文章编号: 1001-0920 (2015) 09-1597-06 DOI: 10.13195/j.kzyjc.2014.0843

基于特征空间变换的纠错输出编码

雷 蕾[a], 王晓丹[a], 罗 玺[b], 宋亚飞[a], 薛爱军[a]

(空军工程大学 a. 防空反导学院，b. 信息与导航学院，西安 710051)

摘 要: 针对基于纠错输出编码多类分类中如何保证基分类器差异性的问题, 提出一种基于特征空间变换的编码方法. 该方法引入特征空间, 将编码矩阵扩展成三维矩阵; 然后基于二类划分, 利用特征变换得到不同的特征子空间, 从而训练得到差异性大的基分类器. 基于公共数据集的实验结果表明: 该方法能够比原始的编码矩阵获得更优的分类性能, 同时增加了基分类器的差异性; 该方法适用于任何编码矩阵, 为大数据的分类提供了新的思路.

关键词: 纠错输出编码；特征空间；基分类器独立性

中图分类号: TP391 **文献标志码:** A

Error-correcting output codes based on feature space transformation

LEI Lei[a], *WANG Xiao-dan*[a], *LUO Xi*[b], *SONG Ya-fei*[a], *XUE Ai-jun*[a]

(a. School of Air and Missile Defense，b. School of Information and Navigation，Air Force Engineering University，Xi'an 710051，China．Correspondent：LEI Lei，E-mail：wendyandpaopao@163.com)

Abstract: The independency between each dichotomizer trained by coding matrix's bi-partition is the key to using error-correcting output codes(ECOC) to solve multiclass problems. Therefore, an error-correcting output codes method based on feature space transformation(FST) is proposed. Inspired by the ensemble learning theory, a third feature space dimension is introduced into the coding matrix. Then, different subspaces are obtained by feature space transformation based on different positive and negative subclasses, so that the diversity between different binary classifiers are promoted to make the classification performance better. The experiment results based on UCI datasets show that the codes based on FST are better than the original codes. Besides, the proposed method can be applied to any kind of coding matrix, and provides new thought to large dataset for its quick training time and simplicity.

Keywords: error-correcting output codes；feature space；independence of dchotomizer

0 引 言

多类分类是模式识别在实际应用中所面临的难题, 已受到众多机器学习研究者的重视. 而纠错输出编码 (ECOC)[1-2]作为一种 "divide-and-conquer" 策略, 能利用其特殊的分解框架, 有效地将多识别问题转化为一系列模式识别常用的二类分类子问题进行求解, 从而简化多模式分类问题的复杂性, 并且使多类分类问题能利用较为成熟的二类模式分类方法加以解决. 而ECOC 框架和集成学习有诸多相似之处, 例如都包含多个基分类器, 并且需要对每一个基分类器进行学习, 在决策过程中同样都存在对各基分类器结果融合的问题, 甚至可以将ECOC 看成是一种特殊的集成学习框架. 已有很多学者将集成结论应用于ECOC 分类中. Kong[1]认为与一般的分类器集成相比, 基于ECOC 的多类分类集成能在基分类器为非本地分类器 (若一个基分类器为稳定的分类器则称该类型分类器为非本地分类器) 时同时减少偏差和方差, 从而减少分类错误率, 而非本地分类器则保证了此类分类器所产生的错误是不相关的. 当基分类器为本地分类器时, Francesco 等[3]讨论了一种通过选择不同特征集来训练本地分类器以减少分类器错误率的相关性, 从而实现了基于本地分类器的ECOC 多类分解方法. Masulli 等[4]分析了ECOC 的结构后认为: 编码矩阵的结构与基于该编码所产生的基分类器之间的相关性是影响此类方法性能的主要因素. Matthew 等[5]讨论了利用多层感知器作为二分器来获得差异性大的基分类器, 从而能有效地减少过拟合带来的误差. Garcia 等[6]在对ECOC 框架进行了大

收稿日期: 2014-05-27；**修回日期:** 2014-08-12.

基金项目: 国家自然科学基金项目(60975026, 61273275).

作者简介: 雷蕾(1988−), 女, 博士生, 从事智能信息处理、多类分类的研究； 王晓丹(1966−), 女, 教授, 博士生导师, 从事智能信息处理、机器学习等研究.

