A Comparative Study on Machine Learning Algorithms for Predicting Breast Cancer Prognosis in Improving Clinical Trials

Neetu Sangari
School of Computer Science
Colorado Technical University
Colorado Springs, CO, USA
neetu.sangari@student.ctuonline.edu

Yanzhen Qu
School of Computer Science
Colorado Technical University
Colorado Springs, CO, USA
yqu@coloradotech.edu

Abstract—In recent years, machine learning algorithms have been more and more used in healthcare industry, especially in research areas involving human participants such as clinical trials and areas where the data is too expensive to collect. This research project has conducted a comparative study on three well-known machine learning methods: Logistic Regression (LR), Support Vector Machine (SVM), and Naïve Bayes (NB) against the same dataset for predicting breast cancer prognosis in improving clinical trials. The experiment results have provided a comprehensive view of the patient’s risk levels and risk factors to clinicians that benefit in effective and efficient treatment. This research has also demonstrated that different machine learning algorithms against the same dataset for breast cancer prognosis can have a difference in both performance and accuracy. Therefore, the comparative study on different machine learning algorithms can identify the most suitable machine learning algorithm to achieve cost-effective clinical trials.

Keywords—data augmentation, machine learning, overfitting, Logistic Regression, Support Vector Machine, Naïve Bayes

I. INTRODUCTION

The context of this research is that a machine learning (ML) based tool can be developed to assist breast cancer prognosis and to improve clinical trials. Breast cancer is the second leading common cancer in women. In the US, it is estimated that there are about 250,000 cases of breast cancer are diagnosed in women and about 2,300 in men [1]. There are about 42,000 women, and 510 men in the US die from breast cancer per year [1]. Many factors cause breast cancer, but some factors that can’t change are: Getting Older, family history, but taking care of health can lower breast cancer risk. Routine check-up, including breast cancer screening, is vital for early detection [2]. Many machine learning algorithms have been applied to build tools to assist breast cancer prognosis. In this research, we aim to identify the difference in performance and accuracy among three well-known machine learning algorithms against the same dataset, which is used in the prediction of breast cancer. The data used by this research can be collected in routine blood analyses - notably, Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin, MCP-1, Age, and Body Mass Index (BMI) [3]. Patricio et al. stated that resistin glucose, age, and BMI data might be used to predict the presence of breast cancer [4].

II. RELATED WORKS

Accurate breast cancer prediction models help early diagnosis help clinicians to develop personalized treatments and reduce death rates. Goldstein et al. mentioned that many emphases were placed on the development and application of machine learning algorithms, not much effort spent on model performance evaluation [5]. For a simple ML algorithm like the linear regression, there are simply explainable available to evaluate model performance. But the machine learning algorithms like support vector machines, naive bayes, etc., there are no uniform and acceptable metrics available for the model performance evaluation [6]. Additionally, most of the research in ML-based prognostic tools in oncology has come from technology-focused labs, and there are several concerns about their applicability in the real world, and the statisticians and clinicians are not comfortable in adopting machine learning model results. This is also noted in Parikh et al., where authors highlight the need to investigate if ML prognostic algorithms improve traditional regression models, especially in oncology settings [7].

A clinical trial is time-consuming as it relies on recruiting patients that also delay reporting the results of the clinical trial [8]. Clinical trials are too expensive and challenging as each stage of clinical trials needs to generate accurate and sufficient results to move on to the next phase [9]. Artificial intelligence and machine learning can help improve clinical trials. Novel machine learning techniques can help the state-of-the-art in several pattern recognition tasks [10].

III. PROBLEM, HYPOTHESIS, AND RESEARCH QUESTION

A. Problem Statement

The breast cancer death rates are high as some cancer patients do not have symptoms at all until the later stage and do not have robust detection tools to find the signs at the early stage. The early detection of cancer using machine learning algorithms could save the patient’s life by providing treatment at the early stage of cancer.
We have discovered that presently there is no reported research work on which machine learning algorithm is most accurate and effective in predicting breast cancer prognosis.

Two possible reasons for such a gap exist are (a) expertise: whoever to conduct such research must have the knowledge and experience in both oncology and computer science; (b) complexity: it requires significant effort to evaluate many possible machine learning algorithms against many possible cancer indicators, and compare the performance and accuracy of all the evaluated machine learning algorithms to discover the most effective one.

B. Hypothesis Statement

There is the most effective machine learning algorithm among all evaluated machine learning algorithms for accurately predicting breast cancer.

C. Research Question

Among all evaluated machine learning algorithms, which one is the most effective machine learning algorithm for predicting breast cancer accurately?

