EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

Improved naive Bayes classification algorithm for traffic risk management

Hong Chen[1†], Songhua Hu[2†], Rui Hua[3*] and Xiuju Zhao[4]

* Correspondence: earlyhua@hbust.edu.cn
[†]Hong Chen and Songhua Hu contributed equally to this work and should be considered co-first authors.
[3]School of Mathematic and Statistic, Hubei University of Science and Technology, Xianning, China
Full list of author information is available at the end of the article

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.

Keywords: Improved naive Bayesian classification algorithm, Discrimination analysis, Multivariate logistic regression, Feature weighted, Traffic risk management

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 knowledge. 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 [1–3], telecommunications [4, 5], artificial intelligence [6], linguistics [7, 8], gene technology [9], precision

instruments [10], 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 [11], fuzzy association naive Bayes classification [12], complex network naive Bayes classification [13], feature selection naive Bayes classification [14], tree augmented naive Bayes classification [15], etc.

At the same time, the study found that with the promotion of urbanization, the improvement of transportation facilities, and the popularity of family cars, "road killers" are more and more, and the problem of traffic risk is becoming increasingly prominent. Therefore, before the drivers implement the driving behavior, how to carry out the risk management and implement the classified early warning in advance and realize the source management has become a hot topic in the industry and academia. From the perspective of research fields, the research of traffic risk management has involved many fields of traffic risk, including traffic accidents [16], water safety [17], extreme weather [18], etc. In terms of research methods, scholars have actively used a large number of different methods to classify, manage, analyze, and predict traffic risks, including signal control [19], spatiotemporal analysis [20], etc. In particular, with the maturity of big data technology and the improvement of database, AI-related methods are more and more used in the field of traffic risk management, including support vector machine [21], RBF neural network [22], deep learning [23], fuzzy rule base [24], etc.

From the above analysis, it is found that the existing research has the following shortcomings:

First, naive Bayes classification has an obvious defect: it is based on the assumption of attribute independence, but in most cases, this assumption does not conform to the reality [25]. At the same time, this assumption makes the redundant, irrelevant, interactive, and noise-contaminated features have the same status as the really important features, which eventually leads to the reduction of classification accuracy.

Second, there are few researches on driver's risk. The existing literature on the risk of traffic scenes is more common, but the risk of drivers is less. The driver is the most important factor leading to traffic accidents, and more than 90% of traffic accidents are related to driver behavior. Therefore, it has great research prospects to establish relevant risk management models for drivers, especially for some characteristics of drivers (such as gender, driving age, personality, etc.). The purpose of this study is to carry out risk research on the personal characteristics of drivers and realize source management.

Third, a machine learning algorithm is rarely used in the field of traffic risk management. With the rapid growth of traffic data and the improvement of its computing power, the machine learning algorithm has become a potentially important means to deal with traffic risk management [26].

Based on the above shortcomings, this paper improves the naive Bayes classification algorithm by combining feature weighting and Laplace calibration. The improved naive Bayes classification algorithm can overcome the above shortcomings and make full use of the information of the training set to greatly improve the accuracy of the original naive Bayes classification algorithm. At the same time, the improved naive Bayes classification algorithm is applied to the scene of traffic risk management to effectively predict and classify the driver's driving risk and finally implement effective risk management.

The rest of the paper is organized as follows: the improved naive Bayes classification algorithm is established in section 2. In Section 3, numerical simulation is used to verify the accuracy of the improved naive Bayes classification algorithm. At the same time, this method is applied to big data of traffic risk for robustness analysis. There are some discussion in the end. Conclusions are given in section 4.

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.

Ω is a complete set, C_1, C_2, $\cdots$, $C_n \in \Omega$, C_i denotes the ith category, $P(C_i) > 0$, $i = 1, 2, \cdots, n$. Any two categories are incompatible with each other, and $\bigcup_{i=1}^{n} C_i = \Omega$. For any set$X$, if $P(X) > 0$, so

$$P(C_i|\mathrm{X}) = \frac{P(\mathrm{X}|C_i)P(C_i)}{\sum_{i=1}^{n} P(\mathrm{X}|C_i)P(C_i)} \tag{1}$$

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 [27], that is:

$$P(C_i|\mathrm{X}) = Max\{P(C_1|\mathrm{X}), P(C_2|\mathrm{X}), \cdots P(C_n|\mathrm{X})\} \tag{2}$$

Suppose the sample $X = (A_1, A_2, \cdots, A_k)$ is an attribute vector, A_j is the jth attribute which may have several different values x_j.

