기계학습/텍스트마이닝과 생명과학

경북대학교 박성배

1. 개 요

인간 유전체 프로젝트(human genome project)의 첫 번째 사업 종료 이후에, DNA 서열의 분석을 위해 생물학이나 의학 연구에서 컴퓨터를 사용하는 것은 이제 거의 필수가 되었다. 또한, 최근에는 DNA 마이크로어레이(DNA microarray) 분석을 위해 컴퓨터의 사용이 더욱 늘어나고 있다. DNA 마이크로어레이 분석연구나 질량 분석기의 사용은 엄청난 양의 데이터를 필요로 하고 기존의 생물학적 접근방법론으로는 이런 대량의 데이터를 처리할 수 없다. **데이터의 수가 적고 문제가 간단할 때에는 기존에 전산학에서 널리 알려져 있던 알고리듬들이 잘 적용되었고 유용하게 사용되었다. 하지만, 이런 방식으로는 생물체 본연의 복잡성으로 인해 생명체를 분자 수준에서 이해하기 위한 모델을 만드는 것이 거의 불가능하다. 따라서 이런 대량의 데이터를 컴퓨터로 처리하기 위해, 기계학습(machine learning) 혹은 데이터마이닝(data mining) 기법이 생물학이나 의학 연구에서 점점 더 중요하게 되었다.**

기계학습은 컴퓨터의 속도가 빨라짐에 따라 복잡하면서도 양이 많은 데이터를 다루는데 적합한 방법으로 인식되고 있다. 1980년대 이후, 컴퓨터의 발전 속도가 크게 증가하면서 기계학습은 매년 거의 두 배씩 증가하고 있는 생물학 데이터를 다룰 수 있는 거의 유일한 방법론으로 여겨지고 있다. 또한, 기계학습은 데이터마이닝의 가장 주요한 방법론으로서도 작동하고 있다. 데이터마이닝은 일반적으로 "대량의 데이터로부터 유용한 정보를 추출하여 이해하기 쉬운 형태로 변환하여 실제의 의사결정과정에 적용하는 과정"으로 생각되어 지는데, 여기서 대량의 데이터에는 생물학 데이터베이스(database)도 포함된다. 이미 존재하는 데이터로부터 인지적인 방법으로 필요한 지식을 추출하기 어려울 때 데이터마이닝 기법이 사용되는데, 아직까지 인간은 생물체의 생명 활동에 대한 정확한 지식을 가지고 있지 못하므로 생물학적 지식을 찾기 위해 생물학 실험 데이터로부터 필요한 지식을 뽑아내는 데이터마이닝 기법이 생명과학 분야에서 매우 유용하다.

최근에는 생의학 논문으로부터 필요한 지식을 자동으로 추출하는 연구가 많이 진행되고 있다. 생의학 관련 논문들은 1966년 이래 MEDLINE과 같은 데이터베이스에 이미 잘 정리되고 있으므로, 이런 데이터베이스로부터 필요한 지식을 추출하는 연구의 필요성도 크게 증대되고 있다. **논문은 일반적으로 텍스트(text) 형태로 되어 있는데, 논문이나 신문 기사로부터 지식을 추출하는 텍스트마이닝(text mining)에는 기존의 데이터마이닝 기법 이외에도 정보검색(information retrieval), 자연언어처리(natural language processing)에서 사용되는 방법들이 함께 사용된다.**

본 논문은 생명 과학 분야에 기계학습 또는 텍스트마이닝 기법이 어떻게 적용될 수 있는지 소개하는 것을 목적으로 하며, 다음과 같이 구성된다. 2절에서는 여러 가지 기계학습 방법들에 대한 소개를 하고 각 방법들이 생의학 문제에 어떻게 적용되었는지 살펴본다. 3절에서는 특별히 텍스트마이닝 문제가 생명 과학의 어떤 문제를 다룰 수 있는지 소개하고, 끝으로 4절에서 결론을 맺는다.

2. 생명 과학과 기계학습

기계학습의 목적은 주어진 데이터로부터 좋은 가설(hypothesis) 혹은 모델(model)을 만드는 것이다. 크게 보아 기계학습 방법들은 (i) 감독 학습(supervised learning)과 (ii) 비감독 학습(unsupervised learning)으로 나누어진다. 감독 학습은 그림 1과 같은 형태로 학습이 일어난다. 환경으로부터 샘플링 된 문제 x에 대해, 학습자(learner)가 답 y를 출력하면 감독자(supervisor)의 정답 d와 비교하여 학습자에게 다시 피드백을 주는 과정을 통해 학습자가 문제를 배우게 된다. 비록 학습자가 틀린 답을 제시하였다 하더라도 피드백을 받아서 자신의 가설을 변경하여 다음에는 틀리지 않게 한다. 이에 비해, 비감독 학습에서는 학습자가 비록 환경과 상

호 작용을 통해 학습을 하기는 하지만 감독자 없이 학습을 수행한다. 따라서 생명 과학에서 널리 쓰이는 기계학습 기법은 주로 감독 학습이지만, 비감독 학습도 데이터 가시화 측면에서 널리 쓰이고 있다.

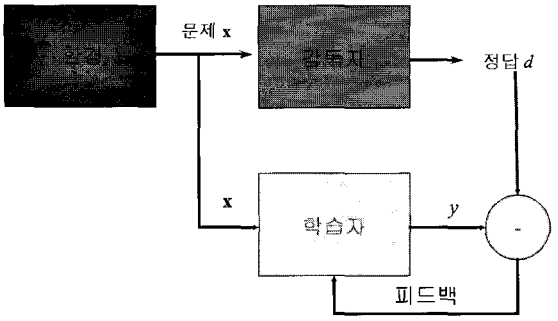


그림 1 감독자 학습의 학습 모델

기계학습 방법을 생명 과학 분야에 적용시킬 때 부딪히는 주요한 문제 중 하나는 과연 생물학 데이터가 기계학습에 필요한 형태로 표현될 수 있는가 하는 점이다. 일반적으로 생명체의 활동이나 진화와 관련된 정보들은 디지털 심볼들의 연속된 열로서 표현가능하다. 즉, 흔히 DNA, RNA, 혹은 단백질 속의 뉴클레오티드(nucleotide)나 아미노산 단량체(amino acid monomer)들은 알파벳 심볼로 표현된다. 따라서 대부분의 생명 과학 문제들은 기계학습 방법으로 푸는데 큰 어려움이 없다.

