

TRAFFIC DATA ANALYSIS USING DECISION TREE AND NAÏVE BAYES CLASSIFIER

Poonam Rani¹, Navpreet Rupal²

¹Student, Dept of Computer Science Engineering, GIMET Amritsar, Punjab, India ²Assistant Professor, Dept of Computer Science Engineering, GIMET Amritsar, Punjab ,India ***

Abstract-Normally, High Traffics and accidents are the leading causes of death worldwide. The Classification and Characterization for the algorithms are essential for these traffic . Data mining techniques have been used in real time applications due to its artificial intelligence nature. It is highly used in domain as it helps in better predictions and supports in decision making. This paper analyses the data using data mining technique called WEKA. The modified J48 classifier is used to increase the accuracy rate. Some data mining classification algorithms (like J48, naïve-Bayes, decision tree and random-forest) are implemented on the dataset .

Key Words: Data-Mining, Classification Algorithms, Traffic prediction, Net Beans, WEKA.

1. INTRODUCTION

Traffic prediction is an important sector in terms of datamining application [1, 2]. A wide variety of data sets differing in volume, variety and velocity are generated. The Traffics like low, High and Medium are the leading causes of death worldwide. Traffic on road networks is nothing but slower speeds, increased trip time and increased queuing of the vehicles. When the number of vehicles exceeds the capacity of the road, traffic occurs [3]. Nowadays, many countries suffer from the traffic problems that affect the transportation system in cities and cause serious dilemma [4,5].

It provides the step by step guide to build a classifier model using on data and then the model is tested using the test data and helps in making predictions. With the advancements of computing facility provided by computer science technology, it is now possible to predict traffic more accurately [6].

1.1 WEKA Tool

WEKA is a powerful tool as it contains both supervised and unsupervised learning techniques. WEKA is an efficient approach and outperforms other data mining approaches [7]. We use WEKA because it helps us to evaluate and compare data mining techniques (like Classification, Clustering, and Regression etc.) conveniently on real data [8]. WEKA is a well-known machine learning software written in Java, developed at Waikato University in New Zealand. The WEKA works and contains a collection of visualization tools and algorithms i.e.]. random forest, naïve Bayes, decision tree, j48 classifier for solving real-world data mining problems and helps in traffic prediction.

1.2 Net Beans

The Net Beans project consists of an open source IDE and an application platform that help developers to rapidly create web, enterprise, and mobile applications [9]. It offers a full-fledged IDE that runs on multiple platforms. Many of the users and students were equally happy and comfortable with the net beans IDE [10,11]. The main aim is to evaluate the performance of various classifiers and improvement is made by fusion of algorithms so as to make better prediction.

2. RELATED WORK

Mr. Chintan Shah et.al [1], explains discussion of various classification algorithms based on certain parameters like time taken to build the model, accurately and inaccurately classified instances etc.

Dr. H.S. Sheshadri et.al[2], "Traffic employing and classification technique", IEEE, DOI: 10.1109/ICITCS.2015.7292973.

Qiu jing et.al [3],Research and development of network traffic prediction model computer enginerring and design,3rd ed,vol.33 pp.865-869,2012.

Feng Huali et.al[4],Prediction of traffic based on Elman neural network,Guilin Chinese control and decision conference,2009,pp.1248-1252.

- A. Quinn et.al[5]. Traffic flow monitoring in crowded cities. In AAAI Spring Symposium on artificial intelligence for development,2010.
- B. C Park et.al [6], Prediction of Network Traffic by Using Dynamic Billinear Recurrent Network(IEEE,2009)978-0-7695-3736-8.

Hornik, K.et.al[7],Open-Source Machine Learning: R Meets Weka, Journal of Computational Statistics - Proceedings of DSC 2007, Volume 24 Issue 2, May 2009 [doi>10.1007/s00180-008-0119-7]. Hunyadi, D., Rapid Miner E-Commerce, Proceedings of the 12th WSEAS International Conference on Automatic Control, Modelling & Simulation, 2010.

IRIET

Han, J.et.al[8], Data Mining Concepts and Techniques . San Fransico,CA:Morgan Kaufmann Publishers, 2011.

