Abstract

Breast cancer is one of the major threats to women nowadays. Early detection of breast cancer decreases mortality rate. Machine learning algorithms are used for this purpose. Accuracy is the most popular measure for evaluating machine learning algorithms for breast cancer diagnosis. However, it does not make a distinction between the performance of the classifier on malignant and benign test cases. This paper studies sensitivity and specificity along with accuracy to differentiate between KNN performance on malignant and benign classes for the different number of neighbors. Additionally, the standard deviations of sensitivity and specificity are studied to show KNN stability in malignant and benign classes. This study is critical because the cost of false negative is more than the cost of false positive in breast cancer detection. This study is conducted on Wisconsin breast cancer dataset (WBCD) from UCI repository. Stratified 10-fold cross-validation is used in this paper. Additionally, in order to increase the correctness of outcome, validation method repeated 100 times by considering that the samples are randomly reassigned to the folds again. The results show that KNN does not work well on malignant
samples compared to the benign test cases, especially for higher values of neighbors. Also, the results for malignant samples are less reliable than benign ones. Furthermore, accuracy is more representative of specificity than sensitivity. It seems that the imbalance distributions of malignant and benign classes make difference between KNN performance on malignant and benign samples. It is recommended that a new study to be conducted to show the effect of imbalance numbers of positive and negative samples and also the difference between standard deviations of positive and negative classes on KNN performance.

References


Index Terms

Computer Science Biomedical

Keywords

Breast Cancer Diagnosis, K-Nearest Neighbors, Imbalance Dataset, Sensitivity, Specificity