Cancer Classification using Fuzzy C-Means with Feature Selection

Arvan Aulia Rachman
Department of Mathematics
Universitas Indonesia
Depok, Indonesia
arvan.aulia@ui.ac.id

Zuherman Rustam
Department of Mathematics
Universitas Indonesia
Depok, Indonesia
rustam@ui.ac.id

***Abstract*— For many years, cancer classification to detect cancer at early stage of treatment has improved. Cancer classification is used for the treatment of cancer has entered the challenge to target specific therapy for each type of cancer pathogens in an effort to maximize efficacy and minimize toxicity. In general, cancer data consists of many features. However, not all of these features are informative. Therefore, among these features, Fisher's Ratio is applied to select the most informative features which form new data. Data on which feature selection has not been and has been performed are classified using Fuzzy C-Means. The experiment reveals that optimization which based on classification with feature selection increases the accuracy. Results show that, without doing feature selection, the accuracy is 82.92 % while with feature selection, the best accuracy is 89.68 % obtained by using 150 features. The results show the difference between all the dataset used and the dataset using feature selection.**

Keywords— classification; cancer; feature selection; Fisher's Ratio; Fuzzy C-Means

I. INTRODUCTION

Cancer is a disease caused by abnormal growth of cells of body tissues. During its development, the cancer cells can spread to other body parts. Cancer is one of the leading causes of death worldwide. In 2012, about 8.2 million deaths were caused by cancer. Furthermore, in Indonesia, about 1.4% of its population, or around 347,492 people were suffering from cancer in 2013 [1].

Cancer classification has been based primarily on morphological appearance of the tumor, but this has limitations. Tumors with similar appearance can follow different clinical courses significantly and show different responses to therapy. In a few cases, such clinical heterogeneity has been explained by dividing similar tumors morphologically into subtypes with distinct pathogeneses. Moreover, cancer classification has been difficult in part because it has historically relied on specific biological insights, rather than systematic and unbiased approaches for recognizing tumor subtypes. The challenge of cancer treatment is to target specific therapies to pathogenically distinct tumor types, and to maximize efficacy and minimize toxicity [2]. Cancer classification can help maximizing efficacy and minimizing toxicity of cancer treatment by specifying target.

Classification methods are used to classify data set $X = \{x_1, x_2, \dots, x_n\}, x_i \in \mathbb{R}^d, i = 1,2, \dots, n$ into several clusters. Data will be placed at the same class C_j if they have similar characteristics. One of the classification techniques is Vector Quantization. It maps the data $X = \{x_1, x_2, \dots, x_n\}$ into $V = \{v_1, v_2, \dots, v_n\} \subset \mathbb{R}^d$.V is called a medoid or signature. One of clustering techniques that based on Vector Quantization is Fuzzy C-Means [3].

In general, cancer data consists of many features. However, not all of these features are informative. Existing research proved that if all the features of the data used, there will be an impact on performance or accuracy of the classification results. In addition, number of features used is directly proportional to the time and cost of processing required. The more features used, the longer time and the greater cost of processing required. For that reason, feature selection needs to be used. Feature selection is selection method to reduce data set. This selection is crucial because there is uninformative part of the dataset which can reduce the accuracy of classification. In this paper, Fisher's Ratio is used as feature selection. Fisher's Ratio method is a technique that measures the strength of the features in the data set. Fisher's Ratio is obtained by looking at the ratio of the mean difference of the Class I (cancer) and class II (healthy) by variance of Class I and Class II. In this classification, Fuzzy C-Means is used. This method is the most stable clustering method, as the cluster center and the results of grouping do not change if there are new extreme data.

II. MATERIALS AND METHODS

A. Microarray Data

Colon cancer microarray data are used in this research. A gene expression data set for colon cancer obtained from [4] which is publicly available. The data consists of 7457 genes and each gene has 36 samples where 18 of them are healthy and the other 18 are cancerous [5].

