CANCER CLASSIFICATION AND PREDICTION USING MULTIVARIATE ANALYSIS

Ho Sun Shon, Heon Gyu Lee, Keun Ho Ryu
Database/Bioinformatics Laboratory, Chungbuk National University
{shon0621, hglee, khryu}@dblab.chungbuk.ac.kr

ABSTRACT...Cancer is one of the major causes of death; however, the survival rate can be increased if discovered at an early stage for timely treatment. According to the statistics of the World Health Organization of 2002, breast cancer was the most prevalent cancer for all cancers occurring in women worldwide, and it account for 16.8% of entire cancers inflicting Korean women today. In order to classify the type of breast cancer whether it is benign or malignant, this study was conducted with the use of the discriminant analysis and the decision tree of data mining with the breast cancer data disclosed on the web. The discriminant analysis is a statistical method to seek certain discriminant criteria and discriminant function to separate the population groups on the basis of observation values obtained from two or more population groups, and use the values obtained to allow the existing observation value to the population group thereto. The decision tree analyzes the record of data collected in the part to show it with the pattern existing in between them, namely, the combination of attribute for the characteristics of each class and make the classification model tree. Through this type of analysis, it may obtain the systematic information on the factors that cause the breast cancer in advance and prevent the risk of recurrence after the surgery.

KEY WORDS: Classification, Discriminant Analysis, Decision Tree, Breast Cancer

1. INTRODUCTION

According to the WHO'S statistic in 2002, breast cancer took the first place among women's cancers outbreak frequency all over the world and it has become an issue in the women's health. In our country, breast cancer took the first place among women's cancers in 2001, after then, it has rapidly increased 40.5persons per 100,000 and now it has taken 16.8% of total women's cancers. The reason why breast cancer has increased so fast like this is Westernization of lifestyle that makes to take animal fat. The animal fat makes female hormone which mainly causes breast cancer and there are early menarche, a declining birth-rate, breast-feeding evasion which make hormone open. It makes breast cancer outbreak-out high. In fact, there have been strengthening primary factors now like Korean women's daily calorie intake is 3000 kcal(2500 kcal in 1981), the birth-rate is 1.19ersons (2.0 in 1980), the age of marriage is 27.3(24. 9 in 1990), and the age of menarche is 12.7(13.5 in 1988) [1].

However, it is known that breast cancer patients do not prepare well after operations and also receive suitable information, either [9]. It could also cause a lot of problems like a mental burden, hard cure process, and so on, so it is necessary to have systematic information on primary factors which can be a cause of breast cancer in advance and to prevent the risk of return with proper preparation after operation so that one should keep healthy life.

Therefore, in this paper, in order to classify the cancer if it is benign or malignant by using breast cancer data, it used discriminant analysis which is a multivariate statistic technique and decision tree which is a method of data mining. Having done that, it compared and analyzed the precision of two.

2. RELATED WORK

Discriminant analysis is a statistic method that finds out some discriminant standard or discriminant function which is able to separate the populations the best by observation value gotten from the populations of two or more and then allots the observation value to these populations. From the classified table classified by this procedure, it evaluates the error rate of the discriminant function and goes through the process that classifies new individuals form the unknown its group into some population. Therefore, the discriminant analysis searches for the discriminant function and classifies not only established individuals, but also new unknown individuals into a certain population by discriminant function [2].

The purposes of this discriminant analysis are as followed. First, it shows the individuals from a few of well-known populations on the low dimension below 3-dimension graph or it decides discriminant rules such as discriminant standard or function by logarithmic method. Second, when it focuses on classifying individuals into two or more groups, the purpose is the lead of the optimal rule in order to allot new individuals into classification population. In that case, it is expressed classification or allocation.

Therefore, in this paper, the discriminant analysis for two groups has been used to discriminate if the cancer is benign or malignant by using breast cancer data.

Decision tree method is very widely used in relation to data mining. It is also closely related to divisive community analysis which is one of data mining methods. Every individual is belonged to a united group at first, and then it is divided up into 2 groups by some variable value. For example, if the variable value is the certain or more, it goes to group1, otherwise it goes group2. And then each of two groups will be divided up into two by the second variable value. It will continue to repeat these processes until it is satisfied with the suitable ending standard. Variables would not be mattered if they are in an order category or not [4].

