基于线性判别分析的Choquet积分的符号模糊测度提取

王灯桂 杨 蓉

(深圳大学机电与控制工程学院 广东 深圳 518060)

摘 要 在解决分类问题时,建立在Choquet积分上的分类器以其非线性和不可加性的特点,扮演着越来越重要的角色。由于Choquet积分中的符号模糊测度可以描述各特征对结果的影响,因此Choquet积分在解决数据分类及融合问题方面具有显著的优势。但是,关于Choquet积分符号模糊测度值的求解,学术界一直缺乏有效的方法。目前最常用的方法是遗传算法,但是遗传算法在解决符号模糊测度值的优化问题时存在算法较为复杂、耗时较长等缺陷。由于符号模糊测度值在Choquet积分分类器中是决定性的重要参数,因此设计出一种有效的符号模糊测度提取方法十分必要。文中提出基于线性判别分析的Choquet积分符号模糊测度的提取方法,推导出在分类问题下Choquet积分的符号模糊测度值的解析式表达,其能够有效、快速地得出关键性参数。分别在人工数据集及基准实际数据集上进行测试与验证,实验结果表明所提方法能有效解决Choquet积分分类器中符号模糊测度的优化问题。

关键词 线性判别分析,Choquet积分,模糊测度,分类器

中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.02.040

Retrieving Signed Fuzzy Measure of Choquet Integral Based on Linear Discriminant Analysis

WANG Deng-gui YANG Rong

(College of Machatronics and Control Engineering, Shenzhen University, Shenzhen, Guangdong 518060, China)

Abstract For solving classification problems, Choquet integral classifier plays an increasingly important role by its nonlinear and nonadditivity. Especially, in the domain of solving the problem of data classification and fusion, Choquet integral has obvious advantages, because its signed fuzzy measure provides an effective representation to describe the interaction among contributions from predictive attributes to objective attributes. However, there is lack of an effective method to extract the signed fuzzy measure of Choquet integral. Currently, the most common used method is genetic algorithm, but the genetic algorithm is complex and time-consuming. Since the values of signed fuzzy measure are critical parameters in the Choquet integral classifier, it is necessary to design an efficient extraction method. Based on linear discriminant analysis, this paper proposed an extraction method for retrieving the values of signed fuzzy measure in the Choquet integral based on linear discriminant analysis, and derived the analytic expression of the signed measure value in Choquet integral under the classification problem, so that the key parameters can be obtained quickly and efficiently. This method was tested and validated on artificial data sets and benchmark data sets, respectively. The experiment results show that this method can effectively solve the optimization problem of signed fuzzy measure in Choquet integral classifier.

Keywords Linear discriminant analysis, Choquet integral, Fuzzy measure, Classifier

1 引言

在数据挖掘问题中,常用加权平均法处理各个特征属性对目标属性的影响。但是,使用加权平均这一线性方法时必须满足一个前提条件,即各个属性特征对某一给定目标特征的贡献之间没有任何交互作用。然而,在大量实际问题中,属性间的交互作用是广泛存在的。建立在测度理论上的符号模糊测度及其扩展具备描述此类非线性交互作用的强大能力[1-2]。因此,建立在符号模糊测度理论上的Choquet积分作为一种非线性的回归工具,经常被用于信息融合和数据挖掘问题中[3-7]。Choquet积分中的符号模糊测度由于可以描述各属性之间的交互作用,因此能合理解释属性间的关系,并把在线性模型中被误认为是随机的一部分影响转化成确定性影响,从而提高解的精度和可信度。

在解决信息融合和数据挖掘问题时,Choquet积分常作为聚合工具,把高维的特征属性投影到低维的属性空间上,进

到稿日期:2017-12-12 返修日期:2018-03-19 本文受国家自然科学基金项目(61773266),深圳市知识创新计划基础研究项目(JCYJ20170818144254033)资助。

王灯桂(1993-),男,硕士生,主要研究方向为模式识别与图像处理;杨 蓉(1976-),女,副教授,主要研究方向为模式识别、非线性积分及数据挖掘,E-mail:ryang@szu.edu.cn(通信作者)。

