INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 03, March 2024, pp: 410-412

NAIVE BAYES CLASSIFICATION ALGORITHM FOR TRAFFIC RISK MANAGEMENT

Sharmitha T[1]

[1]M. Sc Department Of Computer Science, Fatima College, Madurai.

ABSTRACT

Naive Bayesian classification algorithm is widely used in big data analysis and other fields because of its simple and fast algorithm structure. Aiming at the shortcomings of the naive Bayes classification algorithm, this paper uses feature weighting and Laplace calibration to improve it, and obtains the improved naive Bayes classification algorithm. Through numerical simulation, it is found that when the sample size is large, the accuracy of the improved naive Bayes classification algorithm is more than 99%, and it is very stable; when the sample attribute is less than 400 and the number of categories is less than 24, the accuracy of the improved naive Bayes classification algorithm is more than 95%. Through empirical research, it is found that the improved naive Bayes classification algorithm can greatly improve the correct rate of discrimination analysis from 49.5 to 92%. Through robustness analysis, the improved naive Bayes classification algorithm has higher accuracy.

1. INTRODUCTION

There are many ways to construct classifiers, such as the Bayesian method, decision tree method, case-based learning method, artificial neural network method, support vector machine method, genetic algorithm method, rough set method, fuzzy set method, and so on. Among them, the Bayesian method is becoming one of the most attractive focuses of many methods because of its unique form of uncertain knowledge expression, rich probability expression ability, and the incremental learning characteristics of integrating prior knowle dge. Naive Bayesian classification algorithm (NBC) is one of the classic Bayesian classification algorithms, which has a simple algorithm structure and high computational efficiency. One advantage of a naive Bayes classifier is that it only needs to estimate the necessary parameters (mean and variance of variables) based on a small amount of training data. Due to the assumption of independent variables, only the method of estimating each variable is needed, and the whole covariance matrix is not needed. Based on the above excellent properties, the naive Bayesian classification algorithm has a wide range of applications, such as clinical medicine, telecommunications, artificial intelligence, linguistics, gene technology, precision instruments, and other fields. At the same time, naive Bayes classification algorithm has strong compatibility, which can form more powerful algorithms when combined with other methods, such as double-weighted fuzzy gamma naive Bayes classification, fuzzy association naive Bayes classification complex network naive Bayes classification

4 3, March 2024 March

2. MODEL

2.1 Bayes theory

Bayesian theory is an important part of subjective Bayesian inductive theory. Bayesian decision-making is to estimate the subjective probability of some unknown states under incomplete information, then modify the occurrence probability with the Bayesian formula, and finally make the optimal decision by using the expected value and modified probability.

2.2 Naive Bayesian classification

Naive Bayes classification is to use the maximum likelihood estimation principle to classify the sample into the most likely category .

2.3 Feature-weighted naive Bayes classification algorithm

It is generally believed that the more an attribute feature appears, the more important it is, and the greater the corresponding weight in the model.

2.4 Laplace calibration

There maybe a potential problem in formula when the number of training samples is small and the number of attributes is large, the training samples are not enough to cover so many attributes, so the number of samples of Aj =xj maybe 0, and the whole category conditional probability P(Ci | X) will be equal to 0. If this happens frequently, it is impossible to achieve accurate classification.

Therefore, it is very fragile to simply use the proportion to estimate the category conditional probability. The way to solve the problem is to use Laplacian calibration (Laplacian estimation), which can completely solve the problem that the category conditional probability is 0. At the sametime, this slight change does not change sample'sclassification.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 03, March 2024, pp: 410-412

3. RESULT AND DISCUSS

3.1.1 Impact of sample size

Suppose that the number of attributes is k = 5, the number of values of each attribute is q = 5, and the number of categories is C = 2. Ten thousand samples are randomly selected from the standard normal distribution N (0,1), and the accuracy of the model is tested by gradually increasing the sample size.

It can be seen from that when the sample size is small, the accuracy rate of discrimination analysis fluctuates greatly, but with the increase of the sample size, the fluctuation gradually becomes smaller, and the overall trend tends to be stable, with the accuracy reaching more than 99%.

3.1.2 Impact of sample attributes

In the standard normal distribution N (0,1), 1000 samples are randomly selected, assuming that the number of categories is C = 2, and the number of values of each attribute is q = 5.

As can be seen from when the sample attribute is less than 400, the accuracy is above 95%, which remains at a high level, and the trend is stable; when the sample attribute is between 400 and 600, the accuracy drops precipitously; when the sample attribute is more than 600, the accuracy drops to about 50%, and the overall trend is stable.

3.1.3 Impact of category

In the standard normal distribution N (0,1), randomly select 1000.

samples, assuming that the number of attributes is m = 5, and each attribute value is q = 5.

when the number of categories is small (< 24), the accuracy remains above 95%, and the trend is stable; when the number of categories is large (24 – 60), the accuracy fluctuates greatly, and the stability is poor; when the number of categories further increases (> 60), the accuracy rate quickly drops to zero.

3.2.1 Data collection and processing

Based on the random sampling of traffic violation cases in a city from January 2019 to December 2019, a total of 115,482 samples were selected, including 30,340 samples with complete data. There are two kinds of traffic violations: speeding and running red lights. In this paper, speeding without running red lights is set as the first category, running red lights without speeding is set as the second category, speeding with running red lights is set as the third category, respectively, assigned to 0, 1, and 2; there are five reasons for traffic violations: whether driving with a license, gender, vehicle type, driving age, and weather. Among them, unlicensed driving is 0, licensed driving is 1; female driver is 0, male driver is 1; small car is 0, medium bus is 1, and large truck is 2; driving experience less than 1 year is 0, driving experience between 1 and 3 years is 1, and driving experience more than 3 years is 2. It is 0 in sunny days, 1 in rainy days, 2 in foggy days, and 3 in snowy days.

