Jurnal Kajian
Komputasi Inovatif Vol.16 No.02, Feb 2025

DISCOVER GENE ASSOCIATION IN COLON CANCER DISEASE USING SUPPORT VECTOR MACHINE AND MULTI-OBJECTIVE OPTIMIZATION TECHNIQUE

I Gede Wahyu Surya Dharma[1], I Gede Karang Komala Putra[2]
Universitas Bali Internasional
E-mail: suryadharma@iikmpbali.ac.id[1],
igdkarang@iikmpbali.ac.id[2]

Abstract

gene expression to detect association of gene in cancer disease is still promising. Each gene have dominant and less dominant impact to construct a disease. To analyze the cause and prevent further spread of cancer tissue, gene disease association based on gene expression is needed to discover correlation between gene that cause cancer and does not have correlation to cancer. Commonly, to discover gene association in colon cancer disease is lack of patient and number of gene that scrambled between up-regulated and down-regulated to this disease. In this study, Support Vector Machine conducted to separate between mutated colorectal cancer (MCC) and normal patient. Recursive feature elimination (RFE) is conducted to rank and select most informative gene. Furthermore, multi-objective optimization technique named non dominated sorting genetic algorithm (NSGA-II) is applied to find minimal solution of gene selection. To validate classification of gene expression data, Support vector machine obtain 99.6% of accuracy, while precision reach 100% and 98.6% of recall.

Keywords *— Colon Cancer, NSGA-II, Support Vector Machine, Recursive Feature Elimination.*

1. INTRODUCTION

Cancer, a disease that is infecting many people, which is still difficult to recognize when it reaches at an advanced stages. Hence, early treatments is required to be able to detect the early development of cancer. Treatment itself can be early detection of genes that have a role in determining the cancer itself [1]. By using gene expression, a genes that associated with a disease will be able to recognize and find how close its relationship. To view the associated gene, a combination of software and experts (biologist, doctor) is needed [2].

In analytical stage, software is formed in order to make more in-depth analysis process. To get an in-depth analysis of genes association with a disease, a research that conducted by Martinez Ballesteros, M, et al [3], [4] was able to mine a rule that able to look into the genes that might potentially cause of Alzheimer's disease, where at the beginning of the study, they were using the yeast cell microarray data. In his prior research, they introduce multi-objective evolutionary algorithm named GarNet[4], this algorithm is able to mine quantitative associative rule that correlated with a disease.

In this research, support vector machine as classification method[1], aim to obtain initial threshold that classify between patient and control. Classification stage will be validate using 5-fold cross validation, in order to obtain classified dataset that truly valid and suitable. Moreover, NSGA-II[4] will be applied to mine the quantitative association rule due to limitation and close resources of GarNet itself.

However, although NSGA-II have slightly problem in weight objective scheme. In this research, several analysis and optimization that aim to overcome weight objective scheme will be applied. Regarding to literature review and analysis that have been done before, this research aimed at modify the method of determining initial threshold and also NSGA-II method optimization to overcome the weight objective scheme in fitness function.

2. METHOD

Colon cancer is one of common cancer disease which significantly increasing due to fast and dynamic lifestyle, estimated 93.090 newly diagnosis cases. Surgical remain the common treatment for colon cancer patient. However, treatment for colon cancer mostly lead to failure due to poor gene cancer prognosis [1]. To discover associate genes that correlated to colon cancer disease, this study is using dataset which taken from National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) dataset. Those colon cancer dataset are labelled with accession number of GSE17538 and GSE38832. GSE17538 contain 238 samples while GSE38832 contain 122 samples.

3. RESULT AND DISSCUSION

Dataset with high number of feature will lead to long and frustrated process time. Those dataset are containing 54000 gene expression features while respectively 122 samples and 238 samples. First of all, those dataset is reduced with condition that only up-regulated gene expression feature with value more than 7 and contain only more than 60%. Features that contain less than 60% of up-regulated genes is deleted. Finally, total of gene expression features is decreased into 13934. Furthermore, after dataset's gene expression features is decreased, 100 times support vector machine with 5 fold cross validation is applied on reduced dataset. This phase aim to find initial threshold that will used to compare with non-dominated sorted genetic algorithm (NSGA-II). From a hundred times of support vector machine with 5 fold cross validation is applied, it reached 99.6% of accuracy, while precision reach 100% and 98.6% of recall, and shown in figure 4.