而把 n 维的分类问题降到低维或者一维的分类问题上[8-10]。文献[11]给出了模糊化的 Choquet 积分，可以处理模糊的被积函数并得出精确的结果，它可以解决含不同形式的非确定性数据集的分类问题。文献[12]提出了一种新的判别训练算法来训练基于融合算子的 Choquet 积分，并被成功应用于地雷检测。文献[13]提出了基于 Choquet 积分的一种更好的学习方法，其中，符号模糊测度被认为是间隔最大化问题，并通过割平面算法来求解。文献[14]提出了一种广义的加权 Choquet 积分，它通过遗传算法找到对结果贡献最大的特征属性值。

上述各种 Choquet 积分的成功应用都需要满足一个重要的前提，即得到准确的符号模糊测度值。Choquet 积分最先用于解决回归问题，由于回归问题的目标特征是连续的实值而非分类值，因此无法用线性判别分析的方法求解 Choquet 积分的符号模糊测度值。目前，常用的方法有两种：1)根据经验和实际问题的需求人为事先进行设定；2)利用全局的搜索方法，譬如遗传算法，根据学习样本进行参数的优化。显然，第一种方法依赖主观评判，因此不具有通用性；而第二种方法由于使用全局搜索的优化方法，存在算法复杂且时耗过大的缺陷。随后，Choquet 积分越来越多地被应用于分类问题，鉴于 Choquet 积分的非线性特性，更多的学者开始深究它的应用，沿用回归问题中求解符号模糊测度值的传统方法，忽视了 Choquet 积分重要参数求解的快速性及高效性，因此一直未提出直接解析的判别方法。本文提出了一种新颖的基于线性判别分析的 Choquet 积分符号模糊测度提取方法。建立于统计判别上的参数优化方法，能够推导出在分类问题下 Choquet 积分中符号模糊测度值的解析式表达，这样就摆脱了使用遗传算法求解时的繁琐计算过程，能快速且有效地求解出关键性参数。分别在人造数据集及基准实际数据集上进行测试与验证，结果证明所提方法能够有效且快速地获得在分类问题下的 Choquet 积分中的符号模糊测度。

本文第 2 节简要介绍 Choquet 积分的定义以及 Choquet 积分分类器；第 3 节详细介绍基于线性判别分析的 Choquet 积分的符号模糊测度的提取过程；第 4 节讨论实验结果；最后总结全文并探讨了该方法推广的可能性。

2 Choquet 积分及其分类器

本节将简要介绍模糊测度、符号模糊测度、Choquet 积分以及 Choquet 积分分类器的基本概念。

2.1 模糊测度与符号模糊测度

设 $X=\{x_1,x_2,\cdots,x_n\}$ 为特征属性集合。所有 X 的子集的集合称为 X 的幂集，表示为 $P(X)$。

定义 1 如果集函数 $\mu:P(X)\to[0,\infty)$ 在 $P(X)$ 上满足：$\mu(A)\leqslant\mu(B)$，$\forall A\subseteq B$，且 $\mu(\varnothing)=0$，则称 μ 为建立在 X 上的模糊测度。

模糊测度为集合中的每一个元素和每个可能的元素组合分配一个实数值。如果把 X 中的元素作为预测一个特定目标的一组特征属性，那么建立在 X 上的模糊测度可以清楚地阐述各个独立的特征属性以及它们之间所有可能的组合对目标属性的决策影响。由于模糊测度的非可加性，任意组合的特征属性对目标的影响不是简单地把各独立的特征属性的影响进行累加。因此，定义在 X 上的模糊测度具有明确的物理意义，它能表示特征属性之间的交互作用。

模糊测度的性质是单调性和空集上为零，这意味着模糊测度只允许它的值是非负的。模糊测度的单调性和非负性在实际应用中受到很多限制。因此，文献[15]定义了符号模糊测度，它是模糊测度的推广。

