Fast and Effective Network Intrusion Detection Technique Using Hybrid Revised Algorithms

D. Shona
Assistant Professor, Dept. of Computer Science, Sri Krishna Arts & Science College, Coimbatore, India.

A.Shobana
M.Phil. Scholar, Dept. of Computer Science, Sri Krishna Arts & Science College, Coimbatore, India.

Abstract – Intrusion is an abnormal malicious behavior in the network/ host. The intrusion detection using data mining techniques are very prevalent mechanism nowadays. This thesis work implements unique and hybrid architecture for Intrusion detection using new revised algorithm combinations. The new hybrid architecture for intrusion detection is named as H-RAID. The hybrid algorithms are selected from various analysis and research. The H-RAID includes (i). Sequential pattern analysis algorithms and (ii) Multi Class semi supervised SVM (support vector machine), which is a modified SVM with semi supervised nature. The proposed work reduces the training data collection risk. The system also facilitates different type of behavior detection strategies; hence it will find the types of attacks under IDS. Unlike prior intrusion detection approaches, this does not need the entire training dataset. This hybrid framework improved the existing algorithms, so it is significantly will produce higher results than the other algorithm in training and the detection speed, and have a high enhance of the detection rates of attacking sample.

Index Terms – Data mining, Intrusion Classification System, Support Vector Machine, Pattern Mining.

1. INTRODUCTION

Intrusion detection is the process of observing and finding abnormal activities and security violations in the network [1]. It finds the patterns in a dataset whose behavior is not normal on expected. These unexpected behaviors are also termed as intrusion or anomalies [2]. The Intrusion cannot always be categorized as an attack but it can be a surprising behavior. And these behaviors are previously not identified or recognized. The Intrusion detection system provides very important and accurate information in network systems regarding its behavior. These kinds of intrusion behaviors should be detected earlier without false alarm. When data has to be analyzed and reported to predict known or unknown behavior, then data mining techniques are used to do that. In data mining, clustering, classification and machine based learning techniques were used to handle intrusions in the network. This analyzes, detects and summarizes the intrusion activities by its types. Several Hybrid approaches [3] [4] are also being created in order to attain higher level of accuracy on detecting intrusions. Several research works tried to combine numerous traditional data mining algorithms to derive better results and accuracy. This type of detection helps to predict and thwart the new type of attacks or other type of attacks in the network. In general the network log size is huge and distorted, so analyzing from those type of data consumes much time and effort. This paper attempts to provide a better understanding among the various types of data mining approaches towards intrusion detection and finally proposes a new type of algorithm with various parameter considerations.

In specific, the Intrusion detection system framework includes three divisions such as information collection, analysis and response systems. Each stage of the IDS should have a better data mining algorithm [5].

Information collection or Data collection: the source of these collected data can be separated into host, network and application, according to the position.

Analysis engine: Analysis engine is able to analyze whether or not there are symptom of any intrusion.

Response: Take actions after analysis, record analysis results, send real-time alarm, or adjust intrusion detection system, and so on.

The proposed work aims to develop an algorithm that combines the logic of both methods to produce a high-performance of semi supervised clustering with Multi Class. In this paper, we aim at detecting the presence of intrusion from a large amount of data via a MCSSVM classifier.

2. PROBLEM DEFINITION

The problem of intrusion classification has been widely studied in the data mining community. The addition of noise to the data makes the problem far more difficult from the perspective of uncertainty. The detection of an object as an intrusion is affected by various factors, many of which are of interest for practical applications. The intrusion detection problem is similar to the classification problem, where this can be handled by the data mining algorithms. A specific characteristic of the past, however, is that the great majority of the database objects being analyzed are not intrusions. Moreover, in many cases, it is not a priori known what objects are intrusions. More over many intrusion detection techniques identified the objects as intrusions which are not sequential [6]. Such methods like association mining based approaches consider the frequent or infrequent items the data set. For instance, the objects with few frequent items or many infrequent items are more likely to be considered as intrusion objects than others. It is well known that anomaly-based IDS suffer from the high rate of false alarms. Continuous efforts are being made to reduce the high false positive rate. And the intrusion detection is a data analysis process and can be studied as a problem of classifying data perfectly. Any classification scheme need well training samples to yield better accuracy and more concentration on cleaning the data, it implies that if we can extract features that distinguish normal data from abnormal one accurately then the false positive rate can be reduced to a great extent.

