International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol. 10, Issue.4, April 2022 E-ISSN: 2347-2693

Feature Selection and Ensemble Method Analysis for Breast Cancer Datasets

Jyoti Negi[1*], K.L. Bansal[2]

[1,2]Department of Computer Science, Himachal Pradesh University, Shimla, India

[*]*Corresponding Author: negiashu9459@gmail.com*

DOI: https://doi.org/10.26438/ijcse/v10i4.1115 | Available online at: www.ijcseonline.org

Received: 17/Mar/2022, Accepted: 10/Apr/2022, Published: 30/Apr/2022

Abstract—Breast cancer has become the most common cause of death in women. Early detection of breast cancer helps out to reduce the risk factors. Three classification algorithms (NB, DT, and KNN) were used on two different Breast cancer datasets using the WEKA tool. The main purpose of this paper is to compare the results of the classification algorithms using voting and feature selection methods. The experimental result shows that voting of three classifiers gives the highest performance accuracy on the Breast cancer dataset. The ensemble method is used to increase the accuracy of the data mining algorithms. We also compare the performance accuracy of classifiers using feature selection methods (IG and PCA) on breast cancer datasets.

Keywords— J48,NaïveBayes,KNN,Voting classifier, feature selection

I. Introduction

Breast cancer is the most commonly occurring type of cancer among women. It has become a second leading cause of death in women. Early prediction of breast cancer is the best approach to increase the survival rates of patients and helps to reduce the death rates. When the growth of cells in breast tissue becomes intractable then it forms a mass of tissue called a tumor. The abnormal cells form a group together that may be a lump. A tumor can be benign (not cancerous) or malignant (cancerous). Breast cancer is recurrent and non-recurrent. Recurrence is when cancer comes back after some time and when cancer is unlikely to happen again then called non-recurrence. It may return in the same breast area or near the actual tumor. There are many risk components for breast cancer recurrence: age, tumor size, cancer type etc. Early detection of breast cancer is quite significant as cancer can circulate to other parts of the body.

Data mining extracts relevant information from the large dataset. Classification, clustering, association, and regression are some important data mining and machine learning techniques. This paper uses the WEKA tool to predict breast cancer and compares the performance of three classification techniques. We used the breast cancer dataset from the UCI repository. The actual dataset contains 10 and 32 features for BCD and WDBC. Ranker method index the features as the top priority attributes consider rank one. The Ranking method ranked the high-performance attributes by using the two different attribute evaluators, PCA and Info gain for better classification accuracy.

Also, the results of the three classification algorithms combine using the vote ensemble method. The ensemble method is used to increase the performance accuracy of the machine learning algorithms. We applied three datamining classification algorithms to the breast cancer dataset. Those datasets that contain high instances and fewer attributes can provide good results [1]. Therefore we used the concept of the feature selection method in our research.

This paper is divided into five different sections. Section 2 explains the related work on Breast cancer dataset using data mining and machine learning techniques. Section 3 is the methodology. Section 4 is the implementation and result part and the last part of the paper is the conclusion presented in Section 5.

II. Related Work

1. D. lavanya et al. [2] used a decision tree algorithm (CART) on three different breast cancer datasets with and without using feature selection. The result shows that specific feature selection using CART improves the accuracy of a particular dataset.

2. Ravi kumar et al. [3] used the WBC dataset from the UCI repository to compare six classification algorithms for detecting breast cancer. This experiment was on WEKA software. The result indicated that the SVM is the best predictor with 97.59%.

3. Chaurasia et al. [4] compared three classification algorithms, SMO, IBK, and BF tree for predicting breast cancer using the WEKA tool. The result shows that SMO gives higher prediction accuracy.

4. Kumar et al. [5] compares the performance accuracy of classification algorithms Naive Bayes, J48, and SVM for

breast cancer prediction. The author also used the voting approach which is one of the ensemble methods that combines the results of classifiers into one to give the best accuracy. They found that the combination of all three algorithms gives the highest prediction accuracy of 97.13%.

5. Bharati et al. [6] performed Breast Cancer prediction as recurrence or non-recurrence using a UCI machine learning Repository dataset of 286 instances and 10 attributes. They showed the performance of NB, IBK, RF, LR, and MLP classification techniques. They calculated some performance parameters like KP statistics, TP rate, FP rate, and Precision in their paper using the WEKA tool. The research results showed that KNN gives the highest accuracy among others.

6. Chaurasia et al. [7] performed prediction of Breast cancer using three classification techniques NB, RBF, J48. The experimental result shows that Naive Bayes classifier is the best predictor.

7. V.nanda et al. [8] implemented three machine learning algorithms including RF, MLP, and LR for breast cancer diagnosis and applied feature selection on the breast cancer dataset.

