Analysis of Feature Selection Techniques: A Data Mining Approach

Sheena
M.Tech Scholar
CSE, SBSSTC
Moga road, Ferozepur

Krishan Kumar
Associate Professor
CSE, SBSSTC
Moga road, Ferozepur

Gulshan Kumar
Assistant Professor
MCA, SBSSTC
Moga road, Ferozepur

ABSTRACT

Feature Selection plays the very important role in Intrusion Detection System. One of the major challenge these days is dealing with large amount of data extracted from the network that needs to be analyzed. Feature Selection helps in selecting the minimum number of features from the number of features that need more computation time, large space, etc. This paper, analyzed different feature selection technique on the NSL-KDD dataset by using C45 classifier, compared these techniques by various performance metrics like classifier accuracy, number of features selected, a list of features selected, elapsed time.

Keywords

Intrusion Detection System, Feature Selection, NSL-KDD, Data Mining, Classification.

1. INTRODUCTION

Due to availability of large amounts of data from the last few decades, the analysis of data becomes more difficult manually. So the data analysis should be done computerized through Data Mining. Data Mining helps in fetching the hidden attributes on the basis of pattern, rules, so on. Data Mining is the only hope for clearing the confusion of patterns. Before applying any mining techniques we need data pre-processing, i.e. data come from different sources may consist of noise, irrelevant attribute, missing data, etc. That needs to be pre-processed before applying any filtering process. Data pre-processing includes four steps: Data Integration that removes inconsistency, Data Cleaning that detects and corrects errors, Discretization that helps to take only a few discrete values and lastly attribute selection that selects only relevant features. Then we use the feature selection technique after the data is pre-processed. Specifically Feature Selection is used in Intrusion Detection System. Basically, the data gathered from the network are a raw data and contains large log files that need to be compressed. So we used the various feature selection techniques for eliminating the irrelevant or redundant features from the dataset.

Intrusion Detection System: Intrusion is basically the anomaly or malicious data. So an IDS is the system or a software is being developed to protect the system from the malicious activities. The IDS needs to analyze the large amount of data that contains the number of features that is quite more time consuming and occupies the large storage space. All the features available in the dataset extracted from the network may not be relevant. So this large data needs to be compressed to reduce the computation time and storage space as well. For this Feature Selection is being used in IDS.

Feature Selection: Data contains many features, but all the features may not be relevant so the feature selection is used so as to eliminate the irrelevant features from the data without much loss of the information. Feature selection is also known as attributes selection or variable selection [7].

The feature selection is of three types:

- Filter approach
- Wrapper approach
- Embedded approach

Filter approach or Filter method shown in Figure 1. It selects the feature without depending upon the type of classifier used. The advantage of this method is that, it is simple and independent of the type of classifier used so feature selection need to be done only once & Drawback of this method is that it ignores the interaction with the classifier, ignores the feature dependencies, and lastly each feature considered separately.

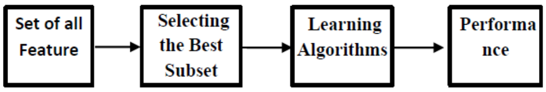

Fig 1: Filter Approach

Wrapper approach or method is shown in Figure 2. The feature is dependent upon the classifier used, i.e. It uses the result of the classifier to determine the goodness of the given feature or attribute. The advantage of this method is that it removes the drawback of the filter method, i.e. It includes the interaction with the classifier and also takes the feature dependencies & Drawback of this method is that it is slower than the filter method because it takes the dependencies also. The quality of the feature is directly measured by the performance of the classifier.

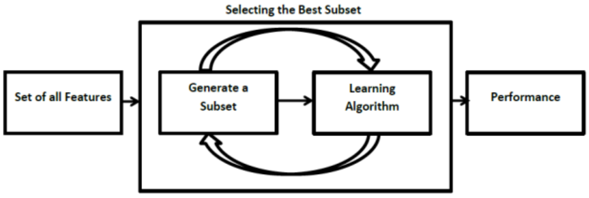

Fig 2: Wrapper Approach

The embedded approach or method shown in Figure 3. It searches for an optimal subset of features that is built into the classifier construction. The advantage of this method is that it is less computationally intensive than a wrapper approach.

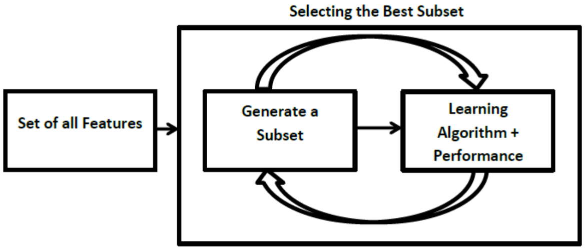

Fig 3: Embedded Approach

2. NEED OF FEATURE SELECTION

- Reduces the size of the problem.
- To improve the classifier by removing the irrelevant features and noise.
- To identify the relevant features for any specific problem.
- To improve the performance of learning algorithm.
- Reduce the requirement of computer storage.
- Reduce the computation time.
- Reduction in features to improve the quality of prediction.