기계학습 방법들을 나누는 또 다른 방법 중의 하나는 (i) 통계적인 방법과 (ii) 비통계적인 방법으로 나누는 것이다. 통계적인 방법은 대체로 널리 알려진 베이스 정리(Bayesian Theorem)에 기반을 두고 있다. 베이지안 방법론은 새로운 모델이나 가설을 창조하는 것이 아니라 주어진 지식과 데이터를 사용해서 가설을 평가하는 것이지만, 불확실성이 존재할 때 추론하는 일관된 방법을 제시하므로 기계학습 분야에서 널리 쓰인다.

비통계적인 방법은 신경망(artificial neural networks), 결정트리(decision trees), support vector machines처럼 여러 가지 형태가 있다. 이들은 그 특성에 따라, 잘 적용되는 분야와 문제가 결정된다. 표 1은

표 1 여러 가지 기계학습 방법의 분류

학습 방법	감독/비감독	통계적/비통계적
결정트리 (decision tree)	감독	비통계적
신경망 (neural networks)	감독	비통계적
자기구성지도 (self organizing map)	비감독	비통계적
나이브 베이스 (naive Bayes)	감독	통계적
주성분 분석 (PCA)	비감독	통계적
베이시안 네트워크 (Bayesian Network)	비감독	통계적
은닉마코프모델 (hidden Markov Model)	감독	통계적
유전자 알고리듬 (genetic algorithms)	감독	비통계적
SVM (Support Vector Machine)	감독	비통계적

몇 가지 대표적인 기계학습 방법들을 분류해 놓은 것이다. 이 절에서는 이 중에서 자주 쓰이는 대표적인 학습 방법이 어떻게 생명 과학 문제에 적용되는지 알아본다.

2.1 결정트리

결정트리는 그림 2처럼 트리 구조로 표현되며, 트리의 각 내부 노드(internal node)는 한 특성(attribute)에 대한 검사를 나타내고, 각 가지(branch)는 그 검사의 결과를 표현한다. 그리고 잎 노드(leaf node)는 분류하고자 하는 클래스를 나타낸다. 그림 2의 결정트리는 자동차 운전자의 사고 위험성을 나타내고 있는데, 나이가 31세 이하이거나 자동차 종류가 스포츠카이면 사고의 위험이 높음을 보이고 있다. 결정트리는 데이터마이닝에서 비교적 널리 쓰인다. 이는 결정트리가 쉽게 if-then 형식의 규칙으로 표현될 수 있어서 학습 결과를 이해하기가 쉽기 때문이다. 예를 들면, 그림 2의 결정트리는 아래의 그림과 같은 3개의 규칙으로 표현될 수 있다.

(i) **if** (Age 〈 31) **then** Output = High
(ii) **if** (Age ≥ 31 **and** Car Type is Sports) **then** Output = High
(iii) **if** (Age ≥ 31 **and** Car Type is **NOT** Sports) **then** Output = Low

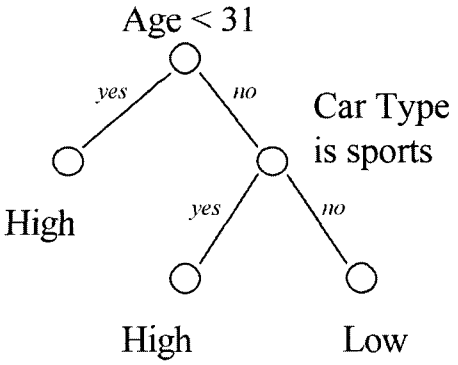


그림 2 결정트리의 예

데이터만 주어지면 결정트리는 데이터로부터 직접 구축될 수 있는데, 이 때 가장 널리 사용되는 알고리듬이 ID3[1]이다. 또한, 결정트리는 C4.5[2]처럼 패키지화되어 있는 도구가 있어 비교적 쉽게 여러 가지 문제에 적용해 볼 수 있다. Park et al.은 인유두종 바이러스(human papillomavirus)의 위험 군을 분류하는데 결정트리 학습을 사용하였고[3], Beerenwinkel et al.은 HIV-1의 약제내성(drug resistance)을 모델링하기 위해 결정트리를 사용하였다[4]. Dettling과 Buehlmann은 종양의 종류를 구분하기 위해서 결정트리에 부스팅 기법을 적용하였다[5]. 또한, Lin et al.은 유전자 이름을 알려진 것과 알려지지 않은 것으로 구분하기 위해 결

정트리를 텍스트 분류에 적용하였다[6].

2.2 신경망

신경망은 가장 널리 알려져 있는 기계학습 방법 중의 하나이다. 초기에 인간의 두뇌를 흉내 내기 위해 만들어졌지만(그림 3), 현재는 생물학적 뉴런(neuron)과 크게 관련이 있지는 않다. 그림 3에서 알 수 있듯이, 가장 간단한 형태의 신경망인 perceptron은 두 단계의 유닛으로 구성된 그래프 형태로 해석될 수 있다. 입력 유닛 p에서 출력 유닛 k 사이에는 가중치(weight)라고 불리는 w_{kp} 값이 존재한다. 기본적으로 perceptron은 $u_k = x_0 + \sum_{i=1}^{p} x_i w_{ki}$일 때, 아래와 같이 출력을 내보낸다.

$$y_k = \begin{cases} 1 & \text{if } u_k > \theta_k \\ 0 & \text{otherwise} \end{cases}$$

이런 perceptron은 비선형 문제를 풀지 못하기 때문에, 보다 일반적인 형태인 MLP(Multilayer Perceptron)가 실제 문제에서는 사용된다. MLP는 다층 유닛과 sigmoid 함수를 이용하여 비선형 문제를 해결한다. 신경망에서 학습이란 주어진 구조의 신경망에 대해, 그 신경망 내의 유닛 간 가중치를 결정하는 것이다. 신경망 학습의 기본적인 방법은 LMSE(least mean squared error)이다[7].