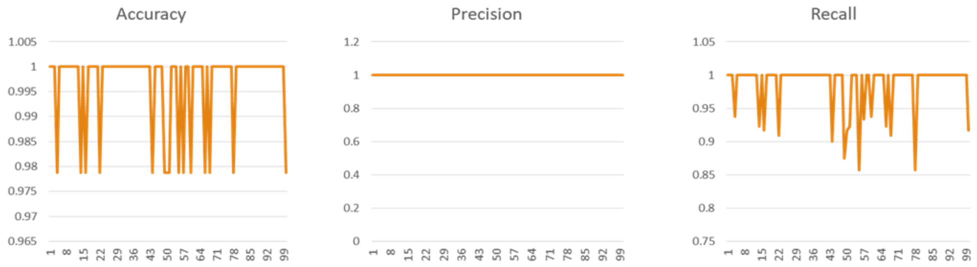

Figure 1.overall accuracy, precision, and recall of 100 times support vector machine with 5 fold cross validation

Next step, NSGA-II is applied. The NSGA-II in this study is using Kursawe's objective function (KUR)[5]. The reason, KUR able to separate region into k discontinuous region in the Pareto-optimal front. Due to ability to separate optimal solution into separated region, KUR able to separate genes expression into two region where each region contain each type of colon cancer disease.

In the optimal solution given by NSGA-II, on the x axis or f2(x) is a decision variable while on the y axis or f1(x) is a stepsize variable. Decision variable or stepsize variable is a vector matrix containing the pareto value generated through the sorting process using Kursawe optimation function. The result graph is a new objective generated from the dominant and recessive genotype of each gene expression feature[8]. Figure 4.2 shows the NSGA-II optimal solution obtained from gene expression features.

After first NSGA-II and 100 times SVM applied, initial threshold is obtained from SVM, and used for optimize the second run of NSGA-II.

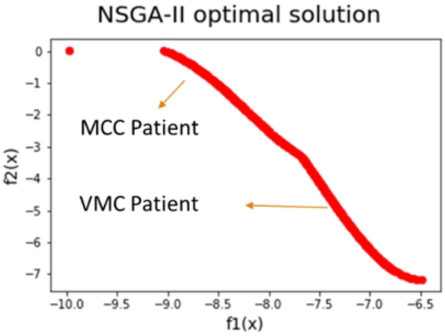

Figure 4-2. Optimal solution of NSGA-II

However, the result remain same with first run of NSGA-II, therefore no additional process need to be applied for NSGA-II to mining quantitative association rule from colon cancer gene expression. It shows that despite initial threshold is already initiated, optimal solution remain same, it is indicated to optimal solution that is reached since first run of NSGA-II.

Further step for this study is gene extraction and selection. To do gene expression feature extraction, recursive feature elimination extracted gene feature while indicated that these gene expression feature have supported among all features or not, additionally do an ascending sorting regarding to each feature ranking, from less significant into most significant. Later on, 100 top genes is selected to select only top up-regulated feature that correlated to colon cancer disease selected.

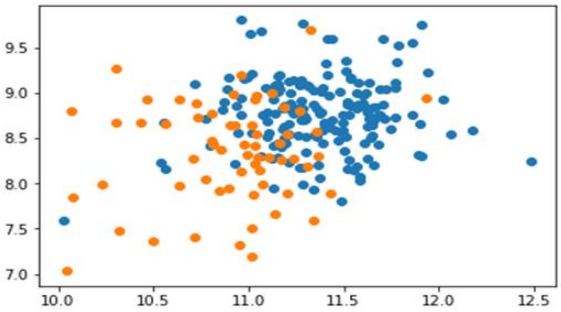

Figure 4-3. Clustering result of hierarchical clustering

Finally, to visualize the obtained genes, hierarchical cluster is selected as clustering method to cluster gene expression feature and sample, due to its specialty to cluster gene expression feature[3].

Using hierarchical clustering as a clustering method produces a clustered cluster in the middle with multiple outliers from the MCC and VMC patient classes. Patient MCC is denoted by orange color while patient VMC is denoted by blue color.

