A Comparison of Feature Selection Algorithms for Cancer Classification Through Gene Expression Data: Leukemia Case

Aslı TAŞÇI [1], Türker İNCE[2], Cüneyt GÜZELİŞ [3]

[1] Department of Electrical and Electronics Engineering, İzmir Institute of Technology
aslitasci@iyte.edu.tr
[2]Department of Electrical and Electronics Engineering, İzmir University of Economics
turker.ince@ieu.edu.tr
[3]Department of Electrical and Electronics Engineering, Yaşar University
cuneyt.guzelis@yasar.edu.tr

Abstract

In this study, three different feature selection algorithms are compared using Support Vector Machines as classifier for cancer classification through gene expression data. The ability of feature selection algorithms to select an optimal gene subset for a cancer type is evaluated by the classification ability of selected genes. A publicly available micro array dataset is employed for gene expression values. Selected gene subsets were able to classify subtypes of the considered cancer type with high accuracies and showed that these feature selection methods were applicable for bio-marker gene selection.

1. Introduction

The number of patients diagnosed with cancer is increasing rapidly [1]. Currently, cancer diagnosis is practiced by using image processing techniques, blood analysis and biopsies. Cancer is caused by the accumulation of excessive amount of damaged cells. The behavior of the cancer differs from patient to patient and can only be explained by studying the origin. Cancer begins in the cell, and cell structure is unique to each individual. Therefore, there is not one specific drug, vaccine or treatment to cure cancer permanently for all cancer patients. A genetic approach to cancer is contributive in understanding the relation between gene and their products, identifying bio-marker genes for targeted drug therapies and the effects of genes on certain cell signaling pathways. Gene expression provides the information of how active a gene is. Micro array is one of the widely used measurement methods for gene expression. Gene expression values obtained by micro arrays can be employed in cancer diagnosis and the classification of cancer types [2]. Micro array datasets are employed for these purposes in many studies. Different feature selection algorithms are employed for the selection of bio-marker gene subsets. Statistical and machine learning classification methods are applied to micro array dataset with and without feature selection. Many studies were able to identify bio-marker genes for specific cancer types or classify cancer types with high accuracies [3], [4]. In this study, the aim is to classify and select the optimal gene subsets for two subtypes of leukemia (Acute Myeloid Leukemia and Acute Lymphoblastic Leukemia). Therefore, three filter type feature selection methods and Support Vector Machines are employed. Feature selection algorithms and classifier are represented in the second section of the paper. The experimental results obtained from the feature selection algorithms and classifier is represented in the third section.

2. Feature Selection Algorithms and Classifier

Optimal gene subsets or bio-marker genes for a specific type of cancers are important to understand the characteristics of cancer and produce alternative drugs. There are three main types of feature selection algorithms; filter, wrapper and ensemble type. Filter type feature selection algorithms consider the statistical relations in between features or feature to class to select a feature or a subset of features. Wrapper type feature selection algorithms select a feature or a subset of features up to a criterion and test these selected features with a learning algorithm. Then, determine the value of feature subset according to learning rate of the classifier. Ensemble feature selection algorithms are a combination of filter and wrapper types. A variety of feature selection algorithms are applicable to micro array datasets [5]. Non parametric tests, information theoretic approaches, probabilistic feature selection methods and genetic algorithms are practiced by many scientists to select the optimal genes [6-9]. Micro array datasets consist of small sample size but thousands of gene expression values as features. This characteristic of micro array data is a disadvantage in terms of data analysis but suitable for the application of filter type feature selection algorithms. Since, micro array datasets are high dimensional datasets, several feature selection algorithms are applied, filter based feature selection algorithms, frequently. In this study, ReliefF algorithm, Correlation based Feature Selection algorithm and t test statistic are employed as filter feature selection methods. Even though, several feature selection methods are employed in a more comprehensive study by the authors of this paper, results of the best three feature selection methods are represented in this paper [10].

2.1. ReliefF Algorithm

Relief is a weight-based feature selection algorithm. ReliefF algorithm is more accurate and applicable to the incomplete or multi class data version of the Relief algorithm. ReliefF algorithm finds k nearest neighbors to calculate the near-hit and near-miss values for the feature and makes an estimation using the average information k-nearest neighbors. ReliefF algorithm calculates a weight value to determine the significance of the

feature by using the equation 1. Then, each feature is ranked according to calculated weight [11].

$$S_i = S_i - diff\left(x_i, near - hit_i\right)^2 + diff\left(x_i, near - miss_i\right)^2 \quad (1)$$

2.2. Correlation Based Feature Selection Algorithm

Correlation based feature selection algorithm calculates a weight by using the equation 2 and ranks the features according to their class to feature or in between feature relations [12].

$$CFS = \frac{k\bar{r}_{cf}}{\sqrt{k + k\left(k-1\right)\bar{r}_{ff}}} \quad (2)$$

2.3. T test

T test statistic is a statistical value which defines the differences between two datasets. It is more suitable for the datasets which have normal distribution and independent [13]. In this study,absolute value of the t-test with pooled variance estimate used as ranking criterion. The t value for the t-statistic is calculated by equation 3.

