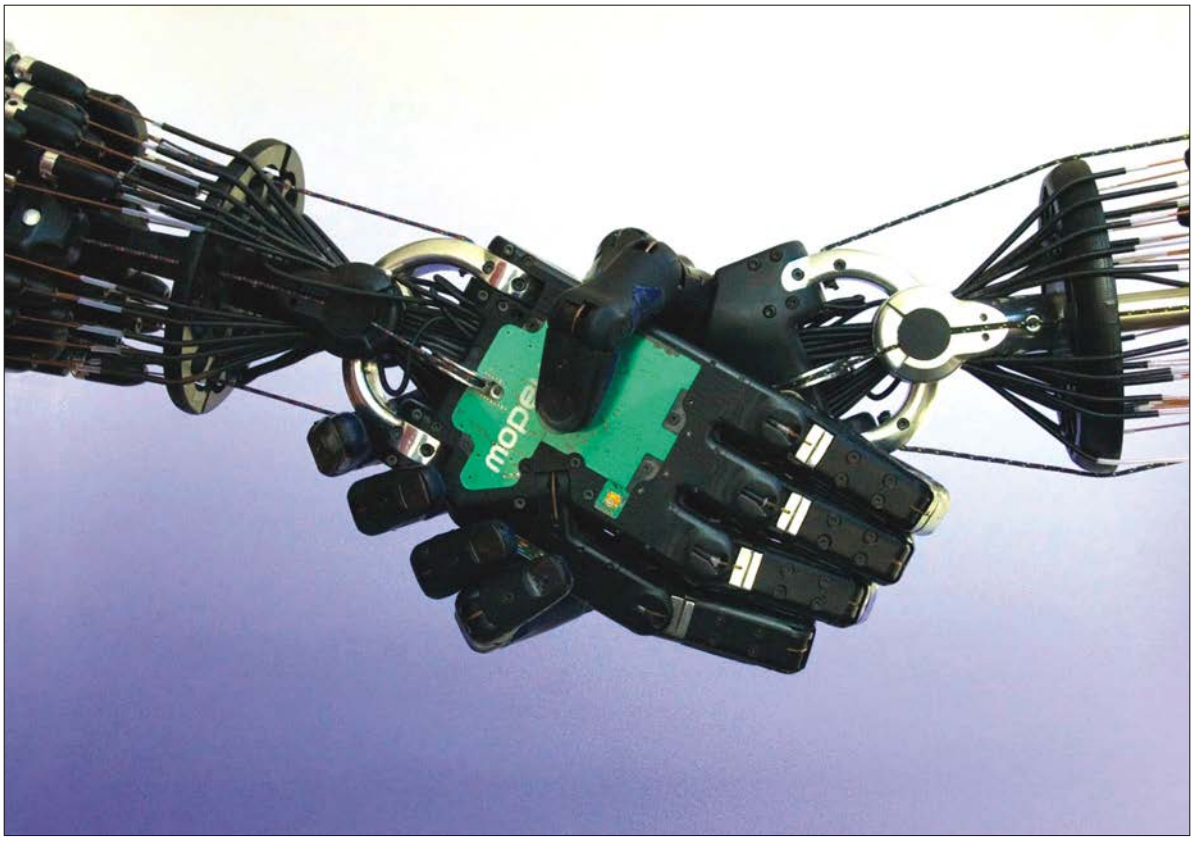

Journal of Artificial Intelligence

ISSN 1994-5450

science alert

ANSInet
an open access publisher
http://ansinet.com

OPEN ACCESS Journal of Artificial Intelligence

ISSN 1994-5450
DOI: 10.3923/jai.2016.1.11

Research Article
Anomaly Based Intrusion Detection in Mixed Attribute Dataset Using Data Mining Methods

B.A. Manjunatha and Prasanta Gogoi

Nitte Meenakshi Institute of Technology, Bangalore, India

Abstract

Background: In network security system anomaly detection is extremely important part in mixed attribute dataset. Detection of unexpected behavior in the network, by using outdated methods anomaly detection becomes inapt because data naturally occurs as mixture of numerical and categorical attributes. **Methodology:** The proposed algorithm, Minimum Threshold Support Count (MTSC) and modified Canberra method is used to detect mainly anomalies in categorical and numerical attributes (mixed attributes) that deals with sparse high-dimensionality of currently available dataset. Enhanced adaptive boosting classifier is very sensitive to anomalies and infrequent data. The accuracy and performance of the proposed method is comprehensively improved by using enhanced adaptive boosting classifier. **Results:** Results show that the classification True Positive Rate (TPR), precision, recall, F-measure and ROC area are more and false positive rate is less for the proposed method when compared to existing method. **Conclusion:** Proposed method gives effective classification accuracy and less computation time.

Key words: Anomaly detection, support count, high-dimensionality, false positive rate, F-measure, ROC

Received: April 26, 2016 Accepted: June 01, 2016 Published: June 15, 2016

Citation: B.A. Manjunatha and Prasanta Gogoi, 2016. Anomaly based intrusion detection in mixed attribute dataset using data mining methods. J. Artif. Intel., 9: 1-11.

Corresponding Author: B.A. Manjunatha, Nitte Meenakshi Institute of Technology, Bangalore, India

Competing Interest: The authors have declared that no competing interest exists.

Data Availability: All relevant data are within the paper and its supporting information files.

INTRODUCTION

The present study trend is facing important challenges and future study work in network anomaly detection by using mixed attributes. Anomaly detection is the detection of unexpected behavior in the network traffic, this unobserved activity is bypassing through the network. Intrusion Detection Systems (IDSs) attempt to identify invasion, suspicious behavior and report unobserved traffic activities. The IDSs have network-based and host-based types. Network based IDSs capture the packets from available network router or switch and analyze the packets as normal or abnormal. Host-based IDSs observed the single host for labeled packets or unlabeled packets. Network based IDSs is alienated into misuse based IDSs and behavior based IDSs (anomaly based IDSs)[1,2].

Misuse based IDSs detection is based on the pre-defined knowledge. A database that consists of mainly rules of past invasions and flaw knowledge is matched with the network packets. When the database rules are matched with network packets, it discovers active signatures meeting the requirements then it will issue a report to the administrator.

