ЎЗБЕКИСТОН РЕСПУБЛИКАСИ АХБОРОТ ТЕХНОЛОГИЯЛАРИ ВА КОММУНИКАЦИЯЛАРИНИ РИВОЖЛАНТИРИШ ВАЗИРЛИГИ

МУҲАММАД АЛ-ХОРАЗМИЙ НОМИДАГИ ТОШКЕНТ АХБОРОТ ТЕХНОЛОГИЯЛАРИ УНИВЕРСИТЕТИ ҚАРШИ ФИЛИАЛИ

АХБОРОТ-КОММУНИКАЦИЯ ТЕХНОЛОГИЯЛАРИНИНГ РИВОЖЛАНИШ ИСТИҚБОЛЛАРИ

мавзусидаги

Республика илмий-амалий анжуман
МАЪРУЗАЛАР ТЎПЛАМИ

(20-21 апрель)

ҚАРШИ-2018

- фойдаланиш даражаси;
- мустахкамлилик;

Тахлил натижалари 1-жадвалда келтирилган.

1-жадвал

NGN тармоғида мастахкамликни таъминлаш усулларининг таққослаш

Т/Р	Кўрсаткичлар номи	Softswitch X параллел улашли тармоқ		
		Статистикли	**Динамикли**	**Умумий юкланишли динамикли**
1	Тармоқ ишини ташкил қилиш	оддий	мураккаб	мураккаб
2	Маълумотларни конфигурациялаш ва техник хизмат кўрсатиш	оддий	оддий	мураккаб
3	Қайта улаш тезлиги	паст	баланд	баланд
4	Кириш тармоғининг ускуналари параллел ишига талаблар	йўқ	бор	бор
5	Фойдаланиш даражаси	паст	паст	баланд
6	Мустахкамлик	паст	баланд	баланд

NGN тармоғининг элементлари ишининг мастахкамлилигини таъминловчи кўриб чиқилган усуллардан қуйидагича хулоса қилиш мумкин. Мустахкамлиликни ошириш, техник хизмат сифатини яхшилаш ва тизимни ишлатиш даражасини кўтариш учун умумий юкланишдаги динамик режимда ишлаш тавсия этилади.

ФОЙДАЛАНИЛГАН АДАБИЁТЛАР

1. Соколов Н.А. Пути преобразования телефонных сетей в NGN-сети. – «Connect! Мир связи», №5, 2007.
2. Методические указания по проведению практических занятий по курсу «Сети следующего поколения NGN», Усманова Н.Б., 2011
3. Пинчук А.В., Соколов Н.А. Прагматическая стратегия перехода к NGN. – Вестник связи, 2006, №6.

COMPARISON OF INTERNET TRAFFIC CLASSIFICATION ALGORITHMS

Narzullayev Shoxrux Erkin o'g'li TUIT named after Muhammad al-Khwarizmi, 2-course master
Shokirov Akmal Tohirjon o'g'li, TUIT named after Muhammad al-Khwarizmi, 2-course master

1. Introduction

Nowadays, intelligent mobile robots are the subject that has received large attention. It is a topic of great research concern arising from the possibility of real applications in many areas, such as manufacturing, aerospace, civil engineering, transportation, agriculture, military operations exploration, help for disabled, and medical surgery and in other areas of science and technology research [1].

The main motivation here is to obtain a better under - standing of the characteristics of the network traffic. One of the approaches used for the preventive control is to predict the near future traffic in the network and then take appropriate actions such as controlling buffer sizes [2]. Several works developed in the literature are interested to resolve the problem of improving the efficiency and effectiveness of network traffic monitoring by forecasting data packet flow in advance. Therefore, an accurate traffic prediction model should have the ability to capture the prominent traffic characteristics, e.g. short- and long- range dependence, self-similarity in large-time scale and multifractal in small-time scale [3]. Several traffic prediction schemes have been proposed [4]. Among the proposed schemes on traffic prediction, neural network (NN) based schemes brought our attention since NN has been shown more than acceptable performance with relatively simple architecture in various fields [5].

