Network Traffic Control Using AI

N. SelvaKumar 1*, M. Rohini 2, C. Narmada 3, M. Yogeshprabhu 4

1,2,3,4 Dept. of CSE, Coimbatore Institute of Engineering and Technology Tamil Nadu, India.

*Corresponding Author: selvaa618@gmail.com Tel: +918870346752

Received: 1/Feb/2020, Accepted: 19/Mar/2020, Published: 30/Apr/2020

Abstract—The core of next generation 5G wireless network is heterogeneous network. The existing traditional 4G technology approaches are centrally managed and reactive conception-based network which needs additional hardware for every update and when there is a demand for the resources in the network. To reduce traffic in 4G network techniques like Port based, payload and feature based approach are used. They do not provide efficient resource management and it does not classify all type of network traffic effectively. 5G helps in giving solution to the problem of 4G network using prediction and traffic learning to increase performance and bandwidth. Heterogeneous network provides more desirable Quality of Service (QOS) and explores the resources of the network explicitly. The assortment of heterogeneous network brings difficulty in traffic control of the network. The problem in heterogeneous network is network traffic which cannot be controlled and managed due to different protocols and data transfer rate. Artificial Intelligence is deployed in the network. This work aims to explore artificial intelligence inspired network traffic scheme for reducing traffic in a heterogeneous network. In this work the proposed system consists of clustering of traffic data set, statistical feature extraction of the network data, prediction of network traffic and classification of network traffic. The clustering of network data is implemented using Enhanced Density-based spatial clustering of applications with noise (DBSCAN) algorithm. The statistical feature of the network data is extracted and they are given as input into the Modified Backpropagation model for prediction and regression tree method is used for classification of network traffic. Finally caching and pushing of network is included to make use of the network resource effectively and also to provide finer Quality of Service (QOS) in a network.

Keywords—5GNetwork; Artificial Intelligence; Machine Learning; Data Clustering; Statistical Feature; Traffic Prediction; Traffic classification.

I. INTRODUCTION

In recent years with the prosperous development of the Internet, networking has allured a lot of recognition in both industry and academia. The improvement of mobile communication has increased the data transfer rate significantly, for large amount of data with multimedia communication service. The mobile communication is now stepping into 5G. To satisfy the data traffic demand the network technology are moving towards heterogeneous network which provides ubiquitous internet access and enhanced public services [1].

The next generation network is service-driven where a single infrastructure should efficiently provide different service such as low latency communication, enhanced mobile broadband and immense machine type communication for heterogeneous network. The heterogeneous network different layers of cells like femto, macro, micro, pico, relays, diverse user devices and application.

The diversity of the heterogenous network results as a barrier for adopting numerous protocols and also it cannot manage the network resource effectively. The diversity of network leads to network traffic.

Artificial Intelligence plays a major role in overcoming the network traffic problem in heterogeneous network. In every possible field, machine learning has been used to leverage its astonishing power.

In variety of application such as speech recognition, bioinformatics and computer vision, Machine Learning (ML) techniques have been used efficiently. Machine learning is mainly used for prediction and classification and also in
networking it is mainly used for performance prediction and intrusion detection. To make decision directly Machine learning constructs models that can learn themselves from data without being explicitly programmed or without following some set of rules.

Machine learning enables the model to get into self-learning mode without being explicitly programmed. The model can be trained by providing data sets to them, when exposed to new data, models are enabled to learn, predict and develop by themselves. Machine learning algorithm can be classified into three categories. They are supervised learning, unsupervised learning, reinforcement learning [3][5].

In Supervised learning the model is trained on a labeled data set which then learns on its own and when new testing data is given it compares with the training data set and predicts the output. Supervised learning is mainly used for regression and classification problems.

In unsupervised learning the training data set is unlabelled, and it finds pattern and relationship among data. It is mainly used in clustering and association problems. In reinforcement learning the model learns on its own without any training data.

This work mainly focuses on detection and classification of network traffic using machine learning techniques. The dataset is taken from network database. The data regarding the network flow are clustered depending on the flow of data and the number of clusters decided. The statistical features of the clustered data are extracted and given as input for classification model where the training and testing process is done and finally network traffic is predicted and classified.