Anomaly based IDS (AIDS) is based on unexpected behavior in the network, where invasion can bypass through the network, it changes in flow behaviors[3]. The AIDS has two learning methods, supervised learning, where the IDS learns the network dataset from its known or labeled data and detects anomalies based on that model and the unsupervised learning, which is capable of handling unknown or unlabeled data by using data mining techniques, where false positive rate are more.

Data in the network segments occurs as combination of numerical and categorical data[4,5], detection of anomalies in combination of data can indicate harm of information[6]. Therefore, the comprehensive aspiration of analysis is required on mixed attributes datasets. To outrival the obstacles of the existing methods and to give a new proposed method comprehensive study is mandatory.

Adaptive boosting[7] is very sensitive to anomaly and infrequent objects in the network, this algorithm constructs a strong classifier by using fragile classifier repeatedly in a series of rounds. At each iteration, the training data are reweighted according to the fragile classifiers misclassification of the network packets. Again it's undergone into a various level of rounds until classification accuracy is obtained. The accuracy of the proposed enhanced AdaBoost algorithm is increased as well as false positive rate is decreased and it includes less memory and computation obligations.

Several previous researchers have addressed on anomaly detection in network datasets using data mining methods. Faria *et al.*[2] have discussed about novelty detection in data stream. Here this study consists of single classifiers, multi classifiers and ensemble of both classifiers is done on online mode and offline mode. This study ignores the traditional methods and is very sensitive to noise and outliers. Koufakou and Georgiopoulos[6] presents an outlier detection methods in mixed attribute datasets. This ODMAD method detects infrequent data or noise in mixed attribute space. The points with high score1 are more infrequent in categorical data. Hence the score1 do not have high values, so numerical score are considered and they use the cosine functions only to those points. Otey *et al.*[4] also supports an apriori approach of outlier detection method, which is based on frequent itemsets. This study is mainly based on apriori concepts and the execution time increases with the number of records. The covariance matrix calculation for each itemsets is required to handle continuous attributes, which requires more memory space. In this study the methods first mines the categorical attributes then numerical attributes and calculates an outlier score for data record. Radovanovic *et al.*[8] discussed about the detection of outliers in high dimensionality. To improve the accuracy of outlier anti-hubs are used, which improves the execution speed. Xu and Dong[9] proposes an Apriori method used to mine frequent patterns in a single minimum support threshold, which is not feasible for all items. To mine frequent patterns in different items per data they allow users to specify multiple minimum supports. The modified Apriori-based method called MS-Apriori uses multiple support count on real dataset[9], so that leads into decrease in complexity and overhead on the system. It keeps on changing in each dataset. Rastogi *et al.*[10] discussed unsupervised classification on mixed continuous, categorical, ratio and ordinal dataset using genetic algorithm. To group large datasets data mining concepts of clustering is used. Data clustering performs a job on huge datasets with mixed all types of attributes (ordinal, continuous, categorical, nominal, ratio and binary). Kumar and Mathur[11] discuses an efficient data mining outlier detection on unsupervised data in cloud environment. For unsupervised method training dataset or any previous knowledge of data is not required. The density-based method is used to work on unsupervised method for intrusion detection in large amount of data. In order to avoid multiple times scanning[12], the support of candidate itemset is determined depending on whether it is a frequent itemset or infrequent itemset. This is not based on the support of candidate itemset but it is based on the prior knowledge of

Apriori algorithm. This computation takes less scan and memory can reduce. Wu and Nagahashi[7], this method adaptively boost the weak methods by increasing their weights. Weak classifier has been combined with non-appropriate decision rules to generate accomplishment in records classification and object detection. Krawczyk *et al.*[13] discussed about ensemble decision trees, where they propose an ensemble of imbalanced classification on cost-sensitive decision trees. By using cost matrix base classifiers are constructed, which are trained on random feature subspaces. To achieve more accuracy through ensemble methods sufficient diversity of the members is ensured. Ahmad[14], proposes ensemble based decision tree on kernel features. Here, researchers consider all kernel parameters of different hybridized methods like bagging, random forest, random subspace and AdaBoost M1 are applied. In this study, selection of kernel parameters is pretty vigorous.

In any IDS datasets like KDD cup data set mixed attribute datasets are present. Apriori algorithm takes huge candidate itemset I and the computation of support (supersets) count consumes a lot of CPU time[4]. Multiple scans[11] will increase the input/output load on memory and it's again more time consuming.

The KDD cup 10% corrected dataset address denial of service (DoS) and probe invasions are occurring in huge amounts and remote to local (R2L) and a user to remote (U2R) invasions occur rarely. This motivated the use of proposed MTSC, modified Canberra and enhanced adaptive boosting classifiers to detect anomalies.

MATERIALS AND METHODS

Anomalies or the rare objects are those objects which have less count in datasets. The traditional methods like Euclidean distance or nearest neighbor methods are not suitable in high dimentional dataset[15]. Hence, Minimum Threshold Support Count (MTSC) method and modified Canberra methods are used to retrieve these anomaly or infrequent objects in mixed attribute dataset. In this study, step by step procedure to represent a data pre-processing method is explained. Their basic properties and terminologies of the symbols are described in Table 1.

Data pre-processing methods:

Step 1: Any dataset DS = {R1, R2, ... ,Rn} of n records, let R_i be the ith record I = 1.....n in R_i.

Step 2: In each data record Ri in the dataset, R_{ic} is categorical attribute, R_{in} is numerical attribute and R_{idc} is the class label of R_i.

Step 3: This is the representation of Ri = [R_{ic}, ..., R_{in},..., R_{idc}].

Step 4: To calculate score for each record, it extracts only categorical attributes in mixed attributes, the idea of an item set I from DS.

Step 5: The collection of data having a transaction TR = {tr_1, tr_2, ⋯tr_n}, where tr_{id} is a transaction id, which is consisting of a subset of R_i.

