개념학습을 위한 DNA 컴퓨팅 기반 커널의 설계

서울대학교 인지과학 협동과정 [1], 서울대학교 심리학과[2], 서울대학교 컴퓨터공학부[3]
노영균[1] (yknoh@bi.snu.ac.kr), 김청택[1,2](ctkim@snu.ac.kr), 장병탁[1,3](btzhang@bi.snu.ac.kr)

Design of Kernels Based on DNA Computing for Concept Learning

Seoul National University, Interdisciplinary Program in Cognitive Science[1], Department of Psychology [2], School of Computer Science and Engineering[3]
Noh Yung-Kyun[1] (yknoh@bi.snu.ac.kr), Kim Cheong-Tag[1,2] (ctkim@snu.ac.kr), Zhang Byoung-Tak[1,3] (btzhang@bi.snu.ac.kr)

요 약

기계학습에서 커널을 이용한 방법은 그 응용범위가 기계학습의 전반에 걸쳐 다양하게 이용되고 있으며, 그 성능 또한 기존의 방법들을 앞지르고 있다. 이는 기존의 비선형적 접근을 커널을 이용한 고차원 공간에서의 선형적 접근법으로 바꿈으로써 가능하게 되는 것이다. 다양한 분야에 적용되는 많은 커널들이 존재하며 각 커널들은 특별한 분야에 적용되기 쉽도록 다른 형태를 띠고 있기도 하지만, 커널로서 작용하기 위해 양한정 조건(positive definiteness)을 만족해야 한다. 본 연구에서는 DNA 문제에 직접 적용시킬 수 있는 방법으로서의 새로운 커널을 제시한다. 또한 매트로폴리스(Metropolis) 알고리즘을 이용하여 DNA의 hybridization과정을 모사함으로써 새로운 종류의 커널이 양한정(positive definite) 조건을 만족시킬 수 있는 방법을 제시한다. 새로 만들어진 커널이 행렬값을 형성해 나가는 과정을 살펴보면 인간이 예(instance)로부터 개념을 형성해 나가는 과정과 흡사한 양상을 보이는 것을 알 수 있다. 개념을 나타내는 좋은 예로서의 표본(prototype)으로부터 개념이 형성되어 가는 과정은 표본(prototype)이 아닌 예로부터 개념이 형성되는 과정과 다른 양상을 띠는 것과 같은 모양을 보인다.

1. 서 론

분자컴퓨팅 방법은 수많은 분자들의 병렬처리 과정을 이용해 기존의 실리콘 기반 컴퓨터에서 계산능력의 한계로 인해 풀지 못하거나 오랜 시간이 걸리던 문제를 해결할 수 있는 새로운 대안으로 제시되었다. 분자컴퓨팅의 방법인 DNA 컴퓨팅은 1994년 Adleman[10]이 Hamiltonian path 문제를 풀어냄으로써 DNA를 이용한 새로운 개념의 컴퓨터의 개발의 가능성을 제시한 후, 여러 가지 컴퓨터에서 행하는 계산의 유비를 취해 DNA로 컴퓨터를 만들려는 시도로서 현재도 계속되고 있다. 하지만 이런 순차적 처리에 기반한 컴퓨터와 같은 역할을 하도록 만들려는 시도는 구조적으로 많은 수의 분자를 이용한 병렬적 처리에 강한 DNA의 특성을 제대로 이용하지 못하게 되는 약점이 있다. 이 논문에서는 순차적 처리가 아닌 패턴을 인식하는 방법으로서 DNA 컴퓨팅의 장점을 살려 기계학습에 적용시켜 본다. 특별히 DNA를 이용하여 커널과 같은 작용을 구현해 판별문제에 사용할 수 있는 방법을 제시한다. 이렇게 구현한 커널의 작동방식과 생성모형은 원형(prototype), 혹은 비원형의 예들로부터 개념이 학습되는 과정을 모사하는 것같이 보인다. 따라서 본 연구에서는 분류문제에 사용할 수 있는 DNA를 통해 구현된 커널을 가지고 예시를 통한 학습 과정에서 제시된 기존의 연구 모델들에 어떤 유비를 취할 수 있으며, 적용될 수 있는지 살펴보는 것을 목표로 한다.

