2014 22nd International Conference on Pattern Recognition

Dual Fuzzy Hypergraph Regularized Multi-label Learning for Protein Subcellular Location Prediction

Jing Chen[†,‡], Yuan Yan Tang[†,‡], C. L. Philip Chen[†], Yuewei Lin[§]
[†]*Faculty of Science and Technology, University of Macau, Macau, China*
[‡]*Chongqing University, Chongqing, China*
[§]*University of South Carolina, Columbia, USA*
{*chenjingmc, ywlin.cq*}*@gmail.com*, {*yytang, philipchen*}*@umac.mo*

***Abstract*—With the explosion of newly found proteins, it is necessary and urgent to develop automated computational methods for protein subcellular location prediction. In particular, the problem of predictor construction for multi-location proteins is challenging. Considering the main limitations of the existing methods, we propose a hierarchical multi-label learning model FHML for both single-location proteins and multi-location proteins. In this model, feature space is firstly decomposed onto a set of nonnegative bases under the nonnegative data factorization framework. The nonnegative bases act as latent feature concepts and the corresponding coefficients on these bases are views as the new feature representation on the latent feature concepts. The similar decomposition is later performed in label space, and then the latent label concepts are extracted. Using these latent concepts as hyperedges, we construct dual fuzzy hypergraphs to exploit the intrinsic high-order relations embedded in both feature space and label space. Finally, the subcellular location annotation information is propagated from the labeled proteins to the unlabeled proteins by performing dual fuzzy hypergraph Laplacian regularization. In this work, our proposed method is evaluated on eukaryotic protein benchmark dataset, and the experimental results have shown its effectiveness.**

I. Introduction

Proteins play an important role for organisms' physiological actions. In particular, the knowledge of proteins' functions will do great help for biology research and drug discovery. The number of newly found proteins is dramatically increasing in the last two decades. However, for a large part of these known proteins, we do not know their functions. Furthermore, this gap is becoming sharply wide with the explosion of newly found proteins. To analyze a protein's functions, the determination of its subcellular locations is a greatly helpful step. This is because proteins perform their appropriate functions only when they are located in the correct subcellular locations [1]. The traditional way to determine subcellular location of proteins is performed by the three experimental approaches: cell fractionation, electron microscopy and fluorescence microscopy. However, these biochemical tests are time-consuming, costly and subjective [2]. To tackle this problem, it is extraordinarily desirable to develop automated methods to predict subcellular locations of proteins accurately.

Many efforts have been paid for protein subcellular location prediction in the past few years. These researches mainly focus on how to effectively represent a protein and how to construct prediction models. For feature extraction, most researches extract three types of feature representation: the amino acid composition, the sequence order and the physical chemistry character. The first two feature types and their combination are more commonly used. In this work, we adopt the direct combination of the first two types of features to construct an original feature space. So far, many computational methods, such as K-nearest neighbor [3], support vector machines [4], neural networks [5], and hidden Markov models [6] have been applied for protein subcellular localization prediction. However, the main limitations of these existing intelligent techniques could be summarized as the following three points:

(1) The traditional methods assume that each protein resides at only one subcellular location, and then they transform the subcellular location prediction task into a single-label classification problem. However, we notice that some proteins may simultaneously exist in, or move between two or more different subcellular locations. It is necessary to take multi-location proteins into account when constructing subcellular location predictors.
(2) Most of the methods which can deal with multi-location proteins directly transform the multi-label problem into multiple binary classification sub-problems for each class label and finally integrate multiple independent outputs. Obviously, this scheme ignores the label correlation. In fact, each subcellular location is not isolated physiologically, and furthermore, they are correlated with each other. We need to consider intra-label similarity and inter-label diversity, which involves both feature space and label space.
(3) The prediction models are constructed based on the direct relation from extracted features to labels. In other words, the traditional methods usually construct the simple two-layer models. In fact, the hierarchical multi-layer prediction models have been evaluated effectively in many other patter recognition fields. It would be promising to consider a hierarchical prediction model for protein subcellular location prediction.