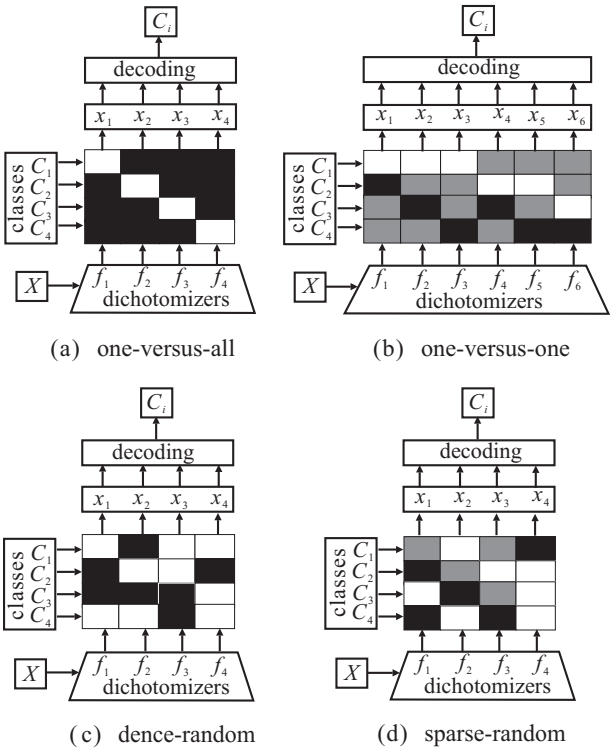
量实验后得出结论: ECOC框架的有效性基于其基分类器的差异性. Mohammad等[7]为提高框架中二分器的差异性, 将特征子空间的思想引入ECOC编码矩阵中对特征进行编码, 从而形成不同的特征子空间训练分类器, 在得到的一系列编码矩阵中选择最小Hamming距离最大的编码矩阵作为最佳选择. 为了缩短在整个特征空间的搜索时间, 其又将遗传算法引入ECOC编码矩阵的寻优过程中[8]. 针对遗传算法在空间搜索上的复杂性和盲目性, Miguel等[9]研究了新的ECOC编码矩阵个体的交叉和变异方法, 保证交叉和变异后的编码矩阵不仅具有较高的差异性, 而且不出现新的无效编码, 从而简化了特征空间, 提高了收敛速度. 周进登等[10]提出的基于混淆矩阵的自适应纠错输出编码方法利用混淆矩阵计算多类问题中各类别的相关性, 从而找出有利于分类的二类划分, 本质上也是通过形成最佳的类别划分构成差异性大的基分类器. 相关的研究还有文献[11-13]等, 这些成果都有力的促进了基于ECOC的多类分类的发展。

针对如何在编码矩阵构造中生成差异性较大的基分类器的问题, 受集成理论特征空间变换的思想启发, 本文提出一种基于特征变换的纠错输出编码(FST-ECOC). 该编码矩阵被扩展成具有特征空间的三维矩阵, 通过对二类划分进行特征变换形成特征子空间, 从而利用这些不同的特征子空间训练得到差异性较大的基分类器. 该方法能使同一类别在不同二类划分中得到不同的特征子空间, 增加了基分类器的差异性, 同时也避免了优化算法在搜索最优编码矩阵时面临的时间复杂度问题, 并且该方法能扩展到任意的编码矩阵中, 不受事前编码和基于数据编码的影响. 实验结果表明, 该方法能在大大减少训练时间的情况下, 有效地提高基分类器间的差异性, 从而提高整体的分类性能.

1 纠错输出编码(ECOC)

ECOC框架用一种二元$(-1,+1)$或三元$(-1,0,+1)$的编码矩阵实现多类类别分解和基分类器集成. 其中: -1代表一类, $+1$代表另一类, 0表示该码字位所对应的类在其列所形成的二类划分中被忽略(即不参与由该列所产生的基分类器的训练)[8]. 图1给出了4种常见的基于ECOC的分类, 以编码矩阵来区分, 分别是: “一对多”编码阵(one-versus-all)、“一对一”编码阵(one-versus-one)、密集随机阵(dense random)和稀疏随机阵(sparse random).

图1中所有编码矩阵的每一行对应着某一类$C_i(i=1,2,3,4)$的码字, 每一列代表训练样本的一个二类分类问题. 训练时, 每一个基分类器$f_i(i=1,2,\cdots,6)$先根据其所在列对应的编码将训练样本集划分为二类分类问题, 然后进行训练, 最后按照某种解码策略进行融合.



