병렬 단백질 상호작용 예측 시스템

김세영° 정유진
한국외국어대학교 컴퓨터공학과
{sayoung°, chungyj}@hufs.ac.kr

A Parallel System for predicting protein-protein interactions

Seyoung Kim° Yoojin Chung
Dept. of Engineering, Hankuk Univ. of Foreign Studies

요 약

최근 단백질간의 상호작용의 중요성의 이해와 함께 축적되어가는 단백질 정보들 간의 상호작용을 예측하기 위하여 통계학적 모델인 Support Vector Machine(SVM)을 사용한 예측 실험이 활발하다. 하지만 이는 거대한 생물 데이터를 처리하기위해 많은 연산시간을 필요로 한다. 즉, 방대하게 존재하는 데이터를 처리하기위해 SVM을 통한 실험은 정확한 결과뿐만 아니라 빠른 처리속도를 요구하게 되었다. 따라서 본 논문에서는 SVM의 개선을 통해 빠른 처리속도로 데이터를 처리하는 incremental SVM과 이를 병렬화 하여 더욱 빠른 처리시간을 가지는 Parallel SVM(PSVM)을 소개하고 실험해 본다. 즉, 단백질 상호작용에 사용되어지는 데이터를 PSVM을 사용한 실험을 통하여 정확성과 처리속도를 측정, 비교함으로써 단백질 상호작용 예측에 적합한지를 검증해본다.

1. 서 론

최근 단백질간의 상호작용의 중요성의 이해와 함께 축적되어가는 단백질 정보들 간의 상호 작용을 예측할 수 있도록 예측시스템에 대한 실험이 활발하다. 그중 Support Vector Machine(SVM)을 사용과 새로운 feature model의 등장에 따른 단백질 상호작용의 예측에 대한 연구가 활발하다[1][2]. 하지만 SVM는 대량의 데이터를 처리하기위해 많은 연산시간 필요로 한다. 즉, 방대하게 존재하는 생물 데이터를 처리하기위해 SVM[5][7]을 통한 실험은 정확한 결과뿐만 아니라 빠른 처리속도를 요구하게 되었다. 이에 따라 많은 처리시간을 소요하는 알고리즘을 간단하고 빠른 처리속도를 가지는 알고리즘으로 개선한 Proximal SVM[8]이 등장하였지만 이는 연산을 위한 많은 메모리를 필요로 하는 단점이 있다. 따라서 적은 메모리를 사용하고, 빠른 연산이 가능한 Incremental Support Vector Machine[6]이 나타나게 되었다. 이는 데이터를 나눠서 처리함으로써 적은 메모리를 사용하고, 더욱 빠른 처리속도를 나타낸다. 또한 이를 병렬화시킨 Parallel SVM(PSVM)은 더더욱 빠른 처리시간을 갖게 되었다[4]. 따라서 이 논문에서는 Incremental Support Vector Machine[6]을 병렬화한 PSVM을 소개하고, 실험을 통하여 PSVM[4]이 단백질 상호작용 예측시스템에 사용되어지는 것이 적합한지를 검증해 본다. 즉, 순차적으로 수행되는 standard SVM인 Tiny-SVM[3]과 PSVM의 실험에 따른 정확도(accuracy)를 비교 제시하고, PSVM에서 노드수에 따른 처리 속도를 비교해 본다.

* 본 연구는 한국과학재단 목적기초연구(R01-2003-000-10860-0) 지원으로 수행되었음

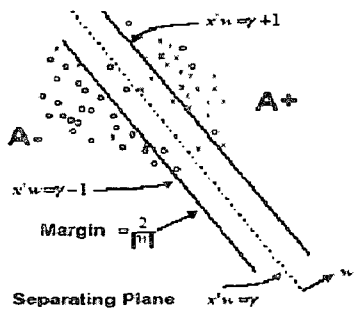
2. 실험개론

2.1 Support Vector Machine(SVM)

SVM은 기본적으로 두 범주를 갖는 객체들을 분류하는 방법이다. 단백질 상호작용 예측에서는 상호작용이 '있다', '없다'라는 구분을 가능하게 해야 하는데 SVM 적절한 모델로써 최근에 그 성능을 인정받아 다양한 예측시스템에서 적용되고 있는 개념이다.