Perceptron은 1982년에 처음으로 아미노산의 시퀀스를 입력으로 하여 리보솜의 결합 부위(ribosome binding site)를 예측하는데 적용되었다[8]. 그 이후, Qian과 Sejnowski는 단백질의 이차 구조를 MLP를 사용해서 예측하였고[9], 다른 수많은 문제에 적용되었다[10,11].

2.3 Support Vector Machines

SVM(Support Vector Machines)은 통계적 학습 이론을 기반으로 하여 1995년 Vapnik에 의해서 개발되었고[12], 여러 가지 분류 문제(classification task)에 성공적으로 적용되었다. SVM은 기본적으로 이진 선형 분류자(binary linear classification)에서 기인한다. 그림 4는 x와 o를 나누는 선형 분류자의 한 예를 보이고 있다. x와 o를 구분할 수 있는 분류자는 그림 4의 분류자 외에도 수없이 많이 있을 수 있는데, 이런 수없이 많은 분류자 중에서 어떤 것을 선택할 것인가를 결정하는 것이 SVM의 학습이다. 여러 가지 분류자 중에서 가장 좋은 것은 분류자로부터 가장 가까운 데이터 포인트까지의 거리가 가장 먼 것이다. SVM에서는 이런 가장 가까운 데이터들을 support vector라고 부르며, 가장 먼 거리를 마진(margin)이라고 한다. 그림 4에서는 마진을 γ로 표시하고 있다.

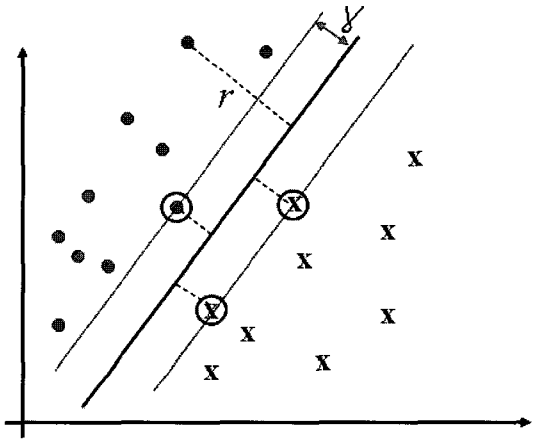


그림 4 이진 선형 분류자

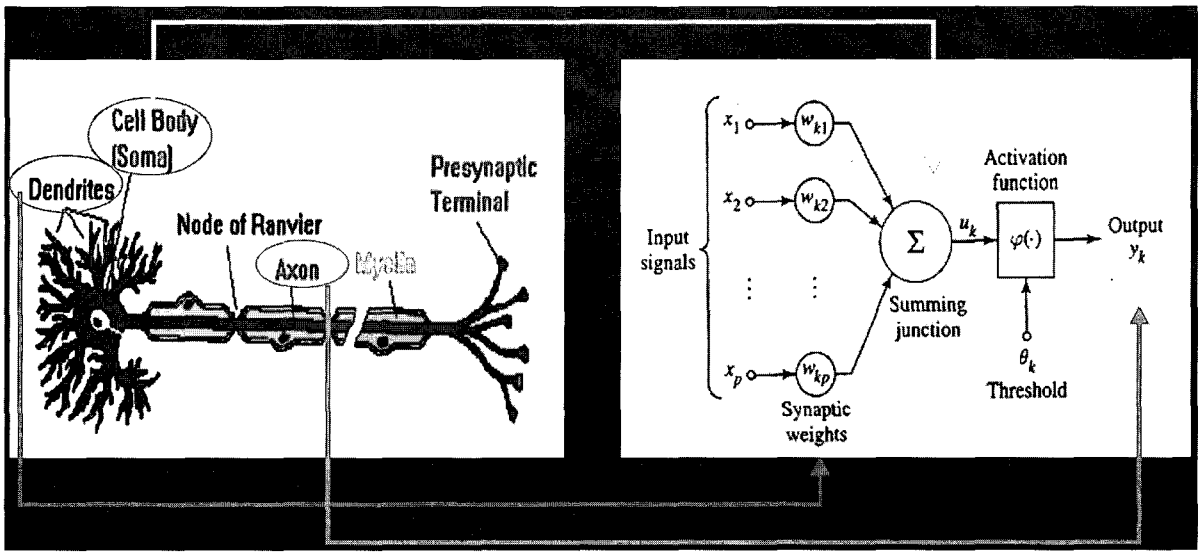


그림 3 생물학적 뉴런과 인공 뉴런과의 관계

선형 분류자는 다음 식과 같이 나타낼 수 있다.

$$f(\mathbf{x}) = \langle \mathbf{w} \cdot \mathbf{x} \rangle + b$$
$$= \sum_{i=1}^{n} w_i x_i + b$$

가장 좋은 선형 분류자를 찾을 때에는 support vector만 고려하면 되고 선형 분류자와 한 데이터 x와의 거리는 $r = \mathbf{w}^T\mathbf{X} + b/\|\mathbf{w}\|$의 관계를 가지므로, support vector x에 대해 $\mathbf{w}^T\mathbf{X} + b = 1$라고 하면 마진은 $\gamma = 1/\|\mathbf{w}\|$이다. 따라서 SVM 학습은 주어진 n 개의 데이터를 사용하여 마진을 최대로 만드는 벡터 w를 결정하는 것이다. 즉, 정형화된 SVM 학습 문제는 선형적으로 구분되는 데이터의 집합 $S = \{(\mathbf{x}_1, y_1), \cdots, (x_n, y_n)\}$에 대해, 다음의 최적화 문제를 해결하는 고차평면(hyperplane) $(\mathbf{w}, b)$를 결정하는 것이다.