In common, the research on Multi Class classification for intrusion detection has been limited in the research. Most of the existing research studied the selection of a set of initial training dataset prior to performing classification. Several studies do not deal the Multi Class classification process in intrusion detection domain, which incurs more training overhead. The problem addressed in this paper is how to effectively perform self-training to produce an accurate classification result.

3. RELATED WORKS

Several classification techniques have been applied for intrusion detection, which are optimized to find clusters rather than intrusions. So that produced the following basic problems. Accuracy of intrusion detection depends on how good the classification algorithm captures the structure of objects. A set of many abnormal data objects that are similar to each other would be recognized as a same group rather than as noise/intrusions. The existing system discovers attributes or properties based on the given training which are called as training dataset. That needs to solve unsupervised problem which is yet unbalanced data learning problem. And several drawback of the existing approach is the system considers frequent items are considered as normal behavior. Several existing systems are based on one class SVM.

Existing Methods:

1. Probabilistic model :

a. This type of model uses a procedure that determines whether a particular object is an intrusion or not.
b. The probabilistic model is not considered as an effective method for intrusion detection due to its less accurate output [7].

2. Statistical modal:

a. Another interesting approach to detecting intrusions by statistical methods is implemented in the SmartSifter algorithm [8]. The basic idea of this algorithm is to construct a probabilistic data model based on observations.
b. In this statistical model, the important data from the statistical view is constructed rather than the entire dataset is stored as the training set. The objects are processed successively, and the model learns while processing each data object.
c. A data object is considered to be an intrusion if the model changes considerably after processing it. For these purposes, a special metrics, the intrusion factor, is introduced to measure changes in the probabilistic model after adding a new element.

Drawbacks of statistical approaches:

- First, they require either construction of a probabilistic data model based on empirical data, which is a rather complicated computational task, or a priori knowledge of the distribution laws.
- It is not guaranteed that the data being examined match the assumed distribution law if there is no estimate of the distribution density based on the empirical data.
- This approach only concentrated on the available dataset samples.

3. Distance based approaches:

a. The distance based approach finds the distance between two objects and calculates the intrusions.
b. Drawbacks:
i. Complexity in heterogeneous data grouping
ii. Time and computation cost was high.

4. PROPOSED SYSTEM

The existing iterative framework requires more training dataset and time. This can be computationally demanding for large data sets. To address this problem, it would be interesting to consider a batch semi-supervised classifier method that updates the existing clustering solution based on the previous assignment for the new point. An alternative way to lower the computational cost is to reduce the number of iterations by applying a batch approach that selects a set of points to query in each iteration. The proposed system aims at increasing the classification performance through the hybrid approach which is named as H-RAID. The goal of the proposed system is applying the Multi Class learning process to identify the best class of objects and classifying them accordingly. This aims at producing least false alarm rate and improving the classification performance. This reduces training data by applying the historical data as an input. So this aims at reducing the training overhead. Multi class classification aims to identify a small group of instances which deviate remarkably from the existing data.

The followings are the contributions of the proposed system.

- A new multi class semi supervised SVM approach has been proposed.
- SVM based self learning approach has been applied to select the top k points that have the highest normalized uncertainty to query their neighborhoods.
- MC-SVM algorithm has been used with multi class active learning for fast intrusion detection. This reduces the need of training dataset.
- The proposed system also updates the previous dataset and performs the Top-k result.
- The proposed work reduces the training data collection risk. The system also facilitates different type of behavior detection strategies; hence it will find the types of attacks under IDS.

