文章编号: 1001-0920(2011)11-1726-05

一种简洁局部全局一致性学习

王雪松, 张晓丽, 程玉虎

(中国矿业大学 信息与电气工程学院，江苏 徐州 221116)

摘 要: 针对局部全局一致性学习 (LLGC) 算法的分类精度在很大程度上取决于控制参数的合理设置问题, 提出一种少参数的简洁局部全局一致性学习 (BB-LLGC). 简化图上的目标函数, 使其不受参数 α 的影响. 另外, 在标签传递过程中, 仅将未标记样本的标签根据相似度传递给其近邻, 而将已标记样本的标签强制填回以确保标签传递源头的准确性. UCI 数据集的实验结果表明, 与 LLGC 相比, BB-LLGC 不仅控制参数少、使用简单, 而且分类精度高、收敛速度快.

关键词: 半监督学习；局部全局一致性学习；参数选择；标签传递

中图分类号: TP18 **文献标识码:** A

Barebones learning with local and global consistency

WANG Xue-song, *ZHANG Xiao-li*, *CHENG Yu-hu*

(School of Information and Electrical Engineering，China University of Mining and Technology，Xuzhou 221116，China. Correspondent：WANG Xue-song，E-mail：wangxuesongcumt@163.com)

Abstract: Based on the analysis of the classification accuracy of the learning with local and global consistency(LLGC) algorithm being influenced greatly by a suitable setting of parameters, a kind of barebones LLGC(BB-LLGC) algorithm with less parameters is proposed. The objective function defined on a graph is simplified to make it not be influenced by parameter α. During the label propagation process, only the predicted labels of unlabeled samples are propagated to its neighbors according to a similarity metric, while the labels of labeled samples are kept unchanged so as to ensure the correctness of the source of label propagation. Experimental results concerning on several UCI datasets show that, compared with LLGC, the BB-LLGC has advantages of less control parameters, simple operation procedure, high classification accuracy and fast convergence speed.

Key words: semi-supervised learning; learning with local and global consistency; parameter selection; label propagation

1 引 言

半监督学习在减少人工标注代价和提高学习器性能方面具有突出的优势. 近年来, 随着机器学习理论在数据分析和数据挖掘中的广泛应用, 半监督学习获得了长足的发展, 其研究成果已经广泛应用于网页检索和文本分类、数字图像处理、邮件分类和医学数据处理等领域[1-2]. 现有的半监督算法主要包括: 在原有监督算法基础上改进的自训练算法, 基于产生式模型的算法[3], 基于多视图的算法[4], 半监督支持向量机[5]和基于图的方法[6-9]等.

基于图的半监督学习算法本质上是非参数的、直推的和判别的, 而且这类方法更加直观, 具有更一般的解释性和良好的分类性能, 已成为当前半监督学习研究中的热点. Blum 等人[6]将半监督学习看成是最小图切割问题; Zhu 等人[7]利用高斯随机场和调和函数来解决图上半监督学习问题; Joachims[8]利用谱图分割来解决半监督分类问题. 最具代表性的是 Zhou 等人[9]提出的局部全局一致性学习 (LLGC) 算法, 该算法将优化目标项的权值取值范围约束到有限值, 从而使得算法对错误标注具有一定的容错能力. 但是, LLGC 的算法性能对控制参数 α 的设置比较敏感, 并且在标签传递过程中使得已标记样本的标签随着循环传递而改变, 这便导致标签传递的源头改变, 因而并不能保证良好的分类结果. 为此, 本文提出一种少

收稿日期: 2010-07-20；**修回日期:** 2010-11-01.

基金项目: 国家自然科学基金项目(60804022, 60974050, 61072094)；教育部新世纪优秀人才支持计划项目(NCET-08-0836)；霍英东教育基金会青年教师基金项目(121066)；江苏省自然科学基金项目(BK2008126).

作者简介: 王雪松(1974−), 女, 教授, 博士, 从事机器学习、生物信息学等研究；张晓丽(1984−), 女, 硕士生, 从事半监督学习的研究.

