一种基于多图的集成直推分类方法

余国先*[①] 张国基[②] 韦 佳[①] 任亚洲[①]
[①](华南理工大学计算机科学与工程学院 广州 510006)
[②](华南理工大学理学院 广州 510640)

摘 要：基于图的直推分类器依赖于图结构。高维数据通常具有冗余和噪声特征，在其上构造的图不能充分反映数据的分布信息，分类器性能因此下降。为此，该文提出一种多图构建方法并把它应用到直推分类中。该方法首先生成多个随机子空间并在每个子空间上进行半监督判别分析，其次在每个判别子空间上构造图并训练一个直推分类器，最后投票融合这些分类器为一个集成分类器。实验结果表明，对比其它直推分类器，该文的集成分类器具有分类正确率高、对参数鲁棒等特点。

关键词：信息处理；直推分类器；图结构；随机子空间；投票
中图分类号：TP18 **文献标识码**： A **文章编号**：1009-5896(2011)08-1883-06
DOI: 10.3724/SP.J.1146.2010.01424

A Multi Graphs Based Transductive Ensemble Classification Method

Yu Guo-xian[①] Zhang Guo-ji[②] Wei Jia[①] Ren Ya-zhou[①]
[①](*School of Computer Science and Engineering*, *South China University of Technology*, *Guangzhou* 510006, *China*)
[②](*School of Sciences*, *South China University of Technology*, *Guangzhou* 510640, *China*)

Abstract: Graph based transductive classifiers are dependent on graph structure. Because of redundant and noisy features in high dimensional data, a graph, constructed from these data, can not reflect their distribution information faithfully. Consequently, the performance of a transductive classifier is downgraded. To address this problem, a multiple graphs construction scheme is introduced and applied into transductive classification. The scheme generates firstly several random subspaces and applies semi-supervised discriminative analysis in each subspace. Next, it trains a transductive classifier in each discriminative subspace. And finally, by voting rule, it fuses these classifiers as an ensemble classifier. Empirical results show that, in comparison with other transductive classifiers, the proposed ensemble classifier is more precise and robust to parameters selection.

Key words: Information processing; Transductive classifier; Graph structure; Random subspace; Voting rule

1 引言

随着信息技术的发展，积累的数据与日俱增，如网络文本和图像等。这些数据通常只有少量类别属性已知，并具有很高的维度。高的维度意味着分类器需要更多的训练样本来获得一个理想的分类器，而仅利用稀少的有标记样本并不能保证分类器具有好的泛化性能。半监督学习[1]通过平衡利用有标记样本和无标记样本获取一个性能较好的学习器，受到越来越多的关注。基于图的直推分类[1]是半监督学习的一个重要分支，它把所有参与训练的样本(包括有标记和无标记样本)均作为图上的顶点，并利用图上的边权重描述样本之间的相似度信息，再在图上利用有标记样本预测这些无标记样本。如高斯随机场与和谐函数(GFHF)[2]，局部与全局一致性学习(LGC)[3]，线性邻域标签传播(LNP)[4]，鲁棒多类直推方法(RMGT)[5]等都是具有代表性的直推分类器。大部分基于图的直推分类方法都可以看作是有标记样本的标签信息在图上的传播过程[1]，定义一个结构良好的图是决定这些方法有效性的关键[1,4–6]。由于高维数据通常包含噪声和冗余特征，常用的距离度量(如欧氏距离，余弦距离等)并不能很好地描述这类数据之间的近邻结构信息[7]，直推分类器性能因此下降。GFHF, LGC 和 LNP 等在低维数据上均可获得较高的分类正确率，但在高维数据上的分类效果并不显著。因而，很有必要在低的维度上构造图再进行直推分类。

集成学习可以推进半监督学习器的性能[8]。本文

2010-12-27 收到，2011-04-21 改回
国家自然科学基金(60973083, 61003174)和广东省自然科学基金(9451064101003233, 10451064101004233)资助课题
*通信作者：余国先 guoxian.yu@mail.scut.edu.cn

