第52卷第2期 大连理工大学学报 Vol. 52, No. 2
2012年3月 Journal of Dalian University of Technology Mar. 2012

电子与信息工程、管理工程

文章编号：1000-8608(2012)02-0264-08

基于非负稀疏表示的标签繁殖算法

杨南海*[1]， 桑媛媛[2]， 赫 然[2]， 王秀坤[1,2]

（1.大连理工大学 软件学院，辽宁 大连 116620；
2.大连理工大学 计算机科学与技术学院，辽宁 大连 116024）

摘要： 提出了一种基于非负稀疏表示(nonnegative sparse representation，NSR)的半监督学习标签传播算法.该算法首先构造一个稀疏概率图(sparse probability graph，SPG)，其权重由非负稀疏表示算法计算的非负系数组成，自然地反映了各样本之间的聚类关系，避免了传统半监督学习算法中的邻居选择和参数设置过程；然后通过对未标记样本的标签进行迭代繁殖至收敛而获得所有样本的标签.在人脸识别、物体识别、UCI机器学习和TDT文本数据集上的实验结果表明采用非负稀疏表示的标签传播算法比典型的标签繁殖算法具有更好的分类准确率.

关键词： 非负稀疏表示；半监督学习；稀疏概率图；聚类关系；标签繁殖
中图分类号： TP181 **文献标志码：** A

0 引 言

在模式分类和数据挖掘任务中，人们往往面临有标记数据缺乏的问题.获取有标记数据通常需要付出较大的人力和时间成本，而获得无标记数据却非常容易，传统的监督学习只利用了标记数据中的信息，而无监督学习则只利用了无标记数据中的信息，为利用标记数据的同时有效利用无标记数据中的信息，提出了半监督学习算法.

与传统机器学习算法相比，半监督学习算法通过同时利用已标记样本的信息和未标记样本的信息来提高学习算法的性能.半监督学习算法中，基于图的半监督学习算法在最近几年备受关注，它将整个数据集用一个图 $G=(V;E)$ 来代表，其中 V 表示图的节点集合，每个节点代表数据集中的一个实例，E 表示节点之间的边的集合，每条边带有相连节点的关联权重.基于图的半监督学习算法的依据是聚类假设：(1)距离相近的点可能具有相同的标签.(2) 相同结构的点(如一个聚类或一个流形)可能具有相同的标签.前一个假设是局部性的，后一个假设则带有全局性. 聚类假设表明在图构建的过程中既要充分利用数据集所包含的局部信息，还要充分利用其中的全局信息.

基于图的半监督学习算法研究中，Wang 等提出了标签的线性邻居繁殖(linear neighborhood propagation，LNP)算法[1]. 该算法在图的构建中，假定每个样本点可以用其邻居进行线性重构，将样本点用其邻居线性表示并令重构误差最小，从而求解多个最小化约束二次规划问题得到图的权重矩阵.LNP 算法利用已构建好的图通过邻居进行迭代标签繁殖，直至标签收敛即可获得所有样本点的标签.LNP 算法还提出了一种排除数据集中桥点和噪声的数据预处理方法.LNP 算法的优点是：与传统算法相比具有更强的参数稳定性.其缺点是：首先，该算法仍然需要手动设置近邻参数 k，因而对参数有一定的敏感性，一定程度上影响了图的结构和算法的鲁棒性.其次，该算法只利用了样本的局部信息，不能更全面地反映样本间

收稿日期： 2009-12-04； **修回日期：** 2012-02-04.
基金项目： 高等学校博士学科点专项科研基金资助项目(20100041120009)；国家自然科学基金资助项目(60873054)；大连理工大学引进人才启动经费资助项目.
作者简介： 杨南海*(1970-)，男，博士生，讲师，E-mail：nanhai@dlut.edu.cn；王秀坤(1945-)，女，教授，博士生导师.

的聚类关系，从而影响了机器学习的性能和效率. 另外，LNP 算法排除桥点和降低噪声的策略采用启发式的思想，缺乏理论依据. Yan 等[2,3]提出了 l^1-graph 算法用于半监督学习. 该方法通过 l^1 最优化问题建立权重矩阵，该矩阵包括正的和负的权重系数. Belkin 等提出了高维数据的低维嵌入表示[4].

稀疏表示框架[3]近年来在信号分析、图像处理和机器学习等领域受到广泛关注. Donoho 提出了基于 l^1 最优化问题的稀疏信号表示[5]，He 等提出了基于相关熵稀疏表示的人脸识别算法[6]. 受此启发，本文提出一种基于 NSR 和鲁棒 NSR 的标签传播算法用于半监督学习. 算法通过非负稀疏表示以无参方式构造一个稀疏概率图(sparse probability graph，SPG). 然后在该图上对未标记样本的标签进行迭代繁殖直至收敛，从而得到未标记样本的标签. 将所提出的算法与 l^1-graph 算法[2]、LNP 算法[1]和 KNN 算法[7]在 ORL、COIL、Aust、Heart 及 TDT 数据集上进行比较实验并分析结果，以评价所提算法的有效性.

