Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Improved Naive Bayes Classification Algorithm for Traffic Risk Management

Rui Hua (earlyhua@hbust.edu.cn)
Hubei University Of Science and Technology https://orcid.org/0000-0001-7450-4293
songhua hu
Nankai University
Hong Chen
Hubei University Of Science and Technology
Xiuju Zhao
Hubei University of Arts and Science

Research Article

Keywords: Improved Naive Bayesian Classification Algorithm, Discrimination Analysis, Multivariate Logistic Regression, Feature weighted, Traffic Risk Management

Posted Date: April 7th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-355037/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Improved Naive Bayes Classification Algorithm for Traffic Risk Management

Hua Rui[1]
School of Mathematic and Statistic, Hubei University of Science and Technology, Xianning, China

Songhua Hu[1]
School of Statistics and Data Science, Nankai University Tianjin, China

Hong Chen*
School of Clinical Medicine, Hubei University of Science and Technology Xianning, China

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 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.

KEYWORD: Improved Naive Bayesian Classification Algorithm; Discrimination Analysis; Multivariate Logistic Regression; Feature weighted; Traffic Risk Management

1. Introduction

Naive Bayesian classification algorithm (NBC) has a simple algorithm structure and high computational efficiency, which is one of the classic Bayesian classification algorithms. It 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

*Correspondence: Rui Hua, earlyhua@hbust.edu.cn; Hong Chen ,chenhong2020@hbust.edu.cn
[1] These authors contributed equally to this work and should be considered co-first authors.

network naive Bayes classification [13], feature selection naive Bayes classification [14], tree augmented naive Bayes classification [15], etc.

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 [16]. 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.

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.

The rest of the paper is organized as follows. In Section 2, the improved naive Bayes classification algorithm is established. In Section 3, Numerical simulation is used to verify the accuracy of improved naive Bayes classification algorithm. In Section 4, this method is applied to big data of traffic risk for robustness analysis. Finally, conclusions are given in Section 5.

2. The establishment of the model

(1) Bayes theory

Ω is a complete set，C_1，C_2,L，$C_n \in \Omega$，C_i denotes the i th category，$P(C_i) > 0$，

$i = 1, 2, \mathrm{L}\ , n$，Any two categories are incompatible with each other，and $\bigcup_{i=1}^{n} C_i = \Omega$. For any

X ,if $P(X) > 0$，so

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

(2) Naive Bayesian classification

Naive Bayes classification is to use the maximum likelihood estimation principle to classify the sample into the most likely category [17], that is:

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

Suppose the sample $X = (A_1, A_2, \mathrm{L}\ , A_k)$ is an attribute vector，A_j is the j th attribute which may have several different values x_j.

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

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

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

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

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

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

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

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

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

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

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

(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 [18-19]. 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.The formula (8) can be improved to：

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

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

(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 $P(C_i \mid X)$ will be equal to 0 [20-21]. 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：

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

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

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

$$P(C_i \mid X) = \alpha \frac{N(C_i)}{N(D)} \prod_{j=1}^{k} \frac{N(A_j = x_j) + 1}{N(D) + q_j} \cdot \frac{N(C = C_i, A_j = x_j) + 1}{N(C_i) + q_j} \quad i = 1, 2, \mathrm{L}\ , n \quad (12)$$

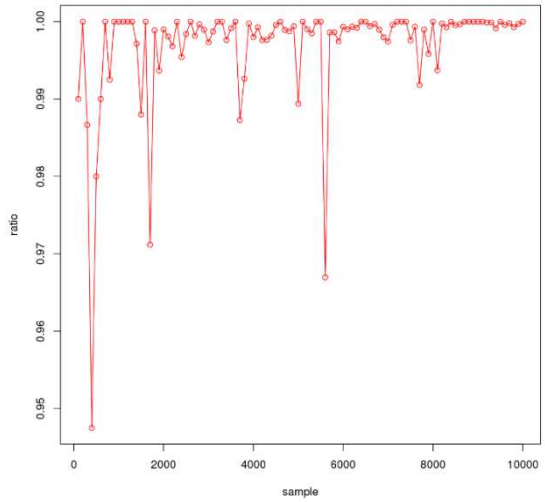
3. Numerical simulation

(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.

