COMPARING SUPPORT VECTOR MACHINE AND NAIVE BAYES ALGORITHMS IN CLASSIFYING BREAST CANCER CASES

A

PREDICTIVE ANALYSIS PROJECT

Presented to

Department of Information Technology

Of the College of Computer Studies

Mindanao State University - Iligan Institute of Technology

In Partial Fulfillment

of the Requirements for ITD105

BIG DATA ANALYTICS

ABDULRAHMAN U. LINGGA

2021

CHAPTER 1

INTRODUCTION

1.1 Rationale

Cancer remains to be the leading cause of death in the world with breast cancer among the most common (World Health Organization [WHO], 2021). This forms when the cells found in the breast grow out of control (Centers for Disease Control and Prevention [CDC], 2021). Although this is not a transmissible and infectious disease, certain factors increase the risk of developing breast cancer such as increasing age, genetics, and unhealthy lifestyle (WHO, 2021).

According to WHO (2021), there were around 2.3 million women diagnosed with cancer and 685,000 deaths caused by the said disease in 2020 alone. In February 2021, WHO confirmed that breast cancer has become the most common form of cancer accounting for almost 12% of new cases each year (Thomson & Nebehay, 2021). From the same report, the number of cancer patients is said to expect a rise to around 30 million new cases every year in 2040. In the Philippines, 16% of cancer diagnoses were accounted by breast cancer, and it was estimated 3 out of 100 Filipino women would develop this type of disease in their lifetime (Cudis, 2019).

In the fight against cancer, early detection plays a very important role for increase in chances of survival. WHO (2021) stated that patients diagnosed early are more likely to have better results in treatment and higher chances of survival. Breast cancer can be confirmed by going through tests like ultrasound, mammogram, magnetic resonance imaging, and including biopsy (CDC, 2021). By these means, cancer can be diagnosed as benign (noncancerous) or malignant (cancerous). With the advancement of technology, computer algorithms like machine learning are widely used in different industries including in health to generate useful information and help authorized people make decisions.

Classification is one of the fundamentals of machine learning and data science (Zhang, 2004). This aims to construct a classifier model from clearly labeled training examples (Zhang, 2004). With numerical scores recorded from previous studies or cases, this

can be used to create accurate classification to help doctors and other health workers easily detect breast cancer in its earlier stage.

Therefore, this project utilized machine learning algorithms with data gathered from digitized images of fine needle aspirate (FNA) of breast mass of breast cancer patients. Using Support Vector Machines (SVM) and naive Bayes, two of the most common classification methods, comparisons were performed to select the best fit model in classifying cases as benign or malignant. This was to help doctors create accurate diagnoses and decisions that are vital for increasing the chances of survival of breast cancer patients. In general, this project holds significance in the fight against breast cancer as well as in support of existing studies related to cancer detection using SVM and naive Bayes.

1.2 Dataset Description

This project used the Breast Cancer Wisconsin dataset to create a classification model. The dataset is publicly available in the UCI Machine Learning Repository (Mangasarian & Wolberg, 1990). It contains 699 instances and 11 attributes with 16 missing values. Further, 241 of the instances are classified as malignant and the other 458 as benign cases.

The values and features of the dataset describe the characteristics of the cell nuclei obtained from the digitized image of a breast mass in an FNA biopsy. The attributes are ID number, clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, mitoses, and the class. The class is valued at 2 for benign and 4 for malignant, while the rest of the attributes, except for the ID number, are valued between 1 and 10.

The dataset is used for the project as it has the qualities of a good dataset. Features and the class are clearly determined and also contain numerical values. It is also distinctly identified that the dataset is for classification and is meant to classify benign and malignant cases in breast cancer. Finally, information on how the data was gathered had also been well provided.

