Online Traffic Prediction with Big Data: A Naive Bayesian Classification

Dr. N.SADHASIVAM[*1] **Dr. C.VEERAMANI** [*2] **Dr.M. PANDI**[*3]
[*1,2,3]Professor Department of Computer Science & Engineering
[*1,2] Sri Indu College of Engineering & Technology, Hyderabad, Telangana.
[*3]Dr. Mahalingam College of Engineering & Technology, Tamilnadu.

Abstract— **Traffic information can be derived from various sources for developing traffic prediction techniques, which in turn improve navigation of the route, traffic regulation and urban area planning. One key challenge for predicting traffic is how much to depend on prediction models that are constructed using historical data in real time traffic situations which may differ from that of the historical data and change over time. Existing approach is used for predicting traffic models that are learned offline or they are retrained after long periods and thus they cannot adapt to dynamically changing traffic situations. To overcome this problem, online traffic prediction with Big Data: A Naive Bayesian Classification algorithm has been proposed. From the current traffic situations in real time, future traffic can be predicted by matching the current traffic situations to the most effective prediction model trained using historical data. When real time traffic data arrives the traffic context space is adaptively partitioned using machine learning algorithm in order to efficiently estimate the effectiveness of each base predictor in different situations. The proposed approach also obtains and proves both short term and long term performance for online algorithm. The proposed algorithm also works effectively in scenarios like, when the true labels are missing or become available with delay.**

Keywords— Naïve Bayesian classification; Machine Learning; Traffic Prediction.

I. INTRODUCTION

Traffic has been growing in major cities around the world. Traffic congestion causes tremendous loss in terms of both time and energy. Traffic congestion is caused when the traffic demand approaches or exceeds the available capacity of the traffic system. Traffic prediction should take place in order to avoid traffic congestion. Several traffic prediction techniques have been used. The majority of these techniques focus on predicting traffic in typical conditions like peak hours, weather conditions, etc. Various machine learning techniques are available. Naive Bayesian classification is chosen as one of the machine learning techniques. In this work, Naive Bayesian classification has been applied to predict the typical traffic conditions and the impact of accidents.

Traffic prediction can be made from the current traffic situation in real time data that is constructed using historical data and predicts the future traffic by matching the current traffic situation to the most effective prediction model. First, a finite number of traffic predictors are constructed for the same number of representative traffic conditions using historical data. Using a minimum number of base predictors, it reduces the training and maintenance costs. The most effective predictor can be selected that best suits the current traffic situation in real time.

Many features can be used to identify a traffic situation called context. The feature content include location, time of day, weather, number of lanes, area type etc., .The context space is a multidimensional space with D dimensions, where D is the number of feature content. Since the context space can be very large, learning the most effective predictor in each individual context using reward estimates becomes extremely slow.

The proposed approach also obtains and proves both long term and short term performance guarantees for online algorithm. Not only it will converge over time to the optimal predictor for each possible traffic situation but also provides a bound for the speed of convergence of the algorithm to the optimal predictor. Various machine learning techniques have been applied to impact of an accident. . The majority of these techniques are predicting traffic in typical conditions and more recently in the presence of accidents. Online ensemble learning is represented by prediction with expert advice and weight update. This technique assigns weights to experts and makes a final prediction by combining the expert's predictions according to the weights. The weights which are updated may enable regret bounds to be derived.

When establishing the regret bound of the proposed algorithm, the adapted techniques from multi armed bandit problems. Since techniques used for ensemble learning problems, such as weighted majority type algorithms, lead to weak regret bounds for the considered contextual learning scenario. In these settings, the prediction action does not have an explicit impact on reward realization and the learner can observe the realized rewards of all predictors. However, this would lead to weak regret bounds. To get strong regret bounds of the reward estimates in some slots than in others and use the different ways to bound the learning loss in different slots.

II. LITERATURE SURVEY

B.Pan, U.Demiryurek, and C.Shahabi and C.Gupta [1] have proposed "Forecasting spatiotemporal impact of traffic incidents on road networks". They proposed to use two real world transportation datasets such as, incident data and traffic data. Incident on traffic include any non-recurring events on road networks, including accidents, weather hazard, road construction or work zone closures. By analyzing archived data, incidents can be classified based on their features. The incidents are analyzed by archived traffic data at the time and location of the incidents. This info in turn can help drivers to effectively avoid impacted areas in real-time. To be useful for such real-time navigation application, and unlike current approaches and model the impact as a quantitative time varying spatial spanking addition to utilizing incident features improve classification approach further by analyzing traffic density around the incident area and the initial behavior of the incident.

