J Inf Process Syst, Vol.14, No.5, pp.1167~1175, October 2018
https://doi.org/10.3745/JIPS.04.0087

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

A Clustering Approach for Feature Selection in Microarray Data Classification Using Random Forest

Husna Aydadenta* and Adiwijaya*

Abstract

Microarray data plays an essential role in diagnosing and detecting cancer. Microarray analysis allows the examination of levels of gene expression in specific cell samples, where thousands of genes can be analyzed simultaneously. However, microarray data have very little sample data and high data dimensionality. Therefore, to classify microarray data, a dimensional reduction process is required. Dimensional reduction can eliminate redundancy of data; thus, features used in classification are features that only have a high correlation with their class. There are two types of dimensional reduction, namely feature selection and feature extraction. In this paper, we used k-means algorithm as the clustering approach for feature selection. The proposed approach can be used to categorize features that have the same characteristics in one cluster, so that redundancy in microarray data is removed. The result of clustering is ranked using the Relief algorithm such that the best scoring element for each cluster is obtained. All best elements of each cluster are selected and used as features in the classification process. Next, the Random Forest algorithm is used. Based on the simulation, the accuracy of the proposed approach for each dataset, namely Colon, Lung Cancer, and Prostate Tumor, achieved 85.87%, 98.9%, and 89% accuracy, respectively. The accuracy of the proposed approach is therefore higher than the approach using Random Forest without clustering.

Keywords

Classification, Clustering, Dimensional Reduction, Microarray, Random Forest

1. Introduction

Cancer is a known deadly disease around the world. According to the World Health Organization (WHO) data in 2015, 8.8 million deaths were caused by cancer, with the number set to increase every year if diagnosis was not resolved early [1]. There are many ways to detect cancer, including one that is known as the microarray technique. Microarray analysis plays an essential role in diagnosing a disease because it can be used to look at the level of gene expression in a particular cell sample and examine thousands of genes simultaneously [2]. Therefore, microarray has high data dimensionality. The bigger the size of the data and the number of fixed observations, then the accuracy of classification at a certain point will be smaller. To overcome this problem, a reduction process is conducted.

Researchers have performed cancer detection based on microarray data classification. Several algorithms have been proposed, with some papers examining each algorithm separately in a closed

Manuscript received April 18, 2018; first revision June 7, 2018; accepted July 14, 2018.
Corresponding Author: Husna Aydadenta (aydadenta@student.telkomuniversity.ac.id)
*School of Computing, Telkom University, Bandung, Indonesia (aydadenta@student.telkomuniversity.ac.id, adiwijaya@telkomuniversity.ac.id)

condition. These algorithms have advantages and weaknesses. The main problem of microarray data is that it has more variables than the samples. However, to get an accurate model using classification, a lot of sample data and variables that have correlation with classes in the dataset are required. Therefore, the variables or relevant features must first be determined.

In the work of Moorthy and Mohamad [3], the Random Forest algorithm was used for gene selection and classification. This is because this algorithm can look for essential variables in a dataset. It is also suitable for use on datasets that have more numbers of variables than the number of samples. However, the Random Forest algorithm is only able to detect significant variables, but not redundant variables. Redundant variables are variables or features that are similar. In [4], if the same or similar characteristics or variables are used in the classification process, the accuracy of the model could be decreased. Hence, an approach to remove redundancy of data is required.

Some studies apply two approaches to find relevant variables and remove redundancy in the dataset, i.e., using feature extraction and feature selection [5]. Feature selection works by removing irrelevant features and redundancy. The purpose of feature selection is to get rid of irrelevant and noisy genes from the input data set, to speed up the processing of data by reducing data dimensionality, and to avoid overfitting of the classifier [6]. Meanwhile, feature extraction works by transforming the original data into a new representation. Feature extraction shares the same goal as feature selection in that it eliminates irrelevant or noisy features in the data and removes redundancy in the data in order to increase the value of classification accuracy [7].

Both approaches have weaknesses and advantages. Feature selection can preserve data characteristics for interpretability, but still maintain discriminative power, shorter training times, and reduce overfitting. Meanwhile, feature extraction yields higher discriminating power, but at a loss of data interpretability and transformation, which may be expensive. To avoid change or perhaps costly measures and loss of data interpretability, this research proposes a method for feature selection to remove redundancy features from the data.

