

一种基于几何关系的多分类器差异性度量及其在多分类器系统构造中的应用

梁绍一¹ 韩德强¹ 韩崇昭¹

摘要 多分类器系统是应对复杂模式识别问题的有效手段之一。当子分类器之间存在差异性 or 互补性时, 多分类器系统往往能够获得比单分类器更高的分类正确率。因而差异性度量在多分类器系统设计中至关重要。目前已有的差异性度量方法能够在一定程度上刻画分类器之间的差异, 但在应用中可能出现诸如“差异性淹没”等问题。本文提出了一种基于几何关系的多分类器差异性度量, 并在此基础上提出了一种多分类器系统构造方法, 同时通过实验对比了使用新差异性度量方法和传统方法对多分类器系统融合分类正确率的影响。结果表明, 本文所提出的差异性度量能够很好地刻画分类器之间的差异, 能从很大程度上抑制“差异性淹没”问题, 并能有效应用于多分类器系统构造。

关键词 多分类器系统, 差异性度量, 差异性淹没, 几何中心

引用格式 梁绍一, 韩德强, 韩崇昭. 一种基于几何关系的多分类器差异性度量及其在多分类器系统构造中的应用. 自动化学报, 2014, 40(3): 449–458

DOI 10.3724/SP.J.1004.2014.00449

A Novel Diversity Measure Based on Geometric Relationship and Its Application to Design of Multiple Classifier Systems

LIANG Shao-Yi¹ HAN De-Qiang¹ HAN Chong-Zhao¹

Abstract The multiple classifier system is one of the effective means to resolve pattern recognition under complicated environments. When the member classifiers are diverse or complementary, multiple classifier systems can usually obtain higher classification accuracy compared with a single classifier. Thus, diversity measures are crucial to multiple classifier systems design. Though the existing diversity measures can, to some degree, describe the difference among classifiers, they may lead to problems like “diversity submergence” in some cases. In this paper, a novel multiple classifier diversity measure based on geometric relationship and a multiple classifier system constructing method based on the new diversity measure are proposed. It is experimentally shown that the proposed diversity measure can well describe the diversity among classifiers and effectively suppress the problem of “diversity submergence”. It can also be effectively used in designing multiple classifier systems.

Key words Multiple classifier system, diversity measure, diversity submergence, geometric center

Citation Liang Shao-Yi, Han De-Qiang, Han Chong-Zhao. A novel diversity measure based on geometric relationship and its application to design of multiple classifier systems. *Acta Automatica Sinica*, 2014, 40(3): 449–458

收稿日期 2012-10-08 录用日期 2013-06-14
Manuscript received October 8, 2012; accepted June 14, 2013
国家重点基础研究发展计划 (2013CB329405), 国家自然科学基金 (61104214, 61203222), 国家自然科学基金创新群体 (61221063), 中国博士后科学基金 (201104670), 中央高校基本科研业务费专项资金 (xjj2012104) 资助
Supported by National Basic Research Program of China (973 Program) (2013CB329405), National Natural Science Foundation of China (61104214, 61203222), Foundation for Innovative Research Groups of the National Natural Science Foundation of China (61221063), China Postdoctoral Science Foundation-Special Fund (201104670), and the Fundamental Research Funds for the Central Universities (xjj2012104)

本文责任编辑 王聪
Recommended by Associate Editor WANG Cong
1. 西安交通大学电子与信息工程学院综合自动化研究所智能网络与网络安全教育部重点实验室 西安 710049
1. Ministry of Education Key Lab For Intelligent Networks and Network Security (MOE KLINNS Lab), Institute of Integrated Automation, School of Electronics and Information Engineering,

由于不同的分类方法具有自身的优势和局限性, 它们的精度和适用范围也有一定的限度, 故在解决复杂模式识别问题时, 很难选择出一种分类器使之能够在所有的应用中都有良好表现。为解决这样的问题, 许多研究者进行了有关多分类器系统的研究^[1-3]。通过多分类器系统不仅可提高分类准确性, 还能提高系统的效率和鲁棒性。目前, 多分类器系统已被广泛应用于生物认证^[4]、遥感地物分类^[5]、医疗诊断^[6]和自动目标识别^[7]等诸多领域。

针对多分类器系统中的融合规则和方法 (即如何将多个子分类器分类结果有效地进行组合), 1992 年, Xu 等^[8]按照子分类器的不同输出级别进行了研究分析。在他们工作的基础上, 研究者提出了多

Xi'an Jiaotong University, Xi'an 710049

种分类器融合方法,如贝叶斯法^[9]、行为知识空间法^[10]、逻辑回归法^[11]、投票法^[12]以及 D-S 证据理论方法^[13]等, Kittler 等^[14]对各种融合规则,特别是 Bayes 框架下的规则做了很好的梳理。

在多分类器系统中除融合规则或方法之外,子分类器之间的关系尤其是差异性显得更为重要。因为多个相同的或具有相近错分区域的分类器(即没有差异性的分类器)之间的融合是没有意义的。Ali 等^[15]指出只有当分类器集合中各个子分类器具有显著的互补性时,它们的融合效果才能够充分体现出来。Tumer 等^[16]在研究中也给出了证明,当各子分类器的错分正相关时,多分类器系统仅能略微减少错分率。2000 年, Dietterich^[17]最早提出了分类器差异性度量的概念,并指出只有参与组合的分类器之间存在差异性才能够获得融合分类性能的提升。差异性度量已成为多分类器系统研究中的热点, Kuncheva 等^[18]对现有的差异性度量方法进行了归类,同时文献 [19] 通过大规模的实验探索了差异性度量与分类器融合分类正确率之间的关系。2005 年 Elsevier 出版的 *Information Fusion* 杂志刊出一期 “A Special Issue on Diversity Measure in Multiple Classifier System”, 专门针对差异性度量的定义以及差异性度量对多分类器系统分类性能的预测能力进行了特别关注。其中, Windeatt^[20]对比了已有的一些成对差异性度量方法对于分类器集合分类正确率的预测能力, Brown 等^[21]对现有产生差异性的方法做了归类和研究, Gal-Or 等^[22]研究了在面对不同的任务时,应该如何选择差异性度量方法来评估一个多分类器集合。在之后的研究中,许多研究者提出了基于传统差异性度量方法的一些改进以及新的差异性度量方法。2008 年, Fan 等^[23]对输出不是决策级的分类器差异性度量做了研究。2009 年, Trawinski 等^[24]初步研究了同时利用正确率和差异性度量指标构造多分类器系统,同时在分类器选择方法上引入了模糊规则。2011 年, Nascimento 等^[25]研究了现有的数种获得高差异性分类器集合的方法,并尝试了将这些方法进行组合以获得更好的分类器集合。