(a) one-versus-all (b) one-versus-one

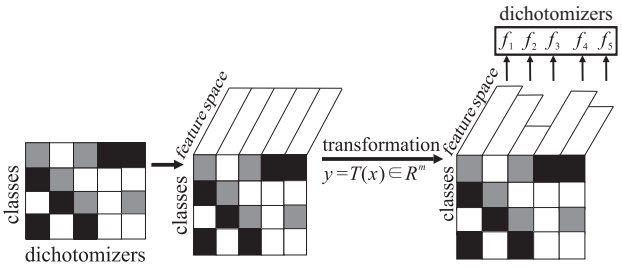
(c) dence-random (d) sparse-random

图1 4种常见的ECOC

2 基于特征空间变换的纠错输出编码FST-ECOC

大量的研究表明, 在编码矩阵的设计过程中, 一个关键特性是基分类器之间的独立性, 从而使基分类器具有不同的分类能力, 产生的分类错误率不发生叠加, 才能保证其在解码阶段的有效性. 很多研究通过设定目标函数的优化方法, 设计出行列分离程度较大的编码矩阵, 从而使训练得到的基分类器差异性增大[8-9]. 但这类方法都面临着优化算法时间复杂性的问题, 不可避免地增加了训练的难度.

针对基分类器差异性问题, 同时为避免搜索算法带来的庞大计算量, 受集成理论的启发, 将特征空间的概念引入编码矩阵构成三维矩阵, 在特征空间维上, 对不同二类划分生成的训练数据集进行特征空间变换, 从而得到不同的特征子空间, 在这些子空间上对分类器进行训练. 在进行特征变换以后, 每个基分类器对应的训练数据空间不同, 从而提高了基分类器之间的差异性. 即使同一类别经过特征变换后特征空间也发生了变化, 从而大大地减少了分类器之间的相关性. 与原始的编码矩阵相比较, 基于特征空间变换的编码矩阵不仅增加了特征空间, 而且在训练二分类器时增加了特征空间的差异性, 如图2所示.



(a) original ECOC (b) three dimensional ECOC (c) FST-ECOC

图 2　编码矩阵

图 2 中 $\boldsymbol{X} \in \boldsymbol{R}^n$, $\boldsymbol{y} = T(\boldsymbol{X}) \in \boldsymbol{R}^m$ 为特征变换, 表示将样本从 n 维空间映射到 m 维新空间, T 为线性变换或非线性变换. 从图 2 中可以看出, 原始编码矩阵通过扩展特征空间维形成三维编码矩阵, 利用特征变换算法对特征空间进行变换, 再利用变换后的特征空间完成基分类器的训练, 从而提高基分类器的差异性, 由此得到的二分器的相关性也随之降低. 同时, 特征空间作为编码矩阵的第 3 维, 可适用于任何编码矩阵. 本质上而言, 基于特征空间变换的编码是在编码矩阵确定后, 通过对训练数据集的特征扰动来提高基分类器的差异性, 属于集成理论在 ECOC 多类分类中的应用, 其具体实现步骤如下.

算法 1　基于特征空间变换的纠错输出编码方法.

输入: 初始编码矩阵 $M_{k\times l}$, 训练样本集, 特征变换算法;

输出: 3 维的 FST-ECOC 矩阵.

Step 1: 训练样本集根据初始的编码矩阵 $M_{k\times l}$, 形成二类划分;

Step 2: 利用特征变换算法对每个二类划分中的正负类样本进行特征空间变换, 形成基分类器的特征子空间;

Step 3: 将特征子空间并入编码矩阵中, 形成第 3 维, 利用每一列的子空间训练基分类器 f_i;

Step 4: 将 f_i 加入最后解码的分类器中.

至此, 便形成了基于特征空间变换的纠错输出编码方法, 下面将重点通过实验来验证该方法在分类中的有效性和实用性.

3　实验分析

为了比较基于 FST 的编码方法与原始编码方法的优劣, 采用 UCI 数据集来验证本文方法的分类效果. 将从实验数据、设计和结果及分析分别加以阐述.

3.1　实验数据

实验所用的 UCI 数据集及各类数据描述如表 1 所示 (C 代表 continous, B 代表 binary, N 代表 nominal), 并给出了其在不同编码下的长度.

