RESEARCHERID
THOMSON REUTERS

[*Vishwakarma* et al., 4(5): May, 2017*] ISSN: 2349-5197
Impact Factor: 2.715

INTERNATIONAL JOURNAL OF RESEARCH SCIENCE & MANAGEMENT

INTRUSION DETECTION SYSTEM USING KDD'99 CUP DATASET

Satyendra Vishwakarma* & Vivek Sharma
Information technology, Samrat Ashok Technology Institute, Vidisha, (M.P.), INDIA
DOI: 10.5281/zenodo.574456

Keywords: DoS, High Dimensional, Intrusion detection, KDD'99, Reliability, User to Root.

Abstract
Today we are going up against with the issue of high dimensionality and outsized measure of information, network intrusion detection is always the focus of current research in the network security field. It is the spoiling of data security rules by pernicious exercises. Interruption discovery (ID) is a progression of strategies for distinguishing and perceiving incredulous activities that make the move acknowledgment of benchmarks of protection/classification, prominence, unwavering quality, and accessibility of a PC based system framework. The KDD Cup 99 dataset has been the purpose of fascination for some analysts in the field of interruption discovery from the most recent decade. Numerous scientists have contributed their endeavors to break down the dataset by various methods. It grants recognizing Denial of organization (DoS), User to root (U2R), Remote to login (R2L) and Probe assault. For the identification of interruption/dangers distinctive information mining calculation has been connected by different creators. In this paper, we present the literature study of the previous work done in the field of intrusion detection with their merits and demerits.

Introduction
In this associated world, at present time because of continually expanding interest of the web and web applications, the measure of movement coursing through the system has expanded altogether. As per report distributed by Cisco Worldwide web activity in 2012 has ended up tremendous, at 43.6 exabytes for every month and it will develop to achieve 120.6 exabytes for every month by 2017[1]. With developing system movement, system assaults are additionally been seen expanding massively.With growing network traffic, networkattacks are also been seen increasing enormously.The Association for Computing Machinery (ACM) has been gathered different network malicious and non-malicious behaviour data in a Knowledge Discovery and Data mining (KDD) platform [2] for the data mining understudies and experts. They have provided set KDD Cup99 data sets for network intrusion detection [3].Network Intrusions are defined as an attempt tocompromise the integrity or availability of computer ornetwork resources. Intrusion detectionsystems (IDSs) are software or hardware systems thatmechanize the procedure of scrutinize the eventsoccurring in a computer system or network, analyzing them for signs of security problems. Forthe analysis of the intrusion detection the dataset needto be analyzed and classification need to be done.Host-based IDS is used tomonitors the host and its objective is to detect the malicious activity on that hostonly by performed local analysis. Network-based IDS operates on network data for a segment of the network. It monitors the network to detecting malicious activity.
The misuse IDS works on the offline data and the other is Anomaly Detection which can detect any abnormal behavior and hence can work well on online data3. The KDD data set is a standard data set used for the research on intrusiondetection systems.

KDDCUP'99 Dataset
Since 1999, KDD'99 has been the most wildly used data set forthe evaluation of anomaly detection methods. This data set.and is built based on the data captured in DARPA'98 IDSevaluation program. DARPA'98 is about 4 gigabytes ofcompressed raw (binary) tcpdump data of 7 weeks of networktraffic, which can be processed into about 5 million connectionrecords, each with about 100 bytes. The two weeks of test datahave around 2 million connection records. KDD trainingdataset consists of approximately 4,900,000 single connectionvectors each of which contains 41 features and is labeled aseither normal or an attack, with exactly one specific attacktype [4].

Network Attack
The data set contains a total of 23 attack, these are groupedinto 4 major categories [5]:

1. Denial-of-Service (DoS)

In this type of attack, the attacker has limits or denies theservice presented to the user, computer or network. Attackertries to prevent genuine users from using a service. It isusually done by making the resources either too busy or toofull and overflow.

2. Probing or Surveillance

Probing or Surveillance attacks have the main aim of gainingknowledge of the existence or configuration of a computersystem or the network. The attacker then tries to harm orretrieve information about resources of the victim network.