控制参数的简洁局部全局一致性学习 (BB-LLGC) 算法, 将图中每个节点的标签根据其与相邻节点的相似度大小传递给其近邻, 如此将标签反复传递; 同时将已知的标记样本的标签在每次传递过程中都强制填回, 从而保证标签传递源头的准确性, 直至达到一个全局稳定的状态为止. UCI 数据集上的实验结果表明, 与 LLGC 算法相比, 该算法不仅能够获得更高的分类精度, 而且需事先确定的控制参数少、使用简单.

2　局部全局一致性学习算法性能分析

LLGC 属于图上半监督学习算法. 基于图的半监督学习算法的主要算法步骤是[1]: 1) 根据数据之间的相似性关系构造图; 2) 构造图上的一个函数; 3) 优化这个函数, 预测未标记样本的类别.

假设 $X=\{x_{i'}\}_{i'=1}^{n}$ 表示数据集所有样本的集合. 其中: n 表示样本总数; $C=\{c_j\}_{j=1}^{c}$ 表示所有样本的类别标签集合, c_j 表示某一样本的类别, 而 c 表示所有类别的总数. 根据半监督学习的框架, 可以建立已标记数据样本集和未标记数据样本集, 具体如下:

已标记样本集 $X_L=\{(x_1,y_1),\cdots,(x_l,y_l)\}$, 即 X 中的前 l 个样本被标注上标签 $y_i\,(y_i\in C)$, $Y_L=\{y_i\}_{i=1}^{l}$;

未标记样本集 $X_U=\{x_{l+1},\cdots,x_n\}$, 即 X 中剩下的 $n-l$ 个样本作为无标记样本, 且令 $u=n-l$, $l\ll u$.

学习的目标是根据 X 和 Y_L 预测未标记样本集 X_U 的类别标签 Y_U.

在学习过程中, 为了将上述已标记样本信息与未标记样本信息有效地结合, 模型中定义了一个无向加权图 $G=\langle V,E\rangle$. 节点集合 V 代表数据集中各个标记样本点和未标记样本点, 连接任意 2 个节点 $x_{i'}$ 和 $x_{j'}$ 的边 E 的权值 $w_{i'j'}$ 描述了这 2 个样本间的相似度[1]

$$w_{i'j'}=\mathrm{e}^{-\frac{\|x_{i'}-x_{j'}\|^2}{\beta}},\tag{1}$$

其中 β 是一个可调参数, 可以近似设置为所有样本对之间距离的平均值.

综上所述, 基于图的半监督学习算法可看作在图上估计一个标注函数

$$F=\arg\min\Big\{\sum_{i=1}^{l}(f_i-y_i)^2+\frac{1}{2}\sum_{i',j'=1}^{n}w_{i'j'}(f_{i'}-f_{j'})^2\Big\},\tag{2}$$

使得已标记样本都能够得到正确的分类, 并保证相邻样本之间的标签具有足够的相似性, 同时使得标签分布在整个图上, 并具有足够的平滑性[9]. 式 (2) 中: y_i 表示已标记样本的真实标签; $f_{i'}$ 和 $f_{j'}$ 表示预测标签值, 有 $f_{i'}=\arg\max\limits_{j\leqslant c}F_{i'j}$, $F_{i'j}$ 表示预测样本 i' 属于类别 j 的概率, $F=\{F_{i'j}\}_{n\times c}$.

通过对式 (2) 推导, 可得到[9]

$$F^*=(I-\alpha S)^{-1}\tilde{Y}.\tag{3}$$

其中: $S=D^{-1/2}WD^{-1/2}$, $W=\{w_{i'j'}\}_{n\times n}$, D 是一个对角矩阵, 其对角线元素

$$D_{i'i'}=\sum_{j'=1}^{n}w_{i'j'};$$

$$\tilde{Y}_{i'j}=\begin{cases}Y_{ij},\ 1\leqslant i'\leqslant l;\\ 0,\ l+1\leqslant i'\leqslant n;\end{cases}$$

$$Y_{ij}=\begin{cases}1,\ y_i=c_j;\\ 0,\ \text{otherwise};\end{cases}$$

α 是需要事先确定的系数, 且 $\alpha\in(0,1)$.