Figure 1 The impact of sample size on the accuracy of the model

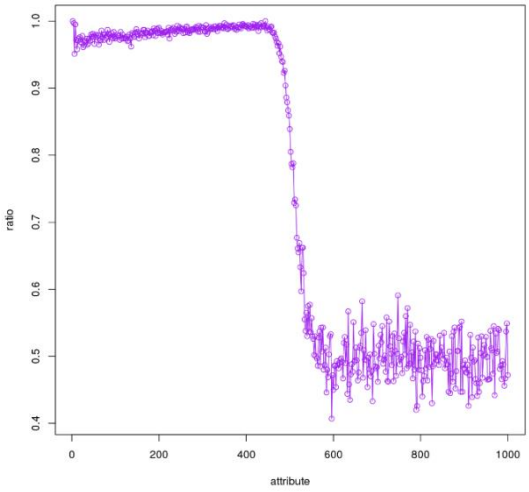


It can be seen from Figure 1 that when the sample size is small, the accuracy rate of discrimination analysis fluctuates greatly, but with the increase of the sample size, the fluctuation gradually becomes smaller, and the overall trend tends to be stable, with the accuracy reaching more than 99%.

(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.

Figure 2 The impact of sample attributes on model accuracy

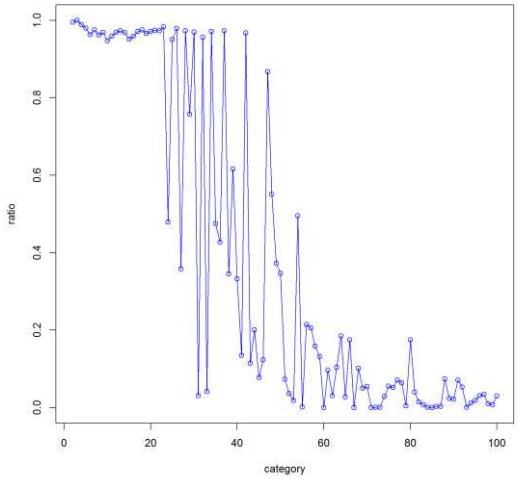


As can be seen from Figure 2, when the sample attribute is less than 400, the accuracy is above 95%, which remains at a high level, and the trend is stable; when the sample attribute is between 400-600, the accuracy drops precipitously; when the sample attribute is more than 600, the accuracy drops to about 50%, and the overall trend is stable.

(3) 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.

Figure 3 The impact of category on model accuracy



As can be seen from Figure 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.

4. Improved Bayesian classification algorithm for traffic risk management

(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 115482 samples were selected, including 30340 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, 2; there are five reasons for traffic violations: whether

driving with a license, gender, vehicle type, driving age, 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, large truck is 2; driving experience less than one year is 0, driving experience between one year and three years is 1, driving experience more than three years is 2. It is 0 in sunny days, 1 in rainy days, 2 in foggy days and 3 in snowy days.

Table 1 Descriptive statistics of data

	First class (= 0)	second class (= 1)	third class （=2）	fourth class （=3）
Traffic violations	7298	17524	5518	
Licensed driving	5518	24822		
Gender	11473	18868		
Vehicle type	24556	3037	2747	
Driving age	12406	13089	4845	
Weather	17195	10210	1423	1512

According to the above statistics, red light running accounts for nearly 60% of violations, and 75% of speeding drivers will also run red lights. 20% 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.

(2) Improved naive Bayes classification algorithm

Table 2 Discriminatory analysis of improved naive Bayes classification algorithm

Actual	Predictive			
	1	2	3	Accuracy
1	5097	2201	0	69. 8%

2	213	17311	0	98.8%
3	9	8	5501	99.7%
Ratio				92.0%

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, 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%, which shows that the improved naive Bayes classification algorithm has a very high correct rate, especially in the second and third category.

(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, the result as follows:

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%

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.

