International Journal of Future Generation Communication and Networking
Vol. 14, No. 1, (2021), pp.456–461

Comparison of Subset Evaluation Feature Selection Algorithm using KDD Cup'99 Dataset for Mobile Ad-hoc Network

Mr. V.Asaithambi[*1], Dr. N.Rama [2]

[*1]*Research Scholar, PG & Research Department of Computer Science, Presidency College, Chennai*

[2]*Research Supervisor, Former Head, PG & Research Department of Computer Science, Presidency College, Chennai*
[*1]*vsasai@yahoo.com*
[2]*nramabalu@gmail.com*

Abstract

Now a days many researches have been carried out on the feature selection process. This feature selection process is more important for any research, since the research will be using any one of the available dataset. These dataset will have many features and incomputable instances. The data in a dataset is not necessary to be more suited for any research process. The features selection algorithms are extract the suitable features for a particular research from the whole dataset. This article provides such feature selection algorithms which are working based on the correlation and consistency. Also the correlation and consistency based algorithms are applied on the standard dataset called KDD Cup'99 and by which these algorithms are compared with each other. The results are analyzed and the performance of the feature selection algorithms concluded. This conclusion could be used for any intrusion detection system on a mobile ad-hoc networks.

Keywords: *MANET, KDD, correlation-based feature selection.*

1. Introduction

Feature selection is one of the most important process in any research work which uses a dataset with enormous instances and many number of attributes. Without feature selection task the research process will provide inaccurate results and leads to incorrect conclusion. There are many algorithms used for feature selection. These feature selection algorithms use many strategies like random selection, correlation based selection, consistency based selection, information gain based selection, etc. This article provides a brief idea about correlation based subset evaluation and consistency based subset evaluation. A correlation based feature selection algorithm is an algorithm which uses along with a search algorithm. A consistency based feature selection algorithm is also an algorithm which uses along with a search algorithm. In this article, these two subset evaluation algorithms are considered for implementation with the combination of two search algorithms. The implementation is conducted by weka. The results and reports are compared and analyzed to obtain better conclusion.

2. Existing System

There are many algorithms available for feature selection. Based on the dataset used, the feature selection algorithms can be applied. Most of these algorithms are available in weka software for testing with any dataset. This software is used for implementation of any data mining and machine learning algorithms. Many packages available in this software for testing the algorithms. Opeyemi Osanaiye et. al. 2019 has

ISSN: 2233-7857 IJFGCN

designed a feature selection algorithm for intrusion detection system in cluster based wireless sensor. James P Anderson has proposed a concept of IDS in 1980. Senthilnayaki Balakrishnan et. al. 2014 has developed an intrusion detection system using feature selection and classification methods. In which an algorithm has been proposed for optimal feature selection using gain ratio. Also most of the researchers using KDD Cup'99 dataset for the research on intrusion detection.

3. Proposed System

The proposed system is a feature selection system. This system provides a comparative study on the feature selection using correlation based feature selection methods and consistency based feature selection methods. The figure-1 shows the system architecture of the proposed feature selection technique.

