基于自适应图的半监督学习方法①

梅松青

(广州医科大学 信息管理与信息系统系, 广州 510182)

摘 要: 基于图的半监督学习方法中, 图结构经常要预先设定, 这就导致了在标签传递过程中, 算法不能自适应地学习一个最优的图. 为此, 提出了一种基于自适应图的半监督学习方法. 该方法通过迭代的优化方法同时学习到最优的图和标签. 而且, 在少量标记样本的情况下该方法也可以得到较高的分类准确率, 并通过实验证明了该方法的有效性.

关键词: 半监督学习; 标签传递; 优化算法; 分类准确率

Adaptive Graph-Based Semi-Supervised Learning Method

MEI Song-Qing

(Dept of Information management and information system Guangzhou Medical University, Guangzhou 510182, China)

Abstract: In most graph-based semi-supervised methods, graph structure is often set in advance, which leads to the fact that the algorithm can't learn an optimal graph in the process of label propagation. Therefore, this paper proposes a method called Adaptive Graph-based Semi-supervised Learning Method (AGSSLM). This method can learn the optimal graph and label simultaneously by using the iterative optimization method. Moreover, this method can also obtain higher classification accuracy with fewer labeled samples. The experimental results validate the effectiveness of this method.

Key words: semi-supervised learning; label Propagation; optimization algorithm; classification accuracy

在现实应用中, 因为人力和时间的约束有标签的数据通常十分有限, 未标记的数据非常容易收集. 特别是在因特网高速发展的今天, 网络上未标记标签的数据更是成千上万. 如果只是使用少量的有标签的样本, 学习到的模型很难具有较好的泛化能力; 另一方面, 如果仅仅使用没有标签的样本, 例如传统的 K-means 聚类算法, 不仅得不到好的分类模型, 而且浪费了宝贵的有标记的样本. 半监督学习算法[1]的出现在一定的程度上解决了这个难题, 利用大量的没有标记的样本来辅助有标记的样本, 使学习到的模型具有较强的泛化能力. 目前, 在半监督学习领域, 涌现了大量优秀的算法, 如近邻的传播算法[2], 判别半监督的聚类分析[3], K-means 多关系数据聚类算法[4]等. 半监督学习不仅要为未标记的样本分配正确的标签, 而且还要求学习到一个决策函数使得错误率最小.

基于图的半监督学习方法先将总体样本点映射成联通带权的无向图, 然后在图上设计一系列的算法. 例如陈利用基于图的半监督学习从文本中识别出模式之间的关系[5], 基于图的特征提取算法[6], zhou[7]等人利用 K-近邻图使算法平滑. 以上列出的算法均在部分数据集上取得了可观的效果. 但是, 由于原始数据中存在噪声或者一系列冗余的特征, 使得构造的图不准确. 对于基于图的半监督学习方法, 图的构造直接影响算法的性能. 如何构造一个正确的权重图已成为一个研究重点, 例如文献[8]自动地计算数据之间的相似度, 且通过对图的权重矩阵的调整使得设计的算法对权重不敏感, 文献[9]利用 e_1-norm 对噪声不敏感的特性设计基于 e_1-norm 图的半监督学习方法等.

Gaussian-Laplacian 正则化(GLR)[10]框架是一个半监督学习的经典框架, 目前大部分的半监督学习算法都是基于此框架. GLR 框架是基于以下聚类假设: 相近点的标签有可能相同; 相同结构中的点标签相同.

① 收稿时间:2013-07-18;收到修改稿时间:2013-09-09

该假设是局部与全局相结合的.

针对原始数据构建图不准确的缺点, 结合GLR框架, 本文提出一种基于自适应图的半监督学习方法(AGSSLM). 一方面利用虚拟的图拉普拉斯学习一个更为正确的分类函数来为未标记的样本分派标签, 另一方面用该虚拟的图拉普拉斯来逼近根据原始数据所构建的图拉普拉斯, 使学习到的图拉普拉斯不会太偏离于正确的图拉普拉斯. 为达到该目的, 提出了一个迭代的优化方法使得半监督学习方法能够同时学习到一个最优的图拉普拉斯和一个分类函数.

