Privacy Preserving Big Data publishing- A scalable K-anonymization approach using MapReduce

Deepika Sharma
Research Scholar
Department of CSE
IET Alwar

Rohit Kumar Singhal, PhD
Professor
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. In the previous study authors proposed various model for privacy preserving which are group based records, K-anonymity etc. In the existing models there are various problems like it affect data utilities, harm the data identifiers. In the research work, the hybrid approach has been designed to ensure data privacy which is based on attribute and data identifiers.

Keywords
VoIP, SVM, KNN, Intrusions.

1. INTRODUCTION
In today's age of information, big data has become a big game changer in most or all types of modern industries over last few years. It has brought revolution in the world of big data analytics. Numerous opportunities have been created by big data for researchers to process huge amount of data generated. This data does not have fixed data types and sensitive information is stored in this data due to which its privacy is needed to be ensured. Since big data consists of both structured and unstructured data, mature analytic tools exist for structured data but analytic tools for mining unstructured data are nascent and developing. In this paper, Network traffic classification of VoIP traffic has been done which contains huge amount of unstructured and unpreserved data. To preserve the privacy of this data, we have proposed SVM and KNN classifiers which are based on prediction analysis of data classification. Existing privacy preserving techniques like anonymization requires having dataset divided in the sets of attributes like Sensitive Attributes, Quasi-Identifiers and non-sensitive attributes. The proposed techniques of SVM and KNN will classify the data into malicious and non-malicious.

Data Analysis can be defined as the process of reviewing and evaluating the data that is gathered from different sources. Data cleaning is very important as this will help in eliminating the redundant information and reaching to the accurate conclusions. Data analysis is the systematic process of cleaning, inspecting and transforming data with the help of various tools and techniques. The objective of data analysis is to identify the useful information which will support the decision-making process. There are various methods for data analysis which includes data mining, data visualization and Business Intelligence. Analysis of data will help in summarizing the results through examination and interpretation of the useful information. The statistical techniques are divided into parametric test and non-parametric test [9].

2. 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 [7].

3. 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. 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 [10]. 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.

3.1 Approaches Of Intrusion Detection

Clustering Approaches

As the data is increasing rapidly each year a huge amount of data is gathered and stored within databases all across the world. 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 [11]. 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.

4. LITERATURE SURVEY

Youyang Qu, Shui Yu, et.al (2017) proposed **Big Data Set Privacy Preserving through Sensitive Attribute-based Grouping.** Attacks on database privacy are increasing as great value of privacy information stored in big data set. Public's privacy are under threats as adversaries are continuously cracking their popular targets such as bank accounts. Earlier models such as k-anonymity, group record based on quasi-identifiers harmed data utility a lot. Due to this, they proposed sensitive attribute based privacy model. In this paper [1], authors have proposed three contributions as: Firstly, grouping data according to sensitive attributes rather than quasi-identifiers. This method maintains the marginal distribution unchanged to keep data utility. Secondly, random shuffle is introduced into every group. With this, the records have the maximum entropy so that it is harder for an adversary to breach the privacy. Lastly, they build the mathematical model. This model decouples the correlations among attributes properly to find a trade-off between privacy and data utility.

K. Sree Divya1, et.al (2018) explained machine learning algorithms in big data analytics and machine learning challenges to take decisions where there is no known 'right path' for the specific problem. Machine Learning describes how to assemble a process framework that enhances consequently through experience. Machine Learning issue refers to the issue of learning from past experience with respect to some tasks and performance measures. In this paper [2], authors have described the types of big data i.e. structured and unstructured, machine learning algorithms include parametric and non-parametric algorithms and tools for big data analytics include Map-Reduce frameworks, Apache spark, Python, etc.

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 [3], 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 [4], 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 [5], 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 [6], 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 [8], 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.

5. RESEARCH METHODOLOGY

This work is based on the network traffic classification to classify the traffic into malicious and non-malicious classes. 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:-

1. 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.
2. 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.
3. Classification technique, SVM classifier is 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.

5.1 Support Vector Machine

SVM classification method is used because it has a predictive model. It performs text categorization of present data in order to predict the future data. 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. 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.

