A Correlation Based Way to Predict the Type of Breast Cancer for Diagnosis

Shahidul Islam KHAN
Department of Computer Science and Engineering (CSE)
International Islamic University Chittagong (IIUC)
Chittagong, Bangladesh
nayeemkh@gmail.com

Nowadays, breast cancer is considered one of the most common causes of death among adult women. At the same time, the bright side is that among all the types of cancer, breast cancer is more curable, if diagnosed in the early stages. In this paper, the diagnosis of breast cancer has been proposed using the least possible number of features based on correlation. In the proposed method, we have used correlation to find the strength between the input and the target features. Then we provided a way to create a new subset that consists of only the most relevant features. We have used the Wisconsin breast cancer data set (WBCD) for the experiments. The performance of the model is justified using classification accuracy and the f-score. The result shows that our proposed method obtained the highest classification accuracy (95.26%) with the Random Forest classification using only 4 features from 29 available features, which led to a reduction of 86% in data set size.
Keywords: Health Data; Feature Selection; Correlation; Breast Cancer; Classification

1 Introduction

The use of healthcare information and information technology to organize and analyze health data to improve the quality and safety of patient health care is broadly known as Health Informatics. It deals with the resources from the healthcare sector, machinery and methods used in healthcare to acquire, store, retrieve, and use knowledge in health and medicine. It provides a way to access medical records digitally for health information enthusiasts. In this era of big data, health informatics is a fast-growing field that plays a vital role in the reformation of healthcare [1, 2].

Cancer is the deadliest disease that mankind is facing now. It is a kind of disease in which body cells grow or change out of control. Cancer may form in almost every major part of a human body and is named after the body part it affects. Hence, breast cancer is a term that refers to the abnormal growth in the breast cells. The extra masses caused by abnormal growth are called tumors. The two types of tumors are malignant, which is supposed to be breast cancer, and benign, which is not cancerous and not life-threatening. Today, breast cancer is one of the most common causes of death among middle-aged women (40-55 years). As indicated by the World Health Organization, 2.1 million women are affected each year by breast cancer. In 2018, 627,000 women were estimated to have died from breast cancer, which is roughly 15% of all cancer-caused death among women [3].

Health informatics may play an important role in improving the survival rates of breast cancer patients. Early detection of the type of a tumor by performing the least possible diagnostic test may help to push toward achieving the maximum survival rate.

The feature is a distinguishable attribute or perceptible characteristic of something that is being observed. Feature selection is a way that removes unimportant and redundant features. It generates a subset of features with less dimensionality than the original data set and still provides good prediction results.

The use of machine learning approaches in the health care field is increasing gradually. In the diagnosis of a disease, the most important factors are the patient's data and the result obtained from the diagnostic tests.

In general, patients have to take a lot of diagnostic tests based on which doctors lead to a decision on whether the patients have benign or malignant type tumors. Current information systems for detecting the type of breast cancer need a lot of features, which is time-consuming. By selecting the most relevant features from the original set, the features needed for the detection of breast cancer type are reduced, which will automatically reduce the number of tests needed as well as the time consumption. Hence, feature selection can be a supportive tool for a doctor in diagnosis and decision-making.

Many authors have used a different type of approach for selecting features, but very few of them have used the correlation method. So, it is a research issue to select features from breast cancer data sets using correlation with the target variable.

In this paper, we have presented a brief overview of a feature selection technique that uses a feature-ranking method based on the correlation of input features with the target. We have provided a way to mitigate the problem of redundant features. We have used two types of correlation methods to find the strength of association between an input feature and the target. Pearson's correlation was used to calculate the correlation between the numeric input and the numeric output. Point Biserial correlation was used to find the correlation between numeric input and binary output and vice versa. We showed that the accuracy of a model can be preserved or improved even after selecting a small subset of features from the original set.

The remainder of the paper is organized as follows. In Section II, we have briefly presented research works related to the diagnosis of breast cancer. Section III describes the feature selection. In Section IV, we have presented the methodology and experiments of our proposed method. The results obtained by applying the proposed method are presented in Section V. Section VI finally concludes the paper.

2 Related works

There has been a great deal of research on breast cancer diagnosis, where the data set was the same as ours in the literature. In [4], the authors have combined two techniques: an evolutionary algorithm and fuzzy systems to automatically report the system of diagnosis. They obtained a classification accuracy of 97.36%.