The proposed approach is the clustering method. Ismi et al. [8] proved that the clustering approach can be used to remove feature redundancy by grouping features that are similar into the same cluster. After each group is clustered, one sample will be taken from the same cluster to represent each cluster as a subset of features. This will be used in the classification process. Previous works have discussed many feature selection algorithms such as Relief and information gain, but the algorithms cannot classify features that have similar characteristics in a microarray data. By using the proposed approach, similar characteristics in microarray data can be clustered; hence, improving classification performance [6].

A ranking system was used in this study to select a sample that will represent each cluster. The ranking process aims to determine the feature with the highest correlation to the dataset class. This process design is expected to produce absolute accuracy for the classification model being constructed.

Based on the advantages of the Random Forest algorithm described in previous research, this research will analyze the performance of the Random Forest algorithm for classification. For feature selection, a clustering approach involving several data microarray datasets is applied by conducting a development dimension reduction process. To observe the performance of the algorithm, some scenarios are tested, such as using parameters of the algorithm, and dividing between training data and testing data.

2. Existing Device Discovery Scheme

2.1 Data

This research used microarray data obtained from Aydadenta [9]. Three datasets were used in this study, as outlined in Table 1. Each dataset was divided into two parts, namely training data and testing data, where training data was used for the learning process, and testing data was used in the testing process of the model obtained.

Table 1. DNA microarray dataset

Data	Class	Sample	Feature
Colon cancer	2	62	2,000
Lung Cancer	2	181	12,533
Prostate Tumor	2	136	12,600

2.2 k-Means Algorithm

k-means is an unsupervised learning clustering algorithm (no class labels required). This algorithm groups given objects into multiple clusters '*k*'. k-means selects *k* at random as a central point (centroid) [10]. Data is grouped by calculating the closeness of the object with the centroid using Euclidean distance.

After that, the centroid is recalculated to generate a new centroid by calculating the actual data rate based on the cluster it occupies. This process is repeated until the new centroid does not change against the old centroid. This algorithm aims to minimize the objective function known as sum square error.

2.3 Relief Method

Kira and Rendell [11] developed the Relief algorithm. The main idea of this algorithm is to estimate attributes based on how well the different values between instants are close to one another. In this algorithm, we need to find the two nearest neighbor values of one of the same class (nearest hit) and one from a different class (nearest miss).

A useful attribute must be different from a different class and should be the same value from an instant if it comes from the same category. The Relief algorithm is capable of handling discrete and continuous attributes and is limited to two classes. However, the Relief algorithm has not yet mastered how to handle incomplete data and problems of more than two categories.

2.4 Random Forest Algorithm

Random Forest is a classifier consisting of several decision trees. Random Forest is included in the ensemble method using the decision tree. This method is random because it uses a random way of making a decision tree. This random approach can help to eliminate correlations between decision trees so that the accuracy of the method used can be improved, such as the ensemble method characteristics.

Each decision tree is built using a random vector. The general approach used to include a random vector in a tree-building process is to select an arbitrary value '*F*', as many as *F* attributes input to be

split on each node in the decision tree to be formed. The benefit of choosing a random value *F* is that one does not have to check all the attributes, but only look at selected *F* attributes. The parameters to regulate the strength of the Random Forest algorithm lie in the selection of *F* values and the number of trees to be built in the forest.

3. Proposed Approach

This research proposed a method combining a clustering algorithm and a classification algorithm i.e., the k-means and Random Forests. Based on evidence from previous research [12], Random Forest has some advantages when it comes to microarray data classification. However, this algorithm has not been able to properly remove redundancy, so a clustering algorithm is needed to remove redundancy dimensions [13].