B.Stephen, F.David, and W.S.Travis [2] have proposed, "Naive Bayesian classifier for incident duration prediction". They have proposed to choose the appropriate response to an incident, it is important to predict the potential impact of an incident, including its duration, as accurately as possible. They have developed a probabilistic model based on a naïve Bayesian classifier to assist with prediction of incident duration. Two significant advantages of this model are its ability to readily accommodate incomplete information. Although incident duration prediction remains a difficult and complex problem, the Naive Bayesian classifier is demonstrated to provide a simpler, more flexible and more useful approach than regression which can provide without sacrificing accuracy.

M.Miller and C.Gupta [3] have proposed “Mining Traffic Incidents to forecast impact”. They have proposed to use sensor data from fixed highway traffic detectors, as well as data from highway patrol logs and local weather situations. In their research they have shown a practical system for predicting the cost and impact of highway incidents using classification models trained on sensor data and police reports. These models built on by understanding the spatial and temporal patterns of the expected state of traffic at different times of day and locations and past incidents. With high accuracy, this model can predict the false reports of incidents that are made to the highway patrol and classify the duration of the incident induced delays and the magnitude of the incident impact. Equipped with these predictions of traffic incident costs and relative impacts, highway operators and first responders will be able to more effectively respond to reports of highway incidents, ultimately improving driver’s welfare and reducing urban congestion.

B.Pan, U.Demiryurek, C.Shahabi [4] have proposed “Utilizing real-world transportation data for accurate traffic prediction”. They have proposed to use real-time high fidelity spatiotemporal data on transportations networks of major cities. This gold mine of data can be utilized to learn about traffic behavior of different times and locations, potentially resulting in major savings in time and fuel. The spatiotemporal behaviours of rush hours and events to perform a more accurate prediction of both short term and long term average speed on road segments, even in the presence of infrequent events.

W.Kim, S.Natarajan, and G.L.Chang [5] have proposed “Empirical Analysis and Modeling of Freeway incident duration”. They have proposed to present a methodology for developing a model to identify the variables influencing incident duration to estimate and predict incident duration. Classification Trees were employed for a preliminary analysis to understand the influence of the variables associated with an incident. Based on the findings from CT, the Rule-Based Tree Model has been constructed. The overall confidence for the estimated model was several remarkable findings regarding the association between the identified factors and incident duration. A discrete choice model was developed as a supplemental model. It is deduced that supplemented models along with better quality database are required to improve the prediction accuracy of the duration of a detected incident.

A.Fern and R.Givan [6] have proposed “Online ensemble learning an Empirical Study”. They have proposed ensemble methods such as boosting and bagging that has been shown to provide significant advantages in offline learning settings. Ensemble learning algorithms provide methods for invoking a base learning multiple time and combining the results into an ensemble hypothesis. Many empirical investigations have shown that ensemble learning methods often lead to significant improvements across a wide range of learning problems. The main goal of their research is to demonstrate that similar performance gains can be obtained in online learning settings by using time and space efficient online ensemble algorithms. Secondary goal include designing and evaluating appropriate online base learners for use in ensembles, and measuring the value of ensembles in reducing the space requirements needed to achieve a given classification accuracy. Online decision tree method that does not store a large number of training instances, but ID4 method improve performance in single trees and are critical to good performance in tree ensemble.

III. NAIVE BAYESIAN CLASSIFICATION

In machine learning, naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Baye's theorem with strong (naive) independence assumptions between the features. Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set.

It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 10 cm in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the colour, roundness, and diameter features.

A. Machine Learning Algorithm: Loading Data

Step 1: The first four buttons at the top of the pre-process section enable to load data into WEKA.

Step 2: Open file brings up a dialog box that allows browsing the data file on the local file system.

Step 3: Using the Open file button files can read in a variety of formats. WEKA’s ARFF format, CSV format, C4.5 format or Serialized instances format. ARFF files typically have a.arff extension. CSV files a .csv extension, C4.5 files a data and .names extension and Serialized instances objects a .bsi Extension.

