결정 트리에 의한 인유두종 바이러스의 위험군 분류

Classification of the Risk Types of Human Papilloma Virus by Decision Trees

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

Abstract

The high-risk type of HPV is the main etiologic factor of cervical cancer, which is a leading cause of cancer deaths in women worldwide. Therefore, classifying the risk type of HPV is very useful and necessary to the diagnosis and remedy of cervical cancer. In this paper, we classify the risk type of 73 Human Papilloma Virus (HPV) and predict the risk type of 4 HPVs of which the type is difficult to determine. As a machine learning method, we use Decision Trees providing a practical method for concept learning and discrete-valued functions. According to the experimental results, 59 HPVs are classified correctly among 73 HPVs and roughly 3 HPVs are predicted right.

1. 서론

자궁 경부암은 세계 3 대 여성암의 하나로, 우리나라에서도 가장 흔한 부인암(婦人癌)이다. 직접적인 병인은 인유두종(人乳頭腫) 바이러스, Human Papilloma Virus(HPV)로 알려져 있고, 조기 발견될 경우 완치가 가능하므로 조기 진단이 매우 중요하다[1]. HPV 는 이중 나선 DNA 암 바이러스(double-strand DNA tumor virus)로 papovavirus 과에 속해 있다. 현재 HPV 감염은 피부, 상기도, 폐, 식도, 외음항문부, 자궁질·경부, 방광 등 다양한 장기에서 보이는데 가장 중요한 것은 피부와 자궁 경부이다. 지금까지 발견된 HPV의 종류(type)는 약 100 여 개에 이르는데, 자궁 경부와 관련된 HPV는 악성 종양 유발 가능 위험도에 따라 고위험군(high risk type)과 저위험군(low risk type)으로 나뉜다[2]. 그러므로 자궁 경부암을 진단할 때에, 환자가 고위험군 HPV 를 가지고 있는지 그리고 어떤 HPV가 고위험군인지를 아는 것이 매우 중요하다.

HPV 의 위험군을 분류하는 한 가지 방법은 실제 생물학자들이 실험 전에 문헌 데이터들로부터 정보를 수집하는 방법과 비슷하게 text mining 기법을 이용하는 것이다. HPV 와 자궁 경부암과 관련해서 매우 많은 연구가 이루어지고 있기 때문에, HPV 와 자궁 경부암에 관련된 문서 자료들을 쉽게 구할 수 있다. 이 논문에서는 HPV 종류의 특성에 대해 기술하고 있는 문서들을 실험 데이터로, 결정 트리 방법을 학습 알고리듬으로 이용하여 HPV 의 위험군을 분류하였다. 이 논문의 결과는 자궁 경부암 관련 HPV 감염여부 진단을 위한 DNA-chip 을 제작하는 데에 참고자료로 유용하게 사용될 수 있다. HPV DNA-chip 을 제작하기 위해서는 여러 HPV 중 자궁 경부암과 관련된 고위험군 바이러스를 찾아야 하므로, 이 논문의 결과는 생물학 문헌 데이터로부터 관련 지식들을 모으고 읽어서 정리하는 데 필요한 시간들을 절약하게 해 주고, 또한 임상 실험 결과에 대한 참고자료로서 유용하게 사용될 것이다.

본 논문의 구성은 다음과 같다. 2 장은 HPV 위험군 분류 방법으로 사용된 결정 트리 학습의 원리와 구성 방법에 대해 설명하였다. 3 장은 실험에 사용한 데이터를 선택한 이유와 실험 데이터의 구성 그리고 실험 결과와 비교하기 위해 만든 대조군 데이터를 만든 방법 등에 대해 기술하였다. 4 장은 실험 결과와 분석 방법에 대해 기술하였고, 마지막으로 5 장에서 결론을 맺는다.

2. 결정 트리 학습

자궁 경부암 관련 고위험군 HPV 를 분류하기 위해서 결정 트리(decision tree)를 사용하였다. 결정 트리 학습은 널리 사용되며 귀납적 추론(inductive inference)에 매우 실용적인 방법 중 하나이다. 결정 트리가 노이즈에 강하고 논리합 표현을 학습하는 이산 함수를 유도하는 방법이고 본 논문에서는 분류 클래스가 이산적으로 표현되기 때문에, 결정 트리가 이 데이터로부터 규칙을 학습하는 데 적합하다고 할 수 있다.