The fusion of SVM and Multi class group based semi supervised learning process helps to improve the intrusion classification performance. The proposed system overcomes the delay in intrusion classification problem by applying the Sequence Databases and Sequential Pattern Analysis process. The system also performs the dataset increment, which is named as oversampling incremental process. The proposed semi supervised method, which utilizes optimal sample method of classification.

The proposed method, which is a semi supervised technique, utilizes the previous top K labels as training data for data learning. This performs MCSSVM algorithm for finding best class for fast intrusion detection. Using the above the proposed system reduces the training phase and improves the classification speed and accuracy. When compared to the existing method or other popular semi supervised clustering algorithms, the required computational costs and memory requirements are significantly reduced and thus the proposed method is especially preferable in batch grouping, streaming data, or large scale problems.

The idea is that MCSSVM maps inputs vectors nonlinearly into the high dimensional feature space and construct the optimum separating hyperplane for multi class detection. In some cases, the technique uses linear and sometime nonlinear. So this is known as optimal hyper plane selection.

Semi Supervised Multiclass SVM steps:

Step 1: collect training data samples and test samples.

Step 2: According to data collection, constructs training sample set and test sample set.

Step 3: Set up parameters, initializes the initial support vector object position, every position corresponding a set of attributes (*a1,a2..an*) in MCSSVM model, builds up SVM prediction model by parameters and samples.

Step3: From the parameters calculate every class threshold value, and then analyze the hyperplane value.

Step4: Randomly select *P* objects from initial cluster, find out the optimal object position *best X* based on the hyperplane. Set it up as individual target *obj X* .

Step5: The non-optimal objects in the initial cluster moving to target class position and make the overall search.

Step6: The optimal object make overall search according to its neighborhood.

Step7: Update every objects class

Step8: Apply the optimal parameter (*a1,a2…an*) and training sample to build up BSVM classification model.

