고차원 데이터의 분류를 위한 서포트 벡터 머신을 이용한 피처 감소 기법

고석하, 이현주
광주과학기술원 정보통신공학과
e-mail : sukdo@gist.ac.kr, hyunjulee@gist.ac.kr

Feature reduction for classifying high dimensional data sets using support vector machine

Seok-Ha Ko, Hyun-Ju Lee
Department of Information and Communications
Gwangju Institute of Science and Technology

Abstract

We suggest a feature reduction method to classify mouse function data sets, which integrate several biological data sets represented as high dimensional vectors. To increase classification accuracy and decrease computational overhead, it is important to reduce the dimension of features. To do this, we employed Hybrid Huberized Support Vector Machine with kernels used for a kernel logistic regression method. When compared to support vector machine, this approach shows the better accuracy with useful features for each mouse function.

I. 서론

생물학에서 대용량의 데이터가 기하급수적으로 증가하면서, 이러한 데이터를 활용한 단백질의 기능 예측 또한 그 중요성이 증가하고 있다. 현재 관련 분야의 연구가 폭넓게 수행되고 있으며, 그 예로서 마우스의 단백질 기능을 예측하는 연구 (Mouse Gene Function Prediction Project) [1] 등이 있다. 단백질 기능과 관련성이 높은 다양한 생물학적 데이터들을 통합하는 방법론을 개발하는 것이 이러한 연구들의 주요 주제이다. 특히, 모든 데이터들을 통합하는 것보다 특정한 데이터들을 선별하여 분류 (classification)할 경우 식별 성능을 높일 수 있을 뿐만 아니라 계산 비용을 줄일 수 있으므로, 피처 선택 및 감소 (feature selection and reduction)에 대한 연구들이 활발히 수행중이다.

본 논문에서는 마우스의 단백질 기능을 예측하기 위해서, support vector machine (SVM) 을 사용하여 Pfam과 Mouse Atlas of Gene Expression (Sage), Phylogenetic Profile, Diseases [1]의 네 가지 데이터들을 통합하였다. 이 데이터들은 고차원의 벡터들로 표현되므로, 피처 감소를 위하여 커널 로지스틱 회귀 분석 (kernel logistic regression) 방법론 [2]에 기반을 두어 SVM의 커널을 구성하였으며, Hybrid Huberized Support Vector Machine (HHSVM) [3]을 사용하여 중요성이 높은 데이터들을 선발하였다.

II. 본론

2.1 Hybrid Huberized Support Vector Machine

SVM은 현재 알려져 있는 많은 분류 모델 중에서 가장 인식 성능이 뛰어난 학습 모델의 하나이며, 마이크로 어레이 (microarray) 데이터를 이용한 암환자의 분류 등 다양한 분야에서 폭넓게 사용된다.

기본적인 SVM은 중요성이 높은 피처들을 자동적으로 선택할 수 없다는 단점이 있다. 이를 개선하기 위해 L1-norm SVM이 개발되었다. 하지만, 이 역시 선택되는 피처의 수가 훈련 데이터 (training data)에 의해 제한되고 매우 높은 상관관계를 갖는 유전자들의

경우, 이 중 몇 개만이 선택된다는 단점이 있다. 이 후, huberized hinge loss function과 elastic-net penalty를 이용한 HHSVM이 제안되었다. HHSVM은 분류하는 데이터의 클래스와 연관이 높으며, 서로 매우 높은 연관관계를 갖는 유전자들을 함께 선택하므로 마이크로어레이를 사용한 분류에 매우 유용하다 [3].

2.2 커널 (Kernel) 정의

Gene Ontology (GO) 중 Biological Process의 GO function 예측을 위해 앞서 언급한 4개의 데이터를 사용하였다. 각 데이터는 많게는 3,186개의 피처들을 갖고 있으므로 효율적인 분류를 위해 피처 감소 기법이 필요하다. 이를 위해 이산 값 (discrete value)을 갖는 Pfam, Phylogenetic Profile, Diseases 3개의 데이터는 다음과 같이 피처를 감소시켰다 [2]. Pfam의 경우 i-th 단백질이 k-th 도메인 (domain)을 가지고 있으면 v_{ik} = 1 이며, 그 외의 경우에는 v_{ik} = 0 이다. 단백질 i와 j의 커널은 $K^{pfam}(i,j)=v_i v_j$ 로 정의한다. Phylogenetic profile과 Diseases 데이터의 커널도 Pfam과 유사하게 정의하였다. 연속 값 (continuous value)을 갖는 Sage의 커널은 Pearson Correlation Coefficient (PCC)를 사용하여 $K^{sage}(i,j)=PCC(i,j)$로 정의한다.