Step 4: These list of formats can be extended by adding custom file converters to the weka.core.converters package.

B. Classification

Step 1: S’= S with weights assigned to be 1; m = n;

Step 2: Consider i = 1 to T

Step 3: Ci= I (S’)

Step 4: Ai=1/m*sum (weight(s));

Step 5: If Ai>1/2, set S’ to a bootstrap sample from S with weight 1 for every instance go to Step 3.

Step 6: βi= Ai/1-Ai

Step 7: For each x j∈S’. If Ci (xj) = yj then weight (xj) = weight (xj). Bi

Step 8: Normalise the weights of instances so that the total weight of S’ is m

Step 9: C(x) = arg max y ∈Y*sum(log(1/ βi))

C. Context Aware Traffic Prediction Algorithm

Step 1: Initialize p0= {a}, rf{a}=0,Ma0=0.

Step 2: For each traffic prediction request (time slot t) do

Step 3: Determine the level of .C

```
Step 4: Generate the predictions results for all predictors.
Step 5: Select the final prediction according to (3)
Step 6: The true traffic pattern is revealed.
Step 7: Update the sample mean reward.
Step 8: Mct=Mct+1
Step 9: if Mct>=A2pl then
Step 10: C is further partitioned.
Step 11: end if
Step 12: end for
```

D. Algorithm Description:

1. The Algorithm works in two parts predictor selection and reward estimates update.

2. When a traffic prediction request comes, the traffic speed vector xt along with the traffic context information is sent to the system.

3. The Algorithm first checks which active subspace C in the current partition Pt and the level l of this subspace.

4. The Algorithm activates all predictors and obtains their predictions f(xt),the given input is xt.

5. yt=f(xt) .This formula is used for the prediction selection.

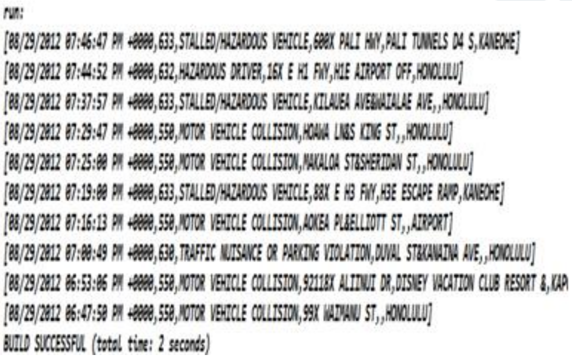
6. The second part of the algorithm is adaptive context space partitioning.

