基于直方图的 mass 函数构造方法*

李文艺[1]，刘 春[2]，李 彪[1]
（1. 宿州学院机械与电子工程学院，安徽 宿州 234000；
2. 河南大学计算机与信息工程学院，河南 开封 475000）

摘 要：针对多特征融合的模式识别问题，给出了一种利用样本特征分布的直方图构造 mass 函数的方法。首先做出样本特征的直方图；在特征直方图的重叠区域，特征的不确定性较大；在特征直方图的非重叠区域，特征的确定性较大。然后，对于一个新的对象，若它的某一特征落入直方图的重叠区，由该特征构造的 mass 函数有较大的不确定性；若该特征落入直方图的非重叠区，则由该特征构成的 mass 函数确定性较大。把不同特征的 mass 函数进行融合得到最终的融合结果。对鸢尾属植物进行分类实验的正确率达到 96.64%，实验结果表明了该方法的有效性。

关键词：证据理论；基本概率赋值函数；直方图
中图分类号：TP391 **文献标志码**：A **文章编号**：0529－6579（2014）06－0155－04

A Mass Function Construction Method Based on Histogram

LI Wenyi[1], *LIU Chun*[2], *LI Biao*[1]
（1. School of Mechanical and Electronic Information Engineering, Suzhou University, Suzhou 234000, China;
2. School of computer and information, Henan University, Kaifeng 475000, China）

Abstract: A method based on the histogram of the sample feature distribution is presented to construct the mass function, for the problem of pattern recognition using the multi-feature fusion. Firstly, the sample feature distribution is established. In the overlapping area of the histogram, the feature is uncertain; while in the no-overlapping area of the histogram the feature is determinate. Then, for a new object, if one of its features falls into the overlapping region of the histogram, the mass function constructed by this feature has a larger uncertainty; if the feature falls into non-overlapping region of the histogram, the mass function constructed by this feature has a greater certainty. The mass functions of different features are fused to get the fusion result. The correct ratio of the iris-plant classify experiment is 96.64%, and the result shows that this method is feasible.

Key words: evidence theory; basic probability assignment function; histogram

证据理论是一种不确定推理方法[1－2]，获取有效的 mass 函数是该理论应用于实际的关键所在。一旦获取了该函数，接下的工作就是利用 Dempster 公式对多个 mass 函数进行合成运算，再根据合成的结果进行判决。从目前来看，已有的获取 mass 函数的方法可以分为两类：一类是利用专家的经验来构造 mass 函数；另一类是根据已知的信息根据一定的条件自动生成函数。前一类方法容易获取 mass 函数，但是由于每个专家的偏好不同、经验不同，给出的 mass 函数有很大的主观性。不同的专家可能会给出相反的证据，此时利用 Dempster 公式进行合成时可能出现错误的结果。而后一类方

* **收稿日期**：2014－01－13
基金项目：国家杰出青年自然科学基金资助项目（61300035）；宿州学院科研启动基金资助项目（2009yss08）
作者简介：李文艺（1980 年生），男；**研究方向**：模式识别、信息融合；E-mail：leets@qq.com

法，比如模糊方法，熵函数方法，粗糙集方法等[3-12]，采用自动生成的方法可以不受个人主观因素的影响，比较客观的获取 mass 函数，在一定程度上解决了 mass 函数的获取。

针对 mass 函数的获取方法，本文给出了一种基于直方图的 mass 函数的构造方法。该方法首先获取样本特征的直方图，再用特征的直方图构造出 mass 函数。其基本思想是不同的样本特征的直方图可能会有重叠部分，则在直方图重叠部分样本提供的信息具有一定的不确定性，重叠的程度大说明不确定性就大，重叠程度小说明确定性就较大，在直方图不重叠部分，样本提供的信息具有较大的确定性，所以可以利用直方图重叠程度来确定 mass 函数的确定程度。该方法的优点是在样本较少，或者较大时都可以得到有效的 mass 函数。通过对鸢尾属植物进行分类实验，显示本文所提出的方法正确分类率达到 96.64%，这说明了本文方法的有效性。

1 证据理论

设非空集合 Θ 是一个完备集合，应包含问题的所有可能，称为 Θ 识别框架。2^Θ 为 Θ 的幂集，函数 m 为 2^Θ 到 [0, 1] 的映射，即 $m:2^\Theta \to [0, 1]$，则映射 m 称为基本概率分配函数，又称 mass 函数。若 $A \in 2^\Theta$，m 应满足以下条件

$$\begin{cases} m(\varnothing) = 0, \\ 0 \leqslant m(A) \leqslant 1, A \subseteq \Theta, \\ \sum\limits_{A \in 2^\Theta} m(A) = 1 \end{cases}$$

