A robust model-free Bayesian classifier

Zirui Chen[1]

[1]*Department of Statistics and Data Science, School of Economics, Xiamen University*

This manuscript was compile on June 1, 2024

Abstract

Many algorithms have been proposed for classification tasks in machine learning. One of the simplest methods, the Naive Bayes classifier (NBC), is often found to provide good performance despite its fundamental assumptions (of independence and a normal distribution of the variables) potentially being violated. Although relationships between attributes can be analyzed using Bayesian networks (BN), the network structure and parameters need to be trained. Research has shown that learning Bayesian networks is NP-hard, and each continuous attribute is assigned prior probability distributions in NBC and BN. If the actual probability distribution of the sample does not align with the prior probability distribution, the classification results are poor. The nearest neighbor (NN) strategy can obtain information about the sample space and joint probability density. The model-free Bayesian classifier (MFBC), proposed by some researchers, can handle discrete or continuous attributes even in the absence of prior probability distributions for the attributes. It does not require the establishment of a network structure to obtain a unified computational framework for the attributes. However, when tested using UCI datasets, the MFBC model exhibited instability in handling high-dimensional data, with significant fluctuations in prediction accuracy. Its performance on some datasets was even inferior to that of the canonical Naive Bayes classifier.Therefore, we considered using Principal Component Analysis (PCA) to preprocess the data and employing cross-validation to select the number of principal components. Additionally, we estimated the prior probabilities using the concept of M-estimation, which significantly enhanced the robustness of the model. Based on this, we proposed the robust model-free Bayesian classifier (RMFBC). We tested our new algorithm on several UCI datasets and found that our algorithm outperformed the Naive Bayes classifier in all four cases and outperformed the MFBC in two cases. Our conclusion is that when encountering datasets with non-normal distributions and high-dimensional attributes with high correlations, our method provides a competitive alternative to both MFBC and the Naive Bayes classifier.

Keywords: *Naive Bayes, Classification, Model-free, Nearest Neighbor, UCI datasets, Robustness*

E-mail address: 15220212202842@stu.xmu.edu.com.
DOI: https://github.com/RayChen200318

1. Introduction

The initial Bayesian classifier, known as the Naive Bayes classifier (NBC), has been widely applied in many fields. However, NBC assumes that all attributes of the sample are independent and does not consider the dependency information between attributes. When certain attributes in the dataset are interdependent, the traditional Bayesian classifier becomes unsuitable.

Despite the seemingly unreasonable assumption of conditional independence, numerous studies have shown that this assumption is not as impractical as initially believed[2]. Even though almost no real datasets meet the conditional independence assumption, the Naive Bayes classifier still performs well, sometimes even outperforming other classification tools[9]. Many vital studies have addressed the issues of dependent attributes and the assumption of attribute independence. Related work can be broadly divided into five main categories[8]:

1. *Structure extension*. Extending the structure of naive Bayes to represent the dependencies among attributes.
2. *Feature selection*. Selecting an attribute subset from the whole space of attributes.
3. *Attribute weighting*. Assigning different weights to attributes in building naive Bayes.
4. *Local learning*. Employing the principle of local learning to build a local naive Bayes.
5. *Data expansion*. Expanding training data and building a naive Bayes on the expanded training data.

The nearest neighbor (NN) rule was first proposed by Fix and Hodges in 1989 and was one of the simplest and most popular pattern classification methods [1], [11]. Moreover, the NN method is shown to be asymptotically optimal when the categories do not overlap[6]. As a result, more and more research has developed algorithms based on attractive NN methods. For real-life datasets, the actual probability distribution of the dataset to be classified is usually unknown. However, Bayesian classifier methods require some prior knowledge to obtain the probability distribution. Therefore, there is a need for a technique that can obtain the probability distribution with a finite number of samples. Sequential Monte Carlo (SMC) methods are a set of simulation-based methods that are both convenient and attractive in computing posterior distributions [4]. In order to determine the relationship between attributes, Geng et al. proposed a simple, unified and effective method to build a probability estimator using the NN method[11].