In each record to make an itemset it ensures basic properties:

- Infrequent property: If itemset I is a minimum frequent itemset, then supp (I) <TR
- A transaction of categorical itemset $tr_i \in R_i$ is said to contain itemset I, if $I \subseteq tr_i$.DS (I) = {$tr_i \in R_{ic}$: $I \subseteq TR$}
- Support property: The support count of categorical itemset I is stated as: supp (I), that means support the sum of transactions that include an itemset I in TR
- Support of maximum frequent set property: maximum frequent itemset is defined as: Maximum (R_i, Maximsupp) = {$I \in R_i$ | supp (I) $\geq \sigma$} for Tr_i transaction
- Support of minimum frequent set property: Minimum infrequent itemset is defined as: Minimum (R_i, Minimsupp) = {$I \in R_i$| supp (I) $\leq \sigma$} for Tr_i transaction

The minimum frequent itemset are the anomaly intrusions of tr_i, which is denoted by score1 that is used for categorical itemset as given in Eq. 1:

$$\text{score1}\ (R_{ic}) = \sum_{I \varepsilon tri, R_{ic} \subseteq I} \frac{\text{Min}(\text{supp}(I))}{\|TR\|} \leq \sigma \quad (1)$$

The score1 equation includes minimum subset and all these subsets are also infrequent item sets because all have minimum support count. A transaction with regular data is

Table 1: Terminology of work

Terms	Description
I	Itemset
tr_{id}	Transaction identifiers
supp (I)	Support of itemset
DS	Dataset
Maximsupp	Maximum Support count
Minimsupp	Minimum Support count
σ	User defined minimum threshold support count
R_i	ith data record in DS
N	The number of attributes in Ri
R_{ic}	Categorical itemset/attributes
R_{in}	Numerical itemset/attributes
R_{idc}	Labeled attribute
‖TR‖	Total transaction of DS
x_{ai}	Numerical i^{th} attribute of x_a
x_{bi}	Numerical i^{th} attribute of x_b

more likely to be a normal transaction, some times it may be DoS and DDoS attacks because items are large distinguishable frequent patterns given in Eq. 2:

$$\text{score2}\ (R_{in}) = 1 - \frac{\sum_{i=1}^{n} |x_{ai} - x_{bi}|}{\sum_{i=1}^{n} (x_{ai} + x_{bi})} \leq \sigma \quad (2)$$

Numerical candidates need to be calculated by using modified Canberra Eq. 2 for infrequent records. The above score1 equation, which calculates scores of those candidates having minimum score1 allows, testing Eq. 2, where x_{ai} is a numerical ith attribute of X_a minimum frequent record and x_{bi} is numerical ith attribute of X_b minimum frequent record.

The modified Canberra Eq. 2 acts as a score2 for numerical attributes. The results of score2 are low when attribute scores are unlike, which means they are close to zero. When attribute scores are like this means that they are close to one. Anomaly or infrequent itemset is calculated using the Eq. 3 that gives the combination of numerical and categorical scores below:

$$\text{score} = \frac{\text{score1} + \text{score2}}{2} \quad (3)$$

The proposed study is justified by using synthetic dataset shown in Table 2 and it has been illustrated in Table 3-6. Step by step procedure is used to find, whether a given record is infrequent or anomaly in dataset. For example dataset DS, the count of each categorical candidate with support count is {M} = 4, {N} = 2, {O} = 5, {P} = 4 and {R} = 6.

Table 3 shows the MTSC score1 calculated using Eq. 1. Thus, which score1 are high induced these are more frequent. Hence, the score1 that does not have high values induced are infrequent. Users also define minimum threshold support count to set minimum threshold to get more accuracy.

Minimum frequent patterns that are present in the data transactions with a frequency minimum threshold support count i.e., item {N} will have support count 2 by using supp (x). Score1 is computed using Eq. 1 which is equal to 0.33.

Table 4 shows the MTSC score1 calculated for 2-itemset patterns. In general, in Apriori algorithm (Xu and Dong, 2013) if the infrequent itemset patterns are found then its associated superset patterns also must be infrequent and infrequent patterns are consider as an anomaly. A pruning strategy is applied based on Apriori principle because it takes less computation and memory. It is a standard calculation for all supersets: {M, N} = 1, {M, O} = 3, {M, P} = 3, {M, R} = 4, {N, O} = 2, {N, P} = 0, {N, R} = 2, {O, P} = 3, {O, R} = 5 and {P, R} = 3. This proposed methods do pruning strategy during computation of infrequent itemset shown in Table 4.

Next, MTSC score1 computes the supersets of 3-itemset. It is a standard calculation for all supersets: {M, N, O} = 1, {M, N, P} = 0, {M, N, R} = 1, {M, O, P} = 2, {M, O, R} = 2, {M, P, R} = 3, {N, O, P} = 0, {N, O, R} = 2 This {O, P, R} = 3. This proposed methods do pruning strategy during computation of infrequent itemset shown in Table 5.

Next the given Table 4 shows, MTSC score, which calculates the supersets of 4-itemset. All itemsets are non-common patterns itemset. It is a standard calculation for all supersets: {M, N, O, P} = 0, {M, N, O, R} = 1 and {N, O, P, R} = 0.

With correspond to above infrequent itemsets record, the associated continuous record is extracted to compute score2. Modified Canberra method uses numerical attributes in each record, which do not have high score1 Eq. 1 (infrequent). The modified Canberra Eq. 2 is used to calculate score2. Anomaly

Table 2: Synthetic dataset

t_{id}	itemset
1	N, 3, O, R, 1, 2
2	M, 1, O, P, R, 2, 4
3	O, 4, P, R, 5
4	M, 5, N, P, R, 2
5	M, 1, O, R, 2
6	M, 2, O, P, R, 5, 3

Table 3: 1-itemsets

1-itemset	Supp (l)	Score1
P	4	0.66
R	6	1
M	4	0.66
N	2	0.33
O	5	0.83

Table 4: 2-infrequent itemsets

2-itemset	Supp (x)	Score1
N, P	0	0.00
N, O	2	0.33
N, R	2	0.33

Table 5: 3-infrequent itemsets

3-itemset	Supp (x)	Score1
M, N, R	1	0.16
N, O, P	0	0.00
M, O, P	2	0.33
M, O, R	2	0.33
M, N, O	1	0.16
M, N, P	0	0.00

Table 6: 4-infrequent itemsets

4-itemset	Supp (l)	Score1
M, N, O, R	1	0.16
N, O, P, R	0	0.00
M, N, O, P	0	0.00

or infrequent itemsets can be calculated using inclusion of score1 and score2 in Eq. 3.

