Using Feature Selection Techniques to Improve the Accuracy of Breast Cancer Classification: Special Issue on Data and Security Engineering

Chapter · January 2019

DOI: 10.1007/978-3-030-11196-0_28

CITATIONS
5

READS
1,142

4 authors, including:

Hajar Saoud
Abdelmalek Essaâdi University
6 PUBLICATIONS **18** CITATIONS

SEE PROFILE

M. Ghailani
Ecole de Sciences Appliqués de Tanger
16 PUBLICATIONS **58** CITATIONS

SEE PROFILE

Boudhir Anouar Abdelhakim
Abdelmalek Essaâdi University
88 PUBLICATIONS **233** CITATIONS

SEE PROFILE

Using Feature Selection Techniques to Improve the Accuracy of Breast Cancer Classification

Hajar Saoud[1(✉)], Abderrahim Ghadi[1], Mohamed Ghailani[2], and Boudhir Anouar Abdelhakim[1]

[1] LIST Laboratory, University of Abdelmalek Essaadi (UAE), Tangier, Morocco
{saoudhajar1994, ghadi05, boudhir.anouar}@gmail.com
[2] LabTIC Laboratory, University of Abdelmalek Essaadi (UAE), Tangier, Morocco
ghalamed@gmail.com

Abstract. Classification is a data mining process that aims to divide data into classes to facilitate decision-making; it is therefore an important task in medical field. In this paper we will try to improve the accuracy of the classification of six machines learning algorithms: Bayes Network (BN), Support Vector Machine (SVM), k-nearest neighbors algorithm (Knn), Artificial Neural Network (ANN), Decision Tree (C4.5) and Logistic Regression using feature selection techniques, for breast cancer classification and diagnosis. We examined those methods of classification and techniques of feature selection in WEKA Tool (The Waikato Environment for Knowledge Analysis) using two databases, Wisconsin breast cancer datasets original (WBC) and diagnostic (WBCD) available in UCI machine learning repository.

Keywords: Breast cancer · Diagnostic · Machines learning algorithms
Feature selection · Classification · WEKA

1 Introduction

Breast cancer is one of the diseases that make higher number of incidence and mortality in the word. It represents the second cause of death for women after lung cancer [1]. Early diagnosis can reduce the breast cancer mortality rate by 40% [2].

Breast cancer can be defined as dangerous disease where cancer cells form in the tissue of the breast of the women and can spread to the others organs of the body. Machines learning algorithms and feature selection techniques will be interesting tools to predict and diagnose breast cancer also to classify it into its two categories either benign or malignant tumor.

We examined accuracy of six machines learning algorithms in the classification and diagnosis of the breast cancer: Bayes Network (BN), Support Vector Machine (SVM), k-Nearest Neighbors Algorithm (Knn), Artificial Neural Network (ANN), Decision Tree (C4.5) and Logistic Regression.

After that we want to improve the accuracy of those classifier using feature selection techniques that can be defined as the process of eliminating irrelevant and redundant features to improve the accuracy of classification.

M. Ben Ahmed et al. (Eds.): SCA 2018, LNITI, pp. 307–315, 2019.
https://doi.org/10.1007/978-3-030-11196-0_28

The rest of this paper is structured as follows. Part two is a presentation of breast cancer. Part three gives a vision about similar research. Part four is a theoretic presentation of machine learning algorithms. Part five give the definition of feature selection techniques. Part six describes the database used. Part seven explain confusion matrix. Part height shows the experiments performed by WEKA software on Wisconsin breast cancer dataset and results of these experiments and finally conclusion and perspectives in part nine.

2 Breast Cancer

Breast cancer is an abnormal production of cells in the breast that grow in an anarchic way. The masses of cells formed in the breast are called tumors. The cancer cells can stay in the breast or spread to other organs of the body. This is called metastasis.

2.1 Types of Breast Cancer

Breast cancer is decomposed into two types benign and malignant tumors:

- Benign tumors are non-dangerous tumors, they have well-defined contours. They develop slowly in the organ where they appeared without producing metastatic cases. Benign tumors are composed of cells that resemble to normal cells of the breast tissue.
- Malignant tumors are dangerous tumors, because they spread to other organs of the body and can produce metastatic cases. Cancer cells of malignant tumors have several abnormalities compared with normal cells in shape, size and contours where cells lose their original characteristics.

2.2 Causes of Breast Cancer

The first risk factor that can increase the probability of breast cancer is the age factor, the risk of breast cancer increases with age. Other factors that can intervene like:

Family or genetic factors. Gender: women are the most infected with breast cancer.