In this work, we deal with the task of subcellular location prediction of multi-location proteins. To this goal, we construct a three-layer hierarchical model as Fig. 1, which consists of feature layer, latent layer and label layer. The middle layer acts as the link between the feature layer and the label layer. The extracted latent concepts perform as the dictionary items which are commonly used in document analysis. Two normal graphs are constructed within the feature layer and the label layer, respectively. In the feature layer, the original features are decomposed onto the latent concepts. A fuzzy hypergraph is used to regularize the consistency between the original features

1051-4651/14 $31.00 © 2014 IEEE
DOI 10.1109/ICPR.2014.97

and the intermediate latent codes. The other hypergraph is constructed to regularize the annotation lists and the latent codes. Above all, the annotation information is propagated from the labeled proteins to the unlabeled proteins by the dual fuzzy hypergraph regularized multi-label learning.

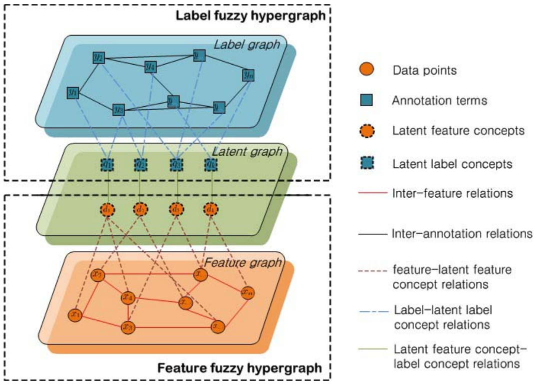

Fig. 1. The diagram of the proposed three-layer model

II. The Proposed FHML method

A. Problem Formulation

Given a protein database $\mathcal{D} = \{\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_n\}$ of n protein sequences, its corresponding annotation vocabulary $\mathcal{V} = \{\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_k\}$ of k subcellular location labels in a multi-label protein subcellular location prediction task. Each protein $\mathcal{I}_i$ is represented by its original feature vector $x_i \in \mathbb{R}^m$ for $i = 1, 2, \ldots, n$. Then we have the protein dataset $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{m\times n}$. Among these proteins in the database, l proteins are annotated with one or more subcellular location labels of the vocabulary $\mathcal{V}$, and other u proteins are not annotated. Here, $l + u = n$. Without loss of generality, we assume that the first l proteins are labeled in advance by the label indicator matrix $\widetilde{Y}_L = [\widetilde{y}_1, \widetilde{y}_2, \ldots, \widetilde{y}_l] \in \{0, 1\}^{k\times l}$. Each $\widetilde{y}_i$ is a multi-dimensional vector. The value of 1 indicates that the protein $\mathcal{I}_i$ resides at the corresponding subcellular location and the value of 0 indicates $\mathcal{I}_i$ has no probability to exist in that location. We denote the output real-valued label score matrix as $Y = [Y_L\ Y_U]$, and the final 0-1 label matrix as $Y^* = [Y^*_L\ Y^*_U]$

B. Latent Concept Learning

To construct the hierarchical structure, we need to extract latent concepts first. Here, we use a similar technique of dictionary learning originally applied in document analysis. At first, we decompose the original multi-dimensional feature space onto a set of nonnegative bases under the nonnegative data factorization framework. The set of nonnegative bases can be viewed as the learned latent feature concepts. Meanwhile, the corresponding new feature representation on these latent bases is also learned. As the case in face recognition, we know that the extracted latent feature concepts are relevant to the parts of the original holistic features. Thus, the obtained new feature representation is localized on the latent feature concepts. With the help of the obtained latent feature concepts, the intrinsic relations embedded in feature space, i.e., feature correlation, can be represented and exploited. Using the latent concepts as the hyperedges, we construct a hypergraph in feature space to capture the embedded intrinsic high-order relations. In addition, a similar trick is applied in label space to exploit its intrinsic relations, i.e., label correlations. And then a similar hypergraph can be constructed in label space based on the extracted latent label concepts. Finally, the three-layer hierarchical model is constructed to deal with the multi-label subcellular location prediction problem.

So, for the protein dataset $X = [x_1, x_2, \ldots, x_n]$, formally, we reconstruct it by using the linear combination of latent feature concepts as $X = DZ$, where $D = [d_1, d_2, \ldots, d_r] \in \mathbb{R}^{m\times r}$ is the latent feature concept basis matrix and $Z = [z_1, z_2, \ldots, z_n] \in \mathbb{R}^{r\times n}$ is the new feature representation over the latent basis. In this work, the basis matrix D and the coefficient matrix Z are both constrained as nonnegative matrices. Each d_i acts as a latent feature concept and the nonnegative column vector z_j is used as the weight coefficient vector of the jth protein belongs to each latent feature concept. D and Z could be obtained under the dictionary learning framework as follows,

$$\begin{aligned} \min_{D,Z}\ & ||X - DZ||_F^2 \\ s.t.\ \ & D, Z \geq 0, \\ & \mathbf{1}^T Z = \mathbf{1}^T \end{aligned} \tag{1}$$

The constraint $\mathbf{1}^T Z = \mathbf{1}^T$ enforces each column z_j to be a normalized weight vector. Here, we call the new feature representation Z as latent codes.