The authors of [5] have developed a knowledge-based system. The system uses the clustering, noise removal, and classification technique. To cluster the data, they have used Expectation-Maximization, Classification, and Regression Tree to generate the fuzzy rule and PCA to overcome the multiple collinearity issue. They obtained 93.20% accuracy on the WDBC data set.

A combination of an Artificial Immune Recognition System and Synthetic Minority Over-Sampling Technique in a system is presented in [6]. They have compared their result with other classifiers like AIRS, BPNN, C4.5, etc. They obtained 96.53% accuracy using their method.

A novel fuzzy model structure that is an extension of the quadratic Bayes classifier has been presented in [7]. The authors analyze the clusters using Fisher's interclass separability criteria to select the relevant input variable. They obtained 95.57% accuracy by applying the supervised fuzzy clustering technique.

RIAC, a method that stands for Rule Induction through Approximate Classification, was presented in [8]. This method was used for inducing rules from examples, which is based on the theory of rough sets, and obtained 94.99% accuracy.

In [9] they have used 10-fold-cross-validation along with the C4.5 decision tree method. They gained 94.74% classification accuracy. The authors of [10] have presented a convenient approach for learning fuzzy classifiers from data. They obtained 95.06% accuracy using the method called neuron-fuzzy techniques.

In summary, the above-cited works have gained promising results, but none of them

have used the feature ranking based on correlation.

3 Feature selection

Feature selection is a process that removes irrelevant and redundant features with little or no predictive information from data sets before an algorithm is applied, generates a subset of features with less dimensionality than the original data set had, and still provides good prediction results.

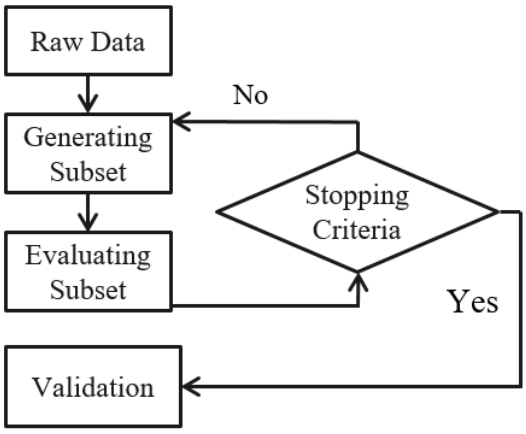

Fig. 1. Basic Procedure of Feature Selection

Fig. 1 demonstrates the progression of the normal technique adopted in feature selection. The dimension space is scaled down to a subset of features that are evaluated based on the criterion. Then, the selected features are validated by the validation process. Stopping criteria are used as an end process indicator; the process may stop if any of the following criterion is fulfilled [11]:

i. Some predefined features have reached;
ii. New features addition/deletion does not bring an improved result;
iii. The selected feature meets the best possible outcome according to the evaluation criterion.

A large number of algorithms have already been proposed to solve the 'curse of dimensionality. Here some of the methods are presented briefly:

3.1 Filter method

The filter method is one of the simplest and computationally less expensive approaches, in this approach feature selection is performed as a preprocessing step. It uses a ranking method to score the features and then the compares the score with a predefined threshold value. If the score is less than the threshold value, then the feature is considered to be irrelevant and gets eliminated. There are three common measures to rank/score a feature; they are distance metrics, correlation, and mutual information.

3.2 Wrapper method

The wrapper method is an approach that evaluates all possible combinations of features to select the subset that leads to the best output [12]. As this method tests all the possible combinations, this can become computationally expensive when the data set is very large. Since the wrapper method evaluates 2n subsets, it becomes an NP-hard problem. To find the subset on the wrapper method, several search algorithms are used. These algorithms can be categorized into Exhaustive Search, Non-Exhaustive Search and Heuristic Search.

3.3 Embedded method

The embedded method is a feature selection method where features are not gets selected or rejected [13]. In this approach, feature selection is integrated into the learning algorithm. Features with less importance are given low weight, which is also called regularization. There are a few types of embedded techniques: Decision tree, LASSO Regression, RIDGE Regression, etc.

