第57卷 第2期 西 安 交 通 大 学 学 报 Vol. 57 No. 2
2023年2月 JOURNAL OF XI'AN JIAOTONG UNIVERSITY Feb. 2023

利用证据理论的多分类支持向量数据描述算法

张世醒，韩德强，范晓婧
（西安交通大学电信学部，710049，西安）

摘要：针对原始多分类支持向量数据描述(SVDD)算法及其拓展算法忽略超球体之间的差异，且未能充分利用超球体的输出信息等问题，提出一种利用证据理论的多分类支持向量数据描述(证据SVDD多分类)算法。首先，为每一类样本训练一个超球体，并计算每个超球体的正确率与紧密程度；接着使用上一步得到的正确率与紧密程度计算每个超球体的可靠程度；然后，根据超球体的输出信息与可靠程度计算样本的信度函数，信度函数的生成方式采用三焦元法和基于评价矩阵的方法；最后，根据Dempster组合规则融合上一步得到的信度函数，使用Pignistic法将融合后的信度函数转换为概率做出最终的判决。在两个人工数据集和多个UCI数据集上进行实验，结果表明，证据SVDD多分类算法相较传统算法可以获得更好的分类性能；在多个数据集上的仿真结果表明，证据SVDD多分类算法比传统的SVDD多分类算法有3%的精度提升。

关键词：证据理论；支持向量数据描述；多属性决策；信度函数
中图分类号：TP391 **文献标志码**：A
DOI：10.7652/xjtuxb202302016 **文章编号**：0253-987X(2023)02-0151-10

Multi-Class Support Vector Data Description Based on Evidence Theory

ZHANG Shixing, HAN Deqiang, FAN Xiaojing
(Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: Original multi-class support vector data description (SVDD) algorithm and its extension algorithm ignore the differences among hyperspheres and fail to make full use of the output information of hyperspheres. To address these problems, a multi-class support vector data description algorithm based on evidence theory (evidential multi-class SVDD algorithm) is proposed. Firstly, a hypersphere is trained for each class of samples, and the accuracy and closeness of each hypersphere are calculated. Secondly, the accuracy and closeness obtained in the previous step are used to calculate the reliability of hypersphere. After that, the output information and reliability of hypersphere are used to calculate the belief functions of samples, which are generated by two methods: the triple focal element method and the method based on payoff matrix. In the end, Dempster combination rule is used to fuse the belief functions, and Pignistic method is used to convert the fused belief functions into probability for final decision. Extensive experimental results on two artificial datasets and multiple UCI datasets show that the evidential multi-class SVDD algorithm achieves better classification performance than the traditional algorithm. Simulation results on multiple datasets show that the evidential multi-class SVDD algorithm has a 3% improvement in accuracy compared with the traditional multi-class SVDD algorithm.

收稿日期：2022-07-11。 作者简介：张世醒(1994—)，男，博士生；韩德强(通信作者)，男，教授，博士生导师。
网络出版时间：2022-10-13 网络出版地址：https://kns.cnki.net/kcms/detail/61.1069.T.20221012.1642.004.html

Keywords: evidence theory; support vector data description; multiple attribute decision making; belief functions

支持向量数据描述(support vector data description,SVDD)是一种著名的单分类算法,该算法运算复杂度低,训练对样本数量要求不高。近年来SVDD算法已经广泛应用于异常检测[1]、故障诊断[2]和图像分类[3]等领域。经典SVDD算法适用于单分类和二分类问题,实际应用场景中的模式分类问题,除二分类之外,更多的是多分类问题。

近年来,学者们针对多分类SVDD算法进行了研究[4-8]。原始多分类SVDD算法(O-SVDD)在训练阶段分别为每类样本训练一个超球体;在测试阶段,比较待测样本到每个超球体球心的距离实现分类。杨晨等[9]除考虑样本到每个超球体球心的距离,还额外考虑了每类样本的占比。Azami等[10]将每类样本的占比作为先验,实现了多分类SVDD算法的概率输出。Mu等[11]将SVDD算法的输出作为新特征,使用线性判别分析及近邻准则实现了多分类。对于每类样本,林雄等[12]训练了不止一个超球体,期望对每类样本进行更精准的描述。蔡金燕等[13]使用待测样本的K近邻辅助多分类SVDD算法进行判决。Hao等[14]通过改进后的相对距离定义样本对每个超球体的隶属程度,进而实现多分类。

虽然上述方法对O-SVDD算法进行了改进,但是都以同等的信任对待每个超球体,并没有考虑这些超球体的输出信息是否可靠,这将导致算法容易受可靠程度较低的超球体的影响;上述方法均只计算样本对每个超球体的隶属程度,丢弃了大量的不确定信息。

为了对样本进行更精确的描述,并减少超球体可靠程度较低的不利影响,应该保留样本的不确定信息并对超球体的输出信息进行修正。在证据框架下,可以通过对混合类赋值保留样本的不确定信息,对全集赋值来减少不可靠超球体的不利影响。证据理论[15-17]已广泛应用于模式分类[18-20]、聚类[21-23]和回归[24]等问题。本文提出基于证据理论的多分类SVDD算法:首先在证据框架下对样本进行建模来保留样本的不确定信息,并使用超球体的正确率与紧密程度对其可靠程度进行表征;然后使用超球体的可靠程度修正其输出信息并生成证据;最后使用组合规则融合多条证据做出判决。

1 SVDD算法简介

1.1 单分类SVDD算法

受支持向量机的启发,Tax等[25]第一次提出SVDD算法,该算法通过寻找一个能够紧密包裹正样本的超球体来实现正负样本的分类。支持向量数据描述示意图如图1所示,图中红色曲线代表训练得到的超球体,其中 x_1、x_2 分别表示样本的第一维特征、第二维特征。待测样本落在曲线(曲面)内部,则判定为正样本;否则,判定为负样本。该曲线(曲面)仅由支持向量决定,故称其为支持向量数据描述。

