International Journal of Computer Science and Informatics

Volume 3 | Issue 2 Article 12

October 2013

A FUZZY BASED DIVIDE AND CONQUER ALGORITHM FOR FEATURE SELECTION IN KDD INTRUSION DETECTION DATASET

ANISH DAS
Department of Computer Science, Pondicherry University, Pondicherry, India, anishdas09@gmail.com

S. SIVA SATHYA
Department of Computer Science, Pondicherry University, Pondicherry, India, ssivasathya@gmail.com

A FUZZY BASED DIVIDE AND CONQUER ALGORITHM FOR FEATURE SELECTION IN KDD INTRUSION DETECTION DATASET

ANISH DAS[1] & S. SIVA SATHYA[2]

[1,2]Department of Computer Science, Pondicherry University, Pondicherry, India
E-mail : {ssivasathya, anishdas09}@gmail.com

Abstract - This paper provides a fuzzy logic based divide and conquer algorithm for feature selection and reduction among large feature set of KDD intrusion detection data set, since a reduced feature set will help to evolve better mining rules.This algorithm introduces a fuzzy idea of dividing the normal record by attacks records or vice-versa, and then considers the feature sets for every attack type separately. Actually, this algorithm is applied on KDD CUP 99 dataset having 37 attack types and selecting important feature among 41 feature of KDD dataset. The selected features are used in TANAGRA [11, 12] data mining tool to classify the dataset (i.e. KDD 99) for every attack vs. normal using various classification algorithms [5, 6]. The result for feature selection and classification shows a reduced set and maximized classification rate respectively.

Keywords *- KDD CUP 99 intrusion data set, TANAGRA data mining Tool, intrusion detection, classification, divide-conquer with fuzzy based.*

I. INTRODUCTION

Performance of an intrusion detection system generally depends on the length of dataset having a lot of features used to identify attack or normal type. So, the reduced dataset with relevant features help an IDS [13] to perform its' task easily with a better use of memory and time. During the process of the reduction of data or features, the main work is to find the hidden relationship among data or features. For finding out the hidden relationship among data or features, fuzzy logic based reduction approaches can be used. In this paper, a fuzzy based divide conquer approach is used to find the relationship.

An IDS [13] is responsible to provide a better security for a system from an attack. But it fails sometime to identify the different type of attack, due to the dynamic behavior of attack types. So, for dynamic behavior of attacks, the KDD 99 [3] dataset provide records of data with 41 features [10] of 37 types of attack with their dynamic behavior. KDD dataset for any security work is very useful to gain information for attacks and normal type.

This paper has also found a hidden relation among the dataset of KDD 99 to select the important feature set using a fuzzy based divide conquer approach. It is found that this approach reduced the features and maximized classification rate and minimized misclassification rate using various classification algorithms [5, 6] in TANAGRA.

The rest of this paper is kept as follows: section 2, describing about related work, section 3, giving fuzzy based divide conquer algorithm, section 4, providing fuzzy based experimental results, section 5, concludes about this work.

II. RELATED WORK

Wanli Ma [14] in his paper described a technique for feature selection by grouping the features in the four categories: Group I contain the basic network traffic features; Group II is not network traffic related, but the features collected from hosts; Group III and IV are temporally aggregated features. In this paper [15] Adetunmbi A.Olusola and others proposed an algorithm for selection of relevance feature set of KDD intrusion detection dataset using rough set degree of dependency and dependency ratio of each class employed to determine the most discriminant features for each class. In 2009 Shailendra Singh and Sanjay Silakari in their paper [16] designed a feature selection algorithm using filter phase for highest information gaining and guides the initialization of search process for wrapper phase and finally feature subsets are passed through the K-nearest neighbor classifier for classification of attacks. In paper [17], Gulshan Kumar and Krishan Kumar proposed feature selection methods for (1) reduction in number of features; (2) performance of Naïve Bayes classification model trained on reduced set of features. Iago Porto-D´ıaz and others describes [18] method consists of a combination between feature selection methods and a novel local classification method. This classification method –called FVQIT (Frontier Vector Quantization using Information Theory)– uses a modified clustering algorithm to split up the feature space into several local models, in each of which the classification task is performed independently.

