Naive Bayes 분석기법을 이용한 유방암 진단

박 나 영* · 김 장 일** · 정 용 규***

목 차

요약

선진국형 질병으로만 알려져 있던 유방암이 우리나라 현대 여성들에게 발병률이 꾸준히 증가하고 있다. 유방암은 보통 50대 이상의 여성에서 발병하는 병으로 알려져 있지만 우리나라의 경우 40대의 서양보다 젊은 여성들에게 발병률이 꾸준히 증가하고 있다. 따라서 우리나라 성인여성을 기준으로 유방암에 대한 정확한 진단을 할 수 있는 매뉴얼을 구축하는 것이 시급한 과제이다. 본 논문에서는 데이터마이닝기법을 이용하여 유방암을 예측하는 방법을 제시한다. 데이터마이닝이란 데이터베이스 내에 숨어 있는 일정한 패턴이나 변수들 간의 관계를 정교한 분석모형을 이용하여 쉽게 드러나지 않은 유용한 정보를 찾아내는 과정을 말한다. 실험을 통하여 Deicion Tree와 Naive Bayes 분석기법을 사용하여 유방암을 진단하는 분석기법을 비교분석을 하였다. Deicison Tree는 C4.5 알고리즘을 적용하여 분석하였고 두 알고리즘이 상당히 좋은 분류 정확도를 나타냈다. 그러나 Naive Bayes 분류방법이 Decision Tree방법보다 더 상회하는 정확도를 보였고 이는 의료데이터의 특성에 많이 기인한다고 볼 수 있다.

표제어: 유방암, 진단, 의사결정나무, 나이브베이즈

접수일(2013년 3월 20일), 수정일(1차: 2013년 3월 25일), 게재확정일(2013년 3월 28일)
* 을지대학교 의료IT마케팅학과, parkny42@nate.com
** GOODiT 실장, gold@goodit.co.kr
*** 을지대학교 의료IT마케팅학과 교수, 교신저자, ygjung@eulji.ac.kr

1. 서론

유방암은 선진국형 질병으로 미국의 경우, 가장 흔한 암으로 알려져 있으며, 특히 40세에서 55세 사이의 미국 여성의 제 1의 사망원인이 되고 있다. 한국의 경우 여성 8명 중 1명의 확률로 유방암이 발생하고 있으며 유방암의 환자 수 역시 매년 약 15%씩 증가하고 있다고 보고되고 있다. 그림 1은 유방암 증가 추이를 시각적으로 보여주고 있다.

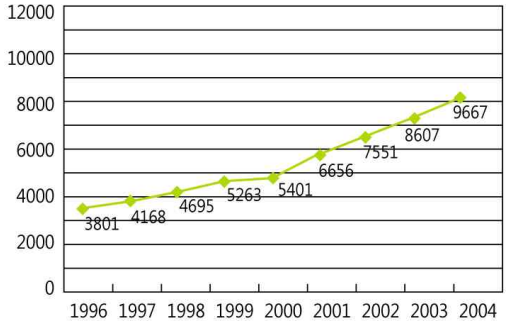

그림 1. 유방암 증가 추이
Fig. 1. Breast Cancer Trends

1990년대에는 한국여성의 유방암은 여성 암환자의 약 11.9%로 자궁경부암, 위암에 이어 세 번째에 유방암이 위치해 있고, 전체적으로는 위암, 간암, 자궁암, 폐암에 이어서 다섯 번째로 높은 암으로 통계에서 보여주고 있다. 이후 급속한 증가를 보여 2000년 보건복지부 통계를 보면 15.1%로 여성 암 중 15.8%의 위암에 이어 두 번째로 가장 흔한 암으로 통계되고 있다. 환자 수 또한 10년 새 2~3배 증가하고 있다고 보고되고 있다. 그림 2는 증가추이를 시각적으로 제시하고 있다[1].