IV. MACHINE LEARNING ALGORITHMS

A. Support Vector Machines

SVMs are a class of highly flexible techniques that can be used to learn a variety of linear and non-linear representations such as splines, radial basis functions, and so on. This makes SVMs highly adaptable methods that can be used on new data inputs with linear or non-linear relationships (Kuhn & Johnson, 2013).

\[ f(u) = \beta_0 + \sum_{i=1}^{n} \alpha_i K(x_i, u) \]  

(1)

Where K(x_i, u) is the kernel function, since the predictor variables have a non-linear structure, we have used if a radial basis function provides the best fitting results.

B. Naïve Bayes

Naïve Bayes is a probabilistic supervised learning method based on the Bayes rule and conditional independence. Let Y denote a vector of response variables such as control (no cancer) and target (cancer), and X denote the explanatory variables, e.g., age, BMI, Glucose, etc. The goal of the machine learning exercise is to learn about the unknown probability distribution of the data generating process and try to find relationships between Y and X. In other words, given the data (X) the Naïve Bayes aims to classify if the patient has cancer or not (Y(\{1,2\})). Using the conditional probability and the total probability, the Bayesian rule is represented as below:

\[ P(Y = y_k | X_1, ..., X_n) \]

\[ = \frac{P(X_1, ..., X_n | Y = y_k) P(Y = y_k)}{\sum_{j=1}^{K} P(X_1, ..., X_n | Y = y_j) P(Y = y_j)} \]

(2)

Where P(Y=y_k) are the prior probabilities and represent the probability of occurrence of each class, e.g., P(Y=target) and P(Y=control), the prior probabilities reflect the subjective beliefs and encode prior knowledge and are usually estimated from the observed proportions from the training set.

The above formulation poses intractable complexity even for a simple classification setting because the model needs to estimate a large number of parameters. Naïve Bayes simplifies this complexity by assuming conditional independence, which implies each X_1 is conditionally independent of all other X_i's given Y. As a result, the Bayesian rule (Equation (2) simplifies to:

\[ \frac{P(Y = y_k | X_1, ..., X_n)}{\sum_{j=1}^{K} P(X_1 | Y_j) P(Y_j)} \]

(3)

C. Logistic Regression

Logistic regression is fitted mainly for parameter selection. This method is widely used in medical statistics to identify predictor variables that could be used in model development. Another use of Logistic regression is to calculate odds, which represent a multiplier of change in the value for a unit change in the predictor value.

\[ \log(\text{odds}) = \logit(P) = \ln(P/1-P) \]  

(4)

V. EXPERIMENT AND RESULTS

In this section, we conduct a quantitative comparison of three well-known machine learning algorithms, such as SVM, Naïve Bayes, and Logistic Regression, to discover the most effective machine learning algorithm for accurately predicting breast cancer prognosis. We ran each of the selected machine learning algorithms against a set of selected breast cancer indicators such as Glucose, Insulin, BMI, and Resistin to discover the correlation between the change of breast cancer indicator and the given machine learning algorithm for predicting breast cancer. Once we got the correlation among the set of selected breast cancer indicators, the study discovered the most effective machine learning algorithm among all the evaluated machine learning algorithms for predicting breast cancer. Comparing the performance and accuracy of all the machine learning algorithms determine the most effective machine learning algorithm.

A. Original Dataset

We used breast cancer data that is archived in the public domain of the University of California, Irvine (UCI) for research. The data source was from the Gynecology Department of the University Hospital Centre of Coimbra (CHUC) between 2009 and 2013 [4]. The data set consisted of 116 total records, of which 52 subjects are healthy volunteers, i.e., the control group, and 64 subjects are breast cancer patients, i.e., the target group. For the exploratory data analysis, we split the original 116 subjects’ data population into training and testing samples for the evaluation of different machine learning algorithms using R software.

Below is the Logistic Regression Final Model, the significant predictor variables.
TABLE 1. Logistic Regression on Original Dataset

B. Augmented Dataset

The original data with 116 records appears to be a small sample for machine learning algorithms as we have found significant overfitting for different training and testing data set compositions. We have decided to augment the original dataset and re-run the model fitting. Even though there are numerous data augmentation methods, most of those are not suitable for the dataset type. The goal was to augment the data set to preserve original features such as the correlation structure and relationships.

TABLE 2. Logistic Regression on Augmented Dataset

We identified three clusters from which we selected one variable as an imputation candidate. This way, the imputation method filled the removed values using the information available from other variables belonging to the same cluster. The initial analysis highlighted Glucose, BMI, and Resistin as significant predictor variables that were used as candidate variables for imputation. Predictive mean matching (PMM) with bootstrapping is used for data imputation and create 300 records needed to augment the original data set.

C. Experiment Results of Support Vector Machine

SVMRB is a maximum margin method that is very sensitive to the scale of data points. Therefore, the input data was scaled...

