International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-04, Issue-04, July 2018

An Efficient Approach for IDS using Minimum Redundancy and Maximum Relevance Feature Selection Algorithm

[1]Hasmitha Dendukuri, [2]Mrs. A. Ramana Lakshmi

[1]PG Scholar, [2]M.Tech (PhD), Associate Professor, [1,2]Prasad V. Potluri Siddhartha Institute Of Technology, Vijayawada, A.P, India.

Abstract: An intrusion detection system refers to the systems which detect the security breaches in a network. The main goal of IDS is to detect the attacks and raise alarms if an attack is found. These IDS has limitation of raising false alarms and this limitation depends on improper classification of attacks due to high dimensionality of data on a network. Improper classification is due to improper feature selection. To overcome this limitation, optimization of feature selection technique is proposed in this paper. Fertilization based streamlining procedure is proposed which in the long run enhances MRMR include determination in high dimensional information of IDS and accomplish decrease in false alerts.

Keywords - Classification, False alarms, High dimensionality, IDS, Pollination.

I. INTRODUCTION

Intrusion, in simple words, is an illegal act of entering in a network or computer system in an unauthorized way. And intrusion detection system is that system which is developed to detect unauthorized use of network or any malicious activities on computer system. These intrusion detection systems are suffering from the problem of false detection of attacks. False detection refers to classifying normal traffic as harmful (false positives) or classifying harmful traffic as normal (false negatives) [1]. Various approaches are being used to overcome the false alarm problem of IDS such as data mining, event correlation and classification of alerts. Data mining is a field which deals with hundreds of data which refer to as high dimensional data. And the main reason for false alarms in IDS is high dimensional data, thus, it can be solved by proper data mining technique. Data mining is a necessary tool by stylish trade to rework knowledge into business intelligence giving an advantage. It's presently used an exceedingly smart range of identification practices, like promoting, surveillance, fraud detection, and scientific discovery. One of the basic categories of data mining is classification. Classification could be a task of categorizing one thing into predefined categories. Classification is done on the basis of features/ attributes/ behavior. Feature selection is that the major issue upon which the classification depends, means, if features are properly selected, then the classification are going to be automatically correct and if features don't seem to be elect in correct approach, then classification can offer dangerous leads to data mining.

II. LITERATURE SURVEY

1. FS has been a lively and providing much detail about the field of analysis area in pattern recognition, machine learning, statistics and data mining communities. The principle target of highlight choice is to select an arrangement of info factors in dispensing with choices, that are inadmissible or of no prophetic information. Feature selection has verified in each theory and applies to be effective in increasing learning efficiency, increasing predictive accuracy and reducing quality of learned results. Highlight choice in directed learning consolidates in a fundamental objective of investigating a list of capabilities which produces higher order precision [3]. Since the spatial property extends, the measure of alternatives increments. Finding AN ideal list of capabilities is intense.

At this point, it's fundamental to clarify antiquated component determination technique, which comprises of 4 essential advances, to be specific, set age, set examination, halting model, and approval. Set generation could be a search method which happens candidate feature subsets for analysis supported a particular search strategy. Every candidate set is inserted and related with the previous best one per a particular analysis. If the new set turns to be higher, then it replaces best one. This method is continual till a given stopping condition is satisfied.

Algorithms for include choice fall under 2 wide classes particularly wrappers that utilization the preparation algorithmic lead itself survey the nature of alternatives and channels that assess highlights steady with heuristics based on general characteristics of the data. Long Sheng Chen and Jhih Siang Syu in [4] used Feature extraction

 DOI : 10.18231/2454-9150.2018.0536

based on local and global latent semantic indexing to improve classification performance.

Work is done in two stages, first for feature extraction and second for optimization of dimension size by SVM, which was not suitable for multidimensional problems. In [5], Dr.Saurabh Mukherjee and Neelam Sharma, connected component choice by utilizing three unique strategies specifically, relationship highlight determination, and data pick up and pick up proportion. Feature vitality based reduction method was used which deletes one feature at one time. The results were compared using NB classifier. Features got reduced to some extent, but it took very large time to select features.

