A Headway to QoS on Traffic Prediction over VANETs using RRSCM Statistical Classifier

ISHTIAQUE MAHMOOD*, AND AHMAD KHALIL KHAN**

RECEIVED ON 08.05.2015 ACCEPTED ON 16.09.2015

ABSTRACT

In this paper, a novel throughput measurement forecast model is recommended for VANETs. The model is based on a statistical technique adopted and deployed over a high speed IP network traffic. Network traffic would always experience more QoS (Quality of Service) issues such as jitter, delay, packet loss and degradation due to very low bit rate codification too. Despite of all such dictated issues the traffic throughput is to be predicted with at most accuracy using a proposed multivariate analysis scheme represented as a RRSCM (Refined Regression Statistical Classifier Model) that optimizes parting parameters. Henceforth, the focus is towards the measurement methodology that estimates the traffic parameters that triggers to predict the accurate traffic and extemporize the QoS for the end-users. Finally, the proposed RRSCM classification model's end-results are compared with the ANN (Artificial Neural Network) classification model to showcase its better act on the projected model.

Key Words: QoS, Regression, Refined Regression Statistical Classifier Model, VANETs, Artificial Neural Network.

1. INTRODUCTION

VANETs are the currently evolving technology which are well-equipped with security, transport efficiency techniques and surplus information. Several tools seem to attempt on integrating the traffic and networks and have claimed to produce stunning results. Multimedia traffic has nowadays been playing an important role over VANETs. Many traffic prediction algorithms have been proposed by researchers so far, to improve the QoS of VANETs for the end-users. Traffic prediction for VANETs enhances the utilization of the network resources by performing dynamic resource allocation and supports QoS. The efficient proficiencies of the network architecture and the multifarious components of the network are depicted in Fig. 1, using a diversity space illustration. The depicted pictorial representation organizes functional capabilities of the network and fits them accordingly into the corresponding dimension. Each and every network component epitomizes a functional ability which in turn represents them as a fact along the dimension. It also explains that such a space assortment helps the researchers to focus on the required appropriate dimension with supreme clarity. The three components that linger over the transport protocol dimension are given utmost focus in this paper and they are UDP, RTP and TCP. Whereas, the communication medium is not explored here under the

* Department of Computer Engineering, University of Engineering & Technology, Taxila.
** Department of Electrical Engineering, University of Engineering & Technology, Taxila.

several dimensions (or physical connection of a network) [1]. The given diversity space diagram of heterogeneous networks helps to inspect the traffic and dissect it based on certain parameters such as protocols, bandwidth, pareto-distribution etc. The network traffic behaviour falls within the classified two regions, SRD (Short-Range Dependencies) and LRD (Long-Range Dependencies). The proposed classification model can classify the SRD and LRD behaviour in a network traffic and predict them with at most accuracy.

2. RELATED WORK

Prediction on a distributed network has always been considered as a difficult task. Some of the well-known time series models which are being used for forecasting down the lane are listed as follows. They are Box - Jerkins Procedure, Holt-Winters Exponential Smoothing, ANN and meek linear regression. Researchers announced that the techniques that are used to forecast by smoothing the exponent or applying linear regression or deploy an adaptive generalized filter are as a whole termed as "extrapolative" projectors [2]. According to Wang, the subsequent models are characterized as the most regularly deployed foretelling models. Among them are Linear Regression, ARMA (Auto Regression Moving Average), and ANN Model. Some investigators [3] revealed that non-parametric practices largely perform better due to their resilient ability to seize the non-deterministic and

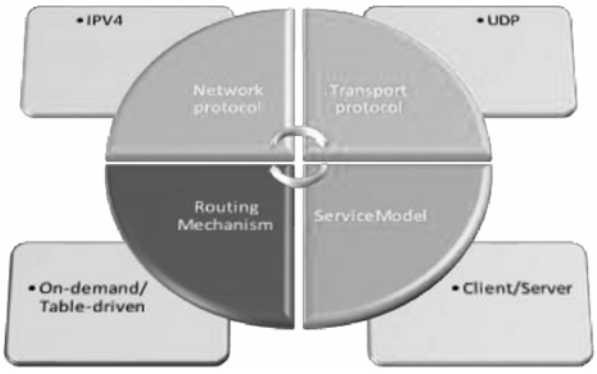

FIG. 1. SPACE DIVERSITY ARCHITECTURE OF A DISTRIBUTED HETEROGENEOUS NETWORK.

multifaceted non-linearity of traffic time series. [4] correlated different natures of Network under BPN (Back Propagation) with economical mathematical models on the data about inflation rate. Altered BPNs considered were: BPN, BPN with ARIMA and BPN with VAR model. The primary glitch with ANN is due to the shortage of clarifying ability and the poor choice made in constructing the approach to describe the architecture of the network. At present, the ANN modeling process is basically empirical. The test result showed hybrid BPN were similar or better than their equivalent econometric model in dynamic forecasting [5]. Researchers [6], made a fair study among ANNs and ARIMA model using 8 years sales data collected from a medium sized enterprise in Brazil which revealed that ANN performs better than ARIMA [7] compared neural network with the regression model OLS (Ordinary Least square) method was used to estimate parameters of regression model using financial stock data. The models performance are henceforth computed by manipulating the appropriate average and variance's square root of the error percentage for which relevant metrics are used to compare them. As summarization of prior research on traffic prediction is done, it is noted that in early 1980's wide variety of parametric approach was deployed such as linear/non-linear regression, ARMA, ARIMA were leading pinnacles in the forecasting domain. Researchers [8] the forecast performance was compared between parametric and non-parametric techniques. The performance analysis indeed guaranteed that least square support vector machines provides a stable and robust approach to predict the domain.[9] stated that among the maximum studied methods of forecasting, the nearby neighbor procedure of non-parametric regression when united with a specific constraint monitoring technique would portray the traffic situations in an improved way. ANN classifier was used by researchers [11] prior to two decades for predicting the traffic. As time surpassed, ANN was coupled with fuzzy-logic to envisage the traffic with better accuracy. An investigation specialist [12] clipped neural network frame with architectures of recurrent radial

