Graph Based Semi-supervised Learning Methods Applied to Speech Recognition Problem

Hoang Trang[1] and Loc Hoang Tran[2(✉)]

[1] Ho Chi Minh City University of Technology-VNU HCM
Ho Chi Minh City, Vietnam
hoangtrang@hcmut.edu.vn
[2] Computer Science Department/University of Minnesota, Minneapolis, USA
tran0398@umn.edu

Abstract. Speech recognition is the important problem in pattern recognition research field. In this paper, the un-normalized, symmetric normalized, and random walk graph Laplacian based semi-supervised learning methods will be applied to the network derived from the MFCC feature vectors of the speech dataset. Experiment results show that the performance of the random walk and the symmetric normalized graph Laplacian based methods are at least as good as the performance of the un-normalized graph Laplacian based method. Moreover, the sensitivity measures of these three semi-supervised learning methods are much better than the sensitivity measure of the current state of the art Hidden Markov Model method in speech recognition problem.

Keywords: Semi-supervised learning · Graph laplacian · Speech recognition · MFCC

1 Introduction

Two of the most noticeable areas of machine learning research are supervised and unsupervised learning. In supervised learning, a learner tries to obtain a predictive model from explicitly labeled training samples. However, in unsupervised learning, a learner tries to mine a descriptive model from unlabeled training samples. Recently, interest has increased in the hybrid problem of learning a predictive model given a combination of both labeled and unlabeled samples. This revised learning problem, commonly referred to as semi-supervised learning, rises in many real world applications, such as text and gene classification [1,2,3,4,5], because of the freely available of unlabeled data and because of the labor-intensive effort and high time complexity to obtain the explicitly labeled data. For example, in text classification, excessive work is required to manually label a set of documents for supervised training while unlabeled documents are available in abundance. It is normal, in this case, to try to exploit the existence of a large set of unlabeled documents and to lessen the number of labeled documents required to learn a good document classifier. Similarly, in the problem of predicting gene function from microarray data and sequence information, the experiments needed to label a subset of the genes are normally very costly to conduct. As a result, there exist only a few hundred labeled genes out of the population of thousands.

© Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2015
P.C. Vinh et al. (Eds.): ICTCC 2014, LNICST 144, pp. 264–273, 2015.
DOI: 10.1007/978-3-319-15392-6_26

Although it is a challenging problem, semi-supervised learning offers acceptable promise in practice that many algorithms have been suggested for this type of problem in the past few years. Among these algorithms, graph based learning algorithms have become common due to their computational efficiency and their effectiveness at semi-supervised learning. Some of these graph based learning algorithms make predictions directly for a target set of unlabeled data without creating a model that can be used for out-of-sample predictions. This process is called transductive learning. Such algorithms avoid many of the requirements of traditional supervised learning and can be much simpler as a result. However, other approaches to semi-supervised learning still create a model that can be used to predict unseen test data.

In this paper, we will present the graph based semi-supervised learning methods, derive their detailed regularization framework, and apply these methods to automatic speech recognition problem. To the best of our knowledge, this work has not been investigated. Researchers have worked in automatic speech recognition for almost six decades. The earliest attempts were made in the 1950's. In the 1980's, speech recognition research was characterized by a shift in technology from template-based approaches to statistical modeling methods, especially Hidden Markov Models (HMM). Hidden Markov Models (HMM) have been the core of most speech recognition systems for over a decade and is considered the current state of the art method for automatic speech recognition system [6]. Second, to classify the speech samples, a graph (i.e. kernel) which is the natural model of relationship between speech samples can also be employed. In this model, the nodes represent speech samples. The edges represent for the possible interactions between nodes. Then, machine learning methods such as Support Vector Machine [7], Artificial Neural Networks [8], or nearest-neighbor classifiers [9] can be applied to this graph to classify the speech samples. The nearest-neighbor classifiers method labels the speech sample with the label that occurs frequently in the speech sample's adjacent nodes in the network. Hence neighbor counting method does not utilize the full topology of the network. However, the Artificial Neural Networks, Support Vector Machine, and graph based semi-supervised learning methods utilize the full topology of the network. Moreover, the Artificial Neural Networks and Support Vector Machine are supervised learning methods.