Naive Bayes classification considers that the attributes are independent of each other, so

$$P(\mathrm{X}|C_i) = \prod_{j=1}^{k} P(A_j = \mathrm{x_j}|C_i) \tag{3}$$

Substituting formula (3) into formula (1), that is:

$$P(C_i|\mathrm{X}) = \frac{\prod_{j=1}^{k} P(A_j = \mathrm{x_j}|C_i)P(C_i)}{P(\mathrm{X})} \tag{4}$$

Let $\frac{1}{P(\mathrm{X})} = \alpha(> 0)$, that is

$$P(C_i|\mathrm{X}) = \alpha \prod_{j=1}^{k} P(A_j = \mathrm{x_j}|C_i)P(C_i) \tag{5}$$

In sample set D, $N(D)$ is the total number of samples, $N(C_i)$ is the number of samples of C_i, $N(C = C_i, A_j = x_j)$ is the number of samples when attribute A_j is x_j in C_i, that is

$$P(C_i) = \frac{N(C_i)}{N(D)} \tag{6}$$

$$P(A_j = \mathrm{x_j} | \mathrm{C} = C_i) = \frac{N(C = C_i, A_j = x_j)}{N(C_i)} \tag{7}$$

Substituting formula (6) and formula (7) into formula (5), then,

$$P(C_i|\mathrm{X}) = \alpha \prod_{j=1}^{k} \frac{N(C = C_i, A_j = x_j)}{N(C_i)} \cdot \frac{N(C_i)}{N(D)} \tag{8}$$

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 [28, 29]. Therefore, the weight coefficient of the feature is set as

$$w_j = \frac{N(A_j = x_j)}{N(D)}$$

w_j represents the proportion of the number of samples in the total number of samples when attribute A_j is x_j. Formula (8) can be improved to:

$$P(C_i|X) = \alpha \prod_{j=1}^{k} w_j \frac{N(C = C_i, A_j = x_j)}{N(C_i)} \cdot \frac{N(C_i)}{N(D)}$$

$$= \alpha \prod_{j=1}^{k} \frac{N(A_j = x_j)}{N(D)} \cdot \frac{N(C = C_i, A_j = x_j)}{N(C_i)} \cdot \frac{N(C_i)}{N(D)} \tag{9}$$

2.4 Laplace calibration

There may be a potential problem in formula (9): 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 A_j=x_j may be 0, and the whole category conditional probability $P(C_i| X)$ will be equal to 0 [30, 31]. 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 same time, this slight change does not change sample's classification.

The specific method is to improve formula (7) as follows:

$$P(A_j = x_j | C = C_i) = \frac{N(C = C_i, A_j = x_j) + 1}{N(C_i) + q_j} \tag{10}$$

$$w_j = \frac{N(A_j = x_j) + 1}{N(D) + q_j} \tag{11}$$

q_j represents the number of possible values of attribute A_j.

By substituting formula (10) and formula (11) into formula (9), we can get

$$P(C_i|X) = \alpha \frac{N(C_i)}{N(D)} \prod_{j=1}^{k} \frac{N(A_j = x_j) + 1}{N(D) + q_j} \cdot \frac{N(C = C_i, A_j = x_j) + 1}{N(C_i) + q_j} \; i = 1, 2, \cdots, n \tag{12}$$

3 Result and discussion

3.1 Numerical simulation

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 Fig. 1 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%.