basis function and echo state network to predict wireless network traffic and quoted that the accuracy ranges between 96.4 and 98.3%.Eventually the fusion of neural network architecture in the domain of prediction had played a vital role and henceforth, some researchers tried to tune them more by applying genetic algorithm within ANN [13]. Investigators [14], had devised a new discontinuous-time parametric random processing model, which was termed to be called as SV (Stochastic Volatility) model to predict short-range traffic. The prognostic enactment of the SV stereotype when related to GARCH (Generalize Autoregressive Conditional Heteroskedasticity) model had predicted the traffic variability with higher accuracy. Researchers [15], had performed a comparative study on various prediction models such as neuro-fuzzy model, the ARMA model and the integrated ARIMA model. The principle of deploying SVR (Support Vector Regression) practice for forecasting were hosted by [16] paved a way to model the traffic physiognomies and envisage the traffic statuses.

3. THE PROPOSED METHOD

The exertion emphasizes on the scheme, the practical appraisal and scrutiny of the conduct which will train the proposed model classifier to predict the throughput which will forecast the arriving input degree in Megabits per second. The proposed action flow involves 5 phases which includes traffic collection, Pre-processing the traffic data, Classifying the data, Clustering, Applying predictor classifier and Visualizing the ROC (Receiver Operating Characteristic) as shown in Fig. 2.

Phase-I: Traffic Collection: There exists numerous ways to collect a trace, where the packets arrival, departure and other associated information are recorded regarding flow of the traffic. A traffic trace which has been used by researchers are collected here for the evaluation purpose. The VANET traces were downloaded from the following link:http://www.lst.inf.ethz.ch/research/ad-hoc/car-traces/index.html#download

Phase-II: Pre-processing: The pre-processing of data involves data reduction and transformation phases, later the traffic is clustered and the relevant classification model is build which gets initially trained and later tested.

The trace file downloaded from the given link is run through NS2 Visual Trace Analyser, an application capable of analysing NS-2 trace files which makes graphics/ statistics of the captured traffic. The graph is then exported with its data to CSV, so that it can imported into the SPSS for analysis and indeed convert the same to ARFF format to run on the proposed RRSCM deployed on WEKA. The tools WEKA and SPSS henceforth use the data files imported as .ARFF and .CSV files. As the data files .ARFF and .CSV are generated from the traffic trace downloaded as specified in phase I, its thereafter connected. The tools deployed to perform pre-processing of the data imported is done with filters. In order to create compatible train and test set, **batch filtering** is necessary which is a filter approach that is straightforward, it just filters the data through the filter and obtains the reduced dataset. The filters tend to discretize by transferring the continuous data into discrete counterparts and are normalized further by eliminating the properties of gross influences. They are later resampled and appropriate attributes/predictor

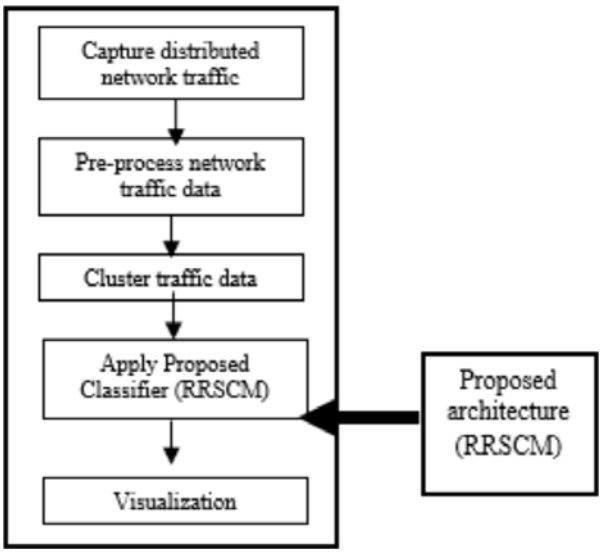

FIG. 2. PROPOSED ACTION FLOW SCHEME.

variables(PV)are selected for the forecasts. Traffic traces are imported as .ARFF and .CSV data files to be used in WEKA and SPSS accordingly. Batch filtering is deployed to reduce the dataset further.

***Phase-III: Clustering*:** It finds and groups similar instances in the dataset and groups them together. Clustering is performed by **EM** (Expectation Maximization)) **algorithm**, the reason why and how is revealed as follows. The Cluster analysis groups the traffic flow based on the flow attributes. An auto-class algorithm otherwise referred as EM algorithm fixes the best set of predictor attributes to be used. Henceforth, it withholds a two-step action that initially guesses the parameters at the expectation-level and later, re-estimates the parameters mean and variance as they converge to the local maximum. It uses a Bayesian score to stamp and hold the best parameters to be used in for the prediction in future. The justification given to choose this clustering approach is based on 2 points.

(i) Time to build is less considerably compared with the other clustering models

(ii) Absolute error is quite low.

***Phase-IV: Classification*:** Classifiers are prototypes which are applied for predicting trivial or numeric quantities. The basic element classifier build in weka adopts the regression function to classify the instances based on its correlation with the DV (Dependent Variable). The proposed predictor classification model RRSCM is applied to overcome issues of linearity, auto-correlation, multi-collinearity and normality. The dataset file is used to train, then test and compare the prediction results with existing ANN classifier model in weka.