Neural Networks plays vital role in recommender systems to identify missing values from the dataset and to identify user's navigation patterns and classification. It has many advantages like robust in noisy environments, high degree of accuracy, improves its performance by learning, parallelized, low error rate. Hence in this work, back propagation neural network approach has been proposed and analysed for personalized recommender systems which outperform in accuracy

The aim of this paper is to use artificial neural networks (ANN) based on the multi-layer perceptron (MLP) for identifying and developing a model that is able to analyze and predict the internet traffic over IP networks by comparing some training algorithms using statistical criteria.

2. TRAINING ALGORITHMS

The MLP network training can be viewed as a function approximation problem in which the network parameters (weights and biases) are adjusted during the training, in an effort to minimize (optimize) error function between the network output and the desired output [9]. Most of the well known ANN training algorithms are based on true gradient computations. Among these, the most popular and widely used ANN training algorithm is the Back Propagation (BP) [8]. The BP method, also known as the error back propagation algorithm, is based on the error correlation learning rul. The BP algorithm uses the gradients of the activation functions of neurons in order to backpropagate the error that is measured at the output of a neural network and calculate the gradients of the output error over each weight in the network.

2.1 The Levenberg-Marquardt method

LM is agreed as a standard technique for solving nonlinear least squares problems. It occurs a combination of gradient descent and Gauss-Newton method. LM exhibits adaptive behavior according to the distance of solution so that it can be guaranteed the solution in many cases (Marquardt, 1963). When BP is gradient descent, the algorithm is far from the solution and it is quite slow (Hagan and Menhaj, 1994). Conversely, in the case that BP is Gauss-Newton, the algorithm is close to correct one. In LM, computation of the approximate Hessian given in (1) is done slightly, and the gradient is computed in the manner given in (2).

$$H = J^T J, \ (1) \qquad g_w = J^T e, \ (2)$$

where J and e indicate the Jacobian matrix and a vector of network errors, respectively. LM algorithm uses this approximation in the manner given in (3) such as Newton.

$$w_{k+1} = w_k - [J^T J + \mu I]^{-1} J^T e (3)$$

In summary, GDAs update the weights and biases in the direction of the negative gradient of the performance function. Unlike GDAs, CGAs search steepest descent direction along conjugate directions. QNAs converge faster than CGAs and give better-generalized results. However, the calculations may take a long time. The conjugate gradient and Quasi-Newton only use the first derivative of the function

2.2. The Conjugate gradient method

This is the well accepted iterative technique for solving huge systems of linear equations. In the 1st iteration, the conjugate gradient algorithm will find the steep descent direction.
Approximate solution, x_k for conjugate gradient iteration is described as formulas below [5]:

$$w_k = w_{k-1} + a_k d_{k-1} \quad (4)$$

The next search direction is determined so that it is conjugated to previous search directions. The general procedure for determining the new search direction is to combine the new steepest descent direction with the previous search direction:

$$d_k = -g_w(k) + \beta_k d_{k-1} \qquad (5), \qquad \beta_k = \frac{g_w^T(k) g_w(k)}{g_w^T(k-1) g_w(k-1)} \qquad (6)$$

where □$_k$ is a parameter to be determined so that d_k becomes the k-the conjugate direction.

3. Results

In this part, we are interested in appling the MPL neural networks for developing a model able to identify and predict the internet traffic. The considered data are composed of 1000 points. The databases were divided in two parts training (750 points) and testing (250 points) data as required by the application of MLP. Additionally, the training data set is used to train the MLP and must have enough size to be representative for overall problem. The testing data set should be independent of the training set and are used to assess the classification accuracy of the MLP after the training process [10]. The error analysis was used to check the performance of the developed model. The accuracy of correlations relative to the measured values is determinated by various statistical means. The criteria exploited in this study were the Root Mean Square Error (RMSE), the Scatter Index (SI), the Relative Error and Mean Absolute Percentage Error (MAPE) given by:

$$R_{error} = E\left[\left\{\frac{(y'_m - y_m)^2}{y_m^2}\right\}^{\frac{1}{2}}\right], \quad RMSE = \sqrt{\frac{1}{N}\sum_{m-1}^{N}(y_m - y'_m)^2}, \quad SI = \frac{RMSE}{\acute{y}_m},$$

$$MAPE = \frac{100}{N}\sum_{i=1}^{N}|y_m - y'_m|$$

where y_m and y'_m represent respectively real and estimated data, $\acute{y}_m$ is the mean values of real data and N represents the sample size. **Table 1** shows the obtained results of each statistical indicator for the different algorithms:

From these results, we conclude that the Levenberg-Marquardt (LM) and the Resilient back propagation (Rp) algorithms give more precision using the statistical criteria than the other training algorithms.