- A woman history: The woman that had already breast cancer in one breast, she has an increased risk to have cancer in the other breast.
- A family history: If several parents of the woman has been diagnosed with breast cancer, especially at a younger age, the risk to develop breast cancer increases.
- Genetic factors: Some genetic mutations increase the risk of breast cancer.

Characteristics of the individual. Obesity: The obesity increases the risk of breast cancer.

- Having period in early age: Having the period before the age of 12 increases the risk of breast cancer.
- Late menopause: Woman that started menopause at a later age, she is more likely to develop breast cancer.

- Having the first child in old age: Women who give birth to their first child after the age of 30 may have an increased risk of breast cancer.
- Women who have never been pregnant: The fact of not having a child increases the risk of developing breast cancer.
- Hormone replacement therapy: (Estrogen and progesterone) increases the risk of having breast cancer after 5 years of treatment.
- Drinking Alcohol: Drinking alcohol increases the risk of breast cancer.

3 Related Works

Several researches have been carried out in this field, some of them used only machine learning algorithms to classify and diagnose cancers others tried to improve the accuracy of classification using feature selection techniques.

Aalaei et al. [3] in this research they have chosen Genetic Algorithm for selecting the best subset of feature, they evaluate it using three classifier ANN, PS-classifier and GA-classifier on Wisconsin breast cancer original, diagnostic and prognosis. The GA improved the accuracy of the classifier, for WBC PS-classifier achieved the beast accuracy for WBCD and WBCP the ANN algorithm that achieved higher accuracy.

Saabith et al. [4] examined the accuracy of three classifiers Decision Tree, Neural Network and Rough set with and without feature selection techniques for breast cancer effective prediction.

Gowri and Ramar [5] they tried to propose a hybrid approach that combines between the best feature selection technique and best machine learning algorithm. The feature selection techniques examined are Correlation based Feature Selection (CFS), Information Gain (IG), Relief (R), Principle Components Analysis (PCA), Consistency based Subset Evaluation (CSE), and symmetrical uncertainty (SU) and techniques of classification examined are Naïve Bayes, K-Nearest Neighbor and Decision Tree on WBC, WBCD and WBCP.

Ahmed Abd El-Hafeez Ibrahim et al. [6] in this research they tried to present multi-classifiers fusion approach that fusion between classifiers to get the best multi-classifier fusion approach for each dataset. Dataset used are WBC, WBCD, WBCP and BCD, in WBCP and WBC 4th fusion level is better than other, for BCD the 3rd is the best and for WDBC the 2nd level is the beast.

Hamsagayathri and Sampath [7] in this research they analyzed the performance of four decision three algorithms in breast cancer dataset original, diagnostic and prognosis, they concluded that Priority based decision tree classifier is the beast classifier it classify data with 93.63% of accuracy.

Lavanya and Usha Rani [8] in this paper they analyzed the performance of decision tree classifier-CART with feature selection techniques and without them, they concluded that classification with feature selection is better than classification without feature selection because feature selection technique enhance the accuracy of classification.

Kaur et al. [9] in this research they compared the existing feature selection methods and compared their performance by calculating TPR, FPR, Classification accuracy, ROC Area, Precision, Kappa Statistic and Training Time.

4 Machine Learning Algorithms

4.1 Bayes Network

Bayes Networks [10], also called (Bayesian belief networks), are methods that are widely used for modeling and presenting knowledge of uncertain domains. Bayes Network is a directed acyclic graph (DAG), consisting of several nodes that represent variables and arcs that represent the probabilistic dependencies between those variables.

4.2 Support Vector Machines (SVM)

Support Vector Machines [11] are supervised learning models that can be used in prediction also in the classification of the linear and nonlinear data. The principle of the SVM algorithm is to use a non-linear mapping to transform the original learning data into a larger dimension. In this new dimension, it looks for the linear hyperplane of optimal separation. SVM algorithm aims to find a hyperplane with the largest margin named maximum marginal hyperplane (MMH) to be more accurate in the classification of future data i.e. it looks the shortest distance between the MMH and the closest training tuple of each class.

4.3 k-Nearest Neighbors Algorithm (Knn)

The k-nearest neighbors classifiers [11] are based on analog learning, i.e. they compare a given test tuple with similar learning tuples. They classify the tuples using more than one nearest neighbor. The principle of the k-nearest neighbors classifier is that it looks in the space model for the K test tuples closest to the unknown tuple. These tuples are named (k nearest neighbors) of the unknown tuple.