Kollinget.al[8], Journal of Computer Science Education, Vol 13,no.4 December 2003.

Stoleret.al[10], "Java Programming Environments", Master's Thesis, Rice University, April 2002.

UCI Machine Learning Repository[11], Available at: http:// http://archive.ics.uci.edu/ml/, (Accessed 22 April 2011).

3. DATA COLLECTION AND PROPOSED METHODOLOGY

The dataset for analysis is collected from UCI (repository dataset)[11]. The UC Irvine Machine Learning Repository is hosted by the Center for Machine Learning and Intelligent Systems at UC Irvine. They maintain data as a service to the machine learning community. Currently maintain 429 dataset which helps in machine learning repository community. Over the last two decades there has been an explosive growth in online data storage of various methods. These large datasets have motivated the rapid development of data mining methods. In this paper, an online repository of large and difficult data sets are being gathered. This repository includes high-dimensional data sets as well as data sets of different data types (time series, spatial data, and so forth). The primary role of the repository is to enable researchers in data mining (including computer scientists, statisticians, engineers, and mathematicians) to scale existing and future data analysis algorithms to very large data sets. This repository will play a substantial role in the gap between research-oriented algorithm development in the laboratory and the real-world practicalities and challenges of very large data sets. Data are available for Annual Average Daily Flows (AADF s) for count points on major roads(motorways and A roads)and minor roads(B,C and unclassified roads). Traffic figures for major roads are also available (in vehicle kilometers and vehicle miles). AADF figures are produced for each junction to junction link on the major road network for every year. Only a sample of points on the minor road network is counted each year and these counts are used to produce estimates of traffic growth on minor roads.

4. RESEARCH WORK WITH DATASET ANALYSIS

A. NAÏVE BAYES: Naïve Bayes classifier is one of the efficient and highly scalable inductive learning algorithm which is trained in a supervised learning strategy [12]. It assumes all the attributes are conditionally independent for evaluating class conditional probability [13].

B. J48 CLASSIFIER: It is a tree based classifier which is used in classification of algorithms [14].It builds decision tree and then applied this tree to all instances of the dataset. It uses the concept of Information gain to choose the best split [15].Root node is selected which has maximum information gain of the attribute.

<pre>bit for big mode that the long is the form took times way bit for big mode of the long of the lon</pre>	Sciencipol Rectingen - (Bibliotegen - (Bicless
Image: Section in the section in th	nchan (Bridan (Bridan (Bridan
<pre>at https://achapariliandaparili</pre>	Mudage (Bittlage (BLGs.)
import import Project Properties. 2 To change this locase lander, choose License Headen in Project Properties. 4 "and open the template in the addre. 5 mport pre-stallagging.Locgt. 6 import pre-stallagging.Locgt. 12 "** 13 "#" 14 "and the pre-stallagging.Locgt. 15 patienter to pre-stallagging.Locgt.	
□ *To change this leaves hander, choose Licosus Henders in Project Properties. 3 *To change this leaves hander, choose Tools) Templates * "and open the transplate in the editor. 7 package workkinset; 0 package workkinset; 10 package is until lagging Lenger; 11 *** 12 *** 13 *** 14 *** 15 *** 16 *** 17 **** 18 *** 19 *** 10 *** 11 **** 12 *** 14 **** 15 ***** 16 pablic chaos Main Class exteends jonce, swing. Frame {	
7 peckage welchinet. 8 ⊖ more inproventilogning Level. 1 → more inproventilogning Legger. 11 ↓ 2 → ** 13 ↓ ** 14 ↓ ** 14 genthere lap ce. 15 ↓ ** 16 perfect the Main_Class extends jong, wing, Frame {	
a U mengri mantalagang Level b Import januatalagang Leger: 11 12 13 14 14 14 16 16 17 16 17 17 17 18 19 19 19 19 19 19 19 19 19 19	
13 * 14 * 15 while class strends jores; swing, Frame { 10 10	
16 public class Main_Class extends javax.swing.JFrame { 17	
18 - /**	
15 ¹ Crates new form Mann_Class 21 ⊂ public Main Class 21 ⊂ public Main Class 21 ⊂ public Main Class	
zzi min, outpolicius (, pa-Mon751 (m)	
Derrently classified horaces 104 N.1227 V Descrively classified horaces 10 A.071 V Appendix and the second of the secon	
Dapit [] Institutions NEGUTEST	(us) syning II (0, 184
	1457

Fig. 1J48 classifier output

accurately which results in higher accuracy rate of 99% and 17 are incorrectly classified.

