IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-6, Issue-4 E-ISSN: 2347-2693

Intruder Attack Detection In Data Network Organization Using Data Mining Techniques

Renu Dewli[1*], Anubhooti Papola[2]

[1*]Computer Science and Engineering, Faculty of Technology, Uttarakhand Technical University, Dehradun, India
[2] Computer Science and Engineering, Faculty of Technology, Uttarakhand Technical University, Dehradun, India

**Corresponding Author: deolirenu@gmail.com, Tel.: 9456391278*

Available online at: www.ijcseonline.org

Received: 20/Mar/2018, Revised: 28/Mar/2018, Accepted: 19/Apr/2018, Published: 30/Apr/2018

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, 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. To increase the accuracy of classification technique Knn classier is applied which increase accuracy, execution time.

Keywords— *Data network, attacks, data mining,, IDS/IPS machine learning*

I. INTRODUCTION

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, 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 (for example, predicting political affiliations in view of online buys and interactions), or we may be interested in recognizing important nodes in the system (for example, basic nodes in correspondence networks). In the vast majority of these situations, an important stride in accomplishing our final goal is characterizing, or labelling, the nodes in the system.

Collective classification alludes to the consolidated classification of an arrangement of interlinked objects utilizing every one of the three sorts of information depicted above. Take note of that, occasionally the expression relational classification is utilized to denote an approach that focuses on arranging system data by utilizing just the initial two sorts of correlations recorded above. Be that as it may, in numerous applications that produces data with correlations between labels of interconnected objects (a marvel once in a while referred to as relational autocorrelation labels of the objects in the neighbourhood are regularly unknown also. In such cases, it ends up noticeably important to all the while gather the labels for every one of the objects in the system. Inside the machine learning community, classification is typically done on each protest independently, without considering any hidden system that associates the objects. Collective classification does not fit well into this setting. For example, in the webpage classification problem where web pages are interconnected with hyperlinks and the task is to assign every webpage with a label that best indicates its topic, it is normal to assume that the labels on interconnected web pages are correlated. Such interconnections happen naturally in data from a variety of applications, for example, bibliographic data, email networks and social networks [2].

Traditional classification procedures would ignore the correlations represented by these interconnections and would be unable to deliver the classification accuracies conceivable utilizing a collective classification approach. Despite the fact that traditional exact inference algorithms, for example, factor disposal and the intersection tree calculation harbour the possibility to perform collective classification, they are practical just when the graph structure of the system fulfill certain conditions. By and large, exact inference is known to be a NP-difficult problem and there is no guarantee that true system data fulfill the conditions that make exact inference tractable for collective classification. As an outcome, the greater part of the research in collective classification has been devoted to the improvement of approximate inference algorithms. There are four well known approximate inference algorithms utilized for collective classification, iterative classification, Gibbs testing, loopy conviction proliferation and mean-field relaxation labeling.

Complex networks assemble ideas from insights, dynamical systems, and graph hypothesis. Fundamentally, they are large-scale graphs with nontrivial connection patterns. In addition, the capacity to capture special, functional, and topological relations is one of their salient characteristics. These days, complex networks appear in numerous situations, for example, social networks, organic networks, Internet and World Wide Web, electric energy networks, and classification and pattern acknowledgment. Subsequently, distinct fields of sciences, for example, physics, arithmetic, science, software engineering, and engineering have contributed to the large advances in complex system think about.

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. 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 are highly dependent of appropriate parameter tunning. 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 stop criterion, the number of neurons, the number of hidden layers, and others in artificial neural networks; and the soft margin, the piece function, the bit parameters, the stopping criterion, and others in SVM. Complex networks have made extensive contributions to machine learning study. Be that as it may, the majority of the researches identified with complex networks are connected to data clustering, dimensionality lessening and semi-directed learning [11].

Classification according to the kinds of databases mined Different types of databases that are to be mined can help in classifying the data mining systems. As per the various criteria, the classification of each database systems is also done. A different data mining technique is to be utilized for each different database system. Thus, the classification also depends on all such factors. When data models are to be classified for instance, they are done within the categories namely relational, transactional, object-oriented, object relational, or data warehousing mining applications. As per the various types of data handled, the classification can be done within spatial, time-series, text as well as multimedia data mining systems or WWW mining systems. There are also the heterogeneous data mining systems as well as the legacy data mining systems involved within these classification types.