결정 트리에서는 각 비단말 노드가 인스턴스의 어떤 속성(attribute) 검사를 뜻하고, 그 노드로부터의 가지(branch)는 이 속성이 가질 수 있는 가능한 여러 값 중에서 하나이다. 각 인스턴스는 트리의 루트 노드에서 시작해서 아래로 내려가면서 분류된다. 비단말 노드에서는 인스턴스의 속성 값을 검사해서 해당 값에 따라 트리의 가지가 선택된다. 이러한 과정이 새로운 노드를 루트로 하는 서브 트리에서도 반복되어 마지막 단말 노드에까지 이르게 된다. 결정 트리 학습에서 가장 중요한 문제는 트리의 각 노드에서 어떤 속성을 검사할 것인지를 선택하는 것이다. 예제 인스턴스를 분류하는데 가장 유용한 속성을 선택하기 위해서, 정량적 측정 단위인 정보이득(*information gain*)이 사용될 수 있다. 이 측정 단위는 결정 트리 학습 알고리듬으로 가장 널리 쓰이는 ID3 알고리듬과 그 후계자라고

할 수 있는 C4.5 알고리듬에서도 채택된 것이다. 목적 속성이 c 개의 서로 다른 값을 가질 수 있을 때 예제 인스턴스 집합 (S)의 엔트로피는

$$Entropy\ (S) \ = \sum_{i=1}^{c} - P_i \log_2 P_i$$

와 같이 정의된다. 그러면, 예제 인스턴스 집합 S 에서의 속성 A 에 대한 정보 이득 $Gain(S,A)$ 은 다음과 같이 정의된다.

$$Gain\ (S,\ A) \ = Entropy\ (S) \ - \sum_{v \in Values\ (A)} \frac{|S_v|}{|S|} Entropy\ (S_v)$$

여기서 $Values(A)$ 는 속성 A 가 가질 수 있는 가능한 모든 값을 뜻하고, S_v 는 A 의 값이 v 인 S 의 부분 집합이다. 그러므로 $Gain(S,A)$은 속성 A 에 따라 학습 예제들을 나눔으로써 얻어지는 엔트로피 감소의 평균으로 해석될 수 있다[3].

T를 학습 예제의 집합, $\{C_1, C_2, ..., C_k\}$를 클래스라고 하자.

i) 집합 T에 있는 모든 학습 예제들이 한 클래스 C_j에 속하면,
- 결정 트리는 클래스 C_j를 나타내는 하나의 노드로 구성된다.

ii) 집합 T에 학습 예제들이 없으면
- 결정 트리는 하나의 노드로 구성되지만, 노드의 클래스는 집합 T 가 아닌 다른 정보에 의해서 결정된다.

iii) 집합 T에 있는 학습 예제들이 여러 클래스에 속하면,
- 정보 이득에 의해서 결정된 하나의 속성 값에 의해서 집합 T 를 부분 집합, $T_1, T_2, ..., T_n$으로 나눈다.
- 각각의 부분 집합마다 동일한 결정 트리 구성 방법을 반복적으로 적용한다.
- i 번째 가지는 학습 예제들의 부분집합 T_i 로부터 구성된 결정 트리를 이룬다.

그림 1. 정보 이득을 이용한 결정 트리 구성 방법

그림 1 은 C4.5 에서 결정 트리를 구성하는 방법에 대한 설명이다[4]. 인스턴스의 속성 중에서 전체 학습 예제들을 가장 잘 분류하는 속성을 정보 이득 방법으로 선택해서 트리의 루트를 만든다. 루트 노드의 하위 노드들은 선택된 속성이 가질 수 있는 값의 개수만큼 만들고, 학습 예제들을 적당한 하위 노드로 분류한다. 각각의 하위 노드들에서 자신에게 할당된 학습 예제를 분류하는 방법은 루트 노드에서의 방법과 동일하다. 자신에게

속한 학습 예제들을 가장 잘 분류하는 속성을 정보 이득 방법으로 선택해서, 그 속성이 가질 수 있는 값의 개수만큼 자신의 노드 아래에 하위 노드들을 만들고 학습 예제들을 분류한다. 위의 설명처럼 예제가 남지 않게 되거나, 모두 동일한 클래스 안에 속하게 될 때까지 반복적으로 동일한 방법을 적용시킨다.