Table 1. Values of different statistical indicators for different algorithms

Training algorithms	Rerror	RMSE	SI	MAPE
LM	0.0230	0.0019	0.0222	4.2563
Ggf	0.1448	0.0128	0.1300	4.2528

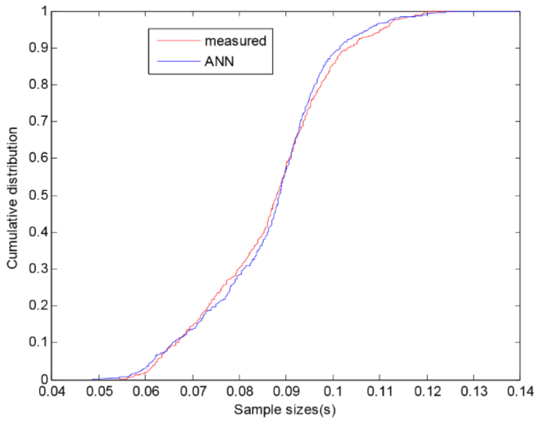

Figure 2. Cumulative distribution of measured and predicted data Rp and LM algorithm

On the other hand, we present in **Figure 2** the cumulative distributions of measured and predicted data. Figure 2 demonstrates clearly the similarity between measured and predicted values. So, the identified ANN model can be used for predicting data of internet traffic.

5. CONCLUSIONS

In this article we present an artificial neural network (ANN) model based on the multi-layer perceptron (MLP) for analyzing internet traffic over IP networks. We used the input and output data to describe the ANN model, and we studied the performance of the training algorithms which are used to estimate the weights of the neuron. The comparison between some training algorithms demonstrates the efficiency of the Levenberg- Marquardt (LM) and the Resilient back propagation (Rp) algorithms using statistical criteria. Consequently, the obtained model using the LM and the Rp can successfully be used as an adequate model for the identification and the management of internet traffic over IP networks. In addition, it can be applied as an excellent fundamental tool to management of the internet traffic at different times, and as a practical concept to install the computer material in a high industrial area.

EFFICIENT ALGORITHMS FOR NETWORK TRAFFIC IDENTIFICATION

Narzullayev Shoxrux Erkin o'g'li, TUIT named after Muhammad al-Khwarizmi, 2-course master,

Traditional classification techniques such as those based on well-known port numbers or packet payload analysis are either no longer effective for all types of network traffic or are otherwise unable to deploy because of privacy or security concerns for the data.

A promising approach that has recently received some attention is traffic classification using machine learning techniques. These approaches assume that the applications typically send data in some sort of pattern; these patterns can be used as a means of identification which would allow the connections to be classified by traffic class. To find these patterns, flow statistics (such as mean packet size, flow length, and total number of packets) available using only TCP/IP headers are needed. This allows the classification technique to avoid the use of port numbers and packet payload information in the classification process.

We apply an *unsupervised* learning technique (EM clustering) for the Internet traffic classification problem and compare the results with that of a previously applied *supervised* machine learning approach. The unsupervised clustering approach uses an Expectation Maximization (EM) algorithm that is different in that it classifies unlabeled training data into groups called "clusters" based on similarity.

Port Number Analysis

Historically, traffic classification techniques used wellknown port numbers to identify Internet traffic. This was successful because many traditional applications use fixed port numbers assigned by IANA. For example, email applications commonly use port 25 and other standart ports can show **Table-1**. This technique has been shown to be ineffective by Karagiannis *et al.* in for some applications such as the current generation of P2P applications which intentionally tries to disguise their traffic by using dynamic port numbers or masquerade as well-known applications. In addition, only those applications whose port numbers are known in advance can be identified.