Several more techniques deal with dimensionality reduction such as PCA, Clustering, Missing value ratio, Boruta, SelectFromModel, etc.

4 Methodology and experiments

4.1. Breast cancer data set

We have utilized a Breast Cancer Wisconsin (Diagnostic) data set which we

took from the Machine Learning Repository of the University of California, Irvine [14]. Researchers who use ML approaches for the diagnosis of breast cancer commonly use this data set. The data set contains 569 data. It consists of 32 features computed from the digitized image of FNA of breast masses. ID was used for identification and diagnosis (M for malignant and B for benign) as the target variable. The rest of the features are 10 real-valued features; these are area, texture, compactness, radius, smoothness, concavity, perimeter, concave points, symmetry, and fractal dimensions. Each of these features was used in three different forms: mean, se (standard error), and worst. For example, radius_mean, radius_se, radius_worst, etc. each of which has a numeric value. In the data set, 357 samples belong to the benign class, and the rest 212 are of the malignant class.

4.2 Correlation

Correlation plays an important role in building a feature selection model. It is advantageous, as it helps to find the input features highly correlated with the target. Measurement of linear dependencies between features correlation coefficient is a widely used method. There are a few types of correlation coefficient measurement methods, such as Pearson's correlation coefficient, Kendall's rank correlation coefficient, etc. we have used Pearson's and Point Biserial Correlation coefficient. Pearson's coefficient can be expressed as:

$$\mathrm{CoR(i)} = \frac{Cov\ (Xi, Y)}{\sqrt{var(Xi) * var(Y)}}$$

Here, CoR(i) is the correlation coefficient, which is obtained by dividing the covariance of the two variables (xi is the ith variable and Y is the output variable) by the product of their variance. Another correlation method is the point Biserial method, which can be expressed as:

$$r_{pb} = \frac{M_1 - M_0}{\sqrt{S_n}}\sqrt{pq}$$

Here, M_1 is the mean for the group that contains a positive binary variable. M_0 is the mean for the group that contains the negative binary variable. Sn is the standard deviation of the complete test. p and q represent the rate of cases in the "0" and "1" groups, respectively.

4.3 Proposed method

In this paper, a feature selection technique based on correlation is presented. We have used the correlation to rank the feature and the sorting method to evaluate all possible ranked features sequentially to select the subset that leads to the best output. We have applied the StandardScalar method to standardize data. To verify the results, we have used different classification techniques such as Random Forest, SVM (Support Vector Machine), Naïve Bayes, and KNN (K-nearest Neighbor). Using these methods, we checked the accuracy of the original data and then the proposed method was applied to make a subset of the reduced feature. Again, we checked the accuracy of the reduced data set using the classification technique mentioned above. Then we compared the result of the original data set and a reduced subset.

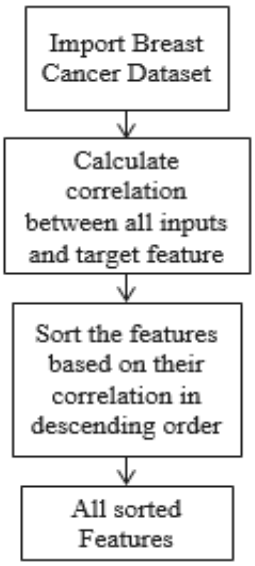

Fig. 2. Flowchart of the proposed method

Fig. 2 demonstrates how the proposed method works. Importations of high-

dimensional data sets to the selection of a reduced subset of features are shown in this flowchart.

The steps of the method are:

i. Import breast cancer data set;
ii. Compute the correlation between inputs and target variable using Pearson correlation and Point biserial correlation;
iii. Sort the features based on their correlation in descending order.

From the sorted feature, we will add one feature at a time to evaluate the result. The features are sorted based on the high correlativity with the target feature. Hence, a few numbers of features meet the desired output.

4.4 Measures for performance evaluation

We have used different equations to measure the performance of the developed method. These measures are precision-recall, accuracy, and F-score. They can be defined using a confusion matrix.

A confusion matrix is a table that helps to visualize the performance of an algorithm. In this 2*2 matrix, where instances in a predicted class are shown in a row, and the actual class is shown in a column.