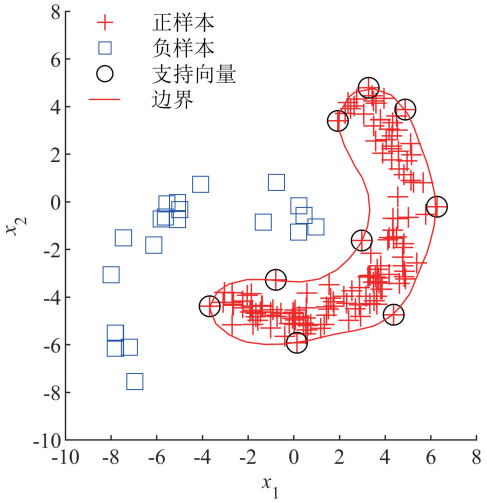

图1 支持向量数据描述示意图

Fig. 1 the schematic of SVDD

SVDD算法的数学表述如下:假设正样本为 $x \in \mathbf{R}^{n\times d}$,其中 n 为样本数,d 为样本的特征维度。先通过非线性变换 $\varphi(x)=z$ 将样本从原始空间映射到核空间,然后在核空间构造一个体积最小的超球体包裹正样本。SVDD算法可表示为以下优化问题

$$\begin{aligned}&\min_{a,R,\xi_i} R^2+C\sum_{i=1}^{n}\xi_i\\ &\text{s.t.}\quad \|\varphi(x_i)-a\|^2\leqslant R^2+\xi_i,\xi_i\geqslant 0,\\ &\qquad \forall i=1,2,\cdots n\end{aligned}\tag{1}$$

式中:R 为超球体的半径;a 为超球体的球心;ξ_i 为松弛因子;C 为平衡超球体体积和样本误差的惩罚因子。通过拉格朗日乘子法,式(1)可转换为以下对偶问题

$$\min_{\alpha_i}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j K(x_i,x_j)-\sum_{i=1}^{n}\alpha_i K(x_i,x_i)$$
$$\text{s.t. } 0\leqslant\alpha_i\leqslant C,\ \sum_{i=1}^{n}\alpha_i=1 \tag{2}$$

式中：α_i 为样本 x_i 对应的拉格朗日系数；$0<\alpha_i<C$，对应的样本被称为支持向量；$K(x_i,x_j)$ 为核函数。常见的核函数包括高斯核、线性核和指数核等，SVDD 算法常选用高斯核作为核函数

$$K(x_i,x_j)=\exp\left(-\frac{\|x_i-x_j\|^2}{2\sigma^2}\right) \tag{3}$$

对偶问题是一个标准的二次规划问题。解出所有的 α_i 后，可得超球体的球心、半径

$$a=\sum_{i=1}^{n}\alpha_i\varphi(x_i) \tag{4}$$

$$R=\left(K(x_v,x_v)-2\sum_{i=1}^{n}\alpha_i K(x_v,x_i)+\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j K(x_i,x_j)\right)^{1/2} \tag{5}$$

式中 x_v 为支持向量。待测样本 x_{test} 距超球体球心的距离可表示为

$$d=\left(K(x_{\text{test}},x_{\text{test}})-2\sum_{i=1}^{n}\alpha_i K(x_{\text{test}},x_i)+\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j K(x_i,x_j)\right)^{1/2} \tag{6}$$

若 $d\leqslant R$，表示待测样本属于正样本，反之表示待测样本属于负样本。α_i 越大，最终的模型受样本 x_i 的影响越大。加权 SVDD 算法就是通过改变优化变量的上界来体现样本的重要性，即 $0\leqslant\alpha_i\leqslant C_i$。

能够获得负样本时，在经典 SVDD 算法的优化问题上增加负样本限制，可以获得更好的分类边界。Tax 等[25]利用负样本限制提出了带有负样本的 SVDD 算法（N-SVDD），该算法构造了一个紧密包裹正样本的超球体，使正样本尽可能落在超球体内部，同时需要使负样本尽可能落在超球体之外。N-SVDD 算法的损失函数为

$$\arg\min_{a,R,\xi_i,\xi_j}\{R^2+C_1\sum_i\xi_i+C_2\sum_j\xi_j\}$$
$$\text{s.t. } \|x_i-a\|\leqslant R^2+\xi_i,\ \xi_i\geqslant 0,\forall i=1,2,\cdots,n_+$$
$$\|x_j-a\|\geqslant R^2+\xi_j,\ \xi_j\geqslant 0,\forall j=1,2,\cdots n_- \tag{7}$$

式中：n_+ 表示正类样本个数；n_- 表示负类样本个数。式(7)的对偶形式为

$$\arg\max_{\alpha}\{\sum_{i=1}^{n_++n_-}\alpha_i y_i k(x_i,x_i)-\sum_{i,j=1}^{n_++n_-}\alpha_i\alpha_j y_i y_j k(x_i,x_j)\}$$
$$\text{s.t. }\sum_{i=1}^{n_++n_-}\alpha_i y_i=1$$
$$0\leqslant\alpha_i\leqslant C_1,\forall i,y_i=1$$
$$0\leqslant\alpha_j\leqslant C_2,\forall j,y_j=-1 \tag{8}$$

1.2　多分类 SVDD 算法

近年来，很多学者针对多分类 SVDD 算法进行了研究，常用的多分类 SVDD 算法如下。

(1)原始多分类 SVDD 算法（O-SVDD）[12]。在训练阶段，为每类样本训练一个超球体；在测试阶段，分析待测样本落入各超球体的情况实现分类。该算法根据样本到每个超球体球心的相对距离进行判决，其判决函数为

$$f(x)=\arg\min_i d_i(x)/r_i \tag{9}$$

式中：$d_i(x)=\sqrt{\|\varphi(x)-a_i\|^2}$ 为核空间内样本 x 到第 i 类超球体球心的距离；r_i 为第 i 个超球体的半径。

(2) Yang-SVDD 算法[9]。除考虑待测样本到每个超球体球心的相对距离之外，额外考虑每类样本的个数，其判决函数为

$$f(x)=\arg\max_i n_i r_i^h/d_i \tag{10}$$

式中：h 为可调参数；n_i 为第 i 类训练样本的个数。Yang-SVDD 算法使用每类样本的数目对相对距离进行加权。该方法在平衡数据集上表现与 O-SVDD 非常类似；当 $h=1$ 时，该算法将退化为 O-SVDD。

(3)概率 SVDD 算法（P-SVDD）[10]。该方法首先使用 sigmoid 函数将 O-SVDD 算法的输出转换为类条件密度，即

$$P(x\mid w_i)=\frac{a}{1+\exp(bd_i/r_i+c)} \tag{11}$$

式中 a、b、c 为需要估计的参数。然后计算样本的先验概率，通常假设先验概率相等或者等于 n_i/n，其中 n_i 为第 i 类样本的个数，n 为样本总数，最后通过贝叶斯判决做出最终判决。

