http://dx.doi.org/10.7236/JIWIT.2012.12.2.175
JIWIT 2012-2-22

베이지안 알고리즘을 이용한 유방암 진단 예측모델

Prediction Model for Breast Cancer Diagnosis using Baysian Algorithm

정용규[*], 이연주[**], 원재강[***]

Yong-Gyu Jung, Yeon-Joo Lee, Jae-Kang Won

요 약 데이터 마이닝은 특정분야에서만 관심을 갖는 분야가 아니라 현재 우리주변 여러 분야에서 많이 사용되고 응용되고 있다. 수많은 데이터 가운데 숨겨져 있는 유용한 상관관계를 발견하여, 미래에 실행 가능한 정보를 예측하여 추출해 내고 추후에 의사 결정에 이용하는 과정을 말한다. 따라서 데이터를 다양한 관점으로 해석하기 위해 데이터를 변환할 수 있다. 의료분야에서의 예를 들면 간단한 질환도 분석의 결과에서 큰 차이를 발견할 수 있다. 유방암에 관련된 속성들에 대해 베이즈 이론을 적용하여 유방암 발병 확률을 예측한다. 이를 통하여 과거 환자진찰 데이터로 얻은 자료를 적용하여 증거기반의 의료서비스를 제공하며, 또한 진찰결과에 대한 신빙성을 증가시킬 수 있다.

Abstract 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. The properties associated with breast cancer by about applying bayesian theory is used to predict the probability. In the past patient data, doctors may be obtaining by applying evidence-based care for patients with the results of examination and By using the the past patient data.

Key Words : Breast-cancer, Cervical Cancer, Baysian Algorithm, 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.[1] 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.

[*]종신회원, 을지대학교 의료IT마케팅학과
[**]정회원, 을지대학교 의료전산학전공
[***]정회원, 경기대학교 전자계산학과(교신저자)

접수일자 2012년 2월 10일, 수정완료 2012년 3월 23일
게재확정일자 2012년 4월 13일

Received: 10 February 2012 / Revised: 23 March 2012 / Accepted: 13 April 2012

[***]Corresponding Author: 06240604@hanmail.net
Department of Computer Science, Kyonggi University

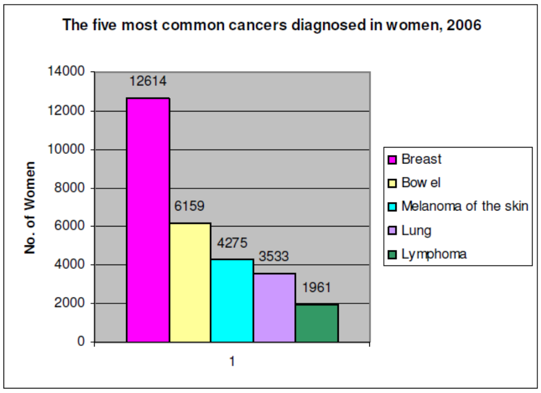


그림 1. 연간 유방암 발병 증가 추세
Fig. 1. Annual increase of breast cancer

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.[2]

II. Related research

1. Breast Canser

Breast cancer tumor cells is made of the comprising palpable lump. breast palpable lump of the symptoms is accounting for approximately 70% as the most common symptoms of breast masses. Typically, these comes mostly from breast duct and lobules in the differential diagnosis with breast cancer

The cause of this cancer does not appear clearly, even though the study of breast cancer patients and healthy people. When it could be found to be compared to the differences, these differences are called risk factors. Risk factors for breast cancer are known as hormone (estrogen), age and birth experience, breastfeeding factors, alcohol, radiation exposure, family history of breast cancer, etc.

It stimulates the growth and division with epithelial cells of breast cancer, such as the female hormone estrogen, which the female hormone estrogen in breast cancer epithelial cells exposed to the longer duration, ie, there is no birth or breast-feeding experience menarche or menopause fast ohraehan delayed menstruation increases the risk of women developing breast cancer. 5-10% of patients with breast cancer have a genetic stamp BRCA1 and BRCA2 genes that are known mutations.[3]