1　基于自适应图的半监督学习方法(AGSSLM)

基于图的半监督学习方法基本设置如下:

假设有 n 个样本, $X=\{x_1,x_2,...,x_l,x_{l+1},...,x_n\}\in\Re^d$, 其中, $X_l=\{x_i\}_{i=1}^{l}$ 是已经标记的样本, 其对应的标签集合 $L=\{1,2,...,c\}$. $X_u=\{x_i\}_{i=l+1}^{n}$ 是未标记的样本, 一般情况下, $l\ll n$. 聚类假设和流形假设是两个常见的假设, 用来建立基于图的半监督学习框架, 这两个假设在本质上是一致的. 对于任意的样本, 可以将样本点映射成带权的无向图 $G=(V,E)$, 其中 V 表示顶点的集合, E 为带权的边值. 每一条边 $e_{i,j}\in E$ 对应一条权重 $w_{i,j}$, 它反映着 x_i 和 x_j 之间的相识度. K-近邻(KNN)算法通常被广泛用来构造图的权重.

如果 x_j 是 x_i 的 K 近邻且 x_i 是 x_j 的 K 近邻, 其对应的权重 $w_{i,j}=\mathrm{e}^{\frac{\|x_i-x_j\|^2}{2\sigma^2}}$; 否则 $w_{i,j}=0$. 在实际应用中, 通常设置 $w_{i,j}=0$. 基于图的半监督学习方法往往被看作在 G 上建立分类函数 f. 为满足聚类假设, 最小化如下目标函数

$$H(f)=\alpha C(x_L)+\beta S(f) \tag{1}$$

其中 $C(x_L)$ 用于度量为已标记的数据的标签与预测标签的差值. $S(f)$ 是惩罚项使算法更加平滑. 基于上述模型, GLR 框架通过最小化以下目标函数来学习一个分类函数:

$$f=\arg\min_f\sum_{i=1}^{l}(f_i-y_i)^2+\lambda(1/2)\sum_{i=1}^{n}\sum_{j=1}^{n}(f_i-f_j)^2w_{i,j} \tag{2}$$

其中 $y_i=[y_{1i},y_{2i},...y_{ci}]^T$ 是有标签样本的标签向量化. $y_{ji}=1$ 如果 x_i 标记为第 j 类, 否则 $y_{ji}=0$. f 为分类函数, 如果第 j 个样本预测的标签为 $k=\arg\max(f_{kj})$, $k\in\{1,2,...c\}$, 该样本标记为第 k 类. 转化(2)式后面项 $(1/2)\sum_{i=1}^{n}\sum_{j=1}^{n}(f_i-f_j)^2w_{i,j}=Tr(fLf^T)$, Tr 为迹范数. 其中 $L=D-W$ 是图拉普拉斯, $W\in\Re^{n\times n}$ 是图的权重矩阵, $D=diag(\sum_i w_{1i},...,w_{mi})$.

由于在数据的采集的过程中, 测量的偏差、光照、不恰当的度量等造成噪声数据的存在, 而且一些高维的数据中一般存在大量冗余的特征, 这些因素都会影响图的构造. 例如数据点 x_i 和 x_j 同时被污染, 或者其中的一个数据点被污染, 那么权重 $\tilde{w}_{i,j}$ 不可能真实地反应 x_i 和 x_j 之间的关系, 设置 $d=\|\tilde{w}_{i,j}-w_{i,j}\|^2$ 表示 $\tilde{w}_{i,j}$ (虚拟的权重)与真实的 $w_{i,j}$ 之间的残差. 如果残差过大, 基于图的半监督算法不可能得到满意的结果. 为此, 提出自适应的图的半监督学习方法, 该算法可以自适应地学习一个正确的图拉普拉斯 $\tilde{L}$. 为防止 $\tilde{L}$ 不会太偏离于真实的图拉普拉斯, 使用 $\tilde{L}$ 逼近原始数据上的图拉普拉斯 L, 即 $\|\tilde{L}-L\|^2$. 为学习完美的 $\tilde{L}$ 和分类函数 f, 最小化下面的目标函数:

$$(f,\tilde{L})=\arg\min_{\tilde{L},f}\{\sum_{i=1}^{l}\|f_i-y_i\|^2+\lambda_1(1/2)\sum_{i=1}^{n}\sum_{j=1}^{n}\|f_i-f_j\|^2W_{i,j}+\lambda_2\|W_{ij}-\tilde{W}_{ij}\|^2\} \tag{3}$$

