

Comparison of Different Machine Learning Models in Breast Cancer

Shuai Wu^{1,*,†}, Wenjia Xiong^{2,†}

¹ Department of Engineering, Changchun University of Science and Technology, Changchun, China

² Department of Engineering, Nanchang Institute of Technology, Nanchang, China

* Corresponding author: 2020003115@mails.cust.edu.cn

†These authors contributed equally.

Abstract. Breast Cancer is mainly found in women and is the main cause of increased mortality among women. Breast cancer diagnosis is time-consuming, and due to the low availability of the system, it is necessary to develop a system that can automatically diagnose breast cancer at an early stage. Various machine learning and Deep Learning Algorithms have been used to classify benign and malignant tumors. This paper focuses on the implementation of various models, such as Logistic regression, random forest and naive Bayes. Each algorithm has measured and compared the accuracy and obtained accuracy. This paper aims to compare the advantages and disadvantages of different regression models in breast cancer prediction. The method proposed in this paper can promote the integration of machine learning and medicine, and improve clinical diagnostic accuracy.

Keywords: Machine Learning, Prediction Model, Breast cancer.

1. Introduction

Breast cancer is a cancer forming in breast, which includes many types depending on the breast cancer cells [1]. Breast cancer accounts for about 12 percent of all new cancer cases and 25 percent of all cancers in women [2]. According to a recent analysis of cancer worldwide by the International Agency for Research on Cancer (IARC), the global cancer burden has increased to 19.3 million new cases and 10 million deaths by 2020. One in five people in the world will develop cancer, and one in eight men and one in 11 women will die from it. The most common type of cancer in the world is female breast cancer, with 2.3 million cases diagnosed in 2020, surpassing the number of new cases of lung cancer for the first time, according to a new report by the International Agency for Research on Cancer. Breast cancer is becoming a cancer that affects a large number of people [3]. Nowadays, mammograms, ultrasound and biopsies are applied to detect breast cancer. However, these techniques are time-consuming and cannot reliably detect early stages of breast cancer, so patients need a computational diagnostic system that recruit machine learning to assist early breast cancer diagnosis. This method includes tumor classification, accurately detect cell algorithm, and more efficient than traditional diagnosis [4]. With the rapid development of machine learning, k-nearest Neighbor (KNN), Support Vector Machine (SVM), Naive Bayes, decision tree, random forest and other algorithms have been developed. It is the latest not only in deep learning algorithms, but also in deep learning algorithms for classification. In depth study, it is currently being developed algorithm, is currently being developed in further study of the algorithm [5]. Breast cancer survival prediction may be extreme influence the selection of the best treatment plan. Have been adopted by many methods, such as statistics, machine learning model to predict survival in patients with prospects, but can test the depth of learning, etc. A new algorithm, in order to improve the model and the prediction accuracy. In previous study, they use the method of machine learning and deep learning medical center at the university of Malaya in 4902 patients with breast cancer registry records in predicting breast cancer. The results showed that the multi-layer perceptron (MLP), random forests (RF) and decision tree (DT) classifier to predict survival, respectively, the test sample accuracy were 88.2%, 83.3% and 82.5%, respectively. Support vector machine (SVM) is falling at the rate of 80.5% [6]. Before this, clinical

doctors to use Microsoft Excel and Cox and Pandas software program analysis the influence factors of breast cancer survival rates. The shortcoming of the traditional statistical analysis of lead to all sorts of machine learning (ML) method is widely used in the field, such as decision tree (DT), random forests (RF), neural networks, extremum ascension, logistic regression, naive bayes and support vector machine (SVM) [7].

This paper aims to compare the advantages and disadvantages of different regression models in breast cancer prediction. To predict the factors of breast cancer, there are many people who use different methods and create different models to predict it, which has been a great help to our breast cancer research. This paper compares three models: naive Bayes algorithm, random forest classifier and KNN algorithm. Our goal is to evaluate the algorithm performance under different situation data, to discuss their sensitivity and accuracy. Naive Bayes are best when the correlation between attributes is low. Of course, there are other algorithms, such as decision trees, that are slightly improved for partial correlation. The rest of this article can be found in the following sections: in the proposed research includes literature review part is section 3, it is included in the proposal itself. This section mainly analyzes the advantages and disadvantages of naive Bayes algorithm, random forest classifier and KNN algorithm, and fully verifies my views with relevant literature. Section 4 discusses the similarities and differences between the three algorithms, which will be explained in more detail later in this article. Section 5 is the summary and conclusion of this paper. Section 6 is the source of all relevant literature.

