Global Journal on Innovation, Opportunities and Challenges in AAI and Machine Learning
Vol. 2, Issue 1 - 2018

ATTRIBUTE ANALYSIS METHODOLOGIES BY FINDING CORRELATIONS AND WEIGHT ESTIMATION FOR OPTIMAL ANALYSIS IN WEKA

VINAY PAL[*], BRIJ KISHORE[*], VISHAL DUTT[**], BHARAT SINGH[***]

ABSTRACT

We'll use the Road Accident Data collected from Data repository available on web stores. The tasks performed in previous studies are to generate the final result on the dataset. We'll also find the result but mainly we analyze that what attribute is more important in decision making process. We'll perform some operation on data, then we'll identify that particular attributes throughout the dataset. For this purposes we would use the Naive Bayesian Classifier. A good accuracy was in the primary motive to proceed further with this work. Naive bayes classifier is a revolutionary approach in the data analysis and classification domain which provides better accuracies. The dataset should be made like so that algorithms can be applied to get optimal results. We have compared the existing work and consider their problem formulation to get overcome with optimizing results. A road accident is something which misshapen with our mistakes or road infrastructure so we have chosen dataset regarding all the issue and attributes and applied Weka over it.

KEYWORDS: Weka, Naïve Bayesian.

INTRODUCTION

Human life is affected with many reasons where traffic and road accident plays a very hazarding role in their life. Mainly accidents happened due to the negligence of driving vehicle on the roads. There are various reasons responsible for the accident like abandon of traffic rules but road conditions and the traffic are considered the one of prime cause of fatality and causality across the globe. Hence traffic accident is a serious global issue of the present time. This thesis discuss about effect of accident on humans life and technique used for the implementation. This thesis also includes the motivation for the work reported in this thesis, scope and objective of the thesis.

Data mining refers to extracting or mining knowledge from large amounts of data. The term is actually a misnomer. Thus, data mining should have been more appropriately named as knowledge mining which emphasis on mining from large amounts of data.

[*]Department of Computer, Science & Engineering, Apex Institute of Engineering & Technology, Jaipur.
[**]Assistant Professor, DEC Ajmer, Visiting Faculty, MDSU Ajmer.
[***]Department of Computer Science & Engineering, St Wilferd Engg College, Ajmer.
Correspondence E-mail Id: editor@eurekajournals.com

 ISSN: 2581-5156

It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. The key properties of data mining are Automatic discovery of patterns Prediction of likely outcomes creation of actionable information Focus on large datasets and databases

BACKGROUND

A literature review is necessary to know about the research area and what problem in that area has been solved and need to be solved in the future. A proper literature review provides a solid background for a noble research work. A good literature review is comprehensive, critical and contextual. It provides a theory base, a survey of published works that pertain to our investigation, and an analysis of that work. It is a critical, factual overview of what has gone before. A good literature review shows awareness of reviewer in the field. One has to start with a broader domain of some area/ sub area and while doing the study of literature narrow down the domain to specific points of issue to decide upon. Literature survey includes the study of various sources of literature in the area of research. It includes finding the related material from magazines, books, research articles, scientific research papers published in various conferences, journals & transactions. One may take a few days to a few weeks to understand a research paper published in standard peer reviewed journals. The researchers need to adopt a certain path for doing literature review of such literature. There has been many procedures and processes defined by the researchers to undergo through and arrive at certain conclusions of research objectives.

A good sized quantity of information is rising inside the difficulty of science, medical and masses of different regions due to the short improvement in computerization and digitalization strategies. These statistics may additionally provide an amazing useful resource for information extraction and desire guide. We one can understand, analyse, and ultimately make efficient use of the large quantity of statistics; a multidisciplinary approach is needed to meet the challenge. Traditionally we have been forceful to rely upon record control tool and manual art work but now we are moving towards a brand new age known as-information age, e.g. online buying, railway fee tag reserving and so forth are becoming relying on laptop which include a huge quantity of Data. Numerous databases and information warehouses are built to seize the statistics. So a flexible and effective device is needed to convert Data into the treasured data. This led to the start of statistics mining or Data dredging. Many applications are based totally on statistics dredging e.g. enterprise intelligence, are trying to find engine and so on. Employer intelligence wishes the beyond data and are looking forward to for future on the idea of the calculation. As gold mining is the exploration for chunks of gold, so statistics mining additionally referred to as statistics dredging is the exploration for chunks of information in time collection data dredging, the ones chunks are called as activities. As gold is buried within the floor, chunks of statistics are masked in data. Energetic efforts had been finished in designing inexperienced mechanisms for extracting Data and policies from huge databases. Data mining methods are collection of critical algorithm for spotting effective, rational, probably beneficial, and previously unknown inclinations in huge databases. Counting on the types of databases processed, Data discovery from database strategies can be divided as transactional databases, temporal databases, and relational databases. Secondly, depending on the training of statistics derived, knowledge discovery from database techniques may be divided as inducing association guidelines, sequential styles,

