

文章编号: 1001-0920(2012)06-0914-05

基于聚类分析与几何的目标特征敏感性评估算法

吴 瑕^a, 周 焰^b, 杨龙坡^b, 崔 建^b

(空军雷达学院 a. 雷达兵器运用工程军队重点实验室, b. 预警监视情报系, 武汉 430019)

摘 要: 为克服在较少先验知识的情况下, 人为选择目标特征进行分类识别的盲目性, 提出一种基于聚类分析与几何学的目标特征评估方法. 该方法在类内类间距离的基础上引入几何学中有关圆的知识, 定义了样本在特征空间中分布的“紧密度”、“分离度”函数和“松紧度”权重, 以考察目标特征对目标样本之间的分类敏感性, 并根据评估结果选取适当特征构成一个新的联合特征, 以提高对目标的分类识别效率. 最后通过仿真验证了该方法的有效性.

关键词: 聚类分析; 几何学; 类距离; 特征评估; 敏感性; 目标识别

中图分类号: TP181

文献标识码: A

Target feature sensitivity evaluation method based on clustering analysis and geometry

WU Xia^a, ZHOU Yan^b, YANG Long-po^b, CUI Jian^b

(a. Key Research Laboratory, b. Department of Early Warning & Surveillance Intelligence, Air Force Radar Academy, Wuhan 430019, China. Correspondent: WU Xia, E-mail: moonshiner007@163.com)

Abstract: To overcome the blindness of subjective selecting dimensionless indicators of target features as sensitive features with less experience, a target feature sensitivity evaluation method based on clustering analysis and geometry is proposed. The method defines “compact degree”, “dissociative degree” and “incompact-compact degree” of samples distributing in features space, and introduces the knowledge of rotundity from geometry, by which the classing ability of target feature to multi-class target samples is reviewed. Then the new combined features are constructed through the sensitivity evaluation results, and the classify recognition efficiency can be improved. Finally, simulation results show the effectiveness of the method.

Key words: clustering analysis; geometry; class distance; feature evaluation; sensitivity; target recognition

1 引 言

弹道目标识别属于多目标多特征融合识别. 美国国防部在 2000 年曾透露, 在飞行中段用于目标识别的各类特征及方法多达 24 种, 并试图用软件将其融合. 随着目标识别技术的发展, 将有更多的目标特征信息被提取出来. 面对众多的特征, 是否需要都对每一个都进行关注? 答案是否定的. 因为选用的特征量越多, 所付出的计算代价也越高, 耗时自然越高, 并且设计的分类器往往也越复杂. 同时通过对弹道目标特性的分析发现, 许多目标特征之间都有着相互关联, 如目标雷达散射截面积(RCS)时间变化序列关系到目标的进动特性与结构特性, 目标质量关系到再入质阻比特性与温度及其变化率特性等^[1-3]. 因此, 选择较多特征进行识别, 会因特征量之间的强相关性带来冗余

计算, 反而造成分类器性能下降. 另一方面, 在弹道目标识别中, 一些特征所发挥作用的重要程度不同, 并有“此消彼长”的现象, 如进入飞行中段后各类目标的轨迹与速度特征基本相同, 此时无法作为主要目标识别依据; 而进入再入段后各类目标的轨迹和速度特征会有较大区别^[1-3], 此时这两个特征的重要程度将显著提高. 因此, 对不同目标在不同时间进行分类识别, 敏感特征往往不同, 凭经验甚至在没有先验知识的情况下盲目地选择特征进行识别判断, 将很难对目标作出高效的识别.

特征评估的基本任务是通过一个定量的准则来衡量目标特征对目标分类的敏感程度, 进而从众多特征中选出那些对识别最有利的特征, 从而实现计算量的减少与高维特征空间维数的压缩. 特征选取是否得

收稿日期: 2010-11-18; 修回日期: 2011-02-15.

基金项目: 国家自然科学基金项目(40101019); 空军雷达学院科研基金项目(LYYY200807).

作者简介: 吴瑕(1979-), 男, 工程师, 博士, 从事军事目标识别、信息融合等研究; 周焰(1966-), 男, 教授, 博士生导师, 从事信息融合、图像处理等研究.

当将直接影响到识别算法的效率、处理时间、所需数据量以及扩展性. 如果能选取适当特征并组合, 则能显著改善识别性能.