2. The different machine learning models

2.1. Naive Bayes

Naive Bayes algorithm is one of the well-known algorithms and has made many great contributions in the field of classifier computing. On the basis of the Bayesian algorithm, some corresponding simplifications lead to the Naive Bayesian algorithm, which assumes that all attribute variables have the same proportion to the decision result. The classification effect of Bayesian algorithm is weakened by this method. However, due to the logic of Bayesian algorithm is simple and stable, for different types of data sets, the classification performance of Naive Bayes algorithm is not much different. The naive Bayes classification algorithm can achieve better results when the relationship between data set attributes is relatively independent. In recent years, researchers have focused on improving naive Bayes classifier. A previous study showed a new NB (weighted NB) classifier is proposed and applied to breast cancer detection [8]. The application effect of weighted NB in breast cancer database was evaluated through experiments. The experiment adopted 5 - fold cross - validation method. In addition, various performance evaluation techniques, namely sensitivity, specificity and accuracy, were considered. The sensitivity, specificity and accuracy of the calculation are 99.11%, 98.25% and 98.54% respectively.

Naive Bayesian classifier in the preparation stage will be to pretreatment of data set, can obtain the training samples. After each category in the samples making probability estimation, estimate the different categories under the condition of different attribute values of probability. Then calculate the probability of each category for each attribute combination, which is the learning stage of Bayesian classifier. Followed by that stage, the biggest probability value as a result that will be a bayesian classifier to select output, then output.

Some studies have shown that the naive Bayes model has a lower error rate than the KNN algorithm and the random forest model. But this is not the case, because the model of Naive Bayes is built on the basis of the independence between attributes, which does not occur in real life. Naive Bayes classifiers do not perform well when there is no mutual independence among attributes. It is also worth mentioning that this algorithm needs to know the prior probabilities, however prior probabilities in most cases depend on the assumptions. There are many and various hypothetical models, so in some cases the predictions will be poor because the probabilistic model used for the

prior is not accurate. Since the prior probability and data calculation determine the posterior probability, there is a certain error rate for classification decisions.

2.2. Random Forest

Random forest is a classic Bagging - based algorithm. The central idea of bagging algorithm is to combine several learners without strong dependence. The performance of each learner may be relatively common, but after combining them together through the bagging algorithm, the advantages of each learner can be used to achieve better performance. As a set of decision trees of traditional learners, the performance of random forest is better than that of decision trees. Traditional decision tree selects an optimal feature (p features as well) from the current node feature set [9]. RF firstly randomly selects a subset containing L features from each base decision tree node feature set, and then selects an optimal feature from the sub-set for partitioning. The degree of randomness is regulated by parameter L : if $l = P$, the construction of decision tree is consistent with that of traditional decision tree. If $l = 1$, a randomly selected feature is divided. Usually, l is equal to $\log_2 p$. Random forest combines the two learning methods of averaging and voting. Average method is for numerical value, voting method is for classification variables, random forest can deal with both types of variables. Voting is the most commonly used combination strategy for classification. Let the class set be $\{\lambda_1, \lambda_2, \dots, \lambda_N\}$, and the predicted output table on sample x may be an N -dimensional vector $(k_1^1(x), k_1^2(x), \dots, k_1^N(x))$, where $k_i^j(x)$ represents the output of k_i for category λ_j .

$$\Lambda(x) = \begin{cases} \lambda_j, & \sum_{i=1}^M k_i^j(x) > 0.5 \sum_{l=1}^N \sum_{i=1}^M k_i^l(x) \\ \text{reject, others} & \end{cases}$$

That is, if a mark gets more than 50% of the votes, it will be predicted as that category. Otherwise, it will be rejected. For numerical output $(x) \in R$, averaging is the most common combination strategy.

$$\Lambda(x) = \sum_{i=1}^M (\Omega_i / M) k_i(x)$$

Where Ω_i / M is the weight factor of individual learner, which is usually required to be greater than or equal to 0, and $\sum_{i=1}^M \Omega_i = 1$.

Compared with some regression algorithms, random forest can process high-dimensional data without feature selection, but it can also get important features. Because of the independent training between trees, random forest is easy to parallelize and can run efficiently in large databases. Random forest uses unbiased estimation and relies on multiple classifiers, which means that it can avoid noise interference and over-fitting as much as possible.