Algorithm 1: Working of the MTSC and modified Canberra methods

Input: Dataset DS (n records, R_{ic}, R_{in} attributes), σ, I, score1, score2, minimsupp
Output: Anomaly detected

1: for each object Ri, I = 1...n do
2: scan_infrequent 1- itemsets (DS);
3: score1 = categorical attributes $R_{ic} = 0$;
4: for each categorical attributes $tr_i \in R_{ic}$ do
5: if supp (I) ≤ σ then
6: score1 (R_{ic}) + = (supp (I))/||TR||;
7: end
8: if minimsupp ≤ supp(I) then
9: pruned infrequent supersets $R_{ic} \in$ I do

10: $$\text{score1}(R_{i\,csc}) + = \frac{\text{Min (supp (I))}}{\|TR\|} \le \sigma$$

11: end
12: end
13: for each score 2 = continuous attribute $R_{in} \in R_{ic} \wedge R_{in}$ do

14: $$\text{score 2}(R_{in}) = 1 - \frac{\sum_{i=1}^{n} |x_{ai} - x_{bi}|}{\sum_{i=1}^{n} (x_{ai} + x_{bi})} \le \sigma$$

15: end

16: $$\text{score} = \frac{\text{score1} + \text{score 2}}{2}$$

17: if score<1
18: X←anomaly;
19: else
20: X←normal;
21: end
22: end

Complexity analysis: In the step-by-step procedure of MTSC, input is DS, where n is the number of records in DS. In order to validate our algorithm switches Minimum Threshold Support Count (MTSC) is calculated for each itemset, the time complexity of the computation of support count is O (n). To compute pruned infrequent supersets of categorical attributes with respect to MTSC. The time complexity of numerical attributes is also O (n), involving of categorical and numerical attributes. Average case complexity is O (n_c)+O (n_n) = O (2n) and worst case complexity is O (3n), (because O (n)+O (n_c)+ O (n_n) = O (3n)).

Adaptive boosting (AdaBoost): The AdaBoost Wu and Nagahashi[7] is very sensitive to anomaly and infrequent objects. This algorithm constructs a strong classifier by using fragile learners repetitively in a sequence of rounds. At each iteration or repeat, training data are reweighted according to the weak classifiers that misclassify the data. This iteration process is repeated until classification performance is obtained as shown in Fig. 1. Here, the classification has been misclassified, hence, weights of the points that have been misclassified are increased, that is the weight of one red point and two blue points are increased as shown below in step 2. Hence, in the next iteration it tries to classify the points correctly. And the procedure is as shown below in step 3. Here red points had been classified correctly but two blue points are in the space of red points. Hence, these points should be classified properly by increasing the weights of those points in step 4. After final iteration the classification gives result as below in step 5.

In enhanced adaptive boosting classifier approach, large number of network intrusion data can be handled. In this proposed approach, it improves the AdaBoost algorithm fragile learner steps, as explained above in Fig. 1. It uses newly approached classifier that contains hybridized random forest, AdaBoost and normalization algorithms are ensemble on the basis of average of their probabilities. In this hybridized method accurate decision-making treesare added by using weighted random objects in normalization algorithm. The assigned weight is used to collect the object for each classifier. If there is less error rate of classifier

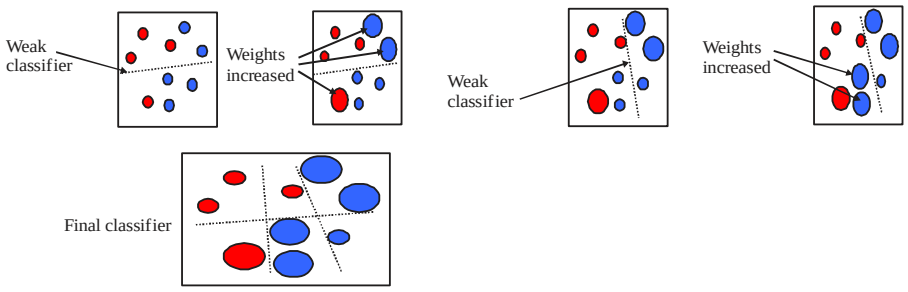

Fig. 1: AdaBoost algorithm each step used to constructing a strong classifier

then more weight assigned to its objects. This kind of data suffers a significant variation due to classification errors and limitations. Presently log transformation is used, which gives satisfactory results but it's still skeptical of possibilities of false positives. So, normalization method is used to normalize the objects, that the ranges of values are normalized to be from 0.0-1.0. If the classification error range between 0.0-0.5 then reweight the objects to 0.2. If the classification error range between 0.5-1.0 then reweight the objects 0.1.

Algorithm 2: Enhanced AdaBoost algorithm

Algorithm:

1: Initialize the weights of data objects;
2: collect the samples of the objects according to the weights;
3: classifier model using hybridized random forest, AdaBoost and normalization combined is derived from the training set;
4: compute the classification error of just selected weak classifier for each data point:

$$\text{Classification error (Me)} = \sum_{j}^{n} w_j + \text{error}(X_j);$$

5: If the compute error is between 0.1 and 0.5 then update the weight of the classified objects

$$\text{W of object} \times \frac{\text{error of Me}}{1-\text{error of Me}} + 0.2;$$

6: If the compute error is between 0.5 and 1.0 then update the weight of the classified objects

$$\text{W of object} \times \frac{\text{error of Me}}{1-\text{error of Me}} + 0.1$$

7: Normalize the weights of each object M (E) using

$$\text{W of object} \times \frac{\text{sum of old weights}}{\text{sum of new weights}};$$

8: repeat step 2;
9: A weight of the classifers vote is

$$W_i = \log \frac{\text{error of M (E)}}{1-\text{error of M (E)}};$$

10: the class prediction for objects X from Me is

$$C = \text{Me}(X);$$

11: Add W_i weights for class C
12: The class having higher sum is the winner class and returns as prediction for object X.