上述方法均以同等的信任对待每个超球体，未判断每个超球体的输出信息是否可靠，并且样本的不确定信息未被充分利用。

1.3　O-SVDD 算法及其拓展算法存在的缺陷

对于分布不均的两类样本，第一类样本分布较为分散，第二类样本分布较为紧密，可靠程度不同的两类超球体如图 2 所示。分别使用两类样本训练超球体，得到第一类超球体（蓝色曲线）和第二类超球体（红色曲线）。由图 2 可以看出：第一类超球体包裹了大量的第二类样本，所以其正确率明显低于第二类超球体，并且第二类超球体相较第一类超球体能更紧密的包裹其对应类别样本；SVDD 算法的训练目标是训练一个能够紧密包裹目标样本的超球

体，并使非目标样本尽可能落在该超球体之外。第二类超球体的可靠程度明显高于第一类超球体。

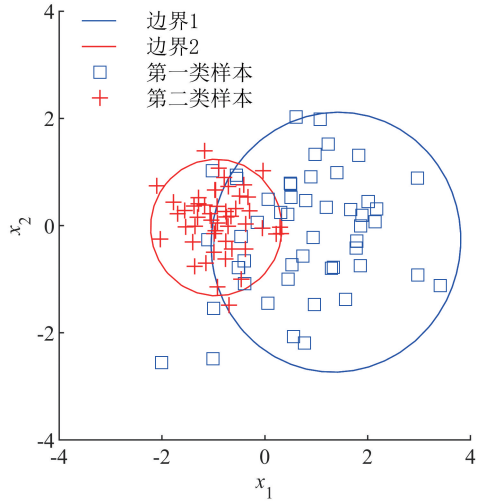

图 2 可靠程度不同的两类超球体

Fig. 2 Two hyperspheres with different reliability

O-SVDD 算法及其拓展算法均以同等的信任对待每个超球体，主要根据样本到每个超球体球心的距离对样本进行分类，即使某些超球体的可靠程度很低。为了缓解上述问题，可以使用超球体的可靠程度对其输出信息进行修正。另外，O-SVDD 算法及其拓展算法只计算待测样本属于每个超球体的隶属度，未充分利用样本的不确定信息。针对上述问题，本文利用证据理论，通过引入混合类与全集来分别表示样本的不确定信息以及每个超球体的可靠程度。基于证据的方法在一定程度上可以抑制图 2 中出现的问题。

2 基于证据的多分类 SVDD 算法

2.1 证据理论简介

信度函数理论也叫 Dempster-Shafer 证据理论(DST)。证据理论的辨识框架为 $\Theta=\{\theta_1,\theta_2,\cdots,\theta_q\}$，$2^\Theta$ 表示辨识框架的幂集。若 $m:2^\Theta\to[0,1]$ 满足

$$\begin{cases} m(\phi)=0 \\ \sum\limits_{A\subseteq\Theta} m(A)=1 \end{cases} \tag{12}$$

则称 m 是辨识框架下的基本信度分配(BBA)。

辨识框架 Θ 上某个命题 A 的信任函数记为 Bel(A)。Bel(A)表示对命题 A 信任程度的下限，定义为

$$\mathrm{Bel}(A)=\sum_{B\subseteq A} m(B),\forall A\subseteq\Theta \tag{13}$$

辨识框架 Θ 上某个命题 A 的似真函数记为 Pl(A)。Pl(A)表示对命题 A 信任程度的上限，定义为

$$\mathrm{Pl}(A)=\sum_{B\cap A\neq\phi} m(B),\forall A\subseteq\Theta \tag{14}$$

当得到多条证据时，可以使用组合规则对多条证据进行组合。Dempster 组合规则是一种常用的组合规则，其计算过程为

$$(m_1\oplus m_2)(B)=\begin{cases} 0, & B=\phi \\ \dfrac{1}{1-Z}\sum\limits_{A\cap C=B} m_1(A)m_2(C), & B\neq\phi \end{cases} \tag{15}$$

式中 $Z=\sum\limits_{B\cap C=\phi} m_1(B)m_2(C)$ 表示两个证据之间的冲突。获得融合后的证据，往往需要将其转换为概率做出最终决策。Pignistic 概率转换是一种常用的转换方法，即

$$\mathrm{Bet}P(\theta_j)=\sum_{B\subseteq\Theta}\frac{|\theta_j\cap B|}{|B|}\frac{m(B)}{1-m(\phi)} \tag{16}$$

该方法将混合焦元的信度平均分配到该混合焦元的各个元素中。

根据证据理论的基本概念，样本属于全集的信度一定程度上可以反映一条证据的不确定程度，也可以反映一条证据的可靠程度。可以先使用每个超球体的可靠程度对其输出信息打折扣，然后将其赋给全集来体生成证据。

2.2 超球体的可靠程度

O-SVDD 算法以同等的信任对待每个超球体，然而每类样本稀疏程度不同，并且每类超球体的分类正确率也不同，但此做法有一定的缺陷。SVDD 算法的目标是训练一个能够紧密包裹目标样本的超球体，所以一个可靠的超球体应该兼顾正确率与紧密度。本小节以第 i 个超球体为例来说明超球体可靠程度的计算过程。

第 i 个超球体的分类正确率为

$$r_{\mathrm{acc}}^i=\frac{n_i'+(\bar{n}_i-\bar{n}_i')}{n} \tag{17}$$

式中：n 为样本总数；n_i 为第 i 类样本数；n_i'为落入第 i 类超球体中第 i 类样本数(简称为分类正确的第 i 类样本)；$\bar{n}_i=n-n_i$ 为不属于第 i 类的样本数目；$\bar{n}_i'$为落入第 i 类超球体中非第 i 类样本的数目；$\bar{n}_i-\bar{n}_i'$为落在第 i 类超球体之外的非第 i 类样本数目。

使用内聚度来定义第 i 个超球体的紧密程度

$$l_i=1-\sum_{\substack{x\in C_i \\ \|x-o_i\|<r_i}}\frac{1}{n_i'}K(x,o_i) \tag{18}$$

式中 $\sum\limits_{\substack{x\in C_i \\ \|x-o_i\|<r_i}}\frac{1}{n_i'}K(x,o_i)$ 表示第 i 类样本到第 i 类

