Incremental Wrapper Based Random Forest Gene Subset Selection for Tumor Discernment

Alia Fatima(✉), Usman Qamar, Saad Rehman,
and Aiman Khan Nazir

National University of Sciences and Technology (NUST), Islamabad, Pakistan
{alia.fatima16,aiman.nazir16}@ce.ceme.edu.pk,
{usmanq,saadrehman}@ceme.nust.edu.pk

Abstract. High-dimensional cancer related dataset permits the researchers to timely diagnose and facilitate in effective treatment of the cancer. Biomedicine application process on the thousands of features. It is challenging to extract the precise statistics from this high-dimensional dataset. This paper presents the Incremental Wrapper based Random Forest Gene Subset Selection of Tumor discernment that mechanisms on the principle of incremental wrapper based feature subset selection with random forest classification algorithm and this algorithm also works as performance validator. Incremental wrapper based feature subset selection is a technique to pick out a finest conceivable subset of genes from the high-dimensional data with low computational cost. Random Forest will increase the overall performance as it works better in cancer related high-dimensional dataset. The efficacy of the random forest classification algorithm as performance validator will significantly improve by working on a selective discriminative subset of prognostic genes as compare to the raw data. We evaluate the proposed methodology on the six publicly available cancer related high dimensional datasets and found that the proposed methodology outperform as compare to standard random forests.

Keywords: Cancer classification · Random forest · IWSS
Incremental wrapper based gene subset selection

1 Introduction

Cancer is the most serious illness in which certain cells of the body grow in an uncontrolled way. It can be cured if timely diagnosed. Nowadays, researchers are extensively working on the early diagnosis and treatment of the cancer [1]. Diverse technologies are used in this respect in which one of the most focused technique is the analysis of the disease related information. Intensive research carries on the analysis of microarrays data [2, 3]. Biology and biomedicine applications are extensively using for this purpose. These applications are generating a large number of genes which is further used for the diagnosis of the disease such as high dimensional datasets are used for the discernment of the cancer. High dimensionality is the fundamental problem for extracting and analyzing the vital information from gene-expression data [4]. These high dimensional datasets have thousands of genes.

M. Elloumi et al. (Eds.): DEXA 2018 Workshops, CCIS 903, pp. 161–167, 2018.
https://doi.org/10.1007/978-3-319-99133-7_13

These datasets are used in two perspectives: first to accurately diagnosis the disease and second to distinguish the particular set of genes which are responsible for the disease [4]. Feature selection techniques are used for selecting the prognostic genes. As, most of the classification algorithms give, the better results by developing the model on fewest number of genes [2].

Feature selection algorithm comprises of two parts, first one is a search technique for new feature subset and the second one is an evaluation measure to score the given feature subset. These techniques are used to increase the efficacy of the classifiers. Microarray datasets are high dimensional datasets with a large number of genes. These all genes are not relevant. Irrelevant and redundant features decreases the performance of the classifier [5].

Bioinformatics community has the deepest interest in Random forest (RF) algorithm [6] for the classification of high-dimensional and microarray dataset from the past few years [7, 8]. Recent work [8] demonstrate that random forest has better performance as compared to other best performing classifier in the cancer microarray gene expression domain. Irrelevant features present in high dimensional data set to deteriorate the performance of the learning algorithm [9].

The effective feature selection method can be used to lessen this problem by selecting a subset of discriminative features from the complete feature set [10, 11]. There are three groups for feature selection method: filter, wrapper and embedded [12]. Filter methods use the intrinsic properties of the training sets to evaluate the features and selecting a valuable feature subset. The generalization ability of the methods is better along with the computational complexity. These methods are flexible to work with diverse classifier. The wrapper method evaluates the quality of the feature subset accordingly to the embedded classifier. These methods gain the better classification performance as compared to proceeding one [13, 14]. The embedded method works on the combine qualities of the preceding two methods.