表 1　UCI 数据集及描述

Dataset	Cases	Classes	F(C,B,N)	Code length of different ECOC			
				OVA	OVO	Dense	Sparse
Ecoli	336	8	7\-\1	8	28	30	45
Glass	214	6	9\-\-	6	15	26	39
Iris	150	3	4\-\-	3	3	3	3
Satimag	6 435	6	36\-\-	6	15	26	39
segment	2 310	7	19\-\-	7	21	7	7
Vowel	990	11	10\-\-	11	55	11	11
Yeast	1 484	10	8\-\-	10	45	10	10
Zoo	101	7	1\15\-	7	21	7	7

3.2　实验设计

实验分别选取了经典的 one-versus-one 编码、one-versus-all 编码、Dense random 密集随机编码和Sparse random 稀疏随机编码进行实验, 在这些编码的基础上通过特征空间变换形成 FST-ECOC 编码. 在选择两种随机编码方法时, 分别从已产生的 2 000 个密集及稀疏随机编码阵集 (对应各码元概率分别为: $p(-1) = 0.5$, $p(+1) = 0.5$; $p(-1) = 1/3$, $p(0) = 1/3$, $p(+1) = 1/3$) 中随机选择所需要的编码阵. 在对解码策略和基分类器选择时, 采用 Hamming 距离解码和两种不同基分类器: 线性逻辑分类器 (LOGLC) 和多项式核函数支持向量机 ($C = 2$), 它们将分别用于各编码矩阵分类效果的比较.

在选择特征空间变换算法时, 采用经典的主成分分析 (PCA) 和线性判别分析 (LDA) 进行特征变换.

PCA 是在最小均方意义下寻找最能有效表示原始数据的投影方法, 是一种无监督的特征变换. PCA 理论认为样本集在 d 空间中形成了一个 d 维椭球体形状的数据堆, 其目的是寻找这个高维椭球体最长的 m 个主轴方向, 从而构成投影矩阵, 完成对原始数据集的特征变换. 具体实现过程描述如下:

设 X 为 d 维样本集, 其中的每个样本 x_i ($i = 1, 2, \cdots, N$) 都是 d 维空间中的一个点, $x_i = (x_{i1}, x_{i2}, \cdots, x_{id})$, m 为提取的主成分个数.

计算样本集 X 的协方差矩阵

$$Q = \frac{1}{N}(X - \mu)(X - \mu)'. \tag{1}$$

其中: μ 为 X 的均值, N 为样本数.

求 Q 的特征值 λ_i, $i = 1, 2, \cdots, d$, 并按 λ_i 的大小进行排列, 取前 m 个特征 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_m$, 并计算其对应的特征向量 $\nu_1, \nu_2, \cdots, \nu_m$, 记其构成矩阵 $\boldsymbol{V}$. 记第 h ($h = 1, 2, \cdots, m$) 个主成分为 t_h, 有 $t_h = \boldsymbol{X}\nu_h = \sum_{i=1}^{d} \nu_{hi} x_i$. 如何决定经过 PCA 变换后保留的主成分个数, 即确定压缩后的特征维数的大小? 本文采用累计贡献率 (ACR) 来决定.

前 m 个主元的累计贡献率 $\eta(m)$ 定义为

$$\eta(m)=\sum_{k=1}^{m}\lambda_k\Big/\sum_{i=1}^{d}\lambda_i. \tag{2}$$

累计贡献率显示了新生成分量对原始数据信息保存的完整程度.

线性判别分析利用了样本的类信息, 将数据从高维输入空间映射到一个低维空间. 在低维空间中, 样本数据具有较好的区分性. 不同于主成分分析法, 线性判别分析不是直接对所有样本的协方差矩阵进行特征值分解, 而是对带有类标签信息的统计量进行分析, 属于有监督的特征变换. LDA 通常基于 Fisher 原理, 寻找最优的变换矩阵 $\boldsymbol{P}$ 使得变换后样本空间的类间散布矩阵与变换后的类内散布矩阵的比值最大[14], 即

$$\boldsymbol{P}_{\text{optimal}}=\arg\max_{\boldsymbol{P}}\frac{\boldsymbol{P}^T\boldsymbol{S}_b\boldsymbol{P}}{\boldsymbol{P}^t\boldsymbol{S}_w\boldsymbol{P}}.$$

将该优化问题转化为一个广义特征向量问题 $\boldsymbol{S}_b\zeta=\lambda\boldsymbol{S}_w\zeta$, 最优变换矩阵 $\boldsymbol{P}$ 的 m 个列向量即为上述最大的 m 个广义特征值对应的广义特征向量.