其中 $\varnothing$ 为空集，若 $m(A) > 0$，则称 A 为证据的焦元。

$Bel(A) = \sum\limits_{B \subseteq A} m(B)$ $(\forall A \subset \Theta)$，称 $Bel(A)$ 为 2^Θ 的信任函数，表示支持证据 A 的程度。$Pl(A) = \sum\limits_{A \cap B \neq \varnothing} m(B)$ $(\forall A \subset \Theta)$，称 $Pl(A)$ 为证据 A 的似真函数；表示不反对证据 A 的程度。区间 $[Bel(A), Pl(A)]$ 构成证据的不确定区间。

假设 m_1, m_2 为识别框架 Θ 下的两个证据的 mass 函数，可以利用 Dempster 合成公式对证据进行合成，Dempster 公式如下

$$m(C) = \begin{cases} 0 & C = \varnothing \\ \dfrac{1}{1-k} \sum\limits_{A_i \cap B_j = C} m_1(A_i) m_2(B_j), & C \neq \varnothing \end{cases}$$

其中 $C, A_i, B_j \in 2^\Theta$，$k = \sum\limits_{A_i \cap B_j = \varnothing} m_1(A_i) m_2(B_j)$，$k$ 为冲突系数，$\dfrac{1}{(1-k)}$ 称为正则化因子。冲突系数 k 表示证据之间相互冲突的程度，$k \in [0,1]$；k 越大表示冲突越严重，$k = 0$ 时表示证据之间完全没有冲突；$k = 1$ 时表示两个证据完全冲突，此时不能利用公式进行合成。

2 基于直方图的 mass 函数构造方法

本节将详细介绍基于直方图的 mass 函数构造方法的基本思想和基本过程。

2.1 基本思想

假如有 A, B, C 三类样本，每一类样本都有 k 个特征可供利用，分别记为 $x_1, \cdots, x_k$。假设 A, B, C 三类样本的特征 x_1 分布区间分别是 $[c, d]$, $[a, e]$, $[b, f]$（如图 1 所示）。在区间 $[a, b]$ 与 $[e, f]$ 中样本不存在重叠，现有一待识别对象为 s，根据 s 的 k 特征确定 s 的归属。若 $x_1 \in [a,b]$，由 x_1 构造的 mass 函数应该对 $s \in B$ 有很大的支持度；若 $x_1 \in [e,f]$，由 x_1 构造的 mass 函数应该对 $s \in C$ 有很大的支持度。区间 $[b, c]$ 与区间 $[d, e]$ 中 B, C 样本的特征存在重叠，若 $x_1 \in [b,c]$ 或 $x_1 \in [d,e]$，由 x_1 构成的 mass 函数应同时支持 $s \in C$, $s \in B$；此时 mass 函数值对 B, C 的支持程度与 x_1 附近 B, C 两种样本的数量紧密相关。若 x_1 附近 C 类样本数比 B 类样本数多，x_1 形成的 mass 函数对 $s \in C$ 的支持程度大于 $s \in B$ 支持程度，反之亦然。如果 $x_1 \in [c,d]$，mass 函数应该 A、B、C 三类都有所支持，具体对 A、B、C 的支持程度同样取决于 x_1 附近样本数目的三类样本数目。x_1 附近那类样本数量多，特征 x_1 生成的 mass 函数应该对此类有较大的支持程度。为了避免某些特征构成的 mass 函数，“一票否决”现象，设定 mass 函数对框架 Θ 函数值不为 0。

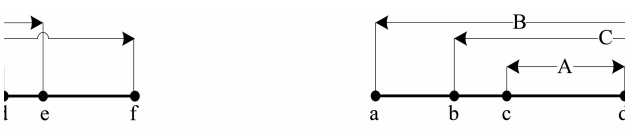

图 1 样本分布区间

Fig. 1 Sample distribution range

2.2 基本过程

假设有 N 个可能的识别结果，辨识框架记为 $\Theta = \{A_1, A_2, A_3, \cdots, A_N\}$。设框架中每个元素的 M 个特征分别记为 $\Theta = \{A_1, A_2, A_3, \cdots, A_N\}$；下面仅

