ANNALS OF "DUNAREA DE JOS" UNIVERSITY OF GALATI
MATHEMATICS, PHYSICS, THEORETICAL MECHANICS
FASCICLE II, YEAR XIV (XLV) 2022, No. 2
DOI: https://doi.org/10.35219/ann-ugal-math-phys-mec.2022.2.04

Performance comparison of two non-parametric classifiers for classification using geometric features

Simona Moldovanu[1,2], Iulia-Nela Anghelache Nastase[1,3], Mihaela Miron[2], Luminita Moraru[4]

[1] *The Modelling & Simulation Laboratory, Dunarea de Jos University of Galati, 47 Domneasca Street, 800008 Galati, Romania*
[2] *Department of Computer Science and Information Technology, Faculty of Automation, Computers, Electrical Engineering and Electronics, Dunarea de Jos University of Galati, 47 Domneasca Street, 800008 Galati, Romania*
[3] *Emil Racovita Theoretical Highschool, 12–14, Regiment 11 Siret Street, 800332 Galati, Romania*
[4] *Department of Chemistry, Physics & Environment, Faculty of Sciences and Environment, Dunarea de Jos University of Galati, 47 Domneasca Street, 800008 Galati, Romania*
[*] *Corresponding author:simona.moldovanu@ugal.ro*

Abstract
This study aims to examine and compare the performances of Random Forest (RF) and k-Nearest Neighbor (k-NN) algorithms used for classification based on certain geometric features. For the purpose of the analysis, the Breast Cancer Wisconsin (BCW) public dataset is used. BCW dataset contains features like area, perimeter, radius, compactness, and symmetry computed from 357 benign, and 212 malignant breast images, respectively. Three different experiments related to the size of training and testing datasets for classification are conducted and different accuracy values are obtained. The best accuracy of 91.9% for RF and 91.3% for kNN, respectively, are reached when 30% of the entire dataset is used as testing dataset. For all experiments, the RF classifier outperformed the kNN.

Keywords: Random Forest, k-Nearest Neighbor, Breast Cancer Wisconsin Data Set

1. INTRODUCTION

In this digital world, machine learning has raised its importance and value. Images contain a lot of useful data from which machine learning can take advantage and extract important features to solve various decision-based problems. Machine learning algorithms are a very useful solution when applied to big data as they offer techniques and tools that are effective in gathering important information from image datasets in a timely manner.

Due to the large number of classification methods/classifiers, the selection of the better classifier for a specific problem becomes an important task because a good prediction based on the meaningful features characterizing either texture or shape of breast lesions can help in some situations saving the life of the patients. El_Rahman [1] proposed a comparative study using different classifiers such as k-NN, RF, logistic regression (LR), support vector machine (SVM), logistic regression (LR) and Naive Bayes (NB). He used four different breast cancer datasets. The reported experimental results showed that the classification based on RF and Wisconsin diagnosis breast cancer dataset gave a higher accuracy value of 96.82%. The same dataset for breast cancer prediction with k-NN and NB, SVM, LR, Multilayer Perceptron (MLP), Softmax Regression (SR) was used in the paper [2]. The MLP algorithm stands out among the implemented algorithms with a test accuracy of 99.04%. Gopal et al. [3] used the IoT devices and machine learning techniques in predicting breast cancer. In this

study, the LR and RF classifiers were used. By comparing these classifiers, the authors have found that MLP yields a higher accuracy of 98% against the LR and RF algorithms. Akay [4] has used the BCW dataset and a SVM classifier, and has reported an accuracy of 99.51%. Nastase et al. [5] integrated six Hu's moments in the analysis of the breast tumour, a k-NN classifier, and a radial basis function neural network (RBFNN) to classify the malignant and benign breast tissues. The first Hu moment was the most relevant, and an accuracy of 85% has been reached using the k-NN classifier.

A machine learning classifier can be either based on parametric or non-parametric methods. A parametric classifier uses the statistical probability distribution of each class under investigation. Non parametric classifiers estimate the unknown quantities or density function and are used to determine the probability density function.