- Fine grained
- Coarse grained

In fine grained classification, exact classification of all network type application is done. The features of the application are selected exactly with relevant to the generated traffic. In coarse classification goals, only the transaction-oriented features are classified [13][14][15].

The traditional classification approaches for network classification using Transmission Control Protocol (TCP) or User Datagram Protocol (UDP) port number which has been used fifteen years ago cannot be used for accurate classification of traffic.

The traditional classification approach that has been used by researchers for network traffic classification are
- Port based IP traffic classification
- Payload based IP traffic classification

For exact classification of traffic new techniques have been used. They are
- Protocol Behavior or Heuristics Based Classification
- Classification based on flow statistics traffic properties [21]

A. Port based IP traffic classification

TCP and UDP give multiplexing of different streams between IP endpoints with the assistance of port numbers. Generally numerous applications use an 'outstanding' port to which different hosts may start correspondently. The application is deduced by looking into the TCP SYN parcel's objective port number in the Internet Assigned Numbers Authority (IANA's) rundown of enlisted ports. In any case, this methodology has constraints. Right off the bat, a few applications might not have their ports enrolled with IANA (for instance, distributed applications, for example, Napster and Kazaa). An application may utilize ports other than its outstanding ports to rescue from working framework get to control confinements. Additionally, at times server ports are powerfully assigned as required. Though port-based traffic grouping is the quickest and straightforward strategy, a few examinations have demonstrated that it performs ineffectively, e.g., under 70% precision in characterizing streams [11][12].

B. Payload based IP traffic classification

This methodology reviews the packet header to decide the applications. Packet payloads are analyzed a little bit at a time to find the bit streams that contain signature. In the event such piece of streams is discovered, at that point bundles can be precisely named. This methodology is regularly utilized for P2P traffic discovery and system interruption identification. Real impediments of this methodology are that the protection laws may not enable directors to assess the payload; it additionally forces huge multifaceted nature and preparing load on traffic ID
gadget; requires significant computationally power and capacity limit since it examines the full payload [2][17]

C. Protocol Behavior Based Classification

In this method the classification of networks is based on connection level patterns and network protocol behavior. This method is based on identifying and observing patterns of host behavior at the transport layer. The advantage of this classification is that packet pay load access is not needed [10][2].

D. Flow statistics Traffic Classification

The existing techniques are restricted by their dependence of packet content (payload and port numbers) on the inferred linguistics of the information gathered through deep review. Newer approaches depend on traffic's statistical characteristics to identify the applying [4][7][6][9] associate degree assumption underlying such ways in which is that traffic at the network layer has mathematical properties that are distinctive definitely classes of applications and modify different applications to be distinguished from each other. It uses network or transport layer that has applied mathematics properties like distribution of flow length, flow idle time, packet interarrival time, packet lengths etc. These are distinctive sure categories of applications and hence facilitate to differentiate different applications from one another.

This methodology is possible to see application sort however not usually the particular consumer type. as an example, it can't verify if flow belongs to Skype or Microsoft Network traveler voice traffic specifically [30][43]. The advantage of this approach is that there's no packet payload security is concerned.

E. Machine learning approach for traffic classification

Machine learning algorithms are effectively used for network traffic. It provides an alternative approach for payload and port-based approach. It is independent of packet length and features. Machine learning algorithm mainly involves testing and training process which gives more accurate result. Machine learning algorithms are classified as supervised, unsupervised, semi-supervised.

Supervised algorithms that are used for traffic classification are: Naïve Bayes, support vector machine, k- nearest neighbor. Unsupervised algorithm used for classification are K-Means and Density-based spatial clustering of applications with noise (DBSCAN).

Moore and Papagiannaki [32] observed no better than 70% byte accuracy for port-based classification using the official IANA list.

Madhukar and Williamson [33] showed that port- based analysis is unable to identify 30-70% of Internet traffic flows they investigated. Sen et al. [34] reported that the default port accounted for only 30% of the total traffic (in bytes) for the Kazaa P2P protocol.

