Big Data Analysis for privacy preservation on VoIP using SVM and KNN Classifier

Dr. Rohit Kumar Singhal
Professor
Department of CSE
IET, Alwar

Deepika Sharma
M.Tech Scholar
Department of CSE
IET, Alwar

ABSTRACT

Networked data contain interconnected entities for which inferences are to be made. For example, web pages are interconnected by hyperlinks, research papers are associated by references, phone accounts are linked by calls, and conceivable terrorists are linked by communications. Networks have turned out to be ubiquitous. Correspondence networks, financial transaction networks, networks portraying physical systems, and social networks are all ending up noticeably progressively important in our everyday life. Regularly, we are interested in models of how nodes in the system influence each other (for example, who taints whom in an epidemiological system), models for predicting an attribute of intrigue in light of observed attributes of objects in the system. The technique of SVM is applied which will classify the data into malicious and non-malicious.

VoIP Traffic

VoIP stands for Voice over Internet Protocol that uses internet or other data network rather than using conventional Public Switched Telephone Network (PSTN). A rapid growth has been seen in use of internet for voice communications that results in reduce cost of equipment, operation and maintenance. The VoIP is a solid technology that allows people to communicate through voice using IP protocol instead of telephone lines. The property standards, high price tag, limited integration with existing telephony environments are some of the factors that have assigned this technology in a niche market. Now a day's situation has been changed due to advent of asterisk as well as low-cost VoIP telephone adapters open source tools. This has become easy and common for internet providers to provide their customers VoIP calls at very low cost, if any in addition to standard xDSL connectivity. Advancement in VoIP also directs the development of convergent networks that support both video and voice services not presented by conventional PSTN. Though VoIP is low cost or almost free technology still various telecom operators try to conceal VoIP traffic intentionally to avoid detection and escape from taxes i.e. Access Promotion Charge (APC) by altering different parameters in VoIP packets [4].

Introduction of Intrusions

A communication platform through which the interaction and transmission of information is provided amongst users is known as a social network. Today, social networks are practically included in each domain with respect to one way or another. The services involved in education, business, excitement etc. all these applications include social networks. For example, there are several such business as well that advertise their brands and products on the social networking platforms such that the products that they wish to sell can gain popularity. The popularity has increased due to the increase in positive feedbacks given online and the flexibility of utilizing these products available. The patterns are carried on and obeyed by the malicious users in very unique manner such that they can become the part of secure networks. For example, as per the association rule, a genuine user sends the messages to normal user on regular basis. However, the data is transmitted randomly to selected audiences only by the malicious users. Further, the track of unpredicted friends in profiles generated in few social network websites such as Facebook and Google+ can confirm that the user is malicious or not. The presence of superfluous friend requests on these websites can also be an indication of malicious users. Any unusual action that shows a different behavior of one user against all other users present in similar platform confirms the presence of an intrusion [7]. There are several studies proposed on the identification of such malicious users and there are several names given for it such as an outlier, abnormality, and anomaly and so on. However, the noise present in the data is not similar to an intrusion in the systems. A noise within the data is a random error or variance that is caused in a variable and the examining of data does not affect it much. For example, the individual's purchasing

activities can be taken as criteria to detect the credit card faults using the behavior. The noise within the data is initially removed before identifying any kinds of intrusions in the systems. A system in which the unobserved new patterns are identified within the data is known as intrusion detection. Initially, these points can be considered similar to the existing points. However, with the detection of new points, these intrusions can be identified. There are several issues caused when an intrusion occurs within a system which are needed to be solved quickly. For example, a set of false identities might be generated by some malicious users due to which the communication of malicious users with genuine users is possible. Thus, it is very important to identify such intruders such that any kinds of losses can be prevented.

Approaches of Intrusion Detection

Clustering Approaches

There is huge amount of data that is gathered and stored within databases all across the world. The data will increase with each year. Within the enterprises and research offices, the databases with Terabytes are difficult to be discovered. Within such databases, there is invaluable data and knowledge hidden and the extraction of such information is impossible without using any automatic strategies. The mining of such information is impossible. For the extraction of data from huge databases, several algorithms have been proposed over the years. Classification, association rule, clustering, etc. are some of the methodologies utilized here. The prediction of particular result on the basis of given input is known as classification. For the prediction of result, a preparation set that includes a set of attributes and respective result is generated which is known as ordinarily called prediction attributes [10]. The result can be predicted with the help of algorithm that identifies relationships amongst the attributes. The data set known as prediction set, that is not seen lately can be provided within the next algorithm. The prediction attribute that is not yet known is also identified here. The input is investigated and prediction is generated by the algorithm. The manner in which the "good" algorithm is to be defined is known as prediction accuracy.

