# Gene Selection for Cancer Diagnosis via Iterative Graph Clustering-based Approach

Mehrdad Rostami[1], Mourad Oussalah[1,2]

[1]*Centre of Machine Vision and Signal Processing, Faculty of Information Technology, University of Oulu, Oulu, Finland*
[2]*Research Unit of Medical Imaging, Physics, and Technology, Faculty of Medicine, University of Oulu, Finland*

### Abstract

The development of microarray devices has led to the accumulation of DNA microarray datasets. Through this technological advance, physicians are able to examine various aspects of gene expression for cancer diagnosis. As data accumulation rapidly increases, the task of machine learning faces considerable challenges for high-dimensional DNA microarray data classification. Gene selection is a popular and powerful approach to deal with these high-dimensional cancer data. In this paper, a novel graph clustering-based gene selection approach is developed. The developed approach has two main objectives, consisting of relevance maximization and redundancy minimization of the selected genes. In this method, in each iteration, one subgraph is extracted, and then among the existing genes in this cluster, appropriate genes are selected using filter-based measure. The reported results on five cancer datasets indicate that the developed gene selection approach can improve the accuracy of cancer diagnosis.

### Keywords

Cancer diagnosis, Microarray data classification, Gene selection, Feature selection, Graph clustering

## 1. Introduction

Worldwide, cancer remains the leading cause of death for both men and women. Cancer diagnosis is important to increase survival chances since early treatment is available to the patients, provided successful early diagnosis. As a result, much research has been conducted to develop better strategies to prevent, diagnose, and treat this disease to decrease the mortality [1, 2]. Using the emerging DNA microarray data, in an experiment, multiple aspects of gene expression can be investigated to diagnose or detect different kinds of cancer.

Existing machine learning and pattern recognition approaches to handle large volume of DNA microarray have been challenged by the high-dimensional structure of a such data [3]. The high data volume makes many genes irrelevant or redundant to cancer diagnosis or classification task. Gene selection is a powerful and efficient technique in microarray data to deal with this challenge [3]. By using this strategy, the training process can be simplified, which, in turn, enhances machine learning performance, and, thereby, the general diagnosis [4].

The main goal of our study is to develop an efficient clustering-based approach to choose a subset of primary genes where one cluster of genes is chosen at each iteration. Then, among the existing genes in this cluster, appropriate ones are selected using a filter-based measure.

*AIxIA 2021 SMARTERCARE Workshop, November 29, 2021, Milan, IT*

Mehrdad.Rostami@oulu.fi (M. Rostami); Mourad.Oussalah@oulu.fi (M. Oussalah)
0000-0001-5710-217X (M. Rostami); 0000-0002-4422-8723 (M. Oussalah)

This process of finding a cluster and selecting a candidate gene from each cluster is iterated until all clusters are selected. We expect that our model, in addition to selecting genes with the least amount of redundancy, will also maximize the relevancy of selected genes.
The remainder of this paper is organized as below: Section 2 reviews some related works. The developed prediction method is presented in Section 3. The experimental results are reported in section 4 and finally, section 5 present the conclusion and future works.

## 2. Related Works

A significant challenge in handling microarray data for cancer diagnosis is their high-dimensionality where the number of genes is much greater than the number of patterns [5, 6, 7], which leads to a well-known problem known as "curse of dimensionality". Gene selection is one popular technique to eliminate irrelevant and/or redundant genes [8]. Previously, gene selection approaches were classified as filter, wrapper, hybrid, and embedded approaches. In the filter techniques, relevant genes are evaluated without a learning model. As a result, these techniques are typically fast [9]. There have been many filter-based approaches for gene selection in cancer diagnosis such as, Filtering and ranking techniques [10], Simplified Swarm Optimization (SSO) [11], Tabu Asexual Genetic Algorithm (TAGA) [12], Feature Clustering and Feature Discretization assisting gene selection (FCFD) [13], Least Loss (LL) [14], etc. The wrapper gene selection strategies employ a learning model to evaluate the efficiency of the chosen gene subset [15]. In this category, an iterative search algorithm-based process is applied to seek the optimum gene subset, and at each step, a subset of original genes are chosen and with a fitness function determining which genes are the best. Despite the fact that wrapper strategies choose a effective subset of original genes, they are computationally complex and may present challenges within analysis of DNA microarray datasets [16]. Hybrid strategies combine the benefits of both filter and wrapper strategies [16]. In addition, the embedded strategies make use of gene selection in the learning process.

