

Application of Bayesian Methods in Cancer Prediction

Yixin Wang

University of Wisconsin-Madison, Madison, USA
wang3493@wisc.edu

Abstract: According to the "Global Cancer Statistics 2020" report released by the International Agency for Research on Cancer, there were approximately 19.3 million new cancer cases worldwide in 2020, and nearly 10 million cancer-related deaths. These figures indicate that cancer has emerged as one of the major public health challenges globally, characterized by extremely high incidence and mortality rates. Cancer is particularly daunting because many forms of the disease exhibit no obvious symptoms in their early stages, often being diagnosed only at advanced stages, thereby complicating treatment efforts. In this context, early prediction of cancer becomes critically important as it enables precision treatment, reduces the overall burden of cancer therapy, alleviates the physical and psychological suffering of patients, and lessens the economic strain on their families, ultimately diminishing the threat posed by cancer to human society. This review systematically examines the use of Bayesian methods in cancer prediction. The paper outlines various Bayesian classification techniques and discusses their unique approaches to handling high-dimensional clinical and molecular data.

Keywords: Bayesian Method, Cancer Prediction, Naïve Bayes, Machine Learning

1. Introduction

In the era of big data and precision medicine, efficient and accurate cancer prediction has emerged as a critical challenge in biomedical research. Bayesian-based classification methods, particularly the Naive Bayes classifier, have gained prominence as essential tools for cancer risk assessment and early diagnosis due to their computational efficiency, interpretability, and excellent performance in handling high-dimensional data. This review provides an overview of the theoretical foundations and implementation strategies of the Naive Bayes method, including applications of Gaussian, multinomial, and Bernoulli models for various data types, as well as technical details such as parameter estimation, handling of missing data, and Laplace smoothing. Moreover, by analyzing case studies on lung, breast, and colon cancer prediction, the review demonstrates the broad applicability and superior performance of this method in processing clinical and molecular data.

The aim of this paper is to offer a comprehensive synthesis of theory and application for researchers in the field, discussing the advantages, limitations, and future prospects of Naive Bayes in cancer prediction.

2. Method and Theory

Let α and y represent the unknown parameter of interest and sample data, respectively. Then the Bayesian formula can be expressed in the following form: $P(C|X) = \frac{P(X|C)P(C)}{P(X)}$, where $P(C|X)$ is posterior probability. It means probability that the sample belongs to class C given feature set X . $P(X|C)$ is likelihood, it shows probability of observing features X given that the sample belongs to class C . $P(C)$ is prior probability and it shows that class C occurs. $P(X)$ is probability of observing features X .

For naïve Bayes assumption, the "naive" assumption is that the features are conditionally independent given the class label. That is:

$$P(X_1, X_2, \dots, X_n|C) = P(X_1|C)P(X_2|C) \dots P(X_n|C) \quad (1)$$

where X_i is individual feature, and C is class label. This simplifies the computation since the joint probability is replaced by the product of individual probabilities, making the algorithm computationally efficient even with high-dimensional data.

Depending on the type of data, different versions of Naive Bayes classifiers are used. The first is Gaussian Naive Bayes (for continuous data). If the features are continuous, they are assumed to follow a Gaussian(normal) distribution:

$$P(X_i|C) = \frac{1}{\sqrt{2\pi\sigma_C^2}} \exp\left(-\frac{(X_i - \mu_C)^2}{2\sigma_C^2}\right) \quad (2)$$

where μ_C is mean of feature X_i for class C , and σ_C^2 is variance of feature X_i for class C . The second is Multinomial Naive Bayes (for categorical data). For discrete or categorical features (e.g., gene mutation status), a multinomial distribution is used:

$$P(X_i|C) = \frac{n_{i,C} + \alpha}{N_C + \alpha d} \quad (3)$$

where $n_{i,C}$ is number of times feature i appears in class C , and N_C is total number of observations in class C , and d is number of features, and α means smoothing parameter. The third is Bernoulli Naive Bayes (for binary data). For binary features (e.g., presence/absence of a mutation), a Bernoulli distribution is used:

$$P(X_i|C) = p_i^{x_i}(1 - p_i)^{1-x_i} \quad (4)$$

where $X_i = 1$ if the feature is present, 0 otherwise, and P_i means probability of observing feature X_i in class C .

To compute prior probabilities for each class based on training data:

$$P(C) = \frac{\text{Number of samples in class } C}{\text{Total number of samples}} \quad (5)$$

The parameters for the likelihood function are estimated based on the type of data used. For Gaussian data, the mean and variance for each feature and class are estimated. For Multinomial and Bernoulli data, feature occurrence probabilities are computed using maximum likelihood estimation (MLE). After estimating the parameters, posterior probabilities for each sample are calculated using Bayes' theorem. The final classification decision is made by predicting the class with the highest posterior probability.