***Phase-V: Visualization*:** Pictorial representations are very useful in comparing the results and helps the investigators further to regulate toils of the learning issue by deploying respective classifiers. Why visualization? The ROC Curve focuses on TPR and FPR (i.e. True Positive/False Positive Rates). The TPR defines exactly how many correct positive outcomes arise amid of all positive trials available during the test. FPR, on influence, describes exactly how many inappropriate positive outcomes arise amongst all negative trials offered during the test.

The ROC space grid is well-defined by false positive and true positive as their corresponding x and y axis as shown in Fig. 3. The prediction results yielded by the RRSCM classifier on the selected 38 attributes may fall in any of the four categories of the space grid as shown below. It can be TP (True Positive), FP (False Positive), FN (False Negative) or TN (True Negative). The dotted diagonal line splits the ROC grid into two slots and the results of the classifier model are interpreted in such a way that the prediction results yielded along the upper diagonal line in the grid space from the left bottom to the top right angles are marked or graded to be a perfect classifier rather than the ones lying low below the diagonal grid space. In context to the ROC curve depicted at Fig. 4. It henceforth reveals that all the results/points yielded upon the classifier RRSCM lies above the upper diagonal line in

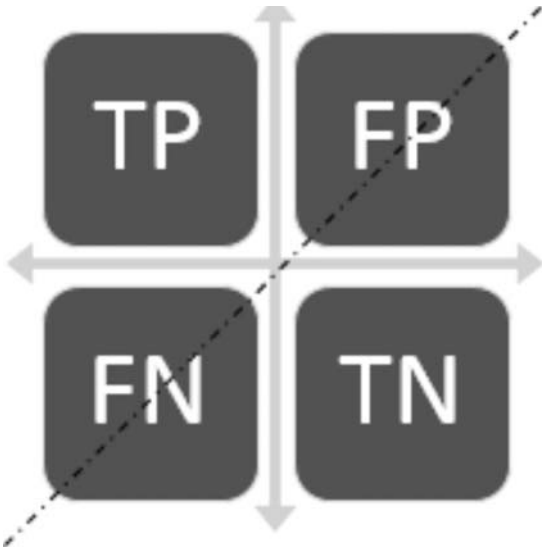

FIG. 3. ROC GRID SPACE

the grid space from the left bottom to the top right angles. So, this henceforward helps us to justify that the classifier yields better prediction result as the TP rate is noted to be high.

The proposed RRSCM involves the following activities, the pictorial representation is shown below in Fig. 5. In OLS model the Equation (1), represents Y as a Dependent variable, PV_i as independent predictor variable for the i^{th} observation (*i* is the number of observation), ε- Error.

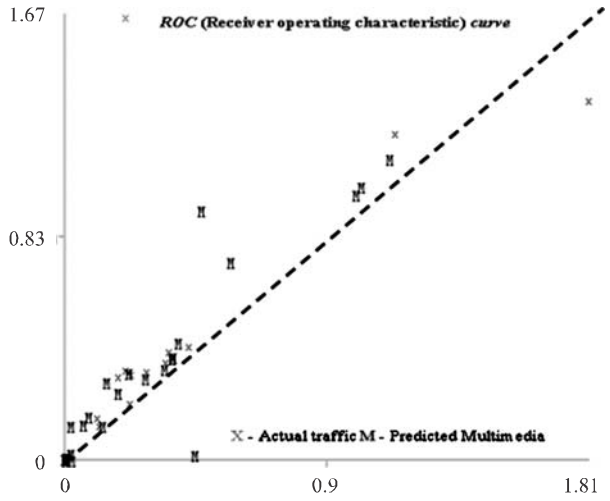

FIG. 4. ROC FOR PREDICTED MULTIMEDIA TRAFFIC

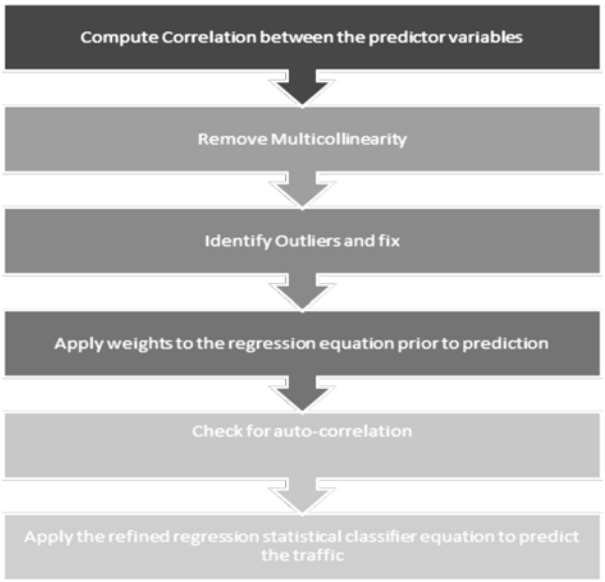

FIG. 5. PROPOSED WORKING ARCHITECTURE OF RRSCM.

(a) Independent variables such as x1, x2, x3…xn with their corresponding influential coefficients â1, β2, β3… ân. á the intercept coefficient. The error term epsilon, represented as å.

(b) $$y_i = (\alpha + \beta_i * PV_i + \varepsilon_i) \quad (1)$$

The proposed RRSCM model uses the same statistical regression equation but revised and refined on applying some postulates. The defined postulates re-transform the parameters influencing the prediction of traffic and help us to yield a perfect forecast. The details are represented with a diagram on Fig. 6.

- The independent variables or predictor variables referred as IV or PV (attributes that contribute the traffic) over the DV (Dependent Variable) are show cased as x1, x2, x3…xn.
- The discriminators/parameters/attributes referred as IVs have corresponding influences represented as β1, β2, β3… ân for the traffic that is to be predicted as Y (dependent variable referred as DV).
- A simple change in any of the independent variables influences β1, β2, β3… βn will reflect on the computation of Y (DVs), the predictable traffic.
- In case, all the attributes at a point do not contribute any influence over the predictor variable Y (DV), then there exists x0 an assumed PV which is always 1.
- Hereafter, x0 hold an influence á otherwise stated as the intercept coefficient on the predictor variable Y.
- á is a constant, intercept coefficient in the equation as it would influence and yield Y.