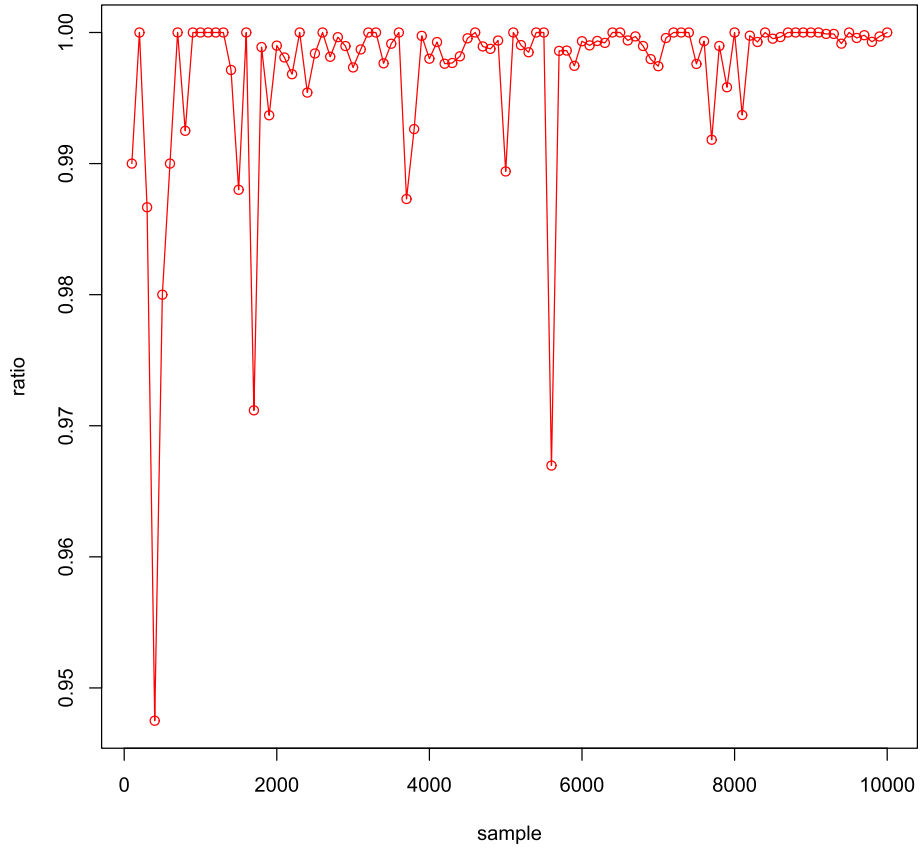

Fig. 1 The impact of sample size on the accuracy of the model

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 Fig. 2, 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.

As can be seen from Fig. 3, 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 Improved Bayesian classification algorithm for traffic risk management