3. RELATED STUDIES

Oreski, D., & Novosel, T. (2014) In this paper, authors have performed the empirical comparison of three feature selection techniques [1]. They have assumed performance differences of different feature selection techniques. Their result has shown that an Information gain technique gives the most accurate subset of features for neural network classification on dataset SPAMBASE. The limitation of this paper is that the classification was done over only first 20 selected features of used techniques, i.e. Relief-F, Information gain and Gain Ratio.

Bart et al. (2014) compared the performance of three different feature selection algorithms Chi-square, Information Gain based and Correlation based with Naive Bayes (NB) and Decision Table Majority Classifier [3]. They also performed DTM classification with CFS. Their results show that significant feature selection can help to design efficient and effective IDS for real world systems. The limitation of this paper is that they have done a classification with all 41 features and also with only 8 selected features, that can be more time consuming and more storage space required.

Kaur, R., Kumar, G., & Kumar, K. (2015) Have done the comparison of feature selection techniques based on various performance metrics like classification accuracy, TPR, FPR, Precision, ROC Area, Kappa Statistic [2]. They selected the best feature selection techniques based on these performance metrics. The limitation of this paper is that they used a less number of instances of the dataset for the experiment.

4. OUR WORK CONTAINING DATASET

4.1 About Dataset

The KDD dataset used for the experiment using KEEL tool which will explain later. Now in this section we will give an introduction about the dataset used.

Since 1999, KDD’99 has been the most widely used data set for the evaluation of anomaly detection methods. This data set is prepared by Stolfo et al. And is built based on the data captured in DARPA’98 IDS evaluation program. 60% data set used in training and 40% data set used in testing purpose. The simulated attacks fall into one of the following four categories:

- Denial of Service Attack (DOS)
- User to Root Attack (U2R)
- Remote to Local Attack (R2L)
- Probing Attack.

4.2 About proposed work

Many researchers have compared feature selection techniques based on their defined performance metrics using WEKA only. In this paper, KEEL tool is used for the comparison. In this paper, the classification of features is not only based on first n selected feature or not on all 41 attributes but on every n selected feature.

Firstly, discretised the test and train data and then imported that dataset into the KEEL and saved the dataset for the experiment. After saving the dataset, \partition it into 10 folds. The data set, then used for applying different feature selection techniques present in the KEEL tool. The C45 classifier used for the experiment. The results extracted by applying different feature selection techniques on the dataset and will choose the best feature selection based on the accuracy. Some of feature selection techniques are Relief-FS, Relief-F-FS, ABB-IEP-FS, ABB-LIU-FS, ABB-MI-FS, Focus-FS, Full-IEP-FS, Full-LIU-FS, Full-MI-FS, SA-IEP-FS, SA-LIU-FS, SA-MI-FS, SBS-IEP-FS, SBS-LIU-FS, SBS-MI-FS & many more.

C45-C

TYPE Classification model by decision trees. To determine a decision tree that on the basis of answers to questions about the input attributes predicts correctly the value of the target attribute. C45 is a decision tree generating algorithm that it induces classification rules in the form of decision trees from a set of given examples. The decision tree is constructed top-down. At each step a test for the actual node is chosen (starting with the root node), which best separates the given examples of classes. C45 is based on ID3 algorithm. The extensions or improvements of ID3 are that it accounts for unavailable or missing values in data, it handled continuous attribute value ranges, it chooses an appropriate attribute selection measure (maximizing gain) and it prunes the result decision trees.

5. EXPERIMENTAL SETUP

The KEEL Tool for the experiment on NSL-KDD dataset. In order to test the classifier, 125973 connection records as a training data set and testing Dataset selected. This section, will describe the metrics used in the experiment methodology, Define the methodology used for Experiment, then how KEEL is being used for the experiment.

5.1 The metrics used

Metrics are basically the standards that are used for the measurement. In this paper, the various Feature Selection techniques being compared using following performance metrics:

- Classification accuracy
- Number of selected features
- List of selected features
- Elapsed time

The best feature selection technique based on the classifier accuracy with minimum number of features selected for the experiment.