最后, 实验对基于原始编码和 FST 编码的基分类器的差异性进行了比较. 在估计分类错误率时为保证估计的准确性, 样本数据小于 500 时采用 5 重交叉验证来进行, 大于 500 时采用 10 重交叉验证, 并利用双边估计 t 检验法来计算置信水平为 0.95 的分类错误率置信区间, 将其作为最终结果. 计算公式如下:

$$\frac{|\overline{x}-\mu|}{\dfrac{\sigma}{\sqrt{(n)}}}\geqslant t_{0.025}(n-1). \tag{3}$$

其中: μ,σ 分别表示 n 重交叉验证的均值和标准差, $t_{0.025}(4)=2.776\,4$, $t_{0.025}(9)=2.262\,2$. 实验中所采用的基分类器和函数均来自于 PRTool 工具箱. 实验机器配置为 2 G 内存, 2.80 G CPU, 算法基于 Matlab7.10 (R2010a) 实现.

3.3 实验结果及分析

3.3.1 分类结果比较

图 3 给出了 3 种数据集 Iris、Ecoli、Wine 在原始情况下和分别经 PCA、LDA 特征变换后的二维数据分布. 从左到右依次为原始数据分布, 经 PCA 作用后的数据分布和经 LDA 变换后的数据分布. 从图 3 中可以看出, 经特征变换算法作用后的数据分布的重叠要明显小于原始数据, 对于类别数较多的 Ecoli 数据集, 特征变换算法也能将其大致分为几个子类, 使得类与类之间的分界线较为清晰. 因此, 通过这样的特征变换后获得的各个基分类器的特征空间差异性将增加, 由此训练得到的基分类器的独立性也将增强. 基于 PCA 特征变换的分类结果将在实验表格中列出.