5.2 K nearest Neighbor

KNN is a lazy learning and simplest amongst all the machine learning algorithms. 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 code, 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.

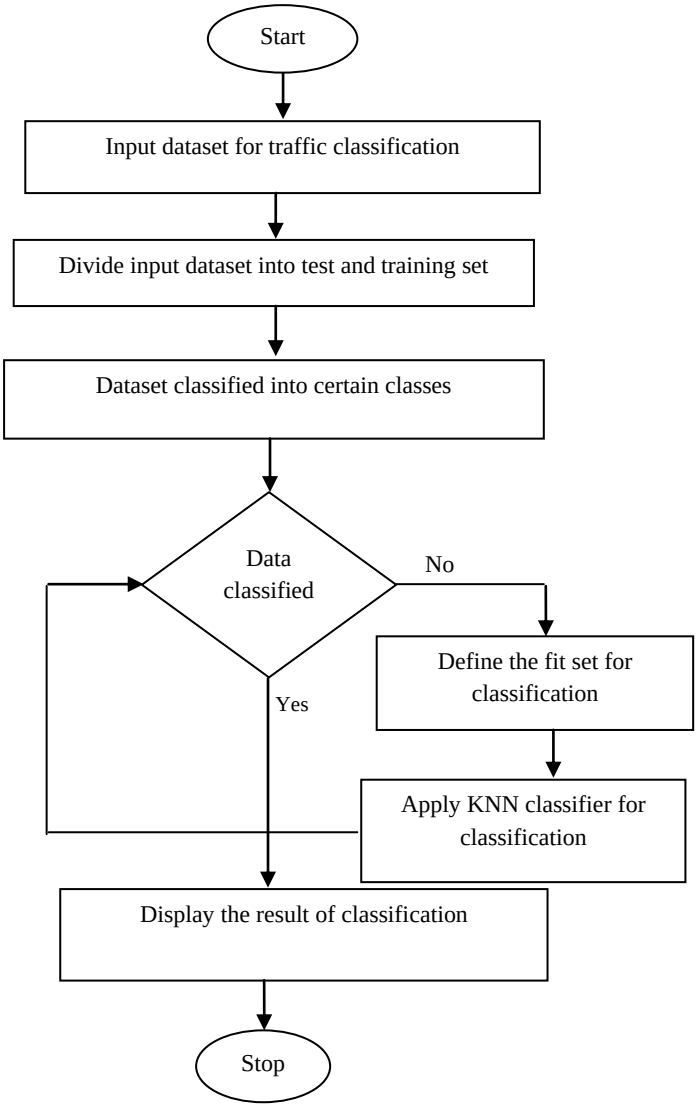

Fig 1: Flow chart of proposed method

6. RESULT AND DISCUSSION

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

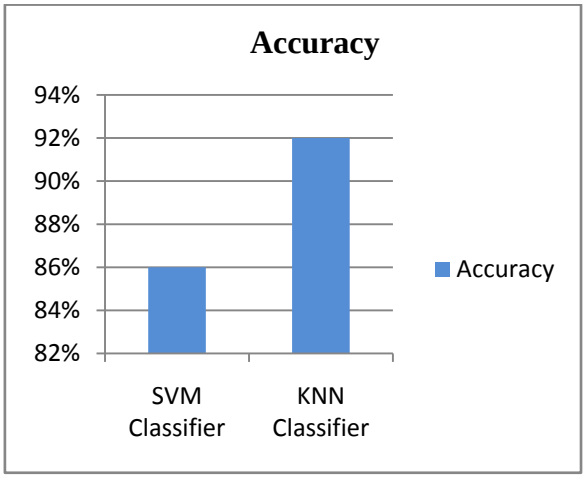

Fig 2. Accuracy Comparison

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 approximately 92 percent and of SVM classifier is approximately 86 percent.

