Kernel 기반 학습을 이용한 HPV의 위험군 분류

정제균[O1,2] 오석준[2] 장병탁[1,2,3]
[1]서울대학교 생물정보학 협동과정, [2]서울대학교 바이오정보기술 연구센터
[3]서울대학교 컴퓨터공학부
{jgjoung[O], sjaugh, btzhang }@bi.snu.ac.kr

HPV Risk Classification Using Kernel Based Learning

Je-Gun Joung[O1,2] Sirk June Augh[2] Byoung-Tak Zhang[1,2,3]
[1]Interdisciplinary Program in Bioinformatics, Seoul National University
[2]Center for Bioinformation Technology, Seoul National University
[3]School of Computer Science and Engineering, Seoul National University

요 약

인유두종바이러스(human papillomavirus: HPV)는 감염되었을 때 각종 악성 종양을 유발할 수 있는 작은 DNA 바이러스이다. 고위험군에 속하는 HPV의 감염은 암으로 진행될 수 있는 가능성이 크다. 본 논문은 HPV를 분류할 수 있는 기계 학습 기법을 제안하고자 한다. 제안된 학습 기법은 단백질 서열을 효과적으로 분류할 수 있는 커널(kernel) 방법에 기반을 두고 있다. 위험군 분류는 감염의 메커니즘의 이해와 유전자칩과 같은 새로운 의학 도구의 개발 등에 있어서 중요한 정보를 제공해 줄 수 있다. 실험 결과는 중요한 부위의 탐색에 의한 커널 기반의 학습 방법이 우수한 성능을 보이는 것으로 나타났다.

1. 서 론

인유두종바이러스(human papillomavirus: HPV)는 약 8kb 의 환상의 이중나선 DNA 바이러스로서 여성의 자궁경부암을 유발하며 여러 가지 악성 종양과 밀접한 관계가 있는 것으로 알려져 있다. HPV 는 지금까지 85 종에 달하는 유전형(genotype)의 염기서열이 완전히 밝혀져 있으며 120 여 개의 새로운 HPV 유전형 구조가 부분적으로 보고되고 있다[1].

특히 자궁 경부암과 관련된 HPV 는 악성 종양 유발 가능성에 따라 고위험군(high-risk type)과 저위험군(low-risk type)으로 나뉜다. 예를 들어, HPV 16, 18, 31 과 같은 고위험 HPV 에 감염될 경우 악성 종양으로 진행될 가능성이 높아진다. 따라서, 감염된 HPV 의 type 을 파악하는 것이 환자의 처방 및 예후에 매우 중요하게 작용하게 된다. 기존에는 이러한 HPV 의 위험군 분류를 수행할 때 생물학자가 직접 수많은 문헌 자료로부터 조사하는 수작업이 필요했다. 하지만 최근에는 텍스트 마이닝(text mining) 연구가 활발해 짐에 따라 컴퓨터를 이용한 자동 분류가 가능해졌다[2]. 그 예로서 결정 트리(decision tree)를 이용한 텍스트 마이닝(text mining) 기법이 HPV 의 자동 분류에 적용된 사례가 있다. 이 시스템은 먼저 대량의 관련 문서에서 자동적으로 위험군을 분류한 다음 해당 분야의 전문가가 검증하는 절차를 거친다는 점에서 분류 작업의 효율을 높일 수 있는 장점이 있다.

텍스트 마이닝 기법을 이용한 HPV분류의 단점은 새로운 유전형이 발견되었을 경우 현재 연구된 논문들이 없다면 수행할 수 없다는 것이다. 따라서 이러한 경우 텍스트 기반이 아닌 새로운 접근 방법이 필요하게 된다.

본 논문은 텍스트 정보가 아닌 단백질 서열 정보를 이용하여 HPV 의 위험군을 분류할 수 있는 새로운 방법을 제시한다. 제시된 커널 기반 분류시스템은 HPV 의 위험군을 판별하기 위해서 내부 핵심 알고리즘으로 문자열 커널(string kernels)을 내장하고 있다. 문자열 커널은 서열의 쌍들을 특성값 공간(feature space)으로 사상(map)하는 기능을 가지고 있다. 문자열 커널 기법은 단백질 서열에서 중요한 정보를 추출할 수 있기 때문에 생물학 관련 서열 데이터 분석에 적합하다.