5.2 Experimental Methodology

Experiment methodology is shown in Figure 4

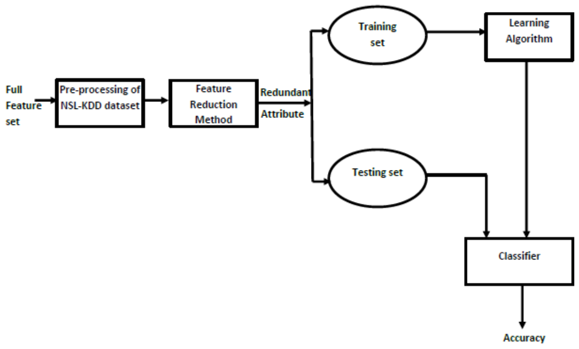

Fig 4: Experimental Methodology

The methodology is as follows:

- Pre-processing of the NSL-KDD data set is done for the input of Feature Reduction method i.e. Feature Selection Technique.
- Applying different feature selection techniques to obtain a reduced feature set.
- Now reduced feature set is used for the preparation of training and test data set.
- Generation of the trained model using C45 classifier with different reduced feature training data set.
- Computational results are based on various performance metrics like classifier accuracy, number of selected features, list of selected features, elapsed time.
- Comparison of different feature selection technique is done on the basis of selected performance metrics.

5.3 Tool Used

KEEL is a software tool developed to build and use different Data Mining models. The interface of KEEL tool. The main features of KEEL are:

- It contains pre-processing algorithms: transformation, discretization, instance selections and feature selections.
- It also contains a Knowledge Extraction Algorithms Library, supervised and unsupervised, remarking the incorporation of multiple evolutionary learning algorithms.
- It has a statistical analysis library to analyze algorithms.
- It contains a user-friendly interface, oriented to the analysis of algorithms.
- This provides a very flexible way for the user to compare its own methods with the ones in KEEL.

5.4 Steps used for Experiment

5.4.1 Data Management:

The data management section brings the operations related to the datasets that are used during the data mining process. Some operations may be the conversion of the dataset files from another dataset format. First import the data into the data management by choosing an input format and then save the data and make partitions of the selected dataset.

5.4.2 Experiment section:

After getting the required data, there comes the experimental section. The experiments section is designed to help a user to create a data mining experiment using a graphical interface. The experiment created can be run on any machine that supports a Java Virtual Machine. After clicking experiment section we can select the type of experiment; Classification is chosen for the experiment.

5.4.3 Datasets:

After selecting the type of experiment, selects the data you want to experiment with. Place the data icon on the white graph plane. In this paper, KDD train data used for the experiment.

5.4.4 Pre-processing method:

After selecting the required data, select the preprocessing method. We selected feature selection method. We applied different feature selection methods on the data set.

5.4.5 Standard methods:

After this step, select the classifier from Data Mining methods of decision tree. C45 classifier is selected.

5.4.6 Visualization modules:

Now to store the result, the Vis-class-check selected for storing the experimental result.

5.4.7 Connections:

Now lastly connect all the modules together with the help of a connection node.

5.4.8 Save and execute the Result:

Now save the experiment as .zip and extract those files and execute on the command prompt using command "java –jar RunKeel.jar". Will get the results after loading all the files correctly and the result will be shown in the result folder in text file format.

6. RESULTS AND DISCUSSIONS

This paper have compared the existing feature selection techniques based on various performance metrics. The Feature Selection Techniques are compared using different performance metrics like classification accuracy, Number of selected features, List of Selected features and Training Time. All the experiments have been evaluated on selected instances from NSL-KDD Dataset using 1 classification algorithm by 10 cross validations. This classifier has been evaluated using KEEL Data mining tool. The best feature selection techniques selected on the basis of Classifier Accuracy with minimum number of features selected. Depending on the requirement, different feature selection techniques that give the best results for different identified metrics. Experimental results indicate that the LVF-IEP-FS gives the best results with 98.79 % accuracy by selecting 15 features. Results are shown in Table 1.

Table 1. Comparison of various Feature Selection Techniques by using C45 classifier