(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 [22-23].

a. Multiple logistic main effect regression

In this section, multiple logistic main effect model was used for regression analysis [24], and the following results were obtained：

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%

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%, 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 [25], and the analysis results are as follows:

Table 5 Discriminant analysis of multiple logistic total factor regression

Actual	Predictive			
	1	2	3	Accuracy
1	3353	3945	0	45.9%
2	1419	16105	0	91.9%
3	153	149	5216	94.5%
Ratio				81.3%

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%, 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. Main conclusions

In view of the shortcomings of 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

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, analyses. Rui Hua and SongHua Hu were

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 2019 philosophy and social science research project of Department of Education of Hubei (19Q175) and 2019 Doctoral start-up fund project of HBUST (BK202025).

Availability of data and materials

Existing datasets cannot be shared for confidentiality.

Consent for publication

not applicable.

Competing interests

The authors declare that they have no competing interests.

Reference

[1]Shakir H, Rasheed H, Khan T M R. Radiomic feature selection for lung cancer classifiers [J]. Journal of Intelligent and Fuzzy Systems, 2020, 38(5):1-9.

[2] Ehsani-Moghaddam B , Queenan J A , Mackenzie J , 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, 2018, 13(12):251-265.

[3]Zhang H, Ding L, Zou Y, et al. Predicting drug-induced liver injury in human with Nave Bayes classifier approach [J]. Journal of Computer-Aided Molecular Design, 2016, 30(10):889-898.

[4]Chu S C , Dao T K , Pan J S , et al. Identifying correctness data scheme for aggregating data in cluster heads of wireless sensor network based on naive Bayes classification[J]. EURASIP Journal on Wireless Communications and Networking, 2020, 20(1):963-982.

[5] Rajalakshmi R, Aravindan C. A Naive Bayes approach for URL classification with supervised feature selection and rejection framework [J]. Computational Intelligence, 2018, 34(1):363-396.

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

[7]Poornima, N, Saleena, et al. Multi-modal features and correlation incorporated Naive Bayes classifier for a semantic-enriched lecture video retrieval system [J]. The imaging science journal, 2018.66(5): 263-277

[8]Shakil A M , Md. S , Masud R M , et al. Robustification of Nave Bayes Classifier and Its Application for Microarray Gene Expression Data Analysis[J]. BioMed Research International, 2017, 2017:1-17.

[9]Maruyama O. Heterodimeric protein complex identification by naïve Bayes classifiers [J]. Bmc Bioinformatics, 2013, 14(1):347.

[10] Karandikar J, Mcleay T, Turner S, et al. Tool wear monitoring using nave Bayes classifiers [J]. International Journal of Advanced Manufacturing Technology, 2015, 77(9-12):1613-1626.

[11] Moraes. A double weighted fuzzy gamma naive Bayes classifier [J]. Journal Of Intelligent & Fuzzy Systems, 2020, 38(1):577-588.

[12] Banchhor. FCNB: Fuzzy Correlative Naive Bayes Classifier with Map Reduce Framework for Big Data Classification [J]. Journal of Intelligent Systems, 2020, 29(1):994-1005

[13] Jiang et al. Fast artificial bee colony algorithm with complex network and naive Bayes classifier for supply chain network management [J]. Soft Computing, 2019, 23(24):13321-13337.

[14] Nitta. LASSO-based feature selection and naive Bayes classifier for crime prediction and its type [J]. Service Oriented Computing and Applications.2019, 13(3), 187-197

[15] Chen. A Classifier Learning Method Based on Tree-Augmented Naive Bayes [J] Journal of Electronics & Information, 2019, 41 (8): 2001-2008

[16] Wong T T. Alternative prior assumptions for improving the performance of nave Bayesian classifiers [J]. Data Mining & Knowledge Discovery, 2009, 18(2):183-213.

[17] Heckerman. Bayesian networks for data mining. Data mining and Knowledge Discovery

[J]. Data Mining & Knowledge Discovery, 1997, 1(1):79-119.

[18] Sun T, Ding S, Li P, et al. A comparative study of neural-network feature weighting [J]. Artificial Intelligence Review, 2019,21(4):167-176.