目前已有的差异性度量方法虽能够从一定程度上刻画分类器之间的关系,然而仍没有某种差异性度量定义能够被广泛地接受和认可,还存在着诸如“差异性淹没”等问题。因此,如何更有效地度量分类器间的差异性并更进一步利用这种差异性指导多分类器系统的设计是一个值得深入研究的问题。本文提出了一种基于几何关系的多分类器差异性度量,在该度量中依据给定规则将样本点映射到圆周上,然后,利用子分类器的分类结果对映射点进行标注。通过圆周上正确分类点的几何中心来计算差异性度量值,能有效抑制“差异性淹没”的问题。本文同时

提出了基于该差异性度量的多分类器系统构造方法。实验结果表明,以上方法是合理有效的。

1 多分类器系统与差异性度量

1.1 多分类器系统

多分类器系统是应对复杂模式识别问题的有效手段之一,即,使用多个不同的分类器进行分类,然后,通过一定的组合机制把多个分类器的分类结果进行融合,以获得更为理想的分类性能。要构造一个多分类器系统,首先需生成多个子分类器以供选择。

生成不同的子分类器有多种方法,例如选取不同的分类模型或参数、不同的特征子空间^[26-27]或训练样本^[28-29]等。多分类器系统的流程如图 1 所示。

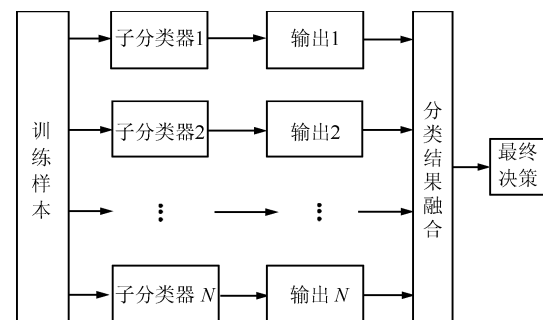


图 1 多分类器系统流程

Fig. 1 Procedure of multiple classifier systems

虽然多分类器技术已被证明可以提高模式识别系统的分类正确率,但并不是任意的分类器组合都能够获得这样的提高。分类器间的差异性和互补性是构造多分类器系统的关键所在,因此,差异性度量的定义对于有效提高多分类器系统融合分类性能来说是至关重要的。一些传统的差异性度量定义介绍如下。

1.2 差异性度量定义

差异性度量的目的是通过某种方法对分类器集合中各子分类器之间的“差异”进行量化,现有的差异性度量方法可以被分为如下两种类型^[30]: 1) 成对差异性度量方法: 成对差异性度量考虑的是两两分类器之间的差异性。表 1 给出了定义两分类器 D_i , D_j 成对差异性度量所需的一些数据。成对差异性度量的代表方法有 Q 统计法 (Q)、相关系数法 (R)、不一致度量法 (D) 以及双错法 (DF) 等,其定义分别如式 (1)~(4) 所示:

$$Q_{i,j} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}} \quad (1)$$

$$R_{i,j} = \frac{N^{11}N^{00} - N^{01}N^{10}}{\sqrt{(N^{11}+N^{10})(N^{01}+N^{00})(N^{11}+N^{01})(N^{10}+N^{00})}} \quad (2)$$

$$D_{i,j} = \frac{N^{01} + N^{10}}{N^{11} + N^{00} + N^{01} + N^{10}} \quad (3)$$

$$DF_{i,j} = \frac{N^{00}}{N^{11} + N^{00} + N^{01} + N^{10}} \quad (4)$$

表1 两分类器的联合输出

Table 1 The joint outputs of two classifiers

	D_j 正确 (1)	D_j 错误 (0)
D_i 正确 (1)	N^{11}	N^{10}
D_i 错误 (0)	N^{01}	N^{00}

对一个由 L 个子分类器构成的集合, 它的成对差异性度量值由所有两两两分类器所得差异性度量值取平均获得, 如式 (5) 所示:

$$\text{Diversity}_{ave} = \frac{2}{L(L-1)} \sum_{i=1}^{L-1} \sum_{j=i+1}^L \text{Diversity}_{i,j} \quad (5)$$

2) 非成对差异性度量方法:

非成对差异性度量方法直接在整个分类器集合上进行计算, 主要从分类器集合的方差、熵或者集合中对随机选择样本错分的分类器比例等^[19] 考察分类器集合的差异性. 式 (6) 所示为基于熵的差异性度量定义.

$$E = \frac{1}{N} \sum_{j=1}^N \frac{1}{\left(L - \left\lfloor \frac{L}{2} \right\rfloor\right)} \min\{l(z_j), L - l(z_j)\} \quad (6)$$

式中, L 为子分类器数目, N 为训练样本个数, $l(z_j)$ 代表对样本 z_j 做出正确分类的子分类器个数, 符号 $\lceil \cdot \rceil$ 代表上取整.

1.3 现有度量存在的“差异性淹没”问题

以上的差异性度量方法都是基于子分类器分类结果的一致性或不一致性, 然而仅仅基于这样的信息, 差异性度量无法对每个子分类器具体错分区域进行量化. 在某些情况下, 这些差异性度量方法还会造成“差异性淹没”问题, 如例 1 所示.

例 1. 假设现有 6 个样本点 $S_1 \sim S_6$, 两个多分类器集合 C_1, C_2 , 并且每个分类器集合由 3 个子分类器 ($C_{1,1} \sim C_{1,3}, C_{2,1} \sim C_{2,3}$) 构成, 在 $C_{i,j}$ 中, i 表示分类器集合, j 表示子分类器. 假设每个子分类器分类结果如下 (某个样本被分对则记为 1, 否则记为

$$0):$$

$$S_1, S_2, S_3, S_4, S_5, S_6$$

$$C_{1,1}: [1, 0, 0, 0, 0, 0]$$

$$C_{1,2}: [0, 1, 0, 0, 0, 0]$$

$$C_{1,3}: [0, 0, 1, 0, 0, 0]$$

$$S_1, S_2, S_3, S_4, S_5, S_6$$

$$C_{2,1}: [1, 0, 0, 0, 0, 0]$$

$$C_{2,2}: [0, 0, 1, 0, 0, 0]$$

$$C_{2,3}: [0, 0, 0, 0, 1, 0]$$

分别利用 Q 统计法 (Q)、相关系数法 (R) 以及双错法 (D) 计算这两个分类器集合的差异性. 在 C_1, C_2 中考察任意两个分类器, 可得 $N^{10} = 1, N^{01} = 1, N^{11} = 0, N^{00} = 4$, 故计算得: $Q_{C_1} = Q_{C_2} = -1, R_{C_1} = R_{C_2} = -0.2, D_{C_1} = D_{C_2} = 0.67$. 该结果说明, 利用以上任意一种方法计算, 都具有相同的差异性. 然而, 两个分类器集合内的子分类器明显具有不同的正确分类区域 (或错误分类区域), 因而可能会具有不同的分类能力. 例如, 我们假设 6 个样本中, 前三个为第一类, 后三个为第二类, 那么分类器集合 C_1 无论利用何种分类器组合方法都无法对第二类样本做出正确分类. 而分类器集合 C_2 虽然在第一类样本上的分类性能不如 C_1 , 但却有可能对第二类样本做出正确的分类. 使用传统的差异性度量方法无法区分出这两个分类器集合的差异, 即造成了“差异性淹没”. 因此, 需要设计更为合理、细致以及区分度更高的差异性度量.

