Application of the Naïve Bayes and SVM classification algorithms for the Breast Cancer Dataset analysis

Dmitriy Chuyko
Lakehead University
Thunder Bay, Ontario, Canada
dchuyko@lakeheadu.ca

ABSTRACT

Breast cancer, which causes the majority of deaths among women with cancer, requires an accurate diagnosis and prognosis for treatment. Since a huge amount of data has accumulated in medical practice, certain tools and methods are required to store and use this data. For this, classification techniques of machine learning are mainly used in which the machine is learned from the past data and can predict the category of new input. This paper is a comparative study on the implementation of two machine learning algorithms, the Naïve Bayes and the SVM, performed on the Wisconsin Breast Cancer Diagnosis dataset imported from the UCI Machine Learning repository. The results showed that the Naïve Bayes algorithm is the best at predicting the type of breast tumor with an accuracy of 0.9356 versus 0.8596 for the SVM.

ACM Reference Format:
Dmitriy Chuyko. 2020. Application of the Naïve Bayes and SVM classification algorithms for the Breast Cancer Dataset analysis. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Breast cancer, the most common cancer among women, affects more than 2 million women every year, and is also the cause of the highest number of cancer deaths among women [2]. Most of the lesions found on mammograms and most of the breast lumps are benign, do not grow uncontrollably or spread, and do not pose a threat to life [3]. Malignant tumors are dangerous because they will metastasize to other organs in the body, such as lungs, brains, bones or livers [5]. Accurate tumor prediction is critical to the diagnosis and treatment of cancer. Among the existing methods, supervised learning methods are the most popular in the diagnosis of cancer [5]. The SVM algorithm was recognized as the best compared to the K-Nearest Neighbor (KNN) and Logistic Regression [5], as well as among the Naïve Bayes, K-Nearest Neighbor (KNN) and Decision Tree classifiers [6]. In other papers, the Naïve Bayes algorithm has shown the highest accuracy and speed for the breast tumor type prediction among other machine learning methods such as the SVM [3]; SMO, Bayes Network and J-48 Decision [4]; Logistic Regression, Bayes Network, Multilayer Perceptron , Sequential Minimal Optimization, J-48 Decision and Instance Based Learner [1]. In this work, two different classifiers, the Naïve Bayes and the SVM, were compared using the Wisconsin Breast Cancer Diagnosis (WBCD) dataset [8] to determine a more efficient classification algorithm.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 SYSTEM DESCRIPTION

2.1 Proposed project architecture

The scheme of the proposed work in relation to the WBCD is shown in Fig.1. The WBCD dataset [8] was taken from the UCI Machine Learning Repository. The total number of instances represented in the WBCD dataset was 569 with 30 attributes, which were calculated based on ten geometric measurements of the Fine Needle Aspirate (FNA) test for each cell nucleus: radius, texture, perimeter, area, smoothness, compactness, concavity, concavity point, symmetry, and fractal dimension of each mass [7].The Naïve Bayes and SVM algorithms were used in this work.

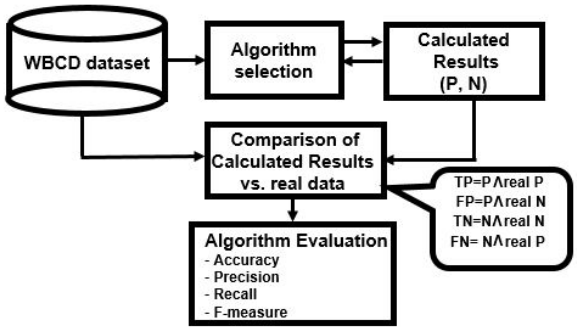

Figure 1: Project architecture

2.2 Naïve Bayes algorithm

The Naïve Bayes algorithm is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) assumptions of symptoms independence.

P(M|F)= P(F|M)P(M)/P(F) - Bayes' theorem (M-malignant tumor, F-feature), where

P(M|F) – the probability of "M" being true given that "F" is true
P(F|M) - the probability of "F" being true given that "M" is true
P(M) - the probability of "M" being true
P(F) - the probability of "F" being true

The Naïve Bayes classifier assumes that the presence of a particular feature fi is unrelated to the presence of any other feature, given the class variable, therefore:

P(F|M) = P(f1|M)*P(f2|M)…P(fn|M) , P(F) = P(f1) P(f2)…P(fn) and

P(M|F)= P(f1|M)*P(f2|M)…P(fn|M) * P(M)/ P(f1) P(f2)…P(fn)

The advantages of the Naïve Bayes classifier are that the Naïve Bayes is one of the fastest and simplest ML algorithms for class prediction of datasets, and it performs well in multi-class predictions compared to other algorithms. The disadvantages of the Naïve Bayes classifier are the assumption that all features are independent or unrelated, so it is impossible to know the relationship between features, and that the results might be biased. The Naïve Bayes classifier is used to classify medical data and might be employed in real-time predictions because this algorithm is an eager learner.