目前, 此类特征评估方法研究尚不多见, 已有的方法主要集中于聚类分析中的类内类间距离、类别可分性判据^[4-12]等, 但是这类方法或者是基于特征向量间的平均距离, 无法涉及异类特征和反映特征时域上的全局信息变化^[4-8]; 或者条件较为苛刻, 需要类别先验概率^[9-12], 不适用于弹道目标特征分析. 对此, 本文提出一种新的利用样本类内类间距离与几何学中圆间距等知识来评价特征聚类有效性的方法, 实现对不同弹道目标特征的时域无量纲下分类敏感性定量评判, 以衡量特征在分类问题上对目标的可行性程度, 并通过敏感特征的优化组合提高目标的识别效率.

2 特征敏感性评估与优选

2.1 特征敏感性评估

该特征评估方法是基于同类样本在不同特征空间的类内类间距离“松紧度”来对不同特征敏感性进行评估的. 其评估原则是: 在某一特征空间下, 同一类的类内特征距离“紧密度”越小, 不同类的类间特征距离“分离度”越大, 则认为这一特征分类效果越明显, 对分类越具有敏感性. 即将几何学中的距离关系与聚类分组相结合, 以统计各类样本的一致性分类. 其步骤如下:

Step 1 数据预处理. 设有 l 类目标, 分别对目标的 m 个特征指标参数进行测量, 则第 i 个目标的特征指标向量集为

$$X_i = \{x_{i1}, x_{i2}, \dots, x_{im}\}.$$

第 k 类中全部目标样本 α 的 m 个特征指标下的特征矩阵可表示为

$$A_k = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{\alpha 1} & x_{\alpha 2} & \dots & x_{\alpha m} \end{bmatrix} = (x_{ij})_{\alpha \times m}.$$

目标融合识别中所构造的特征往往是由性质不同的参数组成, 造成各特征一般具有不同的量纲与量级, 在进行关联分析时, 不具有等效性. 为了消除量纲量级的影响, 进行规格化处理是必要的. 所谓规格化是将不同量级、量纲的数据统一处理为 $[0, 1]$ 区间的数据. 规格化方法很多, 本文采用如下公式:

$$r_{ij} = \left| \frac{x_{ij} - \bar{x}_{ij}}{x_{ij \max} - x_{ij \min}} \right|. \quad (1)$$

其中: $x_{ij \max}$ 为目标 i 的特征指标 j 的最大特征值; $x_{ij \min}$ 为目标 i 的特征指标 j 的最小特征值; \bar{x}_{ij} 为相应均值; r_{ij} 为 x_{ij} 的规格化值, 且 $0 \leq r_{ij} \leq 1$. 根据实

际情况设定的式(1)较全面地包含了 x_{ij} 的取值波动信息, 它是各特征的时域无量纲指标. 若涉及到其他距离运算方法, 则上述绝对值运算将变为相应距离运算公式, 以下同. 在此基础上, 可对不同目标特征在时域展开评估.

Step 2 计算第 k 类中第 j 个特征下的类内样本间距离平均值

$$d_{jk} = \frac{2}{\alpha(\alpha-1)} \sum_{p,q=1}^{\alpha} |r_{jk}(p) - r_{jk}(q)|, \quad (2)$$

$$j = 1, 2, \dots, m, \quad k = 1, 2, \dots, l.$$

其中: α 为样本个数, $r_{jk}(p)$ 和 $r_{jk}(q)$ 分别为第 k 类中的第 p 和 q 个目标样本的特征 j 的规格化值.

Step 3 计算第 k 类中 α 个样本的第 j 个特征平均值

$$J_{jk} = \frac{1}{\alpha} \sum_{q=1}^{\alpha} r_{jk}(q), \quad (3)$$

其中 J_{jk} 为第 k 类的中心.

Step 4 计算每个样本到类中心的距离平均值

$$f_{jk} = \frac{1}{\alpha} \sum_{p=1}^{\alpha} |r_{jk}(p) - J_{jk}|. \quad (4)$$

式(4)表现为样本方差, 体现了样本间的波动情况, 其值越小, 样本间波动越小.