In this study, we will present a novel methodology for the discernment of the high dimensional cancer dataset. The proposed methodology will alleviate the curse of dimensionality and will increase the performance of the random forest in the high dimensional dataset for the accurate discernment of tumor. Remaining paper is arranged as follows: Sect. 2 presents literature review, Sect. 3 presents the proposed methodology, Sect. 4 presents the experimental results and paper is concluded in Sect. 5.

2 Literature Review

2.1 Random Forest

Random forest is an ensemble which is composed by a group of classification trees. Breiman first introduced the concept of random forest in early 2001 [6]. Random forest is developed by the following procedure:

n number of instances are randomly selected from the total N number of instances. These n instances sample sets are used as training dataset for developing the decision tree model.

m number of variables are selected from the total M number of variables. These variables are selected randomly. These m number of variables are used at each node for splitting the node on the basis of best splitting criteria. m remains constant throughout the growth of the trees. Trees are grown on the maximum possible size and there is no pruning.

New objects are classified on the basis of input vector. This input vector is evaluated on the basis of each already developed a decision tree model in the forest. Each tree gives a classified class result and final result selected on the basis of the majority voting.

Forest error rate depends on following two attributes according to the original work.

Forest error rate rise with the increase of the correlation value or similarity between any pair of trees in the forest.

A tree is considered a good classifier with a low error rate.

2.2 Incremental Wrapper Based Feature Subset Selection

Ruiz et al. [15] proposed the incremental wrapper based feature subset (IWSS) algorithm. IWSS algorithm consists of two phases. In the first phase, it uses the feature ranking algorithm for scoring function. The scoring function assigns the score to the individual feature on the basis of computing the score from the values of each feature and class label. Features are ranked in the list in decreasing order as, feature with the highest score is first in the list and so on.

In the second phase, algorithm selects the feature with the highest score and make it the best feature subset after computing the accuracy with the selected learner. It computes the accuracy of the next feature subset which is developed by inducing the next highest scoring feature in the existing subset. This subset is selected as the best feature subset if its accuracy is greater than the previously computed accuracy. This Process is repeated until the last feature is evaluated.

```
Input: D training U-measure, L-classifier
Output: Best Subset
-----------------------------------------
List R={}
For each gene gi ϵ D
   Score = compute (gi, U, D)
   append gi to R according to Score
BestClassif = 0
BestSubset = ∅
For i =1 to N
    TempSubset = BestSubset ∪{gi}(gi ϵ R)
    TempClassif = WrapperClassif(TempSubset, L)
    if(TempClassif > BestClassif)
       BestSubset = TempSubset
       BestClassif = TempClassif
```

Algorithm. 1. Incremental Wrapper Based Feature Subset Selection

2.3 OneRAttributeEval

OneRAttributeEval evaluates the worth of an attribute by using the OneR classifier.

2.4 OneR

OneR [16] is a classification algorithm. It is a short form of One Rule. Holte proposed it in 1993. It is a simple for humans to interpret, but accurate classification algorithm, which performs well on most commonly used datasets. It generates one rule for each predictor in the data. Then, it selects the rule with the smallest total error as its "one rule". A frequency table is constructed for each predictor against the target to create a rule for a predictor.

OneR Algorithm.

```
   For each predictor,
   For each value of that predictor, make a rule as follows;
       Count how often each value of target (class) appears
       Find the most frequent class
       Make the rule assign that class to this value of the predictor
   Calculate the total error of the rules of each predictor
Choose the predictor with the smallest total error.
```

3 Proposed Methodology

Random forest performs well on high-dimensional dataset as compare to other classifiers. In this work, we are proposing a novel methodology for the cancer classification on the basis of high dimensional dataset. This work proposes to use the incremental wrapper based feature subset selection with OneRAttributeEval and Random Forest for the improved and accurate results of prognosis of cancer. This methodology comprises of two phases. In the first phase, it incrementally selects the feature subset and in a second phase, it evaluates the accuracy of the classifier with a selected subset of features for tumor discernment.

