EasyChair Preprint

№ 5685

Application Research of Naive Bayes Classification Algorithm in Weather Website

Chaoning Li, Liang Chen, Shenghong Wu, Yunyin Mo and Liying Chen

June 10, 2021

Application Research of Naive Bayes Classification Algorithm in Weather Website

Chaoning Li[1],Liang Chen[1,2],Shenghong Wu[1],Yunyin Mo[1],Liying Chen[1]

1.Hainan Province Meteorological Service Center, Haikou Hainan 570100, China;

2.South China Sea Key Laboratory of Meteorological Disaster Prevention and Mitigation, Haikou Hainan 570203, China.

Abstract

The update of the weather website travel products provides tourists with a reference to the weather conditions of the destination. However, due to various reasons, the weather website travel products cannot be updated on time. It is necessary to manually monitor whether it is updated. If it is not updated, it must be manually updated manually, which undoubtedly increases the business. The burden of personnel, in the intelligent era, It need to find a solution that saves time and effort to solve this problem. The Naive Bayes algorithm is widely used because of its advantages such as high classification accuracy and simple model. For this purpose, the Naive Bayes classification prediction algorithm combined with Python crawler is used to update the forecast of tourism products on the weather website. The algorithm combines the weather website's historical update data mining in the past month to calculate the a priori probability, and then calculates the classification result based on the Python program to capture the data of the day. By recording 16 sample data sets in the future, this model is used for calculation and analysis. 15 pieces of data conform to the results of the model calculation classification, and the accuracy rate reaches 93.7%. The results show that the high accuracy of algorithm classification prediction can remind business personnel in time, better guarantee the timely update of tourism products, thereby improving the efficiency of business personnel, and providing practical application reference value for the automation of meteorological service business.

Keywords:Meteorological;Naive Bayes;python;crawler;classification prediction;weather website

0 INTRODUCTION

In the era of rapid development of information modernization, busy weather data is being processed and transmitted every day. In order to ensure the timeliness, completeness, and accuracy of weather information data, a lot of manpower needs to be spent to verify and correct the data, as well as timely transmission of the corresponding data. The data is sent to government decision-making departments, disaster prevention, mitigation and flood prevention units. The weather service for the public is mainly to upload the processed weather data to the corresponding website platform on time, so that the public can easily and quickly check the weather conditions to arrange their own life, work and travel. However, in the process of transmitting weather data to the

---

Fund Project: South China Sea Key Laboratory of Meteorological Disaster Prevention and Mitigation Open Fund (SCSF201907).

First author:Chaoning Li,(1990-), male, master, junior engineer, research direction meteorological information technology.

Corresponding author:Liang Chen, (1981-), male, senior engineer, research direction meteorological service and system development, Email: c_ray16@163.com.

website platform, due to various reasons, the data cannot be updated to the website platform in time, which requires human resources to monitor to ensure that the data is updated in time. This is undoubtedly an increase in the busy weather daily work. The workload of the salesman. How to use big data classification prediction technology to simplify the daily work of salespersons, the classification prediction algorithm of data mining obtains a classification result by training and analyzing a large amount of historical data. The naive Bayes algorithm model assumes that the attributes are independent of each other. It is best to use the naive Bayes algorithm when the attribute correlation is small. We choose the naive Bayes algorithm to affect the history of the website platform weather product data update Data factors for data analysis and forecasting.

1 PRINCIPLES OF NAIVE BAYES CLASSIFICATION ALGORITHM

1.1 Naive Bayes Method

The naive Bayes method is based on the Bayes principle and uses the knowledge of probability statistics to classify the sample data set. Due to its solid mathematical foundation, the misjudgment rate of Bayesian classification algorithm is very low. The characteristic of Bayesian method is the combination of prior probability and posterior probability, which avoids the subjective bias of using only the prior probability, and also avoids the overfitting phenomenon of using sample information alone [1]. Naive Bayes Classification (NBC) is a method based on Bayes' theorem and assuming that the feature conditions are independent of each other. First, the prior probability is calculated through the given training set, and then the largest one is obtained through calculation and comparison. Probability.

1.2 Naive Bayes Classification Process

1.2.1 Bayes Theorem

P(A|B) represents the probability of event A occurring under the premise that event B has occurred. It is called the probability of event A under event B. Its formula is as follows:

$$P(A \mid B) = \frac{P(AB)}{P(B)} \qquad (1)$$

Based on conditional probability, we can get P(B|A) through P(A|B):

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)} \qquad (2)$$

Among them, A represents the attribute set, B represents the class variable, P(B) is the prior probability, P(A|B) is the class conditional probability of the sample attribute A relative to the class variable B, P(A) is the evidence factor, P (B|A) is the posterior probability, and the Bayesian classification model expresses the posterior probability through the prior probability P(B), the class conditional probability P(A|B) and the evidence P(A).