C. RANDOM FOREST: This algorithm builds a randomised decision tree in each iteration of the algorithm and produces excellent predictors. Every sub tree gives a classification and provides the tree votes for that class [16][17]. Then the classification is done which is having the most votes. The following are the basic steps in the algorithm:

i) Suppose the number of samples in the training set is X, then these samples will be the training set for building the tree.

ii)Suppose there are Y input variables, a number y<<Y is specified as at every node, m variables are selected randomly out of Y and then the best split on these y is used to split the

e-ISSN: 2395-0056 p-ISSN: 2395-0072

node. The value of y is unchanged during the forest build [18].

iii) Every sub tree is built to the largest extent possible.Out of 1187instances, 1170are classified accurately which results in higher accuracy rate of 97% and 17 are incorrectly classified.

Base Date, Weing Bard, 48,978,214,948,312,914,92 Image State S	The particle - Disconcise - Reserve	djaa X 🕽 forni jaa X 🛔	NexiFormijee X 👌 Nex	ojaa X 🖞 🕮 Lijaa X 🗍 s	airform.java 🗶 🖞 GLEjava 👋	jeelpisteljee × 🛔	NGENTEST, java 🗴 🗟 Mar	,Cesjea × §SM,Cb	ajaa × 👌 S_Cjaa.	
Image: The information for the state of	Sauraz Design History 🔯 🗟 - 🎚 - 🕄 🖓 🧞	84/268 99	0 U (d 🖬							
16 polic class Main_Class extends joers: wing_Frame { 17 18 ** 19 *Contex sets fram Main_Class 20 y jii 21 Diffici Main_Class (jii) 22 Infici Main_Class (jii) 23 Infici Main_Class (jii) 24 Infici Main_Class (jii) 25 projici Main_Class (jii) 26 projici Main_Class (jii) 27 monetal; function terms 28 monetal; function terms 29 monetal; function terms 20 monetal; function terms 21 Diffici Main 22 monetal; function terms 23 Diffici Main 24 Diffici Main 25 Diffici Main 26 Diffici Main 27 Diffici Main 28 Diffici Main 29 Diffici Main 20 Diffici Main 20 Diffici Main 21 Diffici Main 22 Diffici Main 23 Diffici Main 24 Diffici Main 25 Diffici Main 26 Diffici Main 27 Diffici Main 28<	1 □ □ 2 "To change this isome 3 "To change this isome 4 "and open the template 5 □ 6 □ 7 package withinket, 8 □ 11 □ 12 □ 13 □ * □ 14 □ 15 □	hender, choose Lic te file, choose Tool in the editor.	ense Headers in Pri 1 Templates	oject Properties.						
Image: Second	16 public class Main_Class	extends javax.swin	ų,JFrame {							
and Antonio a	19 * Creates new form) 20 */ 21 B public Main_Class() 22 initComponents():	fain_Class								
mended standing braness LHK H.S.B.F. & parended fractifications L H.S.F. & parended standard number L.S.B. H.S.F. & max standard number L.S.B. & H.S.F. & haber set fractions L.S.B. & H.S.F. & haber set fractions L.S.B. & H.S.F. & haber set fractions L.S.B. & H.S.F. &	Copie - MERIATET (M)									X
	Crrwetly classified Instances Incorrectly Classified Instances Rope statistic Name shallote error Root mean spared error Root molative spared error Total Number of Instances	1186 1 0.5980 0.0098 0.0460 1.6985 % 5.2507 % 1187	99.5128 % 0.6942 %							

Fig. 2 Random forest classifier output.

D. SUPPORT VECTOR MACHINE: In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. The support vector clustering algorithm created by Hava Siegelmann and Vladimir Vapnik applies the statistics of support vector machines algorithm, to categorize unlabeled data, and is one of the most widely used clustering algorithms in applications. Out of 1187instances, 1170are classified accurately which results in higher accuracy rate of 98.56% and 17 are incorrectly classified.