While nearest-neighbor classifiers method, the Artificial Neural Networks, and the graph based semi-supervised learning methods are all based on the assumption that the labels of two adjacent speech samples in graph are likely to be the same, SVM does not rely on this assumption. Graphs used in nearest-neighbor classifiers method, Artificial Neural Networks, and the graph based semi-supervised learning method are very sparse. However, the graph (i.e. kernel) used in SVM is fully-connected.

In the last decade, the normalized graph Laplacian [2], random walk graph Laplacian [1], and the un-normalized graph Laplacian [3, 5] based semi-supervised learning methods have successfully been applied to some specific classification tasks such as digit recognition, text classification, and protein function prediction. However, to the best of our knowledge, the graph based semi-supervised learning methods have not yet been applied to automatic speech recognition problem and hence their overall sensitivity performance measure comparisons have not been done. In this paper, we will apply three un-normalized, symmetric normalized, and random walk graph Laplacian based semi-supervised learning methods to the network derived from

the speech samples. The main point of these three methods is to let every node of the graph iteratively propagates its label information to its adjacent nodes and the process is repeated until convergence [2].

We will organize the paper as follows: Section 2 will introduce graph based semi-supervised learning algorithms in detail. Section 3 will show how to derive the closed form solutions of normalized and un-normalized graph Laplacian based semi-supervised learning from regularization framework. In section 4, we will apply these three algorithms to the network derived from speech samples available from the IC Design lab at Faculty of Electricals-Electronics Engineering, University of Technology, Ho Chi Minh City. Section 5 will conclude this paper and discuss the future directions of researches of this automatic speech recognition problem utilizing hypergraph Laplacian.

2 Algorithms

Given a set of feature vectors of speech samples $\{x_1, \dots, x_l, x_{l+1}, \dots, x_{l+u}\}$ where $n = l + u$ is the total number of speech samples in the network, define c be the total number of words and the matrix $F \in R^{n*c}$ be the estimated label matrix for the set of feature vectors of speech samples $\{x_1, \dots, x_l, x_{l+1}, \dots, x_{l+u}\}$, where the point x_i is labeled as sign(F_{ij}) for each word j ($1 \le j \le c$). Please note that $\{x_1, \dots, x_l\}$ is the set of all labeled points and $\{x_{l+1}, \dots, x_{l+u}\}$ is the set of all un-labeled points. The way constructing the feature vectors of speech samples will be discussed in Section IV.

Let $Y \in R^{n*c}$ the initial label matrix for n speech samples in the network be defined as follows

$$Y_{ij} = \begin{cases} 1 \; if \; x_i belongs \; to \; word \; j \; and \; 1 \le i \le l \\ -1 \; if \; x_i \; does \; not \; belong \; to \; word \; j \; and \; 1 \le i \le l \\ 0 \; if \; l+1 \le i \le n \end{cases}$$

Our objective is to predict the labels of the un-labeled points $x_{l+1}, \dots, x_{l+u}$. We can achieve this objective by letting every node (i.e. speech sample) in the network iteratively propagates its label information to its adjacent nodes and this process is repeated until convergence.

Let W represents the network.

Random walk graph Laplacian based semi-supervised learning algorithm

In this section, we slightly change the original random walk graph Laplacian based semi-supervised learning algorithm can be obtained from [1]. The outline of the new version of this algorithm is as follows

1. Form the affinity matrix W. The way constructing W will be discussed in section IV.
2. Construct $S_{rw} = D^{-1}W$ where $D = diag(d_1, d_2, \dots, d_n)$ and $d_i = \sum_j W_{ij}$
3. Iterate until convergence
   $F^{(t+1)} = \alpha S_{rw} F^{(t)} + (1 - \alpha)Y$, where α is an arbitrary parameter belongs to [0,1]

4. Let F^* be the limit of the sequence $\{F^{(t)}\}$. For each word j, label each speech samples $x_i\,(l+1 \le i \le l+u)$ as $\text{sign}(F^*_{ij})$

