International Journal of Advanced Technology in Engineering and Science
Vol. No.4, Special Issue No. 01, February 2016
www.ijates.com
ijates
ISSN 2348 - 7550

TO IMPLEMENT LEAST MEAN SQUARES USING INCREASING THE SECURITY OF BIG DATA

P.Manju[1] M.Ashwin[2]

[1]PG Scholar, [2]Associate Professor, Department of CSE,
Adhiyamaan College of Engineering, Hosur-635109, Tamil Nadu, (India)

ABSTRACT

Data mining is the analysis step of the "Knowledge discovery in databases (KDD) and relatively new term, because the goal is the extraction of patterns & knowledge from large amount of data. In existing system, there are used multitier ensemble classifiers for security of big data. In this method providing of accuracy and performance is less. So, I will use least mean squares. The LMS algorithm for providing more security of big data and increasing the performance & accuracy of classification.

Keywords: big data.LIME classifier, LMS (Least Mean Squares) algorithm, security.

I. INTRODUCTION

Big data is a broad term for data sets, so Large (or)Complex that traditional data processing application extremely large data sets that may be analyzed computationally to Uncover patterns, trends and associations . Especially relating to human behavior &interactions. It has become particularly important in view of the rapid growth of the Cloud creates new opportunities for the users requires further research to address novel tasks &requirements. Security has been one of the major issues required for the use of big data. In previous papers used LIME Classifiers technique. [5] The LIME Classifiers Technique is to develop as a general technique that may be useful for the analysis of big data in various application domains. If a dataset is not Large enough, then the LIME Classifier will revert to using only base classifiers (or) just a small part of the whole system and will focus of this papers Malware detection represents a significant challenge for security of big data. The term of malware refers to malicious software (or) to malicious computer programs. Malware has been an ever growing threat to security for a big time. The main problems of behavior analysis is that it can be applied in practical situations to detect malware only once the later has been executed, which means that it has already had a change to perform its malicious function & spread further possibly in a new modified form. To overcome this problem least mean square algorithm are used.

In this algorithmLeast mean squares (LMS) algorithms are a class of adaptive filter used to mimic a desired filter by finding the filter coefficients that relate to producing the least mean squares of the error signal (difference between the desired and the actual signal). It is a stochastic gradient descent method in that the filter is only adapted based on the error at the current time. The realization of the causal Wiener sifter looks a lot like the solution to the least squares estimate, except in the signal processing domain. The least squares solution, for input matrix $\mathbf{x}$and output vector $\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\boldsymbol{y}.$

The FIR least mean squares filter is related to the Wiener filter, but minimizing the error criterion of the former does not rely on cross-correlations or auto-correlations. Its solution converges to the Wiener filter solution.

II. RELATED WORKS

J. Abawajy[1]Phishing attacks continue to pose serious risks for consumers and businesses as well as threatening global security and the economy. Therefore, developing countermeasures against such attacks is an important step towards defending critical infrastructures such as banking. Although different types of classification algorithms for filtering phishing have been proposed in the literature, the scale and sophistication of phishing attacks have continued to increase steadily. In this paper, we propose a new approach called multi-tier classification model for phishing email filtering. We also propose an innovative method for extracting the features of phishing email based on weighting of message content and message header and select the features according to priority ranking. We will also examine the impact of rescheduling the classifier algorithms in a multi-tier classification process to find out the optimum scheduling.

Laura Auria, Rouslan [2]multiclass classification is a major requirement in field of science and engineering because multiclass discrimination of objects is a serious problem in science and engineering. Multiclass classification is always considered complex than binary classification. In binary classification, only the decision boundaries of 1 class are to be known and rest (complement of first class) is considered as second class where as in multiclass classification, several boundaries are essential for that reason. This may lead to increase the probability of error because of constructions of many decision boundaries.

H. Abawajy, Andrei Kelarev, Morshed Chowdhury[3] As the new-generation distributed computing platform, cloud computing environments offer high efficiency and low cost for data intensive computation in big data applications. However, these advantages come at a price people no longer have direct control over their own data. Based on this view, data security becomes a major concern in the adoption of cloud computing. Authenticated Key Exchange (AKE) is essential to a security system that is based on high efficiency symmetric-key encryption. With virtualization technology being applied, existing key exchange schemes such as Internet Key Exchange (IKE) becomes time-consuming when directly deployed into cloud computing environment. In this paper we propose a novel hierarchical key exchange scheme, namely Cloud Background Hierarchical Key Exchange (CBHKE).