From this all objects need to have the similar positive measure standardization. Drawback of this normalization is lost for the outliers from the dataset. A weight of the classifiers vote is calculated and the class for model prediction i.e., on the basis of error rate more weight to class is added that will give better class for the prediction. Weak classifier has been combined with non-appropriate decision rules to generate high prediction rule. The class having higher sum is the winner class as described in algorithm 2.

Complexity analysis: In enhanced AdaBoost classifiers algorithm for each step is used to construct strong classifiers. Time complexity to collect the training samples of n records is O (n). In the proposed approach, weights of data are raised and data is collected according to the weights. Time to compute the classification error and normalize the weights is O (n). Average case complexity is O (n)+ O (n) = O (2n).

RESULTS

A general definition of dataset is a tabular form of collection of records/data normally. Each column represents a particular record. Mainly two types of data set are present which are: Test dataset and training dataset. In dataset majority portion is used for training and a minor portion is used for testing. To detect anomaly in the dataset KDD Cup 10% corrected IDS datasets is used.

The KDD Cup 10% corrected dataset[16]: It is labeled the data as either specific type of invasion or normal. The types of invasion are:

- Denial of service (DoS): The intruder trying to avoid legitimate users from a service
- Probe: Intruder tries to access information about the particular machine
- User to remote (U2R): The intruder is trying to access normal user account on the system and exploits some vulnerability to the system
- Remote to local (R2L): The intruder access as a user of that system without any knowledge of original user. Intruder can exploits vulnerability to gain local access as a user of that system

Measuring tool: The results are measured using confusion matrix, it has 4 components False Positive (FP) rate, False Negative (FN) rate, True Positive (TP) rate and True Negative (TN) rate. Here accuracy and performance is also very important to measure results.

Accuracy: Accuracy is defined as the fraction of correct prediction out of all predictions shown in Eq. 4:

$$\text{Accuracy} = \frac{\text{No. of correct pridictions}}{\text{Total number of pridictions}} = \frac{TP+TN}{TP+FN+FP+TN} \quad (4)$$

Standard measures: Precision, recall, F-measures and Return Out Characteristics (ROC) area are defined and measures to compare the performance. Precision[17] is defined as the ratio of correctly classified instances (true positive) and classified as all positive as shown below:

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}$$

Recall[17] is defined as the ratio of correctly classified instances (true positive) and positive to the positive element and negative element:

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}$$

F-measure/F-score[17] formula as shown below:

$$\text{F}-\text{measure} = 2\frac{\text{Precision}\times\text{recall}}{\text{Precision}+\text{recall}}$$

The ROC area[17] formula as shown below:

$$\text{ROC} = \frac{\text{P}(\text{x}|\text{positive})}{\text{P}(\text{x}|\text{negative})}$$

Experimental setup: Experimental setup was conducted using Dell precision server R5500-intel Xeon processor E5620 (Quad core, 2.40 Ghz turbo, 12MB, 5.86 GT/s)/24GB (3×8GB) DDR3 RDIMM memory, 1333Mhz, ECC/2×250 GB 2.5 inch SATA hard drive, Microsoft VS-ultimate 2010, OS-windows 8.1.

The MTSC and modified Canberra methods is implemented by using Microsoft VS-ultimate 2010 tool. Enhanced AdaBoost classifiers method is implemented to increase the classification accuracy and performance of the proposed and existing methods. All these methods were performed on the KDD cup 10% corrected dataset. While, doing experiment initially dataset was retrieved from the database. Then whether all data is similar to structured data or not is found out. The MTSC and modified Canberra methods are used to extract anomalies or infrequent data by applying user defined minimum threshold support count. Initially entire output results are varied because of applying a single minimum threshold support count to 10%. The huge number of itemset contained many frequent attacks because of which DoS and PROBE attacks are classified partially. Then the user minimum threshold support count is increased to 30% then rarely occurring attack of R2L and U2R attacks are properly classified but DoS and probe attacks are partially classified. This algorithm takes less data scan and resources but is unable to get comprehensive set of output as shown in above dataset example. By allowing multiple minimum threshold support count[9] scan of dataset through MTSC and modified Canberra method detect anomalies or infrequent itemset. The comparison of proposed and existing method with respect to TPR, FPR, precision, recall, F-measure and ROC area on KDD cup10% corrected dataset is shown in Fig. 2-6.

This KDD cup 10% corrected dataset has exactly one specific type of DoS attack record with 41 attributes and one labeled attribute. This DoS attack includes back, land, pod, smurf, teardrop and Neptune i.e., about 391458 attacks. This proposed method provides better classification compared to existing method as shown in Fig. 2. The minimum threshold support count is set to 50% to obtain better classification in MTSC and modified Canberra methods. To improve the better classification accuracy and performance enhanced AdaBoost classifiers are used.

This KDD cup 10% corrected dataset has exactly one specific type of PROBE attack record that has 41 attributes and one labeled attribute. This PROBE attack includes satan, ipsweep, nmap and portsweep i.e., about 4107 attacks. This proposed method provides better classification compared to existing method as shown in Fig. 3. The minimum threshold support count is set to 30% to obtain better classification in MTSC and modified Canberra methods. To improve the better classification accuracy and performance enhanced AdaBoost classifiers are used.

This KDD cup 10% corrected dataset has exactly one specific type of R2L attack record that has 41 attributes and one labeled attribute. This R2L attack includes guess_password, ftp_write, imap, phf, multihop, warezmaster, warezclient and spy i.e., about 1126 attacks. This proposed method provides better classification compared to existing method as shown in Fig. 4. The minimum threshold support count is set to 30% to obtain better classification in MTSC and modified Canberra methods. To improve the better classification accuracy and performance enhanced Adaboost classifiers are used.