3. User-to-Root (U2R)

User-to-root attack is attempts by an unauthorized user togain administrative privileges. The attacker starts outs withaccess to a normal user account on the system (perhaps gainedby sniffing password, a dictionary attack, or socialengineering) and is able to exploit some vulnerability to gainroot access to the system.

4. Remote-to-Local (R2L)

Remote-to-local attack is the kind of intrusion attack where theremote intruder consistently sends packets to a local machineover a network but who does not have an account on thatmachine exploits some vulnerability to gain local access as auser of that machine.

Table 1: Classwise attack on KDD'99 dataset

Class of Attack	Attack Name
Normal	Normal
DoS	Neptune, Smurf Pod, Teardro, Landback
Probe	Ipsweep, nmap, satan, portsweep
R2L	ftp_write, guess_passwd, imap, multihop, phf_spy
U2R	Perl, buffer_overflow, rootkit, loadmodule

The remaining section of the research work is arranged as follows:Section II describes the various data mining techniques to detect or monitor each and every activity performs over network. In III gives literature of the previous work for intrusion detection. Section IV Application of Intrusion Detection System and last section gives overall conclusion of the research work.

Data Mining Techniques For Ids

Back Propagation Neural Network (BPNN)

Researchers have found that the human ability of thinking,reasoning and learning can be imitated to some extent bycomputer [6]. Neural network has the ability to imitate somebehavior of human brain. Fish [6] also pointed out that "neural network is capable enough of approximate matching", where incomplete patterns could be recognized also. Neural network is composed of 'nodes' which are nothing but processing elements and some weighted connections betweenthe nodes. These nodes operate independently. BPNN is aspecial type of neural network. A BPNN has multiple layers.Each layer consists of one or more than one interconnected nodes with some 'activation function'. The left-most layer is known as 'Input layer' and the right-most layer is known as 'Output layer'. Between these two layers there may be one or more than one hidden layers. Patterns are presented as inputto the network via the input layer, which in turn communicatesto the hidden layers where the actual processing is donethrough a set of weighted connections. The network startswith a set of fresh pattern as input data and set of pre-definedweights in each connection. It works through a forwardcalculation from input

layer to output layer through hiddenlayers followed by a backward calculation from output toinput layer for error rectification by adjusting the old weightsin the connections. Every set of forward and backward operations are termed as single 'Epoch'. For every epoch, a fresh set of pattern is given to the network as input. Thenetwork is trained in this way with a training set for certainnumber of epochs. After the training phase, the network iscapable of identifying the unknown pattern according to itstraining [7].

Advantages

1. It has the ability of the neural network to "learn" the characteristics of misuse attacks
2. Flexibility that the network would provide

Disadvantages

1. It requires much training time to train the BPNN
2. It will increase the complexity of the system and hence will reduce the convergence rate

Random Forest

The random forest is an ensemble of unpruned classification or regression trees [8]. Random forest generatesmany classification trees and each tree is constructed by a different bootstrap sample from the original data usinga tree classification algorithm. After the forest is formed, a new object that needs to be classified is put downeach of the tree in the forest for classification. Each tree gives a vote that indicates the tree's decision about theclass of the object. The forest chooses the class with the most votes for the object. The random forests algorithm(for both classification and regression) is as follows [9] [10]:

1) From the Training of n samples draw ntree bootstrap samples.
2) For each of the bootstrap samples, grow classification or regression tree with the following modification:at each node, rather than choosing the best split among all predictors, randomly sample mtry of the predictorsand choose the best split among those variables. The tree is grown to the maximum size and notpruned back. Bagging can be thought of as the special case of random forests obtained when mtry = p, thenumber of predictors.
3) Predict new data by aggregating the predictions of the ntree trees (i.e., majority votes for classification, averagefor regression).