앞으로 진행될 내용은 다음과 같다. 2장에서는 DNA를 통해 커널을 만들어 분류기로 만들 수 있는 방법을 소개하고, 3장에서는 이를 위해 만족시켜야 하는 양한정 조건을 가질 수 있는 방법을 제시한다. 4장에서는 예시를 통한 개념의 학습 과정이 DNA를 이용한 커널의 생성 과정과 어떤 유비점을 찾을 수 있는지를 논하고, 5장에서 결론을 맺는다.

2. DNA를 이용한 커널

커널을 이용한 방법은 고차원 공간에서의 내적

(inner product) 계산을 가능하게 해 주는 방법으로 제시가 되지만, 데이터의 개수와 같은 차원의 제곱개의 요소(element)를 가짐으로써 커널 행렬을 구성하는데 역시나 많은 계산을 필요로 한다. 하지만, 다음에 제시하는 DNA를 이용한 커널의 구현은 병렬적으로 일어나는 DNA의 hybridization을 이용하여 커널 행렬을 요소별로 계산하지 않고 한꺼번에 구성해 내게 된다. 두 개의 ssDNA (single strand DNA) 분자가 hybridization에 의해 dsDNA (double strand DNA)를 만드는 과정이 DNA가 가지는 염기서열에 의해 결정되는 결합 에너지에만 관련이 있다고 가정했을 때, 만들어지는 dsDNA의 양은 ssDNA의 초기량과 결합 에너지를 통해 계산될 수 있는 볼츠만 분포에 의해 결정된다고 할 수 있다. i번째 DNA와 j번째 DNA의 결합확률을 다음과 같이 표현할 수 있다.

$$P(i,j) = P(i)P(j)\frac{1}{1+\exp(-\Delta_{ij}/kT)} \qquad (1)$$

$P(i)$: i번째 ssDNA가 선택될 확률

Δ_{ij} : i번째 DNA와 j번째 DNA의 Gibbs 자유에너지.

k : 볼츠만 상수

T: 절대온도

이 분포를 만들어가는 과정을 매트로폴리스 알고리즘을 통해 구현할 수 있다. 임의로 선택된 분자들이 ssDNA면, 다음 과정을 통해 hybridization을 시도하고, dsDNA라면 다음 과정을 통해 dehybridization을 시도한다. [6]

ssDNA일 때,

$$\begin{cases} \exp(-\Delta_{ij}) \geq 0 : hybridization \\ \exp(-\Delta_{ij}) < 0 : \exp(-\Delta_{ij})\text{의 확률로} \\ \qquad\qquad hybridization \end{cases}$$

dsDNA일 때,

$$\begin{cases} \exp(\Delta_{ij}) \geq 0 : dehybridization \\ \exp(\Delta_{ij}) < 0 : \exp(\Delta_{ij})\text{의 확률로} \\ \qquad\qquad dehybridization \end{cases}$$

Gibbs 자유 에너지를 구할 때 사용된 식과 상수는 다음과 같다.

Δ_{ij} : (DNA 가닥당 결합에너지) - (절대온도) X(DNA 가닥당 결합 엔트로피)

DNA 가닥당 결합 엔트로피	(결합 베이스쌍 개수)X0.023 (kcal/mol)
아보가드로수	6.0221367×10^{23} (1/mol)
볼츠만상수(k)	1.380657×10^{-23} (J/K)
1(cal)	4.186 (J)
절대온도(K)	273.16+섭씨온도 (°C)
1mol의 베이스쌍(base pair)당 결합에너지	A-T/T-A: -7.2 (kcal/mol) G-C/C-G: -9.0(kcal/mol) 나머지: -5.3(kcal/mol)