Next, we look for the closed-form solution of the random walk graph Laplacian based semi-supervised learning. In the other words, we need to show that

$$F^* = \lim_{t\to\infty} F^{(t)} = (1-\alpha)(I-\alpha S_{rw})^{-1}Y$$

Suppose $F^{(0)} = Y$, then

$$\begin{aligned}
F^{(1)} &= \alpha S_{rw}F^{(0)} + (1-\alpha)Y \\
&= \alpha S_{rw}Y + (1-\alpha)Y \\
F^{(2)} &= \alpha S_{rw}F^{(1)} + (1-\alpha)Y \\
&= \alpha S_{rw}(\alpha S_{rw}Y + (1-\alpha)Y) + (1-\alpha)Y \\
&= \alpha^2 S_{rw}^2 Y + (1-\alpha)\alpha S_{rw}Y + (1-\alpha)Y \\
F^{(3)} &= \alpha S_{rw}F^{(2)} + (1-\alpha)Y \\
&= \alpha S_{rw}(\alpha^2 S_{rw}^2 Y + (1-\alpha)\alpha S_{rw}Y + (1-\alpha)Y) + (1-\alpha)Y \\
&= \alpha^3 S_{rw}^3 Y + (1-\alpha)\alpha^2 S_{rw}^2 Y + (1-\alpha)\alpha S_{rw}Y + (1-\alpha)Y \\
&\cdots
\end{aligned}$$

Thus, by induction,

$$F^{(t)} = \alpha^t S_{rw}^t Y + (1-\alpha)\sum_{i=0}^{t-1}(\alpha S_{rw})^i Y$$

Since S_{rw} is the stochastic matrix, its eigenvalues are in [-1,1]. Moreover, since 0<α<1, thus

$$\lim_{t\to\infty} \alpha^t S_{rw}^t = 0$$

$$\lim_{t\to\infty} \sum_{i=0}^{t-1}(\alpha S_{rw})^i = (I-\alpha S_{rw})^{-1}$$

Therefore,

$$F^* = \lim_{t\to\infty} F^{(t)} = (1-\alpha)(I-\alpha S_{rw})^{-1}Y$$

Now, from the above formula, we can compute F^* directly.

The original random walk graph Laplacian based semi-supervised learning algorithm developed by Zhu can be derived from the modified algorithm by setting $\alpha_i = 0$, where $1 \le i \le l$ and $\alpha_i = 1$, where $l+1 \le i \le l+u$. In the other words, we can express $F^{(t+1)}$ in matrix form as follows

$$F^{(t+1)} = I_\alpha S_{rw}F^{(t)} + (I-I_\alpha)Y, \text{ where}$$

I is the identity matrix and

$$I_\alpha = \begin{bmatrix} 0 & \cdots & 0 & & & \\ \vdots & \ddots & \vdots & & 0 & \\ 0 & \cdots & 0 & & & \\ & & & 1 & \cdots & 0 \\ & 0 & & \vdots & \ddots & \vdots \\ & & & 0 & \cdots & 1 \end{bmatrix} (I_\alpha \textit{ is the diagonal matrix})$$

Normalized graph Laplacian based semi-supervised learning algorithm

Next, we will give the brief overview of the original normalized graph Laplacian based semi-supervised learning algorithm can be obtained from [2]. The outline of this algorithm is as follows

1. Form the affinity matrix W
2. Construct $S_{sym} = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$ where $D = diag(d_1, d_2, \ldots, d_n)$ and $d_i = \sum_j W_{ij}$
3. Iterate until convergence
   $F^{(t+1)} = \alpha S_{sym} F^{(t)} + (1-\alpha)Y$, where α is an arbitrary parameter belongs to [0,1]
4. Let F^* be the limit of the sequence $\{F^{(t)}\}$. For each word *j*, label each speech samples $x_i\ (l+1 \le i \le l+u)$ as sign(F^*_{ij})

Next, we look for the closed-form solution of the normalized graph Laplacian based semi-supervised learning. In the other words, we need to show that

$$F^* = \lim_{t\to\infty} F^{(t)} = (1-\alpha)\left(I - \alpha S_{sym}\right)^{-1} Y$$

Suppose $F^{(0)} = Y$, then

$$\begin{aligned} F^{(1)} &= \alpha S_{sym} F^{(0)} + (1-\alpha)Y \\ &= \alpha S_{sym} Y + (1-\alpha)Y \\ F^{(2)} &= \alpha S_{sym} F^{(1)} + (1-\alpha)Y \\ &= \alpha^2 S_{sym}^2 Y + (1-\alpha)\alpha S_{sym} Y + (1-\alpha)Y \\ F^{(3)} &= \alpha S_{sym} F^{(2)} + (1-\alpha)Y \\ &= \alpha^3 S_{sym}^3 Y + (1-\alpha)\alpha^2 S_{sym}^2 Y + (1-\alpha)\alpha S_{sym} Y + (1-\alpha)Y \\ &\ldots \end{aligned}$$

