2014

ISSN : 0974 - 7435 Volume 10 Issue 20

BioTechnology
An Indian Journal

FULL PAPER

BTAIJ, 10(20), 2014 [12660-12666]

A comparative study on car evaluation forecast based on data mining

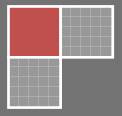
Yuqing Yuan[1]*, Lijun Ling[2], Xiangling Kuang[1], Qinggang Zuo[3]
[1]School of Economics and Management, Hubei University of Automotive Technology, Shi Yan, 442002, (CHINA)
[2]Institute of Science and Technology, Hubei University of Automotive Technology, Shi Yan, 442002, (CHINA)
[3]Department of Science and Technology, Shiyan Central Sub-branch of the People's Bank of China Shiya, People's Republic of China, Shi Yan, 442002, (CHINA)
E-mail: yyq115805@163.com

ABSTRACT

Logistic regression (LR), artificial neural network (ANN), decision trees (DT), and support vector machine (SVM) were used in forecasting car acceptability, and their accuracy, sensitivity and specificity were compared. The results show that support vector machine (SVM) model can well predict the car acceptability evaluation with 99.62 percent of accuracy rate and 100 percent of sensitivity and specificity. The factor of security has the most influence on car acceptability evaluation. Comparative study method is suitable for the evaluation of car acceptability forecasting, can also be extended to all other areas.

KEYWORDS

Data mining; Car acceptability; Logistic regression; Support vector machine; Artificial neural networks; Decision tree.



INTRODUCTION

The automotive industry is the pillar industry of the national economy, and millions of people are closely related to it. For most people, buying a car is to buy a house outside in addition to a maximum consumption, can better reflect consumer demand and the real market behaviour. When consumers consider when buying a car, there are many factors that could restrict their choice, such as price, car performance, the car's comfort and safety, etc. These factors form the right car evaluation. The fierce market competition forced auto companies in a very short development cycle of continuous improvement and innovation in product design to meet the needs of highly diverse target market[4-6].

Therefore, a reasonable evaluation method is equally important for car consumers and producers. It can not only reduce the burden on dealers, but also increase sales. In addition, it plays a strategic role, can improve customer service levels in a highly competitive market environment[1-3]

Byun proposed extension of AHP select vehicle purchase patterns. Lai and some have proposed a method to help designers improve the quality feel of automotive products. Alnoukari and Alhussan proposed using data mining techniques to predict the future of the automotive market demand. Chen and some people make use of artificial intelligence methods; the practical problems faced by the auto parts industry in product performance objectively classify[19-21].

Data mining techniques from the hidden data found useful information to facilitate smart summary and future decisions, so it has great visibility in areas of research and commercial areas, in various applications, including manufacturing, marketing, finance, health care and other fields have outstanding performance, is a data conversion indispensable tool for information. Advantages of data mining technology is its ability to handle massive traffic data in order to adapt to market changes, can provide decision makers with a powerful tool. It is widely used in business management, government administration, scientific and engineering data management of large amounts of data processing. With the explosive growth of data, data mining techniques and tools have become an urgent need, it will be processed data intelligently and automatically converted into useful information and knowledge. It improves their competitive advantage and increases the company's revenue, but also enables enterprises to provide better service to retain customers. In the past few years, the classic data mining technology such as logistic regression (LR), artificial neural networks (ANN), decision trees (DT) and Support Vector Machine (SVM) have been successfully applied in many fields to solve practical problems of production, sales and research in emerging.[18] However, no comparative study to assess a product's acceptability for prediction. Accurate assessment of the development of product acceptability has become an important research topic.

The purpose of this study is to provide a method for comparative evaluation of the study to assess the predictive evaluation of the automotive, and then extended to other fields.

By modelling respectively using logistic regression (LR), artificial neural network (ANN), decision trees (DT), support vector machine (SVM) these technology forecasting automotive, and their accuracy, sensitivity and specificity were compared. The results show that support vector machine (SVM) technology has made the best assessment and prediction. Largely due to the performance of SVM depends on the choice of kernel function, the paper finally linear kernel, polynomial kernel, radial basis (RBF) kernel and the S-shaped core kernel comparative study, polynomial kernel has achieved the best results.

