JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지(J. Korea Inst. Inf. Commun. Eng.) Vol. 18, No. 4 : 992~999 Apr. 2014

단백질의 세포내 위치 예측을 위한 다중레이블 분류 방법의 성능 비교

지상문*

A Performance Comparison of Multi-Label Classification Methods for Protein Subcellular Localization Prediction

Sang-mun Chi*

School of Computer Science and Engineering, Kyungsung University, Busan 608-736, Korea

요 약

단백질이 존재하는 세포내의 다중 위치를 정확하게 예측하기 위하여 다중레이블 학습 방법을 광범위하게 비교한다. 이를 위하여 다중레이블 분류의 접근 방법인 알고리즘 적응, 문제 변환, 메타 학습의 여러 방법을 비교 평가한다. 다양한 관점에서 다중레이블 분류 방법의 특성을 평가하기 위하여 12가지 평가 척도를 사용하였고, 최적의 성능을 보이는 방법을 찾기 위하여 새로운 요약 척도를 사용하였다. 비교 실험 결과, 흔하지 않은 다중레이블 집합을 가지치기 하는 멱집합 방법과, 관련 레이블들을 추가된 특징으로 나타내는 분류기-체인 방법의 성능이 높았다. 또한, 이들 방법들로 구성된 여러 개의 분류기를 조합하면 더욱 성능이 향상되었다. 즉, 세포내 위치간의 연관관계를 사용하는 것이 예측에 효과적인데, 특정 생물학적 기능을 수행하는 단백질의 세포내 위치들의 관계는 독립적이지 않고 서로 관련되어 있기 때문이라 판단된다.

ABSTRACT

This paper presents an extensive experimental comparison of a variety of multi-label learning methods for the accurate prediction of subcellular localization of proteins which simultaneously exist at multiple subcellular locations. We compared several methods from three categories of multi-label classification algorithms: algorithm adaptation, problem transformation, and meta learning. Experimental results are analyzed using 12 multi-label evaluation measures to assess the behavior of the methods from a variety of view-points. We also use a new summarization measure to find the best performing method. Experimental results show that the best performing methods are power-set method pruning a infrequently occurring subsets of labels and classifier chains modeling relevant labels with an additional feature. futhermore, ensembles of many classifiers of these methods enhance the performance further. The recommendation from this study is that the correlation of subcellular locations is an effective clue for classification, this is because the subcellular locations of proteins performing certain biological function are not independent but correlated.

키워드 : 다중레이블 분류, 다중레이블 평가 척도, 다중위치 단백질, 단백질 세포내 위치

Key word : Multi-label classification, Multi-label evaluation measures, Multiplex proteins, Protein subcellular localization

접수일자 : 2014. 02. 25 심사완료일자 : 2014. 03. 27 게재확정일자 : 2014. 04. 07

* **Corresponding Author** Sang-Mun Chi (E-mail:smchiks@ks.ac.kr,Tel:+82-51-663-5146)
School of Computer Science and Engineering, Kyungsung University, Busan 608-736, Korea

Open Access http://dx.doi.org/10.6109/jkiice.2014.18.4.992 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

단백질은 생명체내에서 효소, 영양 저장, 호르몬, 운동, 면역, 정보교환, 구조형성 등의 많은 기능을 수행한다. 동물, 식물, 곰팡이와 같은 진핵생물은 세포 내부의 정교한 구획과 세포소기관이 존재하는데, 이들 지역은 서로 다른 생화학적 환경이 생겨 세포내 위치에 따라 특정한 대사 기능을 수행한다. 따라서 단백질의 기능을 알기 위한 기초 지식은 단백질이 존재하는 세포내의 위치를 알아내는 것이다. 단백질의 세포내 위치를 예측하는 많은 연구들은 오직 하나의 세포내 위치에 존재하는 단백질만을 대상으로 하였지만, 여러 세포내 위치에 동시에 존재하는 단백질의 생물학적 기능이 중요하므로, 이를 예측하려는 시도가 커지고 있다[1-8]. 다중 세포내 위치 예측은 하나의 단백질에 대해 세포내 위치를 나타내는 레이블 집합의 부분집합을 예측하는 다중레이블 분류문제이다. 세포내 위치 개수 Q개에 대하여 지수적 조합인 2^Q개의 가능한 예측결과가 존재하므로 기존의 분류 방법으로는 처리하기 어렵다.

다중 세포내 위치 예측에 적용할 수 있는 기계학습 방법인 다중레이블 분류 방법은 최근 들어 활발히 연구되고 있다[9-11]. 이는 이미지, 비디오, 텍스트, 음악, 마케팅, 생물학 분야에서 하나의 입력 자료에 대해 여러 가지 분류에 동시에 속하는 상황이 발생하기 때문이다. 다중레이블 분류 방법을 (1)알고리즘 적응, (2)문제 변환, (3)메타 학습으로 나누어 볼 수 있다[9-11]. 알고리즘 적응 방법은 기존의 단일 분류 알고리즘인 최근접-이웃 분류기, 신경망, 결정 트리, 지지 벡터 기계를 다중레이블에 맞도록 변형한 방법이다. 문제 변환 방법은 다중레이블 분류문제를 다수의 단일레이블 분류문제로 변환한 후에 단일레이블 분류 알고리즘을 적용하는 방법이다. 메타 학습 방법은 알고리즘 적응방법이나 문제 변환 방법의 여러 개를 동시에 적용하여 조합하여 분류기를 구성하는 방법이다.

다중레이블 분류를 이용하여 단백질의 세포내 위치를 예측 방법들을 살펴보면, 최근접-이웃 분류기의 앙상블을 사용하는 방법[1, 7], 가우시안 과정 모델과 공분산 행렬로 레이블간의 연관성을 표현하는 방법[3]이 있다. 문제 변환 방법을 사용하는 예로, 세포내 위치의 모든 쌍들에 대한 분류기를 구성하여 투표하여 최종 결과를 얻는 방법[2]과 특정 레이블에 관련된 사례들과 관련되지 않은 사례들로 학습하는 BR(binary relevance)을 사용하는 방법[4,6]이 있다. 또한, Q개 이진 분류기를 체인으로 연결하고, k-번째 분류기는 $k-1$ 까지 레이블의 관련성을 이용하는 문제 변환 방법의 일종인 CC(classifier chain)를 앙상블로 사용하는 방법[5]이 있다. 이밖에 LP(label powerset) 방법처럼 레이블 부분집합을 무작위로 만들고, 사례와 관련된 레이블들을 CC처럼 사례의 속성으로 사용하는 방법[8]이 있다.

최근에 개발된 다양한 다중레이블 분류 방법이 단백질의 다중 세포내 위치 예측의 적용을 위하여 충분히 비교 분석되지 않았다. 본 논문에서는 광범위한 다중레이블 분류 방법의 비교를 통하여, 단백질 세포내 위치 예측에 효과적인 방법을 알아내고, 그 방법들의 특징을 분석한다. 또한, 다중레이블 분류의 복잡한 예측결과를 다양한 측면에서 살펴보기 위하여 12개의 평가 척도를 사용하고, 새로운 요약 척도를 사용하여 최적의 방법들을 찾는다.

Ⅱ. 다중레이블 분류 방법

단백질 세포내 위치 예측에 효과적인 방법을 찾고자 다중레이블 분류 방법을 알아본다. 다중레이블 분류는 사례(example)와 관련된 다중 레이블을 찾는다. 즉, 레이블들의 집합을 $L = \{\lambda_1, \lambda_2, ..., \lambda_Q\}$로 나타낼 때, 각 사례와 관련된 레이블 부분집합 $y_i \subseteq L$을 예측한다. 따라서 각 사례에 대해 하나의 레이블로만 분류하는 단일레이블 분류에 비하여 알고리즘의 복잡성과 분류의 정확성을 평가하는 척도가 다양하다.

이번 장에서는 IV장의 비교 실험에 사용되는 다중레이블 분류 알고리즘을 중심으로 하여 (1)알고리즘 적응, (2)문제 변환, (3)메타 학습으로 나누어 살펴 본다[9-11]. 알고리즘 적응 방법은 단일레이블 분류 알고리즘인 최근접-이웃 분류기, 트리 분류, 신경망, 지지 벡터 기계 등을 다중레이블 분류에 적합하도록 확장하거나 변형한다. Ml-knn의 경우에는 각 사례에 제일 근접하는 이웃을 찾고, 평가 자료의 레이블집합을 결정하는 과정에서 각 레이블에 대하여 k 개의 최근접 이웃의 사전확률과 사후확률을 사용한다[12]. BRkNN도 최근접-이웃 분류기를 사용하고, 다중레이블 분류문제를 아래에 설명할 문제 변환인 BR(binary relevance)을 사용한다

[13]. IBLR_ML은 로지스틱 회귀와 최근접-이웃 분류기를 결합한 방법이다[14]. BPMLL은 기존의 역전파 학습을 사용한 신경망을 다중레이블 분류를 고려하여 새로운 오류 함수를 도입한 방법이다[15].

문제 변환 방법은 다중레이블 분류를 다수의 단일레이블 분류로 바꾸는데, BR(binary relevance), LP(label power-set)와 PW(pair-wise)로 나누어 볼 수 있다[9-11]. BR은 각 레이블에 관련된 사례로 양성 집합(positive set), 이외의 사례로 음성 집합(negative set)을 구성하여 학습하고, 각 레이블에 해당하는 분류기의 결과를 조합한다. CC(classifier chains)는 BR과 유사한 방법으로 Q개의 이진 분류기를 체인으로 연결한 것이고, k-번째 분류기는 $k-1$까지의 레이블이 관련되었는지 아닌지에 따라 0 또는 1의 레이블을 자료의 속성에 추가한다[16]. LP는 각 사례와 관련된 레이블 집합을 묶어서 새로운 단일 레이블로 나타낸다. 이는 직접적으로 레이블 간의 연관관계를 이용할 수 있지만, 새로 만든 레이블들의 수가 매우 커질 수 있으므로 적은 빈도의 레이블은 제거하여 PS(pruned sets)을 만드는 방법이 있다[17]. PW는 모든 레이블 쌍에 대해 분류기를 구성하고, 투표를 사용하여 분류한다. PW의 일종인 CLR(calibrated label ranking)은 추가로 인공적인 레이블을 도입하여, 이 레이블보다 높은 점수를 갖는 레이블을 관련된 사례로 예측한다[18].

메타 방법은 여러 다중 레이블 분류기를 배깅(bagging), 부스팅(boosting)과 스태킹(stacking)을 사용하여 조합한다. 배깅은 동일한 종류의 여러 분류기를 조합하는 방법으로 투표를 사용 하는 경우에 각 방법에 동일한 가중치를 부여한다. ECC(ensemble of classifier chains)와 EPS(ensemble of pruned sets)는 각각 CC[16]와 PS[17]로 다수의 분류기의 구성하는 배깅 방법이다. HOMMER는 여러 레이블들 간의 계층적 관계를 구성하고, 각 계층에서 분류기를 구성하는 방법이다. 각 분류기는 Q개의 모든 레이블보다 적은 레이블들을 처리하며, 자료수가 더 균형 있게 배분되고 유사한 레이블은 하나의 부분집합에 속하도록 분할한 후에 학습을 수행한다[19]. ClusteringBased는 자료를 군집화를 통하여 몇 개의 군집으로 분리한 후에, 각 군집에 대하여 다중분류기를 적용하는 배깅 방법이다[20]. RAkEL은 크기가 k인 부분 레이블 집합을 여러 개 만들고, 이를 단일 레이블로 간주하여 기본 분류기를 구성하여 배깅을 적용한다[21]. 부스팅 방법은 새로운 학습 모델은 앞서 구성된 학습 모델의 분류 결과를 이용한다. 즉, 이전에 잘못 처리된 사례들에 대하여 더 높은 가중치를 부여하여 새로운 모델을 구성한다. AdaBoostMH는 부스팅 방법을 사용하여 다음 장에서 알아 볼 평가척도인 hamming_loss를 최소화하도록 방법이다[22]. 스태킹은 여러 분류기의 결과를 투표를 사용하지 않고 다른 학습 알고리즘을 사용하여 조합 한다. MultiLabelStacking은 먼저 각 사례에 대한 기반이 되는 여러 분류기의 출력을 얻고, 이 출력을 다시 메타 학습기에 입력하여 최종 결과를 얻는다[23].

Ⅲ. 다중레이블 분류의 평가 척도

다중레이블 분류기에 대한 성능 평가는 기존의 단일레이블 분류기에 사용되는 성능 척도를 그대로 사용할 수 없다. 즉, 예측된 레이블이 실제 레이블과 일치하는 것만을 판단하면 지나치게 엄격한 평가 척도가 되므로, 일부만 일치하는 경우도 고려한다. 따라서 여러 관점에서 예측 정확도를 판정할 수 있는 방법들이 사용되고 사례기반(example-based)과 레이블기반(label-based)으로 나눈다[9-11]. 식 (1)~(6)의 사례기반 방법은 각 사례에 대해 실제 레이블과 예측된 레이블간의 차이를 평균하고, 식 (7)~(12)의 레이블기반 방법은 각 레이블에 대해 개별적으로 예측성능을 구하고 이를 평균한다.

사례 x_i의 실제 다중레이블을 y_i로, 예측된 레이블을 $h(x_i)$로 나타낼 때, *hamming_loss*는 실제 레이블이 아닌 것이 예측된 수와 실제 레이블이 예측되지 않은 수를 평균한 것이고, 작을수록 성능이 높다. 식 (1)에서 $\triangle$는 두 집합의 대칭차집합, $|\ |$은 집합의 원소수, N은 사례의 총 개수이다.

$$hamming\_loss(h) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{Q}|h(x_i)\triangle y_i| \quad (1)$$

다음에 설명할 척도들은 값이 클수록 정확한 예측이다.

$$accuracy(h) = \frac{1}{N}\sum_{i=1}^{N}\frac{|h(x_i)\cap y_i|}{|h(x_i)\cup y_i|} \quad (2)$$

*precision*과 *recall*은 각각 예측된 레이블 $h(x_i)$과 실제 다중레이블 y_i 중에서 정확히 예측된 비율이다.

$$precision(h) = \frac{1}{N}\sum_{i=1}^{N}\frac{|h(x_i)\cap y_i|}{|h(x_i)|} \quad (3)$$

$$recall(h) = \frac{1}{N}\sum_{i=1}^{N}\frac{|h(x_i)\cap y_i|}{|y_i|} \quad (4)$$

F_1는 *precision*과 *recall*의 조화평균으로, 완전한 예측이 되었을 때는 1이고, 최저는 0이다.

$$F_1 = \frac{1}{N}\sum_{i=1}^{N}\frac{2\times|h(x_i)\cap y_i|}{|h(x_i)|+|y_i|} \quad (5)$$

*subset_accuracy*는 예측된 레이블이 실제 레이블과 정확히 일치하는지를 평가한다. $I(h(x_i)=y_i)$는 $h(x_i)$와 y_i가 같으면 1이고, 아니면 0이다.

$$subset\_accuracy(h) = \frac{1}{N}\sum_{i=1}^{N} I(h(x_i)=y_i) \quad (6)$$

*macro_precision*은 각 레이블에 대해서 *precision*을 구한 후에, 이를 평균한다. 식 (7)에서 tp_j(true positive)와 fp_j(false positive)는 레이블 λ_j와 이외의 레이블을 이진 분류하는 것에서 계산된다.

$$macro\_precision = \frac{1}{Q}\sum_{j=1}^{Q}\frac{tp_j}{tp_j+fp_j} \quad (7)$$

*macro_recall*은 각 레이블에 대해서 *recall*을 구하여 평균한다. 식 (8)에서 fn_j(false negative)은 레이블 λ_j와 이외의 레이블을 이진 분류하는 것에서 계산된다.

$$macro\_recall = \frac{1}{Q}\sum_{j=1}^{Q}\frac{tp_j}{tp_j+fn_j} \quad (8)$$

*macro_*F_1은 레이블 $\lambda_j \in y_i$에 대해서 계산한 *precision*과 *recall*인 p_j와 r_j를 이용하여 조화평균을 구하고, 이들을 레이블에 대해서 평균한다.

$$macro\_F_1 = \frac{1}{Q}\sum_{j=1}^{Q}\frac{2\times p_j\times r_j}{p_j+r_j} \quad (9)$$

*micro_precision*과 *micro_precision*은 다음과 같다.

$$micro\_precision = \frac{\sum_{j=1}^{Q} tp_j}{\sum_{j=1}^{Q} tp_j+\sum_{j=1}^{Q} fp_j} \quad (10)$$

$$micro\_recall = \frac{\sum_{j=1}^{Q} tp_j}{\sum_{j=1}^{Q} tp_j+\sum_{j=1}^{Q} fn_j} \quad (11)$$

*micro_*F_1은 *micro_precision*과 *micro_precision*의 조화평균이다.

$$micro\_F_1 = \frac{2micro\_precision\times micro\_recall}{micro\_precision+micro\_recall} \quad (12)$$

본 연구에서는 식 (1)~(12)의 많은 평가 척도를 사용하므로 각 방법의 비교가 쉽지 않다. 이를 요약하기 위한 새로운 척도로서 합계를 구하는데, *hamming_loss*는 작은 값 일수록 성능이 높으므로 1-*hamming_loss*를 더하였다. 즉, 각 방법의 성능을 통합해서 비교할 경우에는 다음의 *S-measure* (Sum of measures)를 사용하였다.

S-measure = 1-hamming_loss + accuracy + precision + recall + F_1 *score + subset_accuracy + macro_precision + macro_recall + macro_*F_1 *+ micro_precision + micro_recall + micro_*F_1

Ⅳ. 단백질 세포내 위치 예측 비교 실험

단백질의 세포내 위치 예측에 효과적인 다중레이블 분류방법을 찾기 위해 비교 하였다. 실험에 사용한 자료는 인간 단백질 자료로 세포내 위치는 14개(centriole, cytoplasm, cytoskeleton, endoplasmic reticulum, endosome, extracell, golgi apparatus, lysosome, microsome, mitochondrion, nucleus, peroxisome, plasma membrane, synapse)이고, 2,580개의 단백질은 하나의 세포내 위치, 480개는 두 개의 위치, 43개는 3개의 위치, 3개는 4개의 위치에 존재한다[1]. 단백질 서열들은 25% 이하의 작은

서열 동일성을 가지고 있으므로, 서열 유사성만을 이용하여 단백질의 세포내 위치를 예측하기는 어려운 자료이다.

비교 실험에는 5겹 교차검증을 수행하기 위해서 실험 자료를 균등하게 5개로 나누어 사용하였다. 단백질 자료를 다중레이블 분류를 위한 특징 벡터로 변환하기 위해서 논문[1,2,4-7]처럼, 유전자 온톨로지를 가진 데이터베이스(http://www.ebi.ac.uk/GOA)를 탐색하여 주어진 단백질 자료와 가장 유사한 단백질의 유전자 온톨로지를 사용하였다. 유전자 온톨로지는 분자적 기능, 생물학적 과정, 세포 요소의 관점에서 특징화한 용어로 유전자를 표현한 것으로, 유전자 해당하는 단백질의 특징을 나타낸다.

다중레이블 분류기는 Mulan 라이브러리로 구현하였고, 기본 설정을 사용하였다[9]. 각 방법의 성능은 III장의 평가척도 식(1)~(12)로 측정하였고, 비교를 위해 간략한 척도 *S-measure*로 나타내었다.

표 1은 II장에서 설명한 알고리즘 적응 방법의 실험 결과이다. 알고리즘 적응 방법은 이후의 비교 방법들보다 성능이 저조하므로, 다중레이블 분류를 위해서 보다 개선된 알고리즘의 확장이 필요하다.

표 1. 알고리즘 적응 방법의 성능 비교
Table. 1 Comparison of algorithm adaptation methods

Ml-knn	BRkNN	IBLR_ML	BPMLL
8.71	7.97	8.69	3.87

표 2. 문제 변환 방법의 성능 비교
Table. 2 Comparison of problem transformation methods

BR(NaiveBayes)	8.67
BR(SMO)	9.28
BR(J48)	9.37
CC(NaiveBayes)	8.72
CC(SMO)	9.30
CC(J48)	9.26
LP(NaiveBayes)	9.09
LP(SMO)	**9.45**
LP(J48)	8.98
PS(NaiveBayes)	9.11
PS(SMO)	**9.47**
PS(J48)	8.91
CLR(NaiveBayes)	8.68
CLR(SMO)	9.33
CLR(J48)	9.38

문제 변환 방법은 다중레이블 분류문제를 다수의 단일레이블 분류문제로 변환하고 단일레이블 분류를 하므로, 단일 분류 방법이 필요하다. 이를 위해 기본적인 문제 변환 방법인 BR에 대해서 단일 분류 방법들을 비교하여 우수한 단일 분류기를 이후에 사용하였다. Mulan 라이브러리가 기반한 Weka[24]에서 베이즈 분류기(NaiveBayes, NaiveBayesMultinomial), 지지 벡터 기계(SGD, SMO), 예제 기반 lazy 분류기(IBk, KStar, LWL), 신경망(MultilayerPerceptron), 트리 기반 분류기(J48, LMT, RandomForest, REPTree)를 비교하였다. 베이즈 분류기에서는 NaiveBayes, 지지 벡터 기계에서는 SMO, 트리 기반 분류기에서는 J48 방법이 우수하였다. 예제 기반 lazy 분류기는 성능이 저조하고, 신경망은 실험 시간이 오래 걸려서 문제 변환 방법에 적용하기 어려웠다. 표 2는 문제 변환 방법의 성능이고, 괄호 안에 사용한 단일 분류 방법을 표시하였다.

메타 방법 중에서 HOMMER, ClusteringBased, RAkEL은 다중레이블 분류기를 조합하는 방법이다. 표 1과 표 2에서 성능이 9.3 이상인 BR(J48), CC(SMO), LP(SMO), PS(SMO), CLR(SMO), CLR(J48)에 대하여 조합한 결과를 표 3에 나타내었다. 표 3에서 - 표시는 메모리 부족으로 실행 실패를 나타낸다. 본 논문에서는 16 기가바이트의 메모리를 사용하였고, 5겹 교차 검증을 병렬로 수행하였다.

표 3. 다중레이블 방법을 조합한 메타방법의 성능 비교
Table. 3 Comparison of meta algorithms using combination of multi-label classification methods

메타방법 / 다중레이블 분류	HOMMER	Clustering Based	RAkEL
BR(J48)	9.37	9.17	-
CC(SMO)	9.30	9.28	9.12
LP(SMO)	**9.45**	9.34	9.22
PS(SMO)	**9.46**	9.34	-
CLR(SMO)	9.33	9.32	-
CLR(J48)	9.38	9.19	-

표 4는 메타 방법 중에서 배깅을 사용하는 방법인 ECC와 EPS의 결과이다. 이 방법들에서도 단일분류로 NaiveBayes, SMO, J48을 적용하였다.

표 4. 배깅을 사용한 메타방법의 성능 비교
Table. 4 Comparison of meta algorithms using bagging

메타방법 / 단일 분류	ECC	EPS
NaiveBayes	8.81	9.17
SMO	**9.46**	**9.55**
J48	**9.55**	9.34

메타 방법에서 부스팅을 사용하는 AdaBoostMH는 *S-measure*가 8.72로 성능이 저조하였다. 이 밖에 메타 방법에서 스태킹을 사용하는 MultiLabelStacking의 실험을 위해, 각 사례에 대해 적용되는 기반 분류기로 NaiveBayes, SMO, J48로 사용하였고, 분류기들의 출력을 조합하는 메타 학습기로 역시 NaiveBayes, SMO, J48을 사용하여 총 9개의 조합을 비교하였다. 이 방법은 *S-measure*가 모두 9.3이하로 성능이 저조하였다.

비교실험한 모든 다중레이블 분류 방법에서 성능이 9.4이상인 방법들에서 대해서, 부가적으로 단백질의 특징을 보다 효과적으로 표현하면 분류 성능이 향상되는지를 실험하였다. 카이제곱 검정값을 변형하여 단백질의 세포내 위치를 판별력이 높게 나타내는 유전자 온톨로지 가중하는 방법을 적용하였다[25]. 원래의 방법이 가장 유사한 서열의 유전자 온톨로지만을 사용하는 것에 반하여, 본 연구에서는 가장 유사한 두 개의 서열에서 나타나는 유전자 온톨로지의 빈도를 이용하였다.

표 5. 유전자 온톨로지 가중을 사용한 다중레이블 분류기 성능
Table. 5 Performance of multi-label classifiers using weighted gene ontology terms

표 2의 LP(SMO)	9.59
표 2의 PS(SMO)	9.64
표 3의 HOMMER LP(SMO)	9.59
표 3의 HOMMER PS(SMO)	9.64
표 4의 ECC(SMO)	9.69
표 4의 ECC(J48)	9.58
표 4의 EPS(SMO)	9.69

표 6은 표 5에서 성능이 가장 높은 ECC(SMO)와 EPS(SMO)를 식 (1) ~ (12)의 평가 척도로 다양한 관점의 성능을 보여준다. 표 6의 두 가지 방법은 전체적인 성능은 유사하지만, *precision, recall*이 상반된 값을 갖는다. 이러한 경향은 레이블기반 평가척도에서 유사한 것으로부터, EPS에 비해 ECC가 실제 레이블보다 더 많은 레이블을 예측함을 알 수 있다.

표 6의 결과를 실험 방법은 약간씩 다르지만, 동일한 단백질 자료를 사용하는 다른 방법들과 비교한다. 논문[3]의 실험결과에 따르면, 최근접-이웃 분류기를 배깅 으로 조합하는 Hum-mPLOC 2.9[1]은 *recall,* F_1*, subset_accuracy*가 0.519, 0.541, 0.294이고, 알고리즘 적응 방법인 논문[3]은 0.643, 0.506, 0.202로 성능이 저조하다. 논문[4]에서는 문제 변환 방법 BR을 사용하고 *subset_accuracy*가 0.45이하이다. ECC를 사용하는 방법[5]는 *accuracy, precision, recall,* F_1값이 0.7913, 0.8249, 0.8404, 0.8191로서 표 6의 실험 결과보다 약간 저조하다. 각 논문들에서 단백질 자료를 특징벡터로 변환하는 유전자 온톨로지 이용 방법이 본 논문과 동일하지 않으므로 정확한 다중레이블 분류기의 성능 비교는 아니다. 하지만, 논문을 통한 비교에서 보듯이 알고리즘 적응방법이나 BR방법에 비하여 ECC가 효과적인 경향은 유사하다.

본 연구에서는 기본 설정의 파라미터를 가진 다중레이블 분류방법을 사용하였음에도 불구하고, 비교한 논문들의 방법보다 성능이 높았다. 따라서 우수한 접근법인 EPS, ECC를 문제에 적합하게 최적화하면 더욱 향상된 성능을 얻을 수 있다고 판단된다.

표 6. 다양한 평가 척도를 사용한 다중레이블 분류기의 성능
Table. 6 Performance of multi-label classifiers using various evaluation measures

메타방법 / 평가 척도	ECC(SMO)	EPS(SMO)
hamming_loss	0.0318	0.0310
accuracy	0.7990	0.7992
precision	0.8297	0.8529
recall	0.8494	0.8273
F_1	0.8394	0.8399
subset_accuracy	0.6929	0.7205
macro_precision	0.8112	0.8266
macro_recall	0.7164	0.6831
*macro_*F_1	0.7496	0.7331
micro_precision	0.8108	0.8374
micro_recall	0.8149	0.7862
*micro_*F_1	0.8128	0.8110
S-measure	9.6854	9.6861

Ⅴ. 결 론

본 논문에서는 단백질의 세포내 위치 예측을 위하여 여러 다중레이블 분류방법을 광범위하게 비교하였다. 또한, 다중레이블 분류방법을 비교하기 위하여 일부분 척도가 보다 다양한 척도를 사용하여 다양한 관점에서 요약할 수 있는 척도로 쉽게 비교하여 적합한 방법 선택할 수 있게 하였다.

비교 실험을 통하여 살펴보면, 단백질을 세포내 위치 예측에는 세포내 위치간의 연관관계를 학습 모델에 포함하는 방법이 성능이 높았다. 이러한 연관관계를 자료의 속성에 추가하는 CC(classifier chain)나 관련된 레이블 부분 집합으로 직접적으로 레이블간의 연관관계를 표현하고 적은 빈도의 집합을 제거하는 PS(Pruned sets)를 사용하여 여러 분류기를 구성하고 이를 배깅으로 조합하는 것이 가장 성능이 좋았다.

본 논문에서는 많은 수의 분류기를 비교하기 위하여 기본적인 설정을 사용하였으나, 향후에는 효과적인 것으로 밝혀진 CC나 PS로 구성된 여러 분류기를 조합하는 방법을 최적화하고, 분류기를 구성할 때 사용되는 단일 분류기의 파라미터를 최적화하는 것이 필요하다.

감사의 글

이 논문은 2014년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NRF-2012R1A1A4A01010310)

REFERENCES

[1] H.-B. Shen and K.-C. Chou, “A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0,” *Anaytical Biochemistry,* vol. 394, no. 2, pp. 269-274, 2009.

[2] S.-M. Chi and D. Nam, “WegoLoc: accurate prediction of protein subcellular localization using weighted gene ontology terms,” *Bioinformatics,* vol. 28, no. 7, pp. 1028-1030, 2012.

[3] J. He, H. Gu, and W. Liu, “Imbalanced multi-modal multi-label learning for subcellular localization prediction of human proteins with both single and multiple sites,” *Plos One,* vol. 7, no. 6, e37155, 2012.

[4] S. Mei, “Multi-label multi-kernel transfer learning for human protein subcellular localization,” *Plos One,* vol. 7, no. 6, e37716, 2012.

[5] G.-Z. Li, X. Wang, X. Hu, J.-M. Liu, and R.-W. Zhao, “Multilabel learning for protein subcellular location prediction,” *IEEE transactions on Nanobioscience,* vol. 11, no. 3, pp. 237-243, 2012.

[6] S. Wan, M.-W. Mak, and S.-Y. Kung, “mGOASVM: multi-label protein subcellular localization based on gene ontology and support vector machines,” *BMC Bioinformatics,* 13:290, 2012.

[7] W.-Z. Lin, J.-A. Fang, X. Xiao, and K.-C. Chou, “iLoc-Animal: a multi-label learning classifier for predicting subcellular localization of animal proteins,” *Molecular BioSystems,* vol. 9, no. 4, pp. 634-644, 2013.

[8] X. Wang and G.-Z. Li, “Multilabel learning via random label selection for protein subcellular multilocations prediction,” *IEEE transactions on computational biology and bioinformatics,* vol. 10, no. 2, pp. 436-446, 2013.

[9] G. Tsoumakas, I. Katakis, and I. Vlahavas, “Mining multi-label data,” in *Data Mining and Knowledge Discovery Handbook.* Boston, MA: Springer, ch. 34, pp. 667-685, 2010.

[10] G. Madjarov, D. Kocev, D. Gjorgjevikj, and S. Dzeroski, “An extensive experimental comparison of methods for multi-label learning,” *Pattern Recognition,* vol. 45, no. 9, pp. 3084-3104, 2012.

[11] M.-L. Zhang and Z-H. Zhou, “A review on multi-label learning algorithms,” *IEEE transactions on knowledge and data engineering,* http://doi.ieeecomputersociety.org/10.1109/TKDE.2013.39.

[12] M.-L. Zhang and Z-H. Zhou, "Ml-knn: A lazy learning approach to multi-label learning,” *Pattern Recognition,* vol. 40, no. 7, pp. 2038-2048, 2007.

[13] E. Spyromitros, G. Tsoumakas, and I. Vlahavas, "An Empirical Study of Lazy Multilabel Classification Algorithms," *in Proceeding of the 5th Hellenic Conference on Artificial Intelligence,* pp. 401-406, 2008.

[14] W. Cheng and E. Hullermeier, "Combining instance-based learning and logistic regression for multilabel classification," *Machine Learning,* vol. 76, no. 2-3, pp. 211-225, 2009.

[15] M.-L. Zhang and Z-H. Zhou, “Multi-label neural networks with applications to functional genomics and text cate-

gorization," *IEEE transactions on knowledge and data engineering,* vol. 18, no. 10, pp. 1338-1351, 2006.

[16] J. Read, B. Pfahringer, H. Geoff, and F. Eibe, "Classifier Chains for Multi-label Classification," *Machine Learning,* vol. 85, no. 3. pp. 335-359, 2011.

[17] J. Read, B. Pfahringer, and H. Geoff, "Multi-Label Classification using Ensembles of Pruned Sets," *in Proceeding of the 8th IEEE International Conference on Data Mining,* pp. 995-1000, 2008.

[18] J. Furnkranz, E. Hullermeier, E. L. Mencia, and K. Brinker, "Multilabel classification via calibrated label ranking," *Machine Learning,* vol. 73, no. 2, pp. 133-153, 2008.

[19] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Effective and Efficient Multilabel Classification in Domains with Large Number of Labels," *in Proceeding of ECML/PKDD 2008 Workshop on Mining Multidimensional Data (MMD'08),* pp. 30-44. 2008.

[20] G. Nasierding, G. Tsoumakas, and A. Kouzani, "Clustering Based Multi-Label Classification for Image Annotation and Retrieval," *in Proceeding of 2009 IEEE International Conference on Systems, Man, and Cybernetics,* pp. 4514-4519, 2009.

[21] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Random k-Labelsets for Multi-Label Classification," *IEEE transactions on knowledge and data engineering,* vol. 23, no. 7, pp. 1079- 1089, 2011.

[22] R. E. Schapire and Y. Singer, "BoosTexter: A boosting-based system for text categorization," *Machine learning,* vol. 39, no. 2-3, pp. 135-168, 2000.

[23] G. Tsoumakas, A. Dimou, E. Spyromitros, V. Mezaris, I. Kompatsiaris, and I. Vlahavas, "Correlation-Based Pruning of Stacked Binary Relevance Models for Multi-Label Learning," *in Proceeding of ECML/PKDD 2009 Workshop on Learning from Multi-Label Data (MLD'09),* pp. 101-116, 2009.

[24] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," ACM SIGKDD *explorations newsletter,* vol. 11, no.1, pp. 10-18, 2009.

[25] S.-M. Chi, "Prediction of protein subcellular localization by weighted gene ontology terms," *Biochemical and biophysical research communications,* vol. 399, no. 3, pp. 402-405, 2010.

지상문(Sang-Mun Chi)

1991년 서울대학교 수학교육학과 졸업(이학사)
1993년 한국과학기술원 수학과 졸업(이학사)
1998년 한국과학기술원 전산학과 졸업(공학박사)
1993년 ~ 2000년 삼성전자 무선사업부 선임연구원
2001년 ~ 현재 경성대학교 컴퓨터공학부 교수
※관심분야 : 생물정보학, 기계학습, 비선형최적화