<pre>Bit 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0</pre>	dit View Navigate Source Refactor Run Debu	file Team Tools Window Help	Q. Search (Od+1)	
a her ber ' generation i Beneration i Beneratio i Beneration i Beneration i Beneration i Beneration i Benerat	BB®® C (deutorip ∨ 9*8)	-B-0-		
<pre>min Now min () = 10 - 0 - 0 - 0 - 0 - 0 - 0 - 0 - 0 - 0</pre>	.ne Start Page 🗴 🚮 SiliCaain,jane 🛪 🗟 Demociou	× 🖟 formijena × 🖓 Hendifernetijena × 🗟 Mannijena × 🗟 Olijena ×	ijen × 🖟 WERKITEST,jan × 🖟 Mon_Class.jan × 🗟 SMLClass.jan × 🗟 S_C.jan	0
<pre></pre>	urte Desge Hotory 12 11 - 11 - 14 15 47	1 <u>4</u> 8 8 8 8 8 8 8 8 8		
<pre>2 of the charge that it comes it cannot the comes it report investmes. ** of charge that it comes it cannot the other. ** of equal to it up the field one of the other. ** of equal to it up the other equal to it up the other. ** of equal to it up the other equal to it up the other. ** of equal to it up the other equa to it up the other equa to it up the other equa</pre>	1 87*	the state that we have a		
<pre>3</pre>	2 * To change this license.	der, choose License Headers in Project Properties.		
<pre>4 ************************************</pre>	3 To change this templat	e, choose Tools Templates		
<pre>produge weblantst; import jou and logging Level; import jou and logging Level; import</pre>	4 * and open the template	ne editor.		
<pre>pockage verkinet; import journal logging Longe; import journal logging Longe; import</pre>	5 - */			
peckage welchnet; import journal logging Level; import journal logging Level; if and bar logging Level; if additional logging Level; <t< td=""><td>6</td><td></td><td></td><td></td></t<>	6			
import jacand logging Level: import jacand logging Level: impor	7 package wekkatest;			
import journal loging Level: import journal	8			
import jurn will keyne Lagger, if in the set of the s	🔒 🖯 import java.util.logging.L	<u>k</u>		
11 <	s import iava.util.logging L	er.		
Image: Second	11			
image: state in the sector of the sector	12 E /**			
**@utubro top c: **@utubro top c:	13 *			
and the second space is a second processing. Frame (14 * @author hone			
proble class Main Class extends jowns.wing.Finme { * Clasts new from Main Class * () * Clasts new from Main Class * () * () * proble Main Class * ()	15 . */			
A Construction of the set of	16 mblie elses Main Class	ande interest miner Terma (
Image: Second	10 public class mani_class	ana lavarawitikat tame /		
Image: Section provides and the section of the se	10 2 /##			
Image: Second	10 U /			
Image: Second	15 Cleates new total of			
Control Number (Lasson) (ar Wear The Control Number (Lasson) (ar Wear The Control Number (Note) (Control Number (Contro Number (Control Number (20			
22 I mail Anomphalis L de martine Consenti plandin formanne 196 (F. 7001 % Decembri plandin formanne 196 (F. 7001 % Decembri plandin formanne 196 (F. 7001 % Decembri plandin formanne 197 (F. 7001 % Decembri plandin formanne 197 (F. 7001 % Decembri plandin formanne 197 (F. 7001 %) Decembri plandin formanne 197 (F. 700	21 D public Main_Class() {			
Consetly Castified Intensors 196 07.2015 6 Decremently Castified Intensors 196 07.2015 6 Rege statistic Brief Rege statistic Bri	Apart - MORATEST (sue)			
Consult Gandida Ennecon 196 07.001 % Exemutid Gandida Ennecon 196 07.001 % Equational Section 2010 1975 % Equation Section 2010 % Refer Antiler Section				
Recentral Gaudidation Statesses 14 L.2.309 % Regeneration Constrained States series 4.121 Bond which series 4.121 Recent Market States States 4.121 Recent Market States States 4.121 Recent Market States 5.121 Recent Mark	Correctly Classified Instances	1041 87.7001 %		
All and an and a set of the set o	Incorrectly Classified Instances	146 12.2999 8		
Inde an speed ence 1.5.100 Radelle de hollone ence 2.5.100 5 Radelle de hollone ence 3.5.100 5 fond Namier of Lansance 1107 Radel Quadrature Manager (Manager Manager Manager Manager Manager Mana	Nean absolute error	0.123		
Nadari Berlande entre 25.559 5 Bon Statiste egne entre 75.051 5 Doub Basier of Instance LLD Aggel Quantumer (LLD) Aggel Quantumer (LLD)	Root mean squared error	0.3507		
Anno matatare againe tenta multa ' boda Neutra Julia Sale (Neutras Julia O Laudonataranch III A A A A A A A A A A A A A A A A A	Relative absolute error	24.5550 %		
	Root relative squared error Total Number of Instances	78.2525 N		
Agel ('untrame	the second of the second			
🔿 Inalwerts auch 🛛 🛛 🔒 🔒 📮 📮 📮 😭 🚓 🐧 📰 🕼 着 🤶 👔 🛄 🖉 🛆 🖓 🗰 🕅	Subput @Notifications		WERNTEST (san)	18:8
	O June here to coarch	1 10 🔒 🖬 🔺 🐂 🐺 🚓 🌒 🗾 📾 📻 🔅		

