Mansoura Journal of Computers and Information Sciences

A Wrapper Feature Selection Technique for Improving Diagnosis of Breast Cancer

Amal F. Goweda
Faculty of computers and information systems , I.S dep.
Mansoura University, Egypt
amal_goweda@yahoo.com

Mohammed Elmogy
Faculty of computers and information systems , I.T dep.
Mansoura University, Egypt
melmogy@mans.edu.eg

Sherif Barakat
Faculty of computers and information systems , I.S dep.
Mansoura University, Egypt
sherifiib@yahoo.com

ABSTRACT

Nowadays, cancer is considered as a fairly common disease. Regarding the number of newly detected cases, breast cancer is ranked as one of the most leading cancer types to death in women. It can be cured, if it is identified and treated in its early stages. Therefore, this study explores a proposed integrated wrapper feature selection method called wrapper naïve-greedy search (WNGS) to improve the accuracy of the breast cancer diagnosis. WNGS is based on a wrapper method, which is blended with a greedy forward search to select optimal feature subset. WNGS method integrates a wrapper method based on Naïve Bayes (NB) classifier as a learning scheme with a forward greedy search method. Then, the selected feature subset is fed to a classifier to determine breast cancer. In addition, K-nearest neighbor-greedy search (KNN-GS) is used for comparison. In KNN-GS method, k-nearest neighbor (KNN) classifier is used as a learning scheme while a forward greedy search method is used to search through features. NB is used as the classifier for classification process for both methods. By applying these two methods, data features are reduced, and the classification rate is improved. Both methods are tested on two different benchmark breast cancer datasets. Accuracy results showed that WNGS method outperformed KNN-GS method. Also, WNGS method overcame KNN-GS regarding precision, recall, F-measure, and sensitivity.

Keywords: Cancer Classification; Feature Selection; Naïve Bayes (NB); Forward Greedy Search.

1. INTRODUCTION

Classification is one of the most critical tasks in real-world problems, especially in medical diagnosis models. It is defined as the process of allocating a class to an object. Before providing a classifier with a dataset, several considerations are required. It is better to consider only relevant features and eliminate irrelevant ones. Irrelevant features cause a workload on the classifier. For the classification process improvement, feature selection methods should be applied. There is a need to identify a feature selection method that maximizes classification accuracy and minimize data features. Feature selection goal is to determine only relevant features and excluding nonfunctional features from data domain. As known, feature selection is the improvement key of classifier performance. So, it can be considered as a prior step that must be passed before solving a classification task. There are two main feature selection groups. One is independent of the induction algorithm and known as filter methods. The other is dependent on the induction algorithm and known as wrapper methods. Embedded methods is another feature selection group, which efficiently use wrappers idea. Induction algorithm participates in feature selection step as wrapper methodologies. Embedded methods use available data without the need of splitting data into training and testing sets. It reaches a solution faster as no retraining predictor assesses for every variable subset.

Feature selection based on a wrapper method is an attractive feature selection approach. Features are selected depending on a decision from a particular learning scheme. Feature selection cycle consists of a search method embedded within a learning scheme. Search method penetrates feature domain to search for candidate feature subsets. There are many search strategies as best-first, branch-and-bound, genetic algorithms (GA), simulated annealing, and greedy strategies. Learning scheme evaluates candidate feature subsets and selects the best one. To employ wrapper methods in a right way, three components must be applied. First, search strategy is used to search for candidate feature subsets through feature domain. Second, the evaluation function is used to assess candidate feature subsets. The last is the performance function that is used to validate the best-selected subset. To improve the efficiency of the wrapper methods, a feature selection method called wrapper naïve - greedy search (WNGS) is introduced. Naïve Bayes (NB) classifier integrated with a forward greedy search method is applied as induction algorithm with a search method for feature selection step. Learning scheme (NB classifier) is supported by the help of a forward greedy search method to be back with the effective feature subset while ignoring the rest. Then, the selected feature subset is fed to NB classifier for cancer classification task.

The rest of the paper is structured as follows. Related work is inspected in Section 2. Section 3 details out the fundamental principles of wrapper approaches, forward greedy strategy, and NB classifiers. Section 4 shows the proposed integrated method for feature selection and sets overall framework perception for the classification task. Section 5 demonstrates datasets description in addition to discussing experimental results. Finally, the conclusion of this study is given in Section 6.