超球体球心的平均距离。紧密程度不同的两类超球体如图 3 所示，对于第一类超球体，式(18)中求和部分表示分类正确的第一类样本(除箭头所指外，所有其他的红色五角星)到红色超球体球心距离的平均值，该值越小，表示超球体越紧密。由图 3 可以看出，红色超球体相较于蓝色超球体更加紧密，这与式(18)得到的结果相一致。

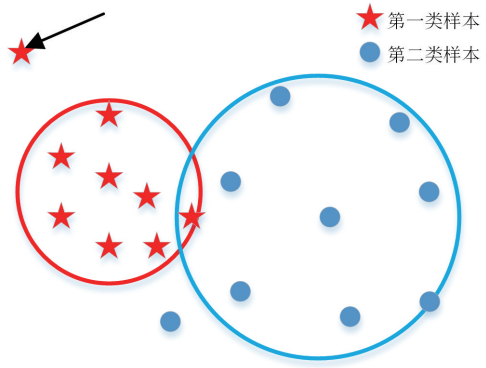

图 3　紧密程度不同的两类超球体

Fig. 3　Two hyperspheres with different compactness

当紧密程度与正确率的量纲相差较大时，为了统一二者的量纲，可以使用归一化的紧密程度

$$l'_i = \frac{l_i}{l_{\max}} \tag{19}$$

式中 $l_{\max}$ 为紧密程度最大的超球体所对应值。

综上可得超球体可靠程度的计算式

$$\gamma_i = r^i_{\mathrm{acc}} l'_i \tag{20}$$

2.3　基于证据的多分类 SVDD 算法

使用证据理论的关键在于信度分配，证据生成没有固定的方法，需要结合具体应用。以多分类 SVDD 算法为例，每个超球体有 2 个输出，即样本属于该超球体的隶属度和样本不属于该超球体的隶属度。以上输出正好可以对映证据中样本属于某类的信度和属于混合类的信度。当超球体的可靠程度较低时，其输出信息应该被高度怀疑，此时无法根据其输出信息做出决策，应该将大部分信度赋给全集表示样本归属无法判断。

基于以上思路生成证据，本文使用两类方法进行信度分配，并根据两类信度分配方法分别提出三焦元证据 SVDD 算法和基于证据多属性决策的 SVDD 算法。

2.3.1　三焦元证据 SVDD 算法

三焦元证据 SVDD 算法(TriE-SVDD)根据每个超球体的可靠程度对其输出信息打折扣，然后生成证据。对于第 i 个超球体，将其不可靠程度分配给全集，然后计算待测样本属于第 i 类的信度 $m(\theta_i)$，剩余信度为待测样本不属于第 i 类的信度，即

$$\begin{cases} m_{x_i}(\theta_j) = \gamma_j s_{ij} \\ m_{x_i}(\bar{\theta}_j) = \gamma_j (1 - s_{ij}) \\ m_{x_i}(\Theta) = 1 - \gamma_j \end{cases} \tag{21}$$

$$s_{ij} = \frac{1}{1 + \mathrm{e}^{a(d_{ij} - r_j) + b}} \tag{22}$$

式中：s_{ij} 为超球体 j 对于样本 i 的评价；d_{ij} 为样本 i 到第 j 个超球体球心的距离；r_j 为第 j 个超球体的半径。

对于 q 分类问题，可以产生 q 组证据。通过式(15)融合 q 组证据，得到融合后的证据，再将融合后的证据转换为概率做出最终的判决。

2.3.2　基于证据多属性决策的 SVDD 算法

基于证据多属性决策的 SVDD 算法将每个超球体的输出修正后作为一种决策，通过多个超球体的决策获得对应样本的评价矩阵，然后根据评价矩阵得到证据。依据不同超球体 $A = \{\mathrm{svdd}_1, \mathrm{svdd}_2, \cdots, \mathrm{svdd}_J\}$ 对不同候选集 $\theta = \{\theta_1, \theta_2, \cdots, \theta_q\}$ 做出评价，可构成评价矩阵 $\boldsymbol{V}$，评价矩阵 $\boldsymbol{V}$ 中的元素 v_{ji} 表示第 j 个超球体对待测样本属于类别 i 的评价，具体计算式为

$$\boldsymbol{V} = \begin{bmatrix} v_{11} & \cdots & v_{1i} & \cdots & v_{1q} \\ \vdots & & \vdots & & \vdots \\ v_{j1} & \cdots & v_{ji} & \cdots & v_{jq} \\ \vdots & & \vdots & & \vdots \\ v_{J1} & \cdots & v_{Ji} & \cdots & v_{Jq} \end{bmatrix} \tag{23}$$

$$\begin{cases} v_{jj} = \gamma_j s_{ij} \\ v_{ji} = (1 - v_{jj}) \dfrac{r_i / d_i}{\sum\limits_{l, l \neq j} r_l / d_l} \end{cases} \tag{24}$$

得到评价矩阵后，可以通过谨慎加权平均证据推理[26](COWA-ER)、模糊谨慎加权证据推理[27](FCOWA-ER)两种方法得到信度函数并做出判决。

COWA-ER 综合最悲观和最乐观两种态度，选出矩阵 $\boldsymbol{V}$ 中每一列的最大值与最小值，得到矩阵

$$\boldsymbol{E} = \begin{bmatrix} \min(v_{\cdot 1}) & \cdots & \min(v_{\cdot i}) & \cdots & \min(v_{\cdot q}) \\ \max(v_{\cdot 1}) & \cdots & \max(v_{\cdot i}) & \cdots & \max(v_{\cdot q}) \end{bmatrix} \tag{25}$$

然后，用 $\boldsymbol{E}$ 的每一个值除以矩阵中的最大值进行归一化，得到矩阵

$$\boldsymbol{E}_{\mathrm{N}} = \begin{bmatrix} a_1 & \cdots & a_i & \cdots & a_q \\ b_1 & \cdots & b_i & \cdots & b_q \end{bmatrix} \tag{26}$$

式中：a_i 为样本属于 θ_i 的置信函数，即 $\mathrm{Bel}(\theta_i) = a_i$；

b_i 为样本属于 θ_i 的似真函数，即 $\mathrm{Pl}(\theta_i)=b_i$。可以通过以下方式得到证据函数，即

$$\begin{cases} m(\theta_i)=a_i \\ m(\bar{\theta}_i)=1-b_i \\ m(\theta_i \bigcup \bar{\theta}_i)=m(\Theta)=b_i-a_i \end{cases} \tag{27}$$

