Analysis of Tourists' Satisfaction with Scenic Spots

Xinyu Guo, Yuxi Su

School of Applied Mathematics, Beijing Normal University, Zhuhai, Guangdong, 519087, China

Abstract: *This article analyzes the tourist's comments on the scenic spot using TF-IDF, principal component analysis and logistic regression, and obtains the factors that influence tourists' satisfaction with the scenic spot. First, pre-process the data is needed, and then use the precise mode in jieba word segmentation to segment the text, and calculate the top 20 high-frequency words for each scenic spot and hotel. Then merge the 20 popular words of 50 scenic spots (hotels) that were mined together as a data pool, use the TF-IDF algorithm to calculate the feature the lexical item weight is reduced by the kernel principal component method (KernelPCA) to obtain the weight matrix. After that, the data is processed by classification and regression. In terms of classification processing: combine the scenic spot (hotel) score as the classification result and supervised learning using the naive Bayes algorithm, the support vector product machine algorithm, the B P neural network method and the logistic regression method. In terms of regression processing: model evaluation according to the mean squared error (Mean Squared Error, MSE), and finally the classification processing MSE index is better than regression processing.*

Keywords: *jieba, TF-IDF, PCA, SVM, naive Bayes, BP neural network*

1. Background introduction

Tourist satisfaction is closely related to the reputation of the destination. The higher the tourist satisfaction, the greater the reputation of the destination. Therefore, mastering the influencing factors of destination tourist satisfaction, effectively improving tourist satisfaction, and ultimately enhancing the reputation of the destination, can not only ensure the stability of the source of tourists, but also have a long-term and positive effect on the scientific supervision of tourism enterprises, the optimal allocation of resources, and the continuous market development. Effect.

2. Establish a comprehensive evaluation model

2.1 Data Preprocessing

2.1.1 Word frequency matrix

Scenic popular word label as the column of sieves to repeat a total of the remaining two words 51 is valid words, the most popular word hotel sieve column labels to repeat words in total remainder one 57 is valid words.

2.1.2 TF-the IDF algorithm with weighting matrix

(1) Applied to text word frequency:

In TF-IDF algorithm is applied to the heat words, word frequency calculating TF. The formula is as follows:

$$TF_w = \frac{Number\ of\ times\ a\ word\ W\ appears\ in\ a\ class}{Number\ of\ all\ words\ in\ this\ class} \tag{1}$$

Inverse document frequency (IDF):

$$IDF = \log(\frac{Total\ number\ of\ requests\ for\ the\ corpus}{Number\ of\ documents\ containing\ the\ entry\ w+1}) \tag{2}$$

2.1.3 Kernel Principal Component Analysis (Kernel PCA)

Using principal component analysis to reduce dimensionality, PCA is known to be linear, and it is often powerless for nonlinear data. Select feature attributes. It is obvious that the feature attributes are nonlinear structures, so choose the kernel principal component analysis method reduces the dimension of the weight matrix. (Keep 85% of the data, reduce the dimensionality to get 34 main components represented by T):

2.2 Model Established

2.2.1 Naive Bayes classifier

Step 1: For Naive Bayes, $X=(x_1,x_2,\dots x_D)$represents a data object with D-dimensional attributes. The training set S contains K categories, which is expressed as $y=(y_1,y_2,\dots y_k)$: Knowing the data object N to be classified, predicting the category of N, the calculation method is as follows :

$$y_k = arg_{y_k \in y} \max(P(y_k|X)) \tag{3}$$

Step 2: According to Bayes' theorem, the $P(y_k|X)$calculation method is as follows:

$$P(y_k|X) = \frac{P(X|y_k)\times P(y_k)}{P(X)} \tag{4}$$

Step 3: According to Bayes' theorem, the $P(y_k|X)$calculation method is as follows:

$$P(X|y_k) = \prod_{d=1}^{D} P(x_d|y_k) = P(x_1|y_k)P(x_2|y_k)\dots P(x_D|y_k) \tag{5}$$

Step 4: The calculation method of $P(y_k|X)$

If the property Ad is a discrete property or a classification property. Data objects belonging to category $y_k y_k$ in the training set, with n phase different attribute values under attribute AdAd; data objects with category y_k, and attribute value *xdxd* under attribute Ad in the training set are m.Therefore, the calculation method of $P(y_k|X)$ is as follows:

$$P(x_d|y_k) = \frac{m}{n} \tag{6}$$

If the attribute *AdAd* is a continuous attribute. It is usually assumed that the continuous attributes all obey the Gaussian distribution with the mean $\mu$and the standard deviationσ: Therefore, the calculation method is as follows:

$$P(x_d|y_k) = G(x_d,\mu_{y_k},\sigma_{y_k}) \tag{7}$$

2.2.2 SVM

Based on the previous analysis, it can be obtained that the characteristic attributes are non-linear, so the non-linear SVM is used for the model to establish the non-linear function to map the input data to the high-dimensional space and then the linear SVM can be used to obtain the non-linear SVM. The nonlinear SVM has the following optimization problems:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + c\sum_{i=1}^{iv} \xi_t \tag{8}$$

$$s.t\ \ y_i[w^T\phi(x_i)+b] \geq 1-\xi_i,\quad \xi_i$$

Analog soft margin SVM, nonlinear SVM has the following dual problems:

$$\max_{a} \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N} \sum_{j=1}^{N}[\alpha;y_i\phi(x_j)y_j \cdot \partial_j] \tag{9}$$

$$s.t.\ \sum_{i=1}^{N}\alpha_i y_i = 0, \qquad 0 \leq \alpha_i \leq C$$

Noting the product of formula present in the mapping function, it is possible to use nuclear methods, i.e. directly select kernel function: $k(x_i,x_j)=\phi(x_i)^T\phi(x_j)$. The KKT condition of the dual problem of nonlinear SVM can be similarly analogous to the soft-margin linear SVM.