Testing step conducted between clustering results obtained by hierarchical clustering with the truth label reached a result of hierarchical clustering homogeneity value of 1 which means the cluster variance caused by hierarchical clustering is exactly same with the cluster results from the truth label, in addition to completeness and v-score was worth 1 which means v-score, the harmonic mean between homogeneity and completeness has exactly similar between hierarchical clustering and truth label. Hence, that the hierarchical clustering result becomes exactly the same as truth label.

Pearson correlation is clustering gene expression feature, and samples is clustered using Spearman correlation. In heat map, gene expression is located in row, while samples is located in column. In the heatmap, the lower left portion denoted with green dendrograms is the top 100 genes that have been sequenced in the previous gene selection process, while the red dendrogram-containing genes are the other genes that are less linked because after the sorting process has a ranking Above 100.

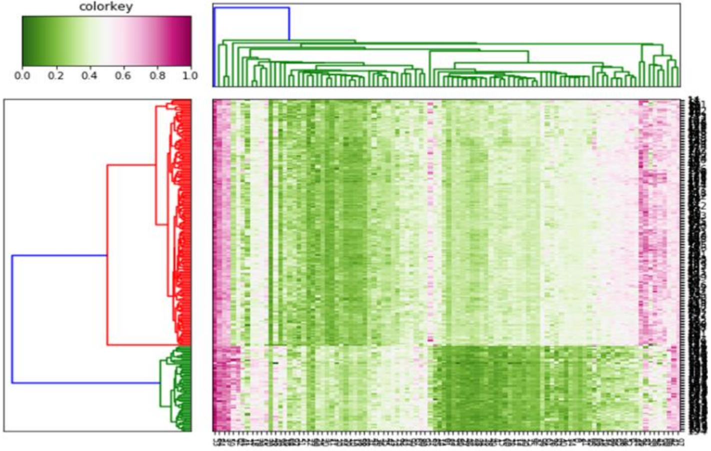

Figure 4-4. Heat map of top genes and hierarchical cluster for colon cancer disease

4. CONCLUSION

First run of non-dominated sorting genetic algorithm giving slightly similar optimal solution to second run of NSGA-II where the second NSGA-II is reconfigured using initial threshold obtained from a hundred times SVM with 5 fold cross validation. The SVM reach 99.6% of accuracy, while precision reach 100% and 98.6% of recall lead that optimal solution is already the best optimal solution. In addition, Recurrence feature elimination ranked and eliminate less significant gene expression features, moreover, a hundred of top gene expression feature is select to discover a hundred gene expression that correlated to colon cancer disease.

5. REFERENCES

C. Cortes and V. Vapnik, “Support-Vector Networks,” Mach. Learn., vol. 20, no. 3, pp. 273–297, 1995.

F. Kursawe, “A variant of evolution strategies for vector optimization,” in Parallel Problem Solving from Nature, Berlin/Heidelberg: Springer-Verlag, pp. 193–197.

H. Byun and S.-W. Lee, “a Survey on Pattern Recognition Applications of Support Vector Machines,” Int. J. Pattern Recognit. Artif. Intell., vol. 17, no. 3, pp. 459–486, 2003.

K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, “A fast and elitist multiobjective genetic algorithm: NSGA-II,” IEEE Trans. Evol. Comput., vol. 6, no. 2, pp. 182–197, 2002.

M. Martínez-Ballesteros, I. A. Nepomuceno-Chamorro, and J. C. Riquelme, “Discovering gene association networks by multi-objective evolutionary quantitative association rules,” J. Comput. Syst. Sci., vol. 80, no. 1, pp. 118–136, 2014.

M. Martinez-Ballesteros, J. M. Garcia-Heredia, I. A. Nepomuceno-Chamorro, and J. C. Riquelme-Santos, “Machine learning techniques to discover genes with potential prognosis role in Alzheimer’s disease using different biological sources,” Inf. Fusion, vol. 36, pp. 114–129, 2017.

S. Huband, P. Hingston, L. Barone, and L. While, “A review of multiobjective test problems and a scalable test problem toolkit," IEEE Trans. Evol. Comput., vol. 10, no. 5, pp. 477–506, Oct. 2006.

T. Latkowski and S. Osowski, "Gene selection in autism – Comparative study," Neurocomputing, vol. 0, pp. 1–5, Feb. 2017.

X. Qin et al., "A 15-gene signature for prediction of colon cancer recurrence and prognosis based on SVM," Behav. Brain Res., vol. 324, no. 1, pp. 33–40, 2017.