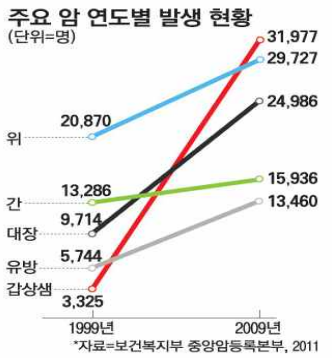

그림 2. 주요 암 증가 추이
Fig. 2. Major Cancer Trend

한국 여성의 유방암 증가추세는 식사나, 서구 방식의 생활화, 빠른 월경, 아이를 적게 낳고 수유 회피, 피임약의 사용 등 생활패턴의 변화가 유방암 증가 원인으로 지적되고 있기는 하지만 유방암의 원인이 아직 정확하게 알려진 것이 없으므로 유방암 증가 원인을 정확하게 파악하기 어렵다. 한국의 유방암의 특징은 서양의 여성들이 보통 50대에 발병하는 것과는 달리 한국의 여성들은 40대에 발생빈도가 높아 비교적 젊은 여성층에서 유방암이 발생한다는 점이다. 따라서 서구의 유방암 지침과 다른 한국 현실에 맞는 유방암에 관리 지침서를 통한 관리가 필요하다[2, 3]. 본 논문에서는 weka 소프트웨어를 사용하여 실험하였으며, Decision Tree와 Naive bayes 2가지의 분류방법을 사용하여 어떤 분석방법이 유방암을 진단하는데 더 정확한지에 대하여 비교분석을 목적으로 실험한다.

데이터마이닝이란 데이터베이스 내에 숨어 있는 일정한 패턴이나 변수들 간의 관계를 정교한 분석모형을 이용하여 쉽게 드러나지 않은 유용한 정보를 찾아내는 과정을 말한다. 데이터마이닝은 방대한 자료 속 에서 의미 있는 정보를 추출해내는 과정이라고 할 수 있다. 본 논문에서는 실험을 통하여 Deicion Tree와 Naive Bayes 분석기법을 사용하여 유방암을 진단하는 분석기법을 비교분석을 하였다. 본 실험에서 사용한 Deicison Tree는 C4.5 알고리즘을 적용하여 분석한다[4, 5].

2. 문헌 연구

2.1 Decision Tree

Decision Tree는 의사결정 규칙을 도표화하여 관심대상이 되는 집단을 몇 개의 소집단으로 분류하거나

예측을 수행하는 계량적인 분석방법을 말한다. 분석 결과는 '조건이 A이고 조건이 B이면 결과집단 C'라는 형태의 규칙으로 표현되므로 분류 또는 예측을 목적으로 하는 다른 계량적 분석방법에 비해 쉽게 이해할 수 있다는 장점이 있다[6].

나무를 어느 정도의 크기로 성장시킬 지를 결정 하는 것이 나무모형을 결정하는데 있어서 중요한 역할을 한다. 이때 나무 모형의 크기를 결정하는데 정지규칙과 가치치기 방법 등이 있다. 정지규칙이란 사용자가 미리 지정한 조건에 해당될 때에 나무의 성장을 정지시키는 방법을 말한다. 이러한 규칙에는 최대 나무 높이, 자식마디의 최소 관측치 수, 또는 카이제곱 검정 통계량, 지니계수, 엔트로피 지수 등이 될 수 있다[7].

가지치기란 나무모형을 크게 만든 다음에 불필요한 가지들을 제거하여 최적의 나무모형을 구축하는 방법으로 Breiman이 제시한 비용-복잡성 가지치기가 대표적이다[8, 9]. 이 방법은 우선 나무모형을 최대한 크게 만든 후, 이를 이용하여 부나무(subtree)를 선택하고, 선택된 부나무에 교차결정법(crossx-validation)을 이용하여 적절한 크기의 나무를 최종적으로 선택하는 방법이다.

2.2 Naive Bayes

Naive Bayes분류는 지도학습(Supervised Learning)을 사용한 간단한 분류 중 하나이다. 이 분류는 베이즈 룰(Bayes' rule)을 기본적으로 사용하고 있다. 베이즈 룰을 사용하는 가장 큰 이유는 조건부 확률을 구할 때 베이즈 룰을 이용할 때 더욱 손쉽게 값을 구할 수 있기 때문이다 분류를 위해 d를 입력값, c를 분류한 class중 하나라고 가정했을 때, 이는 나이브 룰로 아래와 같이 나타낸다.