The rest of this paper is organized as follows. In Section 2, we introduced the Naive Bayes classifier algorithm. In Section 3, we discussed the nearest neighbor strategy and the probability estimator based on Geng et al.'s work[11]. In Section 4, we presented the robust model-free Bayesian classifier algorithm. In the following Section 5, we empirically compare our RMFBC algorithm with other classifiers using UCI datasets and provide the final conclusions.

2. The naive Bayesian classifier

In this section, we introduce the basic assumptions of naive Bayesian classifiers and give a concrete form of the model.

We first provide and discuss the assumptions of the naive Bayesian classifier,

Assumption 1. *For an observation X with r attributes $X = (x_1, x_2, \dots, x_r)^T$, the attributes are independent with each other given the classification variable C , which is defined on the space $\mathcal{C} = \{c_1, c_2, \dots, c_q\}$.*

Assumption 1 gives the conditional independence assumption for naïve Bayesian classifiers. The second assumption is the normality assumption.

Assumption 2. *For an observation X with n attributes $X = (x_1, x_2, \dots, x_r)^T$, it follows $X \sim N_r(\mu, 0)$.*

Assumption 2 states the normality assumption for continuous attributes. Then under assumption 1, the posterior probability of variable C given an instance X is obtained by Eq.(1):

$$P(C|X) = \frac{P(C)P(X|C)}{P(X)} \propto P(C)\prod_{i=1}^{r} P(x_i|C). \tag{1}$$

Then we can derive the naive Bayesian classifier by Eq.(2):

$$C = \arg\max_{C\in\mathcal{C}} P(C)\prod_{i=1}^{r} P(x_i|C). \tag{2}$$

Then, we will introduce a probability estimator based on the nearest neighbor strategy.

3. The probability estimator using limited samples

3.1. The nearest neighbor strategy

Supposed that the sample size is N, with dimensionality r, $X = \begin{pmatrix} x_{11} & \dots & x_{1r} \\ \vdots & \ddots & \vdots \\ x_{N1} & \dots & x_{Nr} \end{pmatrix} = (X_1, \dots, X_N)^T$, and a corresponding class label $Y = (y_1, \dots, y_N)^T$ where $y_k \in \mathcal{C} = \{c_1, \dots, c_q\}$. The nearest neighbor X_{near} to an arbitrary sample X_i under the same class is obtained by Eq.(3):

$$X_{near} = argmin_{X_j, j\neq i \text{ and } y_j = y_i} d(X_i, X_j) \tag{3}$$

where $d(X_i, X_j)$ is the Euclidean distance between X_i adnd X_j, which is obtained by Eq.(4):

$$d(X_i, X_j) = \sqrt{\sum_{k=1}^{r}(x_{jk} - x_{ik})^2}. \tag{4}$$

Then, by applying the NN rule to find the nearest neighbors, probability estimates in the multidimensional space can be obtained.

3.2. Estimating the approximate probability distribution via the NN rule

Each sample in the given N samples under a certain class is viewed as a point on the high-dimensional space. The probability $\varphi(\tau) = P(d \geq \tau)$ represents the likelihood that the distance d between the sample X_k and its nearest neighbor X_{near} is at least τ. Therefore, $\varphi(0) = P(d \geq 0) = 1$ and $\varphi(\infty) = P(d \geq \infty) = 0$.$\varphi(\tau)$ decrease with increasing τ. The decreased speed of $\varphi(\tau)$ depends on the other $N-1$ samples, which are distributed on the superficial area enclosed by the r -dimensional spherical centred at X_k. As the distance τ increases by $\Delta\tau$, the decrease of $\varphi(\tau)$ is proportional to either dimension of X_k, with the interval $\Delta\tau$ for any other $N-1$ samples. The variety of $\varphi(\tau)$ depends on several factors: the number of sides, the other $N-1$ samples, the probability of the selected sample X_k, and $\varphi(\tau)$. Moreover, there are relationships, such as $\mathrm{d}\varphi/\mathrm{d}\tau \propto (N-1)$, $\mathrm{d}\varphi/\mathrm{d}\tau \propto p(X_k)$ and $\mathrm{d}\varphi/\mathrm{d}\tau \propto \varphi(\tau)$. Therefore, $\mathrm{d}\varphi/\mathrm{d}\tau$ is obtained through Eq.(5):

