비용 의존 학습에 의한 인유두종 바이러스의 분류

황소현° 박성배 장병탁
서울대학교 컴퓨터 공학부
{shhwang°, sbpark, btzhang}@bi.snu.ac.kr

Human Papillomavirus Risk Classification by Cost-Sensitive Learning

Sohyun Hwang° Seong-Bae Park Byoung-Tak Zhang
Biointelligence Laboratory, School of Computer Science and Engineering,
Seoul National University

요 약

인유두종 바이러스는 표피세포에 감염되는 DNA 바이러스로 자궁경부암을 일으키는 가장 큰 요인이다. 현재까지 100 여개의 종류가 알려져 있고 악성종양 유발 가능성에 따라 위험군을 나누는데, 여기서 중요한 것은 고위험군을 저위험군으로 잘못 분류하는 것을 최소화하는 것이다. 본 논문에서는 분류를 위한 데이터로 인유두종 바이러스에 관한 문서 자료들을, 기계 학습 방법으로 분류 비용을 고려해 줄 수 있는 비용 의존 학습을 이용하였다. 실험결과, 비용을 고려해 주는 것이 고려하지 않았을 때보다 더 좋은 성능을 나타내었다.

1. 서 론

자궁경부암은 세계 3대 여성암의 하나로, 주요 병인은 인유두종(人乳頭腫, Human Papillomavirus, HPV) 바이러스이다. HPV 감염이 조기 발견될 경우 암의 완치가 가능하므로 조기 진단이 매우 중요하다[1]. HPV는 이중 나선 DNA 암 바이러스(double-strand DNA tumor virus)로 papovavirus 과에 속해 있다. HPV 감염은 피부, 상기도, 폐, 식도, 외음항문부, 자궁질·경부, 방광 등 다양한 장기에서 보이는데 가장 중요한 것은 피부와 자궁 경부이다. HPV의 종류(type)는 약 100 여 개에 이르고 자궁 경부와 관련된 HPV는 악성 종양 유발 가능 위험도에 따라 고위험군(high risk type)과 저위험군(low risk type)으로 나뉜다[2].

100 여개의 HPV 중에서 고위험군 HPV를 분류해 내는 것은 자궁경부암 진단용 DNA chip 제작하는 데에 필요하고, 저위험군을 고위험군으로 잘못 분류하는 경우보다 고위험군을 저위험군으로 잘못 분류하는 경우의 오류가 위험하므로 후자의 경우의 실수를 더 줄여야 한다. 그러므로 이 실험에서는 쉽게 구할 수 있는 HPV 특성에 대해 기술하고 있는 문서들을 실험 데이터로, 학습 데이터의 비용을 고려해 줄 수 있는 비용 의존 학습 방법 (Cost-Sensitive Learning)을 기계학습 방법으로 이용하여서 HPV의 위험군을 분류하였다. 실험의 결과는 임상 학자가 자궁경부암 진단용 DNA chip을 만들기 위한 연구를 하기 위해 생물학 문헌 데이터로부터 관련 지식들을 모으고 읽어서 정리하는데 필요한 시간과 노력을 절약해 주고, 임상 실험 전 후의 참고 자료로 유용하게 사용될 수 있다.

2. 비용 의존 학습 알고리듬

비용 의존 학습 방법으로 AdaCost algorithm[3]을 선택하였다. AdaCost는 트레이닝 분포를 매번 갱신하기 위해 잘못 분류된 것의 비용을 사용하는 AdaBoost의 한 변형이다. 알고리듬은 그림1에 나타나 있다. weak learner로는 naive bayes classifier를 사용하였다.

Input: • $S = \{(x_1, c_1, y_1), \ldots, (x_m, c_m, y_m)\}$:

$x_i \in \chi, c_i \in R^+$ and $y_i \in \{-1, +1\}$

• weak learning algorithm **WeakLearn**

• integer T specifying the number of iterations

Initialize $D_1(t) = c_i / \sum_{i=1}^{m} c_j$ for *all i.*.

For $t = 1, \ldots, T$.

1. Call **WeakLearn**, providing it with the distribution D_t.
2. Get back a hypothesis $h_t : \Box \rightarrow \{-1, +1\}$.
3. Choose $\Box_t \in R$ and $\Box(i)$.
   where $\beta(i) = (sign(y_i h_t(x_i)), c_i)$
4. Update distribution D_t:

$$D_{t+1}(i) = \frac{D_t(i) exp(-\alpha_t y_i h_t(x_i) \beta(i))}{Z_t}$$

where Zt is a normalization constant.