On the basis of classified data that is created from set of data that has previously known exact classes is known as supervised learning. The classification method is included in several scenarios. For example, lesser information related to financial and other personal data is assigned and diseases are treated here. With the advancement in technology, there are several problems arising each day. The statistical, machine learning and neural network are the three major categorizations included here. On the basis of technologies, there are various goals included. On the basis of each technology various issues are also highlighted here.

a. Statistical Procedure Based Approach: The classical and modern phase is two different phases within the statistical classification. In the classical phase, the linear discrimination is improved. Within the modern phase, the flexibility of classes of models is higher. For each class, the estimation related to joint distribution of features is provided. Using this, the classification rule of the system is defined. To categorize the statistical procedures, a precise fundamental probability model is used. The limited classification is provided through probability and in each class, this probability is provided. Various techniques are utilized by several statisticians and a variable and transformation is provided by the humans. Thus, the whole system faces issues during the structuring phase [11].

b. Machine Learning Based Approach: In the machine learning process, the automatic computing procedures are included using logical or binary operations. Various sets of examples relevant to existing work are provided by these steps. Sequences of steps are provided for classification on the basis of decision tree approaches. Complex issues are properly represented using particular data with this type of classification method. Several enhancements have been made within some of the approaches like genetic algorithms and LDP techniques. There are different types of attributes required to deal with more genuine data types. Humans can easily understand the classification of expressions using machine learning techniques. Within background, the statistical approaches are included in the development process. In this operation, humans are not involved however.

c. Neural Network: The understanding and emulation of human brain along with high range of issues relevant to copying human properties includes the presence of huge range of sources. The human properties are copied here within these issues. For classifying the data as being intrusive or normal, there are various fields available in this technique. A non-linear function is generated by each node from provided input on the basis of various interconnected nodes available within neural networks. For providing data to other nodes, the input data or other nodes are available. With the help of other nodes present in the network, the output of network is also provided. Within the complete applications that include neural networks, several patterns and decisions related to them are identified. To enhance

the controls of plane within proper way in the autopilot modes of airplanes, several tools and output units are utilized. Neural networks are used for all such tasks. These systems also help in maintaining the quality control.

Literature Survey

Mazhar Rathore, et.al, (2016), have analyzed that telecommunication authorities and Internet service providers (ISPs) are interested in detecting VoIP calls either to block illegal commercial VoIP or prioritize the paid users VoIP calls. Signature-based, port-based, and pattern-based VoIP detection techniques are not more accurate and not efficient due to complex security and tunneling mechanisms used by VoIP. In this paper [1], authors have proposed a new scheme based on generic rule, robust and efficient statistical analysis that helps in identify encrypted, non-encrypted, tunneled VoIP media flows using traditional approach. It meets the need of any organization to detect VoIP flows to either prioritize or block. They have tested their solution on many traces of more than 10 VoIP applications.. This technique has 97.54% TP and .00015% FP. It is the better choice for telecommunication authorities and ISPs to detect VoIP calls in high-speed big Data environment.

Muhammad Shafiq, et.al, (2016), has recommended network traffic classification as a central topic for researchers in the field of computer science. The most common technique used these days is Machine Learning (ML) technique. This is used by many researchers and got very effective accuracy results. In this paper [2], authors have discussed step by step techniques of network traffic classification and develop a real time internet data set using network traffic capture tool. Then the features are extracted from the capture traffic using tools of feature extraction the applied a Support Vector Machine, C4.5 decision tree, Naive Bays and Bayes Net machine learning classifiers. The experimental and simulation results show that C4.5 classifiers prove to be good in terms of accuracy as compared to other existing classifiers.

Aboagela Dogman, et.al, (2014), have presented managing quality of service (QoS) as a important network operation mainly in hybrid wired and wireless multimedia networks. In this paper [3], authors given a reviewed and developed an approach based on two stages to intelligently manage QoS for multimedia traffic. As a typical multimedia application they have considered VoIP and applied an adaptive statistical sampling technique in initial stages. In order to assess the VoIP provided for QoS the FCM information is used by multilayer perceptron (MLP) neural network. The simulation results show that traffic are represented more correctly by developed adaptive statistical sampling than the systematic, stratified and random non-adaptive sampling methods. The combination of statistical sampling followed by FCM and MLP are more accurately indicated the QoS for VoIP.