先期实验表明，类似决策森林方法[9]，在随机子空间上训练多个直推分类器，再通过投票获得的集成分类器性能并不显著，因为这些随机子空间上基础分类器的准确率较低。基础分类器的准确率和它们之间的差异性是决定集成分类器性能的两个重要指标[10]，高准确率要求基础分类器在预测样本时一致性较高，而差异性意味着基础分类器对同一样本的分类结果不一致。构造既有较高准确率也有较高差异性的基础分类器是提高集成分类器性能的关键[10]。鉴于此，本文提出一种基于多图的集成直推分类方法(Multi Graphs based Transductive Ensemble Classification, MGTEC)。MGTEC 首先生成多个随机子空间，其次在子空间上进行半监督判别分析空间(SDA)[11]，再在每个判别的子空间上构造一个近邻图并训练一个直推分类器，最后投票融合这些分类器。在人脸图像上的实验表明，MGTEC 的基础分类器具有较高的精度和差异度；对比其它直推分类方法，MGTEC 不仅表现出更好的分类效果，对参数的选取也较鲁棒。MGTEC 的多图构建策略非常显式直观，可直接应用到其他基于图的半监督学习算法中。

2 拉普拉斯谱图

基于图的直推分类方法如 GFHF, LGC, LNP 等和基于图的谱嵌入如 SDA, LPP[12], MFA[13]等都同 Laplace 谱图有着紧密的联系。令 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_l,\boldsymbol{x}_{l+1},\cdots,\boldsymbol{x}_{l+u}\}$，$\boldsymbol{X}\in R^{D\times(l+u)}$，其中前 l 个样本的标签已知，其余 u 个未知，D 为样本维度。图 Laplace 矩阵[12]定义如下：

$$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W} \tag{1}$$

其中 $\boldsymbol{D}$ 为对角矩阵，定义如下：

$$\boldsymbol{D}_{ii}=\sum_{j=1}^{l+u}\boldsymbol{W}_{ij} \tag{2}$$

$\boldsymbol{W}$ 通常定义为简单的 0-1 相似度：

$$\boldsymbol{W}_{ij}=\begin{cases}1, & \boldsymbol{x}_i\in kNN(\boldsymbol{x}_j)\text{或}\boldsymbol{x}_j\in kNN(\boldsymbol{x}_i)\\ 0, & \text{其他}\end{cases} \tag{3}$$

或高斯热核相似度[12]：

$$\boldsymbol{W}_{ij}=\begin{cases}\exp\left(-\dfrac{d(\boldsymbol{x}_i,\boldsymbol{x}_j)}{\sigma}\right), & \boldsymbol{x}_i\in kNN(\boldsymbol{x}_j)\text{或}\boldsymbol{x}_j\in kNN(\boldsymbol{x}_i)\\ 0, & \text{其他}\end{cases} \tag{4}$$

其中 $\boldsymbol{x}_i\in kNN(\boldsymbol{x}_j)$ 表示 $\boldsymbol{x}_i$ 是 $\boldsymbol{x}_j$ 的 k 个最近邻之一，$d(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 为某种距离度量，如欧几里得距离，马氏距离等，σ 为高斯热核宽度。

从式(1)-式(4)可以看出，$\boldsymbol{L}$ 和 $\boldsymbol{W}$ 最终都依赖于 k 近邻图结构。然而随着样本维数的增多，常用的距离度量并不能反映样本之间的近邻结构信息[7]，定义的近邻图不足以刻画样本的分布信息，基于图的谱嵌入和直推分类器的性能因此下降。

3 基于多图的集成直推分类

基于多图的集成直推分类分为两步：基于随机子空间的半监督判别分析(Random Subspace based Semi-supervised Discriminate Analysis, RSSDA)和 MGTEC。

3.1 RSSDA

SDA 通过在 k 近邻图上定义一个 Laplace 矩阵 $\boldsymbol{L}$ 来描述样本之间的流形结构信息，再结合线性判别分析(LDA)[14]获得一个 $C-1$ 维的判别子空间，其中 C 为样本类别数。如同绝大多数基于图的降维方法，该方法的有效性也依赖于定义一个结构良好的近邻图[13]。由于高维数据通常包含噪声和大量冗余特征，定义一个结构良好的近邻图比较困难。为此，本文在低维的随机子空间上构造图，进而提出 RSSDA, RSSDA 具体流程如表 1 所示。