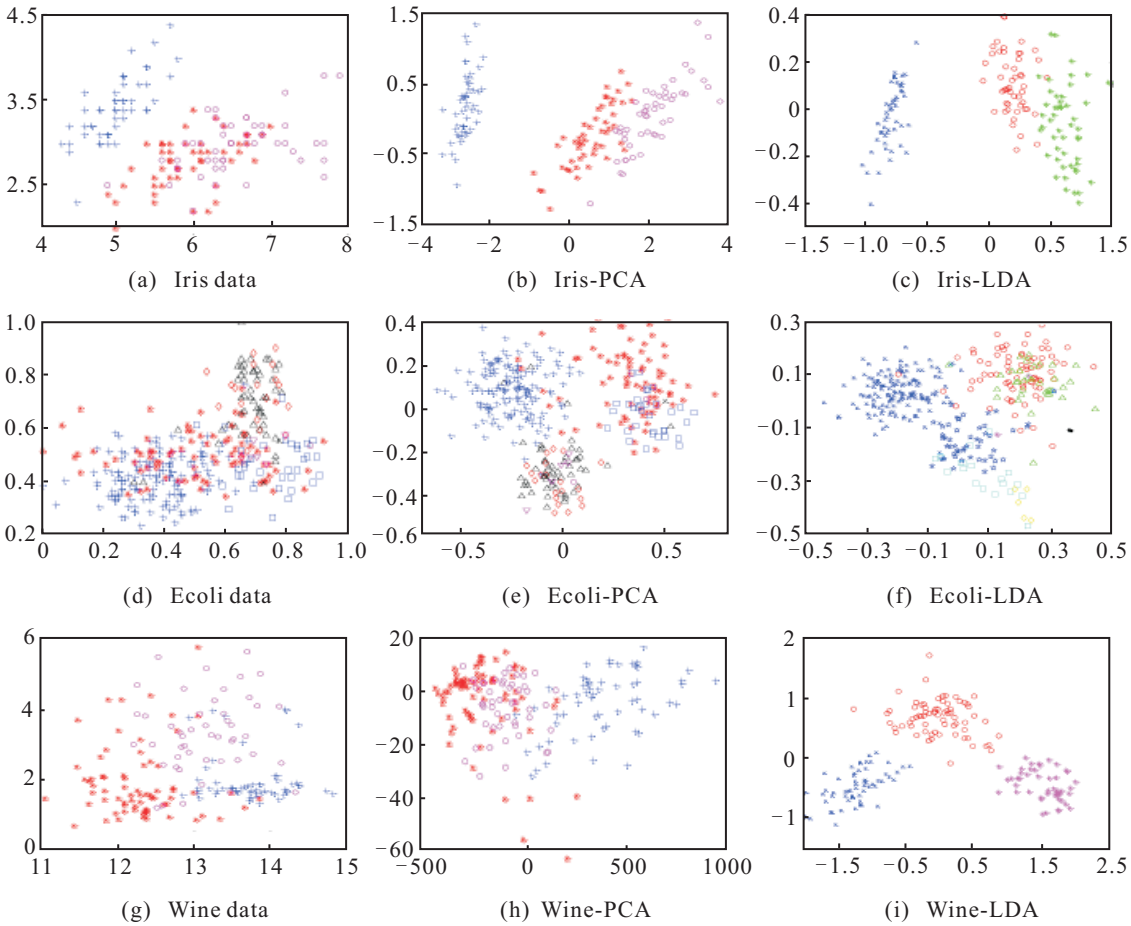


图 3 原始数据和经 PCA、LDA 变换作用后的数据分布

表 2　基于 LOGLC 和 Hamming 距离解码的各数据集分类正确率及置信区间为 0.95 的置信区间　%

数据集	OVA	OVA-FST	OVO	OVO-FST	Dense	D-FST	Sparse	S-FST
Ecoli	57.19±5.51	57.44±4.78	34.77±4.12	40.50±4.95	76.81±3.01	88.68±5.74	77.40±7.06	77.07±8.10
Glass	60.30±5.85	59.78±5.32	39.18±9.62	55.11±9.61	91.60±3.46	92.99±7.78	82.10±7.06	90.17±2.49
Iris	80.71±9.22	82.65±2.35	80.80±14.62	74.32±4.32	94.07±3.85	96.53±1.09	97.17±7.04	96.32±0.23
Satimag	71.18±0.04	76.18±0.34	46.88±0.41	64.64±6.30	89.08±0.04	90.27±0.24	89.08±0.34	87.83±2.45
Segment	85.71±0.09	87.21±0.13	70.56±3.73	74.85±3.97	83.57±0.07	85.71±0.26	81.99±1.10	85.67±0.10
Vowel	89.49±3.70	90.91±0.25	72.42±2.34	89.90±1.02	90.91±0.23	90.92±0.22	95.15±1.59	90.51±0.51
Yeast	67.99±1.88	67.18±1.80	56.13±2.88	68.46±2.90	84.98±5.77	71.09±1.72	96.97±0.89	97.31±1.64
Zoo	54.55±8.64	—	30.73±8.60	—	94.00±5.00	—	89.09±6.28	—

表 3　基于 SVM 和 Hamming 距离解码的各数据集分类正确率及置信区间为 0.95 的置信区间　%

数据集	OVA	OVA-FST	OVO	OVO-FST	Dense	D-FST	Sparse	S-FST
Ecoli	57.38±6.51	56.84±5.35	31.90±5.39	42.87±7.13	90.46±4.92	76.77±5.20	91.71±5.04	94.96±4.58
Glass	59.35±7.01	61.26±5.08	45.54±8.99	47.13±9.43	88.77±5.05	90.78±5.07	90.63±3.23	89.00±9.37
Iris	87.18±10.3	89.36±2.63	83.26±2.13	84.10±0.64	86.20±8.81	87.28±2.06	83.79±1.63	82.36±0.86
Satimag	76.18±0.04	75.26±0.32	77.73±0.06	78.96±0.74	65.78±2.78	89.01±0.03	78.90±0.03	90.27±0.04
Segment	83.73±2.98	85.11±0.76	82.75±2.06	78.80±0.39	86.17±1.77	85.29±0.19	81.47±1.82	85.05±0.79
Vowel	88.86±0.91	90.91±0.00	88.69±1.56	89.60±1.60	88.41±1.52	89.39±4.25	88.36±1.57	90.10±0.13
Yeast	67.79±2.28	67.79±2.47	56.07±1.70	67.65±2.75	68.80±3.52	96.84±2.49	71.09±1.53	83.56±1.84
Zoo	58.27±6.00	—	28.73±8.26	—	59.45±13.55	—	63.45±11.46	—

实验将采用 4 种经典的编码方法与在它们基础上形成的基于 FST 的编码对公共数据集的分类结果进行比较. 表 2 和表 3 分别列出了在 Hamming 距离解码的基础上, 以多项式 SVM 和 LOGLC 作为基分类器的 4 组解码方法的分类结果, 其中 OVA-FST 表示在一对多编码基础上形成的基于特征空间变换的编码方法. 以此类推, 得到 OVO-FST、D-FST 和 S-FST. 表 2 和表 3 为基于 PCA 的特征变换.

