DOI: 10.54254/2755-2721/6/20230907

Price prediction of used cars using ISOMAP and SVM

Yan Bao

School of Computing, University of Birmingham, Birmingham, West Midlands, B15 2TT, United Kingdom

yxb978@student.bham.ac.uk

Abstract. This paper conducts model training and testing for estimating the price of used cars. In the second-hand car market, the residual value is one of the most important reasons for the second-hand car circulation. Because much of the way used cars are traded is traditional used car appraisers rely on their experience with used cars to determine the value of the car. At present, there is no comprehensive evaluation model based on objective factors of second-hand car characteristics. This study is based on the establishment of an objective evaluation of the effectiveness of the price prediction model and the objective factors affecting the second-hand car. In this research, the author wants to replace traditional evaluation methods with deep learning methods. The code written in this study includes the training of the data preprocessing model, flexibly uses NumPy to perform precision calculations, and uses Sklearn to delete highly relevant features for feature screening. The non-linear isomap algorithm is used for dimensionality reduction to keep the data at the same latitude, and finally the SVM algorithm is used for model training. Very unsatisfactory results have been obtained by training the model. The train accuracy and test accuracy were 22.5% and 14.3%, respectively. Through the analysis of the model, it may be that the dimensionality reduction of the isomap algorithm deleted the required influencing factors, resulting in the low training accuracy of the entire model.

Keywords: Used cars, price prediction, SVM, ISOMAP

1. Introduction

The pricing of second-hand cars mainly depends on the brand premium of the vehicle, the condition of the vehicle, the choice of the model, the configuration of the vehicle, whether there is an accident, the maintenance status of the vehicle, and other factors [1, 2]. In reality, there are still many problems in second-hand car trading, such as the excessively high prices of second-hand car dealers or the random pricing of second-hand car trading platforms.

Therefore, it is necessary to use machine learning to replace manual labor in the price of used cars. [3, 4]. By training the model, machine learning can obtain relatively stable and accurate judgments. Machine learning includes many features that humans don't notice or pay much attention to, so it can better predict used cars. The used car can be accurately judged, so that both buyers and sellers can get the maximum benefit, the seller can sell the car at the maximum return price, the buyer can also buy a suitable used car at a moderate price. Only in this way can the balance of the used car market be maintained. This is the purpose of this study.

This study uses the public data set of Aliyun [5]. The dataset contains 400,000 data volumes, 16 characteristic variables, and 15 anonymous variables. The test set and training set in this study are from the total training set. To avoid the influence of irrelevant features in the test set. Then, the data were preprocessed, including data unification and data dimension reduction. The data dimension reduction adopted Isomap algorithm. Why choose Isomap algorithm [6] is selected for dimensionality reduction is that the Isomap algorithm can analyze high-dimensional manifolds to obtain low-dimensional embedding so that adjacent structures between data points on high-dimensional manifolds can be obtained exactly corresponding to each other in low-dimensional embedding [7]. After the feature selection is completed, the support vector machine (SVM) can be used as a learning model for training data analysis [8, 9]. It can solve some machine learning problems under samples and also deal with the interaction between nonlinear features, which is the most important reason for choosing support vector machines. The SVM algorithm is chosen because This is the most classical algorithm, it provides a good basis to avoid overfitting. As long as you find the kernel function you can use, even in the nonlinear feature space, it can be very easy to run. It's perfect for high dimensional space text classification problems. However, the predicted results of the model were far from the expected results, and the predicted R-value [10] were also very unsatisfactory, accounting for only 0.225 and 0.143 on the training and test data respectively. After analysis, this may be due to the loss of some major influence features in the process of dimension reduction by Isomap algorithm, and the features collected by the data set itself may not be perfect, such as the policy support for second-hand car trading, different demands for different models in different regions, whether there have been car accidents and other features. Or it could be that the training results are not ideal and therefore the training results are not ideal. This paper analyzes and the traditional used car price appraisal can only rely on the experience of the appraiser to judge the value subjectively. Subjective judgments lead to used car prices that are not evaluated in a scientific way. The used car surplus value is also an important factor to predict the price in the car circulation. There is no comprehensive evaluation of the price of used cars based on objective factors. This study establishes a prediction model for objectively evaluating the price of used cars. In this research, the author wants to replace traditional evaluation methods with deep learning methods. The code written in this study includes the training of the data pre-processing model, flexibly uses NumPy to perform precision calculations, and uses the SKLearn to delete highly relevant features for feature screening. The non-linear Isomap algorithm is used for dimensionality reduction to keep the data at the same latitude, and finally, the SVM algorithm is used for model training. Very unsatisfactory results have been obtained by training the model. The training accuracy and testing accuracy were 22.5% and 14.3%, respectively. Through the analysis of the model, it may be that the dimensionality reduction of the Isomap algorithm deleted the required influencing factors, resulting in the low training accuracy of the entire model.