定义 2 集函数 $\mu:P(X)\to(-\infty,\infty)$ 称为符号模糊测度，其必须满足 $\mu(\varnothing)=0$。

符号模糊测度允许它的值是负数，且不受单调性的限制，因此能更灵活地描述特征属性及其联合对目标决策贡献的影响。

2.2 Choquet 积分

定义 3 令 $(X,P(X))$ 为测度空间，μ 为定义在 $P(X)$ 上的符号模糊测度，实值函数 $f:X\to(-\infty,+\infty)$ 的 Choquet 积分定义为：

$$\int f\mathrm{d}\mu=\int_{-\infty}^{0}[\mu(F_\alpha)-\mu(X)]\mathrm{d}\alpha+\int_{0}^{\infty}\mu(F_\alpha)\mathrm{d}\alpha \tag{1}$$

其中，$F_\alpha=\{x\mid f(x)\geqslant\alpha\}$，$\alpha\in(-\infty,+\infty)$。

为了方便计算，一般把 $f(x)$ 按升序排序，即使得不等式 $f(x_1')\leqslant f(x_2')\leqslant\cdots\leqslant f(x_n')$ 成立，这里 $(x_1',x_2',\cdots,x_n')$ 为集合 $\{x_1,x_2,\cdots,x_n\}$ 的某一种排序，这样式(1)就可以表示为：

$$\int f\mathrm{d}\mu=\sum_{i=1}^{n}[f(x_i')-f(x_{i-1}')]\cdot\mu(\{x_i',x_{i+1}',\cdots,x_n'\}) \tag{2}$$

其中，$f(x_0')=0$。

文献[16]提出了一种更为方便的利用内乘的计算方法，在实际编程中，Choquet 积分值可以根据式(3)计算：

$$\int f\mathrm{d}\mu=\sum_{j=1}^{2^n-1}z_j\mu_j \tag{3}$$

其中，

$$z_j=\begin{cases}\min\limits_{i:frc(\frac{j}{2^i})\in[\frac{1}{2},1)}f(x_i)-\max\limits_{i:frc(\frac{j}{2^i})\in[0,\frac{1}{2})}f(x_i), & \text{if it is}>0\text{ or }j=2^n-1\\ 0, & \text{otherwise}\end{cases} \tag{4}$$

其中，$j=1,2,\cdots,2^n-1$。

2.3 Choquet 积分分类器

在 n 维特征空间中的某一点 $(p_1,p_2,\cdots,p_n)$ 可以被认为是定义在全集 X 上的一个函数 f，其中 $p_i=f(x_i)(i=1,2,\cdots,n)$，它代表了各个特征属性的观察值。点 $(f(x_1),f(x_2),\cdots,f(x_n))$ 通过 Choquet 积分可以投影到一个实数轴上，得到一个相应的虚拟变量值 $\hat{Y}$，其定义如下：

$$\hat{Y}=\int f\mathrm{d}\mu \tag{5}$$

其中，μ 为符号模糊测度。

图 1 演示了从二维特征空间 $[0,1]\times[0,1]$ 到实数轴 L 的投影，映射函数 $\hat{Y}=\int f\mathrm{d}\mu$，其中 $\mu(\varnothing)=0$，$\mu(\{x_1\})=0.1$，$\mu(\{x_2\})=0.3$，$\mu(\{x_1,x_2\})=1.0$。$\hat{Y}$ 为函数 $f=(f(x_1),f(x_2))$，简记为 $f=(f_1,f_2)$。因此，不同的 $\hat{Y}$ 在特征空间

$[0,1]\times[0,1]$中由不同的几何曲面表示。当$\hat{Y}$分别等于 0.1，0.3，0.5 和 0.7 时，其几何投影线如图 1 所示。这些投影线由平行的折线组成。投影线为折线是因为模糊测度 μ 具有非可加性，它反映了特征属性之间对分类属性的相互作用。每个投影线都有两个分支，它们在实数轴 L 上交于一个顶点，实数轴 L 称为投影轴，由方程 $f_1=f_2$ 确定。实数轴 L 把特征空间$[0,1]\times[0,1]$分为两部分。与传统的只使用一个方向将特征空间的点投影到实轴上不同，该方法中，位于特征空间中被投影轴分成两部分的特征点沿着投影线的两个方向投射到投影轴 L 上。在上半部分，位于投影线 $0.7f_1+0.3f_2=\hat{Y}$ 上的点投影到投影轴 L 上对应的$\hat{Y}$；而在下部分，位于投影线 $0.1f_1+0.9f_2=\hat{Y}$上的点投影到投影轴 L 上对应的$\hat{Y}$。