In addition, we also decompose the annotation vectors onto the latent label concepts. For the protein dataset X, denote the corresponding subcellular location annotation matrix as Y. For Y, we define the prediction model from the latent codes to the annotation vectors as follows:

$$Y = QZ \tag{2}$$

where $Q \in \mathbb{R}^{k\times r}$ and $Q \geq 0$. Thus, the column vectors $Q = [q_1, q_2, \ldots, q_r]$ are regarded as the latent label concepts, and Q is used as the codebook in label space. Here, we assume $Q = PD$, where $P \in \mathbb{R}^{k\times m}$ and $P \geq 0$. Then, P is the relation matrix which shifts the latent components from feature space to label space.

Herein, the Y can be predicted by $Y = PDZ$. In addition, the predicted labels of labeled data should be enforced to be consistent with original labels. Mathematically, we should first optimize the following objective function:

$$\begin{aligned} \min_{P,D,Y}\ & \lambda_1 ||Y - PDZ||_F^2 + \lambda_2 ||Y_L - \tilde{Y}_L||_F^2 \\ s.t.\ \ & P, D, Y \geq 0 \\ & \mathbf{1}^T Y = \mathbf{1}^T \end{aligned} \tag{3}$$

The last constraint $\mathbf{1}^T Y = \mathbf{1}^T$ normalizes each annotation vector to avoid the scaling problem. Moreover, this normalization constraint ensures that we can substitute the standard inner for the cosine similarity.

C. Dual Fuzzy Hypergraph Laplacian Regularization

We can view the above decomposition as this way: the sample i is related to the latent feature concept j with the non-zero weight z_{ji}, and the sample i is unrelated to the latent feature concept j when the z_{ji} is zero. Herein, each protein sequence feature vector could be reconstructed by some latent feature concepts; on the other hand, each latent feature concept covers a subset of samples. The z_{ji} acts as a membership degree of the protein i to the latent feature concept j. The decomposition of label space could be explained in the similar way. Naturally, the latent feature concept i could viewed to be belonged to itself group completely. So we define its weight z_i^D as a column vector with 1 in i-th entry and 0 elsewhere. The the latent codes of the latent feature concepts can be define as $Z_D = \{z_i^D\}_{r \times r}$, and the latent label concepts share the same codes Z_D.

This viewpoint motivates us to employ a hypergraph to represent these relations, in which a hyperedge covers a subset of vertices. We construct fuzzy hypergraphs in feature space and label space, respectively. Each latent concept corresponds to a hyperedge, and the instances (i.e., feature vectors in feature space, annotation vectors in label space) connected to the latent concept belong to its corresponding hyperedge. Here, the instance i is connected to the latent concept j if its weight z_{ji} is non-zero. In feature space, we construct a fuzzy hypergraph $G_F = (V_F, E_F, W_F)$, where V_F is the set of vertices associated to protein features, E_F is the set of hyperedges associated to latent feature concepts and W is the fuzzy degrees of vertices to hyperedges. Here, let $W_F = Z$. In this way, all the protein samples are organized by using latent feature concepts on the fuzzy hypergraph. In label space, we also construct a fuzzy hypergraph $G_S = (V_S, E_S, W_S)$, where V_S is the set of vertices associated to protein annotation vectors, E_S is the set of hyperedges associated to latent label concepts and W_S is the fuzzy degrees of vertices to hyperedges. Here, let $W_S = Z$. In this way, all the protein annotations are also organized by the fuzzy hypergraph.

To capture the embedded intrinsic correlation, we perform a novel regularization on this fuzzy hypergraph. The regularization is based on the assumption that the proteins in the same feature hyperedge have similar latent codes and the similar latent codes yield similar annotations. This type of intrinsic relations could be preserved by performing hypergraph Laplacian regularization.

