International Journal of Advanced Smart Convergence Vol.2 No.2 36-39 (2013)
DOI 10.7236/IJASC2013.2.2.8

IJASC 13-2-8

Analysis of Feature Variables for Breast Cancer Diagnosis

Yong Gyu Jung[1], Jang Il Kim[2], Sung Chul Sihn[3], Jun Heo[4]

[1,2]Department of Medical IT Marketing, Eulji University, Korea
ygjung@eulji.ac.kr, gold@goodit.co.kr
[3]Health Care Solutions Division, Fujitsu Korea Co. Ltd., Korea
scsihn@kr.fujitsu.com
[4]Dept. of Information and Communication, Kyungmin University, Korea
heojun@kyungmin.ac.kr (corresponding author)

Abstract

It is becoming more important as the growing of health information and increasing in cancer patients diagnose over the time gradually. Among the various types of cancer, we focuses on breast cancer diagnosis. The accuracy of breast cancer diagnosis is increasing when the diagnosis is based on evidence and statistics. To do this we use the weka data mining tools and analysis algorithms significantly associated with the decision tree uses rules. In addition, the data pre-processing and cross-validation are used to increase the reliability of the results. The number and cause of the disease becomes important to increase evidence-based medical doctors. As the evidence-based medical, the data obtained from patients in the past through the disease by calculating the probability for future patients to diagnose and predict disease and treatment plan. It can be found by improving the survival rate plays an important role.

Key words : diagnostic criteria, Breast-cancer, Cervical Cancer, 10-fold Validation

1. INTRODUCTION

According to recent studies, the Korean incidence of breast cancer was resulted the first rank in the world. Korea also increased by average 7% yearly while the growth rate decreased in high-incidence countries, such as the United States and Europe. The growth rate of Korea was ranked to 91% of the first during the onset of the OECD 34 member states. Breast cancer is now emerging as another serious disease. Due to the lack of data about currently known breast cancer diagnosis, even women were not significantly interested in breast cancer.

When the disease was discovered early saying that progression of cancer to be detected early, treatment can be most desirable in terms. The early detection and cure rates increase. If early detection of breast cancer, 5-year survival rate can be increased up to 95%, but the Quaternary as a 5-year survival rate is reduced 20%. In addition, early detection of medical technology during recent hospitalization or general anesthesia, with the development of breast tumors even without treatment unscathed is possible. Therefore, we use information from the experimental results to calculate the prior probability of breast cancer for each property values to obtain the bayesian posterior probability theory to predict the outcome of the diagnosis can help in early detection. In this paper, female cancer incidence data for breast cancer using publicly influential properties in breast cancer can be found. Using bayesian theory, which values each property will be found for the cancer diagnosis by obtaining the posterior probability values can predict the outcome.

Manuscript received: Sept. 26, 2013 / revised : Nov. 20, 2013
Corresponding Author: scsihn@kr.fujitsu.com
Tel:+82-2-3787-6000, Fax: +82-2-3787-6071
Health Care Solutions Division, Fujitsu Korea Co. Ltd., Korea

2. NURTURE GAMES

2.1 Decision Tree

When it is faced with an uncertain future and events, decision-making and combining different kinds of results are used. This situation shows that the shapes of bifurcation called a decision tree, which determine the bifurcation points of the branches is determined and uncertain history that two pronged fork is called the uncertainty. Complexity to the problem of decision-making and uncertainty in the sense that such a decision has many branches parted, and a few steps across the junctions that connect branch can be expressed as the picture. Each point of uncertainty for each branch point in history thought of probability of occurring. It is uncertain that each of the expected benefits and expected costs in the two calculated. A decision tree decision and uncertainty can be represented as a chain of thought and decision-making are used in the analysis of the problem.

2.2 Association Rules

Association rules are expressed as means useful patterns, which is conditions between data items. It is the form of purchasing behavior of customers to be analyzed. Association rules are exploration and selection of the appropriate set, which is the item X and Y can be seen as a problem. It is considering a few measures. First, a set of items X and support for the rule R (support) is defined as follows, respectively.

supp (X) = X of the items at the same time, which includes a set of total number of transactions (n) ratio for the

supp (R) = supp (X ∪ Y)

In other words, the set of rules, support for R set of X or Y to include the item in the same time represents the ratio of the number of transactions.