This KDD cup 10% corrected dataset has exactly one specific type of U2R attack record that has 41 attributes and one labeled attribute. This U2R attack includes buffer_overflow, loadmodule, perl and rootkit i.e. about 52 attacks. This proposed method provides better classification compared to existing method as shown in Fig. 5. The minimum threshold support count is set to 10% to obtain better classification in MTSC and modified Canberra methods.

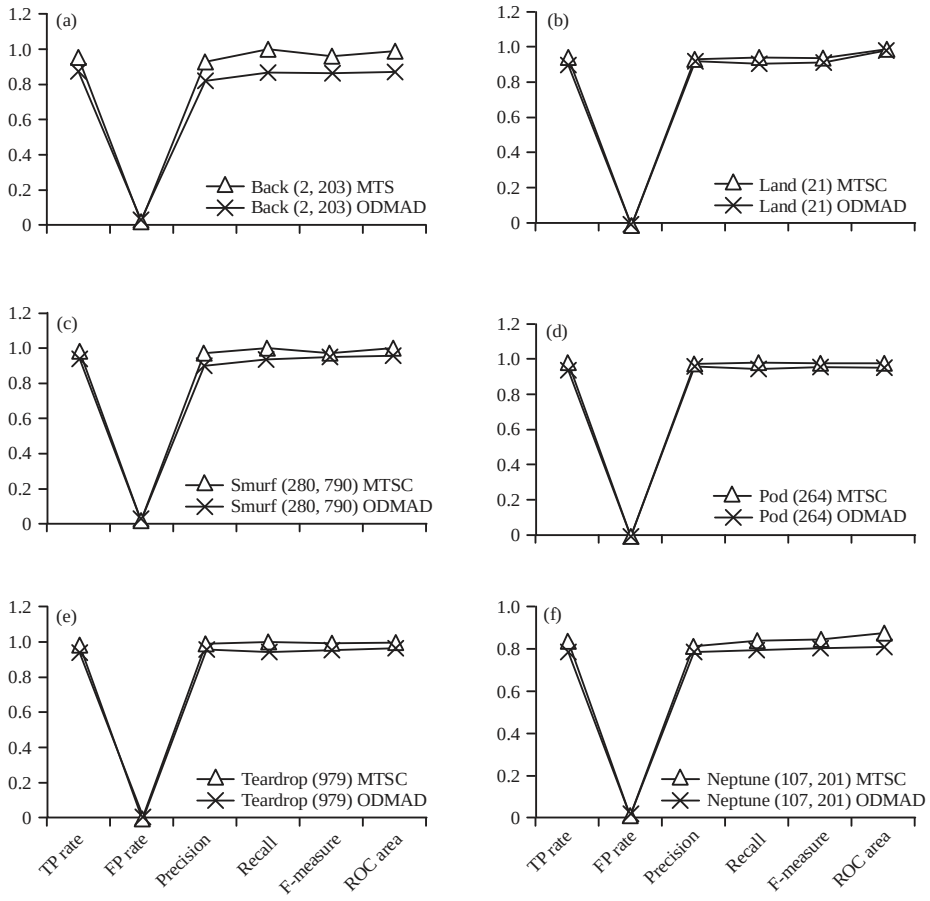

Fig. 2(a-f): Comparison of DoS attacks with respect to TPR, FPR, precision, recall, F-measure and ROC area between proposed and existing method

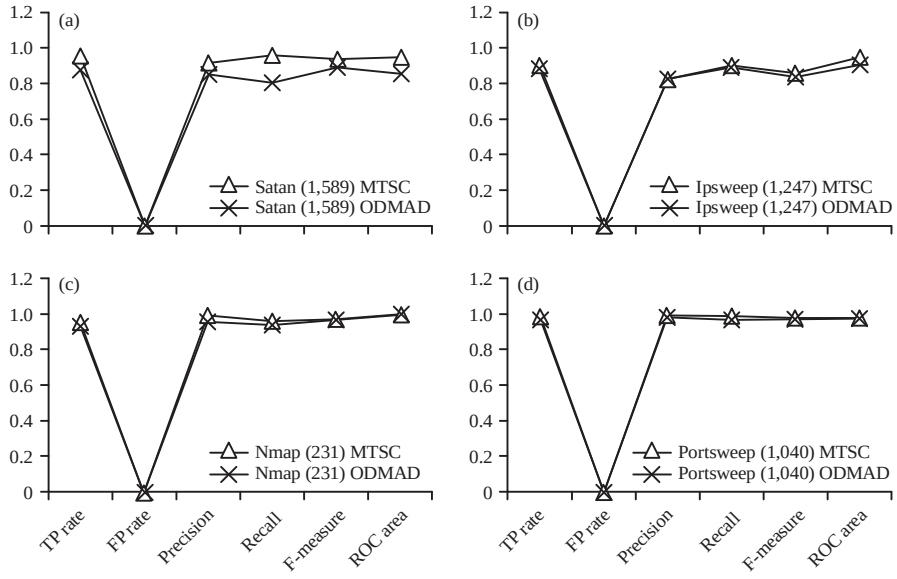

Fig. 3(a-d): Comparison of PROBE attacks with respect to TPR, FPR, precision, recall, F-measure and ROC area between proposed and existing method