3.2.2 Import naive Bayes classification algorithm

Using the improved naive Bayes classification algorithm for analysis this paper can draw the following conclusions: in the first, second, and third classes of traffic violations, 5097, 17,311, and 5501 samples are correct; the correct rate is 69.8%, 98.8%, and 99.7%; and the overall correct rate is 92.0%, which shows that the improved naive Bayes classification algorithm has a very high correct rate, especially in the second and third category.

3.2.3 Naive Bayes classification algorithm

In order to compare with the improved naive Bayesian classification algorithm, this paper uses the original naive Bayesian classification algorithm.

3.2.4 Robustness test

In order to continue to compare the efficiency of the improved naive Bayesian classification algorithm, this paper uses logistic regression to compare. Because all variables are discrete selection variables and there are three values for dependent variables, multivariate logistic regression.

4. DISCUSSION

Through numerical simulation, we found that, when the sample size is small, the accuracy rate of discrimination analysis of improved naive Bayesian classification algorithm fluctuates greatly, but with the increase of the sample size, the fluctuation gradually becomes smaller, and the overall trend tends to be stable, with the accuracy reaching more than 99%; when the sample attribute is less than 400, the accuracy is above 95%, which remains at a high level, and the trend is stable; when the sample attribute is between 400 and 600, the accuracy drops precipitously; when the sample attribute is more than 600, the accuracy drops to about 50%, and the overall trend is stable; when the number of categories is small (< 24), the accuracy remains above 95%, and the trend is stable;

when the number of categories is large (24 – 60), the accuracy fluctuates greatly, and the stability is poor; when the number of categories further increases (> 60), the accuracy rate quickly drops to zero.

Through empirical analysis, this paper found that, using the improved naïve Bayes classification algorithm for analysis, in the first, second, and third classes of traffic violations, 5097, 17311, and 5501 samples are correct; the correct rate is 69.8%, 98.8%, and 99.7%; and the overall correct rate is 92.0%; using the naive Bayes classification algorithm, the accuracy of the first, second, and third classes is 52.8%, 41.5%, 69.7%, respectively, and the overall accuracy of the discriminatory analysis is 49.4%. All the indexes are far lower than the results of the improved naive Bayesian classification algorithm.

Through robustness analysis, we find that, using multiple logistic main effect regression, the correct rates of the first, second, and third classes are 37.7%, 90.0%, and 93.5%, respectively, and the overall correct rate is 78.1%; using the multiple logistic total factor regression, the correct rates of the first, second, and third classes are 45.9%, 91.9%, and 94.5%, respectively, and the overall correct rate is 81.3%. Therefore, the multiple logistic total factor regression has a higher accuracy than the main effect regression, but it is still far lower than the improved naive Bayes classification algorithm.

5. CONCULSION

In view of the shortcomings of the naive Bayesian classification algorithm, this paper improves the algorithm by using the feature weighting and Laplace calibration and obtains the improved naive Bayesian classification algorithm. The results show that when the sample size is large, the improved naive Bayesian classification algorithm has a high accuracy of 99% and is very stable. When the sample attribute is less than 400, the accuracy rate is over 95%, and when the sample attribute is greater than 600, the accuracy rate of discrimination decreases to about 50%, and the trend is stable; when the number of categories is less than 24, the accuracy rate of discrimination analysis is maintained at least 95%, and the trend is stable; when the number is more than 60, the accuracy of discrimination is reduced to zero rapidly. Through empirical research, it is found that, compared with the original naive Bayesian classification algorithm, the improved naive Bayesian classification algorithm greatly improves the accuracy of discrimination analysis from 49.5 to 92%. Compared with the multivariate logistic main effect regression and multivariate logistic total factor regression, the improved naive Bayesian classification algorithm has higher accuracy.

6. REFERENCES

[1] H. Shakir, H. Rasheed, T.M.R. Khan, Radiomic featureselection for lung cancer classifiers [J]. J. Intell. Fuzzy Syst. 38(5), 1–9 (2020)

[2] B. Ehsani-Moghaddam, J.A. Queenan, J. Mackenzie, et al., Mucopolysaccharidosis type II detection by naïve Bayes classifier: an example of patient classification for a rare disease using electronic medical records from the Canadian Primary Care Sentinel Surveillance Network [J]. PLoS One 13(12), 251– 265 (2018)

[3] H. Zhang, L. Ding, Y. Zou, et al., Predicting drug-induced liver injury in human with naïve Bayes classifier approach [J]. J. Comput. Aided Mol. Des. 30(10), 889–898 (2016)

[4] S.C. Chu, T.K. Dao, J.S. Pan, et al., Identifying correctness data scheme for aggregating data in cluster heads of wireless sensor network based on naive Bayes classification [J]. EURASIP J. Wirel. Commun. Netw. 20(1), 963 –982 (2020)

[5] R. Rajalakshmi, C. Aravindan, A Naive Bayes approach for URL classification with supervised featureselection and rejection framework [J]. Comput. Intell. 34(1), 363 –396 (2018)

[6] W. Xu, L. Jiang, An attribute value frequency-based instance weighting filter for naive Bayes [J]. Journal of Experimental & Theoretical Artificial Intelligence 31(4), 225–236 (2019)