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)} \tag{1}$$

부류가 여러 개일 경우 각각의 부류에 대한 해당 조건부 확률을 계산할 수 있으며 입력값은 고정되어 있으므로 가장 높은 확률을 가지는 부류가 입력값에 속할 부류가 되게 된다.

$$\begin{aligned} C_{MAX} &= \arg\max_{c \in c} P(c|d) \\ &= \arg\max_{c \in c} \frac{P(c|d)P(c)}{P(d)} \\ &= \arg\max_{c \in c} P(c\,d)P(c) \end{aligned} \tag{2}$$

따라서 최종 나이브 베이지안 분류식은 다음과 같다.

$$\begin{aligned} C_{MAX} &= P(x|c)P(c) \arg\max_{c \in c} \\ & P(x_1, x_2, \cdots, x_n|c)P(c) \\ &= \arg\max_{c \in c} \prod_{x \in X} \end{aligned} \tag{3}$$

3. 모델제안

분석에 사용된 데이터는 UCI에서 제공하는 유방암진단에 관한 데이터 699개를 사용하였고, 11개의 속성을 바탕으로 분석하였다. 표 1은 데이터 속성들을 표로 설명하고 있다. 나무 모형의 크기를 결정하는 방법으로는 교차결정방법(cross-validation)을 사용하여 결정하였고 이에 따른 나무의 크기는 27이고, 트리의 리프(leaves) 개수는 14개로 이루어져 있다. Tree 분석의 정확도는 94.5637%이고, 오분류율(misclassified rate)은 5,4363%로 분석되었다.

표 1. 데이터 속성 설명
Tab. 1. Property Description data

데이터	데이터 형
Sample Code Number	Numeric
Clump Thickness	Numeric
Uniformity of Cell Size	Numeric
Uniformity of Cell Shape	Numeric
Marginal Adhesion	Numeric
Single Epithelial Cell Size	Numeric
Bare Nuclei	Numeric
Bland Chromatin	Numeric
Normal Nucleoli	Numeric
Mitoses	Numeric
Class	Numeric

4. 실험 및 결과평가

본 논문의 분석을 위하여 C4.5 알고리즘을 적용하였다. C4.5 알고리즘은 1993년 J.Ross Quilnlan에 수정 발전된 의사결정 알고리즘이다. 이것은 초기버전인 ID3 알고리즘으로 기계학습 분야에 많은 영향을 주었다. CART가 각 마디에 이원분할을 형성하며 이지분리 나무구조를 만드는데 반하여 C4.5는 연속형 예측변수에 관해서는 이지분리를 하지만, 명목형 예측변수에 관해서는 각 범주가 하나의 마디를 가지는 다지 분류 구조를 갖는 나무로 구성된다[10]. weka 소프트웨어를 통해 도출된 결과는 그림 3과 같다.

그림 3을 시각적으로 표현하면 그림 4와 같은 결과를 얻을 수 있다. 그림 4의 Decision Tree는 C4.5 알고리즘을 이용하여 분석하였기 때문에 각 범주가 하나의 마디를 차지하고 있는 구조를 갖고 있다. 정확도는 전체 661/699개로 94.5637%로 나타나고 있다. Naive Bayes로 실험했을 때 Cross-Validation은 10으로 두고 실험하였다 그 결과 정확도는 671/699, 95,9943%로 Decision Tree로 실험했을 때 보다는 약 1% 정도 높은 정확도를 보여주고 있다. 다음 표 2는 정확도와 부정확도 그리고 오분류 Instance의 개수를 표로 보여주고 있다.

```
Decision Tree:
Cell_Size_Uniformity <= 2
|   Bare_Nuclei <= 3: benign (405.39/2.0)
|   Bare_Nuclei > 3
|   |   Clump_Thickness <= 3: benign (11.55)
|   |   Clump_Thickness > 3
|   |   |   Bland_Chromatin <= 2
|   |   |   |   Marginal_Adhesion <= 3: malignant (2.0)
|   |   |   |   Marginal_Adhesion > 3: benign (2.0)
|   |   |   Bland_Chromatin > 2: malignant (8.06/0.06)

Cell_Size_Uniformity > 2
|   Cell_Shape_Uniformity <= 2
|   |   Clump_Thickness <= 5: benign (19.0/1.0)
|   |   Clump_Thickness > 5: malignant (4.0)
|   Cell_Shape_Uniformity > 2
|   |   Cell_Size_Uniformity <= 4
|   |   |   Bare_Nuclei <= 2
|   |   |   |   Marginal_Adhesion <= 3: benign (11.41/1.21)
|   |   |   |   Marginal_Adhesion > 3: malignant (3.0)
|   |   |   Bare_Nuclei > 2
|   |   |   |   Clump_Thickness <= 6
|   |   |   |   |   Cell_Size_Uniformity <= 3: malignant (13.0/2.0)
|   |   |   |   |   Cell_Size_Uniformity > 3
|   |   |   |   |   |   Marginal_Adhesion <= 5: benign (5.79/1.0)
|   |   |   |   |   |   Marginal_Adhesion > 5: malignant (5.0)
|   |   |   |   Clump_Thickness > 6: malignant (31.79/1.0)
|   |   Cell_Size_Uniformity > 4: malignant (177.0/5.0)
```

