학습을 위한 네거티브 데이터가 존재하지 않는 경우의 microRNA 타겟 예측 방법

이제근[O1,2] 김수진[1,2] 장병탁[2,3]
서울대학교 생물정보학 협동과정
서울대학교 바이오정보기술 연구센터 (CBIT)
서울대학교 컴퓨터공학부
jkrhee@bi.snu.ac.kr, sjkim@bi.snu.ac.kr, btzhang@bi.snu.ac.kr

microRNA target prediction when negative data is not available for learning

Je-Keun Rhee[O1,2] Soo-Jin Kim[1,2] Byoung-Tak Zhang[2,3]
Graduate program in Bioinformatics, Seoul National University
Center for Bioinformation Technology (CBIT), Seoul National University
Department of Computer Science & Engineering, Seoul National University

요 약

기존의 알려진 데이터에 기반하여 분류 알고리즘을 통해 새로운 생물학적인 사실을 예측하는 것은 생물학 연구에 매우 유용하다. 하지만 생물학 데이터 분류 문제에서 positive 데이터만 존재할 뿐, negative 데이터는 존재하지 않는 경우가 많다. 이와 같은 상황에서는 많은 경우에 임의로 negative data를 구성하여 사용하게 된다. 하지만, negative 데이터는 실제로 negative임이 보장된 것이 아니고, 임의로 생성된 데이터의 특성에 따라 분류 성능 및 모델의 특성에 많은 차이를 보일 수 있다. 따라서 본 논문에서는 단일 클래스 분류 알고리즘 중 하나인 support vector data description (SVDD) 방법을 이용하여 실제 microRNA target 예측 문제에서 positive 데이터만을 이용하여 학습하고 분류를 수행하였다. 이를 통해 일반적인 이진 분류 방법에 비해 이와 같은 방법이 실제 생물학 문제에 보다 적합하게 적용될 수 있음을 확인한다.

1. 서 론

알려지지 않은 새로운 생명 현상을 밝혀내기 위한 생물학 연구에서는 최근 들어 컴퓨터를 이용한 분석이 중요한 위치를 차지하고 있다. 특히 기계학습(machine learning) 기술은 이러한 목적으로 매우 유용하게 이용될 수 있다. 기계학습 기법 중 대표적인 한 분류 중 하나인 감독 학습(supervised learning)에서는 주로 기존의 알려진 데이터를 기반으로 학습하여 새로운 데이터를 적절한 클래스로 분류(classification)하는 문제를 다루고 있다. 의사 결정 나무(decision tree), 인공 신경망(artificial neural networks), 베이지안망(Bayesian networks), 서포트 벡터 머신(support vector machine, SVM) 등이 이를 위해 일반적으로 사용되는 방법들이다. 생물학 연구에서는 이와 같은 분류 방법을 이용하여 새로운 사실을 예측할 수 있다. DNA 서열 정보에 기반한 유전자 예측, 각종 임상 정보 및 유전자 발현 데이터를 이용한 질병 진단 및 예측, 서열 및 구조적 특성을 이용한 단백질 상호작용 정보 예측 등이 생물학 연구에서 기계학습 기술이 일반적으로 많이 이용되는 대표적인 예라고 할 수 있다[1, 2, 3].

이와 같은 기계학습 기반의 분류 문제를 위해서는 적합한 학습 데이터가 필요하다. 일반적으로 학습 데이터는 positive data와 negative data로 구성된다. 기계학습 기술은 데이터의 특성에 기반한 방법이므로 학습 데이터의 구성에 따라 분류 모델 및 분류 성능에 큰 차이를 보이게 된다. 하지만 생물학 문제에서는 negative 데이터 구성에 어려움이 존재한다. Positive 데이터는 기존에 알려진 사실들에 기반하여 학습 데이터를 구성하는 것이 가능하다하지만 특정 현상에 대한 어떤 데이터가 실제로는 전혀 일어날 수 없는 사실인지, 혹은 실제로는 가능하지만 아직 실험적인 어려움으로 밝혀내지 못한 것인지를 구분하는 것이 어렵다. 따라서 많은 경우에 negative 데이터를 명확하게 정