The advantages of this decision tress analysis method are that it forms easy rules to understand and classifying process without lots of computing work and it can use both sequence variables and categorical-typed variables. However, it does binary division, so it has too long divided branches and it costs too much when it forms a tree. There is another problem that it should consider the both cases of pruning and the time of setting up the dividing standard.

3. DATA PREPROCESSING

This chapter describes the pre-process in order to classify Wisconsin breast cancer data. The attributes of data is shown [Table1]. Preprocessing stage for classification used entropy based measure in order to divide sequenced data into numerical attribute values [5].

Table.1. Data Attribute

Data	Attribute
Class	No-recurrence-events Recurrence-events
Attribute	Age, Menopause, Tumor-size, Inv-nodes, Node-caps, Deg-malig, Breast, Breast-quad, Irradiate

The result of entropy based measure is discretization. The numerical concept stratum of the attributes is formed through this discretization. When S, the group of datatuple, is given, the basic method for the entropy based discetization is as followed.

First, the value A can be regarded as potential section critical value T. For instance, some value v of A can be divided up into two subgroups which satisfy the samples in S with the conditions, A < v and A ≧ v. From that, binary discretization will be created. Second, the critical value chosen about S given is the value that maximizes information gains which are happened from the succeeding division. Information gains can be achieved by formula 1 followed and S_{1a} and S_2 in S are the samples that satisfy each condition A < T and A ≧ T.

$$E(A,T;S) = \frac{|S_1|}{|S|} Ent(S_1) + \frac{|S_2|}{|S|} Ent(S_2) \quad (1)$$

Ent, Entropy function for the given group, is calculated in the base of class distribution of the samples in the group. For example, entropy formula of S_1 for given m pieces of classes is the same as equation 2. p_i is the probability of class i, it decides by dividing up the numbers of samples of class i into total number of samples in S_1. Ent(S_2) can be calculated similarly.

$$Ent(S_1) = -\sum_{i=1}^{m} p_i \log_2(p_i) \quad (2)$$

The process which decides critical value applies to each division by itself until some stopping rule meets the satisfaction like equation 3 [5].

$$Ent(S) - E(A,R;S) > \delta \quad (3)$$

Therefore, entropy based information is able to reduce the size of data and uses class information. So, breast cancer data has been discredited by using entropy based measure. Figure 1 followed is the results through pre-process of sequenced variables.

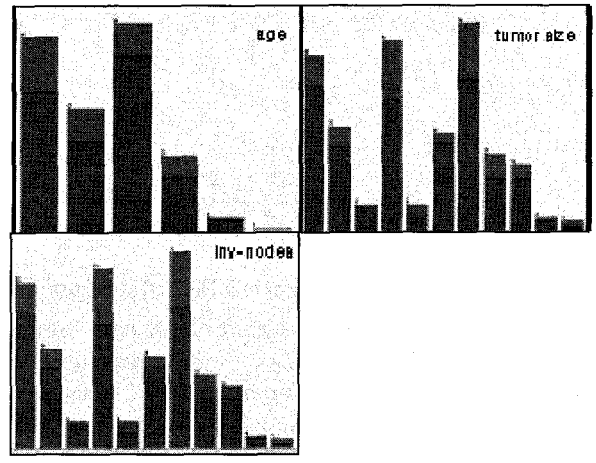

Figure.1. Discretization of Continuous variables

4. BREAST CANCER DATA CLASSIFICATION

To classify data after preprocess, discriminant analysis, a multivariate statistic analysis method, and decision tree, a data mining method, were used to classify.