TABLE 3. SVM on Original Dataset

and centered on improving model performance. The model fitted on the original data set had clear indications of massive overfitting which were reflected in a very high number of support vectors - 53 out of 80 data points. Also, the cost function (C) value was very low, indicating a very low penalty for incorrectly classifying training data. The performance improved significantly when the model was fitted on augmented data: the proportion of support vectors used for the model fit reduced from 66% to 36%.

Analysis of the confusion matrix also confirmed that the model fit and performance improved with the augmented data set. The accuracy, a measure of correctly classified test cases, has improved from 83% to 88%, and the confidence intervals have tightened as well, indicating lower variation in the predictive model performance. This improved performance is mainly the result of sensitivity, which increased to 91% from 78% of the original data set model. This clearly indicates the model fitted with the augmented data set is much better at classifying the control cases than the one with a smaller data set.
TABLE 4. SVM on Augmented Dataset

TABLE 5. Naïve Bayes on Original Dataset

The model assumes conditional independence to make it tractable. This assumption proved too restrictive for our data set, which had highly correlated predictor variables causing poor model performance. For example, No Information (NI) rate is the accuracy that can be achieved without a model [11]. It was higher than the model’s accuracy with the original data set, suggesting a very weak model predictive power. The situation improves when Naïve Bayes is fitted with an augmented data set: the accuracy improves and exceeds the NI rate but still very low.

TABLE 6. Naïve Bayes on Augmented Dataset

The confusion matrix appears the performance has deteriorated when the model is fitted with augmented data. The accuracy has decreased to 75% from 81%. This is a consequence of a reduction in sensitivity and specificity. The sensitivity has decreased by around 10%, suggesting the model fitted with augmented data has difficulties in classifying control cases. The reduction in specificity has been quite small - around 3%. However, the 95% confidence interval has improved slightly, indicating some reduction in the model variability.

Kappa, a measure of agreement between two label generation processes or classifiers, has significantly decreased to 50% from 60%. One possible explanation could be that the augmented data set does not satisfy the logit assumption very well, resulting in a poor model fit.
F. Summary

The research uses biomarker data of routine blood analyses - Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin, MCP-1, Age, and Body Mass Index (BMI) to predict the presence of breast cancer. We have conducted a quantitative comparison on a set of well-known machine learning algorithms such as SVM, Logistic Regression, and Naïve Bayes on the data. The study discovers SVM, the most effective machine learning algorithm for predicting breast cancer prognosis accurately as per the data analysis. We have also compared the predictor variables’ estimates with clinical research findings. We have shown the variable rankings based on the original and augmented data.
  - Glucose is the topmost indicator in all the methods in the augmented data.
  - Resistin is on top #2 or #3 indicator in all the methods.
  - BMI is on top #4 or #5 indicator in all the methods.
  - Insulin is on top #4 or #5 in all the methods except Logistic regression.
  - Only SVMRB and NB ranks HOMA on the top #3 indicator.

We have concluded that the research study results support breast cancer indicators, i.e., Glucose, Resistin, BMI, HOMA, and Insulin, which impacted the outcome of the success rate of predicting breast cancer of a machine learning algorithm. The important predictor variables are also in line with those reported in Patricio et al., 2018. SVMRB is the best among all of the machine learning algorithms as per metrics on the performance measure, and SVMRB is flexible and simpler among all evaluated algorithms. We have also concluded that the research study results supported the hypothesis that SVMRB is the most effective machine learning algorithm among all evaluated machine learning algorithms to accurately predict breast cancer.

VI. CONCLUSION

The prediction of breast cancer prognosis using machine learning algorithms collected on the routine blood test lab provides crucial information for the patients and clinicians that saves a life. We have used only 116 records, i.e., a small dataset using machine learning...
methods, but there was difficulty in training and testing data splitting. The model results trained on the small dataset indicates severe overfitting and a lack of generalizability of models. We have augmented the data around three times of the original dataset using three machine learning methods employed for the research study. The model fitting improved for most of the machine learning algorithms except LR, for which the performance deteriorated. There are noticeable improvements in SVMRB on model prediction. The early detection of breast cancer using machine learning algorithms contributes a major role in clinical trials. The collaboration of researchers, statisticians, clinicians, and industries can expedite drug development and increase care quality and save patients from cancer. In conclusion, this comparative study has indeed demonstrated that different machine learning algorithms against the same dataset for breast cancer prognosis can have a difference in both performance and accuracy. Therefore, we have approved our hypothesis and answered our research question.

For future work, we are looking into robust our model, including electronic health records, to predict breast cancer. We will also apply machine learning algorithms to another type of cancer or disease through working with more statisticians and clinicians in the future.

REFERENCES