2. RELATED WORK

The following discusses previous works related to hybrid feature selection methods. The objective of all researchers attempts was to eliminate features that have no significance and improve classification results. Selecting the most distinctive features leads to performance improvements of the prediction model. The motivation is to find a method, which uses the fewest possible features, improves classification activity, and reduces execution time. According to that, the diagnosis process becomes more accurate and seems to be ideal.

Chen et al. [1] realized the importance of using mutual information in feature gene selection. However, mutual information cannot deal with continuous features directly. To solve the problem, two direction feature selection methods were proposed. First, a reliefF algorithm [2] was exploited to eliminate genes space and obtain candidate subset of genes. Then, a neighborhood mutual information manner combined with a forward greedy search strategy was proposed. The manner was capable of dealing with continuous features and selecting feature genes from genes subset. The returns on six different microarray cancer datasets illustrated that the proposed manner achieved higher accuracy rate using few genes.

Karthikeyan and Thangaraju [3] submitted a diagnostic refinement framework for solving hepatitis diagnosis problem. Two proposed methods using Correlation Feature Selection (CFS) as feature evaluator with two different search strategies were introduced. One dubbed BFSCFS-NB, and it used CFS with a best first method for feature selection process. The other method dubbed GSCFS-NB and it used CFS with greedy search method for feature selection process. BFSCFS-NB method used best first engine search as it allowed backtracking through search path and made local changes to the current feature subset. GSCFS-NB method used greedy search, which started with an initial state and selected only the best local change from all possible local changes. NB was used as a classifier and applied on hepatitis disease dataset. The proposed method proved effective in the accuracy rate term and in running time reduction term. Finally, the proposed method performance overcame methods, such as Multilayer Perceptron (MLP) and radial basis function (RBF).

Soufan et al. [4] recognized that efficient classification models require no attention to irrelevant features. A web tool dubbed DWFS was developed for efficient feature selection step. Wrapper method embedded with genetic search method was the base of DWFS tool. GA parameters can be adjusted according to the problem addressed. The proposed tool was applied on different biomedical datasets. Experiments demonstrated that proposed tool was fast and leads to features space minimization without performance giving up.

Saripan et al. [5] proposed in their study an integrated framework for gene selection. The proposed genes selection approach included multiple phases. Firstly, gene selection initiated with genes ranking. An independent evaluation criterion was used to rank genes and outputted a ranked genes matrix. Then the ranked matrix separated into same size partitions. The process of dividing the ranked matrix was to simplify wrapper feature selection method. After that, a sequential forward feature selection method was applied to each part. Finally, the returned genes subsets were combined and purified to produce the final subset of genes. The integrated framework was applied on two different datasets and validated using two different classifiers.

Aruna and Rajagopalan [6] developed a constrained search sequential floating forward search (CSSFFS) based on support vector machine (SVM) for breast cancer detection. It was a greedy mechanism based on constrained search strategy to select minimal feature subset with balanced error rate (BER) minimization. SVM acted as a feature ranking measure for discarding irrelevant features. A sequential floating forward search (SFFS) acted as wrapper method to extract the optimal subset of features. One of its advantages is that it is a hybrid algorithm between filters and wrappers. Attributes were ranked with the square value of weights estimated by SVM. Attribute ranking acted as filters to eliminate irrelevant features. With remaining features, SFFS with SVM was used to select the optimum feature subset, and this represented wrapper stage to remove irrelevant features if any yields. The objective was to select optimum features with minimal BER. The experiments are conducted in WEKA [7]. The WDBC dataset with 32 features is used for the experiment. CSSFFS algorithm is applied on breast cancer domain.

Liu et al. [8] proposed a statistical measure named LW-index to evaluate the feature subsets. Then, a new feature selection method was presented. The new method was the combination of LW-index with sequence forward search algorithm (SFS-LW). The experiments were conducted on nine UCI datasets [9]. The experimental results indicated good classification accuracy and reduced the computation cost compared to other wrapper methods. As an achievement, the proposed method can replace the expensive cross-validation scheme as an evaluation measure.

Waad Bouaguel [10] demonstrated a new wrapper method for feature selection in big data. The proposed method was based on a random search using GA and prior information. The new method was tested on two biological datasets and also compared to two well-known wrapper feature selection approaches. The outcome results showed that the new approach extracted the best performances.