Following the star expansion algorithm, we transform the initial fuzzy hypergraphs G_F and G_S into the two bipartite graphs $\hat{G}_F = (\hat{V}_F, \hat{E}_F)$ and $\hat{G}_S = (\hat{V}_S, \hat{E}_S)$ with the adjacency matrices as $\hat{W}_F$ and $\hat{W}_S$ by introducing a new vertex for each hyperedge. Then we could transform the dual fuzzy hypergraph Laplacian regularization into the traditional graph Laplacian regularization. The vertex set $\hat{V}_F$ consists of the initial vertices corresponding to protein feature vectors and the new vertices corresponding to latent feature concepts, i.e. $\hat{V}_F = \hat{X} = [X, D]$. The weight of each edge in G_F is inherited from the fuzzy membership degree of each vertex in the hypergraph G_F, i.e., the weight of each edge is defined as the inner of the two joint vertices. The similarity matrix $\hat{W}_F$, whose entry $\hat{W}_{ij} = \hat{x}_i^T \hat{x}_j$ measures the similarity between a vertex pair $(\hat{x}_i, \hat{x}_j)$, i.e.,

$$\hat{W}_F = \hat{X}^T \hat{X} = \left[\begin{array}{cc} X^T X & X^T D \\ D^T X & D^T D \end{array} \right] \tag{4}$$

We define the degree matrix as $\hat{D}_F$, which is a diagonal matrix with $\hat{D}_{ii}^F = \sum_j \hat{W}_{ij}^F$.

In the other hand, the vertex set of the bipartite graph $\hat{G}_S$ in label space is $\hat{Y} = [Y, Q] = [y_1, y_2, \ldots, y_n, q_1, q_2, \ldots, q_r]$. The pairwise similarity is measured by the inner $\hat{y}_i^T \hat{y}_j$ for the pair $(\hat{y}_i, \hat{y}_j)$. Thus, these pairwise similarity measures constitute the following similarity matrix $\hat{W}^S$ with $\hat{y}_i^T \hat{y}_j$ to be the entry W_{ij}^S.

$$\hat{W}_S = \hat{Y}^T \hat{Y} = \left[\begin{array}{cc} Y^T Y & Y^T Q \\ Q^T Y & Q^T Q \end{array} \right] \tag{5}$$

Thus, we can extend the optimization problem (1) by adding a graph Laplacian regularization term as follows:

$$\begin{array}{ll} \min\limits_{D,Z} & ||X - DZ||_F^2 + \lambda_3 tr(\hat{Z}\hat{L}_F\hat{Z}^T) \\ s.t. & D, Z \geq 0 \\ & \mathbf{1}^T Z = \mathbf{1}^T \end{array} \tag{6}$$

Define the Laplacian matrix $\hat{L}_F = \hat{D}_F - \hat{W}_F$.

In label space, the optimization problem (3) is extended by adding a graph Laplacian regularization term as follows:

$$\begin{array}{ll} \min\limits_{P,D,Y} & \lambda_1 ||Y - PDZ||_F^2 + \lambda_2 ||Y_L - \tilde{Y}_L||_F^2 \\ & +\lambda_4 tr(\hat{Z}\hat{L}_S\hat{Z}^T) \\ s.t. & P, D, Y \geq 0 \\ & \mathbf{1}^T Y = \mathbf{1}^T \end{array} \tag{7}$$

where the Laplacian matrix $\hat{L}_S = \hat{D}_S - \hat{W}_S$

D. Multi-label learning formulation

By integrating all of the above two folds, the semisupervised multilabel learning problem for protein subcellular location prediction is formulated as a dual fuzzy hypergraph regularized nonnegative data factorization problem in the following form:

$$\begin{array}{ll} \min\limits_{P,D,Z,Y} & ||X - DZ||_F^2 + \lambda_1 ||Y - PDZ||_F^2 \\ & +\lambda_2 ||Y_L - \tilde{Y}_L||_F^2 + \lambda_3 \mathrm{Tr}(\hat{Z}\hat{L}_F\hat{Z}^T) \\ & +\lambda_4 \mathrm{Tr}(\hat{Z}\hat{L}_S\hat{Z}^T) \\ s.t. & P, D, Z, Y \geq 0 \\ & \mathbf{1}^T Z = \mathbf{1}^T, \quad \mathbf{1}^T Y = \mathbf{1}^T \end{array} \tag{8}$$

The parameters λ_1, λ_2, λ_3 and λ_4 are used to balance the contribution of each objective terms to the solution.