Thus, by induction,

$$F^{(t)} = \alpha^t S_{sym}^t Y + (1-\alpha)\sum_{i=0}^{t-1} (\alpha S_{sym})^i Y$$

Since $D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$ is similar to $D^{-1}W$ which is a stochastic matrix, eigenvalues of $D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$ belong to [-1,1]. Moreover, since 0<α<1, thus

$$\lim_{t\to\infty} \alpha^t S_{sym}^t = 0$$

$$\lim_{t\to\infty} \sum_{i=0}^{t-1} (\alpha S_{sym})^i = (I - \alpha S_{sym})^{-1}$$

Therefore,

$$F^* = \lim_{t\to\infty} F^{(t)} = (1-\alpha)(I - \alpha S_{sym})^{-1} Y$$

Now, from the above formula, we can compute F^* directly.

Un-normalized graph Laplacian based semi-supervised learning algorithm

Finally, we will give the brief overview of the un-normalized graph Laplacian based semi-supervised learning algorithm [3]. The outline of this algorithm is as follows

1. Form the affinity matrix W
2. Construct $L = D - W$, where $D = diag(d_1, d_2, \dots, d_n)$ and $d_i = \sum_j W_{ij}$
3. Compute closed form solution $F^* = \gamma(L + \gamma I)^{-1} Y$, where γ is any positive parameter
4. For each word j, label each speech samples $x_i\ (l+1 \le i \le l+u)$ as $\text{sign}(F_{ij}^*)$

The closed form solution F^* of un-normalized hypergraph Laplacian based semi-supervised learning algorithm will be derived clearly and completely in Regularization Framework section.

3 Regularization Frameworks

In this section, we will develop the regularization framework for the normalized graph Laplacian based semi-supervised learning iterative version. First, let's consider the error function

$$E(F) = \left\{ \frac{1}{2} \sum_{i,j=1}^{n} W_{ij} \left\| \frac{F_i}{\sqrt{d_i}} - \frac{F_j}{\sqrt{d_j}} \right\|^2 \right\} + \gamma \sum_{i=1}^{n} \|F_i - Y_i\|^2$$

In this error function $E(F)$, F_i and Y_i belong to R^c. Please note that c is the total number of words, $d_i^{(k)} = \sum_j W_{ij}^{(k)}$, and γ is the positive regularization parameter. Hence

$$F = \begin{bmatrix} F_1^T \\ \vdots \\ F_n^T \end{bmatrix} \; and \; Y = \begin{bmatrix} Y_1^T \\ \vdots \\ Y_n^T \end{bmatrix}$$

Here $E(F)$ stands for the sum of the square loss between the estimated label matrix and the initial label matrix and the smoothness constraint.

Hence we can rewrite $E(F)$ as follows

$$E(F) = trace\left(F^T \left(I - S_{sym}\right) F\right) + \gamma trace((F - Y)^T (F - Y))$$

Our objective is to minimize this error function. In the other words, we solve

$$\frac{\partial E}{\partial F} = 0$$

This will lead to

$$\begin{aligned}\left(I - S_{sym}\right)F + \gamma(F - Y) &= 0\\ F - S_{sym}F + \gamma F &= \gamma Y\\ F - \frac{1}{1+\gamma}S_{sym}F &= \frac{\gamma}{1+\gamma}Y\\ \left(I - \frac{1}{1+\gamma}S_{sym}\right)F &= \frac{\gamma}{1+\gamma}Y\end{aligned}$$

Let $\alpha = \frac{1}{1+\gamma}$. Hence the solution F^* of the above equations is

$$F^* = (1-\alpha)(I - \alpha S_{sym})^{-1}Y$$

Also, please note that $S_{rw} = D^{-1}W$ is not the symmetric matrix, thus we cannot develop the regularization framework for the random walk graph Laplacian based semi-supervised learning iterative version.