The denominator P(A) in the above formula (2) can be decomposed into:

$$P(A) = \sum_{i=1}^{n} P(B_i)P(A \mid B_i) \qquad (3)$$

1.2.2 Naive Bayes Classification Model

The Bayesian classification method will have problems such as sparse samples, discrete samples, and complex calculations in practical applications. In response to such problems, experts and scholars have proposed the assumption of independence on the conditional probability distribution of attributes, that is, by calculating the order of the probability values of various variables It is not necessary to use a completely accurate probability value to calculate the optimal value; secondly, the dependency between attributes sometimes has the same influence on all categories, and sometimes the influence brought by this dependency can offset each other, so The application of naive Bayes classifier can often get better and more accurate results [2][3].

Naive Bayes is an algorithm that relies on the Bayes rule, assuming that the conditions of each attribute are independent; the attribute vector set D={X1, X2,..., Xn}, the class vector set C={C1, C2 }, C1 is expressed as a positive category, C2 is expressed as a negative category, as in formula (4).

$$P(c \mid d) = \frac{P(d \mid c)P(c)}{P(d)} \qquad (4)$$

For the distinction of distinguishing attribute classes, P(d) are all the same, that is, we get:

$$C = \operatorname{argmax} P(c \mid d), c \in C \qquad (5)$$

$$C = \operatorname{argmax} \frac{P(d \mid c)P(c)}{P(d)}, c \in C \qquad (6)$$

In formula (6), the denominators are all the same constant, and the comparison is negligible, that is, formula (7):

$$C = \operatorname{argmax} P(d \mid c)P(c), c \in C \qquad (7)$$

The attribute class can be expressed as:

$$C = \operatorname{argmax} P(x_1, x_2, ..., x_n \mid c)P(c), c \in C \qquad (8)$$

$$C = \operatorname{argmax} P(c)\prod_{i=1}^{n} P(x_i \mid c), c \in C \qquad (9)$$

As in formula (9), the comparison of the judgement attribute category is the comparison of the maximum value of the product of the prior probability.

2 EXPERIMENT AND ANALYSIS

2.1 Algorithm implementation

In the implementation of the naive Bayes algorithm, the classification results can be divided into whether the travel products of Hainan Weather Net are updated in time or not. Therefore, the product update results can be classified with 0, 1 classification problems. . The specific algorithm training application flowchart is as follows:

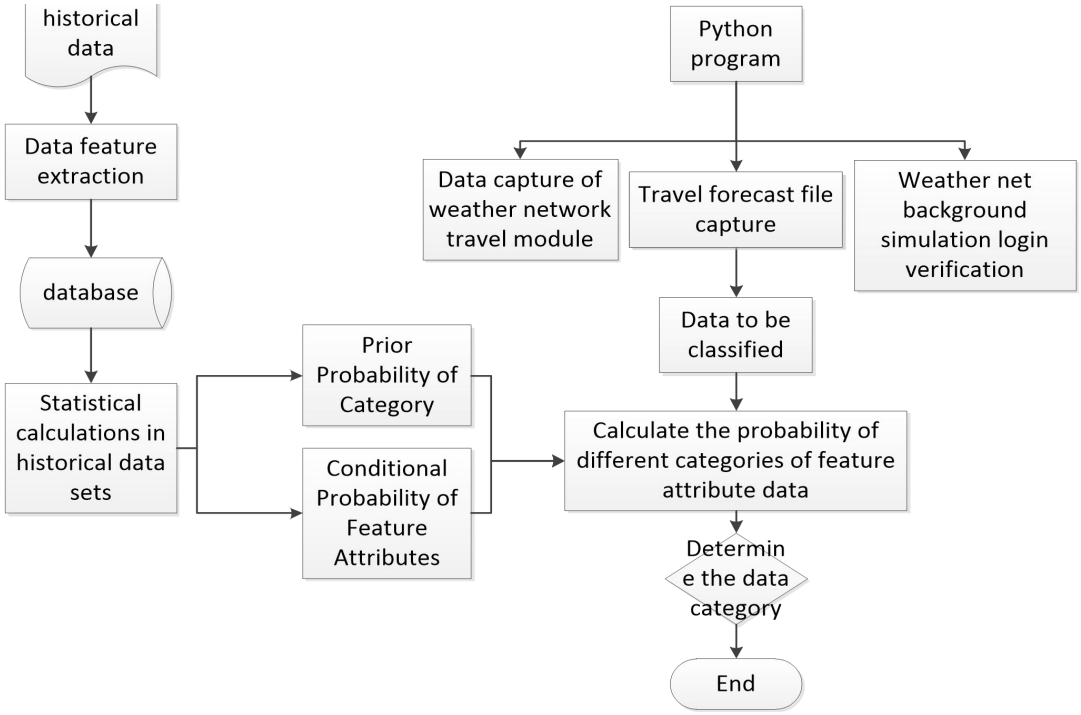

Figure 1:Naive Bayes algorithm classification process

As shown in Figure 1, during the training of the Naive Bayes algorithm, the recent historical data is first extracted into the database for processing, and the class conditional probability of each attribute and the prior probability of different categories are calculated statistically. The Python program passes three functional modules The latest data to be classified is automatically captured, and the posterior probability is calculated by the algorithm according to the currently acquired attribute feature value, and the maximum posterior probability is obtained by comparison, thereby obtaining the classification result.

2.2 Data preparation

The classification data source of this application comes from the data update record of the travel weather module of Hainan Weather Network in the past month. Data collection is carried out for the data update of the travel weather module at 19:00 every day. The training sample includes 31 data, and the attribute is a travel forecast file. The update status, the login status of the weather network

background, and the update status of the weather network travel module are shown in Table 1 below.

Table 1:31 training data sets of tourism weather module

No	Update status of travel forecast documents	Weather net background login situation	Update status of weather net tourism module
1	File generation 1	Login normal 1	Data update 1
2	File generation 1	Login exception 0	Data not update 0
3	File not generation 0	Login normal 1	Data not update 0
4	File generation 1	Login normal 1	Data update 1
5	File generation 1	Login normal 1	Data update 1
6	File generation 1	Login exception 0	Data update 1
7	File generation 1	Login normal 1	Data update 1
8	File generation 1	Login exception 0	Data not update 0
9	File generation 1	Login normal 1	Data update 1
10	File generation 1	Login exception 0	Data not update 0
11	File generation 1	Login normal 1	Data update 1
12	File generation 1	Login normal 1	Data update 1
13	File not generation 0	Login normal 1	Data not update 0
14	File generation 1	Login normal 1	Data update 1
15	File generation 1	Login exception 0	Data not update 0
16	File generation 1	Login normal 1	Data update 1
17	File generation 1	Login normal 1	Data update 1
18	File generation 1	Login normal 1	Data update 1
19	File generation 1	Login normal 1	Data not update 0
20	File not generation 0	Login exception 0	Data not update 0
21	File generation 1	Login normal 1	Data update 1
22	File generation 1	Login normal 1	Data update 1
23	File generation 1	Login normal 1	Data not update 0
24	File generation 1	Login normal 1	Data update 1
25	File generation 1	Login normal 1	Data update 1
26	File not generation 0	Login normal 1	Data not update 0
27	File generation 1	Login normal 1	Data update 1
28	File generation 1	Login normal 1	Data update 1
29	File not generation 0	Login normal 1	Data not update 0
30	File generation 1	Login normal 1	Data update 1
31	File generation 1	Login normal 1	Data update 1

Table 2: Category cj and Sample statistics of attribute xi under cj condition

Update status of travel forecast	Weather net background login	Update status of weather net

documents		situation		tourism module	
generate	not generate	normal	abnormal		
20	0	19	1	update	20
6	5	6	5	Not update	11

Table 3: Types of conditional probability P(xi|cj) and prior probability P(cj)

Update status of travel forecast documents		Weather net background login situation		Update status of weather net tourism module	
generate	not generate	normal	abnormal		
1.00	0.00	0.95	0.05	update	0.65
0.55	0.45	0.55	0.45	Not update	0.35

2.3 Application of Naive Bayes Model

Now the data obtained one day is X={file generation, abnormal login}, predict whether the travel data of that day will update the classification situation. According to Bayesian formula (7)(9):

$$P(c_{\text{update}}|X)=P(X\,|\,c_{\text{update}})P(c_{\text{update}})=\prod_{i=1}^{n}P(x_i\,|\,c_{\text{update}})*P(c_{\text{update}})$$

$$=P(x_{\text{file generation}}|c_{\text{update}})*P(x_{\text{login exception}}|c_{\text{update}})*P(c_{\text{update}})$$

$$=20/20*1/20*20/31=0.0322$$

$$P(c_{\text{not update}}|X)=P(X\,|\,c_{\text{not update}})P(c_{\text{not update}})=\prod_{i=1}^{n}P(x_i\,|\,c_{\text{not update}})*P(c_{\text{not update}})$$

$$=P(x_{\text{file generation}}|c_{\text{not update}})*P(x_{\text{login exception}}|c_{\text{not update}})*P(c_{\text{not update}})$$

$$=6/11*5/11*11/31=0.0879$$

Therefore， $P(c_{cap}\,|\,X)=\max(0.0322,0.0879)=0.0879$ ， it is predicted that the classification of tourism data for that day will not be updated.