表 1 RSSDA 算法

输入：随机子空间大小 p 及个数 T, $\boldsymbol{X}$, C

输出：判别投影矩阵 $\boldsymbol{A}_t\in R^{p\times(C-1)}(1\le t\le T)$

For $t=1$ to T

(1)生成一个大小为 p 的随机子空间，登记第 t 个随机子空间对应的样本集为 $\boldsymbol{S}_t=\{\boldsymbol{S}_t^1,\boldsymbol{S}_t^2,\cdots,\boldsymbol{S}_t^{(l+u)}\}$，$\boldsymbol{S}_t\in R^{p\times(l+u)}$，$\boldsymbol{S}_t$ 选中的特征索引 $\boldsymbol{R}_t$ 满足：

$$\left.\begin{aligned}&\sum_{j=1}^{D}\boldsymbol{R}_{tj}=p\\ &\text{s.t. }\boldsymbol{R}_{tj}\in\{0,1\}\end{aligned}\right\} \tag{5}$$

$\boldsymbol{R}_{tj}=1$ 表示第 j 个特征在 $\boldsymbol{S}_t$ 中被选中，$\boldsymbol{R}_{tj}=0$ 表示未被选中。

(2)在 $\boldsymbol{S}_t$ 上构造图 Laplace 矩阵 $\boldsymbol{L}^t$：

$$\boldsymbol{L}^t=\boldsymbol{D}^t-\boldsymbol{W}^t \tag{6}$$

$\boldsymbol{W}^t$ 定义如下：

$$\boldsymbol{W}_{ij}^t=\begin{cases}1, & \boldsymbol{S}_t^i\in kNN(\boldsymbol{S}_t^j)\text{或}\boldsymbol{S}_t^j\in kNN(\boldsymbol{S}_t^i)\\ 0, & \text{其他}\end{cases} \tag{7}$$

(3)在 $\boldsymbol{S}_t$ 上求解关于 $\boldsymbol{a}$ 的广义特征值问题：

$$\boldsymbol{S}_t\boldsymbol{W}\boldsymbol{S}_t^{\mathrm{T}}\boldsymbol{a}=\lambda(\boldsymbol{S}_t(\tilde{\boldsymbol{I}}+\alpha\boldsymbol{L}^t)\boldsymbol{S}_t^{\mathrm{T}}+\beta\boldsymbol{I})\boldsymbol{a} \tag{8}$$

关于式(8)的具体推导和各参数的定义详见文献[11]。令 $\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_{C-1}$ 为式(8) $C-1$ 个最大特征值对应的特征向量，$\boldsymbol{A}_t=\{\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_{C-1}\}$, $\boldsymbol{A}_t\in R^{p\times(C-1)}$，则 $\boldsymbol{A}_t$ 可以把 $\boldsymbol{S}_t$ 投影到 $C-1$ 维的判别子空间。

End for

不同于 SDA, RSSDA 目标方程中的 $\boldsymbol{L}^t$ 在随机子空间上构造，随机子空间的大小 p 远小于 D，它

较少受噪声和冗余特征的影响，定义的近邻图能较好地反映子空间的分布信息。RSSDA 的广义特征向量在随机子空间上求解，其时间复杂度为 $O(p^3)$，而 SDA 在原始空间上求解，其时间复杂度为 $O(D^3)$。为改善图嵌入结果，LGSSDR[15]定义一个奖励图和一个惩罚图，再在图上构造 Laplace 矩阵进行基于图的半监督降维。EGGLE[16]利用测地线距离和广义高斯函数构造多个图，再在每个图上进行谱嵌入，最后在这些嵌入结果上集成分类。SSDA[17]利用流形距离来刻画样本之间的距离，再进行半监督判别分析。这 3 种方法都是基于原始空间上的近邻图进行优化，此时近邻图易受冗余和噪声特征的影响，并不能保证得到优化的图，另外，特征向量求解的时间复杂度均为 $O(D^3)$。

3.2 MGTEC

基于图的直推分类方法依赖于图结构。尽管 LGC, GFHF 和 LNP 拥有不同的优化目标方程，但是它们的有效性更多依赖于近邻图及其上定义的边权重[1,4−6]。实验表明，类似决策森林[9]的方法构造的集成直推分类器精度并不高。通过在 RSSDA 的判别子空间上训练的基础直推分类器不仅具有较高的精度，由于判别子空间之间的差异性，这些基础分类器之间还具有一定的差异性，从而保证了 MGTEC 具有更高的精度。下文以 GFHF 为例，论述如何在判别子空间上训练基础直推分类器并通过投票策略集成它们。GFHF 的目标方程如下[2]：

$$\arg\min_{f\in R}\left(\infty\sum_{i=1}^{l}(f_i-y_i)^2+\frac{1}{2}\sum_{i,j=1}^{l+u}\boldsymbol{W}_{ij}(f_i-f_j)^2\right) \tag{9}$$

