DOI:10.11721/cqnuj20170209

非对称稀疏图的半监督学习研究*

刘建峰[1]，吕　佳[2]
(1. 重庆三峡学院 教务处，重庆 万州 404100；2. 重庆师范大学 计算机与信息科学学院，重庆 401331)

摘要:【目的】如何构造一个有效的数据图，是半监督学习领域中一个重要的研究方向，为了更好地研究数据样本之间的结构关系，提高基于图的半监督学习算法性能。【方法】利用数据的稀疏表示，构造数据样本的非对称图，并在标准数据集上进行半监督学习实验。【结果】在半监督学习框架中建立了异类数据和同类数据之间距离、内部结构和数据的稀疏表示关系，构造了非对称稀疏的数据图。【结论】通过在标准数据集上进行实验说明非对称稀疏图可以利用半监督学习数据特点，有效地对数据样本进行分类。

关键词:数据图；半监督学习；稀疏表示；非对称图
中图分类号:TP391.4；TP301.6　　**文献标志码**:A　　**文章编号**:1672-6693(2017)02-0076-05

在机器学习中通常根据训练样本是否含有类标记将学习问题分成 3 类：训练样本全部含有标记并参加训练的学习方式称为监督学习[1-2]；训练样本集中少量样本含有类标记，大量样本没有类标记，同时这些无标记样本可以用来提高分类性能的学习方式，称为半监督学习；所有的训练样本都没有类标记的学习方式，称为无监督学习。从现实数据中获得样本的类标记是一件耗时费力的事情，因此通过少量的样本进行初始分类器的训练，而用大量的无标记样本提高分类器的泛化性能的半监督学习得到越来越多的研究。

到目前为止，已有很多学者提出了各种半监督学习算法，其中基于图的方法是一种重要的、有效的半监督学习算法，它建立在完备的数学基础之上，并随着基础数学的发展而不断改进。基于图的方法重点是图的构造，图上的结点表示样本点，图上的边表示样本点之间的关系，通常由权重决定样本点之间是否有边相连接，样本点之间有边则代表样本有相似的类标记[2]。

基于图的半监督方法大部分都比较类似，图的构造完成之后，建立目标函数 f，通过最优化技术求得目标函数的最优解，在求解目标函数的时候，大部分算法使用的是拉普拉斯算子。

Blum 和 Chawla[3] 提出基于图的半监督学习是最小切图的问题，并给出了无标记样本的最小分割图算法，但是该算法没有考虑边界数据的分类概率，在分类时没有给出相应的置信度，因此它是一种硬分类的方法，另外该方法不适宜扩展到多类问题。Savelli 等人[4]将图上的每个样本点标记由离散值松弛到连续域，并将随机场上的能力函数优化问题转化成了高斯场上的能量函数优化问题。Zhou[5] 等人受高斯场启发，提出了一种基于局部和全局的半监督学习，基本思想是让每个样本的标记信息向其邻近的样本迭代传播，直到全局稳定的状态。Tomioka 等人[6] 提出了一种流行正则化框架，目标函数带有两个正则项，许多基于图的半监督学习都是在此基础上进行改进的。

1 图的构造

图的构造有很多种方法，不同的构造方法对算法的性能会产生较大的影响。尽管图的构造问题是基于图方法的一个核心问题，但是并没有得到深入的研究。目前为止，图的构造方法有以下几种：全连图、近邻图、局部自适应图和稀疏图，这些方法在计算权重的时候采用的是高斯核函数[7-8]。

下面给出基于图方法的一些符号，训练集 $\boldsymbol{X}=\boldsymbol{L}\cup\boldsymbol{U}$，$\boldsymbol{X}=\{x_1,\cdots,x_l,x_{l+1},\cdots x_{l+u}\}$。$\boldsymbol{L}$ 是少量有标记样本，$\boldsymbol{L}=\{x_1,x_2,\cdots,x_l\}$，$\boldsymbol{U}$ 大量无标记样本，$\boldsymbol{U}=\{x_{l+1},x_{l+2},\cdots,x_n\}$。$\boldsymbol{G}=(\boldsymbol{V},\boldsymbol{E})$，$\boldsymbol{V}$ 表示每个样本点，$\boldsymbol{E}$ 代表样本点之

