Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/5/20230555

Comparison between Bayesian and SVM model for breast cancer risk prediction

Jiatong Jiang[1, †], Haodan Li[2, 5, †], Houhua Qi[3, †], Wenbo Wu[4, †]

[1]School of Information Science and Engineering, Dalian Polytechnic University, Dalian, Liaoning, 116000, China
[2]School of Electrical Engineering, University of South China, Hengyang, Hunan, 421000, China
[3]School of Electronics and Information Engineering, Lanzhou Jiaotong University, Lanzhou，Gansu, 730000, China
[4]School of International Exchange, Ningbo University of Technology, Ningbo, Zhejiang, 315000, China

[5]16011010228@stu.suse.edu.cn
[†]These authors contributed equally

Abstract. Breast cancer ranks first in the incidence of cancer worldwide. It is a kind of malignant disease with incidence increasing year by year. Recently, machine learning algorithms widely used in cancer prediction and other fields to relieve the burden of doctors and accelerate the diagnostic process. In this work, two representative models, the support vector machine (SVM) and Bayesian classification algorithm are leveraged for breast cancer risk prediction. It is found that the effect and accuracy of the two algorithms are very different when they are used for breast cancer prediction. However, there is still a research gap in the comparison of the two algorithms in practical application. Therefore, the research topic of this paper is the comparison between the Bayesian classification algorithm and the SVM algorithm in breast cancer prediction. The research methods of this paper are as follows: firstly, the dataset is collected, then applying the Bayesian classification to process the dataset, and then the SVM algorithm is used to process the dataset. Finally, the processing results of the two algorithms are compared to comprehensively analyze and compare which algorithm is more efficient and more suitable for actual breast cancer prediction. The test results support that the SVM outperforms the Bayesian classification algorithm in the actual target tracking problem. Therefore, it is suggested to choose the SVM classification algorithm in the actual target tracking problem to boost the accuracy and efficiency of prediction results to the greatest extent.

Keywords: Machine Learning, Bayesian Model, Support Vector Machine, Breast Cancer Risk Prediction.

1. Introduction

Breast cancer is a widely seen disease of the female breast, ranking first in the incidence of female cancer worldwide, and the incidence is increasing year by year. [1] Breast cancer screening is an

DOI: 10.54254/2755-2721/5/20230555

effective, simple, and economical breast cancer screening measure for asymptomatic women to achieve early detection, early diagnosis, and early treatment, aiming at reducing breast cancer mortality [2]. Delayed diagnosis and inaccurate prediction are important reasons for the high mortality rate of breast cancer [3]. Bayesian classification algorithm, SVM and other methods were leveraged to predict breast cancer risk, which greatly improved the accuracy of prediction, shortened the prediction time, was conducive to early detection and treatment of patients, and greatly improved the survival rate of patients.

SVM [4] is a widely used linear classifiers. It learns from labelled data and classify novel data bivariate [5]. Support vector machine algorithm can estimate almost any shape function, and automatically estimate the optimal fitting function according to the sample distribution. In terms of estimation accuracy and applicability, support vector machine algorithm is superior to traditional function estimation methods such as least square method, which has a good application prospect [6].

Bayesian classification algorithm is a representative generative classification model. It is constructed based on statistical classification methods, exploiting probability and statistical knowledge [7]. It has many advantages, such as easy implementation, could be extend to large scale dataset and high classification accuracy [8]. The idea of Bayes' Theorem requires extensive data computation and inference to be effective. It is very useful in many computer applications such as natural language processing, machine learning, recommendation systems, image recognition, game theory, etc [9]. This method is based on Bayes theorem. Given a specific dataset, this method first calculates the joint probability distribution based on the independent assumption [10]. Afterwards, the well-trained model could be used to predict novel samples and generate the probabilities [11].

In this paper, SVM algorithm and Bayesian classification algorithm will be introduced respectively in the Method section, and the prediction results of the two methods will be analysed and compared in the Result section**.**

2. Dataset Source introduction

Breast cancer is common in various types of cancer. It is also the one of the deadliest cancers and can be determined by examining the lump in the breast. The experiments in this paper are conducted on the dataset collected from the University of Wisconsin Hospitals, Madison. For this dataset six features are exploited, which are mean_radius, mean_texture, mean_perimeter, mean_perimeter, mean_area, mean_smoothness, diagnosis. This information on breast cancer obtained through the collection will be used as a dataset in the next prediction methods. In this experiment, we will use SVM and Bayesian methods to process the indicators to obtain the final prediction of the probability of breast cancer.