Phase 1: Incremental wrapper based feature subset is used to select the relevant features by eliminating the redundant features. It incrementally selects the feature subset and OneRAttributeEval algorithm evaluates the worth of an attribute by using the OneR Classifier and features are ranked on the basis of their worth. The performance of the features is evaluated on the basis of the algorithm accuracy. Random forest is used to evaluate the given subset of the features.

Phase 2: In this phase, a selected subset of features is evaluated and accuracy of the dataset is evaluated by using the random forest (Fig. 1).

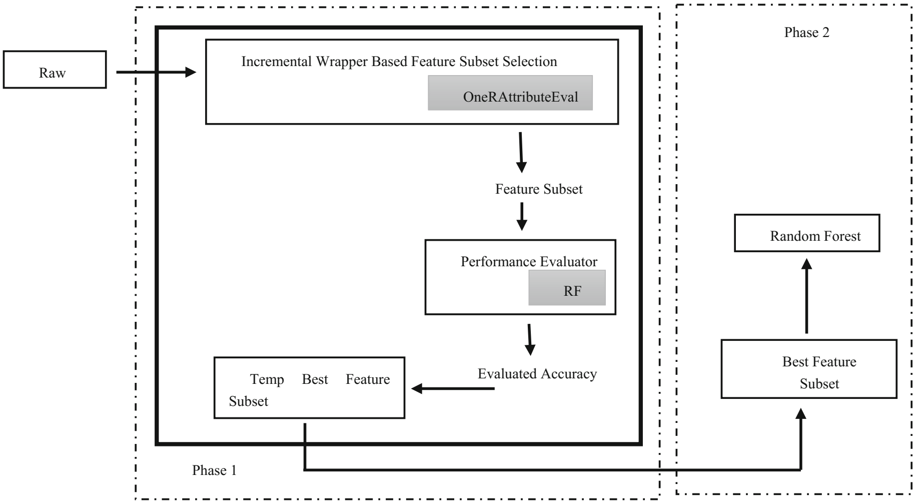

Fig. 1. Framework of proposed approach: wrapper based RF gene subset selection for tumor discernment

4 Experimental Results

We have performed the experimental evaluation of our proposed approach on Weka 3-8. Our evaluation environment worked on the machine which has i3 core processor, 6 GB RAM, 64 bit operating system and Windows 8.1. Table 1 presents the features of the selected cancer related datasets, such as no. of attributes and no. of instances.

Table 1. Dataset used.

Dataset name	No. of attribute	No. of instances
Colon	2000	62
GCM	16064	190
Leukemia	7130	72
Lung-Cancer	56	32
Lymphoma	4027	96

Leukemia dataset is obtained from [17], Lung-Cancer datasets are obtained from UCI Repository [18], and Colon, Lymphoma and GCM datasets are obtained from [19]. We used accuracy parameter as the performance measure of our proposed methodology.

Accuracy is used to predict the class label. Accuracy is defined as,

$$Accuracy = \frac{TruePositive + TrueNegative}{TruePositive + TrueNegative + FalsePositive + FalseNegative}$$

The experimental results of our proposed approach are shown in Table 2 on the basis of the above mentioned cancer related datasets.

Table 2. Results of the proposed approach.

Dataset name	Selected no. of attribute	Accuracy
Colon	4	87.096%
GCM	23	62.1053%
Leukemia	4	91.67%
Lung-Cancer	3	68.75%
Lymphoma	13	85.4167%

We compared the results of our proposed approach with the standard Random Forest Algorithm. Comparison of the standard Random Forest Algorithm and our proposed Approach is shown in Table 3.

Table 3. Comparison of the proposed approach and standard random forest.

Dataset name	Standard RF	Proposed approach
Colon	83.87%	87.096%
GCM	61.0526%	62.1053%
Leukemia	90.27%	91.67%
Lung-Cancer	46.75%	68.75%
Lymphoma	84.375%	85.4167%

Above mentioned results demonstrate that our proposed methodology performs well in cancer related high-dimensional dataset as compared to the standard Random Forest by selecting the minimum prognostic genes.