## 3. Proposed Method

Our novel gene selection algorithm is introduced by combining the notions of Graph Clustering with Feature Weighting (GCFW). We can group our algorithm under the category of filter gene selection technique and this algorithm measure relevancy and redundancy notions in its selection mechanism. GCFW consists of two main phases including (1) Gene similarity calculation, (2) Iterative subgraph finding and gene selection. In the reminder of this section the details of these two phases are explained.

### 3.1. Gene similarity calculation

Initially, microarray datasets are represented as a weighted graph. In this demonstration, each gene is indicated by a node and the value of each edge shows by gene similarities. In this graph representation, the Pearson correlation coefficient criterion [17] is used to measure the similarity values between two genes. This similarity measure maps the gene space of a microarray dataset into a fully weighted and connected graph. To make the graph sparser, the edges with values lower than the threshold $\theta$ are deleted. $\theta$ is an adjustable parameter and takes its value in the range [0 1]. In our experiment $\theta$ is set to 0.6.

### 3.2. Iterative subgraph finding and gene selection

In order to avoid choosing similar genes, initial genes are divided to several groups to to reduce the possibility of selecting redundant genes. Moreover to select a subset of relevant genes, a feature weighting strategy is developed. In the proposed method, in each iteration, a subgraph is identified and then among these present genes in this subgraph, the suitable ones are chosen using gene relevance.

The purpose of the subgraph discovery is to divide the genes into clusters so that the genes of each cluster are as similar as possible. Previous gene clustering algorithms suffer from inherent shortcomings, like the need to specify the number of clusters, ignoring the distribution of genes and equal consideration of all genes. To deal with these issues, quite different from existing feature clustering algorithms, a fast algorithm for subgraph discovery [18] is employed for gene clustering. This algorithm yields grouping faster than previous gene clustering method.

Moreover, in this proposed method, Fisher Score weighting technique is used to rank the genes of each subgraph and select the relevant genes for representing the genes of that cluster. Therefore, it can claim that the genes selected satisfy both qualities: maximum relevancy and minimal redundancy. In other words, the use of subgraph discovery criteria avoids to choose the redundant genes and utilization of the notion of feature weighting results in the selection of appropriate genes.Then, from each subgraph, the relevant gene is selected by performing the feature weighting technique. The purpose of feature weighting is to select a representative gene that is most relevant to the cancer diagnosis task. Fisher Score (FS) is a supervised univariate filter which gives higher values to those genes that have a separation property. The Fisher Score of gene k is calculated as below:

$$FS(G_k) = \frac{\sum_{v \in V} n_v \left(\bar{G_k^v} - \bar{G_k}\right)^2}{n_v \left(\sigma_v (G_k)\right)^2} \tag{1}$$

where $\bar{G_k}$ is the average value of all the patterns related to the gene $G_k$, $V$ is a set of all classes in a dataset, $n_v$ is the size of samples on the class v, and $\sigma_v (G_k)$ and $\bar{G_k^v}$ indicates the variance and mean of gene $G_k$ on class v, respectively.

According to the algorithm design, a subgraph of genes is extracted in each iteration using a repetitive process, and then the appropriate genes are selected from the existing genes in this subgraph using a feature-weighting strategy. This process is repeated again for all remaining genes after deleting the genes present in the extracted subgraph.

## 4. Experimental results

To investigate the performance of our cancer diagnosis algorithm, various experiments are performed.The efficiency of our algorithm is compared with three new methods of filter-based

**Table 1**
Characteristics of the used microarray datasets

| Dataset | # Genes | # Classes | # Patterns |
|---|---|---|---|
| Colon | 2000 | 2 | 62 |
| SRBCT | 2328 | 4 | 83 |
| Leukemia | 7129 | 2 | 72 |
| Prostate Tumor | 10509 | 2 | 102 |
| Lung Cancer | 12600 | 5 | 203 |