978-1-5090-3385-0/16/$31.00 ©2016 IEEE

B. Feature Selection

Feature selection is crucial for cancer classification, as for each cancer type only a small number of genes are informative, and the presence of other genes reduces the classification accuracy [6]. The score of each feature that we get after feature selection examines the level of contribution of each gene to cause discrimination between cancerous (class I) and healthy (class II) tissues. In this study, the distributions of class I and class II is known. Fisher's Ratio method is used to choose the most informative features. The formula is shown as follows [7].

$$Fisher's\ Ratio\ (X_i) = \frac{(\mu_1 - \mu_2)}{(\sigma_1^2 - \sigma_2^2)} \quad (1)$$

where μ_i and σ_i are mean and variance of class *i* respectively. Fisher's ratio scores are calculated in all features where the lowest score is the best feature.

C. Classification

Fuzzy C-Means is development method from method of K-Means. Fuzzy C-Means was proposed by Dunn [8] and developed by Bezdek [8]. This method is included in Non-Hierarchical Clustering. Fuzzy C-Means is the most stable clustering method, because the cluster center and the results of grouping do not change if new extreme data appear. In general, the objective function of Fuzzy C-Means is

$$J_{FCM}(V, U, X, c, w) = \sum_{l=1}^{c}\sum_{k=1}^{N}(u_{lk})^w d_{lk}^2(X_k, V_l) \quad (2)$$

subject to,

$$\sum_{i=1}^{c} u_{ik} = 1, \forall k \in \{1,2,\dots,N\} \quad (3)$$

where V & U are two variables which find for optimal condition; the optimal conditions for the matrix signifies the convergence of group memberships in Fuzzy C-Means, X is matrix data in the cluster, c is number of clusters, w is fuzzy degree for grouping, d_{ik}^2 is the distance data to the centroid, calculated by

$$\|x_k - v_i\|^2 = (x_k - v_i)^T(x_k - v_i) \quad (4)$$

and V is matrix centroid.

Algorithm 1. Fuzzy C-Means

Step 1. Initialization

a. X with size $n \times m$, where n is number of dataset training, m is number of parameter.
b. Number of cluster, $c \geq 2$.
c. $w > 1$.
d. Max iteration N.
e. Iteration stop criterion
f. Initial value of Objective Function
g. Initial centroid.
Iteration begins $t = 1,2,\dots,N$

Step 2. Update degree of fuzzy membership for each data in each cluster (correct membership matrix) $u = [u_{ik}]$, $k = 1,2,\dots,n$

$$u_{ik} = \left[\sum_{j=1}^{c}\left(\frac{d_{ik}^2}{d_{jk}^2}\right)^{\frac{1}{w-1}}\right]^{-1} \quad (5)$$

where, $d_{ik}^2 = \|x_k - v_i\|^2 = (x_k - v_i)^T(x_k - v_i)$

Step 3. Calculate centroid V for each cluster i.

$$V_i = \frac{\sum_{k=1}^{n}(u_{ik})^w X_k}{\sum_{k=1}^{n}(u_{ik})^w} \quad (6)$$

Step 4. Calculate the value of Objective Function

$$J = \sum_{l=1}^{c}\sum_{k=1}^{N}(u_{lk})^w d_{lk}^2(X_k, V_l) \quad (7)$$

Step 5. Stopping criterion

$$\Delta = \|J^t - J^{t-1}\| \quad (8)$$

where J is the value of objective function.
If $\Delta < \varepsilon$, where ε is a small positive number, iteration stops. Else, go to step 2.

D. Performance Evaluation

In this section, the performance of the classifier is evaluated. The evaluation is carried out in the form of accuracy. There are 4 possible outcomes from the classifier. The first possibility is true positive (TP), which refers to the case that a diseased sample is correctly diagnosed. The second possibility is false positive (FP), which means that a healthy sample is incorrectly identified as a diseased case. The third possibility is true negative (TN), which indicates the case where a healthy sample is correctly spotted. Final possibility is false negative (FN), which refers to the case that diseased sample is incorrectly identified as healthy [10]. The percentage value for the evaluation criteria (accuracy) can be calculated using following formula, and the accuracy criteria (9).

$$Accuracy = \frac{n_{TP} + n_{TN}}{n_{TP} + n_{TN} + n_{FP} + n_{FN}} \times 100 \quad (9)$$

III. MAIN RESULT

A. Data

The data consist of 7457 genes and each gene has 36 samples where 18 of them are healthy and the other 18 are cancerous. The row of data explains the name of features and

their expression. The column is a sample of cancer and normal where T1-T18 are cancer samples and N19-N36 are normal samples. Look at Table I.