4.1 Discriminant Analysis

In order to classify two populations, standard average vector $\overline{X}_i$ of observation vector X_{ij}(i = 1, 2: j = 1, 2,..., N_j) which is belonged to G_1, G_2 that are two of subgroups with sizes of N_1, N_2, and a sample covariance matrix S_i are each declared as equation 4 and equation 5.

$$\overline{X}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} X_{ij} \quad (4)$$

$$S_i = \frac{1}{N_i - 1}\sum_{j=1}^{N_i}(X_{ij} - \overline{X}_i)(X_{ij} - \overline{X}_j)' \quad (5)$$

When it happens, Fisher's discriminant method is that it linear-transforms equation 6 followed instead of discriminant variable vector $X = (X_1, X_2, \cdots, X_p)'$ and separates two groups by change variable Y.

$$Y = dX = d_1X_1 + d_2X_2 + \cdots + d_pX_p \quad (6)$$

From here, if equation 6 above is partly linear-transformed the *jth* observation vector $X_{ij} = (X_{ij_1}, X_{ij_2}, \cdots, X_{ijp})'$ which is belonged to subgroup G_1, then next formula 7 can be adjusted.

$$Y_{ij} = dX = d_1X_{ij1} + d_2X_{ij2} + \cdots + d_pX_{ijp} \quad (7)$$

On the assumption that the variance of the variable Y of between two groups is the same, 2 samples t-test statistic of two changed subgroups is equation 8 followed.

$$t = \frac{d(X_1 - X_2)}{[dS_p d(\frac{1}{N_1} + \frac{1}{N_2})]^{\frac{1}{2}}} \quad (8)$$

$$where\ ,\ S_p = \frac{(N_1 - 1)S_1 + (N_2 - 1)S_2}{N_1 + N_2 - 2}$$

Fisher's discriminant function is a linear-transformation that maximizes t^2 which is squared of standard distance between sample averages of two subgroups for transformed Y values, and it can be achieved by using the formula followed. After transforming equation 9 followed, it is called canonical discriminant function. Also, $dX = (d_1X_1 + d_2X_2 + \cdots + d_pX_p)$ which used this is the canonical discriminant coefficient.

$$d = S_p^{-1}(\overline{X}_1 - \overline{X}_2) \quad (9)$$

The maximum value of t^2 calculated by using the canonical discriminant coefficient will be the same as Hotelling T^2 statistic like formula 10.

$$(\frac{1}{N_1} + \frac{1}{N_2})(\overline{X}_1 - \overline{X}_2)'S_p^{-1}(\overline{X}_1 - \overline{X}_2) \quad (10)$$

If the sample averages of two transformed subgroups are $d\overline{X}_1, d\overline{X}_2$, new individual X_0 with no group information will be classified subgroup G_1 or G_2 by equation 11 and equation 12 as followed.

$$dX_0 > \frac{d\overline{X}_1 + d\overline{X}_2}{2} \quad (11)$$

If X_0 satisfies equation 11, it classifies it subgroup G_1.

$$dX_0 \leq \frac{d\overline{X}_1 + d\overline{X}_2}{2} \quad (12)$$

Also if X_0 satisfies equation 12, it classifies it subgroup G_2.

4.2 Decision Tree

Decision tree, one of data mining methods, analyzes data record collected before and shows the pattern, that is to say the attribute of each class, among them as the compounding of attributes. And then it makes a classification model as a tree type. There are various methods to operate decision tree and some other things like dividing standard, stopping rules, and pruning proposed. And a different decision tree method will be made how to unite these. If you look at decision tree algorithm, a tree begins with a single node of training sample. If the samples are from all the same class, a node will be a leaf and it classifies labels to corresponded class. Otherwise, it uses heuristic method to choose the attribute which can classify samples to each class the best by using entropy based measure, information gain. This attribute will be test or decision attribute in the node and it supposes that all attributes have a discrete value. A test attribute owned will be generated for each value, and samples will be divided properly. In order to form a decision tree for the divided samples, the same process will be repeated reflexively and if each attribute happens in one node, it should not be considered child node of the node. The most general decision tree algorithm is CART that looks for first dividing standard and perfect tree next and then decides pruning and candidate sub-tree after measuring off the error rate of each node. CART does binary division, but C4.5 is able to vary the numbers of branch. Also, C4.5 treats categorized division character and it does not use test group when it evaluates sub-tree. On the contrary to pruning of CART, it checks the error rate at the very end branch and does pruning to get the minimum error rate[7][8]. C4.5 algorithm has been used in this paper.

5. EXPERIMENTS AND EVALUATION

The contents of experiment on the algorithm suggested in this paper are as followed.

The data used in this experiment is from breast cancer data of Wisconsin university. There were 85 of breast

cancer patients with return and 200 of ones with no return. They were analyzed by using SPSS statistic package and WEKA mining tool as experimental tools.