2. HPV유전형 분류를 위한 커널 기반 시스템

그림 1 은 HPV 유전형 분류를 위한 전체적인 시스템의 개념을 보여 주고 있다. 실선 화살표는 학습 처리 과정의 흐름을 나타낸다. 여기서 데이터 전처리는 HPV 데이터베이스로부터 서열데이터를 검색해서 학습 데이터를 생성하기까지의 단계이다. 시스템은 우선 HPV 데이터베이스에서 각 유전형에 대하여 라벨이 붙어 있는 위험군에 해당 하는 단백질 서열들을 추출하게 된다. 다음 단계로 각 서열들의 위치가 서로 동일하도록 다중 서열 정렬(multiple sequence alignment)을 수행한다. 이 결과에서 고위험군과 저위험군 서열들을 분리한 후, 은닉 마코프 모델로 학습을 하게 된다. 이 단계는 전체 길이의 서열상에서 두 위험군을 분류하기 가장 용이한 지점(point)을 알아 내기 위한 작업이다. 전체 서열 중에서 이 지점들에 해당하는 특정 길이 또는 윈도우(window) 크기를 가진 부서열(subsequence)들이 최종적으로 학습 데이터에 사용된다.

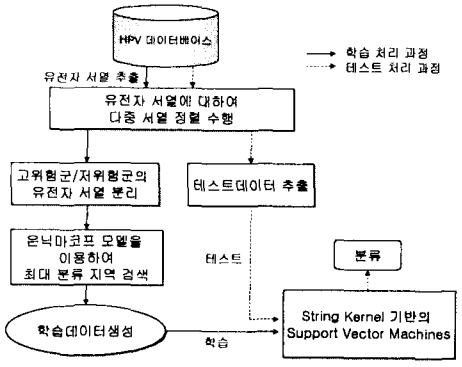

그림 1. HPV 위험군 분류를 위한 시스템의 전체 개념도

은닉 마코프 모델을 이용한 지점 탐색에 있어서 부서열 집합 S의 분류 가능성 정도를 score(S)라고 정의한다. 여기서 스코어(score)는 저위험군이 고위험군의 유전형 모델에 적합할 확률의 상대적인 비율, $\log(P_M(S_{Pos})/P_M(S_{Neg}))$을 계산한 값이다. 스코어가 높을수록 위험군 분류가 용이한 지점이라고 할 수 있다.

이렇게 생성된 학습 데이터는 단백질 서열인 문자열들의 집합으로 표현된다. 문자열 커널 함수는 문자열들을 특정 수치로 변환하여 특성값 공간으로 사상시킨다. SVM은 문자열 커널 함수를 이용하여 분류를 위한 최적화된 파라미터를 학습하게 된다. 일단 학습이 완료되면 미지의 HPV서열에 대하여 예측을 수행할 수 있다.

3. Support Vector Machines

Support Vector Machines는 통계적 학습 이론을 기반으로 1995년 Vapnik에 의해서 개발되었고, 훈련된 모델을 기반으로 하여 주어진 데이터를 분류하는 문제에 적합하다[3]. 이 학습 기법은 기계 학습 분야에 있어서 전형적으로 2개 또는 그 이상의 그룹으로 분류(classification)되는 문제에 대하여 먼저 훈련 데이터를 학습하고 테스트 데이터를 예측하는 감독 학습 기법으로 취급된다.

SVM 기법은 이런 분류자를 이용하여 입력 공간의 비선형적인 높은 차수를 특성값 공간(feature space)에서 선형적으로 투영하여 해석할 수 있도록 하며, 각 특성값 사이의 최적의 경계(boundary)를 제시한다. 입력 공간에서 특성값 공간 ϕ로 사상시키는 작업은 먼저 학습 데이터 $S=\{x_i, y_i\}$, $i=1,\ldots,n$ 을 $\Phi(S)=\{\Phi(x_i), y_i\}=\{z_i, y_i\}$, $i=1,\ldots,n$ 으로 사상시키는 과정이다. 특성값 공간에서 SVM은 선형 분리 함수(linear discriminant function) $f(z)=\langle w\cdot x\rangle+b$ 를 학습한다. 그래서 hyperplane $f(z)=0$은 음성 데이터로부터 양성 데이터를 분리해 낸다. 어떤 hyperplane $f(z)=\langle w\cdot x\rangle+b=0$에 대하여 최근접 $z^*\in\{z_1,\ldots,z_n\}$의 유클리드 거리(Euclidean distance)를 hyperplane의 마진(margin)이라고 부른다. 만약 hyperplane을정규화(normalization)한다면 hyperplane의 마진은 $1/\|w\|$이 된다.