3. 실험 데이터

일반적으로 생물학 실험의 연구는 PubMed 에서 이전 실험 연구 논문들을 조사하고 분석함으로 시작된다. PubMed 는 미국 국립 보건원(National Institute of Health)에 있는 국가 생물공학 센터(National Center for Biotechnology Information)에서 제공하는 생물문헌 데이터에 관한 자료들을 검색하는 도구이다[5]. 문헌을 다루는 대부분의 전산생물학자들은 PubMed 가 방대한 생물 관련 문헌 데이터의 대부분을 다루고 있기 때문에 PubMed 에서의 검색 데이터를 가지고 연구를 시작한다.

그러나, PubMed 데이터로부터 자궁 경부암 관련 고위험군 HPV 를 분류해 내는 것은 쉬운 일이 아니다. 첫째, PubMed 데이터의 전체 양에 비해 정보의 양이 매우 적기 때문이다. 예를 들어 설명하면 PubMed 에서 HPV 와 Cervical Cancer(자궁 경부암)을 주요 단어로 검색을 하면 3,797 개의 관련 논문이 검색된다. 그러나 그들 대부분은 자궁 경부암 관련 HPV 고위험군 바이러스의 종류(type)에 대해 직접적으로는 거의 언급하지 않는다. 둘째, 자연언어 처리(Natural Language Processing)의 현재 기술이 아직은 문헌을 이해하는 수준에 이르지 못하기 때문이다. 그러므로 전체 양에 비해 정보량이 그리 많지 않은 검색 논문 데이터로부터 우리가 원하는 정보를 얻어 내기 위해서는 좀 더 세밀화된 연구가 필요하다.

이 논문에서는 PubMed 의 문헌 자료 대신 미국 로스 알라모스 국립 연구소(Los Alamos National Laboratory)에서 만든 HPV 서열 데이터베이스(HPV Sequence Database)에 있는 텍스트를 사용하였다[6]. HPV 서열 데이터베이스는 1994, 1995, 1996 그리고 1997 년 요약본을 확장한 것으로 HPV 종류의 목록과 각각의 특성과 관련 데이터들을 모아서 정리해 놓았다. 그림 2 는 이 데이터베이스로부터 만든 HPV 데이터의 예이다. 이는 HPV type 43 에 대한 예로서 각각의 데이터는 <definition>, <source> 그리고 <comment>의 세 부분으로 이루어져 있다. <definition>은 HPV 의 종류(type)와 유전자 구성에 대한 정보를, <source>는 HPV DNA 유전자의 출처를 그리고 <comment>는 HPV 의 특성들과 관련 문헌 데이터의 주석을 보여준다.

<definition>
Human papillomavirus type 43 (HPV-43), E6 region.
</definition>
<source>
Human papillomavirus type 43 DNA recovered from a vulvar biopsy with hyperplasia.
</source>
<comment>
HPV-43 was classified by Lorincz et al. [435] as a "low-risk" virus. Prevalence studies indicate that HPV-44 and HPV-43 have been found in 4% of cervical intraepithelial neoplasms, but in none of the 56 cervical cancers tested. During an analysis of approximately 1000 anogenital tissue samples, two new HPV types, HPV-43 and HPV-44, were identified. The complete genome of HPV-43 was recovered from a vulvar biopsy and cloned into bacteriophage lambda. The biopsy was taken from a woman living in the Detroit Michigan area. The DNA consisted of two fragments: a 6.3 kb BamHI fragment and a 2.9 kb HindIII fragment. The total quantity of unique DNA was 7.6 kb. Only the E6 region of the cloned sample has been sequenced, although all positions of the ORFs have been deduced and are consistent with the organization of DNA from HPV-6b. A possible feature of HPV types associated with malignant lesions is the potential to produce a different E6 protein by alternative splicing. This potential has been found in types HPV-16, HPV-18, and HPV-31. HPV-43 has both the potential E6 splice donor site at nt 233 and the potential splice acceptor at nt 413.
</comment>

그림 2. HPV 서열 데이터베이스로부터 만든 HPV 43 데이터의 예

실험 결과의 정확성을 측정하기 위해서, 미국 알라모스 국립 연구소에서 만든 HPV 서열 데이터베이스 1997 년 요약본과 우리가 만든 데이터의 <comment>를 사용하여서, HPV 위험군 분류 실험 결과와 비교될 수 있는 대조군을 만들었다. 결과는 표 1 에 나타나 있다. 이 표를 만든 과정은 다음과 같다.