Classification according to the kinds of knowledge mined As per the knowledge that is mined, the systems are classified within the data mining. The criteria might be on the basis of functionalities like characterization, discrimination, association, classification, clustering, etc. All such characteristics help in determining the different categories. There are multiple as well as integrated data mining functionalities present within the comprehensive data mining system.

Classification according to the kinds of techniques utilized As per the employed data mining techniques, the data mining systems can also be classified. As per the involvement of degree of user interaction various techniques are to be described within the data mining. Also there are various methods employed within the analysis of data mining which include the visualization, pattern recognition, neural networks, machine learning, etc. An integration of techniques that help in combination of advantages of specific approaches is done with the help of variety of data mining techniques in a sophisticated data mining system.

Anomalies of Social networks have turned into a communication platform where distinctive users with a personalized user profile interact and impart information to each other. At present, practically every domain is linked in one frame or the other with the social networks. Be it excitement, education, exchanging, business, communication and so on., OSN has made an influence on each of them. For instance, for the most part companies have begun advancing their brands and products on social networking sites to expand the popularity of their products which thus enhances their sales. Opposite, to the positive side of social networking sites, its expanding popularity and open and free utilize have likewise led to their extensive misuse. Malicious users are utilizing it diversely by carrying on and obeying patterns uniquely in contrast to their peers. For instance, a normal user regularly send messages to set of users which as a rule have association among themselves however an anomalous user picks its audience at random which are probably not going to have a relation in the middle of them .

In the base paper, we designed a model using data mining technique to detect anomalies in network traffic data. We assembled new rules from the automatic learning thought data mining and them translated and reprogrammed the IDS/IPS to assess the efficiency of the mode. we have to extend analysis to the implementation of an algorithm that is able to predict future attack on this data network. Select the algorithm to improve the security and increase the accuracy of classification technique Knn classier is applied which increase accuracy, execution time.

After the introduction in Section I the rest of this dissertation is organized as follows: Section II contains the related work, section III explains the methodology with flow chart, Section IV describes results and discussion, Section V concludes research work with future directions.

II. Related Work

Pedro Amaral, et.al, “Machine Learning in Software Defined Networks: Data Collection and Traffic Classification”, 2016

Software Defined Networks (SDNs) gives a separation between the control plane and the forwarding plane of networks. The software implementation of the control plane and the inherent data collection mechanisms of the OpenFlow protocol are the excellent tools for implementing the Machine Learning (ML) network control applications. An initial phase toward that path is to comprehend the sort of data that can be gathered in SDNs and how information can be learned from that data. In this work we portray a straightforward architecture sent in an enterprise network that gathers traffic data utilizing the OpenFlow protocol [11]. We display the data-sets that can be acquired and indicate how a few ML methods can be applied to it for traffic classification. The results indicate that high accuracy classification can be acquired with the data-sets utilizing supervised learning.. As future work the study of unsupervised methods or semi-supervised methods can be explored for Traffic classification. The study of different types of valuable information that can be learned from the gathered data like, client usage profiles, network utilize profiles or traffic predictions is likewise a next stride.