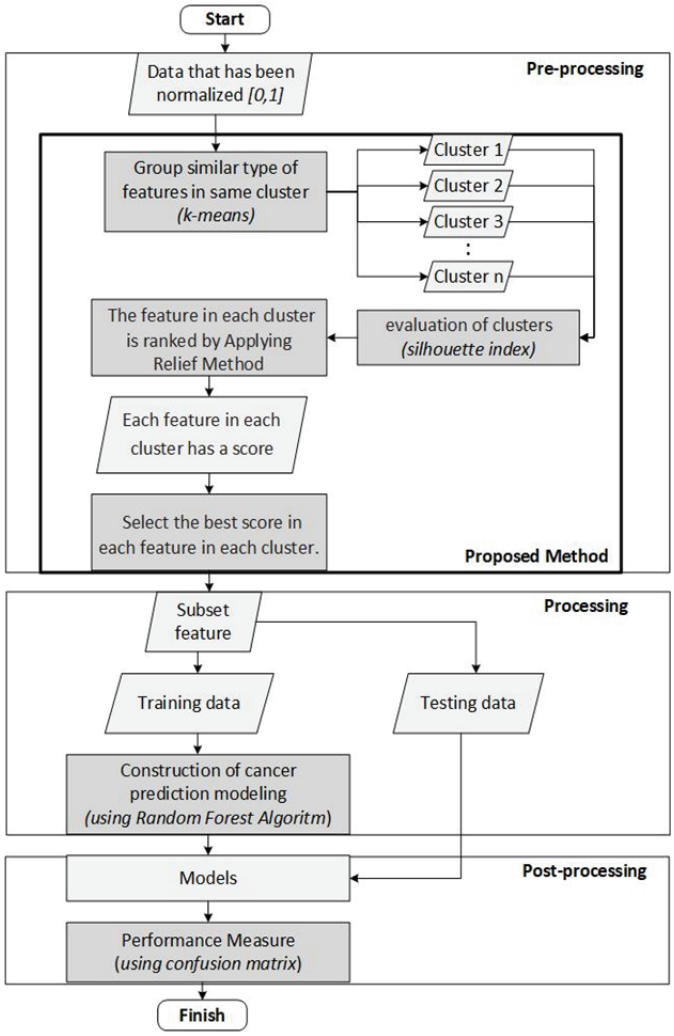

Fig. 1. Research design process.

Fig. 1 shows the three main design processes in this study, which are preprocessing, processing, and post-processing. Preprocessing is a process for removing redundant features, in which a clustering approach using the k-means algorithm is employed, the goal of which is to group features that are similar in a cluster. This is one way to remove redundant features (i.e., the duplicate features in datasets).

The process of reducing redundancy dimension is shown in Fig. 2. The process uses feature selection relying on two algorithms, which are the k-means and Relief algorithms. The steps in this process are given below:

1. Features of data are clustered by applying the k-means clustering algorithm. By applying the clustering technique, similar types of features can be grouped in the same cluster to remove redundancies in microarray data and to generate the centroid in the k-means algorithm randomly.
2. The Relief method for ranking the feature in each cluster is applied.
3. The feature in each cluster, which has a n-top ranking, is selected.
4. Finally, the subset of features obtained in Step 3 will be used in the dataset for the classification process.

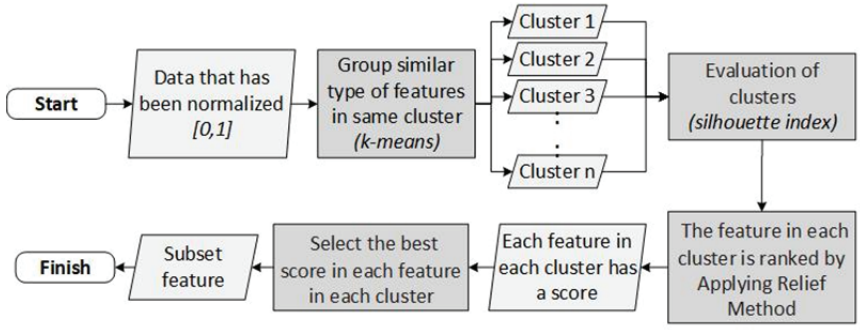

Fig. 2. Flowchart of dimensional reduction.

Then, further still in the preprocessing, clusters formed from the k-means algorithm will take one feature for each cluster as a cluster representative within the feature subset. To select the features to be chosen as representatives, the Relief method is used. The result of the Relief method is weight, so each feature in the cluster will have a specific weight. The weights obtained for each cluster are ranked, so each cluster will have a ranking. The top-ranking cluster will be selected and a feature subset will be made to build the model in the classification process.