7. At the end of each slot t, the algorithm decides whether to further partition the current subspace C.

8. Mct>=A2pl, then C will be further partitioned, where l is the subspace level of C,A>0 and p>0.

IV. RESULTS AND DISCUSSION

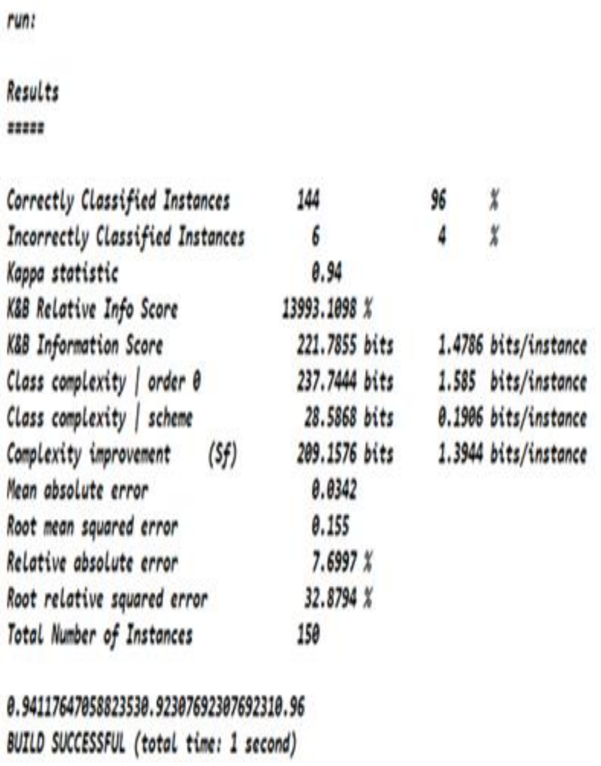
In the WEKA Explore first read the traffic data.



```
run:
[08/29/2012 07:46:47 PM +0000,633,STALLED/HAZARDOUS VEHICLE,600X PALI HWY,PALI TUNNELS D4 S,KANEOHE]
[08/29/2012 07:44:52 PM +0000,632,HAZARDOUS DRIVER,16X E H1 FWY,H1E AIRPORT OFF,HONOLULU]
[08/29/2012 07:37:57 PM +0000,633,STALLED/HAZARDOUS VEHICLE,KILAUEA AVE&WAIALAE AVE,,HONOLULU]
[08/29/2012 07:29:47 PM +0000,550,MOTOR VEHICLE COLLISION,HOAWA LN&S KING ST,,HONOLULU]
[08/29/2012 07:25:00 PM +0000,550,MOTOR VEHICLE COLLISION,MAKALOA ST&SHERIDAN ST,,HONOLULU]
[08/29/2012 07:19:00 PM +0000,633,STALLED/HAZARDOUS VEHICLE,88X E H3 FWY,H3E ESCAPE RAMP,KANEOHE]
[08/29/2012 07:16:13 PM +0000,550,MOTOR VEHICLE COLLISION,AOKEA PL&ELLIOTT ST,,AIRPORT]
[08/29/2012 07:00:49 PM +0000,630,TRAFFIC NUISANCE OR PARKING VIOLATION,DUVAL ST&KANAINA AVE,,HONOLULU]
[08/29/2012 06:53:06 PM +0000,550,MOTOR VEHICLE COLLISION,92118X ALIINUI DR,DISNEY VACATION CLUB RESORT &,KAP
[08/29/2012 06:47:50 PM +0000,550,MOTOR VEHICLE COLLISION,99X WAIMANU ST,,HONOLULU]
BUILD SUCCESSFUL (total time: 2 seconds)
```

Fig.1 Read the Traffic data

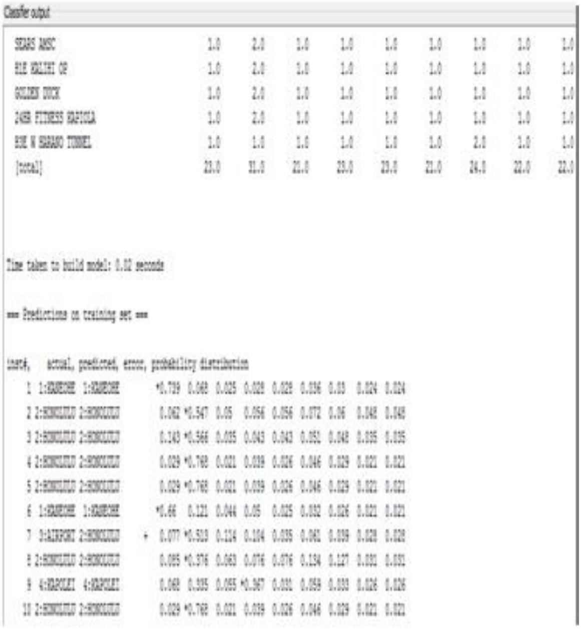
The machine learning algorithm called the Naive Bayesian classification is used to predict the area. Probability calculation is straight forward conditioning on the observed attributes and want to find the probability that I belongs to each category.
Applying Baye's theorem
P(I∈Ci/X1,X2,..Xn)=P(I∈Ci)P(X1,X2,,Xn/
I∈Ci)}/P(X1,X2,..,Xn)
Need to only calculate the numerator
I*∈ arg max P(I∈Ci)/P(Xj/ I∈Ci)



```
run:

Results
=====

Correctly Classified Instances         144               96      %
Incorrectly Classified Instances         6                4      %
Kappa statistic                          0.94
K&B Relative Info Score              13993.1098 %
K&B Information Score                  221.7855 bits      1.4786 bits/instance
Class complexity | order 0             237.7444 bits      1.585  bits/instance
Class complexity | scheme               28.5868 bits      0.1906 bits/instance
Complexity improvement     (Sf)        209.1576 bits      1.3944 bits/instance
Mean absolute error                      0.0342
Root mean squared error                  0.155
Relative absolute error                  7.6997 %
Root relative squared error             32.8794 %
Total Number of Instances              150

0.94117647058823530.92307692307692310.96
BUILD SUCCESSFUL (total time: 1 second)
```

