Mining model of purchase intention based on support vector machine algorithm

Mengtong Zhang
School of International Education
Wuhan University of Technology
Wuhan,China
zhang_000.t@whut.edu.cn

Jichang Dong
School of Science
Wuhan University of Technology
Wuhan,China
2504157235@qq.com

Wenkan Huang
School of Computer Science and Artificial Intelligence
Wuhan University of Technology
Wuhan,China
769397149@qq.com

***Abstract*—The automobile industry is an important pillar industry of the national economy, while new energy automobile industry is a strategic emerging industry. Firstly, the Spearman correlation coefficient of each index of each brand is calculated, and then the influencing factors of each brand are obtained, and the customer mining model of each brand is established. Finally, it tests the willingness of 15 target customers to buy electric vehicles and conducts k-fold cross-verification. The results show that the purchase user's number of different brands is NO.2 and 4 for brand 1, NO.7 for brand 2 and NO.12 for brand 3. The comprehensive prediction accuracy is as high as 0.94, indicating that the model has high sensitivity and the test results are accurate.**

Keywords-New energy automobile; Spearman correlation coefficient; SVM support vector machine model; K-fold cross-examination; Customer mining model

I. INTRODUCTION

The automobile industry is an important pillar industry of the national economy, and the new energy vehicle industry is a strategic emerging industry. Vigorously developing new energy vehicles represented by electric vehicles is an effective way to solve energy and environmental problems, with broad market prospects. However, for electric vehicles, as an emerging thing, consumers still have some doubts in some areas compared with traditional vehicles, and their market sales require scientific decision-making[1-2].There are many influencing factors that determine whether the target customer is going to buy electric vehicles, including the factors of the electric vehicles themselves and the personal characteristics of the target customer. Therefore, quantitative statistics on the data are conducted[3-4](The data in this paper comes from: http://shumo.neepu.edu.cn/).

II. ANALYSIS OF THE FACTORS INFLUENCING THE SALES OF THREE BRANDS OF ELECTRIC VEHICLES

By observing the data, it can be learned that its problems basically revolve around the "economy", "region" and "number of families" of car buyers. According to be seen from 0 / 1 data, this is a second classification problem. This article can divide the car buyers according to different brands and analyze the correlation of a1－a8, B1－ B17 and purchase willingness respectively. The stronger the correlation, the greater the impact, so as to determine the impact of a given factor on the purchase behavior. The solution idea is shown in Figure 1.

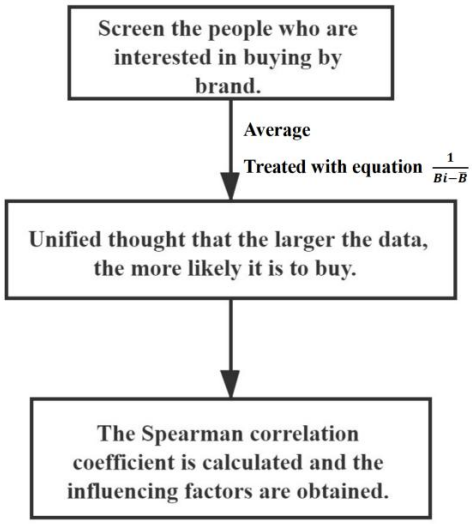

Figure 1. Thought diagram

For the cleaned data, all users willing to buy electric vehicles are selected, and then those willing to buy will be divided according to different brands. Learn through the information:

- If the data quantity is large enough, according to the large number law, the central limit theorem can guarantee an approximate normal distribution;
- If the continuous data meet the normal distribution and meet the linear relationship, the pearson correlation coefficient is the most appropriate.

If any of the above conditions are not met, the spearman correlation coefficient is used. In this paper, the data were analyzed with SPSS(Statistical Product and Service Solutions) to obtain linear regression analysis of 25 factors of three brands, among which the factor significance level is below 0.05 after regression, so there is a reason to think that it does not conform to the normal distribution and its data is discrete, so that the spearman correlation coefficient was selected for testing.

The analysis shows that the greater the B2, B3, B5, B7, B9, B10, B13, B14, B15 data, the more likely the people to buy electric vehicles; the closer the B1, B4, B6, B8, B11 data to the median, the more likely to buy the electric

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00018

vehicle; the smaller the B12, B16, B17 problem corresponding data, the more likely it is to buy an electric vehicle. For this, process the data:

- For B1, with only 1,2,3, changing 3 to 1 makes the greater the value, the greater the willingness to buy.
- For other factors, the paper calculates the average of one problem of the people willing to buy, thinking that the closer to this average, the more likely to buy.

Use the following formula for analyze:

$$\frac{1}{|\mathrm{Bi}-\overline{\mathrm{B}}|}(\mathrm{I}=1,4,6,8,11) \quad (1)$$

The spearman correlation coefficient is calculated, and the influencing factors of the three brands are shown in Table 1.

