Review on Meta Classification Algorithms using WEKA

[1]Rausheen Bal, [2]Sangeeta Sharma
[1]*Student of M-Tech and Computer Science & Engineering, Lovely Professional University Phagwara, Punjab, India*
[2]*Faculty of Computer Science & Engineering Department, Lovely Professional University Phagwara, Punjab, India*

Abstract*—*This paper is having a comparative review on different classifiers used for prediction of attack risks on environment having network. In total there are 19 classifiers explained in this paper and the three best or efficient classifiers have been evaluated by three different authors as mentioned in this paper. The data of those three authors has been used in this paper for doing comparison between different classification algorithms. Comparison are taken on the fields of TP-Rate, FP-Rate, Precision, Recall, F-measure etc. Anlaysis was done by those mentioned authors on WEKA tool .

Keywords—*Classification Algorithms; Intrusion Detection System; Meta Classifier; Decision Trees; Machine Learning; Data Mining; WEKA*

1. INTRODUCTION

Intrusion is categorised as any set of operation that can bargain the confidentiality,availability and integrity of system resources. Intrusion Detections are of two types :

- Anomaly Detection
- Misuse Detection.

Anomaly Detection indicates the prediction of various new patterns in data for intrusion. Misuse Detection indicates the classification of the previously recognised intrusion patterns in the data set. Intrusion Detection System is a machine that is positioned inside a secured network to supervise what has appeared within the network.The main ambition of IDS is:

- To resolve the true attacks from false alarms.
- To observe abnormal network behaviour or wastage of resources precisely.
- To inform the network administrators about the activity taken place.

Data Mining is a technique of intrusion detection which has a significant application area to examine the massive volume of inspected data and realizing performance for the enhancement of detection rules.

Developing technologies and online social networks have become abundant for new generation.Number of operators are growing quickly on online social web.The huge amount of online social web are fascinating the green-eyed users for several purposes like deframing,data-theft,spams,snooping etc due to this it has become important for Internet Service providers to identify the unseen relationships in online social web.To separate malicious nodes from authentic nodes it is important to discover the communication relationships of the users.

2. DATA MINING TAXONOMY

Data Mining is the evaluation stage of the "knowledge discovery in databases" process (KDD). It is the discovery procedure of important non-spontaneous correlations and patterns composing possibly to fetch high-level knowledge information from low-level data. It is relevant method of detecting effective and unusual useful and logical patterns of data. Data Mining includes evaluation and likelihoods.

Taxonomy is the data mining task which is also known as classification that maps the data into already defined clusters and classes. It is also referred as supervised learning.

There are two steps in data mining:

- Model formation
- Model usage

Model Formation: It comprises the set of predefined classes and every single tuple is assumed to fit in predefined class. Training Set is a set of tuple consumed for model formation.

Model usage: It is the second stage consumed for classifying upcoming or unfamiliar objects.

3. CLASSIFICATION ALGORITHMS

Classifiers such as OneR, BayesNet, Meta-Bagging, ZeroR, IBk, Random Forest, Adaboost, Simple cart, Naïve Bayes, J48, Adaboost.M1, Attribute selected classifier, Filtered classifier, Logiboost, Multiclass classifier, Bagging, classification via regression, REP-tree, Naïve Bayes Multinomial Updateable, Complement Naïve Bayes, Classification Via Clustering are to be discussed as follows:

3.1 ONER Classifier

- OneR classifier is known for "One Rule" is not complex, prior to accurate classification algorithm.
- It produce one rule for every single analyst in the data and after that it choose the rule with the smallest total error as its one rule.
- Art classification algorithm produces rules marginally more accurate than rules produced by OneR but OneR rules are simple for humans to interpret.

3.2 BAYES NET Classifier

- It is a probabilistic graphical model that signifies a set of arbitrary variables as well as their provisional dependencies via DAG (Directed acyclic graph)
- The other names of BayesNet classifier are Bayesian network, Bayes Network, Bayesian Model and Probabilistic directed acyclic graphical model.