Fig.1 - Machine Learning

Naïve Bayesian classification is applied to predict the typical traffic conditions and the impact of accidents. Figure.3shows the time prediction.



```
Classifier output

SEARS AMSC                  1.0   2.0   1.0   1.0   1.0   1.0   1.0   1.0   1.0
H1E KALIHI OF               1.0   2.0   1.0   1.0   1.0   1.0   1.0   1.0   1.0
GOLDEN DUCK                 1.0   2.0   1.0   1.0   1.0   1.0   1.0   1.0   1.0
24HR FITNESS KAPIOLA        1.0   2.0   1.0   1.0   1.0   1.0   1.0   1.0   1.0
H3E W HARANO TUNNEL         1.0   1.0   1.0   1.0   1.0   1.0   2.0   1.0   1.0
[total]                    23.0  31.0  21.0  23.0  23.0  21.0  24.0  22.0  22.0



Time taken to build model: 0.02 seconds


=== Predictions on training set ===


inst#,   actual, predicted, error, probability distribution
   1 1:KANEOHE  1:KANEOHE      *0.739  0.068  0.025  0.028  0.028  0.036  0.03   0.024  0.024
   2 2:HONOLULU 2:HONOLULU      0.062 *0.547  0.05   0.056  0.056  0.072  0.06   0.048  0.048
   3 2:HONOLULU 2:HONOLULU      0.143 *0.566  0.035  0.043  0.043  0.051  0.048  0.035  0.035
   4 2:HONOLULU 2:HONOLULU      0.029 *0.768  0.021  0.039  0.026  0.046  0.029  0.021  0.021
   5 2:HONOLULU 2:HONOLULU      0.029 *0.768  0.021  0.039  0.026  0.046  0.029  0.021  0.021
   6 1:KANEOHE  1:KANEOHE      *0.66   0.121  0.044  0.05   0.025  0.032  0.026  0.021  0.021
   7 3:AIRPORT  2:HONOLULU   +  0.077 *0.513  0.124  0.104  0.035  0.041  0.039  0.028  0.028
   8 2:HONOLULU 2:HONOLULU      0.085 *0.376  0.063  0.076  0.076  0.134  0.127  0.031  0.031
   9 4:KAPOLEI  4:KAPOLEI       0.068  0.335  0.055 *0.367  0.031  0.059  0.033  0.026  0.026
  10 2:HONOLULU 2:HONOLULU      0.029 *0.768  0.021  0.039  0.026  0.046  0.029  0.021  0.021
```

Fig.3- Time Prediction

The Figure 4 shows the visualization of traffic data comprises of traffic speed, time, date, location and area. In the visualization stage is used to traditional map as an interactive medium to perform the simulation. Simulation is done by placing the sensor together with the position of latitude and longitude respectively. Each sensor is represented in the form of point base in geographic position in map.

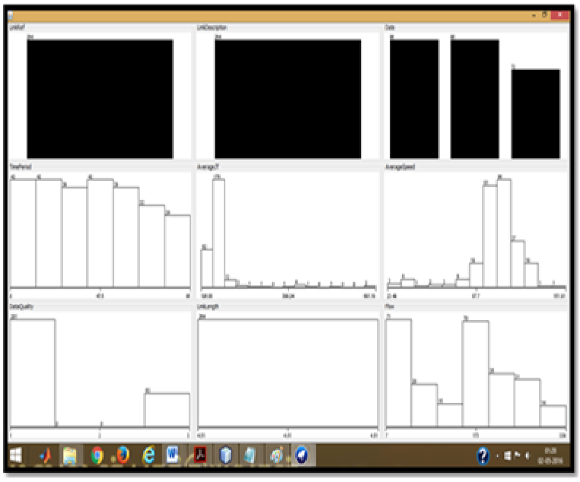


Fig.4: Visualize the Traffic Data

Figure 5 shows the traffic flow prediction is used to predict the traffic with yearly details. This data series provides flow information, average traffic speed, and average trip time per a period of fifteen minutes. Travel time and average speed is calculated using combination of sources. Travel time is derived from a real vehicle observation and calculated by using adjacent time periods.

