis. A good, simple and fast classifier that is able to detect novel attacks is hard to build. Usually simplicity and speed are traded for accuracy and machine learning methods are complemented by traditional signature based methods.

REFERENCES

[1] KDD Cup 1999 Task Description, http://kdd.ics.uci.edu/databases/kddcup99/task.html

[2] Bernhard Pfahringer, Winning the KDD99 Classification Cup: Bagged Boosting, ACM SIGKDD Explorations Newsletter, Volume 1, Issue 2, p. 65-66 January 2000.

[3] Itzhak Levin, KDD-99 Classifier Learning Contest LLSoft's Results Overview, ACM SIGKDD Explorations Newsletter, Volume 1, Issue 2, p. 67-75 January 2000.

[4] Vladimir Miheev, Alexei Vopilov, Ivan Shabalin, The MP13 Approach to the KDD'99 Classifier Learning Contest, SIGKDD Explorations Newsletter, Volume 1, Issue 2, p76-77 January 2000.

[5] Tsong Song Hwang, Tsung-Ju Lee, Yuh-Jye Lee, A Three-tier IDS via Data Mining Approach, MineNet'07, June 12, 2007, San Diego, California, USA

[6] W. Lee. A Data Mining Framework for Constructing Features and Models for Intrusion Detection Systems. PhD thesis, Columbia University, 1999.

[7] Computer Security and Intrusion Detection, http://www.acm.org/crossroads/xrds11-1/csid.html

[8] H. Gunes Kayacik, Nur Zincir-Heywood, Malcolm I. Heywood, Selecting Features for Intrusion Detection: A Feature Relevance Analysis on KDD '99 Benchmark,http://www.unb.ca/pstnet/pst2005/Shaughnessy%20Room/Oct13/GK_FeatRelevance.ppt#256,1,Selecting Features for Intrusion Detection: A Feature Relevance Analysis on KDD 99 Benchmark

[9] Results of the KDD Cup 1999 Classifier Learning Contest,http://wwwcse.ucsd.edu/users/elkan/clresults.html

[10] TextGarden–Text Mining Tools, http://kt.ijs.si/Dunja/textgarden/

[11] C. Cortes and V. Vapnik, Support-Vector Networks, Machine Learning, 20(3):273-297, September 1995.

[12] Y.-J. Lee and O. L. Mangasarian. SSVM: A smooth support vector machine. Computational Optimization and Applications, 20:5–22, 2001. Data Mining Institute, University of Wisconsin, Technical Report 99-03.