There are four postulates framed to address the issues that may arise in as we use the deploy the equation to predict the traffic results for our VANETs. The issues are henceforth addressed in order as dictated below.

Postulate-1: Addresses the linearity issue

Postulate-2: Reports the homoscedasticity, or equal Y variance across the values of X (independent variables)

Postulate-3: Uncorrelated standard errors terms

Postulate-4: Normality of the standard error terms

All the 4 postulates are framed to check and rectify the issues in case of violations during predicting the traffic.

Postulate-1: The postulate on linearity check is safeguarded by calculating the standard error term on every observation. The standard error is measured as shown in Equation (2).

$$\varepsilon = \sqrt{p(1-p)} / n \tag{2}$$

Where n is the sample size and p is the reported standard error proportion. As a check on such standard error equated to zero, it guarantees linearity and henceforth is showed that the expected standard error is to be assumed zero as shown in Equation (3). If the condition is violated the corresponding Independent variable (IV) will be transformed by the proposed RRSCM classifier model.

$$E(\varepsilon) = 0 \tag{3}$$

Postulate-2: For a particular value of x (IV) there are possibilities to yield several values of y, which when plotted over the graph will reveal variability (variance, σ_y). It is observed that the variability of y (σ_y) should be equal as shown below in Equation (4). The squared variance of individual y(σ^2_y)components with respect to every independent variable x should be the same and are finally equated to a constant term.

$$\sigma y_1^2 = \sigma y_2^2 = \sigma y_3^2 \ldots = \sigma y_i^2 = \sigma y_n^2 = \sigma^2 \tag{4}$$

$$E(\varepsilon^2) = \sigma^2 \tag{5}$$

This shall also ensure the constant variability on the standard error terms, as its variance is squared. As shown in Equation (5) the Y variance remains the same across the X (IVs). In case if it is violated, then the attributes will be transformed to avoid heteroskedasticity. The issue of heteroskedasticity raises, as the variance of Y differs accordingly with the corresponding IVs.

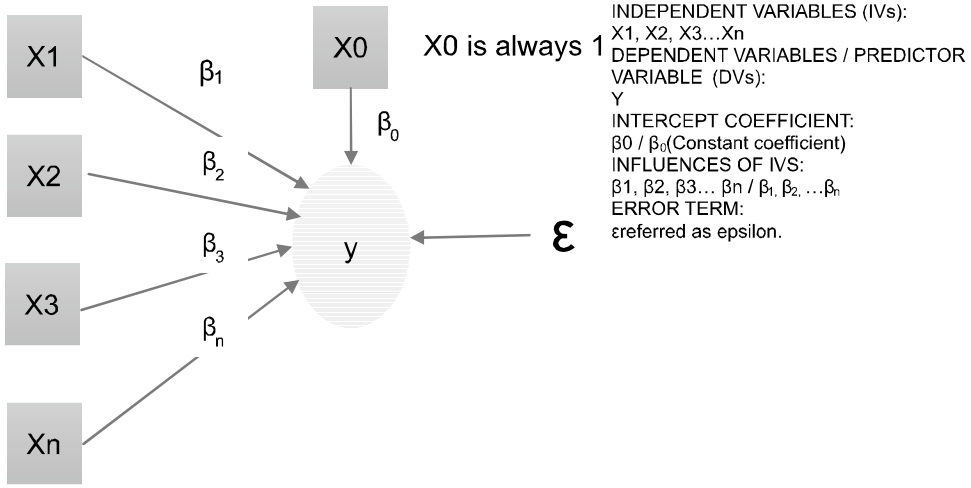

FIG. 6. PARAMETERS FLOW THAT IMPACT THE FORECAST

Postulate-3: The standard error terms for an observation i and observation k yields error terms of $å_i$ and $å_k$ respectively. The COV (Co-Variance) between the error terms ε_i and ε_k is expected to be zero to guarantee uncorrelated error existence as shown below in Equation (6).

$$E(COV(\varepsilon_i, \varepsilon_j)) = 0 \quad (6)$$

As there are n errors, the co-variance between the errors shall have a matrix with diagonal elements, whose variance will be zero and the error will follow a normal distribution as stated in postulate 4.

Postulate-4: As there are n observations, there exists n error terms ($\varepsilon_1 \ldots \varepsilon_i \ldots \varepsilon_n$) which is said to follow normality or normal distribution (N). The error terms take care of the variability issue, helps to acquire a good generalization and achieve better prediction results. Henceforth, this shall come true if and only if normality is achieved as specified in Equation (7).

$$\varepsilon i \approx N(0, \sigma_y^2) \quad (7)$$

The standard error rate ε_i will follow a normal distribution within a range of minimum zero and the variance (sigma/ σ) squared for y (σ_y^2. All these postulates when violated may result in discrepancies during the forecast. Henceforth, our RRSCM proposed classification model algorithms are deployed in weka that overrides all these issues and yields perfect classification results as shown in ROC diagram.

If all the illustrated four hypothesis are held good or fulfilled then this proposed model will yield the best unbiased estimation. If the assumptions are violated, if the first assumption is not met or violated then it means that there is a biased element estimation made. If postulate 2 is violated, is a crucial assumption. The variance of some of the observations will be different with the variance of the other observations which leads to the issue of Heteroscedasticity. If postulate 3 is violated, relates to issue of auto correlation. If postulate 4 is violated, find that the error term does not influence the independent variable. The norms and corresponding issues when violated are listed in **Table 1**. The adopted algorithms to overcome the violations are briefly discussed below in detail. Estimation of β^o for the equation is obtained as:

$$\beta^o = \frac{\sum XY}{\sum X^2} \quad (8)$$

β^o as shown in Equation (8), customs mean and variance and the expectation of β^o, β^o as is equal to β then it concludes that it's not a biased estimate, otherwise is dictated to be biased.