2 基于几何关系的差异性度量

本文提出一种基于几何关系的差异性度量, 在该差异性度量中, 首先将所有样本点根据一定的规则映射到圆周上, 再求取圆周上所有正确分类点的几何中心. 一个分类器集合内的子分类器所得中心相互之间越分散, 该分类器集合的差异性越大.

2.1 样本点的映射与标注

首先, 将所有样本映射到二维平面的单位圆上. 为简单起见, 将样本点按照其序号, 顺时针等间隔映射在圆周上, 使所有样本点将圆周等分. 下面, 将所有圆周上的点根据“分类结果向量”进行标注. 分类结果向量定义如下:

$$A_j = [A_{j1}, A_{j2}, \dots, A_{jm}], \quad j = 1, 2, \dots, c \quad (7)$$

式中 m 为样本数目, c 为集合中子分类器数目, j 表示子分类器序号. 当第 i 个样本被第 j 个分类器分类正确, $A_{ji} = 1$; 否则, $A_{ji} = 0$. 若 $A_{ji} = 1$, 则分类器 j 标注样本点 i 为“有效点”; 否则, 标注该样本点为“无效点”.

例 2. 现假设有 6 个样本 $S_1 \sim S_6$, 分类器集合 C_1, C_1 中有 3 个子分类器 $C_{1,1} \sim C_{1,3}$. 各子分类器的分类结果如下:

$$\begin{aligned} & S_1, S_2, S_3, S_4, S_5, S_6 \\ C_{1,1} &: [1, 1, 0, 0, 0, 1] \\ C_{1,2} &: [0, 1, 1, 1, 0, 0] \\ C_{1,3} &: [0, 0, 0, 1, 1, 1] \end{aligned}$$

图 2 所示为样本点映射与标注的示意图 (实心为有效点, 空心为无效点).

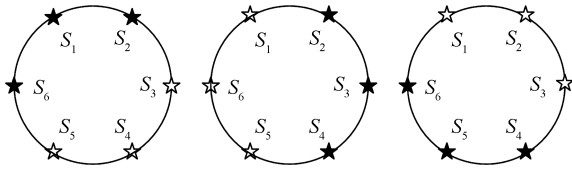


图 2 样本点映射与标注示意图

Fig. 2 Illustration of mapping and marking

2.2 求各子分类器所标注的有效点的几何中心

本文基于各子分类器样本点映射图中有效点的几何中心对该分类器集合的差异性进行度量. 这里仍使用第 2.1 节中的例 2. 图 3 中各子分类器所标注的有效点的几何中心分别表示为 $Z_{C_{1,1}} \sim Z_{C_{1,3}}$.

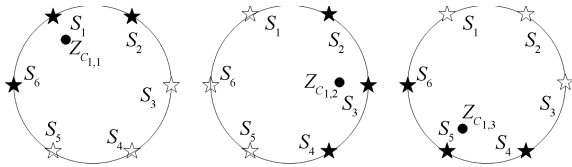


图 3 有效点的几何中心

Fig. 3 Geometric centers of the "valid" points

2.3 基于几何中心的差异性度量定义

对于分类器集合 C 中的每一个分类器 C_i , 都可以得到类似图 2 所示的样本点映射与标注, 并求得如图 3 所示的中心位置. 某个分类器标注的“有效点”的几何中心可以从一定程度上反映出该分类器正确分类样本点 (或错误分类样本点) 的分布情况, 故可认为, 当分类器集合中的各个子分类器所得到的中心位置相互之间越分散, 这些子分类器的正确分类区域 (错误分类区域) 差别就越大, 也即分类器集合具有越大的差异性. 图 4 显示了几何中心反映出的不同正确/错误分类区域.

定义多分类器集合 C_i 的差异性为

$$Diversity(C_I) = \sum_j d(Z_{C_{I,j}}, Z_{C_{I,ave}}) \quad (8)$$

其中, $d(\cdot)$ 为两点间的欧氏距离, $Z_{C_{I,j}}$ 为第 I 个分类器集合中, 第 j 个子分类器所得中心, $Z_{C_{I,ave}}$ 为该分类器集合中所有子分类器所得中心的平均位置.

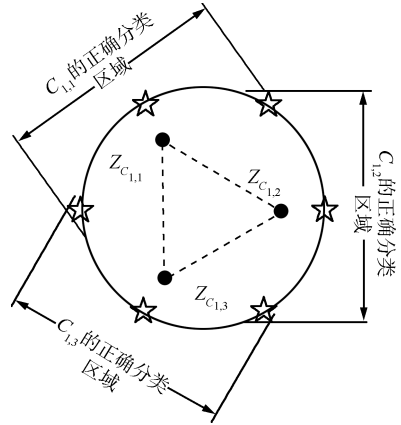


图 4 几何中心反映出的子分类器正确分类区域的差异
Fig. 4 Different correct classification regions reflected by geometric centers

以下基于新差异性度量重新对例 1 进行分析. 对于分类器集合 C_1 来说, 每个成员分类器仅有一个正确分类样本, 其样本点映射如图 5 所示.

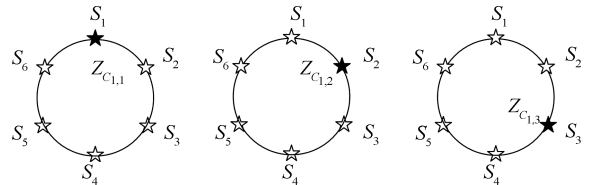


图 5 分类器集合 C_1 中的样本点映射及标注
Fig. 5 Mapping and marking in ensemble C_1

对于分类器集合 C_2 来说, 每个成员分类器仅有一个正确分类样本, 其样本点映射如图 6 所示.

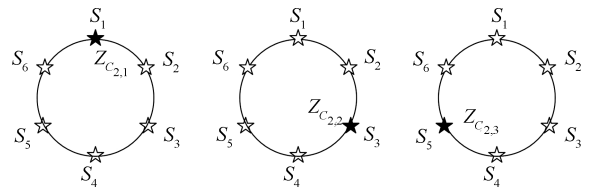


图 6 分类器集合 C_2 中的样本点映射及标注
Fig. 6 Mapping and marking in ensemble C_2

对于分类器集合 C_1 来说, 几何中心分布如图 7 所示.

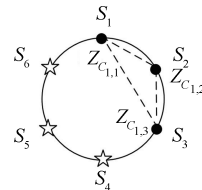


图 7 分类器集合 C_1 几何中心分布

Fig. 7 Distribution of geometric centers in C_1

对于分类器集合 C_2 来说, 几何中心分布如图 8 所示.