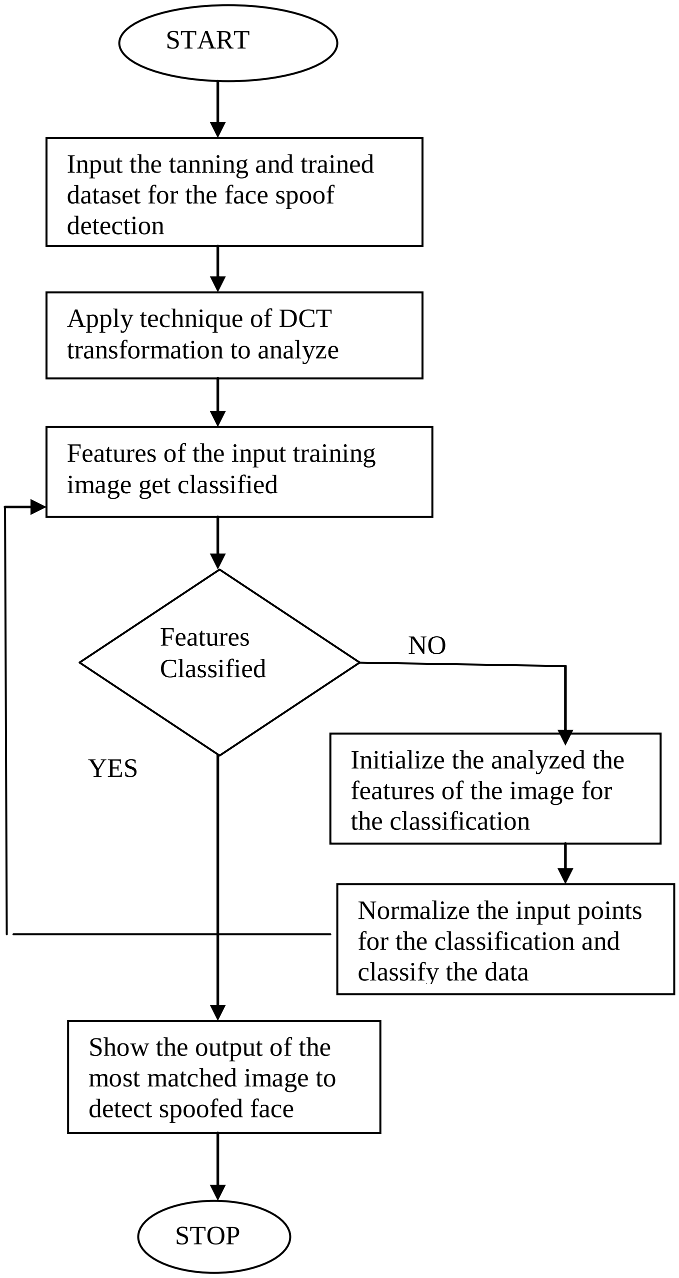
Zhizhong Kang, et.al, “A Bayesian-Network-Based Classification Method Integrating Airborne LiDAR Data With Optical Images”, 2016
Point cloud classification is of incredible importance to applications of airborne Light Detection And Ranging (LiDAR) data. Lately, airborne LiDAR has been incorporated with different sensors, e.g., optical imaging sensors, and thus, the fusion of multiple data types for scene classification has turned into a hot point. Experiments demonstrate that the BN classifier can effectively recognize four types of fundamental ground objects, including ground, vegetation, trees, and buildings, with a high accuracy of more than 90%. In addition, compared with different classifiers, the proposed BN classifier can accomplish the highest overall accuracies, and specifically, the classifier demonstrates its advantage in the classification of ground and low vegetation points.
Nikola K. Kasabov, et.al, “Mapping, Learning, Visualization, Classification, and Understanding of fMRI Data in the NeuCube Evolving Spatiotemporal Data Machine of Spiking Neural Networks”, 2016
This paper introduces another methodology for dynamic learning, visualization, and classification of functional attractive reverberation imaging (fMRI) as spatiotemporal brain data. The method is based on an evolving spatiotemporal data machine of evolving spiking neural networks (SNNs) exemplified by the NeuCube architecture [13]. The method consists of a few steps: mapping spatial coordinates of fMRI data into a 3-D SNN 3D shape (SNNc) that represents a brain template; input data transformation into trains of spikes; profound, unsupervised learning in the 3-D SNNc of spatiotemporal patterns from data; supervised learning in an evolving SNN classifier; parameter improvement; and 3-D visualization and model interpretation. Two benchmark case study problems and data are utilized to illustrate the proposed methodology—fMRI data gathered from subjects when reading affirmative or negative sentences and another—on reading a sentence or seeing a picture.

Sahan L. Maldeniya, et.al, “Network Data Classification Using Graph Partition”, 2013
Utilization of network classification can be seen in numerous domains. These changes are from safeguarding the quality of network to breaking down personal characteristics of network users. However present methods applied for network data classification does not meet the expectations. This is on account of networks are dynamic which are prone to rapid changes, while methods utilized for the classification has been either trained utilizing examples or defined utilizing heuristics. World Wide Web itself is a big graph which is made out of number of URLS connecting each other by means of hyper-connections [14]. Henceforth in this work we have utilized this graph nature of WWW and applied graph theories to partition the network to classify network data. We have utilized results acquired by classifying the network traffic utilizing k-implies algorithm to evaluate the performance and usability of proposed method.