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

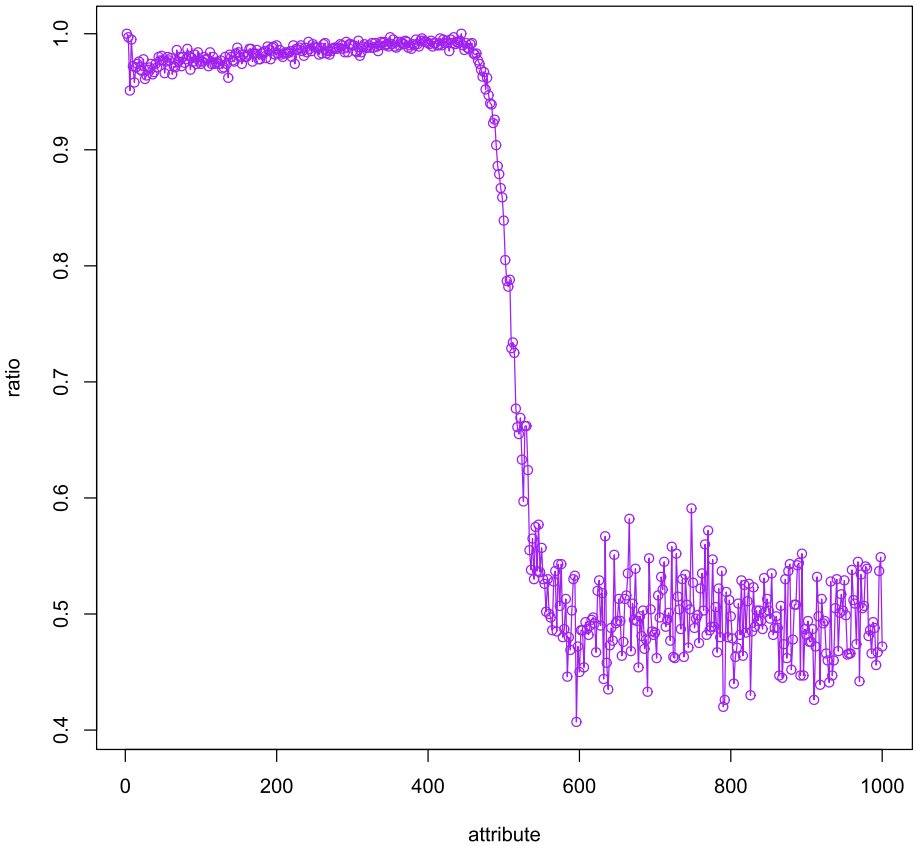

Fig. 2 The impact of sample attributes on the accuracy of the model

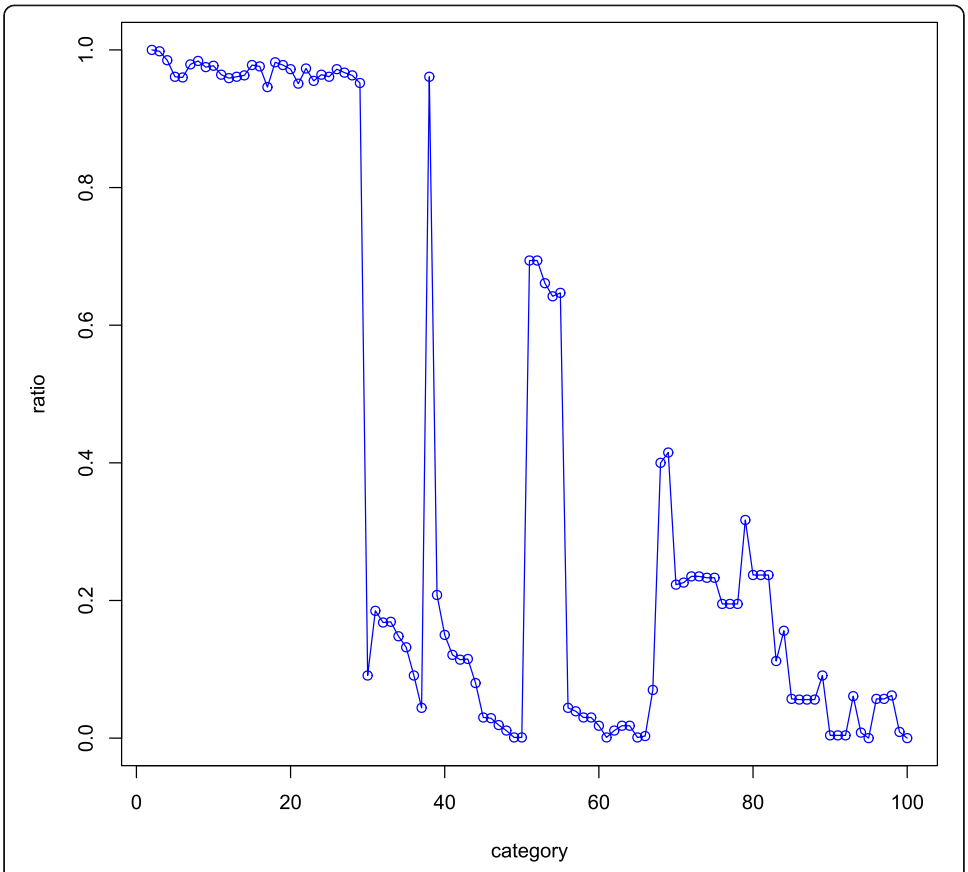

Fig. 3 The impact of category on the accuracy of the model

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.

According to the above statistics (Table 1), red light running accounts for nearly 60% of violations, and 75% of speeding drivers will also run red lights. Twenty percent of the violations are caused by unlicensed drivers, which shows that unlicensed driving is a very dangerous driving behavior. Men account for more than 60% of violations, indicating that there is no reason for discrimination against female drivers. From the perspective of driving experience, there is a reverse relationship between violation and driving experience. The smaller the driving experience, the more violation. From the perspective of weather, nearly 60% of the violations occurred in sunny days, and bad weather is not the main reason for violations.

Table 1 Descriptive statistics of data

	First class (= 0)	Second class (= 1)	Third class (= 2)	Fourth class (= 3)
Traffic violations	7298	17,524	5518	
Licensed driving	5518	24,822		
Gender	11,473	18,868		
Vehicle type	24,556	3037	2747	
Driving age	12,406	13,089	4845	
Weather	17,195	10,210	1423	1512

3.2.2 Improved naive Bayes classification algorithm

Using the improved naive Bayes classification algorithm for analysis (Table 2), 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 to carry out the back analysis, and the result is as follows (Table 3):

From the above results, 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. Therefore, the efficiency of the improved naive Bayesian classification algorithm is greatly improved.