其中 y_i 为第 $i\,(1\le i\le l)$ 个样本的已知标签，f_j 为第 j 个样本的预测标签，$\boldsymbol{W}_{ij}$ 的定义同式(4)。令 $\boldsymbol{f}=[\boldsymbol{f}_l,\boldsymbol{f}_u]^{\mathrm{T}}$，其中 $\boldsymbol{f}_l$ 为 l 个有标记样本的标签值，$\boldsymbol{f}_u$ 为 GFHF 在 u 个未标记样本上的预测标签值。根据文献[2]，

$$\boldsymbol{f}_u=(\boldsymbol{D}_{uu}-\boldsymbol{W}_{uu})^{-1}\boldsymbol{W}_{ul}\boldsymbol{f}_l \tag{10}$$

其中

$$\boldsymbol{W}=\begin{bmatrix}\boldsymbol{W}_{ll} & \boldsymbol{W}_{lu}\\ \boldsymbol{W}_{ul} & \boldsymbol{W}_{uu}\end{bmatrix} \tag{11}$$

$\boldsymbol{D}$ 为对角矩阵，$\boldsymbol{D}_{ii}=\sum_{j=1}^{l+u}\boldsymbol{W}_{ij}$，其分块同 $\boldsymbol{W}$ 一样。从式(10)可以看出 $\boldsymbol{f}_u$ 也依赖于一个有效的相似性度量 $\boldsymbol{W}$。在 RSSDA 的基础上，本文提出 MGTEC，其流程如表 2 所示。

通过在判别分析后的子空间上构建图，再在图上训练直推分类器，可有效降低噪声和冗余特征对分类器的影响。另外，由于这些判别的子空间具有较好的区分能力和差异性，因而易于在这些子空间上获得具有高准确率和较好差异性的基础分类器，这正是 MGTEC 的有效性的基础。事实上，MGTEC 可以看做是降维和直推分类结合的一种混合方法。本文中的 SDA 也可以用其它的降维方法代替，如 LGSSDR；同样，式(14)也可以采用 LGC, LNP 等方法的目标方程替换。LapRLSC[18]在核主成分分析(KPCA)后的低维嵌入上定义一个近邻图，再在近邻图上定义一个 Laplace 矩阵，并把它作为一个正则项引入到半监督分类中，但该方法需要选择合适的核宽度和主成分个数。RASCO[19]通过在随机子空间上训练协同训练所需的子分类器再集成分类，但它对子空间的依耐性较高且子分类器的精度也较低。基于子空间的集成方法也被应用到人脸识别中，Semi-RS[10]首先把人脸图像分割为多个子模块，其次在这些子模块上进行主成分分析(PCA)[14]，再在这些 PCA 子空间上产生多个随机子空间，最后在这些子空间上训练分类器进行投票集成，但它需要选择合适的子模块大小和较多的基础分类器。RLDA[20]先把图像投影到 $N-1$ 维(N 为样本数量)的 PCA 子空间，再选取前 N_0 个最大主成分对应的特征，再从剩余的 $N-N_0-1$ 个特征中多次随机选择 N_1 个特征，构成大小为 N_0+N_1 的子空间，最后在这些子空间上训练 LDA 分类器并集成，但它较难选择合适的 N_0 和 N_1。不同于 Semi-RS 和 RLDA, MGTEC 不需 PCA 进行预处理，而且比较易于选择合适大小的随机子空间和核宽度。

表 2 MGTEC 算法

输入：随机子空间个数 $T, S_t(1\le t\le T)$，样本类别数 C

输出：未标记样本预测标签值 $\boldsymbol{f}_u$

For $t=1$ to T

(1)根据式(8)求得判别投影矩阵 $\boldsymbol{A}_t$，由 $\boldsymbol{A}_t$ 获得 $\boldsymbol{S}_t$ 的 $C-1$ 维嵌入 $\boldsymbol{Z}_t$：

$$\boldsymbol{Z}_t=\boldsymbol{A}_t^{\mathrm{T}}\boldsymbol{S}_t \tag{12}$$

(2)在 $\boldsymbol{Z}_t$ 上定义对应的 $\boldsymbol{W}^t$：

$$\boldsymbol{W}_{ij}^t=\begin{cases}\exp\left(-\dfrac{d(\boldsymbol{Z}_t^i,\boldsymbol{Z}_t^j)}{\sigma}\right), & \boldsymbol{Z}_t^i\in kNN(\boldsymbol{Z}_t^j)\\ & \text{或}\boldsymbol{Z}_t^j\in kNN(\boldsymbol{Z}_t^i)\\ 0, & \text{其他}\end{cases} \tag{13}$$

