Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 191 (2021) 481–486

Procedia Computer Science

www.elsevier.com/locate/procedia

International Workshop on Edge IA-IoT for Smart Agriculture (SA2IOT)
August 9-12, 2021, Leuven, Belgium

Breast Cancer Prediction and Diagnosis through a New Approach based on Majority Voting Ensemble Classifier

Mohammed Amine Naji[a, *], Sanaa El Filali[b]
Meriem Bouhlal[c], EL Habib Benlahmar[d], Rachida Ait Abdelouhahid[e] , Olivier Debauche[f]

[a,b,c,d,e]*Faculty of Science Ben M'sik, Hassan 2 University, Casablanca, Morocco*
[f]Faculty of Engineering, University of Mons, Mons, Belgium

Abstract

Researchers have extensively used machine learning techniques and data mining methods to build prediction models and classify data in various domains such as aviation, computer science, education, finance, marketing and particularly in medical field where those methods are applied as support systems for diagnosis and analysis in order to make better decisions. On this subject, our research paper attempts to assess the performance of Individual and Ensemble machine learning techniques based on the effectiveness and the efficiently, in terms of accuracy, specificity, sensitivity and precision to choose the most effective. The main object of our research paper is to define the best and effective machine learning approach for the Breast Cancer diagnosis and prediction. To achieve our objective, we applied individual based level machine learning algorithms Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naïve Bayes (NB), Decision tree (C4.5), Simple Logistic and well known ensembles methods like Majority Voting and Random Forest with 10 cross field technique on the Breast Cancer Diagnosis Dataset obtained from UCI Repository. The experimental results show that the Majority Voting Ensemble technique based on 3 top classifiers SVM, K-NN, Simple Logistic gives the highest accuracy 98.1% with the lowest error rate 0.01% and outperformed all other individual classifiers. This study demonstrates that our proposal approach based on Majority Voting Ensemble technique was the best classification machine learning model with the highest level of accuracy for breast cancer prediction and diagnosis. All experiments are effectuated within a simulation environment and realized in Weka data mining tool.

Keywords: Brest Cancer; Prediction; Diagnosis; Classficiation; SVM; NB; C4.5; K-NN; Simple Logistic; Majority; Voting; Ensembles;

* Corresponding author. Tel.: Tel.: +212-634-32-83-78;
E-mail address: aminenaji55@gmail.com

1877-0509 © 2021 The Authors. Published by Elsevier B.V.

10.1016/j.procs.2021.07.061

1. INTRODUCTION

The most frequent and deadliest cancer among Women is breast cancer, each year, 2.1 million women are impacted. In 2018, it is estimated that 627 000 women died from breast cancer – which is approximately 15% of all cancer deaths among women. Although breast cancer rates are higher among women in more advanced regions, rates are increasing in nearly every single region globally [1].

To improve breast cancer outcomes and survival, early detection is well needed. Among the early detection strategies for breast cancer, there is the diagnosis. The diagnosing and treatment of breast cancer have speedily evolved throughout the past 3 decades. A part of this evolution is thanks to individual or organized breast screening programs and progress of breast imaging techniques [2]. Indeed, a sub-domain of artificial intelligence called "machine learning" makes it potential to create algorithms able to accumulate data and intelligence from experiments, while not being human-guided throughout their learning, not expressly programmed to manage a specific task, thence their central role within the data value chain. The application of machine learning approach in medical science topics rise speedily due to their high performance in predicting outcomes, personalizing the treatment of illness for each patient, helping the doctors to make the right decision, improving diseases detection and diagnostics, and finally reducing the risk of death. For prediction and classification of Breast Cancer, several algorithms are applied. This current paper presents a comparison between the performance of various individual classifiers and ensembles approaches such as Random Forest and Majority Voting which are among the foremost influential data mining algorithms within the research community. The dataset was retrieved from the UCI repository. Firstly, the effectiveness of Support Vector Machine (SVM), k Nearest Neighbours (K-NN), Simple Logistic, Naïve Bayes (NB), Decision tree (C4.5) and Random Forrest is assessed in terms of accuracy and f-measure. Secondly, a majority voting-based ensemble of top 3 best performing classifiers is constructed to predict Breast cancer. Indeed, these 3 best performing classifiers are constructed using ensemble technique namely Majority Voting. The voting ensemble technique is an example of the multi- expert approach, which helps to combine the classifiers in a parallel fashion. Subsequently, each classifier trained on all data and contributes to a decision. Finally, the voting technique helps to generate the final solution and results.