由式 (3) 可以看出, LLGC 算法在标签的传递过程中引入了参数 α. α 用以表明相对大量的信息是来自于近邻, 还是初始标记样本信息. 一般而言, 如果与近邻关系大, 则 α 取大些; 否则, α 取小值. 参数 α 的选取方法主要有经验选择法和实验试凑法. 经验选择要求对所研究问题拥有很好的经验和十足的知识, 否则不易获得合适的参数. 实验试凑是通过大量的数字仿真实验来获得较优的参数, 比较费时, 而且获得的参数也不一定是最优的. 但是, LLGC 算法对参数 α 的选择比较敏感, 如果选择不当, 则将无法获得良好的分类性能. 下面以“双月形”Toy 数据集为例来说明 α 的选取对算法性能的影响, 如图 1 所示. 初始 Toy 数据集如图 1(a) 所示, 共有两类, 每一类有 100 个样本, 其中每一类只有一个标记样本, 其余均为未标记样本. 期望得到的分类结果如图 1(b) 所示, 上半月为 +1 类, 下半月为 −1 类, 此时 α 为 0.4. 但是, 从图 1(c), 图 1(d) 和图 1(e) 可以看出, 当 α 分别取 0.6,0.8,0.99 时, LLGC 算法都不同程度地将 +1 类样本误分为 −1 类样本; 当 $\alpha=0.99$ 时, +1 类的所有样本均被错误分为 −1 类, 分类正确率仅为 50%.

3　简洁局部全局一致性学习

针对 LLGC 算法存在的参数 α 的选择问题, 对标签传递过程中的目标函数进行修改, 使其不受参数 α 的影响, 并在计算邻接矩阵时利用 k 近邻图代替完全连接图[10]. 这样做的好处是: 1) 用 k 作为限制近邻数的条件, 将不相似节点间的多余连接去掉, 从而增大相似节点间的传播概率; 2) 计算速度快、代价低.