(3)在 $\boldsymbol{Z}_t$ 上求解 u 个未标记样本的预测标签值：

$$\boldsymbol{f}_u^t=(\boldsymbol{D}_{uu}^t-\boldsymbol{W}_{uu}^t)^{-1}\boldsymbol{W}_{ul}^t\boldsymbol{f}_l \tag{14}$$

End for

(4)通过投票策略，集成预测第 j 个未标注样本：

$$f_j=\arg\max_{1\le k\le C}\left(\sum_{t=1}^{T}(f_j^t=k)\right) \tag{15}$$

4　实验及分析

本节通过在两个标准人脸数据集上的分类实验

来验证 MGTEC 的有效性。参与比对的算法有 GFHF, LGC, LNP, RMGT, SDAGFHF(先在原始空间上 SDA，再在判别子空间上训练 GFHF)和 RSGFHF(先在每个随机子空间上训练一个基础分类器 GFHF，再通过投票集成它们)，前 5 个比对算法都是在单图上的直推分类，最后 1 个对比算法也是基于多图的集成直推分类。RSGFHF 与 MGTEC 的主要区别是 MGTEC 在经过判别分析后的子空间上训练基础分类器，而 RSGFHF 则在随机子空间上训练基础分类器。实验采用的两个数据集为 AR[21] 和扩展的 yaleB[22]。在 AR 的实验中，共选取 2600 张(50 名男性，50 名女性)图像作为实验数据集，其中每个实例 26 张图像，在实验前，先对这些图像进行配准，再剪裁到大小为 42×30 的灰度图像。在 yaleB 的实验中，选择其正面图像子集，共计 2414 张，共 38 个实例，并配准再剪裁到大小为 32×32 的灰度图像。为了便于分析，做如下符号约定，Lp: 每类有标记样本数量，C: 样本的类别数，D: 样本的原始维度，p: 随机子空间大小，EnsK: 随机子空间的个数，N: 训练样本数量，k: 近邻大小。实验中，式(8)中参数 α,β 均取 0.1, σ 取为训练样本间欧几里得距离的均值，k 取为 5。算法的评价指标为算法在测试样本上的分类正确率，测试样本为训练样本中的未知标记样本。为避免随机性，报告的实验结果为 20 次独立运行结果的平均值。

4.1 不同数量标记样本下的分类性能

为了分析算法在不同数量有标记样本下的分类精度，本文分别在 AR 和 yaleB 上执行实验。在这两个实验中，EnsK 为 20，p 为类别数的 2 倍。图 1 和图 2 给出了各比对算法在不同 Lp 下的分类精度变化情况。

从图 1 和图 2 可以看出，相对于其他算法，MGTEC 通常可以获得更高的分类正确率。这是由于 MGTEC 在判别嵌入的子空间上构造的图较少受噪声和冗余特征的影响，从而更好的描述了子空间的分布信息。SDAGFHF 获得了仅次于 MGTEC 的性能，这说明有效的降维可以提高分类器的性能。LNP 通过利用线性邻域传播来优化近邻图边权重，获得了比 GFHF, LGC 更高的精度，这说明对图进行优化可以提高直推分类器的精度。RMGT 利用非参数学习方法改进近邻图边权重，但它没有利用同类样本之间的相似度信息，在迭代优化过程中需要较多的矩阵运算损失了运算精度，分类正确率因此下降，在 yaleB 上甚至低于 GFHF 和 LGC。RSGFHF 在 Lp 增加时分类精度不断提高，但其精度低于 SDAGFHF 和 MGTEC，同 GFHF 也没有显著的区别，这正好验证了在随机子空间上训练多个分类器再集成，并不能保证得到一个准确率较高的集成分类器，而在判别嵌入的子空间上训练基础分类器，可以得到一个准确率较高的集成分类器(LGC 作为基础分类器也有同样结论，限于篇幅没有报告)。