We now claim that previous works carried a lot of useful advantages in classification accuracies and featured elimination. Previous studies also led to the gradual development of wrapper methods use for feature selection. To enhance the using of wrapper methods in feature selection, we try to demonstrate a method that can help in making an accurate diagnostic decision. A mixed feature selection method for breast cancer classification problem is presented. The method mix a forward greedy search with a wrapper method that uses NB classifier as a learning scheme. The use of search method help in NB decision to evaluate feature subset. The forward greedy search candidates subsets for NB classifier. While NB evaluates candidate feature subsets and decides to select the best feature subset depending on subsets best accuracies.

From this point of view, WNGS objective is to integrate a forward greedy search method within NB estimator. WNGS is exploited to select informative feature subset from data domain. Then, the selective feature subset is tested and fed to a classifier to help in solving classification problems. WNGS is also compared with KNN-GS mechanism to measure performance quality. KNN-GS is also a mixed feature selection method for breast cancer classification problem. The method mixes a greedy forward search with a wrapper method that uses KNN classifier as a learning scheme instead of NB classifier.

3. BASIC CONCEPTS

3.1 Wrapper Approach

Feature selection process passed through four stages which described in [11]. First, candidate feature subsets are generated using a search strategy. The search strategy is like best first, forward selection, backward selection, and GA. Second, the subset generated is then evaluated by using a filter, wrapper or embedded mechanisms. Filter, wrapper, and embedded mechanisms are the three classes of feature selection methodology. Filter and wrapper mechanisms vary in their relation with the induction classifier. The filter approaches are isolated from the learning scheme. Unlike filter ones, the induction learning scheme participated in estimating the merit of feature subsets in wrapper methods. Third, the process still works until a stopping criterion reached. Finally, a validation procedure is to check the validation of the feature subset being selected.

In wrapper methods, learning classifier is used as a black box in the process of feature selection as shown in Fig. 1. To extract ideal feature subset, the learning classifier is used as a portion of feature selection procedure. It is used as an evaluation measure for evaluating candidate feature subsets extracted by search methods. The accuracy rate of the learning scheme is evaluated using estimation measurements [12]. Wrapper methods are based on hypothesis. In wrapper method, a weight vector is associated with feature subsets. The features weights are assessed by its performance degree in classification learning. The learning scheme iteratively adjusts feature weights according to its performance.

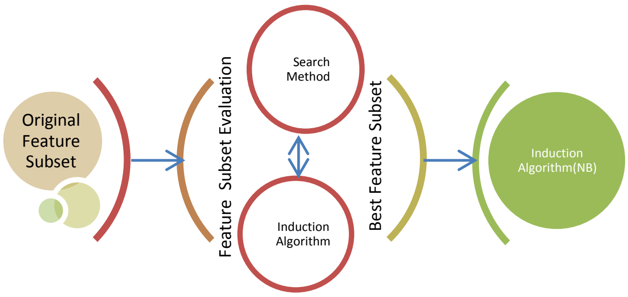


Fig 1. The wrapper steps for feature selection.

3.2 Forward Greedy Approach

Greedy algorithms are simple and straightforward. It is an easy and quick way to be implemented. Most of the greedy algorithms are used to solve optimization problems. A greedy search is to decide on local optimum at each stage while finding a global optimum. Greedy stepwise strategy perspective covers two deterministic directions. One is forward, and the other is backward elimination through features space. The greedy approach starts with no attributes in case of forwarding direction. Unlike backward elimination, it starts with all attributes. It must stop when the addition or deletion of any remaining attributes decreases performance evaluation. Fig. 2 lists greedy forward steps.

```
(0) $= \emptyset$ ; $F^{(0)} = \{f_1, f_2, f_3, \dots, f_n\}$;
=0; iter=0;
$< n$)  K=size ($F^{(i)}$) max = 0;feature = 0;
 For j from 1 to k
    score = eval( $F_j^{(i)}$);
    If ($score > max$)
        max = score;
       feature = $F_j^{(i)}$ ;
              Endif;
Endfor;
    If ($max > opt$)
    opt = max;    iter =i;
    Endif
                 $FS^{(i+1)} = FS^{(i)} + feature$;

                  $F^{(i+1)} = F^{(i)} - feature$;

; Endwhile;
```

Fig. 2. The forward greedy search algorithm [13].