2.2 Proximal Support Vector Machine

Standard SVM은 두 객체를 분리하는 최적의 하이퍼플레인을 찾는 문제로 귀착되어지며 [그림 1]과 같이 나타내어진다. 여기에서 하이퍼플레인은 [식 1]과 같이 나타내어진다.

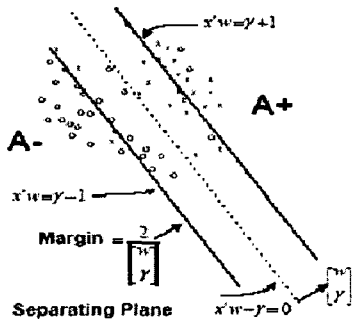


[그림 1] Standard SVM classifier

$$\min_{(w,\gamma,y)\in\mathbb{R}^{n+1+m}} \{ve'y + \frac{1}{2}w'w\}$$
$$\text{s.t. } D(Aw - e\gamma) + y \geq e$$
$$y \geq 0 \quad (1)$$

$A \in \mathbb{R}^{m\times n}$, $D \in \{-1,+1\}^{m\times 1}$, $e = 1^{m\times 1}$

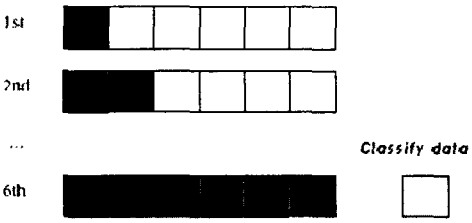
Fung and Mangasarian[8]은 [식 1]을 [식 2]와 같이 대체함으로써 많은 연산시간을 발생하게 만들었던 알고리즘을 간단하고 빠른 알고리즘으로 바꾸었다[그림 2]. 후에 *Fung and Mangasarian*[6]에 의해 [식 2]는 increments와 decrements를 가능하게 개선되었다.



[그림 2] Proximal SVM classifier

$$\min_{(w,\gamma,y)\in R^{n+1+m}} \quad \frac{\nu}{2}\|y\|^2 + \frac{1}{2}(w'w+\gamma^2)$$
$$\text{s.t.} \quad D(Aw-e\gamma)+y = e \quad (2)$$

이는 전체 데이터를 한번에 처리하는게 아닌 [그림 3]과 같이 일정한 크기로 데이터를 나눈 후 각 각 처리하는 방법으로써 많은 연산시간을 줄이게 하였다.



[그림 3] Incremetal SVM classifier

따라서 [식 2]는 ω와 γ를 구하기 위해 [식 3]과 같이 전개 된다.

$$\begin{bmatrix} w \\ \gamma \end{bmatrix} = \begin{bmatrix} A'A+\frac{I}{\nu} & -A'e \\ -e'A & \frac{1}{\nu}+m \end{bmatrix}^{-1} \begin{bmatrix} A'De \\ -e'De \end{bmatrix} = \left(\frac{I}{\nu} + \begin{bmatrix} A' \\ -e' \end{bmatrix} \begin{bmatrix} A & -e \end{bmatrix} \right)^{-1} \begin{bmatrix} A'De \\ -e'De \end{bmatrix}.$$

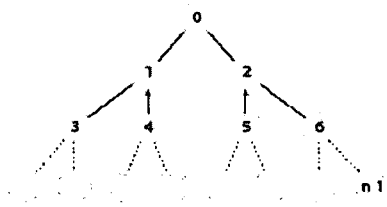
Defining

$$E = [A \quad -e].$$

$$\begin{bmatrix} w \\ \gamma \end{bmatrix} = (\frac{I}{\nu} + E'E)^{-1}E'De. \quad (3)$$