위의 알고리즘을 가지고 커널의 구성은 다음과 같이 만들 수 있다. ssDNA 하나를 데이터 하나로 코딩했을 때, 각각의 동일한 수의 ssDNA에 대해서 동일한 수의 상보적(complementary)인 ssDNA를 만들어, i번째 데이터에 해당하는 ssDNA와 j번째 데이터에 해당하는 상보적 ssDNA와의 결합량을 세어(Counting) 커널 행렬 K의 (i,j)번째 구성요소(element)로 만들게 된다.

$$K_{ij} = count(dsDNA(i, complementary(j))) \qquad (2)$$

이렇게 만들어진 커널 행렬을 가지고 커널 회귀(Regression)나 커널PCA (Principal Component Analysis)등 여러가지 기계학습 기법을 사용할 수 있지만, 개념학습에 적용하는 예로서 판별(discrimination)할 수 있는 방법인 SVM(Support Vector Machine)과 비슷하게 작동하는 방법을 소개하면 다음과 같다.

SVM의 듀얼 폼(dual form)의 데이터 개수의 Lagrange multiplier α_i 들이 만들어진 커널 행렬에 대해 다음 조건을 만족하도록 정할 수 있다.

$$\vec{\alpha} = \arg\max_{\vec{\alpha}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K_{ij} \qquad (3)$$

단, $\alpha_i \geq 0, i = 1,2,...,n$, $\sum_{i=1}^{n} \alpha_i y_i = 0$

$y_i \in \{+1, -1\}$: 각 데이터 i의 분류값

각각의 i번째 ssDNA들과 이것과 상보적인 ssDNA들의 개수를 α_i 에 비례하도록 초기분포를 두면, 새로운 데이터 x 를 표현하는 가닥에 대해 각 분류마다 다음 식을 만족하는 dsDNA를 얻을 수 있다.

$$count(dsDNA(x, others \in y)) = \sum_{i \text{ where } y_i \in y} \alpha_i K_i(x) \quad (4)$$

단, $K_i(x)$ 는 새로운 데이터 x 가 이전 분포에서 i번째 데이터와 붙는 양.

식 (4)의 값이 가장 큰 y값이 바로 분류값이 된다. 이는 2개의 범주{+,-}를 가진 분류에서 다음의 식을 푸는 것에 해당된다.

$$f(x) = sign(\sum_{i=1}^{n} \alpha_i y_i K_i(x)) \quad (5)$$

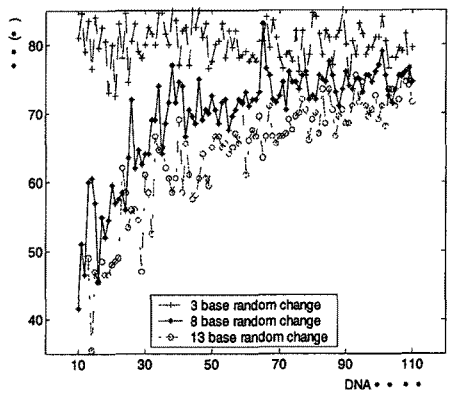
단, $f(x) \in \{+,-\}$

이를 실제 DNA에 적용시킬 때는 다음과 같은 과정이 적용되게 된다.

1. 각각의 데이터에 해당되는 DNA를 구성한다. 이 때, 유사한 DNA일수록 염기배열이 비슷하게 해서 상보적 가닥과의 결합 에너지가 크게 한다.
2. 만들어진 DNA 모음을 시험관에서 섞고, 온도를 조절해서 DNA 가닥들이 서로 붙게 한다.
3. i번째 데이터에 해당되는 DNA 가닥과 j번째 데이터에 해당되는 DNA 가닥이 붙은 dsDNA의 양에 비례해 커널 행렬의 성분 K_{ij}를 구성한다.