1　传统基于图的半监督学习算法

基于图的半监督学习问题中，训练样本数据集表示为 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$，其中 $\boldsymbol{x}_i\in\boldsymbol{R}^d$，$n$ 为训练样本总数. 在基于图的学习算法中，图的构建是最核心的问题. 传统方法中，将图构建过程分为两个步骤：首先选择邻居；然后计算图的权重矩阵.

常用的邻居选择方法有两种[2]：ε-ball 最近邻和 k-最近邻. 在这两种方法中，通常采用欧几里得距离来度量样本点 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 之间的距离. 因此，两种方法定义的都是无向图，两者都需要人为设置参数，参数的大小往往对分类的准确性产生很大的影响. ε-ball 最近邻的优点是容易揭示样本数据间的几何关系，缺点是不能保证图的连通性，不合理的参数设置往往形成很多独立的子图. k-最近邻的优点是容易保证图的连通性，缺点是不能很好地揭示样本数据间的几何关系.

图的权重计算有下面 3 种方法：

(1)径向基核函数

$$W_{ij}=\begin{cases}e^{-\|x_i-x_j\|^2/2\sigma^2}; & i\text{ 和 }j\text{ 是邻居}\\ 0; & i\text{ 和 }j\text{ 不是邻居}\end{cases}\tag{1}$$

(2) 逆欧几里得距离[8]

$$W_{ij}=\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^{-1}$$

其中 $\|\cdot\|$ 表示欧几里得距离.

(3) 局部线性嵌入[9]

Roweis 等提出样例可以通过其邻居进行局部线性重构，重构损失函数由式(2) 定义. 通过添加一个归一化约束条件并求解重构损失的最小约束优化问题，即可以获得图的权重矩阵.

$$\begin{aligned}&\xi(\boldsymbol{W})=\sum_i\left\|\boldsymbol{x}_i-\sum_j W_{ij}\boldsymbol{x}_j\right\|^2\\ &\text{s. t.}\quad\sum_j W_{ij}=1\end{aligned}\tag{2}$$

式中：对于任意的 i，如果 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 不是邻居，则 $W_{ij}=0$. 文献[1]提出的 LNP 算法即采用了 k-最近邻方法选择邻居，然后对式(2) 增加一个非负约束条件 $W_{ij}\geqslant 0$，并最小化该目标函数进行权重计算，从而得到整个样本空间上的关系图.

在基于图的半监督学习中，选择一个合理、有效的图至关重要. 从机器学习的角度考虑，图的构建应满足 3 个要求[3]：高辨别力、稀疏性和自适应邻居选择. 传统半监督学习方法并不能满足这 3 个要求，所有样本的邻居选择都依赖于同一个参数 k，不能实现样本的自适应邻居选择，因而在邻居的选择过程中即将邻居的结构固定下来了，这在一定程度上限制了利用邻居关系进行的权重计算. 邻居选择和权重计算是相互联系的，传统算法中将这两个过程分离开，不能充分利用数据中包含的信息. 因此研究能同时进行邻居选择和权重计算的无参半监督学习算法显得十分必要.

2　基于 NSR 的标签繁殖算法

2.1　SPG 的构建

近年来，非负稀疏表示[10,11]问题在理论和实际应用方面得到了越来越多的关注. 非负稀疏表示可以用以下 l^0 优化问题描述：

$$\begin{aligned}&\min_{\boldsymbol{w}_i}\|\boldsymbol{w}_i\|_0\\ &\text{s. t.}\quad\boldsymbol{X}_i\boldsymbol{w}_i=\boldsymbol{x}_i;\ w_{ij}\geqslant 0\end{aligned}\tag{3}$$

$\boldsymbol{X}$ 表示有 n 个训练样本的矩阵，$\boldsymbol{X}_i$ 是 $\boldsymbol{X}$ 除去第 i 列 $\boldsymbol{x}_i$ 后的矩阵.

Bruckstein 等[11]和 Donoho 等[10]研究了非负稀疏表示的 l^0 和 l^1 理论，并指出式(3) 的优化问题是一个 NP 困难问题[10]. 但如果解足够稀

疏，则式(3)描述的优化问题的解唯一，且可以通过如下的 l^1 优化问题求解：

$$\begin{aligned}&\min_{\boldsymbol{w}_i}\|\boldsymbol{w}_i\|_1\\&\text{s.t. } \boldsymbol{X}_i\boldsymbol{w}_i=\boldsymbol{x}_i;\ w_{ij}\geqslant 0\end{aligned}\tag{4}$$

采用拉格朗日乘子法将式(4)的第一个约束加到目标函数中，即得式(5)所示的NSR模型，该模型可以用非负最小二乘法进行求解[12].

$$\begin{aligned}&\sum_i(\min\|\boldsymbol{x}_i-\boldsymbol{X}_i\boldsymbol{w}_i\|_2^2+\lambda\|\boldsymbol{w}_i\|_1)\\&\text{s.t. } w_{ij}\geqslant 0\end{aligned}\tag{5}$$

