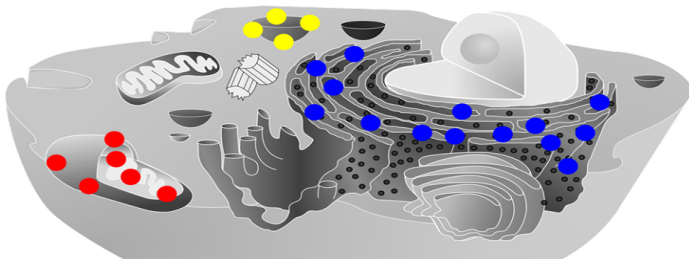


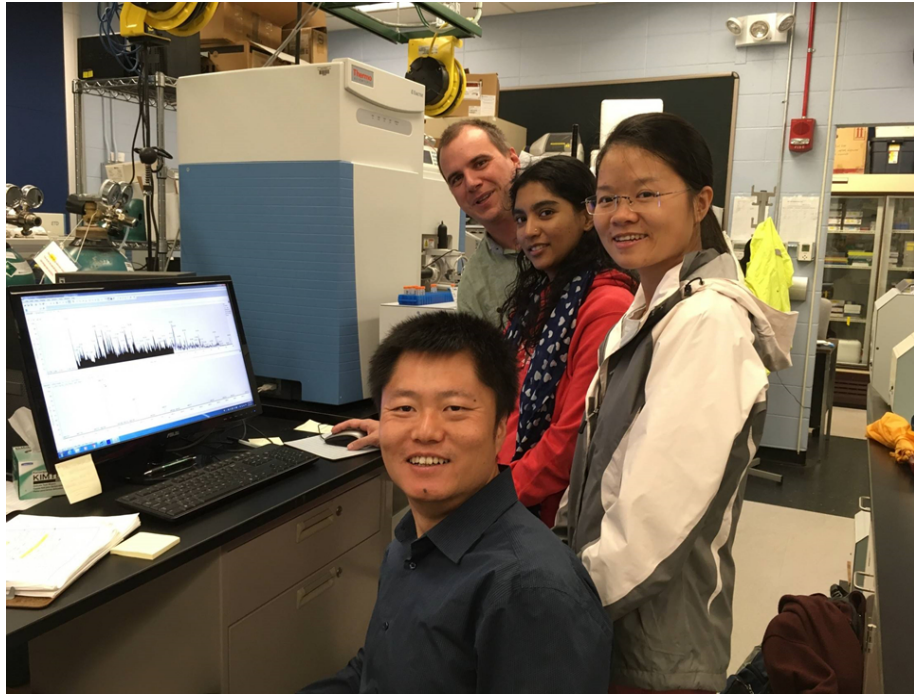
Organelle Proteomics Classification using Graph-Based Semi-Supervised Learning

Nicholas Carruthers; Afsana Kabir; Tingli Hu

Wayne State University

12-7-2017





Outline

1 Problems of the data set

2 Semi-supervised learning

- Basic idea
- Graph-based semi-supervised learning

3 Experiments

- Create negatives with KNN graph type models
- Train and test the model

Problems of the data set

The data set has 8 numeric variables and 6 independent categories.

- Insufficient labeled samples

Table: Number of labeled samples of each category

Category	ER	Golgi	Mito	Nucleus	Perox	PM
l_{2015}	368	269	520	1297	74	907
l_{2017}	920	757	1266	3302	117	1947
R_{2015}	0.038	0.028	0.054	0.134	0.0076	0.094
R_{2017}	0.095	0.078	0.131	0.341	0.012	0.204
m	9683					

- Absence of negative labeled samples

Semi-supervised learning

Basic idea



Figure: Classification with only labeled samples



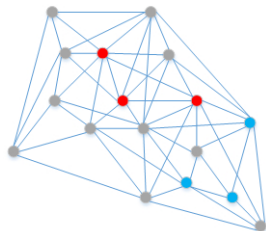
Figure: Classification with a complete set of samples

Semi-supervised learning

Graph-based semi-supervised learning

Affinity matrix

$$\mathbf{w}_{i,j} = \begin{cases} \exp\left(-\frac{\|x_i - x_j\|_2^2}{2\sigma^2}\right), & \text{if } i \neq j, \\ 0, & \text{otherwise.} \end{cases}$$



Semi-supervised learning

Graph-based semi-supervised learning

Energy function

$$\begin{aligned} E(f) &= \frac{1}{2} \sum_{i,j} w_{i,j} (f(x_i) - f(x_j))^2 \\ &= \frac{1}{2} \left(\sum_{i=1}^m d_i f^2(x_i) + \sum_{j=1}^m d_j f^2(x_j) - 2 \sum_{i=1}^m \sum_{j=1}^m w_{ij} f(x_i) f(x_j) \right) \\ &= \sum_{i=1}^m d_i f^2(x_i) - \sum_{i=1}^m \sum_{j=1}^m w_{i,j} f(x_i) f(x_j) \\ &= \mathbf{f}^T (\mathbf{D} - \mathbf{W}) \mathbf{f} \end{aligned}$$

where $w_{i,j}$ is an entry of $\mathbf{W}$, $\mathbf{f} = (\mathbf{f}_l^T \mathbf{f}_u^T)^T$, $\mathbf{f}_l = \{f(x_1), \dots, f(x_l)\}$ and $\mathbf{f}_u = \{f(x_{l+1}), \dots, f(x_{l+u})\}$. The real-valued function $f : V \rightarrow \mathbb{R}$ is learnt from the complete samples.