우선 HPV 서열 데이터베이스 1997 년 요약본에 의해 구분된 그룹으로 각각 HPV 를 분류하였다. 그림 3 은 이 그룹들을 보여준다. 이 그림은 108 개의 Papilloma 바이러스의 L1 단백질 고정 프라이머 부분(L1 consensus primer region)을 neighbor joining 방법과

변형된 Kimura 2-parameter model 을 이용해 만든 거리 행렬(distance matrix)을 사용해서 만들어 졌다. Neighbor-joining 방법은 많은 수의 종을 연관 지어서 그룹을 만들어 주는 편리하고 빠른 방법이다. 제일 밖에 있는 회색의 원호는 Papilloma 바이러스가 A 부터 E 까지 5 개의 큰 그룹으로 이루어져 있음을 보여준다. 각각 나무의 가지들은 HPV 바이러스의 이름을 나타낸다. 대부분의 경우에는 숫자만 있는데, 40 으로 이름 붙여진 가지는 HPV40 을 의미한다.

둘째, 우리가 알고자 하는 것은 자궁 경부암과 관련된 고위험군 HPV 이므로, HPV 의 그룹이 피부와 관련된 것이면 그 그룹의 멤버들은 모두 저위험군으로 분류하였다. 셋째, HPV 의 그룹이 자궁 경부암 관련 고위험군 HPV 으로 알려진 것들만 그 그룹의 멤버들을 고위험군으로 분류하였다. 마지막으로, 그룹이 자궁 경부암과 관련된 HPV 인데, 그룹 전체적으로 위험군을 분류할 수 없는 것들의 멤버들은 우리가 만든 데이터의 <comment> 부분을 참고해서, 각각의 멤버들을 분류하였다. 표 1 은 위의 방법에 의해서, 위험도에 따라 HPV 바이러스를 분류한 것이다. Low 는 저위험군을, High 는 고위험군을 그리고 'Don't know'는 위에 있는 자료들만 가지고는 구분할 수 없는 것들을 나타낸다.

종류 (type)	위험군 (risk)	종류 (type)	위험군 (risk)	종류 (type)	위험군 (risk)	종류 (type)	위험군 (risk)
PV1	Low	HPV2	Low	HPV3	Low	HPV4	Low
HPV5	Low	HPV6	Low	HPV7	Low	HPV8	Low
HPV9	Low	HPV10	Low	HPV11	Low	HPV12	High
HPV13	Low	HPV14	Low	HPV15	Low	HPV16	Low
HPV17	Low	HPV18	High	HPV19	Low	HPV20	Low
HPV21	Low	HPV22	Low	HPV23	Low	HPV24	Low
HPV25	Low	HPV26	Don't know	HPV27	Low	HPV28	Low
HPV29	Low	HPV30	Low	HPV31	High	HPV32	Low
HPV33	High	HPV34	Low	HPV35	High	HPV36	Low
HPV37	Low	HPV38	Low	HPV39	High	HPV40	Low
HPV41	Low	HPV42	Low	HPV43	Low	HPV44	Low
HPV45	High	HPV47	Low	HPV48	Low	HPV49	Low
HPV50	Low	HPV51	High	HPV52	High	HPV53	Low
HPV54	Don't know	HPV55	Low	HPV56	High	HPV57	Don't Know
HPV58	High	HPV59	High	HPV60	Low	HPV61	High
HPV62	High	HPV63	Low	HPV64	Low	HPV65	Low
HPV66	High	HPV67	High	HPV68	High	HPV69	Low
HPV70	Don't know	HPV72	High	HPV73	Low	HPV74	Low
HPV75	Low	HPV76	Low	HPV77	Low	HPV80	Low