Step 5 通过前面的分析得知, 第 j 个特征因为重要性与分类能力等原因, 对不同类别的聚类效果可能不同, 即不同类的类内样本分布可能相差很大. 为科学合理地反映第 k 类中的类内样本的聚类有效性, 特定义“紧密度”为

$$v_{jk} = d_{jk} + f_{jk}. \quad (5)$$

式(5)从样本之间的距离关系及样本与类中心的距离关系两个方面考察了同类样本的紧密程度, 于是第 j 个特征对 l 个类的类内样本平均“紧密度”函数为

$$D_j = \frac{1}{l} \sum_{k=1}^l v_{jk}. \quad (6)$$

该“紧密度”反映了全体样本间的波动情况. 其值越小, 样本间波动越小, 类内之间的样本紧密程度越高.

Step 6 通过计算第 j 个特征下 l 个类的类间距离平均值

$$d_j^* = \frac{2}{l(l-1)} \sum_{s,z=1}^l |J_{js} - J_{jz}|, \quad (7)$$

$$s, z = 1, 2, \dots, l; \quad s \neq z.$$

其中: J_{js} 和 J_{jz} 分别为第 s 个和第 z 个类对 α 个样本的第 j 个特征的平均值.

Step 7 计算所有类的中心

$$\bar{J}_j = \frac{1}{l} \sum_{k=1}^l J_{jk}. \quad (8)$$

Step 8 计算每个类中心到所有类的中心的距离平均值

$$\bar{f}_j = \frac{1}{l} \sum_{k=1}^l |J_{jk} - \bar{J}_j|. \quad (9)$$

Step 9 定义“分离度”函数为反映不同类之间的差异性,它应包含类与类之间的关系,以及个体类与整体类之间的关系.

在此将两个类看成两个圆,两个类分离性好坏可看成两个圆是否相离还是相交或相切,即

相离时

$$\delta_j(s, z) = |J_{js} - J_{jz}| - d_{js} - d_{jz} > 0;$$

相交时

$$\delta_j(s, z) = |J_{js} - J_{jz}| - d_{js} - d_{jz} < 0;$$

相切时

$$\delta_j(s, z) = |J_{js} - J_{jz}| - d_{js} - d_{jz} = 0.$$

因为所有值均由规格化后得到,故 $-1 \leq \delta \leq 1$.

由此定义类间“松紧度”权重为

$$\mu_j = 1 + \frac{2}{l(l-1)} \sum_{s,z=1}^l \delta_j(s, z). \quad (10)$$

引入特征分类“松紧度”权重,可以更好地反映特征所携目标信息的分类能力,以有效区分开特征的分类敏感性.因此定义第 j 个特征对 l 个类的类间的平均分离度为

$$D_j^* = \mu_j(d_j^* + \bar{f}_j). \quad (11)$$

Step 10 评估第 j 个特征的敏感因子

$$\varphi_j = D_j^*/D_j. \quad (12)$$

为了更直观形象地说明上述方法的原理,以图 1 中对 3 种不同类别的物体分类为例,通过 3 类目标在不同特征空间投影上的类内类间距离“松紧度”来反映特征的敏感性.设有特征 1 和特征 2 ($m=2$) 对 3 种物体分类 ($l=3$), 每个类别各有 6 个样本 ($\alpha=6$).

通过图 1 可以看出,特征 1 能够明显地对 3 种类别进行分离,因为特征 1 对每种类别的类内样本间距

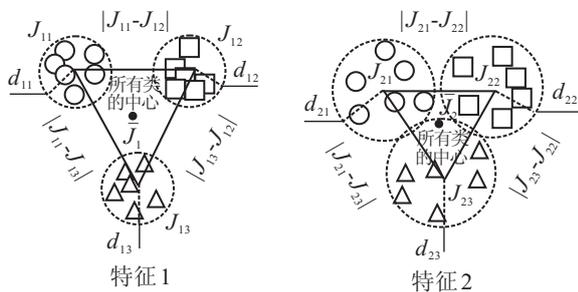


图 1 特征敏感性比较示意

离平均值(虚线圆圈的半径 d_{jk})与每个类中的样本到类中心 J_{jk} (圆心)的平均距离 f_{jk} 要比特征 2 小,所以显得集中,因此,特征 1 比特征 2 的类内平均紧密度值小 ($D_1 < D_2$).同时,因为特征 2 下的类圆存在着相交 ($\mu_2 < \mu_1$),特征 1 比特征 2 的 d_j^* 和 \bar{f}_j 要大,所以特征 1 下的类与类间显得较分散,因此,特征 1 比特征 2 的类间平均分离度大 ($D_1^* > D_2^*$).由式(12)可知,特征 1 在这 3 类别分类问题上比特征 2 更敏感.