3. EXPERIMENTS

3.1 Data set

Experimental data are provided by the breast cancer Wisconsin data in UCI repository. They are used for the public obtained from the Wisconsin hospitals, which has 699 properties with 11 pieces of data. In this paper, Variables used in this paper the characteristics of the data, as shown in Table 1 has a total of 11 properties. It is in the case of breast cancer cells to deliver proteins accounted for a large proportion. However, if these cells into cancer cells spread to facilitate the delivery of proteins made nor can know about the spread of cancer cells. Thus, associated with cell thickness, size, shape, adhesion, etc. under the various properties, such as breast cancer is found, the probability can be obtained.

Table 1. Attribute of breast cancer dataset

Attribute	Definition
Sample code number	Id value that identifies
Clump Thickness	Thickness of cell aggregation
Uniformity of Cell Size	Similarity in cell size
Uniformity of Cell Shape	Similarity of cell shape
Marginal Adhesion	Partial degree of adhesion
Single Epithelial Cell Size	Single epithelial cell size
Bare Nuclei	Exposure of the nucleus
Bland Chromatin	Bland chromatin
Normal Nucleoli	Normal nuclear
Mitoses	Mitosis
Class	2(benign), 4(malignant)

3.2 Preprocessing

Data mining technologies usually are based on large amounts of data leading to dozens of gigabytes to tens of megabytes. This can have a huge data, but consumes a lot of time. By performing sampling and selection process, It changed to resembling a small amount of data from vast amounts of population extracted.

3.3 Cross-validation

Each attribute is repeated with 10-fold cross validation. As the result of performing, the following error rate are shown in Figure 1.

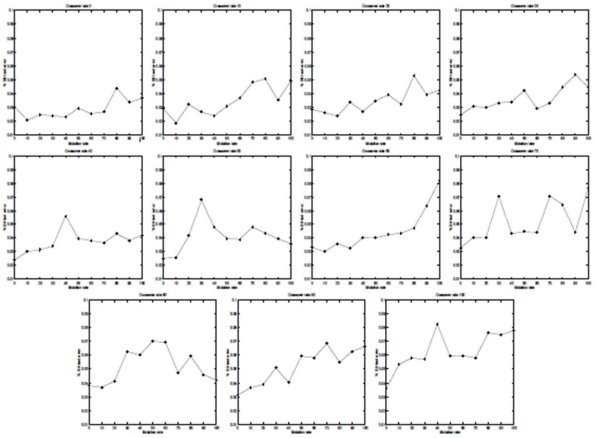

Figure 1 10-fold cross validation results

4. EXPERIMENTAL RESULT

In the first, experiment data mining techniques to calculate the probability of class 2 and class 4 were to detect

differences. The class 2 levels of all property values were higher than class 4.

Table 2. No. of cases by attribute values

attribute	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape
class : 2	27	21	31
class : 4	125	100	92
total	152	121	123
attribute	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei
class : 2	15	15	27
class : 4	83	51	159
total	98	66	186
attribute	Bland Chromatin	Normal Nucleoli	Mitoses
class : 2	73	21	10
class : 4	59	96	21
total	132	117	31

Total of 11 variables on the properties, except for the two remaining nine property values in the breast-cancer properties that have the greatest effect on the probability of to find the values were obtained.

Table 3. Posterior probabilities by attribute values

attribute	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape
class : 2	0.18	0.17	0.25
class : 4	0.82	0.83	0.75
attribute	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei
class : 2	0.15	0.23	0.15
class : 4	0.85	0.77	0.85
attribute	Bland Chromatin	Normal Nucleoli	Mitoses
class : 2	0.55	0.18	0.32
class : 4	0.45	0.82	0.68

1) Threshold : Attribute value = 7
2) class 2 : 0.35, class 4 : 0.65

In Table 3, three decimal digits are the results from the rounding in breast-cancer properties affected the risk ranking Marginal Adhesion, Bare nuclei, Uniformity of Cell Size, Clump Thickness, Normal Nucleoli, Single Epithelial Cell Size, Uniformity of Cell Shape, Mitoses, Bland. Chromatin is the same as Marginal Adhesion. This is a partial adhesion of cells separated from each other, and inflammation of the skin or membrane adhering to each other called phenomenon. It can be seen duration of these adhesions outdated higher incidence of breast-cancer.