There are two ways to evaluate the error rate. One is to split the dataset into training part and test part. We canemploy the training part to build the forest, and then use the test part to calculate the error rate. Another way isto use the Out-of-Bag (OOB) error estimate. Because random forests algorithm calculates the OOB error duringthe training phase, therefore to get OOB error, we do not need to split the training data. In our work, we haveused both ways to evaluate the error rate.There are three tuning parameters of Random Forest: number of trees (ntree), number of descriptors randomlysampled as candidates for splitting at each node (mtry) and minimum node size [10]. When the forest is growing,random features are selected at random out of the all features in the training data. The number of features employedin splitting each node for each tree is the primary tuning parameter (mtry). To improve the performance ofrandom forests, this parameter should be optimized. The number of trees should only be chosen to be sufficientlylarge so that the OOB error has stabilized. In many cases, 500 trees are sufficient (more are needed if descriptor'simportance or intrinsic proximity is desired). In contrast to other algorithms having a stopping rule, inRF, there is no penalty for having "too many" trees, other than waste in computational resources. Another parameter,minimum node size, determines the minimum size of nodes below which no split will be attempted. Thisparameter has some effect on the size of the trees grown. In Random Forest, for classification, the default valueof minimum node size is 1, ensuring that trees are grown to their maximum size and for regression, the defaultvalue is 5 [10].

Advantages

1. It achieve high detection rate when false positive rate is low
2. Having low cost and possibility of scaling based on number of parallel nodes

Disadvantages

1. It is not able to detect novel attack
2. Produces overhead due to the dataset allocation

Conditional Random Fields for Intrusion Detection

Conditional models are probabilistic systems that are usedto model the conditional distribution over a set of randomvariables. Such models have been extensively used in thenatural language processing tasks. Conditional models offera better framework as they do not make any unwarrantedassumptions on the observations and can be used to modelrich overlapping features among the visible observations.Maxent classifiers [11], maximum entropy Markov models[12], and CRFs [13] are such conditional models. Theadvantage of CRFs is that they are undirected and are, thus,free from the Label Bias and the Observation Bias. Thesimplest conditional classifier is the Maxent classifier basedupon maximum entropy classification, which estimates theconditional distribution of every class given the observations[11]. The training data is used to constrain this conditionaldistribution while ensuring maximum entropy and hencemaximum uniformity. We now give a brief description of theCRFs, which is motivated from the work in [13].

Advantages

1. Avoid observation bias and Label bias problem
2. This allow model arbitrary relationship among different features

Disadvantages

1. It is not the distribution of interest, since the observations are completely visible and the interest is in finding the correct class for the observations.
2. Inferring the conditional probability from the modeled joint distribution, using the Bayes rule, requires the marginal distribution.

Genetic Algorithm

Genetic algorithms [14] are employed as chromosome-like data structures. Figure 3 adopted from represent the structure and processing in a genetic algorithm. A genetic algorithm has various parameters, operators and processes which decide its arrival to an optimal solution. A short description of theparameters, operators and processes as depicted in figure 3, isas: Fitness Function: The fitness function is the measure of thesuperiority of a meticulous solution. The fitness function isused to conclude the mainly optimal solution from a numberof solutions in a population. Selection: This process in geneticalgorithms is used to opt for the most optimal solutiondetermined by using the fitness function. The solutions whichare not most favorable are discarded. Crossover: The crossoverprocedure in genetic algorithms is used to substitutecharacteristics among two dissimilar solutions. The pairs ofsolutions to swap characteristics are selected randomly andremain exchanging characteristics, until a completely newgeneration of solutions is obtained. Mutation: The mutationprocess in genetic algorithms transforms some random bits ina solution. The modification in the bits results in the geneticdiversity of the mutated algorithms.

Advantages

1. The detection rate of this is very high
2. False alarm is minimal if the fitness function is doing well

Disadvantages

1. It cannot locatethe attack in audittrail
2. It cannot detectnovel attacksas it requires moredomain specificknowledge
3. No capability to perceivemultipleSimultaneous andit is complex todesign.

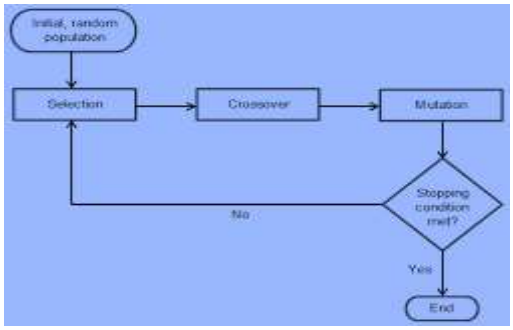

Fig. 1 Flowchart for Genetic Algorithm

k-NN: k-Nearest Neighbor

k-NN classification is an easy to understand and easy to implement classification technique[15]. Despite its simplicity, it can perform well in many situations. K-NN is particularly well suited for multi-modal classes as well as applications in which an object can have many class labels.