The forward greedy search method initially starts with an empty set. The feature subset is to be filled with features gradually. Each unused feature is added to the feature subsets one at a time to train the model. Learning scheme is used to evaluate feature weight. If feature score exceeds maximin value, then set the feature score to be the maximum value. The feature from the unused group is added to the set when it provides the best performance.

3.3 NB Algorithm

NB classifier is a supervised learning method with strong independence assumptions for classification. NB algorithm is one of the simplest algorithms to be implemented as it does not need any complicated parameter settings. It is considered to be one of the most applied classifiers due to its simplicity [14]. Also, it performs well in diagnostic problems. It is a quicker implementation algorithm and has no problem to deal with any size of data dimension space. As a summary, it is very easy to be constructed and to be implemented.

The main advantage of NB classifier is its swiftness of use. Because of its swiftness, it can handle many attributes of a data set. To develop accurate parameter estimations, NB classifier needs an only small set of training data because it requires only attributes frequencies calculation and attribute outcome pairs in the training dataset [21]. NB learning classifier uses Bayes theorem concept to calculate the most likely class label of the new instance. The basic assumption of

NB classifier is that different attributes are independent of each other concerning the class [16]. Assuming independence of features can be a major defect of using the NB classifier. As real-world data may contain relations among attributes. To overcome this limitation, attributes are selected with the help of search strategies.

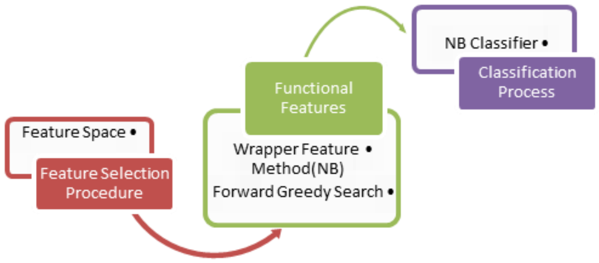


Fig. 3. The flow diagram of WNGS method framework

4. PROPOSED METHOD

In this study, a feature selection method that tries to reduce features space is explored. The integrated manner is presented in two-phase schemes. Firstly, feature selection scheme is to pick out the most suitable features from overall feature space. The second scheme concentrates on breast cancer classification task. The first phase combines NB classifier with a forward greedy search method to take off the best feature subset from overall features space. At each iteration, NB classifier role in feature selection step is to evaluate the accuracy of generated feature subsets. Candidate feature subsets are generated by the support of a forward greedy search method. Forward greedy search method is one of simplest greedy search algorithms. Greedy search method starts with just an initial feature and gradually adding in all other features. Each time a new feature is added in, the feature set is evaluated by NB classifier. The new feature added is only kept if there is a noticeable change in accuracy. The addition process still works, and a new round is initialized with the modified feature subset. The best feature subset that achieves the best accuracy and the best performance is selected. Then NB classifier is applied to perform classification procedure. The overall framework for cancer classification problem is shown in Fig. 3. The major objective is to disregard irrelevant features and hold only relevant ones to facilitate NB classifier mission. The target of the combination of NB classifier with a greedy forward strategy is to get back with optimal features from overall features. The returned features are then applied to a classification task.

In this study, another feature selection method called K-nearest neighbor – greedy search (KNN-GS) is introduced. KNN-GS method is introduced to be compared with WNGS method. KNN-GS is going on WNGS route as it uses K-nearest neighbor (KNN) as a learning scheme instead of NB for feature selection. KNN-GS also tries to reduce feature space and improve classification accuracy. The integrated manner is shown in two-phase schemes. Firstly, feature selection scheme is to pick out the most suitable features from overall feature space.The second scheme concentrates on breast cancer classification task. The first phase combines KNN classifier with a forward greedy search method to take off the best feature subset from overall features space. KNN classifier role in feature selection step is to evaluate the generated feature subset. Candidate feature subsets are generated by the support of a forward greedy search method.

Greedy search method starts with just an initial feature and gradually adding in all other features. Each time a new feature is added in, the feature set is evaluated by KNN classifier. The new feature added is only kept if there is a noticeable change in accuracy. The addition process still works, and a new round is initialized with the modified feature subset. The best feature subset that achieves the best accuracy and the best performance is selected. Then NB classifier is applied to perform classification procedure. The overall KNN-GS framework for cancer classification problem is shown in Fig. 4.