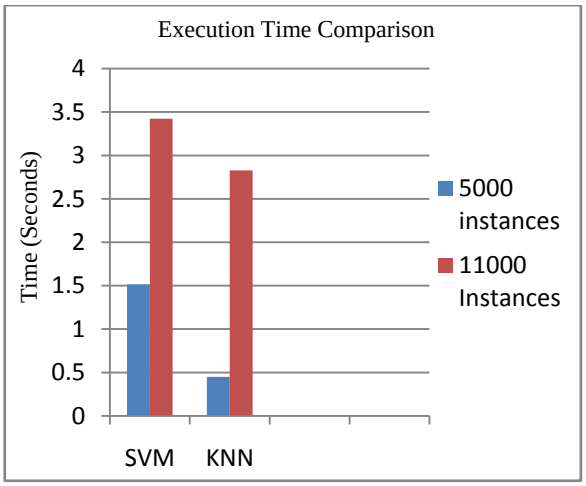

Fig 3. Execution Time

As shown in figure 3, 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.

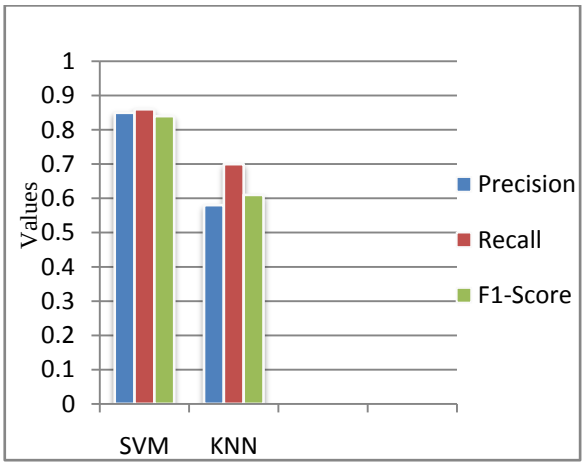

Fig. 4: Comparative result analysis of Performance Measuring Values

As shown in figure 4, the comparative result analysis of the SVM and KNN classifiers have been done with the performance measuring values obtained by them. After analyzing the obtained values, it is concluded that KNN gives the best performance as compared to the SVM. KNN classifies network traffic more accurately in less time than SVM.

7. CONCLUSION

Data classification is an important task in machine learning. It is identified with developing 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. A portion of the well-known ones are decision trees, instance-based learning, e.g., the K-nearest neighbors algorithm (KNN), artificial neural networks, Naive-Bayes, and support vector machines (SVM). All things considered, the greater

part of them is highly dependent on appropriate parameter tuning. Examples include the confidence factor and the minimum number of cases to partition a set in C4.5 decision tree; the K value in KNN; the number of hidden layers and others in artificial neural networks; and the piece function, the bit parameters, the stopping criterion, and others in SVM.

In the past years, various techniques of classification have been proposed which were 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. In the present dissertation to implement predictive analysis, techniques of classification SVM and KNN are implemented which ensure privacy of unstructured data in terms of accuracy and execution time. It is concluded that KNN gives the best performance as compared to the SVM. In network traffic classification, accuracy of KNN is approximately 92% and of SVM is approximately 86% with SVM taking more time to execute as compared to KNN.

Future work includes more improvement in big data analytics, predictive analytics i.e. clustering and classification techniques. New research is always in need for implementing more new methods and tools, finding the one with best results. In future, this tool can be used for datasets other than VoIP. The proposed technique can be compared with the other techniques of classification.

8. REFERENCES

[1] Youyang Qu, Shui Yu, Longxiang Gao, and Jianwei Niu "Big Data Set Privacy Preserving through Sensitive Attribute-based Grouping" IEEE ICC Communication and Information Systems Security Symposium, 2017.

[2] K. Sree Divya1, P.Bhargavi, S. Jyothi, "Machine Learning Algorithms in Big data Analytics," International Journal of Computer Sciences and Engineering vol.6 (1), Jan 2018.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Seonghoon Moon, Juwan Yoo, and Songkuk Kim. Exploiting Adaptive Multi-interface Selection to Improve QoS and Cost-efficiency of Mobile Video Streaming. IEEE International Conference on Mobile Services, (pp. 134-141), 2015.

[10] 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.

[11] 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, pp. 444–447, Aug 2015.