$$t = \frac{\mu_1 - \mu_2}{s_p\sqrt{\frac{1}{N1} + \frac{1}{N_2}}} \quad (3)$$

where s_p is the pooled variance estimate calculated by equation 4.

$$s_p^2 = \frac{(N_1 - 1) * s_1^2 + (N_2 - 1) * s_2^2}{N_1 + N_2 - 2} \quad (4)$$

The adequacy of selected feature selection algorithms to select optimal gene subset for cancer diagnosis and classification is tested by employing Support Vector Machines for classification. Support vector machines are supervised learning algorithms and widely employed for text, image recognition and bioinformatics. They are applicable to linear or nonlinear datasets. In this study, LibSVM library is employed to perform classification with Support Vector Machines. LibSVM is a library which provides the application Support Vector Machine on several software platforms and available for public utilization [14].

3. Experimental Results

The performance of feature selection algorithms and classifier to classify two subtypes of leukemia is evaluated with respect to classification accuracy and the number of selected features. Even though high classification accuracy is important in the case of cancer classification, the number of genes used for classification is important as well. Most of the studies in this research area were able to classify cancer types using less than fifty genes as features for cancer type classification and cancer diagnosis with high accuracies. The micro array dataset used in this study is comprised of 7,129 gene expression values which are collected from 72 samples [15]. Experiments performed on both raw and pre-processed data. Feature selection algorithms were stable enough to choose same features for raw and pre-processed data and pre-processing had no effect on the feature selection. Therefore, only the results of raw data is represented here. K-fold cross validation is employed to divide the dataset into train and test sets with K being equal to ten. Gene expression values in the micro array dataset are ranked by using the feature selection algorithms. Top ranked hundred genes are selected to classify the types of leukemia. Classification is performed by taking chunks of genes from top ranked hundred genes subset and testing the trained support vector machine with several number of different gene subsets. The best classification accuracy and least number of genes is accepted as successful classification result.

Table 1. CLASSIFICATION ACCURACY OF T TEST STATISTIC FOR LEUKEMIA

Number of Features	Kernel Type	Classification Accuracy
1-10	Linear	100%
1-82	Polynomial	98.57%
1-20	Radial Basis Function	65.47%
1-10	Sigmoid	65.53%

Table 2. CLASSIFICATION ACCURACY OF RELIEFF ALGORITHM FOR LEUKEMIA

Number of Features	Kernel Type	Classification Accuracy
1-19	Linear	97.32%
1-55	Polynomial	96.25%
1-46	Radial Basis Function	65.53%
1	Sigmoid	65.47%

Table 3. CLASSIFICATION ACCURACY OF CORRELATION BASED FEATURE SELECTION ALGORITHM FOR LEUKEMIA

Number of Features	Kernel Type	Classification Accuracy
1-19	Linear	97.32%
1-55	Polynomial	96.25%
1-46	Radial Basis Function	65.53%
1	Sigmoid	65.47%

The employed feature selection algorithms performed with accuracies above 90%. Even though each algorithm performed with high classification accuracies, the best feature selection method for leukemia subtype classification is t test statistic. T test statistic reached a 100 % classification accuracy, using only ten selected genes. Furthermore, rankings of three feature selection algorithms show us 26 similar genes commonly selected by the algorithms and important for the classification of leukemia subtypes.

4. Discussions and Conclusion

In this study, optimal gene subset for cancer classification is obtained with three different feature selection algorithms and support vector machine is employed for the performance evaluation. Even though, classification accuracy and the number of selected genes for classification are the two main criteria for the performance. biological relevancy of the selected genes is important as well. Most commonly selected genes are researched to find any biological relation to cancer development or leukemia and several reference studies are found for most of the genes. Further, selected genes can be studied to find more biological connections between genes and leukemia. Figure 1 shows the gene expression values of most commonly selected genes by all feature selection algorithms. The classification ability of most commonly selected genes are represented clearly in the figure 1. CD33 has an important role in the cell cycle of myelomonocytic-derived cells and it can be seen that CD33 is more active (yellow color/ 0.6 mean gene expression value) for AML patients [16]. Similarly, transcription factor 3 (TCF3) has

an important role in the production of E protein [17]. This protein is essential for B and T lymphocyte development and this gene is more active for ALL patients as seen from the figure.