Algorithm-1, guides the selection of attribute via the perfect pathway. The multicollinearity issue is measured with the formula given in line number 3. It can be measured by tolerance and VIF (Variance Inflation Factor). Tolerance is the variance percentage that is unaccounted by the other IVs (Independent Variables). Tolerance is a multiple regression analysis were the IV is regressed with the other IVs in the MR analysis, which derives R^2 value is obtained. $1\text{-}R^2$, leftover of variance which not accounted for is considered as tolerance. Tolerance values are mostly 0.10 or less are sighted as problematic. As 0.20 is the lowest suggested tolerance value, which may further lead to instability. VIF is the reciprocal of tolerance. VIF is correlated with tolerance, it indicates the degree of inflation of the standard error rates due to the levels of collinearity or multicollinearity. VIF values of 10 and above

TABLE 1. NORMS AND ISSUES FILTERED IN RRSCM

Violated Norms	Issues
Postulate 1	Biased estimation
Postulate 2	Heteroscedasticity
Postulate 3	Auto-correlation
Postulate 4	Multicollinearity

will sighted as problematic. Here the standard error rates are supposed to be inflated by factor of 10 or above where there wouldn't be any correlation between the independent variables.

Stability is achieved in the selection of predictor variables by avoiding multicollinearity are through the removal of redundant variable, aggregating similar independent variables/attributes of choice and by increasing the sample size.

Algorithm-2, concentrates on identifying the PVs that may mislead the estimation by producing higher/ increased coefficients. A standard regression model is mathematically represented as depicted in Equation (9), in which Y_j is the j^{th} observed predictor variable value, PV_{ij} is the j^{th} observed predictor variable value for i^{th} variable and CE_i is the regression coefficient to be defined for a dependent variable. M will be the number of data points and N signifies the number of terms in the regression equation.

$$S = \sum_{j=1}^{M} \left(Yj - pf\left(x\right)\right)^2 \tag{9}$$

Where

$$pf\left(x\right) = \sum_{i=1}^{N} \left(CE_i * PV_{ij}\right) \tag{10}$$

Algorithm-3, it is noted that all errors have the same significance, henceforth the variation in error significance is showcased by using a weighted regression as represented in Equation (10).

$$S = \sum_{j=1}^{M} WF_j \left(Yj - f\left(x\right)\right)^2 \tag{11}$$

Therefore, the squared difference between the perceived and forecast value is multiplied with the weights (WF) computed accordingly as formulated in line number 2, to have more accurate estimation.

Algorithm-4, verifies for the auto-correlation issue. In Algorithm-4, the error correlation on various observations are noted and those PVs are transformed further to yield appropriate accurate prediction results. These boosting techniques are introduced into traffic prediction by exploiting all possible regression equations to have the best forecast of the traffic using our proposed RRSCM model. The general form which is used to predict the traffic as shown in Equation (11). The variables pv1, pv2, pv3…pvn are the independent-variables that help us to predict the dependent variable. The dependent variable dataset relies on the dissected traffic on the network biased on the protocols, which leads us to the generation of the prediction equation. Table 2 projects the issues that has to be overcome to have an unbiased prediction estimation. The formula for calculating Coefficient Correlation (r or R) that dictates the degree of association as shown in Table 3, between the dependent and the independent variables (PVs) are given in Equation (12). Everywhere p and q are independent and dependent variables respectively.

$$R = \frac{n\sum\left(p * q\right) - \sum p * \sum q}{\sqrt{\left[n\sum p^2 - \left(\sum p\right)^2\right] * \left[n\sum q^2 - \left(\sum q\right)^2\right]}} \tag{12}$$

TABLE 2. VALIDATED RRSCM ISSUES

Validated Refined Regression Statistical Classification Model	Plaid	Algorithm
Data are not ill-conditioned.	Variance Inflation Factor (VIF)	Removal of multicollinearity
Strange/ Unusual observations or outliers.	FIT	Identify outliers and removal
Heteroscedasticity	Weight Factor	Weighted Regression
Auto-correlation	Dublin-Watson statistic	Measure Significance

4. DISCRIMINATOR AND CLASSIFICATION MODELS – SPSS

A traffic trace of a shorter period is taken into account to evaluate the prediction using MLP (Multilayer Perceptron) and RRSCM classification model in SPSS. The discriminators that measure the dependent variables are 38 which are referred as $PV_1 \ldots PV_{38}$. The summary of the prediction using MLP is shown in Table 4. Investigations on the outcomes of reliable predictor variable values and the scattered correlation relying on each dependent element on the forecasted precision rate leads to probe in further to develop the predictor equation using multiple regression method and is done. The upshots of the examination front-runs to find the consequence of the reliant variable values spread on prediction precision that probes and clues us to generate an equation. Finally, the generated equation would envisage the probable traffic based on the liberated variable-values dispersal using the modelling tool SPSS as shown in Table 5. The MLP architecture is chosen in ANN model to predict the traffic.