Table 1. Summary of related work

Reference	Year	Technique	Accuracy

[2]	2011	CART and feature selection	SVM Attribute Eval = 73.07% (BCD)
			PCA = 96.99%(WBCO)
			SymUncert= 94.72%(WDBC)

[3]	2013	J48, SVM, MLP, NB, LR, KNN	SVM = 97.59%
[4]	2014	DT, SMO, KNN	SMO = 96.2%
[5]	2017	DT,NB,SVM, voting classifier	SVM = 97.13 %
[6]	2018	NB,RF,LR,ML, KNN	KNN = 97.90%
[7]	2018	NB, RBF, J48	NB = 97.36%
[8]	2021	MLP, RF, LR	MLP = 98%

Table 1. Shows the summary of the literature review. Various interests are found throughout this literature review. Many researchers have concentrated on the performance of data mining algorithms and the impact of the datasets on the overall performance of algorithms. Therefore, this research focused on data mining algorithms for different datasets by applying the feature selection and ensemble method.

III. METHODOLOGY

The breast cancer dataset is collected from the UCI machine learning repository. Preprocessing is an important stage in data mining. Before applying data mining techniques, the feature of the dataset is checked from the pre-process tab on the WEKA to remove all the missing values. This paper uses data mining classification algorithms and ensemble method on the Breast cancer dataset by using WEKA tool. Vote combines the output of multiple classifiers into one and generate a single result. Fig. 1 shows the process design of the voting method. This paper also analyzes the performance of data mining algorithms using feature selection in terms of accuracy and time to build a model on two different breast cancer dataset. PCA attribute evaluator with Info gain for feature selection that selects the attributes by using the ranking method.

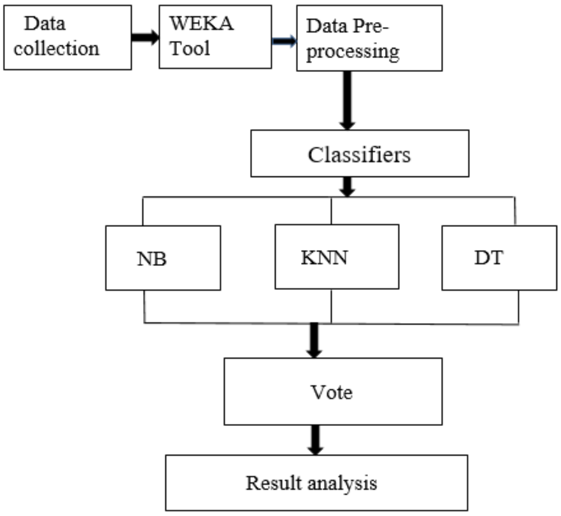

Fig.1 Process design

Decision Tree

A decision tree is similar to a tree-like structure and it has a root node, leaf node, and internal node. The root node is on the top of the tree also it is the first node in the tree and it can further split into other internal nodes. The internal nodes represent a test on various attributes, each branch represents the result of the test and the leaf node represents the end value of the target or the class label. The action of splitting the dataset into subsets continues until it reached the end values of the target. Numerical and categorical features can be classified by the decision tree and also it is a supervised learning technique. Decision tree J48 is an extension of ID3 algorithm and the C4.5 algorithm is treated as a Decision tree Classifier.

Naïve Bayes

Naïve Bayes is a probabilistic classification algorithm that uses Bayes theorem for finding the probability of each attribute of a class and solves classification problems. All attributes values are independent of each other i.e. the effect of an attribute value will be independent of another value of the attribute on the given class. This assumption is called as conditional independence.

Bayes theorem formula

$$P(A/B) = \frac{P(B/A) * P(A)}{P(B)}$$

K- Nearest Neighbor

KNN classifier is a simple and most common classification technique in Machine learning. It classifies the input data based on the connection of its neighbor. It classifies the input data into the category of the most similar nearest neighbor. In KNN, the K defines the number of close neighbors to use. KNN algorithm is named as IBK in WEKA. It is a supervised learning approach that uses learning techniques based on likeness. Generally, the Euclidean distance is used as distance measure in the KNN classification.

Vote

Voting is one of the ensemble techniques where it combines the prediction results of multiple classifiers into one that will help to improve the performance accuracy. The basic idea behind this technique is to attain a high level of accuracy. Voting is a good ensemble technique when one classifier algorithm drawback can be helpful for another classifier [5]

WEKA TOOLKIT

Weka (Waikato Environment for Knowledge Analysis) is an open-source tool that was developed at the University of Waikato, New Zealand. It provides data preprocessing that removes the inconsistent data from the dataset. Weka consists of numerous Data Mining algorithms such as classification, clustering, and association. These algorithms can be used in a dataset for developing the model. Another feature that Weka provides is Attribute Selector which helps to reduce the dataset. Weka contains a visualization component to examine the data. WEKA support some file extension such as ARFF, CSV.