CHAPTER 2

METHODOLOGY

2.1 Procedures

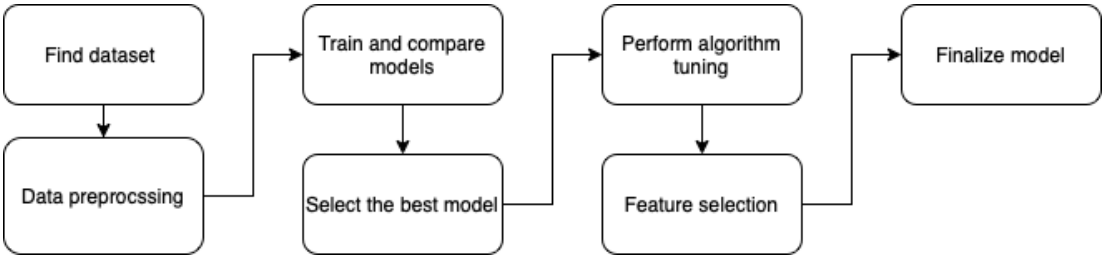

Figure 1. Proposed Methodology for Accurate Breast Cancer Cases Classification

The first step in this project was to find the dataset to be used and perform data preprocessing. In this project, the dataset was collected from the UCI Machine Learning Repository. With 16 missing values in the *bare_nuclei* attribute, rows were deleted which made the number of instances to be 683. The UCI Machine Learning Repository was first created in 1987 and now acts as a repository for databases, theories, and data generation that are widely used in the community of machine learning (Dua & Graff, 2019).

Next, models were trained with machine learning algorithms namely, SVM and naive Bayes. The dataset was divided using *k*-cross validation where it is split into 10 folds. The 9 folds were used for training while the other was used for testing. This is repeated until each fold is used for testing. The performance of each model was then measured for comparison to select the best fit model.

After choosing the model, this was followed by improving its performance using algorithm tuning. In the case of SVM, the parameters kernel and *c* would be adjusted whereas the smoothing parameter would be tuned for naive Bayes. Next, features selection using the algorithm of univariate selection was used to select four features from the dataset. Features that obtained the four highest scores were used to train the model. Finally, the final trained model was measured for its accuracy.

2.2 Classification Methods

In this project, SVM and naive Bayes were utilized to find the best fit model for breast cancer detection.

2.2.1 Support Vector Machines

Support Vector Machine developed by Vapnik in 1995, is a statistical learning theory that aims to find the best line or hyperplane that could separate the data according to its class (Satapathy et al., 2019; Dukart, 2015). For example, for a binary classification problem, a line is drawn in a two-dimensional plane to separate training data while also keeping the misclassification to a minimum. There can be infinite lines that provide a fine separation between the data, however, the best line is chosen to one that maximizes the margin between the two classes (Dukart, 2015). In other words, the line with the biggest distance from the closest data points in each class, or also known as support vectors, is selected. This concept extends to higher dimensional space where the dataset contains more than just two features and the optimal hyperplane is chosen instead of a line (Dukart, 2015). SVM has been widely used in machine learning to solve a wide range of classification problems (Satapathy et al., 2019).

2.2.2 Naive Bayes

Naive Bayes is the simplest probabilistic classifier in the Bayesian network (Zhang, 2004). This considers all features as independent given the value of the class variable (Shobha & Rangaswamy, 2018; Zhang, 2004). According to Zhang (2004), the probability of an example $E = (x_1, x_2, \dots, x_n)$ being class c from the set of two classification variables C is,

$$p(c|E) = \frac{p(E|c)p(c)}{p(E)} \quad (1)$$

Equation (1) aims to use available evidence to calculate the validity of a hypothesis. For example, an apple is considered to be red, round, and around 10 cm in diameter. Regardless of any possible relationships between the color, shape, and diameter, a naive

Bayes classifier examines each of these parameters separately to the likelihood that the given element is from a particular class variable (Shobha & Rangaswamy, 2018).

2.3 Performance Measure

To determine the best fit model, performance metrics namely, accuracy, logarithmic loss, and area under the receiver operating characteristic (ROC) curve (AUC) were used. Accuracy describes the number of correct predictions a model has made over the total number of predictions. In the case with binary classification, accuracy is defined as

$$\frac{TP + TN}{TP + TN + FP + FN}, \qquad (2)$$

where TP, FP, TN, and FN stand for true positive, false negative, true negative, and false negative respectively. The true positive refers to the result that the model correctly identified the positive class whereas the true negative is the result that correctly categorized the negative class. On the other hand, false positive and false negative are the outcomes when the model incorrectly predicts the positive and negative classes respectively.