SVM의 최적화는 2차 프로그래밍(quadratic programming) 문제와 동일시된다. 최대 마진 분류자(maximal margin classifier)는 다음 식의 α를 최적화함으로써 발견할 수 있다.

$$\text{maximize} \quad \sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j\langle z_i\cdot z_j\rangle,$$

$$\text{subject to} \quad \alpha_i\geq 0 \ \ (1\leq i\leq n), \sum_{i=1}^{n}\alpha_i y_i=0.$$

여기서 α는 Lagrange multiplier 라고 불린다. 최적의 $\alpha_1^*,\ldots,\alpha_n^*$ 을 풀기 위해서 최대 마진 hyperplane $f^*(z)=0$은 아래와 같은 파라미터들 관점에서 이중적인 표현으로 기술될 수 있다.

$$f^*(z)=\sum_{i=1}^{n}\alpha_i^* y_i\langle z_i\cdot z\rangle+b^*$$

$$b^*=y_s-\sum_{i=1}^{n}\alpha_i^* y_i\langle z_i\cdot z_s\rangle \text{ for some } \alpha_s^*\neq 0$$

이러한 이중 표현은 다음과 같이 단지 특성값 사상 Φ의 내적에 의한 다양한 커널 기법들을 허락할 수 있게 하였다.

$$\langle z_i\cdot z_j\rangle=\langle\Phi(x_i)\cdot\Phi(x_j)\rangle,$$

$$K(x_i\cdot x_j)=\langle\Phi(x_i)\cdot\Phi(x_j)\rangle$$

여기서 이러한 함수들 K를 커널 함수라고 한다.

4. HPV위험군 분류를 위한 Mismatch 문자열 커널 기법

Mismatch 문자열 커널은 spectrum 커널을 개선한 것이다[4]. 먼저 생물학 서열에 있어서 k-spectrum 이란 k-길이를 가진 부서열(subsequences)들의 집합을 의미한다. 생물정보학에 있어서 spectrum 이란 용어는 k-길이를 가진 부서열을 이용하여 교배(hybridization)에 의한 시퀀싱 기법 연구에 언급된 바 있다[5].

이러한 k-길이를 가진 부서열을 이용하여 단백질 서열 데이터로부터 중요한 정보를 추출하여 특성값 벡터로 사상시키는 커널을 spectrum커널이라고 한다.

Spectrum 커널은

$$\Phi_k(x)=(\phi_\alpha(x))_{\alpha\in A^k}$$

과 같이 특성값으로 사상시키는 식으로 표현된다. 여기서 $\phi_\alpha(x)$는 서열 x에서 k-mer (k-길이를 가진 부서열) α가 나온 빈도수를 나타낸다. α는 20 개의 아미노산(amino acid)들로 구성된 k-mer 의 알파벳 A^k 의 원소이다. 그래서 두 서열 x_i, x_j의 k-spectrum:

$$K_k(x_i\cdot x_j)=\langle\Phi_k(x_i)\cdot\Phi_k(x_j)\rangle$$

는 내적을 취함으로써 계산된다. 만약 두 서열이 같은 k-mer 를 많이 공유하고 있다면 큰 수의 k-spectrum 값을 가지게 된다.

Mismatch 문자열 커널은 이를 확장하여 k-mer 에 있어서 m 개의 알파벳에 대하여 mismatch 를 허용할 수 있게 한다. 이러한 방법을 적용한 커널을 (k, m)-mismatch 커널로 정의한다. 여기서는 α의 빈도수 벡터를 계산하는 것이 아니라 k-mer 인 α가 알파벳열 a_1

$a_2...a_k$ 이라면 α 로 인하여 만들어 질 수 있는 (k, m)의 패턴들 β의 확률값을 계산한다. 즉, 커널은 β의 발생 확률을 계산하는 $\phi_\beta(\alpha) = P(b_1 \mid a_1)P(b_2 \mid a_2)...P(b_k \mid a_k)$로 교체해 주면 된다. 여기서 아미노산 a에서 아미노산 b로 일치 또는 불일치될 확률 $P(b \mid a)$는 아미노산 a에서 아미노산 b 의 치환(substitution) 확률로 생각할 수 있으며 PAM[12]이나 BLOSUM[6] 행렬을 이용할 수 있다. 따라서 β에 대한 α의 mismatch 문자열 커널은

$$\Phi_{(k,m)}(\alpha) = \left(\phi_\beta(\alpha)\right)_{\beta \in A^k}$$

로 나타낼 수 있고 입력 서열 x는 다음 식

$$\Phi_{(k,m)}(x) = \sum_{k-mers\ \alpha \text{ in } x} \Phi_{(k,m)}(\alpha)$$

에 의하여 특성값 백터로 사상된다. 따라서 최종적인 mismatch 문자열 커널의 함수 식은

$$K_{(k,m)}(x_i \cdot x_j) = \left\langle \Phi_{(k,m)}(x_i) \cdot \Phi_{(k,m)}(x_j) \right\rangle$$

로 정의할 수 있다.