TABLE 3. CORRELATION COEFFICIENTS OF DIFFERENT DEPENDENT VARIABLES

Correlation Coefficient	Dataset-Combinations
0.988	RRSCM_1(y):(Predictor Variables (PVs), with the selected 38 attributes)
0.953	RRSCM_2(y):(PVs of random choice)
0.904	RRSCM_3(y):(All PVs)

TABLE 4. MODEL-SUMMARY – ANN-MLP

Classifier	ANN-MLP
Cross Entropy Error	1.187
Percent Incorrect Predictions	12.5%

TABLE 5. MODEL-SUMMARY – RRSCM

Classifier	R/r	R2	R2-Adjusted	Estimation on Standard Error degree
RRSCM	.994a	.988	.987	.185

5. DISCRIMINATORS AND CLASSIFICATION MODELS – WEKA

Among the various IVs influencing the dependent variable of prediction on the throughput (discriminator 111), 38 discriminators have been used by the RRSCM classifier model to get along with the prediction measures. The discriminators are defined in detail as shown in Table 6. The Table 7 shows reports on the data-set being casted-off for traffic training and testing phases deployed over various classification models to forecast the distributed short term network traffic. The discriminators which are acting as independent variables are being pre-processed, clustered to be classified under the MLP and regression classification model as shown below and their corresponding prediction results are show cased in Tables 8-9. The day trace has been fragmented into several blocks. The discriminator centred classification is planned for all runs of foreseeing the traffic. The total number of instances of the traffic is 21648 and the number of fields taken into account is 249.

6. EXPERIMENTAL RESULTS

6.1 RMSE, MAPE and Correlation Coefficient Metrics

The RMSE (Root Mean Square Error) is one of the universally deployed metric [17]. The SE (Square Error) is the central component of RMSE. RMSE rubrics out the issue of negative and positive inaccuracies annulling out each other by captivating the square of each error. In conclusion, taking square root of the average squares of error among predicted and real values acquires the accurate RMSE. The RMSE Equation (13) is represented as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} n\left(D_t - F_t\right)^2}{n}} \tag{13}$$

TABLE 6. DISCRIMINATORS AND DEFINITIONS

Number	Short	Long
12	median_data_eth_pkt	Average bytes in packets
19	total_median_data_ip	Average of the aggregate bytes in IP packets
26	median_data_cntrl	Median of control bytes in packet
31	total_pkts_cs	The total number of packets perceived
38	server_ack_pkts	Acknowledged (Server to Client)
42	max_server_ack_blk	Maximum-Acknowledged (Server to Client)
43	unique_byts_directed	The amount of unique bytes directed which the bytes retransmitted.
45	count_actual_data_pkts	The packets with a minimum byte is also counted on TCP data payload transmitted from client to server.
47	total_actual_data_byts	The over-all bytes of information perceived which includes the bytes retransmitted too. (Client to Server)
59	push_data_pkts	The packets grasped on setting thePUSH bit-set in the TCP header.(Client to Server)
74	server_acks_directed	Acknowledgement directed (Server to Client)
111	throughput_cs	The average throughput. (Client to Server)
113	RTT_trials	The total amount of Round-Trip-Time (RTT) trials are noted from Client to Server.
125	RTT_full_size_trials_cs	The entire number of full-size RTT samples are calculated from the RTT trials of standard segments from Client to Server.
126	RTT_full_size_trilas_sc	The entire number of full-size RTT samples are calculated from the RTT trials of standard segments from Server to Client.
133	RTT_full_size_SD_cs	The RTT trials standard deviation from Client to Server is noted.
135	inf_loss_acks_cs	The number of acknowledged packets acquired after lost ones are noted and detected, for which re-transmission occurs from Client to Server.
137	seg_cumm_ack_cs	The amount of segments that were cumulatively acknowledged and the o nes not directly acknowledged from Client to Server are noted.
139	redundant_acks_cs	Number of duplicated acknowledgments acquired from Client to Server)
141	triple_duplicate_ack_cs	The amount of threefold duplicate bylines acquired from Client to Server.
154	first_quartile_data_cs	First quartile of bytes in (Ethernet) packet
155	median_data_cs	Average bytes in (Ethernet) packet
161	first_quartile_data_ip_cs	Bytes in first quartile of IP packet
171	third_quartile_data_cntrl_cs	Control bytes in the third quartile
210	transitions_bulk_trans-mode	The total transitions between transaction and bulk-transfer mode.
211	Time_spnt_bulk	Total time expended in bulk transfer mode
213	Percent_bulk	Time in percentage, expended in bulk transfer
Discriminators 221,223,225 and 227 acquires the FFT (Fast Fourier Transform) of the packets inter-arrival time with the top frequencies ranked by the magnitude are noted.		
Discriminators 231,233,235 and 237 acquires the FFT (Fast Fourier Transform) of the packets inter-arrival time with the top frequencies ranked by the magnitude from client to server.		
Discriminators 241,243,245 and 247 acquires the FFT (Fast Fourier Transform) of the packets inter-arrival time with the top frequencies ranked by the magnitude from server to client.		

Wherever D_t denotes the factual significance, F_t dictates the estimated significance and n notifies the observations. Since the errors are squared before they are averaged, the RMSE contribute reasonable high weightage to more errors, but this distress is somewhat meticulously controlled by hosting the square root at the end. The superior the value of RMSE signifies an inferior estimation on ANN classification model as shown in Table 10. Mean Absolute Percentage Error (MAPE) is the most shared and widespread error gauging metric for forecasting. MAPE evaluates the mean of absolute percentage error which is stress-free to comprehend and estimate. The MAPE is epitomized by the Equation (14).

$$MAPE = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{D_t - F_t}{D_t}\right| \quad (14)$$

TABLE 7. DATA SET REPORT TABLE

Data-set details	Description
Data-set used by researchers Andrew W. Moore and Denis Zuev	Instances: 21648 Fields: 249 Selected PVs: 38

TABLE 8. EVALUATION – SUMMARY (ANN-MLP)

Correlation-Coefficient	0.644
Average magnitude of error (MAE)	11.3959
Quadratic mean (RMSE)	0.5419
Relative vs the absolute error	32.60%
Root Relative-Squared Error	0.26%
Total time to construct the model: 0.21 seconds	