When calculating the conditional probability of the sample, we found that when the i-th attribute value x_i is included, its conditional probability value is 0, if the attribute value of the sample to be estimated is x_i (the conditional probability value is 0) , Then the calculation result of the entire Bayesian formula will be 0. For example, we set the sample value as X={file not generated, login is normal}, and the value of $P(x_{\text{file generation}}|c_{\text{update}})$ is 0 in the calculation of the posterior probability. In order to calculate the conditional probability more accurately, Laplacian correction is introduced to solve this problem. The meaning is to add a prior distribution, corresponding to the case where the

conjugate prior parameter takes 1, and expand the DC actual observation categories. The total number of categories in the training set is represented by N; the number of possible values of the Di attribute is represented by Ni, so the revised conditional probability calculation formula is:

$$P(x_i|c) = \frac{D_{c,xi} + 1}{D_c + Ni} \qquad (10)$$

The conditional probability corrected by Laplace is:

$$P(c_{\text{update}}|X) = P(X \mid c_{\text{update}})P(c_{\text{update}}) = \prod_{i=1}^{n} P(x_i \mid c_{\text{update}}) * P(c_{\text{update}})$$

$$= P(x_{\text{file not generated}}|c_{\text{update}}) * P(x_{\text{login normal}}|c_{\text{upda te}}) * P(c_{\text{update}})$$

$$= 1/22 * 19/20 * 20/31 = 0.0278$$

$$P(c_{\text{not update}}|X) = P(X \mid c_{\text{not update}})P(c_{\text{not update}}) = \prod_{i=1}^{n} P(x_i \mid c_{\text{not update}}) * P(c_{\text{not update}})$$

$$= P(x_{\text{file not generated}}|c_{\text{not update}}) * P(x_{\text{login normal}}|c_{\text{not update}}) * P(c_{\text{not update}})$$

$$= 5/11 * 6/11 * 11/31 = 0.0879$$

Therefore, $P(c_{cap} \mid X) = \max(0.0278, 0.0879) = 0.0879$,the result of the forecast classification is that the tourism data has not been updated, which is consistent with the previous calculation and classification results.

The following table 4 compares the update status of the travel forecast products recorded daily at 19:00 for the next 16 days and the calculation and classification results of the naive Bayes algorithm as shown in Table 4 (in the table, 1 means update, normal, 0 means not updated, abnormal). 11 of the results predicted by the algorithm under the condition that the travel forecast file is updated and the weather network back-end login is normal, are consistent with the actual weather network travel module update; the travel forecast file is not updated and the weather network back-end login is normal in the case of the algorithm prediction results One item is consistent with the fact that the tourism module of the actual weather network is not updated; the tourism forecast file is not updated and the weather network background login is abnormal. The result of the algorithm prediction is consistent with the situation of the tourism module of the actual weather network is not updated; the tourism forecast file is updated and In the case of abnormal weather network background login, one of the three predicted results of the algorithm is inconsistent with the actual weather network travel module update status. Analysis found that due to a short-term network failure in the business environment that day, the weather network background login verification program judged It was abnormal. After the network returned to normal, the weather network's travel module data was updated normally, which caused the algorithm prediction result to be inconsistent with the actual update of the weather network on that day. The prediction and classification results of 15 algorithms in 16 sample data are consistent with the

actual situation, and the accuracy of model prediction reaches 93.7%.

Table 4 :Comparison of tourism product update results in the next 16 days

No	Update status of travel forecast documents	Weather net background login situation	Naive Bayes algorithm classification forecast update situation	The actual update of the weather net tourism module
1	1	1	1	1
2	1	1	1	1
3	0	1	0	0
4	1	1	1	1
5	1	0	0	0
6	1	1	1	1
7	0	0	0	0
8	1	1	1	1
9	1	1	1	1
10	1	1	1	1
11	***1***	***0***	***0***	***1***
12	1	1	1	1
13	1	1	1	1
14	1	0	0	0
15	1	1	1	1
16	1	1	1	1