의하기에는 어려움이 있다
Negative data가 존재하지 않는 경우 일반적으로는 임의(random)로 negative data를 발생시켜서 사용한다 하지만 임의로 만들어진 데이터가 실제로 negative에 적합한 데이터임을 증명할 수는 없다 또한 임의의 negative 데이터가 실제 positive 데이터와는 특성이 매우 많이 다른 경우, 실제 문제에서의 예측 성능은 많이 떨어지게 된다 따라서 생물학적 예측 모델을 만들 시 positive 데이터와 유사하면서, negative로 사용할 수 있는 데이터를 만들고자 많은 노력을 기울이고 있지만 negative 데이터 구성을 위한 보장된 방법은 없는 것이 사실이다
이와 같은 문제점을 해결할 수 있는 방법 중 하나는 단일 클래스 분류기(one-class classifier)를 사용하는 것이다 일반적인 분류 모델이 이진 분류 혹은 다중 분류 문제를 다룰 수 있는 데에 비해 단일 클래스 분류기는 positive 데이터만을 이용하여 특이점을 구분해내는 역할을 수행한다. 따라서 negative 데이터가 존재하지 않거나 소량만 존재하는 경우에도 효과적으로 예측하는 것이 가능하다
본 논문에서는 단일 클래스 분류기 중 하나인 SVDD (support vector data description) 방법을 이용하여 microRNA target 유전자를 예측한다(그림 1). SVDD는 대표적인 이진 분류기인 SVM의 변형된 형태로 positive 데이터만을 이용하여 분류할 수 있다[4]. 이 방법을 통해 기존의 연구와는 달리 임의의 negative 데이터를 생성하는 일 없이도 효율적인 분류가 가능함을 확인한다

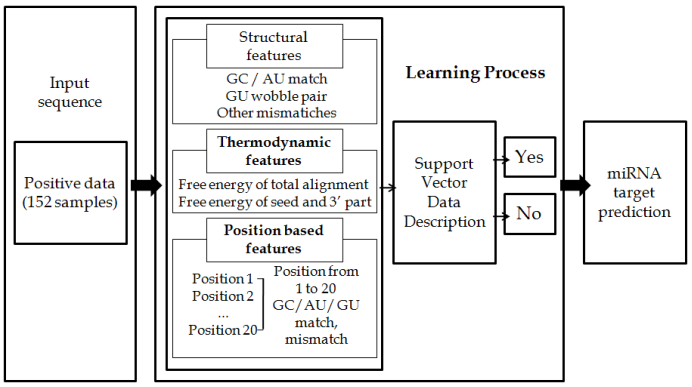

그림 1 전체적인 실험 개요

2. 실험 방법

2-1. Support vector data description (SVDD)

SVDD는 분류 대상이 되는 하나의 학습 클래스에 속한 데이터만을 이용하여 학습을 수행할 수 있는 단일 클래스 문제(one-class classification problems)를 해결하는데 유용한 기법 중 하나이다 단일 클래스 분류 알고리즘 중 가장 대표적인 SVDD는 그림 2와 같이 특이점(outlier)을 검출하여 주어진 목적(target) 데이터 대부분을 포함하는 경계선을 찾는다[4]. 이와 같은 경계선은 목적 데이터를 최대한 포함하고 특이점을 가장 적게 포함하는 중심 a와 반경 R로 구성된 최소한의 구(hypersphere)이다. 이는 d-차원 입력 공간에 존재하는 n개의 데이터로 구성되는 학습 데이터의 집합 $D=\{x_i \mid i=1,2,..,n\}$에 대해서 R^d 위에 정의되는 중심이 a이고 반경이 R인 구 H를 이용하여 학습 클래스의 영역을 표현하는 것이다 또, 각 학습 데이터 x_i와 중심 a 사이의 거리가 R을 초과하는 경우에는 벌점(penalty)을 부과하는 방법을 사용한다 에러 함수는 식 (1)과 같이 정의하며 이를 최적화하여 최소한의 반지름을 가지는 구를 구한다

$$\min F(R,a) = R^2 + C\sum_i \xi_i \qquad (1)$$
$$s.t.\ \|x_i - a\| \le R^2 + \xi_i, \xi_i \ge 0, \forall i$$