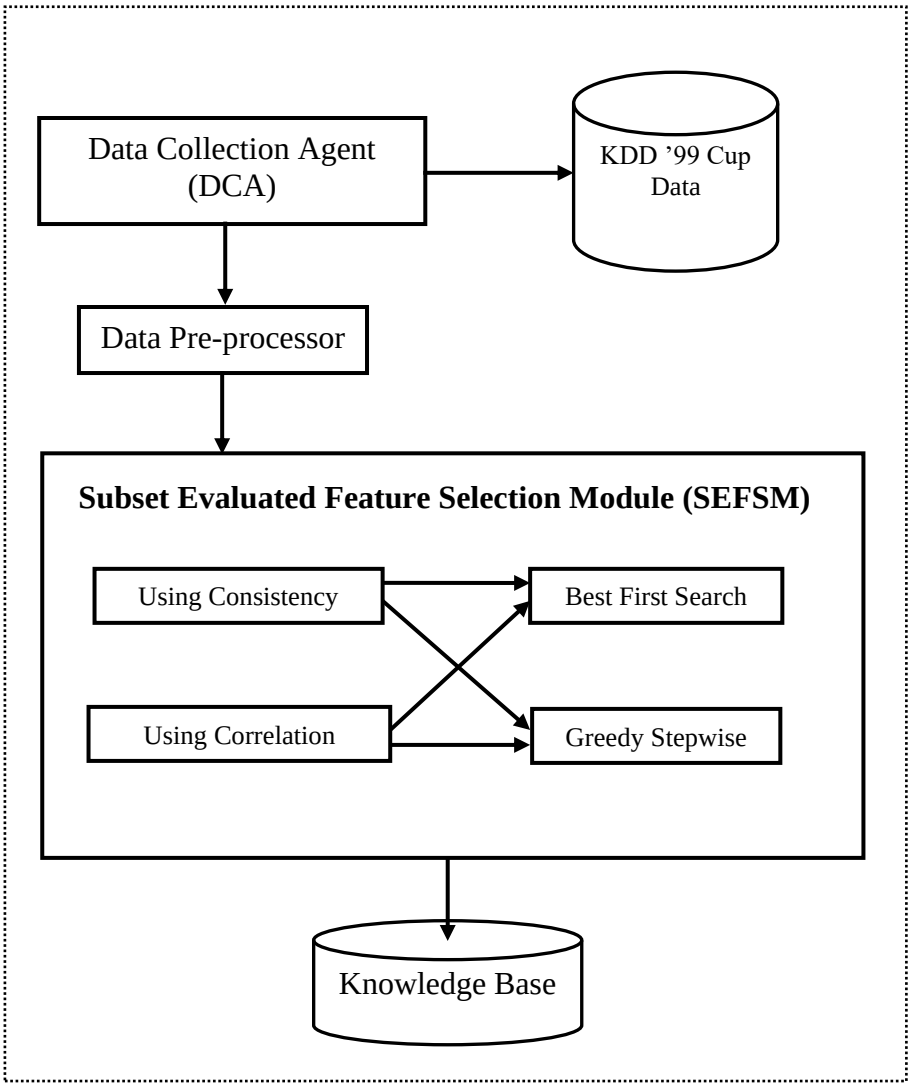

Figure-1: System Architecture

The data collection agent collect the instances of data from the KDD Cup'99 dataset database. The preprocessing unit is used for conducting all preprocessing tasks

ISSN: 2233-7857 IJFGCN

like data cleansing, data validating and enriching the data. The preprocessed data is send to the feature selection module. In which the subset evaluation techniques are used. The correlation based sub set evaluation is executed along with the search algorithms best fit search and greedy step-wise algorithm. The consistency based sub set evaluation is also executed along with the search algorithms best fit search and greedy step-wise algorithm. The various reports are generated using the software weka.

4. Results and Reports

The benchmark dataset KDD Cup'99 has 41 features. All these features are network related features. The table-1 provides the list of 41 features.

Table-1: List of Features Available in KDD Cup'99 Data set

S.No	Feature Name	Description	Type
1.	Duration	length (number of seconds) of the connection	Continuous
2.	Protocol_type	type of the protocol, e.g. tcp, udp, etc.	Discrete
3.	Service	network service on the destination e.g. http, telnet, etc.	Discrete
4.	Src_bytes	number of data bytes from source to destination	Continuous
5.	Dst_bytes	number of data bytes from destination to source	Continuous
6.	Flag	normal or error status of the connection	Discrete
7.	Land	1 if connection is from/to the same host/port; 0 otherwise	Discrete
8.	Wrong_fragment	number of ``wrong'' fragments	Continuous
9.	Urgent	number of urgent packets	Continuous
10.	Hot	number of ``hot'' indicators	Continuous
11.	Num_failed_logins	number of failed login attempts	Continuous
12.	Logged_in	1 if successfully logged in; 0 otherwise	Discrete
13.	Num_compromised	number of ``compromised'' conditions	Continuous
14.	Root_shell	1 if root shell is obtained; 0 otherwise	Discrete
15.	Su_attempted	1 if ``su root'' command attempted; 0 otherwise	Discrete
16.	Num_root	number of ``root'' accesses	Continuous
17.	Num_file_creations	number of file creation operations	Continuous
18.	Num_shells	number of shell prompts	Continuous
19.	Num_access_files	number of operations on access control files	Continuous
20.	Num_outbound_cmds	number of outbound commands in an ftp session	Continuous
21.	Is_host_login	1 if the login belongs to the ``host'' list; 0 otherwise	Discrete
22.	Is_guest_login	1 if the login is a ``guest''login; 0 otherwise	Discrete
23.	Count	number of connections to the same host as the current connection in the past two	Continuous