THE BASIC CONCEPT AND RESEARCH FRAMEWORK

Logistic regression

Logistic regression is a popular non-linear statistical model, and is widely used in many fields. Compared with the multiple regression models, Logistic regression model can simulate two or more dependent variables. For binary variables can be defined as an event of interest coding and coding are not interested in the event of 0[7-11] A logistic regression model can be written as:

$$\log(\frac{p(Y=1)}{1-p(Y=1)}) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Equation can be modified as follows:

$$p(Y=1) = \frac{1}{1+e^{-z}}$$

Where

$$z = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

The logistic regression model enables us to calculate the probability of event Y=1 occurring for each case. The predictors, X_k can be a mixture of continuous and categorical variables.

Decision tree

A decision tree is a predictive model; It represents a mapping between object attributes and object values. Each node in the tree represents an object, and a possible attribute values for each forked paths are represented, each leaf node corresponds to the value of the path from the root node to the leaf node experienced represented object. Tree only a single output, if it want to have a plurality of output, it can establish an independent decision tree to handle different output. Data Mining Decision Tree is a technique often used, can be used to analyze the data, also can be used to make predictions[12-13]. Common tree algorithm CHAID (Chi-squared Automatic cross-checking), CART (Classification and Regression Trees) and C5.0.CART algorithm uses the Gini as a standard decision tree split,C5.0 entropy as the split criteria, CHAID using chi-square test as segmentation criteria. Through these algorithm will generating tree diagram,, splitting rule and important information can be reflected from the Figure out.

Artificial neural networks

Artificial neural network is an application similar to the structure of the brain-Fi mathematical model of information processing, constituted by a large number of interconnected nodes (or neurons) between. Each node represents a specific output function, called activation function. Each connection between two nodes represents a weighted value of the connection for the signal, referred to the weight, it is equivalent to the memory of artificial neural networks. The output of the network according to the different network connections, weights and excitation functions and different. Multilayer Perception (MLP) is the most widely used neural network model in the data analysis, it will enter multiple data sets are mapped to a single output data set. Artificial neural network can identify and study the pattern of association between input data set and the corresponding target value[14-15]. However, artificial neural networks (ANNs) for its "black box" approach and interpretation difficulties suffer criticism. Nevertheless, compared with other comparative classification techniques, artificial neural network to provide alternative models. After training, the artificial neural network can be used to predict independent input data of the new.

Support vector machine

Support vector machine (SVM) is a supervised learning method can be widely used in statistical classification and regression analysis. Support vector machine is a class classifier with different kind of samples can be separated in the sample space hyper plane. That is a given number of marked well training samples. SVM algorithm outputs an optimized separating hyper plane. The essence of SVM algorithm is to find a can be a value maximizing hyper plane, this value is the minimum distance hyper plane distance of all the training samples. The minimum distance is called interval (margin)[16-17].

The following equation defines a hyper plane expression:

$$f(x) = \beta_0 + \beta^T x$$

Which β is called the weight vector β_0 called bias. Wherein x represents from those points closest hyper plane, these points are called support vectors.

The key to SVM is the kernel function. Vector set of low-dimensional space is often difficult division, so the solution is to map them to the high-dimensional space. But the difficulty of this approach is to bring the computational complexity increases, while kernel just ingenious solve this problem[15]. In other word, as long as the selection of appropriate kernel function, you can get a high-dimensional space classification function. In SVM theory, using different kernel function will lead to a different SVM algorithm to get a different output.

Research framework

Comparative studies of different algorithms to assess predictive capabilities, this study provide a very good solution to this practical problem car acceptability evaluation. Research framework shown in Figure 1, each stage of the process is as follows:

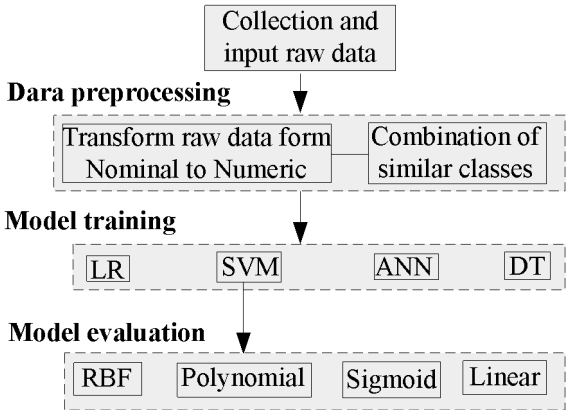


Figure 1 : Research Framework

Collection and input raw data: It comprises a collection of the original data and selects a car acceptability evaluation characteristic parameter.

Data pre-processing: First, the data used to calculate the nominal data format. Secondly, the car assess the acceptability of the data set is divided into four categories (unacceptable, acceptable, good and very good),in the present study, in order to simplify the complexity of the acceptability of the research into two categories, acceptable and unacceptable, the good kind of the same nature and very good class merging to acceptable class.

Modelling Research: Studies using logistic regression (LR), artificial neural network (ANN) and decision tree (DT), support vector machine (SVM) of four kinds of algorithm to calculate the evaluation of forecasting accuracy rate, sensitivity and specificity. Accuracy, sensitivity and specificity of the test method is as follows:

Process 1 collection and input raw dataset: It includes the collection of raw data, selecting the data and focusing on the features influence the car evaluation.

Process 2 pre-processing the dataset: This step includes three parts. Firstly, the data are transferred to forms "nominal to numeric" for calculating. Secondly, there are four classes (unacceptable, acceptable, good, and very-good) in car evaluation dataset. In this study, we combined the similar classes (acceptable, good and very-good) into one class. The four classes were combined to form two classes (unacceptable, acceptable).

Process 3 modelling training: Studies using logistic regression (LR), artificial neural network (ANN), decision tree (DT) and support vector machine (SVM) in total four kinds of algorithm to calculate the evaluation of forecasting accuracy rate, sensitivity and specificity. Accuracy, sensitivity and specificity of the test method is as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

If an instance of the class is positive and also predicted positive class, that is the true positive, if the instance of the class is predicted negative positive class, called false positive. Accordingly, if the instance is negative class is predicted to become the negative class, called true negative, the positive class is predicted to become the negative class is false negative.

THE EMPIRICAL RESEARCH

Data description

In this work, a real-world car evaluation database was taken from the UCI repository of machine learning database as described in TABLE 1. It contains 1728 instances and classified into four classes, there is no missing value in the dataset. The car evaluation database contains six attributes examples with a car (Buying, Main, Doors, Persons, Lug boot and Safety)

TABLE 1 : Car attribute Description

Characteristic	Attribute	Attribute description	Nominal values
Price	Buying	Buying price	vhigh, high, med, low
	Maint	Price of the maintenance	vhigh, high, med, low
Tech	Doors	Number of doors	2, 3, 4, 5, more
	Persons	Capacity in terms of persons to carry	2, 4, more
	Lug_boot	The size of luggage boot	small, med, big
	Safety	Estimated safety of the car	low, med, high

Clementine modelling

The pentagon-shaped nodes show the construction of the models using logistic regression, decision trees (CART) and neural network. The diamond-shaped nodes show the model outputs of the respective models. For the logistic regression model, four selection methods (ENTER, STEPWISE, FORWARDS, BACKWARDS) were compared using the Analysis and Evaluation nodes. While for decision tress, the C5.0, CHAID and CART models were generated and compared. Then, the three predictive models which are stepwise logistic regression, CART and neural network are connected to the "analysis" node which provides the computation of accuracy rates, while the evaluation node produces the lift charts.

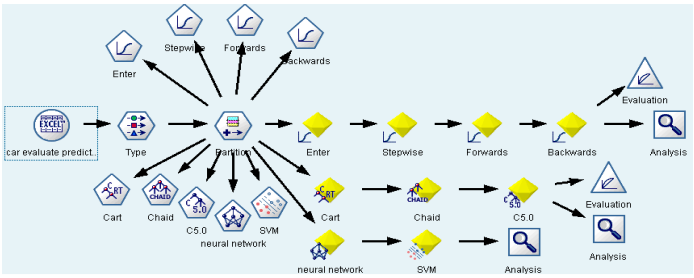


Figure 2 : Data mining process flow diagram

Comparison of results

The different modelling algorithm results as shown in the TABLE 2:

TABLE 2 : Comparison of modelling results

Model	Sample	Accuracyrate	Sensitivity	Specificity
LR	Training	95.40%	93.26%	96.31%
	Validation	96.25%	94.40%	97.04%
NN	Training	98.66%	98.60%	98.69%
	Validation	99.06%	98.15%	99.46%
SVM	Training	99.92%	100.00%	99.88%
	Validation	99.62%	100.00%	100.00%
CHAID	Training	95.06%	98.60%	93.56%
	Validation	92.31%	100.00%	88.94%
CART	Training	92.55%	87.08%	94.87%
	Validation	90.43%	88.27%	91.37%
C5.0	Training	96.07%	92.98%	97.68%
	Validation	93.06%	91.36%	93.80%

Four different logistic regression methods have the same accuracy, sensitivity and specificity. It also shows that the logistic regression model to assess the car acceptability is not significant.

Decision tree is the most easily understood model, and can be easily converted into a set of rules. In addition, decision tree algorithm can handle both discrete and continuous data, without the need for data to make a priori assumptions. Because of these advantages, the method of decision tree is widely used for classification and prediction. The table shows the difference between three kinds of decision tree model accuracy, sensitivity and specificity. Sensitivity considered true positive rate, refers to the actual acceptability of the probability is determined to be accepted. And specificity is considered the true positive rate, refers to the actual acceptability of not to accept the probability is determined unacceptable. The results of three decision tree algorithms are very close, but CART model has the best testing and prediction.

Artificial neural networks can be used to predict complex systems is determined by the relationship between the number of training samples, the training samples and the test samples, meanwhile, the forecast performance also depends on the choice of data structures, data quality and variables. In the present study, artificial neural network achieved good performance, second only to SVM.

In all these models, SVM has better predictive ability, which showed the best accuracy, sensitivity and specificity. The SVM prediction results are directly related to the choice of kernel, Secondary modelling four different kernel functions of SVM shown in Figure 3, experimental results show that the polynomial kernel function has the best predictive ability.

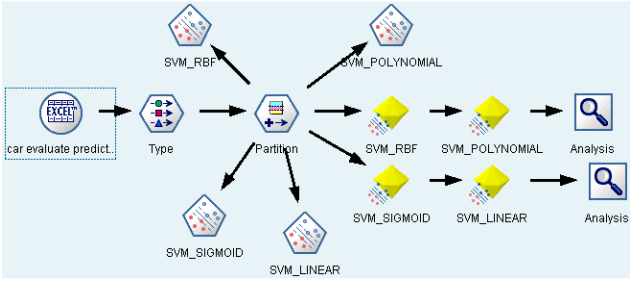


Figure 3 : SVM kernel function modeling flow

In summary, this study SVM polynomial kernel function to predict the car acceptability, with a very good performance, training and test sets of accuracy were 99.92%, 99.62%, sensitivity is 100%, specificity of 99.88% and 100% as shown in TABLE 2.

TABLE 3 : Performance of the different kernels

SVM kernel	Training	Validation
RBF	99.92%	99.62%
Polynominal	100%	99.62%
Sigmoid	70.21%	69.61%
Linear	94.31%	94%

The importance of the variables analyzed, four model results were almost identical, car security parameters are considered the most significant impact on the acceptability of the car the most, followed by the car can accommodate the number. The two parameter on consumer acceptability of the effects are less significant is the number and size of the trunk door.

TABLE 4 : Comparison of the importance of auto variables

Important	NN	SVM	LR	CARDT	CHAID	C5.0
Safety	0.319	0.453	0.451	0.419	0.383	0.334
Persons	0.274	0.281	0.272	0.288	0.205	0.327
Buying	0.142	0.226	0.276	0.181	0.204	0.226
Maint	0.133	0.001	0	0.098	0.131	0.091
Lug_boot	0.084	0.039	0	0.007	0.066	0.023
Doors	0.048	0	0	0.007	0.009	0

Conclusion

Four models used in the empirical car acceptability evaluation were compared and researched, the results showed that all four models have similar good predictive ability, and support vector machine model (polynomial kernel) showed the best accuracy, sensitivity and specificity, with the best predictive ability, can be very good for car acceptability evaluation. In the six attributes of the car, the safety has the largest influence on car acceptability followed by occupancy, however, consumers is less sensitive to these two factors of the size of the trunk and the number of door. This can help companies make better policy, targeted, improved methods to improve the car's acceptability and consumer satisfaction. Summing up the appeal, the use of data mining modelling method can accurately predict the acceptability of the car in order to build a good bridge between consumers and businesses, for the enterprise profits and consumers' satisfaction.

CONCLUSIONS

Acceptance of the product of growing concern, the manufacturer must know which factors influence consumers' buying decisions. In recent years, the product has been in-depth evaluation of the acceptability of research, unfortunately, manufacturers often misunderstand the real needs of consumers, and how to better evaluate the acceptability of the product is the key issue of product development. In this thesis, for the evaluation of vehicles lines empirical research, using four models for the evaluation of automotive forecasting a comparative study, the experimental results show that, using polynomial kernel SVM model can best be assessed on car evaluation prediction. In the evaluation of the car, the safety performance of the most significant factors, occupancy second, but customer have no special requirements of the size and the number of door. All in all, a comparative study different four models bade on data mining to evaluate car acceptability, results showed, SVM model can better solve the car evaluation, in turn, and the method can be extended to other industries to solve the evaluation of the product.

REFERENCES

[1] Libo Li, Frank Goethals, Antonio Giangreco, Bart Baesens; Using Social Network Data to Predict Technology Acceptance.ICIS2013. 12 **(2013)**.

[2] F.Liebana-Cabanillasa, R.Noguerasb, L.J.Herrerac, A.Guillenc; Analysing user trust in electronic banking using data mining methods.Expert Systems with Applications. **40(14)**, 5439–5447 **(15 October 2013)**.

[3] S.Makki, A.Mustapha, J.M.Kassim, E.H.Gharayebeh; Employing Neural Network and Naive Bayesian Classifier in Mining Data for Car Evaluation.ICGST, 4 **(2011)**.

[4] Hemanta Kumar Bhuyan, Maitri Mohanty, Smruti Rekha Das; Privacy Preserving for Feature Selection in Data Mining Using Centralized Network.IJCSI International Journal of Computer Science, **9(3-2)**, **(May 2012)**

[5] Shu-Ting Luo, Chwen-Tzeng Su, Bor-Wen Cheng; Developing a hybrid evaluation process for product acceptability: An Empirical Study in Automobile Industry. **5(7)**, 2708-2715 **(April 2011)**.

[6] Roisin McNaney, John Vines, Daniel Roggen; tec. Exploring the acceptability of google glass as an everyday assistive device for people with parkinson's.CHI '14 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2551-2554. **(2014)**.

[7] P.K.Yadav, K.L.Jaiswal, S.B.Patel, D.P.Shukla; Intelligent Heart Disease Prediction Model Using Classification Algorithms.IJCSMC, **2(8)**, 102-107 **(August 2013)**.

[8] Anuj Sharma, Prabin Kumar Panigrahi; A Review of Financial Accounting Fraud Detection based on Data Mining Techniques.Computer Science. **(2013)**.

[9] Chiming Chang, K.U.Leuven, Leuven; A Comparison of Classifiers for Intelligent Machine Usage Prediction.Intelligent Environments (IE), 2014 International Conference. **(July 2014)**.

[10] Xiuli Li, Rui Zhao, Yan Xiao; Electronic Commerce Data Mining using Rough Set and Logistic Regression.Journal of Multimedia, **9(5)**, 688-693 **(May 2014)**.

[11] Imran Kurt Omurlu, Mevlut Ture, Mustafa Unubol, Merve Katranci, Engin Guney; Comparing Performances of Logistic Regression, Classification & Regression Trees and Artificial Neural Networks for Predicting Albuminuria in Type 2 Diabetes Mellitus.IJSBAR, **(2014)**.

[12] Yunpeng Li, Jie Liu, Qiuchen Bao, Wenxiao Xu, Rehan Sadiq, Yong Deng; A new method of mapping relations from data based on Artificial Neural Network. **(Nov 2013)**.

[13] J.Gaurav Sawale, Dr. R.Sunil Gupta; Use of Artificial Neural Network in Data Mining For Weather Forecasting.International Journal Of Computer Science And Applications. **6(2)**, **(Apr, 2013)**.

[14] Shamsher Bahadur Patel, Pramod Kumar Yadav, Dr.D.P.Shukla. Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques.IOSR Journal of Agriculture and Veterinary Science (IOSR-JAVS). 61-64 **(Aug. 2013)**.

[15] Chun Fu Lina, Yu-Chu Yehb, Yu Hsin Hungc, I.Ray Changa; Data mining for providing a personalized learning path in creativity: An application of decision trees.Computers & Education, **68**, 199–210 **(October 2013)**.

[16] Shen, Runjie, Yang, Yuanyuan, Shao, Fengfeng; Intelligent Breast Cancer Prediction Model Using Data Mining Techniques. 26-27 **(Aug. 2014)**.

[17] M.A.H.Farquad, Vadlamani Ravi, S.Bapi Raju; Churn prediction using comprehensible support vector machine: An analytical CRM application.Applied Soft Computing. **19**, 31-40 **(June 2014)**.

[18] D.Akay, M.Kurt; A neuro-fuzzy based approach to affective design. Int. J. Adv. Manufact. Technol., **40**, 425-437.

[19] G.M.Alam; Can governance and regulatory control ensure private higher education as business or public goods in Bangladesh Afr. J. Bus. Manag., **3(12)**, 890-906 **(2009)**.

[20] M.Alnoukari, W.Alhussan; Using data mining techniques for predicting future car market demand; DCX case study. International Conference on Information and Communication Technologies: From Theory to Applications. IEEE Conference. **(2008)**.

[21] A.Arauzo-Azofra, J.M.Benitez, J.L.Castro; A feature set measure based on Relief. Proceedings of the 5th International Conference on Recent Advances in Soft Computing, 104-109 **(2004)**.