该NSR模型假定训练集中每个样本都能够用其他样本稀疏表示线性重构，从而构造一个SPG. 非负稀疏向量 $\boldsymbol{w}_i$ 反映了样例 $\boldsymbol{x}_i$ 是如何用其他样例稀疏表示的，因而揭示了数据间的聚类关系. 为获得一个灵活、实用的SPG构建方法，本文中令 $\lambda=0$[12]，式(5)即变换为式(6).

$$\begin{aligned}&\sum_i\min\|\boldsymbol{x}_i-\boldsymbol{X}_i\boldsymbol{w}_i\|_2^2\\&\text{s.t. } w_{ij}\geqslant 0\end{aligned}\tag{6}$$

在实际应用中往往存在各种噪声，这将对分类结果产生较大的影响. 为有效消除噪声的影响，本文提出了一种基于相关熵的非负稀疏表示框架. 在信息论学习中，两个随机变量 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的相似程度可以用相关熵来衡量，相关熵的定义如下：

$$V_\sigma(\boldsymbol{A},\boldsymbol{B})=E[k_\sigma(\boldsymbol{A},\boldsymbol{B})]\tag{7}$$

其中 $k_\sigma(\cdot)$ 是核函数，$E[\cdot]$ 是期望运算. 本文中 $k_\sigma(\cdot)$ 采用高斯核函数，即 $k_\sigma(\boldsymbol{x})=g(\boldsymbol{x},\sigma)$. 在实际问题中，很难估计数据的分布情况，因此文献[13]提出了基于样本的相关熵描述，其定义如下：

$$\hat{V}_{\sigma,n}(\boldsymbol{A},\boldsymbol{B})=\frac{1}{n}\sum_{i=1}^{n}g(\boldsymbol{a}_i-\boldsymbol{b}_i,\sigma)\tag{8}$$

把 $\boldsymbol{a}_i=\boldsymbol{x}_i$ 和 $\boldsymbol{b}_i=\boldsymbol{X}_i\boldsymbol{w}_i$ 代入相关熵的定义式(8)，得到每个样本的基于相关熵的NSR模型：

$$\begin{aligned}&\max_{\boldsymbol{w}_i}\sum_{k=1}^{d}g(x_{ik}-(\boldsymbol{X}_i\boldsymbol{w}_i)_k)-\lambda\sum_{j=1}^{n-1}w_{ij}\\&\text{s.t. } w_{ij}\geqslant 0\end{aligned}\tag{9}$$

这里 $g(\boldsymbol{x})=\exp\left(-\frac{\|\boldsymbol{x}\|^2}{2\sigma^2}\right)$ 是一个高斯核函数，x_{ik} 表示 $\boldsymbol{x}_i$ 的第 k 个分量，$(\boldsymbol{X}_i\boldsymbol{w}_i)_k$ 表示 $\boldsymbol{X}_i\boldsymbol{w}_i$ 的第 k 个分量. 这里的高斯核函数实际上是一个鲁棒函数，依据相关熵的性质，如果样本 $\boldsymbol{x}_i$ 是桥点或噪声，那么不能用线性模型 $\boldsymbol{X}_i\boldsymbol{w}_i$ 表示它，两者的距离在高斯核函数的作用下会得到一个较小的值，因而对目标函数贡献较小，因此通过式(9)得到的是一个鲁棒的度量矩阵. 在获得权重矩阵后，用式(10)进行归一化处理，将权重值限定在一定范围内，即可得到稀疏概率图.

$$W(i,j)=\begin{cases}w_{ij}\Big/\sum_{j'=1}^{n-1}w_{ij'}; & j<i\\ w_{i(j-1)}\Big/\sum_{j'=1}^{n-1}w_{ij'}; & j>i\\ 0; & j=i\end{cases}\tag{10}$$

$W(i,j)$ 反映了数据 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 的相似程度. 由于 $\sum_{j=1}^{n}W(i,j)=1$，$W(i,j)$ 也可看作 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 属于同一分类的概率.

由SPG的构建过程可知，它与LNP以及传统的图构建方式有很大的不同. 首先SPG构建不需要手动设置参数，如 k-最近邻中邻居个数 k、ε-ball最近邻中的 ε 等，设置这些参数往往需要一定的经验，不同的参数设置对图的邻接关系、稀疏程度和分类结果往往产生很大影响. LNP算法采用 k-最近邻选择，通常取 k 值小于样本总数 n，此时算法只反映了样本空间的局部信息而未反映样本空间的全局信息，因而未能充分反映聚类假设中的从局部和全局角度考虑样本结构的要求. 当 $k=n$ 时，尽管反映了样本空间的全局信息，且也能得到稀疏解，但其计算复杂度却很高，很难有效用于解决实际的半监督学习问题. 其次，SPG通过NSR和鲁棒NSR来构建稀疏概率图，将邻居选择和图权重计算合二为一，两个步骤相互作用很自然地决定了图的稀疏性，避免了LNP及传统方法将二者分步进行，过早固定样本的紧邻结构而影响权重计算的弊端. 最后，鲁棒NSR算法通过相关熵的鲁棒原理排除桥点和噪声，相比LNP算法中采用启发式方法排除桥点具有更好的理论基础.

l^1-graph[2,14]半监督学习中，图的权重系数可以是正的，也可以是负的，然而负的权重系数并不能较好地反映某一样本在线性重构其他样本时的重要性，以致不能合理地进行样本间的标签传播过程. 与 l^1-graph不同，SPG是通过NSR获得的

图，利用了样本空间的局部和全局信息，其权重系数是非负的，因此能更好地描述数据间的聚类关系，使得标签传播更为合理. 此外，与 l^1-graph 相比，SPG 的计算模型更简单，具有更低的计算复杂度.