Jaiswal Rupesh Chandrakant, et.al, (2013), have analyzed that internet traffic recognition techniques has become very important for researchers because these techniques are independent of TCP or UDP port numbers. In traffic recognition ML techniques has been used which are the subset of artificial intelligence. The Classification, clustering, Numeric prediction and Association are the four types of Machine Learning. In this paper [4], authors have implemented traffic recognition through classification process. AdaboostM1, C4.5, Random Forest tree, MLP, RBF and SVM are six ML algorithms that are used for IP traffic classification with Polykernel function classifiers. The simulation and implementation results show that Tree based algorithm are more effective ML techniques for internet traffic classification in terms of achieved accuracy of 99.7616%.

Riyad Alshammari, et.al, (2015), have analyzed the performance of C5.0, AdaBoost and Genetic programming (GP) like three different machine learning algorithms that generate robust classifiers to identify VoIP encrypted traffic. In this paper [6], authors have found it very challenging to find robust rules specifically to detect encrypted VoIP Skype network traffic. The authors have investigated how to form a training set when machine learning based approach is used for classifying network traffic without including port numbers, IP addresses, or payload information. Given the results obtained in this research paper, one of the future directions which can be followed would be to explore whether a similar trend for other network applications.

Research Methodology

This work is based on the network traffic classification to classify the traffic into malicious, non-malicious. The network traffic analysis is the technique which is applied to predict the malicious activities of the users which are active on the network. To classify the network traffic three steps has been followed in the methodology, in the first step technique of k-mean clustering is been applied in which similar and dissimilar type of data will clustered. The dataset which is taken as input will be refined by removing redundancy and missing values. In the second step, technique of k-mean clustering is applied in which arithmetic mean of the whole dataset is calculated which will be the central point of

the dataset. The Euclidian distance from the central point is calculated which define the similarity and dissimilarity of the points. The points which are similar will be clustered in one cluster and other in the second cluster. In the last step of classification technique, SVM classifier will be applied which classify the data into two classes. To improve the performance of the existing system technique of Knn classifier will be applied which will cluster the uncluttered points and increase accuracy of classification. The Knn classifier the nearest neighbor classifier in which Euclidian distance is calculated and points which have similar distance will be clustered in one class and other in the second class.

Support Vector Machine

In order to perform text categorization, popularly used mechanism is the SVM classification method which has a predictive model. The data is taken as input here and classified data is given in two categories as output. For the text corpus in which each training example belongs to one of the two classes, a model is implemented best by using SVM training algorithm. Further, by constructing N-Dimensional hyperplane, the data is partitioned into two categories. In order to separate the data, two parallel hyper planes are generated on each side of the hyper plane. Here, the distance between the two hyper planes is maximized through the separation of hyper plane. In correspondence to the partitioning hyper plane f(X) which passes across the middle of two classes and divides them, there is a linearly separable data set for which a linear classification function is created. The classification of a new data instance, Xn, is done very easily through the testing of sign of function f(Xn) once the function is determined:

Where Xn belongs to a positive class if f(Xn) > 0

For larger distance or margin, the error of the classifier can be generalized in better way. On the high dimensional feature set, this algorithm performs well and the kernel trick is utilized for creating a new linearly separable data through the transformation of non-linearly separable data. In order to perform numerical calculations and also to calculate the regression analysis, SVM can be utilized. Further, the elements can be ranked with the help of this algorithm as well. Another benefit of SVM is that the performance of SVM on the datasets that include numerous attributes is very good even through only specific cases can be accessed for training purpose. However, during the training and testing phase of SVM, the speed and size might be the issues. Also, choosing the kernel function parameters is not an easy task and thus is a disadvantage of this approach.