3. 양한정 조건

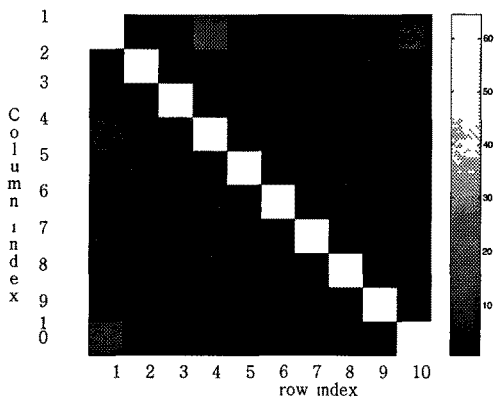
위에서 제시한 커널이 분류기에 사용되는 커널로서 역할을 하기 위해서는 양한정 조건을 만족시켜야 한다.[8] 매트로폴리스 알고리즘을 DNA가닥 10개를 가지고 온도 T를 변화시키며 실험해 보면, 일정 온도 이상에서만 양한정 조건을 만족시키게 되는데, DNA 가닥의의 길이에 따라 양한정 조건을 만족시키기 시작하는 온도는 (그림 1)과 같다. DNA 가닥의 생성은 전체 길이가 l 로 임의로 생성된 가닥에서 다시 임의로 3개, 8개, 13개의 베이스를 변화시킨 n-1개의 데이터를 가지고 n개의 데이터를 만들어 하나의 실험 군을 생성했는데, 이러한 작업을 매 번 실험마다 실시했다.



(그림 1) DNA 가닥의 길이에 따른 양한정 조건을 만족하기 시작하는 온도

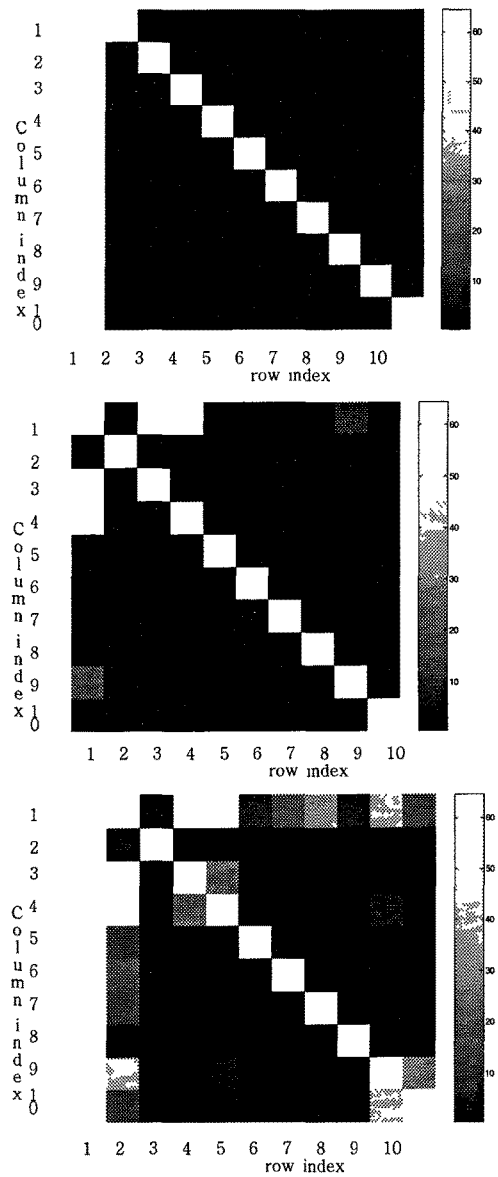
DNA 가닥의 길이에 따른 온도 변화가 일정치 않은 데서 양한정 조건을 만들기 시작하는 온도가 염기의 순서열에 많은 영향을 받음을 알 수 있지만, 염기의 순서열에 변이를 준 개수에 따라서 양한정 조건을 만족시키는 구간에 확연한 차이를 보여주는 것을 알 수 있다.