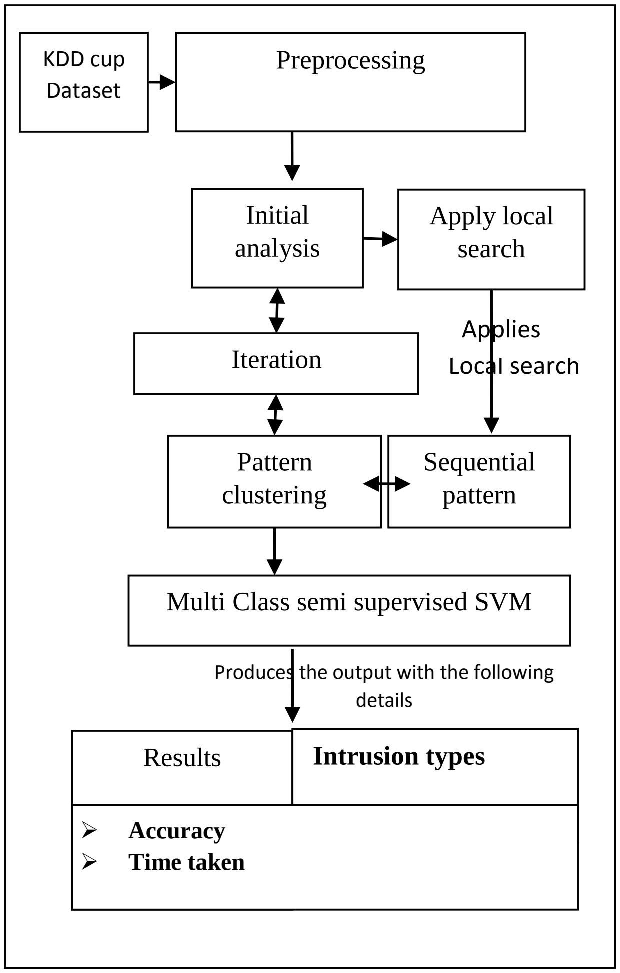


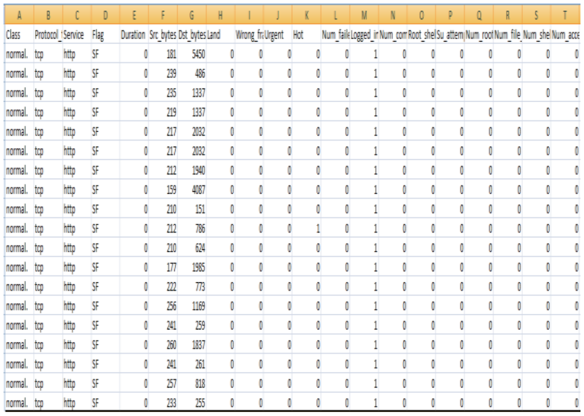
Fig 1.0 Overall Architecture

The improved MCSSVM makes separating hyperplane quickly find the right class position basis optimized hyperplane function, makes the algorithm substantially higher than the previous methods for intrusion class detection.

5. EXPERIMENT AND RESULTS

A. DATASET COLLECTION AND UPLOAD PROCESS:

The first module is the process of uploading datasets. The experiments collect dataset for MCSSVM implementation from UCI repository dataset.



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
Class	Protocol_t	Service	Flag	Duration	Src_bytes	Dst_bytes	Land	Wrong_fr	Urgent	Hot	Num_faile	Logged_ir	Num_com	Root_shel	Su_attem	Num_root	Num_file_	Num_shel	Num_acce
normal.	tcp	http	SF	0	181	5450	0	0	0	0	0	1	0	0	0	0	0	0	0
normal.	tcp	http	SF	0	239	486	0	0	0	0	0	1	0	0	0	0	0	0	0
normal.	tcp	http	SF	0	235	1337	0	0	0	0	0	1	0	0	0	0	0	0	0
normal.	tcp	http	SF	0	219	1337	0	0	0	0	0	1	0	0	0	0	0	0	0
normal.	tcp	http	SF	0	217	2032	0	0	0	0	0	1	0	0	0	0	0	0	0
normal.	tcp	http	SF	0	217	2032	0	0	0	0	0	1	0	0	0	0	0	0	0
normal.	tcp	http	SF	0	212	1940	0	0	0	0	0	1	0	0	0	0	0	0	0
normal.	tcp	http	SF	0	159	4087	0	0	0	0	0	1	0	0	0	0	0	0	0
normal.	tcp	http	SF	0	210	151	0	0	0	0	0	1	0	0	0	0	0	0	0
normal.	tcp	http	SF	0	212	786	0	0	0	1	0	1	0	0	0	0	0	0	0
normal.	tcp	http	SF	0	210	624	0	0	0	0	0	1	0	0	0	0	0	0	0
normal.	tcp	http	SF	0	177	1985	0	0	0	0	0	1	0	0	0	0	0	0	0
normal.	tcp	http	SF	0	222	773	0	0	0	0	0	1	0	0	0	0	0	0	0
normal.	tcp	http	SF	0	256	1169	0	0	0	0	0	1	0	0	0	0	0	0	0
normal.	tcp	http	SF	0	241	259	0	0	0	0	0	1	0	0	0	0	0	0	0
normal.	tcp	http	SF	0	260	1837	0	0	0	0	0	1	0	0	0	0	0	0	0
normal.	tcp	http	SF	0	241	261	0	0	0	0	0	1	0	0	0	0	0	0	0
normal.	tcp	http	SF	0	257	818	0	0	0	0	0	1	0	0	0	0	0	0	0
normal.	tcp	http	SF	0	233	255	0	0	0	0	0	1	0	0	0	0	0	0	0

Fig 2.0 KDD cup 99 training dataset

B. Preprocessing

The dataset will be preprocessed before starting the implementation. This step eliminates the duplicate and missing items in the uploaded dataset.