		Predicted Class: P	Predicted Class: N
Actual Class	P	True Positive (TP)	False Negative (FN)
	N	False Positive (FP)	True Negative (TN)

Fig. 3. Confusion Matrix

To demonstrate the result, the following equations are used:

$$Accuracy\ (\%) = \frac{TP + TN}{P + N} \ldots i$$

$$Precision\ (\%) = \frac{TP}{TP + FP} \ldots ii$$

$$Recall\ (\%) = \frac{TP}{TP + FN} \ldots iii$$

$$F - Score(\%) = \frac{2 * Precision * Recall}{Precision + Recall} \ldots iv$$

5 Results and discussion

We have conducted experiments on the Breast Cancer data set to justify the effectiveness of our approach. The significance of each feature is calculated by its correlation with the target feature. Table 1 shows the results obtained by applying five different classification models. The result shows that the reduction of features does not affect the outcome to a greater extent. Among the five models, naïve Bayes achieved the highest classification accuracy; 96.31%, but in terms of selecting features random forest is more promising which selects only 4 features with an accuracy of 95.26%.

Fig. 4 and Fig. 5 are the performance v/s feature numbers after applying the Random Forest and KNN classification algorithm, where no. of features is on the x-axis and performance is on the y-axis. From the curve, it is seen that the accuracy has been increased initially and been almost the same in the middle and later part of the curve, which indicates that a smaller number of features are enough for obtaining the highest possible outcome.

Table. 1. Summary of the results of Breast Cancer data

Classification Algorithm	No. of Features		Accuracy	F-Score
Random Forest	Original	29	94.21%	90.85%
	Reduced	04	95.26%	93.13%
SVM	Original	29	95.78%	94.28%
	Reduced	07	95.26%	93.53%
Decision Tree	Original	29	92.63%	90.02%
	Reduced	04	93.16%	90.02%
KNN	Original	29	94.74%	92.75%
	Reduced	05	93.68%	91.04%
Naïve Bayes	Original	29	92.11%	88.89%
	Reduced	06	96.31%	94.81%

Fig. 4. Performance vs Feature numbers on Breast Cancer data set for Decision Tree

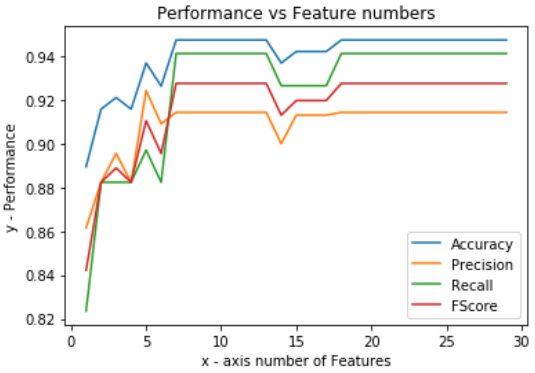

Fig. 5. Performance vs Feature numbers on Breast Cancer data set for KNN

The proposed method helps to reduce data set size. From Table 1 we can see that the difference between the result obtained from the original data set and the reduced data set is negligible. But what impacts the outcome to a greater extent is the size of the data. Table 2 shows a summary of the reduction in data size for different classification algorithms after applying the proposed method. In terms of reduction in size, Random Forest and Decision Tree outperformed other algorithms by reducing the size approximately by 86%. So when we will work with a larger data set in the future, we will get better results with only a few features.

Table. 2 Summary of the data set size-reduction

Classification Algorithm	No Original Feature	No of the Selected Feature	Reduction in Data Size
Random Forest	29	04	86.20%
SVM		07	75.86%
Decision Tree		04	86.20%
KNN		05	82.76%
Naïve Bayes		06	79.31%

From the results above, we presume that in classifying the potential breast cancer patients, promising outcomes have been obtained by the proposed method.

6 Conclusion

A method based on correlation and sorting of features according to their correlation has been applied to the task of predicting the type of breast cancer using the least possible number of features. As we have found the correlation between input and target features and sorted them, hence we were able to detect those features that make the most impact on a particular output. We observed that our proposed method has gained classification accuracies of 95.26%, 95.26%, 93.16%, 93.68%, 96.31% using 04, 07, 04, 05, 06 no. of features for Random Forest, SVM, KNN, decision tree, and naïve Bayes classification, respectively, without using the original (29 features) set. In terms of data set size reduction, the developed model outperformed all other models by reducing approximately 89% of the data set size.