图上半监督学习算法的最终目的是使标签在未标记样本中平滑分布, 即对于任意 $x_{i'}\in X_U$, 其类别 $f_{i'}$ 均应满足目标函数

$$\min\frac{1}{2}\sum_{x_{j'}\in N(x_{i'})}w_{i'j'}(f_{i'}-f_{j'})^2,\tag{4}$$

其中 $N(x_{i'})$ 表示未标记样本 $x_{i'}$ 的 k 个近邻组成的数

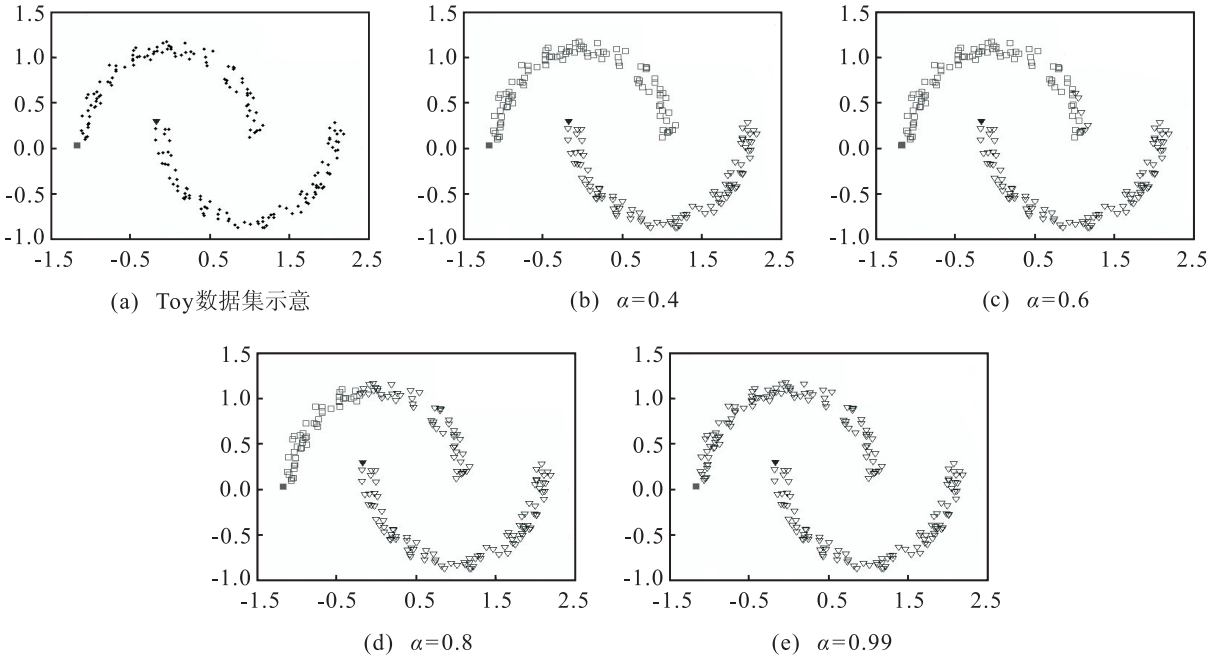


图 1 α 的选取对 Toy 数据集分类结果的影响

据集. 对于已标记样本, 在每一次迭代过程中保持其初始标记不变, 便可保证未标记样本标签以标记样本标签为起点, 逐次传播到其近邻的未标记样本, 直到所有未标记样本类别不再变化为止, 即标签在整个样本集上实现平滑分布.

与局部全局一致性算法一样, 定义一个 $n\times c$ 的非负 F 矩阵来表示每个样本的标注概率. 例如, 对于任意样本 $x_{i'}$, 其对应的 $F_{i'j}$ 表示 $x_{i'}$ 属于第 c_j 类的概率, $x_{i'}$ 的预测标签 $f_{i'}=\arg\max\limits_{j\leqslant c} F_{i'j}$. 初始化阶段, 若 $x_{i'}\in X_L$, 则 F 的第 j 列元素的值定义为

$$F_{i'j}=\begin{cases}1,\ x_{i'}\in c_j;\\ 0,\ x_{i'}\notin c_j.\end{cases}\tag{5}$$

将其记为 $F_L(0)$, 它是一个 $l\times c$ 的矩阵. 若 $x_{i'}\in X_U$, 初始设定 $F_{i'}$ 的每一列元素的值为 0, 记为 $F_U(0)$, 这时它是一个 $(n-l)\times c$ 的矩阵. 在标签传递过程中, 未标记样本 $x_{i'}$ 属于第 c_j 类的概率范围为 $F_{i'j}\in[0,1]$.

简洁局部全局一致性算法步骤如下:

Step 1: 计算所有样本间的欧氏距离 $\mathrm{Dist}_{i'j'}=(\|x_{i'}-x_{j'}\|^2)^{1/2}$, 根据 Dist, 从中选取每个样本的 k 个近邻, 构造 k 近邻图.

Step 2: 计算邻接矩阵 W, 即

$$w_{i'j'}=\begin{cases}\mathrm{e}^{\frac{-\|x_{i'}-x_{j'}\|^2}{\beta}},\ x_{j'}\in N(x_{i'});\\ 0,\ \text{otherwise}.\end{cases}\tag{6}$$

Step 3: 计算正则化的图拉普拉斯矩阵

$$S=D^{-1/2}WD^{-1/2},$$

其中 D 表示度矩阵, 它是一个对角矩阵, 其对角线元素 $D_{i'i'}=\sum\limits_{j'=1}^{n}w_{i'j'}$.

Step 4: 根据 $F(t+1)=SF(t)$, 更新样本点的标签概率:

Step 4.1: $t=0$, $F(0)=[F_L(0);F_U(0)]$;

Step 4.2: $t=t+1$, $F(t+1)=SF(t)$;

Step 4.3: 限制 $F_L(t)=F_L(0)$.

Step 5: 重复 Step 4.2 和 Step 4.3, 直到 F 收敛到一个确定的值 F^* 为止.

Step 6: 为每个未标记样本 $x_{i'}$ 标出标签

$$f_{i'}=\arg\max_{j\leqslant c}F^*_{i'j}.$$

在 Step 4.3 中, 将初始标签信息 F_L 强制填回, 参与下一轮标签传递, 从而确保已标记样本不会随着标签的全局传递而改变.