3.3 META-BAGGING Classifier

- Meta-Bagging is also known as only Bagging also. Bagging is defined as bootstrap aggregation.
- Bagging produces bootstrap samples of the training data.
- It constructs the unique training set consists of numerous data sets.
- Numerous data sets are constructed by unsystematic sampling occurrences with replacement.
- Each single bootstrap sample is used for training a regression function or a classifier.
- Classification results are taken on maximum number of votes for classification purpose.
- For regression average of expected values are taken.

Advantage of bagging:

- Variation is reduced and performance is improved for unsteady classifiers which differ meaningfully with small changes in the dataset.

3.4 ZEROR Classifier

- It disregards all the predictors and only relies on the target.
- ZeroR Classifier is the classification method which is simplest of the all.

3.4 IBk Classifier

- K-Nearest Neighbor (k-NN) is a pattern based learning or lazy learning.
- It is the one of the simplest algorithms of machine learning.
- An item is organized by maximum number of votes of its neighbors with the item being allocated to the class that is known among its k-NN.
- If k is assigned value 1 then the item is directly assigned to the class of its nearest neighbor.

3.5 ADABOOST Classifier

- Boosting as well as Bagging are the meta algorithms that collects the decision from various classifiers.
- This algorithm repetitiously determine from weak classifiers.
- Weighted summation of the outcome of weak classifier is considered as final result.

3.6 NAIVE BAYES Classifier

- It is based on theorem named as bayes theorem with autonomous assumptions between analysts .
- Naïve Bayesian model is simple to construct with no complex repetitive parameter estimation which make it predominantly useful for huge amount of datasets.
- It is widely used as it often leave behind more refined classification methods and yet it is simple too.

3.7 J48 Classifier

- It is somewhat improved C4.5 in WEKA .
- C4.5 algorithm produces a classification-decision tree for specified set of data by iterative partitioning of data.
- Depth first approach is used for decision growth.
- The algorithm reflects all the likely tests that can divide the data set and chooses a test that gives the finest data gain.
- There are two attributes: discrete and continuous
- For discrete attribute: one test with results as countless as the amount of dissimilar values of the attribute is ruminated.

- For continuous attribute: two tests relating each and every dissimilar values of the attribute is ruminated.
- The instruction to collect the entropy gain of each and every two tests competence the training data set that is appropriate to the node in thought is arranged for the values of the attributes that are continuous. The process continues for each continuous attribute.

3.8 SIMPLE CART Classifier

- Binary decision tree is produced by the technique of classification known as classification and regression tree as well as simple cart classifier.
- Meanwhile outcome is binary tree, it produces two offspring only.
- The usage of entropy is to select the best separating attribute.
- Simple cart disregards that record which is missing.

3.9 RANDOM FOREST Classifier

- This is the collaborative learning technique for classification , regression and other activities that functions by creating a gathering of decision trees at training time and producing the class that is the method of the classes known as classification or mean prediction known as regression for every single tree.

3.10 ADABOOST.M1 Classifier

- Adaboost.M1 is broadly implemented boosting algorithm and is well known because it is used for boosting a multiclass base classifier as if there is problem in a multiclass classification.
- Though it does not work if the base classifier is too weak but it can be made usable by doing modification of adaboost.M1 in one line only.

3.11 ATTRIBUTE SELECTED Classifier

- The range of the training data and testing data is lessened by this algorithm before being departed onto the classifier. Presently, base classifiers are used.
- Since,the classifier is raised various search approaches are used during the phase of attribute selection.

3.12 CLASSIFICATION VIA REGRESSION

- Regression approaches are applied for classification under this classifier.
- Single regression model is constructed for every single value of the class.

3.13 FILTERED Classifier

- Retaining the architecture of the training and testing data similar this classifier is used with numerous type of filters.

3.14 LOGIBOOST Classifier

- This classifier is the continuation of the Adaboost algorithm as it substitutes the exponential loss of Adaboost Algorithm to provisional Bernoulli probability loss.
- The usage for this class is for accomplishment of additive logistic regression.
- This classifier practices on regression structure as the base learner as well as handles problems of multi class.

3.15 MULTICLASS Classifier

- Error rectification codes are adapted with this classifier for attaining for more accuracy as this classifier is used for classifying occurrences additional to two classes.