This study proposes two non-parametric classifiers which rely on decision trees or clustering, namely the RF and k-NN classifiers. The basic concepts are to compare the performance of both non-parametric classifiers. The input of these classifiers contains relevant features like area, perimeter, radius, compactness, and symmetry. These features characterize the lesion of the breast from a geometrical point of view. This article is organized as follows: Section II discusses on the datasets, used features and the classification methods and classifier's performance. The experimental results are given in Section III. Finally, the conclusions are pointed out in Section IV.

2. EXPERIMENTAL

The Breast Cancer Wisconsin Dataset is a public one and belongs to the University of Wisconsin Hospital, Madison created by Dr. William H. Wolberg [7]. The features, such as area, perimeter, radius, compactness, and symmetry are computed from a digitized image. All these features are extracted from 357 benign and 212 malignant images.

The classifiers are implemented in Python programming language, version 3.10.7 by using the following libraries: numpy, pandas, matplotlib, seaborn and sklearn.

The hardware used was a computer with the following specifications: Inter (R) Core (TM) i7-8550U CPU @ 1.80 GHz; Memory (RAM) 8 GB DDR4.

The following features, area, perimeter, radius, compactness, and symmetry feed the k-NN and RF classifiers, in order to analyze their performances. The accuracy value is computed with the results from the confusion matrix [7]:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}(\%) \tag{1}$$

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

The k-NN classifier is a supervised classification algorithm which clusters the instances based on their features similarity. A data/case/sample item is classified based on its similarity with a majority of its neighbors. In our study, two classes are considered, benign and malignant breast lesion.

The second classifier is RF and it is also a supervised classification algorithm. RF fits a number of decision tree classifiers on different sub-samples of the dataset. In our study, 100 decision trees and single tree step precision are choosen.

3. RESULTS AND DISCUSSION

To evaluate the proposed classifiers, three experiments are performed as follows:

Experiment (1): the WBC dataset has been split in 85% training data and 15% testing data. The Figure 1(a) shows the features importance comparison provided by RF classifier. The most important feature is concavity with 0.229 feature importance. The accuracy values obtained for both classifiers are 89.5% for RF and 89.3% for k-NN, respectively.

Experiment (2): the WBC dataset has been split in 70% training data and 30% testing data. The Figure 1(b) shows the features importance comparison provided by RF classifier. The most important feature is concavity with 0.229 feature importance. The accuracy values obtained for both classifiers are 91.9% for RF and 91.3% for k-NN, respectively.

Experiment (3): the WBC dataset has been split in 55% training data and 45% testing data. The Figure 1(c) shows the features importance comparison provided by RF classifier. The most important feature is concavity with 0.257 feature importance. The accuracy values obtained for both classifiers are 89.4% for RF and 88.7% for k-NN, respectively.