3. Method

3.1. Naive bayes classifier

Bayes method is widely leveraged in machine learning field. It is implemented based on Bayesian principle. It exploits the statistical knowledge for learning and classifying. The Bayes method could combine the prior probability and the likelihood probability for learning. By doing so, it could alleviate the drawbacks of leveraging the prior probabilities only. Therefore, the Bayes classification algorithm could obtain a relative high accuracy. When dealing with large-scale dataset, it could achieve high accuracy with relatively low computational burden.

Naive Bayes is an improved version of the aforementioned Bayes algorithm. It is designed based on a solid mathematical foundation and stable classification efficiency. Meanwhile, the parameters in this model are relatively low, so that it is helpful for avoiding the overfitting problem. Moreover, there are many other advantages. For example, it is easy to implement. However, the naive Bayes method assumes that the features are conditionally independent of each other. No feature accounts for a too large or small proportion in the decision-making results. Methods of this kind will weaken the classification effect of Bayes classification model to a certain extent, since the independent features assumed in Naive Bayes are not usually tenable in practical applications. Although Bayes classification sacrifices the

DOI: 10.54254/2755-2721/5/20230555

classification effect, it is helpful for constructing a simple but effective model, which makes it more usable and available in many practical cases.

Conditional probability：There are two events x, y. Their respective probabilities P(x) and P(y) denote the occurrence of event x and y. P(y|x) could be regarded as the conditional probability of occurrence of event y. Given the event x is true, P(x) represents the prior probability and P(y|x) represents the likelihood probability. The basic formula is shown in (1):

$$P(y|x) = \frac{P(xy)}{P(x)} \tag{1}$$

It is known from Bayes Theorem that formula (1) can be converted into (2):

$$P(y|x) = \frac{P(xy)}{P(x)} = \frac{P(y)P(x|y)}{P(x)} \tag{2}$$

Assuming that each sample x includes the i-dimensional features, that is, $x = (x_1, x_2 \ldots x_i)$, then formula (3) can be obtain:

$$P(x|y) = P(x_1, x_2, \ldots, x_i\ |y) \tag{3}$$

Since the Naive bayes makes an independent assumption on the conditional probability, that is assuming that the features, such as x_1, $x_2 \ldots x_i$, are assumed to be independent of each other, formula (4) can be obtained:

$$P(x|y) = P(x_1, x_2, \ldots, x_i\ |y) = \prod_{i=1}^{n} P(x_i|y) \tag{4}$$

Formula (5) can be obtained by bringing formula (4) into formula (2):

$$P(y|x) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x)} \tag{5}$$

In the Naive bayes classifier we do not need the evidence P(x), the evidence remains constant for a given input, we can remove this term and just maximize the numerator. As is shown in formula (6):

$$f(x) = argmax(P(y)\prod_{i=1}^{n} P(x_i|y)) \tag{6}$$

3.2. SVM

Predicting the direction of the data by the trained model requires high accuracy in a limited time. The basic version of the SVM is trained by a supervised machine learning model to provide a hyperplane to double-classify the data set, and the segmenting results is the prediction of the base SVM, which is simple and efficient when the amount of data is too large. The SVM finds an optimal divider, a boundary (hyperplane) such as a dividing line, a dividing surface, so that samples on both sides are far enough away from the interface. When the distance is far enough, the classification can be made: assuming that the one on the left side of the boundary is class A, then the one on the other side is class B. Can be well distinguished. SVM (Support Vector Machine) is a linear classifier. It seeks the largest spacing in the feature embeddings, which maximizing the spacing between different classes.

If x represents sample data and y represents the category of data, then this linear classifier aims at learning a hyperplane in n-dimensional feature embeddings. The learned hyperplane can be represented by this equation:wtx+b=0. This hyperplane can be represented by f(x), and when f(x) is equal to 0, x could be regarded as the point in the hyperplane. If there is a point with f(x) greater than 0, it indicats that the label of the data point is positive or 1. Similarly, points with f(x) less than 0 means points have negative ground truth. In SVM, there is a concept called function interval: (y(w*x+b)). where x is a function variable; w is a normal vector. Normal refers here to a vector represented by a line perpendicular to the plane, which determines the direction of the hyperplane.

That is, we use this positive and negative to determine whether the classification is correct or not. However, there is still some deficiencies, if the w and b are changed proportionally (such as doubling them), the value of the function interval will also be doubled meanwhile leave the original hyperplane unchanged. As a result, maximizing the function interval merely is insufficient. In practical situations,

DOI: 10.54254/2755-2721/5/20230555

some constrains should be implemented on the normal vector w, so that researches could achieve true distance from the point to the hyperplane.