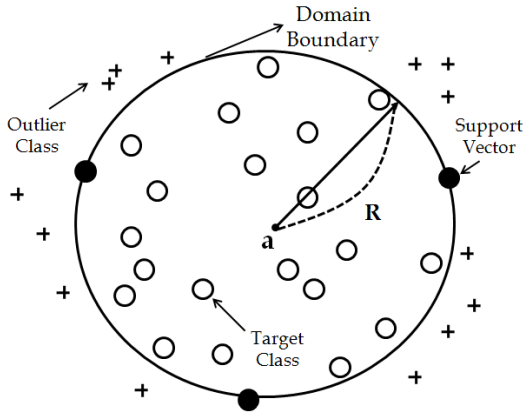

그림 2 SVDD의 기본 개념

여기에서 특이점과 구의 크기는 변수 ξ_i (slack variable)에 의해 조절된다. ξ_i는 i번째 학습 데이터 x_i가 구 H에서 벗어나는 벌점이며, C는 구의 크기, 즉 반지름과 에러의 상대적 중요성을 조절하는 상수(trade-off constant)이다. 이를 통해 성능 관점에서 C값이 클 경우 검출률이 높아지지만 이에 따라 에러 검출률도 높아지는 결과를 얻을 수 있고, C값이 작을 경우에는 작아진 구로 인해 검출률은 낮아지지만 에러 검출률은 줄어드는 결과를 얻을 수 있다 최적화 하는데 연산을 편리하게 하기 위해 Lagrangean multiplier를 도입하여 위의 식을 dual problem으로 변환하였다. Lagrangean multiplier를 이용하여 정리한 함수 L의 식은 식 (2)와 같다.

$$L(R,a,\alpha_i,\gamma_i,\xi_i) = R^2 + C\sum_i \xi_i \qquad (2)$$
$$-\sum_i \alpha_i R^2 + \xi_i - (\|x_i\|^2 - 2a \cdot x_i + \|a\|^2)$$
$$-\sum_i \gamma_i \xi_i$$
$$where, \alpha_i \ge 0 \text{ and } \gamma_i \ge 0$$

식 (2)에서 a_i, γ_i는 Lagrangean multiplier를 나타내며 a_i, ≥ 0, $\gamma_i \ge 0$의 조건을 가진다. 이와 같은 조건에서 a_i, γ_i는 최대화하면서 각 R, a, ξ_i에 대해서 최소화 하는 값으로 최적화한다. 식 (2)를 변수 R, a, ξ_i에 관하여 각각 편미분한 식을 0으로 하여 등식을 구성하여 $0 \le a_i \le 0$ 이라는 새로운

조건을 얻어 최소한의 구를 찾기 위해 학습해야 하는 최대화해야 하는 함수 L을 식 (3)과 같이 정리할 수 있다.

$$Max\, L = \sum_i \alpha_i (x_i \bullet x_i) - \sum_{i,j} \alpha_i \alpha_j (x_i \bullet x_j) \quad (3)$$
$$where, 0 \le \alpha_i \le C, \sum_i \alpha_i = 1, i = 1, ..., n$$

L 함수를 최대가 되도록 하는 Lagrangean multiplier, a_i를 구하여 구하고자 하는 구의 반경R과 중심 a를 구할 수 있다.
또, 입력 공간 위에서 정의되는 구는 매우 간단한 형태의 영역만을 표현 할 수 있으므로 커널 k를 통하여 정의되는 고차원 특징 공간(Feature space)으로의 변환을 통해 비선형적으로 분류를 가능하게 하여 보다 좋은 성능을 얻을 수 있다. 데이터들의 내적인 ($x_i \bullet x_j$)에 커널을 적용하여 다시 쓰면 식 (4)와 같다.

$$Max\, L = \sum_i \alpha_i K(x_i, x_i) - \sum_{i,j} \alpha_i \alpha_j K(x_i, x_j) \quad (4)$$
$$where, 0 \le \alpha_i \le C, \sum_i \alpha_i = 1, i = 1, ..., n$$

커널을 적용한 K($x_i \bullet x_j$)은 Mercer's 이론을 만족해야 한다. 또, 커널은 다항식(polynomial) 커널, Radial basis function (RBF) 커널, 가우시안 (Gaussian) 커널 등 여러 가지가 있다. 각각의 커널은 문제에 따라 다른 성능을 보이므로 풀고자 하는 문제에 적절한 커널을 이용하여야 한다.