4 收敛性证明

将图拉普拉斯矩阵 S 分解为四块子矩阵, 即

$$S=\begin{bmatrix}S_{LL} & S_{LU}\\ S_{UL} & S_{UU}\end{bmatrix},$$

则有

$$\begin{bmatrix}F_L(t+1)\\ F_U(t+1)\end{bmatrix}=\begin{bmatrix}S_{LL} & S_{LU}\\ S_{UL} & S_{UU}\end{bmatrix}\begin{bmatrix}F_L(t)\\ F_U(t)\end{bmatrix}.\tag{7}$$

由于 F_L 在迭代过程中保持不变, 即

$$F_L(t+1)=F_L(t)=F_L(0),\tag{8}$$

$$F_U(t+1)=S_{UL}F_L(t)+S_{UU}F_U(t),\tag{9}$$

当 $t\to\infty$ 时, 式 (9) 可表示为

$$F_U(t+1)=$$

$$(S_{UU})^{t+1}F_U(0)+S_{UL}F_L(0)\sum_{r=1}^{t+1}(S_{UU})^{r-1}. \quad (10)$$

由于S的特征值在区间$[-1,1]$内取值, 所以当$t\to\infty$时有

$$\lim_{t\to\infty}(S_{UU})^{t+1}=0, \quad (11)$$

$$\lim_{t\to\infty}\sum_{r=1}^{t+1}(S_{UU})^{r-1}=(I-S_{UU})^{-1}. \quad (12)$$

由式(10)～(12)可得

$$F_U^*=\lim_{t\to\infty}F_U(t+1)=(I-S_{UU})^{-1}S_{UL}F_L(0). \quad (13)$$

由式(13)可以看出, 算法能够收敛到唯一一个确定的值. 在执行算法时, 可不通过迭代而直接计算标注概率矩阵F^*, 而且迭代结果不依赖于$F_U(0)$的值.

5　实验研究

为评价算法的有效性, 选用表1所示的5个UCI数据集作为实验对象. 实验用计算机的硬件配置如下: Pentium CPU主频1.73GHz, 内存1GB, 采用Mtlab7.0.1软件编程.

分别采用监督学习K近邻(KNN), 半监督学习LLGC和简洁LLGC算法解决这5个数据集的分类问题. 此处采用2种类型的简洁LLGC算法, 一种是基于完全连接图的简洁LLGC (BB-LLGC 1)算法, 另一种是基于近邻图的简洁LLGC (BB-LLGC 2)算法. 选用KNN的原因是, 它不需要从训练数据中通过学习得到模型, 是一种简单直接的非参数监督学习算法, 常被用于半监督学习的对比算法.

表 1　数据集信息

数据集	样本总数	特征维数	类别数
Iirs	150	4	3
Ionosphere	351	34	2
Vowel	990	10	11
Semeion handwritten digits	1593	256	10
Image segmentation	2310	19	7

对于不同的数据集, 各算法的控制参数设置情况如表2所示, 其中Semeion和Image分别表示Semeion handwritten digits和Image segmentation数据集. 在进行参数选取时, 需注意以下2点: 1) 为了比较的公平性, 各算法的共有参数(如β), 均设为一致; 2) 为获得较好的分类性能, 将KNN算法中的近邻数K取为1, 通过反复实验试凑对LLGC中的α和BB-LLGC 2中的k进行选择. 图2给出了不同α取值情况下, LLGC算法在Ionosphere数据集上的分类准确率. 由图2可以看出, 当$\alpha=0.4$时, LLGC的分类性能最佳, 因此这里α取为0.4.

表 2　控制参数设置情况

数据集	KNN K	LLGC		BB-LLGC 1 β	BB-LLGC 2	
		α	β		k	β
Iirs	1	0.6	180.5	180.5	4	180.5
Ionosphere	1	0.4	3 125	3 125	9	3 125
Vowel	1	0.99	0.5	0.5	30	0.5
Semeion	1	0.99	4.5	4.5	50	4.5
Image	1	0.99	512	512	20	512