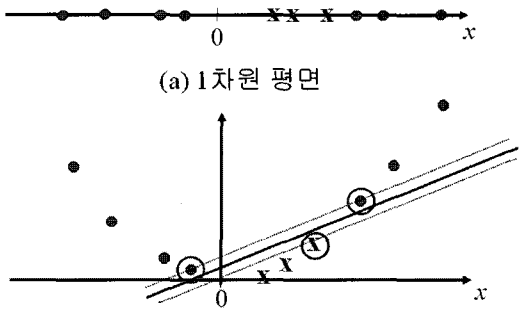
Minimize $\langle \mathbf{w} \cdot \mathbf{w} \rangle$
subject to $y_i(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) \geq 1$ for all $i = 1, \ldots, n$

이 문제는 quadratic programming으로 해결할 수 있는 최적화 문제이지만, 실제로는 Lagrangian programming으로 변환하여 문제를 푼다[13]. Lagrangian programming 문제에서는 모든 데이터 포인트 쌍 $\mathbf{x}_i$와 $\mathbf{x}_j$에 대해 $\mathbf{x}_i$와 $\mathbf{x}_j$의 내적을 계산하여야 한다.

비선형 문제를 SVM으로 풀기 위해서는 원래 $\mathbf{x}_i$가 존재하던 R^d상의 데이터를 또 다른 공간 H로 사상(mapping)하는 함수 Φ를 가정한다.

$$\Phi : R^d \rightarrow H$$

공간 H 상에서 $\mathbf{x}_i$들이 선형적으로 분리가능하다면, $\mathbf{x}_i$와 $\mathbf{x}_j$의 내적 대신에 $\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$를 구하면 된다. 또한, $\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$를 계산해 주는 커널(kernel) 함수 $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$가 있다면, Φ를 직접 사용하지 않고도 커널 함수 K만 사용하여 비선형 문제에서 SVM을 학습시킬 수 있다. 일반적으로, H의 차원은 d보다 크며, 그림 5에서 보듯이 저차원 공간에서는 비선형인 문제가 고차원 공간에서는 선형이 될 수 있다. 그림 5(a)에서는 한 점을 잡아 1차원 평면상에 존재하는 o와 x들을 두 영역으로 나눌 수 없다. 따라서 이 문제는 주어진 1차원 공간상에서는 비선형인 문제이다. 하지만 주어진 데이터들을 그림 5(b)처럼 2차원 공간으로 사상시키면 직선에 의해 두 개의 영역으로 구분될 수 있다. 즉, 2차원 공간상에서는 주어진 문제가 선형으로 구분된다.



(a) 1차원 평면

(b) 2차원 평면

그림 5 저차원 공간에서 비선형 문제가 고차원 공간에서 선형 문제가 되는 예

Brown et al.이 gene expression data 분석을 위해 SVM을 적용한 이후[14], SVM은 수많은 유전자, 세포 분류 문제, DNA의 전사 시작 위치를 찾는 문제, protein fold 인식 등에 적용되었다[15,16,17] 특히, J.-G. Joung et al.은 인유두종 바이러스의 위험군 분류에 SVM을 적용하였다[18]. 이 인유두종 바이러스의 위험군 분류 실험에서는 HPV 단백질의 시퀀스를 입력으로 삼아 위험군을 분류하였다.

2.4 자기 구성화 지도(Self-Organizing Maps)

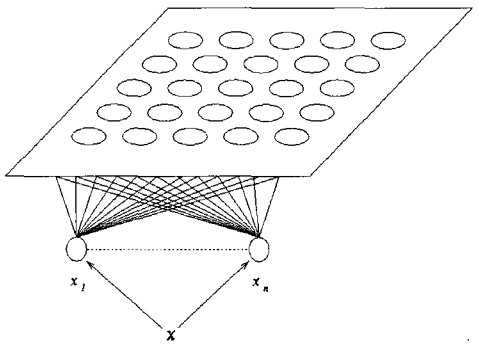


그림 6 자기 구성화 지도(SOM)의 구조

자기 구성화 지도(SOM)는 비감독 신경망의 일종이다. 즉, 입력받은 데이터들을 유사한 정도에 근거하여 군집화(clustering)시켜주는 기능을 한다. 예를 들어, 두 입력 데이터 사이의 거리를 유클리디안 거리(Euclidean distance)로 측정한다면, 이 거리가 작을수록 두 데이터는 유사하다고 판단하여 같은 군집 내에 속하도록 한다. SOM의 기본 구조는 그림 6과 같다. SOM의 구조는 용도에 따라 조금씩 달라질 수 있지만, 일반적으로는 두 개의 층(layer)으로 구성된다. 아래층은 입력층(input layer)이고, 위층은 출력층(output layer)이다. 입력층에는 입력 데이터가 들어가는데, 입력 데이터를 m차원 벡터라고 한다면 입력층에 있는 노드의 개수는 m이

된다. 즉, 입력 데이터 x는 다음과 같이 주어진다.

$$\mathbf{x} = [x_1, \ldots, x_m]^T$$

출력층의 노드는 2차원 형태로 배열되어 있으며, 출력층의 각 노드와 입력층의 모든 노드는 서로 연결되어 있다. 이 연결은 perceptron에서처럼 가중치(weight)를 가진다. 이를 테면, 출력 계층 j 노드의 가중치 벡터는 다음과 같이 표현할 수 있다.

$$\mathbf{w}_j = [w_{j1}, w_{j2}, \ldots, w_{jm}]^T$$

SOM의 학습은 데이터가 잘 군집화 되도록 가중치들을 결정하는 것으로, 순차적으로 진행된다. 우선 입력층에 들어온 데이터 x와 가장 거리가 가까운 가중치 값을 갖는 출력층 노드가 승자(winner)로 선택된다. 아래 식에서 c는 선택된 승자 출력노드의 인덱스이다.