R. Islam, R. Tian, L. M. Batten, and S.Versteeg [4] Collection of dynamic information requires that malware be executed in a controlled environment; the malware unpacks itself as a preliminary to the execution process. On the other hand, while execution of malware is not needed in order to collect static information, the file must first be unpacked manually. In this paper, we present the first classification method integrating static and dynamic features into a single test. Our approach improvess on previous results based on individual features and reduces by half the time needed to test such features separately.

Algorithm Methodology

III. LEAST MEAN SQUARES (LMS) ALGORITHMS

The Least Mean Squares (LMS) algorithm is recursive technique used to identify transfer function of unknown systems by using input and output data, which is corrupted by noise. This paper evaluates the accuracy and

stability of the LMS algorithm. The LMS algorithm demonstrates a better performance over the Least Mean Square (LMS) and Least Square (LS) algorithms. [11] It provides accurate results and faster computational time as compared to conventional approaches.

Least mean squares (LMS) algorithms are a class of adaptive filter used to mimic a desired filter by finding the filter coefficients that relate to producing the least mean squares of the error signal (difference between the desired and the actual signal). It is a stochastic gradient descent method in that the filter is only adapted based on the error at the current time. The realization of the causal Wiener filter looks a lot like the solution to the least squares estimate, except in the signal processing domain. The least squares solution, for input matrix $\mathbf{X}$and output vector $\mathbf{y}$is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\boldsymbol{y}.$$

The FIR least mean squares filter is related to the Wiener filter, but minimizing the error criterion of the former does not rely on cross-correlations or auto-correlations. Its solution converges to the Wiener filter solution. One of the ways to find the optimal solution of the impulse response of the unknown system when it has the interference presented in both the input and the output is by using the Least Mean Squares (LMS) algorithm.

Instead of basing the approach on the minimum mean squares error as the LMS algorithm does, the LMS algorithm is based on the Least mean squares or the minimum Raleigh quotient approach. Like the LMS algorithm, the LMS algorithm is the unsupervised learning algorithm.

The LMS algorithm was derived from Oja and Xu's learning algorithm, which is used for extracting only the minor features from the data sequence, unlike the LMS algorithm. By extracting only the subsidiary information, the effect of the interference can be eliminated. Moreover, the LMS algorithm has also an advantage over the Least Squares (LS) algorithm in the computation time. The LMS algorithm's computation time for N-by-N matrix is 4N per iteration, whereas the LS algorithm's is $6N^3$ per iteration.

3.1 Definition of symbols

$n$is the number of the current input sample

$p$is the number of filter taps

$\{\cdot\}^H$(Hermitian transpose or conjugate transpose)

$\mathbf{x}(n) = [x(n), x(n-1), \ldots, x(n-p+1)]^T$

$\mathbf{h}(n) = [h_0(n), h_1(n), \ldots, h_{p-1}(n)]^T, \quad \mathbf{h}(n) \in \mathbb{C}^p$

$$y(n) = \mathbf{h}^H(n) \cdot \mathbf{x}(n)$$

$$d(n) = y(n) + \nu(n)$$

$\hat{\mathbf{h}}(n)$Estimated filter; interpret as the estimation of the filter coefficients after n samples