2.3 Support vector machine (SVM)

The SVM is a binary classifier meaning that it can use a decision function that will return "yes" or "no" for a given input data point. The goal of the SVM algorithm is to find a hyperplane in N-dimensional space (N is the number of functions that clearly classify data points). Hyperplanes, which size depends on the number of elements, are decision boundaries that can help to classify data points. The advantages of the SVM are high accuracy, less overfitting and using small and clean datasets; the disadvantages are that training time is high for large data and less efficient for noisy data.

The Naïve Bayes and SVM algorithms were converted and executed via VS Code using Python Development Environment. With respect to the result of Accuracy, Precision, Recall and F-Measure the efficiency of each algorithm was measured and compared.

2.4 Measures for performance evaluation

2.4.1 Accuracy. Accuracy measure represents how far the set of tuples are being classified correctly in according to the equation: Accuracy = (TP+TN)/(TP+FP+TN+TP).

2.4.2 Recall. The recall or true positive rate measures the proportion of positives that are correctly identified as such: Recall = (TP/TP+FN).

2.4.3 Precision. The precision or positive predictive value is the proportion of positive results in statistics and diagnostic tests that are true positive: Precision = (TP/TP+FP).

2.4.4 F-measure. The F-measure is a measure of the accuracy of the test and is calculated based on the precision and recall of the test according to the equation: F-measure = (2TP/2TP+FP+FN),

where TP is the number of true positives,

TN is the number of true negatives,

FP is the number of false positives,

FN is the number of false negatives.

3 RESULTS

The WBCD dataset is composed of 569 observations, where 357 are benign and 212 are malignant breast masses. The values for measuring the methods' performance (i.a. Accuracy, Precision, Recall, F-measure) are presented in Table 1 and plotted in Fig.2. The data obtained showed that the Naïve Bayes algorithm produced the higher accuracy, 0.9356, compared to the SVM accuracy of 0.8596. Other performance measures' values were found to be similar for the Naïve Bayes and the SVM algorithms, respectively.

Table 1: Algorithm performance comparison

	Naïve Bayes	SVM
Accuracy	0.9356	0.8596
Precision	0.9787	0.9672
Recall	0.9261	0.9365
F-measure	0.9517	0.9516

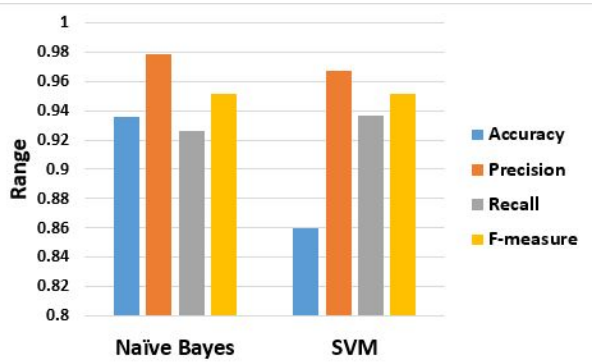

Figure 2: Classifiers performance measures evaluation

4 CONCLUSION

In this work, the Naïve Bayes and SVM were used to evaluate the best classification algorithm for breast cancer based on the WBCD dataset. The Naive Bayes Algorithm is considered the best method because it provides better accuracy than the SVM classifier. Further research is needed in this area to improve the efficiency of classification methods for predicting more variables.

REFERENCES

[1] E Jide et al. 2019. Breast cancer predictive analytics using supervised machine learning techniques. *International Journal of Advanced Trends in Computer Science and Engineering* 8, 6 (2019), 3095–3104.

[2] Shweta Kharya and Sunita Soni. 2016. Weighted naive bayes classifier: A predictive model for breast cancer detection. *International Journal of Computer Applications* 133, 9 (2016), 32–37.

[3] A Khatija and N Shajun. 2016. Breast cancer data classification using SVM and Naive Bayes techniques. *International Journal of Innovative Research in Computer and Communication Engineering* 4, 12 (2016), 21167–21175.

[4] R Megha and KS Arun. 2012. Breast Cancer Prediction using Naïve Bayes Classifier. *International Journal of Information Technology & Systems* 1, 2 (2012), 77–80.

[5] Ch Shravya, K Pravalika, and Shaik Subhani. 2019. Prediction of Breast Cancer Using Supervised Machine Learning Techniques. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)* 8, 6 (2019), 1106–1110.

[6] Ankita Sinha, Bhaswati Sahoo, Siddharth Swarup Rautaray, and Manjusha Pandey. 2019. Analysis of Breast Cancer Dataset Using Big Data Algorithms for Accuracy of Diseases Prediction. In *International Conference on Computer Networks and Inventive Communication Technologies*. Springer, 271–277.

[7] S Sountharrajan, M Karthiga, E Suganya, and C Rajan. 2017. Automatic classification on bio medical prognosisof invasive breast cancer. *Asian Pacific Journal of Cancer Prevention: APJCP* 18, 9 (2017), 2541–2544.

[8] William H Wolberg, W Nick Street, and Olvi L Mangasarian. 1992. Breast cancer Wisconsin (diagnostic) data set. *UCI Machine Learning Repository [http://archive.ics.uci.edu/ml/]* (1992).