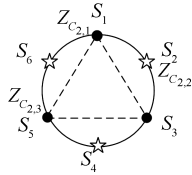


图8 分类器集合 C_2 几何中心分布

Fig 8 Distribution of geometric centers in C_2

设图7和图8中的圆周半径为1, 则依式(8)结合平面几何知识, 不难得到: $Diversity(C_1) = (1 + 2\sqrt{7})/9$, $Diversity(C_2) = 1$, 显然基于新差异性度量, 两个分类器集合的差异性程度并不相同. 因此, 就例1而言, 新度量抑制了“差异性淹没”问题.

2.4 一种抑制“中心重叠”问题的映射方法

在第2.1节中为了便于读者理解, 基于样本点映射的新差异性度量方法, 使用了一种简单的等分圆周的方式. 然而该映射方法在某些情况下可能会出现一些问题. 例如, 在一个分类器集合中各子分类器具有不同的分类结果, 但经过映射与标注后, 得到重叠的几何中心, 如例3所示.

例3. 现假设有6个样本 $S_1 \sim S_6$, 分类器集合 C_1 , C_1 中有3个子分类器 $C_{1,1} \sim C_{1,3}$, 各子分类器分类结果如下:

$$\begin{aligned} &S_1, S_2, S_3, S_4, S_5, S_6 \\ C_{1,1} &: [1, 0, 0, 1, 0, 0] \\ C_{1,2} &: [0, 1, 0, 0, 1, 0] \\ C_{1,3} &: [0, 0, 1, 0, 0, 1] \end{aligned}$$

各子分类器所得中心 $Z_{C_{1,1}} \sim Z_{C_{1,3}}$ 如图9所示. 根据第2.3节中差异性度量方法的定义, C_1 的差异性度量值为0.

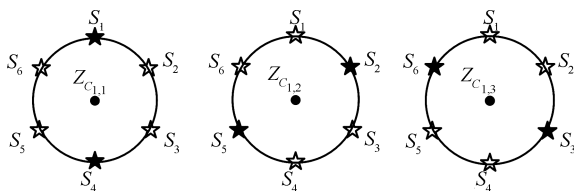


图9 中心重叠的问题

Fig. 9 The problem of “overlapped centers”

该问题是由圆周的对称性导致的, 因此本文提出一种能够抑制此问题的样本点映射方法.

首先, 将所有类别从 $1 \sim CL$ 编号, CL 代表最大类别数目. 圆周被 CL 所等分, 将每个类别的中心点按照其类别标号顺序放置于每一段弧的中点. 每个样本点映射到圆周上的位置由该样本点与其所属类别相邻的两个类别中心的距离决定. 例如, 样本共有3个类别, 样本点 S 属于第2类. 样本点 S 与第1类样本中心 (Cen_1) 之间的距离

为 d_1 , 同样, 也可以得到 d_2 和 d_3 . 样本点 S 将被映射到弧 $Cen_1Cen_2Cen_3$ (从 Cen_1 顺时针至 Cen_3) 上, Cen_1 与样本点 S 在圆周上的映射点 MP_S 间的弧长和 Cen_3 与 MP_S 间的弧长满足关系, $AL(Cen_1, MP_S)/AL(MP_S, Cen_3) = d_1/d_3$, 其中 $AL(a, b)$ 表示弧 ab (a 顺时针至 b) 的长度. 映射算法描述如下:

- 1) 设圆周长为 Len .
- 2) 将所有类别从 $1 \sim CL$ (类别数目) 编号, 将圆周等分为 CL 段 (弧), 将所有类别的中心 (Cen_{cl}) 按照类别标号顺序放置在每个段的中点.
- 3) For $i=1$, 训练样本个数
 - a) 得到样本点 i 的类别标签 cl .
 - b) $PreC$ 表示类别标签 $cl-1$ (如 $cl-1 = 0$, 则设 $PreC$ 为 CL), $PosC$ 表示类别标签 $cl+1$ (如 $cl+1 > CL$, 则设 $PosC$ 为 1). 计算样本点 i 与 Cen_{PreC} 之间的距离 $d_{i,PreC}$, 计算样本点 i 与 Cen_{PosC} 之间的距离 $d_{i,PosC}$.
 - c) 将样本点 i 映射到弧 $Cen_{PreC}Cen_{cl}Cen_{PosC}$ (Cen_{PreC} 顺时针至 Cen_{PosC}) 上, 并满足关系 $AL(Cen_{PreC}, MP_i)/AL(MP_i, Cen_{PosC}) = d_{i,PreC}/d_{i,PosC}$.

End
为了更清楚地说明该映射算法, 现举例如下.

例4. 假设从样本空间中取出6个样本点 $S_1 \sim S_6$, 样本共有3个类别且均具有一维属性. 三个类别的样本中心分别为3.0, 5.0和7.0. 样本 S_1, S_2 属于类1, 样本 S_3, S_4 属于类2, 样本 S_5, S_6 属于类3. 每一个样本的属性值为: 样本 S_1 : 3.2, 样本 S_2 : 4.0, 样本 S_3 : 4.7, 样本 S_4 : 5.7, 样本 S_5 : 6.5, 样本 S_6 : 6.0. 分类器集合 C_1 包含3个成员, 且各成员的分类结果与例3中相同. 在本例中, 我们设 $Len = 6$. 考察样本 S_1 . 样本 S_1 到 Cen_2 和 Cen_3 的距离分别为1.8和3.8, 因此, 样本点 S_1 被映射到弧 $Cen_3Cen_1Cen_2$ (从 Cen_3 顺时针至 Cen_2) 上, 且满足 $AL(Cen_3, MP_{S_1})/AL(MP_{S_1}, Cen_2) = d_3/d_2$, 其中 $AL(Cen_3, MP_{S_1}) = 4 \times [3.8/(1.8 + 3.8)] = 2.71$. 其他样本点以相同方法映射到圆周上, 如图10所示.

将样本点进行映射后, 利用各子分类器的分类结果, 对样本点进行标注, 并求各子分类器所标注的“有效点”的几何中心. 如图11所示, 该映射方法避免了例3中心点重叠的问题. 该映射方法在一定程度上基于各样本点在样本集内的空间位置, 可以更客观地反映样本间的关系. 由于实际情况下样本完全对称均匀分布的情况出现的概率极低, 通过这种映射方法, 圆周上的样本映射点将是疏密不均的, 故能够很大程度上抑制“中心点重叠”问题的出现. 当然也不排除极端情况下仍可能出现“中心点重叠”

问题 (但概率极低).