For example, for the assignment of functions to genes based on expression profiles, some researchers found that k-NN outperformed SVM, which is a much more sophisticated classification scheme. The 1-Nearest Neighbor (1NN) classifier is an important pattern recognizing method based on representative points [16]. In the 1NN algorithm, whole train samples are taken as representative points and the distances from the test samples to each representative point are computed. The test samples have the same class label as the representative point nearest to them. The k-NN is an extension of 1NN, which determines the test samples through finding the k nearest neighbors.

Advantages

1. It leads to a very simple to use and design.
2. Robust to noisy training dataset

Disadvantages

1. Slow as you need to scan entire training data to make each prediction.
2. It's really difficult to stay local because of the curse of dimensionality. In high dimension

Support Vector Machine

The SVM is already known as the best learning algorithm for binary classification. The SVM, originally a type of pattern classifier based on a statistical learning technique for classification and regression with a variety of kernel functions, has been successfully applied to a number of pattern recognition applications. Recently, it has also been applied to information security for intrusion detection. Support Vector Machine has become one of the popular techniques for anomaly intrusion detection due to their good generalization nature and the ability to overcome the curse of dimensionality .Another positive aspect of SVM is that it is useful for finding a global minimum of the actual risk using structural risk minimization, since it can generalize well with kernel tricks even in high-dimensional spaces under little training sample conditions. The SVM can select appropriate setup parameters because it does not depend on traditional empirical risk such as neural networks [17]. One of the main advantage of using SVM for IDS is its speed, as the capability of detecting intrusions in real-time is very important. SVMs can learn a larger set of patterns and be able to scale better, because the classification complexity does not depend on the dimensionality of the feature space. SVMs also have the ability to update the training patterns dynamically whenever there is a new pattern during classification [18].

Advantages

1. Its speed, as the capability of detecting intrusions in real-time is very significant
2. It has the ability to update training pattern dynamically

Disadvantages

1. SVM is computationally costly for resource-limited ad hoc network
2. It is complex to design and decreases the accuracy of detection Intrusion

Review Of Literature

Senthilkumar and Shona[19], proposed work weighted minkowski based Firefly algorithm is applied to eliminate redundant record set and enhanced KNN based imputation method with the help of bagging technique to handle missing value is introduced. The experimental results shows that after preprocessing there is more improvement in the accuracy of learning algorithm during classification of normal and abnormal packets.

Yan et al. [20],focusing at false negative rate and false alarm rate which exist generally in the intrusion detection system. They have proposed an intelligent intrusion detection model. Based on the characteristics of global superiority of genetic algorithm and locality of nerve, the model optimizes the weights of the neural network using genetic algorithm. Their experiment results show that the intelligent way can improve the efficiency of the intrusion detection.

Waghet al. [21], proposed Network security is a key a portion of web enabled systems in the present world circumstance. According to the makers as a result of confusing chain of PCs the open entryways for intrusions and strikes have extended. Along these lines it is need of extraordinary significance to find the best courses possible to secure our structures. So the makers propose intrusion distinguishing proof structure is expecting essential part for PC security. The best methodology used to handle issue of IDS is machine learning. Thy watched that the rising field of semi managed learning offers an ensured course to correspond investigation. So they proposed a semi-oversaw framework to reduce false ready rate and to improve revelation rate for IDS.

Ambusaidi et al.[22] , considered the feature selection problem for data classification in the absence of data labels. It first proposed an unsupervised feature selection algorithm, which is an enhancement over the Laplacian score method, named an Extended Laplacian score, EL in short. Specifically, two main phases are involved in EL to complete the selection procedures. In the first phase, the Laplacian score algorithm is applied to select the features that have the best locality preserving power. In the second phase, EL proposes a Redundancy Penalization (RP) technique based on mutual information to eliminate the redundancy among the selected features. This technique is an enhancement over Battiti's MIFS. It does not require a user defined parameter such as β to complete the selection processes of the candidate feature set as it is required in MIFS. After tackling the feature selection problem, the final selected subset is then used to build an Intrusion Detection System. The effectiveness and the feasibility of the proposed detection system are evaluated using three well-known intrusion detection datasets: KDD Cup 99, NSL-KDD and Kyoto 2006+ dataset. The evaluation results confirm that our feature selection approach performs better than the Laplacian score method in terms of classification accuracy.