2.2.3 Logistic regression

Log linear regression tries to approximate the true label y with the predicted value of the linear model, so the linear regression model can be abbreviated as $\mathbf{\ln \boldsymbol{y} = \boldsymbol{w}^T\boldsymbol{x}+\boldsymbol{b}}$, try to let $\boldsymbol{e^{w^T x+b}}$approximate y . On this basis, we need to find a monotonically differentiable function to label the true label of the classification task y is connected with the predicted value of the linear regression model, and on this basis, the log probability function is introduced. Substituting the logarithmic probability function into it, we get $y = \frac{1}{1+e^{-(w^T x+b)}}$, the two of y.The result is regarded as a class posterior probability estimate, which can be expressed as the following two formulas:

$$p(y=1|x) = \frac{e^{W^T x+b}}{1+e^{W^T x+b}} \qquad p(y=0|x) = \frac{1}{1+e^{W^T x+b}} \tag{10}$$

The objective function of logarithmic probability regression is an arbitrary-order derivable convex function, which has good mathematical properties. The data optimization algorithm can be used to find the optimal solution, that is, to determine the values of w and b in the formula. The common ones are gradient descent method, maximum likelihood method estimation and so on.

2.2.4 Neural Network algorithm

Without determining the mathematical equations for the mapping relationship between input and output, artificial neural networks learn some rules only through their own training and get the results closest to the desired output value at a given input value. As an intelligent information processing system, the core of the artificial neural network realizing its function is the algorithm. The BP network is the addition of several layers (one or multiple layers) neurons between the input and output layers, called hidden units, they have no direct contact to the outside world, but their state changes, can affect the relationship between input and output, each layer can have several nodes.

2.2.5 Linear regression

Use $X = (x_1, x_2, \ldots.. x_n)^T \in R^{np}$ to epresent the data matrix, which $x_i \in R^p$ represents a p-dimensional long data sample; $y = (y_1, y_2, \ldots.. y_n)^T \in R^n$represents the label of the data , here only one type of case for each sample is considered. The linear regression model is as follows. For a sample x_i, its output value is a linear combination of its features:

$$f(x_i) = \sum_{m=1}^{p} w_m x_{im} + w_0 = w^T x_i \tag{11}$$

The goal of linear regression is to fit the target label as much as possible with the predicted result , using the most common *Least square* as *Loss function*:

$$J(w) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 = \frac{1}{n}||y - X_w||^2 \tag{12}$$

The goal of linear regression optimization is to minimize the sum of the distance to the segmentation plane, which is the least square; linear regression is the objective function of convex , and the predicted value for each sample point is $f(x_i) = \hat{y_i} = \omega^T x_i$, so there are:

$$\hat{y} = X\omega = X(X^T X) - 1X^T y \tag{13}$$

3. Conclusion

40% of the data was selected as the test set each time and run the model, and SPSS software was used for statistical analysis of the results, shown as the follow:

Table 1: Descriptive statistic

	N	minimun	maximum	Mean value	Standard deviation
Naive Bayes	50	.0115	.1925	.096960	.0376448
BP Neural Network	50	.0230	.2070	.126600	.0470334
SVM	50	.0150	.1765	.092350	.0417983
Linear regression	50	.3892	16.4180	4.627329	10.0327916
Logistical regression	50	.0011	.5500	.328000	.0953832

In the five algorithms of the two categories (classification algorithm, regression algorithm), we summarized the following conclusions through visual analysis and statistical indicator analysis:

1) In the classification algorithm, the naive Bayes classification has the smallest minimum value, and the support vector machine (SVM) classification has the smallest maximum value.

2) In the classification algorithm them, five 0 mean in terms of experiments, M SE mean largest B P neural network, mean that the minimum support vector machine (the SVM) classification, and support vector machines and Naive Bayes gap small, while B P mean both the neural network and the rest have

a certain difference.

3)_ In the classification algorithm them, five 0 experiments in terms of standard deviation, M SE standard deviation is greatest B P neural network classifiers, naive Bayes classifier is a minimum.

4) In the comparison of the two algorithms, the classification algorithm is superior to the regression algorithm in terms of mean, and the classification algorithm of the SVM support vector machine is the best, followed by the naive Bayes algorithm.

References

[1] Sun Jinwen, Xiao Jianguo. Research on keyword learning in text classification based on SVM [J]. Computer Science, 2006(11): 182-184.

[2] Liu Yang. Research on text classification based on support vector machines [D]. Lanzhou University of Technology, 2007.

[3] Shi Fenggui. Implementation of Chinese text corpus preprocessing module based on jieba Chinese word segmentation [J]. Computer Knowledge and Technology, 2020, v.16(14): 254-257+263.

[4] Huo Shandong. Using BP neural network to realize Chinese text classification [J]. Computer Times, 2015, No.281 (11): 58-61.

[5] Fu Yeqin, Wang Xinjian, Zheng Xiangmin. Research on tourism image based on network text analysis ——Taking Gulangyu as an example [J]. Tourism Forum, 2012, 05(4): 59-66.

[6] Hu Xiqiang. Evaluation of Tourist Satisfaction in Changsha City Tourism Destinations Based on Internet Comments [D]. Xiangtan University.