S. No	**Feature Selection Techniques**	**No. Of Selected Feature s**	**List of Selected Features**	**Accuracy**	**Elapsed time**
1	Relief-FS [9]	17	service - src_bytes - dst_bytes - hot - count - srv_count - serror_rate - rerror_rate - diff_srv_rate - dst_host_count - dst_host_srv_count - dst_host_same_srv_ratedst_host_diff_srv_rate dst_host_same_src_port_rate- dst_host_srv_diff_host_rate - dst_host_serror_rate - dst_host_rerror_rate -	99.24%	00:00:02
2	Relief-F-FS [10]	27	Duration - protocol, type - service, flag - src_bytes, dst_bytes - hot - num_failed_login - NUM, compromised - count - srv_count - error, rate - srv_serror_rate - rerror_rate - same_srv_rate - diff_srv_rate - srv_diff_host_rate, dst_host_count - dst_host_srv_count - dst_host_same_srv_rate- dst_host_diff_srv_rate - dst_host_same_src_port_rate - dst_host_srv_diff_host_rate - dst_host_serror_ratedst_host_srv_serror_rate: - dst_host_rerror_rate - dst_host_srv_rerror_rate-	98.10%	00:00:02
3	ABB-IEP-FS [11]	41	All Features	98.14%	00:00:03
4	ABB-LIU-FS [11]	41	All Features	98.99%	00:00:05
5	LVF-FS [12]	18	service - src_bytes - dst_bytes - wrong_fragment - urgent - logged_in - root_shell - num_root - num_shell - num_access_files - count - serror_rate - same_srv_rate - dst_host_srv_count - dst_host_same_srv_rate - dst_host_diff_srv_rate - dst_host_same_src_port_rate - dst_host_srv_serror_rate: -	98.90%	00:00:02
6	**LVF-IEP-FS** [13]	**15**	service - flag - src_bytes - dst_bytes - hot - num_failed_login - num_compromised - num_root - num_outbound_cmds - count - srv_diff_host_rate - dst_host_same_src_port_rate - dst_host_srv_serror_rate: dst_host_rerror_rate - dst_host_srv_rerror_rate -	**98.79%**	00:00:02
7	LVW-FS [14]	22	duration - protocol_type - service - flag - src_bytes - dst_bytes - land - urgent - num_failed_login - logged_in - num_compromised - num_root - num_shell - num_access_files - is_hot_login - is_guest_login - srv_count - srv_serror_rate	98.97%	00:00:02

			- dst_host_count - dst_host_diff_srv_rate - dst_host_rerror_rate - dst_host_srv_rerror_rate -		
8	SBS-IEP-FS [15]	41	All Features	98.99%	00:00:04
9	SBS-LIU-FS [15]	41	All Features	98.98%	00:00:03

7. CONCLUSION

The various Feature Selection techniques are being compared using different performance metrics like Number of features selected, a list of features, Classifier accuracy, elapsed time. All the features present in the dataset gives the less accurate results and needs more computation time and storage space as well. So this paper, have applied different feature selection techniques on the KDD dataset by using single classifier C45. The feature selection helps in reducing the number of features by ignoring the features that are irrelevant or redundant. The best Feature Selection technique selected based on high accuracy with the minimum number of features. In this experiment LVF-IEP-FS is the best technique with an accuracy of 98.79% by selecting 15 features out of 41. The future scope is to apply these feature selection techniques by using the different classifier.

8. REFERENCES

[1] Oreski, D., & Novosel, T. Comparison of Feature Selection Techniques in Knowledge Discovery Process.

[2] Kaur, R., Kumar, G., & Kumar, K. A Comparative Study of Feature Selection Techniques for Intrusion Detection.

[3] Barot, V., Chauhan, S.S., & Patel, B. (2014). Feature Selection for Modeling Intrusion Detection

[4] Shah, B., & Trivedi, B. H. (2015, February). Reducing Features of KDD CUP 1999 Dataset For Anomaly Detection Using Back Propagation Neural Network. In Advanced Computing & Communication Technologies (ACCT), 2015 Fifth International Conference on (pp. 247-251). IEEE.

[5] Derrac, J., et al. "KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework." (2015).

[6] Kumar, K., Kumar, G., & Kumar, Y. (2013). Feature Selection Approach for Intrusion Detection System. International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE), 2(5).

[7] Beniwal, S., & Arora, J. (2012). Classification and feature selection techniques in data mining. International Journal of Engineering Research & Technology (IJERT), 1(6).

[8] Shardlow, Matthew. "An Analysis of Feature Selection Techniques."

[9] Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. InProceedings of the ninth international workshop on Machine learning (pp. 249-256).

[10] Robnik-Šikonja, M., & Kononenko, I. (1997, July). An adaptation of Relief for attribute estimation in regression. In Machine Learning: Proceedings of the Fourteenth International Conference (ICML'97)

[11] H. Liu, L. Yu. Toward Integrating Feature Selection Algorithms for Classification and Clustering. IEEE Transactions on Knowledge and Data Engineering 17:4 (2005) 491-502.

[12] H. Liu, R. Setiono. A Probabilistic Approach to Feature Selection: A Filter Solution. 13th International Conference on Machine Learning (ICML96). Bari (Italy, 1996) 319-327.

[13] H. Liu, L. Yu. Toward Integrating Feature Selection Algorithms for Classification and Clustering. IEEE Transactions on Knowledge and Data Engineering 17:4 (2005) 491-502.

[14] H. Liu, R. Setiono. Feature Selection and Classification: A Probabilistic Wrapper Approach. 9th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA-AIE'96). Fukuoka (Japon, 1996) 419-424.

[15] H. Liu, L. Yu. Toward Integrating Feature Selection Algorithms for Classification and Clustering. IEEE Transactions on Knowledge and Data Engineering 17:4 (2005) 491-502.