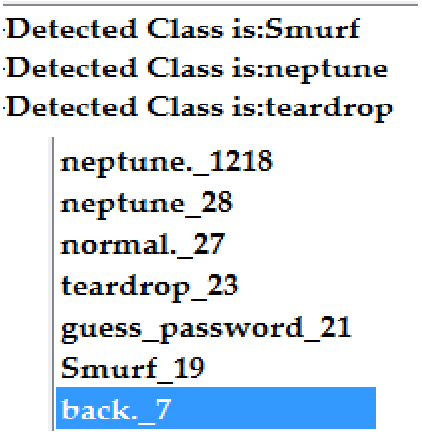
C. MCSSVM Process:

The investigating data have 65000 observing sample, there exists missing value in these sample. After eliminating the missing element, the system performs the MCSSVM for every attribute. The MCSSVM implementation process identifies the frequency of every value from the dataset. Sequential pattern analysis on Multi class semi supervised SVM has been implemented to identify the class of the given test data property and its label. This is based on the SVM based approach which performs the best segmented portion for identification process and class analysis. MCSSVM based labeling has been created in this process. The user can give the partition threshold. A set of data instances in the original data set is taken as predefined input. This data may be contaminated by noise and incorrect data labelling etc., this data might be error free, because this is going to be used as training data. So the cleaning is done using before updating the data.

D. Detecting classes (Results)

This is for detecting the cluster label from the user test data. When the user gives the input to the system, the system calculates the threshold value for every attribute value for the new input.

And then compare that new S_t value with the threshold value which is calculated in earlier. Final results will be identified individually and updated in the database using oversampling method.



```
Detected Class is:Smurf
Detected Class is:neptune
Detected Class is:teardrop
```

```
neptune._1218
neptune_28
normal._27
teardrop_23
guess_password_21
Smurf_19
back._7
```

Fig 3.0 Detected intrusion class with its count

6. RESULTS

Precision: Precision for a class is the number of true positives (i.e. the number of instances correctly labeled as belonging to the positive class) divided by the total number of elements labeled as belonging to the positive class (i.e. the sum of true positives and false positives, which are items incorrectly labeled as belonging to the class). The equation is:

$$\text{Precision} = TP / (TP + FP)$$

Recall: Recall in this context is defined as the number of true positives divided by the total number of elements that actually belong to the positive class (i.e. sum of true positives and false negatives, which are items which were not labeled as belonging to the positive class but should have been.) The Recall can be calculated as:

$$\text{Recall} = TP / (TP + FN)$$

Accuracy: The percentages of the predicted values are match with the expected value for the given data. The best system is that having the high Accuracy, High Precision and High Recall value. The performance of the proposed system is tested with the 65000 instances, from each instance the precision and recall values are gathered and that is plotted in the fig 4.0.

	Real Data Value of Project Status	
Predicted class	Success	Failure
Success	**TP**	**FP**
Failure	**FN**	**TN**

Table: 1 Confusion Matrix of Prediction Outcomes

With help of the confusion matrix values measurement of the precision and recall values are calculated and plotted as a graph below:

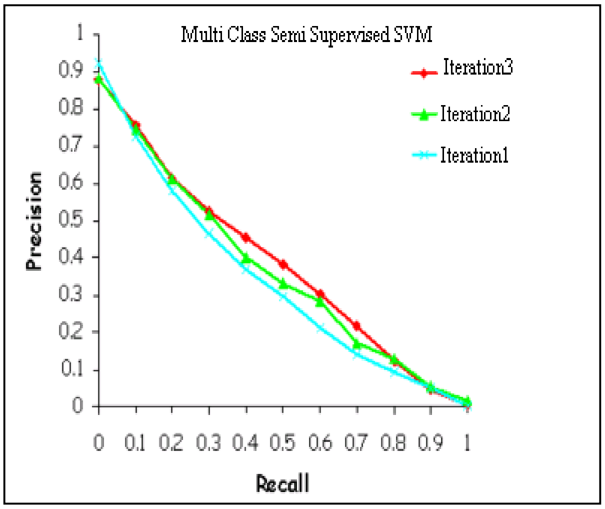


Fig: 4.0 precision and recall analysis after every iteration.