Ouput: the final hypothesis

$$f(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x))$$

그림 1. AdaCost algorithm

3. 실험 데이터

미국 로스 알라모스 국립 연구소(Los Alamos National Laboratory)에서 만든 HPV 서열 데이터베이스(HPV Sequence Database)에 있는 문서들을 사용하였다[4]. HPV 서열 데이터베이스는 1994, 1995, 1996 그리고 1997 요약본을 확장한 것으로 HPV 종류의 목록과 각각의 특성과 관련 데이터들을 모아서 정리해 놓았다. 그림 2는 이 데이터베이스로부터 만든 HPV 데이터의 예이다. HPV type 80에 대한 예로써 각각의 데이터는 <definition>, <source> 그리고 <comment>의 세 부분으로 이루어져 있다. <definition>은 HPV의 종류와 유전자 구성에 대한 정보를, <source>는 HPV DNA 유전자의 출처를 그리고 <comment>는 HPV 종류의 특성들과 관련 문헌 데이터의 주석을 보여준다.

<definition>
Human papillomavirus type 80 E6, E7, E1, E2, E4, L2, and L1 genes.
</definition>
<source>
Human papillomavirus type 80.
</source>
<comment>
The DNA genome of HPV80 (HPV15-related) was isolated from histologically normal skin, cloned, and sequenced. HPV80 is most similar to HPV15, and falls within one of the two major branches of the B1 or Cutaneous/EV clade. The E7, E1, and E4 orfs, as well as the URR, of HPV15 and HPV80 share sequence similarities higher than 90%, while in the usually more conservative L1 orf the nucleotide similarity is only 87%. A detailed comparative sequence analysis of HPV80 revealed features characteristic of a truly cutaneous HPV type [362]. Notice in the alignment below that HPV80 compares closely to the cutaneous types HPV15 and HPV49 in the important E7 functional regions CR1, pRb binding site, and CR2. HPV 80 is distinctly different from the high-risk mucosal viruses represented by HPV16. The locus as defined by GenBank is HPVY15176.
</comment>

그림 2. 실험 데이터 HPV 80의 예

실험 결과의 정확성을 측정하기 위해서, 미국 알라모스 국립 연구소에서 만든 HPV 서열 데이터베이스 1997년 요약본과 우리가 만든 데이터의 <comment>를 사용하여서, HPV 위험군 분류 실험 결과와 비교될 수 있는 대조군을 만들었다.

우선은 HPV 서열 데이터베이스 1997년 요약본에 의해 구분된 그룹으로 각각 HPV를 분류하였다. 둘째, 우리가 알고자 하는 것은 자궁경부암과 관련된 고위험군 HPV이므로, HPV의 그룹이 피부와 관련된 것이면 그 그룹의 멤버들은 모두 저위험군으로 분류하였다. 셋째, HPV의 그룹이 자궁경부암 관련 고위험군 HPV으로 알려진 것들만 그 그룹의 멤버들을 고위험군으로 분류하였다. 마지막으로, 그룹이 자궁경부암과 관련된 HPV인데, 그룹 전체적으로 위험군을 분류할 수 없는 것들의 멤버들은 우리가 만든 데이터의 <comment> 부분을 참고해서, 각각의 멤버들을 분류하였다.

실험에서는 HPV 데이터 중 <comment> 부분을 텍스트 마이닝을 할 문서로 이용하였다. 각각의 HPV는 $tf \cdot idf$ 값을 원소로 갖는 벡터로 표현된다. $tf \cdot idf$ 에서 문서 d_i에 나타나는 각 단어 w_j의 가중치는 $N(w_j, d_i)$로 나타낸다.

$$N(w_j, d_i) = tf_{ij} \cdot \log_2 \frac{N}{n}$$

tf_{ij} 는 문서 d_i에 나타나는 w_j의 빈도를 나타내고 n은 w_j가 적어도 한번 나타나는 문서의 수를 의미한다. 텍스트에서 Porter's algorithm[5]으로 불필요한 어미를 제거하고 추출해 낸 단어의 수는 1,434개였다. 그러므로 각각의 HPV 데이터는 1,434 차원의 벡터로 표현된다.

4. 실험

4.1 성능 측정 방법

텍스트 분류 문제에서 성능을 측정할 때 다양한 방법들이 사용된다. 이 실험에서는 contingency table method를 사용하여서 분류 성능을 측정하였다. Recall과 Precision은 다음과 같이 정의된다.

$$recall = \frac{a}{a+c} \cdot 100\,\%$$

$$precision = \frac{a}{a+b} \cdot 100\,\%$$

$$accuracy = \frac{a+d}{a+b+c+d} \cdot 100\,\%$$

a, b, c, d 는 표 1에 정의 되어 있다.

