

단순한 커널 갱신을 통한 분류기의 설계

노영균⁰¹ 김청택² 장병탁³
서울대학교 인지과학 협동과정¹, 서울대학교 심리학과², 서울대학교 컴퓨터공학부³
yknoh@bi.snu.ac.kr, ctkim@snu.ac.kr, btzhang@bi.snu.ac.kr

Modeling of Classifiers by Simple Kernel Update

Yung-Kyun Noh⁰¹ Cheong-Tag Kim² Byoung-Tak Zhang³
Interdisciplinary Program in Cognitive Science¹, Department of Psychology²,
School of Computer Science and Engineering³,
Seoul National University, Seoul 151-742, Korea

요 약

커널(Kernel)을 이용한 분류 방법은 넓은 마진(large margin) 분류기로서 SVM(Support Vector Machine)을 주로 사용하게 된다. 하지만, 이 방법은 라그랑제 파라미터(Lagrange Parameter)의 최적화 과정을 포함함으로써 학습 과정을 쉽지 않게 만든다. 이 최적화 과정은 특히 DNA computing과 같은 단순한 과정의 설계를 통해 결과를 얻어야 하는 새로운 계산 모델에 커널을 적용하고자 했을 경우 큰 장벽이 된다. 본 논문에서는 넓은 마진을 목표로 하는 최적화 과정이 아닌 다른 라벨(label)의 데이터간의 경계 파악을 위한 간단한 커널 갱신 방법의 도입을 통해 분류기를 설계한다. 이 방법을 가우시안 커널에 적용시켜 본 결과, 반복을 통해 데이터의 구조를 찾아갈 수 있는 특성을 보여주며, 결국 넓은 마진의 최적화된 파라미터를 찾게 됨을 보여준다. 본 논문에서는 이 최적화 방법을 DNA 분자를 이용한 커널 생성 모델인 DNA 커널에 적용시켰을 때 잘 알려진 AML/ALL 데이터를 잘 분류해 봄을 보여준다.

1. 서 론

최근의 분자 컴퓨팅 모델링의 경우 DNA 분자간 반응의 통계적 특성을 이용하여 얻은 double strand 분자들의 통계적 분포로부터 정보를 얻어내는 방법을 시도하고 있다.[1,2] 이 방법의 장점은 순차적 계산 과정이 들어가야 하는 병목 과정을 제거함으로써 수많은 분자간의 평행 반응 시간이 분자의 양이나 데이터의 양에 구애받지 않고 이루어지게끔 하였다는데 있다. 따라서 이 방법은 분자 컴퓨팅의 원래 목적과 장점을 잘 살려 이용하는 계산 방법이라고 할 수 있다. 하지만 이러한 방법은 분자들의 평행 작용의 장점을 그대로 이용하기 위해 분자의 단순한 작용만으로 분포를 얻을 수 있도록 단순한 설계를 해야 한다는 어려움을 안고 있다. 커널을 DNA 분자들을 이용하여 구현하는 최근의 연구[2]는 DNA의 간단한 작용만으로 양한정 조건을 만족하며 데이터간의 유사도를 측정하는 커널(DNA 커널)을 구현한다는 면에서 DNA 분자를 이용한 분류기를 만들 수 있는 가능성을 보여준다. 하지만 실제로 분자 컴퓨팅의 장점을 만족시키는 분류기를 구현하기 위해서는 커널을 만든 후 최종 분류된 결과까지도 DNA들의 작용으로 얻을 수 있어야 바람직하다. 하지만, 커널(Kernel)을 이용한 대표적 분류기인 SVM(Support Vector Machine)의 경우 최대 마진을 만족시키는 파라미터들을 얻기 위해 분자 컴퓨팅으로 구현하기에 불가능한 복잡하면서 정교한 최적화 과정을 거치게 된다.

한편, SVM의 최대 마진 최적화 과정을 실리콘 컴퓨터에서 구현하는 것도 데이터가 많아졌을 경우 쉽지 않다. 최대 마진 최적화 과정의 경우 QP(Quadratic Programming)방법을 통해 데이터의 수와 같은 양만큼의 Lagrange multiplier들의 값을 구해 최적화된 특성 공간(feature space)에서의 초평면(hyperplane)을 찾아 주게 되는데, 이 최적화 과정이 많은 데이터에 대해서 매우 시간을 많이 들이게 하는 작업이다.

본 연구에서는 구현이 간단한 커널 갱신 과정을 통해 만들어진 데이터 개수의 파라미터가 SVM의 Lagrange multiplier와 같은 용도로 쓰였을 때, 분류기로서 역할을 할 수 있음을 보여주고, 이것이 DNA 커널에 응용되었을 때 DNA만을 사용하여 구현이 가능한 단순한 과정을 통해 결과를 낼 수 있는 분류기를 만들 수 있음을 보여준다.