E. Solution

The cost function is not convex with respect to D, Z, P and Y together. Thus, it is not realistic to find the global minima. However, the cost function is strictly convex with respect to each matrix variable block respectively. So, here we adopt the common method which is to iteratively optimize the objective function by alternatively minimizing over one matrix variable while keeping the other three blocks fixed. Together with the strict convexity of the objective function, we can deduce that each subproblem has a unique minimum. Here, for the nonnegative constraint, we employ the multiplicative iterative algorithm used for NMF. For the sum-to-one constraint, an effective technique in [8] is employed here. We use the matrices $\bar{X}$ and $\bar{D}$ to take the place of X and D as inputs, which are defined as

$$\bar{X} = \left[\begin{array}{c} X \\ \delta 1^T \end{array} \right], \qquad \bar{D} = \left[\begin{array}{c} D \\ \delta 1^T \end{array} \right] \tag{9}$$

where δ adjusts the effect of the sum-to-one constraint. Similarly, We use the following equation to replace the original decomposition assumption $Y = PDZ$.

$$\left[\begin{array}{c} I \\ \delta 1^T \end{array} \right] Y = \left[\begin{array}{c} PDZ \\ \delta 1^T \end{array} \right] \tag{10}$$

Here, we denote $\bar{I} = \left[\begin{array}{c} I \\ \delta 1^T \end{array} \right]$ and $\bar{S} = \left[\begin{array}{c} PDZ \\ \delta 1^T \end{array} \right]$. In this work, $\delta = 20$ is selected.

The Eqn. (9) and (10) are substituted into the problem (8). We solve the new optimization problem by optimizing Z, D, Y and P alternately with a set of multiplicative updating rule, which guarantee the nonnegativity of the solution. Finally, we can obtain the update rules for all variable matrix as follows:

$$Z_{ij}^{t+1} = Z_{ij}^{t} \times \frac{(\bar{D}^T \bar{X} + A_1^Z + A_2^Z + A_3^Z)_{ij}}{(\bar{D}^T \bar{D} Z + A_4^Z)_{ij}} \tag{11}$$

$$D_{ij}^{t+1} = D_{ij}^{t} \times \frac{(A^D)_{ij}}{(D(A^D)^T A)_{ij}} \tag{12}$$

$$Y_{ij}^{t+1} = Y_{ij}^{t} \times \frac{(A_1^Y + A_2^Y)_{ij}}{(A_3^Y)_{ij}} \tag{13}$$

$$P_{ij}^{t+1} = P_{ij}^{t} \times \frac{(A^P)_{ij}}{(P(A^P)^T P)_{ij}} \tag{14}$$

where

$$\begin{aligned}
A_1^Z &= \lambda_1 D^T P^T Y \\
A_2^Z &= \lambda_3 (Z X^T X + Z_D D^T X) \\
A_3^Z &= \lambda_4 (Z Y^T Y + Z_D D^T P^T Y) \\
A_4^Z &= \lambda_1 D^T P^T P D Z \\
A^D &= X Z^T + \lambda_1 P^T Y Z^T + \lambda_3 X Z^T Z_D + \lambda_4 P^T Y Z^T Z_D \\
A_1^Y &= \lambda_1 \bar{I}^T \bar{S} + \lambda_2 \tilde{Y}_L \bar{I}^T \\
A_2^Y &= \lambda_4 (Y Z^T Z + P D Z_D^T Z) \\
A_3^Y &= \lambda_1 \bar{I}^T \bar{I} Y + \lambda_2 Y \left[\begin{array}{cc} I & 0 \\ 0 & 0 \end{array} \right] \\
A^P &= \lambda_1 Y Z^T D^T + \lambda_4 Y Z^T Z_D D^T
\end{aligned}$$

From the above alternative update procedure, we obtain the real-valued label score matrix Y. Then we need a cut-off threshold to transform the score matrix into the 0-1 matrix Y^*. Thus, the final predicted label subset for each protein is obtained. In this work, we employ the S-Cut technique to optimize the threshold based on the Hamming distance between the actual label matrix $\tilde{Y}_L$ and the predicted label matrix Y_L^* of the labeled proteins. The whole proposed algorithm can be summarized as following steps:

Algorithm — FHML

```
Input: protein dataset $X$, annotated label matrix $\tilde{Y}_L$
Initialization: Randomly choose $D^0$, $Z^0$, $P^0$ and $Y^0$ as nonnegative matrices.
A. For $t = 0, 1, 2, \ldots, T_{\text{max}}$, do
    1) For given $D = D^t, P = P^t, Y = Y^t$, update the latent codes $Z$ as Eqn. (11);
    2) For given $Z = Z^t, P = P^t, Y = Y^t$, update the latent feature concept basis matrix $D$ as Eqn. (12);
    3) For given $Z = Z^t, D = D^t, P = P^t$,, update the label ranking matrix $Y$ as Eqn. (13);
    4) For given $Z = Z^t, D = D^t, Y = Y^t$, update the relation matrix $P$ as Eqn. (14);
    5) If $\| Z^{t+1} - Z^t \| < \epsilon, \| D^{t+1} - D^t \| < \epsilon, \| Y^{t+1} - Y^t \| < \epsilon, and \| P^{t+1} - P^t \| < \epsilon$ ($\epsilon$ is set as $10^{-3}$ in this work), then break.
  end
B. Optimize the threshold $\theta$ and perform cut-off on $Y_U$
Output: The predicted label matrix $Y_U^*$
```

III. Experiments

The dataset $\mathbb{S}$ of Euk-mPLoc from the well-known package Cell-Ploc 2.0 [9] with experimentally determined protein subcellular localization is used as the benchmark dataset for the current study. The dataset is built for eukaryotic proteins specially. It includes 7,766 different eukaryotic protein sequences, covering 22 corresponding subcellular locations. In the dataset, 6,687 belong to one subcellular location, 1,029 to two locations, 48 to three locations, and 2 to four locations. Each protein in the dataset has less than 25% sequence similarity to any other in the same subcellular location group, which makes it more reliable to compare our proposed method with others. The dataset is obtained from the Online Supporting Information S1 in [9].

We perform the 3-fold, 5-fold and 10-fold cross validation. Each n-fold cross validation is repeated for ten times, where all the proteins are randomly divided into n mutually exclusive parts with approximately equal size and approximately equal class distribution. The averaged results are reported in this work.

As the case study of [10] suggested, we use the two types of performance measures, i.e., example-based and label-based measures. The example-based measures are *F-measure* and *Accuracy*, and the label-based measures are *Precision* and *Recall*. The definitions of these four measures follow those presented in [10], which are different from the single-label measures.

For a given protein sequence, the two types of features, i.e., PseAAC and PSSM-ACT, are extracted and concatenated serially as its original high-dimension feature vectors. These two types of features involve not only the amino acid composition information but also the protein sequence order information and sequence evolution information, which have been demonstrated effective in many bioinformatics fields. The detailed description of these two types of features can be found

TABLE I. PERFORMANCE COMPARISON OF THE DIFFERENT METHODS IN THE THREE CORSS-VALIDATIONS

	Method	*Precision*	*Recall*	*F-Measure*	*Accuracy*
3-fold	mKNN	0.5227±0.0013	0.4937±0.0011	0.7561±0.0019	0.7596±0.0020
	mSVM	0.5190±0.0126	0.4965±0.0083	0.7424±0.0104	0.7663±0.0112
	FHML	**0.5285±0.0011**	**0.5082±0.0017**	**0.7709±0.0013**	**0.7748±0.0011**
5-fold	mKNN	0.5226±0.0016	0.4965±0.0013	0.7688±0.0022	0.7619±0.0017
	mSVM	0.5354±0.0074	0.5097±0.0105	0.7719±0.0128	0.7574±0.0124
	FHML	**0.5563±0.0012**	**0.5113±0.0007**	**0.7934±0.0010**	**0.7635±0.0013**
10-fold	mKNN	0.5460±0.0009	0.5035±0.0010	0.7968±0.0011	0.7763±0.0013
	mSVM	0.5395±0.0121	0.5081±0.0138	0.7834±0.0114	0.7798±0.0125
	FHML	**0.5642±0.0010**	**0.5279±0.0009**	**0.8160±0.0007**	**0.7913±0.0010**

in [11]. In this work, the PseAAC and PSSM-ACT feature vectors are both in 140-dimension.