K nearest Neighbor

One of the simplest algorithms amongst all the learning machine algorithms is the K-Nearest Neighbor (KNN) algorithm. Since there are no assumptions made on the underlying data distribution, KNN is known to be a non-parametric supervised learning algorithm. Here, on the basis of nearest training samples present within the feature space, the samples are classified. The feature vectors are stored along with the labels of training pictures within the training process. Towards the label of its k-nearest neighbors, the unlabelled question point is doled out during the classification process. Through majority share cote, on the basis of labels of its k nearest neighbors, the object is characterized. The object is classified essentially as the class of the object that is nearest to it in the event when k=1. k is known to be an odd integer in case when there are only two classes present. During the performance of multiclass categorization, there can be tie in case when k is an odd whole number. The classification of samples on the basis of majority class of its nearest neighbor is the major task of KNN algorithms.

$$Class = arg_v max \sum_{(X_i, y_i) \in D_z} I(v = y_i) \quad \ldots\ldots(1)$$

Here, the class label is represented by v. The class label for i^{th} nearest neighbors is denoted by y_i. An indicator function is denoted by I, in which if the argument is true, the value of 1 is returned and otherwise, 0 value is returned. Thus, within the class of its K nearest neighbors, the samples are assigned. A set of labeled objects, a distance or similarity metric that calculates the distance amongst objects and the number of nearest neighbors that is the value of k, are the three important elements within the KNN approach. In order to make the recognition task successful, the selection of an appropriate similarity function as well as value for parameter k is important. For understanding as well as implementation of classification techniques, KNN classification is known to be simple and easy.

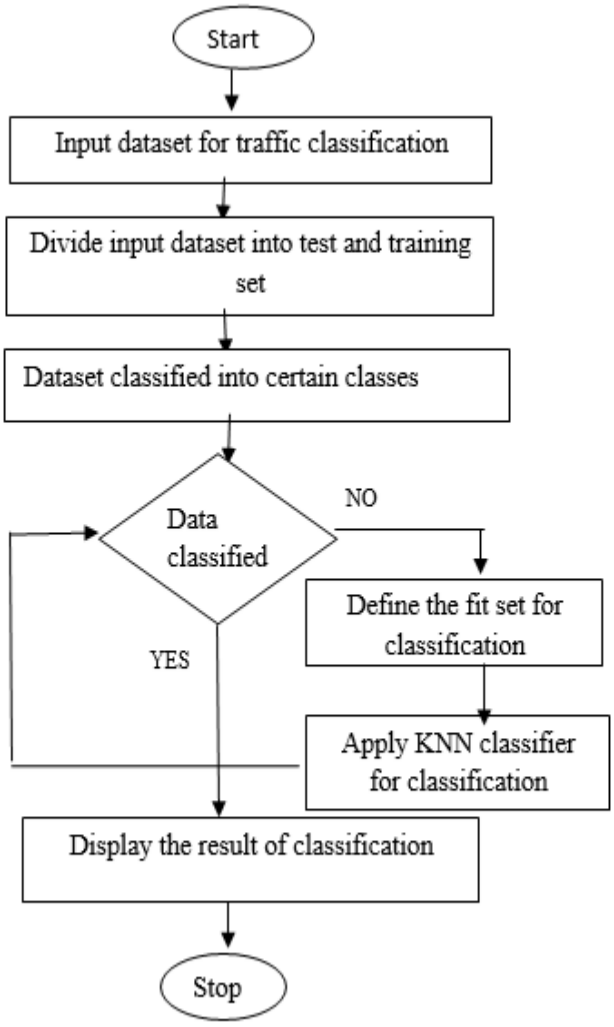

Fig 1: Flow chart of proposed method

Result and Discussion

The proposed and existing algorithms are implemented using anaconda python and results are analyzed in terms of accuracy, execution time

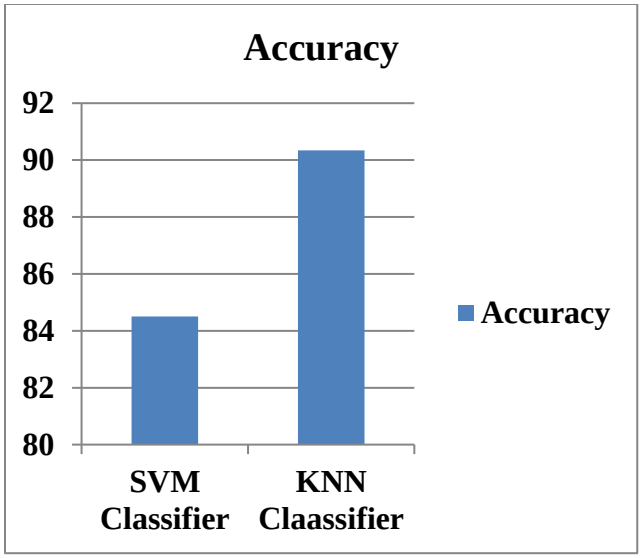

Fig 2. Accuracy Comparison

As shown in figure 2, the value of accuracy of SVM classifier is compared with the KNN classifier for the network traffic classification. It is been analyzed that accuracy of KNN classifier is high as compared to SVM classifier