2.2　基于 NSR 的标签繁殖

若 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_l,\boldsymbol{x}_{l+1},\cdots,\boldsymbol{x}_n\}$ 表示有 n 个训练样本的数据集，前 l 个已标记，记为 $\boldsymbol{X}_{\mathrm{l}}$，后 $n-l$ 个未标记，记为 $\boldsymbol{X}_{\mathrm{u}}$. 图构建完成后，下一步任务是预测未标记样本的标签. 本文中采用迭代繁殖方法，通过对已标记样本的标签在图上进行迭代繁殖直至收敛，从而获得未标记样本的标签.

若 $\mathscr{F}$ 表示定义在 $\boldsymbol{X}$ 上的分类函数集，$\forall\, \boldsymbol{f}\in\mathscr{F}$ 均能赋予每个样本点 $\boldsymbol{x}_i$ 一个实数 f_i，未标记样本点 $\boldsymbol{x}_{\mathrm{u}}$ 的标签可以由 $f_{\mathrm{u}}=f(\boldsymbol{x}_{\mathrm{u}})$ 的符号确定(这里仅考虑二分类问题). 样本点标签信息的传播包括两部分：一部分获自它的邻居的标签信息，另一部分则保留其初始的标记信息. 因此经过 $m+1$ 次传播后，$\boldsymbol{x}_i$ 的标签值可表示为

$$f_i^{m+1}=\alpha\sum_{j:\,\boldsymbol{x}_j\in\mathbf{N}(\boldsymbol{x}_i)}w_{ij}f_j^m+(1-\alpha)t_i \tag{11}$$

其中 $0<\alpha<1$，表示 $\boldsymbol{x}_i$ 从其邻居获得标签信息所占的比例. 令 $\boldsymbol{t}=(t_1\quad t_2\quad\cdots\quad t_n)^{\mathrm{T}}$，若 $\boldsymbol{x}_i\in\boldsymbol{X}_{\mathrm{l}}$，则 $t_i\in L$(L 为类标集)，若 $\boldsymbol{x}_i\in\boldsymbol{X}_{\mathrm{u}}$，则 $t_i=0$. $\boldsymbol{f}^m=(f_1^m\quad f_2^m\quad\cdots\quad f_n^m)^{\mathrm{T}}$ 表示样本在第 m 次迭代后的标签向量，并令 $\boldsymbol{f}^0=\boldsymbol{t}$，则式(11)可重写为

$$\boldsymbol{f}^{m+1}=\alpha\boldsymbol{W}\boldsymbol{f}^m+(1-\alpha)\boldsymbol{f}^0 \tag{12}$$

利用式(12)可以对样本的标签进行迭代更新直到收敛. 这里的收敛表示经过若干次迭代后，$\boldsymbol{f}^m$ 的值不再发生变化，文献[1]证明了序列$\{\boldsymbol{f}^m\}$的极限值为 $\boldsymbol{f}^*=(1-\alpha)(\boldsymbol{I}-\alpha\boldsymbol{W})^{-1}\boldsymbol{y}$，其中 $\boldsymbol{y}=\boldsymbol{f}^0$，$\boldsymbol{I}$ 是单位阵. 下面将它作为引理表述.

引理 1　当 $w_{ij}\geqslant 0$，$\sum_{j=1}^{n}\mathrm{W}(i,j)=1$，且 $m\to\infty$ 时，按 $\boldsymbol{f}^{m+1}=\alpha\boldsymbol{W}\boldsymbol{f}^m+(1-\alpha)\boldsymbol{f}^0$ 计算的序列$\{\boldsymbol{f}^m\}$收敛于极限值 $\boldsymbol{f}^*=(1-\alpha)(\boldsymbol{I}-\alpha\boldsymbol{W})^{-1}\boldsymbol{y}$，其中 $\boldsymbol{f}^0=\boldsymbol{y}$.

证明　参见文献[1].

标签繁殖方法可以很容易扩展到多分类问题. 若样本点有 c 类，标签集合 $L=\{1,2,\cdots,c\}$，$\boldsymbol{M}$ 是一个 $n\times c$ 的非负矩阵，$\boldsymbol{F}=(\boldsymbol{F}_1^{\mathrm{T}}\quad\boldsymbol{F}_2^{\mathrm{T}}\quad\cdots\quad\boldsymbol{F}_n^{\mathrm{T}})^{\mathrm{T}}\in\boldsymbol{M}$，是数据集 $\boldsymbol{X}$ 上的分类函数. 数据 $\boldsymbol{x}_i$ 的标签为 $y_i=\arg\max_{j\leqslant c}F_{ij}$. 初始设 $\boldsymbol{F}_0=\boldsymbol{T}$，当 $\boldsymbol{x}_i$ 是已标记数据时，若 $\boldsymbol{x}_i$ 标为 j 类，$T_{ij}=1$，否则 $T_{ij}=0$；当 $\boldsymbol{x}_i$ 是未标记数据时，$T_{ij}=0(1\leqslant j\leqslant c)$. 整个的算法过程概括如下.