TABLE 9. EVALUATION – SUMMARY (RRSCM)

Correlation-Coefficient	0.99
Average magnitude of error (MAE)	0.1916
Quadratic mean (RMSE)	0.4392
Relative vs the absolute error	0.0508 %
Root Relative-Squared Error	0.0905 %
Instances	21648
Fields	249
Total time to construct the model: 0.2 seconds	

Wherever D_t is the real value and F_t is the anticipated value. To conclude, the MAPE for the classification models discussed above are computed and equated accordingly as shown in Table 11. However, in here three diverse error metrics are applied for evaluating the classification models. They are the RMSEs, MAPEs and correlation coefficient (r). Henceforth the same are evaluated and summarized to show that RRSCM classification model can outperform ANN classification model as the attribute selection for prediction are made precisely based on the architecture proposed. In order to compare the forecasting accuracy in normal conditions numerically, there are two measures of effectiveness: the root mean squared error performance metric and the mean absolute percentage error performance metric [18]. The RMSE is a representative of the size of a typical error. The MAPE is another commonly used measure of effectiveness for purposes of reporting because it is expressed in percentage terms, which give us a general sense of the error even without knowledge of what constitutes a big error for the data set. From these results, one can conclude that RRSCM model performs best among the other model as the attribute selection and non-missing data situations are rectified. The relationships between the considered errors matrices (i.e. MAPE, RMSE and MSE) are examined considering the linear correlation between the actual traffic and the forecasting multiple regression equation model is shown below in Table 12.

TABLE 10. SUMMARIZED RMSE VALUES OF THE CLASSIFICATION MODELS

Classification Model	SPSS (%)	WEKA (%)
ANN	15.99	7.66
RRSCM	11.78	4.12

TABLE 11. SUMMARIZED MAPE VALUES OF THE CLASSIFICATION MODELS

Classification Model	SPSS (%)	WEKA (%)
ANN	6.61	2.7
RRSCM	4.61	1.24

The actual traffic is graphically represented with the blue color and the predicted traffic showcased in orange color, using the ANN classifier in weka does not correlate much as shown in Fig. 7. The variations are also hereby showcased using the standard error corrections existing over the classifiers. The other Fig. 8 depicts that the actual traffic in blue line correlates and seems to browse the predicted traffic line in the orange color. The RMSE henceforth paves a way to justify that the average squared error rates are low in range and is eventually depicted in the experimentation summary table as 0.4 for the RRSCM classifier. The average magnitude of the error is computed to be less for the proposed RRSCM classifier when compared with the other ANN classifier. As we debate on the performance of the ANN classifier over the proposed RRSCM, the model RRSCM seems to reveal better outcomes as depicted in Fig. 9. The graph denotes a better correlation rate for RRSCM with respect to actual and the predicted traffic and lower error rates on RMSE, MAPE and relative vs absolute loss. Henceforth, it is approved through proper experimentation that RRSCM is a better classifier over ANN as it is tuned to produce better outcomes predicting the traffic for VANETs. Any metric which processes the error should possess five elementary potentials which are listed as follows: validity, easy to interpret, reliability, presentable and have statistical equation.

6.2 Visualize ROC

ROC (Receiver Operating Characteristic) *curve* [19], for the tested prediction results on the dataset are run under various classifiers and multiple ROC's are analyzed for the 38 attributes and the ROC for discriminator 143 is analysed. A ROC curve for multimedia traffic prediction on a plotted time series is depicted in visualization phase. It helps to evaluate the performance of classifiers by its visual representation. To generate these ROC flow, the knowledge flow environment in weka is explored. The flow constitutes of the ARFF Loader followed by the Class assigner and picker which are cross validated and further pushed forward to opt the corresponding classifier of ANN and RRSCM accordingly. The cross validation fold maker otherwise referred as CV maker validates the dataset and forwards it as 2 sets to both the classifiers one dictated as the training set and other the test set. To evaluate the performance of the classifiers it forwards the result as a batch classifier to the performance evaluator that toil as a middle man. The evaluated threshold data is further forwarded to the model performance chart which

TABLE 12. CORRELATION COEFFICIENT BETWEEN THE ACTUAL AND THE FORECASTED TRAFFIC

Correlation Coefficient (Actual vs. ANN Classifier)	0.926608
0.926608 Correlation Coefficient	0.999733

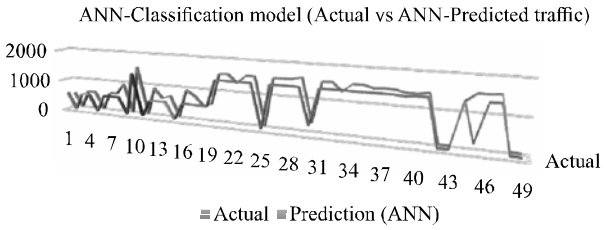

FIG. 7. ACTUAL VS PREDICTED TRAFFIC ANN CLASSIFIER.

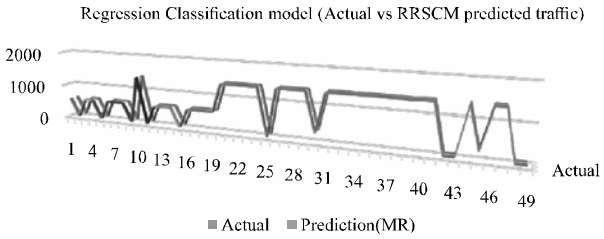

FIG. 8. ACTUAL VS PREDICTED TRAFFIC RRSCM CLASSIFIER.