Erman et al. in [31] in early 2007 proposed a semi-supervised traffic classification approach which combines unsupervised and supervised methods. Motivations to the proposal are due to two main reasons: Firstly, labeled examples are scarce and difficult to obtain, while supervised learning methods do not generalize well when being trained with few examples in the dataset. Secondly, new applications may appear over time, and not all of them are known as a priori, traditional supervised methods map unseen flow instances into one of the known classes, without the ability to detect new types of flows [31].

The relationship between the class of traffic and its observed statistical properties has been noted in [36] (where the authors analyzed and constructed empirical models of connection characteristics - such as bytes, duration, arrival periodicity - for a number of specific TCP applications), and in [37] (where the authors analyzed Internet chat systems by focusing on the characteristics of the traffic in terms of flow duration, packet inter-arrival time and packet size and byte profile). Later work (for example [38][39] and [40]) also observed distinctive traffic characteristics, such as the distributions of packet lengths and packet inter-arrival times, for a number of Internet applications. The results of these works have stimulated new classification techniques based on traffic flow statistical properties.

III. PROPOSED SYSTEM

In this paper an effective approach for network traffic detection and classification using machine learning approaches has been introduced. The proposed system includes methods such as data clustering, statistical feature extraction, detection and classification of network traffic. Caching and pushing of the traffic are implemented using collaborative filtering to make use of the resource.
The network dataset is taken from the network database which does not contain any unwanted data. Initially network data is clustered using DBSCAN algorithm and then the statistical feature of the network data are extracted and unwanted flow of data is removed. The extracted features are given as input to the neural network classification model using testing and training process and finally, the network traffic is predicted, classified using regression tree and traffic control is made using the caching and pushing operations.

IV. METHODOLOGY

The proposed methodology for detection of network traffic using flow data packet is as follows:

Step 1: Network data set acquisition.
Step 2: Load data with flow construction
Step3: Clustering of flow data packets using Enhanced DBSCAN clustering algorithm.
Step 4: Statistical feature extraction of clustered data
Step5: Prediction classification of feature extracted network data using modified neural network and regression tree.

A. Network Data set Acquisition
The dataset for network traffic classification is collected online from the network traffic website. The dataset consists of traffic flows which are randomly selected from the wide trace and carefully recognized by the manual inspection. It consists of 3416 instances with 7 classes such as (CHAT, FT, BROWSING, MAIL, STREAMING, VOIP, P2P) and it has 16 attributes.

B. Clustering of Network Dataset Using Modified DBSCAN

Density Based Spatial Clustering of Application of Noise is a density-based clustering algorithm which uses dense area of objects. The parameters that are used in DBSCAN algorithm are eps, min points. The clusters in DBSCAN are

<table>
<thead>
<tr>
<th>Table 1: NETWORK TRAFFIC ATTRIBUTES</th>
</tr>
</thead>
<tbody>
<tr>
<td>min_active</td>
</tr>
<tr>
<td>mean_active</td>
</tr>
<tr>
<td>max_active</td>
</tr>
<tr>
<td>std_active</td>
</tr>
<tr>
<td>min_idle</td>
</tr>
<tr>
<td>mean_idle</td>
</tr>
<tr>
<td>max_idle</td>
</tr>
<tr>
<td>min_biat</td>
</tr>
<tr>
<td>max_biat</td>
</tr>
<tr>
<td>mean_biat</td>
</tr>
<tr>
<td>mean_flowiat</td>
</tr>
<tr>
<td>max_flowiat</td>
</tr>
<tr>
<td>mean_flowiat</td>
</tr>
</tbody>
</table>

Figure-5: Network Traffic Dataset

Figure-4: Flow Diagram of Proposed System
formed from the core point which are directly-density reachable and density reachable [18].

The parameters that are used for accurate classification are Threshold parameter, Minimum points, Distance measure function

The data are collected from online tools and features are selected based on the packets and then they are fit into the model for testing and training. Finally, they are predicted and classified.