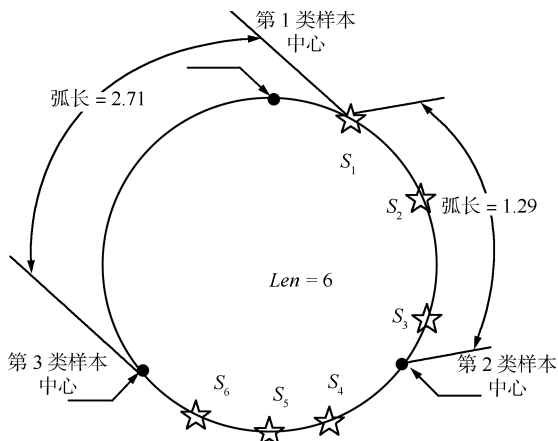


图 10 新样本点映射方法示意图

Fig. 10 Illustration of the new mapping method

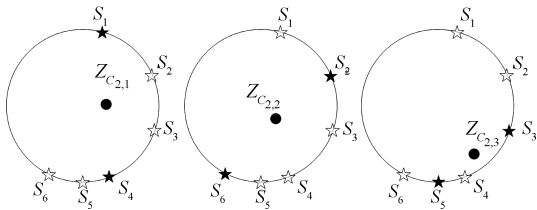


图 11 利用新样本点映射方法解决中心点重叠问题

Fig. 11 Illustration of the solution of the “overlapped centers” problem stated in Example 3

3 基于差异性度量的多分类器系统构造

差异性度量方法将多分类器系统中子分类器之间的差异进行了量化,但其目的并不仅仅是分析各个子分类器之间的关系.差异性度量更重要的作用在于作为指标函数,指导分类器集合的选择.为验证第 2 节中所提出的差异性度量方法对多分类器系统设计的指导作用,本文设计了如下几个算例.在算例中首先采用“过度生成-再选择”策略^[13],过度产生以不同属性作为输入的子分类器.其次,以优化求解方式(如遗传算法等)从生成的子分类器中依据式(9)定义的指标函数选择一定数目的子分类器组成一个分类器集合进行分类与融合,在算例中,子分类器分类结果的融合采用相对多数投票方法.

$$\text{Fitness} = \text{Diversity} + \alpha \cdot \text{Accuracy}_{ave} \quad (9)$$

其中, Accuracy_{ave} 表示所选择分类器集合内各子分类器在训练样本集上的平均分类正确率.

3.1 基于 UCI-iris 数据集的算例

在本算例中选择 UCI^[31] 中的 iris 数据集作为实验对象.iris 数据集共有 3 个类别,每个类别各 50 个样本,所有样本均有 4 维属性.子分类器选择 k -NN 分类器,考虑到数据集样本个数,在此算例中

k -NN 的参数 k 选择为 9.在过度生成子分类器的过程中,从 4 维属性中选取 2 维生成子特征空间,共产生 $C_4^2 = 6$ 个子分类器.设定某个分类器集合可以含有 3 个或 5 个子分类器,以避免在投票中出现平局,即有 $C_6^3 + C_6^5 = 26$ 个分类器集合可供选择.考虑到子分类器数目较少,本算例中使用遍历搜索以及式(9)所示的指标函数寻找最优分类器集合,在指标函数中,使用了不同的差异性度量,包括 Q 统计法(Q)、相关系数法(R)、不一致度量法(D)以及基于几何中心的差异性度量方法(G)等.实验重复进行 20 次.每次实验中样本点被随机分为 3 份,选取其中一份作为训练样本集,其余部分用作测试.依据不同标准选择的最优分类器集合融合分类正确率如表 2 所示.

表 2 基于 UCI-iris 数据集的分类正确率比较

指标	平均正确率 (%)	集合中分类器数目
所有分类器	95.07	6
基于几何中心 (G)	96.00	3
Q 统计 (Q)	94.67	3
相关系数 (R)	94.67	3
不一致度量 (D)	94.00	3

3.2 基于 UCI-Blood Transfusion Service Center 数据集的算例

Blood Transfusion Service Center 数据集由 748 个样本组成,共分为两类.每个样本包含 4 维属性,分别为 R (Recency - 距离上次献血的月数), F (Frequency - 总献血次数)、 M (Monetaryx - 总献血量 c.c.)、 T (Time - 距离第一次献血的月数).其两个类别所占比例分别为 24% 与 76%,代表被调查者是否曾经在 2007 年 3 月进行献血.子分类器选择 k -NN 分类器,考虑到数据集中各类样本的数目,取参数 $k = 15$.其余的设置均与第 2.1 节中的算例相同.依据不同标准选择的最优分类器集合融合分类正确率如表 3 所示.

表 3 基于 UCI-Blood Transfusion Service Center 数据集的分类正确率比较

指标	平均正确率 (%)	集合中分类器数目
所有分类器	77.51	6
基于几何中心 (G)	78.21	3
Q 统计 (Q)	75.94	3
相关系数 (R)	75.94	3
不一致度量 (D)	78.07	3

如表 2 和表 3 所示,相比于传统的差异性度量方法,利用基于几何中心的差异性度量,所选择的最

优分类器集合在实验数据集上可获得更高的融合分类正确率.

3.3 基于分类器聚类的多分类器系统成员个数确定方法

利用遍历搜索的方式可以寻找到最优的分类器集合进行分类与融合, 然而, 在很多应用中人们常常会面对数目庞大的分类器需要进行选择. 在如此大的搜索空间上构造多分类器系统, 首先面对的问题是如何确定该系统子分类器的个数. 假设有 50 个子分类器待选择, 在无法确定分类器集合大小 (即集合中子分类器数目) 的情况下, 将有多达 $\sum_{i=1}^{50} C_{50}^i = 1.1259 \times 10^{15}$ 种可能的组合, 这将导致搜索算法收敛速度极慢, 且很容易陷入局部极值. 而如果能够通过某种方法确定分类器集合的大小, 如假设子分类器个数 $C = 5$, 那么可能的组合将减少为 $C_{50}^5 = 2.12 \times 10^6$, 较之前大大减少. 依据此思想本文提出通过分类器聚类确定多分类器系统成员 (子分类器) 个数的方法, 在该方法中, 使用某个分类器所标注的“有效点”的几何中心作为特征来代表这个分类器, 并使用该特征进行分类器的聚类.

3.3.1 通过分类器聚类减小搜索空间

第 2 节中提出的基于几何中心的差异性度量方法可以通过各个子分类器所标注的“有效点”的几何中心反映出分类器正确或错误分类区域的不同. 显然, 具有相近中心位置的分类器也会具有相近的正确或错误分类区域, 因此, 具有相近中心位置的分类器如果大量同时出现在一个分类器集合中, 那么该分类器集合的差异性就会较小. 将该中心位置作为子分类器的特征代表分类器, 并依据此特征对原始分类器集合进行聚类, 选择分类器的搜索空间将大为缩小. 首先, 经过聚类可确定分类器集合的大小, 即为聚类所得聚类团的数目, 同时, 经过聚类后只需从每个聚类团中选取一个分类器就可以构成多分类器系统, 也即确定了构成系统的每个子分类器应当从原始分类器集合的哪个子集中进行选择, 更进一步缩小了搜索空间. 在该搜索空间上使用诸如遗传算法等进行搜索, 可以更快, 更准确地找到较优的分类器集合.