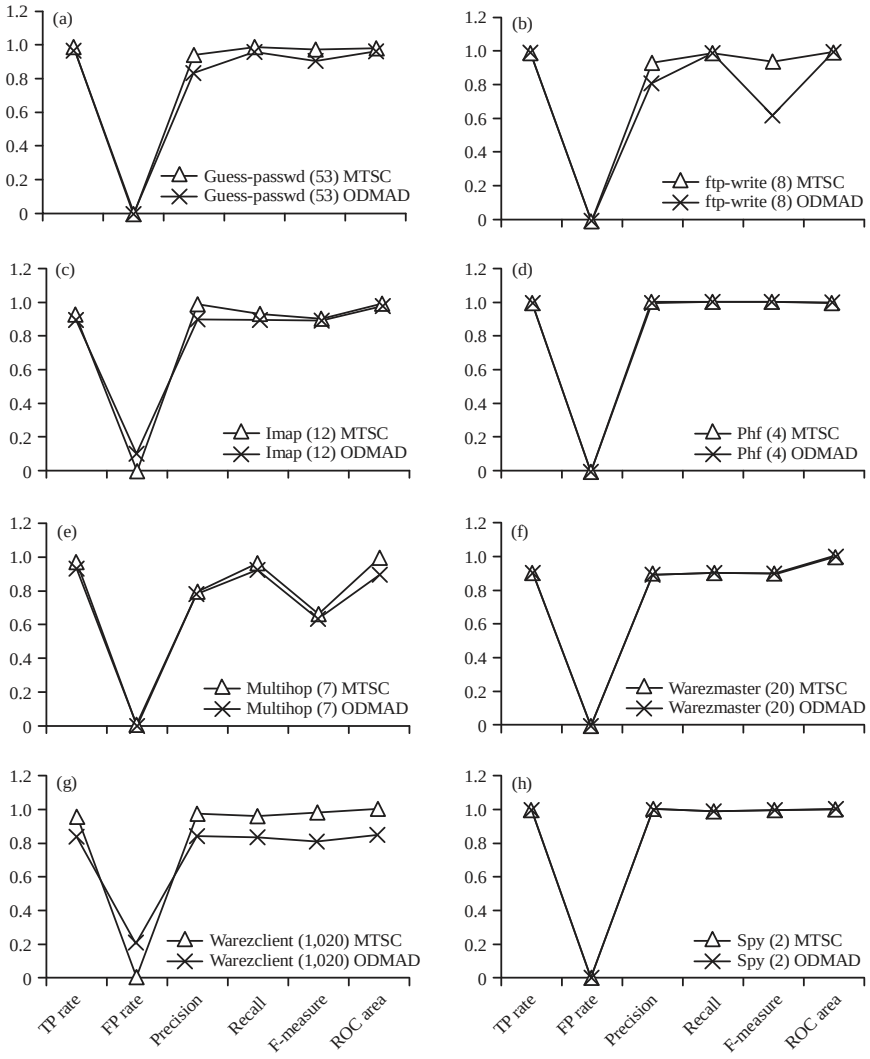

Fig. 4(a-h): Comparison of R2L attacks with respect to TPR, FPR, precision, recall, F-measure and ROC area between proposed and existing method

To improve the better classification accuracy and performance enhanced AdaBoost classifiers are used.

This KDD cup 10% corrected dataset has normal record that has 41 attributes and one labeled attribute. This normal record consists of 97,277 records. This proposed method provides better classification compared to existing method as shown in Fig. 6. The minimum threshold support count is set to 30% to obtain better classification in MTSC and modified Canberra methods. The difference between proposed and existing method is shown in Table 7 and 8 and it shows proposed method is better than existing method.

From the entire diagram it is observed that the proposed method performs well for the training set and it is verified with an accuracy of 97.85%. This proposed study has advantage of less memory consumption because of infrequent data

Table 7: Proposed method

Correctly classified record	97.8586%
Incorrectly classified record	2.1414%
Mean absolute error	0.0476
Root mean squared error	0.1125
Relative absolute error	76.6968%
Root relative squared error	60.4159%
Total number of record	494021
Time taken	3.3647 sec

Table 8: Existing method

Correctly classified record	95.2863%
Incorrectly classified record	4.7137%
Mean absolute error	0.0834
Root mean squared error	0.2467
Relative absolute error	96.7654%
Root relative squared error	76.5643%
Total number of record	494021
Time taken	8.6734 sec

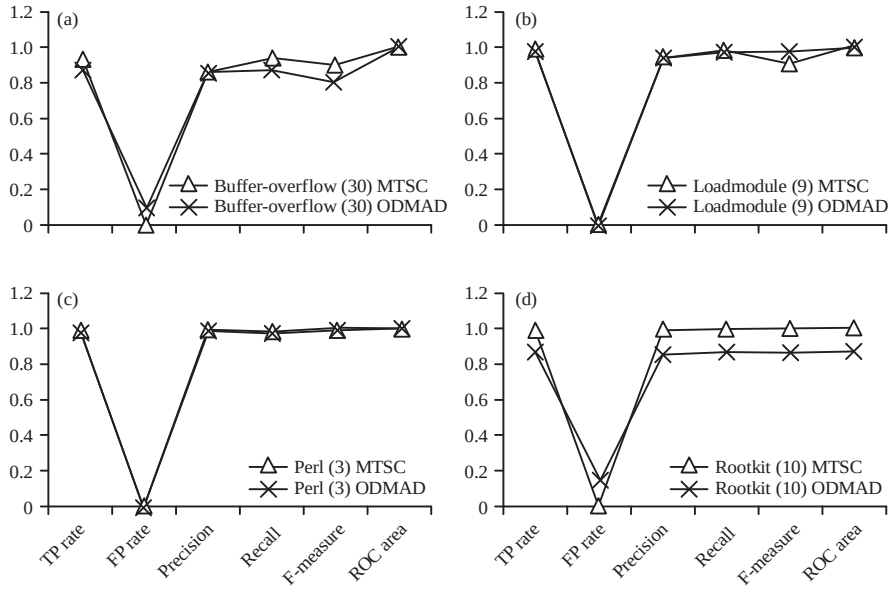

Fig. 5(a-d): Comparison of U2R attacks with respect to TPR, FPR, precision, recall, F-measure and ROC area between proposed and existing method

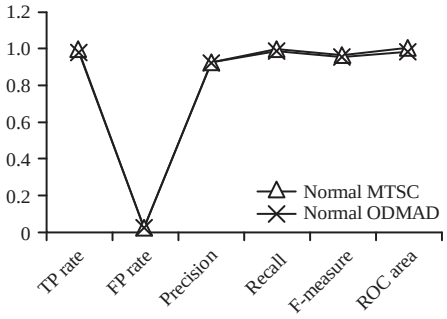

Fig. 6: Comparison of normal packets with respect to TPR, FPR, precision, recall, F-measure and ROC area between proposed and existing method

computation. It is observed that proposed methods use less CPU computation because of infrequent data. The enhanced AdaBoost classifiers method is used to improve accuracy nearly from 1-3% and time is reduced from 5-10 msec. But existing method performs more computation and hence it takes more CPU resources for frequent data and they have used single threshold support count for comprehensive dataset. Enhanced AdaBoost classifier indicates that the accuracy of the proposed method is improved by using hybridized random forest and AdaBoost and normalization methods.