该方法具有明显的物理意义,它反映了特征 j 对 l 个类的分类难易程度.显然,类间距离平均分离度越大,类内样本平均紧密度越小,越好分类.相比于文献 [7,9,11] 中的类内类间距离模型,由于它采用了全局信息和更好的几何意义,而非传统的基于两点间距离可分离性判据,避免了类间距离较小、类内距离较大时,可分离性判据仍有可能取得较大值的情况.该计算过程更加直观与明晰,而且因其运算过程相对简单、计算量不大,故时间复杂性不高,能够更好地符合敏感性评估原则.

2.2 特征组合

通常特征优化组合需要从 m 个特征中挑选出 λ 个对分类最有利的特征,但是在实际应用中, λ 是未知的.通过本文方法在得到 φ_j 后按其大小顺序进行排序,在不丢失主要目标信息的情况下,逐一选取敏感特征作为贝叶斯网络的输入,并考虑两种确定 λ 的终止条件:

- 1) 分类准确率达到某一设置门限值,其值需要根据具体问题的要求进行相应设置;
- 2) 特征个数连续增加若干个后,分类准确率没有明显提高.

根据弹道目标分类识别的特点,要求分类准确率越高越好,所以本文选用第 2 个条件.即通过 λ 个组合特征后, l 类目标分类识别率再无显著提高,进而避免了再次提取和选取特征的重复性.

此处目的在于从整体最优角度来选择敏感特征组合,以减少冗余计算量和提高效率.

3 实验与分析

下面对目标特性评估进行仿真.设有 4 类目标依次是:弹头、大型碎片、气球诱饵、锥球形诱饵,它们基本上代表了弹道目标威胁群中几种典型的目標类型,它们在各种特征上的差异已有相关分析^[1-3],在此不多述.目标运动场景设置为中段初期某时间段作为观测区,取 120 s 观测时间,假定每 2 s 提取一次观测样本,共获得 60 个样本.对于每一个样本,本文选择惯量比和温度变化两个特征参数为例,并假定理想情况下,观测过程中没有受到其他干扰和欺骗等对抗措施.

通过查阅相关文献并参考美国 Firefly 实验中的目标特性^[1-3], 以此为基础进行理论计算. 设置弹道场景: 射程为 1000 km, 进动角 $\theta = 10^\circ$, 中段速度为

5 km/s, 弹道最高点为 180 km. 现以中段初期为例, 部分数据规格化后如表 1 所示 (设导弹是在日照区运行).

表 1 特征规格化值 (中段初期)

类别	温度变化				惯量比			
	弹头	大型碎片	气球诱饵	锥形球诱饵	弹头	大型碎片	气球诱饵	锥形球诱饵
采样 1	0.478	0.661	0.553	0.568	0.558	0.159	0.001	0.139
采样 2	0.474	0.614	0.445	0.460	0.362	0.726	0.259	0.046
采样 3	0.471	0.608	0.439	0.455	0.032	0.108	0.002	0.273
采样 4	0.465	0.681	0.431	0.442	0.032	0.108	0.603	0.558
采样 5	0.460	0.653	0.426	0.431	0.063	0.065	0.201	0.043
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
采样 26	0.449	0.491	0.486	0.468	0.460	0.110	0.442	0.476
采样 27	0.451	0.490	0.485	0.464	0.104	0.060	0.001	0.211
采样 28	0.456	0.492	0.486	0.468	0.374	0.778	0.404	0.406
采样 29	0.454	0.493	0.485	0.471	0.001	0.110	0.357	0.194
采样 30	0.454	0.493	0.486	0.468	0.173	0.009	0.104	0.211

将表 1 数据代入式 (2)~(12), 计算评估因子, 所得结果如表 2 所示.

表 2 特征敏感性评估结果 (中段初期)

类别	温度变化	惯量比
弹头	$d_{11} = 0.011, J_{11} = 0.375$	$d_{21} = 0.007, J_{21} = 0.206$
大型碎片	$d_{12} = 0.032, J_{12} = 0.496$	$d_{22} = 0.075, J_{22} = 0.239$
气球诱饵	$d_{13} = 0.014, J_{13} = 0.434$	$d_{23} = 0.009, J_{23} = 0.303$
锥形球诱饵	$d_{14} = 0.015, J_{14} = 0.416$	$d_{24} = 0.008, J_{24} = 0.369$
类内紧密度	$D_1 = 0.034$	$D_2 = 0.039$
类间分离度	$D_1^* = 0.099$	$D_2^* = 0.164$