B. Feature Selection

Feature selection is conducted by using Fisher's Ratio. The result is obtained by Fisher's Ratio's value of each feature where the feature with the lowest Fisher's Ratio's value is the best feature. Furthermore, the data is sorted based on the value of Fisher's Ratio.

C. Classification

In this process, first, we make a new data with 10, 15, 25, 50, 75, 100, 150, 200,500,1000 genes selected and classify each data. After that we do classification without feature selection or all genes used. The result is summarized in Table 2. On each experiment, a percentage of p from the data, where $p = 10,20,\dots,90$ is used as a training data and the remaining $(100 - p)\%$ is used as testing data. The process is repeated 10 times.

D. Performance Evaluation

In this section, accuracy of classification is calculated. Table II shows the average accuracy using a various percentage p of training data. Then, the classification results with and without feature selection are compared. Without feature selection, the accuracy is 82.92%. With feature selection, the accuracy increased and decreased based on a number of features used.

TABLE I. DATA EXAMPLE

No	Gene	T1	T2	…	T18	N19	N20	…	N36
1	"Human liver cytochrome P-450 mRNA, complete cds" (Fitur ke-1)	1.9902	7.1735	…	-0.4804	5.9931	0.7386	…	-0.7502
2	"Human liver cytochrome P-450 mRNA, complete cds" (Fitur ke-2)	-5.4731	1.4715	…	-0.9608	21.7931	2.0311	…	14.2543
…	…	…	…	…	…	…	…	…	…
7457	H.sapiens mRNA for hBD-1 (Fitur ke-7457)	-3.1125	0.1665	…	-2.3212	0.6582	-2.0527	…	-3.2814

TABLE II. ACCURACY

Number of Features	Accuracy
All (without Feature Selection)	82.92194
10	74.83052
15	75.55994
25	83.64533
50	83.22533
75	86.6189
100	87.12253
150	**89.67763**
200	83.20999
500	76.19364
1000	79.44136

IV. CONCLUSION

Accuracy results show that feature selection can improve the accuracy of classification compared to the cases with no feature selection. Furthermore, there is no rule to determine minimum number of features needed to do classification. The number of features used directly proportional to the time and cost of processing required. The best result when you used this data is to use 150 features.

References

[1] *InfoDATIN*, Indonesian Health Ministry, Jakarta: Indonesian Health Ministry, 2015.

[2] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439,1999, pp. 531–538.

[3] Z. Rustam and A. S. Talita, "Fuzzy kernel k-medoids algorithm for multiclass multidimensional data classification," *J. Theoretical Appl. Inform. Technology*, vol. 80, no. 1, 2015, pp. 1817-3195.

[4] Notterman, et al., "Princeton University Gene Expression Project," [Online]. Available: http://genomics-pubs.princeton .edu/oncology/

[5] V. Elyasigomari, M. S. Mirjafari, H. R. C. Screen, M. H. Shaheed, "Cancer classification using a novel gene selection approach by meansof shuffling based on data clustering with optimization," *Appl. Soft Computing*, 2015.

[6] D. B. Allison, X. Q. Cui, C. P. Page, M. Sabripour, "Microarray data analysis:from disarray to consolidation and consensus," *Nat. Rev. Genet.*, vol. 7, no. 1, 2006, pp. 55–65.

[7] A. Idris, et al., "Churn prediction in telecom using random forest and PSO based data balancing in combination with various feature selection strategies," *Comput. Elect. Eng*., 2012, pp. 1808-1819

[6] J. C. Dunn, "*A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters*," *J. Cybernetics*, vol. 3, 2012, pp. 32-57

[7] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algoritms.* New York: Plenum Press, 1981.

[8] A. Rahideh, M. H. Shaheed, "Cancer classification using clustering based gene selection and artificial neural networks," *in Int. Conf. Control, Instrumentation and Automation (ICCIA)*, Shiraz, 2011, pp. 1175–1180.