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 is adopted [32, 33].

a. Multiple logistic main effect regression

In this section, a multiple logistic main effect model was used for regression analysis [34], and the following results were obtained (Table 4):

Table 2 Discriminatory analysis of the improved naive Bayes classification algorithm

Actual	Predictive			
	1	2	3	Accuracy
1	5097	2201	0	69.8%
2	213	17,311	0	98.8%
3	9	8	5501	99.7%
Ratio				92.0%

Table 3 Discriminant analysis of naive Bayes classification algorithm

Actual	Predictive			
	1	2	3	Accuracy
1	3855	3443	0	52.8%
2	6075	7276	4173	41.5%
3	199	1475	3844	69.7%
Ratio				49.5%

According to the results of the above table, 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%. It can be seen that the correct rate of multiple logistic main effect regression is much lower than the improved naive Bayes classification algorithm.

b. Multiple logistic total factor regression

The multivariate logistic main effect regression is only considered in the whole factor regression, and the interaction effect of each factor is not considered. Therefore, this section continues to analyze the multiple logistic total factor regression [35], and the analysis results are as follows (Table 5):

It can be seen from the above table that in 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.

3.3 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.

Table 4 Discriminant analysis of multiple logistic main effect regression

Actual	Predictive			
	1	2	3	Accuracy
1	2753	4545	0	37.7%
2	1745	15779	0	90.0%
3	161	197	5160	93.5%
Ratio				78.1%

Table 5 Discriminant analysis of multiple logistic total factor regression

Actual	Predictive			
	1	2	3	Accuracy
1	3353	3945	0	45.9%
2	1419	16,105	0	91.9%
3	153	149	5216	94.5%
Ratio				81.3%

Through empirical analysis, this paper found that, using the improved naive 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.

Through the research of this paper, it is found that the improved naive Bayes algorithm has greatly improved the original algorithm, but unfortunately, there are some limitations in this paper, such as unable to consider the interaction of features, sample size, category and other factors, and so on.

4 Main conclusions

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.

Abbreviations

NBC: Naive Bayesian classification algorithm; RBF: Radial basis function

Acknowledgements
The authors would like to thank HBUST for this support and anyone who support this paper to be published.

Authors' contributions
All authors made contributions in the discussions and analyses. RH and SHH contributed equally to this work and should be considered co-first authors. All authors read and approved the final manuscript.

Funding
This work is funded by the 2019 philosophy and social science research project of the Department of Education of Hubei (19Q175) and the 2019 Doctoral start-up fund project of HBUST (BK202025).

Availability of data and materials
Existing datasets cannot be shared for confidentiality.

Declarations

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]School of Clinical Medicine, Hubei University of Science and Technology, Xianning, China. [2]School of Statistics and Data Science, Nankai University, Tianjin, China. [3]School of Mathematic and Statistic, Hubei University of Science and Technology, Xianning, China. [4]School of Mathematic and Statistic, Hubei University of Arts and Science, Xiangyang, China.

Received: 31 March 2021 Accepted: 8 June 2021
Published online: 22 June 2021

References

1. H. Shakir, H. Rasheed, T.M.R. Khan, Radiomic feature selection 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 feature selection 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)
7. V. Jafarizadeh, A. Keshavarzi, T. De Rikvand, Efficient cluster head selection using Naïve Bayes classifier for wireless sensor networks [J]. Wirel. Netw **23**(3), 1–7 (2016)
8. V.L. Jong, P.W. Novianti, K.C.B. Roes, M.J.C. Eijkemans, Selecting a classification function for class prediction with gene expression data. *Bioinformatics*. **32**(12), 1814–1822 (2016)
9. O. Maruyama, Heterodimeric protein complex identification by naïve Bayes classifiers [J]. Bmc Bioinformatics **14**(1), 347 (2013)
10. J. Karandikar, T. Mcleay, S. Turner, et al., Tool wear monitoring using naïve Bayes classifiers [J]. Int. J. Adv. Manuf. Technol. **77**(9-12), 1613–1626 (2015)
11. Moraes, A double weighted fuzzy gamma naive Bayes classifier [J]. Journal Of Intelligent & Fuzzy Systems **38**(1), 577–588 (2020)
12. Banchhor, FCNB: fuzzy correlative naive Bayes classifier with Map Reduce framework for big data classification [J]. J. Intell. Syst. **29**(1), 994–1005 (2020)
13. Jiang et al., Fast artificial bee colony algorithm with complex network and naive Bayes classifier for supply chain network management [J]. Soft. Comput. **23**(24), 13321–13337 (2019)
14. G.R. Nitta, B.Y. Rao, T. Sravani, N. Ramakrishiah, M. Balaanand, LASSO-based feature selection and naive Bayes classifier for crime prediction and its type [J]. SOCA **13**(3), 187–197 (2019)
15. A. Meehan, C.D. Campos, Averaged extended tree augmented naive classifier [J]. Entropy **17**(7), 5085–5100 (2015)
16. J. Zhang, T. Shi, Spatial analysis of traffic accidents based on WaveCluster and vehicle communication system data [J]. EURASIP J. Wirel. Commun. Netw. **32**(1), 278–403 (2019)
17. M.A. Jun, D. Reckhow, Y. Xie, Drinking water safety: science, technology, engineering and policy [J]. Frontiers of Environmental Science & Engineering **9**(1), 1124–1142 (2015)
18. P. Levi Kangas, S.S. Michaeli De, Transport system management under extreme weather risks: views to project appraisal, asset value protection and risk-aware system management [J]. Nat. Hazards **72**(1), 263–286 (2014)
19. B.C. Ezell, R.M. Robinson, P. Foytik, et al., Cyber risk to transportation, industrial control systems, and traffic signal controllers [J]. Environment Systems & Decisions **33**(4), 508–516 (2013)