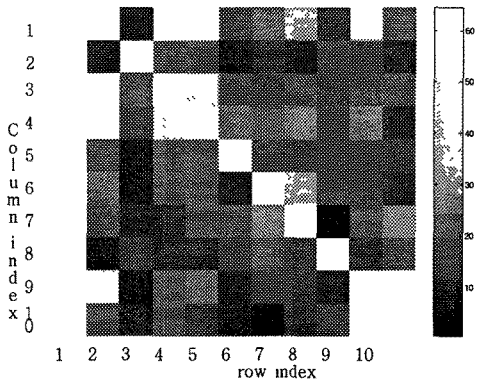
양한정 조건을 만족시키는 구간에서의 행렬의 요소를 살펴보면, (그림 2)와 같이 대각요소의 우세(dominance)가 양한정 조건을 만족시키는 데 큰 역할을 하는 것을 알 수 있다. 대각요소를 제외하고는 매우 희박한(sparse) 형태를 보여주는데, 이는 커널 방법에서 높은 차원의 공간을 이용해서 데이터를 서로 수직(orthogonal)하게 만들어 주는 것에 해당되는 것으로 일반화(generalization) 성능에 안 좋은 영향을 미치게 된다. 이는 양한정 조건을 만족시키기 위해 hybridization 온도를 높여 주었을 때 오는 불가피한 결과다.



(그림 2) 일정한 온도에서 양한정 조건을 만족시키도록 형성된 커널(65.5℃, 50mer, 8개 베이스 변화). 커널 행렬의 각 요소 값을 색깔로 표시하였다.

양한정 조건을 만족하도록 대각 요소가 만들어진 행렬에 온도를 천천히 낮춰 커널 행렬의 대각원소가 아닌 원소들이 생기게끔 하면, 대각원소는 비슷하게 유지되면서 희박성이 없어지는 특성이 나타나게 된다. 이렇게 만들어진 행렬을 데이터를 분류하는 양한정성을 만족하는 커널 행렬로서 사용 가능하게 된다. 이 커널 행렬을 가지고 데이터의 라벨에 따른 Lagrange multiplier 들을 계산함으로써 데이터의 분류에 이용할 수 있게 된다.





(그림 3) 온도를 내리는 과정중 커널 행렬의 변화.(80℃->40℃, 50mer, 8개 베이스 변화) 위에서부터 각각 74.4℃, 67.7℃, 60℃, 40℃

4. 원형과 개념 학습

예시를 사용한 개념 학습에 있어서 분류 항목의 원형(prototype)에 해당되는 예시를 사용한 학습과 비원형에 해당되는 예시를 사용한 학습간에 차이를 보인다는 연구(Attneave[3])가 있다. 원형은 분류 항목을 가장 잘 나타내주는 예시이고, 비원형은 원형의 변이(distortion), 혹은 변형(transformation)이다. 일반적으로 원형을 예시로 사용한 학습이 비원형을 예시로 사용한 학습보다

분류의 오류율도 낮았으며, 분류에 걸리는 반응시간도 적게 걸린다고 알려져 있다. 하지만, 다양성(variability)이 제시되는 예시가 다양성이 덜 제시되는 예시보다 학습에 더 유리하다는 연구 결과가 있으며[2], 데이터의 평균이 원형으로서 기여하지 못한다는 연구도 있는 등[1] 예시를 통한 학습은 보다 일반적인 어떤 일정한 규칙에 의해 조절이 되지 않는 것으로 보였다.

본 연구를 개념 학습의 모사로 생각했을 때, 대각 행렬의 요소들은 데이터 자체의 두드러짐 정도를 나타내고, 대각 행렬의 요소가 아닌 값들이 데이터 사이의 유사도를 나타낸다. 예시를 통한 개념 학습에서의 예시에 해당하는 것들이 데이터에 해당되고, 다른 데이터와의 관계에서 희박성(sparsity)정도가 크게 나타나는 데이터가 원형, 그렇지 않은 데이터가 비원형에 해당된다고 할 수 있다.