[19] Wettschereck D, Aha D W, Mohri T. A Review and Empirical Evaluation of Feature Weighting Methods for a Class of Lazy Learning Algorithms [J]. Artificial Intelligence Review, 2000, 11(1):412-421.

[20] Cardona A V, Vilhena M T, Bodmann B, 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]. Journal of Engineering Mathematics, 2010, 67(3):193-204.

[21] Cassia M, Shah P, Bruun E. A Novel Calibration Method for Phase-Locked Loops [J]. Analog Integrated Circuits & Signal Processing, 2004, 42(1):77-84.

[22] Maanen L V, KaTsImpokis D, Campen AV. Correction to: Fast and slow errors: Logistic regression to identify patterns in accuracy–response time relationships [J]. Behavior Research Methods, 2019, 51(6):1471-1493.

[23] Zkale M R, Lemeshow S, Sturdivant R. Logistic regression diagnostics in ridge regression[J]. Computational Statistics, 2018, 33(2):563-593.

[24] Boning D. Multinomial logistic regression algorithm [J]. Annals of the Institute of Statal Mathematics, 1992, 44(1):197-200.

[25] Huang H H, Tu X, Yang J. Comparing logistic regression, support vector machines, and permanental classification methods in predicting hypertension [J]. Bmc Proceedings, 2014, 28(S1):96-102.