$$e(n) = d(n) - \hat{y}(n) = d(n) - \hat{\mathbf{h}}^H(n) \cdot \mathbf{x}(n)$$

IV. ALGORITHM DESCRIPTION

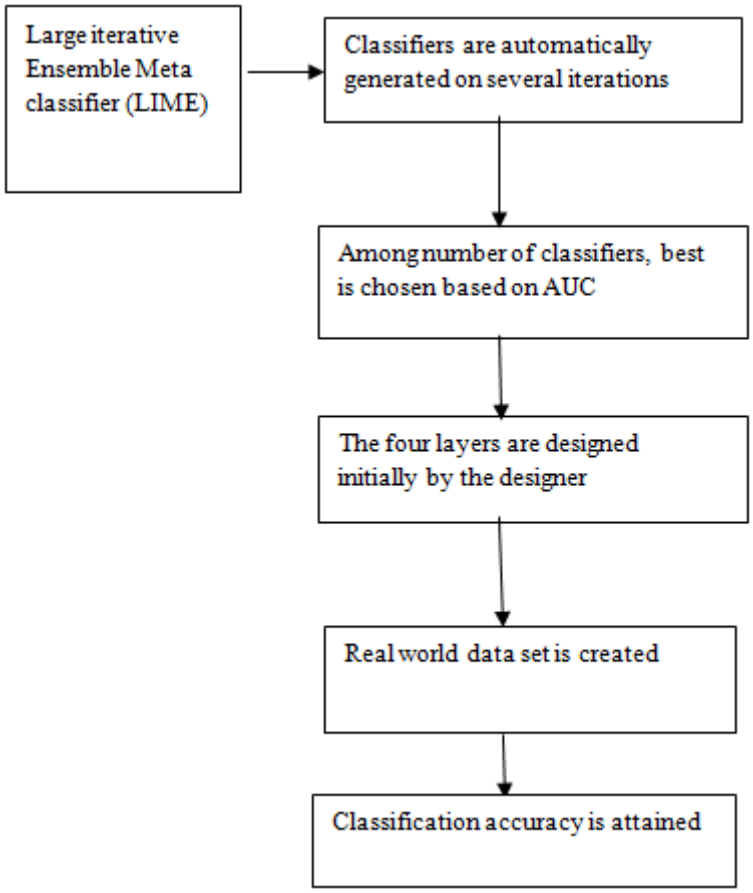

Fig: 1classification diagrams

From large number of classifiers the best classifiers is obtained based on AUC.then this best classifier is applied to the real world data set. And by using least mean square algorithm the classification is done. Thus the classification accuracy is improved.

V. OUTPUT ANALYSIS

In the first simulation, used the LMS and LMS algorithms to find the transfer function of an unknown system from the varied length of inputs and outputs. As mentioned in the methodology section, the last 100 values of the results are used to compute the averages, which represent thecomputation of the LMS and LMS algorithms for each number of samples. In order to determine the performance of the LMS and LMS algorithms, we have to find the errors between the averaged results of both algorithms and compare those to the ideal solution for every number of samples.

Regression is the easiest technique to use, but is also probably the least powerful (funny how that always goes hand in hand). This model can be as easy as one input variable and one output variable (called a Scatter diagram in Excel, or an XYDiagram in OpenOffice.org). Of course it can get more complex than that, including dozens of input variables. In effect, regression models all fit the same general pattern.