由表 2 计算温度变化的敏感因子 $\varphi_1 = 2.912$, 惯量比的敏感因子 $\varphi_2 = 4.205$. 这表明, 在中段初期惯量比特征比温度变化特征具有更好地可分性. 分析原因, 在中段初期各类目标温度经过一段时间的下降后会逐渐进入一个动态平衡区, 此时温度变化特征不明显, 相反惯量比特征则相对比较稳定, 具有较好的表现性, 这与文献 [1-3] 的定性分析相吻合. 经仔细观察后发现, d_{22} 相对于其他类内样本间距离平均值要大许多, 且 $f_{22} = 0.044$ 也为最大, 说明惯量比特征对碎片类目标的类内紧密度较低, 聚类效果较差. 这是因为碎片在空间存在着翻滚现象, RCS 序列变化无规律, 由其提取的参数估计惯量比自然不稳定; 相比之下, 其在温度变化特征上具有较好的可分性. 所以, 相同目标在不同特征空间上具有不同的可分性.

另外, 表 2 中温度变化特征下的弹头与锥形球诱饵的类间距离 $|J_{11} - J_{14}| = 0.041$, 远小于其在惯量比特征下的类间距离 $|J_{21} - J_{24}| = 0.163$, 表明弹道中段初期惯量比特征对这两类目标具有更好的可分性. 这说明通过本文方法可以评估某一特征对某些特定目标的分类能力, 同时也说明对于特定目标 (尤其是弹头和较复杂诱饵), 单纯采用某一特征识别方法, 效果将难以保证; 而根据评估结果选择另一最优特征区分

特定目标, 效果可能更明显, 因此需要对多种特征进行优化组合.

现设置 8 个目标一维特征向量分别为: 辐射强度 (1), 体态比 (2), 温度变化 (3), RCS 均值 (4), 惯量比 (5), 平均灰度值 (6), RCS 变化周期 (7) 以及目标径向长度变化 (8). 同样以中段初期某一时间段为例, 还是对上述 4 类目标进行分类, 各参数设置同上, 计算出各特征敏感性如图 2 所示.

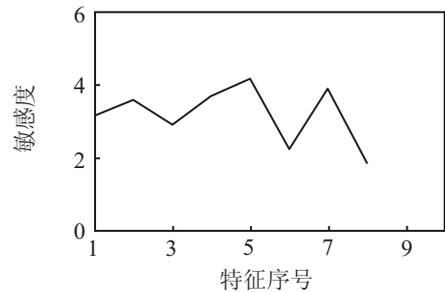


图 2 特征评估敏感度

分析各特征敏感性差异原因: 辐射强度 (1) 与温度变化 (3) 因为特性相差不大, 其敏感程度也基本相同; 而平均灰度值 (6) 是通过温度、辐射强度和图像处理等转换而来, 其精确性受到一定影响, 进而敏感性有所降低; 体态比 (2), RCS 变化周期 (7) 和 RCS 均值 (4) 较全面地反映了目标的结构和运动信息; 相比之下, 目标径向长度变化 (8) 所含信息相对有些单一, 如它对碎片和气球有较好的分辨率, 但对锥形诱饵和弹头则效果不佳.

对各特征按敏感性大小重新排序后逐一作为贝叶斯分类器的输入, 进行训练和测试, 得到分类结果如图 3 所示. 通过图 3 发现, 当增加的识别特征达到前 6 个时, 分类准确率达到最高值, 即前 6 个特征为最优特征组合. 随后增加的平均灰度值 (6) 和目标径向

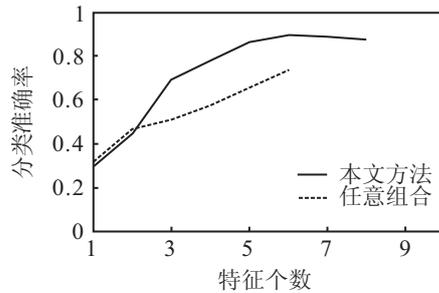


图 3 分类准确率与输入特征之间的关系

长度变化(8)这2个特征对分类率无显著影响,分析原因,主要是因为其特征信息已包含在前面相应的特征里,它们的加入对分类准确率已起不到进一步的提高。

任意组合6个特征(特征6和特征8替换掉最优特征组合中的任意2个特征)对目标的最高分类准确率与本文方法相比,明显要低一些(约17%),这表明通过本文得到的特征组合的确是最优特征属性约简集,克服了传统方法在特征选择上的盲目性,提高了目标分类识别的准确率与效率。