* **收稿日期**:2016-05-04　**网络出版时间**:2017-03-13 11:07
资助项目:重庆市自然科学基金(No. cstc2014jcyjA40011)；重庆市教委科学技术研究项目(No. KJ1400513；No. KJ1400519)
第一作者简介:刘建峰，男，研究方向为机器学习、数据挖掘，E-mail：397272175@qq.com；**通信作者**:吕佳，教授，E-mail:lvjiacqnu@126.com
网络出版地址:http://kns.cnki.net/kcms/detail/50.1165.N.20170313.1108.040.html

间的边，边 $e\in \boldsymbol{E}$ 由 $w(e)$ 确定，$w(e)=w_{ij}$，w_{ij} 表示样本点 x_i 和 x_j 之间的相似性，在这里需要指出的是 $w_{ij}=w_{ji}$。此时数据图 G 的权重矩阵 $\boldsymbol{W}$ 可以通过以下方式计算得到。$W_{ij}=\begin{cases}w_{ij}, \text{如果 } e(i,j)\in \boldsymbol{E}\\ 0, \text{否则}\end{cases}$。不难知道，$\boldsymbol{W}$ 是一个对称矩阵。$D_{ij}=\sum_i W_{ij}$，$\boldsymbol{D}$ 是对角矩阵。因此图的拉普拉斯可以得到以下形式：$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W}$，$\boldsymbol{L}$ 的归一化形式是 $\boldsymbol{L}=\boldsymbol{I}-\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$，其中 $\boldsymbol{I}$ 是单位矩阵。这类方法主要有两个方面的缺点：

(a) 只考虑了样本点之间的高斯距离，而忽视了类内的数据结构。

(b) 分类的结果对参数 K 和 σ 非常敏感。当标记样本很少时，没有足够的假设供模型选择[9]。

Cheng 等人[10]和 Yan 等人[11]提出了 L_1 图构造方法的两个基本假设：线性性和稀疏性。线性性：任何一个样本都可由同类的其他所有样本线性表示出来。稀疏性：给定一个样本 x_i，当所有基向量与 x_i 来自同一类时，可以获得该样本的稀疏表达。

这两个性质符号化如下：T_k 表示除 x_k 之外的列样本矩阵。$\tilde{a}$ 表示稀疏分解系数。其中 $T_k=[x_1,\cdots,x_{k-1},x_{k+1},\cdots,x_n]$，$\tilde{a}=[\tilde{a}_1,\cdots,\tilde{a}_{k-1},\tilde{a}_{k+1},\cdots,\tilde{a}_n]^{\mathrm{T}}$。

通过上述稀疏性的两个基本假设可以得到 $\tilde{a}_k$，并且 $\tilde{a}_k$ 可以通过以下优化问题进行求解：

$$\begin{aligned}&\min \|a\|_0\\ &\text{s.t. } T_k a=x_k,\end{aligned}\tag{1}$$

其中 $\|a\|_0$ 是 L_0 模的向量，这是一个 NP-困难问题，很难得到它的近似解。因此，往往转化为 L_1 模问题。考虑到数据中可能存在干扰噪声问题和避免解不适定问题，这里提出化 L_0 模问题为 L_1 模问题。即求解以下最优化问题：

$$\begin{aligned}&\min \|q\|_1\\ &\text{s.t. } \boldsymbol{P}q=x_k,\\ &\quad q_i\geqslant 0, i=1,2,\cdots,k-1,k+1,\cdots,n,\end{aligned}\tag{2}$$

其中 $\boldsymbol{P}=[G_k \quad I_d]\in \mathbf{R}^{d\times(d+n-1)}$，$q=[a \quad e]^{\mathrm{T}}$。这个问题可以通过线性规划问题进行求解[12]。对于每一个训练样本 $x_k\in X$，求解(1)式。此时权重矩阵 $\boldsymbol{W}=(W_{ij})_{n\times n}$ 表示如下：

$$W_{kj}=\tilde{a}_j, j=1,2,\cdots,k-1,k+1,\cdots,n。\tag{3}$$

注意，通过(3)式不难发现权重矩阵 $\boldsymbol{W}$ 并不一定是对称的，因此 $W_{ij}\neq W_{ji}$，下面定理 1 给出详细的证明。

定理 1　若 $\boldsymbol{X}$ 是 $n\times m$ 矩阵，其中 n 表示数据集样本点的个数，m 表示每个样本点具的属性个数，则权重矩阵 $\boldsymbol{W}$ 为非对称阵。