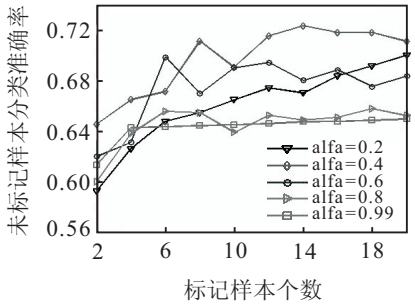
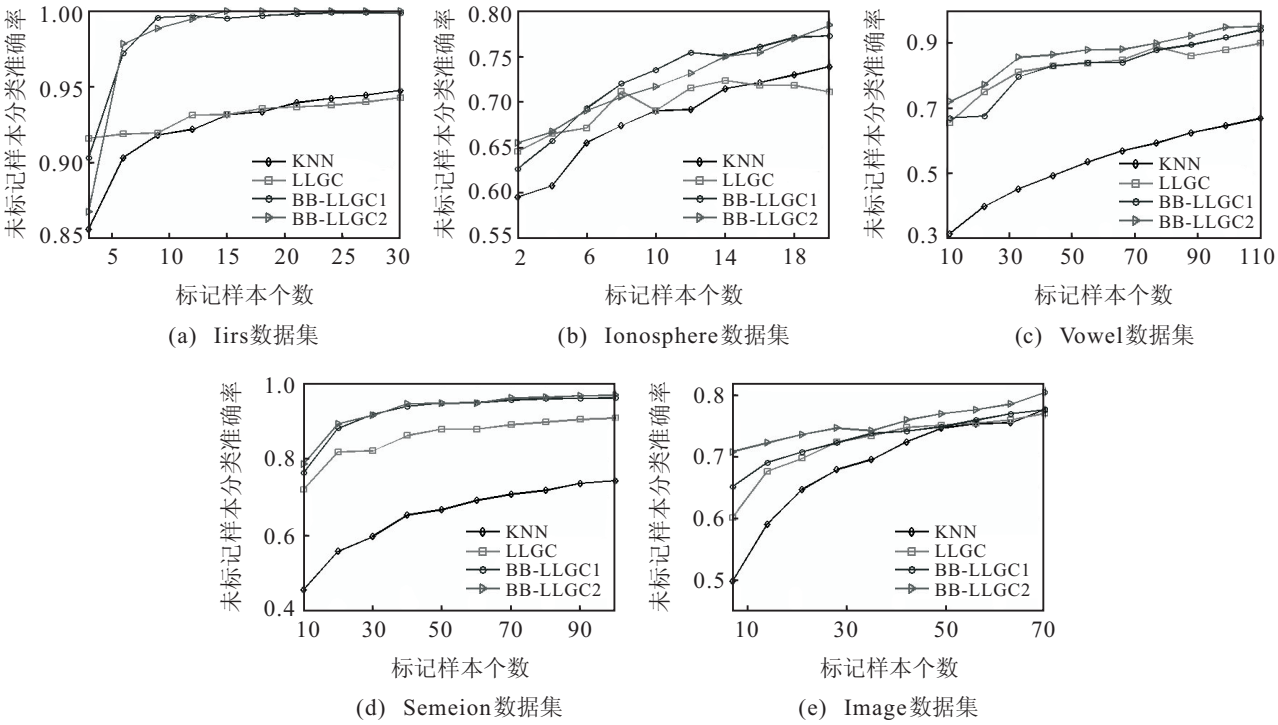


图 2　α 对 LLGC 分类结果的影响 (Ionosphere)

对于每个实验数据集, 随机抽取l个样本组成已标记样本集(限定已标记样本集中的每一类样本标记个数相等), 剩下的$n-l$个样本组成未标记样本集. 以图3(a)所示的Iris数据集为例, 标记样本的个数分别取3, 6, 9, 12, 15, 18, 21, 24, 27和30, 其中'3'意味着Iris数据集中每一类样本仅有1个为已标记样本(由表1可知, Iris数据集的类别数为3). 独立重复上述样本选择过程50次, 作为50次随机实验的输入数据集, 分别采用KNN, LLGC, BB-LLGC 1和BB-LLGC 2算法对其进行分类, 各数据集50次实验的平均分类准确率比较如图3所示. 图3所示的实验结果显示, 对于所有实验数据集, 当已标记样本较少时, 半监督学习算法的分类性能均优于监督学习KNN. 对于半监督学习算法, 当标记样本数量达到一定程度时, 其分类准确率将不再有明显改进.