4 结 论

本文提出的基于聚类分析与几何学的目标特征分类敏感性评估方法,一方面可以比较各类特征的优劣,为特定目标和时机进行特征选择和特征优化组合提供指导意见,这对分类器设计、识别算法设计、构建合理的多传感器网络具有极大的参考价值;另一方面,特征的敏感性也可作为相应识别技术的先验静态信息,结合各传感器判决包含的动态信息对特征识别技术进行赋权,参加融合识别。这表明,对于弹道目标识别这一复杂动态过程,以及将来可能出现的针对某些目标特征的反识别措施和对新提取目标特征的分类效果评判,都迫切需要对各类特征进行“适时评价与优化组合”。

参考文献(References)

- [1] 刘永祥,朱玉鹏,黎湘,等. 导弹防御系统中的目标综合识别模型[J]. 电子与信息学报, 2006, 28(4): 638-643.
(Liu Y X, Zhu Y P, Li X, et al. Integrated target discrimination model in mission in missile defense system[J]. J of Electronics & Information Technology, 2006, 28(4): 638-643.)
- [2] 吴瑕,周焰,崔建,等. 导弹防御系统中红外光电识别技术分析[J]. 红外与激光工程, 2009, 38(5): 759-766.
(Wu X, Zhou Y, Cui J, et al. Analyses on infrared optoelectronics recognition technology in missile defense system[J]. Infrared and Laser Engineering, 2009, 38(5): 759-766.)
- [3] Frank Hanson, Guy Beagher. Discriminating interceptor technology program(DITP) laser radar[C]. Laser Radar Technology and Applications IV. Orlando, 1999, 3707: 372-380.
- [4] 李海峰,张建洲,游志胜. 基于类距离的可分离性判据[J]. 计算机工程与应用, 2003, 39(26): 97-99.
(Li H F, Zhang J Z, You Z S. A separative criterion based on class distance[J]. Computer Engineering and Application, 2003, 39(26): 97-99.)
- [5] Yang B S, Han T, An J L. ART-KOHONEN neural network for fault diagnosis of rotating machinery[J]. Mechanical Systems and Signal Processing, 2004, 18(3): 645-657.
- [6] Marcelloni F. Feature selection based on a modified c-means algorithm with supervision[J]. Information Sciences, 2003, 151(6): 201-226.
- [7] 赵德滨,宋利利,闫纪红. 基于模糊聚类分析的特征识别法及其应用[J]. 计算机集成制造系统, 2009, 15(12): 2417-2423.
(Zhao D B, Song L L, Yan J H. Feature recognition method based on fuzzy clustering analysis and its application[J]. Computer Integrated Manufacturing Systems, 2009, 15(12): 2417-2423.)
- [8] 杨延光,周智敏,宋千,等. 一种VFGPIR序列图像特征评估与选择新方法[J]. 信号处理, 2009, 25(10): 1574-1581.
(Yang Y G, Zhou Z M, Song Q, et al. A novel feature evaluation and selection method for sequence images of VFGPIR[J]. Signal Processing, 2009, 25(10): 1574-1581.)
- [9] 孙雷,王新. 一种基于遗传算法和类内类间距离判据理论的特征选择方法[J]. 计算机工程与应用, 2004, 40(21): 178-181.
(Sun L, Wang X. A feature selection method based on genetic algorithm options and euclidean distance among all instances of different class[J]. Computer Engineering and Application, 2004, 40(21): 178-181.)
- [10] Lei Y G, He Z J, Zi Y Y. A new approach to intelligent fault diagnosis of rotating machinery[J]. Expert Systems with Applications, 2008, 35(4): 1593-1600.
- [11] 张岐龙,单甘霖,段修生,等. 基于特征空间中类别可分性判据的特征选择[J]. 火力与指挥控制, 2010, 35(6): 118-120.
(Zhang Q L, Shan G L, Duan X S, et al. Feature selection based on separability in feature space[J]. Fire Control & Command Control, 2010, 35(6): 118-120.)
- [12] Wang Y J, Lee H S. A clustering method to identify representative financial ratios[J]. Information Sciences, 2008, 178(4): 1087-1097.