2-2. microRNA target 예측

microRNA (miRNA)는 약 22nt 크기의 작은 RNA 분자로서, 유전자 발현양 조절에 중요한 역할을 수행하는 것으로 알려져 있다[5]. 유전자는 일반적으로 전사(transcription) 와 번역(translation) 과정을 차례로 거치면서, 단백질(protein)로 발현된다. 기존에는 유전자의 발현양이 전사 조절 인자(transcription factor)들에 의해 주로 조절받는 것으로 생각되어 왔다. 하지만 최근 들어 miRNA에 의한 조절 기작이 밝혀지면서 miRNA 연구는 생물학 분야에서 매우 중요한 위치를 차지하게 되었다.
세포 내에서 miRNA는 유전자가 mRNA로 전사된 후, mRNA의 3' UTR 영역에 결합하여, mRNA가 단백질로 번역되는 것을 억제하는 역할을 한다(그림 3). 따라서 특정 miRNA들이 어떤 유전자의 발현을 억제할 수 있는지 확인하는 것은 중요한 문제이다.

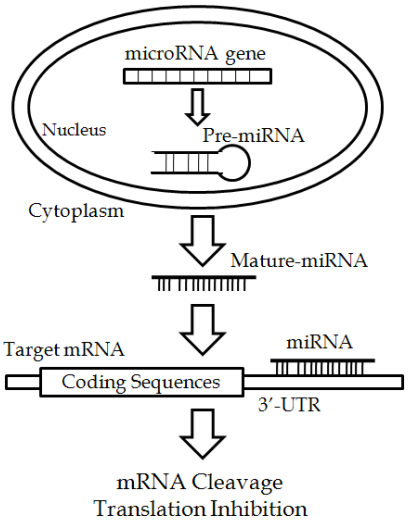

그림 3 miRNA의 세포내 조절 기작

miRNA의 target 유전자 예측은 많은 연구자들에 의해 활발하게 수행되고 있다. 대표적인 target 예측 방법들은 표 1에 보인다. miRNA target 예측은 일반적으로 상보적인 결합 서열 정보와 그 때의 구조 및 에너지 정보들에 기반하여 target 가능성을 예측한다.

표 1 대표적인 microRNA target 예측 연구

Target연구	웹사이트
miRanda[6]	http://www.microrna.org//miranda.html
PicTar[7]	http://pictar.bio.nyu.edu
TargetScan[8]	http://genes.mit.edu/targetscan
miTarget[9]	http://cbit.snu.ac.kr/~miTarget
RNAhybrid[10]	http://bibiserv.techfak.uni-bielefeld.de/rnahybrid/welcome.html

2-3. 실험 데이터 및 실험 설정

microRNA target 예측을 위한 데이터는 기존의miTarget 연구에서 사용된 데이터를 그대로 사용하였다[9]. miTarget 연구에서는 SVM 모델을 이용하여 miRNA target 유전자를 예측하였다. 이 데이터는 총 41개의 feature set으로 구성되며, 각각에 대한 정보는 표 2에 보인다.

표 2 실험에서 사용된 feature 구성

Feature 구분	Feature의 특성
Structural Features	miRNA와 유전자가 결합하였을 때의 구조적 특성 정보
Thermodynamic Features	결합 에너지에 기반하여 얻어진 특성
Position-based Features	각 위치별 서열 조성 정보

SVDD 실험은 Gaussian RBF kernel을 이용하였고, sigma 값은 3으로 하여 실험하였다.