$$\frac{\mathrm{d}\varphi}{\mathrm{d}\tau} \approx -\mathcal{K}_\gamma \times \tau^{(r-1)} \times (N-1) \times P(X_k) \times \varphi(\tau). \tag{5}$$

where $\mathcal{K}_\gamma = r\pi^{r/2}/\Gamma(\frac{r}{2}+1)$. By solving Eq.(5), the probability under the current class can be obtained as shown by Eq.(6):

$$P(X_k) = \frac{2^{-\left(r\log_2(\tau+1)+\log_2(\frac{\mathcal{K}_\gamma(N-1)}{r})+\frac{\gamma}{\ln(2)}\right)}}{Z}. \tag{6}$$

where Z is a normalization term, which is the sum of the probabilities for all class types, Y is the Euler–Mascheroni constant, and $\Gamma(\cdot)$ is the Gamma function.

In conclusion, when the probability distribution of data samples is unknown, joint probability values can be obtained through such probability estimator. This also offers a new improvement strategy for algorithms like NBC, replacing marginal probabilities with joint probabilities[11].

4. Robust model-free Bayesian classifier

This section will provide a detailed introduction to the RMFBC algorithm. Research indicates that using PCA for data preprocessing significantly enhances classification accuracy and model robustness [7]. We have adopted this approach. First, we use PCA to preprocess the data, then embed the previously mentioned probability estimator into the NBC, removing the assumptions of attribute independence and the prior probability distribution of continuous attributes in the samples[11]. Additionally, we employ Laplace smoothing to estimate the prior probabilities of the classes, which further contributes to the robustness of the model[12].

Given N samples with dimensionality r, $X = \begin{pmatrix} x_{11} & \dots & x_{1r} \\ \vdots & \ddots & \vdots \\ x_{N1} & \dots & x_{Nr} \end{pmatrix} = (X_1, \dots, X_N)^T$, and a corresponding class label $Y = (y_1, \dots, y_N)^T$ where $y_k \in \mathcal{C} = \{c_1, \dots, c_q\}$. The class label of a certain sample X_k can be predicted by Eq.(7):

$$y_k = \arg\max_{C\in\mathcal{C}} P(C)P(X_k). \tag{7}$$

The prior probability $P(C)$ is estimated using the Laplace smoothing estimation as follow:

$$P(C) = \frac{F(C)+1}{N+q}. \tag{8}$$

where $F(\cdot)$ is the frequency with the class appears in the training data.

In summary, the pseudo-code of the RMFBC algorithm is presented in Algorithm 1.

Algorithm 1 The RMFBC Algorithm

Input: training samples $\{X^{(\text{train})}, Y^{(\text{train})}\}$ and testing samples $\{X^{(\text{test})}, Y^{(\text{test})}\}$
Output: class type O
N is the number of samples $X^{(\text{train})}$
r is dimensions of samples $X^{(\text{train})}$
q is the number of the classes $\mathcal{C}$
OBTAIN class labels $C = \{c_1, c_2, \dots, c_q\}$ of $X^{(\text{train})}$
Ψ is a list, $\Psi\{k\}$, $(k = 1, 2, 3, \dots, q)$ is a set
for $i \leftarrow 1$ **to** *row*(train) **do**
  $X_i^{(\text{train})}$ add in $\Psi\{k\}$ if $Y_i^{(\text{train})}$ equals to c_k
end
for $i \leftarrow 1$ **to** q **do**
  $N_i \leftarrow \text{Length}(\Psi\{i\})$
  $\phi_i \leftarrow \frac{N_i+1}{N^{(\text{train})}+q}$
end
for $i \leftarrow 1$ **to** *row*(test) **do**
  for $k \leftarrow 1$ **to** q **do**
    $\tau \leftarrow \text{find\_NN}(X_i^{(\text{test})}, \Psi\{k\})$
    $P(i,k) \leftarrow 2^{-(r\log_2(\tau+1)+\log_2(\frac{\mathcal{K}_\gamma(N_k-1)}{r})+\frac{\gamma}{\ln 2})}$
  end
  $Z \leftarrow \sum_{k=1}^{q} P(i,k)$
  $B(i,k) \leftarrow \frac{P(i,k)\phi_k}{Z}$
  $\theta \leftarrow \arg\max_k(B(i,k))$
  $O_i \leftarrow c_\theta$
end