输入：$\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_l,\boldsymbol{x}_{l+1},\cdots,\boldsymbol{x}_n\}\in\boldsymbol{R}^d$，$\{\boldsymbol{x}_i\}_{i=1}^{l}$ 是标记的数据，$\{\boldsymbol{x}_i\}_{i=l+1}^{n}$ 是未标记数据，标签繁殖常量 α.

输出：所有样本点的标签.

步骤 1　for $i=1$ to n do

(1) $\boldsymbol{X}_i=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_{i-1},\boldsymbol{x}_{i+1},\cdots,\boldsymbol{x}_n\}$；

(2) 求解式(6)基于 NSR 的极小二次规划问题得到权重向量 $\boldsymbol{w}_i$(若样本中存在噪声，则按式(9)的鲁棒 NSR 模型求解).

end for

步骤 2　依式(10)对权重归一化，即得权重矩阵 $\boldsymbol{W}$.

步骤 3　按 $\boldsymbol{F}_{t+1}=\alpha\boldsymbol{W}\boldsymbol{F}_t+(1-\alpha)\boldsymbol{Y}$ 迭代至收敛于 $\boldsymbol{F}^*$.

步骤 4　样本 $\boldsymbol{x}_i$ 的标签为 $y_i=\arg\max_{j\leqslant c}F_{ij}^*$.

下面给出算法的收敛性证明.

定理 1　基于 NSR 的标签繁殖算法是收敛的.

证明　因为 $w_{ij}\geqslant 0$ 且 $\sum_{j=1}^{n}\mathrm{W}(i,j)=1$，根据引理 1，算法收敛，且按式(12)计算的序列$\{\boldsymbol{f}^m\}$的极限为 $\boldsymbol{f}^*=(1-\alpha)(\boldsymbol{I}-\alpha\boldsymbol{W})^{-1}\boldsymbol{y}$. 证毕.

基于 NSR 的标签繁殖算法和标准的标签繁殖算法使用相同的标签传播方式，因此具有相同的时间复杂度. 其与标准的标签繁殖算法的不同在于权重矩阵的计算. 根据非负最小二乘理论[15]，当使用主动集算法时，非负解的计算复杂度随样本数 n 的增加而增加；而使用 BLOCK 或 PRECOR 算法求解时，其计算复杂度不依赖于样本数 n.

3　实验及结果分析

3.1　数据集选取

为了全面比较 l^1-graph 、LNP 和传统 KNN 算法与本文提出的基于非负稀疏表示和鲁棒非负稀疏表示的标签繁殖算法在半监督学习中的分类性能，本文从不同的机器学习领域，根据不同的数

据特点选取了6个数据集进行实验:ORL、COIL、Aust、Heart、TDT60和TDT100数据集.其中ORL是人脸识别数据集,COIL是物体识别数据集[16],其数据特点为非负离散数据.ORL包含了采自40个不同人的400个图像,这些图像是在不同时间、不同光照、不同表情下采集的.每幅图像被手动缩放成32像素×32像素大小.COIL包含了100个不同种类的物体图像,每一类选12个样本,共1 200个样本.Aust和Heart是UCI中的两个数据集,其数据特点为实数.Aust(Australian sign language signs)是一个澳大利亚人手语信息数据集.Heart是一个对心脏病进行诊断统计的数据集.TDT是文本分类数据集,其数据特点是具有稀疏性.TDT包含11 201个样本,共分96类,本文使用分类数最高的9类来进行实验,TDT60和TDT100表示分别在每类数据中选取60个、100个样本.表1显示了各个数据集的详细信息.

表1 实验中使用的数据集

Tab.1 Datasets used in experiment

数据集	维数	类别数	样本数	数据特点
ORL	1 024	40	400	非负
COIL	1 024	100	1 200	非负
UCI_Aust	14	2	690	实数
UCI_Heart	13	2	270	实数
TDT60	500	9	540	稀疏
TDT100	500	9	900	稀疏

3.2 算法设置

实验采用KNN、LNP、l^1-graph算法和本文中提出的两种算法进行比较,这些算法分别采用不同的模型构建图,即运用不同的策略进行邻居选择和权重计算.成功构建图之后,标签繁殖的过程是相同的,标签繁殖中的参数 $\alpha=0.4$.

(1)KNN算法.首先用 k-最近邻方法选择 k 个邻居,然后使用式(1)的高斯核函数计算邻居之间的权重构建图,这里设 $k=20$,$\sigma=0.01$.

(2)LNP算法.首先用 k-最近邻方法选择 k 个邻居,然后利用式(2)并增加约束条件 $w_{ij}\geqslant 0$ 来计算邻居之间的权重.这里设 $k=20$.