证明　根据线性规划理论求解(3)式结果为：$w_{ii}=0, i=1,2,\cdots,n$；$w_{12}=a_2$；$w_{13}=a_3$；$\cdots$；$w_{1n}=a_n$；$w_{21}=a_1^{(2)}$；$w_{23}=a_3^{(2)}$；$\cdots$；$w_{2n}=a_n^{(2)}$；$w_{n1}=a_1^{(n)}$；$w_{n2}=a_2^{(n)}$；$\cdots$；$w_{(n-1)\times 1}=a_{n-1}{}^{(n)}$。

以上 $w_{ii}=0, i=1,2,\cdots,n$，组成矩阵 $\boldsymbol{W}=\begin{bmatrix}w_{11} & w_{12} & \cdots & w_{1\times(n-1)} & w_{1\times n}\\ w_{21} & w_{22} & \cdots & w_{2\times(n-1)} & w_{2\times n}\\ \cdots & \cdots & \cdots & \cdots & \cdots\\ w_{n1} & w_{n2} & \cdots & w_{n\times(n-1)} & w_{n\times n}\end{bmatrix}$。矩阵中每一个元素都是通过求解线性规划问题得到，在这里 $w_{ij}\neq w_{ji}$。非对称稀疏图不仅给出了每个样本点的稀疏表达，而且给出了样本点之间的表示关系，即它们之间的位置矩阵，为稀疏矩阵 $\begin{bmatrix}0 & a_2 & \cdots & a_n\\ a_1^{(2)} & 0 & a_3^{(2)} & a_n^{(2)}\\ \vdots & \vdots & & \vdots\\ a_1^{(n)} & a_2^{(n)} & \cdots & 0\end{bmatrix}$。因此，可得 $\boldsymbol{W}$ 为非对称矩阵。

证毕

当需要对某个类别的样本数据进行分类时，即用训练样本的线性组合来描述此样本的类别，对该未知类别的样本，最合适最准确的表示必是稀疏的[13]。另外，计算时大多数系数都是零，这样计算速度也会更快。

在一个足够大的训练样本空间内，构建如上所述稀疏矩阵 $\boldsymbol{W}$，$\boldsymbol{W}$ 矩阵直接关联拉普拉斯项 $\boldsymbol{L}$ 的计算，如(6)式所示。对于一个未知类别的样本数据，按照稀疏性的核心思想，均可以由训练样本中同类数据样本子空间大致的线性表示，当该物体由整个样本空间表示时，它表示的系数是稀疏的。它实际上是把很多不同类别的对象放入训练集中，此时计算式子形如 $\boldsymbol{y}=\boldsymbol{A}\boldsymbol{x}$ 中向量 $\boldsymbol{x}$ 的值。其中 $\boldsymbol{y}$ 是目标数据样本，$\boldsymbol{A}$ 是训练样本空间。

从计算的角度来看,分类的最终依据是目标函数 $\boldsymbol{F}$ 的实数值,通过最优化计算对(4)式进行求解计算得到 $\boldsymbol{F}$。实验证明,$\boldsymbol{F}$ 的求解过程中稀疏非对称阵 $\boldsymbol{W}$ 直接影响样本的分类准确率。

2 非对称稀疏图的半监督学习

2.1 算法描述

基于图的半监督学习目标函数通常是由两项组成,本研究亦采用由损失函数和正则项组成的目标函数,表达式如下:

$$\arg\min \frac{1}{l}\sum_{i=1}^{l}C(x_i,y_i,\boldsymbol{f})+\gamma_A\|\boldsymbol{f}\|^2+\gamma_I\|\boldsymbol{f}\|_I^2, \tag{4}$$

其中 C 是损失函数,第二项 $\|\boldsymbol{f}\|^2$ 是 RKHS 空间的惩罚项,限制目标函数的复杂性,第3项 $\|\boldsymbol{f}\|_I^2$,表示数据的内部流形特征,约束有标记和无标记的正则项,保证图上数据分布的光滑性,最后两项通常称为局部和全局一致性的正则项。第3项表达式为 $\|\boldsymbol{f}\|_I^2=\frac{1}{l+u}f^{\mathrm{T}}\boldsymbol{L}f$,其中 $\boldsymbol{f}=[f(x_1),\cdots,f(x_{l+u})]$,$\boldsymbol{f}$ 是对应于 $x_1,\cdots,x_n$ 实值向量。(4)式的优化问题通过求解可以得到如下表达式:

$$\min_{F}((\boldsymbol{F}-\boldsymbol{Y})^{\mathrm{T}}\boldsymbol{D}(\boldsymbol{F}-\boldsymbol{Y})+\frac{\gamma_1}{n^2}\boldsymbol{F}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{F}+\boldsymbol{F}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{F}), \tag{5}$$