Handling missing data can be approached through various methods. Imputation techniques involve replacing missing values with the mean, median, or mode of the observed data. Alternatively, the

Expectation-Maximization (EM) algorithm can be used to estimate missing values iteratively by maximizing the posterior probabilities based on the available data.

To avoid zero probabilities, smoothing is commonly applied using Laplace smoothing. In this method, a small positive value α (often set to 1) is added to the numerator and a corresponding adjustment is made to the denominator when calculating the conditional probability. This technique ensures that no feature probability becomes zero, which improves model robustness, especially when dealing with sparse data. The Laplace-smoothed conditional probability is defined as:

$$P(X_i|C) = \frac{n_{i,C} + \alpha}{N_C + \alpha d} \quad (6)$$

where $n_{i,C}$ is the number of times feature X_i appears in class C , N_C is the total number of observations in class C and d is the number of features.

3. Results and Applications

Gaussian Naive Bayes has been effectively applied to lung cancer prediction using clinical and molecular data, including gene expression, smoking status, and air pollution levels. The model utilized key features such as smoking history, environmental exposure, and genetic mutations to predict lung cancer incidence. The model demonstrated strong predictive performance, achieving an accuracy of 90% and an area under the curve (AUC) of 0.88, highlighting its effectiveness in handling high-dimensional data and complex dependencies in lung cancer risk assessment [1].

It has been successfully applied to breast cancer prediction using the Wisconsin Breast Cancer Dataset (WBCD). The model leveraged features such as tumor size, shape, texture, and other biological markers to distinguish between benign and malignant breast tumors with high sensitivity and specificity, achieving an accuracy of 98% and an area under the curve (AUC) of 0.97 [2].

It also has been applied to colon cancer prediction using gene expression data from colon cancer patients. The model utilized gene expression levels as predictive features and demonstrated high accuracy in distinguishing colon cancer cases, achieving an accuracy of 95.24%. This result underscores the importance of gene expression patterns in accurately identifying colon cancer.

Moreover, Naive Bayes classifiers have shown excellent performance in colon cancer prediction. Salmi and Rustam used Gaussian Naive Bayes on clinical data from colon cancer patients, utilizing biomarkers including carcinoembryonic antigen (CEA), haemoglobin, leukocyte counts, haematocrit levels, and thrombocyte counts [3]. Their model demonstrated robust classification accuracy, reaching 95.24% with 80% training data. Additional evaluations from their confusion matrix showed precision of 100%, recall of 94%, and an F1-score of 96%, emphasizing the model's capability to precisely identify colon cancer cases while effectively excluding non-cancer cases [3].

Bayesian methods have been widely applied in cancer prediction and prognosis due to their ability to handle high-dimensional, noisy, and incomplete data while incorporating prior knowledge from biological or clinical research. The ability to update and refine predictions as new data becomes available makes Bayesian models especially useful for personalized medicine and cancer research. According to Zhai et al. [4], Bayesian N-of-1 trials provide rich posterior information for clinical decision-making. The fundamental approach begins with the classic N-of-1 trial design, where an individual patient undergoes multiple treatment cycles with randomized, crossover, and blinded protocols. By leveraging the within-subject control inherent in N-of-1 designs and incorporating prior information, the Bayesian model can achieve reliable estimates even with limited sample sizes [4].

The article by Ye and Lin addresses the challenge of constructing sparse regression models for high-dimensional cancer data by improving upon traditional sparse methods such as the Lasso [5]. Ye and Lin compared the performance of three Bayesian approaches—the traditional Laplace prior-based method, the semi-mixed LAZE prior model, and the complete LAZE prior model—against six

common sparse methods (including Least Squares, Best Subset Regression, Ridge, Lasso, Principal Component Regression, and Partial Least Squares), discovered the Bayesian framework can achieve a higher degree of sparsity while simultaneously reducing prediction error—a significant advantage for handling high-dimensional clinical data [5].

Bayesian approaches provide rich posterior distributions that allow researchers to quantify uncertainty in parameter estimates, making them particularly valuable in the context of small-area analyses and sparse data. Models such as the Bayesian spatio-temporal and hierarchical Bayesian models enable the accurate estimation of lung cancer risk across different regions and time periods. These methods facilitate the identification of high-risk clusters and allow for adjustment of confounding variables through multilevel modeling. Bayesian models have significantly advanced lung cancer research, systematic approaches and further methodological development are needed to address challenges such as computational complexity and model identifiability.