Table 1. Description of data sets used in the experiment

Dataset	Samples	Attributes	Classes	Missing data	Continuous data	Numeric data
Iris	150	4	3	N	Y	Y
Car	1728	5	4	Y	N	N
Drybean	13611	16	7	N	Y	Y
Glass	214	10	6	N	Y	Y
Ionosphere	351	33	2	N	N	Y
Letter	20000	16	26	N	N	Y
Redwinequality	1599	11	6	N	Y	Y
Segment	2310	19	7	N	N	Y
Wine	178	13	3	Y	Y	Y

Table 2. The detailed experimental results on accuracy and standard deviation

Dataset	NN	kNN	SVM	NBC	HNB	MFBC	RMFBC
Iris	0.9714 ±0.0356	0.9714 ±0.0356	0.9810 ±0.0325	0.9524 ±0.0325	0.9438 ±0.0236	0.9714 ±0.0356	0.9238 ±0.0600
Car	0.6511 ±0.0305	0.6956 ±0.0361	0.9358 ±0.0160	0.2117 ±0.0286	0.9053±0.0201	0.3229 ±0.0344	0.3621 ±0.0344
Drybean	0.9056 ±0.0052	0.9197 ±0.0046	0.9262 ±0.0071	0.8971 ±0.0083	0.8293 ±0.0896	0.9056 ±0.0052	0.9104 ±0.0057
Glass	0.9776 ±0.0528	0.9680 ±0.0561	1.0000 ±0.0000	0.7972 ±0.1038	0.5933 ±0.0833	0.9776 ±0.0528	0.9776 ±0.0528
Ionos-	0.9103 ±0.0525	0.8906 ±0.0419	0.9000 ±0.0394	0.9244 ±0.0416	0.9208 ±0.0424	0.9103 ±0.0525	0.9295 ±0.0500
Letter	0.7619 ±0.0016	0.7590 ±0.0024	0.8575 ±0.0042	0.6491 ±0.0082	0.8231±0.0174	0.9609 ±0.0015	0.9119 ±0.0039
Redwin-	0.5922 ±0.1802	0.5719 ±0.0171	0.5912 ±0.0167	0.5683 ±0.0327	0.4785 ±0.0606	0.5922 ±0.1802	0.6470 ±0.0298
Segm-	0.9685 ±0.0082	0.9604 ±0.0068	0.9685 ±0.0105	0.8071 ±0.0100	0.9472 ±0.0142	0.9685 ±0.0082	0.9685 ±0.0082
Wine	0.9564 ±0.0541	0.9626 ±0.0395	0.9564 ±0.0541	0.9753 ±0.0293	0.9608 ±0.0246	0.9564 ±0.0541	0.9626 ±0.0395

5. Experiments and results

We ran our experiments on 9 UCI data sets[13]. They covers a wide range of data sources, data dimensions, and features, which are described in Table 1. In our experiments, we adopted the following data preprocessing steps:

1. *Replacing missing attribute values.* We used K-Nearest Neighbors Imputation to fill in missing values based on the K nearest data points[10].
2. *Converting categorical attribute values.* We converted categorical variables into dummy variables.
3. *Removing useless attributes* We removed variables that were equal for all instances, as they do not contribute to the classification training.
4. *Principle component analysis.* We propose the principle component analysis on the data, then we use the cross-validation method to select the optimal number of principal components.

Table 3. The compared results of two-tailed t-test on accuracy with the 95 precent confidence level

	NN	kNN	SVM	NBC	HNB	MFBC
kNN	4/2/3					
SVM	0/5/4	0/7/2				
NBC	6/2/1	6/2/1	7/2/0			
HNB	5/2/2	6/3/0	7/2/0	4/3/2		
MFBC	1/1/7	2/4/3	3/1/5	2/6/1	2/6/1	
RMFBC	2/4/3	3/4/2	4/3/2	2/7/0	1/5/3	2/4/3

We conducted empirical experiments to compare RMFBC with naive Bayes classifier (NBC), HNB[11], nearest neighbor (NN), k nearest neighbor (kNN) and support vector machine (SVM)[3]. We evaluated the selected algorithms based on classification accuracy. To validate the performance of RMFBC, we used the 10-fold cross-validation method to estimate classification error. Specifically, for all experiments on each dataset, the cross-validation folds were identical. Finally, we compared the relevant algorithms using a two-tailed t-test with a 95% confidence level. According to statistical theory, we only consider the results of two datasets to be "significantly different" if the probability of a significant difference is at least 95% [5].