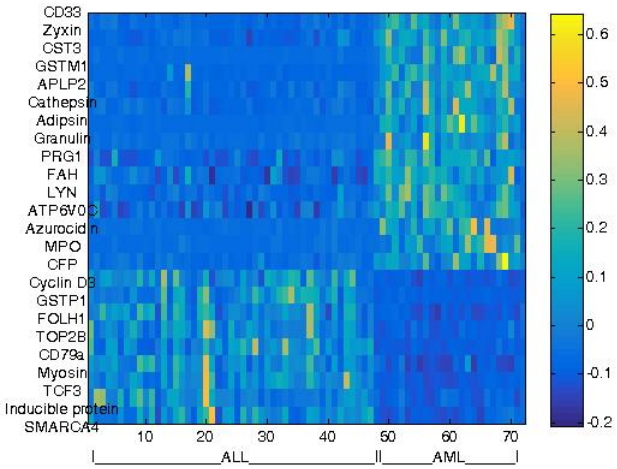

Figure 1. Most Commonly Selected Gene Expression Values

The proposed method in this study and experimental results conclude that filter feature selection methods might be useful for determining the set of relevant genes for a given type of cancer and improve generalization accuracy of the classifier. The commonly selected genes by the three feature selection algorithm includes CD33, Cathepsin D and Cyclin D3 genes. CD33 is active in the pathway for apoptosis of myelomonocytic-derived cells and further research of this gene may yield to targeted drugs for AML patients [18]. Cathepsin D and Cyclin D3 have a role in p53 pathway. Since p53 is a tumor suppressor gene and it is seen in more than 50% cancers, genes that affect the regulation of p53 is important to understand the cancerous cell dynamics [19-21]. Further research of all other selected genes' relation with leukemia may have great importance for the diagnosis, treatment and the prognosis phases of cancer.

5. References

[1] Stewart, B., & Wild, C. P. (2016). World cancer report 2014. World.

[2] Mikkilineni, V., Mitra, R. D., Merritt, J., DiTonno, J. R., Church, G. M., Ogunnaike, B., & Edwards, J. S. (2004). Digital quantitative measurements of gene expression. Biotechnology and bioengineering, 86(2), 117-124.

[3] Begum, S., Chakraborty, D., & Sarkar, R. (2016, January). Identifying cancer biomarkers from leukemia data using feature selection and supervised learning. In 2016 IEEE First International Conference on Control, Measurement and Instrumentation (CMI) (pp. 249-253). IEEE.

[4] Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine learning, 46(1-3), 389-422.

[5] Asyali, M. H., Colak, D., Demirkaya, O., & Inan, M. S. (2006). Gene expression profile classification: a review. Current Bioinformatics, 1(1), 55-73.

[6] Troyanskaya, O. G., Garber, M. E., Brown, P. O., Botstein, D., & Altman, R. B. (2002). Nonparametric methods for identifying differentially expressed genes in microarray data. Bioinformatics, 18(11), 1454-1461.

[7] Breitling, R., & Herzyk, P. (2005). Rank-based methods as a non-parametric alternative of the T-statistic for the analysis of biological microarray data. Journal of bioinformatics and computational biology, 3(05), 1171-1189.

[8] Jain, A. K., Duin, R. P. W., & Mao, J. (2000). Statistical pattern recognition: A review. IEEE Transactions on pattern analysis and machine intelligence, 22(1), 4-37.

[9] Bontempi, G., & Meyer, P. E. (2010). Causal filter selection in microarray data. In Proceedings of the 27th international conference on machine learning (icml-10) (pp. 95-102)

[10] Taşçı A. A comparative evaluation of feature selection algorithms for cancer classification through gene expression data. 10138688.

[11] Kononenko, I., Å imec, E., & Robnik-Å ikonja, M. (1997). Overcoming the myopia of inductive learning algorithms with RELIEFF. Applied Intelligence, 7(1), 39-55.

[12] Hall, M. A. (1999). Correlation-based feature selection for machine learning(Doctoral dissertation, The University of Waikato).

[13] Breitling, R., & Herzyk, P. (2005). Rank-based methods as a non-parametric alternative of the T-statistic for the analysis of biological microarray data. Journal of bioinformatics and computational biology, 3(05), 1171-1189.

[14] Chang, C. C., & Lin, C. J. (2011). LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST), 2(3), 27.

[15] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., ... & Bloomfield, C. D. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. science, 286(5439), 531-537.

[16] Malik, M., Chiles, J., Xi, H. S., Medway, C., Simpson, J., Potluri, S., ... & Crane, P. (2015). Genetics of CD33 in Alzheimer's disease and acute myeloid leukemia. Human molecular genetics, ddv092.

[17] National Center for Biotechnology Information, 2016 https://www.ncbi.nlm.nih.gov/gene/5199

[18] Marin, V., Pizzitola, I., Agostoni, V., Attianese, G. M. P. G., Finney, H., Lawson, A., ... & Biagi, E. (2010). Cytokine-induced killer cells for cell therapy of acute myeloid leukemia: improvement of their immune activity by expression of CD33-specific chimeric receptors. Haematologica, 95(12), 2144-2152.

[19] Wu, G. S., Saftig, P., Peters, C., & El-Deiry, W. S. (1998). Potential role for cathepsin D in p53-dependent tumor suppression and chemosensitivity. Oncogene, 16(17), 2177-2183.

[20] Sicinska, E., Aifantis, I., Le Cam, L., Swat, W., Borowski, C., Yu, Q., ... & Sicinski, P. (2003). Requirement for cyclin D3 in lymphocyte development and T cell leukemias. Cancer cell, 4(6), 451-461.

[21] Fimognari, C., Nüsse, M., Berti, F., Iori, R., Cantelli-Forti, G., & Hrelia, P. (2002). Cyclin D3 and p53 mediate sulforaphane-induced cell cycle delay and apoptosis in non-transformed human T lymphocytes. Cellular and molecular life sciences, 59(11), 2004-2012.