Considering the results, the SVM and Random Forest-based model obtain the best results in classifying breast cancer using the developed method. We have high hope that the method proposed here can be very supportive for the health researchers in their ultimate decisions. They can make a decision within the least possible time using such a tool. Further exploration with a larger data set and finding the reduced subset for data with no target variable can yield more useful results. We will focus on these for our future work.

Bibliography

[1] P. A. Bath, (2008). Health informatics: current issues and challenges. *Journal of information science*, *34*(4), 501-518.

[2] "What Exactly is “Health Informatics?", Healthcare-management-degree.net, 2019. [Online]. Available: https://www.healthcare-management-degree.net/faq/what-exactly-is-health-informatics/. [Accessed: 14- Oct- 2021].

[3] "Breast cancer", World Health Organization, 2019. [Online]. Available: https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/. [Accessed: 14- Oct- 2021].

[4] C. Peña-Reyes and M. Sipper, "A fuzzy-genetic approach to breast cancer diagnosis", *Artificial Intelligence in Medicine*, vol. 17, no. 2, pp. 131-155, 1999. Available: 10.1016/s0933-3657(99)00019-6.

[5] M. Nilashi, O. Ibrahim, H. Ahmadi, and L. Shahmoradi, "A knowledge-based system for breast cancer classification using fuzzy logic method", Telematics and *Informatics*, vol. 34, no. 4, pp. 133-144, 2017. Available: 10.1016/j.tele.2017.01.007.

[6] K. J. Wang and A. M. Adrian, "Breast cancer classification using hybrid synthetic minority over-sampling technique and artificial immune recognition system algorithm," *Int J Comput Sci Electron Eng (IJCSEE)*, vol. 1, pp. 408-412, 2013.

[7] J. Abonyi and F. Szeifert, "Supervised fuzzy clustering for the identification of fuzzy classifiers", *Pattern Recognition Letters*, vol. 24, no. 14, pp. 2195-2207, 2003. Available: 10.1016/s0167-8655(03)00047-3.

[8] H. J. Hamilton, N. Cercone, and N. Shan, RIAC: a rule induction algorithm based on approximate classification: Citeseer, 1996.

[9] J. Quinlan, "Improved Use of Continuous Attributes in C4.5", *Journal of Artificial Intelligence Research*, vol. 4, pp. 77-90, 1996. Available: 10.1613/jair.279.

[10] D. Nauck and R. Kruse, "Obtaining interpretable fuzzy classification rules from medical data", *Artificial Intelligence in Medicine*, vol. 16, no. 2, pp. 149-169, 1999. Available: 10.1016/s0933-3657(98)00070-0.

[11] L. Ladha and T. Deepa, "Feature selection methods and algorithms," *International journal on computer science and engineering*, vol. 3, pp. 1787-1797, 2011.

[12] U. Malik, "Applying Wrapper Methods in Python for Feature Selection", Stack Abuse, 2019. [Online]. Available: https://stackabuse.com/applying-wrapper-methods-in-python-for-feature-selection/. [Accessed: 14- Oct- 2021].

[13] S. Rawale, "Feature Selection Methods in Machine Learning.", Medium, 2019. [Online]. Available: https://medium.com/@sagar.rawale3/feature-selection-methods-in-machine-learning-eaeef12019cc. [Accessed: 14- Oct- 2021].

[14] UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set", Archive.ics.uci.edu, 2019. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Breast%20Cancer%20Wisconsin%20(Diagnostic). [Accessed: 14- Oct- 2021]

Dr. Shahidul Islam KHAN obtained his B.Sc. Engineering Degree in Computer Science and Engineering (CSE) from Ahsanullah University of Science and Technology (AUST) in 2003. He obtained his M.Sc. and Ph.D. from Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh in 2011 and 2020. His current fields of research are Data Science, Database Systems, Machine Learning, Information Security, and Health Informatics. Currently, he is serving as the Head of the IIUC Data Science Research Group. He has more than fifty published papers in peer-reviewed journals and at reputed international conferences. He is also an Associate Professor in the Dept. of CSE, International Islamic University Chittagong (IIUC), Bangladesh.