ISSN: 2233-7857 IJFGCN

		seconds	
24.	Serror_rate	% of connections that have ``SYN'' errors	Continuous
25.	Rerror_rate	% of connections that have ``REJ'' errors	Continuous
26.	Same_srv_rate	% of connections to the same service	Continuous
27.	Diff_srv_rate	% of connections to different services	Continuous
28.	Srv_count	number of connections to the same service as the current connection in the past two seconds	Continuous
29.	Srv_serror_rate	% of connections that have ``SYN'' errors	Continuous
30.	Srv_rerror_rate	% of connections that have ``REJ'' errors	Continuous
31.	Srv_diff_host_rate	% of connections to different hosts	Continuous
32.	Dst_host_count	count of connections having the same destination host	Continuous
33.	Dst_host_srv_count	count of connections having the same destination host and using the same service	Continuous
34.	Dst_host_same_srv_rate	% of connections having the same destination host and using the same service	Continuous
35.	Dst_host_diff_srv_rate	% of different services on the current host	Continuous
36.	Dst_host_same_src_port_rate	% of connections to the current host having the same src port	Continuous
37.	Dst_host_srv_diff_host_rate	% of connections to the same service coming from different hosts	Continuous
38.	Dst_host_serror_rate	% of connections to the current host that have an S0 error	Continuous
39.	Dst_host_srv_serror_rate	% of connections to the current host and specified service that have an S0error	Continuous
40.	Dst_host_rerror_rate	% of connections to the current host that have an RST error	Continuous
41.	Dst_host_srv_rerror_rate	% of connections to the current host and specified service that have an RST error	Continuous

From the above 41 features, some of the features are selected by using the feature selection algorithms evaluated by the subsets. The table-2 shows the list of features extracted from these 41 features by applying two subset evaluation feature selection algorithms with the combination two search algorithms.

For both the subset evaluation algorithms the best fit search is selecting the subset with the more number of features than the number of features using greedy step-wise search algorithm. Also based on the requirement of the research any one of the combination of the algorithms shown in the table-2 can be applied. The common features selected from these four subsets will be very less number of features. So the hybrid version of these algorithms can be used only when it is essential. Otherwise the hybrid method of algorithm is not advisable.

ISSN: 2233-7857 IJFGCN

Table-2: Comparison of Feature Selection using subset evaluation

Sl.No	Feature Selection Algorithm	Search Algorithm	Number of Features Selected	List of Features Selected
1	Consistency Based	Best Fit	7	service, flag, src_bytes, lnum_root, dst_host_srv_count, dst_host_same_srv_rate, dst_host_same_src_port_rate
2	Consistency Based	Greedy Stepwise	6	service, flag, src_bytes, hot, dst_host_same_srv_rate, dst_host_same_src_port_rate
3	Correlation Based	Best Fit	9	protocol_type, src_bytes, wrong_fragment, hot, logged_in, lnum_root, lnum_access_files, dst_host_srv_diff_host_rate, dst_host_srv_serror_rate
4	Correlation Based	Greedy Stepwise	8	protocol_type, src_bytes, wrong_fragment, hot, lnum_root, lnum_access_files, dst_host_srv_diff_host_rate, dst_host_srv_serror_rate

The figure-2 shows the pictorial representation of the results. It provides the performance of the feature selection algorithm in four combination. Depending on the requirement of the research work carried out, the combination of the algorithm can be selected.