With the logarithmic loss, the performance of the classifier is measured by penalizing false predictions. This is defined by the following formula:

$$-\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\, log\, p_{ij} \qquad , \qquad (3)$$

where N and M are the numbers of instances and class values respectively, y_{ij} is the correct or wrong predictions for instance i for label j, and lastly, p_{ij} is the probability of the model in assigning class value j to instance i. On the other hand, the AUC calculates the area under the true positive rate $\frac{TP}{TP + FN}$ and the true negative rate $\frac{FP}{FP + TN}$ to tell how a model can distinguish between classes. Finally, the model with the highest cumulative scores was selected and chosen to be the fit model in detecting breast cancer cases.

CHAPTER 3

RESULTS AND DISCUSSION

Table 1 shows the results of performance scores of each model from each scoring. These were summarized by taking the mean of each result of the 10-fold cross-validation. Based on the comparisons, both models do not have big dissimilarities with each other with only a few differences from each scoring. However, it is shown that SVM dominated each scoring compared to naive Bayes. Overall, the SVM gained a cumulative score of 1.86 and naive Bayes with 1.24. Hence, SVM was chosen as the best model.

	Support Vector Machine	**Naive Bayes**
Accuracy	0.967860	0.963427
Logarithmic Loss	-0.100739	-0.824068
Area Under ROC Curve	0.988249	0.984277
Total	1.86	1.24

Table 1. Performance Scores

To improve the performance of the selected model, algorithm tuning was conducted. Based on the results, the radial basis function (RBF) was chosen as the best kernel among linear, polynomial, and sigmoid kernels. Further, 0.01 was chosen as the regularization parameter against other values. The accuracy scoring obtained from the hyperparameters was 0.969309, a 0.001449 difference from the original score.

In feature selection, the attribute cell_size, cell_shape, bare_nuclei, and normal_nuclei acquired the highest scores among others. Meaning, these features were selected as the best ones using the univariate selection. Table 2 shows the summary of the results.

Features	**Score**
clump	624.135704
cell_size	1370.064587
cell_shape	1279.767704
marg_adhesion	986.417879

se_cell_size	497.536763
bare_nuclei	1729.066174
bland_chromatin	682.978239
norm_nuclei	1143.866712
mitoses	228.994346

Table 2. Scores of Each Feature from Univariate Selection

Finally, results from the previous methods were used to finalize the model. By training and testing the model with 77% and 33% of the instances respectively, it performed well by obtaining a 0.93 final accuracy score. The figure below presents the confusion matrix performed by the final model whereas table 3 shows its reports on precision and recall.

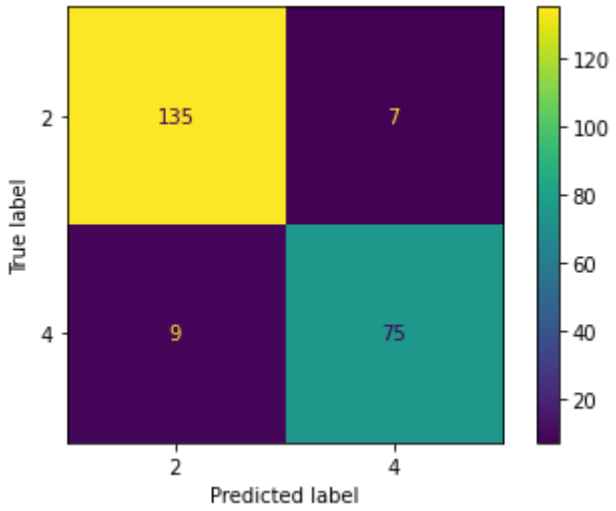

Figure 2. Confusion Matrix of the final SVM model

From the 226 cases in the testing set, 142 of them are classified as benign whereas the other 84 are malignant cases. This means that of the 84 malignant cases, the model classified 75 of them correctly while incorrectly labeling the other 9. Similarly, 135 benign cases were classified correctly by the model out of 142 cases but mislabeled the other 7 benign cases.