일반적으로 커널을 이용한 SVM의 분류 현상이 원형과 같이 두드러진 데이터일수록 데이터에 해당되는 듀얼 폼의 상수 α_i 값은 0값을 가지기 어렵고, 이는 원형이 데이터의 분류에 일반적으로 많이 기여하게 된다는 설명이 된다. 위의 실험에서 DNA 길이에 비해 DNA간 변이를 많이 둔 데이터가 낮은 온도에서도 양한정성을 유지한다는 것은 이러한 데이터들간의 상관성을 줄였을 때,

데이터가 제대로 분류될 수 있는 가능성을 유지하기가 쉽다는 것을 의미한다. 하지만, 데이터가 너무 두드러져 커널 요소가 희박해지는 경우 이것 또한 분류를 못하게 하는 요소가 되는데, 이것이 학습에 있어서 데이터의 다양성(variability)이 확보되어 데이터간 관계가 설정되어야 하는 이유이다.

또한, α_i 값을 가지는 값들은 경계값이지 평균값이 아니다. 이것은 분류 항목에 해당되는 예시의 평균을 나타내는 예시가 이 분류 항목을 분류해 내는데 도움이 되는 예시가 되지 못하는 현상을 설명해 주기도 한다.

높은 온도에서 시작해 온도를 내려야 학습에 적합한 커널을 만들 수 있는 것은 대각 행렬값을 유지시켜주기 위한 작업이며, 예시들끼리의 관계가 만들어지기 전에 예시 자체의 개념이 먼저 두드러지게 적립되는 단계가 필요하다고 해석될 수 있다.

5. 결 론

DNA의 기작을 가지고 개념학습이라는 인지 작용에 적용시켜 보기 위해 커널 방법을 사용한 학습 방법을 제시하였다. 이는 패턴을 분류하는 도구로서 DNA 컴퓨팅을 이용하는 방법으로서 의의를 찾을 수 있을 뿐만 아니라 개념 분류에 있어서의 원형(prototype)과 관련된 연구 결과들이 설명하는 바와 상당히 유사한 현상을 지닌다는 것을 알 수 있다.

또한, 이는 실리콘 컴퓨터의 순차적 처리에 기반한 모델이 아닌 분자컴퓨팅의 병렬적 처리를 염두에 둔 모델이다. 이러한 병렬처리의 특성은 실재 DNA를 이용할 경우 많은 숫자의 기재가 참여하여 동시에 실리콘 컴퓨터가 해내기 힘든 양의 동작(operation)을 해 낼 수 있다는 점에서 기존의 시뮬레이션 방법들과 완전히 다른 방법이며, 심리현상을 모사하는 새로운 방법으로 제시될 수 있다.

6. 참고 문헌

[1] Neumann, P.G. (1977) Visual prototype ınformation with discontinuous representation of dimensions of variability. *Memory & Cognition, 5*, 187-197.

[2] Posner, M.I., & Keele, S.W. (1968) On the genesis of abstract ideas. *Journal of Experimental Psychology, 77*, 353-363.

[3] Attneave A. (1957) Transfer of experience with a class schema to identification learning of patterns and shapes. *Journal of Experimental Psychology, 54*, 81-88.

[4] A.L. Glass. & K.J. Holyoak (1986) *Cognition, 2nd Ed.* Random House.

[5] D.V. Howard (1983) *Cognitive Psychology - Memory, Language, and Thought*, Macmillan Publishing Co., Inc.

[6] Emile A. & Jan K. (1989) *Simulated Annealing and Boltzmann Machines*, John Wiley & Sons.

[7] J. Kım, J.J. Hopfield, & E. Winfree. (2004). Neural network computation by ın vitro transcriptional circuits, *Advances in Neural Information Processing Systems 2004.*

[8] Schoelkopf, B., & Smola, A. (2001) *Learning with Kernels*, MIT Press.

[9] R. I. Kondor, & J. Lafferty (2002) Diffusion Kernels on Graphs and Other Discrete Input Spaces. *ICML 2002.*

[10] L.M. Adleman (1994) Molecular Computation of Solutions To Combinatorial Problem. *Science, 266*, 1021-1024.