$$c = \arg\max_j \{\| \mathbf{x} - \mathbf{w}_j \|\}$$

승자가 정해지면, 아래의 식과 같이 가중치를 변경하여 다음번에 x를 보더라도 c가 승자가 되도록 한다.

$$\mathbf{w}_j(n+1) = \mathbf{w}_j(n) + \eta(n) h_{jc}(n)(\mathbf{x} - \mathbf{w}_j(n))$$

$$\eta(n) = \eta_0 \left(\frac{\eta_T}{\eta_0} \right)^{n/T}$$

η는 학습률로서 가중치 값이 변경되는 정도를 조절하는 역할을 한다. 보통, 시간이 지남에 따라 줄어드는 함수로 표현되며, 위의 식에서 η_T는 최종 학습률이다. $h_{jc}(n)$은 입력 x에 가장 근접한 노드 c와 j번째 출력 노드 사이의 거리에 의해 값이 결정되는 이웃 함수(neighborhood function)이다. 이웃 함수로는 여러 가지가 있을 수 있지만, 일반적으로 멕시코 모자 함수(Mexican hat function)가 쓰인다. 이 과정을 모든 입력 데이터에 대해 반복하여 비슷한 입력 데이터가 비슷한 위치의 출력 노드에 위치되도록 하여 군집화를 수행한다.

Eisen et al.과 Tamayo et al.은 gene expression data를 SOM으로 표현하였으며[19,20], Zhang et al.은 SOM의 잠재 변수 모델(latent variable model) 변형인 GTM으로 gene expression data를 가시화하였다[21]. 군집화 문제에서는 군집(cluster)의 수를 결정하는 것이 중요한 문제인데, Sharan과 Shamir는 gene expression data를 군집화 할 때 군집의 개수를 자동으로 결정하는 방법을 제시하였다[22].

3. 생명 과학과 텍스트마이닝

생물학이나 의학 관련 논문은 MEDLINE과 같은 데이터베이스에 비교적 잘 정리되어 있다. 따라서 생의학 연구자들은 선행 연구나 관련 연구를 데이터베이스 검색을 통해 쉽게 찾고 자신의 연구 방향과 비교해 볼 수 있다. 하지만 데이터베이스 내의 논문 수가 점차 증가함에 따라, 사람의 힘으로 모든 관련 연구를 다 조사해 보는 것이 점점 힘들어 졌다. 또한, 논문은 특정한 분야에 국한된 것이 많으므로 전체를 조망할 수 있는 기능도 떨어진다. 따라서, 생의학 논문이나 관련 저서, 기사 등으로부터 필요한 지식을 추출하거나 전체적인 관계를 파악하는 기법의 필요성이 증대되고 있다.

텍스트마이닝(text mining)은 텍스트로 된 데이터에 데이터마이닝 기법을 적용하는 것을 말한다. 논문, 저서, 기사 등이 일반적으로 텍스트로 되어 있기 때문에, 이런 정보원으로부터 정보를 추출하거나 정보를 요약하기 위해서는 데이터마이닝 기법에 텍스트를 가공하고 처리할 수 있는 기법이 추가된 텍스트마이닝 기법이 필요하다. 즉, 텍스트마이닝에서는 데이터마이닝 기법 이 외에도 정보 검색이나 자연언어처리 기술을 추가로 요구한다. 이 절에서는 이에 따라 정보 검색 기술과 자연언어처리 기술이 텍스트로 된 정보원으로부터 필요한 지식을 추출하는데 어떻게 사용되는지 알아본다.

3.1 문서 분류 (Text Classification)

정보 검색(information retrieval)의 목적은 일반적으로 사용자의 요구에 적합한(relevant) 문서를 찾는 것이다. 월드와이드웹(WWW)에서는 야후(Yahoo)나 구글(Google)처럼 검색엔진들이 이런 역할을 대신해 주고 있고, MEDLINE과 같은 생의학 관련 데이터베이스에서도 여러 가지 검색 기법들을 제공하고 있다. 하지만, 이와 같은 정보 검색은 단순한 검색 결과만 보여줄 뿐, 필요한 정보를 추출하였다고 말하기 힘들다. 예를 들면, MEDLINE의 사용자가 HIV를 직접적으로 다루는 논문을 찾고 싶을 때, MEDLINE의 검색을 통하면 HIV를 주 내용이 아니라 부가적인 토픽으로 다루는 논문조차도 검색될 것이다. 이 경우, 사용자의 의도와 부합되는 문서를 다른 문서들로부터 분리해 주는 기능이 필요하다.

문서 분류는 주어진 문서에 대해서 그 문서가 속하는 클래스를 결정하는 것이다[23]. 기계학습 관점에서 보면 대표적인 감독 학습 문제이며 분류 문제이다. 기계학습 방법을 문서 처리에 적용하기 위해서는 문서를 기계학습에 적합하도록 표현하여야 하는데, 일반적으로 문서 공간을 벡터 공간(vector space)으로 보아 문서를 벡터로 표현한다. 이 때, 벡터의 각 요소는 어휘(vocabulary) 내의 한 단어가 된다. 각 요소가 가지는 값은 정보 검색

에서는 보통 $tf \cdot idf$로 결정된다.

Park et al.은 인유두종 바이러스의 위험군을 분류하기 위해 문서 분류 기법을 사용하였다[24]. 이들은 Los Alamos National Laboratory의 "The HPV Sequence Database"에 있는 인유두종 바이러스 설명문에 기초하여 인유두종 바이러스의 위험군을 분류하였다. 그림 7은 인유두종 바이러스인 HPV80에 대한 설명문의 예이다. 이 논문은 위양성(false positive)과 위음성(false negative)의 비용이 크게 다름을 감안하기 위하여 학습 비용을 고려하는 기계학습 방법을 제시하였다. Kowalczyk와 Raskutti는 특정한 유전자를 제거하는

<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 7 인유두종 바이러스에 대한 설명문의 예

것이 AHR(Aryl Hydrocarbon Receptor) signaling에 어떤 영향을 미치는지를 예측하기 위해서 MEDLINE에 있는 abstract들을 SVM으로 분류하였다[25].