4.2 参数敏感性分析

本节通过在 AR 上进行实验来分析 MGTEC 的参数敏感性。在实验中，如未作特别说明，Lp 设置为 10，$p=2C$，EnsK 为 20。

实验 1 高斯热核相似度对热核宽度 σ 较敏感，为了研究各种算法在不同 σ 时的性能，在 AR 上执行一个实验。在实验中，σ 的取值为 1, 10, 100, 1000, 10000, 100000, 1000000, 10000000, 100000000，对应的实验结果如图 3。

从图 3 可以看出，依赖于高斯热核相似度的直推分类方法，都需要选择一个合适的 σ。LNP 不需要设定 σ，故没有报告其结果。当 σ 过小时，所有的 $\boldsymbol{W}_{ij}$ 趋近或等于 0，此时的 σ 一般不选取。无论选择哪个指标下的 σ, LGC, GFHF 均没有获得显著的分类性能，而 MGTEC 可以在绝大多数指标下保持

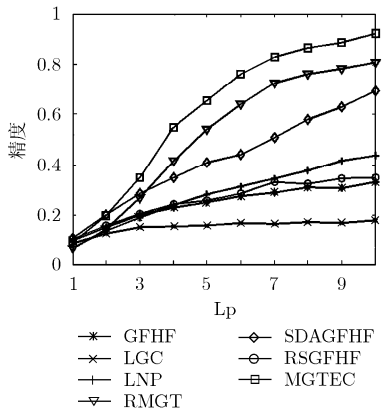

图 1 AR 上不同 Lp 下的分类准确率

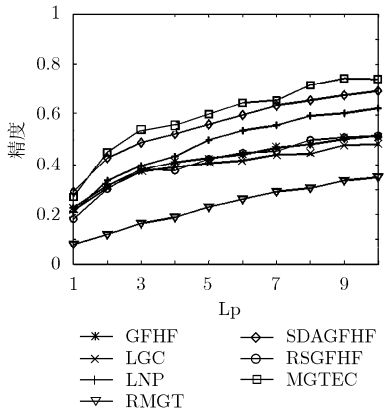

图 2 yaleB 上不同 Lp 下的分类准确率

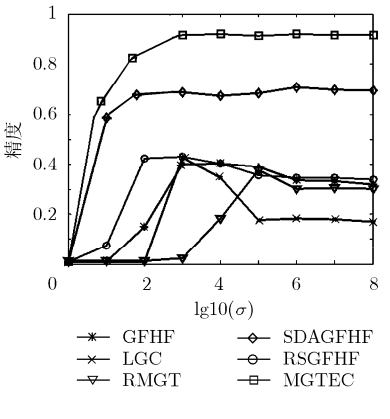

图 3 不同 σ 下的分类正确率

稳定，这说明 MGTEC 对参数选取的鲁棒性比这些对比方法要好。SDAGFHF 由于运用了判别分析，获得的性能也较稳定，这说明降维不仅可以提高分类正确率，在一定程度上还可以稳定分类器性能。

实验 2　基于随机子空间的集成分类方法对 p 依赖性较高。为了分析不同的 p 对 MGTEC 的影响，本文在 AR 上执行一个实验。在本实验中，p 从 20 增大到 300，图 4 为不同 p 时分类器的性能变化情况。

从图 4 可以看出，无论何种 p 下，RSGFHF 的性能都没有大的变化，这说明通过直接在随机子空间上训练基于图的直推分类并不能保证获得一个稳定且性能较好的分类器。由于 SDA 的判别子空间大小为 $C-1$，因而通常 $p \geq C$。当 $p \geq 2C$ 时，MGTEC, RSGFHF 的性能均有所下降，原因是随着 p 的增大，噪声和冗余特征也随之增多。在 $2C \geq p \geq C$ 时，MGTEC 均表现出较高和稳定的正确率，这表明 MGTEC 可在一定的原则下选择合适的 p，可选择的 p 也较多。

实验 3　基础分类器的个数(此处等同于 EnsK)影响着基于随机子空间的集成分类器性能。为此本文继续在 AR 上执行一个实验，在实验中，EnsK 从 1 增加到 50，图 5 为不同 EnsK 时的分类结果。