I

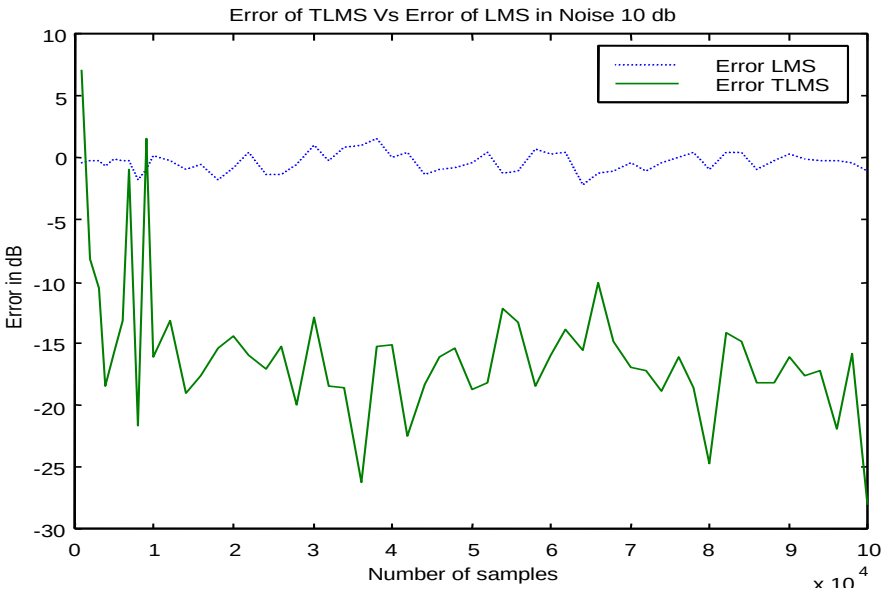

Fig 2: The error of LMS and LMS algorithm

The new addition of this article is in generating new large systems as repetitive ensembles of by linking a fifth tier ensemble meta classifier to a different fourth tier ensemble meta classifier rather than a base classifier and linking the fourth tier ensemble meta classifier to a third tier ensemble meta classifier, second tier ensemble meta classifier that successively are connected to their base classifiers. During this manner the fifth tier ensemble Meta classifier will generate the full system.

These classifiers are a new construction within the framework of this approach for the following two reasons. First, classifiers embody totally different ensemble Meta classifiers on many tiers. Second, they use these ensemble Meta classifiers iteratively to get the whole classification system automatically. In this automatic generation capability includes many large ensemble Meta classifiers in several tiers simultaneously and auto combines them into one hierarchical unified system so that one ensemble Meta classifier is an integral part of another one.

VI. CONCLUSION

The Least Mean Squares (LMS) Algorithm is the unsupervised learning adaptive linear combiner based on the least mean squares or the minimum Raleigh quotient, which isused to extract minor features from training sequences. This algorithm can be best used in situations when there is interference presented in both the input and output signals of the system. According to the simulations in this study, I found that the LMS algorithm provides better ability and stability and a more accurate way to identify the transfer function of the unknown system than the LMS algorithm does. In addition, the LMS algorithm's computation time, 4N per iteration, is also faster than the LS algorithm's, $6N^3$ per iteration. However, one of the problems of using the LMS algorithm is the difficulty of selecting the suitable learning rate,μ.In conclusion, the advantages of using the LMS algorithm are its ability to perform in a system, which is heavily corrupted with noise, and the stability of its results.

VII. ACKNOWLEDGMENT

This is a great pleasure and immense satification to express my deepest sense of gratitude and thanks to everyone who has directly or indirectly helped me in this project work successfully.

REFERENCES

[1]. R. Islam and J. Abawajy, "A multi-tier phishing detection and filtering approach," J. Newt. Compute. Appl., vol. 36, no. 1, pp. 324-335, 2013.

[2]. Laura Auria, Rouslan A. Moro, "Support Vector Machines (SVM) as a Technique for Solvency Analysis", in Berlin, August 2008.

[3]. Jamal H. Abawajy, Andrei Kelarev, Morshed Chowdhury," Large Iterative Multitier Ensemble Classifiers for Security of Big Data",in IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTING, 30 October 2014.

[4]. R. Islam, R. Tian, L. M. Batten, and S.Versteeg, "Classification of malware based on integrated static and dynamic features," J. New. Compute. Appl., vol. 36, no.

[5]. Cloud computing," in Proc. 12th IEEE Int. Conf. Trust Security Privacy Compute. Common. Melbourne, Australia, Jul. 2013, pp. 9-16.

[6]. X. Zhang, C. Liu, S. Nepal, C. Yang, and J. Chen, "Privacy preservation over big data in cloud systems," in Security, Privacy and Trust in Cloud Systems. Berlin, Germany: Springer-Verlag, 2013, pp. 239-0257.

[7]. R. Islam, J. Abawajy, and M. Warren, "Multi-tier phishing email classification with an impact of classifier rescheduling," in Proc. 10th ISPAN, 2009, pp. 789-793.

[8]. L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5-32,2001.

[9]. T. K. Ho, Random decision forest," in Proc. 3rd Int. Conf. Document Anal. Recognit., 1995, pp. 278-282.

[10]. Y. Amit and D. Geman, "Shape quantization and recognition with randomized trees," Neural Compute., vol. 9, no. 7, pp. 1545-1588, 1997

[11]. Abatzoglou, T.J., Mendel, J.M., and Harada G.A. (1991), The Constrained Least Squares Technique and its Application to Harmonic Super solution, IEEETransactions on Signal Processing, vol. 39, pp. 1053-1086.

[12]. Mendel, J.M. (1995). Lessons in Estimation Theory for Signal Processing, Communications, and Control. Englewood Cliffs, NJ: Prentice-Hall.