그림 3. 실험결과
Fig. 3. Results Obtained Experiments

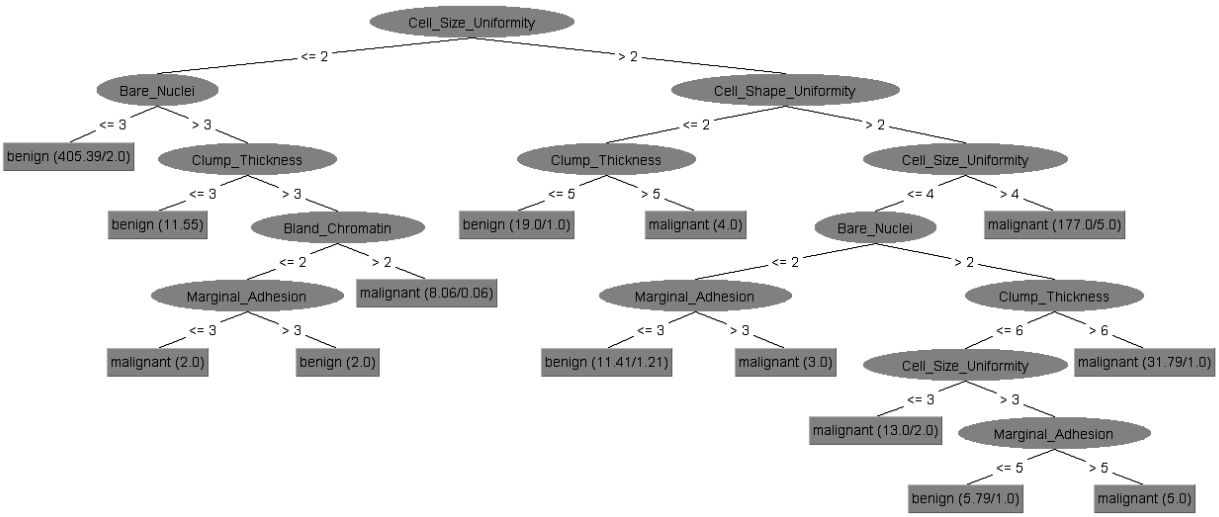

그림 4. 유방암 진단 Decision Tree
Fig. 4. Decision Tree for Diagnosis of Breast Cancer

표 2. 정확도 비교

Tab. 2. Accuracy Comparison

	Decision Tree	Naive Bayes
정확도	94.5637%	95.9943%
부정확도	5.4363%	4.0057%
오분류개수	38/699	28/699

제안하는 Decision Tree, Naive Bayes 분류는 weka 소프트웨어를 이용하여 실험하였다.

5. 결론

본 논문의 실험은 동일한 유방암 진단에 대한 데이터에 대하여 Decision Tree에는 C4.5 알고리즘을 적용하여 실험하였고 Naive Bayes에서는 Cross-Validation을 10으로 두고 실험한 결과 Decision Tree의 정확도는 94.5637%, Naive Bayes의 정확도는 95.9943%로 약 1%의 미세한 차이로 Naive Bayes를 통한 분류가 더 정확하다는 결론을 얻었다. 약 1%의 차이로 Naive Bayes를 통한 분류가 더 정확하다고 판단하기 어렵지만, 의료문제에서는 1%의 확률로도 암을 진단하는데 매우 중요하므로 1%도 간과 할 수 없다고 판단하여 Naive Bayes 분류방법이 더 정확하다는 결론을 내렸다. 향후에는 Decision Tree와 Naive Bayes에 다른 알고리즘을 적용하여 정확도를 높일 수 있는 방법을 고찰할 필요가 있다.