j-th 단백질이 특정 GO function을 가지고 있을 경우 Xj = 1, 아닐 경우 Xj = 0로 나타낸다면, Pfam을 사용한 SVM의 피처는 다음과 같이 구성한다.

$$Feature^{pfam}{}_1(i) = \sum_{j<>i, X_j known} K^{pfam}(i,j) I\{X_j=1\} / \sum_{j<>i, X_j known} I\{X_j=1\}$$

$$Feature^{pfam}{}_0(i) = \sum_{j<>i, X_j known} K^{pfam}(i,j) I\{X_j=0\} / \sum_{j<>i, X_j known} I\{X_j=0\}$$

Phylogenetic profile과 Diseases, Sage 데이터의 피처도 Pfam과 유사하게 정의하였다.

Ⅲ. 구현

Pfam과 Sage, Diseases, Phylogenetic Profile 4개의 데이터는 각각 2개의 피처로 감소된다. 총 8개의 피처를 가지는 841개의 단백질로 구성된 데이터로 20개의 GO function을 예측하였다. R 프로그램의 SVM (패키지 e1071)과 HHSVM [3]을 사용하였다. 5-폴드 크로스 밸리데이션 (5-fold cross-validation)을 한 후, AUC (Area Under the ROC Curve) 값을 구해 정확도를 비교하였다 (표 1).

GO Function	GO function을 가지는 단백질 수	정확도 (%)	
		SVM	HHSVM
GO:0006461	7	71.4	49.7
GO:0035239	15	67.6	72.2
GO:0045045	6	74	77
GO:0019932	12	95.3	94.7
GO:0019226	30	82.7	78.7
GO:0030097	23	77.7	84.9
GO:0048598	28	63.1	73
GO:0031175	19	73.8	76.9
GO:0007243	26	59.3	76.5
GO:0016070	6	70.9	75.2
GO:0045893	46	79.3	83.8
GO:0006519	39	85.6	66.7
GO:0009056	43	71.1	69.1
GO:0007049	28	81.8	80
GO:0007155	24	77.9	79.2
GO:0006091	32	77.1	75.7
GO:0046907	32	71.4	75.4
GO:0006793	41	65.7	85.1
GO:0006796	41	83	90.2
GO:0016265	67	48.7	60.7
평 균		73.87	76.26

표 1. GO Function 예측의 정확도

Ⅳ. 결론 및 향후 연구 방향

대용량의 데이터를 이용한 단백질의 기능 예측은 매우 중요하며, 이를 위해서는 피처 선택 및 감소가 필수적다. 본 연구를 커널 (Kernel) 기법 [2]과 HHSVM [3]을 활용하여 마우스의 단백질 기능을 예측하였다. HHSVM이 SVM에 비해 뛰어난 정확도를 보였고, 실험에 사용된 5개의 GO function을 분석한 결과 대체적으로 Diseases 데이터는 예측의 기여도가 낮은 반면 Pfam 데이터와 Sage 데이터의 기여도가 높았다. 향후 이러한 피처 선택 및 감소 방법을 로지스틱 회귀 분석 등에 적용하고자 한다.

감사의 글

본 연구는 광주과학기술원의 GIST faculty start-up fund의 지원에 의한 것입니다.

참고문헌

[1] Pena-Castillo, L., et al,. (2008) A critical assessment of M. musculus gene function prediction using integrated genome evidence. Genome biology, In press.

[2] Lee, H., et al., (2006) Diffusion Kernel-Based Logistic Regression Models for Protein Function Prediction. OMICS: A Journal of Integrative Biology, 10(1): 40-55.

[3] Ji Zhu and Hui Zou, (2008) Hybrid huberized support vector machines for microarray classification and gene selection, Bioinformatics, 24(3): 412-419.