Semi-supervised learning

Graph-based semi-supervised learning

By minimizing the energy function $E(f)$, the propagation function can be obtained

$$\mathbf{f}_u = (\mathbf{I} - \mathbf{P}_{uu})^{-1} \mathbf{P}_{ul} \mathbf{f}_l$$

“Label Propagation”

where $\mathbf{P}_{uu}$ and $\mathbf{P}_{ul}$ are constant matrices derived from $\mathbf{D}$ and $\mathbf{W}$

Experiments

Create negatives with KNN graph type models

Take the “ER” category for example :

- Use semi-supervised learning method on data2015 with KNN graph type with $k=20$
- Label all samples that have f values less or equal than 0.3 “FALSE” (0.3 is selected randomly, when 0.3 is chosen the number of negative samples is close to the number of positive samples, 235 negative labels are newly labeled)
- With the same semi-supervised learning method on data2017, $k=20$
- Label all samples that have f values less or equal than 0.5 “FALSE” (746 samples are newly labeled)

Experiments

Train and test the model

Train the model with the created data2015 with the semi-supervised learning method, and then test the model on samples newly labeled in 2017

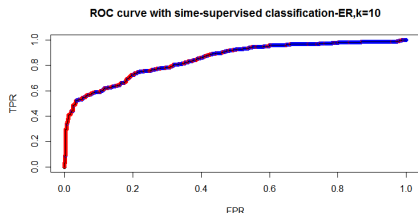


Figure: AUC=0.848

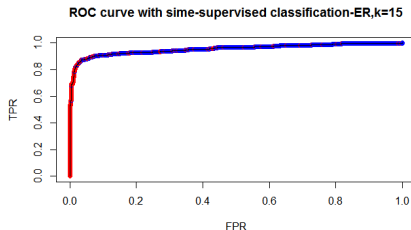


Figure: AUC=0.953

Experiments

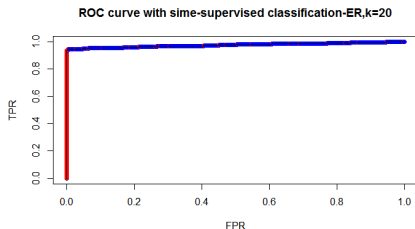


Figure: AUC=0.976

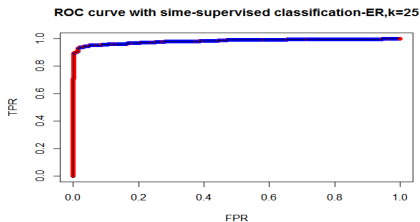


Figure: AUC=0.980

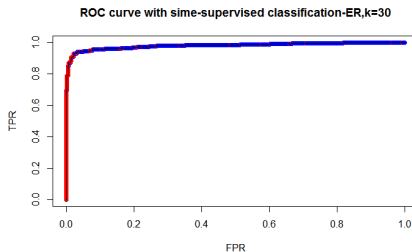


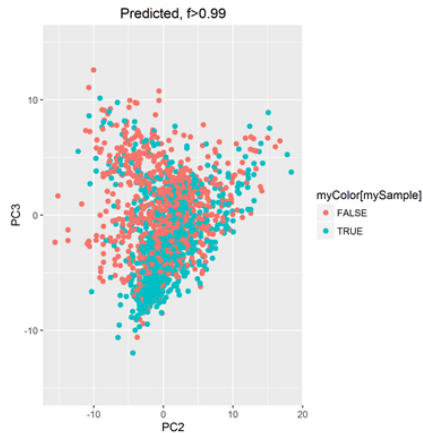
Figure: AUC=0.978

Experiments

KNN graph type semi-supervised learning model features:

- Compared to the exponential-weighted graph, KNN type model greatly reduces the dimensions
- For ER, the best model parameters are $k=25$ ($AUC = 0.980$)

PCA Visualization



Conclusions

- Our approach was successful in identifying novel organelle members as measured by classification accuracy in our test set;
- The explicit consideration of the unlabeled samples in modeling is an advance over other machine learning approaches in organelle proteomics;
- Still remaining to be done is it's application to canine heart data.