Performance evaluation

The most commonly used performance measures for machine learning models are accuracy, precision, recall, ROC curve, Kappa statistic, and confusion matrix. The confusion matrix is a table like structure having two parts actual and prediction. It helps to visualize the performance of the model.

Accuracy = (TP+TN)/ (TP+ TN + FP+FN)
Sensitivity = TP / (TP + FN)
Specificity = TN / (TN + FP)

IV. EXPERIMENTAL SETUP

This section analyzes the implementation of three data mining algorithms (NB, J48, and KNN) on the two different breast cancer dataset using ensemble and feature selection methods. The description of dataset is shown in Table2. The breast cancer dataset is collected from the UCI machine learning repository.

(i) Dataset description

- Breast Cancer Dataset
- Wisconsin Diagnostic Breast Cancer (WDBC)

Table 2. Summary of selected dataset

Data set	Instances	Attributes	Class
Breast Cancer Dataset (BCD)	286	10	Recurrence and non-recurrence
Wisconsin Breast Cancer Dataset (Diagnostic)	569	32	Benign and malignant

Breast Cancer Dataset

The Breast Cancer Dataset contains 286 instances and 10 attributes. The Principal component analysis and Info gain attribute evaluator selects 7 attributes by using the ranker method. One attribute consists of a class attribute and it has two values such as "recurrence" and "no recurrence". Attributes of breast cancer dataset is shown in Table 3.

Table 3. Attributes of BCD

S.no	Attribute
1.	Class
2.	Age
3.	*Menopause*
4.	*Tumor size*
5.	*Inv-nodes*
6.	*Node caps*
7.	*Degree of malignancy*
8.	*Breast*
9.	*Breast quadrant*
10.	*Irradiant*

Wisconsin Diagnostic Breast Cancer (WDBC)

Wisconsin Breast cancer dataset from UCI machine learning contains 569 instances and 32 attributes shown in the Table 4. The dataset contains two class values "benign" and "malignant". The principal component analysis and Info gain attribute evaluator for feature selection selects only 11 attributes by using the rankers method (search technique) in the WEKA tool. All the irrelevant attributes in the dataset are excluded by using an attribute evaluator. The problem of over-fitting is also solved by PCA.

Table 4. Attributes of WDBC

S.no	Attribute
1.	ID number
2.	Diagnosis (M = malignant, B = benign)
3.	Ten real-valued features are computed for each cell nucleus
	a) radius
	b) texture
	c) perimeter
	d) area
	e) smoothness
	f) compactness
	g) concavity
	h) concave points
	i) symmetry
	j) fractal dimension

(ii) Experiment Result

In Table 5. We have calculated the performance of the individual classifier's accuracy on the BCD and WDBC dataset with and without applying the feature selection technique. Dataset is split into training and testing set for every single data mining classification algorithms using ten-fold cross validation.

Table 5. Individual classifiers accuracy

Classifiers	**Before feature selection**		**After feature selection**	
	BCD	WDBC	BCD	WDBC
NB	71.67%	92.61%	72.72%	93.67%
KNN	72.37%	96.13%	71.67%	94.2%
J48	75.52%	93.14%	75.5%	94.72%

We have compared two feature selection methods Info Gain attribute evaluator and Principal components using the Ranker search method on two different breast cancer dataset shown in Table 6. In some cases using the feature selection method, we are receiving a little more accurateness for every individual classifier which is shown in Table 5.

Naïve Bayes classifier accuracy on BCD and WDBC, with feature selection is 71.67%, 92.61% which turns to 72.72%, 93.67% after applying Info gain and Principal components. By applying J48 to the Breast cancer dataset there is no change in the accuracy of 75% on BCD with and without using feature selection but it turns to 94.72% after applying principal components on WDBC

Table 6. Performance of classifiers with feature selection

Classifiers	**Feature selection**	**Search Technique**	**BCD (Accuracy)**	**WDBC (Accuracy)**
NB	Information Gain	Ranker	72.37%	**93.67%**
	Principal component		**72.72%**	91%
KNN	Information Gain	Ranker	**71.67%**	93.49%
	Principal component		59.09%	**94.2%**
J48	Information Gain	Ranker	**75.5%**	92.4%
	Principal component		72.02%	**94.72%**

KNN classifier accuracy on BCD and WDBC is 71.67%, 94.2%, which gives a declining result after applying features election methods. After using Information gain and Principal components, we do not get the outcome with a vast difference. When we compare individual classifiers with each other using Vote, it gives more accurate results. A combination of three algorithms using the voting method helps to improve the accuracy of a particular algorithm shown in Table 7. NB+KNN+J48 (without feature selection) gives the best accuracy with 96.30%.