Fig. 3 Support vector machine classifier output

This section compares the classification accuracy of the supervised algorithms namely Naïve Bayes, J48, Random Forest and support vector machine. All simulations were performed using WEKA machine learning environment which consists of collection of popular machine learning techniques that can be used for practical data mining.

WEXXTEST - NetBeans IDE &0	- 0 X
File Edit View Navigate Source Refactor Run Debug Profile Team Tools Window Help	.th (Ctr(+))
전 월 월 월 70 (* induktionly) · · · · · · · · · · · · · · · · · · ·	
🖉ad Startege × 👸 Staatupe × 👸 Staatupe × 🗒 Stortiges × 👸 StarTentiges × 👸 Starteges × 🗒 Starteges × 🗒 Starteges × 🗒 Starteges × 🗑 Starteges × 🖗 Starteges × 🖏 Starteges × S	S_Cjea. + + = #
월 Saura Desja Hatay (2) 등 - 10, 10, 20 월 6 년 20 월 9 년 20 월	± 🕯
	^ • §
2 * To change this license header, choose License Headers in Project Properties.	rter
3 * To change this template file, choose Tools Templates	
6 4 * and one the template in the editor.	
/ puckage weikanesi,	
a 🛛 import Java util logging Level;	
3 umport java.unil.logging.Logger,	_
12 🗄 /**	
13 *	-
14 ° @author hp pc	
15 L */	
16 public class Main_Class extends javax.swing.JFrame {	
17	-
18 🖯 /**	
19. * Creates new form Main Class	
20 4/	-
21 🗄 public Main Class() {	
22 initComponents():	
(dapar-1666/1517 par)	×e
	^
P Correctly Classified Statements 1000 04.3.2 % Tocompute Unsuffied Statements 100 15.02 k	
Tage statistic 0.4900	
Mean absolute error 0.1704	
Root mean equared error 0.3464	
bearstre adducte error 3-0-011 T Root making smared error 61.0789 %	
Total Number of Instances 1187	
2 The function of the second sec	¥
	1500
📲 🕐 Type here to search 🛛 🔮 🖪 🔮 😌 😅 🕷 😭 🔮 🚺 📲 😫 🚱 👘 🚱	NG 29/04/2018

Fig. 4 Naïve Bayes classifier output

5. CONCLUSION AND FUTURE SCOPE

This work evaluates the Traffic n using different machine learning algorithms by WEKA Tool. Compare the results in terms of time taken to build the model and its accuracy. WEKA is an efficient approach and outperforms other data mining approaches. This work shows the Random Forest is best classifier for Traffic of WEKA tool because it runs efficiently on large datasets. In future we will use different classifiers on different datasets and evaluating the performance of each classifier.



Fig. 5 Graphical representation of classifiers

ACKNOWLEDGEMENT

I would like to thank the partial support funded by the computer science department of the Global institute of management and emerging technologies at Sohian Khurd, Amritsar PTU University.