以特征 x_1 为例说明由样本特征构造 mass 函数的具体步骤。

步骤 1　对样本进行筛选，删除样本中的“野点”，记 $x_{1\max} = \max_i\{x_i\}$，$x_{1\min} = \min_i\{x_i\}$，则样本特征 x_1 的取值范围为 $[x_{1\min}, x_{1\max}]$。

步骤 2　把样本特征 x_1 的取值范围 $[x_{1\min}, x_{1\max}]$ 等分成 r 等份，形成 r 个子区间，每个区间的长度为 δ，$\delta = \dfrac{x_{1\max} - x_{1\min}}{r}$。每一个子区间分别记为：$\Delta 1, \Delta 2, \cdots, \Delta r$。

步骤 3　假设样本 A_i 的特征 x_1 落入区间 Δ_j 的个数记为 k_j^i。统计每一类样本的特征 x_1 落入每个子区间中的数目 k_1^i，k_2^i，$\cdots$，k_{k-1}^i，k_k^i（$i = 1, 2, \cdots, M$）做出每一类的样本直方图。

步骤 4　对子区间数进行扩展，$\Delta 1$ 左边扩展一个区间记为 $\Delta 0$，Δr 左边扩展一个区间记为 $\Delta(r+1)$，$\Delta 0$ 与 $\Delta(r+1)$ 中样本特征 x_1 的个数分别记为 k_0^i，k_{r+1}^i（$i = 1, 2, \cdots, N$）。令 $k_0^i = \dfrac{k_1^i}{2}$，$k_{r+1}^i = \dfrac{k_r^i}{2}$。根据 k_0^i，k_1^i，$\cdots$，k_k^i，k_{r+1}^i 做出特征分布的直方图。

步骤 5　特征 x_1 构成的 mass 函数记为 m_1，$m_1(A_i)$ 表示 x_1 支持 A_i 的程度。现有一个未知对象 a，若 a 的特征 $x_1 \in \Delta j$ 此时特征 x_1 构成的 mass 为

$$m_1(A_i) = (1-\alpha)\frac{k_j^i}{\sum_{i=1}^{N} k_j^i}, \quad m(\Theta) = \alpha$$

对余下的 $M-1$ 个特征按照以上步骤可以分别生成 $m_2, m_3, \cdots, m_M$。容易验证对于由 x_1 产生的 mass 函数之和为 1，这完全满足 mass 函数的条件。α 应是一个较小的数值，通常 $\alpha \in (0, 0.3]$。利用 Dempster 合成公式对多个特征进行融合，可以完成多特征融合的分类器设计。

3　仿真实验

鸢尾属植物样本的数据集中共有三类植物分别是 Iris-Setsoa，Iris-Versicolor，Iris-Virgninca；每类样本 50 个，共有 150 个样本。每个样本包含四个特征，分别是含萼片长度，萼片宽度，花瓣长度，花瓣宽度[13]。本文利用该样本数据进行仿真实验来验证所提方法的可行性。

具体的仿真实验过程如下：对样本数据进行预处理，去除“野点”后余下 149 个样本。样本的特征范围（单位 cm）以及每个特征直方图的区间个数如表 1 所示。首先利用本文方法构造出每个特征对应的 mass 函数；然后采用 Dempster 公式融合不同鸢尾属植物的四个特征；再根据融合结果完成对三种鸢尾属植物的分类工作，判决规则采用最大化 mass 函数值的原则。

对样本数据采用“留一法”进行测试，采用不同的方法进行实验，结果如表 2 所示。对待识别对象的萼片长度、萼片宽度、花瓣长度、花瓣宽度分别加上不同方差的干扰信号（干扰信号的平均值等于该类样本特征的平均值），使用本文方法进行分类的结果如表 3 所示。

表 1　样本数据

Table 1　Data of sample

特征	特征范围	子区间数
鄂片长度	[4.3, 9.7]	18
鄂片宽度	[1.9, 4.4]	13
花瓣长度	[1.0, 6.9]	15
花瓣宽度	[0.1, 2.5]	15

表 2　不同方法的实验结果

Table 2　Experiment results of different method

采用方法	正确识别率/%
BP 神经网络	67.32
支持向量机	89.62
聚类分析	82.96
本文方法	96.64

表 3　不同干扰情况下的实验结果

Table 3　Experiment result in different disturbance

干扰信号的方差				正确识别率/%
萼片长度	萼片宽度	花瓣长度	花瓣宽度	
0.09	0.04	0.16	0.025 6	96.64
0.81	0.36	1.44	0.230 0	95.30
3.24	1.44	5.76	0.921 6	85.23