표 1. 자궁 경부암 관련 HPV 바이러스의 위험군 분류

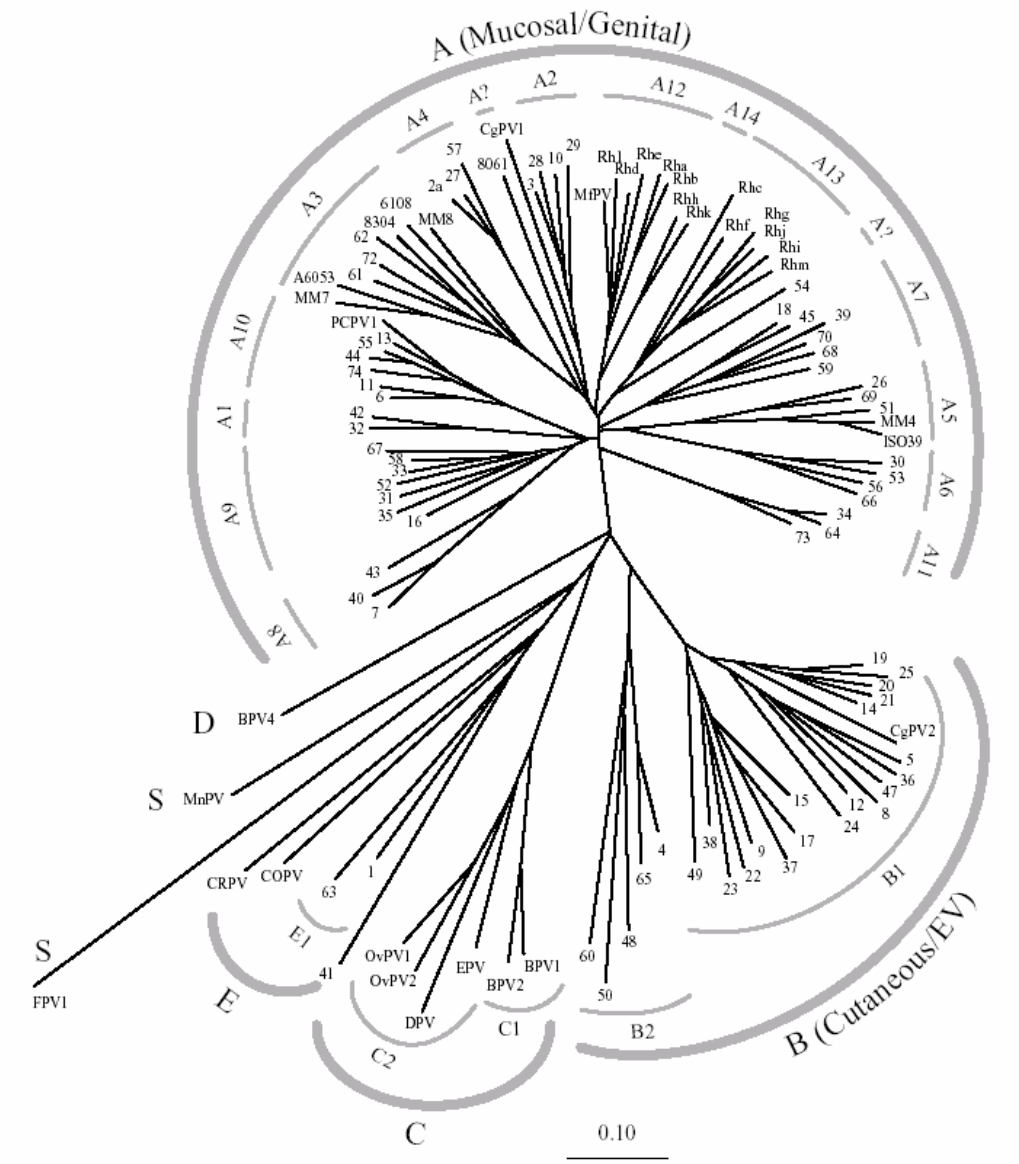

그림 3. 106 개의 Papilloma 바이러스의 L1 CPR 의 neighbor joining 분류 트리

실험에서는 HPV 데이터 중 <comment>를 text mining 할 문서로 이용하였다. 각각의 HPV 는 $tf \cdot idf$ 값을 원소로 갖는 벡터로 표현된다. $tf \cdot idf$ 에서 문서 j 에 나타나는 단어 i 의 가중치는 다음과 같이 표현된다[7, 8].

$$Weight_{ij} = tf_{ij} \cdot \log \frac{N}{df_i}$$

tf_{ij} 는 문서 j 에 나타나는 단어 i 의 빈도를 나타내고 df_i 는 단어 i 를 적어도 한번 이상 나타난 문서의 수를, N 은 전체 문서의 수를 의미한다.