	고위험군 HPV	저위험군 HPV
고위험군으로 분류된 것	a	b
저위험군으로 분류된 것	c	d

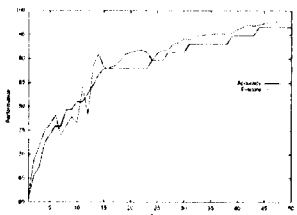
표 1. Contingency Table

$F_\beta - score$의 값은 아래와 같이 precision과 recall을 결합한 값이다.

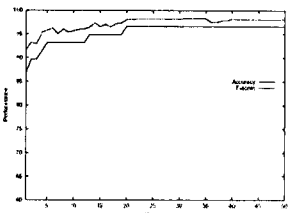
$$F_\beta = \frac{(\beta^2+1)\cdot recall \cdot precision}{\beta^2 \cdot recall + precisiion}$$

β는 precision에 대한 recall의 값으로, 실험에서는 모두 β=1로 사용하였다.

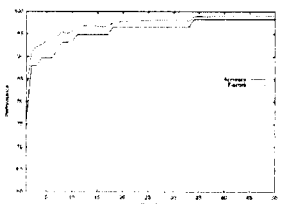
4.2. 실험 결과



(a) high : low = 0.1 : 0.9



(b) high : low = 0.5 : 0.5



(c) high : low = 0.8 : 0.2

그림 3. AdaCost의 HPV 위험군 분류 성능 측정 그래프

실험에서 데이터로 HPV 종류 74개를 구할 수 있었기 때문에, 5-fold cross validation을 사용하였다.

그림 3은 비용 값을 다르게 했을 경우의 결과이다. (a)는 가장 결과가 좋지 않았을 때, (b)는 AdaBoost의 경우, 그리고 (c)는 가장 결과가 좋았을 때의 F-score와 Accuracy 그래프이다. 위 그래프들은 비용 값을 올바르게 설정하는 것이 매우 중요함을 보여준다. F-score의 값이 accuracy보다 높게 나왔는데, 이는 고위험군 HPV를 더 많이 찾았음을 의미한다. 비율이 동일한 AdaBoost의 경우보다 (C)의 경우의 결과에서 F-score 값이 높게 나왔다.

5. 결론

이 논문에서 자궁경부암 관련 HPV의 위험군을 분류하는 실용적인 방법을 제안하였다. 분류할 때, 저위험군 HPV의 일부를 고위험군으로 잘못 분류하더라도, 고위험군 HPV를 저위험군으로 잘못 분류하는 것을 줄이는 것이 중요하므로, 비용을 고려해 줄 수 있는 AdaCost 알고리듬을 선택하였고, 실험 결과도 AdaCost가 AdaBoost보다 좋은 결과를 보여주었다. 그리고 F-score 값이 accuracy보다 높게 나와서 고위험군 HPV를 더 많이 정확하게 찾았음을 보여 주었다. 이 결과는 또한 실제적으로 자궁경부암 진단 검사용 고위험군 HPV 검색 DNA-chip 제작 연구 실험 전 후의 참고자료로 유용하게 사용될 수 있으며, 관련 자료를 찾고 읽고 정리하는데 필요한 시간과 노력을 절약해 줄 수 있다.

감사의 글

본 논문은 교육부 BK21 사업, 과기부 NRL, BrainTech 프로그램에 의하여 지원되었음.

참고 문헌

[1] Schiffman, M.H., H.M. Bauer, R.N. Hoover, A.G. Glass, D.M. Cadell, B.B. Rush, D.R. Scott, M.E. Sherman, R.J. Kurman, S. Wacholder, "Epidemiologic Evidence Showing That Human Papillomavirus Infection Causes Most Cervical Intraepithelial Neoplasis," *Journal of the National Cancer Institute*, 85, pp. 958-964, 1993.

[2] Janicek, M.F., H.E. Averette, "Cervical Cancer: Prevention, Diagnosis, and Therapeutics," *Cancer Journal for Clinicians*, 51, pp. 92-114, 2001.

[3] W. Fan, S. Stolfo, J. Zhang, and P. Chan, "AdaCost: Misclassification Cost-Sensitive Boosting," In Proceedings of the 16th International Conference on Machine Learning, pp 97-105, 1999.

[4] http://hpv-web.lanl.gov/stdgen/virus/hpv/index.html

[5] M. Porter, "An Algorithm for Suffix Stripping," Program Vol 14, No.3 pp. 130-137, 1980.