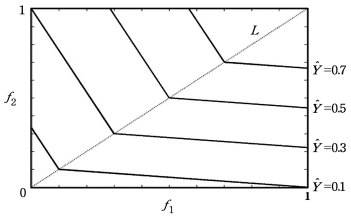

图 1　二维空间中 Choquet 积分的投影线和投影轴

Fig. 1　Contours of Choquet integral projection in two-dimensional

对于分类这样的决策问题，这些投影线发挥着边界的作用，每个决策边界都可以由特定常数$\hat{Y}$($\hat{Y}=\int f\mathrm{d}\mu$)来表示。相应地，通过 Choquet 积分把 n 维特征空间投影在投影轴 L 上后，分类边界可降低至一维，即在投影轴 L 上的一点$\hat{Y}$。一种基于 Choquet 积分的两类分类问题的分类策略可以表示为：

$$\text{若}\int f\mathrm{d}\mu\geqslant c,\text{则 } f\in A;\text{否则 } f\in B \tag{6}$$

其中，A 和 B 为两类的标签。在式(6)中，μ 为符号模糊测度且 $\mu(X)=1$；c 表示分类阈值。显然，模糊测度 μ 与分类阈值 c 均为未知的参数，需要从数据集中获取。这些参数在过去常用遗传算法来求取，需要消耗大量的时间来计算基于全局的进化过程，因此，本文提出一种可以由基于线性判别分析算法进行优化提取的方法。

3　基于线性判别分析的模糊测度提取

为了计算模糊测度 μ，将式(3)写成(2^n-1)维向量的内积形式：

$$\int f\mathrm{d}\mu=\sum_{j=1}^{2^n-1}z_j\mu_j=\boldsymbol{z}^{\mathrm{T}}\boldsymbol{\mu} \tag{7}$$

其中，$\boldsymbol{z}=(z_1,z_2,\cdots,z_{2^n-1})$，$\boldsymbol{\mu}=(\mu_1,\mu_2,\cdots,\mu_{2^n-1})$。

在模糊测度提取过程中，考虑 3 个空间。

1)模式空间：n 维向量特征空间，记为：$\boldsymbol{f}=(f(x_1),f(x_2),\cdots,f(x_n))$。

2)Choquet 空间：根据式(4)从模式空间映射到(2^n-1)维空间，将每个向量记为 $\boldsymbol{z}=(z_1,z_2,\cdots,z_{2^n-1})$。

3)投影空间：根据式(7)把特征向量从 Choquet 空间投影到一维实轴。

由于任何多类分类问题都可以分解为多组两类分类问题，因此下文重点讨论两类分类问题。

为了寻找投影方向 μ，我们需要在 Choquet 空间和投影空间中定义一些关键参数。

在 Choquet 空间中，令 Z_i 为 i 类的样本集，其中 i 为 1 或 2。$\boldsymbol{m}_i$ 为 i 类(2^n-1)维的样本均值向量，定义为：

$$\boldsymbol{m}_i=\frac{1}{N_i}\sum_{\boldsymbol{z}\in Z_i}\boldsymbol{z},\qquad i=1,2 \tag{8}$$

其中，N_i 为 i 类样本的数量。类内散布矩阵 $\boldsymbol{S}_i$ 可由式(9)计算：

$$\boldsymbol{S}_i=\sum_{\boldsymbol{z}\in Z_i}(\boldsymbol{z}_i-\boldsymbol{m}_i)(\boldsymbol{z}_i-\boldsymbol{m}_i)^{\mathrm{T}} \tag{9}$$

因此，总的类内散布矩阵为：

$$\boldsymbol{S}_w=\sum_{i=1}^{C}\boldsymbol{S}_i \tag{10}$$

类间散布矩阵为：

$$\boldsymbol{S}_b=(\boldsymbol{m}_1-\boldsymbol{m}_2)(\boldsymbol{m}_1-\boldsymbol{m}_2)^{\mathrm{T}} \tag{11}$$