2. 커널의 갱신을 통한 파라미터 획득

학습을 위해 주어진 n 개의 데이터 \mathcal{D} 는 다음의 집합으로 표시되고,

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$$

n 개의 원소(element)는 각각 d 차원의 벡터 $\mathbf{x}_i \in \mathbb{R}^d$ 와 라벨값 $y_i \in \{1, -1\}$ 으로 구성된다. 아래는 이렇게 주어진 n 개의 데이터로부터 새로 주어지는

데이터의 라벨 값을 예측하는 학습 모델에 대한 기술이다.

각 데이터마다 주어진 파라미터값 p_i 에 대해서, p_i 의 n 번째 갱신값 p_i^n 이 다음과 같이 주어지고,

$$p_i^n = p_i^{n-1} \sum_{y_k \neq y_i} K_{ik}^n / \sum_m K_{im}^n \quad (\text{단, } p_i^0 = 1)$$

여기서, K_{ij}^n 이 커널값 $K(\mathbf{x}_i, \mathbf{x}_j) \equiv K_{ij}$ 의 n 번째 갱신값으로서, n 이 2 이상일 때, K_{ij}^n 는 라벨이 다른 데이터끼리의 정보만을 얻기 위해 다음과 같이 주어진다고 했을 때, p_i^n 와 K_{ij}^n 를 교대로 갱신해 가는 것이 가능하게 된다.

$$K_{ij}^n = P_i^{n-1} P_j^{n-1} K_{ij}$$

$$= P_i^{n-2} P_j^{n-2} K_{ij} \left(\sum_{y_k \neq y_i} \sum_{y_l \neq y_j} K_{ik}^{n-1} K_{jl}^{n-1} \right) / \sum_{k,l} K_{ik}^{n-1} K_{jl}^{n-1}$$

(단, $K_{ij}^1 = K(\mathbf{x}_i, \mathbf{x}_j)$)

다른 데이터끼리의 정보만을 얻는 이유는 다른 라벨의 데이터끼리의 경계 정보를 반영하기 위한 어림법(heuristic)이다.

테스트 데이터에 대한 라벨값은 다음과 같이 계산된다.

$$y_{new}^k = \text{sign} \left(\sum_i p_i^k K(\mathbf{x}_{new}, \mathbf{x}_i) y_i \right)$$

위의 간단한 방식을 가우시안(Gaussian) 커널과 함께 사용했을 때 2차원상의 임의의 12개의 학습 데이터에 대해 입력 공간을 어떻게 분류하는가를 살펴보면 다음과 같다.

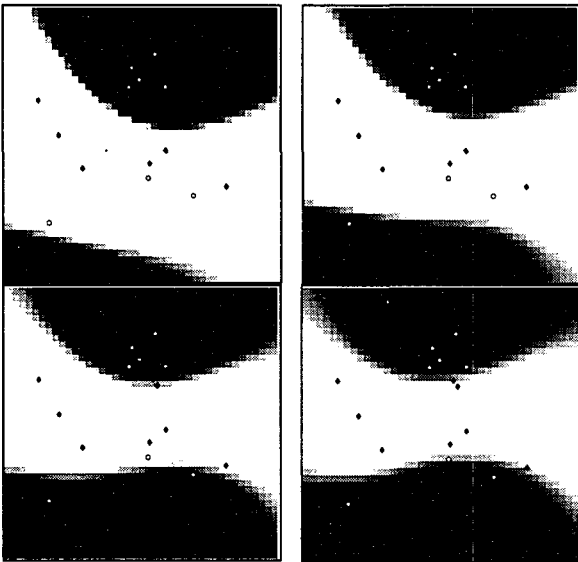


그림 1. 한 라벨당 8개씩의 데이터에 대한 반복 횟수에 따른 공간 분류. 왼쪽 위 0회, 오른쪽 위 1회, 왼쪽 아래 4회, 오른쪽 아래 8회. 점점 데이터의 구조를 잡아갈 수 있다.

이 때, 갱신 횟수 k 에 따른 파라미터 벡터 $\mathbf{p}^k = (p_1^k, p_2^k, \dots, p_n^k)$ 와 SVM을 통해 얻은 최적화된 라그랑제 파라미터 벡터와의 상관도는 다음과 같다.

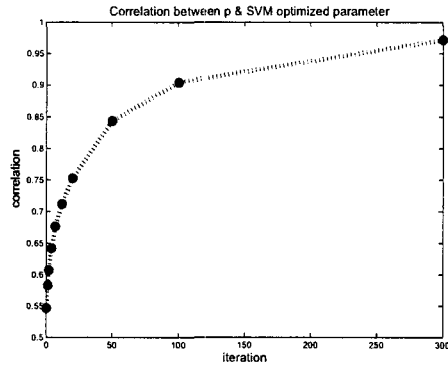


그림 2. 커널 갱신에 따라 얻어진 파라미터의 갱신 횟수에 따른 SVM 최적화된 파라미터와의 상관도.