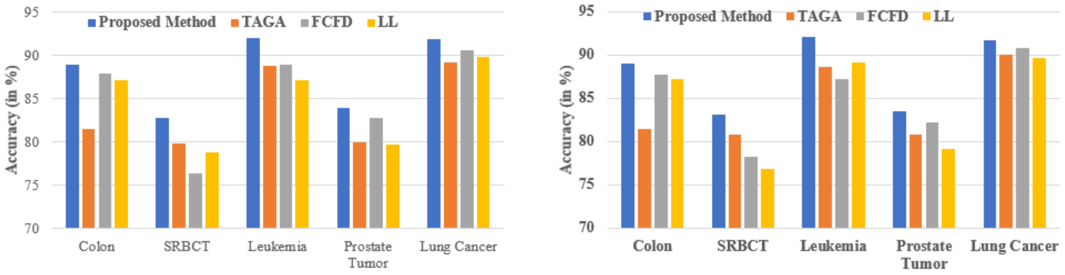

**Figure 1:** Average accuracy of different methods on (a) SVM classifiers (b) AB classifiers.

cancer diagnosis: Tabu Asexual Genetic Algorithm (TAGA) [12], Feature Clustering and Feature Discretization assisting gene selection (FCFD) [13], Least Loss (LL) [14]. Moreover, the experiments in this study use a variety of datasets with different properties to demonstrate the effectiveness of the developed algorithm. These microarray data include of Colon, Leukemia, SRBCT, Prostate Tumor, and Lung Cancer [6]. The primary characteristics of these datasets are detailed in Table 1. Additionally, to assess the flexibility of the proposed algorithm on different classifiers, we also examined the performance of two frequent used classifiers, including Support Vector Machine (SVM) and AdaBoost (AB).

In our experiments, the efficiency of our algorithm is measured using the mentioned classifiers. Figure  1 summarizes the average classification accuracy over ten separate and autonomous runs of the different gene selection algorithms. The reported results indicate that in all cancer datasets, the developed gene selection performs better than those of other alternative approaches. For example, for the Colon dataset on SVM classifier, the proposed algorithm obtained a 88.91% accuracy while for TAGA [12], FCFD [13], LL [14], this value is 81.52%, 87.82% and 87.24%, correspondingly. Furthermore, the reported results for AB classifier were similar to SVM classifier, and in all cases the developed algorithm was more precise than the other compared algorithms.

## 5. Conclusion

In this paper, an efficient dimensionality reduction algorithm in cancer diagnosis has been developed utilizing the subgraph discovery and feature weighting. The main aim of our algorithm is to choose a subset of appropriate and non-redundant genes that are most closely associated to the target class of microarray data classification. The proposed algorithm has been compared to the recent gene selection algorithms on the cancer microarray datasets. The experimental results show that our cancer diagnosis algorithm gained the highest performance.
In future work, our goals are (1) integrate our proposed gene selection approach along with deep learning techniques for accurate cancer diagnosis, and (2) study explainable artificial intelligence in detail to see how explainable artificial intelligence can improve further interpretability and transparency of diagnosis.

## Acknowledgments

This project is supported by the Academy of Finland Profi5 DigiHealth project, which is gratefully acknowledged.

## References

[1] S. Yang, H. Xiao, L. Cao, Recent advances in heat shock proteins in cancer diagnosis, prognosis, metabolism and treatment, Biomedicine & Pharmacotherapy 142 (2021) 112074.

[2] R. Daneshjou, B. He, D. Ouyang, J. Y. Zou, How to evaluate deep learning for cancer diagnostics – factors and recommendations, Biochimica et Biophysica Acta (BBA) - Reviews on Cancer 1875 (2021) 188515. doi:https://doi.org/10.1016/j.bbcan.2021.188515.

[3] Y. Liu, F. Nie, Q. Gao, X. Gao, J. Han, L. Shao, Flexible unsupervised feature extraction for image classification, Neural Networks 115 (2019) 65–71. doi:https://doi.org/10.1016/j.neunet.2019.03.008.

[4] S. Forouzandeh, K. Berahmand, M. Rostami, Presentation of a recommender system with ensemble learning and graph embedding: a case on movielens, Multimedia Tools and Applications 80 (2021) 7805–7832. URL: https://doi.org/10.1007/s11042-020-09949-5. doi:10.1007/s11042-020-09949-5.

[5] M. Rostami, K. Berahmand, E. Nasiri, S. Forouzandeh, Review of swarm intelligence-based feature selection methods, Engineering Applications of Artificial Intelligence 100 (2021) 104210. URL: https://www.sciencedirect.com/science/article/pii/S0952197621000579. doi:https://doi.org/10.1016/j.engappai.2021.104210.