其中 $\boldsymbol{L}$ 是拉普拉斯项,本研究采用稀疏权重的表示 $\boldsymbol{L}$ 计算如下:

$$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W}\in\mathbf{R}^{n\times n}。 \tag{6}$$

w_{ij} 通过(3)式求解得到,因此二次规划问题(5)式可以表示为下式:

$$\boldsymbol{F}=((\boldsymbol{I}-\boldsymbol{A})^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{A})+\frac{\gamma_1}{n^2}(\boldsymbol{D}-\boldsymbol{W})+\boldsymbol{D})^{-1}\boldsymbol{D}\boldsymbol{Y}。 \tag{7}$$

Wu 等人[14]在文献中推导了正则项矩阵 $\boldsymbol{A}$ 的计算方法,$\boldsymbol{A}$ 可由 $\boldsymbol{P}$ 和 $\boldsymbol{Q}$ 表示,如下:

$$\boldsymbol{A}=\boldsymbol{P}-\boldsymbol{P}\boldsymbol{Q}+\boldsymbol{Q}, \tag{8}$$

$$\boldsymbol{P}=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}(\lambda\boldsymbol{I}+\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}, \tag{9}$$

$$\boldsymbol{Q}=1\,\frac{1^{\mathrm{T}}-1^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}(\lambda\boldsymbol{I}+\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}}{n-1^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}(\lambda\boldsymbol{I}+\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}}。 \tag{10}$$

符号函数定义如(11)式所示:

$$\mathrm{sign}(f)=\begin{cases}+1, & f\geqslant 0\\ -1, & f<0\end{cases}。 \tag{11}$$

2.2 算法的执行过程

非对称稀疏图的半监督学习算法框架如下:输入:训练集 $\boldsymbol{X}$,测试集 $\widetilde{\boldsymbol{X}}$,参数 λ,δ,γ;输出:测试集的标记和分类正确率;步骤1,输入 $\boldsymbol{X}$,通过求解(2)式,得到最优解;步骤2,计算(3)式得非对称的权重矩阵 $\boldsymbol{W}$;步骤3,利用(6)式计算 $\boldsymbol{L}$;步骤4,通过计算(8),(9) 和(10)得到矩阵 $\boldsymbol{A}$;步骤5,计算(7)式得到样本的实值函数值;步骤6,通过符号函数(11)式, 得到样本的类标记;步骤7,计算分类的正确率 accuracy。

3 实验

实验过程中,本研究利用机器学习算法对比实验中常采用的 UCI 标准数据集,分别是 adult,abalone,student,yeast,Breast 和 Credit-rating,在表1中给出了它们属性的描述,使用本研究算法和改进加权贝叶斯分类算法(WB),基于局部和全局的半监督学习算法(LGS)和基于信息熵的主动选择算法(EAS),进行对比。

表1 数据集描述

Tab.1 Descriptions of date sets

数据集	数据集规模	数据集特征
adult	10 391	14
abalone	4 177	7
student	649	33
yeast	1 484	8
Breast	699	10
Credit-rating	690	15

3.1 数据集准备

实验中采用6个标准数据集,数据集详细描述表1所示[15-18]:

3.2 分类正确率

分类算法的优劣通常是以分类精度来度量的,在实验的过程中对于参数的控制,本研究采用留一法,针对每一组参数,每次将第 i 个有标记的样本改为无标记的样本后再对训练集进行训练,然后判

断第 i 个样本是否被错分，最后选择错分最少的参数为最优参数。

本研究算法参数设置如下：每个训练集随机选取 10%作为有标记样本，其他为无标记样本；$D_l \in \{0.1, 0.5, 1, 10, 50, 100\}$；$K \in \{5, 20, 50, 100, 200\}$；$\gamma \in \{2^{-6}, 2^{-5}, 2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 2^{0}, 2^{1}, 2^{2}, 2^{3}, 2^{4}, 2^{5}, 2^{6}\}$；$\delta \in \{2^{-6}, 2^{-5}, 2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 2^{0}, 2^{1}, 2^{2}, 2^{3}, 2^{4}, 2^{5}, 2^{6}\}$；$\lambda \in \{0.1, 0.5, 1, 10, 100\}$。

实验重复 20 次，得到平均分类正确率和标准误差，如表 2 所示。从平均正确率和标准误差两项评价标准来看，本研究提出的非对称稀疏图的半监督学习算法表现出了一定的优势和较好的稳定性。