Let \( Y = \{Y_1, Y_2, Y_3, \ldots \} \) be the set of data points
- Initially the process has to be started with an arbitrary starting point which is not visited already.
- Extract the neighbour of arbitrary point using \( \varepsilon \).
- Then clustering process starts if there are sufficient neighbourhood and point is marked as visited or it is noted as noise data.
- If a point is found to be a part of the cluster then its \( \varepsilon \) neighbour is also the part of the cluster and the above procedure from step 2 is repeated for all \( \varepsilon \) neighbour points. Until all the cluster point is determined this process is repeated.
- The unvisited new point is retrieved.
- This process stops when all points are marked as visited. The proposed method also includes Fixed Radius Nearest Neighbour Calculation of the clustering the network data which increases the accuracy of the clustering process.

Here the \( \text{frNN} \) is calculated from the output that is obtained from \( \text{res}<\text{rdbscan} \).

\[
\begin{align*}
\text{res} & \leftarrow \text{rdbscan}(\text{data}, \varepsilon=0.5, \\
\text{MinPts} & = 7) \ \text{res} \\
\#\text{pairs}(\text{data}, \text{col} = \\
\text{res}\text{cluster} + 1) & \ \text{fr} \leftarrow \\
\text{frNN}(\text{data}, \varepsilon = .5) & \ \text{rdbscan}(\text{fr}, \\
\text{minPts} & = 7)
\end{align*}
\]

C. Statistical Feature Extraction of the Network Data
After clustering of network data, the statistical features of the network data are extracted which is then move forward to classification process. The statistical features help in classifying and predicting the network data in more accurate manner. There are many types of statistical features some of them have been extracted in this method [23][28].

The statistical feature contains
- Packets
- Bytes
- Packet size
- Inter packet time

<table>
<thead>
<tr>
<th>Types Of Features</th>
<th>Feature Description</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Packets</td>
<td>Number of packets transferred</td>
<td>2</td>
</tr>
<tr>
<td>Bytes</td>
<td>Volume of bytes transferred</td>
<td>2</td>
</tr>
<tr>
<td>Packet Size</td>
<td>Min, Max, Median and Standard. Deviation of packet size</td>
<td>8</td>
</tr>
<tr>
<td>Inter Packet Time</td>
<td>Min, Max, Median and Standard. Deviation of inter packet time</td>
<td>8</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>20</strong></td>
</tr>
</tbody>
</table>

D. Network Traffic Classification Using modified Backpropagation and Regression Tree
Artificial neural network is one of the learning algorithms

\[
\Delta W_{ij} = -\varepsilon \frac{\partial E}{\partial W_{ij}} = \varepsilon \frac{\partial E}{\partial \alpha_i}
\]

where which is used within machine learning techniques. It consists of many layers for learning and analysing data [43][42]. It learns like human brain and it is mainly used for pattern recognition and data classification. Neural networks are trained using examples. They can be programmed explicitly. It contains three layers
- Input layer
- Hidden layer
- Output layer

It may also contain multiple hidden layer. Hidden layer is mainly used for feature extraction and calculation. Feed forward and feedback are two topologies in neural network [22].

It is the most important algorithm for training a neural network. It is mainly used to network traffic effectively in heterogeneous network. For training the weights in multi-layer feed forward network, backpropagation algorithm is used [27].

The neuron has weights that has to be maintained. Then the forward propagation is classified as neuron activation, neuron transfer, forward propagation [26][29].
Next is backpropagate error where the error is calculated and error is then back propagated through the hidden layer. It involves transfer derivative and error propagation. Then the network has to be trained by propagating the error and forwarding inputs [16].

The proposed backpropagation algorithm includes In the context of learning, backpropagation is commonly used by the gradient decent optimization algorithm to adjust the weight of neurons by calculating the gradient of the loss function.