Next is processing, namely the process of classification using the Random Forest algorithm. The feature subset obtained from the previous process will be used as input in this process. The microarray dataset is divided into two, the training data and test data, which proportionally, the training data is always more than the test data. This is because the training process in machine learning requires more data for modeling exercises. Then, after the next model is obtained, the evaluation stage (post-processing) is conducted. The evaluation stage uses a confusion matrix. Test data obtained from previous dataset divisions is used to obtain the accuracy of the model. From the results of the evaluation measure, we can determine whether or not the model developed from the design of this research process is feasible.

4. Performance Analysis

To evaluate the performance of the proposed scheme, we use a silhouette index and a confusion matrix. The function of the silhouette index is to test the quality of the resulting cluster. This method is a cluster validation method that combines Cohesion and Separation methods. To calculate the silhouette index value, the distance between documents is calculated using a Euclidean distance formula. The value of the silhouette index is divided into four categories, namely Strong Structure, Medium Structure, Weak Structure, and No Structure. On the other hand, the confusion matrix is a method used to perform accuracy calculations for data mining concepts. This formula shows calculations with four outputs, namely recall, precision, accuracy, and error rate. Precision defines the number of records that are correctly classified from all records classified by the classifier. Meanwhile, recall establishes the number of records classified appropriately by the classifier for all records that should be appropriately classified by the classifier. Precision defines the number of records that are exact classifications of all records classified by the classifier. Recall establishes the number of records classified appropriately by the classifier for all records that should be appropriately classified by the classifier. This research uses three parameters, i.e., the *k*-Cluster, *n*-Ranking, and *n*-Trees, where the parameters used are outlined in Table 2.

The parameter selection in this research was done via an empirical study. From several scenarios that were conducted, the optimal parameters for each dataset are obtained, as per Table 3.

Table 2. Parameters

Data	*k*-Cluster
k-Cluster	2, 4, 6, 8, 10, 12, 14, 16,18, 20, 22, 24, 26, 28, 30, 32, 34, 36, and 38
n-Ranking	1, 2, 3, 4, 5, 6, 7, 8, 9, and 10
n-Trees	20, 30, 40, 50, 60, and 70

Table 3. Summary of optimal parameter

Data	*k*-Cluster	*n*-Ranking	*n*-Trees
Colon cancer	2	2	60
Lung Cancer	8	3	60
Prostate Tumor	6	8	60

From the optimal parameters in Table 3, the optimal *n*-Cluster for Colon Cancer is 2, which means that for the Colon Cancer dataset there are 2 feature groups, and from the two groups of features, the optimal *n*-Ranking taken for each cluster is two features. In other words, for the feature subset of Colon Cancer there are only 4 features, with 62 samples. Therefore, the optimal *n*-Trees are 60 trees, so, for the classification model using Colon Cancer, there are 60 classifiers with 8-fold cross-validation, or there are eight variants of training data and testing data to assess or validate the accuracy of the classification model.

Next, the optimal *n*-Cluster for lung cancer is 8, meaning for the lung cancer dataset, there are 8 clusters containing the features of the dataset, and from the 8 clusters, the optimal *n*-Ranking taken for each feature is 3. In other words, for the lung cancer feature subset, there are only 24 features, with 181 samples. Subsequently, the optimal *n*-Trees are 60 trees, so, for the classification model using lung

cancer, there are 60 classifiers and 3-fold cross-validations, or there are three variants of training data and testing data to assess or validate the accuracy of the classification model.

The optimal *n*-Cluster for Prostate Tumor is 6, meaning for the Prostate Tumor dataset, there are 6 clusters containing the features of the dataset, and from the 6 clusters, the optimal *n*-Ranking taken for each feature is 8. In other words, for a subset of Prostate Tumor features there are only 48 features, with 136 samples. Therefore, the optimal *n*-Trees are 60 trees, so for the classification model using Prostate Tumor, there are 60 classifiers with 8-fold cross-validation, or there are 8 variants of training data and testing data to assess or validate the accuracy of the classification model. Hence, based on the optimal parameters obtained for each dataset, the model is trained to achieve these settings. Therefore,the obtained accuracy for each dataset are listed in Table 4.