Next, we will develop the regularization framework for the un-normalized graph Laplacian based semi-supervised learning algorithms. First, let's consider the error function

$$E(F) = \left\{\frac{1}{2}\sum_{i,j=1}^{n} W_{ij}\left\|F_i - F_j\right\|^2\right\} + \gamma\sum_{i=1}^{n}\|F_i - Y_i\|^2$$

In this error function $E(F)$, F_i and Y_i belong to R^c. Please note that c is the total number of words and γ is the positive regularization parameter. Hence

$$F = \begin{bmatrix}F_1^T\\ \vdots\\ F_n^T\end{bmatrix} \; and \; Y = \begin{bmatrix}Y_1^T\\ \vdots\\ Y_n^T\end{bmatrix}$$

Here $E(F)$ stands for the sum of the square loss between the estimated label matrix and the initial label matrix and the smoothness constraint.

Hence we can rewrite $E(F)$ as follows

$$E(F) = trace(F^T L F) + \gamma trace((F-Y)^T(F-Y))$$

Please note that un-normalized Laplacian matrix of the network is $L = D - W$. Our objective is to minimize this error function. In the other words, we solve

$$\frac{\partial E}{\partial F} = 0$$

This will lead to

$$\begin{aligned}LF + \gamma(F - Y) &= 0\\ (L + \gamma I)F &= \gamma Y\end{aligned}$$

Hence the solution F^* of the above equations is

$$F^* = \gamma(L + \gamma I)^{-1}Y$$

4 Experiments and Results

In this paper, the set of 4,500 speech samples recorded of 50 different words (90 speech samples per word) are used for training. Then another set of 500 speech samples of these words are used for testing the sensitivity measure. This dataset is available from the IC Design lab at Faculty of Electricals-Electronics Engineering, University of Technology, Ho Chi Minh City. After being extracted from the conventional MFCC feature extraction method, the column sum of the MFCC feature matrix of the speech sample will be computed. The result of the column sum which is the R^{26*1} column vector will be used as the feature vector of the three graph Laplacian based semi-supervised learning algorithms.

There are three ways to construct the similarity graph from these feature vectors:

a. The ε-neighborhood graph: Connect all speech samples whose pairwise distances are smaller than ε.
b. k-nearest neighbor graph: Speech sample *i* is connected with speech sample *j* if speech sample *i* is among the k-nearest neighbor of speech sample *j* or speech sample *j* is among the k-nearest neighbor of speech sample *i*.
c. The fully connected graph: All speech samples are connected.

In this paper, the similarity function is the Gaussian similarity function

$$s\big(f(:,i), f(:,j)\big) = \exp\left(-\frac{d\big(f(:,i), f(:,j)\big)}{t}\right),$$

where $f(:,i)$ is the feature vector of speech sample *i*.

In this paper, t is set to 10^6 and the 5-nearest neighbor graph is used to construct the similarity graph from this dataset.

In this section, we experiment with the above three methods in terms of sensitivity measure. All experiments were implemented in Matlab 6.5 on virtual machine. The sensitivity measure Q is given as follows

$$Q = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) are defined in the following table 1

Table 1. Definitions of TP, TN, FP, and FN

		Predicted Label	
		Positive	Negative
Known Label	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

In these experiments, parameter α is set to 0.85 and $\gamma = 1$. For this dataset, the table 2 shows the sensitivity measures of the three methods and HMM method (i.e. the current state of the art method of speech recognition application) applying to network for 50 words.

Table 2. Comparisons of symmetric normalized, random walk, and un-normalized graph Laplacian based methods and HMM method

Sensitivity Measure (%)			
Normalized	Random Walk	Un-normalized	HMM (8 states, 4 mixtures)
97.60%	97.60%	97.60%	89%

The following figure 1 shows the sensitivity measures of the conventional HMM method and the three graph Laplacian based semi-supervised learning methods:

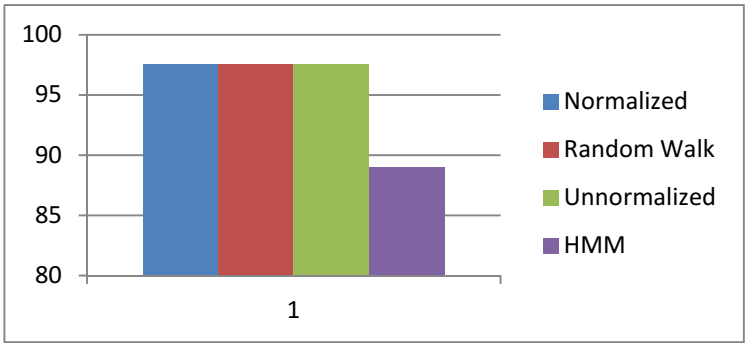

Fig. 1. Sensitivity measures of the three graph based semi-supervised learning methods and conventional HMM method