[6] M. Rostami, S. Forouzandeh, K. Berahmand, M. Soltani, Integration of multi-objective pso based feature selection and node centrality for medical datasets, Genomics 112 (2020) 4370–4384. URL: https://www.sciencedirect.com/science/article/pii/S088875432030224X. doi:https://doi.org/10.1016/j.ygeno.2020.07.027.

[7] M. Rostami, K. Berahmand, S. Forouzandeh, A novel method of constrained feature selection by the measurement of pairwise constraints uncertainty, Journal of Big Data 7 (2020) 83. URL: https://doi.org/10.1186/s40537-020-00352-3. doi:10.1186/s40537-020-00352-3.

[8] M. Rostami, K. Berahmand, S. Forouzandeh, A novel community detection based genetic algorithm for feature selection, Journal of Big Data 8 (2021) 2. URL: https://doi.org/10.1186/s40537-020-00398-3. doi:10.1186/s40537-020-00398-3.

[9] M. Labani, P. Moradi, F. Ahmadizar, M. Jalili, A novel multivariate filter method for feature selection in text classification problems, Engineering Applications of Artificial Intelligence 70 (2018) 25–37. URL: https://www.sciencedirect.com/science/article/pii/S0952197617303172. doi:https://doi.org/10.1016/j.engappai.2017.12.014.

[10] W. De Smet, K. De Loof, P. De Vos, P. Dawyndt, B. De Baets, Filtering and ranking techniques for automated selection of high-quality 16s rrna gene sequences, Systematic and Applied Microbiology 36 (2013) 549–559. URL: https://www.sciencedirect.com/science/article/pii/S0723202013001495. doi:https://doi.org/10.1016/j.syapm.2013.09.001.

[11] C.-M. Lai, H.-P. Huang, A gene selection algorithm using simplified swarm optimization with multi-filter ensemble technique, Applied Soft Computing 100 (2021) 106994. URL: https://www.sciencedirect.com/science/article/pii/S1568494620309339. doi:https://doi.org/10.1016/j.asoc.2020.106994.

[12] S. Salesi, G. Cosma, M. Mavrovouniotis, Taga: Tabu asexual genetic algorithm embedded in a filter/filter feature selection approach for high-dimensional data, Information Sciences 565 (2021) 105–127. URL: https://www.sciencedirect.com/science/article/pii/S0020025521000475. doi:https://doi.org/10.1016/j.ins.2021.01.020.

[13] H.-Y. Lin, Feature clustering and feature discretization assisting gene selection for molecular classification using fuzzy c-means and expectation–maximization algorithm, The Journal of Supercomputing 77 (2021) 5381–5397. URL: https://doi.org/10.1007/s11227-020-03480-y. doi:10.1007/s11227-020-03480-y.

[14] F. Thabtah, F. Kamalov, S. Hammoud, S. R. Shahamiri, Least loss: A simplified filter method for feature selection, Information Sciences 534 (2020) 1–15. URL: https://www.sciencedirect.com/science/article/pii/S0020025520304242. doi:https://doi.org/10.1016/j.ins.2020.05.017.

[15] M. Rostami, P. Moradi, A clustering based genetic algorithm for feature selection, in: 2014 6th Conference on Information and Knowledge Technology (IKT), 2014, pp. 112–116. doi:10.1109/IKT.2014.7030343.

[16] P. García-Díaz, I. Sánchez-Berriel, J. A. Martínez-Rojas, A. M. Diez-Pascual, Unsupervised feature selection algorithm for multiclass cancer classification of gene expression rna-seq data, Genomics 112 (2020) 1916–1925. URL: https://www.sciencedirect.com/science/article/pii/S0888754319304811. doi:https://doi.org/10.1016/j.ygeno.2019.11.004.

[17] M. M. Kabir, M. Shahjahan, K. Murase, A new local search based hybrid genetic algorithm for feature selection, Neurocomputing 74 (2011) 2914–2928. URL: https://www.sciencedirect.com/science/article/pii/S0925231211002748. doi:https://doi.org/10.1016/j.neucom.2011.03.034.

[18] M. Bressan, Faster algorithms for counting subgraphs in sparse graphs, Algorithmica 83 (2021) 2578–2605. URL: https://doi.org/10.1007/s00453-021-00811-0. doi:10.1007/s00453-021-00811-0.