Fig 2. Execution Time

As shown in figure 2, the execution time of the proposed algorithm is compared with the existing

algorithm. It is analyzed that execution time of KNN classifier is less as compared to SVM classifier

CONCLUSION

Data classification is an important task in machine learning. It is identified with develop computer programs ready to gain from labeled data sets and, in this way, to predict unlabeled instances. Because of the vast number of applications, numerous data classification systems have been developed. In the past years various techniques have been proposed which are based on attribute type of privacy preservation. Due to this technique when data is not in the structured form security of the data get compromised. Hence, SVM and KNN classifiers have been used which ensure data privacy of unstructured data in terms of accuracy and execution time. Accuracy of KNN classifier is higher than SVM classifier and SVM takes more time to execute as compared to KNN.

References

[1] M. Mazhar, U. Rathore, "Threshold-based generic scheme for encrypted and tunneled Voice Flows Detection over IP Networks", Journal of King Saud University Computer and Information Sciences, vol. 27, pp. 305–314, 2015.

[2] Muhammad Shafiq, Xiangzhan Yu, Asif Ali Laghari, Lu Yao, N abin Kumar Karn, F oudil Abdessamia, "Network Traffic Classification Techniques and Comparative Analysis Using Machine Learning Algorithms", 2016 2nd IEEE International Conference on Computer and Communications, vol. 8, pp. 2451-2455, 2016.

[3] Aboagela Dogman, Reza Saatchi, "Multimedia traffic quality of service management using statistical and artificial intelligence techniques", The Institution of Engineering and Technology 2014, vol. 8, pp. 367–377, 2014.

[4] Jaiswal Rupesh Chandrakant, Lokhande Shashikant. D., "Machine Learning Based Internet Traffic Recognition with Statistical Approach", 2013 Annual IEEE India Conference (INDICON), vol. 7, pp. 121-126, 2013.

[5] Uzma Anwar, Ghulam Shabbir, Malik Ahsan Ali, "Data Analysis and Summarization to Detect Illegal VOIP Traffic with Call Detail Records", International Journal of Computer Applications (0975 – 8887), vol. 89, pp. 1-7, 2014.

[6] Riyad Alshammari, A. Nur Zincir-Heywood, "Identification of VoIP encrypted traffic using a machine learning approach", Journal of King Saud University – Computer and Information Sciences, vol. 27, pp. 77–92, 2015.

[7] I. Martinez-Yelmo, I. Seoane, and C. Guerrero, "Fair QoE measurements related with networking technologies," WWIC 2010, LNCS 6074, Springer-Verlag Berlin Heidelberg, pp. 228–239, 2010.

[8] O. Hersent, J.P. Petit, and D. Gurle, "Beyond VoIP Protocols. Understanding Voice Technology and Networking Techniques for IP Telephony," John Wiley & Sons Ltd, 2005.

[9] C. Olariu, J. Fitzpatrick, P. Perry, and L. Murphy, "A QoS based call admission control and resource allocation mechanism for LTE femtocell deployment," in Consumer Communications and Networking Conference (CCNC), 2012 IEEE. IEEE, 2012, pp. 884–888.

[10] M. Afaq, S. U. Rehman, and W. C. Song, "Visualization of elephant flows and qos provisioning in sdn-based networks," in Network Operations and Management Symposium (APNOMS), 2015 17th Asia-Pacific, Aug 2015, pp. 444–447.

[11] C. Xu, B. Chen, and H. Qian, "Quality of service guaranteed resource management dynamically in software defined network," Journal of Communications, vol. 10, no. 11, 2015.

[12] M. F. Bari, S. R. Chowdhury, R. Ahmed, and R. Boutaba, "Policycop: an autonomic qos policy enforcement framework for software defined networks," in Future Networks and Services (SDN4FNS), 2013 IEEE SDN for. IEEE, 2013, pp. 1–7.

[13] Z. A. Qazi, J. Lee, T. Jin, G. Bellala, M. Arndt, and G. Noubir, "Application-awareness in sdn," in ACM SIGCOMM Computer Communication Review, vol. 43, no. 4. ACM, 2013, pp. 487–488.