The article by Yi, Guo, Zhuang, Xu, and He reviews the application of various Bayesian models in the investigation of lung cancer incidence and mortality [6]. These models incorporate both spatial and temporal correlations to analyze the evolution of lung cancer risk across regions and time periods. They are implemented using Markov chain Monte Carlo (MCMC) techniques and are evaluated using criteria such as the deviance information criterion (DIC) [6].

The integration of high-dimensional omics data using Bayesian methods, especially through multi-omics approaches, has been shown to yield more accurate prognostic models than those based on single-omics. These models provide improved survival prediction by leveraging the complementary strengths of various data types [7].

Khozama and Mayya propose a novel range-based breast cancer prediction model that integrates Bayes' theorem with ensemble learning to produce a probabilistic risk score rather than a simple binary outcome [8]. The model is developed using a modified version of the Breast Cancer Surveillance Consortium (BCSC) dataset. The methodology begins with the computation of various statistics derived from the BCSC dataset. After the probabilistic computation, the newly modified BCSC dataset (with added attributes such as cancer score and non-cancer score) is used to train an ensemble learning model. The ensemble is based on the AdaBoost algorithm with decision trees as base learners. Hyperparameter optimization is performed using Bayesian techniques to fine-tune parameters such as the maximum number of splits, number of learners, and learning rate. This ensemble approach refines the range-based prediction by combining multiple weak classifiers to produce a robust final model. The evaluation of the proposed model is conducted under two training scenarios—using a subset and the entire BCSC dataset. A comparative analysis (See Table 1 below) shows that the proposed range-based prediction model has the unique advantage of providing a risk score as a percentage [8]. Moreover, the model outperforms earlier methods that yield only binary outcomes by offering detailed, probabilistic insights into the likelihood of breast cancer presence, thus enhancing the interpretability and utility of the predictions.

Ni, Xu, and Wang investigate the spatiotemporal distribution of female lung cancer incidence in China using a Bayesian spatiotemporal model [9]. The study employs a hierarchical Poisson regression framework to model the standardized incidence ratio (SIR) of lung cancer across 33 provincial administrative units over the period 1992–2016. The model detects significant space–time interaction effects. These interactions indicate that while the overall temporal trend of lung cancer incidence does not change markedly at the national level, local regions exhibit dynamic variations over time. The authors conclude that from 1992 to 2016, although the overall temporal trend in female lung cancer incidence in China remained relatively stable, there is clear spatial clustering and significant local spatiotemporal interaction effects. Moreover, among the various examined risk factors, household kitchen ventilator ownership appears to be a key factor influencing the spatial distribution of lung cancer risk among Chinese women [9].

Zhong and Xue demonstrate that a joint model combining Bayesian network uncertainty inference with Cox regression can provide robust and well-calibrated lung cancer risk predictions [10]. This integrated approach overcomes limitations of traditional risk models that require complete datasets and offers a promising tool for identifying high-risk individuals for lung cancer prevention and control.

Bazila Banu et al compare several Bayes classifiers for breast cancer classification using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset [11]. Before applying gradient boosting, the Naïve Bayes-based classifiers (NB, TAN, and BAN) achieve an approximate accuracy of 90.1% when classifying benign versus malignant cases. When integrated with the GB technique, all classifiers exhibit improved performance. In particular, the TAN classifier demonstrates the best performance: The mean squared error (MSE) for TAN with GB is reduced to approximately 0.04%, and its misclassification rate to 0.05%. In comparison, the BAN classifier shows error rates of 0.07% and 0.09%, and the standard Bayesian Network shows error rates of 0.07% and 0.10%. These improvements underscore the benefits of combining Bayesian classifiers with ensemble boosting techniques for breast cancer classification tasks.

Table 1: Comparison of different methods [8].

Researcher	Methods	Used dataset	Output cancer score	Results/Limitations
Ramkumar et al.	Naïve Bayes classifier	BUPA research lab 20 cases	Yes or No	Accuracy: 50% Low dataset size, low accuracy. No parameter tuning.
Al-Jawad et al.	SVM, Bayesian network	Haberman's survival 306 cases	Yes or No	SVM: Recall: 73.78% Precision: 74.77% BM: Recall: 78.22% Precision: 64.47% Fixed learning parameters. Low dataset size.
Yang et al.	Bayesian models Markov models ANN	36000 cases	Yes or No	Accuracy: ANN: 73.55% CBM: 76.07% Fusion: 75.63% Low F1-score due to the nature of their dataset.
Li et al.	Logistic Regression, SVM, Naïve Bayes, and Bayesian Network	154899 samples	Yes or No	Naïve Bayes has the best accuracy. No parameter optimization.
Kurian and Jyo	K-NN, SVM, Naïve Bayes, AdaBoost, Decision Trees, Random Forests, gradient boosting	NCBI dataset 1580samples	Yes or No	Decision trees accuracy: 94.3% No parameter optimization. Low dataset size.