From the existing publications, the commonly-cited methods able to deal with multi-location proteins in subcellular location prediction are multi-label KNN (mKNN) in iLoc-Euk [12] and multi-label SVM (mSVM) in [13]. These two methods are constructed as the same in the references, while their inputs are the PseAAC and PSSM-ACT features, which is the same as our method, for a reliable comparison. Their parameter selection is the same as the original references. For our FHML method, the parameters λ_i's and the number of latent concepts r are optimized by using 3-fold cross validation on the labeled set. The λ_i's are tuned from 10^{-5} to 10^{-3}. r is tuned from 50 to 500. We uniformly select twenty values for each parameter range and select the highest one to finetune. Here, $\lambda_1 = 0.00047$, $\lambda_2 = 0.00182$, $\lambda_3 = 0.00131$, $\lambda_4 = 0.00064$ and $r = 120$.

Table I reports the experimental results of the three compared methods on the four performance measures. From this result, we find that mKNN and mSVM perform similarly, and the proposed method outperforms the compared predictors on the four measures in the three types of cross-validations. This result evaluates the effectiveness of label correlation exploring. In particular, the superiority of FHML is more significant when the model receives more training samples. This fact would suggest us that the more samples provide the more accurate relations embedded in feature space and label space.

IV. Conclusion

We construct a three-layer hierarchical multi-label learning model with dual fuzzy hypergraph regularization. We explore the intrinsic relations not only in feature space but also in label space. We conduct comparable experiments on the eukaryotic protein dataset, and the experimental results have shown that our work outperforms state-of-the-art protein subcellular location prediction methods in terms of the four measures. Further work would be focusing on the imbalance of class distribution which frequently occurs in biochemical datasets.

Acknowledgment

This work is supported by the National Natural Science Foundations of China (60873092, 90820306 and 61100114), the Natural Science Foundation of Chongqing (CSTC, 2009AB5002 and CSTC, 2010BB2217), and the Fundamental Research Funds for the Central Universities (No.CDJXS10182216).

V. References

References

[1] F. Eisenhaber, P. Bork, Wanted: subcellular localization of proteins based on sequence. *Trends. Cell Biol.*, 8:169-170, 1998.

[2] R. F. Murphy, M. V. Boland, M. Velliste, Towards a systematics for protein subcellular location: quantitative description of protein localization patterns and automated analysis of fluorescence microscope images. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 8:251C259, 2000.

[3] T. Wang, J. Yang, Predicting subcellular localization of gram-negative bacte rial proteins by linear dimensionality reduction method. *Protein Pept. Lett.*, 17: 32-37, 2010.

[4] B. Liao, J. B. Jiang, Q. G. Zeng, W. Zhu, Predicting Apoptosis Protein Subcellular Location with PseAAC by Incorporating Tripeptide Composition. *Protein Pept. Lett.*, 18: 1086-1092, 2011.

[5] L. Zou, Z. Wang, J. Huang, Prediction of subcellular localization of eukaryotic proteins using position-specific profiles and neural network with weighted inputs. *J. Genet. Genomics* 34: 1080-1087, 2007.

[6] T. Lin, R. Murphy, Z. Bar-Joseph, Discriminative motif finding for predicting protein subcellular localization. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 8: 441-451, 2011.

[7] A. Edelman, T. Arias, and S. T. Smith, The geometry of algorithm with orthogonality constraints, *SIAM J. Matrix Anal. Appl.*, 20(2):303-353, 1998.

[8] X. Lu, H. Wu, and Y. Yuan, Double constrained NMF for hyperspectral unmixing, *IEEE Trans. Geosci. Remote S.*, 2013.

[9] Kuo-Chen Chou, and Hong-Bin Shen, Cell-PLoc 2.0: an improved package of web-servers for predicting subcellular localization of proteins in various organisms, *Nat. Sci.*, 2: 1090-1103, 2010.

[10] S. Nowak, H. Lukashevich, P. Dunker, and S. Ruger, Performance measures for multilabel evaluation: A case study in the area of image classification, *Proc. Int. Conf. Multimedia Inf. Retrieval*, 35-44, 2010.

[11] D. Yu, X. Wu, H. Shen, J. Yang, Z. Tang, Y. Qi, and J. Yang, Enhancing membrane protein subcellular localization prediction by parallel fusion of multi-view features, *IEEE Trans. Nanobioscience*, 11(4):375-385, 2012.

[12] K.-C. Chou, Z.-C. Wu, and X. Xiao, iLoc-Euk: a Multi-Label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins, *PLoS ONE*, 6(3):e18258, 2011.

[13] L. Zhu, J. Yang and H. Shen, Multilabel learing for prediction of human protein subcellular localizations, *Protein J.*, 28: 384-390, 2009.