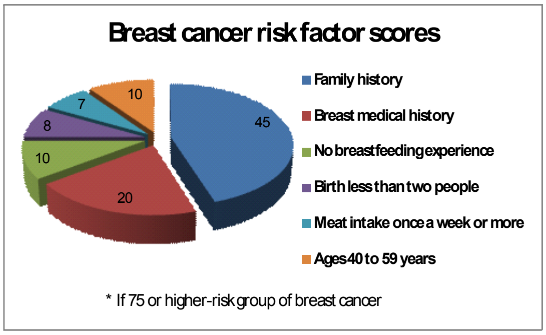


그림 2. 유방암 위험인자 점수
Fig. 2. Breast cancer risk factor scores

2. Baye's Theorem

Bayes′ theorem is also known as bayesian theory, Thomas Bayes an accident under the principles of essays on how to solve the problem published in the theory. Some experimental results are obtained from the information to improve the probability, while an event can be the first time. Bayes′ theorem is also known as bayesian theory, Thomas Bayes an "accident under the principles of essays on how to solve the problem" is published in the theory.[4],[5]

$$P(A|B) = \frac{P(B|A)\cdot P(A)}{P(B)} \quad (1)$$

3. Bayesian Network

This probabilistic dependencies between variables representing the graph consists of the conditional probability of each parametric screening. Thus, a Bayesian network each node has a conditional probability tables can be defined as a non-cyclic graph and the values assigned to other nodes have a value based on a particular node can be used to calculate the conditional probability. In other words, from one set of data to learn Bayesian network, each node for each attribute in the dataset, and each arc to express dependencies between the properties, and thus classified on the basis of the learned Bayesian network to predict the probability of class. The Bayesian network to solve more complex problems to calculate the conditional probability using Bayesian theory is a mechanism that can be automatically extended.[6]

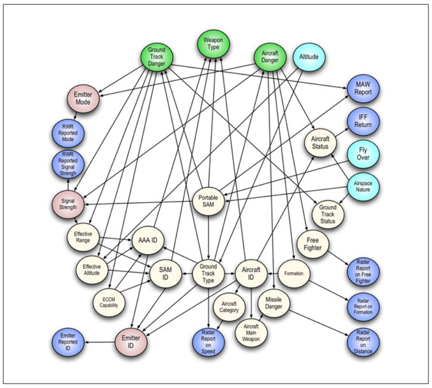


그림 3. 베이지안 네트워크
Fig. 3. bayesian network

III. Experiments

1. Data set

UCI repository for experimental data provided by the breast cancer Wisconsin data were used for the public. Breast cancer dataset using the data obtained from the Wisconsin hospitals has 699 properties with 11 pieces of data are included. In this paper, we take the probability of breast cancer bayesian theory to calculate the characteristic values of the variables that can not be obtained except two variables Matlab variable by using a tool for the prediction of breast cancer test results were obtained.

2. Features variable

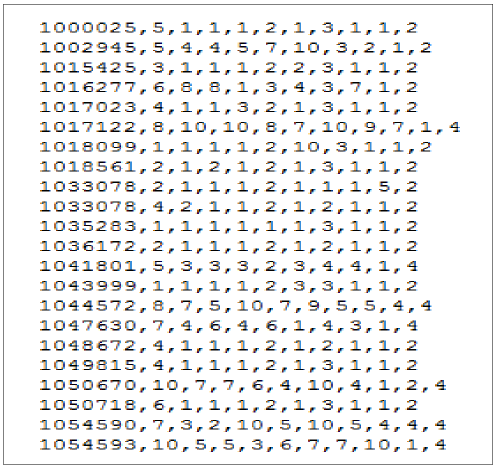
Variables used in this paper the characteristics of the data, as shown in Table 1 has a total of 11 properties. That exist in the case of breast cancer cells to deliver proteins accounted for a large proportion. However, if these cells into cancer cells spread to facilitate the delivery of proteins made 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.

표 1. breast cancer dataset의 속성
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. The range of data

Data for the above properties, look at the range of values of the properties as shown in Figure 4 below, except for sample code number and class of nine has a value of attribute values from 0 to 10.



```
1000025,5,1,1,1,2,1,3,1,1,2
1002945,5,4,4,5,7,10,3,2,1,2
1015425,3,1,1,1,2,2,3,1,1,2
1016277,6,8,8,1,3,4,3,7,1,2
1017023,4,1,1,3,2,1,3,1,1,2
1017122,8,10,10,8,7,10,9,7,1,4
1018099,1,1,1,1,2,10,3,1,1,2
1018561,2,1,2,1,2,1,3,1,1,2
1033078,2,1,1,1,2,1,1,1,5,2
1033078,4,2,1,1,2,1,2,1,1,2
1035283,1,1,1,1,1,1,3,1,1,2
1036172,2,1,1,1,2,1,2,1,1,2
1041801,5,3,3,3,2,3,4,4,1,4
1043999,1,1,1,1,2,3,3,1,1,2
1044572,8,7,5,10,7,9,5,5,4,4
1047630,7,4,6,4,6,1,4,3,1,4
1048672,4,1,1,1,2,1,2,1,1,2
1049815,4,1,1,1,2,1,3,1,1,2
1050670,10,7,7,6,4,10,4,1,2,4
1050718,6,1,1,1,2,1,3,1,1,2
1054590,7,3,2,10,5,10,5,4,4,4
1054593,10,5,5,3,6,7,7,10,1,4
```

그림 4. breast cancer dataset의 데이터 일부
Fig. 4. part of breast cancer dataset

The data above the very last, as you can see the property values of 2 and 4 are divided into two classes. Diseased cells, if the class attribute value of 2, class 4, if the cancer is. The data used for the experiment with data from a total of 699 patients, 458 of 699 people tested and breast cancer, 65.5% master cells, have the disease and the remaining 241 patients have cancer. Patients with disease of the cells belonging to class2, class4 belong to patients with cancer, the distribution of each class are listed below.