 ISSN: 2581-5156

clustering rules and classification rules. Inducing association policies in transaction databases is maximum commonplace software program in Data mining. an association rule can be symbolized in the form p → q, wherein p and q are item devices, in this sort of way that the presence of p in a transaction will constitute on the presence of q. Measures, aid and self-belief, are evaluated to determine whether a rule ought to be stored. Guide and self- belief, are evaluated to decide whether or no longer a rule wants to be kept. The useful resource of a rule is the percentage of the transactions which encompass all the gadgets in p and q.

The self-assurance of a rule is the conditional possibility of the occurrences of devices in p and q over the occurrences of gadgets in p. the assist and the self- assurance is an thrilling rule must be large than or equal to a client-sure minimal useful resource and a minimal self-assurance respectively. Maximum of the previous methods set a unmarried minimal manual threshold for all the gadgets or object units. But in actual programs, one of a kind items may additionally moreover have incredible criteria to choose its significance. For instance, the minimum allows for less costly objects can be set better than the ones for greater steeply-priced objects..

Mayanka Katyal et al., 2014 [9] In this scrutiny paper Current date endured need for resource solutions which can be hungry demands inside it industry has grasped to advance of Cloud processing. Cloud processing nature involves elevated worth groundwork on one side and demand elevated scale computational sources on the hand that is supplementary. These assets require progress toward becoming provisioned (assignment and planning) to the complete clients in most way that is effectual that the unfathomable capacities of cloud can be utilized effectually and efficiently. In this report they face off regarding a recognizing calculation for designation of cloud assets to end-clients on-request premise. This calculation is organized on max-min and min-min calculations. These are two normal undertaking calculation that is orchestrating. The perceiving calculation utilizes heuristics being exact select in the midst of the two calculations so completed make traverse of employments on the components is limited. The occupations have a tendency to be anticipated on systems in whichever space age or area form. They survey their provisioning heuristics holding a cloud test system, hollered Cloud Sim. They likewise differentiated their way to deal with the insights acquired subsequently provisioning of assets was done in First-Cum-First-Serve(FCFS) way. The aftermath this is certainly experimental that completed make span of jobs on given collection of VMs minimizes considerably in disparate scenarios.

Mohamed Abu Sharkh et al., 2013 [10] In this scrutiny paper Cloud processing is a computing this is certainly progressively consented, nowadays elucidating absolutely essential for utility computing services. Every single solitary provider propositions a exceptional skill portfolio alongside a scope of resource configurations. Resource provisioning for cloud services in a comprehensive method is critical to every single resource allocation model. Every single flawless must to ponder both computational resources and web resources to precisely embody and assist functional needs. One extra aspect that must to be trusted as provisioning resources is manipulation usage. This aspect has become attention that is additional industry and states parties. Phone calls of prop when it comes to clouds which can be green obtaining energy. Understanding that, resource allocation algorithms target to complete the duty of organizing adjacent systems on information center servers and subsequent link that is arranging on the net tracks available as complying alongside the setback limitations. Countless external and facets that are internal change the presentation of resource allocation models are provided in this paper. These

 ISSN: 2581-5156

elements are discussed at length and spaces which are scrutiny revealed. Design exams are debated alongside the mark of bestowing a reference becoming used afterward arranging a manipulation that is comprehensive resource allocation flawless for cloud computing data facilities.

SIMULATION AND RESULTS

In previous studies, they work is done on comparing the algorithm performances. The Data is taken and any respected or desired algorithms or classifiers were implemented on that data. Then each algorithm result was stored separately. Then they compared all the results and Statistical information among them. And finally the best resulting algorithm was identified. In most of the cases this methodology was used. In other more cases, only result was calculated on the particular data.