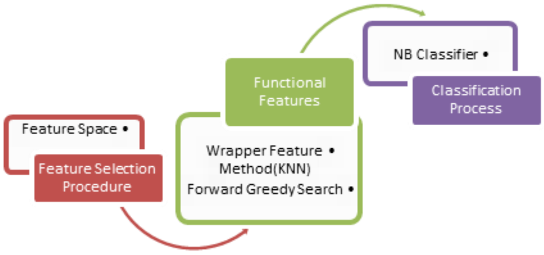


Fig. 4. The flow diagram of KNN-GS method framework.

5. EXPERIMENTAL RESULTS

5.1 Datasets Description

Table 1 described datasets names, number of original features, data instances and data classes. The following paragraph shows the characteristics of each used dataset. Wisconsin breast cancer dataset is obtained from the University of Wisconsin hospitals [9] [17]. Benign case or malignant case is the two possible probabilities for each instance. A number of case instances is 699 and number of attributes are ten plus class attribute. This breast cancer dataset includes 16 missing values. Benign cases represent 65.5% of all cases where malignant cases represent 34.5% of overall cases. Wisconsin diagnostic breast cancer (WDBC) dataset is obtained from UCI repository [9] [18]. It contains 568 breast samples, 32 attributes, and no missing attribute values. The malignant class has 212 instances, and benign one holds 357 (62.75%). The class attribute can take M value or B value. M value represents a malignant case where B value represents a benign case. Descriptive attributes are attributes from 3 to 32. The first attribute is removed as it represents ID number for a patient case in the two datasets. For WDBC, the second attribute represents class attribute. A summary of breast cancer datasets is presented in table 1.

Table 1: The characteristics of UCI datasets

Datasets	No. of instances	No. of attributes	No. of classes
Wisconsin breast cancer	699	10	2
WDBC	568	31	

5.2 Experimental Setup

WNGS and KNN-GS methods were implemented in WEKA [7] toolkit version 3.7.12 on an Intel Core i3 processor and 4 GB RAM machine. WEKA [19] is a unified tool for data pre-processing, classification, regression, clustering, association rules and visualization.

5.2.1 Methodology of Proposed System

The proposed approaches are incorporated in two stages. Firstly, WNGS method is used on two different datasets. Secondly, KNN-GS method is used on the same datasets and is used for comparison with WNGS. For Wisconsin breast cancer dataset, all the number of features was reduced to 5 from 9 by WNGS method which based on NB and forwards greedy search. Then, Wisconsin breast cancer dataset is classified using NB classification algorithm. The block diagram of the proposed method is shown in Fig. 3. For WDBC dataset, the number of features was reduced to 3 from 30 by WNGS method which based on NB and forwards greedy search. Then, WDBC dataset is classified using NB classification algorithm.

For Wisconsin breast cancer dataset, the number of features was reduced to 7 from 9 by KNN-GS method which based on KNN and forwards greedy search. Then, Wisconsin breast cancer dataset is classified using NB classification algorithm. The block diagram of the proposed method is shown in Fig. 4. For WDBC dataset, the number of features was reduced to 4 from 30 by KNN-GS method based on KNN and forward greedy search. Then, WDBC dataset is classified using NB classification algorithm.

Table 2: Best accuracies of WNGS and KNN-GS for breast cancer datasets.

Datasets	WNGS	KNN-GS
Wisconsin breast cancer dataset	0.97568	0.9728
WDBC	0.97007	0.96302

For WDBC dataset, WNGS method outperformed KNN-GS method regarding the minifying number of original attributes. The two methods achieved robust features returns. WNGS method returned three attributes, and KNN-GS returned four attributes from 30 attributes to be passed to the classifier.
For Wisconsin breast cancer dataset, five functional attributes were returned by WNGS method where seven informative attributes were returned by KNN-GS method. For WDBC dataset, three functional attributes were returned by WNGS method where four attributes were returned by KNN-GS method. The informative attributes from the two methods are then passed to the classifier for the classification task.