4.4 Decision Tree (C4.5)

The decision tree algorithm [12] is a classification algorithm that is similar to an organizational chart where the internal nodes (not- leaf) of a decision tree represent the tests on the attributes, the branches represent the results of the test and the external nodes (leaves) represent the predicted results. At each node, the algorithm chooses the best attribute to partition the data into individual classes. At the end a tree will be built when selecting subsets from the data provided.

4.5 Logistic Regression

Logistic regression [13] is one of the generalized linear models much used in machine learning. Logistic regression predicts the probability of a result that can take two values

from a set of predictor variables. Logistic regression is mainly used for prediction and also to calculate the probability of success.

4.6 Artificial Neural Network (ANN)

Neural Network [14] can be defined as a reasoning model based on the human brain. An Artificial Neural Network is a set of processors (or neurons) very simple, very interconnected by weighted connections to pass signals from one neuron to another and they operate in a parallel manner. These neurons are similar to the biological neurons of the human brain. Artificial Neural Network consists of three layers: input layer, output layer and between them they are extra layers called hidden layers.

5 Feature Selection

Feature selection also called variable selection is the process of choosing the most relevant features to improve the process of the classification. They are three types of feature selection:

5.1 Filter Method

Filter methods select the variables independently of the chosen classification model; they select the variables by correlating the predictors and the answer variable. They take the variables that are relevant for the classification and delete the others.

5.2 Wrapper Method

Wrapper methods choose the variables by doing a combination between them, unlike the filter methods, wrapper methods try to find the interactions between the variables.

5.3 Embedded Methods

Embedded methods have been recently proposed, they try to combine the advantages of wrapper and filter methods. The choice of feature is done at the same time with the execution of the algorithm.

6 Description of the Dataset

In this paper we used two databases: Wisconsin breast cancer original (WBC) [15] and Wisconsin breast cancer diagnostic (WBCD) [16] available in UCI machine learning repository. WBC contains 699 records (458 benign tumors and 241 malignant tumors). It is composed of 10 variables 9 predictor variables and one result variable that shows whether the tumor is benign or malignant. The predictive attributes vary between 0 and 10. The value 0 corresponds to the normal state and the value 10 corresponds to the most abnormal state.

WBCD contains 569 records (357 benign tumors and 212 malignant tumors). It is composed of 33 variables 32 predictor variables and one result variable that shows also whether the tumor is benign or malignant.

7 Confusion Matrix

Confusion matrix gives the possibility to evaluate the performance of each classifier by calculating its accuracy, Sensitivity, Specificity. It contains information about real classifications or (current) and predicted (Table 1):

Table 1. Confusion matrix

	Predicted benign	Predicted malignant
Actual benign	TP (true positives)	FN (false negatives)
Actual malignant	FP (false positives)	TN (true negatives)

TP the cases predicted as benign tumors, they are in fact benign tumors
TN the cases predicted as malignant tumors, they are in fact malignant tumors
FP the cases predicted as benign tumors but in the reality they are malignant tumors
FN the cases predicted as malignant tumors but in the reality they are benign tumors

From the confusion matrix we can calculate:

- Accuracy = $\frac{TP+TN}{TP+FP+TN+FN}$
- Sensitivity = $\frac{TP}{TP+FN}$
- Specificity = $\frac{TN}{TN+FP}$

8 Experimentations and Results

8.1 WEKA Tool

The platform that we used to apply the machine learning algorithms on the breast cancer database is WEKA [17], because WEKA is a collection of open source machine learning algorithms, which allows realizing the tasks of data mining to solve real world problems. It contains tools for data preprocessing, classification, regression, grouping, and association rules. Also it offers an environment to develop new models.

8.2 K-Fold Cross-Validation

To evaluate the performance of machine learning algorithms based on breast cancer data we used the K-fold cross validation test method. This method aims to divide the database in two sets, the training data to run the model and the test data to evaluate the performance of the model. This is the most used method in the evaluation of machine learning techniques.

8.3 Feature Selected

Wrapper methods are used for feature selecting using Best first as search method with classifier Subset Evaluator technique to improve the accuracy of the classification:

- Bayes Net:

	Feature selected
WBC	1,2,3,4,5,6,7,8
WBCD	1,9,22,23,28

- Support Vector Machines:

	Feature selected
WBC	1,2,3,4,6,7,8
WBCD	4,5,22,23,24,25,26,28,30

- k-Nearest Neighbors Algorithm

	Feature selected
WBC	1,2,6,7
WBCD	1

- Decision Tree

	Feature selected
WBC	1,2,3,4,6
WBCD	1,22,23,29,31

- Logistic Regression

	Feature selected
WBC	1,2,3,5,6,8,9
WBCD	3,21,24,26

- Artificial Neural Network

	Feature selected
WBC	1,2,3,4,5,6,7,9
WBCD	3,6,10,12,16,17,21,25,26,28,31

8.4 Results

- WBC (Table 2 and 3):
- WBCD:

Table 2. Classification accuracy in WBC

	Without FS (%)	With FS (%)
BN	97.2818	97.4249
SVM	97.2818	95.279
KNN	95.279	95.8512
DT	95.1359	95.7082
LR	96.5665	96.7096
ANN	95.422	95.8512

Table 3. Classification accuracy in WBCD

	Without FS (%)	With FS (%)
BN	95.2548	96.1336
SVM	97.891	97.3638
KNN	96.1336	96.1336
DT	92.9701	94.9033
LR	94.2004	95.6063
ANN	96.1336	95.6063

9 Conclusion

To conclude, in this paper we tried to improve the accuracy of the classification of breast cancer using feature selection techniques. We use to databases Wisconsin breast cancer dataset original (WBC) and diagnostic (WBCD). The feature selection technique improved the accuracy of some classifier like Bayes net in both WBC and WBCD but we see the opposite for some classifier like SVM the feature selection technique has reduced the accuracy of classification. The best model to classify breast cancer in WBC is Bayes Network with feature selection and the beast one for WBCD is Support Vector Machines without feature selection. In the feature work we will try to propose methods that can improve more the accuracy of the classification of breast cancer.

References

1. «Breast cancer statistics»: World Cancer Research Fund, 22 Aug 2018. Available on: https://www.wcrf.org/dietandcancer/cancer-trends/breast-cancer-statistics
2. Ganesan, K., Acharya, U.R., Chua, C.K., Min, L.C., Abraham, K.T., Ng, K.-H.: Computer-aided breast cancer detection using mammograms: a review. IEEE Rev. Biomed. Eng. **6**, 77–98 (2013)
3. Aalaei, S., Shahraki, H., Rowhanimanesh, A., Eslami, S.: «Feature selection using genetic algorithm for breast cancer diagnosis: experiment on three different datasets». Iran J. Basic Med. Sci. **19**(5), 7 (2016)
4. Saabith, A.L.S., Sundararajan, E., Bakar, A.A.: «Comparative Study on Different Classification Techniques for Breast Cancer Dataset», p. 8 (2014)
5. Gowri A.S. Ramar, D.K.: «A novel approach of feature selection techniques for image dataset». **3**(2), 5
6. Abd El-Hafeez Ibrahim, A., Hashad, A.I., El-Deen Mohamed Shawky, N. Maher, A., Arab Academy for Science, Technology Maritime Transport, Cairo, Egypt: «Robust breast cancer diagnosis on four different datasets using multi-classifiers fusion». Int. J. Eng. Res., **V4**(03) (Mars 2015)
7. Hamsagayathri P., Sampath, P.: «Performance analysis of breast cancer classification using decision tree classifiers». Int. J. Curr. Pharm. Res. **9**(2), 19 (Mars 2017)
8. Lavanya, D., Usha Rani K.: «Analysis of feature selection with classification: breast cancer datasets». Indian J. Comput. Sci. Eng. (IJCSE) **2**(5), 9 (2011)
9. Kaur, R.: «Study and comparison of feature selection approaches for intrusion detection». Int. J. Comput. Appl. **7**
10. Mahmood, A.: «Structure Learning of Causal Bayesian Networks: A Survey», p. 6
11. Han, J., Kamber, M.: Data mining: concepts and techniques, 2nd ed., [Nachdr.]. Elsevier/Morgan Kaufmann, Amsterdam (2010)
12. Han, J., Kamber, M.: Data mining: concepts and techniques, 3rd edn. Elsevier, Burlington, MA (2011)
13. Yusuff, H., Mohamad, N., Ngah, U., Yahaya, A.: «Breast cancer analysis using logistic regression». Int. J. Res. Appl. Stud. **11** (2012)
14. Negnevitsky, M.: Artificial intelligence: a guide to intelligent systems. 2nd ed. Addison-Wesley, Harlow, England; New York: (2005)
15. «UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set»: Available on: https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)
16. «UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set»: Available on: https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(Diagnostic)
17. «Machine Learning Project at the University of Waikato in New Zealand»: Available in: https://www.cs.waikato.ac.nz/ml/index.html
18. Saoud, H., Ghadi, A., Ghailani, M.: Analysis of evolutionary trends of incidence and mortality by cancers. In: Ben Ahmed M., Boudhir A. (eds.) Innovations in Smart Cities and Applications. SCAMS 2017. Lecture Notes in Networks and Systems, vol 37. Springer, Cham (2018)