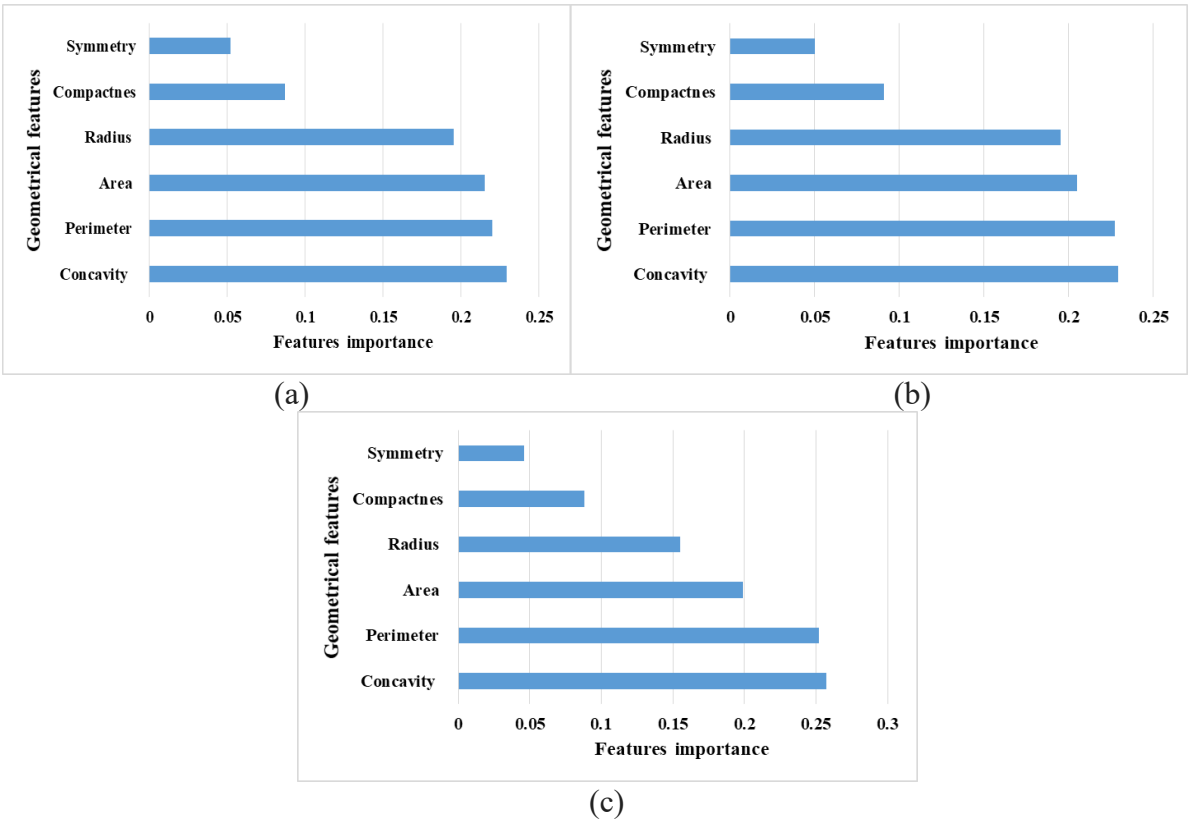

Fig. 1. Geometrical features and their importance

Our experiments demonstrated that, for all selected testing dataset size (i.e., 15%, 30% and 45% from the whole database), the classification accuracy values exceeded the limit of 88%. From the data in Fig. 1, it is clear that the most important feature is concavity. The poor importance among all these features belongs to symmetry. The order of the feature's importance is preserved for all the analyzed experimental cases.

4. CONCLUSIONS

In this study, we have analyzed the features' importance of the geometric features provided by WBC dataset and the classification accuracy of non-parametric RF and k-NN classifiers in the framework of three experiments. The best classification result has been obtained when the whole dataset was split in 70% training data and 30% testing data, for both classifiers. The reported accuracy was 91.9% for RF and 91.3% for k-NN, respectively. Statistics shows that the RF classifier has provided the best performance of prediction for breast cancer recurrence for all three experiments.

Further exploration of the WBC dataset can yield more interesting results by taking into consideration a larger number of classifiers. This will be the focus of our future work.

References

1. El_Rahman S.A., Predicting breast cancer survivability based on machine learning and features selection algorithms: a comparative study, Journal of Ambient Intelligence and Humanized Computing 12(8) (2021) 8585-8623.
2. Abien Fred M. Agarap, On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset, Proceedings of the 2nd International Conference on Machine Learning and Soft Computing (2018) 5-9.
3. Gopal V. N., Al-Turjman F., Kumar R., Anand L., Rajesh M., Feature selection and classification in breast cancer prediction using IoT and machine learning, Measurement 178 (2021) 109442.
4. Akay M.F., Support vector machines combined with feature selection for breast cancer diagnosis, Expert Systems with Applications 36(2) (2009) 3240-3247.
5. Anghelache Nastase I. N., Moldovanu S., Moraru L., Image Moment-Based Features for Mass Detection in Breast US Images via Machine Learning and Neural Network Classification Models, Inventions 7(2) (2022) 42.
6. Wolberg W. H., Mangasarian O. L., Multisurface method of pattern separation for medical diagnosis applied to breast cytology, Proceedings of the National Academy of Sciences 87(23) (1990) 9193–9196.
7. Ahmmed R., Swakshar A. S., Hossain M. F., Rafiq M. A., Classification of tumors and it stages in brain MRI using support vector machine and artificial neural network, Proceedings of the 2017 International Conference on Electrical, Computer and Communication Engineering (2017) 229–234.