5.2.2 Performance Evaluation

Accuracy is the measure for evaluating methods performance. Accuracy is defined as correct classified cases divided by the total number of cases. Best accuracies of WNGS and KNN-GS methods for breast cancer datasets are shown in Table 2 and Table 3. Accuracy is measured according to this equation [21]:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

For Wisconsin breast cancer dataset, the best accuracy found by WNGS method for a feature subset is 97.56% as shown in Table 2. KNN-GS achieved 97.28% as the best accuracy for a subset. For WDBC dataset, the best accuracy found by WNGS method for a feature subset is 97%. KNN-GS achieved 96.3% as the best accuracy for a feature subset. Note that 10 fold cross-validation had been used to validate each method. Cross-validation [21] as known divides the data into k subgroups and each one is tested via classification rule constructed from the remaining (k -1) groups. The test accuracy is evaluated according to an average of the algorithm.

Table 3: The correct and incorrect classification cases.

Datasets	WNGS	Percentage	KNN-GS	Percentage
Correct Classified Cases	682	97.568%	680	97.281%
Incorrect Classified Cases	17	2.432 %	19	2.718
Wisconsin breast cancer dataset				
Datasets	**WNGS**	**Percentage**	**KNN-GS**	**Percentage**
Correct Classified Cases	551	97.007%	547	96.302
Incorrect Classified Cases	17	2.993 %	21	3.697
WDBC dataset				

From Table 3, number of misclassified instances for WNGS method was less than KNN-GS method on Wisconsin breast cancer dataset. WNGS returned 17 misclassified samples and 682 correctly classified ones. KNN-GS method is back with 19 misclassified samples and 680 correctly classified ones. For WDBC dataset, number of misclassified instances for KNN-GS was greater than WNGS method. WNGS returned 17 misclassified samples and 551 correctly classified ones. KNN-GS method is back with 21 misclassified samples and 547 correctly classified ones.

On Wisconsin breast cancer dataset, WNGS achieved 0.967 TP rate for class 2 and 0.991 for class 4. KNN-GS achieved 0.967 TP rate for class 2 and 0.983 for class 4. On WDBC dataset, WNGS achieved 0.938 TP rate for class M, and 0.989 for class B. KNN-GS achieved 0.938 TP rate for class M and 0.978 for class B. The two values were needed to calculate the precision value. Precision and recall are the basic measures used in evaluating search strategies. Error rate on the two datasets are shown in Table 4. Precision, recall, specificity and error rate were calculated according to the following equations [21]:

$$\text{Precision}=TP/(TP+FP) \qquad (2)$$

$$\text{Recall}=TP/(TP+FN) \qquad (3)$$

$$\text{Specificity} = TN/(TN+FP) \qquad (4)$$

$$\text{Error rate} = (FP+FN)/(TP+FN+FP+TN) \qquad (5)$$

Table 4: The error rate of WNGS and KNN-GS on breast cancer datasets

Error Rate	WNGS	KNN-GS
Wisconsin breast cancer dataset	0. 024	0.027
WDBC	0.0299	0.0369

Precision is the mathematical measure of relevant samples retrieved. Specificity (TN) rate is a test measure of how accurate a test is against FP. Recall (TP) rate fraction is the measure of the test to identify correctly those who have the disease. Precision can be seen as a measure of exactness or quality, whereas recall is a measure of completeness or quantity. High recall means that used method returned most of the relevant results. High precision means that used method returned more informative results than irrelevant. F1 score or F-measure conveys the balance between the precision and the recall. It can be calculated according to the following equation [21]:

$$\text{F-measure}= 2 * \frac{Precision*Recall}{Precision+Recall} \qquad (6)$$

Table 5: The statistical measures on Wisconsin breast cancer dataset.

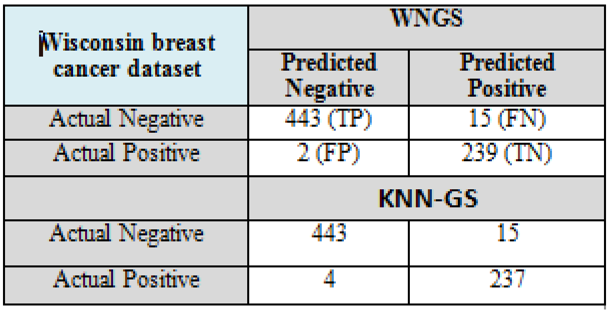
Wisconsin breast cancer dataset			
	WNGS	**KNN-GS**	**Class**
TP Rate	0.967	0.967	2
	0.991	0.983	4
FP Rate	0.008	0.017	2
	0.033	0.033	4
Precision	0.996	0.991	2
	0.941	0.940	4
Recall (Sensitivity)	0.967	0.967	2
	0.992	0.983	4
F-measure	0.981	0.979	2
	0.966	0.961	4

Table 6: The statistical measures on WDBC dataset.