The remainder of this work is organized as following. In the second section, we present related work. Then, the third section explains in detail the experimental procedure. Afterwards, the fourth section discusses experiments results obtained. Finally, in the fifth section, we present conclusion and future research directions

2. Related Work

Classification is an essential and very important task for machine learning and data mining. Several researches have been carried out for machine learning and data mining on multiple and different medical datasets, the objective is to classify breast cancer. Many of them have good classification accuracy.

The author S. Aruna compared the performance of Support Vector Machine (SVM), Naïve Bayes, K- Nearest Neighbor (K-NN) and C4.5 to get the best machine learning algorithm in WBC. SVM has showed to be the most performing classifier with an accuracy of 96.99% [3]. The author V. Chaurasia compared the performance criteria of supervised learning classifiers such as Naïve Bayes, SVM-RBF kernel, RBF neural networks, decision tree (J48) and simple CART to find the best classifier element in breast cancer data sets. The experimental result showed that the SVM-RBF core is more accurate than other classifiers obtaining 96.84% accuracy in the (original) Wisconsin breast cancer data sets [4]. The author N. Khuriwal used voting algorithm method for evaluating an ensemble of two machine learning algorithm ANN and logistic regression on the Wisconsin Breast Cancer database for detecting and diagnosed breast cancer. When compared to related work from the literature and existing research. It is shown that the ANN approach with logistic algorithm is achieved 98% accuracy from another machine learning algorithm [5]. The author M. I. Faisal used and compared the performance of individual classifiers including MLP, Neural Network, Decision Tree, Naïve Bayes, Gradient Boosted Tree, and SVM. Random forest and majority voting-based ensembles are also analyzed. Based on performance evaluations, the author Find that Gradient-boosted Tree outperformed all other individual as well as ensemble classifiers and achieved 90% accuracy [6]. The author H. When applied artificial neural network and several traditional machine learning techniques to SEER (the Surveillance, Epidemiology, and End Result program) database to classify mortality rate in two categories including less than 60 months and more than 60 months and obtained the result as that neural network has the best accuracy 85.64% in predicting survivability of prostate cancer patients [7]. Many other authors also have done research for detection and diagnosis breast cancer using various machines learning algorithm.

Our research is focused on assessing such machine learning algorithms and approaches in order to conclude the best methodology for breast cancer prediction and diagnosis. With respect to all related work mentioned above, our work compares the behavior of base level classifiers, Random Forest and Majority Voting based ensembles « SVM, KNN, Simple Logistic » using Breast Cancer Wisconsin Diagnostic datasets in both diagnosis and analysis to make decisions. The objective is to reach the best accuracy with the lowest error rate in analyzing data. For that reason, we compare efficiency and effectiveness of these approaches in terms of many criteria, including accuracy, precision, sensitivity, specificity, correctly, incorrectly classified instances and time to build model, among others.

3. Methodology

Our methodology begins with data acquisition followed by pre-processing which contains four steps viz: cleaning and editing data, select attributes, set target Role and features extraction. The selected classifiers are then trained and tested on the Breast Cancer Dataset using standard 10- fold cross-validation approach which are training options that split the dataset into a training set to train classifier and a testing set to evaluate it.

In other terms, our research answers the question of identifying the most predictive and effective algorithm for the detection of breast cancer, and selection the algorithm that provides the high up accuracy and exploits better effectiveness and efficiently.

3.1. Experiment Environment

All experiments on the machine learning algorithms described during this paper were conducted using libraries from Weka machine learning environment. Weka is defined as a set of machine learning algorithms for data mining tasks. It contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization [8].

Machine learning techniques implemented in Weka are applied to a variety of world issues. The program offers a well-defined framework for experimenters and developers to create and evaluate their models.