3. 실험 결과

SVDD를 이용한 miRNA target 예측은 miTarget 연구에서 사용한 데이터 중 152개의 문헌에서 증명된 positive 데이터만을 이용하여 학습하였다. 이 데이터로부터 5-fold cross-validation을 통해 얻어진 예측 정확도는 75.5%로 나왔다. 또한 본 연구에서는 miTarget 연구에서 사용된 negative 데이터를 일부 추출하여 특이점 정보로 넣고 학습을 수행해보았다. 표 3은 152개의 positive 데이터만을 이용하여 학습한 결과와 negative 데이터를 그 개수를 변화시켜가면서 추가하였을 경우의 예측 성능을 5-fold cross-validation을 통하여 얻은 결과를 보여준다.

표 3 negative 데이터의 개수 변화에 따른 SVDD 성능 변화

Negative data 수	False Positive	False Nagative
0	0.24494	NA
10	0.24494	0.0000
30	0.2385	0.0666
50	0.2514	0.0800

일반적으로 SVDD는 positive 데이터만을 이용하여 학습하지만, 일부 negative 데이터를 특이점 정보로 넣어줌으로서 보다 정확한 결과를 얻을 수 있다. 본 실험에서는 문헌에서 알려진 실제로 microRNA가 target 하지 않는다는 사실이 생물학 실험적으로 알려진 정보를 특이점으로 사용하여 학습 결과를 비교하였다. 본 실험에서는 negative 데이터의 구성에 따라 그 성능에 조금씩 차이를 보이기는 하였으나 현저히 많은 차이를 보이지는 못하였다. 이는 특이점으로 사용된 negative 데이터가 그 수도 적고 대체로 비슷한 조건하에서 실험된 결과로 그 특성이 크게 다르지 않은 경우가 많이 포함되어 있기에, 최종 성능에 큰 영향을 주지는 못한 것으로 여겨진다.

표 4 SVDD와 SVM의 분류 성능 비교

	SVDD	SVM
TP	0.7885	0.9808
TN	0.6774	0.0645

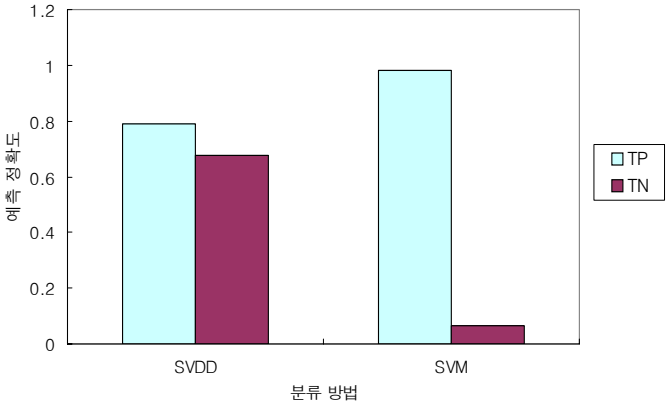

그림 4 SVDD와 SVM에서의 True Negative 차이 비교

표 4는 독립적인 테스트 데이터를 구성하여 SVDD와 SVM을 이용한 예측 결과를 비교한 표이다. 이 실험에서는 100개의 positive 데이터와 10개의 negative 데이터로 학습한 후, 총 84개의 독립적인 테스트 데이터로 예측 성능을 비교한 것이다.

일반적으로 SVM은 이진 클래스 분류 문제에서 우수한 성능을 보이는 것으로 알려져 있다. 하지만 현재의 문제와 같이 이용 가능한 negative 데이터가 부족한 경우에는 표 4에서 보는 것과 같이 SVM에서는 그 결과를 정확하게 분류해내지 못하고 있다. 표 4에서 보는 것과 같이, SVM을 이용한 예측에서는 true negative의 경우 약 6.45%의 값을 보이고 있다. 이는 실제로 microRNA의 target이 안되는 결과에 대해서도 SVM 모델에서는 대부분 positive로 예측함을 의미한다. 이에 비해 SVDD 방법은 SVM과는 달리 true negative의 경우에서도 70%에 가까운 성능을 보이고 있다. 따라서 SVM과 같은 이진 분류기는 본 논문에서 다루고 있는 것과 같은 문제를 해결하기에는 적합하지 않은 방법이다. 즉 실제로 negative임이 밝혀진 사례가 절대적으로 부족한 상황에서는 SVM과 이진 분류기를 사용하는 것에 문제가 있음을 명확히 알 수 있다.