TABLE I. TABLE TYPE STYLES

Brand name	Influencing factor
Brand name 1	a1-a8，B16，B17
Brand name 2	a1-a8，B12，B13，B15，B16，B17
Brand name 3	a1-a8，B13，B16

III. CUSTOMER MINING MODEL OF DIFFERENT BRANDS OF ELECTRIC VEHICLES

A. Overall idea of the model

In fact, the customer mining model to be established in this section is a binary classification model for whether to buy or not, so we consider the SVM (Support Vector Machine) algorithm. The basic model of the SVM is to find the best separation hyperplane on the feature space with the largest positive and negative sample intervals on the training set. Considering that this problem requires machine learning processing, this article changes the purchase intention to be 0 to-1, -1 means no purchase, and 1 means purchase. In addition, because it is a binary classification problem, and its small classification dimension and small data volume, we consider using the SVM algorithm to establish a shopping willingness prediction model and solve it. The advantage of the SVM algorithm is that it can handle classification and even multiple classification problems well, but because the SVC (Support Vector Classification) used in this paper is a time complexity O(n^2) algorithm, its operation speed is greatly reduced for high dimensions (e. g., 1000 dimensions) and high data volume (e. g., 1 million bars).However, it is calculated that the data selected in this paper meet the computable range of SVC on an ordinary computer, so this model is used.SVC is a support vector machine for classification, so this article is not going to go into much detail.

B. Model establishment

As for data segmentation, the first thing that comes to mind in this paper is the perceptron. The principle is to find a straight line and separate the data. If it rises to the high latitude, it is found that a hyperplane separates the data at the high latitude. The number of hyperplane of the perceptron may be at infinity, so the support vector machine model can be understood as finding the best one[5].

a) Functional interval

Function interval is the expression of the loss function of the perceptron, and can be expressed as:

$$\gamma' = \mathrm{y}(\omega^{\mathrm{T}}\mathrm{x}+\mathrm{b}) \quad (2)$$

b) Geometric interval

Geometric interval is the true distance from the point to the hyperplane, namely the function interval divides the normal vector:

$$\gamma = \frac{\mathrm{y}(\omega^{\mathrm{T}}\mathrm{x}+\mathrm{b})}{||\omega||_2} = \frac{\gamma'}{||\omega||_2} \quad (3)$$

As shown in Figure 2, the classification hyperplane $\omega^{\mathrm{T}}\mathrm{x}+\mathrm{b}=0$ can not only separate all samples, but also maintain a certain functional distance from the nearest sample point (support vector) (this function distance is 1), then such a classification hyperplane is better than the classification hyperplane of the perceptron. It can be shown that there is only one such hyperplane[6].

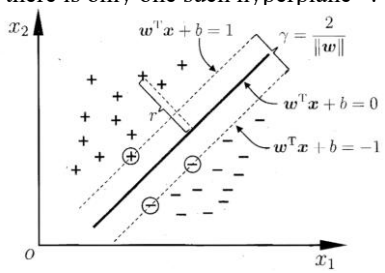

Figure 2. Support vector and interval

The SVM algorithm is to make all distances to the hyperplane greater than a certain distance, that is, all classification points are to be on both sides of the support vector of the respective categories. The mathematical formula is expressed as:

$$\max \gamma = \frac{\mathrm{y}(\omega^{\mathrm{T}}\mathrm{x}+\mathrm{b})}{||\omega||_2} \ \mathrm{s.t}\ y_i(\omega^{\mathrm{T}}x_i+\mathrm{b}) = \gamma'^{(\mathrm{i})} \geq \gamma'(\mathrm{i}=1,2,\ldots,\mathrm{m}) \quad (4)$$

Generally, this paper takes the function interval $\mathrm{y}'=1$, so that the optimization function of this paper is defined as:

$$\max \frac{1}{||\omega||_2} \ \mathrm{s.t}\ y_i(\omega^{\mathrm{T}}x_i+\mathrm{b}) \geq 1(\mathrm{i}=1,2,\ldots,\mathrm{m}) \quad (5)$$

The upper equation is equal to:

$$\min \frac{1}{2}||\omega||_2 \ \mathrm{s.t}\ y_i(\omega^{\mathrm{T}}x_i+\mathrm{b}) \geq 1(\mathrm{i}=1,2,\ldots,\mathrm{m}) \quad (6)$$

C. Oversampling of the samples

In the selected sample data, the number of negative samples far exceeds the number of positive samples (954: 20), so you cannot use the random sample segmentation of svc_train_test_split for learning samples directly. In random cases, the samples marked as supported may not be selected in the training set. In view of this situation, the SMOTE method provided by sklearn is adopted in this paper. For this serious imbalance, predictions are often biased that classification results are biased towards more observed classes. The easiest way to deal with this problem is to construct 1: 1 data and then remove some of the more categories (i. e., undersampled), or the less categories are Bootstrap sampled (i. e., oversampled). It's a problem. Firstly, the excluded data will lead to the loss of some hidden information. In the second, a simple

replication of the put-back sampling makes the model overfitting.