20. D. Pavlyuk, Feature selection and extraction in spatiotemporal traffic forecasting: a systematic literature review [J]. Eur. Transp. Res. Rev. **25**(6), 215–226 (2019)
21. Y. Zhu, Y. Zheng, Traffic identification and traffic analysis based on support vector machine [J]. Neural Comput. & Applic. **32**(7), 1903–1911 (2020)
22. D. Shi, R. Li, Traffic identification method based on multiple probabilistic neural network model [J]. Neural Comput. Applic. **31**(1), 1–15 (2017)
23. S. Khatri, H. Vachhani, S. Shah, et al., Machine learning models and techniques for VANET based traffic management: implementation issues and challenges [J]. Peer-to-Peer Networking and Applications **45**(3), 618–634 (2020)
24. S. Nemet, D. Kukolj, G. Ostojic, et al., Aggregation framework for TSK fuzzy and association rules: interpretability improvement on a traffic accidents case [J]. Appl. Intell. **49**(11), 3909–3922 (2019)
25. T.T. Wong, Alternative prior assumptions for improving the performance of naïve Bayesian classifiers [J]. Data Min. Knowl. Disc. **18**(2), 183–213 (2009)
26. X. Hu, X. Zhang, N. Lovrich, Public perceptions of police behavior during traffic stops: logistic regression and machine learning approaches compared [J]. Journal of Computational Social Science **3**, 1–26 (2020)
27. D. Heckerman, Bayesian networks for data mining. Data mining and knowledge discovery [J]. Data Min. Knowl. Disc. **1**(1), 79–119 (1997)
28. T. Sun, S. Ding, P. Li, et al., A comparative study of neural-network feature weighting [J]. Artif. Intell. Rev. **21**(4), 167–176 (2019)
29. D. Singh, B. Singh, Hybridization of feature selection and feature weighting for high dimensional data [J]. Appl. Intell. **45**(1), 1023–1046 (2018)
30. A.V. Cardona, M.T. Vilhena, B. Bodmann, et al., An improvement of the double discrete ordinate approximation solution by Laplace technique for radiative-transfer problems without azimuthal symmetry and high degree of anisotropy [J]. J. Eng. Math. **67**(3), 193–204 (2010)
31. M. Cassia, P. Shah, E. Bruun, A novel calibration method for phase-locked loops [J]. Analog Integr. Circ. Sig. Process **42**(1), 77–84 (2004)
32. L.V. Maanen, D. KaTsImpokis, A.V. Campen, Correction to: Fast and slow errors: logistic regression to identify patterns in accuracy–response time relationships [J]. Behav. Res. Methods **51**(6), 1471–1493 (2019)
33. M.R. Zkale, S. Lemeshow, R. Sturdivant, Logistic regression diagnostics in ridge regression [J]. Comput. Stat. **33**(2), 563–593 (2018)
34. D. Boning, Multinomial logistic regression algorithm [J]. Annals of the Institute of Statal Mathematics **44**(1), 197–200 (1992)
35. H.H. Huang, X. Tu, J. Yang, Comparing logistic regression, support vector machines, and permanental classification methods in predicting hypertension [J]. BMC Proc. **28**(S1), 96–102 (2014)

Publisher's Note