5 Conclusion

Cancer is found to be a most killing disease over the globe. Cancer diagnosis and treatment related biomedicine applications work on the high dimensional dataset. These datasets consist of the number of redundant features. It is important to remove the redundant and irrelevant features from the high dimensional dataset to get the valuable statistics. This paper presents a novel methodology: Incremental wrapper based random forest gene subset selection for tumor discernment. This methodology comprises of two phases. In the first phase, minimum relevant prognostic subset of genes is selected and in the second phase, Random Forest is used for the classification of the dataset as, it is found to perform well in the classification of disease related high dimensional dataset. The performance of the proposed methodology is evaluated on the basis of six high dimensional cancer related datasets. Accuracy is used as the evaluation measure of the performance. We compare the results of the proposed approach with the standard

Random forest and found that proposed methodology is giving the accurate results as compare to standard random forest with the less number of genes. In the future, we will prefer to work on the embedded random forest feature subset selection to decrease the computational cost and time complexity.

References

1. Ahmad, F., Isa, N.A.M., Hussain, Z., Osman, M.K., Sulaiman, S.N.: A GA-based feature selection and parameter optimization of an ANN in diagnosing breast cancer. Pattern Anal. Appl. **18**, 861–870 (2015)
2. Mishra, D., Sahu, B.: Feature selection for cancer classification: a signal-to-noise ratio approach. Int. J. Sci. Eng. Res. **2**, 1–7 (2011)
3. Deng, L., Pei, J., Ma, J., Lee, D.L.: A rank sum test method for informative gene discovery. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 410–419 (2004)
4. Dasgupta, S., Saha, G., Mondal, R.: A comparison between methods for generating differentially expressed genes from microarray data for prediction of disease. In: Proceedings of the 2015 Third International Conference on Computer, Communication, Control and Information Technology (C3IT) (2015)
5. Hua, J., Tembe, W.D., Dougherty, E.R.: Performance of feature-selection methods in the classification of high-dimension data. Pattern Recogn. **42**, 409–424 (2009)
6. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)
7. Wu, B., et al.: Comparison of statistical methods for classification of Ovarian cancer using mass spectrometry data. Bioinformatics **19**, 1636–1643 (2003)
8. Díaz-Uriarte, R., De Andres, S.A.: Gene selection and classification of microarray data using random forest. BMC Bioinform. **7**, 3 (2006)
9. Shu, W., Shen, H.: Incremental feature selection based on rough set in dynamic incomplete data. Pattern Recogn. **47**, 3890–3906 (2014)
10. Prabhakar, S., Jain, A.K.: Decision-level fusion in fingerprint verification. Pattern Recogn. **35**, 861–874 (2002)
11. Gheyas, I.A., Smith, L.S.: Feature subset selection in large dimensionality domains. Pattern Recogn. **43**, 5–13 (2010)
12. You, W., Yang, Z., Ji, G.: PLS-based recursive feature elimination for high-dimensional small sample. Knowl. Based Syst. **55**, 15–28 (2014)
13. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artif. Intell. **97**, 273–324 (1997)
14. Inza, I., Larrañaga, P., Blanco, R., Cerrolaza, A.J.: Filter versus wrapper gene selection approaches in DNA microarray domains. Artif. Intell. Med. **31**, 91–103 (2004)
15. Ruiz, R., Riquelme, J.C., Aguilar-Ruiz, J.S.: Incremental wrapper-based gene selection from microarray data for cancer classification. Pattern Recogn. **39**, 2383–2392 (2006)
16. Holte, R.C.: Very simple classification rules perform well on most commonly used datasets. Mach. Learn. **11**, 63–91 (1993)
17. Cancer program data sets (2010). Broad Institute. http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi
18. Frank, A., Asuncion, A.: UCI machine learning repository (2010). http://archive.ics.uci.edu/ml
19. Dataset repository in ARFF (weka) (2010). BioInformatics Group Seville. http://www.upo.es/eps/bigs/datasets.html