Table 2 shows the accuracy of each model on each data set, and Table 3 shows the compared result of two-tailed t-test, in which each entry $w/l/t$ means that the model in the corresponding column wins in w data sets, loses in l data sets and ties in t data sets, compared to the model in the corresponding row.

From our experiments, we can observe that RMFBC shows advantages over other algorithms when dealing with medium to high-dimensional data and datasets with imbalanced class distributions. However, for low-dimensional data and datasets that better satisfy the attribute normality assumption, the advantages of RMFBC over other classifiers are not significant. Additionally, we can see that the RMFBC algorithm significantly enhances the accuracy and robustness of the MFBC algorithm. Overall, RMFBC is a great classifier with significant advantages for high-dimensional continuous data.

References

[1] E. Fix and J. L. Hodges, "Discriminatory analysis - nonparametric discrimination: Consistency properties," *International Statistical Review*, vol. 57, p. 238, 1989. [Online]. Available: https://api.semanticscholar.org/CorpusID:120323383.

[2] P. Langley, W. Iba, and K. Thompson, "An analysis of bayesian classifiers," in *AAAI Conference on Artificial Intelligence*, 1992. [Online]. Available: https://api.semanticscholar.org/CorpusID:21634132.

[3] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Processing Letters*, vol. 9, pp. 293–300, 1999. [Online]. Available: https://api.semanticscholar.org/CorpusID:207579947.

[4] A. Doucet, S. J. Godsill, and C. Andrieu, "On sequential monte carlo sampling methods for bayesian filtering," *Statistics and Computing*, vol. 10, pp. 197–208, 2000. [Online]. Available: https://api.semanticscholar.org/CorpusID:16288401.

[5] C. Nadeau and Y. Bengio, "Inference for the generalization error," *Machine Learning*, vol. 52, pp. 239–281, Jan. 2003. DOI: 10.1023/A:1024068626366.

[6] J.-G. Wang, Neskovic, and Cooper, "An adaptive nearest neighbor algorithm for classification," in *2005 International Conference on Machine Learning and Cybernetics*, vol. 5, 2005, 3069–3074 Vol. 5. DOI: 10.1109/ICMLC.2005.1527469.

[7] L. Fan and K. L. Poh, "A comparative study of pca, ica and class-conditional ica for naïve bayes classifier," in *International Work-Conference on Artificial and Natural Neural Networks*, 2007. [Online]. Available: https://api.semanticscholar.org/CorpusID:21304301.

[8] L. Jiang, H. Zhang, and Z. Cai, "A novel bayes model: Hidden naive bayes," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, pp. 1361–1371, 2009. [Online]. Available: https://api.semanticscholar.org/CorpusID:17505286.

[9] T.-T. Wong, "Alternative prior assumptions for improving the performance of naïve bayesian classifiers," *Data Mining and Knowledge Discovery*, vol. 18, pp. 183–213, 2009. [Online]. Available: https://api.semanticscholar.org/CorpusID:12261093.

[10] C.-H. Cheng, C.-P. Chan, and Y.-J. Sheu, "A novel purity-based k nearest neighbors imputation method and its application in financial distress prediction," *Eng. Appl. Artif. Intell.*, vol. 81, pp. 283–299, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:88496304.

[11] Z. Geng, Q. Meng, J. Bai, *et al.*, "A model-free bayesian classifier," *Inf. Sci.*, vol. 482, pp. 171–188, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:67791655.

[12] D. Pradana and E. Sugiharti, "Implementation data mining with naive bayes classifier method and laplace smoothing to predict students learning results," *Recursive Journal of Informatics*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258748008.

[13] K. N. Markelle Kelly Rachel Longjohn, *The UCI Machine Learning Repository*, UCI Machine Learning Repository, DOI: https://archive.ics.uci.edu, 2024.