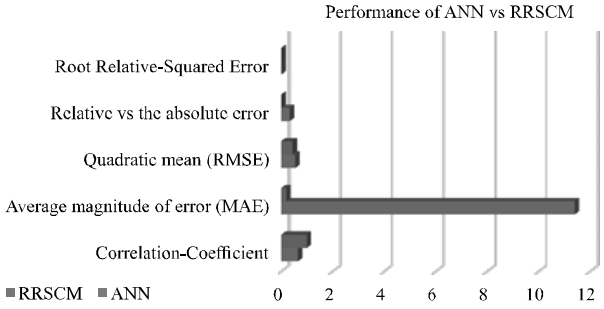

FIG. 9. PERFORMANCE ANN VS RRSCM

helps to visualize the ROC curves. Henceforth analyzing the multiple ROC's for the dataset based on 38 different discriminators the actual vs predicted traffic using both the classification models of ANN and regression are compared using the graphical plots accordingly.

7. CONCLUSIONS

The investigated outcomes exhibits that:

(i) The RRSCM outperforms the ANN classification model and is labelled to be the prominent effective model to forecast the traffic with ease.

(ii) The projected prediction equation and standard error based on R (correlation coefficient) modifies the pattern of effective probability of the forecasted traffic. In a precise way, it keeps track of all qualities over the performance metrics as discussed prior in the experimental results section. The future direction of this exploration is to, deploy the regression classification approach to a long-term traffic and predict the results with the similar accuracy as in short-term traffic data, then compare the throughput and queue size with the collected traffic trace as input. The work can also be envisaged to assimilate the selected predictor classification model into a network management system and assess it on real-time.

ACKNOWLEDGEMENT

The authors would like to thank "**almighty ALLAH**", for providing every support to accomplish this independent research work successfully. We would also like to express our sincere gratitude towards the cooperation and encouragement rendered by the Department of Computer Engineering, University of Engineering &Technology Taxila, Pakistan.

REFERENCES

[1] Andersen, C.M., and Bro, R., "Variable Selection in Regression—A Tutorial", Journal of Chemometrics, Volume 24, Nos. 11#12, pp. 728-737, 2010.

[2] Aneiros-Pérez, G., and Vieu, P., "Nonparametric Time Series Prediction: A Semi-Functional Partial Linear Modeling", Journal of Multivariate Analysis, Volume 99, No. 5, pp. 834-857, 2008.

[3] Simpson, S.L., DuBois, B.F., and Paul, J.L., "Analyzing Complex Functional Brain Networks: Fusing Statistics and Network Science to Understand the Brain", Statistics Surveys, Volume 7, 2013.

[4] Basak, D., Pal, S., and Patranabis, D.C., 'Support Vector Regression", Neural Information Processing Letters and Reviews, Volume 11, No. 10, pp. 203-224, 2007.

[5] Camargo, M.E., Gassen, I.M., Russo, S.L., dosSantos, D.A.I, and Moiseichyk, A.E., "Performance of the Two Approaches Forecasting Model of Sales".

[6] Clarke, J., and Subramanian, A., "Dynamic Forecasting Behavior by Analysts: Theory and Evidence", Journal of Financial Economics, Volume 80, No. 1, pp. 81-113, 2006.

[7] Dashevskiy, M., and Luo, Z., "Time Series Prediction with Performance Guarantee", Communications, IET, Volume 5, No. 8, pp. 1044-1051, 2011.

[8] Yoon, B., and Chang, H., "Potentialities of Data-Driven Nonparametric Regression in Urban Signalized Traffic Flow Forecasting", Journal of Transportation Engineering, Volume 140, No. 7, 2014.

[9] Zheng, Z., and Su, D., "Short-Term Traffic Volume Forecasting: A k-Nearest Neighbor Approach Enhanced by Constrained Linearly Sewing Principle Component Algorithm", Transportation Research Part-C: Emerging Technologies, Volume 43, pp. 143-157, 2014

[10] Weichang, H., Yucheng, D., and Feifeng, Z., "Forecasting Urban Traffic Flow by SVR with Continuous ACO Original Research Article", Applied Mathematical Modelling, Volume 35, No. 3, pp. 1282-1291, 2011.

[11] Guang, S., "Network Traffic Prediction Based on the Wavelet Analysis and Hopfield Neural Network", International Journal of Future Computer and Communication, Volume 2, No. 2, pp. 101-105, 2013.

[12] Khashei, M., and Bijari, M., "An Artificial Neural Network (p,d,q) Model for Time Series Forecasting", Expert Systems with Applications, Volume 37, No. 1, pp. 479-489, 2010.

[13] Shirsath, P.B., and Singh, A.K., "A Comparative Study of Daily Pan Evaporation Estimation Using ANN, Regression and Climate Based Models", Water Resources Management, Volume 24, No. 8, pp. 1571-1581, 2010.

[14] Erdem, E., and Shi, J., "ARMA Based Approaches for Forecasting the Tuple of Wind Speed and Direction", Applied Energy, Volume 88, No. 4, pp. 1405-1414, 2011.

[15] Faruk, D.Ö., "A Hybrid Neural Network and ARIMA Model for Water Quality Time Series Prediction", Engineering Applications of Artificial Intelligence, Volume 23, No. 4, pp. 586-594, 2010.

[16] Hu, W., Yan, L., Liu, K., and Wang, H., "A Short-Term Traffic Flow Forecasting Method Based on the Hybrid PSO-SVR", Neural Processing Letters, pp. 1-18, 2015.

[17] Jolliffe, I.T., and David, B.S., (Editors), "Forecast Verification: A Practitioner's Guide in Atmospheric Science", John Wiley & Sons, 2012.

[18] Montgomery, D.C., Peck, E.A., and Vining, G.G., "Introduction to Linear Regression Analysis", John Wiley & Sons, Volume 821, 2012.

[19] Harrell, F.E., "Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis", Springer Science & Business Media, 2013.