在投影空间，让 Y_i 从(2^n-1)维的 Choquet 空间中投影到一维投影空间，投影后的值用 y 表示，其中 i 为 1 或 2。$\tilde{m}_i$ 为 i 类样本在一维中的均值，定义为：

$$\tilde{m}_i=\frac{1}{N_i}\sum_{y\in Y_i}y,\qquad i=1,2 \tag{12}$$

类内散布测度为：

$$\tilde{S}_i=\sum_{y\in Y_i}(y-\tilde{m}_i)^2,\quad i=1,2 \tag{13}$$

总的类内散布测度为：

$$\tilde{S}_w=\sum_{i=1}^{2}\tilde{S}_i \tag{14}$$

判别分析算法的目的是寻找投影方向 μ^*，使得在投影空间中，不同类别的点尽可能分离，而同一类别的点尽可能聚集。合理的 Choquet 判别标准定义为：

$$J_c(\boldsymbol{\mu})=\frac{(\tilde{m}_1-\tilde{m}_2)^2}{\tilde{S}_w} \tag{15}$$

把式(12)与式(13)代入式(15)，可得：

$$J_c(\boldsymbol{\mu})=\frac{\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{S}_b\boldsymbol{\mu}}{\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{S}_w\boldsymbol{\mu}} \tag{16}$$

为求得使 $J_c(\boldsymbol{\mu})$取极大值的 μ^*，可以采用 Lagrange 乘数法。若 $\boldsymbol{S}_w$ 非奇异，可得：

$$\boldsymbol{S}_w^{-1}\boldsymbol{S}_b\boldsymbol{\mu}^*=\lambda\boldsymbol{\mu}^* \tag{17}$$

方程(17)表明，$J_c(\boldsymbol{\mu})$最大化时的解向量 μ^* 等于求特征值问题中的广义特征向量。

根据式(11)中 $\boldsymbol{S}_b$ 的定义，式(17)中等号左侧的 $\boldsymbol{S}_b\boldsymbol{\mu}^*$ 可写成：

$$\begin{aligned}\boldsymbol{S}_b\boldsymbol{\mu}^*&=(\boldsymbol{m}_1-\boldsymbol{m}_2)(\boldsymbol{m}_1-\boldsymbol{m}_2)^{\mathrm{T}}\boldsymbol{\mu}^*\\&=(\boldsymbol{m}_1-\boldsymbol{m}_2)R\end{aligned}$$

其中 $R=(\boldsymbol{m}_1-\boldsymbol{m}_2)^{\mathrm{T}}\boldsymbol{\mu}^*$，是一个标量。因为我们要求最佳的投影方向，所以这个标量的大小不会影响最终的结果。式(17)可写成：

$$\lambda\boldsymbol{\mu}^*=\boldsymbol{S}_w^{-1}(\boldsymbol{m}_1-\boldsymbol{m}_2)R$$

化简得：

$$\boldsymbol{\mu}^*=\frac{R}{\lambda}\boldsymbol{S}_w^{-1}(\boldsymbol{m}_1-\boldsymbol{m}_2)$$

其中，R/λ 为尺度因子，可以忽略。最后得出模糊测度的求解结果为：

$$\boldsymbol{\mu}^* = \boldsymbol{S}_w^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_2) \tag{18}$$

当模糊测度 $\boldsymbol{\mu}^*$ 确定后，在投影空间中分类阈值可由等式：

$$c = (N_1 \cdot \tilde{m}_1 + N_2 \cdot \tilde{m}_2)/(N_1 + N_2)$$

或

$$c = (\tilde{m}_1 + \tilde{m}_2)/2$$

计算处理。

4 实验结果与分析

为了证明 Choquet 判别分析算法的可行性，我们在一系列的人工数据集和真实数据集上进行了实验。实验平台为 MATLAB2012b。

4.1 人工数据集

人工数据集是由计算机随机产生的一个服从正态分布的两维特征属性、两类目标属性的数据集，共有 400 个样本。将其随机分成 200 个样本的训练数据集和 200 个样本的测试数据集。先预设 $\mu(\{x_1\})=0.1$，$\mu(\{x_2\})=0.5$，$\mu(\{x_1,x_2\})=1.0$，分类阈值 $c=0.5$。