3.2. Data Acquisition

In this paper, we used a publicly available data set namely the Breast Cancer Wisconsin Diagnostic dataset from the UCI Machine Learning Repository [9].The datasets has 569 instances (Benign : 375 ; Malignant : 212), 2 classes (62.74% Benign and 37.26% Malignant) with 32 attributes. There is one class attribute in addition to other attributes. One of the other numeric value columns is the instance ID column. Our data set includes two classes, as mentioned earlier. They are benign (B) and malignant (M). The results of our experiment are computed and evaluated based on the effectiveness, and the efficiency of different base level algorithms and well-known ensemble methods such as Random Forest and Majority Voting.

In this study, Weka software was utilized to perform experiments with base level classifiers namely Naïve Bayes (NB), Support Vector Machine (SVM), Simple Logistic, K-Nearest Neighbors (K-NN) and Decision Tree (C4.5), Random Forest and Majority Voting based ensembles are also evaluated using the same software.Majority Voting based Ensemble method.

3.3. Majority Voting based Ensemble method

Our approach is constructed using a widely used ensemble technique namely Majority Voting. The voting ensemble technique is a common example of the multi- expert approach, which helps to combine the classifiers in a parallel fashion.

Subsequently, each classifier trained on all data and contributes to a decision. Finally, the voting technique helps to generate the final solution.

4. RESULTS, EVALUATIONS AND DICUSSION

In this section, we discuss the Breast Cancer Diagnostic dataset, experiments, the evaluation scheme, and we compare the performance of base-level classifiers to each other. The results are shown in the following.

4.1. Workflow

We performed several experiments. However, workflow of experiments is generalized and can be shown through the following steps:

- Step-1: Extraction of datasets through an online repository.
- Step-2: Application of pre-processing for data cleaning.
- Step-3: Standard 10-fold cross validation is applied for training and testing.
- Step-4: Computation of results for all individual classifiers
- Step-5: Select top-3 classifiers based on comparison of the performance measure such as accuracy and compose majority voting-based ensemble.
- Step-6: Compute results for Random Forest and Majority Voting based ensemble.
- Step-7: A Performance comparison is conducted for all individual as well as ensemble classifiers in order to identify the best classifier for breast cancer detection.

4.2. Results

After applying and finishing the workflow step, we try to analyze the results and figure out the distribution of values in terms of effectiveness and efficiency

4.2.1. Effectiveness

In This section, we evaluate the effectiveness of base level classifiers, Random Forest and Majority Voting-based ensembles « SVM, KNN and Simple Logistic » in terms of time to build the model, correctly classified instances, incorrectly classified instances and accuracy as shown in Table 1.

In order to better the measurement of classifier performance, the simulation error is also considered in this study. To do this, we assess the effectiveness of our classifier in terms of: Kappa as a randomly corrected measure of agreement between classifications and actual classes, Mean Absolute Error as the way in which predictions or predictions approximate possible results, Root Mean Squared Error, Relative Absolute Error, Root Relative Absolute Error, and Root Relative Squared Error as shown in Table 2.

Table 1. Performance of the classifiers.

Evaluation Criteria	SVM	NB	C4.5	KNN	Simple Logistic	Random Forest	Majority Voting
Time to build model (s)	0.03	0.01	0.15	0.01	0.10	0.13	0.11
Correctly classified instances (%)	97.891	92.6186	93.1459	96.8366	97.1889	95.9578	98.0668
Incorrectly classified instances (%)	2.109	7.3814	6.8541	3.1634	2.812	4.0422	1.9332
Accuracy (%)	97.8	92.6	93.1	96.8	97.1	95.9	98.1

Table 2. Training and simulation error.