WDBC dataset			
	WNGS	**KNN-GS**	**Class**
TP Rate	0.938	0.938	M
	0.989	0.978	B
FP Rate	0.011	0.022	M
	0.068	0.062	B
Precision	0.980	0.961	M
	0.964	0.964	B
Recall	0.938	0.938	M
	0.989	0.978	B
F-measure	0.959	0.950	M
	0.976	0.971	B

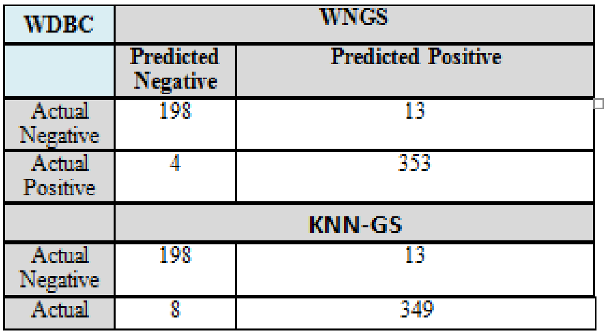
Table 7 and Table 8 showed WNGS and KNN-GS confusion matrices for Wisconsin breast cancer and WDBC datasets.

Table 7: The confusion matrices of WNGS and KNN-GS on Wisconsin breast cancer dataset**.**



Wisconsin breast cancer dataset	WNGS	
	Predicted Negative	Predicted Positive
Actual Negative	443 (TP)	15 (FN)
Actual Positive	2 (FP)	239 (TN)
	KNN-GS	
Actual Negative	443	15
Actual Positive	4	237

Table 8: The confusion matrices of WNGS and KNN-GS on Wisconsin breast cancer dataset.



WDBC	WNGS	
	Predicted Negative	Predicted Positive
Actual Negative	198	13
Actual Positive	4	353
	KNN-GS	
Actual Negative	198	13
Actual	8	349

From confusion matrices, we evaluated lift measure as shown in Table 9. Lift measure is the ratio of confidence to expected confidence. Lift measures the degree to which classification model predictions are better than randomly-generated predictions. The Lift is applied to binary classification only, and it requires the designation of a positive class. The lift can be defined as a ratio of two percentages. It is the percentage of model correct positive classifications to the percentage of actual positive classifications in the test data. Based on the confusion matrices above, we compute lift as follows:

$$\text{Lift measure} = (TP/(TP+FN))/((TP+ FP)/(TP+FN+FP+TN)) \qquad (7)$$

Table 9: The WNGS and KNN-GS lift measures.

Lift measure	Wisconsin breast cancer dataset	WDBC
WNGS	1.519	2.6386
KNN-GS	1.512	2.5874

In fact, WNGS method acted in two directions. Firstly, WNGS method improved feature domain by adding features gradually and then evaluated the accuracy on each iteration. Finally, the classifier was back with the optimal feature subset. Hence, WNGS worked in both directions effectively.

It minimized feature space and achieved satisfying results regarding classification accuracy. So, WNGS method is a powerful dimensionality lowering method that can work well on different size datasets.

6. Conclusion

Feature selection is treated as a prior step that must be passed before making a classification decision. An integrated feature selection method was introduced to pick best features from overall ones. After selecting ideal feature subset, feature subset is now ready for the breast cancer classification task. An integrated method that combined NB learning scheme with a forward greedy search method for feature selection task was introduced. The attachment of a forward greedy search strategy with a wrapper method helped in improving classification accuracy. Experimental results clarified that WNGS feature selection method achieved better classification performance. Also, it is very simple, speedy way to be applied to the problem domain. It is an ideal way to be used for feature selection tasks. WNGS method is suitable for the nature of the problem domain. NB as a learning classifier proved its strength and durability in dealing with a learning problem.