2.3 Incremental SVM의 병렬화

incrementals$((E^i)'E^i)$, $((E^i)'d^i)$ 요소는 Heap-based Tree Topology를 이용한 병렬화 계산을 가능하게 하였다. 각 각의 $(E^i)'E^i$와 d^i 를 leaf node에서 구하고 마지막 연산을 top node가 수행하게 된다면 incremental SVM과 같은 알고리즘을 수행하게 하는 것과 같게 되어진다. 이에 따라 병렬처리가 가능한 [그림 4]와 같은 구조를 가지는 PSVM[4]이 구현 되었다.

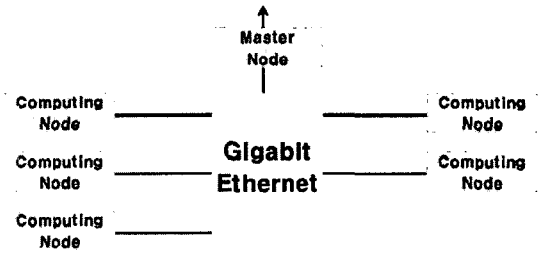


[그림 4]Parallel heap-based tree topology for Incremental PSVM

3. 실험 및 평가

3.1 실험 환경

본 논문에서의 테스트 환경은 순차적으로 수행되는 Standard SVM인 Tiny-SVM의 성능 측정 실험 시에는 Dual 2.8GHz의 CPU와 2G의 메인 메모리를 사용하는 Linux 시스템을 이용하였다. PSVM의 성능 측정시에는 866MHz의 CPU와 256MB의 메인 메모리를 가진 컴퓨터 6대를 기가비트 이더넷 기반의 PC-Cluster[그림 5]로 구성하였다. 그리고 Linux 시스템을 이용하였고 병렬처리를 위한 라이브러리로 MPI를 사용하였다.



[그림 5] Gigabit Ethernet 기반의 PC-Cluster

3.2 실험 데이터 생성

기존의 연구[9]는 단백질의 소수성을 이용한 1차구조로부터 상호작용을 예측하였기 때문에 기존의 data형식을 사용하였다. 즉, Database of Interacting Proteins(DIP)[10]에서 제공하는 yeast종의 상호작용 리스트를 사용하였고, 상호작용 리스트가 1500개인 8개의 test set를 만들어서 7개를 training 할 때 사용, 8번째 data를 classify시 사용하였다.

3.3 실험 및 평가

본 논문에서는 단백질 상호작용 예측에 있어서 Standard SVM중 하나인 Tiny-SVM과 병렬화한 PSVM에 대한 성능을 비교 평가하기 위한 것이므로 실험 데이터 처리에 대한 정확도(accuracy)를 구하고 비교하였다. 그리고 PSVM에 대한 노드수에 따른 처리속도의 변화를 측정하고 평균을 구하여 비교하였다.

실험 결과는 [표 1], [표 2]와 같다. [표 1]은

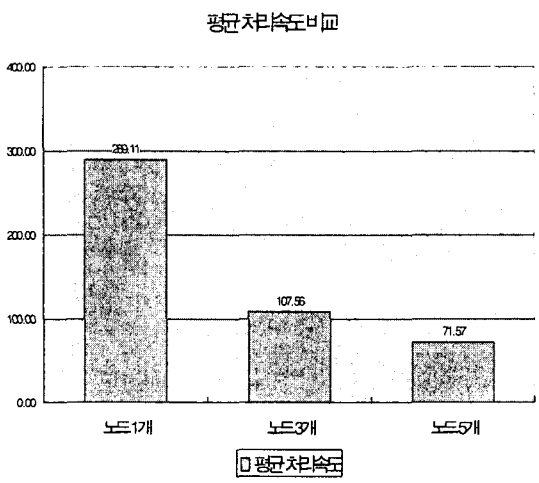
Tiny-SVM과 PSVM의 정확도를 비교한 것이고, [표 2]는 PSVM의 노드의 사용 개수에 따른 처리 속도를 구하고 비교한 것이다. 그리고 [도표 1]은 [표 2]에 대한 평균을 그래프로 나타낸 것이다.