Evaluation Criteria	SVM	NB	C4.5	KNN	Simple Logistic	Random Forest	Majority Voting
Kappa statistic	0.9545	0.8418	0.8541	0.9321	0.9395	0.9128	0.9583
Mean absolute error	0.0211	0.0732	0.0741	0.0527	0.0444	0.0757	0.0193
Root mean squared error	0.1452	0.2648	0.2574	0.1601	0.1408	0.1731	0.139
Relative absolute error %	4.5095	15.6565	15.8345	11.2735	9.5043	16.1855	4.1337
Root relative squared error %	30.0354	54.7597	53.2317	33.1115	29.1122	35.8076	28.7567

4.2.2. Efficiency

Once the predictive model is constructed, we can check how efficient it is. For that, we compare the performance measures based on TP rate, FP rate, precision, recall and F-Measure values for base level classifiers and Majority Voting-based ensembles « SVM, KNN, Simple Logistic » as shown in Table 3.

Table 3. Comparison of accuracy measures for Classifiers.

Classifiers	TP	FP	Precision	Recall	F-M
SVM	0.979	0.034	0.979	0.979	0.979
NB	0.926	0.086	0.926	0.926	0.926
C4.5	0.931	0.073	0.932	0.931	0.932
KNN	0.968	0.040	0.968	0.968	0.968
Simple Logistic	0.972	0.038	0.972	0.972	0.972
Random Forest	0.960	0.055	0.960	0.960	0.959
Majority Voting	0.981	0.031	0.981	0.981	0.981

Since confusion matrices are a useful way to assess the classifier, each row in Table 4 represents the rates in an actual class while each column displays the predictions

Table 4. Confusion matrix for Classifiers.

	Malignant	***Benign***	
SVM	***201***	***11***	***Malignant***
	1	***356***	***Benign***
NB	***190***	***22***	***Malignant***
	20	***337***	***Benign***
C4.5	***195***	***17***	***Malignant***
	22	***335***	***Benign***
KNN	***201***	***11***	***Malignant***
	7	***350***	***Benign***
Simple Logistic	***201***	***11***	***Malignant***
	5	***352***	***Benign***
Random Forest	***196***	***16***	***Malignant***
	7	***350***	***Benign***
Majority Voting	***202***	***10***	***Malignant***
	1	***356***	***Benign***

4.3. Discussion

We can notice from Table 1 that Majority Voting based ensemble takes about 0.11 s to build its model unlike k-NN that takes just 0.01 s. It can be demonstrated by the fact that k-NN is a lazy learner and does not do much during training process contrary to other classifiers that construct the models.

In other side, the accuracy got by Majority Voting 98.1% is better than the accuracy got by base level classifiers and Random Forest that have an accuracy that varies between 93.1% and 97.8%. It can also be clearly viewed that Majority Voting has the highest value of correctly classified instances and the lower value of incorrectly classified instances than the other classifiers.

From Table 2, we can better understand that the chance of obtaining a best classification 0.95% with the least warning error rate 0.01 is produced by Majority Voting. We can also notice that Majority Voting has the best compatibility between the validity of the data collected and their reliability. C4.5 and random Forest has the highest value of error rate, which explains the large number of incorrectly classified instances for each algorithm (6.8541% incorrect instances for C4.5 and 4.0422% incorrect instances for RF) (see Table 1).

After building the predicted model, we can at analyze results got in assessing efficiency of our algorithms. In effect, Table 3 shows that Majority Voting and SVM got the highest value (98%, 97%) of TP for Weighted AVG of benign and malignant class. The FP rate is lower when applying Majority Voting (0.031 for Weighted AVG of benign and malignant class), and the other algorithms pursue: SVM 0.034, Simple Logistic 0.038 and K-NN 0.040. From these reasons, we can realize why Majority Voting has outperformed other classifiers.

Let us now compare actual class and predicted results gotten using confusion matrix as shown Table 4. Majority Voting ensemble technique predicts correctly 558 cases out of 569 cases constituted of 202 malignant cases that are actually malignant and 356 benign cases that are actually benign, and 11 cases incorrectly predicted including 10 cases of malignant class predicted as benign and 1 case of benign class predicted as malignant. That is why the accuracy of Majority Voting is better than other classification techniques used with lower error rate value.

In summary, Majority Voting ensemble technique was able to show its power in terms of effectiveness and efficiency based on accuracy and recall. Compared to an important number of researches on Breast-cancer-Wisconsin retrieve in literature that treat and compare classification accuracies of data mining algorithms, our experimental results make the highest value of accuracy (98.1%) in classifying breast cancer diagnostic dataset. It can be remarked that Majority Voting outperforms other classifiers with regard to accuracy, sensitivity, specificity and precision in classifying breast cancer diagnostic dataset.