从表 2 和表 3 中的结果可以看出, 基于 FST 变换的编码方法在大多数情况下能获得更优的分类结果, 其中以特征维数较大的数据集尤为明显, 比如: Satimag、Segment 和 Vowel 数据集, 特征维数越高, 作用越明显. 这是因为, PCA 能从大量的特征数据中提取最能表征数据的特征, 从而减少训练时间. 对于数据特征维数较低的数据集, 如 Ecoli 和 Glass 数据集, 其分类效果略有提升, 但不是很显著. 而对于某些特殊的数据集, 例如 Zoo 数据集, 原始编码和基于特征空间变换的编码方法的分类效果基本无差异, 这是由数据集的特殊性决定的 (Zoo 数据集中, 符号表示的特征维数居多, 在将符号转换为数据表示时会对 PCA 的矩阵变换产生一定的影响).

3.3.2　基分类器差异性比较

为进一步总结 FST 方法的优势, 将从统计学的角度采用 Yule 的 Q 统计量[5]对基分类器之间的差异性进行比较.

对于分类器 C_i 和 C_k, 两者之间的差异性度量值可用 Q 统计量来表示, 即

$$Q_{ik}=\frac{N^{11}N^{00}-N^{01}N^{10}}{N^{11}N^{00}+N^{01}N^{10}}. \tag{4}$$

其中 N^{ab} 的含义见表 4. N^{11} 表示正类样本被正确划分为正类的个数, N^{10} 表示正类被错分为负类的样本数; N^{01} 表示负类样本被错分为正类的个数, N^{00} 表示负类被正确划分为负类的样本数.

表 4　N^{ab} 的含义

	$D_k\text{correct}(1)$	$D_k\text{wrong}(1)$
$D_i\text{correct}(1)$	N^{11}	N^{10}
$D_w\text{rong}(1)$	N^{01}	N^{00}

由式 (4) 可以看出, 对于识别同一类别的基分类器, 其 Q 统计量的值为正, 否则为负; 相互独立的基分类器, 其 Q 值为零. 对于 L 个基分类器, 可用平均值来衡量, 即

$$Q_{av}=\frac{2}{L(L-1)}\sum_{i=1}^{L-1}\sum_{k=i+1}^{L}Q_{ik}. \tag{5}$$

表 5 给出了在所有数据集上的基分类器差异性比较的结果. 其中第 1 行为 4 种基分类器; Q_{av} 为 L 个基分类器在所有数据集上的 Q 平均值; "S" 对应的行给出的是 win\tie\loss 统计量, 分别表示所比较的两种算法前一行的 Q_{av} 大于后一行, 等于和小于后一行的数据集个数. FST 采用的是 LDA 变换.

从表 5 中的实验结果可以看出, 基于特征空间变换 (FST) 的编码方法在大部分训练数据集上的基分类器均值要比原始的编码方法更接近于 0, 同时在数据集的比较结果上占优, 从而可以认为基于 FST 编码方法训练得到的不同基分类器都具有最大的差异性. 这

是因为, 由编码矩阵的二类划分得到的正负类数据集在扩展到特征空间后, 经过特征变换增加了不同训练数据集之间的差异性, 即使是同一类样本在不同二类划分的基础上得到的特征变换也是有所差异的, 由此得到的基分类器之间的差异性也就应该更明显.

表 5 基分类器差异性比较

algorithms		C4.5	LOGLC	RBFSVM	Poly-SVM
OVA	Q_{av}	0.002 3	0.000 0	0.326 5	0.044 1
OVA-FST	Q_{av}	0.002 0	0.000 0	0.318 9	0.040 7
	S	1\0\7	2\0\6	3\0\5	2\0\6
OVO	Q_{av}	0.004 7	0.123 5	0.489 0	0.200 7
OVO-FST	Q_{av}	0.002 9	0.102 5	0.473 0	0.200 5
	S	3\0\5	2\0\6	2\0\6	3\0\5
Dense	Q_{av}	0.032 5	0.052 6	0.213 4	0.828 5
D-FST	Q_{av}	0.028 8	0.042 3	0.146 9	0.808 0
	S	1\0\7	2\0\6	1\0\7	1\0\7
Sparse	Q_{av}	0.001 5	0.012 5	0.337 0	0.205 0
S-FST	Q_{av}	0.000 6	0.012 6	0.333 3	0.200 7
	S	2\0\6	1\0\7	2\0\6	2\0\6

基于FST的纠错输出编码的时间复杂性因其采用的特征空间变换算法的不同而不同. 以本文的LDA特征选择算法为例, 其时间主要消耗在基分类器训练阶段, 其训练时间复杂度为$O(d^3 + nd^2)$, 测试时间复杂度为$O(md)$, 其中d为样本特征维数, n和m分别为训练和测试样本数.