Porter's algorithm [9]으로 불필요한 어미를 제거하고 Stopword list 로 불용단어를 제거했을 때, 전체 텍스트에서 추출해 낸 단어는 1,434 개였다. 그러므로 각각의 HPV 데이터는 1,434 차원의 벡터로 표현된다.

4. 실험 결과

실험은 크게 두 가지를 하였다. 하나는, 결정 트리 학습의 HPV 위험군 분류 성능을 측정하는 것이고, 다른 하나는 don't know 로 분류되기 어려운 데이터를 결정 트리는 어느 위험군으로 분류하는 지를 알아보는 것이다.

결정 트리의 HPV 위험군 분류 성능 측정, 즉 결정 트리의 정확도는, 앞에서 언급하였듯이, 기계가 다룰 데이터를 사람이 먼저 읽어서 HPV 위험군 분류를 해 놓은 대조군 데이터와 결정 트리가 분류해 놓은 실험 결과를 비교하여서 정확도가 어느 정도 되는 지 측정하였다. 전체 실험 데이터는 HPV type 이 73 개이고, 학습 예제로 58 개를, 테스트 데이터로 15 개를 사용하였다. 데이터의 양이 많지 않아서 5-fold cross validation[1]을 사용하였다. 결정 트리는 C4.5 release 8 을 사용하였다[4].

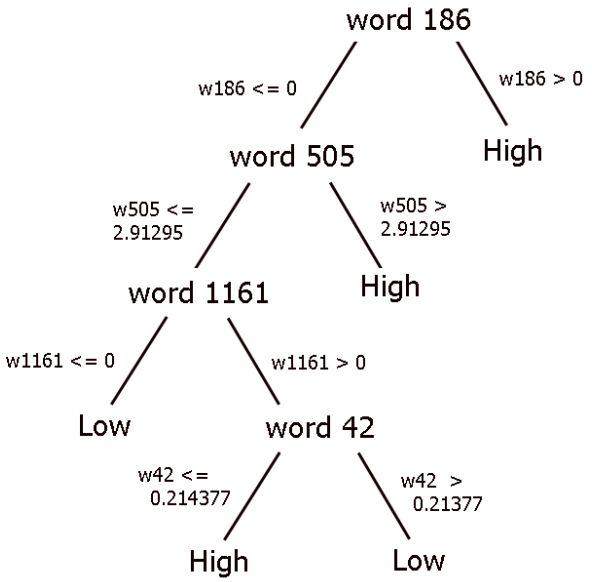

그림 4. C4.5 release 8 에 의해 만들어 진 결정 트리

[1] *k*-fold cross validation은 전체 데이터 집합을 k개의 부분 집합으로 나누어서 실험을 k번 반복하는 방법이다. 실험이 시행될 때마다, k개의 부분 집합 중 하나가 실험 데이터로 사용되고 나머지 k-1 개의 부분집합은 학습 예제로 사용된다. 성능(performance)은 k 번 시도에 대한 평균 값으로 계산한다.

그림 4 는 결정 트리의 한 예이다. 추출한 1434 개의 단어 중 186, 505, 1161 그리고 42 번째 단어가 사용되어서 분류되었음을 알 수 있다. 트리에서 분류 기준으로 사용된 단어들은 cervix, error, neoplasia, patient, risk, vagina, wart 처럼 자궁 경부암이나 피부에 생기는 사마귀 등에 직접 관련 있는 단어들도 있고, bind, codon, grade 처럼 직접적인 관련이 없는 단어들도 있었는데, 자궁 경부암과 관련 있는 단어들이 많은 것으로 보아 결정 트리의 분류가 사람이 보기에도 어느 정도는 논리적으로 이루어 지고 있음을 알 수 있었다.