III. FUZZY BASED DIVIDE AND CONQUER ALGORITHM

The proposed algorithm is named as divide and conquer as it divides the normal record by attack record or vice-versa depending on the length of the records i.e. greater length record would be divided by lesser length record, then important features are selected taking deviation for every feature set using the result of matching values for every feature of attack with respect to normal. So, this algorithm selects the important feature set based on the non-similarity of attack vs. normal using the concept of deviation. Initially, the KDD dataset is preprocessed as follows.

(a) Divide the whole KDD data set into attack vs. normal records.

(b) Separate each attack and normal records in KDD 99 dataset, so that total number of files should be 38 (i.e. one normal record + 37 attack record).

Algorithm

1. Input the file containing normal records and another file containing attack records.
2. Calculate or find the greater length among the two input file (i.e. normal and attack file). Let Solution [] be a vector that will contain selected features, Mismatch [] will contain number of mismatch, deviation [] will have value for every set of mismatch count and min_support will be initialized before execution of algorithm as a threshold.
3. Divide the greater length file by using the size of the lesser length file, i.e. if the greater length file is G and lesser length file is L, then divide G by L resulting in $G_1,G_2,\ldots,G_n$ each of length L. Here each division G_d or file L have 41 features and they are represented as G_1 ($G_{d1},G_{d2},\ldots G_{d41}$), $G_2(G'_{d1},G'_{d2},\ldots,G'_{d41}),\ldots,G_n$ ($G'''_{d1},G'''_{d2},\ldots,G'''_{d41}$) and $L_1,L_2,L_3,\ldots,L_{41}$ respectively.
4. Perform One-to-one comparison with every attribute value of each record in G_d and L, i.e. $L(L_1,L_2,L_3,\ldots,L_{41})$ to G_1 ($G_{d1},G_{d2},\ldots,G_{d41}$), $L(L_1,L_2,L_3,\ldots,L_{41})$ to G_2 ($G'_{d1},G'_{d2},\ldots,G'_{d41}$),…, and $L(L_1,L_2,L_3,\ldots,L_{41})$ to G_n ($G'''_{d1},G'''_{d2},\ldots,G'''_{d41}$).

   For i←1 to size of lesser file L do
   If ($L(L_i) \neq G_i(G_{di})$) then ,do
   Mismatch[i]←Mismatch[i] + 1;
   End if
   End for
5. Calculate deviation of Mismatch [] for every features which we got from step 4 and keep it in deviation [] array.
6. Considering or keeping the features in the Solution [] array, should obey the following consideration:-
   1. Value of deviation [] > min_support and deviation []<= 1and Minimum value of Mismatch [] – deviation [] > min_support.
   2. Value of deviation []=minimum value of Mismatch[].
   3. Value of deviation [] =0.
7. Repeat step 4 until all records have finished their one to one corresponding comparison.
8. Display the features in Solution [] which are taken using the step 6 consideration.
9. End.

IV. EXPERIMENTAL ANALYSIS AND RESULT

The result of fuzzy based divide and conquer algorithm give the important features for every attack with respect to the normal. The result is analyzed in two stages:

- Using the fuzzy based divide conquer algorithm to the KDD dataset to select the important and reduced feature set.
- Classifying the KDD records into attack and normal records using the reduced feature set with the help of various classification algorithms [5, 6] in Tanagra.

4.1 Feature Reduction Using Fuzzy Based Divide and Conquer Algorithm:

The reduced feature set for the 37 attacks in KDD 99 dataset are selected using proposed fuzzy based divide and conquer algorithm, given in Table1.

Table1. List of selected features using fuzzy based divide conquer algorithm.