Given a point x, and its projection on to the hyperplane, the mapped point is denoted as x0. w is a perpendicular vector to the hyperplane, and the γ is the distance from the sample x to the hyperplane. When classifying, the higher the confidence level, the more accurate the classification. Therefore, the classification probabilities could be measured by the distance of the points to the decision hyperplane. Therefore, in order to obtain a higher degree of confidence, people also want to maximize this interval by making this hyperplane. In the SVM, there will be results that are very different from the prediction, and then the relaxation variable is used. The use of this relaxation variable can be seen as a noise of "inclusive acceptance". Making this one outlier does not cause the hyperplane to move and the spacing to shrink. Model was adjusted sensitively by changing the relaxation variable. Corresponding to the relaxation variable is the penalty factor C. The penalty coefficient refers to the degree of punishment for classification errors, when the C is larger, the better the effect of the classifier, but it is easy to overfit and the generalization ability becomes worse. Conversely, when C is smaller, the generalization ability will increase but the accuracy will be slightly reduced.

4. Result

4.1. Evaluation Matrix

In this work, the confusion matrix is leveraged as the measurement of the performances of the model. It is one of the mostly used indexes for evaluating the classification model, which reveals many valuable information of the classification results.

In the matrix, the number of true values classified as true by the model is represented as True Positive (TP). The number of true values considered as negative is denoted as False Negative (FN). Error of this kind is also regarded as the statistical Type I error. Moreover, the number of negative values but classified to positive is regarded as False Positive (FP). Statistically, it is understood as the statistical type II error. The number of negative values predicted as negative is decided as True Negative (TN). By demonstrating these four values simultaneously in the same matrix, we can obtain a matrix that is the confusion matrix. The confusion matrix also requires secondary indicators, which are accuracy, precision, and sensitivity-specificity. The various indicators are calculated as follows:

Accuracy=(TP+TN)/(TP+TN+FP+FN)

Precision=TP/(TP+FP)

Sensitivity=TP/(TP+FN)

Specificity=TN/(TN+FP)

On this basis, a three-level index can be obtained by further expansion. It's called the F1 Score. The F1-Score metric considers both the Precision and Recall value. Its value is larger than 0 and smaller than 1. The 1 represents the optimal output of the model and 0 is the worst output of the model.

4.2. Result comparison

Comparing the effects of SVM and Nave Bayes models against the training breast cancer prediction dataset, the resulting confusion matrix study found that the SVM model was less correct in predicting the true positive data than the Naive Bayes model, where the wrongly diagnosis rate of the SVM model was higher. In the prediction of true negative data as shown in Figure1 and Table 1, the SVM has a higher accuracy rate than the confusion matrix, that is, a lower misdiagnosis rate. This is the different effect of the two models working with the same data.

DOI: 10.54254/2755-2721/5/20230555

Figure 1. Comparison of confusion matrices of different algorithms.

Table 1. Compares of ACC PVV TPR and FN.

	accuracy score	precision score	recall score	F1-score
SVM	0.9	0.89	0.82	0.86
Naive Bayes	0.89	0.91	0.74	0.82

The analysis of four related parameters of the confusion matrix found that the accuracy score of the SVM model was higher than that of The Nave Bayes, but because this dataset was predicted for breast cancer, the positive and negative ratios were imbalanced, so the accuracy score here was not very informative, and the precision score and call score were parsed, and the precision of the SVM model was found The score is higher than that of Nave Bayes, while the call score, on the contrary, in the parameter F1-score parsing, which can better exclude the influence of the data itself, it is found that the value of the SVM is higher than that of Naive Bayes, which proves that the SVM model prediction is more applicable and works better for this type of data set

In real life, for the specific actual situation, the emergence of misdiagnosis rate will lead to the patient's time and mood is affected but will not cause damage to the patient's body, so misdiagnosis is usually tolerated, its tolerance is high, but the emergence of missed diagnosis may lead to untimely treatment and then affect the patient's health, its tolerance is low, so this data set as a breast cancer prediction, its False negative is given a higher importance, In other types of data sets such as the detection of flammable and explosive materials or animal recognition, human expression recognition, etc., specific predictions need to be specifically analysed for specific situations, and the reference of which model to use cannot only look at one or a few items of data in figure, but need to be weighted and balanced, and the calculation process of F1-score also involves the specific balance of the precision score and recall weighting coefficients, if precision Score is more important, the β in F1-score is less than 1 and the final value is smaller, otherwise the final value is larger, the prediction effect is good or bad in the data set where there is no imbalance in the positive and negative ratios, and the prediction tolerance is not too low, F1-score can make the data have a more intuitive reflection

5. Discussion

The current SVM method is a mature classification model, and the corresponding results can be obtained when comparing with Bayesian methods. It could be observed that both algorithms yielded appropriate conclusions for the prediction of breast cancer, but there were differences in the conclusions obtained by the two models. It could be concluded that both methods yielded appropriate conclusions for the prediction of breast cancer, but there were differences in the conclusions obtained by the two models.