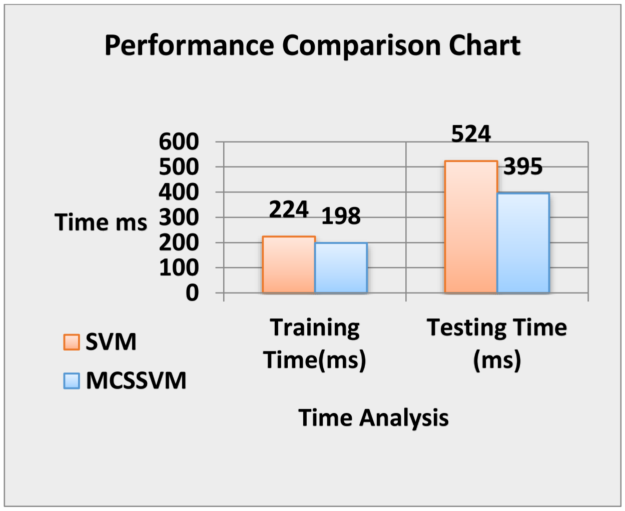


Fig:5.0 Time comparison between existing SVM and proposed MCSSVM

7. CONCLUSION

The proposed MCSSVM has been applied KDD cup 99 dataset. The system expands the existing SVM by applying Multi class and semi supervised sequence feature selection process to achieve two target. The first goal is improving the efficiency of the intrusion classification. Another one is the optimized feature selection for multi class SVM which is the trial of satisfying the research and improving the multi class detection accuracy. Further the study compares the with the existing system under different parameters. Finally the sequential semantic algorithm shows the better result than the existing intrusion classification algorithms.

REFERENCES

[1] Portnoy, Leonid, Eleazar Eskin, and Sal Stolfo. "Intrusion detection with unlabeled data using clustering." *In Proceedings of ACM CSS Workshop on Data Mining Applied to Security (DMSA-2001*. 2001.

[2] Baykara M, Das R. A survey on potential applications of honeypot technology in intrusion detection systems. International Journal of Computer Networks and Applications. 2015; 2(5):203–211

[3] Amoroso EG (1999) Intrusion detection: an introduction to internet surveillance, correlation, trace back, traps, and response. Intrusion.Net Books, NJ

[4] Mukkamala, Srinivas, Guadalupe Janoski, and Andrew Sung. "Intrusion detection using neural networks and support vector machines." Neural Networks, 2002. IJCNN'02. Proceedings of the 2002 International Joint Conference on. Vol. 2. IEEE, 2002.

[5] S. Mukkamala, G. Janoski, A. Sung. Intrusion Detection Using Neural Networks and Support Vector Machines. Proceedings of IEEE International Joint Conference n Neural Networks, pp.1702-1707, 2002

[6] Ahmed Youssef and Ahmed Emam “Network Intrusion Detection using Data Mining and Network Behavior Analysis” International Journal of Computer Science & Information Technology (IJCSIT) Vol 3, No 6, Dec 2011.

[7] S.A.Joshi, Varsha S.Pimprale “Network Intrusion Detection System (NIDS) based on Data Mining” International Journal of Engineering Science and Innovative Technology (IJESIT) Volume 2, Issue 1, January 2013

[8] S. Devaraju, S .Ramakrishnan “Detection of Accuracy for Intrusion Detection System using Neural Network Classifier” International Journal of Emerging Technology and Advanced Engineering(ISSN 2250-2459 (Online), An ISO 9001:2008 Certified Journal, Volume 3, Special Issue 1, January 2013)