3.2 개체명 인식(Named Entity Recognition)

개체명 인식(named entity recognition) 문제는 자연언어처리에서 오래전부터 고민해 오던 문제 중의 하나이다. 사람이 처리하는 텍스트에는 수많은 고유명사가 나타나게 되는데, 개체명 인식은 이런 고유명사가 사람 이름인지, 회사 이름인지, 지역명인지를 결정하는 문제이다. 생의학 쪽에서도 이런 문제가 나타나는데, 예를 들면 논문에서 사용된 고유명사가 유전자 이름인지, 단백질 이름인지를 결정하여야 하는 경우가 있다.

생의학 분야의 개체명 인식이 다른 분야보다 어려운 점은 (i) 용어들이 'EGFR(epidermal growth factor receptor)'처럼 축약형으로 잘 쓰이고, (ii) 용어의 일부분에 로마자, 그리스문자 등이 포함되는 경우(annexin II mRNA)가 잦고, (iii) 숫자가 용어의 일부분으로 사용되며(NIH 3T3 fibroblasts, CYP1A1 promoter), (iv) 'Jurkat T cells'처럼 eponym이 잘 쓰인다는 점이다. 따라서, 바이오메디칼 텍스트로부터 용어를 구분하기 위해서는 위와 같은 특성을 잘 반영하여 처리하여야 한다.

Lee et al.은 생의학 관련 개체명의 경계 인식과 종류 파악을 위해 두단계 SVM을 적용하였고[26], Hatzivassiloglou et al.은 생의학 논문에 나타나는 생물학 용어를 단백질, 유전자, mRNA 중의 하나로 구분하기 위하여 세 가지 기계학습 방법론(나이브베이스, 결정트리, RIPPER)을 적용하였다[27].

개체명 인식은 감독 학습 문제이므로, 이 문제를 학습하기 위해서는 텍스트에 나타난 개체명에 그 종류를 부착해(annotate) 놓은 말뭉치(corpus)가 필요하다. 그동안, 서로 다른 기관에서 많은 비용을 들여 각기 다른 말뭉치를 만들어 왔으나, 최근에 동경대에서 비교적 큰 규모의 GENIA 말뭉치(http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus)를 발표하였다[28]. 이 말뭉치는 2000개 이상의 논문 abstract에 대해 개체명 종류와 단백질 반응에 나타나는 생물학적 위치 등을 표시하였다.

3.3 부분 파싱(Shallow Parsing)

개체명 인식은 그 자체로도 의미를 가지지만, 유전자나 단백질의 관계 파악의 시작이라는 점에서 중요한 의미를 가진다. 그림 8은 생의학 논문에서 유전자와 단백질의 발현 관계를 추출하는 프로그램의 수행 예인데, 이와 같이 유전자나 단백질 간의 관계를 파악하기 위해서는 논문의 내용을 이해하여야 한다. 하지만, 현재의 자연언어처리 기술은 텍스트의 내용을 이해할 수 있을 만큼의 정확한 정보를 제공할 수준에 이르지 못하였을 뿐만 아니라, 구문 정보(syntactic information) 조차도 아주 높은 정확도로 제공하지 못한다. 따라서 순수한 자연언어처리 기법을 사용하여 유전자나 단백질의 관계를 파악하는 것은 불가능하므로, 현실성이 있는 수준의 처리만 하여 필요한 정보를 추출한다.

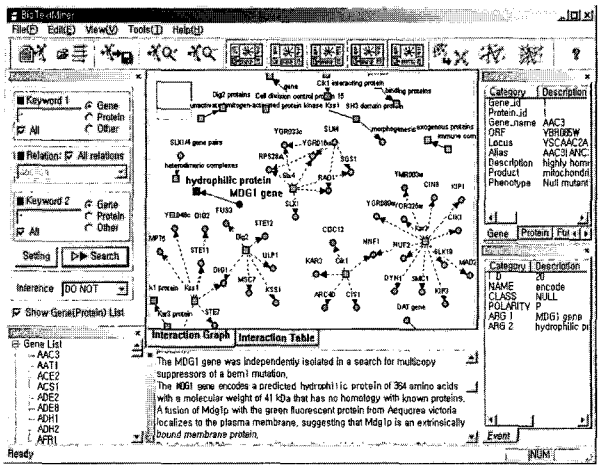


그림 8 유전자/단백질 관계 추출 프로그램의 수행 예

구문 정보를 파악하기 위해, 현재는 전체 파싱(full parsing) 대신에 부분 파싱(partial parsing) 기법을 사용한다. 부분 파싱 방법 중에서 가장 널리 알려진 것은 텍스트청킹(text chunking)이다. 텍스트청킹은 아래의 예처럼 문장을 겹침이 없는 구(phrase)로 나누는 작업을 의미한다[29].

He reckons the current deficit will narrow to only # 1.8 billon in September.
[He] [reckons] [the current deficit] [will narrow] [to] [only # 1.8 billion] [in] [September] [.]

문장을 기본이 되는 구로 나누게 되면, 간단한 패턴 인식 기법을 사용하여 유전자나 단백질의 관계 정보를 추출할 수 있다. 따라서 정확한 정보를 추출하기 위해서는 텍스트청킹의 정확도가 중요한데, Zhang et al.은 regularized Winnow 알고리듬을 이용하여 지금까지 가장 성능이 좋은 텍스트청킹 결과를 제시하였고[30], Park과 Zhang은 최대 엔트로피 모델에 부스팅 기법을 적용하여 전처리나 사전 지식이 거의 없이도 매우 높은 정확도를 보이는 텍스트청킹 모델을 제시하였다[29]. 또한, 텍스트청킹은 그림 10과 같은 생물학 관련 온톨로지를 자동으로 구축하는데도 중요한 역할을 한다.