Attack Name	Selected Feature using Fuzzy based divide and conquer algorithm
Apache2	8,11,13,14,15,16,17
Back	8,11,14,15,16
Buffer Overflow	4,25,26,27,28
Ftp write	23,24,39
Guess passwd	7,9,11,15,18,20,21
Http tunnel	8,17,19
Imap	12,23,24,31,32,36,37
Ip Sweep	8,13,14,16,17
Land	27,28
Load Module	23,24,31,37
Mailbomb	7,9,15,18,20,21
Mscan	8,11,13,14,15,16
Multihop	4,25,26,27,28
Named	25,26
Neptune	7,9,18,20,21,30,40
Nmap	19,28

Perl	23,24,31
Phf	23,24,31,37
Pod	4,19,27,28
Portsweep	8,13,14,16,17
Process Table	8,11,13,14,15,16,17
Ps	4,25,26,27,28
Root Kit	4,25,26
Saint	8,11,13,14,15,16,17
Satan	13,15,16
Sendmail	25,26
Smurf	5,7,9,18,20,21
Snmpget	7,9,15,18,20,21
Snmpguess	7,9,15,16,18,20,21
Sql	23,24,31,37
Teardrop	25,26,39
Udpstorm	2,3,12,23,24,31,37
Warez master	13,15
Worm	23,24,31,37
Xlock	25,26
Xsnoop	24,38,40
Xterm	4,25,26

4.2 Classification in Tanagra:

The selected feature set acquired from fuzzy based divide and conquer algorithm are utilized to classify the KDD records in TANAGRA [2, 12]. In TANAGRA various classification algorithms like KNN, Naïve Bayes, ID3, C4.5, SVM, and LDA [5, 6] are used to classify and results are compared with discriminant based algorithm [1] given in Table2. From the Table2 it is found that the misclassification rate of the proposed algorithm is lesser than the discriminant based techniques.

Table2 List of classification and misclassification rate for fuzzy based divide and conquer and Discriminant analysis based algorithm.

	Fuzzy Based Divide Conquer Algorithm		Discriminant Analysis Based Algorithm	
Attack Name	Classification rate (%)	Misclassification rate (%)	Classification rate (%)	Misclassification rate (%)
Apache2	99.96	0.04	99.7	0.3
Back	99.94	0.06	99.4	0.6
Buffer Overflow	99.99	0.01	68.2	31.8
Ftp write	99.99	0.01	33.3	66.7
Guess passwd	99.83	0.17	100	0
Http tunnel	99.99	0.01	96.8	3.2
Imap	99.99	0.01	100	0
Ip Sweep	99.98	0.02	97.1	2.9
Land	99.99	0.01	100	0
Load Module	99.99	0.01	100	0
Mailbomb	99.80	0.20	100	0
Mscan	99.95	0.05	95.7	4.3
Multihop	99.99	0.01	22.2	77.8
Named	99.99	0.01	17.6	82.4
Neptune	99.91	0.09	100	0
Nmap	99.99	0.01	100	0
Perl	99.99	0.01	100	0
Phf	99.99	0.01	100	0
Pod	99.99	0.01	100	0
Portsweep	99.98	0.02	100	0
Process Table	99.96	0.04	97.2	2.8
Ps	99.99	0.01	31.3	68.8
Root Kit	99.99	0.01	23.1	76.9
Saint	99.96	0.04	81.9	18.1
Satan	99.92	0.08	99.8	0.2
Sendmail	99.99	0.01	35.3	64.7
Smurf	99.78	0.22	100	0
Snmpget	99.72	0.28	100	0
Snmpguess	99.89	0.11	100	0
Sql	99.99	0.01	100	0
Teardrop	99.99	0.01	66.7	33.3
Udpstorm	99.99	0.01	50	50
Warez master	99.93	0.07	95.2	4.8
Worm	99.99	0.01	100	0
Xlock	99.99	0.01	33.3	66.7
Xsnoop	99.99	0.01	50	50
Xterm	99.99	0.01	76.9	23.1

V. CONCLUSION

Thus this paper describes the proposed fuzzy based divide and conquers algorithm and how it has been applied to KDD 99 dataset. The reduced feature set thus obtained yields maximized classification rate in Tanagra tool. The fuzzy based divide conquer technique used here easily finds the relationship among dataset and features. This algorithm is simple, flexible, and generic and could be used for any dataset apart from KDD dataset. The results of the classification algorithms in TANAGRA shows that the proposed approach produces very less misclassification rate i.e. 0.01%. The generated results are better than other existing algorithms in literature.