5. 실험 및 결과

실험 데이터로써 8 개의 유전자 중에서 E6 선택했으며 현재까지 밝혀진 약 80 여 개의 유전형에 대하여 아미노산 서열을 고려하였다. E6 를 선택한 중요한 이유는 기능적으로 몇몇 부위가 위험도에 있어서 직접적인 관련이 있는 것으로 밝혀져 있기 때문이다.

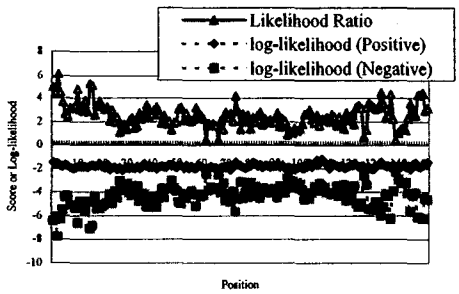

그림 2. E6 유전자에 대하여 은닉 마코프 모델로 각 서열의 위치에 따른 스코어 값

그림 2 는 E6 유전자에 대하여 은닉 마코프 모델로 각 서열의 위치에 따른 스코어를 얻어낸 것이다. 가장 위에 있는 선은 스코어를 나타내고 중간에 출력된 선은 양성에 대한 log-likelihood 를 나타낸다. 그리고 가장 아래에 있는 선이 음성에 대한 log-likelihood 값을 나타낸다. 스코어가 높은 지역의 서열들을 추출하여 SVM 의 데이터로 활용한다.

그림 3 은 전체 서열을 이용하여 분류했을 경우와 은닉 마코프모델에 의한 스코어링 방법을 통하여 분류했을 경우의 ROC (receiver-operating characteristic) 스코어를 보여 주고 있다. 전체 서열을 통해서 분류할 경우, 은닉 마코프 모델에 의한 스코어링 방법에 비해서 좋은 예측 결과를 보이지 못하고 있다. 전체 서열을 이용하여 분류할 경우, 또 한가지 단점은 서열이 길기 때문에 커널

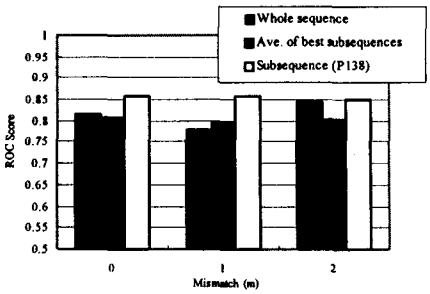

그림 3. 전체 서열을 수행한 것과 likelihood 스코어를 고려한 것과의 비교

행렬을 계산하는 시간이 길다는 것이다. 이러한 계산시간은 은닉마코프모델을 수행하는 시간과 비슷한 반면 좋은 예측 성능을 보이지는 못한다는 것을 알려 준다.

6. 결론

본 논문은 HPV 의 위험 여부를 판별하기 위해서 커널 기반의 분류 시스템을 소개하였다. HPV 위험군 분류를 위한 커널 기반의 분류 시스템은 서열 데이터를 학습하기 위해서 문자열 커널을 사용하였다. 문자열 커널은 서열 정보를 특성값 공간으로 사상시킴으로써 분류의 효율성을 향상시켰다. 대부분의 분류 알고리즘은 서열 정보를 학습하기에 적합하지 않지만, 본 연구에서 적용한 방법에 따르면 SVM 의 커널함수를 간단히 문자열 커널 함수로 교체함으로써 용이한 학습 효과를 가져올 수 있음을 확인하였다.

감사의 글

이 논문은 과학기술부의 국가지정연구실 사업과 IMT-2000 과제에 의하여 지원되었음.

참고문헌

1. zur Hausen, H., Papillomaviruses causing cancer: evasion from host-cell control in early events in carcinogenesis, *Journal of National Cancer Inst* ., Vol. 92, pp.690-698, 2000.
2. 황소현, 박성배, 장병탁,, 결정 트리에 의한 인유두종 바이러스의 위험군 분류, 한국데이터마이닝학회 추계 학술대회 논문집, pp. 148-160, 2002.
3. Vapnik, V. N., *Statistical learning theory*, Springer, 1998.
4. Leslie, C., Eskin, E., Weston, J. and Noble, W., Mismatch string kernels for SVM protein classification, *NIPS* 2002. (to appear)
5. Peer, I. and Shamir, R., Spectrum alignment: efficient resequencing by hybridization, *In ISMB*, pp. 260-268, AAAI Press, 2000.
6. Henikoff, S. and Henikoff, J. G., Amino acid substitution matrices from protein blocks, *PNAS*, Vol. 89, pp. 10915-10919, 1992.