Overall, Bayesian methods have achieved satisfactory prediction outcomes in various cancer predictions through the flexible application of different models (such as Gaussian, Multinomial, and Bernoulli), providing practical data support for clinical decision-making and early diagnosis. Due to their flexible statistical inference framework and the advantage of characterizing uncertainty, Bayesian methods are increasingly applied in the medical field. Firstly, Bayesian models can fully utilize prior knowledge by integrating clinical experience and historical data into the predictions, which enables them to obtain robust results even when sample sizes are limited or the data is noisy. Especially in personalized medicine and precision treatment, by continuously updating the posterior

distribution, the models can be optimized as new data are incorporated, thereby offering patients more accurate risk assessments and treatment recommendations.

4. Conclusion

In conclusion, this study systematically reviewed and analyzed the theoretical foundations and practical applications of Bayesian classification methods in cancer prediction. From specific cases of lung, breast, and colon cancer, one observed the outstanding performance of Bayesian models in handling high-dimensional, complex, and incomplete data. The effective integration of prior knowledge, flexible model construction, and comprehensive characterization of uncertainty has made Bayesian methods an important tool in the medical field, particularly for the early diagnosis and risk assessment of cancer. Looking ahead, as multi-omics data become widely applied and computational technologies continue to advance, Bayesian methods are expected to play an even greater role in personalized medicine, real-time risk prediction, and cross-domain data integration, thereby providing robust data support and theoretical underpinning for precision medicine and clinical decision-making.

References

- [1] Setiawan, Q. S., Laeli, A. R., Rustam, Z., Wirasati, I., & Hartini, S. (2020). Comparison of naive Bayes and decision tree for classifying hepatocellular carcinoma (HCC). In *2020 International Conference on Innovation and Intelligence for Informatics, Computing and Technologies (3ICT)*. IEEE.
- [2] Kamel, H., Abdulah, D., & Al-Tuwaijari, J. M. (2019). Cancer classification using Gaussian Naive Bayes algorithm. In *Fifth International Engineering Conference on Developments in Civil & Computer Engineering Applications 2019 (IEC2019)* (pp. 165). IEEE.
- [3] Salmi, N., & Rustam, Z. (2019). Naïve Bayes classifier models for predicting the colon cancer. *IOP Conference Series: Materials Science and Engineering*, 546, 052068.
- [4] Zhai, J. B., Li, J., Shang, H. C., Wang, H., Dai, M., & Chen, J. (2017). Brief introduction of Bayesian N-of-1 trials. *Chinese Journal of Evidence-Based Medicine*, 17(7), 848–851.
- [5] Ye, Q., & Lin, Y. (2018). Application of machine learning methods based on LAZE priors to cancer data: Take the prostate cancer data set for example. *Journal of South China Normal University (Natural Science Edition)*, 50(4), 115–120.
- [6] Yi, J., Guo, C., Zhuang, J., Xu, X., & He, F. (2024). Application of Bayesian models in studies related to factors influencing lung cancer incidence and mortality. *Modern Oncology*, 32(13), 2485–2491.
- [7] Chu, J., Sun, N., Hu, W., Chen, X., Yi, N., & Shen, Y. (2022). The application of Bayesian methods in cancer prognosis and prediction. *Cancer Genomics & Proteomics*, 19, 1–11.
- [8] Khozama, S., & Mayya, A. M. (2022). A new range-based breast cancer prediction model using the Bayes' theorem and ensemble learning. *Information Technology and Control*, 51(4), 757–770.
- [9] Ni, X., Xu, N., & Wang, Q. (2022). Investigating the incidence risk and influencing factors of female lung cancer in China based on a Bayesian spatiotemporal model. *Journal of Environmental Hygiene*, 12(6), 804–914.
- [10] Zhong, L., & Xue, F. Z. (2023). Lung cancer risk prediction model based on Bayesian network uncertainty inference. *Journal of Shandong University (Health Sciences)*, 61(4), 87–92.
- [11] Bazila Banu, A., & Thirumalaikolundusubramanian, P. (2018). Comparison of Bayes classifiers for breast cancer classification. *Asian Pacific Journal of Cancer Prevention*, 19(10), 2917–2920.