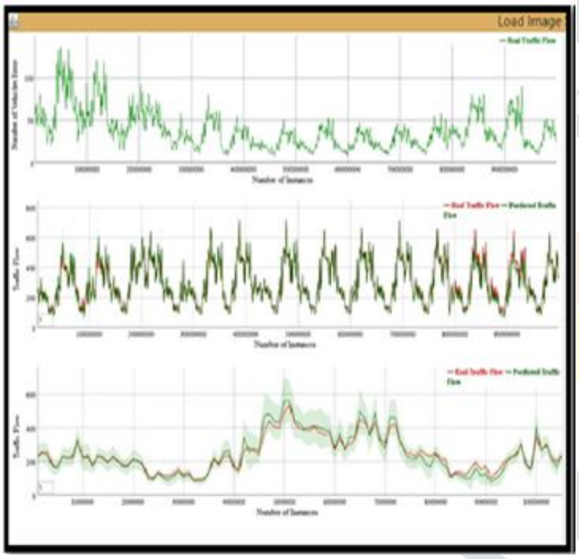


Fig. 5: Traffic Flow Prediction

V. CONCLUSION

The proposed framework for online traffic prediction discovers online, the contextual specialization of predictors to create a strong hybrid predictor from several weak predictors. This framework matches the real-time traffic situation to the most effective predictor constructed using the historical data, thereby self-adapting to the dynamically changing traffic situations. Systematically it has been proved both short term and long term performance guarantees for this algorithm, which provides not only it will converge over time for each possible traffic situation but also provides a bound for the speed of convergence of the algorithm. This project work is using real-world dataset and verified efficiency. The future work of this paper is to extend the current framework to distributed scenarios where the traffic data is gathered by distributed entities which are required to achieve a global traffic prediction goal.

REFERENCES

[1] B.Pan, U.Demiryurek, C.Shahabi and C.Gupta, "Forecasting Spatiotemporal Impact of Traffic Incidents on Road Networks" , In proceedings of the IEEE International Conference on Data Mining, pp. 587-596,2013.

[2] B.Stephen, F.David, and W.S.Travis, "Naive Bayesian Classifier for Incident Duration Prediction", In Proceedings of the IEEE Transaction on Intelligent Transportation Systems, vol. 13, no. 3, pp.1454-1461, Sep 2012.

[3] M.Miller and C.Gupta, "Mining Traffic Incidents to Forecast Impact", In proceedings of the AcmSigkdd Workshop on Urban Computing, pp. 33-40, 2012.

[4] B.Pan, U.Demiryurek, and C.Shahabi, "Utilizing Real-World Transportations Data for Accurate Traffic Prediction", In proceedings of the IEEE 12th International Conference on Data Mining, pp. 595-604, 2012.

[5] W.Kim, S.Natarajan, and G.L.Chang, "Empirical Analysis and Modelling of Freeway Incident Duration", In Proceedings of the Intelligent Transportation Systems, pp.453-457, Oct2008.

[6] A.Fern and R.Givan, "Online Ensemble Learning: An empirical study", In Proceedings of the IEEE 17th International Conference on Machine Learning., vol.53, no.1-2, pp.71-109, 20035.

Dr. N. Sadhasivam received a Bachelor of Engineering in Computer Science in the year 2005 from Anna University, Chennai, Tamil Nadu, India. He completed his Master of Engineering in Computer Science in the year 2009 from Anna University, Coimbatore, Tamil Nadu, India. He received Ph.D in the year 2017 from Anna University, Chennai, Tamil Nadu, India He has presented and published many papers in national, international conferences and journals. Currently he is a Professor in the Department of Computer Science and Engineering at Sri Indu College of Engineering and Technology, Hyderabad, Telangana. His areas of research interest are cloud computing, big data and optimization techniques

Dr. M. Pandi received a B.Tech (IT) from Anna University, Chennai, Tamil Nadu, India. He completed his Master of waqre Engineering in nna University, Tamil Nadu, India. He received 7 from Anna University, Chennai, Tamil Nadu, sistant Professor in the Department of Computer ing at Dr. Mahalingam College of nology, Pollachi, Tamilnadu. His areas of research interest are Datamining,, big data and optimization techniques