Gupta et al. [24],addresses these two issues of Accuracy and Efficiency using Conditional Random Fields and Layered Approach. We demonstrate that high attack detection accuracy can be achieved by using Conditional Random Fields and high efficiency by implementing the Layered Approach. Experimental results on the benchmark KDD '99 intrusion data set show that our proposed system based on Layered Conditional Random Fields outperforms other well-known methods such as the decision trees and the naive Bayes. The improvement in attack detection accuracy is very high, particularly, for the U2R attacks (34.8 percent improvement) and the R2L attacks (34.5 percentimprovement). Statistical Tests also demonstrate higher confidence in detection accuracy for our method. Finally, we show that our system is robust and is able to handle noisy data without compromising performance.

Altwaijry and Algarny [24], presented an intrusion detection system is developed using Bayesian probability. The system developed is a naive Bayesian classifier that is used to identify possible intrusions. The system is

trained a priori using a subset of the KDD dataset. The trained classifier is then tested using a larger subset of KDD dataset. The Bayesian classifier was able to detect intrusion with a superior detection rate.

Janakiraman and Vasudevan [25], presented an intelligent learning approach using Ant Colony Optimization (ACO) based distributed intrusion detection system to detect intrusions in the distributed network. The experimental results on the proposed system with the feature extraction algorithm is effective to detect the unseen intrusion attacks with high detection rate and recognize normal network traffic with low false alarm rate.

Thomasa et al.[26],presented two hybrid approaches for modeling IDS. Decision trees (DT) and support vectormachines (SVM) are combined as a hierarchical hybrid intelligent system model (DT–SVM)and an ensemble approach combining the base classifiers. The hybrid intrusion detectionmodel combines the individual base classifiers and other hybrid machine learning paradigmsto maximize detection accuracy and minimize computational complexity. Empirical resultsillustrated that the proposed hybrid systems provide more accurate intrusion detection systems.

Varma et al. [27],presented an overview of intrusion detection system and a hybridtechnique for intrusion detection based on Bayesian algorithmand Genetic algorithm. Bayesian algorithm classifies the datasetinto various categories to identify the normal/ attacked packetswhere as genetic algorithm is used to generate a new data by applyingmutation operation on the existing dataset to produce a newdataset. Thus this algorithm classifies KDD99 benchmark intrusiondetection dataset to identify different types of attacks with highdetection accuracy. The experimental result also shows that the accuracyof detecting attacks is fairly good.

Conclusion

KDDCUP'99 is widely used dataset for detection of intrusion over internet. Intrusion detection is a critical problem in the networktechnology and lots of work has been done for conformingthe safekeeping. The attack are classified into different categories such as DoS, U2R, R2L and probe etc. Various researchers uses supervised and unsupervised learning algorithm of data mining. This paper presents the related work for the detection of intruders and various data mining techniques which helps in the detection of intruders with their advantages and disadvantages. In this some of the techniques is able to detect only known attack and has low detection rate. So in future, it is necessary to implement such system which enhances the performance, requires less training and reduces the overhead on the network.

References

[1] Cisco, Cisco Visual Networking Index: Forecast and Methodology, 2012-2017, Cisco, 2013.
[2] Tarakanov AO, Kvachev SV, Sukhorukov AV. "A formal immune network and its implementation for on-line intrusion detection", In MMM-ACNS 2005 (pp. 394-405).
[3] Farhaoui Y. "How to secure web servers by the intrusion prevention system (IPS)", International Journal of Advanced Computer Research. 2016; 6(23); 65-71.
[4] KDD Cup 1999. (2014, Nov.) [Online]. Available: http://kdd.ics.uci.edu/databases/kddcup99/
[5] MIT Lincoln Labs. (2014, Nov.). DARPA intrusion detection evaluation[Online]. Available: http://www.ll.mit.edu/mission/communications/ist/corpora/ideval/index.html.
[6] B. D. Fish, "Artificial Intelligence (AI): Intrusion Analysis", Encyclopedia of Information Assurance, 2011, pp. 52-162.
[7] S. Haykin, "Neural Networks – A Comprehensive Foundation", 2nd Edition, Pearson.
[8] Breiman, L. (2001) Random Forests. Machine learning, 45, 5-32. http://dx.doi.org/10.1023/A:1010933404324
[9] Liaw, A. and Wiener, M. (2002) Classification and Regression by Random Forest. R News, 2, 18-22.
[10] Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P. and Feuston, B.P. (2003) Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. Journal of Chemical Information and Computer Sciences, 43, 1947-1958. http://dx.doi.org/10.1021/ci034160g.
[11] A. Ratnaparkhi, "A Maximum Entropy Model for Part-of-Speech Tagging," Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP '96), pp. 133-142, Assoc. for Computational Linguistics, 1996.