(3)l^1-graph算法.通过式(13)的 l^1 范式构造权重图:

$$\begin{aligned}&\min_{\boldsymbol{w}_i}\|\boldsymbol{w}_i\|_1\\&\text{s.t. }\|\boldsymbol{x}_i-\boldsymbol{X}_i\boldsymbol{w}_i\|_2\leqslant\varepsilon\end{aligned}\tag{13}$$

(4)NSR_1(非负稀疏表示算法).利用式(6)的最小二乘模型进行无参构图,并用式(10)归一化.模型求解方法具体参见文献[15、17、18].

(5)NSR_2(鲁棒非负稀疏表示算法).利用式(9)的基于相关熵的鲁棒NSR模型来构图,用主动集算法[6]求解模型获得权重矩阵,并用式(10)归一化.

3.3 实验结果及分析

将本文提出的非负稀疏表示标签繁殖算法(NSR_1)、鲁棒非负稀疏表示标签繁殖算法(NSR_2)与传统的 l^1-graph、KNN以及LNP算法分别在人脸识别、物体识别、UCI数据集和文本数据集上进行了对比实验.为更合理地进行比较,所有实验都独立进行了50次,每一次实验中的标记数据都是随机选取的,最后计算出这50次实验的平均分类错误率及均方差.

3.3.1 *人脸和物体识别* 自动图像识别技术是计算机视觉领域最显著、最具挑战性的应用研究之一,目前已经提出了很多有效的图像识别算法.稀疏表示理论提出一个测试样本可以表示为其他训练样本的稀疏线性组合.本文把这一理论应用于自动图像识别的半监督学习任务中,提出了 NSR_1 和 NSR_2 算法.表2列出了它们与 l^1-graph、LNP以及KNN算法在ORL、COIL数据集上进行半监督学习的分类错误率的比较结果.从表中可以看出:(1)在相同条件下,NSR_1 和 NSR_2 比 l^1-graph、KNN和LNP算法有更低的分

表2 人脸和物体识别数据集上的半监督分类错误率

Tab.2 Semi-supervised classification error rates for face recognition and object recognition datasets

		错误率/%				
		NSR_2	NSR_1	l^1-graph	LNP	KNN
ORL	50%	16.0±2.9	17.4±3.2	17.9±3.0	18.3±3.3	23.2±4.2
	60%	11.7±3.1	13.2±2.9	14.0±3.2	14.1±3.0	20.0±4.4
	80%	6.2±3.1	6.9±3.1	7.5±3.3	8.2±4.0	15.8±4.2
COIL	50%	10.9±0.7	11.2±2.0	16.6±0.9	12.1±0.8	16.8±0.8
	60%	9.4±0.8	9.8±1.6	14.8±0.9	10.0±1.9	15.3±1.0
	80%	8.1±0.6	8.9±1.0	11.9±1.3	9.2±1.7	13.2±1.3

类错误率.(2) 相同条件下，NSR_2 比 NSR_1 具有更低的分类错误率.这可以理解为实验数据集中存在噪声(如部分图像被遮挡或者图像模糊)，NSR_2 利用高斯函数能自动辨别噪声图像，减弱了噪声影响，从而具有更低的分类错误率.由此可见，NSR_2 算法具有更高的鲁棒性.

3.3.2　UCI 数据集　UCI 数据集是标准的机器学习数据集，用于测试机器学习算法的性能.本文选取了其中的两个数据集 Aust 和 Heart，并将所提出的基于 NSR 的标签繁殖算法和其他传统半监督学习算法在这两个数据集上进行了实验比较.表 3 给出了各种算法的分类错误率的比较结果.从表 3 不难得出与人脸识别和物体识别实验相似的结论：(1)相同条件下，NSR_1 和 NSR_2 算法比 l^1-graph、KNN 和 LNP 算法具有更低的分类错误率.(2)受数据集中的噪声与桥点影响，鲁棒算法 NSR_2 比 NSR_1 具有更低的分类错误率.

表 3　UCI 数据集上的半监督分类错误率

Tab. 3　Semi-supervised classification error rates for UCI dataset

		错误率/%				
		NSR_2	NSR_1	l^1-graph	LNP	KNN
Aust	50%	25.6±2.6	26.4±2.8	27.5±2.1	28.8±1.9	30.2±1.8
	60%	24.2±2.8	25.1±2.6	26.4±2.6	27.6±1.4	29.3±2.4
	80%	22.4±3.5	23.4±3.0	24.3±3.2	25.6±3.5	28.7±3.8
Heart	50%	25.6±3.8	27.0±2.7	29.8±3.3	31.3±3.5	34.0±3.4
	60%	24.4±3.7	26.0±3.7	28.6±3.5	30.1±3.7	32.8±4.2
	80%	22.1±6.3	23.8±4.6	26.1±5.7	28.3±5.4	31.3±6.1

3.3.3　文本数据集　为进一步研究本文提出的基于 NSR 的标签繁殖算法在半监督学习中的性能，本实验采用了 TDT 文本分类数据集，表 4 给出了 NSR_1、NSR_2、l^1-graph、LNP 以及 KNN 算法的实验比较结果.从表中可以看出，该结果和人脸识别、物体识别、UCI 数据集上的实验结果有所不同：(1)在相同条件下，NSR_1 算法在 TDT 数据集上比 l^1-graph、KNN 和 LNP 算法具有更低的分类错误率.(2)在相同条件下，鲁棒算法 NSR_2 比其他 4 种算法的分类错误率都要高.导致 NSR_2 性能较差的原因一方面在于文本数据集中几乎不存在噪声；另一方面在于文本数据集中样本特征本来就是稀疏的，NSR_2 算法采用高斯函数计算权重，而高斯函数往往过分注重稀疏项，因而在稀疏样本上产生了很大的权重.