从图 5 可以看出，尽管 RSGFHF 的基础分类器不断增多，但其性能并没有显著变化，而 MGTEC 可以在较少的基础分类器时达到较高且稳定的精度。在只有 1 个基础分类器时，MGTEC 的分类器精度可以达到 0.8，而 RSGFHF 接近 0.3，这说明在随机子空间上训练的分类器精度并不高，通过投票得到的集成分类器性能也不显著。本实验还表明基础分类器的精度对集成分类器的精度起着关键作用，仅增加基础分类器的个数并不能很好的提高集成分类器的性能。

实验 4　为了证明 MGTEC 的基础分类器不仅具有较高的精度还有较好的差异度。图 6 给出了 MGTEC 和 RSGFHF 的 20 个基础分类器各自的分类正确率及其集成后的分类正确率。

图 6 中 RSGFHF(I) 和 MGTEC(I) 分别为 RSGFHF 和 MGTEC 的基础分类器。从图 6 可以发现，通过投票策略获得的集成分类器均获得了较基础分类器更好的性能，这不仅表明基于投票的集成策略是有效的，而且说明这些基础分类器之间具有较大的差异性。由于随机子空间较少受噪声和冗余特征的影响，通过 SDA 获得的判别子空间能把在子空间中的不同类样本区分开来，从而使得在这些子空间上训练的基础分类器拥有较高的精度。同时，随机子空间之间的差异性在判别分析后得以保留，使得判别子空间上的分类器之间也具有较大的差异性。因此，MGTEC 具有比 RSGFHF 更好的分类性能。

5　结论

针对在高维数据上构造的图易受噪声和冗余特征影响，而基于图的直推分类器的性能依赖于图结构的难题，本文提出一种基于多图的集成直推分类方法 MGTEC。MGTEC 首先在随机子空间上进行半监督判别分析，再在这些判别嵌入后的子空间上训练基础分类器，最后通过投票策略集成这些分类器。实验表明，通过这种方式构造的基础分类器具有较高的精度，它们之间还有较好的差异性，获得的集成分类器具有更高的精度，对参数的选取也较鲁棒。本文提出的多图构建策略直观，可应用到绝大多数基于图的半监督学习方法中。

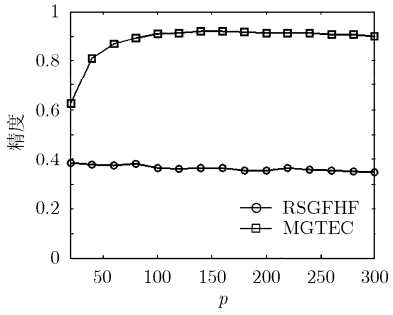

图 4　不同 p 下的分类正确率

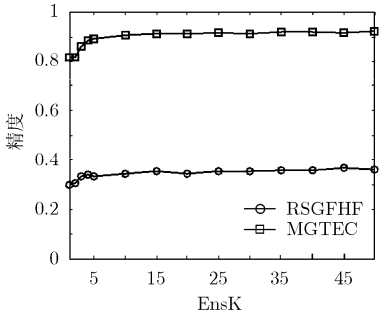

图 5　不同 EnsK 下的分类正确率

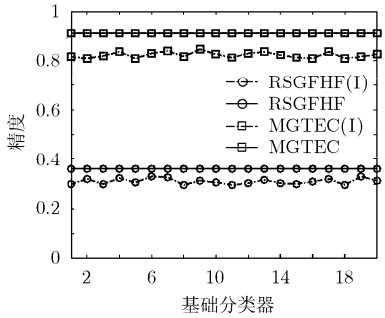

图 6　基础分类器及其集成后的分类正确率

参 考 文 献

[1] Zhu X J. Semi-supervised learning literature [R]. Technical Report 1530, Department of Computer Sciences, University of Wisconsin-Madison, 2008.

[2] Zhu X J, Ghahramani Z, and Lafferty J. Semi-supervised learning using Gaussian fields and harmonic functions [C]. Proceedings of the 20th International Conference on Machine Learning (ICML), Washington, DC, USA, 2003: 912–919.

[3] Zhou D Y, Bousquet O, Lal T N, *et al*.. Learning with local and global consistency [C]. Advances in Neural Information Processing Systems (NIPS), Vancouver and Whistler, British Columbia, Canada, 2003: 257–264.