4 结 论

本文将ECOC框架与集成学习紧密相联, 针对ECOC编码矩阵中基分类器独立性的问题, 提出了基于特征空间变换的编码方法. 首先将特征空间引入编码矩阵中, 从而形成三维的矩阵; 再根据各个二类划分, 对正负类数据进行特征空间变换或扰动, 从而使训练得到的基分类器的差异性增加, 在整体上提高了编码的分类性能. 同时该方法作为第3维的扩展, 能应用到任何编码矩阵中.

参考文献(References)

[1] Kong E, Dietterich T G. Error correcting output codes corrects bias and variance[C]. Proc of the 21th Int Conf on Machine Learning. California, 1995: 313-321.

[2] Dietterich T G, Bakiri G Solving. Multi-class learning problems via error-correcting output codes[J]. J of Artificial Intelligence Research, 1995, 34(2):263-286.

[3] Francesco Ricci, David W Aha. Error correcting output codes for local learners[C]. Proc of the 10th European Conf on Machine Learning. Chemita: Springer, 1998: 280-291.

[4] Masulli F, Valentini G. Effectiveness of error correcting output coding methods in ensemble and monolithic learning machines[J]. Pattern Analysis and Application, 2003, 65(6): 285-300.

[5] Matthew Prior, Terry Windeatt. Over-fitting in ensembles of neural network classifiers within ECOC frameworks[J]. Lecture Notes in Computer Science, 2005, 3541(1): 286-295.

[6] Garcia-Pedrajas N, Ortiz-Boyer D. An empirical study of binary classifier fusion methods for multiclass classification[J]. Information Fusion, 2011, 12(2): 111-130.

[7] Mohammad Ali Bagheri, Gholam Ali Montazer, Ehsanollah Kabir. A subspace approach to error correcting output codes[J]. Pattern Recogniton Letters, 2013, 34(1): 176-184.

[8] Mohammad Ali Bagheri, Qigang Gao, Sergio Escalera. A gentic-based subspace analysis method for improving error-correcting output coding[J]. Pattern Recognition, 2013, 46(10): 2830-2839.

[9] Miguel Angel Bautista, Sergio Escalera, Xavier Baro, et al. On the design of an ecoc-compliant genetic algorithm[J]. Pattern Recogntion, 2014, 47(2): 865-884.

[10] 周进登, 王晓丹. 基于混淆矩阵的自适应纠错输出编码多类分类方法[J]. 系统工程与电子技术, 2012, 34(7): 220-226.
(Zhou J D, Wang X D. Multiclass classification of adaptive error-correcting output codes based on confusion matrix[J]. Systems Engineering and Electronics, 2012, 34(7): 220-226.)

[11] Zhou J D, Wang X D, Zhou H J, et al. Decoding design based on conditional probabilities in ternary error-correcting output codes[J]. Pattern Recognition, 2012, 45(4): 1342-1351.

[12] 雷蕾, 王晓丹, 罗玺, 等. ECOC多类分类研究综述[J]. 电子学报, 2014, 42(9): 1794-1800.
(Lei L, Wang X D, Luo X, et al. An overview of multi-classification based on error-correcting output codes[J]. Acta Electrinica Sinica, 2014, 42(9): 1794-1800.)

[13] 周进登, 王晓丹, 周红建, 等. 基于最小k近邻错分率编码确定方法及其在多类分类中的应用[J]. 控制与决策, 2011, 26(9): 1296-1302.
(Zhou J D, Wang X D, Zhou H J, et al. Designing of output codes based on minimal k nearest neighbor classifying error and its application in multi-class classification[J]. Control and Decision, 2011, 26(9): 1296-1302.)

[14] 边肇祺, 张学工. 模式识别[M]. 第2版. 北京: 清华大学出版社, 1999: 62-65.
(Bian Z Q, Zhang X G. Pattern recognition[M]. 2nd ed. Beijing: Tsinghua University Press, 1999: 62-65.)

(责任编辑: 孙艺红)