2.4 Model evaluation

The Bayesian model algorithm is clear and easy to understand. The time and space overhead in the classification process is small, and the classification accuracy is high for discrete data. We conduct training and analysis by collecting data on different key factors that affect the update of weather network tourism products. The historical data is trained and calculated to get the classification result. Based on 31 training sample sets, the Naive Bayes classification model is used to classify and predict whether the weather network tourism products are updated. By recording the future 16 sample data sets, this model is used for calculation and analysis and the actual update situation is compared. The results show that 15 The results of data calculation and classification are consistent with the actual update situation, and the accuracy rate reaches 93.7%. The classification and prediction of the updates of other forecast products under the weather network, the result accuracy rate can also reach more than 90%, indicating that the model is effective. The predicted classification results through model calculation can achieve the expected results, and can more intuitively provide decision-making information to business personnel, thereby improving work efficiency and promoting the development of weather service business automation to a certain extent.

3 CONCLUSION

This paper discusses the algorithm principle of the naive Bayes model. The naive Bayes classifier is used to update the factors affecting the weather network tourism product data as attributes. The historical data of the weather network is collected as a training set for analysis, and the training data is first analyzed. Extract and delete redundant and edge irrelevant attributes, calculate their prior probability and class conditional probability, and then calculate the classification prediction of the sample to be tested, test and evaluate the model through the future 16 sample data, and the experimental data results show that 15 data predictions The result is consistent with the actual update situation, and the accuracy of the model prediction result and the actual update situation reaches 93.7%, indicating that the naive Bayes classifier can achieve better classification results. The disadvantage is that the training data and sample data are small, the model considers the factors that affect the update of the weather network tourism module is not comprehensive, and the model assumes that each attribute is independent of each other, so the classification effect is not good when the correlation between the attributes is strong. Using this model to classify and predict weather network product updates can effectively improve the work efficiency of business personnel, thereby reducing pressure on business personnel, and providing practical application reference value for weather service business automation.

REFERENCES

[1] 朱军,胡文波.贝叶斯机器学习前沿进展综述[J].计算机研究与发展,2015,52(01):16-26.

[2] 王乐慈,高世臣,林孟雄,李宗贤.基于不同概率密度估计方法的朴素贝叶斯分类器[J]. 中国矿业, 2018, 27(11):177-183.

[3] 叶进,林士敏.基于贝叶斯网络的推理在移动客户流失分析中的应用[J]. 计算机应用,2005,25(3)673-675.

[4] 韩丽娜.贝叶斯分类模型在学生成绩预测中的应用研究[J]. 计算机与数字工程, 2018, 046(010):2039-2041,2056.

[5] 王斌.基于朴素贝叶斯算法的垃圾邮件过滤系统的研究与实现[J]. 电子设计工程, 2018(17):171-174.

[6] 王峻. 基于属性相关性分析的扩展朴素贝叶斯分类器[J]. 平顶山学院学报, 2018, 33(05):70-74.

[7] 王乐慈, 高世臣, 林孟雄,等. 基于不同概率密度估计方法的朴素贝叶斯分类器[J]. 中国矿业, 2018, 27(11):177-183.

[8] 郭勋诚. 朴素贝叶斯分类算法应用研究[J]. 通讯世界, 2019, 26(01):247-248.

[9] 刁海军，尹钊. 一种基于朴素贝叶斯分类算法的数据预测[J]. 电大理工, 2018, 277(04):5-7.

[10] Hermanto D T , Ziaurrahman M , Bianto M A , et al. Twitter Social Media Sentiment Analysis in Tourist Destinations Using Algorithms Naive Bayes Classifier[J]. Journal of Physics Conference Series,2018,1140(1):012037.

[11] Sourav Kunal, Arijit Saha, Aman Varma, Vivek Tiwari.Textual Dissection of Live Twitter Reviews using Naive Bayes[J]. Procedia Computer Science, 2018,132:307-313.

[12] Findawati Y , Taurusta C , Widiaty I , et al. Teacher Performance Assesment Application using Naive Bayes Classifier Method[J]. IOP Conference Series Materials Science and Engineering, 2018, 384:012047.

[13] M Irfan,W Uriawan,O T Kurahman,M A Ramdhani,I A Dahlia. Comparison of Naive Bayes and K-Nearest Neighbor methods to predict divorce issues[J]. IOP Conference Series: Materials Science and Engineering,2018,434(1):012047.

[14] Deng X , Hou L , Wang F . Web advertisement detection using Naive Bayes[J]. Journal of Physics Conference, 2019, 1187(4):042023.

[15] Rahmadani S , Dongoran A , Zarlis M , et al. Comparison of Naive Bayes and Decision Tree on Feature Selection Using Genetic Algorithm for Classification Problem[J]. Journal of Physics: Conference Series, 2018, 978:012087.