对于一个 q 分类问题，将产生 q 条证据。通过 Dempster 组合规则融合 q 条证据函数 $\boldsymbol{m}_j(j=1,2,\cdots,q)$，即

$$\boldsymbol{m}_{\text{COWA-ER}}=\boldsymbol{m}_1 \oplus \boldsymbol{m}_2 \oplus \cdots \oplus \boldsymbol{m}_{\mathrm{q}} \tag{28}$$

这样可以得到融合后的证据，再将其转化为概率，即可得到最终判决。

使用 COWA-ER 算法可知，计算量会随着样本类别数剧增。FCOWA-ER 算法针对上述问题进行改进，与 COWA-ER 算法相同，FCOWA-ER 算法也是基于 $\boldsymbol{E}$ 矩阵生成证据函数，但是它以 $\boldsymbol{E}$ 的每一行作为样本的一个模糊隶属度函数，然后使用 α-cut 法将其转换为 mass 函数。假设辨识框架为 $\Theta=\{\theta_1,\theta_2,\cdots,\theta_q\}$，模糊隶属度函数为 $\mu(\theta_i)(i=1,2,\cdots,q)$，使用模糊隶属度函数生成一个升序 α-cut$(0<\alpha_1<\alpha_2<\cdots<\alpha_M\leqslant 1)$，则有

$$\begin{cases} m(B_j)=\dfrac{\alpha_j-\alpha_{j-1}}{\alpha_M} \\ B_j=\{A_i\in\Theta \mid \mu(A_i)\geqslant\alpha_i\} \end{cases} \tag{29}$$

式中：$\alpha_i(i=1,2,\cdots,M)$为模糊隶属度序列；$B_j(j=1,2,\cdots,M)$为焦元。

根据 $\boldsymbol{E}_{\mathrm{N}}$ 的第一行可以生成悲观态度下的 mass 函数 $\boldsymbol{m}_{\text{pess}}(\cdot)$；第二行可以生成乐观态度下的函数 $\boldsymbol{m}_{\text{opti}}(\cdot)$；最后对两条证据进行融合得到融合后的 mass，即

$$\boldsymbol{m}_{\text{FCOWA-ER}}=\boldsymbol{m}_{\text{opti}} \oplus \boldsymbol{m}_{\text{pess}} \tag{30}$$

COWA-ER、FCOWA-ER 算法根据评价矩阵生成证据，在实际使用过程中，两者的精度相差不大，然而 COWA-ER 算法的计算量随着类别数的增加而增加。所以，在类别数较少时，使用 COWA-ER、FCOWA-ER 算法均可，当类别数较多时，建议使用 FCOWA-ER 算法。

3 基于证据理论的多分类 SVDD 算例

取 Iris 数据集中的一个第二类样本为待测样本来说明证据 SVDD 算法的计算过程。先计算待测样本到三类超球体的距离，并使用式(22)将其转换为概率；然后计算 3 个超球体的正确率以及紧密程度。计算结果如表 1 所示。

表 1 超球体的输出信息及其可靠程度

Table 1 The output and reliability of each hypersphere

超球体	隶属度	正确率	紧密程度	可靠程度
svdd_1	0.32	1.00	0.81	0.81
svdd_2	0.88	0.95	0.84	0.80
svdd_3	0.63	0.94	0.82	0.77

(1)三焦元证据 SVDD 算法。使用式(21)生成证据

$$m_1(\theta_1)=0.26,\ m_1(\bar{\theta}_1)=0.55,\ m_1(\Theta)=0.19$$
$$m_2(\theta_2)=0.70,\ m_2(\bar{\theta}_2)=0.10,\ m_2(\Theta)=0.20$$
$$m_3(\theta_3)=0.49,\ m_3(\bar{\theta}_3)=0.28,\ m_3(\Theta)=0.23$$

接着，使用 Demspter 组合规则融合 $\boldsymbol{m}_1$、$\boldsymbol{m}_2$、$\boldsymbol{m}_3$ 得到融合后的证据函数

$$m_f(\theta_1)=0.09,\ m_{\mathrm{f}}(\bar{\theta}_2)=0.58,\ m_{\mathrm{f}}(\theta_{1,2})=0.02$$
$$m_{\mathrm{f}}(\theta_3)=0.24,\ m_f(\theta_{1,3})=0.00,\ m_{\mathrm{f}}(\theta_{2,3})=0.05$$
$$m_{\mathrm{f}}(\Theta)=0.02$$

最后，使用 Pignistic 公式将融合后的证据转换为概率做出判决

$$\mathrm{Bet}P(\theta_1)=0.11,\mathrm{Bet}P(\theta_2)=0.62,$$
$$\mathrm{Bet}P(\theta_3)=0.27$$

(2)基于证据多属性决策的 SVDD 算法。先根据表 1 中信息，计算评价矩阵

$$\boldsymbol{V}=\begin{bmatrix} 0.26 & 0.44 & 0.30 \\ 0.10 & 0.70 & 0.20 \\ 0.13 & 0.36 & 0.49 \end{bmatrix}$$

然后，根据评价矩阵 $\boldsymbol{V}$，分别使用 COWA-ER 算法和 FCOWA-ER 算法进行决策融合。

(3)COWA-ER 算法。先根据评价矩阵以悲观和乐观两种方式得到有序加权平均评价向量并构成矩阵

$$\boldsymbol{E}=\begin{bmatrix} \boldsymbol{V}_{\min} \\ \boldsymbol{V}_{\max} \end{bmatrix}=\begin{bmatrix} 0.10 & 0.36 & 0.20 \\ 0.26 & 0.70 & 0.49 \end{bmatrix}$$

通过使用 $\boldsymbol{E}$ 中的最大值对矩阵 $\boldsymbol{E}$ 归一化得到

$$\boldsymbol{E}_{\mathrm{N}}=\begin{bmatrix} 0.14 & 0.51 & 0.29 \\ 0.37 & 1.00 & 0.70 \end{bmatrix}$$

然后，依据 $\boldsymbol{E}_{\mathrm{N}}$ 生成证据函数

$$m_1(\theta_1)=0.14,m_1(\bar{\theta}_1)=0.63,\ m_1(\Theta)=0.23$$
$$m_2(\theta_2)=0.51,m_2(\bar{\theta}_2)=0.00,\ m_2(\Theta)=0.49$$
$$m_3(\theta_3)=0.29,m_3(\bar{\theta}_3)=0.30,\ m_3(\Theta)=0.41$$

接着，使用 Demspter 组合规则融合 $\boldsymbol{m}_1$、$\boldsymbol{m}_2$、$\boldsymbol{m}_3$ 得到融合后的证据函数

$$m_{\mathrm{f}}(\theta_1)=0.06,\ m_{\mathrm{f}}(\theta_2)=0.52,\ m_{\mathrm{f}}(\theta_{1,2})=0.04$$
$$m_{\mathrm{f}}(\theta_3)=0.15,\ m_{\mathrm{f}}(\theta_{1,3})=0.00,$$
$$m_{\mathrm{f}}(\theta_{2,3})=0.16,\ m_{\mathrm{f}}(\Theta)=0.06$$

最后，使用 Pignistic 公式将其转换为概率做出判决

$$\mathrm{Bet}P(\theta_1)=0.11,\ \mathrm{Bet}P(\theta_2)=0.64,$$
$$\mathrm{Bet}P(\theta_3)=0.25$$

(4)FCOWA-ER 算法。基于评价矩阵 $\boldsymbol{V}$ 分别以悲观和乐观方式得到有序加权向量 $\boldsymbol{V}_{\min}$、$\boldsymbol{V}_{\max}$

$$\boldsymbol{V}_{\min}=[0.28\quad 1.00\quad 0.56]$$
$$\boldsymbol{V}_{\max}=[0.37\quad 1.00\quad 0.70]$$

然后，使用 α-cut 生成证据函数

$$m_{\mathrm{pess}}(\theta_2)=0.44,\ m_{\mathrm{pess}}(\theta_2\cup\theta_3)=0.28,$$
$$m_{\mathrm{pess}}(\Theta)=0.28$$
$$m_{\mathrm{opti}}(\theta_2)=0.30,\ m_{\mathrm{opti}}(\theta_2\cup\theta_3)=0.33,$$
$$m_{\mathrm{pess}}(\Theta)=0.37$$

接着，使用 Demspter 组合规则融合 $\boldsymbol{m}_{\mathrm{pess}}$、$\boldsymbol{m}_{\mathrm{opti}}$ 得到融合后的证据函数

$$m_{\mathrm{f}}(\theta_2)=0.61,\ m_{\mathrm{f}}(\theta_{2,3})=0.29,\ m_{\mathrm{f}}(\Theta)=0.10$$

最后，使用 Pignistic 公式将其转换为概率做出判决

$$\mathrm{Bet}P(\theta_1)=0.03,\ \mathrm{Bet}P(\theta_2)=0.79,$$
$$\mathrm{Bet}P(\theta_3)=0.18$$

4　基于证据的 SVDD 算法实验分析

4.1　数据集分割

为了验证本文所提方法的有效性，在人工数据集和多个公共数据集上进行实验。对比算法包括 O-SVDD、Yang-SVDD、P-SVDD 算法。额外对比算法包括逻辑回归(LG)、朴素贝叶斯(NB)。对于每个数据集，每次使用 60%作为训练数据集，20%作为测试数据集，20%作为验证数据集；共进行 10 次实验，取平均结果。

4.2　人工数据集上的结果

使用两个人工数据集验证所提算法的有效性。Toy1 为稀疏程度不同的两类人工数据，每类样本有两个维度，每类包含 100 个样本。第一类样本是圆心为(0，0)、半径为 1 的圆所包含的样本；第二类样本是圆心为(−1，0)、半径为 0.3 的圆所包含的样本。样本通过以下公式生成

$$\begin{cases}x_1^k=r_k\cos(\theta)+a_k\\ x_2^k=r_k\sin(\theta)\end{cases}\tag{31}$$

式中：$r_1=1$，$r_2=0.3$；$a_1=0$，$a_2=1$；θ 为$[0,2\pi]$中的任意取值。生成的样本如图 4 所示，图中蓝色方块代表第一类样本，红色叉号代表第二类样本。分别画出每个超球体的分类边界，蓝色曲线表示第一类超球体的边界，红色曲线表示第二类超球体的边界；重叠区域第二类样本比第一类样本多，svdd2 的精度明显高于 svdd1。svdd1 表示使用第一类样本作为正样本，第二类样本作为负样本得到的超球体；svdd2 反之。并且边界 2 相对边界 1 更紧密。O-SVDD 算法及其拓展算法以同等的信任对待两个超球体的输出信息并不全然合理。

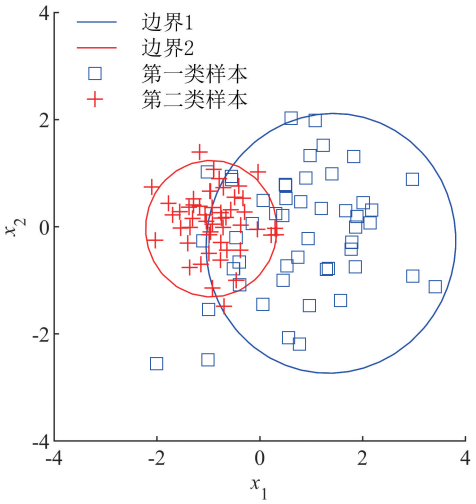

图 4　人工数据集 1

Fig. 4　Toy1 dataset

由图 4 可以看出：由于两类样本的稀疏程度不同，重叠区域的样本更多属于第二类；该区域的样本属于第二类的概率更大。然而传统方法将重叠区域靠右侧的区域判决为第一类，导致分类错误。使用基于证据的方法进行判断，由于红色超球体的可靠程度较高，将会修正原始分类界面使其偏右，这样部分原本被分错的样本会被分对，所以基于证据的方法在一定程度上可以抑制因未考虑超球体之间的差异使部分重叠区域样本被误分的问题。各算法在 Toy1 上的分类正确率如表 2 所示，可知，基于证据的方法能够有效提升原有多分类 SVDD 算法的正确率，并且精度高于朴素贝叶斯、逻辑回归。

表 2　各算法在 Toy1 上的分类正确率

Table 2　The accuracy of each algorithm on Toy1

算法	分类正确率/%
O-SVDD	94.00±0.71
Yang-SVDD	94.80±0.45
P-SVDD	94.00±0.71
TriE-SVDD	97.67±1.32
C-SVDD	97.32±1.57
F-SVDD	96.88±1.30
NB	60.70±1.48
LG	86.15±1.93

Toy2 是沿着两个坐标轴分布的十字数据集，共两类，每类包含 100 个样本。第一类样本围绕 x 轴分布，第一维的变化范围为[−1，1]，第二维的变化范围为[−0.5，0.5]；第二类样本围绕 y 轴分布，第

一维的变化范围为[-0.1,0.1]，第二维的变化范围为[-1,1]。生成的样本如图 5 所示，图中蓝色方块代表第一类样本，红色叉号代表第二类样本。表 3 为各算法的平均正确率，可知在 Toy2 数据集上，基于证据的方法相较传统方法具有优势。

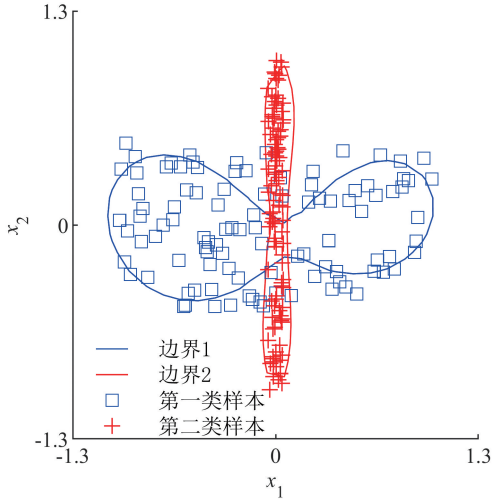

图 5 人工数据集 2

Fig. 5 Toy2 dataset

表 3 各算法在 Toy2 数据集上的分类正确率

Table 3 The accuracy of each algorithm on Toy2

算法	正确率/%
O-SVDD	86.88±1.14
Yang-SVDD	88.43±1.80
P-SVDD	86.56±1.49
TriE-SVDD	93.88±0.96
C-SVDD	94.32±0.94
F-SVDD	95.00±1.79
NB	90.10±2.42
LG	91.30±1.74

4.3 UCI 数据集上的结果

在公共数据集上验证所提出算法的有效性。数据集和实验结果如表 4、表 5 所示，最好的结果已用黑体进行标注。由表 5 可以看出，基于证据的 SVDD 优于传统算法，在 Seed、HR、Vehicle、Sonar 等 4 个数据集上，基于证据的方法相较传统方法有较大的提升。

表 4 数据集信息

Table 4 Information of the datasets

数据集	类别数	属性数
Iris	3	4
Wine	3	13
Seeds	3	7
Hayes-Roth(HR)	3	4
Vehicle	4	18
Heart	2	13
Liver	2	6
Sonar	2	60

这是因为：基于证据的方法通过引入全集，使可靠程度较低的超球体有较小的权重，从而降低该超球体对输出的不利影响；基于证据的方法保留了更多的不确定信息，融合多组证据进行判决相较于传统方法不容易引起误判。在其余数据集上基于证据的方法略有提升。在多数情况下，基于证据的 SVDD 算法能取得较好的效果，TriE-SVDD 算法，C-SVDD 算法和 F-SVDD 算法差异不是很大，三者的平均正确率集中在 83%；传统 SVDD 方法的平均正确率为 81%。与经典多分类算法进行比较。基于证据的多分类 SVDD 算法精度高于逻辑回归和朴素贝叶斯。基于证据的 SVDD 算法保留了待测样本的不确定信息，并且较为充分地利用了每个超球体的输出信息。所以在上述公共数据集上相较于传统方法分类性能得到了提升。

表 5 各算法在 UCI 上的分类正确率

Table 5 The accuracy of each algorithm on UCI

数据集	分类正确率/%							
	O-SVDD	Yang-SVDD	P-SVDD	CSVDD	F-SVDD	TriE-SVDD	NB	LG
Iris	95.20±1.12	95.47±1.41	95.53±1.93	96.33±1.71	96.13±1.17	96.66±1.33	95.87±1.44	96.24±1.31
Wine	94.41±4.54	94.86±5.31	94.24±3.22	95.55±3.99	96.23±3.67	95.66±3.61	94.59±2.26	94.23±2.56
Seeds	92.33±2.88	92.43±1.54	92.95±2.02	94.12±4.95	94.87±1.80	94.00±2.97	90.00±2.88	91.71±2.52
HR	71.85±4.53	72.00±3.28	72.86±2.75	73.85±3.07	74.23±3.94	74.62±3.68	67.54±4.25	65.08±3.08
Vehicle	68.18±2.25	69.01±3.26	69.01±3.26	72.43±1.75	71.94±1.95	71.64±1.94	64.31±3.26	69.01±1.69
Heart	81.11±2.30	81.43±3.21	80.96±4.31	83.15±2.71	83.33±2.65	84.44±3.63	83.70±1.84	82.10±1.44
Liver	64.03±1.92	64.21±3.41	64.52±4.32	68.51±1.85	67.79±2.09	67.73±1.80	60.79±2.62	62.53±1.59
Sonar	78.54±3.47	79.95±3.90	78.63±4.20	82.36±3.54	82.00±4.12	82.55±3.39	68.93±3.68	75.73±3.73
平均值	80.70	81.17	81.09	83.16	83.19	83.16	78.22	79.58

5 结　论

本文提出基于证据理论的多分类 SVDD 算法。在证据框架下，使用超球体的可靠程度作为折扣因子，使用两种方法生成证据。相较于传统方法，基于证据的方法不仅计算样本属于每个超球体的隶属度，还额外考虑样本不属于每个超球体的隶属度以及样本属于全集的隶属度，基于证据的方法更加细致地利用了 SVDD 的输出信息，所以在大多数情况下，本文方法优于传统方法；并且本文方法输出结果为概率，可以方便很多后续处理。本文使用两类证据生成方法生成证据，并将其应用于 SVDD 多分类算法，但证据的生成方法并不唯一。

参考文献：

[1] TU Junling, XU Yong, JIANG Hao, et al. Unknown fault identification method of neutron detector based on SVDD [C]//Proceedings of 2021 Chinese Intelligent Systems Conference. Singapore: Springer Singapore, 2022: 830-838.

[2] QIU Kepeng, SONG Weihong, WANG Peng. Abnormal data detection for industrial processes using adversarial autoencoders support vector data description [J]. Measurement Science and Technology, 2022, 33(5): 055110.

[3] 闫红梅，何明一，梅寒雪. 一种自适应多层结构和空谱联合的高光谱图像异常检测方法 [J]. 西北工业大学学报，2021，39(3)：484-491.
YAN Hongmei, HE Mingyi, MEI Hanxue. Adaptive multi-layer structure with spatial-spectrum combination for hyperspectral image anomaly detection [J]. Journal of Northwestern Polytechnical University, 2021, 39(3): 484-491.

[4] LEE K Y, KIM D W, LEE K H, et al. Density-induced support vector data description [J]. IEEE Transactions on Neural Networks, 2007, 18(1): 284-289.

[5] 刘晟，朱玉全，孙金津. 基于核空间相对密度的 SVDD 多类分类算法 [J]. 计算机应用研究，2010，27(5)：1694-1696.
LIU Sheng, ZHU Yuquan, SUN Jinjin. SVDD multi-class classification algorithm based on relative density in kernel space [J]. Application Research of Computers, 2010, 27(5): 1694-1696.

[6] SUN Qiaoyan, SUN Yumei, LIU Xuejiao, et al. Study on fault diagnosis algorithm in WSN nodes based on RPCA model and SVDD for multi-class classification [J]. Cluster Computing, 2019, 22(3): 6043-6057.

[7] TAO Xinmin, CHEN Wei, LI Xiangke, et al. The ensemble of density-sensitive SVDD classifier based on maximum soft margin for imbalanced datasets [J]. Knowledge-Based Systems, 2021, 219: 106897.

[8] 陶新民，李晨曦，沈微，等. 基于密度敏感最大软间隔 SVDD 不均衡数据分类算法 [J]. 电子学报，2018，46(11)：2725-2732.
TAO Xinmin, LI Chenxi, SHEN Wei, et al. The SVDD classifier for unbalanced data based on density-sensitive and maximum soft margin [J]. Acta Electronica Sinica, 2018, 46(11): 2725-2732.

[9] 杨晨，王婕婷，李飞江，等. 基于概率的支持向量数据描述方法 [J]. 计算机应用，2019，39(11)：3134-3139.
YANG Chen, WANG Jieting, LI Feijiang, et al. Support vector data description method based on probability [J]. Journal of Computer Applications, 2019, 39(11): 3134-3139.

[10] AZAMI M, LARTIZIEN C, CANU S. Converting SVDD scores into probability estimates: application to outlier detection [J]. Neurocomputing, 2017, 268: 64-75.

[11] MU Tingting, NANDI A K. Multiclass classification based on extended support vector data description [J]. IEEE Transactions on Systems, Man, and Cybernetics, 2009, 39(5): 1206-1216.

[12] 林雄，冯海. 基于 SVDD 多类分类新方法的研究 [J]. 信息技术，2008，32(7)：100-103.
LIN Xiong, FENG Hai. Research on new method of multi-class classification based on SVDD [J]. Information Technology, 2008, 32(7): 100-103.

[13] 蔡金燕，杜敏杰. 多分类 SVDD 混叠域识别新方法与故障诊断应用 [J]. 航天控制，2012，30(6)：83-88.
CAI Jinyan, DU Minjie. A novel approach to discriminate the overlap region of multi-class classification SVDD and fault diagnosis application [J]. Aerospace Control, 2012, 30(6): 83-88.

[14] HAO Peiyi, LIN Y H. A new multi-class support vector machine with multi-sphere in the feature space [C]//New Trends in Applied Artificial Intelligence. Berlin: Springer, 2007: 756-765.

[15] DEMPSTER A P. Upper and lower probabilities induced by a multivalued mapping [J]. The Annals of Mathematical Statistics, 1967, 38(2): 325-339.

[16] SHAFER G. A mathematical theory of evidence[M]. Princeton, NJ, USA: Princeton University Press, 1976.

[17] LIU Zhunga, PAN Quan, DEZERT J, et al. Hybrid classification system for uncertain data [J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2017, 47(10): 2783-2790.

[18] HAN Deqiang, LIU Weibing, DEZERT J, et al. A novel approach to pre-extracting support vectors based on the theory of belief functions [J]. Knowledge-Based Systems, 2016, 110: 210-223.

[19] 刘卫兵, 杨艺, 韩德强. 一种基于 FCOWA-ER 的 SVM 多分类方法 [J]. 控制与决策, 2015, 30(10): 1773-1778.
LIU Weibing, YANG Yi, HAN Deqiang. A multi-class SVM based on FCOWA-ER [J]. Control and Decision, 2015, 30(10): 1773-1778.

[20] LIU Zhunga, PAN Quan, DEZERT J. A belief classification rule for imprecise data [J]. Applied Intelligence, 2014, 40(2): 214-228.

[21] LIU Zhunga, DEZERT J, PAN Quan, et al. A new evidential C-means clustering method [C]//2012 15th International Conference on Information Fusion. Piscataway, NJ, USA: IEEE, 2012: 239-246.

[22] MASSON M H, DENŒUX T. ECM: an evidential version of the fuzzy C-means algorithm [J]. Pattern Recognition, 2008, 41(4): 1384-1397.

[23] ZHAO Jianguang, LI Hongbo, ZENG Fanjing, et al. Prognostics of high frequency receiver based on evidential regression [C]//Proceedings of the IEEE 2012 Prognostics and System Health Management Conference. Piscataway, NJ, USA: IEEE, 2012: 1-5.

[24] HE Hongshun, HAN Deqiang, YANG Yi. Cooperative semi-supervised regression algorithm based on belief functions theory [C]//2019 22th International Conference on Information Fusion. Piscataway, NJ, USA: IEEE, 2019: 1-6.

[25] TAX D M J, DUIN R P W. Data domain description using support vectors [C]//ESANN'1999 Proceedings: European Symposium on Artificial Neural Networks. [S. l.: s. n.], 1999: 251-256.

[26] TACNET J M, DEZERT J. Cautious OWA and evidential reasoning for decision making under uncertainty [C]//14th International Conference on Information Fusion. Piscataway, NJ, USA: IEEE, 2011: 1-8.

[27] HAN Deqiang, DEZERT J, TACNET J M, et al. A fuzzy-cautious OWA approach with evidential reasoning [C]//2012 15th International Conference on Information Fusion. Piscataway, NJ, USA: IEEE, 2012: 278-285.

(编辑 赵炜 刘杨)