对训练样本采用第 3 节提出的 Choquet 判别分析算法训练出实际的符号模糊测度 μ 值，训练得到的 μ 值与预设的 μ 值的比较结果如表 1 所列。由表 1 可知，预设值与训练值相当接近，两者的均方误差只有 0.0056。表 1 中的比较也验证了第 3 节所述判别分析方法的可行性，即基于 Choquet 积分分类器的符号模糊测度值通过判别分析算法能够被正确提取出来。

表 1 提取值与预设值的比较

Table 1 Comparison between retrieved values and preset values

参数	预设值	预设值标准化	提取值	提取值标准化
$\mu(\{x_1\})$	0.1	0.0476	0.0143	0.0524
$\mu(\{x_2\})$	0.5	0.2381	0.0666	0.2441
$\mu(\{x_1,x_2\})$	1.0	0.4762	0.1278	0.4685
c	0.5	0.2381	0.0641	0.2350

用上述提取出来的符号模糊测度值构造基于 Choquet 积分的分类器，该分类器在测试集上的分类结果如图 2 所示，其中粗实线是 Choquet 积分的投影方向，两条虚线是分类的分界线，实心圆点表示 A 类，空心圆点表示 B 类，x 表示错分点。在分类阈值为 0.0641 的情况下，错分率只有 0.5%。

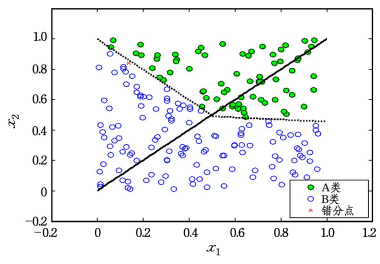

图 2 基于 Choquet 线性判别的测试集的分类结果

Fig. 2 Classification results based on Choquet discriminant analysis

表 2 列出了用十折交叉验证方法得出的符号模糊测度值，整个训练数据集有 200 个样本。预置的模糊测度值与十折交叉验证得出的提取值的均值之间的比较如表 3 所列。由表 3 可知，十折交叉验证的提取值的均值与符号模糊测度的预设值非常接近，二者的均方误差只有 0.0061。

表 2 十折交叉验证下提取的符号模糊测度值

Table 2 Retrieved values of signed fuzzy measure in ten-fold cross-validation

实验序号	$\mu(\{x_1\})$	$\mu(\{x_2\})$	$\mu(\{x_1,x_2\})$	c
1	0.0013	0.0062	0.0131	0.0064
2	0.0014	0.0063	0.0133	0.0065
3	0.0014	0.0063	0.0132	0.0065
4	0.0014	0.0064	0.0132	0.0065
5	0.0014	0.0063	0.0132	0.0065
6	0.0015	0.0063	0.0134	0.0066
7	0.0014	0.0063	0.0133	0.0066
8	0.0015	0.0062	0.0132	0.0065
9	0.0012	0.0061	0.0130	0.0064
10	0.0014	0.0063	0.0134	0.0066

表 3 十折交叉验证下提取值与预设值的比较

Table 3 Comparison between retrieved values and preset values in ten-fold cross-validation

参数	预设值标准化	提取值平均值	提取值标准化
$\mu(\{x_1\})$	0.0476	0.00139	0.0507
$\mu(\{x_2\})$	0.2381	0.00627	0.2283
$\mu(\{x_1,x_2\})$	0.4762	0.01323	0.4828
c	0.2381	0.00651	0.2376

在人工数据集上的实验表明，Choquet 判别分析算法可以高效地提取出 Choquet 分类模型中的符号模糊测量值。该方法优于以往基于遗传算法的优化方法。Choquet 判别分析算法的解析形式，使得符号模糊度量值的提取更简单、省时。