In [6], Hee-suChae, Byung-gracious Jo, Sang-Hyun Choi, Twae-kyung Park recommended another element choice strategy that uses the credited normal of aggregate and each class information. The decision tree classifier evaluated with the NSL-KDD dataset to detect attacks. High accuracy was achieved but time taken was more to detect the attack type. Ayman I. Madbouly, Amr M. Gody, Tamer M. Barakat [7], applied four different learning classifiers separately and compares the results. They used best feature reduction method by gradually adding and deleting the feature and finally obtained the reduced set of 11 features out of 41 features. Ayman I. Madbouly, Tamer M. Barakat [8], again applied the same feature subset selection methodology with different approach of using CFS with seven search methods.

They gained 70% reduction in feature subset which was similar to the previous results. Main limitation was the old KDD99 dataset used as it consists of lot of drawbacks.

S.no	Algorithm	Advantage	Disadvantages
1	SVM	-it can be defined by convex optimization problems for which there are efficient methods. -it uses kernel trick -It has a regularization parameter which makes the user think about avoiding over fitting. -it is an approximation to abound on the test error rate. -it able to handle both numerical and categorical variables. -is a highly accurate classifier	-the kernel models are sensitive to over fitting. - It lies in the choice of the kernel. -Speed and size in training and testing. Choosing the kernel is difficult.
2	IBK	-Optimized to noisy training data. - It is effective for large training data. -it handles multi-class levels. -It is easy to distance choices. -it is better to handle numeric variables.	-it needs to determine the number of the nearest neighbour. - It is difficult to distance based learning to determine clearly. -The computation cost is high. The data storage problems. -it is not good for categorical variables to handle it well.
3	J48	-Less search time. -the values generated with fewer tree approaches.	-it requires the pre-processing namely sorting
4	NBTree	-It mitigates the effects of data loss on test attribute selections. -It is better to improve predictive accuracy	-the combined features of the decision tree and naïve Bayes classifier is somewhat complex to implement. -The pre-discrimination may affect test attribute selection and decreases the predictive accuracy.

Table 2.1 Merits & Demerits of Classification

III. IMPLEMENTATION

- **Data Collection**

One of the crucial and first steps in data collection is intrusion detection. In IDS there are two various factors such as data collection and data location plays the major role in the efficacy of the data. To improve the better-fitted security for the targeted network or hosts, this paper proposes IDS based on the host network. This will execute at the nearest victim to the router and observers the internal traffic. At the time of the training, the data collected from the various sources and this is based on the transport/Internet layer protocols and is stamped opposite to the domain knowledge.

- **Data Preprocessing:**

From the first stage when the data has retrieved the features from the KDD Cup 99 dataset. This stage contains 3 elementary stages appeared as takes once. Data exchanging the ready classifier needs every record within the information data to be taken to as a vector of real variety. Here, each representative part is an exceeding dataset which is initially modified into a numerical esteem. As an example, the KDD CUP ninety nine dataset contains numerical as well as additionally representative highlights. These symbolic highlights include some sort of convention like TCP and UDP,profit composes like HTTP and FTP and TCP standing signal like SF, REJ. The strategy simply replaces the extension of the all-out characteristics with numerical qualities.

 DOI : 10.18231/2454-9150.2018.0536

- **Data normalisation:**

Normalization is the process of converting all symbolic attributes into numerical values after the data pre-processing. Data normalization may be a procedure of scaling the estimation of every attribute into a proportional vary that the inclination for highlights with additional distinguished qualities is disposed of from the dataset. Data used is institutionalized. Every component within every record is normalization by its actual greatest esteem and falls into an identical vary. The exchanging and standardization method can likewise be connected to check data. For KDD Cup 99 and to form a correlation with those frameworks that are assessed on numerous forms of assaults, we have a tendency to develop 5 categories. One amongst these categories contains fully the standard records and therefore the different four hold distinctive kinds of assaults (i.e., DoS, Probe, U2R, R2L), separately.