表 2　分类正确率和标准误差

Tab. 2　Classification accuracy and standard error

数据集	WB		LGS		EAS		NSSS	
	CR	*SE*	*CR*	*SE*	*CR*	*SE*	*CR*	*SE*
credit-rating	84.01	±1.71	85.68	±1.24	83.85	±1.02	85.70	±1.62
adult	78.56	±2.10	84.73	±1.23	82.46	±0.89	83.12	±2.05
abalone	81.42	±0.52	87.25	±0.89	83.78	±1.66	84.39	±1.48
student	79.90	±1.51	80.80	±2.56	79.80	±2.40	79.46	±2.19
yeast	89.26	±2.13	89.51	±1.83	90.13	±1.03	91.56	±1.67
Breast	90.02	±0.24	97.89	±1.01	96.05	±2.29	98.11	±0.45

同时，在实验的过程中发现，标记样本的个数往往会影响分类的正确率和标准误差。当标记样本数目占训练集总数的 10%，20%，30%，40%，50%，60%，70%，80%，90%时，分类正确率的变化如图 1 所示，由图 1 可知随着标记数目的增多，分类正确率也随着提高，但也具有峰值，这正验证了标记数目并不是越多对分类器越有利，相反由于标记样本的类标记含有的噪声对分类器有负面的影响，往往会造成分类器性能随着标记数目的增多而下降。参数 λ 的选择对算法的影响如图 2 所示，不难发现参数的调节在算法的执行过程中有较大的影响，并且具有最大值。在实验的过程中本研究选择的参数是经过反复验证，最后得到一个对分类影响较小，同时又使实验结果最好的值。

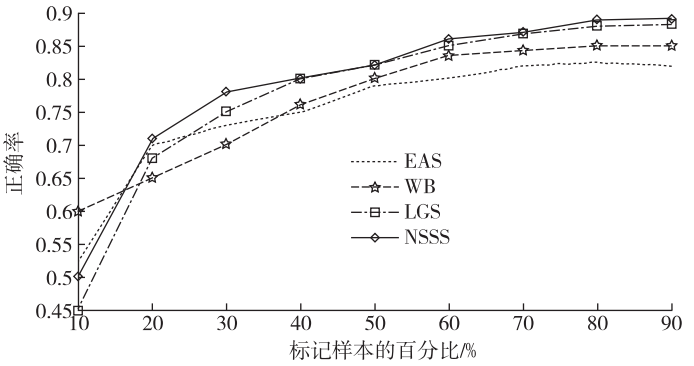

图 1　正确率随标记样本百分比的变化

Fig. 1　The change of accuracy with the percentage of labeled samples

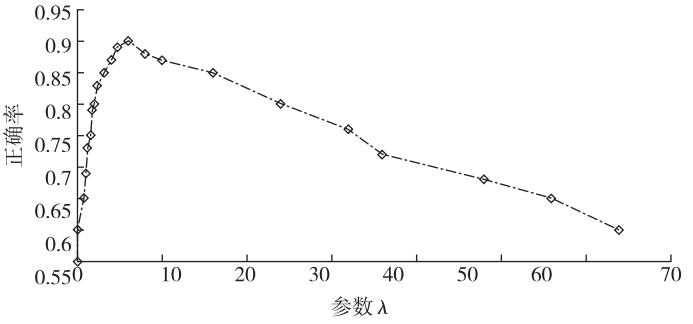

图 2　参数对正确率的影响

Fig. 2　The influence of parameter on the correct rate

4 结语

提出了非对称稀疏图的半监督学习算法，通过构建稀疏图和解线性规划问题得到非对称稀疏的权重矩阵，在此基础上形成了非对称稀疏图的半监督学习，并在 6 个标准数据集上和其他 3 个算法进行对比，实验结果显示，本研究算法在分类正确率和标准误差两项评价指标中取得了较好的效果。将半监督核引入半监督学习、非对称稀疏图上的核函数研究和无参数的学习问题是将是下一步要开展的研究工作。

参考文献：

[1] 曾令伟，伍振兴，杜文才. 基于改进自监督学习群体智能(ISLCI)的高性能聚类算法[J]. 重庆邮电大学学报(自然科学版)，2016，28(1)：131-137.
ZENG L W，WU Z X，DU W C. Improved self-supervised learning collection intelligence based high performance data clustering approach[J]. Journal of Chongqing University of

Posts and Telecommunications (Natural Science Edition), 2016,28(1):131-137.