참 고 문 헌

[국내 문헌]

[1] 패턴인식 (2010), 오일석, 교보문고

[2] 이극노, 이홍철 (2003), “이동통신고객 분류를 위한 의사결정나무와 신경망 결합 알고리즘에 관한 연구”, 한국지능정보시스템학회논문지, 9(1), 139-155.

[3] 김형세, 문호석, 이동근, 황명상, 김영국 (2010), 의사결정나무를 이용한 근접전투전문가시스템.

[국외 문헌]

[4] Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth (1996), “Knowledge Discorvery and DataMining: Towards a Unifying Framework”, Proc.KDD-96, 1996.

[5] Breiman, L, Friedman, J. H., Olshen, R., Stone, C. J. (1984), Claasification and Regression Trees, Champman and Hall, New York, 85-98.

[6] Trong Dung Nguyen, Tu Bao Ho, Hiroshi Shimo-daira (2001), "A Scalable Algorithm for Rule Post-pruning of Large Decision Trees”, Proceedings of the 5th Pacific-Asia Conference on Know-ledge.

[7] J, Ross Quinlan (1993), C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, Inc.

박 나 영 (Na Young Park)

을지대학교 의료IT마케팅학과에 재학 중이며, 빅데이터 실험을 통한 HDFS와 MapReduce의 비교 등을 연구하고 있다. 병원에서 사용하고 있는 OCS, EMR 등 의료정보시스템 구현기술과 임상데이터 분석을 위한 데이터마이닝 기술에 관심이 많다.

김 장 일 (Jang Il Kim)

순천대학교에서 학사를 취득하였고 현재 을지대학교 IT마케팅학과 대학원에 재학 중이다. GOODiT에서 IT관련 컨설팅을 하고 있다. 또한 KISA 피싱센터 자문 컨설턴트로 활동 중이며, 보안 및 의료정보 관련 분야에 관심이 많다.

정 용 규 (Yong Gyu Jung)

서울대학교, 연세대학교, 경기대학교에서 각각 학사, 석사, 박사학위를 취득하였고, 현재 을지대학교 의료IT마케팅학과 교수로 재직 중이다. ISO/TC154, UN/Cefact의 한국대표위원으로 활동하고 있으며, 의료정보, 전자무역, 해상물류, 금융전산에 Semantic Web, Process Modelling, ebXML 등의 표준기술의 적용에 관심이 많다.

Breast Cancer Diagnosis using Naive Bayes Analysis Techniques

Na-Young Park* · Jang-Il Kim** · Yong-Gyu Jung***

ABSTRACT

Breast cancer is known as a disease that occurs in a lot of developed countries. However, in recent years, the incidence of Korea's modern woman is increased steadily. As well known, breast cancer usually occurs in women over 50. In the case of Korea, however, the incidence of 40s with young women is increased steadily than the West. Therefore, it is a very urgent task to build a manual to the accurate diagnosis of breast cancer in adult women in Korea. In this paper, we show how using data mining techniques to predict breast cancer. Data mining refers to the process of finding regular patterns or relationships among variables within the database. To this, sophisticated analysis using the model, you will find useful information that is easily revealed. In this paper, through experiments Deicion Tree Naive Bayes analysis techniques were compared using analysis techniques to diagnose breast cancer. Two algorithms was analyzed by applying C4.5 algorithm. Deicison Tree classification accuracy was fairly good. Naive Bayes classification method showed better accuracy compared to the Decision Tree method.

Keywords: Beast Cancer, Diagnosis, Decision Tree, Naive Bayes

* Eulji University, Department of Medical IT Marketing, parkny42@nate.com
** GOODiT, Director, gold@goodit.co.kr
*** Eulji University, Department of Medical IT Marketing, professor, Corresponding author, ygjung@eulji.ac.kr