The classifying result of discriminant analysis using SPSS is summarized as [Table2] and we can see that 140 cases out of 200 cases from the group with no return had been predicted right. That is, it shows 70 percents of precision and the precision is 65.9% when there is a return. So, when there is no return, the hit rate is higher. Therefore, general precision is 68.8% as the result of classification and 31.2% has not been classified right.

Table.2. Classification Result of Discriminant Analysis

recurrence		Prediction Group		Total
		No	Yes	
Source frequency	No	140	60	200
	Yes	29	56	85
%	No	70.0	30.0	100.0
	Yes	34.1	65.9	100.0

Table.3. Classification Result I of Decision Tree

Precision	Recall	F-Measure	Class
0.923	0.973	0.947	No
0.864	0.679	0.760	Yes

Table.4. Classification Result II of Decision Tree

recurrence	Positive	Negative
True	215	6
False	18	38

Table 3 show the result of C4.5 which is a decision tree algorithm using WEKA. F-Measure which uses the result above as an evaluating method by giving the same importance to Precision and Recall uses equation 13 followed.

$$F = \frac{2 * \Pr ecision * \mathrm{Re}\, call}{\Pr ecision + \mathrm{Re}\, call} \qquad (13)$$

In order to evaluate the performance of classification, F-Measure uses harmonic mean which is united Precision and Recall. Harmonic mean must be smaller the arithmetic mean if two groups have different values, it is smaller than arithmetic mean of group1 and group2. When the groups increase, F-Measure which was got from coupling of two groups can be gotten from coupling two groups again. And also Precision and Recall have relation to reciprocal proportion. We can get Precision and Recall by using Table 4. Precision can be gotten easily, but recall can be gotten when we know the related information in advance. The value of Recall was calculated bigger than the value of Precision because the value of True-Negative is small in Table 4.

6. CONCLUSIONS

In order to classify if the cancer is benign or malignant, I have operated pre-process and a regular process by using breast cancer data. To sort data, I have used discriminant analysis which is a multivariate statistic analysis and decision tree which is a data mining method for the experiment. As the result, discriminant analysis shows 70% of precision for the group with no actual return and the precision shows 65.9% of hit rate in the case of return. Therefore, the total precision of classifying result is 68.8%. Decision tree shows that F-Measure using Precision and Recall classified better in the patients group with no potential return. Also it says that Recall is calculated bigger than Precision.

If we construct database with this classification result, it can be used as basic data when computer-aided medical diagnosis system is designed or implemented. The exact classification is needed to prevent return after operation of breast cancer patient. If we can extract many attributes that affect, it will be good to cure patients a lot.

REFERENCE

[1] http://www.medcity.com

[2] Krzanowski, W.J., 1972. "The Performance of Fisher's Linear Discriminant Function under Non-Optimal Condition", Technometrics, 19(2), pp. 191-200.

[3] Lachenbruch, P.A., 1975. *Discriminamt Analysis*, New York, Hafner Press.

[4] Breiman,L., Freidman, J., Olshen, R., and Stone, C., 1984, *Classification and Regression Tree*, Belmont, California, Wardworth, Inc.

[5] Jiawei, H., Micheline, K., 2001, *Data Mining Concepts and Techiques*, Morgan Kaufmann,

[6] Murray, G.D., 1997, "*A Gautionary Note on Selection of Variables in Discreminant Analysis*", Applied statistics, 26(3), pp. 246-250.

[7] Kim, H., Loh W.Y., 2001, "*Classification trees with unbiased multiway splits*", JASA, 96(456), pp. 589-604.

[8] Kim, H., Loh W.Y., 2001, "*Classification trees with bivariate linear discriminant node models*", 2003, JASA 12, pp.512-530.

[9] Galloway, S., Graydon, J., Harrison, D., Evans-Boyden, B., Palmer-Wickham, S., Burlein-Hall, S., et. al., 1997, "Informational needs of women with a recent diagnosis of breast cancer: Development and initial testing of a tool", Journal of Advanced Nursing, 25, pp. 1175-1183.

ACKNOWLEDGEMENTS

This work was supported by the Regional Research Centers Program of Ministry of Education & Human Resources Development in Korea.