PROBLEM FORMULATION AND PROPOSED SOLUTION

In comparison of pervious work, our main focus would be on the attribute role and their weights. This will help to identify the correlation among variable. To do so, we'll implement the Naive Bayes Classifier to classify the dataset specifically. With the help of these major role attributes, we can reach to our goal i.e. finding the main causes of accident, and what precautions should we keep while driving or on roads.

NAIVE BAYES CLASSIFIER

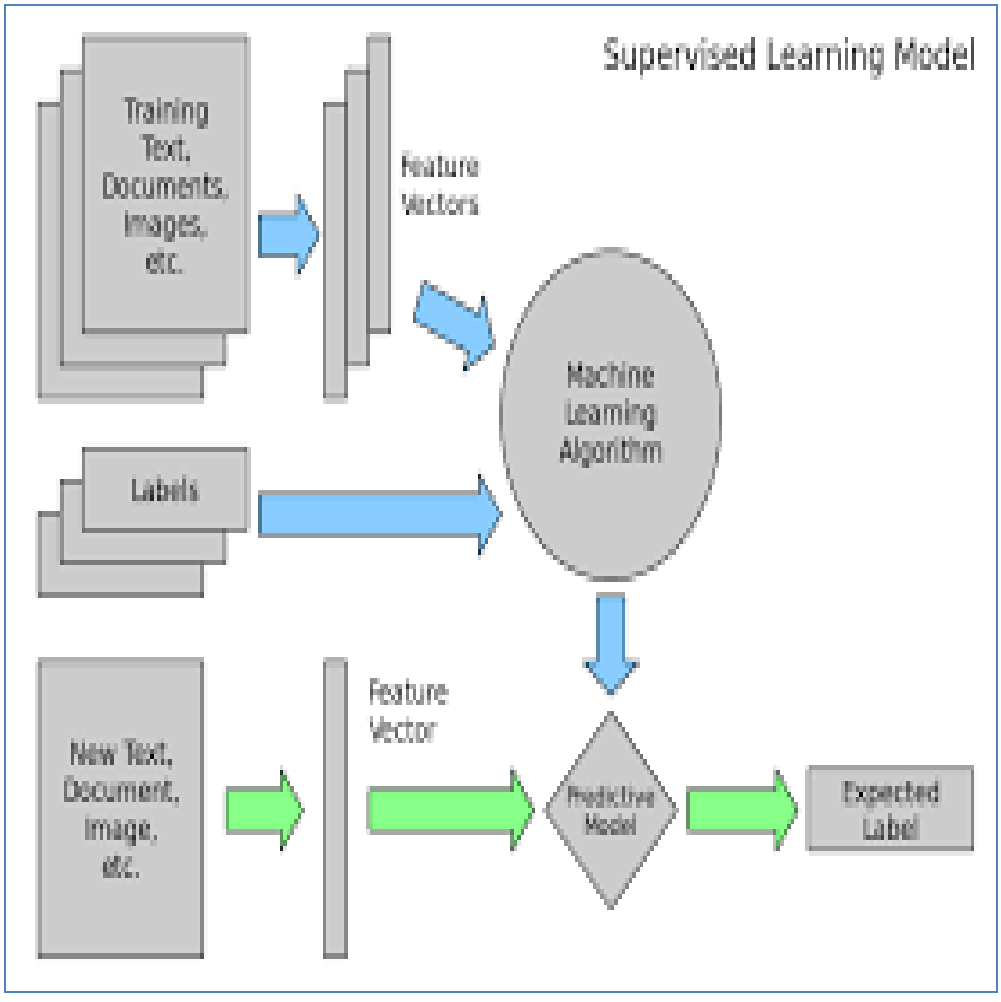

Figure 1.Learning Model

 ISSN: 2581-5156

Step 1: Load the dataset and Select the Naive Bayes Classifier from the Classify Tab as shown in figure given below:

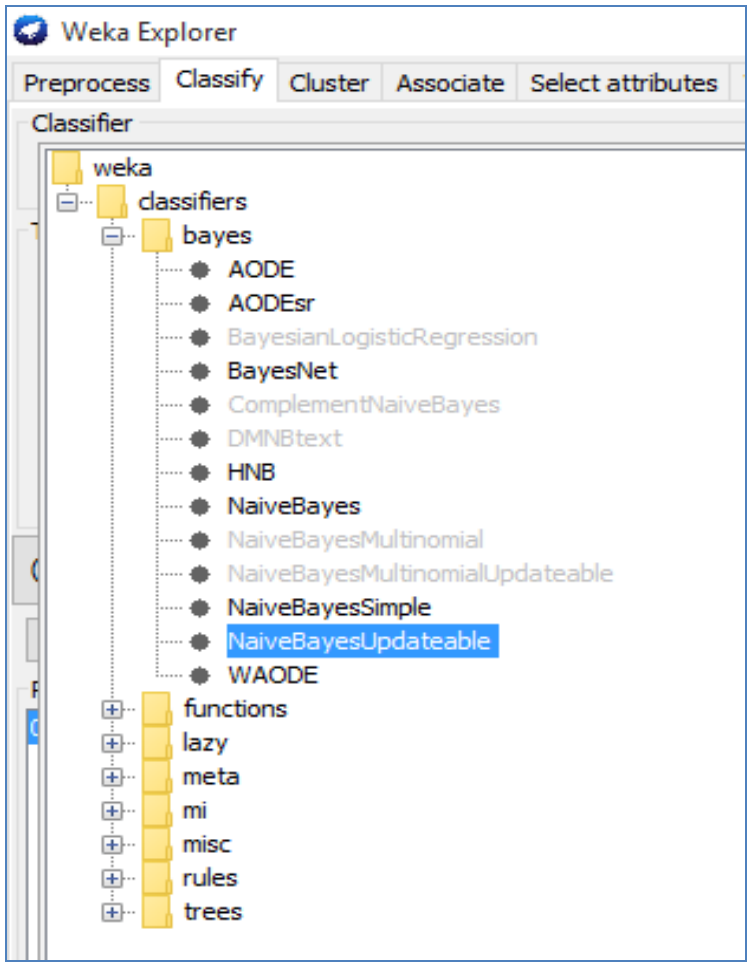

Figure 2.Selecting naïve Bayes Updatable as Classifier

Class for a Naive Bayes classifier using estimator classes. This is the updateable version of Naïve Bayes. This classifier will use a default precision of 0.1 for numeric attributes when build Classifier is called with zero training instances.

Debug: If set to true, classifier may output additional info to the console. Display Model In Old Format -- Use old format for model output. The old format is better when there are many class values. The new format is better when there are fewer classes and many attributes.

Use Kernel Estimator: Use a kernel estimator for numeric attributes rather than a normal distribution.

Use Supervised Discretization: Use supervised discretization to convert numeric attributes to nominal ones.

Capabilities: It can handle various types of Class such as Missing class values, Nominal class, Binary class.

Attributes: Numeric attributes, Missing values, Binary attributes, Nominal attributes, Empty nominal attributes, Unary attributes

Step 2: Set the Classifier output behavior or output specifications of optimizing results, as shown in figure given below:

 ISSN: 2581-5156

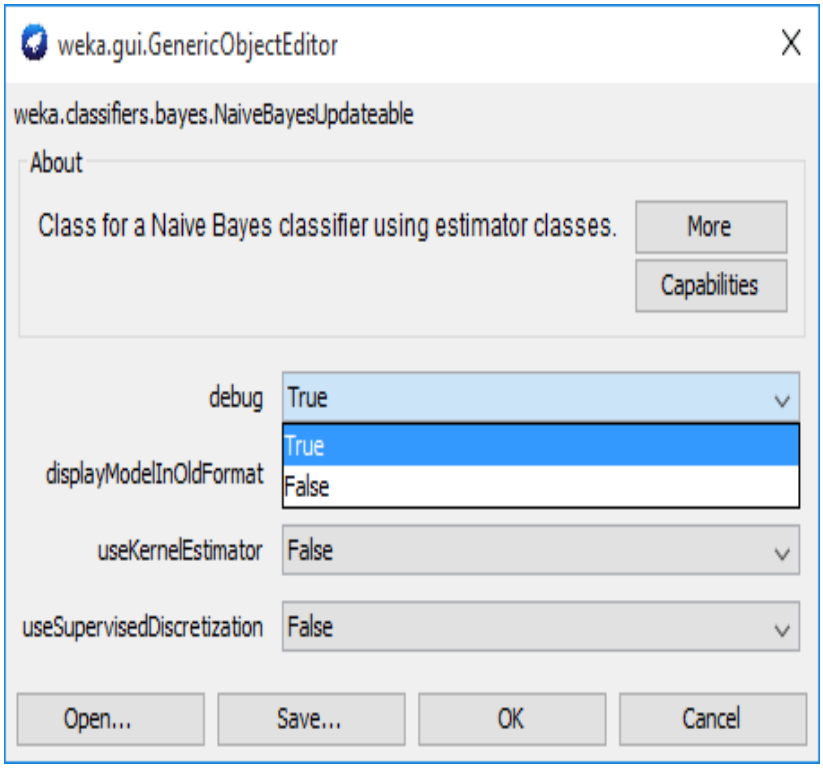

Figure 3.Setting up classifier output properties

Step 3: Start the Algorithm Execution by clicking on start button as shown in figure given below:

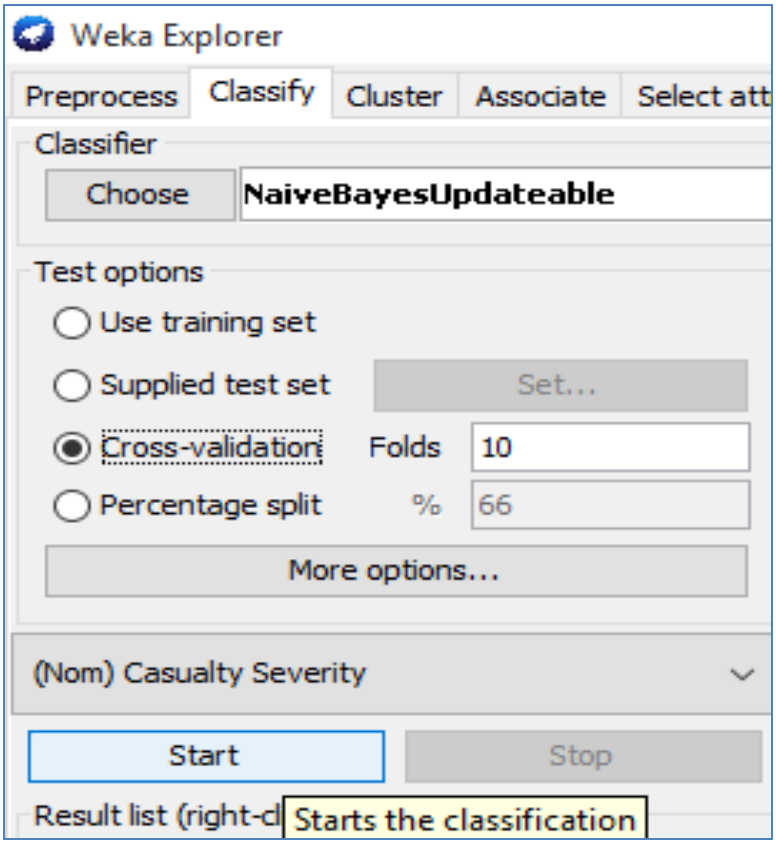

Figure 4.Start the Classification or Execution of classifier

 ISSN: 2581-5156

When execution starts, the output of the classifier is shown in Classifier Output window stated in the right of the interface as shown in figure given below:

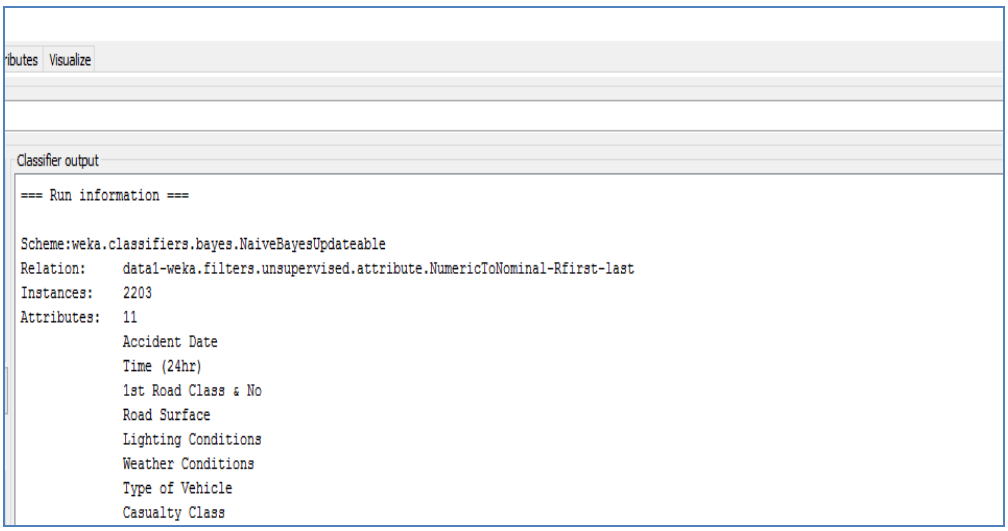

Figure 5.Classifier Output Window

Step 4: Generate the Output.