The basic idea of the SMOTE algorithm is to analyze and simulate a few categories of samples and add new samples simulated manually to the dataset, thus making the categories in the original data not seriously out of balance. The simulation process of the algorithm adopts k-Nearest Neighbors technology, simulated to generate new samples, and artificially adds partial noise to entry, making the data more credible.

Similarly, because of the sample imbalance, after using SMOTE, the data not oversampled can still be weighted with a higher weight, further increasing the training of the model. The dimension reduction diagram of each brand is shown in Figure 3.

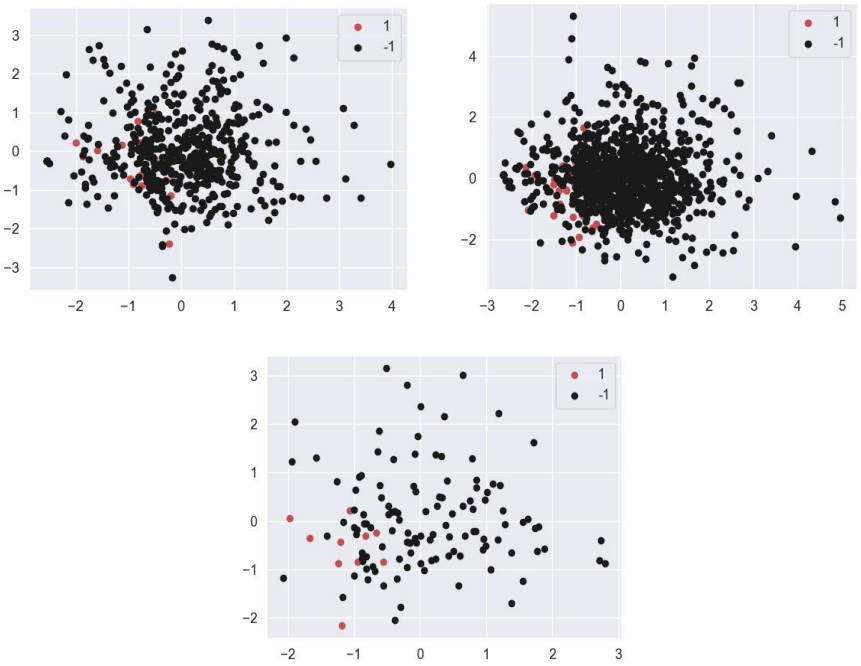

Figure 3. Pca(Principal Component Analysis) dimension reduction diagram of brand 1 (upper left), brand 2 (upper right) and brand 3 (under)

```
[[32  0]
 [ 4 36]]
report:
              precision    recall  f1-score   support

          -1       0.89      1.00      0.94        32
           1       1.00      0.90      0.95        40

    accuracy                           0.94        72
   macro avg       0.94      0.95      0.94        72
weighted avg       0.95      0.94      0.94        72

train_data: 1.0
test_data: 0.9444444444444444
```

```
[[356   4]
 [  2 355]]
report:
              precision    recall  f1-score   support

          -1       0.99      0.99      0.99       360
           1       0.99      0.99      0.99       357

    accuracy                           0.99       717
   macro avg       0.99      0.99      0.99       717
weighted avg       0.99      0.99      0.99       717

train_data: 1.0
test_data: 0.9916317991631799
```

```
[[32  0]
 [ 1 39]]
report:
              precision    recall  f1-score   support

          -1       0.97      1.00      0.98        32
           1       1.00      0.97      0.99        40

    accuracy                           0.99        72
   macro avg       0.98      0.99      0.99        72
weighted avg       0.99      0.99      0.99        72

train_data: 1.0
test_data: 0.9861111111111112
```

Figure 4. Adjustment parameters of brand 1 (upper left), brand 2 (upper right) and brand 3 (under)

D. SVM parameter tuning

We are using support vector machine to complete the entire model building so the parameters of svc should be

made certain adjustments to achieve the best results. The adjustment parameters of each brand are shown in Figure 4.

- Kernel function: the kernel functions used in this paper is chosen between the linear and radial basis function Gauss. Linear kernel: mainly used for linear separable cases. Less parameters, faster speed, the classification effect for general data is ideal. RBF kernel: mainly used for linear non separable cases. There are many parameters, and the classification results depend on them. Many people cross-verify the training data to find the appropriate parameters, but this process is time consuming. Because of the small number of samples, the RBF kernel function used has also achieved good verification results[7-8].
- Sample segmentation: 70% samples were learned and 30% samples were tested.
- Classification mode: because it is a binary classification problem, so the 'One-vs-One' mode is used for learning, which is the one-to-one classification method.
- Penalty parameter: the penalty parameter is processed for linear not separable data points. Since the data in this paper is not separable linear, the general penalty parameter value 10 is used to correct the data beyond the hyperplane.