转化上式为:

$$(f,\tilde{L})=\arg\min_{\tilde{L},f}\{Tr\{(f_l-y_l)(f_l-y_l)^T\}+\lambda_1Tr\left(f\tilde{L}f^T\right)+\lambda_2Tr\{\left(\tilde{L}-L\right)\left(\tilde{L}-L\right)^T\}\} \tag{4}$$

为计算方便, 矩阵 $\tilde{L}$ 和 f 分块, $\tilde{L}=\begin{bmatrix}\tilde{L}_{ll} & \tilde{L}_{lu}\\ \tilde{L}_{lu}{}^T & \tilde{L}_{uu}\end{bmatrix}$, $f=[f_l\quad f_u]$, 其中 $(\cdot)_{ll}$ 对应有标签的矩阵块, $(\cdot)_{uu}$ 对应未标记标签矩阵块. f_1 对应有标签样本对应的预测函数, f_u 对应无标签样本对应的预测函数. 进一步将(4)式转化为

$$(f,\tilde{L})=\arg\min_{\tilde{L},f}\{Tr\{(f_l-Y_l)(f_l-Y_l)^T\}+\lambda_1Tr\left([f_l\ f_u]\begin{bmatrix}\tilde{L}_{ll} & \tilde{L}_{lu}\\ \tilde{L}_{lu}{}^T & \tilde{L}_{uu}\end{bmatrix}\begin{bmatrix}f_l^T\\ f_u^T\end{bmatrix}\right)+\lambda_2Tr\{\left(\tilde{L}-L\right)\left(\tilde{L}-L\right)^T\}\} \tag{5}$$

目标函数(5)式中含有两个未知量, 采用迭代的优化方法寻找最优的 $\tilde{L}$ 和分类函数 f, 其步骤如下:

(1) 给定 f, 目标函数变成

$$\tilde{L}=\arg\min_{\tilde{L}}\{\lambda_1 Tr\left(f\tilde{L}f^T\right)+\lambda_2 Tr\{\left(\tilde{L}-L\right)\left(\tilde{L}-L\right)^T\}\} \quad (6)$$

对 $\tilde{L}$ 求导并等于 0, 有下式成立

$$\lambda_1 f^T f+\lambda_2\left(2\tilde{L}-2L\right)=0 \quad (7)$$

即有 $\tilde{L}=L-(1/2)\left(\lambda_1/\lambda_2\right)f^T f$

(2) 给定 $\tilde{L}$, 并对 $\tilde{L}$ 分块, 目标函数变成

$$f=\arg\min_{f}\{Tr\{(f_l-Y_l)(f_l-Y_l)^T\}+\lambda_1 Tr\left(\begin{bmatrix}f_l & f_u\end{bmatrix}\begin{bmatrix}\tilde{L}_{ll} & \tilde{L}_{lu}\\ \tilde{L}_{lu}{}^T & \tilde{L}_{uu}\end{bmatrix}\begin{bmatrix}f_l^T\\ f_u^T\end{bmatrix}\right)\} \quad (8)$$

对 f_1 求导并等于 0, 有下式成立

$$\left(f_l\tilde{L}_{ll}+f_u\tilde{L}_{lu}\right)\lambda_1+f_l-Y_l=0 \quad (9)$$
$$\Rightarrow f_u=-f_l\tilde{L}_{lu}\tilde{L}_{uu}{}^{-1}$$

把 f_u 代入(9)得

$$\lambda_1\left(f_l\tilde{L}_{ll}-f_l\tilde{L}_{lu}\tilde{L}_{uu}{}^{-1}\tilde{L}_{lu}{}^T\right)+f_l-Y_l=0 \quad (10)$$
$$\Rightarrow f_l=(1/\lambda_1)Y_l\left(\tilde{L}_{ll}-\tilde{L}_{lu}\tilde{L}_{uu}{}^{-1}\tilde{L}_{lu}{}^T\right)^{-1}$$

综合(9)(10)式, 可以先求出 f_1, 再求出 f_u.

AGSSLM 算法总结如下:

步骤 1: 初始化 f 为单位矩阵, 设定 t=0;

步骤 2: Repeat

(a) 按照(7)式计算并且按照分块矩阵要求对 $\tilde{L}$ 分块.

(b) 按照(9)(10)计算 f_1, f_u 并且合并成矩阵 f;

t=t+1

Until

$$\begin{cases}\|f(t)-f(t-1)\|^2\le 0.001\\ \|\tilde{L}(t)-\tilde{L}(t-1)\|^2\le 0.001\end{cases}$$

AGSSLM 算法流程图如图 1 所示.