本文使用迭代自组织数据分析 (Iterative self-organizing data analysis technique algorithm, ISODATA)^[32] 算法进行聚类, 其特点是能够自动地进行类的“合并”和“分裂”, 从而得到较为合理的各个聚类^[33], 而不需事先由人工指定聚类个数, 因此, 能够自动确定分类器集合的大小. 在该算法中, 需要确定如下控制参数: K : 期望得到的聚类数; θ_N : 一个聚类中最少样本数; θ_S : 标准偏差参数; θ_C : 合并参数; L : 每次迭代允许合并的最大聚类对数; I : 允许迭代的次数. 构造多分类器系统的流程如图 12 所

示.

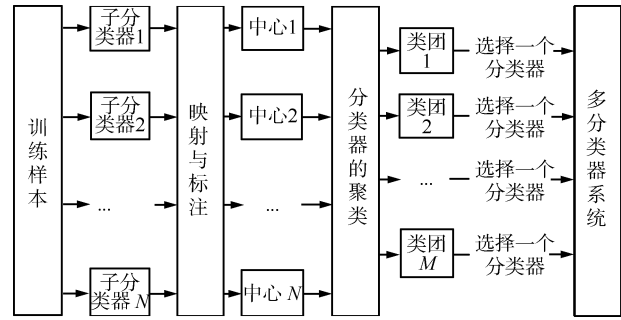


图 12 基于分类器聚类构造多分类器系统流程

Fig. 12 Multiple classifier systems construction based on classifier clustering

3.3.2 基于分类器聚类的多分类器系统构造方法示例

本示例的目的是检验上文中提出的多分类器系统构造方法是否有效, 即是否能够在减小搜索空间的同时选择出较优的分类器集合. 在本文第 3.2 节基于 UCI-Blood Transfusion Service Center 的算例中, 构造了 6 个子分类器的原始分类器集合并使用遍历的方法进行选择, 结果表明, 使用不同的差异性度量方法作为指标选择出的最优分类器集合均具有 3 个子分类器.

现使用基于分类器聚类的方法构造多分类器系统, 分类器的选择和参数设置与第 3.2 节中的算例相同, 并使用分类器所标注的“有效点”几何中心代表分类器. 在 ISODATA 算法中, 设置参数为 $K = 4$, $\theta_N = 1$, $\theta_S = 0.001$, $\theta_C = 10$, $L = 2$, $I = 100$. 聚类结果如图 13 所示. 从每个聚类团中分别选取一个分类器组成多分类器系统. 考虑到每个聚类内分类器数目较少, 使用遍历方式进行搜索, 指标函数如式 (9) 所定义. 在指标函数中使用各差异性度量选出的最优分类器集合分类正确率如表 4 所示.

表 4 在 UCI-Blood Transfusion Service Center 上使用新多分类器系统构造方法的分类性能

Table 4 Classification accuracy on UCI-Blood Transfusion Service Center using new multiple classifier system construction method

指标	平均正确率 (%)	集合中分类器数目
基于几何中心 (G)	78.21	3
Q 统计 (Q)	75.94	3
相关系数 (R)	75.94	3
不一致度量 (D)	78.07	3

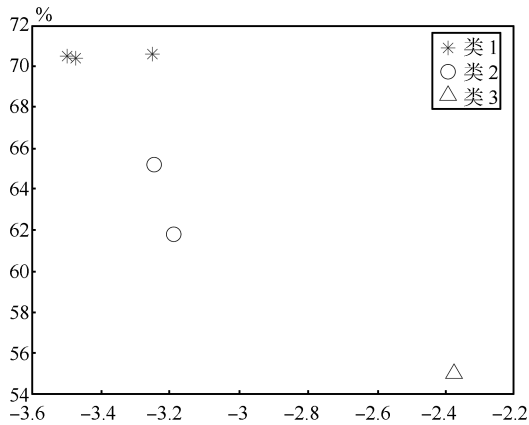


图 13 分类器聚类

Fig. 13 Classifier clustering

经过聚类后, 原始分类器集合被划分为 3 个聚类团, 即多分类器系统应由 3 个子分类器组成, 这与本文第 3.2 节算例中以遍历方式选择出的分类器个数一致. 从这 3 个聚类团中, 各选出一个分类器构成多分类器系统, 选法共有 $C_3^1 \times C_2^1 \times C_1^1 = 6$ 种, 较第 3.2 节中需要遍历的 26 种大为减少, 且使用此多分类器系统构造方法选择出的最优分类器集合的分类正确率与第 3.2 节算例中通过遍历方法找到的最优分类器集合分类正确率相同.

4 基于新差异性度量的多分类器系统构造实验

本实验的目的是对比在预生成子分类器数目较大时, 本文所提出的差异性度量方法, 多分类器系统构造方法与传统方法的优劣. 实验选取 UCI-Wine 数据集.

Wine 数据集共有 178 个 13 维样本, 分为 3 个类别, 每个类别中分别含有 59、71 和 48 个样本. 在实验中, 首先使用“随机子空间法”^[34-35] (从原始特征空间中随机选取一定维数的特征组成子特征空间) 基于原始数据集生成具有随机属性维数的 50 个特征子空间, 并分别使用含有这些特征子空间的数据集训练, 测试分类器, 即建立了 50 个子分类器待选择加入多分类器系统. 在实验中所使用的分类器模型为 k -NN, 根据数据集样本个数, 选择参数 $k = 9$. 在聚类过程中, ISODATA 算法参数选择如下: $K = 4$, $\theta_N = 1$, $\theta_S = 0.001$, $\theta_C = 10$, $L = 2$, $I = 100$. 图 14 所示为原始分类器集合的聚类结果. 共得到 3 个聚类团.

经过聚类后, 从每一个聚类团内选择一个分类器组成多分类器系统. 在这里使用遗传算法进行分类器的选择, 考虑到算法时间性能, 设定最大遗传代数为 100 代. 在遗传算法中, 根据原始集合中某个分类器是否被选择进行编码, 编码长度为 50 位对应

50 个子分类器, 若某个分类器被选择则其对应位置编码为 1, 否则为 0. 使用式 (9) 作为优化指标函数. 需要指出的是, 遗传算法的搜索并不在全空间上进行, 算法选择出的分类器集合中每一个子分类器都来自于一个不同的分类器类团, 因此, 算法的搜索空间被大幅缩小, 算法收敛时间也相应降低.