Table 4. The result of the scenarios

Data	Accuracy (%)	Time	Recall
Colon Cancer	85.87	3.481	0.841
Lung cancer	98.90	1.411	0.993
Prostate Tumor	88.97	4.103	0.855

Based on the results of the final accuracy obtained in this study, a comparison of the model accuracy in this study is done with that of previous research [9], which only used the Random Forest algorithm for gene selection and classification. The results can be seen in Fig. 2.

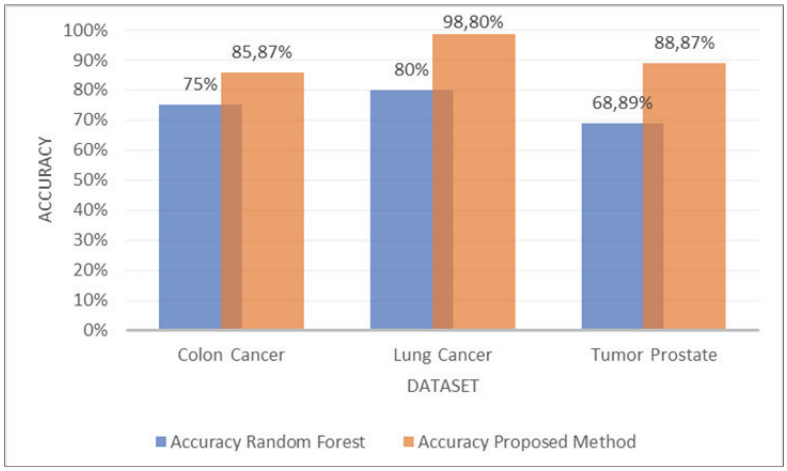

Fig. 3. Comparison of accuracy of Random Forest without clustering and the proposed approach.

Based on Fig. 3, there are three types of microarray data used for accuracy comparison, namely Colon Cancer, Lung Cancer, and Prostate Tumor. The orange color indicates the accuracy of the previous research [9], and the yellow color shows the accuracy of the proposed approach. From Fig. 3, the accuracy of proposed approach is always higher than the accuracy of previous work. This means the accuracy of the proposed approach (this research) is better than previous research, which uses only the RF algorithm for gene selection and classification. The higher accuracy of the proposed approach is due to the method of preprocessing. In the previous research, only the Random Forest algorithm was used,

for which this algorithm used all dataset features to build a model, without deleting redundancy features. Meanwhile, the nature of the Random Forest algorithm is random, so the chances of choosing an irrelevant feature during the construction of the model are huge, and therefore the accuracy obtained is always lower than the accuracy of the model proposed, which has undergone preprocessing (deletion of redundancy features).

The proposed approach has gone through preprocessing in which the redundant features were deleted, so the process of classification using the Random Forest algorithm already utilized a subset of features relevant to the class dataset. Also, when running the algorithm, only relevant features are selected, resulting in higher accuracy compared to the model that did not go through pre-processing (reduction of features).

5. Conclusions

In this research, we reduced the redundancy in microarray data, namely Colon Cancer, Lung Cancer, and Prostate Tumor datasets. We provided four scenarios to observe the effect of some parameters used to model the proposed approach. In this study, four scenarios were conducted to observe the effect of some parameters used to build the model of the proposed approach. In the first scenario, a cluster parameter with measurement evaluation utilizing the silhouette index was used. The second scenario looked at the effect of taking a large number of features for each cluster using a confidence matrix. The third scenario looked at the impact of the tree parameter against the time required for execution and resulting accuracy. The last scenario tested for the effects of training data and testing data on the performance model, using a k-fold cross-validation. Each scenario was run three times with the same parameters, so that for the evaluation of the measure, an average value could be used. Because each dataset has different data characteristics, the optimal settings for each dataset were also varied. Of all the scenarios obtained, the highest accuracy for the Colon Cancer, Lung Cancer, and Prostate Tumor dataset was 85.87%, 98.9%, and 88.97%, respectively. Meanwhile, using the same number of tree parameters, the results of accuracy for previous research were 75%, 80%, and 68.89% for Colon Cancer, Lung Cancer, and Prostate Tumor, respectively. Therefore, the accuracy of this work is higher than previous research, which only used the Random Forest algorithm for gene selection and classification. In addition, after running some scenarios, it can be concluded that the clustering approach applied for the microarray data to remove redundancy could be used and applied in cancer detection. The benefits of this research is that it provides information about other combinations of filter methods to remove redundancy in microarray data by utilizing a clustering approach. For further research, the clustering process can be redeveloped. This is because this research used the k-means algorithm so random centroids were initialized. For future work, it may be possible to add an optimization algorithm to obtain more optimal research parameters.