Table-1

Protocol	Port Numbers
http	80, 8080, 443
smtp	25
dns	53
socks	1080
irc	113
ftp (control)	21
pop3	110
limewire	6346
ftp (data)	20

Payload-based Analysis

Another well researched approach is analysis of packet payloads. In this approach, the packet payloads are analyzed to see whether or not they contain characteristics signatures of known applications. These approaches have been shown to work very well for Internet traffic including P2P traffic. However, these techniques also have drawbacks. First, payload analysis poses privacy and security concerns. Second, these techniques typically require increased processing and storage capacity. Third, these approaches are unable to cope with encrypted transmissions. Finally, these techniques only identify traffic for which signatures are available and are unable to classify previously unknown traffic.

Machine Learning Approaches

Machine learning techniques generally consists of two parts: model building and then classification. A model is first built using training data. This model is then inputted into a classifier that then classifies a data set.

Machine learning techniques can be divided into the categories of unsupervised and supervised. McGregor *et al.* hypothesize the ability of using an unsupervised approach to group flows based on connection-level (i.e., transport layer) statistics to classify traffic. In this method, an EM algorithm is used and McGregor *et al.* draw the conclusion that this approach is promising. In and, Zander *et al.* extend this work by using an EM algorithm called AutoClass and find the optimal set of attributes to use for building the classification model.

Supervised Machine Learning Approach

The Native Bayes classifier is the supervised machine learning approach used in this paper. Assuming that flow attributes are independent and identically distributed, Moore *et al.* applied the Na¨ıve Bayes classifier and found that this approach has good accuracy for classifying Internet traffic. Here we provide an overview of this method and point the interested reader to for details.

The Native Bayes method estimates the Gaussian distribution of the attributes for each class based on labeled training data. A new connection is classified based on the conditional probability of the connection belonging to a class given its attribute values. The probability of belonging to the class is calculated for each attribute using the Bayes rule:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \quad (1)$$

where A is a given class and B is a fixed attribute value. These conditional probabilities are multiplied together to obtain the probability of an object belonging to a given class A. In this paper, we used the Native Bayes implementation in the WEKA software suite version 3.4. This software suite was also used by Moore *et al.* for their analysis.

Unsupervised Machine Learning Approach

The unsupervised machine learning approach is based on a classifier built from clusters that are found and labeled in a training set of data. Once the classifier has been built, the classification process consists of the classifier calculating which cluster a connection is closest to, and using the label from that cluster to identify that connection.

Clustering Process*:* The clustering process finds the clusters in a training set. This is an *unsupervised* task that places objects into groupings based on similarity; this approach is unsupervised because the algorithm does not have a priori knowledge of the true classes. A good set of clusters should exhibit high intra-cluster similarity and high intercluster dissimilarity.

We use an implementation of the EM clustering technique called AutoClass to determine the most probable set of clusters from the training data. AutoClass calculates the probability of an object being a member of each discrete cluster using a finite mixture model of the attribute values for the objects belonging to the cluster. This assumes that all attribute values are conditionally independent and that any similarity of the attribute values between two objects is because of the class they belong to.

When this algorithm is initially run, the parameters of the finite mixture model for each cluster are not known in advance. The EM algorithm has two steps: an expectation step and a maximization step. The initial expectation step guesses what the parameters are using pseudo-random numbers. Then in the maximization step, the mean and variance are used to reestimate the parameters continually until they converge to a local maximum. These local maxima are recorded and the EM process is repeated. This process continues until enough samples of the parameters have been found (we use 200 cycles in our experimental results). A best set of parameters is selected based on the intra-cluster similarity and inter-cluster dissimilarity.

Using Clustering Results as a Classifier*:* Once an acceptable clustering has been found using the connections in a training data set, the clustering is transformed into a classifier by using a transductive classifier. In this approach, the clusters are labeled and a new object is classified with the label of the cluster which it is most similar to.

We labeled a cluster with the most common traffic category of the connections in it. If two or more categories are tied, then a label is chosen randomly amongst the tied category labels.

We also determined that the time required to classify connections can be reduced with the unsupervised clustering technique. The time savings can be achieved because only a portion of the connections in each cluster must be manually identified. Not all clusters are necessarily needed to have fairly accurate results.