DISCUSSION

In latest previous studies data in the network segments occurs as combination of numerical and categorical data[4,5], detection of anomalies in combination of data can indicate loss of information[6]. In proposed method there are 41 attributes and one labeled attribute, i.e., total mixed combination of 42 attribute. The ODMAD method detects infrequent data or noise in mixed attribute space. The points with high score1 are more infrequent in categorical data. Hence, the score1 do not have high values, so numerical score are considered and they use the cosine functions only to those points. A result shows the accuracy of less than 95% in finding anomalies. Otey *et al.*[4] supports an Apriori approach of outlier detection method, which is based on frequent itemsets. This study is mainly based on Apriori concepts and the execution time increases with the number of records. The covariance matrix calculation for each itemsets is required to handle continuous attributes, which requires more memory space. Xu and Dong[9] considered an Apriori method used to mine frequent patterns in a single minimum support threshold, which is not feasible for all items. To mine frequent patterns in different items per data they allow users to specify multiple minimum supports. The modified Apriori-based method called MS-Apriori uses multiple support count on real dataset. In all above studies they considered only frequent itemsets but in this study infrequent itemsets are used to reduce computation. The methods first mines the categorical attributes by using MTSC then numerical attributes by using modified Canberra and calculates an outlier score for data record. To improve the performance and accuracy of proposed method enhanced adaptive boosting classifiers are used which classifies 97.85% of anomalies with in 3.36 sec.

CONCLUSION AND FUTURE RECOMMENDATION

This proposed algorithm, Minimum Threshold Support Count (MTSC) and modified Canberra methods are applied to detect anomaly or infrequent patterns in mixture of categorical and numerical datasets. In the proposed method multiple minimum thresholds are applied to classify other then rare occurring attacks. The proposed enhanced AdaBoost classifiers can improve the classification accuracy as well as reduce the processing time and perform reliably better for defect classification. The benefits of enhanced AdaBoost classifiers include less memory and computation necessities.

Existing method have been tested on our method using KDD cup 10% corrected dataset. Furthermore, in future computational complexity has to be reduced in each dataset scan. If dataset size is huge (Tera bytes) then it takes more memory.

REFERENCES

1. Gogoi, P., B. Borah and D.K. Bhattacharyya, 2011. Network anomaly detection using unsupervised model. Int. J. Comput. Applic. (Special Issue on Network Security and Cryptography), NSC(1): 19-30.
2. Faria, E.R., I.J.C.R. Goncalves, A.C.P.L.F. de Carvalho and J. Gama, 2016. Novelty detection in data streams. Artif. Intell. Rev., 45: 235-269.
3. Gogoi, P., B. Borah and D.K. Bhattacharyya, 2013. Network anomaly identification using supervised classifier. Inform. Int. J. Comput. Inform., 37: 93-105.
4. Otey, M.E., A. Ghoting and S. Parthasarathy, 2006. Fast distributed outlier detection in mixed-attribute data sets. Data Mining Knowl. Discov., 12: 203-228.
5. Tran, K.N. and H. Jin, 2010. Detecting network anomalies in mixed-attribute data sets. Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, January 9-10, 2010, Phuket, pp: 383-386.
6. Koufakou, A. and M. Georgiopoulos, 2010. A fast outlier detection strategy for distributed high-dimensional data sets with mixed attributes. Data Mining Knowl. Discov., 20: 259-289.
7. Wu, S. and H. Nagahashi, 2014. Parameterized AdaBoost: Introducing a parameter to speed up the training of real AdaBoost. IEEE Signal Process. Lett., 21: 687-691.
8. Radovanovic, M., A. Nanopoulos and M. Ivanovic, 2014. Reverse nearest neighbors in unsupervised distance-based outlier detection. IEEE Trans. Knowl. Data Eng., 27: 1369-1382.
9. Xu, T. and X. Dong, 2013. Mining frequent patterns with multiple minimum supports using basic Apriori. Proceedings of the 9th International Conference on Natural Computation, July 23-26, 2013, Shenyang, pp: 957-961.
10. Rastogi, R., S. Agarwal, P. Sharma, U. Kaul and S. Jain, 2014. Unsupervised classification of mixed data type of attributes using genetic algorithm (Numeric, categorical, ordinal, binary, ratio-scaled). Adv. Intell. Syst. Comput., 258: 121-131.
11. Kumar, M. and R. Mathur, 2014. Unsupervised outlier detection technique for intrusion detection in cloud computing. Proceedings of the International Conference for Convergence of Technology (I2CT), April 6-8, 2014, IEEE., Pune, pp: 1-4.
12. Zhang, K., J. Liu, Y. Chai, J. Zhou and Y. Li, 2014. A method to optimize apriori algorithm for frequent items mining. Proceedings of the 7th International Symposium on Computational Intelligence and Design, Volume 1, December 13-14, 2014, Hangzhou, pp: 71-75.
13. Krawczyk, B., M. Wozniak and G. Schaefer, 2014. Cost-sensitive decision tree ensembles for effective imbalanced classification. Applied Soft Comput., 14: 554-562.
14. Ahmad, A., 2014. Decision tree ensembles based on kernel features. Applied Intell., 41: 855-869.
15. Murthy, M.K., A. Govardhan and D.L.S. Reddy, 2013. A model to find outliers in mixed-attribute datasets using mixed attribute outlier factor. Int. J. Comput. Sci. Issues, 10: 215-219.
16. KDD Cup., 1999. Index of /databases/kddcup99. http://kdd.ics.uci.edu/databases/kddcup99/.
17. Sokolova, M., N. Japkowicz and S. Szpakowicz, 2006. Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. In: AI 2006: Advances in Artificial Intelligence, Sattar, A. and B.H. Kang (Eds.)., LNCS. Vol. 4304, Springer, Berlin, Heidelberg, pp: 1015-1021.