Figures

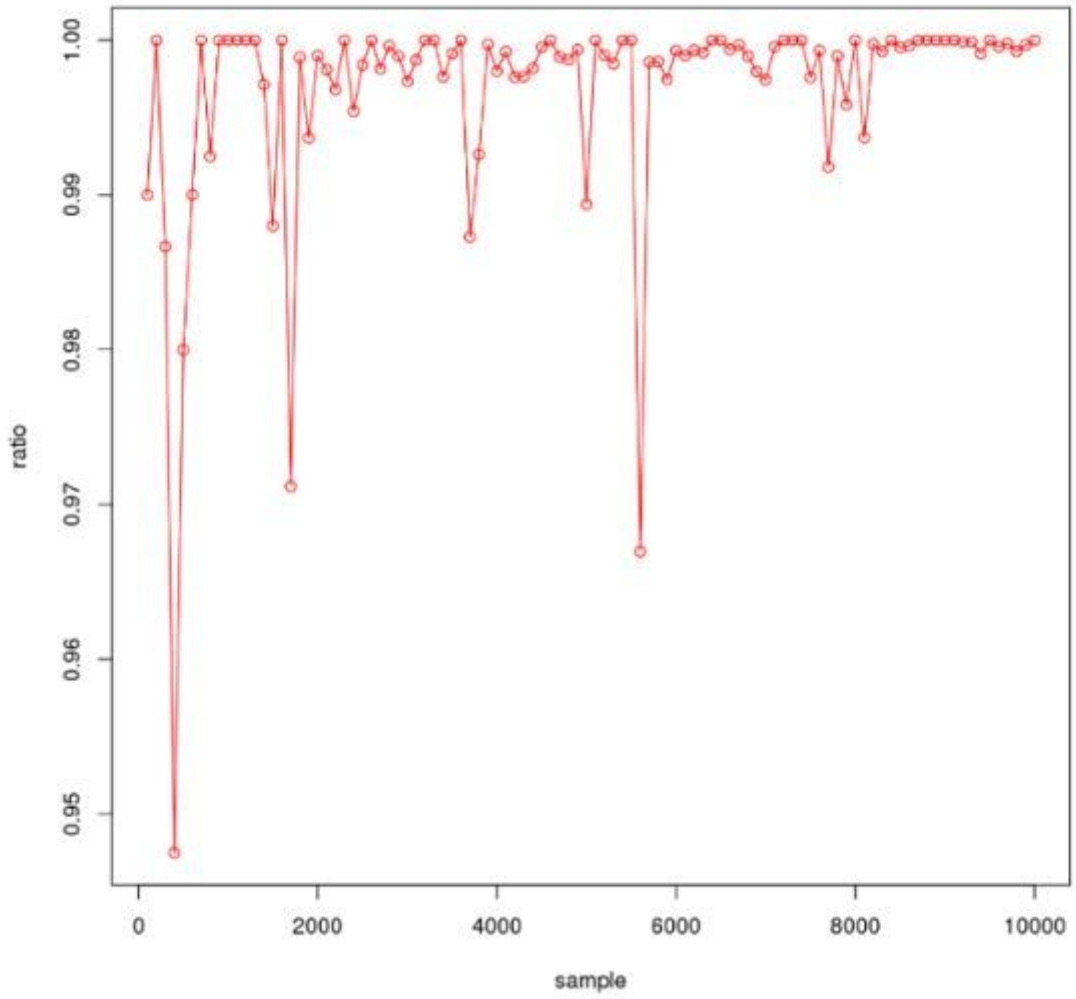


Figure 1

The impact of sample size on the accuracy of the model

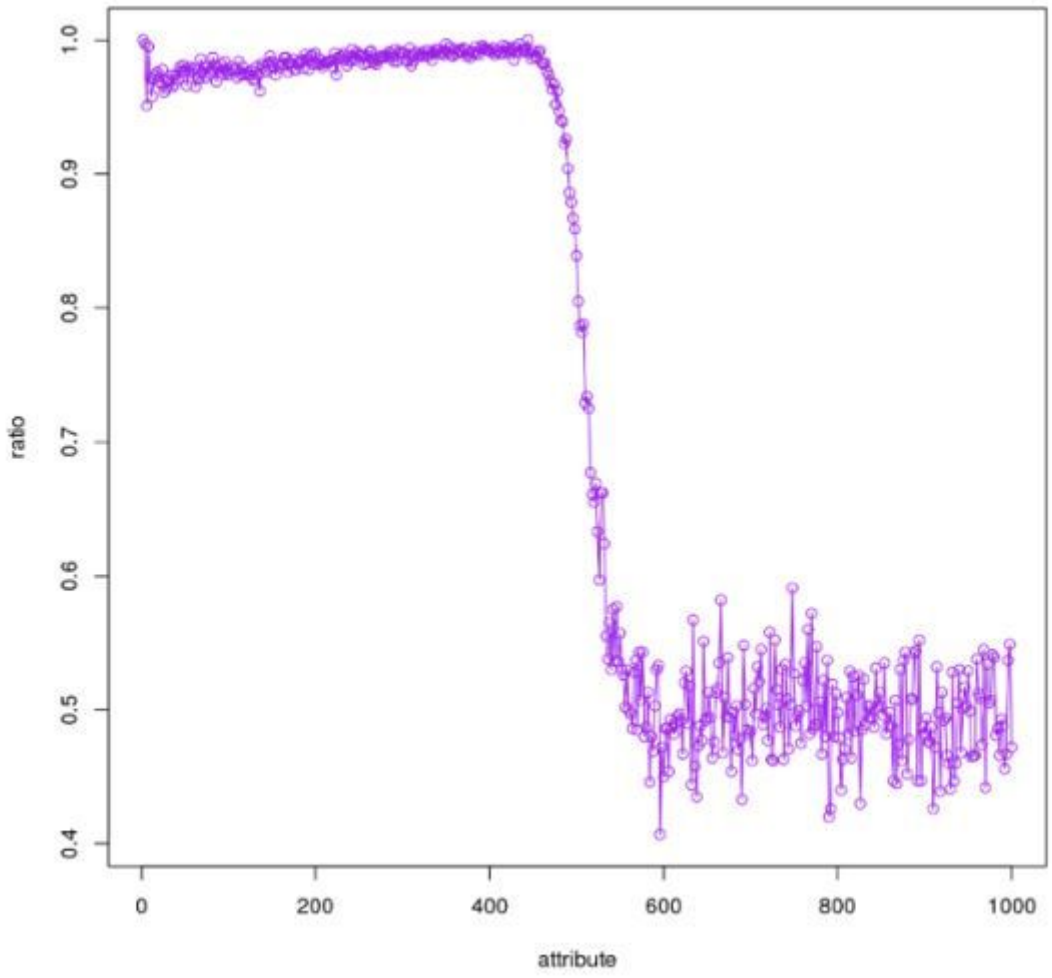