III. 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

Flowchart


START
Input the tanning and trained dataset for the face spoof detection
Apply technique of DCT transformation to analyze
Features of the input training image get classified
Features Classified
NO
YES
Initialize the analyzed the features of the image for the classification
Normalize the input points for the classification and classify the data
Show the output of the most matched image to detect spoofed face
STOP

IV. Result and Discussion

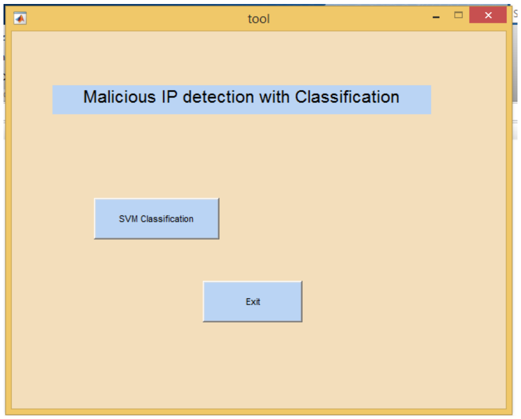


Fig 3.1: Default Interface

As shown in the figure 1, the interface is designed in which buttons are provided for the SVM classifier. These classifiers are compared in terms of accuracy for network traffic classification

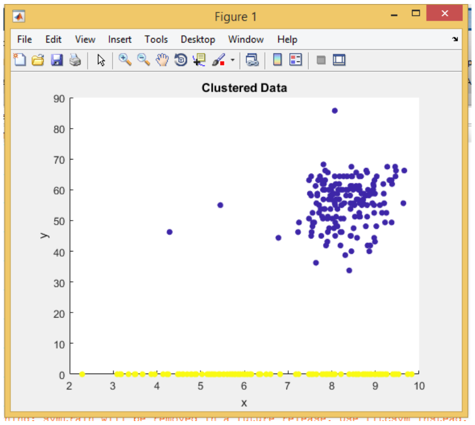
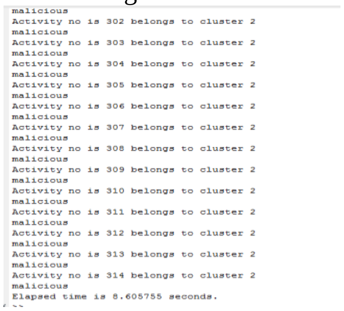


Fig3. 2: Clustered Data

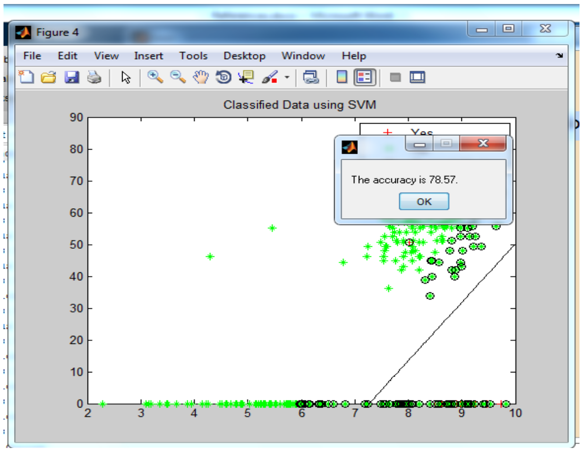
As shown in the figure 2, the first step of network data classification is clustering of the data. The technique of k-mean clustering is applied which will similar and dissimilar type of data is clustered together



```
malicious
Activity no is 302 belongs to cluster 2
malicious
Activity no is 303 belongs to cluster 2
malicious
Activity no is 304 belongs to cluster 2
malicious
Activity no is 305 belongs to cluster 2
malicious
Activity no is 306 belongs to cluster 2
malicious
Activity no is 307 belongs to cluster 2
malicious
Activity no is 308 belongs to cluster 2
malicious
Activity no is 309 belongs to cluster 2
malicious
Activity no is 310 belongs to cluster 2
malicious
Activity no is 311 belongs to cluster 2
malicious
Activity no is 312 belongs to cluster 2
malicious
Activity no is 313 belongs to cluster 2
malicious
Activity no is 314 belongs to cluster 2
malicious
Elapsed time is 8.605755 seconds.
```

Fig 3.3: Classification of Data

As shown in figure 3, the network traffic is clustered which has similar and dissimilar type of data. The network traffic which belongs to which cluster is shown in the above figure .