From the above table 2 and figure 1, we recognized that the symmetric normalized and un-normalized graph Laplacian based semi-supervised learning methods slightly perform better than the random walk graph Laplacian based semi-supervised learning method. Moreover, these three graph Laplacian based semi-supervised learning methods outperform the current state of the art HMM method in speech recognition problem since the graph based semi-supervised learning methods utilize the "relationship" among all speech samples in the datasets (i.e. the kernel's definition) to build the predictive model.

5 Conclusions

The detailed iterative algorithms and regularization frameworks for the three normalized, random walk, and un-normalized graph Laplacian based semi-supervised learning methods applying to the speech recognition problem have been developed. These three methods are successfully applied to this problem (i.e. classification problem). Moreover, the comparison of the sensitivity performance measures for these three methods and the current state of the art HMM method has been done.

Moreover, these three methods can not only be used in classification problem but also in ranking problem. Given a set of genes (i.e. the queries) involved in a specific disease (for e.g. leukemia), these three methods can also be used to find more genes involved in the same disease by ranking genes in gene co-expression network (derived from gene expression data) or the protein-protein interaction network or the integrated network of them. The genes with the highest rank then will be selected and then checked by biologist experts to see if the extended genes in fact are involved in the same disease. These problems are also called biomarker discovery in cancer classification.

Finally, to the best of our knowledge, the normalized, random walk, and unnormalized hypergraph Laplacian based semi-supervised learning methods have not been applied to the speech recognition problem. These methods applied to the speech recognition problem are worth investigated since [10] have shown that these hypergraph Laplacian based semi-supervised learning methods outperform the graph Laplacian based semi-supervised learning methods in text-categorization and letter recognition tasks.

Acknowledgement. This work is funded by the Ministry of Science and Technology, State-level key program, Research for application and development of information technology and communications, code KC.01.23/11-15.

References

1. Zhu, X., Ghahramani, Z.: Learning from labeled and unlabeled data with label propagation. Technical Report CMU-CALD-02-107, Carnegie Mellon University (2002)
2. Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Schölkopf, B.: Learning with local and global consistency. In: Thrun, S., Saul, L., Schölkopf, B. (eds.) Advances in Neural Information Processing Systems (NIPS), vol. 16, pp. 321–328. MIT Press, Cambridge (2004)
3. Tsuda, K., Shin, H.H., Schoelkopf, B.: Fast protein classification with multiple networks. Bioinformatics (ECCB 2005) **21**(Suppl. 2), ii59–ii65 (2005)
4. Tran, L.: Application of three graph Laplacian based semi-supervised learning methods to protein function prediction problem. CoRR abs/1211.4289 (2012)
5. Tran, L.: The Un-normalized graph p-Laplacian based semi-supervised learning method and protein function prediction problem. In: Huynh, V.N., Denoeux, T., Tran, D.H., Le, A.C., Pham, B.S. (eds.) KSE 2013, Part I. AISC, vol. 244, pp. 23–35. Springer, Heidelberg (2014)
6. Rabiner, L., Juang, B.H.: Fundamentals of speech recognition, 507 pp. AT&T (1993)
7. Ganapathiraju, A.: Support vector machines for speech recognition. Diss. Mississippi State University (2002)
8. Marshall, A.: Artificial Neural Network for Speech Recognition, 2nd Annual Student Research Showcase (2005)
9. Labiak, J., Livescu, K.: Nearest neighbor classifiers with learned distances for phonetic frame classification. In: Proceedings of Interspeech (2011)
10. Zhou, D., Huang, J., Schölkopf, B.: Learning with hypergraphs: clustering, classification, and embedding. In: Schölkopf, B., Platt, J.C., Hofmann, T. (eds.) Advances in Neural Information Processing System (NIPS), vol. 19, pp. 1601–1608. MIT Press, Cambridge (2007)