4.2 实际数据集

为了进一步证明本文方法在实际数据集上的有效性，选择了 UCI 数据库中 5 个著名的数据集进行验证。

1)Iris：3 类，4 个特征，每一类有 50 个数据，共有 150 个数据；

2)Breast cancer：2 类，9 个特征，包含 699 个数据；

3)Pima Indians diabetes：2 类，8 个特征，包含 768 个数据；

4)Haberman's survival：2 类，3 个特征，包含 306 个数据；

5)Blood transfusion：2 类，5 个特征，包含 748 个数据。

基于判别分析算法提出的符号模糊测度的 Choquet 积分分类器与其他 9 种分类器在这 5 个数据集上的分类结果比较如表 4 所列。9 种分类器分别为朴素贝叶斯分类器、贝叶斯网络、NBtree、模型树、SMO、RBF 网络、模糊推理分类器、模糊决策树、基于遗传算法的 Choquet 积分分类器。

表 4 中加粗字体为分类器在该数据集中最低的错分率。由表 4 可知，在这 5 个数据集上，本文方法的分类效果的表现较优。与其他非 Choquet 积分分类器相比，本文方法在 Breast cancer，Pima Indians diabetes 数据集上的表现最好，在其他 3 个数据集上仍然达到中上水平，这得益于 Choquet 积分分类器的非线性以及强大的特征属性融合能力。与基于遗传算法的 Choquet 积分分类器相比，本文提出的方法在 Iris，Haberman's survival 数据集上表现得较为优秀，而在其他 3 个数据集上的表现较差。这表明当特征属性的维度越高时，使用线性判别分析求取符号模糊测度的表现越差，这是由于使用线性判别分析求取符号模糊测度时，要求总类内散布矩阵 $\boldsymbol{S}_w$ 非奇异，而当特征属性维数越高时，总类内散布矩阵 $\boldsymbol{S}_w$ 越难满足非奇异这一条件。本文在求解奇异的总类内散布矩

阵 $\boldsymbol{S}_w$ 时先忽略全零行，构造一个非奇异的总类内散布矩阵；求解完符号模糊测度时，再在相应的位置补回零，从而使维度保持一致。但是这种方法影响了本文最终的结果，未来会寻找一种更优的处理奇异矩阵的方法。本文方法虽然在高维数时精度有所下降，但是与遗传算法相比，效率大大提高。采用本文提出的线性判别分析进行求取时，用时只需几秒，而使用遗传算法求取时，需要几百秒以上，这证明了本文方法在提取符号模糊测度上具有较大的优势。

表 4　各种分类器 5 个数据集上的训练和测试的错分率

Table 4　Comparison of misclassification rates of ten classifiers on five real data sets

（单位：%）

分类器	Iris		Breast cancer		Pima Indians diabetes		Haberman's survival		Blood transfusion	
	训练	测试	训练	测试	训练	测试	训练	测试	训练	测试
朴素贝叶斯分类器	4.0	4.2	3.9	4.5	23.7	25.9	24.2	29.8	25.0	28.4
贝叶斯网络	5.3	6.0	2.7	3.0	21.7	26.7	25.8	34.6	24.06	29.7
NBtree	2.7	2.9	2.7	3.1	25.7	28.1	22.9	24.9	20.5	24.7
模型树	**2.0**	**2.1**	2.3	2.3	22.7	26.7	25.5	31.8	19.8	**20.7**
SMO	3.3	4.1	3.0	3.5	22.5	29.7	25.2	34.7	23.8	30.9
RBF 网络	2.7	2.8	3.6	3.9	25.6	30.9	24.8	30.5	21.8	28.1
模糊推理分类器	3.3	4.6	0.7	**0.9**	19.1	28.3	38.2	40.2	32.4	50.6
模糊决策树	4.0	4.8	3.0	3.0	20.1	27.1	**22.8**	**23.8**	19.1	30.3
基于遗传算法的 Choquet 积分分类器	4.0	5.1	3.0	3.5	**2.6**	9.4	29.8	32.4	**16.2**	21.4
本文方法	2.2	3.0	**0.4**	8.7	18.0	**22.4**	25.8	26.3	23.4	23.6