Table 7.Voting of classifiers

Vote	**BCD**	**WDBC**
NB+KNN+J48	73.77%	96.30%
NB+J48	73.7%	92.9%
NB+KNN	73.4%	94.72%

IV. CONCLUSION AND FUTURE SCOPE

We applied three data mining algorithms to two different breast cancer datasets with different attributes. Intend to know which features and individual algorithms are better for breast cancer datasets.

We used classification algorithms and applied feature selection to breast cancer datasets. Using a ten- fold cross validation dataset is split into training and testing set for every single data mining classification algorithms. We mainly focus on which classification algorithm has good accuracy. As per the result, no single classifier perform well. So we combined classifiers for the highest classification accuracy using an ensemble approach. Combining classifiers using vote gives the best result. The fusion of NB, J48, and KNN gives the best result of 96.30% on WDBC and 73.77% on BCD. We used two different feature selection algorithms (Info gain and PCA) on breast cancer datasets. Info gain and PCA both perform well on Breast cancer datasets and gives different results.

In future work, we can consider more data mining algorithms, ensemble methods, and more feature selectors on the breast cancer dataset for attaining high accuracy.

REFERENCES

[1] Ahmed Iqbal Pritom, Shahed Anzarus Sabab, Md. Ahadur Rahman Munshi, Shihabuzzaman Shihab,"Predicting breast cancer recurrence using effective classification and feature selection technique", 19th International conference on computer and information technology, December **18-20, 2016.**

[2] D. Lavanya,K. U. Rani, "Analysis of feature selection with classification: Breast Cancer Datasets", Indian Journal of computer science and engineering (IJCSE), **Vol. 2 NO. 5 Oct-NOV 2011**, ISSN: 0976-5166.

[3] G. Ravi Kumar, Dr. G. A. Ramachandra,K.Nagamani, "An Efficient Prediction of Breast Cancer Data using Data Mining Techniques", International Journal of Innovations in Engineering and Technology (IJIET), **Vol. 2 Issue 4 August 2013**, ISSN: 2319-1058.

[4] Vikas Chaurasia, Saurabh Pal, "A Novel Approach for Breast Cancer Detection using Data Mining Techniques.", International Journal of Innovative Research in Computer and Communication Engineering ,**Vol. 2, Issue 1, January 2014.**

[5] U. Karthik Kumar, M.B. Sai Nikhil and K. Sumangali, "Prediction of Breast Cancer using Voting Classifier Technique", IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), **2 - 4 pp.108-114, August 2017.**

[6] Subrato Bharati, Mohammad Atikur Rahman, Prajoy Podder, "Breast Cancer Prediction Applying Different Classification Algorithm with Comparative Analysis using WEKA.", 4th International conference on electrical engineering and information and communication technology, **2018.**

[7] Vikas Chaurasia, Saurabh Pal and BB Tiwari, "Prediction of benign and malignant breast cancer using data mining

techniques.", Journal of Algorithms & Computational Technology, **Vol. 12(2) 119–126, 2018,** doi:DOI: 10.1177/1748301818756225.

[8] V.Nanda Gopal, Fadi Al-Turjman, R. Kumar, L.Anand, M.Rajesh, "Feature selection and classification in breast cancer prediction usingIOTandmachine learning.", Elsevier, **18 April 2021**, doi:https://doi.org/10.1016/j.measurement.2021.109442.

[9] Ghanchi, Nileshkumar Modi and Kaushar, "A Comparative Analysis of Feature Selection Methods and Associated Machine Learning Algorithms on Wisconsin Breast Cancer Dataset", Proceedings of International Conference on ICT for Sustainable Development, Advances in Intelligent Systems and Computing 408, **2016**, doi:10.1007/978-981-10-0129-1_23.

[10] J. Han, M.K.,"Data Mining Concepts and Techniques", A volume of The Morgan Kaufmann Series in Data Management system, **2012**.

AUTHORS PROFILE

Ms. Jyoti Negi pursued a Diploma in Computer Engineering from Himachal Pradesh Takiniki Shiksha Board of Dharamshala, India in 2016 and Degree in Bachelor of Technology in Computer Science and Engineering from Himachal Pradesh University Shimla, India in 2019. She is currently pursuing in Master of Technology in Computer Science from Himachal Pradesh University Shimla, India. Her main research work focuses on Datamining, classification techniques, Weka Tool, feature selection and ensemble methods.

Dr. K.L. Bansal is currently working as a professor in the Department of Computer Science, Himachal Pradesh University Shimla, India, with a specialization in Data Mining and Data Warehousing, Artificial Intelligence, Computer Networks.