2. Method

2.1. Dataset

This data comes from Aliyun [5], This dataset contains used car transaction records from a used car trading website. The dataset contains 40w of data, including 31 variables, 16 columns of desensitization features and 15 columns of anonymous features. The variables are shown in Table 1. All the data are desensitized and contain 15 anonymous features.

Table 1. Characteristics of this data set(continue).

Field	Description
SaleID	Transaction ID, unique code
name	Automobile transaction name
regDate	Date of car registration
model	Car type code

Table 1. (continued).

bodyType	Body type: limousine: 0, minicar: 1, van: 2, bus: 3, convertible: 4, two-door car: 5, commercial vehicle: 6, mixer: 7
brand	Brand of Automobile
FuelType	Fuel type: Gasoline: 0, diesel: 1, LPG: 2, natural gas: 3, Hybrid: 4, Others: 5, electric: 6
Gearbox	Transmission: Manual: 0, automatic: 1
power	Engine power: Range [0, 600]
notRepairedDamage	The car has unrepaired damage: Yes: 0, No: 1
Kilometer	The car has traveled thousands of kilometers
RegionCode	Area code
OfferType	Offer type: Provided: 0, Request: 1
Seller	Seller: Individual: 0, non-individual: 1
CreatDate	The time when the car goes online is the time when it starts selling
Price	Used car transaction Price (forecast target)
v series features	Anonymous features, including v0-14 and 15 anonymous features

2.2. Dataset pre-processing

For operators and designers to manage the data more convenient and flexible provides a detailed data segmentation overall, the data is divided into small physical units. Small physical units have the advantages of easy reconstruction, free indexing, sequential scanning, easy recombination, easy recovery and easy monitoring. Part of the nature of a data warehouse is flexible access to data, and large chunks of data don't do that. The total training set is separated for training and testing respectively. In order to prevent the total data set data volume is huge and increasing the computational difficulty, only data from 0 to 10,000 was intercepted for analysis.

2.3. ISOMAP

SIOMAP is a conventional manifold learning algorithm. In traditional machine learning methods, the distances and mapping functions between data points and data points Defining Euclidean space, In reality, Data points may not be distributed in Euclidean space, so nonlinear data is difficult to realize in traditional Euclidean space, and assumptions about data distribution are introduced into it. If it is assumed that the learned data point distribution of the processed manifold is embedded in the Euclidean space, a latent manifold is formed through these data points.

The process of the ISOMAP is as follows: in i = 1,2,3,4,5...m do, Determine the k neighboring values of Xi; Euclidean distance is the distance between Xi and K nearest neighbors, set other points for infinite distance, end for, the distance dist(Xi, Xj) between two samples will call the shortest path algorithm to calculate . Use this coordinate point as the input of the MDS algorithm to obtain the output of the MDS algorithm.

$$Z = V_Z A_Z^{1/2} \tag{1}$$

2.4. Model construction

Support Vector Machine (SVM) is a binary classification model. The SVM algorithm can deal with linear and nonlinear data problems. For nonlinear data, kernel functions should be used to implicitly map its input to high-dimensional space. In order to maintain a suitable calculation progress, it is generally necessary to select the kernel function of the problem to map. Make sure to compute the dot product in space. The base model is a linear classifier in feature space. Maximize the interval to form a quadratic programming problem solution. The larger the hyperplane of the classifier, the more stable the data classification. A hyperplane in a multidimensional space can be represented by a linear equation:

$$W^T \cdot X + B = 0 \tag{2}$$

The SVM has many advantages. It can solve high dimensional problems and Machine learning works with small samples. The reciprocity of nonlinear features can be used to process small samples quickly. Compared with other algorithms such as neural network, SVM algorithm does not need to rely on all the data, because there is no local minimum in this algorithm.

The disadvantage of SVM algorithm is that when there are too many samples observed, the efficiency of the algorithm will be greatly reduced. For nonlinear problems, it is necessary to find appropriate kernel function to solve them through research. Only one kernel function corresponds to one solution, and there is no universal kernel function. It is very sensitive to missing data, which will cause data errors. The interpretation of higher dimensional mappings of kernel functions is not strong.