IV. TEST THE POSSIBILITY OF 15 TARGET CUSTOMERS BUYING ELECTRIC VEHICLES

When using support vector machine to solve, it is found that SVC.fit (X, y) cannot achieve ideal results, and its data is not randomly separated, so the test set is recognized by the machine when it is implemented. After some modifications, we amended it to svm.SVC, so that the size and random range of the test set and the training set can be specified. However, in the cross-test, this paper found that the reliability of the data is not high, which leads to the following results: negative attitude is valid, but positive attitude is not. Therefore, analysis of its data set shows that the data has great limitations, with less supporting data and more supporting opposing data. Therefore, the way that decided to use oversampling is to artificially add the noise-containing data, and then balance the raw data in a weighted way, so that the raw data has a higher weight and the manually generated data has a lower confidence. After integrating the data through the above processing, the model has enough data for training. Finally, the predict_proba function is used to judge the supporting level of the model for each data. Then we can judge the new data set. The results are shown in Table 2:

TABLE II. PURCHASE USER FORECAST

Brand name	Purchase user number
Brand name 1	2，4
Brand name 2	7
Brand name 3	12

The prediction model is based on SVM algorithm, those support vectors with good discrimination ability for classification can be automatically found by learning the algorithm. The constructed classifier can maximize the interval between class and class, thus have better adaptation ability and higher classification rate, and the method only needs the category of boundary samples of various domains to determine the final classification results. After the above operation, the data are cross-verified by k-fold. The obtained values mainly observe the results of f1−socre, and the closer to 1, the better the model can judge the label. Thus this comprehensive prediction accuracy of this model is as high as 0.94.

V. CONCLUDING REMARKS

Firstly, the SVM algorithm in this paper can greatly simplify the classification problem. Secondly, its final decision function is determined by only a few support vectors, and the computational complexity depends on the number of support vectors rather than the dimension of the sample space, avoiding the "dimensionality disaster". In addition, a few support vectors determine the final result, which can not only help us capture the key samples and "eliminate" a large number of redundant samples, but also make the method not only algorithmically simple, but also "robust". Follow-up research can be completed by setting the weights of dimension. In each multidimensional dimension trained, each dimension weight is uniform, which against data with unique characteristics. These data can be analyzed, weighted by each vector and then added to learning.

REFERENCES

[1] Chen lei, Wang peiyong. prediction model of primary water supply pipe leakage time based on genetic least squares support vector machine [J]. journal of Zhejiang university of technology, 2021,(05):546-549.

[2] Yang Di, Fang Yangxin, Zhou Yan. Research on new classification based on MEB and SVM [J]. Journal of Guangxi Normal University (Natural Science Edition), 2021,(05):1-10.

[3] Hu Xuan, Li Chun, Ye Kehua. Application of support vector machine optimized by grey wolf algorithm in fault diagnosis of wind turbine gearbox [J]. Mechanical Strength, 2021,(05):1026-1034.

[4] Liu Chenyang, Xu Huang Rong, Duan Feng, Wang Taisheng, Lu Zhenwu, Yu Weixing. Spectral identification of rabbit liver VX2 tumor based on genetic algorithm and support vector machine [J]. Spectroscopy and Spectral Analysis, 2021,(10):3123-3128.

[5] Astuti Suryani Dyah,Tamimi Mohammad H.,Pradhana Anak A.S.,Alamsyah Kartika A.,Purnobasuki Hery,Khasanah Miratul,Susilo Yunus,Triyana Kuwat, Kashif Muhammad, Syahrom Ardiyansyah.Gas sensor array to classify the chicken meat with E. coli contaminant by using random forest and support vector machine[J].Biosensors and Bioelectronics: X,2021,9(9):5-8.

[6] Tan Hongchuang,Xie Suchao,Liu Runda, Ma Wen.Bearing fault identification based on stacking modified composite multiscale dispersion entropy and optimised support vector machine[J].Measurement,2021,186(186):6-.

[7] Xiao Shijie, Wang Qiaohua, Li Chunfang, Zhao Limei, Liu Xinya, Lu Shiyu, Zhang Shujun. Construction of milk purchasing grading model based on random leapfrog and support vector machine [J]. Smart Agriculture (English and Chinese), 2021,(05):1-9.

[8] Li Mengmeng, Liu Jingdang, Liang Tianyi, Tan Liang, Wang Gang, Zhu Xi. Prediction model of sulfur element in magmatic sulfide deposits based on support vector machine algorithm [J]. Journal of Jilin University (Earth Science Edition), 2021,(05):1-15.