In this research, we also found some threats. The first threat is related to a generalization of results since we performed our experiments on a single dataset. Consequently, the result may vary if we consider several experiments with different datasets. The secondly threat is related to the selection of one ensemble technique. We report the results according to the functionality of the majority voting method

5. Conclusion

In this paper, we have provided explanations of different Machine Learning approaches and their applications in breast cancer diagnosis and prognosis used to analyze the data in the benchmark database Breast Cancer Wisconsin Diagnostic.

The application of data mining technologies in the medical field is very important because they certainly help in the decision-making process. Nevertheless, to do this, such algorithms require high performance with great precision and a good choice of methods depending on the working context and the data being processed. The focus of this research is an evaluation of machine learning classifiers as well as ensembles for Breast cancer detection in order to define the best and performed prediction method.

For this purpose, individual classifiers including SVM, K-NN, Decision Tree, Naïve Bayes and Simple Logistic are assessed. Random forest and majority voting-based ensembles are also analyzed for Breast cancer prediction. We compared efficiency and effectiveness of those methods in terms of accuracy, precision, sensitivity, specificity and F-Measure to find the best classification accuracy. Majority Voting Ensemble technique reaches an accuracy of 98.1% and outperforms all other algorithms. In conclusion, our Majority voting-based ensemble has demonstrated its efficiency in Breast Cancer prediction and diagnosis and attained the better performance in terms of precision and low error rate. It must be noted that the results and conclusions obtained from our approach is limited to the Breast Cancer Wisconsin Diagnostic dataset. For this it is essential to reapply the same work on different datasets to confirm the results obtained. In the future, we plan to test our machine learning approach on larger data sets with more disease classes to achieve higher accuracy. In addition, we will evaluate other machine learning techniques on breast cancer and different disease datasets. Similarly, other ensemble technique like Stacking, Adaboost and Bagging will be analyzed.

6. References

[1] 'WHO | Breast cancer', WHO. http://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/ (accessed Feb. 18, 2020)

[2] A. Addioui, F. Benabbou, S. E. Filali, and M. E. Aroussi, 'Breast cancer mammography diagnosis approach using dual tree complex wavelet transform and artificial neural networks', International Journal of Imaging and RoboticsTM, vol. 16, no. 4, pp. 62–68, Sep. 2016.

[3] S. Aruna, D. S. P. Rajagopalan, and L. V. Nandakishore, 'KNOWLEDGE BASED ANALYSIS OF VARIOUS STATISTICAL TOOLS IN DETECTING BREAST CANCER', Computer Science, p. 9.

[4] V. Chaurasia and S. Pal, 'Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability', p. 14, 2014.

[5] N. Khuriwal and N. Mishra, 'Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm', in 2018 IEEMA Engineer Infinite Conference (eTechNxT), New Delhi, Mar. 2018, pp. 1–5, doi: 10.1109/ETECHNXT.2018.8385355

[6] M. I. Faisal, S. Bashir, Z. S. Khan, and F. Hassan Khan, 'An Evaluation of Machine Learning Classifiers and Ensembles for Early Stage Prediction of Lung Cancer', in 2018 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST), Karachi, Pakistan, Dec. 2018, pp. 1–4, doi: 10.1109/ICEEST.2018.8643311.

[7] H. Wen, S. Li, W. Li, J. Li, and C. Yin, 'Comparision of Four Machine Learning Techniques for the Prediction of Prostate Cancer Survivability', in 2018 15th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), Chengdu, China, Dec. 2018, pp. 112–116, doi: 10.1109/ICCWAMTIP.2018.8632577.

[8] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, 'The WEKA data mining software: an update', SIGKDD Explor. Newsl., vol. 11, no. 1, p. 10, Nov. 2009, doi: 10.1145/1656274.1656278.

[9] 'UCI Machine Learning Repository'. https://archive.ics.uci.edu/ml/index.php (accessed Feb. 18, 2020).