정확도 측정은 표 1 에서 만든 대조군과 일치하는 HPV 의 수 c 를 전체 비교한 HPV 데이터의 수 N 으로 나누어서 계산하였다.

$$accuracy = \frac{c}{N} \cdot 100\,\%$$

실험 시행 횟수	정확도(%)
1	86.7
2	86.7
3	66.7
4	73.3
5	92.3
평균(%)	81.14 ± 10.68

표 2. 실험의 정확도 측정

표 2 에 나타나 있듯이 정확도는 81.14 ± 10.68% 이다. 잘못 분류한 HPV 종류를 살펴보면, 저위험군을 고위험군으로 분류한 것이 9 개, 고위험군을 저위험군으로 분류한 것이 5 개이다. 저위험군을 고위험군으로 분류한 9 개 중 4 개(13, 14, 30, 40)는 자궁 경부암과는 상관 없지만, 후두암을 유발하거나 일부가 암으로 진행될 가능성이 있는 것들로 자궁 경부암을 일으키는 것과 비슷한 표현들을 가질 수 있는 것들이다. 고위험군을 저위험군으로 분류한 5 개 중 3 개(59, 62, 72)는 결과를 간단히 하기 위해 고위험군으로 분류하기는 했지만, 임상 학자들은 중간 정도의 위험을 가진 것으로 분류하므로 자궁 경부암과 관련된 직접적인 표현이 그리 많지 않을 수 있는 것들이다. 그러므로 위의 7 개를 제외하면 저위험군을 고위험군으로 분류한 것 5 개(2, 42, 43, 44, 53)와 고위험군을 저위험군으로 분류한 것 2 개(18, 56)를 잘못 분류한 것으로 볼 수 있다. 잘못 분류된 것은 표 3 에 정리되어 있다.

HPV type	위험군	결정 트리에 의해 분류된 위험군	특징
HPV 2	저위험군	고위험군	일반적인 사마귀
HPV 13	저위험군	고위험군	구강 감염, 일부는 암으로 진행
HPV 14	저위험군	고위험군	일부는 암으로 진행
HPV 18	고위험군	저위험군	
HPV 30	저위험군	고위험군	후두암
HPV 40	저위험군	고위험군	후두암
HPV 42	저위험군	고위험군	생식기 사마귀
HPV 43	저위험군	고위험군	
HPV 44	저위험군	고위험군	
HPV 53	저위험군	고위험군	
HPV 56	고위험군	저위험군	
HPV 59	고위험군	저위험군	
HPV 62	고위험군	저위험군	중간 위험군으로 분류되기도 함
HPV 72	고위험군	저위험군	중간 위험군으로 분류되기도 함

표 3. 잘못 분류된 HPV 의 종류

주어진 실험 데이터만으로는 분명히 알 수 없어서 don't know 로 분류되었던 것들(HPV26, 54, 57, 70)을 결정 트리는 어느 위험군으로 분류하는 지 알아보았다. 다른 문헌 데이터[10,11,12,13]를 참고하여서 대조군을 설정하고 결정 트리가 각각의 don't know HPV 를 어느 위험군으로 분류하는 지 알아보았다. 학습 예제는 앞의 실험에서 사용하였던 58 개를 같은 방법으로 5 번 설정해 주었고, 테스트 데이터는 위의 don't know HPV 4 개를 사용하였다. 실험 결과는 표 4 에 나타나 있다. 전체 평균은 55 ± 20.92% 이다. 특히, HPV 70 는 제대로 구분해 내지 못했다. 이는 HPV 70 에 대한 연구가 많이 이루어 지지 않아서, 실험 데이터로 사용했던 HPV70 <comment>에 위험도에 대한 내용은 없고, HPV 70 가 환자의 자궁 경부에서 발견되어 서열이 분석되었다는 정도 밖에 없기 때문이다.

시행 횟수	HPV 26 저위험군	HPV 54 저위험군	HPV 57 저위험군	HPV 70 고위험군	평균 (%)
1 회	저위험군	저위험군	저위험군	저위험군	75
2 회	저위험군	저위험군	고위험군	저위험군	50
3 회	저위험군	저위험군	고위험군	저위험군	50
4 회	저위험군	저위험군	저위험군	저위험군	75
5 회	고위험군	저위험군	고위험군	저위험군	25