	Precision	**Recall**
Benign	0.94	0.95
Malignant	0.91	0.89

Table 3. Classification Report of the final SVM model

Table 3 shows the precision and recall reports of the model in each label. Precision refers to the proportion of the positive classification that was correct. This means that 94% of the time when the model predicts benign cases is correct. On the other hand, when the model predicts malignant cases, it is correct 91% of the time. Recall refers to the proportion of the actual positives that were classified correctly. This means that 95% of all benign cases and 89% of all malignant cases are correctly identified by the model.

The code and the final model for this project could be obtained from this link:

https://drive.google.com/drive/folders/1pMdKld2V0__uaoVKNAKWnbsLdvcQlljQ?usp=sharing

CHAPTER 4

FUTURE APPLICATIONS

Machine learning, specifically in classification, uses a vast range of algorithms to meet the needs of different applications. This includes detecting diseases using data gathered from previous cases. Hence, this project recommends exploring further other algorithms, such as logistic regression and artificial neural networks, for classifying breast cancer cases as cancerous or not.

Furthermore, ensembles, such as voting, bagging, and boosting ensembles, should also be utilized and added in comparison among other classification models. This allows wider options and a better and appropriate selection of the model. The researcher also suggests broadening the values in performing algorithm tuning as well as using other feature selection techniques to further improve the chosen model.

In addition to that, the Wisconsin Breast Cancer dataset only contains hundreds of samples with few features. On top of that, the recent data were last gathered in November 1991. Thus, the researcher recommends that more features and instances should be added to the dataset to help make a better and more accurate classifier. Finally, the findings of this project can be used in future studies in machine learning and healthcare, specifically related to breast cancer detection.

BIBLIOGRAPHY

Benett, K. P., & Mangasarian, O. (1992). "Robust linear programming discrimination of two linearly inseparable sets". *Optimization Methods and Software 1*, 23-34.

Centers for Disease Control and Prevention. (2021, September 22). *How Is Breast Cancer Diagnosed?* CDC. Retrieved December 16, 2021, from https://www.cdc.gov/cancer/breast/basic_info/diagnosis.htm

Cudis, C. (2019, March 6). Top diseases that kill women in PH. *Philippine News Agency*. https://www.pna.gov.ph/articles/1063773

Dua, D., & Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. *Irvine, CA: University of California, School of Information and Computer Science.*

Dukart, J. (2015). Basic Concepts of Image Classification Algorithms Applied to Study Neurodegenerative Diseases,. In A. W. Toga (Ed.), *Brain Mapping: An Encyclopedic Reference*. Elsevier Science.

Mangasarian, O., Setiono, R., & Wolberg, W. H. (1990). "Pattern recognition via linear programming: Theory and application to medical diagnosis", in: "Large-scale numerical optimization", Thomas F. Coleman and Yuying Li, editors. *SIAM Publications, Philadelphia*, 22-30.

Mangasarian, O., & Wolberg, W. H. (1990, September). "Cancer diagnosis via linear programming". *SIAM News*, *23*(5), 1 & 18.

Satapathy, S. K., Dehuri, S., Jagadev, A. K., & Mishra, S. (2019). Chapter 1 - Introduction. *EEG Brain Signal Classification for Epileptic Seizure Disorder Detection*, 1-25.

Shobha, G., & Rangaswamy, S. (2018). Handbook of Statistics (V. Gudivada, Ed.). *Elsevier*, *38*, 197-228.

Thomasson, E., & Nebehay, S. (2021, February 2). Breast cancer overtakes lung as most common cancer-WHO. *Reuters*. https://www.reuters.com/business/healthcare-pharmaceuticals/breast-cancer-overtakes-lung-most-common-cancer-who-2021-02-02/

WHO. (2021, March 26). *Breast cancer*. WHO | World Health Organization. Retrieved December 16, 2021, from https://www.who.int/news-room/fact-sheets/detail/breast-cancer

WHO. (2021, September 21). *Cancer*. WHO | World Health Organization. Retrieved December 16, 2021, from https://www.who.int/news-room/fact-sheets/detail/cancer

Wolberg, W. H., & Mangasarian, O. L. (1990, December). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proceedings of the National Academy of Sciences, U.S.A*, *87*, 9193-9196.

Zhang, H. (2004). The Optimality of Naive Bayes. *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2004*. https://www.cs.unb.ca/~hzhang/publications/FLAIRS04ZhangH.pdf