3.16 REP-TREE Algorithm

- It uses the logic of regression tree and produces numerous trees in several iterations.
- Subsequently it chooses the finest one from all the produced trees and will be reflected as demonstrative.
- The measure used in this algorithm is MEAN-SQUARE-ERROR on the calculations made by tree.
- Ultimately REP-Tree is speedy decision tree learning and it constructs a decision tree established on the information gain as well as reducing the variance.
- Decision trees are created by using the data of information gain and pruning is done by REP. It only arranges the mathematical elements single time and incomplete values are handled using C4.5 algorithm using fractionary occurrences.

3.17 COMPLEMENT NAÏVE BAYES Algorithm

- The Compliment Naive Bayes (CNB) classifier recovers upon the limitations of the Naive Bayes classifier by approximating factors from data in all outlook classes excluding the one which we are going to do execution.

3.18 NAÏVE-BAYES-MULTINOMIAL UPDATABLE

- The term Multinomial Naive Bayes just makes us know that each p(fi|c)p(fi|c) is a multinomial distribution, rather than some other distribution.
- This works well for data which can simply be revolved into counts, such as word counts in text.

3.19 CLASSIFICATION VIA CLUSTERING

- A simple classifier that uses a clusters for classification.
- For cluster algorithms that use a static amount of clusters, like Simple-K-Means, the user has to make it sure that the amount of clusters to produce are the similar as the amount of class labels in the dataset in order to attain a useful model.

4. INTRODUCTION TO WEKA TOOL

Weka is a compilation of machine learning algorithms for tasks used in data mining. The algorithms can be applied in two ways i.e

i. directly to a dataset

ii. or called from your own Java code.

Weka consists of tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also compatible for evolving new machine learning schemes.We can locate Weka on :

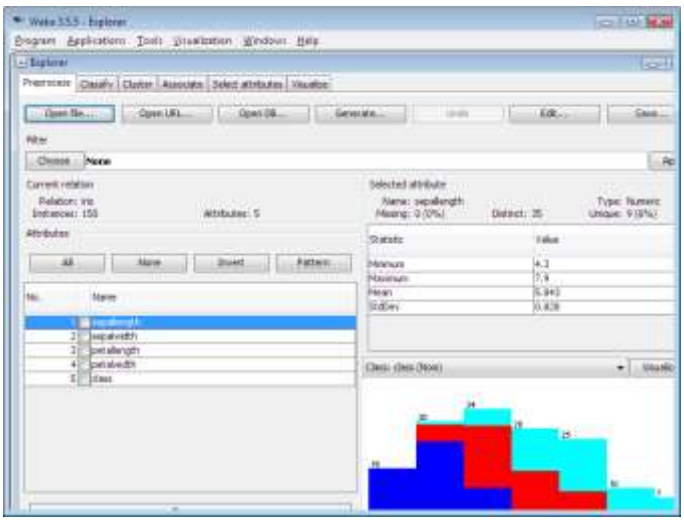
- Weka-website(Latest_version_3.6): –

  – http://www.cs.waikato.ac.nz/ml/weka/

- Weka-Manual: – http://transact.dl.sourceforge.net/sourceforge/weka/WekaManual-3.6.0.pdf

Weka tool consists of four keys viz.

1. Explorer: It is the location where you can explore the data .
2. Experimenter: It is the location where you can execute the experiments and can bear statistical tests between learning outlines.
3. Knowledge flow: It is the location which upkeeps essentially the similar functions as the EXPLORER nevertheless with a drag and drop interface.
4. Simply CLI: It offers an easy command-line interface that permits implementation of WEKA commands for operating systems and does not permit their own command line interface.



5. COMPARISION FIELDS FOR ANALYSIS

The comparison is done by some researchers and the comparison fields are taken as follows in [1].

1. Percent_correct
2. Fmeasure
3. Ir-precision
4. Ir-recall
5. AUC

Comparison fields are taken as follows in [2],[3].

1. TP-Rate
2. FP-Rate
3. ROC Area
4. Precision
5. Recall
6. F-Measure

Parameters are taken as follows in [3].

1. Correctly classified instances
2. Incorrectly classified instances
3. Kappa statistic
4. Mean absolute error

5. Root mean absolute error
6. Relative absolute error
7. Root relative squared error
8. Total no. of instances

6. SUMMARY ABOUT SOME TERMS

TP-Rate : It is known as true positive rate and is calculated as

TP-Rate= TP/(TP+FN)

TN-Rate : It is known as true negative arte and is calculated as

TN-Rate= TN/(TN+FP)

FP-Rate: It is known as false positive rate and is calculated as

FP-Rate= FP/(FP+TN)

FN-Rate: It is known as false negative rate and is calculated as

FN-Rate= 1-TP-Rate

PRECISION = TP/(TP+FP)

RECALL= TP/(TP+FN)

F-MEASURE= 2*precision*recall/(precision + recall)

ROC : Receiver Operating Characteristic Curve is a graphical plot equating the tp-rates and the fp- rates of a classifier as the refinement threshold of the classifier is different.

AUC: The area under the curve is frequently used as a sum-up of the ROC curve and as a measure the performance of the classifier.

Ir-precision: It is known as information retrieval precision and has two components retrieved relevant(a) and retrieved irrelevant(b) is calculated by :

Ir-precision= a / (a U b);U=union

Ir-precision= fraction of retrieved documents that are relevant.

Ir-recall: It is known as information retrieval recall and has two components retrieved relevant (a) and not retrieved relevant (c) is calculated by:

Ir-recall= a / (a U c)

Ir-recall= fraction of relevant documents retrieved.

7. COMPARISION OF CLASSIFIERS BY LITERATURE SURVEY

In [1] author has done an experiment to find the best classification algorithms among reflected to classify the records into normal and abnormal in the KDD data cup 20% training data set using WEKA tool. In [1] training set was having instances 25192 and that too through experiment type of 10 fold cross-validation. The experiment was implemented on WEKA-experimenter and as per the outcomes attained were with the comparison fields viz. percent_correct, fmeasure, AUC, ir-precision and ir-recall. The test of significance was taken as 0.05.Experiment outcomes [1] were as follows:

parameters ⇨	%Correct	f-measure	Ir-precision	Ir-recall	AUC
RF	1.0	1.0	1.0	1.0	1.0
Simplecart	1.0	1.0	1.0	1.0	1.0
J48	.99	1.0	1.0	1.0	1.0
Bagging	.99	1.0	1.0	1.0	1.0
Adaboost	.94	.95	.94	.96	.99
Ibk	.99	1.0	.99	1.0	.99
Naïve Bayes	.90	.90	.89	.91	.97
BayesNet	.97	.97	.95	.99	1.0
ZeroR	.53	.70	.53	1.0	.50
OneR	.96	.96	.99	.94	.96

After the analysis of these classification algorithms author[1] attained the analysis that Simple cart is finest, one of the best classifier with comparison fields percent_correct, f-measure as compared to other classifiers viz. ZeroR, OneR, BayesNet, NaiveBayes, Ibk, & Adaboost. The author [1] concluded that Random-Forest Classifier outperforms all other 9 classifiers with comparison fields percent_correct, f-measure, AUC and considered as best. ZeroR was concluded as the worst classifier in terms of comparison field except ir-recall. To reduce the computational time and efficiency one can do further complete study on only 5 classifiers viz. simple cart, bagging , ibk, j48, random forest.

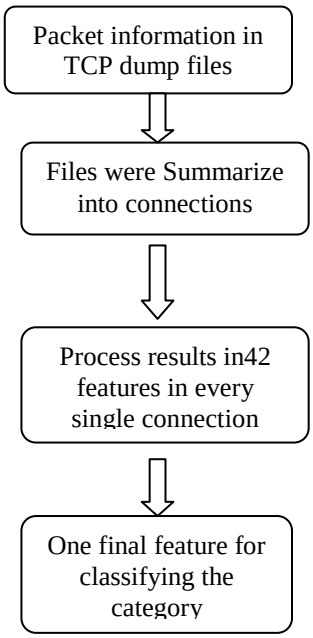
In [2] the author has proposed the learning models. In his research suitable method to calculate the performance of the set of meta classification algorithms viz. adaboost, attribute selected classifier, bagging, classification via regression, filtered classifier, logiboost, multiclass classifier. Author carefully taken a set of supervised machine learning approaches with classifiers that were useful on the chosen data set and that helped to calculate the risk of attacks of the environment having network. The execution was implemented using 10- fold cross validation and the outcomes were equated to attain the accuracy. In [2] author has estimated the set of classifier algorithms on KDD dataset. The attacks were estimated on four types i.e. Probe (information gathering), DOS(deny of service), U2R (user to root), R2L (remote to local). The procedure using recognition of mischievous behavior attacks types had achieved the detection-rate of

probe	99.17%
dos	96.71%
U2R	93.57%
R2L	31.17%

In spite of this due to piece of information that no FP was described by researcher approximately impossible detection rate[v] of 93.57% U2R type. Algorithms based on learning machine and multiclass svm adaptive intrusion detection was proposed by [D] for the establishment of IDS and the outcome of svm were calculated by kdd99 dataset where the performance was recorded as

Probe	81.2%
Dos	76.7%
U2r	21.7%
R2l	11.2%

Though, FP was sustained at comparatively low level of average=0.6% for all the four types.Author[2] proposed that study he used has a very little amount of dataset i.e. approx.10,000 unsystematic records.Yang Li and Li Guo[4] had understood the insufficiencies of KDD dataset. The novel approach is estimated on a KDD subset by arbitrary sampling 49,402 records for training stage and 12350 for testing stage. The standard TP of 99.6% and FP of 0.1% was recorded and further no information was represented by authors. The execution setup for all the experiments performed by [2] was computer with configurations : Intel® Core™ 2 CPU 2.13GHz, 2GB RAM and the operating system platform is Microsoft windows 7 & latest windows version: WEKA 3.7.1. There were two dissimilar kinds of attacks first is Normal and second is anomaly. Binary class content can be extended in[5] i.e.c4.5 programs for machine learning.



Therefore, each instance of data contains 42 features and every instance of them can be directly mapped due to the large audit data records in the original NSL-KDD AND 125971 instances were extracted.

Category of attacks	normal	anomaly	total
Number of records	67342	58629	125971
%ageclass occur	53%	47%	100%

Comparison performance of all seven classifiers were recorded as follows:

algorithm	fp rate	tp rate	Roc area	class
Adaboost.M1	.072	.960	.988	Normal

Attribute-selected	.009	.998	.999	Normal
Classification-via-regression	.003	.999	1.000	Normal
Bagging	.002	.999	1.000	Normal
Filtered	.003	.997	.998	Normal
Logiboost	.038	.979	.996	Normal
multiclass	.032	.973	.989	Normal

algorithm	precision	recall	f-measure	class
Adaboost.M1	.953	.928	.940	Anomaly
Attribute-selected	.997	.991	.994	Anomaly
Classification-via-regression	.998	.997	.998	Anomaly
Bagging	.999	.998	.998	Anomaly
Filtered	.996	.997	.997	Anomaly
Logiboost	.975	.962	.969	Anomaly
multiclass	.969	.968	.968	Anomaly

From the outcomes the author[2] concluded that bagging is the best classifier in prediction as compared to other classifiers.

In [3] the author had tried to deal with the problem of green-eyed users behavior that is necessary to be discriminated from authentic users behavior as it is obligatory to create the social web safe for the users. The author[3] had investigated the performance on two different and real sets of data. First is live journal and second is facebook links using the classification algorithms. The algorithms that author hired in his paper are REP-Tree, Naïve Bayes Multinomial Updateable, Complement Naïve Bayes, Classification Via Clustering on the source parameters as tp-rate, fp-rate, recall, precision, f-measure, roc and after the evaluation the class is generated that which is green-eyed node and which is authentic node.

Analysis is done on REP-Tree is as follows:

Parameters	Facebook		Live Journal	
Correctly Classified Instances	99819		99990	
Incorrectly-Classified-	181		10	
Kappa Statistics	.903		.995	
Mean absolute error	.002		.0002	
Root mean squared error	.04		.01	
Relative absolute error	14.243		.879	
Root relative absolute error	40.243		10.0362	
Total no of instances	100000		100000	
TP-Rate	.860	1	.996	1
FP-Rate	0	.140	0	.004
CLASS	malicious	legitimate	malicious	legitimate

Analysis is done on Naïve Bayes Multinomial updatable is as follows:

Parameters	Facebook		Live Journal	
Correctly Classified Instances	48474		50073	
Incorrectly-Classified-	51526		49927	
Kappa Statistics	-0.001		-0.0038	
Mean absolute error	.51		.5006	
Root mean squared error	.71		.6991	
Relative absolute error	2600.83		2526.859	
Root relative absolute error	718.38		702.5895	
Total no of instances	100000		100000	
TP-Rate	.48	.485	.404	.502
FP-Rate	.515	.52	.498	.596
Precision	.009	.989	.008	.988
Recall	.48	.485	.404	.502
F-measure	.018	.651	.016	.666
ROC	.502	.513	.479	.477
CLASS	malicious	legitimate	malicious	legitimate

Analysis is done on Complement Naïve Bayes as follows:

Parameters	Facebook		Live Journal	
Correctly Classified Instances	46573		43887	
Incorrectly-Classified-	53427		56113	
Kappa Statistics	.0004		-0.0031	
Mean absolute error	.5343		.5611	
Root mean squared error	.7309		.7491	
Relative absolute error	2696.62		2832.4634	
Root relative absolute	734.62		752.8599	
Total no of instances	100000		100000	
TP-Rate	.547	.465	.475	.439
FP-Rate	.535	.453	.561	.525
Precision	.01	.99	.008	.988
Recall	.547	.465	.475	.439
F-measure	.02	.633	.017	.607
ROC	.506	.506	.457	.457
CLASS	malicious	legitimate	malicious	legitimate

Analysis is done on Classification via clustering as follows:

Parameters	Facebook		Live Journal	
Correctly Classified Instances	65580		55216	
Incorrectly-Classified-	34420		44784	
Kappa Statistics	-0.002		-0.0109	
Mean absolute error	.3442		.4478	
Root mean squared error	.5867		.6692	
Root relative absolute error	589.64		672.5798	
Total no of instances	100000		100000	
TP-Rate	.294	.659	.2	.556
FP-Rate	.341	.706	.444	.8
Precision	.009	.989	.005	.986
Recall	.294	.659	.2	.556
F-measure	.017	.784	.009	.711
ROC	.477	.477	.378	.378
CLASS	malicious	legitimate	malicious	legitimate

Outcomes of this [3] paper may be further used for more finer classification of the datasets from online social web. The author concluded that amongst all the classifiers experimented on the two datasets; one of the finest and best algorithm for classification of green-eyed and authentic users is REP-Tree .

8. CONCLUSION AND FUTURE WORK

In reference to [1], [2], [3] it is observed that using WEKA the best classifiers for classification and detection of malicious users are Random Forest Algorithm, REP-Tree, Bagging Classifier are the finest and best algorithms and classifiers as compared to others. Further, experiment can be done by restricting to only these three algorithms and can evaluate the best out of these three algorithms.

REFRENCES

[1] S. Venkata Lakshmi1 and T. Edwin Prabakaran, "Performance Analysis of Multiple Classifiers on KDD Cup Dataset using WEKA Tool*" Indian Journal of Science and Technology, Vol 8(17),* August 2015

[2] G.Michael, A.Kumaravel and A.Chandrasekar "Detection of malicious attacks by Meta classification algorithms" *Int. J. Advanced Networking and Applications Volume: 6 Issue: 5 Pages: 2455-2459* January2015

[3] Pran Dev, Dr. Kulvinder Singh and Dr. Sanjeev Dhawan "Classification of Malicious and Legitimate Nodes for Analysing the Users' Behaviour in Heterogeneous Online Social Networks" *2015 1st International conference on futuristic trend in computational analysis and knowledge management (ABLAZE 2015)* 2015

[4] Langley P. and Simon H. "Applications of machine learning and rule induction", *Communications of the ACM, Vol.38, No. 11, pp. 55–64.* 1995

[5] Morgan Kaufmann, San Mateo "C4.5: Programs for Machine Learning." *Quinlan, J.:* (1993).

[6] A detunmbi AO, Falaki SO, Adewale OS, Alese BK. "Network Intrusion Detection based on Rough Set and k-Nearest Neighbour". *International Journal of Computing and ICT Research. 2008; 2(1):60–6.* Available from: http://www.ijcir.org/volume1number2/article7.pdf

[7] R anjan R, Sahoo G. "A new clustering approach for anomaly intrusion detection." *International Journal of Data Mining and Knowledge Management Process 4(2):29– 38* 2014 Mar.

[8] A zad C, Jha VK." Data Mining based Hybrid Intrusion Detection System." *Indian Journal of Science and Technology. 7(6):781–9.* Jun; 2014

[9] K hor K-C, Ting C-Y, Amnuaisuk S-P. "From Feature Selection to Building of Bayesian Classifiers: A Network Intrusion Detection Perspective." *American Journal of Applied Sciences; 6(11):1948–59* 2009

[10] L ee W, Stolfo SJ, Mok KW. "Algorithms for Mining System Audit Data.Proc KDD"; 1999.

[11] G hali NI. "Feature Selection for Effective Anomaly Based Intrusion Detection." *International Journal of Computer Science and Network Security. ; 9(3):285–9.* 2009 Mar

[12] "K-DD CUP 1999 DATASET": Available from: http://kdd.ics. uci.edu/databases/kddcup99/

[13] S ANS Institute InfoSec Reading Room."Understanding Intrusion Detection Systems"; 2001.

[14] W u X, Kumar V, Ross Quinlan RJ, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Yu PS, Zhou Z-H, Steinbach M, Hand DJ, Steinberg D. Top 10 "algorithms in data mining." *London: Springer-Verlag; 2008. p. 1–3. DOI: 10.1007/ s10115-007-0114-2*

[15] V enkata Lakshmi S, Edwin Prabakaran T." Application of k-Nearest Neighbour Classification Method for Intrusion Detection in Network Data*." International Journal of Computer Applications (0975-8887);; 97(7):34–7. 11. W eka Manual. Available from: http://www.ittc. ku.edu/~nivisid/WEKA_MANUAL.pdf* **;**2014 Jul

[16] Witten, I.H., Frank, E.: "Data Mining: Practical Machine Learning Tools and Techniques," *2nd edn. Morgan Kaufmann, San Francisco* (2005).

[17] Tavallaee M.E, Bagheri W. Lu and Ghorbani A. "A Detailed Analysis of the KDD CUP 99 Data Set", *Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), pp. 53-58*. (2009),

[18] Xu, X.: "Adaptive Intrusion Detection Based on Machine Learning: Feature Extraction, Classifier Construction and Sequential Pattern Prediction*." International Journal of Web Services Practices 2(1-2), 49–58* (2006).

[19] Li, Y., Guo, L.: "An Active Learning Based TCMKNN Algorithm for Supervised Malicious Network node detection". *In: 26th Computers & Security pp. 459–467* (October 2007)

[20] "Nsl-KDD data set for network-based intrusion detection systems." *Available on: http://nsl.cs.unb.ca/NSL-KDD* .

[21] Panda M. and Patra M.R (2008), "A Comparative study of Data Mining Algorithms for Network Intrusion Detection", *Proceedings of the 1st Conference on Emerging Trends in Engineering and Technology, pp. 504-507, IEEE Computer Society, USA.*

[22] Amor N.B, Benferhat S. and Elouedi Z (2004), "Naïve Bayes vs. Decision Trees in Intrusion Detection Systems", *Proceedings of 2004, ACM Symposium on Applied Computing, pp. 420-424*.

[23] G.MeeraGandhi, Kumaravel Appavoo, S.K.Srivatsa," Effective Network Intrusion Detection using Classifiers Decision Trees and Decision rules", *Int. J. Advanced Networking and Applications Volume: 02, Issue: 03, Pages: 686-692* (2010*).*

[24] N. Shrivastva, A. Majumder and R. Rastogi, "Mining (Social) Network Graphs to Detect Random Link Attacks," *Proceedings of the IEEE 24th International Conference on Data Engineering, pp. 486-495*, 2008.