In this step, the output of the classifier is shown in classifier window. It consists of some important information such as:

- Classifier Schema
- Relation name
- Total Instances
- Total Attributes
- Attribute Names

CONCLUSION AND FUTURE WORK

A road accident is something which misshapen with our mistakes or road infrastructure so we have chosen dataset regarding all the issue and attributes and applied Weka over it. A good accuracy was in the primary motive to proceed further with this work. Naive bayes classifier is a revolutionary approach in the data analysis and classification domain which provides better accuracies. The dataset should be made like so that algorithms can be applied to get optimal results. We have compared the existing work and consider their problem formulation to get overcome with optimizing results. A road accident data and its optimal analysis can help the data analytics and government also to take a better care regarding accident prone activities. During the implementation of data analysis in Weka We have applied different algorithms for calculating updated weight in order to find better correlations for optimal solution and meaningful pattern.

REFERENCES

[1]. Mell, Peter, and Tim Grance. "The NIST definition of cloud computing." National Institute of Standards and Technology 53, no. 6 (2009): 50.

[2]. N.Krishnaveni, G.Sivakumar,“Survey on Dynamic Resource Allocation Strategy in Cloud Computing environment”, Dept. of CSE Erode Sengunthar Engineering College Thudupathi, India, International Journal of Computer Applications Technology and Research, Vol. 2, Issue 6, pp. 731-737, 2013.

[3]. Abhishek Kumar, Pramod Singh et al. “An Approach for Classification using Simple CART Algorithm in Weka”, IEEE Sponsored 3rd International Conference on Electronics and Communication Systems (ICECS 2016), 978-1-4673-7832-1/16, 2016 IEEE.

[4]. Armbrust, Michael, Armando Fox, Rean Griffith, Anthony D. Joseph, Randy Katz,

 ISSN: 2581-5156

Andy Konwinski, Gunho Lee et al. "A view of cloud computing."Communications of the ACM 53, no. 4 (2010): 50-58.

[5]. Pramod Singh. Analysis Of Crime Data Using Data Mining Algorithm. International Journal Of Engineering Sciences & Research Technology, 7(2), 675-681.

[6]. Chieu, Trieu C., Ajay Mohindra, Alexei A. Karve, and Alla Segal. "Dynamic scaling of web applications in a virtualized cloud computing environment." In e-Business Engineering, 2009. ICEBE' 09. IEEE International Conference on, pp. 281-286. IEEE, 2009.

[7]. Pramod Singh Rathore. An Contemplated Approach for Criminality Data using Mining Algorithm. International Journal on Future Revolution in Computer Science & Communication Engineering. Volume: 4 Issue: 2 ISSN: 2454-4248. February 2018.

[8]. Sotomayor, Borja, Rubén S. Montero, Ignacio M. Llorente, and Ian Foster. "Virtual infrastructure management in private and hybrid clouds." Internet Computing, IEEE 13, no. 5 (2009): 14-22.

[9]. Anshul Rai, Ranjita Bhagwan, Saikat Guha, "Generalized Resource Allocation for the Cloud", Microsoft Research India.

[10]. Zhang, Qi, Lu Cheng, and Raouf Boutaba. "Cloud computing: state-of-the-art and research challenges." Journal of internet services and applications 1, no. 1 (2010): 7-18.

[11]. Jin, Hai, Guofu Xiang, Deqing Zou, Song Wu, Feng Zhao, Min Li, and Weide Zheng. "A VMM-based intrusion prevention system in cloud computing environment." The Journal of Supercomputing 66, no. 3 (2013): 1133-1151.

[12]. DU, P. and NAKAO, A., "Ddos defense as a network service," pp. 894–897, 2010.

[13]. Durcekova, V., Schwartz, L., And Shahmehri, N., "Sophisticated denial of service attacks aimed at application layer," in ELEKTRO, 2012, pp. 55–60, IEEE, 2012.

[14]. Chan, S.C., Chu, Y.J., Zhang, Z.G.: A New Variable Regularized Transform Domain NLMS Adaptive Filtering Algorithm. IEEE Transactions on Audio, Speech, and Language Processing 21 (4), 868–878 (2013) Cross Ref Google Scholar.

 ISSN: 2581-5156