[12] A. McCallum, D. Freitag, and F. Pereira, "Maximum Entropy Markov Models for Information Extraction and Segmentation," Proc. 17th Int'l Conf. Machine Learning (ICML '00), pp. 591-598, 2000.
[13] J. Lafferty, A. McCallum, and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," Proc. 18th Int'l Conf. Machine Learning (ICML '01), pp. 282-289, 2001.
[14] Parry GowherMajeed and Santosh Kumar, " Genetic Algorithms in Intrusion Detection Systems: A Survey", International Journal of Innovation and Applied Studies, ISSN 2028-9324 Vol. 5 No. 3 Mar. 2014, pp. 233-240.
[15] S. Thirumuruganathan , "A Detailed Introduction to K-Nearest Neighbor (KNN) Algorithm", World Press, May 17, 2010.
[16] X Wu, V Kumar, J Ross Quinlan, J Ghosh, "Top 10 Data mining Algorithm", Knowledge and Information Systems, Volume 14, Issue 1, pp 1-37 ,2008 – Springer
[17] T.Shon, Y. Kim, C.Lee and J.Moon,(2005), A Machine Learning Framework for Network Anomaly Detection using SVM and GA, Proceedings of the 2005 IEEE.
[18] SandyaPeddabachigari, Ajith Abraham, CrinaGrosan, Johanson Thomas (2005). Modeling Intrusion Detection Systems using Hybrid Intelligent Systems.Journal of Network and Computer Applications.
[19] D. Shona, M. Senthilkumar, "An Ensemble Data Preprocessing Approach for Intrusion Detection System Using variant Firefly and Bk-NN Techniques", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 6 (2016) pp 4161-4166.
[20] Yan C. "Intelligent Intrusion Detection Based on Soft Computing" In seventh international conference on measuring technology and mechatronics automation (ICMTMA) 2015, (pp. 577-80). IEEE
[21] Wagh SK, Kolhe SR. Effective intrusion detection system using semi-supervised learning. In international conference on data mining and intelligent computing (ICDMIC) 2014 (pp. 1-5), IEEE.
[22] Mohammed A. Ambusaidi, Xiangjian He and Priyadarsi Nanda "Unsupervised Feature Selection Method for Intrusion Detection System, 2015 IEEE Trustcom/BigDataSE/ISPA.
[23] Kapil Kumar Gupta, BaikunthNath and RamamohanaraoKotagiri "Layered Approach Using Conditional Random Fields for Intrusion Detection", IEEE Transactions on Dependable And Secure Computing, Vol. 7, No. 1, January-March 2010.
[24] HeshamAltwaijry, SaeedAlgarny "Bayesian based intrusion detection system", Journal of King Saud University – Computer and Information Sciences (2012) 24, 1–6.
[25] S. Janakiraman, V. Vasudevan, "ACO based Distributed Intrusion Detection System", International Journal of Digital Content Technology and its Applications Volume 3, Number 1, March 2009.
[26] SandhyaPeddabachigaria, AjithAbrahamb, CrinaGrosanc, Johnson Thomasa "Modeling intrusion detection system using hybrid intelligent systems", Journal of Network and Computer Applications 30 (2007) 114–132-Elsevier.
[27] Y V Srinivasa Murthy, Kalaga Harish and D K Vishal Varma "Hybrid Intelligent Intrusion Detection System using Bayesian and Genetic Algorithm (BAGA): Comparitive Study", International Journal of Computer Applications (0975 8887) Volume 99 - No. 2, August 2014