2 实验对比和分析

为了验证 AGSSLM 算法的有效性, 本节在不同的数据集上对不同的基于图的半监督算法进行比较, 例如 GLR, 线性近邻传递算法(Linear neighborhood propagation, LNP)[11], 高斯随机域(Gaussian random field, GRF)[12], 拉普拉斯最小二乘(Laplacian regularized least squares, LapRLS)[13]和学习局部和全局一致性(Learning with local and global consistency, LLGC)[14].

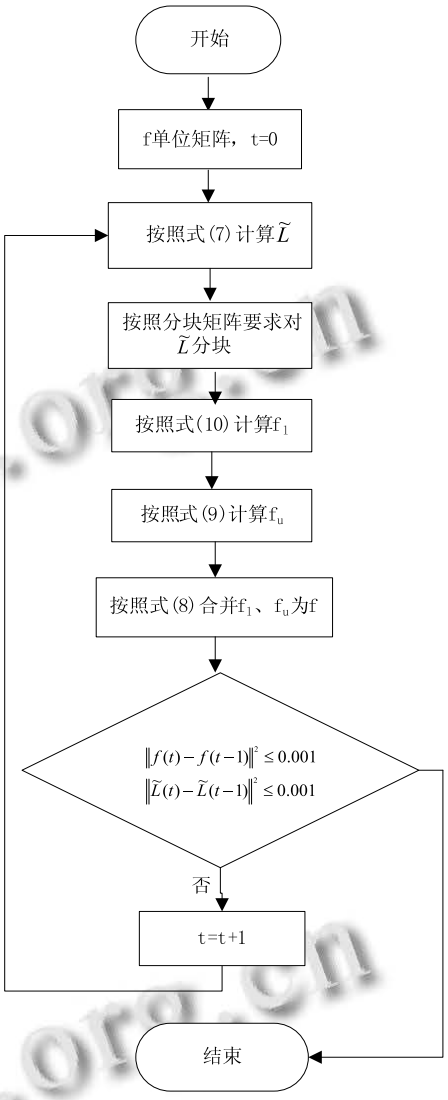

图 1 算法流程图

实验所用的数据集包括 UCI[15]数据集, USPS[16]数据集, COIL[16]数据集和 DIGIT1[17]数据集. 表 1 具体介绍了各个数据集的特征.

表 1 试验中使用的数据集介绍

数据集	维数	类别	数目
Balance-Scale	4	3	625
Sonar	60	2	208
Tic-Tac-Toe(TTT)	9	2	958
Glass	10	6	214
Ionosphere	34	2	351
USPS	241	2	1500
COIL	241	6	1500
DIGIT1	241	2	1500

在试验中用于训练和测试的数据集都是从整个数据集中随机选取的. 对于L的计算分成3步: 为每个数据集选择K个近邻; 计算权重; 计算图拉普拉斯L. 选择近邻采用 KNN 算法. 在所有的试验中 K 从集合$\{5,8,10,12,14,16,18,20,25\}$中选取; 计算权重的核函数使用高斯核, 其核参数$\sigma$从集合$\{2^{-4},2^{-3},2^{-2},2^{-1},2^{0},2^{1},2^{2}\}$中选取; 按照公式$L=D-W$计算图拉普拉斯 L, 其代码来自于蔡的主页(http://www.cad.zju.edu.cn/home/dengcai/Data/MLData.html). 试验中参数λ_1、λ_2从集合$\{0.01,0.05,0.1,0.5,1.0,1.5,2.0,2.5\}$中选取. 因为实验的计算复杂度, 在 UCI 的 Balance-Scale, Sonar 和 TTT 数据集上, 各随机抽取 30%, 50%和 70%的样本作为训练集, 其余的样本作为测试集, 进行 50 次的独立实验, 并且记录平均分类准确率和标准差作为最终的评价标准(所有实验都是在最好的参数组合上就行50次实验). 实验结果见表 2.

表 2　三个数据集上平均分类准确率与标准差

标签	GLR	LNP	AGSSLM
百分数	Balance-Scale	Sonar	TTT
30%	84.50 ± 0.07	85.01 ± 0.77	86.67 ± 0.01
50%	89.02 ± 0.03	90.98 ± 0.05	91.02 ± 0.68
70%	92.18 ± 0.44	93.63 ± 0.90	94.70 ± 0.82

从表 2 可以看出, 在 3 种 UCI 数据集上, 对比使用 GLR 和 LNP 方法, AGSSLM 方法取得了最好的实验效果. 可以看出, AGSSLM 能对不同数据集自适应地学习了一个最优的图拉普拉斯和一个分类函数, 因此 AGSSLM 能够较准确地为未标记样本分配标签.