Because of its excellent performance, random forest has been widely used in many fields. A previous study showed [9], T. L. Octaviani and Z. Rustam used a random forest approach to predict a Wisconsin breast cancer dataset. They first specified N guide tree samples in the original data set, randomly selected M tree variables on each tree node for separation, collected information from N trees, predicted new data, and finally achieved more than 99% classification accuracy. However, random forest cannot be a perfect algorithm, because researchers cannot control the internal operation of the model, random forest is more like an unknown black box for researchers. Moreover, the principle of random forest is to select among different parameters and different seeds, so the structure of random forest may be unreliable in the face of data with fewer features.

2.3. K-nearest Neighbor

KNN is an algorithm that does not use parameters and is one of the simplest algorithms [10]. The principle of KNN is to classify new input data by measuring the distance between input value and each value on the basis of correctly classified training set. If the new data is similar to k data in the training set, the new data and K data are grouped together [11]. The algorithm of KNN determines

that KNN is not sensitive to outliers and can be used for nonlinear classification. In KNN, the only tunable parameter is the k value, which can be selected by grid search. Previous studies showed that [12], the author chose $K=13$ to predict breast cancer, and PCA method was used to reduce the dimension of data samples, reducing the training time and operation complexity of the model, and the final accuracy of survival rate prediction reached 90.44%. The principle of KNN is simple, but its defects are also obvious. The performance of KNN largely depends on the selection of K value. When the selected K value is larger than the optimal value, a classification will contain too much data, resulting in the model being insensitive to outliers and large classification deviation. Because KNN classifies the input values according to the original data, when one category contains a large amount of data, while other categories contain only a small part of the data, a large number of different input values will be divided into the same category. On the contrary, if the selected k value is partial, a model with large variance will be obtained, leading to overfitting. In addition to prediction performance defects, KNN requires a large amount of space to store training samples before running, and also requires a large amount of computing time to calculate the distance between each input value and each training set data, which is not friendly to large samples.

3. Comparison

KNN starts work only during the test/evaluation phase, comparing a given test observation with the most recent training observation [13], which takes a lot of time to compare each test data point. As the dimension increases, the average distance between sample points increases logarithmically. Therefore, with the increase of dimension, KNN algorithm will generate a lot of space, which makes its classification ability weak. KNN dependence of training data is particularly big, when the samples were distributed evenly or gather, KNN cannot give full play to the advantages of its search neighborhood, at the same time, if our training data set, just need to have a small amount of error classification of numerical, the prediction data will be inaccurate and the training data will have poor fault tolerance. [14]. It can be concluded that although KNN has achieved a good prediction effect in this sample, it is not necessarily universal. KNN is suitable for data with small sample size and typical distribution.

The random forest algorithm has a very good effect on classification problems, and in most cases the effect is far better than SVM, log regression, KNN and other algorithms [15]. Random forest has a wide application range and can deal with classification variables and numerical variables at the same time. In general, a random forest combines the predictions of multiple decision trees into a single model [11], and the logic is that a model made up of many mediocre models is still better than a single good model. Random forest greatly reduces the risk of overfitting by means of average decision tree. Compared with KNN, random forest has obvious advantages, mainly including three aspects. First, random forest algorithm is a very stable algorithm, and the appearance of new data points in the data set will not affect the whole algorithm. Only one decision tree will be affected. It is difficult to influence all decision trees. If more than half of the basic classifiers are wrong, the random forest will only make wrong predictions. Second, random forest is easy to parallelize, so it saves a lot of computing time. Third, random forest can process high-dimensional data and can be effectively applied to big data.

The application scenarios of these three algorithms are different. KNN algorithm is applied to: when a particularly easy to explain model is needed, such as the recommendation algorithm that needs to explain the reason to users. Naive Bayes is used when a model that is easy to interpret and has little correlation between dimensions is needed to efficiently process high-dimensional data, although the results may not be satisfactory. Random forest is suitable for: when the data dimension is relatively low (dozens of dimensions) and the accuracy is highly required, it can achieve a good effect without much parameter adjustment. Basically, when you don't know what method to use, you can try random forest first.