2.5. Evaluation index

The R value is used to verify the accuracy of the data model. Slice the matrix using indices [0,1] to get the values of R, which are the correlation coefficients. Square the value of R to obtain the value of R squared. q When r is 1 or -1, it represents a completely linear variable, where r equal to 1 is completely positive correlation and -1 the opposite. When r equals 0, no linearly related variable of village officials, but it cannot be ruled out that there is no nonlinear variable. When -1 is less than or equal to r and less than or equal to 0, there is a negative correlation; when r ranges from 0 to 1, it means there is a positive correlation variable. When r is larger than 0.8 or smaller than -0.8, there is a highly correlated variable. When the absolute value of r is between 0.5 and 0.8, it denotes that there are moderately correlated variables. The absolute value of constant r is between 0.3 and 0.5, indicating a low correlation variable. When r between -0.3 and 0.3, it means that there is no linearly dependent variable.

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{x\sum y^2 - (\sum y)^2}} \tag{3}$$

3. Result

Table 2 demonstrates the model training results.

Table 2. R-value results of the model.

	R value
Train	0.2257860901900868
Test	0.1437823173237496

Obviously, the training results of the model are not ideal, and there is almost no linear correlation rule. The accuracy gap between the model and the predicted model is very large. As the SVM algorithm is one of the most classical algorithms, the model is very mature, but the obtained results are not ideal. This may also be due to data preprocessing, which screens out important features of the. When the total training set is abandoned, some features contained in the total test set will be lost, resulting in a very low model accuracy.

4. Discussion

There may be several reasons for the unsatisfactory results obtained by training the training model. The first point may be that the dimensionality reduction of Isomap algorithm screens out some features that affect model training, thus losing the model accuracy. The second point may be that the data set itself does not contain more features that affect second-hand car trading, such as the color of the vehicle, the use of the vehicle in the sale area, the size of the vehicle accident, and the number of transactions of the vehicle, which will affect the model accuracy of second-hand car trading. The third point may be that the model is not advanced enough. As SVM is difficult to achieve large-scale training samples, the model is not advanced enough and many features cannot be selected and trained, which is also one of the reasons for the unsatisfactory training results of the model. If the characteristics of the data set are more perfect and the model is optimized, better accuracy may be obtained

5. Conclusion

This research uses the machine learning knowledge learned and the most classical support vector machine algorithm to train the data. Although the training accuracy of the model is very low, it also indicates that the machine learning model is not advanced enough, which will greatly reduce the efficiency of the model, and the model training of neural computation cannot be carried out without advanced hardware. In the future, if neural computation can be used to train the model, it may greatly improve the accuracy of the model. Compared with machine learning, neural computing algorithm is more advanced and its performance is better than machine learning algorithm. Nowadays, neural computing has surpassed traditional machine learning in many aspects, such as speech, natural language processing, vision and other aspects. Deep learning can better expand data. For machine learning, this may not have any effect. No complex feature engineering is required for neural computation, which greatly reduces the tedious process. It is possible to improve the predictive accuracy of the used car price by studying the neural computation and the training of the model can be completed only by enlarging the data sets. It does not completely deny the role of machine learning. Machine learning may be better than neural computing on small data sets. Small data cannot improve the training of large amounts of data, so machine learning is more convenient to complete model training. Neural computing also requires a lot of expensive hardware to train. Compared with machine learning, the most common hardware can complete the training, and different algorithms can be tried and trained in a relatively short time. The prediction is more in line with the training and testing of neural computing. If the second-hand car price prediction of neural computing can be completed in the future. This is the direction and plan of the future development of this research.

References

[1] Sun, N., Bai, H., Geng, Y., & Shi, H. (2017). Price evaluation model in second-hand car system based on BP neural network theory. In 2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, 431-436.

[2] Oprea, C. (2011). Making the decision on buying second-hand car market using data mining techniques. The USV Annals of Economics and Public Administration, 10(3), 17-26.

[3] Pudaruth, S. (2014). Predicting the price of used cars using machine learning techniques. Int. J. Inf. Comput. Technol, 4(7), 753-764.

[4] Fathalla, A., Salah, A., Li, K., Li, K., & Francesco, P. (2020). Deep end-to-end learning for price prediction of second-hand items. Knowledge and Information Systems, 62(12), 4541-4568.

[5] Aliyun (2020) Second hand car price prediction. URL: https://tianchi.aliyun.com/competition/entrance/231784/information

[6] Balasubramanian, M., & Schwartz, E. L. (2002). The isomap algorithm and topological stability. Science, 295(5552), 7-7.

[7] Ghojogh, B., Ghodsi, A., Karray, F., & Crowley, M. (2020). Multidimensional scaling, sammon mapping, and isomap: Tutorial and survey. arXiv preprint arXiv:2009.08136.

[8] Noble, W. S. (2006). What is a support vector machine?. Nature biotechnology, 24(12), 1565-1567.

[9] Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. In Machine learning, 101-121.

[10] Asuero, A. G., Sayago, A., & González, A. G. (2006). The correlation coefficient: An overview. Critical reviews in analytical chemistry, 36(1), 41-59.