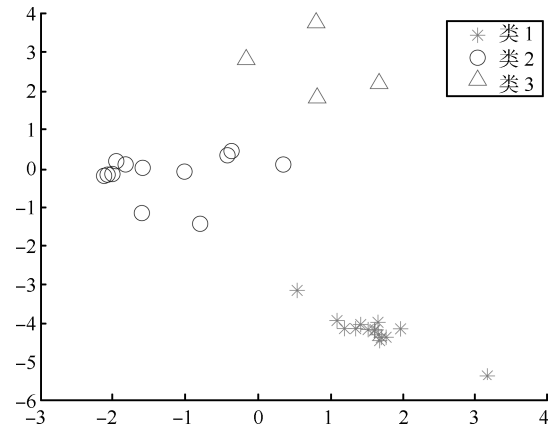


图 14 分类器聚类

Fig. 14 Classifier clustering

实验中分别使用 Q 统计法 (Q)、相关系数法 (R)、不一致度量法 (D) 以及基于几何中心的差异性度量方法 (G), 结合所选子分类器集合的平均分类正确率作为指标函数, 如式 (9) 所示 (为统一标准, Q 统计法等差异性度量值与其所定义的差异性成负相关时, 使用 $1-Q$ 构成指标函数), 从原始分类器集合中选择子分类器组成多分类器系统. 当遗传算法达到指定最大代数后, 从种群中选择出现次数最多的分类器组合作为最终的选择, 并基于此分类器集合对测试样本进行分类和融合, 融合方式采用相对多数投票法. 每组应用不同差异性度量方法作为指标函数的实验分别进行 20 次, 并求取最终融合分类正确率的平均值. 遗传算法经过 100 代后 (某一次运行中) 各指标函数值如图 15 所示. 利用各指标最终选出的分类器集合融合分类性能对比如表 5 所示.

表 5 依据不同指标所选分类器集合分类正确率
Table 5 Classification accuracy of ensembles selected by different diversity measures

指标	平均正确率 (%)	所选集合中子分类器最高正确率 (%)
基于几何中心 (G)	78.09	71.35
Q 统计 (Q)	74.16	72.47
相关系数 (R)	76.40	72.47
不一致度量 (D)	70.79	70.22

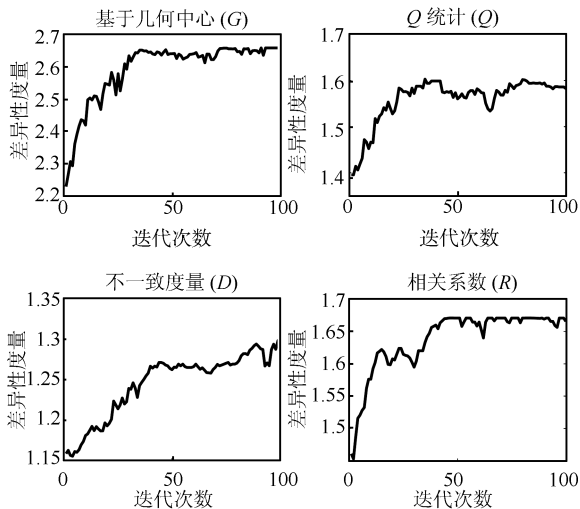


图 15 使用各种差异性度量方法作为指标的遗传算法

Fig. 15 GA based on different diversity measures

以上结果表明,利用基于几何中心的差异性度量方法构成指标函数,能够获得较使用传统差异性度量方法性能更优的分类器集合。

5 结论

本文提出了一种新的差异性度量方法。该方法将样本点映射到圆周上,利用各分类器的分类结果对样本点进行标注,并依据被标注为“有效点”的几何中心求取差异性度量值。该差异性度量方法能够量化分类器集合中各子分类器之间正确或错误分类区域的不同,因而能够很大程度上抑制“差异性淹没”问题。

为解决构造多分类器系统时,由于待选择的子分类器数目庞大,导致搜索空间过大等问题,本文提出了一种基于分类器聚类的多分类器系统构造方法。通过原始分类器集合的聚类,不仅能够确定多分类器系统的成员个数,还能够确定每个成员应当从原始分类器集合的哪个子集中进行选取,大大缩小了搜索空间,节省了运算时间。

需要注意的是,实验中得到的结果与实验中子分类器的选择、分类器参数的选择以及聚类方法与参数的选择都是相关的。比如 k -NN 分类器参数的选择会影响分类器的分类结果,ISODATA 算法参数的选择会影响聚类所得类团数目,进而影响多分类器系统成员的选择。如何根据不同的任务与环境选择合适的方法以及参数将是我们的下一步工作的重点。

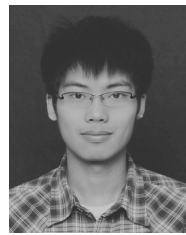
需要指出的是,本文中所提出的方法目前也存在着一些缺陷。相比于传统的差异性度量方法,新方法需要的预处理步骤较为复杂。基于该种差异性度量方法的多分类器构造方法也同样存在这个问题。在基分类器池庞大的情况下,该差异性度量方法所

需要的运算时间较长。目前本文中所提出的样本点映射等方法是为了解决现有问题而做的尝试性工作,其系统性完备性仍有待进一步检验和探索。在后续的工作中需要不断地加以改进。

References

- 1 Didaci L, Fumera G, Roli F. Diversity in classifier ensembles: fertile concept or dead end? In: Proceedings of the 11th International Workshop on Multiple Classifier Systems. Nanjing, China: Springer-Verlag, 2013. 37–48
- 2 Bar A, Rokach L, Shani G, Shapira B, Schclar A. Improving simple collaborative filtering models using ensemble methods. In: Proceedings of the 11th International Workshop on Multiple Classifier Systems. Nanjing, China: Springer-Verlag, 2013. 1–12
- 3 Sciarone F. An extension of the Q diversity metric from single-label to multi-label and multi-ranking multiple classifier systems for pattern classification. In: Proceedings of the 2012 International Conference on Machine Learning and Cybernetics. Xi'an, China: IEEE, 2012. 6–10
- 4 Radtke P, Granger E, Sabourin R, Gorodnichy D. Adaptive ensemble selection for face re-identification under class imbalance. In: Proceedings of the 11th International Workshop on Multiple Classifier Systems. Nanjing, China: Springer-Verlag, 2013. 95–108
- 5 Sun Liang, Han Chong-Zhao, Shen Jian-Jing, Dai Ning. Generalized rough set method for ensemble feature selection and multiple classifier fusion. *Acta Automatica Sinica*, 2008, **34**(3): 298–304
(孙亮, 韩崇昭, 沈建京, 戴宁. 集成特征选择的广义粗集方法与多分类器融合. *自动化学报*, 2008, **34**(3): 298–304)
- 6 Zhang Cai-Po. Fuzzy Integral and Fusion of Multiple Classifiers Applying in Medical Diagnosis [Master dissertation], Tianjin University of Technology, China, 2010
(张彩坡. 模糊积分及多分类器融合在医疗诊断中的应用 [硕士学位论文], 天津理工大学, 中国, 2010)
- 7 Zhang Xue-Feng, Wang Peng-Hui, Feng Bo, Du Lan, Liu Hong-Wei. A new method to improve radar HRRP recognition and outlier rejection performance based on classifier combination. *Acta Automatica Sinica*, 2013, **39**(1): 1–9
(张学峰, 王鹏辉, 冯博, 杜兰, 刘宏伟. 基于多分类器融合的雷达高分辨距离像目标识别与拒判新方法. *自动化学报*, 2013, **39**(1): 1–9)
- 8 Xu L, Krzyzak A, Suen C Y. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, 1992, **22**(3): 418–435
- 9 Lam L, Suen C Y. Optimal combinations of pattern classifiers. *Pattern Recognition Letters*, 1995, **16**(9): 945–954
- 10 Huang Y S, Suen C Y. A method of combining multiple experts for the recognition of unconstrained handwritten numerals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1995, **17**(1): 90–94
- 11 Verlinde P, Ghollet G. Comparing decision fusion paradigms using k -NN based classifiers, decision trees and logistic regression in a multi-modal identity verification application. In: Proceedings of the 2nd International Conference on Audio and Video Based Biometric Person Authentication. Washington D. C., USA: Springer-Verlag, 1999. 188–193
- 12 Lv Yue, Shi Peng-Fei, Zhao Yu-Ming. Voting principle for combination of multiple classifiers. *Journal of Shanghai Jiaotong University*, 2000, **34**(5): 680–683
(吕岳, 施鹏飞, 赵宇明. 多分类器组合的投票表决规则. *上海交通大学学报*, 2000, **34**(5): 680–683)