4. 결론

본 논문에서는 생물학적인 분류 및 예측 문제에서 SVDD의 적용 가능성을 살펴보았다. 유전자 조절 기작에 중요한 역할을 하고 있는 것으로 알려진 microRNA 관련 연구를 통해, 기존의 target 예측 연구보다 SVDD를 이용하는 것이 보다 효율적일 수 있음을 실험적으로 보여주었다.
일반적으로 SVDD와 같은 단일 클래스 분류 알고리즘은 전체 데이터에서 일부의 특이점을 인지하거나 noise를 찾는 문제 등에서 주로 이용되어왔다[11]. 하지만 본 연구에서는 증명된 negative 데이터를 얻기 힘든 생물학 문제에 단일 클래스 분류 알고리즘을 이용함으로서, 기존의 연구들과는 다른 성격의 생물학 연구에도 이와 같은 알고리즘이 유용하게 사용될 수 있음을 보였다.
기계학습 기반의 일반적인 분류 알고리즘들은 적절한 학습 데이터가 존재할 때, 그 성능이 극대화될 수 있다. 하지만 실제 생물학 문제에 적용하기에는 데이터 수도 부족하며, 특히 연구가 많이 진행되지 않았거나 최근 들어 활발히 연구가 진행되고 있는 중요한 문제들의 경우에서는 더더욱 확실하게 증명된 데이터를 얻기 어렵다는 문제가 있다. 따라서 일반적으로 많이 쓰는 이진 분류 알고리즘 혹은 다중 분류 알고리즘을 적용하여 예측하는 경우에는 본 논문에서 보인 것과 같이 유용한 결과를 얻기 어려울 수 있는 것이다.
SVM 방법에 기반한 단일 분류기인 SVDD 외에도 K-nn 기반 단일 분류기, 가우시안 혼합 모델(gaussian mixture model) 기반 단일 분류기 등 다른 형태의 단일 분류 알고리즘에 대한 연구 역시 활발히 진행되고 있다. 이러한 단일 클래스 분류 알고리즘들은 각 데이터의 특성에 따라 유용하게 사용될 수 있을 것이다.

참고문헌

[1] A. Bernal, K. Crammer, A. Hatzigeorgiou, F. Pereira, Global Discriminative Learning for Higher-Accuracy Computational Gene Prediction. *PLoS Comput Biol.,* 3(3): e54, 2007.
[2] T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, D. Haussler, Support vector machine classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics*, 16(10): 906-914, 2000.
[3] L. Nanni and A. Lumini, An ensemble of K-local hyperplanes for predicting protein-protein interactions *Bioinformatics*, 22(10): 1207-1210, 2006.
[4] D.M.J. Tax and R.P.W. Duin, Support Vector Data Description, *Machine Learning*, 54(1): 45-66, 2004.
[5] D.P. Bartel, MicroRNAs: genomics, biogenesis, mechanism, and function, *Cell,* 116(2): 281-297, 2004.
[6] B. John, A.J. Enright, A. Aravin, T. Tuschl, C. Sander, D.S. Marks, Human MicroRNA targets, *PLoS Biol.*, 3(7): e264, 2005.
[7] D. Grün, Y.L. Wang, D. Langenberger, K.C. Gunsalus, N. Rajewsky, microRNA target predictions across seven Drosophila species and comparison to mammalian targets, *PLoS Comput. Biol.,* 1: e13, 2005.
[8] B.P. Lewis, C.B. Burge, D.P Bartel, Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets, *Cell,* 120: 15-20, 2005.
[9] S.-K. Kim, J.-W. Nam, J.-K. Rhee, W.-J. Lee, B.-T. Zhang. miTarget: microRNA target-gene prediction using a Support Vector Machine, *BMC Bioinformatics,* 7(1): 411, 2006.
[10] M. Rehmsmeier, P. Steffen, M. Höchsmann, R. Giegerich, Fast and effective prediction of microRNA/target duplexes, *RNA*, 10: 1507-1517, 2004.
[11] V. Hodge and J. Austin, A Survey of Outlier Detection Methodologies, *Artif. Intell. Rev.,* 22(2): 85-126, 2004.