Forward greedy search strategy had shown good performance in their quest to find the most efficient feature subset. However, some future work is needed to improve classification accuracy. It is possible to use randomized wrapper methods. Most popular randomized wrapper methods use GA and simulated annealing. Another track is to make comparisons and test different classifiers for evaluating the quality of selected feature subset by forwarding greedy search method. Much more experimental efforts and data analysis should be spent on more complex data sets.

REFERENCES

[1] Tao Chen, Zenglin Hon, Hui Zhao, Xiao Yang and Jun Wei (2015). A novel feature gene selection method based on neighborhood mutual information. International Journal of Hybrid Information Technology, vol. 8, no.7, 272-292.

[2] M. Dash and H. Liu (1997). Feature selection for classification, Intelligent Data Analysis, vol. 1.

[3] T. Karthikeyan and P. Thangaraju (2015). Best first and greedy search based CFS-Naive Bayes classification algorithms for hepatitis diagnosis. Biosciences and Biotechnology Research Asia, vol.12, no.1, 983-990.

[4] Othman Soufan, Dimitrios Kleftogiannis, Panos Kalnis and Vladimir B. Bajic (2015). DWFS: A Wrapper feature selection tool based on a parallel genetic algorithm. PLoS One,vol.10,no.2. https://doi.org/10.1371/journal.pone.0117988.

[5] Ahmed A. A., M Mokhtar, M. I. B. Saripan, M. H. B. Abu Bakar (2015). Integrated framework of feature selection from microarray data for classification. Journal of Theoretical and Applied Information Technology, vol.73, no.2.

[6] S. Aruna and S. P. Rajagopalan (2011). A novel SVM based CSSFFS feature selection algorithm for detecting breast cancer. International Journal of Computer Applications, vol. 31, no.8.

[7] WEKA: A multi-task machine learning software developed by Waikato University 2006. http://www.cs.waikato.ac.nz/ml/weka.

[8] Chuan Liu, Wenyong Wang, Qiang Zhao, Xiaoming Shen, Martin Konan (2017). A new feature selection method based on a validity index of feature subset, Pattern Recognition Letters, vol. 92, 1–8.

[9] UCI machine learning repository. http://archive.ics.uci.edu/ml/ (Last accessed on 09/09/2017)

[10] Bouaguel W. (2016). A new approach for wrapper feature selection using genetic algorithm for big data. In: Lavangnananda K., Phon-Amnuaisuk S., Engchuan W., Chan J. (eds), Intelligent and Evolutionary Systems. Proceedings in Adaptation, Learning and Optimization, vol 5. Springer, Cham.

[11] J. C. H. Hernandez, B. Duval, and J. Hao. (2007). A genetic embedded approach for gene selection and classification of microarray data, In: Proceedings of the 5th European Conference on Evolutionary computation, machine learning and data mining in bioinformatics, Springer Berlin Heidelberg.

[12] R. Kohavi and G. H. John (1997). Wrappers for feature selection. Artificial Intelligence, 273–324.

[13] F. Liu, H. Yu. (2014). Learning to Rank Figures within a Biomedical Article, PLoS |ONE, vol. 9, no. 3.

[14] P. Langley, W. Iba, and K. Thompson. (1992). An analysis of Bayesian classifiers. Proceedings of the Tenth National Conference on Artificial Intelligence, 223-228.

[15] Dumitru, D. (2009). Prediction of recurrent events in breast cancer using the Naive Bayesian classification. Annals of the University of Craiova-Mathematics and Computer science series, vol. 36, no. 2, 92-96.

[16] G. I. Webb. (2010). Naïve Bayes. In: Encyclopedia of Machine Learning, C. Sammut, and G. I. Webb, Eds., Springer, New York, NY, USA, 713–714.

[17]UCImachinelearningrepository. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29/ Last accessed on 20/09/2017).

[18]UCImachinelearningrepository. https://archive.cs.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29/ Last accessed on 20/09/2017).

[19] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten (2009). The WEKA data mining software: an update, ACM SIGKDD Explorations Newsletter, vol. 11, no. 1, 10-18.

[20] G. Kowalski (1998). Information retrieval systems: theory and implementation, Comput. Math. Appl., vol. 5, no. 35, 133-134.

[21] D. M. Powers (2011). Evaluation: from Precision, Recall and F-Measure to ROC. Informedness, Markedness & Correlation, Journal of Machine Learning Technologies, vol. 2, no. 1, 37–63.