[4] Wang J D, Wang F, Zhang C S, *et al*.. Linear neighborhood propagation and its applications [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009, 31(9): 1600–1615.

[5] Liu W and Chang S F. Robust multi-class transductive learning with graphs [C]. Proceedings of IEEE 22th Computer Vision and Pattern Recognition (CVPR), Miami, Florida, USA, 2009: 381–388.

[6] Jebara T, Wang J, and Chang S F. Graph construction and b-matching for semi-supervised learning [C]. Proceedings of the 26th International Conference on Machine Learning (ICML), Montreal, Quebec, Canada, 2009: 441–448.

[7] Parsons L, Haque E, and Liu H. Subspace clustering for high dimensional data: a review [J]. *ACM SIGKDD Explorations Newsletter*, 2004, 6(1): 90–105.

[8] Zhou Z H. When semi-supervised learning meets ensemble learning [C]. Proceedings of 8th International Workshop on Multiple Classifier System (MCS), Reykjavik, Iceland, 2009: 529–538.

[9] Ho T K. The random subspace method for constructing decision forests [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998, 20(8): 832–844.

[10] Zhu Y L, Liu J, and Chen S C. Semi-random subspace method for face recognition [J]. *Image and Vision Computing*, 2009, 27(9): 1358–1370.

[11] Cai D, He X F, and Han J W. Semi-supervised discriminate analysis [C]. Proceedings of IEEE 11th International Conference on Computer Vision (ICCV), Rio de Janeiro, Brazil, 2007: 1–7.

[12] He X F and Niyogi P. Locality preserving projections [C]. Advances in Neural Information Processing Systems (NIPS), Vancouver and Whistler, British Columbia, Canada, 2003: 153–160.

[13] Yan S C, Xu D, Zhang B Y, *et al*.. Graph embedding and extensions: a general framework for dimensionality reduction [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007, 29(1): 40–51.

[14] Martinez A M and Kak A. PCA versus LDA [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002, 23(2): 228–233.

[15] 韦佳, 彭宏. 一种基于局部与全局保持的半监督维数约减方法[J]. 软件学报, 2008, 19(11): 2833–2842.
Wei Jia and Peng Hong. Local and global preserving based semi-supervised dimensionality reduction method [J]. *Journal of Software*, 2008, 19(11): 2833–2842.

[16] 曾宪华, 罗四维, 王娇, 等. 基于测地线距离的广义高斯型 Laplacian 特征映射[J]. 软件学报, 2009, 20(4): 815–824.
Zeng Xian-hua, Luo Si-wei, Wang Jiao, *et al*.. Geodesic distance-based generalized Gaussian laplacian eigenmap[J]. *Journal of Software*, 2009, 20(4): 815–824.

[17] 魏莱, 王守觉. 基于流形距离的半监督判别分析[J]. 软件学报, 2010, 21(10): 2445–2453.
Wei Lai and Wang Shou-jue. Semi-supervised discriminate analysis based on manifold distance [J]. *Journal of Software*, 2010, 21(10): 2445–2453.

[18] 张向荣, 阳春, 焦李成. 基于 Laplacian 正则化最小二乘的半监督 SAR 目标识别[J]. 软件学报, 2010, 21(4): 586–596.
Zhang Xiang-rong, Yang Chun, and Jiao Li-cheng. Semi-supervised SAR target recognition based on laplacian regularized least squares classification [J]. *Journal of Software*, 2010, 21(4): 586–596.

[19] 王娇, 罗四维, 曾宪华. 基于随机子空间的半监督协同训练算法[J]. 电子学报, 2008, 36(12A): 60–65.
Wang Jiao, Luo Si-wei, and Zeng Xian-hua. A random subspace method for co-training [J]. *Acta Electronic Sinica*, 2008, 36(12A): 60–65.

[20] Wang X G and Tang X O. Random sampling for subspace face recognition [J]. *International Journal of Computer Vision*, 2006, 70(1): 91–104.

[21] Martinez M and Benavente R. The AR face database [R]. CVC Technical Report 24, 1998.

[22] Georghiades S, Belhumeur P N, and Kriegman D J. From few to many: illumination cone models for face recognition under variable lighting and pose [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2001, 23(6): 643–660.

余国先： 男，1985 年生，博士生，研究方向为半监督学习及其在高维数据中的应用.

张国基： 男，1953 年生，教授，博士生导师，研究方向为计算智能、密码学.