Figure 2

The impact of sample attributes on model accuracy

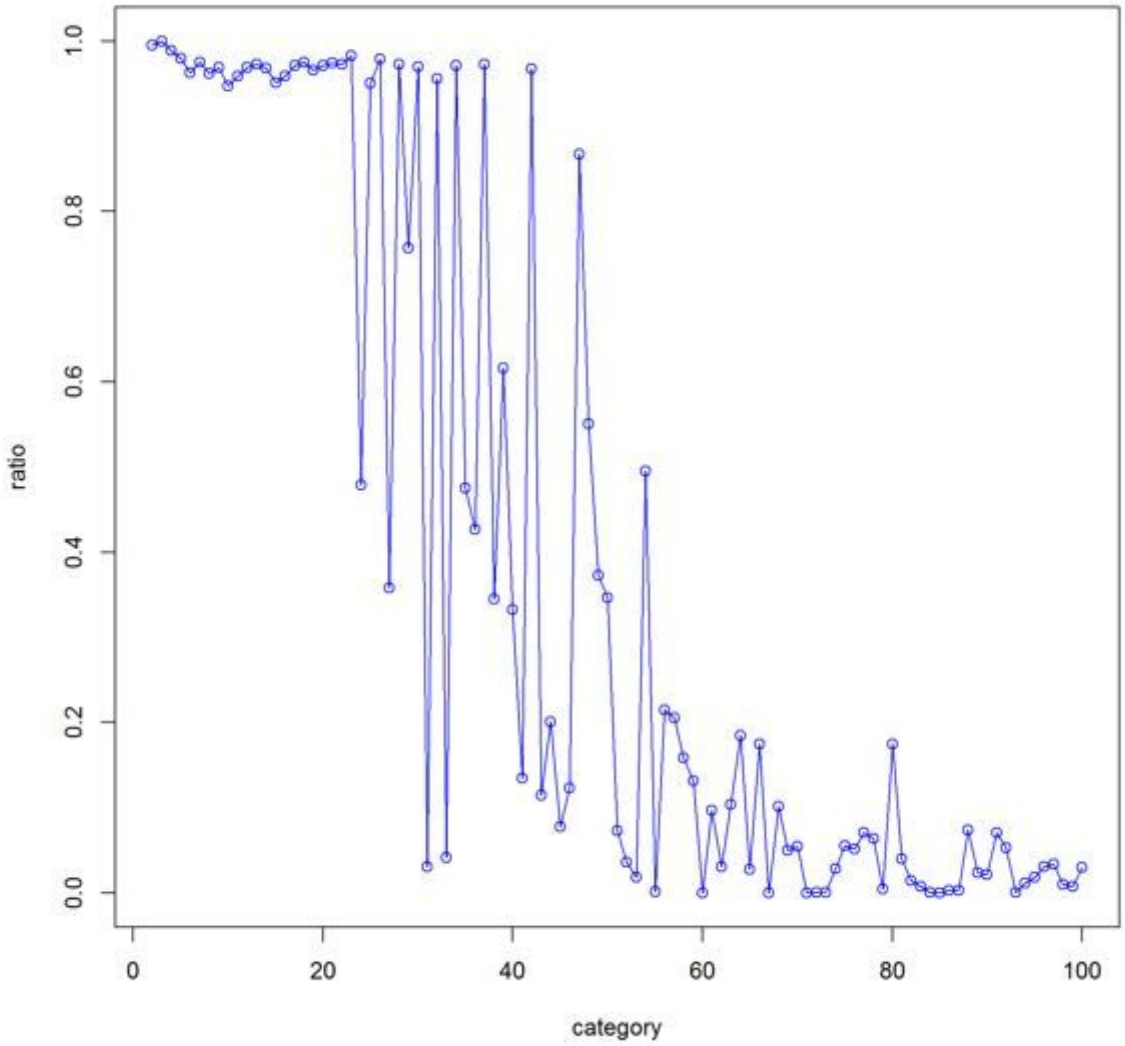


Figure 3

The impact of category on model accuracy