[표 1] 실험 데이터의 정확성(accuracy)비교

test set	Tiny-SVM	PSVM
1	90.82%	78.84%
2	93.16%	82.71%
3	89.74%	81.95%
4	94.93%	82.83%
5	95.95%	84.23%
6	92.21%	83.09%
7	94.81%	88.22%
평균	93.09%	83.12%

[표 2] 실험 데이터의 처리속도 비교(단위:초)

test set	병렬화된 PSVM		
	node 1개	node 3개	node 5개
1	289.7	118.08	73.46
2	307.1	107.27	71.09
3	289.1	109.05	72.45
4	285.1	106.12	70.60
5	291.9	104.59	71.48
6	277.2	107.69	69.66
7	283.6	105.13	72.22
평균	289.11	107.56	71.57



[도표 1] 실험 데이터의 평균 속도변화(단위:초)

4. 결론 및 향후 연구과제

[표 1]에서 Tiny-SVM의 정확성은 평균 93.09% 그리고 병렬화 된 PSVM은 평균 83.12%로 Tiny-SVM의 정확성이 평균 9.97% 조금 더 높게 나타나고 있음을 알 수 있다. [표 2]에서는 병렬화된 PSVM의 노드수에 따른 처리속도를 비교한 것으로써, 노드가 3개일 때는 1개일 때보다 2.68배 빠른 처리속도를 나타내고, 노드가 5개일 때는 1개일 때보다 4.03배 빨라짐을 알 수 있다. 즉, PSVM은 노드의 개수가 증가함에 따라 속도도 또한 비례에서 빨라지고 있다는 것을 확인할 수 있다. 따라서 PSVM이 단백질 상호작용 예측 시스템에 사용되어지기 위해서는 PSVM의 정확성을 높이는 방법이 연구되어야 하고 노드 개수의 증가를 통해 더욱 빠른 속도향상을 기대한다면 상호작용 예측에 사용되어도 무리가 없을것으로 보인다. 그래서 향후 우리는 노드개수를 늘이고, 10000개 이상의 데이터를 가지는 대량의 데이터에 대한 실험으로 상호작용 리스트를 실험하는 연구를 진행 할 계획이다.

Reference

[1] Yoojin Chung, Gyeong-Min Kim, Young-Sup Hwang, And Hoon Park, Predicting Protein-protein Interactions from One Feature Using SVM, IEA/AIE 2004, LNAI 3029, (2004), pp.50-55

[2] Bock,J.R. and Gougj,D.A. Predicting protein-protein interactions from primary structure. Bioinformatics, 17, (2001), pp. 455-460.

[3] TinySVM
http://cl.aist-nara.ac.jp/~taku-ku/software/TinySVM

[4] Amund Tveit and Havard Engum : Parallelization of the Incremental Proximal Support Vector Machine Classifier using a Heap-based Tree Topology

[5] Burges, C.J.C., "A tutorial on support vector machines for pattern recognition", Data Mining and Knowledge Discovery, 2, (1998), pp.121-167

[6] Fung, G., Mangasarian, O.L.:Incremental Support Vector Machine Classification. In Grossman, R., Mannila, H., Motwani, R., eds.:Proceedings of the Second SIAM International Conference on Data Mining, SIAM (2002), pp. 247-260

[7] Vapnik, V., The Nature of Statistical Learning Theory, Springer-Verlag, New York. (1995)

[8] Fung, G., Mangasarian, O.L.:Multicategory Proximal Support Vector Classifiers. Submitted to Machine Learning Journal (2001)

[9] 김철환, 정유진 : 단백질 상호작용 예측을 위한 SVM의 부정예제 생성방법론, 정보과학회 (2004)

[10] DIP (Database of Interacting Proteins)
http://dip.doe-mbi.ucla.edu/