표 4. don't know 의 HPV 위험군 분류 결과

5. 결론 및 고찰

첫째, 결정 트리로 HPV 위험군을 분류한 뒤, 사람에 의해 만들어진 대조군 데이터로 결정 트리의 분류 성능을 측정하여 보았다. 정확도는 약 82 %로 기대했던 것보다 높게 나왔다. 그러나 이 실험 데이터를 참고 자료가 아닌 임상 실험 전의 기초 자료로 사용하려 한다면 정확도가 더 높아야 한다. HPV 저위험군을 고위험군으로 분류하는 경우는 실수(error)의 위험 부담이 그리 높지 않겠지만, HPV 고위험군을 저위험군으로 잘못 분류하는 경우에 위의 실험 결과처럼 에러률이 약 20 % 정도 될 경우에는 위험 부담이 크다. HPV 고위험군 하나의 감염도 자궁 경부암을 유발시킬 수 있는 매우 중요한 원인 중 하나인데, 이 실험 결과를 토대로 만든 HPV 진단용 DNA chip 은 모든 고위험군 HPV 를 진단해 낼 수 없기 때문이다. 만일 이 DNA chip 이 진단용으로 사용된다면 분명 오진을 일으키게 될 것이다. 둘째, 분류되기 어려운 HPV 종류를 결정 트리는 어느 위험군으로 분류하는 지를 알아보았다. 전체 평균은 약 55 %로 그리 높은 편은 아니었지만, 사람이 읽었을 때도 분류해 내기 어려운 것들이었기 때문에, 이 실험 결과는 임상 실험 전 기초 참고 자료로서 믿고 사용할 만한 결과라고 생각된다.

바이러스는 숙주 내에서 살아 남기 위한 생존 수단으로 돌연변이를 쉽게 일으키기 때문에, 대부분의 바이러스는 HPV 처럼 종류가 매우 다양하며 또 각 종류마다 심각한 질병을 유발시키는 지의 여부가 다르다. 그러므로, 질병 관련 여부에 따라 HPV 처럼 바이러스를 종류나 변종을 구별해 내야 하는 필요성이 있으므로 이와 비슷한 연구들이 계속 되어야 할 것이다. 그러나, HPV 의 경우는 연구가 많이 되어 문헌 자료도 많고 데이터베이스도 잘 정리되어 있어서 비교적 text mining 하기 쉬운 경우였다고 생각된다. 따라서 앞으로 바이러스의 변종 분류와 같은 실험을 하기 위해서는 잘 정리되지 않고 엉성한 문헌 자료들을 사용해야 하므로 좀 더 세밀화된 연구가 진행될 필요성이 있다.

감사의 글

이 논문은 교육부 BK21 사업과 과기부 BrainTech 프로그램에 의하여 지원되었습니다.

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, (1993). "Epidemiologic Evidence Showing That Human Papillomavirus Infection Causes Most Cervical Intraepithelial Neoplasis," *Journal of the National Cancer Institute*, 85, pp. 958-964.

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

[3] Mitchell, T.M. (1997). *Machine Learning,* The McGraw-Hill Companies, Inc.

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

[5] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed

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

[7] Joachims, T. (1997). "A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization," *Proceedings of the 14th International Conference on Machine Learning,* pp. 143-151.

[8] Freund, Y., R.E. Schapire, (1996). "Experiments with a New Boosting Algorithm" *In Proceedings of the 13th International Conference on Machine Learning*, pp. 148-156.

[9] Porter, M. (1980). "An Algorithm for Suffix Stripping," *Program*, 14, pp. 130-137.

[10] Nuovo, G.J., C.P. Crum, E.M. De Villiers, R.U. Levine, S.J. Silverstein, (1988) "Isolation of a novel human papillomavirus (type 51) from a cervical condyloma," *Journal of Virology*, 62, pp. 1452-1455.

[11] Favre, M., D. Kremsdorf, S. Jablonska, S. Obalek, G. Pehau-Arnaudet, O. Croissant, G. Orth, (1990). "Two new human papillomavirus types (HPV54 and 55) characterized from genital tumours illustrate the plurality of genital HPVs," *International Journal of Cancer*, 45, pp. 40-46.

[12] Chan, S.Y., S.H. Chew, K. Egawa, E.I. Grussendorf-Conen, Y. Honda, A. Rubben, K.C. Tan KC, H.U. Bernard, (1997). "Phylogenetic analysis of the human papillomavirus type 2 (HPV-2), HPV-27, and HPV-57 group, which is associated with common warts," *Virology*, 239, pp. 296-302.

[13] Meyer, T., R. Arndt, E. Christophers, E.R. Beckmann, S. Schroder, L. Gissmann, E. Stockfleth, (1998). "Association of rare human papillomavirus types with genital premalignant and malignant lesions," *The Journal of Infectious Diseases*, 178, pp. 252-255.