[2] ZHU X. Semi-supervised learning literature survey [J]. Computer Sciences Technical Report,2008,7(1):1530-1542.

[3] BLUM A, CHAWLA S. Learning from labeled and unlabeled data using graph mincuts[J]. 18th International Conference on Machine Learning,2011.

[4] SAVELLI F, KNIERIM J. Hebbian analysis of the transformation of medial entorhinal grid-cell inputs to hippocampal place fields[J]. Journal of Neurophysiology,2010, 103(6):3167-3183.

[5] ZHOU D, BOUSQUET O, LAL T N, et al. Learning with local and global consistency[J]. Advances in Neural Information Processing System,2004,16(1):40-52.

[6] TOMIOKA R, SUZUKI T. Super-linear convergence of dual augmented Lagrangian algorithm for sparsity regularized estimation[J]. Journal of Machine Learning Research, 2011, 12(1):1537-1586.

[7] ROWEIS S T, SAUL L K. Nonlinear dimensionality reduction by locally linear embedding[J]. Science, 2000, 29(1): 2323-2326.

[8] CHENG B, YANG J, YAN S C. Learning with-graph for image analysis[J]. IEEE Trans Image Processing, 2010, 19 (2):858-866.

[9] WANG F, ZHANG C S. Label propagation through linear neighborhoods [J]. IEEE Transaction on Knowledge and Data Engineering, 2008, 20(1):55-67.

[10] CHENG H, LIU Z C, YANG J. Sparsity induced similarity measure for label propagation[C]// IEEE international conference on computer vision. Kyoto, Japan: IEEE, 2009, 1(1):317-324.

[11] YAN S C, WANG H. Semi-supervised learning by sparse representation[C]//SIAM international conference on data mining. Sparks, Nevada, USA: IEEE, 2009, 2(2): 792-801.

[12] DONOHO D, TSAIG Y. Fast solution of l1-norm minimization problems when the solution may be sparse [J]. IEEE Transactions on Information Theory, 2008, 54(11): 4789-4812.

[13] 施志刚,蒋玲. 一种基于近邻稀疏表示的人脸识别新方法[J]. 重庆师范大学学报(自然科学版),2016,33(6):106-113.
SHI Z G, JIANG L. A new method based on the nearest neighbor sparse representation in face recognition [J]. Journal of Chongqing Normal University (Natural Science), 2016, 33(6):106-113.

[14] WU M, SCHOLKOPF B. Transductive classification via local learning regularization[J]. Artificial Intelligence and Statistics, 2007, 5(1):628-635.

[15] CORTEZ P, SILVA A. Using data mining to predict secondary school student performance[J]. Proceedings of 5th Future Business Technology Conference, 2008 (1): 978-813.

[16] PAN X L, LUO Y, XU Y. K-nearest neighbor based structural twin support vector machine[J]. Knowledge-Based Systems, 2015, 88(1):34-44.

[17] SONG L, SMOLA A, GRETTON A. Feature selection via dependence maximization[J]. The Journal of Machine Learning Research, 2012, 13(3):1393-1434.

[18] ZHANG J, YANG J. Linear reconstruction measure steered nearest neighbor classification framework [J]. Pattern Recognition, 2014, 47(2):1709-1720.

Research on Semi-supervised Learning via Non-symmetric Sparse Graph

LIU Jianfeng[1], LÜ Jia[2]

(1. Office of Academic Affairs, Chongqing Three Gorges University, Wanzhou Chongqing 404100;
2. College of Computer and Information Science, Chongqing Normal University, Chongqing 401331, China)

Abstract: [**Purposes**]How to construct an effective data graph, which is an important research direction in semi-supervised learning. In order to study the relationship between data samples, and improve the performance of semi-supervised learning algorithm based on graph. [**Methods**]Using the sparse representation of the data, we construct non-symmetric sparse graph of the data samples, and experiment with the semi-supervised learning on the standard data sets. [**Findings**]In the framework of semi-supervised learning, we establish the distance relationship, the internal structure and the sparse representation of the data between heterogeneous data and similar data, and then the graph is constructed. [**Conclusions**]Experiments on standard datasets show that the non-symmetric sparse graphs can make use of the characteristic of semi-supervised learning, and data samples are classified effectively.

Keywords: data graph; semi-supervised learning; sparse representation; non-symmetric graph

(责任编辑 游中胜)