The performance is compared by 10-fold cross-validation for bayesian networks data. The results in Table 4 can be found through the validation of the results.

Table 4. 10-fold cross-validation Results

	using baye's theorem	
	Benign	Malignant
Benign	406	17
Malignant	5	199
	using NBN	
	Benign	Malignant
Benign	399	19
Malignant	7	192

Based on the theory of bayesian, cross-validation results of total 699 pieces of data, 605 pieces of data classification based on the probability obtained with an accuracy of 0.865%. In the case of NBN 699 pieces, 591 pieces were obtained by classifying the accuracy of 0.844%. With the results of the bayesian theory, diagnostic accuracy of prediction is applied as 0.021% higher performance.

5. CONCLUSIONS

Currently datamining sector is interested and applied in many areas. In other words, datamining is predicting the future to discover hidden correlations and make decisions. To interpret data on various aspects can be converted to real expectation. Analyzing the results even a simple can be found big difference. In this paper, we disclose the data as evidence for breast-cancer database against each property using the calculated probability of cancer found. Thus, each cancer cell properties that affect the ranking of the most influential property priced Marginal Adhesion were found in the result. With the occurrence of inflammation persists long after the change into cancer cells by taking the results of evidence-based resources, doctors have a chance to improve the accuracy in patients. It can be treated as an opportunity.

In the future, cervical cancer related to this paper will be developed as bayesian theory and bayesian network. The actual data will be obtained and error rates be calcurated by a deliberate plans to experiment and research.

REFERENCES

[1] Australian Institute of Health and Welfare & National Breast Cancer Centre, Breast cancer is Australia: an overview, 2006. Cancer series No. 34. Cat. no. CAN 29. Canberra:AIHW.

[2] Shmueli Galit, Patal Nitin R. and Bruce Peter C, "Data Mining for Business Intelligence", John Wiley & Sonc Inc., 2006

[3] Yong Gyu Jung, Song Ei Han, Ranking Methods of Web Search using Genetic Algorithm, IWIT, Vol.10 No.3 p91-p96

[4] Hwan Seung Yong, Introduction to Data Mining , Infinitybooks, p223–p241, 2007

[5] Hag Yong Han, Introduction to Pattern Recognition, Hanbit Press, p86 – p88, 2009

[6] Charniak, E. (1991). Bayesian Networks without Tears. AI Magazine, p50-p63.

[7] Yong Gyu Jung, Bum Jun Lee, Features Reduction using Logistic Regression for Spam Filtering, IWIT, Vol.10 No.2 p13-p18

Yong Gyu Jung
received the B.S. in physics Education from Seoul National University in 1981. And then he got the M.S. and ph.D. degree of Computer Science from Yonsei and Kyonggi University in 1994 and 2003 respectively. Since 1999, he has been a Faculty of Medical IT marketing Dept. in Eulji University. His research interests are in the areas of medical information analysis and international standards including e-Business. He has been a leading member of the UN/ECE/CEFACT and ISO/TC154 for international standardiztion.

Jang Il Kim
obtained a B.S. in physics from the Suncheon National University and now a graduate student at the Department of IT Marketing in Eulji University. He is a member of KISA phishing Center Advisory Consultant and interested in the field of security and medical information

Jun Heo
received his B.S., M.S. degrees and Ph.D. Candidate in computer science and engineering from Sogang University. And he received Ph.D. degree in information and control engineering, Kwangwoon University. Now he joined the faculty of the Department of Information and Communication, Kyungmin College, Uijeongbu, Kyunggi, Korea, as an assistant professor. His research interests include wired and wireless network protocol, military communication network.

Sung Chul Sihn
obtained B.S. and M.S at Hannam University and Yonsei University respectively. And now work for Health Care Solutions Division, Fujitsu Korea Co. Ltd., Korea as a director. He is interested in the field of Health and Medical system and consulting.