Diagnosis With Several ML Techniques In Breast Cancer Data

ONUR BERKAY DOGAN
COMPUTER ENGINEERING
MIDDLE EAST TECHNICAL UNIVERSITY
ANKARA,TURKEY

***Abstract*— Technology is used in almost every area today. One of them is medicine. The use of technology in medicine has become widespread. The use of technology in the diagnosis of cancer patient has become quite popular. The importance of early and accurate diagnosis in cancer disease has been proven by many medical specialists. In this paper, some machine learning methods have been applied to breast cancer data set.**

Keywords— Breast Cancer, Preprocessing, Normalization, Descriptive Analysis

I. INTRODUCTION

Breast cancer continues to be the most common cancer and the first largest cause of cancer deaths among women. The annual mortality rate of approximately 28 deaths per 100,000 women has remained nearly constant over the past 20 years.

The problem is to ensure that the breast cancer cell is determined as benign or malignant by using the information given to us in the breast cancer data set. I aim to diagnose breast cancer faster and more quickly. Also, I searched for answers to these questions; which machine learning model most accurate of classification score and what pathological feature is most important for diagnosing breast cancer.

II. LITERATURE SURVEY

Throughout the research, there are several sample studies on breast cancer that guide me. First,

Jerez, José M., et al. "Missing data imputation using statistical and machine learning methods in a real breast cancer problem." Artificial intelligence in medicine 50.2 (2010): 105-115, paper predicting the accuracy rate for used various algorithms of IBK, Simple Logistic, Naive bayes, Decision table and Multiple perception using WEKA tool. Simple logistics has produced the accuracy of 99. 7612 % in comparison with other algorithms.

Second research,

Vikas, C., Saurabh, P.:A Novel Approach for Breast Cancer Detection using Data Mining Techniques. International Journal of Innovative Research in Computer and Communication Engineering. Vol. 2, their analyse breast cancer data available from Wisconsin dataset from UCI. They compare three classification techniques; Sequential Minimal Optimization (SMO), K Nearest Neighbours Classifiers, BF Tree. This paper empirically compares performance of three classical decision classifiers that are suitable for direct interpretability of their results. In this study mentioned build precise and computationally efficent classifiers in Medical applications is important challenge. Therefore SMO classifier is suggested for this study. This study also show that the most important attributes using Chi-square test, Info Gain test and Gain Ratio Test.

Finally,

In this paper they investigated the generalization performance of RepTree, RBF Network and Simple Logistic in order to enhance the prediction models for decision-making system in the prognosis of breast cancer survivability. The best algorithm based on the patient's data is Simple logistic Classification with accuracy of 74.47% and the total time taken to build the model is at 0.62 seconds.

III. DATA PREPARATION AND DESCRIPTIVE ANALYSIS

A. Descriptive Analysis

I found my breast cancer dataset on the Kaggle website. This dataset has numbers describing each of the features that are monitored when trying to detect breast cancer. The dataset contains 30 features: diagnosis, Radius_mean, texture_mean, perimeter_mean, area_mean, smoothness_mean, compactness_mean, concavity_mean, concave_points_mean, symmetry_mean, fractal_dimension_mean, Radius_se, texture_se, perimeter_se, area_se, smoothness_se, compactness_se, concave_points_se, symmetry_se, fractal_dimension_se, Radius_worst, texture_worst, perimeter_worst, area_wrost, smoothness_worst, compactness_worst, concavity_worst, concave_points_worst, symmetry_worst, fractal_dimension_worst. Breast cancer dataset contains some features equal to zero. The dataset contains 569 samples. The 'diagnosis' feature in the dataset shows which type of tumor it has. If the 'diagnosis' feature is 'M', the sample belongs to the malignant class. If the 'diagnosis' feature is 'B', the sample belongs to the benign class. The 'diagnosis' feature in our dataset is a categorical feature. All other features are continuous. All of 569 samples in the dataset have their own 'id' features. Class distribution: Benign: 357 (%62.7), Malignant: 212 (%37.3)

B. Data Preprocessing

The preprocessing and quality checks techniques I use in this paper are as follows:

The collected data may be in complete, inconsistent or outdated. Data collection methods are generally loosely controlled and occurs problems such that out of range values (Ex: Income= -100), imposible data combinations (Ex: Gender: Male and Pregnant:Yes), missing values etc. that are not carefully scanned can produce misleading results. Therefore, preprocessing should be done for the representation and more quailty of the data. Data Preprocessing is should be done before all methods are applied.

In this paper, I applied normalization to breast cancer dataset. Normalization means reducing the input value. It is used when there is a difference between the data and handling data in a single order. It also often refers to rescaling by the minimum and range of the vector to make all the elements between 0 and 1 thus bringing all the values of numeric columns in the dataset to a common scale.

Other preprocessing techniques I apply to the dataset are to convert the categorical features to numerical features and to divide the data set into two as training and test. I allocated 20% of the data set as a test, %80 of the dataset as a train.

IV. Machine Learning Techniques

The machine learning techniques i use this paper are as follows:

A. PCA

PCA (Principal Component Analysis) is used in high dimensional data. It is a technique whose main aim is to store the dataset with the highest varince in high dimensioanl data, but to achieve dimension reduction while doing so. With the decrease in dimension, some features are lost but the aim is that these lost features contain little information about the population. It combines highly correlated variables and create small number of set of artificial variables called principal component that make up the most variation in data.

$$\sigma^2=(1/n-1)\sum(xi-\bar{x})^2 \quad (1)$$

The basis of PCA (Principal Component Analysis) is based on the spectral features of the covariance and correlation matrix between variables in datasets. Use "(1)", the covariance matrix is calculated. This matrices are symmetrical and positive. The eigenvalues of these matrices are identical to their variances. In other words, PCA is the process of finding eigenvalues and eigenvectors of datasets of covariance and correlation matrices.

PCA generally consist of 5 basic steps: Prepare the dataset, calculating the covariance/correlation matrix, calculating eigenvalues and eigenvectors of the covarince/correlation matrix, selecting principal component, calculating the new dataset.

In this paper, i decided number of components determined two.

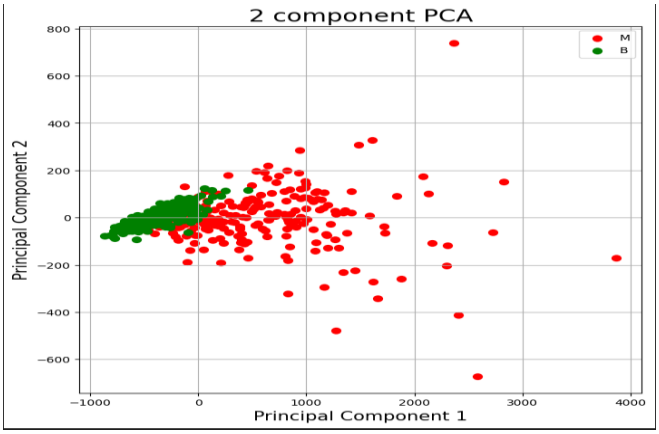

Fig. 1. Plot of 2 Component PCA over "M" and "B" classes

Figure 1. shows the result of PCA over two target class. Red color dots represent Malign class, Green color dots represent Beningn class.

B. KNN

KNN (K Nearest Neighborhood) is one of the easy to implement supervised learning algorithm. KNN is used in classification and regression problems. The algorithm uses data from a sample set with certain classes. The distance of the new data to be included in the sample dataset is calculated according to the existing data and k nearest neighborhoods are checked.

$$D(P,Q) = \text{sqrt}((x1-y1)^2+(x2-y2)^2 \quad (3)$$

$$D= \sum|xi-yi| \quad (4)$$

Generally 2 types of distance functions are used for distance calculations: Euclidean distance, Manhattan distance. Equation "(3)" is formulation of euclidean distance. Equation"(4)" is formulation of manhattan distance.

KNN is one of the most popular machine learning algoithms as it is resistant to old, simple and noisy educational data. However, it also has a disadvantage. For example, it requires a lot of memory space when used for big data.

KNN algorithm: First of all the k parameter is determined. This parameter is the number of neighbors closest to a given point. The distance of the new data to be included in the sample dataset is calculated individually according to the existing data. The closest k neighbors are considered from the related distances. According to the attribute values, k is assigned to neighbor or class of neighbors. The selected class is considered the class of the observation value expected to be estimated. In other word, the new data is labeled.

KNN provide easy to implementation and also provides parallel implementation. It can be monitored analytically. Its performance depends on the number of k neighbors, distance criteria and number of attributes.

In this paper, I implement k-nn with five as a "n_neighbors" parameter. After the implementation, I invest the optimum "k" value for knn algorithm. Implemented Mean Error – Error Rate.

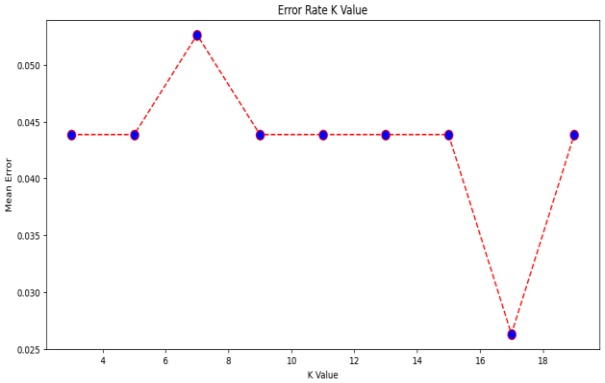

Fig. 2. Mean Error – Error Rate K value plot

Figure 2. shows us best k value is 17. K = 17 parameter has the lowest error rate, on the contrary k = 7 highest error rate.

C. Decision Tree

Decision Tree method is one of the most popular algorithms for machine learning used in both classification and regression problems. It is usually at a human level, so it is very simple to understand the data and make some good comments and visualize it. Decision Tree is a recursively process and a tree structure is used. It starts with a single node and branches to new results, creating a tree structure. When the algorithm runs, the entered value proceeds on a certain path bu looking at the nodes and gives a result.

Decision Trees requires fast data preprocessing. Compared to most alternative techniques, data becomes available with little processing. Preprocessing stage is shorter and simpler than other alternatives.

Most machine learning algorithms are either useful in numerical applications or useful for classification problems. Decision Tree algorithms can be used in both areas.

Decision Tree has low computational complexity. Due to its simple and fast processing, it can process large amount of data in a short time and compared to alternative methods, it becomes more preferable when the amount of data increases. It uses White box model so every step can be viewed and interpreted.

In this paper, Entropy choosed as criterion for DecisionTreeClassifier.

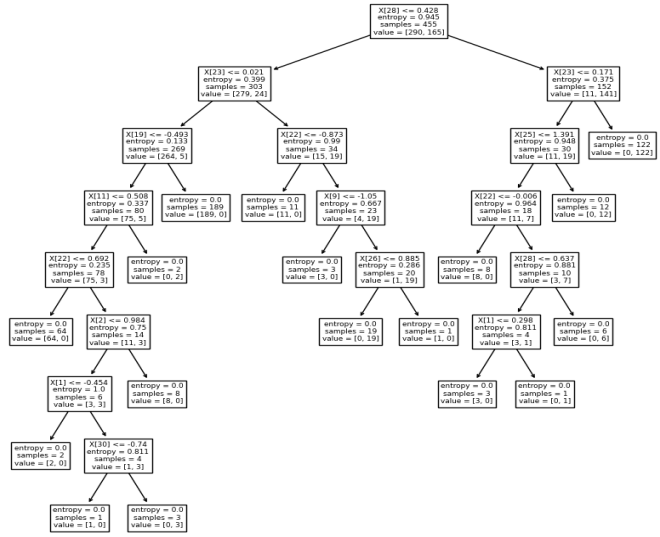

Fig. 3. Visualization of Decision Graph with entropy value

Figure 3. demonstrate the final version of decision tree nodes and leaves. Every single node has indivual entropy value and deision values every features.

V. Evaluation Metrics

The method I use to evaluate techniques is confusion matrix. A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. Calculating a confusion matrix can give you a better idea of what your classification model is getting right and what types of errors it is making. It allows the visualization of the performance of an algorithm.

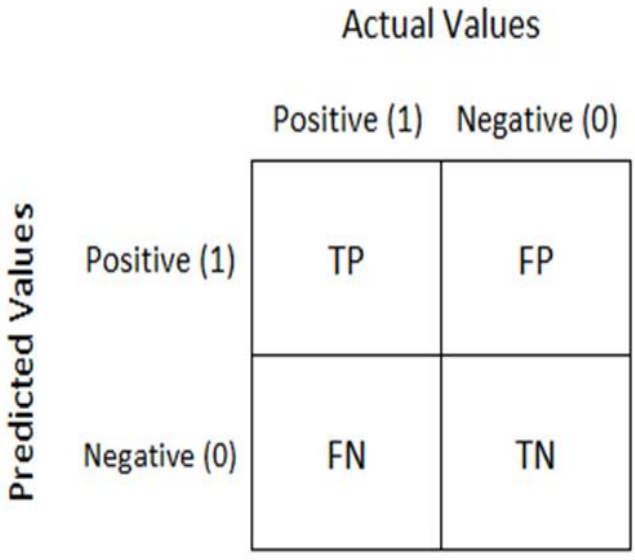

Fig 4. Confusion Matrix

The table shown in Fig. 4 shows the confusion matrix. This matrix helps us calculate the accuracy of the classifier.

True Positive and True Negative values are actually the number of samples that belong to that class and the classifier also distinguishes correctly.

False Positive and False Negative are the sample numbers that the classifier incorrectly classified.

Our first classifier evaluation metric is accuracy, as in

$$\text{Accuracy} = (TP+TN)/(TP+TN+FN+FP) \quad (7)$$

Target classes in our dataset are "B" and "M". The evaluation of different classifier is made with this formula. Accuracy shows the classifier's performance for all samples. A 99% accuracy can be excellent, good, mediocre, poor or terrible depending upon the problem.

$$\text{Sensitivity} = TP/(TP+FN) \quad (8)$$

Sensitivity (SN) is calculated as the number of correct positive predictions divided by the total number of positives. It is also called recall (REC) or true positive rate (TPR). The best sensitivity is 1.0, whereas the worst is 0.0. High Recall indicates the class is correctly recognized. Equation "(8)" is mathematical expression of sensitivity. Our second classifier evaluation metric is sensitivity.

$$\text{Precision} = TP/(TP+FP) \quad (9)$$

Precision is calculated as the number of correct positive predictions divided by the total number of positive predictions. It is also called positive predictive value (PPV). The best precision is 1.0, whereas the worst is 0.0. High Precision indicates an example labelled as positive is indeed positive. Equation "(9)" is mathematical expression of precision.

If Sensitivity is high and **Precision is low, t**his means that most of the positive examples are correctly recognized (low FN) but there are a lot of false positives.

If Sensitivity is **low and Precision is high,** we miss a lot of positive examples (high FN) but those we predict as positive are indeed positive (low FP).

$$\text{Specificity} = TN/(TN+FP) \quad (10)$$

Use "(10)", specificity (SP) is calculated as the number of correct negative predictions divided by the total number of negatives. It is also called true negative rate (TNR). The best specificity is 1.0, whereas the worst is 0.0. Our third classifier evaluation metric is specificity.

VI. Conclusion

Table I. Comparison of KNN with Different k Values

	K = 1	K = 3	K = 5	K = 7
Accuracy	0.94	0.94	0.96	0.96
Sensitivity	0.95	0.97	0.95	0.94
Specificity	0.92	0.88	0.886	0.886

The table shown in Table I shows the classification results made with different k values on 143 sample test data of kNN method i applied to the dataset. The highest accuracy is k=5 and k=7. The highest Sensitivity is k=3 and k=5. However, unlike accuracy and Sensitivity values, the highest specificity is k=1 so it is more successful than the other k values in the correct classification of negative values. I decide to use k=5 in the kNN method.

Table II. Comparison of Classifier Models

	KNN	Decision Trees
Precision	0.97	0.93
Recall	0.95	0.94
Accuracy	0.96	0.94

The table shown in Table II shows comparison of classifier models. Lineer Discriminant Analysis is the highest accuracy of techniques we apply.

The prediction of breast cancer recurrence. In this paper, I analyzed cancer data using two classification techniques to predict the recurrence of the cancer and then compared the results. The results indicated that KNN is the way better classifier predictor with the test dataset. Further studies should be conducted to improve performance of these classification techniques by using more variables and choosing for a longer follow-up duration.

References

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE

Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.