表 4　文本数据集上的半监督分类错误率

Tab. 4　Semi-supervised classification error rates for text dataset

		错误率/%				
		NSR_2	NSR_1	l^1-graph	LNP	KNN
TDT60	50%	30.4±2.9	17.7±2.7	18.1±2.7	18.3±2.2	24.1±3.0
	60%	27.3±3.3	15.2±2.8	16.1±2.7	16.4±2.5	22.9±2.7
	80%	23.5±4.0	12.3±3.1	13.0±3.3	13.9±3.1	19.8±4.0
TDT100	50%	24.7±2.0	14.2±1.6	15.0±1.6	15.0±1.7	18.4±1.7
	60%	22.5±1.8	12.9±2.0	13.6±1.9	14.0±1.7	16.8±2.0
	80%	18.9±3.0	9.8±2.1	11.2±2.4	11.9±2.3	16.1±2.9

3.4　稀疏性分析

NSR_1、NSR_2、LNP 以及 l^1-graph 算法都能获得样本空间的稀疏编码.为探究实验中各半监督学习算法分类准确率存在差异的深层原因，本文进一步研究了这几种算法在各数据集上构建的图的稀疏特性.本文中，图的稀疏性采用相应权重矩阵的 l^0 范式值来衡量，该值越小，图越稀疏.权重矩阵的 l^0 范式定义为

$$\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{W}(i,:)\|_0$$

其中 $\boldsymbol{W}(i,:)$ 表示权重矩阵 $\boldsymbol{W}$ 的第 i 行，衡量图稀疏性的 l^0 范式可以看作概率图上所有样本的邻居个数的均值.图 1 给出了实验中 5 种标签繁殖算法在所采用的 6 个数据集上构建的图的稀疏性.从图中不难看出：(1)KNN 算法由于采用了固定的近邻个数 $k=20$，在 6 个数据集上的 l^0 范式值都是 20.(2)NSR_1 和 NSR_2 算法的 l^0 范式值比 LNP 算法要高.这表明 LNP 算法学习到的图要比 NSR_1 和 NSR_2 算法所学得的图更加稀疏.(3)l^1-graph 的 l^0 范式值比较高.

NSR_1 和 NSR_2 算法在构建概率图时同时利用了样本的局部信息和全局信息，获得的是数据集自适应的邻居个数与聚类关系，因此获得的稀疏概率图更加接近数据集的实际内在稀疏性，反映了每个数据集上较优的邻居个数.KNN 算法采用人工设定 k 值，在前 4 个数据集上设置 $k=20$ 比实际较优的近邻个数值偏大，而在后两个数据集上设置 $k=20$ 则偏小，这也进一步说明手工

设定一个合适的 k 值是非常困难的. LNP 算法也是手工设置近邻个数 k($k \ll \min\{n, d\}$),因此所学得的图只利用了样本的局部信息,导致该图过于稀疏,不能更准确地反映样本间的聚类关系. l^1-graph 同时使用了正的和负的系数表示样本,造成样本邻居过多,并且那些负系数的邻居对表示样本起消极作用. 这都使得 LNP 算法和 l^1-graph 算法的分类性能往往低于 NSR_1 和 NSR_2 算法.

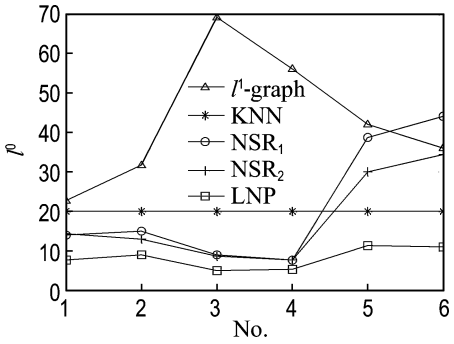

图 1 不同算法的稀疏性

Fig. 1 The sparsity of different algorithms

4 结 论

本文提出了一种基于非负稀疏表示的用于半监督学习的标签繁殖算法. 该算法通过将一个数据点表示为训练集中其他数据点的非负稀疏线性组合,从而构造出一个 SPG. 图的构造同时利用了样本的局部和全局信息,能够更好地反映数据间的聚类关系. 图中的权重通过 NSR 算法和基于相关熵的鲁棒 NSR 算法计算得到. 这两种方法以一种无参的方式同时实现了自适应邻居选择和权重计算过程. 基于相关熵的 NSR 算法削弱了噪声的影响,使该算法具有较强的鲁棒性. 图构建完成后,标签在概率图上进行迭代繁殖直至收敛,从而获得所有未标记样本的标签. 在多个机器学习数据集上的实验结果表明基于 NSR 的标签繁殖算法具有较高的分类准确率.