为评价算法的收敛速度, 以Semeion数据集为例, 表3给出了各半监督分类算法对目标函数评价次数的对比结果. 因为监督型KNN算法主要依靠周围有限的K个已标记近邻样本的标签来直接确定其所属的类别, 因此其在分类过程中不需要优化如式(4)所示的目标函数, 故不存在评价次数的统计. 结合图3和表3可以看出, 因为BB-LLGC使已标记样本的标签在标签传递过程中确保不变, 因此其分类准确率高于传统LLGC且收敛速度快; 由于BB-LLGC 2采用k近邻图代替了BB-LLGC 1中的完全邻接图, BB-LLGC 2的分类精度和收敛指标均略优于BB-LLGC1, 但是BB-LLGC 2比BB-LLGC 1多了一个需要事先给定的控制参数k, 因此在难以确定k取值的情况下, 可考虑采用BB-LLGC1.



(a) Iirs数据集 (b) Ionosphere数据集 (c) Vowel数据集

(d) Semeion数据集 (e) Image数据集

图 3 分类准确率对比

表 3 目标函数评价次数比较(Semeion 数据集)

标记样本个数	LLGC	BB-LLGC 1	BB-LLGC 2
10	66	60	57
20	58	49	46
30	60	46	41
40	50	45	42
50	70	42	38
60	63	54	40
70	48	44	36
80	59	42	38
90	63	39	34
100	62	39	32

6 结 论

局部全局一致性学习是一种典型的基于图的半监督学习算法, 其分类精度对控制参数α的设置比较敏感. 另外, 在实际使用过程中, 合理选择该参数比较困难或需要很大代价. 为此, 本文提出了一种少参数的简洁局部全局一致性学习算法. 首先, 利用图论策略建立数据集的图模型, 将每个标记样本点和未标记样本点作为图的顶点, 样本之间的相似度由连接两点之间边的权值来反映; 然后, 在图上构造一个简化的目标函数, 使其不受参数α的影响; 最后, 遵循一种简单的标签循环传递过程, 仅将未标记样本点的标签根据其与近邻节点的相似度传递给其近邻, 同时将已标记样本的标签在每次传递过程中都强制填回, 直到达到一个全局稳定的状态为止. 从UCI数据集分类实验结果看, 与监督学习K近邻和传统的半监督学习LLGC相比, 该算法性能较优, 是一种非常有效的半监督分类算法.

参考文献(References)

[1] Chapelle O, Scholkopf B. Semi-supervised learning[M]. Cambridge: MIT Press, 2006: 193-196

[2] Casamayor A, Godoy D, Campo M. Identification of non-functional requirements in textual specifications: A semi-supervised learning approach[J]. Information and Software Technology, 2010, 52(4): 436-445.

[3] Nigam K, Mccallum A, Thrun S, et al. Text classification from labeled and unlabeled documents using EM[J]. Machine Learning, 2000, 39(2/3): 103-134.

[4] Maillard O A, Vayatis N. Complexity versus agreement for many views: Co-regularization for multi-view semi-supervised learning[J]. Lecture Notes in Computer Science, 2009, 5809: 232-246.

[5] Chapelle O, Sindhwani V, Keerthi S S. Optimization techniques for semi-supervised support vector machines[J]. J of Machine Learning Research, 2008, 9(2): 203-233.

[6] Blum A, Chawla S. Learning from labeled and unlabeled data using graph mincuts[C]. Proc of the 18th Int Conf on Machine Learning. San Fransisco: Morgan Kaufmann Publishers Inc, 2001: 19-26.

[7] Zhu X J, Ghahramani Z, Lafferty J. Semi-supervised learning using gaussian fields and harmonic functions[C]. Proc of the 20th Int Conf on Machine Learning. Washington: AAAI Press, 2003: 912-919.

[8] Joachims T. Transductive learning via spectral graph partitioning[C]. Proc of the 20th Int Conf on Machine Learning. Washington: AAAI Press, 2003: 290-297.

(下转第1734页)