REFERENCES

[1] Mr. Chintan Shah, Dr. Anjali G. Jivani, "Comparison of Data Mining Classification Algorithms for Traffic Prediction", IEEE,DOI:10.1109/ICCCNT.2013.6726477.

[2] Dr. H.S. Sheshadri S. R. Bhagya Shree2, and Dr. Muralikrishnauses, "Traffic employing and classification technique", IEEE, DOI: 10.1109/ICITCS.2015.7292973.

[3]Qiu jing,Xia jing-bo,Wiu ji-xiang.Research and development of network traffic prediction model computer enginerring and design,3rd ed,vol.33 pp.865-869,2012.

[4] Feng Huali,Liu Yuan, Z Mao hu.Prediction of internet traffic based on Elman neural network,Guilin Chinese control and decision conference,2009,pp.1248-1252.

[5] A. Quinn and R. Nakibuule. Traffic flow monitoring in crowded cities.In AAAI Spring Symposium on artificial intelligence for development,2010.

[6] C Park, D-M Woo, Prediction of Network Traffic by Using Dynamic Billinear Recurrent Network(IEEE,2009)978-0-7695-3736-8.

[7] Hornik, K., Buchta, C., Zeileis, A., Open-Source Machine Learning: R Meets Weka, Journal of Computational Statistics -Proceedings of DSC 2007, Volume 24 Issue 2, May 2009 [doi>10.1007/s00180-008-0119-7]. Hunyadi, D., Rapid Miner E-Commerce, Proceedings of the 12th WSEAS International Conference on Automatic Control, Modelling & Simulation, 2010.

[8]Meenakshi sharma and Dr. Himanshu Aggarwal," Development and Implementation challenges in clinicaldecision-support system",.

[9] Han, J.,Kamber,M.,Jian P., Data Mining Concepts and Techniques. San Fransico,CA:Morgan Kaufmann Publishers, 2011.

[10] Kolling, Michael and Bruce Quing ,Andrew Patterson, John Rosenberg. "The BlueJ System and Its Pedalogy," Journal of Computer Science Education, Vol 13,no.4 December 2003.

[11] Stoler, Brian. "A Framework for Building Pedagogic Java Programming Environments", Master's Thesis,Rice University,April 2002.

[12] UCI Machine Learning Repository, Available at: http:// http://archive.ics.uci.edu/ml/, (Accessed 22 April 2011).

[13] M.Venkat Dass, Mohammed Abdul Rasheed, Mohammed Mahmood Ali," Classification of Data Mining technique", IEEE, 2014, DOI: 10.1109/CIEC.2014.6959151.

[14] Meenakshi sharma and Dr. Himanshu Aggarwal. "Methodologies of legacy clinical decision support system -A review", *Journal of Telecommunication, Electronic* and Computer Engineering(JTEC Journal).



[15] Abeer Y. Al-Hyari, Ahmad M. Al-Taee, Majid A. Al-Taee," Management of Traffic Failure", IEEE, DOI: 10.1109/AEECT.2013.6716440.

[16] Kanika Chuchra and Amit Chhabra," Evaluating the Performance of Tree based Classifiers", IEEE, DOI: 10.1109/NGCT.2015.7375168.

[17] SP.Chokkalingam and K.Komathy," Comparison of Different Classifier in WEKA", IEEE, DOI: 10.1109/ICHCI-IEEE.2013.6887821.

[18] Velu C. M. and Kashwan K. R., "MULTI-LEVEL COUNTERPROPAGATIONTRAFFICNETWORKFORMEDIUMCLASSIFICATION",IEEE,10.1109/ICSIPR.2013.6497986.

[19] Rachana Mishra and Ramjeevan Singh," An efficient approach for supervised learning algorithms using Different Data Mining Tools for spam categorization", IEEE, DOI: 10.1109/CSNT.2014.100.

[20] Meenakshi Sharma, Dr. Himanshu Aggarwal," Evaluation factors for testing and validation of Clinical Reporting System", *International Journal of Computer Science and Engineering*, vol.6, issue-2, e- ISSN:2347-2693 on 2018.