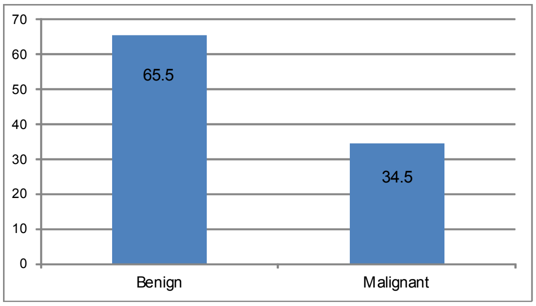


그림 5. 클래스변수 분포도 그래프
Fig. 5. Distribution graph of the class variable

4. Experiment

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.

표 2. Attribute value별 사례 수
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

Threshold : Attribute value = 7

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.

표 3. Attribute value별 사후확률
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.

Ⅳ. Experiments Results and Discussion

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.

표 4. 10-fold 교차 검증 결과
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.

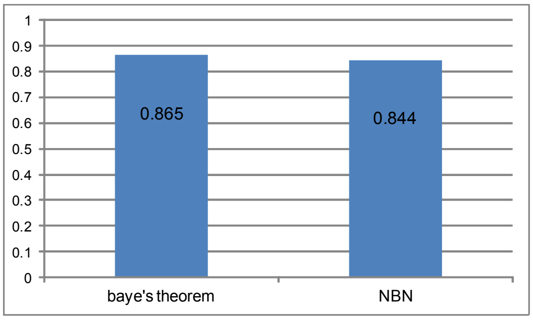


그림 6. 10-fold Validation 결과
Fig. 6. 10-fold Validation Result

In addition, breast cancer dataset is used in the experiment for the NBN result tree. It is generated as Fig 7. Class properties of the data sets is from the parent node of the other properties, other properties except the parent node is independent of each other. Thus, except for the Class property of the lower-level nodes in the same Class attribute is present.

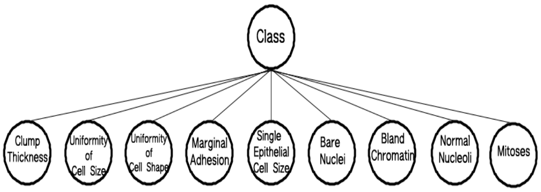


그림 7. NBN 트리
Fig. 7. NBN tree

Ⅴ. Conclusion

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.

Typical examples of these CRI (Centerstone Research Institute) was in community mental health center. It is likely to help patients in the center, the clinician to the most successful treatment through the use of direct feedback is gaining practical clinical information. This information applies to trial again, and actually fit the needs of the patient by applying a variety of treatment options to find.

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

[8] Jeong-Suk Lee, Kyung-Min Ahn, Korean Skin Care on Japanese Tourist's Satisfaction and Revisit, Journal of the Korea Academia-Industrial Cooperation Society, v.12, no.11, pp.4756-4763, NOV-2011

[9] Young-Mee Lee, Organizational Commitment and Its Related Factor among Medium Hospitals of Nurses, Journal of the Korea Academia-Industrial Cooperation Society, v.12, no.11, pp.4764-4769, NOV-2011

[10] Dong-Il Kim, Seung-Il Choi, Analysis of ERP (Enterprise Resource Planning) Implementation and Project Critical Sucess Factor, Journal of the Korea Academia-Industrial Cooperation Society, v.12, no.11, pp.4770-4777, NOV-2011

저자 소개

정 용 규(종신회원)

- 1981년 서울대학교 (이학사)
- 1994년 연세대학교 (공학석사)
- 2003년 경기대학교 (이학박사)
- 1999년~현재 을지대학교 교수

<주관심분야: 임상데이터마이닝, 의료정보시스템, 전자거래표준>

이 연 주(정회원)

- 2009년~현재 을지대학교 의료전산학 전공

<주관심분야: 의료정보시스템, 데이터마이닝>

원 재 강(정회원, 교신저자)

- 1999년 강릉대학교 (이학사)
- 2002년 경기대학교 (이학석사)
- 2009년 경기대학교 (이학박사)
- 2000년~현재 경기대학교 외래강사

<주관심분야: 워크플로우, BPM, 데이터마이닝>