Comparing the results of Bayesian method and SVM method about the project prediction, we can get that SVM method is more accurate than Bayesian method in terms of prediction accuracy, etc. The accuracy score of SVM method can be as high as 0.90, and thus the comparison concludes that the intelligent algorithm supported by a certain amount of data has good model prediction ability, can process the data effectively, and draws valid conclusions. However, the SVM method is inferior to the Bayesian method in terms of prediction scores because of the difference in the mathematical theory of the two models and the difference in the requirements of the data set between them. The SVM method is superior to the Bayesian method in terms of prediction scores because of the difference in the

DOI: 10.54254/2755-2721/5/20230555

mathematical theory of the two models and the difference in the requirements of the data set between them.

Currently, machine learning continues to innovate and breakthrough in the medical field, and more and more machine learning-related algorithms are being applied to the medical field. Biomedical data sets will become increasingly suitable for analysis. The growth in the application of clinical data (including images, records and real-time monitoring data, intelligent algorithms and related mathematical models will greatly expand the availability of phenotypic and environmental data that were previously unavailable at the same scale. Meanwhile, the application of machine learning methods based on genomic, phenotypic, and environmental predictors may also develop new methods for disease classification. This project plays and vital role in the medical field, as the accuracy of the model is high, it can help doctors to diagnose the corresponding diseases, thus reducing the cost and effort of patients in the examination, and it can effectively reduce the time spent by doctors to diagnose patients' diseases, which is a very good tool as a supplementary diagnosis. With the improvement of data depth and methods, this project can be widely used not only in breast cancer prediction, but also in other areas, which can help medical progress and development. In the future, SVM methods can be used as models for detecting the efficacy of anti-cancer drugs, positively impacting the estimation of drug efficacy and the production of new drugs, and more and more deeper applications of deep learning algorithms will be available to contribute to the advancement of human medical technology.

6. Conclusion

In this paper, the effectiveness of SVM and Bayesian classification algorithms is validated in the breast cancer risk prediction problem. SVM is a typical machine learning algorithm, Bayesian classification is a statistical classification algorithm. Both of which are widely used in cancer prediction and other fields because of their high efficiency and accuracy in training prediction data sets. The results show that the SVM could achieve an accuracy of 0.9. It is a satisfactory result. The comparison results demonstrate that the Bayesian classification algorithm is worse than the SVM in many practical problems, and the speed and accuracy of the results are higher. Therefore, it is suggested to choose the SVM classification algorithm in the actual target tracking problem for boosting the accuracy of prediction results to the greatest extent.

In the future, more advanced machine learning algorithms could be exploited, such as the artificial neural network. The increase in accuracy for cancer risk prediction could be quite useful for alleviating the burden of doctors and providing faster and more accurate diagnostic results for patients.

References

[1] Yala A, Mikhael P G, Strand F, et al, 2022, Multi-institutional validation of a mammography-based breast cancer risk model. Journal of Clinical Oncology, vol: 40(16), pp 1732-1740.

[2] Eriksson M, Czene K, Pawitan Y, et al, 2017, A clinical model for identifying the short-term risk of breast cancer. Breast Cancer Res, vol 19, pp 1-8.

[3] Tyrer J, Duffy SW, Cuzick J, 2004, A breast cancer prediction model incorporating familial and personal risk factors. Stat Med, vol 23, pp 1111-1130.

[4] Smith RA, Andrews KS, Brooks D, et al, 2019, Cancer screening in the United States, 2019: A review of current American Cancer Society guidelines and current issues in cancer screening. CA Cancer J Clin, vol 69, pp 184-210.

[5] C. Ingolf, 2018, Implications of pharmacogenomics in Precision medicine, Clinical. Magazine. Therapeutics, vol 103, pp 732-735.

[6] Semenza G L, 2003, Targeting HIF-1 for cancer therapy. Nature reviews cancer, vol 3(10), pp 721-732.

[7] Golub GTR, Slonim DK, Tamayo P, et. al, 1999, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, vol 286(5439), pp 531-537.

DOI: 10.54254/2755-2721/5/20230555

[8] Alizadeh A, et al, 2000, Different types of diffuse large b-cell lymphoma identified by gene expression profiling. Nature, vol 403, pp 503-511

[9] Chiesa M, Maioli G, Colombo GI, et al, 2020, GARS: genetic algorithm for the identification of a Robust Subset of features in high-dimensional datasets. BMC Bioinform, pp 21-54.

[10] Guyon I, Weston J, Barnhill S, Vapnik V, 2002, Gene selection for cancer classification using support vector machines. IEEE Trans Knowl Data Eng, vol 25(1), pp 1-14

[11] Jansi RM, Devaraj D, 2019, Two-stage hybrid gene selection using mutual information and genetic algorithm for cancer data classification. J Med Syst, vol 43, pp 235.