结束语　本文提出了一种新的基于判别分析算法的 Choquet 积分分类器的符号模糊测度提取方法。首先，含符号模糊测度的 Choquet 积分的计算在 Choquet 空间中表现为线性方程，然后借鉴 Fisher 判别分析方法，得出一种有效的提取符号模糊测度的线性判别分析法，从训练集中把符号模糊测度提取出来。该方法的可行性和有效性已在一系列的人工数据集和现实数据集中得到验证。与常用的遗传算法训练出符号模糊测度相比，本文方法能提高基于 Choquet 积分的分类器解决现实世界中的分类问题的能力，使得分类更灵活、高效。

参 考 文 献

[1] DENNEBERG D. Non-additive measure and integral[C]// Theory & Decision Library, 1994, 27(2): 872-879.

[2] WANG Z, KLIR G J. Fuzzy measure theory [J]. Plenum Berlin, 1992, 35(1-2): 3-10.

[3] XU K, WANG Z, HENG P A, et al. Classification by nonlinear integrals projections[C]// IEEE Transactions on Fuzzy Systems. IEEE, 2003: 187-201.

[4] TEHRANI A F, CHENG W, HULLERNRIER E. Preference learning using the Choquet integral: The case of multipartite ranking [C] // IEEE Transactions on Fuzzy Systems. IEEE, 2012: 1102-1113.

[5] WANG H X. On Choquet integral and set-valued Choquet integral with their applications in finance [D]. Beijing: Beijing University of Technology. 2013. (in Chinese)
王洪霞. Choquet 积分和集值 Choquet 积分及其在金融中的应用[D]. 北京：北京工业大学，2013.

[6] CHEN J F, HE Q. Determination of fuzzy measure in Choquet fuzzy integral fusion model[J]. Journal of Hebei University, 2006, 26(4): 354-357. (in Chinese)
陈俊芬，何强. Choquet 模糊积分融合模型中模糊测度的确定[J]. 河北大学学报，2006，26(4)：354-357.

[7] LI T S. Research on multi classifier fusion model based on Choquet integral[D]. Hebei: Hebei University, 2010. (in Chinese)
李铁松. 基于 Choquet 积分的多分类器融合模型研究[D]. 河北：河北大学，2010.

[8] WANG Z, YANG R, SHI Y. A new nonlinear classification model based on cross-oriented Choquet integrals [C]// International Conference on Information Science and Technology Nanjing. China, 2011: 176-181.

[9] FANG H, RIZZO M L, WANG H, et al. A new nonlinear classifier with a penalized signed fuzzy measure using effective genetic algorithm [J]. Pattern Recognition, 2010, 43(4): 1393-1401.

[10] GRACISCH M, MUROFUSHI T, SUGENO M. Fuzzy measures and integral[M] // Fundamentals of Uncertainty Calculi with Applications to Fuzzy Inference. Netherlands: Springer, 1995.

[11] YAGER R R. Evaluating Choquet integrals whose arguments are probability distributions[J]. IEEE Transactions on Fuzzy Systems, 2016, 24(4): 957-965.

[12] XU K, WANG Z, HENG P A, et al. Classification by nonlinear integral projections [J]. IEEE Transactions on Fuzzy Systems, 2003, 11(2): 187-201.

[13] YANG R, WANG Z, HENG P A, et al. Classification of Heterogeneous fuzzy data by Choquet integral with fuzzy-valued integrand [J]. IEEE Transactions on Fuzzy Systems, 2007, 15(5): 931-942.

[14] KOCHI N, WANG Z. Solving nonlinear programming problems based on the Choquet integral by a genetic algorithm [J]Journal of Intelligent and Fuzzy Systems, 2015, 29(1): 437-442.

[15] BUSTINCE H, GALAR M. A new approach to interval-values Choquet integrals and the problem of ordering in interval-values fuzzy set applications [J]. IEEE Transactions on Fuzzy Systems, 2013, 21(6): 1150-1162.

[16] WANG Z. A new genetic algorithm for nonlinear multi-regressions based on generalized Choquet integrals. [C]// 12th IEEE International Conference on Fuzzy Systems. IEEE, 2003: 819-821.