Acknowledgement

The authors would like to thank the Ministry of Research, Technology, and Higher Education (Republic of Indonesia) for financially supporting this research.

References

[1] American Cancer Society, *Cancer Facts & Figures 2015*. Atlanta, GA: American Cancer Society, 2015.

[2] A. Nurfalah and A. A. Suryani, "Cancer detection based on microarray data classification using PCA And modified back propagation," *Far East Journal of Electronics and Communications*, vol. 16, no. 2, pp. 269-281, 2015.

[3] K. Moorthy and M. S. Mohamad, "Random forest for gene selection and microarray data classification," in *Knowledge Technology*. Heidelberg: Springer, 2012, pp. 174-183.

[4] E. Pashaei, M. Ozen, and N. Aydin, "A novel gene selection algorithm for cancer identification based on random forest and particle swarm optimization," in *Proceedings of 2015 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB),* Niagara Falls, Canada, 2015, pp. 1-6.

[5] Z. M. Hira and D. F. Gillies, "A review of feature selection and feature extraction methods applied on microarray data," *Advances in Bioinformatics*, vol. 2015, article ID. 198363, 2015.

[6] P. K. Ammu and V. Preeja, "Review on feature selection techniques of DNA microarray data," *International Journal of Computer Applications*, vol. 61, no. 12, pp. 39-44, 2013.

[7] C. S. Tan, W. S. Ting, M. S. Mohamad, W. H. Chan, S. Deris, and Z. Ali Shah, Z. (2014). A review of feature extraction software for microarray gene expression data," *BioMed Research International*, vol. 2014, article ID. 213656, 2014.

[8] D. P. Ismi, S. Panchoo, and M. Murinto, "K-means clustering based filter feature selection on high dimensional data," *International Journal of Advances in Intelligent Informatics*, vol. 2, no. 1, pp. 38-45, 2016.

[9] H. Aydadenta, "On the classification techniques in data mining for microarray data classification," *Journal of Physics: Conference Series*, vol. 971, no. 1, article no. 012004, 2018.

[10] J. Biesiada, W. Duch, A. Kachel, K. Maczka, and S. Palucha, "Feature ranking methods based on information entropy with Parzen windows," in *Proceedings of International Conference on Research in Electrotechnology and Applied Informatics*, Katowice, Poland, 2015.

[11] K. Kira and L. A. Rendell, "The feature selection problem: traditional methods and a new algorithm" in *Proceedings of the 10th National Conference on Artificial Intelligence (AAAI)*, San Jose, CA, 1992, pp. 129-134.

[12] R. Diaz-Uriarte and S. A. De Andres, "Gene selection and classification of microarray data using random forest," *BMC Bioinformatics*, vol. 7, article no. 3, 2006.

[13] J. Darbon and S. Osher, "Algorithms for overcoming the curse of dimensionality for certain Hamilton–Jacobi equations arising in control theory and elsewhere," *Research in the Mathematical Sciences*, vol. 3, article no. 19, 2016.

Husna Aydadenta https://orcid.org/0000-0002-4024-6545

She is student college from School of Computing, Telkom University. She received a bachelor's degree of computational science at Telkom University and master's degree of Informatics at Telkom University. Her research interests include data mining, big data, machine learning and artificial intelligence.

Adiwijaya https://orcid.org/0000-0002-3518-7587

He is a professor of mathematics at School of Computing, Telkom University. He is interested in the research area of graph theory and its applications, data science, and information science. He joined Telkom University since 2000 and has become professor since 2016.