- 13 Yang Yi, Han De-Qiang, Han Chong-Zhao. A novel diversity measure of multiple classifier systems based on distance of evidence. *Acta Aeronautica et Astronautica Sinica*, 2012, **33**(6): 1093–1099
(杨艺, 韩德强, 韩崇昭. 一种基于证据距离的多分类器差异性度量. *航空学报*, 2012, **33**(6): 1093–1099)
- 14 Kittler J, Hatef M, Duin R P W, Matas J. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998, **20**(3): 226–239
- 15 Ali K M, Pazzam M J. On the Link Between Error Correlation and Error Reduction in Decision Tree Ensembles, Technical Report 95-38, Department of Information and Computer Science, University of California Irvine, 1995
- 16 Tumer K, Ghosh J. Theoretical Foundations of Linear and Order Statistics Combiners for Neural Pattern Classifiers, Technical Report TR-95-02-98, Computer and Vision Research Center, University of Texas, Austin, 1995
- 17 Dietterich T G. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. *Machine Learning*, 2000, **40**(2): 139–157
- 18 Kuncheva L I, Whitaker C J. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 2003, **51**(2): 181–207
- 19 Shipp C A, Kuncheva L I. Relationship between combination methods and measures of diversity in combining classifiers. *Information Fusion*, 2002, **3**(2): 135–148
- 20 Windeatt T. Diversity measures for multiple classifier system analysis and design. *Information Fusion*, 2005, **6**(1): 21–36
- 21 Brown G, Wyatt J, Harris R, Yao X. Diversity creation methods: a survey and categorization. *Information Fusion*, 2005, **6**(1): 5–20
- 22 Gal-Or M, May J H, Spangler W E. Assessing the predictive accuracy of diversity measures with domain-dependent, asymmetric misclassification costs. *Information Fusion*, 2005, **6**(1): 37–48
- 23 Fan T G, Zhu Y, Chen J M. A new measure of classifier diversity in multiple classifier system. In: Proceedings of the 7th International Conference on Machine Learning and Cybernetics. Kunming, China, 2008. 18–21
- 24 Trawinski K, Quirin A, Cordón A. On the combination of accuracy and diversity measures for genetic selection of bagging fuzzy rule-based multiclassification systems. In: Proceedings of the 9th International Conference on Intelligent Systems Design and Applications. Pisa, Italy: IEEE, 2009. 121–127
- 25 Nascimento D S C, Canuto A M P, Silva L M M, Coelho A L V. Combining different ways to generate diversity in bagging models: an evolutionary approach. In: Proceedings of the 2011 International Joint Conference on Neural Networks. San Jose, USA: IEEE, 2011. 2235–2242
- 26 Kamel M S, Wanas N M. Data dependence in combining classifiers. In: Proceedings of the 4th International Workshop on Multiple Classifier Systems. Guilford, UK: Springer-Verlag, 2003. 1–14
- 27 Ho T K, Hull J J, Srihari S N. Decision combination in multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1994, **16**(1): 66–75
- 28 Efron B, Tibshirani R. *An Introduction to the Bootstrap*. New York: Chapman Hall, 1993
- 29 Freund Y, Schapire R E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 1997, **55**(1): 119–139
- 30 Kuncheva L I, Whitaker C J. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 2003, **51**(2): 181–207
- 31 Blake C L, Merz C L. UCI repository of machine learning databases [Online], available: <http://www.ics.uci.edu/~mllearn>, August 1990
- 32 Ma Cai-Hong, Dai Qin, Liu Shi-Bin. A hybrid PSO-ISODATA algorithm for remote sensing image segmentation. In: Proceedings of the 2012 International Conference on Industrial Control and Electronics Engineering (ICICEE). Xi'an, China: IEEE, 2012. 1371–1375
- 33 Zhang Xue-Gong. *Pattern Recognition* (3rd edition). Beijing: Tsinghua University Press, 2010
(张学工. 模式识别 (第三版). 北京: 清华大学出版社, 2010)
- 34 Ho T K. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998, **20**(8): 832–844
- 35 Cheplygina V, Tax D M J. Pruned random subspace method for one-class classifiers. In: Proceedings of the 10th International Conference on Multiple Classifier Systems. Naples, Italy: Springer-Verlag, 2011. 96–105



梁绍一 西安交通大学电信学院硕士研究生。主要研究方向为目标识别与信息融合。

E-mail: shaoyi.liang2@stu.xjtu.edu.cn
(LIANG Shao-Yi Master student at the School of Electronic and Information Engineering, Xi'an Jiaotong University. His research interest covers target recognition and information fusion.)



韩德强 西安交通大学电子与信息工程学院副教授。2008 年获得西安交通大学控制科学与工程博士学位。主要研究方向为证据理论, 信息融合, 目标识别。本文通信作者。

E-mail: deqhan@mail.xjtu.edu.cn
(HAN De-Qiang Associative professor at the School of Electronic and

Information Engineering, Xi'an Jiaotong University. He received his Ph.D. degree in control science and engineering from Xi'an Jiaotong University in 2008. His research interest covers Dempster-Shafer evidence theory, information fusion, and pattern recognition. Corresponding author of this paper.)



韩崇昭 西安交通大学电子与信息工程学院教授。主要研究方向为复杂系统控制与信息融合。

E-mail: czhan@mail.xjtu.edu.cn
(HAN Chong-Zhao Professor at the School of Electronic and Information Engineering, Xi'an Jiaotong University. His research interest covers control of complex system and information fusion.)