VI. ACKNOWLEDGEMENT

This work is a part of the AICTE funded project titled „Bio-inspired Intrusion Response System through feature relevance Analysis on Attack Classification□, Under the Research Promotion Scheme (RPS), Ref. No: 8023/BOR/RID/RPS-59/2009-10.

REFERENCES

[1] S.Sivasathya, R. Geetha Ramani and K. Sivaselvi: "Discriminant Analysis Based Feature Selection in KDD intrusion Dataset": International Journal of Computer Applications (0975-8887) volume 31-No.11, October 2011.

[2] Data Mining Tutorial and guiding available on: http://data-mining-tutorials.blogspot.in.

[3] KDD Cup 1999 dataset available on: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

[4] M. Tavallaee, E. Bagheri, Wei Lu, and Ali A. Ghorbani:"A Detailed Analysis of the KDD CUP 99 Data Set": http://ebagheri.athabascau.ca/ papers/cisda.pdf.

[5] KNN, SVM, ID3, C4.5 and Naïve Bayes algorithms are available on: http://en.wikipedia.org/wiki/{K-nearest_neighbor _algorithm, Support_vector_machine, ID3_algorithm, C4.5_algorithm, Naive_Bayes_classifier}.

[6] Linear Discriminant Analysis available on: http://www.music.mcgill.ca/~ich/classes/mumt611_07/.../lda_theory.pdf

[7] N. Araújo, Ruy de Oliveira and others; "Identifying Important Characteristics in the KDD99 Intrusion Detection Dataset by Feature Selection using a Hybrid Approach": http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=05478852

[8] Feature Selection and Conversion Methods in KDD Cup 99 Dataset: A Comparison of Performance: http://www.actapress.com/Abstract.aspx?paperId=37685.

[9] "Feature selection and classification in multiple class datasets: An application to KDD Cup 99dataset": http://www.sciencedirect.com/science/article/pii/S0957417410012650.

[10] KDD dataset features and details available on: http://kdd.ics.uci.edu/databases/kddcup99/task.html.

[11] Details of Tanagra Data Mining Tool available on: http://eric.univ-lyon2.fr/~ricco/tanagra/ en/tanagra.html.

[12] How to use Tanagra available on: http://data-mining-tutorials.blogspot.in.

[13] About IDS available on: http://en.wikipedia.org/ wiki/Intrusion_detection_system.

[14] Wanli Ma; "A study on the feature selection of network traffic for intrusion detection purpose": http://Fieeexplore.ieee.org/Fstamp/Fstamp.jsp/

[15] Adetunmbi A.Olusola., Adeola S.Oladele. and Daramola O.Abosede: "Analysis of KDD '99 Intrusion Detection Dataset for Selection of Relevance Features": http://www.iaeng.org/publication/WCECS2010/WCECS2010_pp162-168.pdf

[16] Shailendra Singh and Sanjay Silakari; "An ensemble approach for feature selection of Cyber Attack Dataset": published on (IJCSIS) International Journal of Computer Science and Information Security, Vol. 6, No. 2, 2009: http://arxiv.org/ftp/arxiv/papers/0912/0912.1014.pdf

[17] Gulshan Kumar and Krishan Kumar; "An information theoretic approach for feature selection": http://onlinelibrary.wiley.com/ doi/10.1002/sec.303/abstract

[18] Iago Porto-D´ıaz, David Mart´ınez-Rego, Amparo Alonso-Betanzos and Oscar Fontenla-Romero; "Combining Feature Selection and Local Modelling in the KDD Cup 99 Dataset": http://www.springerlink.com/content/134547jq51278313/fulltext.pdf

◈◈◈