[25] J. Karamon, Y. Matsuo and M. Ishizuka, "Generating useful networkbased features for analyzing social networks," *Proceedings of the 23rd national conference on Artificial intelligence, Vol. 2 (AAAI'08), pp. 1162-1168*, 2008.

[26] M. Maia, J. Almeida, and V. Almeida, "Identifying user behavior in online social networks," *Proceedings of the 1st Workshop on Social Network Systems of ACM, pp. 1-6*, 2008.

[27] M. Lahiri and T.Y. Berger-Wolf, "Mining Periodic Behaviour in Dynamic Social Networks," *Proceedings of the Eight IEEE Conference on data mining, pp. 373-382*, 2008.

[28] Markines, C. Cattuto, and F. Menczer, "Social spam detection," *Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web, pp. 41-48*, 2009.

[29] Q. Wang, B. Liang, W. Shi, Z. Liang and W. Sun, "Detecting Spam Comments with Malicious Users' Behavioral Characteristics," *Proceedings of Information Theory and Information Security (ICITIS), IEEE, pp. 563-567*, 2010.

[30] Z. Halim, M. M. Gul, N. Hassan, R. Baig, S. Rehman and F. Naz, "Malicious Users' Circle Detection in Social Network Based on SpatioTemporal Co-Occurrence," *Computer Networks and Information Technology, IEEE, pp. 35-39*, 2011.

[31] J. Cameron, C. Leung and S. Tanbeer, "Finding Strong Groups among Friends in Social Networks," *9th International Conference on on Dependable, Autonomic and Secure Computing, IEEE, pp. 824-831*, 2011.

[32] S. Al-Oufi, H. Kim and A. E. Saddik, "Controlling Privacy with Trustaware Link Prediction in Online Social Netwroks," *Proceedings of the 3rd International Conference on Internet Multimedia Computing and Service, ACM, pp. 86-89*, 2011.

[33] D. Prakash and S. Surendran, "Detection and Analysis of Hidden Activities in Social Networks," *International Journal of Computer Applications, pp. 34-38*, 2013.

[34] D. M. Freeman, "Using Naïve Bayes to Detect Spammy Names in Social Networks," *Proceedings of the ACM Workshop on Artifical Intelligence and Security, pp. 3-12*, 2013.

[35] S.Y. Bhatt, M. Abulasih, "Community-Based Features for Identifying Spammers in Online Social Networks*," Proceedings of the International Conference on Advances in Social Networks Analysis and Mining, ACM, pp. 100-107*, 2013.

[36] P. Dev, K. Singh and S. Dhawan, "Hidden Relationships for Analysing Users' Behaviour in Heterogeneous Social Networks," *Proceedings of the 2nd National Conference on Converging Technologies Beyond 2020, pp. 297-300*, 2014.

[37] H. Yin, B. Cui, L. Chen, Z. Hu and Z. Huang, "A Temporal Contextaware Model for User Behavior Modeling in Social Media Systems," *Proceedings of the International Conference on Management of Data, ACM, pp. 1543-1554*, 2014.

[38] B. Viswanath, M. A. Bashir, M. Crovella, S. Guha, K. P. Gummadi, B. Krishnamurthy and A. Mislove, "Towards Detecting Anomalous User Behavior in Online Social Networks*," Proceedings of the 23rd USENIX Security*, 2014.

[40] M. Jiang, P. Cui, A. Beutel, C. faloutsos and S. Yang, "Detecting Suspecious Following Behavior in Multimillion-Node Social Networks," *Proceedings of the 23rd International Conference on World Wide Web Companion, ACM, pp. 305-306*, 2014.

[41] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explorations, Vol. 11(1), pp. 10-18*, 2009.

[42] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi, "On the evolution of user interaction in Facebook," *Workshop on Online Social Networks, pp. 37–42*, 2009.

[43] L. Backstrom, D. Huttenlocher, J. Kleinberg and X. Lan, "Group Formation in Large Social Networks: Membership, Growth, and Evolution." *Proceedings of the 12th International Conference on Knowledge Discovery and Data Mining, ACM, pp. 44-54*, 2006.

[44] J. Leskovec, K. Lang, A. Dasgupta, M. Mahoney, "Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters," *Proceedings of the Internet Mathematics, Vol 6(1), pp. 29-123*, 2009.