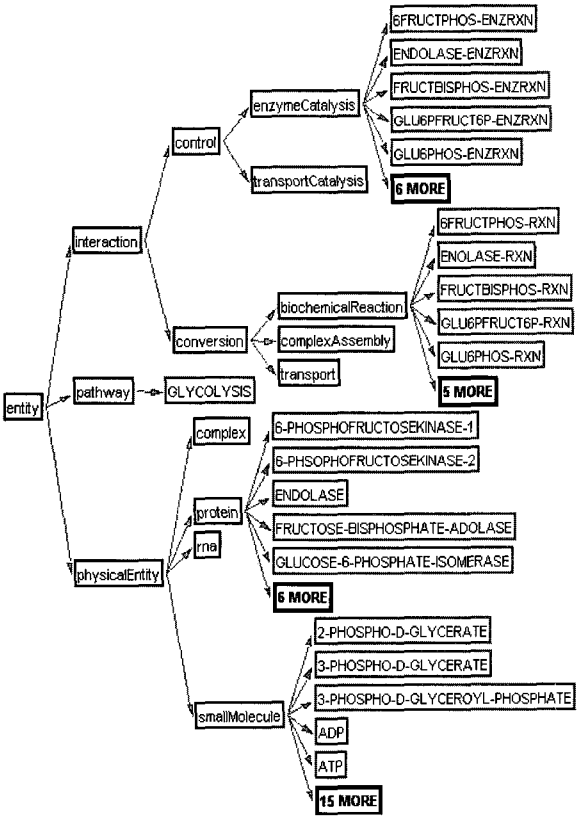


그림 9 BioPAX Pathway Ontology

4. 결 론

본 논문에서는 생명 과학의 여러 문제와 기계 학습, 그리고 텍스트마이닝과의 관계에 대해 알아보았다. 생명 과학, 특히 바이오인포매틱스와 관련된 여러 가지 문제들은 다루고 있는 문제의 복잡성과 데이터의 방대함으로 인해, 생물학적 지식이 많은 전문가에 의해서만 해결될 수 있는 단계를 넘어섰다. 이러한 특징 때문에 생명 과학의 많은 문제들을 기계학습 기법으로 해결하고자 하는 시도가 있어 왔다.

기계학습의 여러 방법들은 그 특성에 따라 잘 적용되는 분야와 잘 적용되지 않는 분야를 가지게 된다. 예를 들면, 결정트리는 이산적인(discrete) 함수를 추정할 때 잘 적용되며, 은닉 마코프 모델(hidden Markov model)은 순서가 있는 데이터를 다룰 때 적합하다. 즉, 모든 문제에 모두 잘 적용되는 만능인 기계학습 방법은 존재하지 않으므로, 방법론의 특성과 자신이 해결하고자 하는 문제를 잘 파악하여 어떤 방법이 자신의 문제에 적합한지를 결정하여야 한다. 이를 위해, 본 논문에서는 몇 가지 주요한 기계학습 방법론들을 소개하였고, 각 방법들이 어떤 문제에 적용될 수 있는지를 소개하였다.

또한, 생의학 논문에서 필요한 정보를 추출하는 바이오 텍스트마이닝 기법을 소개하였다. 논문에 나타나는 용어들을 어떻게 다루어야 하는지, 논문이나 문서의 내용은 어떻게 파악하여야 하는지, 또한 각종 유전자나 단백질의 관계는 자동으로 어떻게 추출할 수 있는지를 소개하였다. 논문처럼 텍스트로 된 데이터는 기계학습 외에도 정보검색이나 자연언어처리 기술이 추가로 필요한데, 하고자 하는 내용에 따라 그에 따른 정보검색, 자연언어처리 기술을 적합하게 선택하여 적용하여야 한다. 본 논문은 이를 위해 몇 가지 언어처리 기술들이 어떻게 적용될 수 있는지를 소개하였다.

생명 과학과 관련된 문제는 앞으로도 계속해서 기계학습의 주요 연구 분야가 될 것이다. 따라서 이 분야에서 다루는 특정 문제를 어떻게 기계학습으로 해결할 수 있는가 하는 것이 앞으로도 계속해서 연구 토픽이 될 것이다. 예를 들면, 마이크로어레이의 데이터양은 많으나 특정한 질병에 대한 마아크로어레이 데이터의 수는 매우 적은 경우가 있다. 즉, 질병의 인과 관계를 파악하기 위해서 파악하여야 하는 자질(feature) 혹은 특성(attribute)의 수는 아주 많은 데 비해 데이터의 수는 부족한 경우가 있다. 이런 경우에는 현재의 기계학습 기법이 잘 적용되지 않는다. 따라서 이 문제를 포함하여 현재 기계학습이 잘 적용되지 않는 문제들을 어떻게 해결해 나갈 것인가 하는 것이 앞으로의 과제이다.

참고문헌

[1] R. Quinlan, "Induction of Decision Trees," *Machine Learning*, Vol. 1, No. 1, pp. 81-106, 1986.

[2] R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, 1993.

[3] S.-B. Park, S.-H. Hwang, and B.-T. Zhang, "Classification of Human Papillomavirus (HPV) Risk Type via Text Mining," *Genomics & Informatics*, Vol. 1, NO. 2, pp. 80-86, 2003.

[4] N. Beerenwinkel, B. Schmidt, H. Walter, R. Kaiser, T. Lengauer, D. Hoffmann, K. Korn, and J. Selbig, "Diversity and Complexity of HIV-1 Drug Resistance: A Bioinformatics Approach to Predicting Phenotype from Genotype," *PNAS*, Vol. 99, No. 12, pp. 8271-8276, 2002.