The accuracy is 78.57.

Fig 3.4: SVM classification result

As shown in the figure 4, the technique of k-mean clustering is applied which will cluster similar and dissimilar type data. In the last step technique of SVM classifier is applied which will classify the data into two class

Table1. : Performance comparison between different algorithms

ALGORITHM	RPTREE ALGO.	J48 ALGO.	Proposed Algo
ACCURACY	78.57	78.57	92.86
ERROR	0.0001	0.0001	0.0085

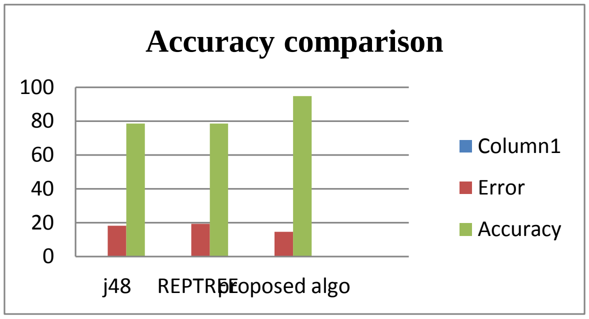


Fig.3.5 Graph showing the accuracy and error rate of different algorithms

V. Conclusion and Future Scope

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. 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 are highly dependent of appropriate parameter tunning. 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 stop criterion, the number of neurons, the number of hidden layers, and others in artificial neural networks; and the soft margin, the piece function, the bit parameters, the stopping criterion, and others in SVM. The technique of knn is applied in the improved work for the network traffic classification. the technique can be further improve for better security. The Knn classifier will classify the traffic which remained un classified . The classification which increasing security accuracy and execution time Performance. The proposed algorithm can be further compared with the other classifiers to check reliability of proposed work.

References

[1] G. Zhang, B. E. Patuwo, and M. Y. Hu, "Forecasting with artificial neural networks: The state of the art," International journal of forecasting, Vol.14(1), pp. 35-62. 1998

[2] L. M. Manevitz, and M. Yousef, "One-class SVMs for document classification," Journal of machine Learning research, Vol. 2, pp. 139- 154, 2002

[3] B. Schölkopf, and A. J. Smola, "Learning with kernels: support vector machines, regularization, optimization, and beyond," MIT press, USA, 2002

[4] C. T. Lin, C. M. Yeh, S. F. Liang, J. F. Chung, and N. Kumar, "Supportvector- based fuzzy neural network for pattern classification," IEEE Trans Fuzzy Systems, Vol.14(1), pp. 31-41, 2006

[5] MF. Amin and K. Murase, "Single-layered complex-valued neural network for real-valued classification problems," Neurocomputing 72, January 2009, pp. 945-955

[6] Doris Hooi-Ten Wong & Selvakumar Manickam, "Intelligent Expertise Classification Approach: An Innovative Artificial Intelligence Approach To Accelerate Network Data Visualization", 2010 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE)

[7] C. S. Dangare, and S. S. Apte, "Improved study of heart disease prediction system using data mining classification techniques," International Journal of Computer Applications, Vol. 47(10), pp. 44-48, 2012

[8] HY. Huang, YJ. Lin, YS. Chen and HY. Lu, "Imbalanced data classification using random subspace method and SMOTE," 2012 Joint 6th International Conference on Soft Computing and Intelligent Systems (SCIS) and 13th International Symposium on

Advanced Intelligent Systems (ISIS) , November 2012, pp.817-820

[9] R. O. Duda, P. E. Hart, and D. G. Stork, “Pattern classification”, John Wiley & Sons, Canada, 2012

[10] AR. Hafiz, AA. Yarub, MF. Amin, and K. Murase, “Classification of Skeletal Wireframe Representation of Hand Gesture using Complex- Valued Neural Network,” Neural Processing Letters 42(3), December 2015, pp.649-664