3. DNA 커널에의 적용

위의 모델을 DNA 커널에 적용시키는 것은 매우 간단하다. DNA 커널을 생성하는 과정은 기존 연구에 자세히 기술되어 있다.[2] 기존 연구대로 적당한 온도 조절 하에서 hybridization을 통해 커널을 생성 후 커널을 갱신하는 과정은 라벨이 다른 dsDNA(double strand DNA)만을 골라내 다시 de-nature시켜 새로운 초기 population을 만들고, 이를 가지고 똑같은 조건으로 다시 hybridization시켜 갱신된 커널을 만드는 것이다. 이 때, p_i 값은 새로 만들어진 population의 각 데이터에 해당하는 ssDNA(single strand DNA)의 양에 해당되는 값이다.

즉, 커널을 갱신하여 결과값을 얻는 과정을 정리하면 다음과 같다.

1. Hybridization을 통해 DNA 커널을 생성한다.
2. 라벨값이 다른 것들끼리 붙은 dsDNA만을 추출한다. ($y_i \neq y_j$ 면 $K_{ij} = 0$)
3. De-nature시켜 다시 ssDNA를 만든다. 이 때, 각각의 ssDNA양은 p_i 값이 된다.
4. 2.~3.과정을 k 번 반복한다.
5. k 번 반복된 후 만들어진 ssDNA 풀(pool)에 판별하고자 하는 ssDNA를 넣고, 이것이 어느 라벨을 가진 ssDNA와 더 많이 붙었는지를 판단하여 라벨을 판단한다.

이 방법을 당뇨병 환자의 gene expression 데이터인 AML/ALL 데이터 가운데 mutual information이 큰 순서로 attribute 10개를 사용하여 데이터를 정리한 후 중복된 데이터를 없애 총 75개의 데이터를 만든 후 적용시켜 보았다. 중복된 데이터를 없앤 75개의 데이터 중 학습 데이터 20개로 p값을 얻어 테스트 데이터 55개에 대해 테스트해 얻은 분류 성공률은 다음과 같다.

DNA 커널(온도 스케줄 80℃→40℃)	
커널 갱신(10회)	.9273
SVM 최적화	.8909
가우시안 커널	
SVM 최적화	.8545

5. 결론 및 토의

본 연구는 간단한 커널의 갱신 과정만을 가지고 복잡한 SVM의 최적화 과정을 대신할 수 있는 가능성을 제시한다. DNA 분자를 사용해서 풀 수 있는 기계학습 문제를 고안하기 위해선 DNA분자들의 단순한 작용(process)만을 사용하여 학습 과정을 기술할 수 있어야 한다. 이 방법은 DNA분자들의 불고, 떨어지고 올라가는 세 가지 과정만을 이용한 커널 최적화 과정의 기술이다. 이 방법을 통해 DNA 커널을 통한 데이터의 분류에 관한 전체 과정의 설계가 완성된다.

감사의 글

본 연구는 과학기술부 국가지정연구실사업, 산업자원부 차세대 연구개발사업, 그리고 산업자원부 뇌신경정보학 사업(김청택)에 의하여 지원 되었음.

참고문헌

- [1] Zhang, B., & Jang, H., DNA, Molecular learning of wDNF formulae. (2005) *Preliminary Proceedings of the Eleventh International Meeting on DNA Computing(DNA11)*.
- [2] 노영균, 김성규, 김청택, 장병탁, (2005) DNA 커널을 이용한 MicroRNA 유전자 예측, *정보과학회 추계학술대회 2005*.
- [2] 김준식, 김종찬, 노영균, 이동운, 장병탁. (2005) DNA 컴퓨팅연산 과정의 통계 물리적 예측, *한국정보과학회 춘계학술대회 2005*.
- [3] 노영균, 김청택, 장병탁. (2005) 개념 학습을 위한 DNA컴퓨팅 기반 커널의 설계, *한국인지과학회 춘계학술대회, 2005*.
- [4] Cheok, M. et al., (2003) Treatment-specific changes in gene expression discriminate *in vivo* drug response in human leukemia cells, *Nature Genetics 35*, 2003.
- [5] Schoelkopf, B., & Smola, A. (2001) *Learning with Kernels*, MIT Press.
- [6] John S. & Nello C. (2004) *Kernel Methods for Pattern Analysis*, Cambridge University Press.
- [7] Schoelkopf, B. et al. (2002) A Kernel Approach for Learning From Almost Orthogonal Patterns, *Proceedings of the 13th European Conference on Machine Learning*.