为了进一步验证 AGSSLM 算法在少量有标签训练样本情况下的有效性, 除 TTT 数据集外, 对每个数据集随机选取$\ell=50$个有标签的样本作为训练集, 余下的作为测试集, 进行 10 次独立的实验, 记录平均分类准确率和标准差作为评价标准. 表 3 和图 2 给出了不同数据集上的分类实验结果, 从表 3 和图 2 中可以看出: 除在数据集 COIL 上 AGSSLM 算法略逊于算法 LLGC 外, 在其他的数据集上, AGSSLM 算法都优于别的算法, 特别是对于 DIGIT1 数据集, DIGIT1 的平均分类准确率高于算法 LLGC 接近 2%. 从表 3 和图 2 的结果可以看出, 在较少的有标签的样本情况下, AGSSLM 算法也能取得比同类算法较优的平均分类准确率. 在所有的实验中采用十字交叉法选择最优的参数组合, 所以试验中只给出相对于较好的分类准确率. 进一步就如何选择最优的参数组合进行深入研究.

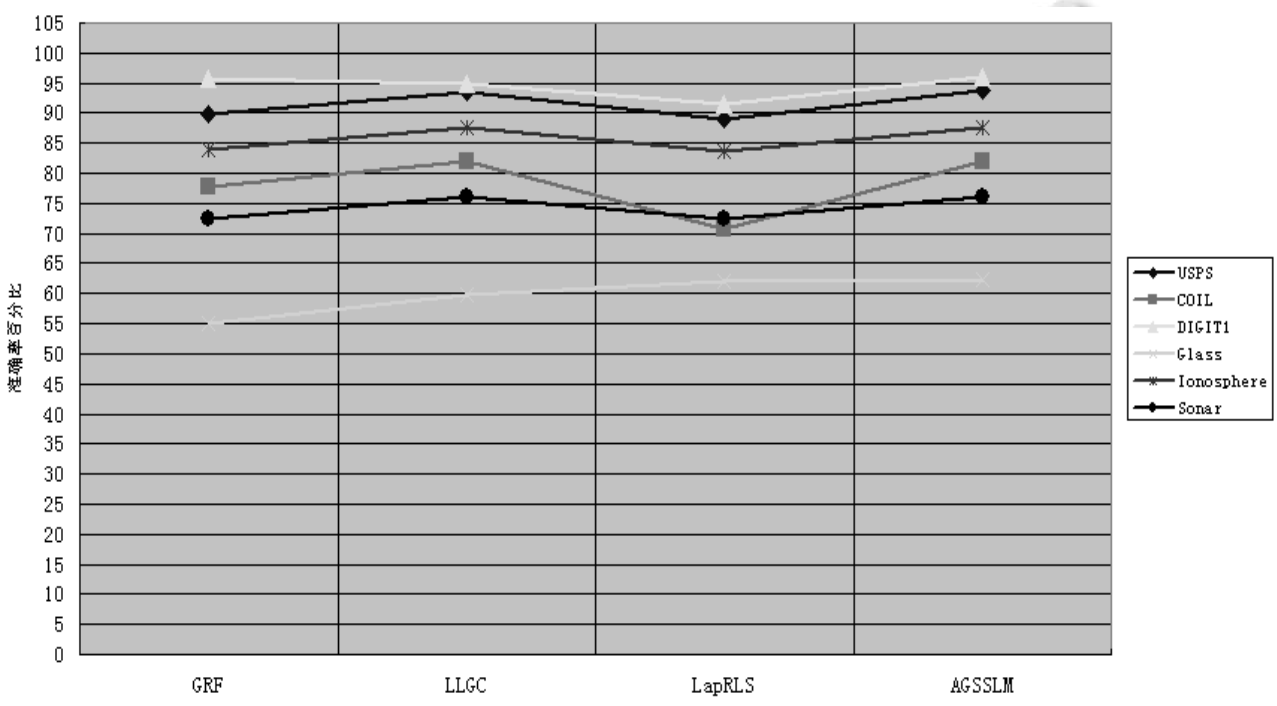

图 2　在不同数据集上不同算法分类准确率对比曲线图

表 3 六个数据集上的平均分类准确率和标准差

数据集	GRF	LLGC
USPS	89.72 ± 0.21	93.55 ± 0.29
COIL	77.90 ± 0.56	82.01 ± 0.43
DIGIT1	95.66 ± 0.71	94.85 ± 0.50
Glass	54.98 ± 2.54	59.80 ± 1.03
Ionosphere	84.00 ± 2.00	87.67 ± 1.01
Sonar	72.44 ± 0.78	75.99 ± 0.82
USPS	88.89 ± 0.56	**93.89 ± 0.22**
COIL	70.76 ± 0.53	82.00 ± 0.05
DIGIT1	91.54 ± 0.48	**96.01 ± 0.29**
Glass	62.00 ± 0.78	**62.33 ± 1.12**
Ionosphere	83.64 ± 0.98	**87.72 ± 1.35**
Sonar	72.43 ± 0.22	**76.05 ± 0.36**

3 结论

本文提出了一种自适应图的半监督学习方法, 以提高基于图的半监督学习方法的分类准确率. 该方法能够自适应地同时学习一个最优的图和分类函数. 为解决算法实施问题, 提出了一种迭代的优化方法. 实验结果很好的显示了该方法能够有效地提高基于图的半监督学习方法的分类准确率, 特别是在较少的训练集下, AGSSLM 也具有较好的分类结果. 进一步研究将如何更正确地学习一个最优的图结构, 例如在图上施加低秩限制; 利用 $\ell_{2,1}$ 范数代替 ℓ_2 范数来度量学习到的图与在真实数据集上训练得到的图之间差异; 最优参数组合等.

参考文献

1 Wei J, Pang H. Local and global preserving based semi-supervised dimensionality reduction method. Journal of Software, 2008, 19(11): 2833–2842.

2 肖宇,于剑.基于近邻传播算法的半监督聚类.软件学报, 2008,19(11):2803–2813.

3 尹学松,胡恩良,陈松灿.基于成对约束的判别型半监督聚类分析.软件学报,2008,19(11):2791–2802.

4 高滢,刘大有,齐红,刘赫.一种半监督 K 均值多关系数据聚类算法.软件学报,2008,19(11):2814–2821.

5 陈锦秀,姬东鸿.基于图的半监督关系抽取.软件学报, 2008,19(11):2843–2852.