[11] Pedro Amaral, et.al, “Machine Learning in Software Defined Networks: Data Collection and Traffic Classification”, 2016

[12] Pedro Amaral, Joao Dinis, Paulo Pinto, Luis Bernardo, Joao Tavares, Henrique S. Mamede, “Machine Learning in Software Defined Networks: Data Collection and Traffic Classification”, 2016, IEEE 24th International Conference on Network Protocols (ICNP), Workshop on Machine Learning in Computer Networks (NetworkML 2016)

[13] Zhizhong Kang, Juntao Yang, and Ruofei Zhong, “A Bayesian-Network-Based Classification Method Integrating Airborne LiDAR Data With Optical Images”, 2016, IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING

[14] Sahan L. Maldeniya, Ajantha S. Atukorale, Wathsala W. Vithanage, “Network Data Classification Using Graph Partition”, 2013, IEEE.

[15] Nikola K. Kasabov, Maryam Gholami Doborjeh, and Zohreh Gholami Doborjeh, “Mapping, Learning, Visualization, Classification, and Understanding of fMRI Data in the NeuCube Evolving Spatiotemporal Data Machine of Spiking Neural Networks”, 2016, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS

[16] Kentarou Matsuda, Kazuyuki Murase, “Single-layered Complex-valued Neural Network with SMOTE for Imbalanced Data Classification”, 2016, Joint 8th International Conference on Soft Computing and Intelligent Systems

[17] Parisa Naraei, Abdolreza Abhari, Alireza Sadeghian, “Application of Multilayer Perceptron Neural Networks and Support Vector Machines in Classification of Healthcare Data”, FTC 2016 - Future Technologies Conference

[18] Joseph D. Prusa, Taghi M. Khoshgoftaar, “Designing a Better Data Representation for Deep Neural Networks and Text Classification”, 2016 IEEE 17th International Conference on Information Reuse and Integration

[19] Lorenzo A. Rossi, Bhaskar Krishnamachari and C.-C. Jay Kuo, “Energy Efficient Data Collection via Supervised In-Network Classification of Sensor Data”, 2016, International Conference on Distributed Computing in Sensor Systems

[20] Justin Salamon and Juan Pablo Bello, “Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification”, 2016, IEEE

[21] Bambang Sugiarto, Rika Sustika, “Data Classification for Air Quality on Wireless Sensor Network Monitoring System Using Decision Tree Algorithm”, 2016, 2nd International Conference on Science and Technology-Computer (ICST)

[22] Balasaheb Tarle, Sudarson Jena, “Improved Artificial Neural Network For Dimension Reduction In Medical Data Classification”, 2016, IEEE

[23] Ran Zhang, Lifeng Wu, Xiaohui Fu and Beibei Yao, “Classification of Bearing Data Based on Deep Belief Networks”, 2016, IEEE

[24] Yinfeng Zhao, Lei Li, “Link Prediction-based Multi-label Classification on Networked Data”, 2016, IEEE First International Conference on Data Science in Cyberspace

[25] Saptarshi Rudra, Soham Mitra, Soumyajit Das, Abhisek Roy, Shibasis Guha, “Gender Classification System from Offline Survey Data Using Neural Networks”, 2016, IEEE

[26] Joao Roberto Bertini Junior, Maria do Carmo Nicoletti, “Functionally Expanded Streaming Data as Input to Classification Processes Using Ensembles of Constructive Neural Networks”, 2016, IEEE

[27] Arun Manicka Raja M., Swamynathan S., “ENSEMBLE LEARNING FOR NETWORK DATA STREAM CLASSIFICATION USING SIMILARITY AND ONLINE GENETIC ALGORITHM CLASSIFIERS”, 2016 Intl. Conference on Advances in Computing, Communications and Informatics (ICACCI)

Authors Profile

Ms Renu Dewli completed Bechuler of Computer Science from Uttarakhand Technical University Dehradun in 2016 and she is cuently pursuing Master of Tchnology under uttarakhand Technical University Dehradun,india.

Mrs Anubhooti Papola working as an assistar Uttarakhand technical University in departmen science .She is doing her PhD in computer science from Uttarakhand technical University. She is pos Graphic Era University and graduate from University.