Gradient Descent learning requires that any change in a particular weight be proportional to the negative of the derivative of the error, the change in a given weight must be proportional to the negative of our prior equation. Replacing the difference between the target and actual activation of the relevant output node by $d$, and introducing a learning rate epsilon, that equation can be re-written in the final form of the Delta Rule [19][20][12]:

The proposed algorithm for prediction of network traffic using neural network

```r
# Implement the gradient descent algorithm
m_new <- m - learn_rate * ((1 / n) * (sum((yhat - y) * x)))
c_new <- c - learn_rate * ((1 / n) * (sum(yhat - y)))

m <- m_new

c <- c_new

yhat <- m * x + c

MSE_new <- sum((yhat - y) ^ 2) / n

if(MSE_new <= conv_threshold) {
  abline(c, m)
  converged = T
  return(paste("Optimal intercept:", c, "Optimal slope:", m))
}

iterations = iterations + 1

if(iterations > max_iter) {
  abline(c, m)
  converged = T
  return(paste("Optimal intercept:", c, "Optimal slope:", m))
}
```

Figure 7: Neural Network Learning Process using modified Backpropagation algorithm - Gradient Decent

The class which are predicted in proposed system are
- CHAT
- FT
- BROWSING
- MAIL
- STREAMING
- VOIP
- P2P
Figure 8: Prediction of different class in network traffic

i) Class VOIP
ii) CHAT
iii) MAIL
iv) P2P
v) STREAMING
vi) BROWSING

Figure 9: Classification of network traffic

The predicted classes are then classified using regression tree algorithm with different traffic parameters.

E. Caching and Pushing using Collaborative filtering.

The main aim of caching and pushing is to use the resource effectively and also to improve the Quality of Service in the network.

By analysing users’ information, habits, and historical data, recommendation algorithms can implement caching and pushing efficiently in the network which increases the accuracy of the system.

There are two types of collaborative filtering algorithm. They are:

- User based
- Item based

The user-based collaborative filtering algorithm will cache and push information based on the similarity between users; for example, caching video play to a user who has high similarity with the target user.

The item-based collaborative filtering algorithm will push and cache data based on the historical data of the base station; for example, caching the next episode of a TV series automatically.

V. EXPERIMENTAL RESULT

The proposed system classifies 3416 instances with seven classes and controls network traffic using Backpropagation classifier. The network traffic dataset is tested and classified. The classification performance is measured by sensitivity, specificity, accuracy, error rate F-measure, Precision and Recall.

\[
\text{Sensitivity} = \frac{TP}{TP + FN}
\]

\[
\text{Specificity} = \frac{TN}{TN + FP}
\]

\[
\text{Accuracy} = \frac{TP + TN}{TP + FN + TN + FP}
\]

\[
F - \text{Measure} = \frac{2TP}{2TP + FP + FN}
\]

Here Sensitivity denotes true positive rate and specificity explores true negative rate. Accuracy represents proportion of true positive and negative rate. Precision represents the ration of True positives over the sum of True Positives and False Positives. Recall indicates the ratio of True Positives over the sum of True Positives and False Negatives.

The sensitivity of the proposed system improved 92% to 97% which is positive proportion of the system.

The specificity of the work reduced from 70% to 65%. Specificity indicates the negative rate of the prediction system.

The error rate of the proposed system gradually decreased from 0.45% to 0.32%.

The F-measure of the system improved from 0.96 to 0.90%.
By applying weight optimization in neural network using soft computing technique gradient decent algorithm, the overall performance of the proposed system improved to 97.6 from 96.2%. The Performance of the proposed and existing system is represented in bar chart.

VI. CONCLUSION
AI is used as a tool to improve 5G technologies in recent technologies. The core of next generation 5G wireless network is heterogeneous network. Network traffic is the biggest problem of heterogeneous network which reduces QOS of network. To provide better service and also to use resource effectively traffic has to be monitored and controlled. The problem of 5G network can be solved using Machine Learning techniques. In this work AI inspired traffic control scheme is introduced. The clustering of network data is implemented using Enhanced Density-based spatial clustering of applications with noise (DBSCAN) algorithm. The statistical feature of the network data is extracted and they are given as input into the Modified Backpropagation model for prediction and regression tree method is used for classification of network traffic. Finally caching and pushing of network is included to make use of the network resource effectively and also to provide finer Quality of Service (QOS) in a network. The performance of the Proposed system achieved is about 97.6 and it is calculated using the parameters accuracy, sensitivity, specificity, error rate and F-Measure.

REFERENCES