6 Yang SH, Zha HY, Zhou SK, Hu BG. Variational graph embedding for globally and locally consistent feature extraction. In: Buntine W, ed. Proc. of the European Cone on Machine Learning and Knowledge Discovery in Databases: Part II(ECML PKDD 2009). Berlin, Springer-Verlag. 2009. 538–553.

7 Zhao Z, Liu H. Spectral feature selection for supervised and unsupervised learning. In: Ghahramani Z, ed. Proc. of the 24th Int'1 Conf. on Machine Learning(ICML 2007). New York, ACM. 2007. 1151–1157.

8 Wang F, Zhang CS. Robust self-tuning semi-supervised learning. Neuro Computing, 2007, 70(16–18): 2931–2939.

9 Yan SC, Wang H. Semi-supervised learning by sparse representation. Proc. of SDM. 2009. 792–801.

10 Belkin M, Matveeva L, Niyogi P. Regularization and semi-supervised learning on large graphs. Lecture Notes in Computer Science, 2004, 3120: 624–638.

11 Wang F, Zhang CS. Label propagation through linear neighborhoods. IEEE Trans. on Knowledge and Data Engineering, 2008, 20(1): 55–67.

12 Zhu X, Ghahramani Z, Lafferty J. Semi- supervised learning using Gaussian fields and harmonic functions. Proc. of the International Conference on Machine Learning. 2003.

13 Belkin M, Niyogi P, Sindhwani V. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. Journal of Machine Learning Research 7, 2006: 2399–2434.

14 Zhou D, Bousquet O, Lal T, Weston J, Scholkopf B. Learning with local and global consistency. Advances in Neural Information Processing Systems 16, 2004: 321–328.

15 Asuncion A, Newman D. UCI Machine Learning Repository. 2007. http://ergodicity.net/tag/machine-learning/.

16 http://www.cad.zju.edu.cn/home/dengcai/Data/MLData.html.

17 Chapelle O, Scholkopf B, Zien A. Semi-supervised Learning. MIT Press. 2006.