参考文献:

[1] WANG Fei, ZHANG Chang-shui. Label propagation through linear neighborhoods [J]. **IEEE Transactions on Knowledge and Data Engineering**, 2008, **20**(1):55-67

[2] YAN S C, WANG H. Semi-supervised learning by sparse representation [C] // **SIAM International Conference on Data Mining SDM**. Philadelphia: Society for Industrial and Applied Mathematics Publications, 2009:792-801

[3] WRIGHT J, MA Y, MAIRAL J, *et al*. Sparse representation for computer vision and pattern recognition [C] // **Proceedings of IEEE**. Piscataway: Institute of Electrical and Electronics Engineers Inc., 2009:1031-1044

[4] BELKIN M, NIYOGI P. Laplacian eigenmaps for dimensionality reduction and data representation [J]. **Neural Computation**, 2003, **15**(6):1373-1396

[5] DONOHO D. Compressed sensing [J]. **IEEE Transactions on Information Theory**, 2006, **52**(4):1289-1306

[6] HE Ran, ZHENG Wei-shi, HU Bao-gang. Maximum correntropy criterion for robust face recognition [J]. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, 2011, **33**(8):1561-1576

[7] ZHOU D, BOUSQUET O, LAL T, *et al*. Learning with local and global consistency [C] // **17th Annual Conference on Neural Information Processing Systems (NIPS)**. Cambridge:MIT Press, 2004:321-328

[8] CORTES C, MOHRI M. On transductive regression [C] // **Twentieth Annual Conference on Neural Information Processing Systems (NIPS)**. Cambridge: MIT Press, 2007:305-312

[9] ROWEIS S, SAUL L. Nonlinear dimensionality reduction by locally linear embedding [J]. **Science**, 2000, **290**(5500):2323-2326

[10] DONOHO D L, TANNER J. Sparse nonnegative solution of underdetermined linear equations by linear programming [C] // **Proceedings of the National Academy of Sciences**. Washington D C: National Academy of Sciences, 2005

[11] BRUCKSTEIN A M, ELAD M, ZIBULEVSKY M. On the uniqueness of nonnegative sparse solutions to underdetermined systems of equations [J]. **IEEE Transactions on Information Theory**, 2008, **54**(11):4813-4820

[12] HE Ran, HU Bao-gang, ZHENG Wei-shi, *et al*.

Two-stage sparse representation for robust recognition on large-scale database [C] // **Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10)**. Menlo Park: American Association for Artificial Intelligence, 2010:475-480

[13] LIU Wei-feng, POKHAREL P P, PRINCIPE J C. Correntropy: Properties and applications in non-Gaussian signal processing [J]. **IEEE Transactions on Signal Processing**, 2007, **55**(11):5286-5298

[14] BELKIN M, MATVEEVA I, NIYOGI P. Regularization and semi-supervised learning on large graphs [C] // **Lecture Notes in Artificial Intelligence**. Berlin:Springer-Verlag, 2004:624-638

[15] PORTUGAL L F, JUDICE J J, VICENTE L N. A comparison of block pivoting and interior-point algorithms for linear least squares problems with nonnegative variables [J]. **Mathematics of Computation**, 1994, **63**(208):625-643

[16] NENE S, NAYAR S, MURASE H. Columbia object image library (COIL-20) [R] // **Technical Report CUCS-006-96**. Columbia: Department of Computer Science, Columbia University, 1996

[17] BJORCK A. A direct method for sparse least-squares problems with lower and upper bounds [J]. **Numerische Mathematik**, 1988, **54**(1):19-32

[18] VO N, MORAN B, CHALLA S. Nonnegative least square classifier for face recognition [C] // **Proceedings of the 6th International Symposium on Neural Networks: Advances in Neural Networks**. Berlin:Springer Verlag, 2009:449-456

Label propagation algorithm based on nonnegative sparse representation

YANG Nan-hai[*1], SANG Yuan-yuan[2], HE Ran[2], WANG Xiu-kun[1,2]

(1. School of Software Technology, Dalian University of Technology, Dalian 116620, China;
2. School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China)

Abstract: A novel label propagation algorithm for semi-supervised learning based on nonnegative sparse representation(NSR) is presented. Firstly, this algorithm derives a sparse probability graph (SPG) from nonnegative weight coefficients which are computed by nonnegative sparse representation algorithm. The weights of SPG naturally reveal the clustering relationship of labeled samples and unlabeled samples, meanwhile avoid the adjacency selection and parameter setting process in traditional semi-supervised learning algorithm. Then, the labels of unlabeled samples are propagated until convergence to obtain all the labels of samples. Extensive experimental results on face recognition, object recognition, UCI machine learning and TDT text datasets show that label propagation algorithm based on NSR achieves the lower error rate as compared with the standard label propagation algorithm.

Key words: nonnegative sparse representation; semi-supervised learning; sparse probability graph; clustering relationship; label propagation