- **Feature selection:**

Despite the very fact that every association during a dataset is spoken to by completely different highlights, not these highlights are expected to build an IDS. Therefore, it's imperative to tell apart the foremost helpful highlights of movement data to accomplish higher execution. within the past space utilizing rule one, associate convertible strategy for the difficulty of highlight determination, FMIFS, is made. In any case, the proposed feature selection algorithms will simply rank highlights concerning their significance bunches with prime request p, let g be a generator of G. A additive guide may be a capability with following properties: Bilinearity, Non-decline, and Computability.

IV. MRMR ALGORITHM

What's more, information smoothing The MRMR is an element determination approach that has a tendency to choose highlights with a high relationship with the class (yield) and a low connection between themselves. For ceaseless highlights, the F-measurement can be utilized to ascertain connection with the class (importance) and the Pearson relationship coefficient can be utilized to compute connection be tween's highlights (excess). From that point, highlights are chosen one by one by applying a covetous hunt to amplify the goal work, which is an element of significance and excess. Two usually utilized sorts of the target work are MID (Mutual Information Difference paradigm) and MIQ (Mutual Information Quotient standard) speaking to distinctionor remainder of importance and repetition, separately. For worldly information, MRMR highlight choice approach requires some pre-processing procedures that level transient information into a solitary framework ahead of time. This may bring about lost perhaps imperative data among transient information, (for example, worldly request data).

Advantages:

1. There are two low-dimensional problems such as relevance and redundancy.
2. Comparatively Speed.
3. Accurate estimation.
4. Best first-order approximation of I (.)
5. Relevance-only ranking only maximizes J (.)!

Algorithm:

Maximum relevance and minimum redundancy feature selection

Input: Discretized data d, class c, and number of feature n, number of features in d is g.

Output: Output feature F.

```
idleft=[1:g]
for i=1 do
    relevance(i) = mutual-info(d(:,i),c);
end for
[R,id]=Max(relevance);
F[1]=id;
Idleft=idleft-F;
for i=2:n do
    Obj=relevance(idleft);
    for j=1 :|idleft|do
        Sum= ∑F k=1(mutual-info(d(:,k),d(:,idleft)));
        Redun(j)=sum/|F|;
end for
end for
```

Fig 4.1: Algorithm for MRMR

Where I (i,j) are the mutual information between ith and jth gene. When we want to retrieve the gene, again we must use MI. The capacities which are discriminant of a gene by the information I (h, gi) is in equation 6. Therefore information is intended h=h1, h2…hk and gi is a measure of relevance of that gene. Then the maximum relevance condition is to maximize the average relevance of all genes in s in equation 6.

$$\text{Maximum } V = 1/|S| \sum_{iEs} I(h,i). \quad (6)$$

Therefore, the redundancy is decreased and relevance of a gene is increased. The two possibilities are grouped inside one function that is in MRMR. As both are crucial, then two criteria's areMax (V-W), and Max (V/W). The MRMR with MID scheme is formulated in equation 7 and MRMR with MIQ is in equation 8.

$$\text{MRMR (MID)} = \max_{iE\Omega s}[I(i,h) - 1/|s| \sum_{jEs} I(i,j)], \quad (7)$$

$$\text{MRMR (MIQ)} = \max_{iE\Omega s}\{I(i,h)/[1/s \sum_{jEs} I(i,j)]\} \quad (8)$$

 DOI : 10.18231/2454-9150.2018.0536

V. RESULTS

Statistics by Class:

	Class: Back	Class: Buffer Overflow	Class: FTPWrite
Sensitivity	0.98824	0.714286	0.250000
Specificity	1.00000	0.999358	1.000000
Pos Pred Value	1.00000	0.833333	1.000000
Neg Pred Value	0.99932	0.998717	0.998082
Prevalence	0.05431	0.004473	0.002556
Detection Rate	0.05367	0.003195	0.000639
Detection Prevalence	0.05367	0.003834	0.000639
Balanced Accuracy	0.99412	0.856822	0.625000

Fig 5.1: Optimized values of back, buffer overflow, FTPWrite

	Class: GuessPassword	Class: Neptune	Class: NMap	Class:Normal
Sensitivity	0.88000	1.00000	0.98765	1.0000
Specificity	1.00000	1.00000	1.00000	0.9783
Pos Pred Value	1.00000	1.00000	1.00000	0.9883
Neg Pred Value	0.99806	1.00000	0.99933	1.0000
Prevalence	0.01597	0.06198	0.05176	0.6473
Detection Rate	0.01406	0.06198	0.05112	0.6473
Detection Prevalence	0.01406	0.06198	0.05112	0.6550
Balanced Accuracy	0.94000	1.00000	0.99383	0.9891

Fig 5.2: Optimized values of guess password, Neptune, NMap, Normal

	Class: Port Sweep	Class: Root kit	Class: Satan	Class: Smurf
Sensitivity	0.97727	0.000000	0.98810	1.00000
Specificity	0.99932	1.000000	0.99932	1.00000
Pos Pred Value	0.98851	NaN	0.98810	1.00000
Neg Pred Value	0.99865	0.998722	0.99932	1.00000
Prevalence	0.05623	0.001278	0.05367	0.05048
Detection Rate	0.05495	0.000000	0.05304	0.05048
Detection Prevalence	0.05559	0.000000	0.05367	0.05048
Balanced Accuracy	0.98830	0.500000	0.99371	1.00000

Fig 5.3: Optimized values of Port Sweep, Root Kit, Satan, Smurf

VI. CONCLUSION

In this paper, the proposed system FBFS methods for temporal gene expression data are implemented. One of the most extreme pertinence and least repetition criteria which were initially presented by the MRMR algorithm.

REFERENCES

[1] M. Blaze, G. Bleumer and M. Strauss, "Divertible Protocols and Atomic Proxy Cryptography", Proc. Advances in Cryptology-EUROCRYPT' 98, Springer, Heidelberg, 1998, pp. 127-144.

[2] A. Boldyreva, M. Fischlin, A. Palacio and B. Warinschi, "A Closer Look at PKI: Security and Efficiency", Proc. PKC 2007 Springer, Heidelberg, 2007, pp. 458-475.

[3] M. Green and G. Ateniese, "Identity-Based Proxy Re-Encryption", Proc. ACNS 2007, Springer, Heidelberg, 2007, pp. 288-306.

[4] T. Matsuo, "Proxy Re-encryption Systems for Identity-Based Encryption", Proc. PAIRING 2007, Springer, Heidelberg, 2007, pp. 247-267.

[5] C.-K. Chu and W.-G. Tzeng, "Identity-Based Proxy Re-encryption without Random Oracles", Proc. ISC 2007, Springer, Heidelberg, 2007, pp. 189-202.

[6] L. Ibrahim, Q. Tang, P. Hartel and W. Jonker, "A Type-and-Identity based Proxy Re-Encryption Scheme and its Application in Healthcare", Proc. SECURE DATA MANAGEMENT 2008, Springer, Heidelberg, 2008, pp. 185-198.

[7] J. Shao, G. Wei, Y. Ling and M. Xie, "Identity-based Conditional Proxy Re-encryption", Proc. IEEE International Conference on Communications (ICC), 2011, pp. 1-5.

[8] Ayman I.Madbouly, Tamer M. Barakat "Relevant Feature Selection Model Using CFS", International Journal of Engineering Trends and Technology (IJETT), 2014, pp. 501-512.

 DOI : 10.18231/2454-9150.2018.0536