Breast Cancer Classification Using Machine Learning

Meriem AMRANE[1]
Saliha OUKID[2]
Computer Science Department,
LRDSI Laboratory, University of Blida 1, Blida, Algeria
[1]amrane.meriem@outlook.fr , [2]osalyha@yahoo.com

Ikram GAGAOUA[3]
Tolga ENSARİ[4]
Computer Engineering, Istanbul University,
Istanbul, Turkey
[3]i.gagaoua@gmail.com , [4]ensari@istanbul.edu.tr

***Abstract*— During their life, among 8% of women are diagnosed with Breast cancer (BC), after lung cancer, BC is the second popular cause of death in both developed and undeveloped worlds. BC is characterized by the mutation of genes, constant pain, changes in the size, color(redness), skin texture of breasts. Classification of breast cancer leads pathologists to find a systematic and objective prognostic, generally the most frequent classification is binary (benign cancer/malign cancer). Today, Machine Learning (ML) techniques are being broadly used in the breast cancer classification problem. They provide high classification accuracy and effective diagnostic capabilities. In this paper, we present two different classifiers: Naive Bayes (NB) classifier and knearest neighbor (KNN) for breast cancer classification. We propose a comparison between the two new implementations and evaluate their accuracy using cross validation. Results show that KNN gives the highest accuracy (97.51%) with lowest error rate then NB classifier (96.19 %).**

Keywords— Breast cancer classification; Bayesian classifier component; K-nearest neibhor

I. INTRODUCTION

Breast Cancer's causes are multifactorial and involves family history, obesity, hormones, radiation therapy, and even reproductive factors. Every year, one million women are newly diagnosed with breast cancer, according to the report of the world health organization half of them would die, because it's usually late when doctors detect the cancer [1]. Breast Cancer is caused by a typo or mutation in a single cell, which can be shut down by the system or causes a reckless cell division. If the problem is not fixed after a few months, masses are formed from cells containing wrong instructions.

Malignant tumors expand to the neighboring cells, which can lead to metastasize or reach other parts, whereas benign masses can't expand to other tissues, the expansion is then only limited to the benign mass [1] [2]. Detection of BC may be hard at the beginning of the disease, due to the absence of symptoms, after some clinical tests, the accurate diagnosis should have the ability to differentiate the benign and malignant tumors. A good detection provides low false positive (FP) rate and false negative (FN) rate[3].

78-1-5386-5135-3/18/$31.00 ©2018 IEEE

Machine learning is a set of tools utilized for the creation and evaluation of algorithms that facilitate prediction, pattern recognition, and classification. ML is based on four steps: Collecting data, picking the model, training the model, testing the model [4]. The relation between BC and ML is not recent, it had been used for decades to classify tumors and other malignancies, predict sequences of genes responsible of cancer and determine the prognostic [5, 6]. The classification's aim is to put each observation in a category that it belongs to. In this study, we used two machine learning classifiers which are Naïve Bayesian Classifier and k-nearest neighbor. The purpose is to determine whether a patient has a benign or malignant tumor. In this study, we customize two techniques of machine learning for classification of breast cancer. We use the Wisconsin breast cancer database. The purpose of this article is developing effective machine learning approaches for cancer classification using two classifiers in a data set. The performance of each classifier will be evaluated in terms of accuracy, training process and testing process.

II. BACKGROUND

In this section, we first introduce the breast cancer classification, then different machine learning techniques used in our cancer classification.

A. Brest cancer classification (BCC)

BCC aims to determine the suitable treatment, which can be aggressive or less aggressive, depending on the class of the cancer. To make a good prognostic, breast cancer classification needs nine characteristics which are: 1.determine the layered structures **(Clump Thickness)**; 2. Evaluate the sample size and its consistency (**Uniformity of Cell Size**); 3. Estimate the equality of cell shapes and identifies marginal variances, because cancer cells tend to vary in shape (**Uniformity of Cell Shape**); 4. Cancer cells spread all over the organ and normal cells are connected to each other (**Marginal Adhesion**); 5. Measure of the uniformity, enlarged epithelial cells are a sign of malignancy (**Single Epithelial Cell Size**); 6. In benign tumors nuclei is not surrounded by cytoplasm **(Bare Nuclei)**; 7. Describes the nucleus texture, in benign cells it has a uniform shape. The chromatin tends to be coarser in tumors (**Bland Chromatin**); 8. In normal cells, the nucleolus is usually invisible and very small. In cancer cells, there are more than one nucleoli and it becomes much more prominent, (**Normal Nucleoli**); 9. Estimate of the number of mitosis that has taken place. The larger the value, the greater is the chance of malignancy (**Mitoses**) [7]. In order to classify BC, pathologists assigned to each of these

characteristics a number from 1 to 10. The likelihood of malignancy needs the nine criteria, even if one of them is very large.

A. Machine learning approaches

Machine learning is branch of artificial intelligence, ML methods can employ statistics, probabilities, absolute conditionality, Boolean logic, and unconventional optimization strategies to classify patterns or to build prediction models[8]. Machine learning can be divided into two categories: supervised learning (classification) and unsupervised learning. Depending on the used data and their availability [9]. In this section, we will see two supervised learning classifiers.

1) Naïve Bayesian Classifier (NBC)

A Bayesian method is a basic result in probabilities and statistics, it can be defined as a framework to model decisions. In NBC, variables are conditionally independent; NBC can be used on data that directly influence each other to determine a model. From known training compounds, active (D) and inactive (H), Given representation B, the conditional probability distribution P(B/D) and P(B/H) are estimated, respectively. Bayesian classifiers are additionally well adapted for ranking of compound databases all with consideration to probability of activity [10].

Bayesian classifiers use Bayes theorem, which is:

$$p(h \mid d) = \frac{p(d \mid h)p(h)}{p(d)} \quad (1)$$

In Eq. 1, P(h) is the priori probability that event h will occur. P(d) is the prior probability of the training data. The conditional probability of d when p (d | h) is given. P(h | d) is the conditional probability of h when given d training data. P (h | d) is the probability of generating instance d given class h. In the equation above Bayesian decision theorem is used to determine whether a given x_i belongs to S_i where S_i represents a class [11]:

$$P(x|S_i)P(S_i) > P(x|S_j)P(S_j) \quad (2)$$

In the Eq. 2, $j \neq I$ which means that Si and Sj are two different classes and X belongs to Si.

2) k-Nearest Neighbors (KNN)

The KNN algorithm is used to predict the class or property of data. Given N training vector, suppose we have A and Z as training vectors in this bidimensional features space**,** we want to classify c which is feature vector. Classifying c depends on its k neighbors, and the majority vote, k is a positive integer, k is generally smaller then 5, if k=1 the class of c is the closest element from the two sets to c [12]. We use the Euclidean distances to evaluate the distance of a sample with other points, Euclidean distance is given in equation 3.

$$d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (3)$$

C. Cross validation

Cross-Validation is a statistical technique; it is generally used to check and evaluate learning algorithms or models, by partitioning data into a learning set to train the model and testing set to evaluate it.

The training and testing sets in cross-validation are randomly divided into partitions (60% of data are in training sets and 40% of data are in testing sets) and go through successive crossover rounds so that each instance is being tested against. K-fold cross validation is the basic form, one of the K partitions it is used as a validation set. There are more complicated forms of cross validation using k-fold as a base. [13].

III. Related works

A lot of studies have been done in the field of BCC and Ml, some of them used mammography images and the issue is that images can miss about 15% of breast cancer [14], some techniques are more specific and used genome or phenotypes to do classification [15, 16]. The breast cancer is classified with serval techniques such as Softmax Discriminant Classifier (SDC), Linear Discriminant Analysis (LDA) [17], and Fuzzy C Means Clustering [18]. The knearest neighbors algorithm is one of the most used algorithms in machine learning [19, 20]. Before classifying a new element, we must compare it to other elements using a similarity measure [14]. In cancer classification, KNN can be used to measure the performance of false positive rates [21, 22]. Naïve Bayesian classifiers are generally used to predict biological, chemical and physiological properties. In cancer classification, NBC are sometimes combined to other classifiers such as decision tree to determine prognostics or classification models.

Different classification techniques were developed for breast cancer diagnosis, the accuracy of many of them was evaluated using the dataset taken from Wisconsin breast cancer database [23]. For example, in [24] the optimized learning vector method's performance was 96.7%, big LVQ method reached, SVM for cancer diagnosis's accuracy is 97.13% is the highest one in the literature .

IV. The proposed algorithms

A. Datasets

The Breast Cancer Dataset (BCD) that we used is donated to the University of California, Irvine (UCI). There are 11 attributes and the first one is ID that we will remove (it is not a feature we actually want to feed in our classification). The nine criterions are as discussed earlier in breast cancer classification section, they are meant to determine if a tumor is benign or malign, the last feature contains a binary value (2 for benign tumor and 4 for malign tumor). The set consists of 699 clinical cases. The initial BCD contains missing data for 16 observations, which limited our dataset to 683 samples.

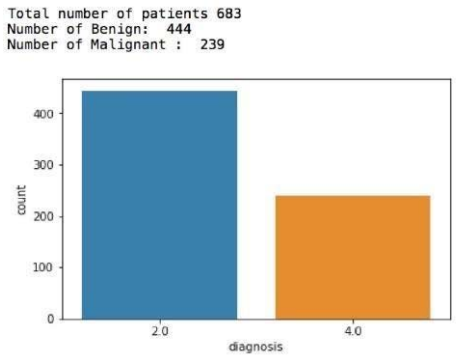

FIGURE 1: WISCONSIN BREAST CANCER DATASETS

Figure 1 shows that 444 (65%) tumors are benign tumors and 239 (35%) tumors are malign.

B.Nearest Neighbors Algorithm (k=3) for breast cancer classification

1) Algorithm:

1- Input the dataset and split it into a training and testing set.

2- Pick an instance from the testing sets and calculate its distance with the training set.

3- List distances in ascending order.

4- The class of the instance is the most common class of the 3 first trainings instances (k=3).

2) Description:

Given a sample of N examples and their classes. We split the data for cross validation and testing stages. The training stage in KNN is nonexistent, as we compare every new instance each time. To predict the outcome of a new instance, we calculate the Euclidean distance between the instance and all the points in the training set.

C. Naive Bayes classifier for breast cancer classification (NB)

1) Algorithm:

1- Separate data into block of 2 classes and 2 sets of features T and classes D.

2- Calculate the mean and standard deviation of each feature and each class.

3- Generate a summary for each feature and for each class. 4- Calculate the probability of each feature using the density of normal distribution.

5- Calculate the probability of each class as a multiplication of the probabilities of all features. 6- To predict the class of an instance from the testing set, calculate the probability of each class.

2) Description:

The algorithm that we used uses the same Naive Bayes primitive, we first divided the dataset into a testing and training sets. The training phase consists first of separating the set into 2 different sets: D is the presence of the tumor and T is a set of features test and then to separate the D set into 2 classes malignant and benign (4 or 2). In the following step, we calculated the mean, standard deviation for each feature from set T and then for each class from set D. We ended up with a summary for each feature and each class that we will use for our prediction see equation (see equation 5).

$$P(T \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(T-\mu)^2}{2\sigma^2}} \quad (5)$$

Our prediction is based on the multiplication of each probability of a feature given a class. The probability of each feature is calculated using the density of normal distribution (6).

$$P(D) = \prod P(T \mid \mu, \sigma^2) \quad (6)$$

μ: The mean is average of each features.

σ: The standard deviation

P (D): The probability of each tumor class P (T):

The probability of each feature.

The testing phase consists of calculating the probability of each class given an instance from the testing set, the class chosen is the class with the largest value.

From Table 1 we can notice that the two algorithms are extremely effective in the diagnosis, all of which show a high level of accuracy despite the small dataset.

Table 1: comparison between KNN and NB

Method	Accuracy	Training process	Test process	Total process
KNN	0.975109	0.000735	0.001744	0.002479
NB	0.961932	0.000759	0.000422	0.001182

For this example, KNN classifiers are ranked first in terms of accuracy and duration. As a result, KNN is the most effective classifier for this cancer classification problem. However, if the dataset is larger, the KNN will lose the first order because of the time complexity of the computation that need to be done.

V. CONCLUSION

On the Wisconsin Breast Cancer datasets, we used our two main algorithms, which are: NB & KNN, since our target and challenge from breast cancer classification is to build classifiers that are precise and reliable. After an accurate comparison between our algorithms, we noticed that KNN achieved a higher efficiency of 97.51%, however, even NB has a good accuracy at 96.19 %, if the dataset is larger, the KNN's time for running will increase.

References

[1] L.A. Altonen, R. Saalovra., P. Kristo, F. Canzian, A. Hemminki, Peltomaki P, R. Chadwik, A. De La Chapelle, "Incidence of hereditary nonpolyposis colorectal cancer and the feasibility of molecular screening for the disease", N Engl J Med, Vol. 337, pp. 1481–1487, 1998.

[2] S.Chakraborty, "Bayesian kernel probit model for microarray based cancer classification", Computational Statistics and Data Analysis, Vol. 12, pp. 4198–4209, 2009.

[3] I. Guyon, J. Weston, S. Barnhill, V. Vapnik. "Gene selection for cancer classification using support vector machines". Machine Learning, Vol. 46, pp. 389–422, 2002.

[4] S. Gokhale., "Ultrasound characterization of breast masses", The Indian journal of radiology &imaging, Vol. 19, pp. 242-249, 2009.

[5] T. Jinshan, R.R., X. Jun, I. El Naqa, Y. Yongyi, "Computer-Aided Detection and Diagnosis of Breast Cancer With Mammography: Recent Advances", Information Technology in Biomedicine. IEEE, Vol. 13, pp. 236-251, 2009.

[6] A. Jemal, R.S., E. Ward, Y. Hao, J. Xu, T. Murray, M.J.
Thun, "Cancer statistics", A Cancer Journal for Clinicians, Vol. 58, pp. 71-96, 2008.

[7] L. Adi Tarca, V.J.C., X. Chen, R. Romero, S. Drăghici, "Machine Learning and Its Applications to Biology", PLoS Comput Biol,, Vol. 3, pp. 116-122, 2007.

[8] JF McCarthy, M.K., PE Hoffman, "Applications of machine learning and high-dimensional visualization in cancer detection, diagnosis, and management", Ann N Y Acad Sci, Vol.62, pp. 10201259, 2004.

[9] AC. Tan, D. Gilbert, "Ensemble machine learning on gene expression data for cancer classification", Appl. Bioinform, Vol. 2, pp. 75-83, 2003.

[10] S. Kanta Sarkar, A.N., "Identifying patients at risk of breast cancer through decision trees", International Journal of Advanced Research in Computer Science. Vol. 08, pp. 88-96, 2017.

[11] JA. Cruz, W.D, "Applications of Machine Learning in Cancer Prediction and Prognosis". Cancer Inform, Vol. 2, pp. 56-77, 2006.

[12] M. Sugiyama, "Introduction to Statistical Machine Learning "1ed, ed. T. Green: Morgan Kaufmann, 2006.

[13] A. Lavecchia, "machine-learning approaches in the context of ligand-based virtual screening for addressing complex compound classification problems and predicting new active molecules". D. Montesano, 49, 2005.

[14] P. Baldi, S.R.B., Bioinformatics: The machine learning approach. 2 ed, ed. S.r.B. Pierre Baldi, 2001.

[15] N. Bhatia, "Survey of Nearest Neighbor Techniques", International Journal of Computer Science and
Information Security, Vol. 8, No. 2, 2010.

[16] A. Francillon, P.R., "Smart Card Research and Advanced Applications": 12th International Conference, CARDIS 2013, Berlin, Germany, 2013.
Revised Selected Papers. 1 ed. Lecture Notes in Computer Science 8419 Security and Cryptology2014: Springer International Publishing, November 27-29.

[17] A. Alarabeyyat, A.M., "Breast Cancer Detection Using K-Nearest Neighbor Machine Learning Algorithm", in 9th International Conference on. IEEE, v.i.e.E. (DeSE), pp. 35-39, 2016.

[18] MF. Akay. "Support vector machines combined with feature selection for breast cancer diagnosis". Expert Syst Appl Vol. 36, Issue. 2, Part. 2, pp. 3240-3247, March 2009.

[19] S.K. Prabhakar, H. Rajaguru, "Performance Analysis of Breast Cancer Classification with Softmax Discriminant Classifier and Linear Discriminant Analysis", In: Maglaveras N., Chouvarda I., de Carvalho P. (eds) Precision Medicine Powered by pHealth and Connected Health. IFMBE Proceedings, vol 66. Springer, Singapore, 2018.

[20] J. S. Snchez, R.A.M., J. M. Sotoca. "An analysis of how training data complexity affects the nearest neighbor classifiers", Pattern Analysis and Applications, Vol. 10, Issue 3, pp 189–201, August 2007.

[21] M. Raniszewski, "Sequential reduction algorithm for nearest neighbor rule", Computer Vision and Graphics, 2010.

[22] P.BhuvaneswariaA, B. Therese, "Detection of Cancer in Lung with K-NN Classification Using Genetic Algorithm", Procedia Materials Science, Vol. 10, pp. 433-440, 2015.

[23] Z. Zhou, Y.J., Y. Yang, S.F. Chen, "Lung Cancer Cell Identification Based on Artificial Neural Network Ensembles Artificial Intelligence", Medicine Elsevier, Vol. 24, pp. 25-36, 2002.

[24] A. Pradesh, A.o.F.S.w.C.B.C.D.," Indian J. Comput. Sci. Eng., vol. 2, no. 5, pp. 756–763, 2011.

[25] D. Delen, G.W., and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," and v. Artif.Intell. Med., pp. 113– 127, 2005.

[26] M.H. Asri, H.A Moatassime, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis". Procedia Comput Sci, Vol. 83, pp. 1064–1073, 2016.