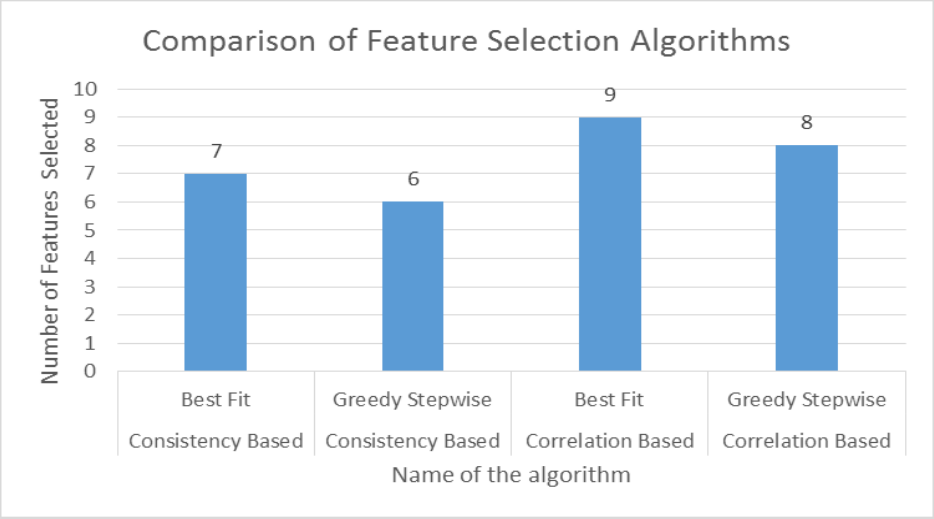

Figure-2: Comparison of Feature Selection Algorithms

ISSN: 2233-7857 IJFGCN

4. Conclusion

Even though there are many feature selection algorithms available, it is mandatory to select a perfect feature selection algorithm for any research work. As per the focus of this article, the two feature selection algorithms using subset evaluation are applied on the dataset KDD Cup'99. Also the list of features are selected and analyzed. The feature selection will be used for any research work to obtain the optimum result. But the suitable feature selection algorithm should be used for selecting the features to get the best result. Provided the hybrid feature selection algorithm is very much useful for the research work which requires the input with multiple class of characteristics.

References

[1] Ebenezer Popoola,Aderemi Adewumi "Efficient Feature Selection Technique for Network Intrusion Detection System Using Discrete Differential Evolution and Decision Tree", international Journal of Network Security, vol. 19, no. 5, (2017), pp. 660-669.

[2] Senthilnayaki balakrishnan, Venkadalakshmi K, Kannan A "Intrusion Detyection System using Feature Selection and Classification Technique", international Journal of Computer Science and Applications(IJCSA), vol. 3, issue. 4, (2013**)**, pp. 145-151.

[3] J. P. Anderson. "Computer security threat monitoring and surveillance", Technical report, James P. Anderson Company, Fort Washington, Pennsylvania, April 1980.

[4] Dorothy E. Denning. "An intrusion-detection model", IEEE Trans. Software Eng., 1987.

[5] Chen, Y., Abraham, A. and Yang, B. "Feature selection and classification using flexible neural tree",Neurocomputing, Vol. 70, pp. 305-313, 2006.

[6] Denning, D.E. "An intrusion-detection model", IEEE Transaction on Software Engineering, Vol.13, pp. 222-232, 1987.

[7] Jain, A. and Zongker, D. "Feature selection: Evaluation, application, and small sample performance", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, pp. 153-158, 1997.

[8] Liu, H. and Yu, L. "Toward integrating feature selection algorithms for classification and clustering", IEEE Transactions on Knowledge and Data Engineering, Vol. 17, No. 491-502, 2005.

[9] Srilatha Chebrolu, Ajith Abraham, Johnson P. Thomas, "Feature Deduction and Ensemble Design of Intrusion Detection Systems", Computers & Security, Vol. 24, pp. 295-307, 2005.

[10] Verikas, A. and Bacauskiene M, "Feature Selection with Neural Networks", Pattern Recognition Letters, Elsevier, Vol.23, pp. 1323-1335, 2002.

[11] Guyon, I., Gunn, S., Nikravesh, M., & Zadeh, L., "Feature extraction, foundations and applications", Springer-German Verlag, 2006.

ISSN: 2233-7857 IJFGCN