4. Conclusion

In this paper, the advantages and disadvantages of three models including KNN, random forest and Naïve Bayse are analyzed from the perspective of breast cancer prediction, which will provide help for the future research of breast cancer data prediction. KNN algorithm is simple and easy to use, easy to understand, and mature in theory. However, when the sample size is large, it takes a long time to calculate, occupies a large space, and is prone to make wrong judgments due to the influence of error values. Naive Bayes algorithm has simple logic and is relatively stable, but when the number of attributes is large or the correlation between attributes is large, the classification effect is not good. Random forest can solve both classification and regression problems, and can better adapt to different types of data with high fault tolerance. This paper analyzed the predictive effectiveness of three classifiers on a Wisconsin breast cancer dataset. We hope to evaluate the characteristics and effects of KNN, random forest and naive Bayes models by analyzing the prediction results, so as to determine the best classifier for this dataset. Machine learning, as a new technology, is more and more widely used in the diagnosis and analysis of diseases. If it is to be used in clinical data diagnosis in practice, it is necessary to select a more suitable classifier according to the characteristics of collected data. In real life, there may be many imperfections in clinical data, such as many missing values or a small number of error information. Therefore, when comparing the performance of several classifiers, the performance of classifiers in such unexpected situations should be included in the scope of examination. From the prediction results, the three classifiers have excellent performance, and their prediction accuracy reaches more than 90%, indicating that they have good adaptability to this data set. However, in order to apply the model to a wider range of fields, the defects of the classifiers need to be optimized respectively. According to the results, random forest is the most stable data set among the three classifiers, which can cope with more situations and thus avoid diagnostic errors to the maximum extent. Therefore, we believe that random forest is the most suitable classifier for this data set. Although the performance of KNN and Naive Bayes in this data set is slightly inferior to that of random forest, it does not mean that these two classifiers are not competent for clinical diagnosis. The algorithms of KNN, stochastic forest and Naive Bayes each have their own advantages, which can be exploited and avoided by subsequent algorithm optimization. In the future, we hope to test the stability and prediction accuracy of the model by applying several classifier models to a wider data set, and optimize the corresponding algorithm according to the performance of several classifiers on the data set, so as to promote the integration of machine learning and medical treatment and help doctors make more accurate judgments.

References

- [1] Spratt J S, Spratt J A. (1985) What is breast cancer doing before we can detect it? [J]. *Journal of Surgical Oncology*, 30 (3): 156 - 160.
- [2] Nounou M, Elamrawy F, Helal N, et al. (2015) Breast Cancer: Conventional Diagnosis and Treatment Modalities and Recent Patents and Technologies [J]. *Breast Cancer Basic & Clinical Research*, 2015: 17 - 34.
- [3] International Agency for Research on Cancer. 4 February 2021. <https://www.iarc.who.int/news-events/world-cancer-day-2021/> [2022-5-4].
- [4] Bhise S, Bepari S, Gadekar S, et al. (2021) Breast Cancer Detection using Machine Learning Techniques [J]. *International Journal of Engineering and Technical Research*, 10 (7): 98.
- [5] Khosravi K, Golkarian A, Tiefenbacher J P. (2022) Using Optimized Deep Learning to Predict Daily Streamflow: A Comparison to Common Machine Learning Algorithms [J]. *Water Resources Management: An International Journal, Published for the European Water Resources Association (EWRA)*, 36.
- [6] Naji M A, Aarika S, Benlahmar E, et al. (2021) Machine Learning Algorithms for Breast Cancer Prediction and Diagnosis [C]. *International Workshop on Edge IA-IoT for Smart Agriculture (SA2IOT)*.
- [7] MD Ganggayah, Taib N A, Har Y C, et al. (2019) Predicting factors for survival of breast cancer patients using machine learning techniques [J]. *BMC Medical Informatics and Decision Making*: 19 (1).

- [8] Karabatak M. (2015) A new classifier for breast cancer detection based on Nave Bayesian [J]. Measurement, 72.
- [9] Song, Q., Liu, X., Yang, L. (2015) The random forest classifier applied in droplet fingerprint recognition [C]. In: 2015 12th International Conference on FSKD. pp.722 – 726.
- [10] Kumar A, Poonkodi M. (2019) Comparative Study of Different Machine Learning Models for Breast Cancer Diagnosis [J].
- [11] Sharma S, Aggarwal A, Choudhury T. (2018) Breast Cancer Detection Using Machine Learning Algorithms [C]. 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS).
- [12] Rajaguru H, Sannasi C. (2019) Analysis of Decision Tree and K-Nearest Neighbor Algorithm in the Classification of Breast Cancer [J]. Asian Pacific Journal of Cancer Prevention, 20 (12): 3777 - 3781.
- [13] Prastyo P H. (2020) Predicting Breast Cancer: A Comparative Analysis of Machine Learning Algorithms [C]. Proceeding International Conference on Science and Technology.
- [14] Xing W, Bei Y. (2019) Medical Health Big Data Classification Based on KNN Classification Algorithm [J]. IEEE Access, 2019 (99): 1 - 1.
- [15] Mashudi N A, Rossli S A, Ahmad N, et al. (2021) Comparison on Some Machine Learning Techniques in Breast Cancer Classification [C]. 2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES).