JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지(J. Korea Inst. Inf. Commun. Eng.) Vol. 18, No. 7 : 1749~1756 July. 2014

다중레이블 조합을 사용한 단백질 세포내 위치 예측

지상문*

Multi-Label Combination for Prediction of Protein Subcellular Localization

Sang-mun Chi*

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

요 약

단백질이 존재하는 세포내 위치에 대한 지식은 단백질의 기능과 관련된 중요한 정보이다. 본 논문은 개선된 레이블 멱집합 다중레이블 분류방법을 제안하여 단백질이 존재하는 세포내의 다중 위치를 예측한다. 다중레이블 분류 방법 중에서 레이블 멱집합 방법은 특정 생물학적 기능을 수행하는 단백질의 세포내 위치간의 연관 관계를 효과적으로 모델링할 수 있다. 본 논문은 다중레이블을 다른 다중레이블들의 선형조합으로 나타낼 때의 조합가중치를 제약조건이 있는 최적화를 통하여 구하고, 이를 사용하여 여러 다중레이블의 예측 확률들을 조합하여 최종적인 예측을 수행한다. 인간 단백질 자료에 대한 실험에서 제안한 방법이 다른 단백질 세포내 위치 예측 방법에 비하여 높은 성능을 보였다. 이는 제안한 방법이 레이블 멱집합 방법에서 사용되는 다중레이블들내에 존재하는 중복 정보를 이용하여 다중레이블의 예측확률을 성공적으로 강화할 수 있기 때문이다.

ABSTRACT

Knowledge about protein subcellular localization provides important information about protein function. This paper improves a label power-set multi-label classification for the accurate prediction of subcellular localization of proteins which simultaneously exist at multiple subcellular locations. Among multi-label classification methods, label power-set method can effectively model the correlation between subcellular locations of proteins performing certain biological function. With constrained optimization, this paper calculates combination weights which are used in the linear combination representation of a multi-label by other multi-labels. Using these weights, the prediction probabilities of multi-labels are combined to give final prediction results. Experimental results on human protein dataset show that the proposed method achieves higher performance than other prediction methods for protein subcellular localization. This shows that the proposed method can successfully enrich the prediction probability of multi-labels by exploiting the overlapping information between multi-labels.

키워드 : 단백질 세포내 다중 위치, 다중레이블 분류, 레이블 멱집합, 제약조건 최적화

Key word : Protein subcellular multiple localization, Multi-label classification, Label power-set, Constrained optimization

접수일자 : 2014. 05. 24 심사완료일자 : 2014. 06. 13 게재확정일자 : 2014. 06. 27

* **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.2014.18.7.1749 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

단백질을 구성하는 아미노산 서열정보로부터 단백질이 존재하는 세포내 위치를 예측하려는 연구 분야에서, 최근에는 여러 세포내 위치에 동시에 존재하는 단백질의 생물학적 기능이 중요하므로, 이를 예측하려는 시도가 커지고 있다[1-9]. 단백질의 세포내 위치에 대한 정보가 단백질의 기능과 효과적인 약물의 발견에 중요한데, 이는 동물, 식물, 곰팡이는 세포소기관으로 구획된 서로 다른 생화학적 환경에서 단백질이 세포내 위치에 특이적인 기능을 수행하기 때문이다[10]. 단백질이 존재하는 다중 세포내 위치의 예측에는 기존의 단일레이블 분류 방법을 적용할 수 없고, 다중레이블 분류를 적용하여야 한다. 다중레이블 분류는 이미지, 비디오, 텍스트, 음악, 마케팅, 생물학 분야에서 하나의 입력 자료에 대해 여러 가지 분류에 동시에 속하는 문제를 모델링하기 위하여 연구되고 있다[11-13]. 이 방법 중에 알고리즘 적응은 단일 분류 알고리즘인 최근접-이웃 분류기, 신경망, 결정 트리, 지지 벡터 기계 등을 다중레이블 분류에 적합하게 변형한 방법이고, 문제 변환은 다중레이블 분류를 여러 개의 단일레이블 분류로 변환하여 단일레이블 분류를 사용하는 방법이고, 메타 학습은 알고리즘 적응이나 문제 변환을 여러 개 조합하여 분류하는 방법이다[11-13].

단백질의 세포내 위치 예측에 적용된 알고리즘 적응은 가우시안 과정 모델과 공분산 행렬로 레이블간의 연관성을 표현하는 방법[3]이 있고, 문제 변환은 세포내 위치의 모든 쌍들에 대한 분류기를 구성하여 분류결과를 투표를 통하여 최종 결과를 얻는 방법[2]과 각 단일레이블에 관련된 사례들과 관련되지 않은 모든 사례들로 학습하고 분류를 위해서 투표를 하는 방법[4, 6]이 있다. 또한, 여러 개의 이진 분류기를 체인으로 연결하고, k-번째 분류기는 $k-1$ 까지의 분류기의 예측결과를 이용하는 분류체인(classifier chain)을 앙상블로 사용하는 방법[5]과 각 사례의 다중레이블 자체를 하나의 레이블로 변환하는 레이블 멱집합방법을 변형하여, 레이블 부분집합을 무작위로 만들고, 사례와 관련된 레이블을 사례의 속성으로 사용하는 방법[8]이 있다. 메타학습 방법으로 최근접-이웃 분류기의 앙상블을 사용하는 방법[1, 7]이 적용되었다.

다중레이블 분류를 단백질의 세포내 위치 예측에 적용한 연구들을 살펴보면, 레이블들의 상호연관성을 모델링에 직접적으로 반영하는 분류체인 방법[14]과 멱집합 방법[15]이 성능이 높다[5, 8, 9]. 이러한 이유는 특정 생물학적 기능을 수행하는 단백질의 세포내 위치들의 관계는 독립적이지 않고 서로 관련되어 있다는 특징을 효과적으로 분류기에 반영하였기 때문이다. 본 논문에서는 기존의 레이블 멱집합 방법을 더욱 개선하기 위하여 다중레이블간의 중복 정보를 활용하여 각 다중레이블의 예측 정확도를 상호 강화한다. 이를 위하여 하나의 다중레이블을 다른 다중레이블들의 선형조합으로 나타낼 때의 조합가중치를 제약조건이 있는 최적화 방법을 사용하여 구하였다. 분류시에는 여러 다중레이블의 예측 확률을 조합가중치로 선형조합하여 최종적인 다중레이블을 예측하였다.

Ⅱ. 멱집합을 사용한 다중레이블 분류

본 논문에서는 단백질의 다중 세포내 위치에 효과적인 다중레이블 분류 방법 중에 하나인 레이블 멱집합(label power-set) 방법[11-13]을 개선한다. 분류기를 학습하는 자료에 나타나는 모든 레이블들의 집합을 $L = \{\lambda_1, \lambda_2, ..., \lambda_Q\}$이라 하면, 다중레이블 분류 방법은 각 사례 x_i에 대하여 이와 관련된 레이블들의 부분집합 $y_i \subseteq L$을 예측한다. 따라서 각 사례를 하나의 레이블로만 분류하는 단일레이블 분류에 비하여 알고리즘의 복잡성이 크게 증가한다. 레이블 멱집합 방법은 문제 변환 방법의 일종으로 각 사례와 관련된 다중레이블 y_i 자체를 새로운 단일 레이블로 처리한다. 이는 직접적으로 학습 자료에 나타나는 레이블간의 연관관계를 나타낼 수 있지만, 새로 만든 단일레이블의 수가 매우 커지는 단점이 있다. PS(pruned sets) 방법[15]은 이러한 단점을 극복하기 위하여 적은 빈도로 발생하는 다중 레이블은 제거하여 새로운 단일레이블로 만들지 않는다.

PS 방법을 구체적으로 살펴보면, 먼저 학습 자료에 나타나는 모든 상이한 다중레이블과 각 다중레이블을 가지는 사례의 수를 계산한다. 최종 선택되는 다중레이블의 수를 축소하기 위하여 파라미터 p를 사용하여 p보다 많은 사례의 수를 가지는 다중레이블과 이와 관련된 사례들로 새롭게 학습 자료를 구성한다. 학습 자료

에서 제외된 사례들은 각 사례가 가지는 다중레이블을 고려하여 새로운 자료로서 다시 학습 자료에 도입한다. 우선, 제외된 학습 자료의 다중레이블을 Y라 할 때, 학습 자료로 선택된 다중레이블 중에서 $y_i \subseteq Y$인 y_i들을 찾는다. 학습 자료에 재도입은 두 가지 전략을 사용하는데 , (A) 선택된 y_i들 중에서 포함하고 있는 레이블이 많고, 이런 다중레이블을 갖는 학습 자료가 많은 b개를 선택하는 방법과 (B) b보다 크기가 큰 y_i는 모두 선택하는 방법이 있다. 선택된 y_i를 학습에 제외되었던 사례의 레이블로 바꾸어 학습 자료에 재도입한다.

EPS(ensemble of PS)[15]는 PS방법 여러 개를 조합하여 분류기를 구성하는 방법이다. 이러한 앙상블 방법은 개개 분류기가 과도하게 적합(over-fitting)하는 것을 완화시킬 수 있고, 학습 자료에 나타나지 않는 새로운 레이블 부분집합을 예측할 수 있는 장점이 있다. EPS를 구성하기 위해서 학습 자료의 부분 집합 (63%가 사용됨)을 표본 추출하여 학습에 사용하여 PS 분류기를 구성하는 과정을 m번 반복한다. 분류 과정은 (1) 학습 과정에서 만들어진 m개의 PS 방법으로 분류 결과인 m개의 다중레이블에 포함된 각 단일레이블의 개수를 구하고, (2) 이러한 개수가 문턱치보다 큰 경우에만 해당 레이블을 분류 결과로 출력한다.

Ⅲ. 기초 다중레이블들의 조합을 사용한 다중레이블 분류

본 논문에서는 다중레이블 분류를 위하여 레이블 멱집합 방법을 이용한다. 즉, 분류의 첫 단계에서 학습에 충분한 개수의 자료와 관련된 다중레이블만을 학습과 분류에 사용하여, 각 다중레이블로 분류될 확률을 얻는다. 두 번째 단계에서 이러한 확률들을 조합하여 최종적으로 분류될 다중레이블을 결정한다.

다중레이블의 분류 확률을 조합하기 위하여 먼저 각 다중레이블을 다른 다중레이블들의 선형조합으로 나타내었고, 이러한 선형조합에서 나타나는 조합가중치를 사용하여 확률을 조합하였다. 학습 자료에 나타나는 모든 레이블의 집합을 $L = \{\lambda_1, \lambda_2, ..., \lambda_Q\}$, 파라미터 p보다 많은 사례와 관련된 다중레이블을 기초 다중레이블이라 하고, $y_1^b, y_2^b, \cdots y_R^b$로 나타내고, p이하의 사례와 관련된 다중레이블을 비기초 다중레이블 $y_1^{nb}, y_2^{nb}, \cdots y_S^{nb}$이라 하자. 다중레이블은 $(0110\cdots 0)^T$ 같은 형태로 0과 1로 구성되고, k번째 값은 레이블 λ_k가 다중레이블의 원소이면 1이고, 그렇지 않으면 0이다. 행렬 A는 k번째 열이 y_k^b인 $Q \times R$ 차원의 기초 다중레이블만으로 구성된 행렬이라 하자. 본 논문에서는 기초 다중레이블을 조합하여 여러 다중레이블을 표현하기 위한 방법으로 제약 조건을 가진 최적화를 사용하였다. 그림 1의 계산 과정으로 조합가중치 $x_i\,(i=1,2,\cdots,R+S)$를 구하였다.

for $i \leftarrow 0$ to R

$x_{ii} = 0$

minimize $\| A\,x_i - y_i^b \|_2$

subject to $\sum_{k=1}^{R} x_{i,k} = 1$

$\| x_i \|_\infty \leq l$

최적화 후 $x_{ii} = w$

for $i \leftarrow R+1$ to $R+S$

minimize $\| A\,x_i - y_{i-R}^{nb} \|_2$

subject to $\sum_{k=1}^{R} x_{i,k} = 1$

$\| x_i \|_\infty \leq l$

그림 1. 조합 가중치의 계산 단계
Fig. 1 Calculation process of combination weights

그림 1에서 기초 다중레이블들에 대한 조합가중치 $x_i\,(i=1,2,\cdots,R)$는 최적화 단계 전에는 i번째 차원 x_{ii}는 0으로 고정하여, y_i^b를 y_i^b 제외한 다른 기초 다중레이블의 선형조합으로 나타낼 수 있게 하였다. 최적화후에 x_{ii}를 w로 바꾸어서, 분류의 최종단계에서 y_i^b으로 분류될 확률을 구할 때에, 첫 단계에서 얻은 y_i^b에 대한 확률을 w만큼 반영할 수 있게 하였다. 이는 각 기초 다중레이블을 자신을 제외한 다른 기초 다중레이블만으로 조합하면 가장 중요한 y_i^b에 대한 확률을 제외하기 때문이다. 비기초 다중레이블은 기초 다중레이블 전체로

조합하므로 x_i의 일부를 바꾸어 주는 처리를 하지 않았다. 따라서 조합가중치 x_i의 합이 기초 다중레이블을 $1+w$이고, 비기초 다중레이블은 1이므로, 학습자료에 존재하는 자료의 수가 반영되어 기초 다중레이블이 더 높은 확률로 예측된다. 그림 1에서 최적화의 제약식에서 조합가중치의 합이 1이 되게 정규화하고, 조합가중치의 각 차원의 절대값이 파라미터 l보다 작게 제한하여 특정 기초 다중레이블의 확률이 지나치게 크게 선형조합에 기여하지 못하게 하였다.

본 논문에서 다중분류의 첫 번째 단계는 기초 다중레이블과 관련된 자료만 학습에 사용하여 다중분류기를 구성하고, 분류를 수행하여 각 평가 자료가 기초 다중레이블로 분류될 확률 $p_1, p_2, \cdots, p_R$을 계산한다. 두 번째 단계는 식 (1)을 사용하여 각 사례가 기초 다중레이블과 비기초 다중레이블로 분류될 확률을 다시 계산한다.

$$p_i^{trans} = \sum_{k=1}^{R} x_{i,k} \circ p_k,\ i = 1, 2, \cdots R+S \qquad (1)$$

이러한 $p_i^{trans}\ (i = 1, 2, \cdots, R+S)$ 중에서 가장 큰 값을 갖는 i에 해당하는 다중레이블이 최종 분류결과가 된다.

Ⅳ. 실험 및 결과

이번 장에서는 단백질의 세포내 위치 예측에 제안한 방법이 효과적인지를 검증한다. 먼저, 실험 자료, 실험 방법, 다중레이블 분류의 평가척도에 대해서 설명한 후에, 실험 결과에 대해서 알아본다.

실험에는 여러 논문[1-5]에서 사용된 인간 단백질 자료[1]를 사용하였고, 14개의 세포내 위치 (centriole, cytoplasm, cytoskeleton, endoplasmic reticulum, endosome, extracell, golgi apparatus, lysosome, microsome, mitochondrion, nucleus, peroxisome, plasma membrane, synapse)로 구성되어 있으며, 2,580개의 단백질은 하나의 세포내 위치, 480개는 두 개의 위치, 43개는 3개의 위치, 3개는 4개의 위치에 동시에 존재한다. 이 자료는 25% 이하의 단백질 서열 동일성을 가지고 있는 서열 유사성이 적은 자료이므로 단백질의 세포내 위치 예측이 어려운 자료이다.

단백질 자료를 분류를 위한 특징벡터로 변환하기 위해서, 실험 자료의 각 단백질 서열과 가장 유사한 단백질을 유전자 온톨로지를 가진 단백질 데이터베이스(http://www.ebi.ac.uk/GOA)에서 찾아, 그것의 유전자 온톨로지로 각 단백질을 나타내었다[1, 2, 4-7, 16]. 유전자 온톨로지는 분자적 기능, 생물학적 과정, 세포 요소의 관점에서 특징화한 용어로 유전자를 표현한 것으로, 각 단백질의 특징을 표현할 수 있다. 본 연구에서는 단백질의 특징을 보다 효과적으로 표현하기 위해서, 단백질의 세포내 위치에 따라 보다 판별력이 높게 나타내는 유전자 온톨로지를 가중하는 방법[16]을 사용하였고, 가장 유사한 두 개의 서열에서 나타나는 유전자 온톨로지의 빈도를 이용하였다[9].

다중레이블 분류의 평가는 단일레이블 분류처럼 예측된 레이블이 실제 레이블과 일치하는 것만을 판단하면 지나치게 엄격한 평가 척도가 되므로, 일부만 일치하는 경우도 고려하는 여러 관점의 평가 척도가 사용된다. 평가 척도는 사례기반(example-based)과 레이블기반(label-based)으로 나눌 수 있다[11-13]. 부록의 식 (S1)~(S6)의 사례기반 방법은 각 사례에 대해 실제 레이블과 예측된 레이블간의 차이를 평균하고, 식 (S7)~(S12)의 레이블기반 방법은 각 레이블에 대해 개별적으로 예측성능을 구하고 이를 평균한다. 본 논문에서는 여러 평가척도를 합한 부록의 *S-measure*도 사용하여 간략한 비교가 가능하게 하였다[13].

본 논문에서 제안한 방법을 CML(Combination of Multi-Labels)이라 하면, CML은 분류의 첫 단계에서 얻은 기초 다중레이블들의 확률을 그림 1의 알고리즘으로 계산한 조합가중치 x_i들을 이용하여 선형조합하여 최종적인 분류결과를 계산한다. 그림 1의 계산과정에서는 제약조건이 존재하는 최적화를 수행하여야 하는데, 이를 위하여 CVX[17]를 사용하였다. 분류실험에는 자료를 균등하게 5개로 나누어, 하나는 평가에 사용하고 나머지 4개는 학습 자료로 사용하는 방법을 5회 반복하는 5겹 교차검증(fivefold cross-validation)을 사용하였다. 이러한 5회의 실험에서 파라미터 $p = 5$, $l = 3$, $w = 1.1$을 사용할 때, 기초 다중레이블의 수는 33, 32, 32, 25, 33개이고, 비기초 다중레이블의 수는 48, 49, 46, 56, 41개였다. CML은 모든 기초 다중레이블과 비기초 다중레이블을 기초 다중레이블의 선형조합으로 표시하므

로, 총 395(33+32+32+25+33+48+49+46+56+41)개의 최적화가 수행되었는데, 10개의 $\| Ax_i - y_i^b \|_2$, $\| Ax_i - y_{i-R}^{nb} \|_2$가 1로 작은 오류를 보였고 나머지 385개는 0이므로, 대부분의 경우에 오류 없이 정확히 최적화되었다.

$$\begin{bmatrix}1\\1\\0\\0\\0\\0\\1\\.\\.\\.\\0\\0\end{bmatrix} = -0.06\begin{bmatrix}0\\0\\0\\0\\0\\0\\0\\.\\.\\.\\0\\1\end{bmatrix} - 0.20\begin{bmatrix}0\\0\\0\\0\\0\\0\\0\\.\\.\\.\\1\\0\end{bmatrix} + 0.06\begin{bmatrix}0\\0\\0\\0\\0\\0\\0\\.\\.\\.\\1\\1\end{bmatrix} + ... + 0.42\begin{bmatrix}1\\1\\0\\0\\0\\0\\0\\.\\.\\.\\0\\0\end{bmatrix}$$

$$y_{41}^{nb} = x_{R+41,1}y_1^b + x_{R+41,2}y_2^b + x_{R+41,3}y_3^b ... + x_{R+41,33}y_{33}^b$$

그림 2. 조합 가중치의 계산 예
Fig. 2 Example of combination weight calculation

그림 2는 5번째 실험에서 41번째 비기초 다중레이블을 33개의 기초 다중레이블의 조합으로 나타낸 것의 일부를 보여준다(표시 안된 세로 . . . 부분은 모두 0임). y_{41}^{nb}와 유사한 y_{33}^b의 계수는 0.42로 크고, y_1^b, y_2^b, y_3^b의 경우에는 -0.06, -0.20, 0.06으로 작으므로 다중레이블의 유사성에 따라 조합가중치가 할당되었음을 알 수 있다. 그러나 y_1^b보다 y_3^b이 더욱 y_{41}^{nb}과 상이하지만 같은 절대치의 조합가중치를 주는 문제점이 존재한다.

단백질 세포내 위치예측에 다중레이블 분류를 적용한 연구[5, 8, 9]에 따르면, 다중레이블을 구성하는 각 레이블들의 연관성을 반영하는 분류체인 방법[14]과 레이블 멱집합 방법[15]을 앙상블로 사용하는 ECC (Ensembel of Classifier Chain), EPS (Ensemble of Pruned Sets)가 성능이 높았다. 이는 특정 생물학적 기능을 수행하는 단백질은 서로 관련되어 있는 세포내 위치에 존재한다는 특징을 효과적으로 분류기에 반영하였기 때문이다. 따라서, 본 연구에서는 ECC, EPS를 비교실험에 사용하였고, Mulan 라이브러리[18]를 사용하였다. 또한, 이 방법들에서 내부적으로 사용되는 분류기로 트리기반 분류기 대신에 실험을 통해 성능이 더 높은 것으로 확인한 지지벡터기계를 사용하였고, 나머지는 기본 설정을 사용하였다. CML의 분류기는 지지벡터기계 라이브러리[19]를 사용하였는데, 가우시안 커널을 사용하였고, 최적의 파라미터 설정을 위해서 그리드 탐색을 통하여 γ와 C를 각각 0.01, 0.05, 0.1, 0.5, 1, 5와 0.1, 1, 10, 100에서 선택하였고, 표 1은 $\gamma = 0.1, C = 10$이고 그림 1에서 구한 조합가중치를 사용할 때의 결과이다.

표 1. ECC, EPS와 CML의 성능비교
Table. 1 Performance comparison of ECC, EPS, and CML

방법 / 평가 척도	ECC	EPS	CML
hamming_loss	0.0318	0.0310	**0.0282**
accuracy	0.7990	0.7992	**0.8175**
precision	0.8297	0.8529	**0.8703**
recall	**0.8494**	0.8273	0.8431
F_1	0.8394	0.8399	**0.8564**
subset_accuracy	0.6929	0.7205	**0.7424**
macro_precision	0.8112	0.8266	**0.8631**
macro_recall	**0.7164**	0.6831	0.7119
*macro_*F_1	0.7496	0.7331	**0.7636**
micro_precision	0.8108	0.8374	**0.8563**
micro_recall	**0.8149**	0.7862	0.8011
*micro_*F_1	0.8128	0.8110	**0.8278**
S-measure	9.6854	9.6861	**9.9253**

표 1에서 각 방법의 성능은 부록에서 설명한 평가척도를 사용하였다. 표 1에서 보듯이 본 논문에서 제안한 방법인 CML이 대부분의 척도에서 ECC나 EPS보다 높았고, 총괄적인 평가 척도인 *S-measure*로도 성능이 높다. 특히, ECC와 EPS의 예측 성능에서는 *precision*은 EPS가 높고, *recall*은 ECC가 높았다. 이와 같은 경향성이 *macro_precision, macro_recall*쌍과 *micro_precision, micro_recall*쌍에서도 나타난다. 하지만, EPS와 같은 레이블멱집합 계열의 방법인 CML도 이러한 경향성은 동일하지만, *precision*, *recall*과 관련된 모든 평가척도에서 향상됨을 볼 수 있다. 또한, 같은 레이블 멱집합 방법인 EPS와 비교하면 모든 평가척도에서 향상된 결과를 얻었다.

표 2. 단백질 세포내 위치예측의 성능 비교
Table. 2 Performance comparison of protein subcellular localization predictions

방법 / 평가 척도	논문 [1]	논문 [3]	논문 [4]	논문 [5]
hamming_loss	-	-	-	-
accuracy	-	-	-	0.7913
precision	-	-	-	0.8249
recall	0.519	0.643	-	0.8404
F_1	0.541	0.506	-	0.8191
subset_accuracy	0.294	0.202	0.45이하	-

표 2는 동일한 평가척도와 단백질 자료를 사용하는 다른 방법들의 성능이다. 논문[3]의 실험결과에 따르면, 여러 최근접-이웃 분류기를 조합하는 방법인 HummPLOC 2.0[1]과 알고리즘 적응 방법인 논문[3]은 표 1의 방법들보다 성능이 크게 저조하다. 이밖에 논문[4]는 문제 변환 방법의 일종인 BR(Binary Relevance)을 사용하였고, 논문 [5]에서는 ECC를 사용하였는데, 표 1의 실험 결과보다 예측정확도가 떨어짐을 알 수 있다.

표 2의 방법에서 단백질 자료를 특징벡터로 변환하여 유전자 온톨로지를 이용하는 방법이 본 논문과 동일하지 않으므로, 정확한 다중레이블 분류기의 성능 비교는 아니다. 하지만, 표1과 표2의 결과에서 보듯이 제안한 방법인 CML이 다른 방법들에 비하여 전반적으로 효과적임을 알 수 있다.

Ⅴ. 결 론

단백질의 다중 세포내 위치 예측에는 세포내 위치간의 연관관계를 학습 모델에 포함하는 방법이 성능이 높은 것으로 알려져 있다[5, 8, 9]. 즉, 이러한 연관관계를 자료의 속성에 추가하거나, 관련된 레이블 부분 집합 자체를 새로운 단일 레이블로 만드는 ECC, EPS 방법이 효과적이었다. 본 논문에서는 EPS 방법처럼 다중레이블 자체를 하나의 단일레이블로 구성하지만, 이러한 새로운 레이블들간에 나타나는 정보의 중복성을 활용하여 분류정확도를 향상시킨다. 기초 다중레이블을 조합하여 여러 다중레이블을 표시하기 위하여 제약 조건을 가진 최적화를 수행하였다. 최적화를 통하여 얻은 조합가중치를 사용하여 기초 다중레이블의 분류확률을 조합하여 최종적인 분류확률을 계산하였다.

제안한 CML 방법은 같은 레이블 멱집합 방법인 EPS에 비교하면 모든 평가척도에서 향상된 결과를 얻었고, ECC에 대해서는 12개 평가척도 중에서 9개가 높았다. 이러한 결과는 레이블 멱집합 방법처럼 레이블간의 연관성을 나타내면서, 동시에 다중레이블간의 중복된 정보를 분류에 활용하였기 때문이다.

본 논문은 다중레이블간의 정보 중복을 활용하여 다중레이블 분류를 시도하였다. 제안한 방법에 ECC나 EPS처럼 앙상블 방법을 추가적으로 적용하면 여러 다중레이블 분류의 응용분야에 적용될 수 있으리라 판단된다. 그러나, 여러 분야에 응용하기 위해서는 조합가중치의 계산에 보다 정교한 제약조건을 사용하여 다중레이블간의 중복도를 효과적으로 추출할 필요성이 있다. 향후에는 제안한 방법을 개선하여 동물, 식물, 곰팡이, 바이러스 등의 여러 영역의 세포내 위치 예측에 적용할 예정이다.

감사의 글

이 논문은 2014학년도 경성대학교 학술연구비지원에 의하여 연구되었음

부 록

다중레이블 평가척도는 다음과 같다[11-13]. 사례 x_i의 실제 다중레이블을 y_i, 예측된 레이블을 $h(x_i)$, $|A|$는 집합 A의 원소 개수, N은 사례의 총 개수, Q는 모든 레이블의 개수라 하자.
*hamming_loss*에서 △는 두 집합의 대칭차집합이다.

$$hamming\_loss(h) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{Q}|h(x_i)\triangle y_i| \quad \text{(S1)}$$

$$accuracy(h) = \frac{1}{N}\sum_{i=1}^{N}\frac{|h(x_i)\cap y_i|}{|h(x_i)\cup y_i|} \quad \text{(S2)}$$

$$precision(h) = \frac{1}{N}\sum_{i=1}^{N}\frac{|h(x_i)\cap y_i|}{|h(x_i)|} \quad \text{(S3)}$$

$$recall(h) = \frac{1}{N}\sum_{i=1}^{N}\frac{|h(x_i)\cap y_i|}{|y_i|} \quad \text{(S4)}$$

$$F_1 = \frac{1}{N}\sum_{i=1}^{N}\frac{2\times|h(x_i)\cap y_i|}{|h(x_i)|+|y_i|} \quad \text{(S5)}$$

*subset_accuracy*에서 $I(h(x_i)=y_i)$는 $h(x_i)$와 y_i가 같으면 1이고, 아니면 0이다.

$$subset\_accuracy(h) = \frac{1}{N}\sum_{i=1}^{N} I(h(x_i)=y_i) \quad \text{(S6)}$$

아래 식들에서 tp_j(true positive), fp_j(false positive), fn_j(false negative)은 레이블 λ_j와 이외의 레이블을 이진 분류하는 것에서 계산된다.

$$macro\_precision = \frac{1}{Q}\sum_{j=1}^{Q}\frac{tp_j}{tp_j+fp_j} \quad \text{(S7)}$$

$$macro\_recall = \frac{1}{Q}\sum_{j=1}^{Q}\frac{tp_j}{tp_j+fn_j} \quad \text{(S8)}$$

*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.

$$macro\_F_1 = \frac{1}{Q}\sum_{j=1}^{Q}\frac{2\times p_j\times r_j}{p_j+r_j} \quad \text{(S9)}$$

$$micro\_precision = \frac{\sum_{j=1}^{Q} tp_j}{\sum_{j=1}^{Q} tp_j + \sum_{j=1}^{Q} fp_j} \quad \text{(S10)}$$

$$micro\_recall = \frac{\sum_{j=1}^{Q} tp_j}{\sum_{j=1}^{Q} tp_j + \sum_{j=1}^{Q} fn_j} \quad \text{(S11)}$$

$$micro\_F_1 = \frac{2micro\_precision\times micro\_recall}{micro\_precision+micro\_recall} \quad \text{(S12)}$$

본 연구에서는 많은 평가척도를 요약하기 위하여 평가척도의 합을 사용하였는데, *hamming_loss*는 작은 값일수록 성능이 높으므로 1-*hamming_loss*를 더하였다. 즉, 각 방법의 성능을 요약하여 나타낼 경우에는 다음 척도를 사용하였다[13].

$$\begin{aligned} S\text{-}measure = {} & 1\text{-}hamming\_loss + accuracy + precision \\ & + recall + F_1\ score + subset\_accuracy \\ & + macro\_precision + macro\_recall + macro\_F_1 \\ & + micro\_precision + micro\_recall + micro\_F_1 \end{aligned} \quad \text{(S13)}$$

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] S.-M. Chi, “A performance comparison of multi-label classification methods for protein subcellular localization prediction,” *Journal of the Korea Institute of Information and Communication Engineering*, vol. 18, no. 4, pp. 992-999, Apr. 2014.

[10] H. Lodish, et al., Molecular cell biology, 6th ed. New York, NY:W. H. Freeman and Company, 2008.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] M. Grant and S. Boyd, CVX: Matlab software for disciplined convex programming, version 2.0 beta. http://cvxr.com/cvx, September 2013.

[18] G. Tsoumakas, E. Spyromitros-Xioufis, J. Vilcek, I. Vlahavas, "Mulan: a java library for multi-Label learning," *Journal of Machine Learning Research*, vol. 12, pp. 2411-2414. 2011.

[19] C.-C. Chang and C.-J. Lin, "LIBSVM : a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, Issue 3, pp. 27:1-27:27, 2011.

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