[5] M. Dettling and P. Buehlmann, "Boosting for Tumor Classification with Gene Expression Data," Bioinformatics, Vol. 19, No. 9, pp. 1061-1069, 2003.

[6] S. Lin, S. Patel, A. Duncan, and L. Goodwin, *"Using Decision Trees and Support Vector Machines to Classify Genes by Names," In Proceedings of the European Workshop on Data Mining and Text Mining for Bioinformatics*, pp. 35-41, 2003.

[7] T. Mitchell, *Machine Learning*, McGraw-Hill, 1997.

[8] G. Stormo, T. Schneider, L. Gold, and A. Ehrenfeucht, "Use of Perceptron Algorithm to Distinguish Translational Initiation Sites in *e.coli*," *Nucl. Acids Res.*, Vol. 10, pp. 2997-3011, 1982.

[9] N. Qian and T. Sejnowski, "Predicting the Secondary Structure of Globular Proteins Using Neural Network Models," *J. Mol. Biol.*, Vol 202, pp. 865-884, 1988.

[10] H. Nielsen, J. Engelbrecht, S. Brunak, and G. Heijne, "Identification of Prokaryotic and Eukaryotic Signal Peptides and Prediction of Their Cleavage Site," *Prot. Eng.*, Vol. 10, pp. 1-6, 1997.

[11] B.-H. Kim, S.-B. Park, and B.-T. Zhang, "PromSearch: A Hybrid Approach to Human Core-Promoter Prediction," *Lecture Notes in Computer Science*, Vol. 3177, pp. 125-131, 2004.

[12] V. Vapnik, *Statistical Learning Theory*, Springer, 1998.

[13] C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, Vol. 2, No. 2, pp. 121-167, 1998.

[14] M. Brown, W. Grundy, D. Lin, N. Cristianini, C. Sugnet, T. Furey, M. Ares, and D. Haussler, "Knowledge-based Analysis of Microarray Gene Expression Data by Using Support Vector Machines," *PNAS*, Vol. 97, pp. 262-267, 2000.

[15] I. Guyon, J Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification Using Support Vector Machines," *Machine Learning*, Vol. 46, No. 1, pp. 389-422, 2002.

[16] A. Zien, G. Raetsch, S. Mika, B. Scholkopf, T. Lengauer, and K. Muller, "Engineering Support Vector Machine Kernels that Recognize Translation Initiation Sites," *Bioinformatics*, Vol. 16, No. 9, pp. 799-807, 2000.

[17] C. Ding and I. Dubchak, "Multi-class Protein Fold Recognition Using Support Vector Machines and Neural Networks," *Bioinformatics*, Vol. 17, pp. 349-358, 2001.

[18] J.-G. Joung, S.-J. O, and B.-T. Zhang, "Prediction of the Risk Types of Human Papillomaviruses by Support Vector Machines," *Lecture Notes in Artificial Intelligence*, Vol. 3157, pp. 723-731, 2004.

[19] M. Eisen, P. Spellman, P. Brown, D. Botstein, "Cluster Analysis and Display of Genome-wide Expression Patterns," *PNAS*, Vol. 95, pp 14863-14868, 1998.

[20] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu,

S. Kitareewan, E. Dmitrovsky, E. Lander, and T. Golub, "Interpreting Patterns of Gene Expression with Self-Organizing Maps: Methods and Application to Hematopoietic Differentiation," *PNAS*, Vol. 96, pp. 2907-2912, 1999.

[21] B.-T. Zhang, J. Yang, and S.-W. Chi, "Self-Organizing Latent Lattice Models for Temporal Gene Expression Profiling," *Machine Learning*, Vol. 52, No. 1/2, pp. 67-89, 2003.

[22] R. Sharan and R. Shamir, "CLICK: A Clustering Algorithm with Applications to Gene Expression Analysis," In *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology*, pp. 307-316, 2000.

[23] 박성배, 자연언어 학습을 위한 최대 엔트로피 부스팅 모델, 서울대학교 전기컴퓨터공학부 박사학위 논문, 2002.

[24] S.-B. Park, S.-H. Hwang, and B.-T. Zhang, "Mining the Risk Types of Human Papillomavirus (HPV) by AdaCost," *Lecture Notes in Computer Science*, Vol. 2690, pp. 403-412, 2003.

[25] A. Kowalczyk and B. Raskutti, *Single Class SVM for Yeast Gene Regulation Prediction*, KDD Cup 2002 Task 2 Winner.

[26] K.-J. Lee, Y.-S. Hwang, S.-H. Kim, and H.-C. Rim, "Biomedical Named Entity Recognition Using Two-phase Model Based on SVMs," *Journal of Biomedical Informatics*, VOl. 37, No. 6, pp. 436-447, 2004.

[27] V. Hatzivassiloglou, P. Duboue, and A. Rzhetsky, "Disambiguating Proteins, Genes, and RNA in Text: A Machine Learning Approach," *Bioinformatics*, Vol. 17, Suppl. 1, pp. 97-106, 2001.

[28] J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii, "GENIA corpus: A Semantically Annotated Corpus for Bio-Textmining," *Bioinformatics*, Vol. 19, Suppl. 1, pp. 180-182, 2003.

[29] S.-B. Park and B.-T. Zhang, "A Boosted Maximum Entropy Model for Learning Text Chunking," In *Proceedings of the 19th International Conference on Machine Learning*, pp. 482-489, 2002.

[30] T. Zhang, F. Damerau, and D. Johnson, "Text Chunking Based on a Generalization of Winnow," *Journal of Machine Learning Research*, Vol. 2, pp. 615-637, 2002.

박 성 배

1990~1994 한국과학기술원 전산학과(학사)
1994~1996 서울대학교 컴퓨터공학과(석사)
1996~2002 서울대학교 전기컴퓨터공학부(박사)
2004~현재 경북대학교 컴퓨터공학과 전임강사
관심분야 : 기계학습, 자연언어처리, 정보검색, 바이오인포매틱스
E-mail : seongbae@knu.ac.kr