分别采用 BP 神经网络、支持向量机、聚类分析以及本文方法进行了分类实验，结果如表 2 所示。由表 2 可见采用本文方法的正确识别率为 96.64%，明显高于支持向量机、神经网络与聚类分析方法。在被识别对象的特征受到干扰时，实验结果如表 3 所示。由表 3 可看出在干扰较小时，本文方法保持了原有的识别率；在干扰增大时，本文方法仍能获得较为理想的实验效果；在干扰信号较大时识别率为 85.23%，此时识别率仍然优于表 2 中的神经网络与聚类分析方法。

在样本较少时可以把直方图的组距设计的大些，并利用 mass 函数构造过程中的步骤 4 与步骤

5，这样在小样本的情况下同样可以生成可用的mass函数。此时不会出现神经网络中的欠学习的问题；其次使用本文方法时不用纠结于神经网络中神经元个数的选择，同时也避免了神经网络结构的选择。在样本数量很大时直方图的组距可以设计的小一些，此时每一个子区间内样本的频率更接近特征分布密度函数在该区间内的平均值，这会使结果更准确而不会出现神经网络中的过学习问题。利用本文方法进行多特征融合时，由于单个特征的mass函数精度对于最终融合结果影响不明显，所以在待识别对象的特征存在干扰时仍能取得较好的识别率。

4 结 论

针对证据理论的使用中mass函数的构造问题，提出了一种利用直方图思想构造mass函数的方法，该方法可以利用样本的特征构造出需要的mass函数，利用Dempster规则合成多个特征的mass函数值，即可实现多特征融合的模式识别方法。把该方法用于鸢尾属植物的分类实验中，在没有干扰的情况下，分类正确率达到96.64%。在被识别样本特征受到干扰时，使用本文方法仍然可以获得较为理想的识别效果。文中的实验说明了该方法可以很好的构造出所需要的mass函数。由于在实际的模式识别中通常都需要利用对象的多个特征进行识别，只要有少量的样本就可以使用本文方法构造出样本中每类特征对应的mass函数。在进行多传感器融合时，若有多个传感器的输出数据作为样本，利用本文方法可以构造出每个传感器的mass函数，可实现多传感器的信息融合。

参考文献：

[1] DEMPSTER A P. Upper and lower probabilities induced by a multi-valued mapping [J]. Annuals of Mathematics Statistics, 1967, 38(4): 325 – 339.

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

[3] 康兵义,李娅,邓勇,等. 基于区间数的基本概率指派生成方法及应用[J]. 电子学报, 2012, 40(6):1092 – 1096.

[4] 王俊林,张剑云. 基于统计证据的Mass函数和D-S证据理论的多传感器目标识别[J]. 传感技术学报, 2006, 19(3): 862 – 864.

[5] 江四厚,王汉功,阳能军. 基于熵的Mass函数算法及在液压泵故障诊断中的应用[J]. 机床与液压, 2007, 35(12):185 – 187.

[6] DENG Y, SHI W K, ZHU Z F, et al. Combining belief functions based on distance of evidence [J]. Decision Support Systems, 2004, 38(3): 489 – 493.

[7] DENG Y, JIANG W, XU X, et al. Determining BPA under uncertainty environments and its application in data fusion [J]. Journal of Electronics (China), 2009, 26(1): 13 – 17.

[8] 肖建于,童敏明,朱昌杰,等. 基于广义三角模糊数的基本概率赋值构造方法[J]. 仪器仪表学报, 2012, 33(2): 429 – 434.

[9] 孔金生,李文艺. 基于模糊集合的mass函数构造方法[J]. 计算机工程与应用, 2008, 44(20): 152 – 154.

[10] 韩峰,杨万海,袁晓光. 基于模糊集合的证据理论信息融合方法[J]. 控制与决策, 2010, 25(3):449 – 452.

[11] 蒋雯,张安,杨奇. 一种基本概率指派的模糊生成及其在数据融合中的应用[J]. 传感技术学报, 2008, 21(10): 1717 – 1720.

[12] 刘雷健,扬静宇. 基于融合信息的物体识别[J]. 模式识别与人工智能, 1993, 6(1): 27 – 33.

[13] IrisData Set. Famous database for pattern recognition from Fisher [OL]. [2011 – 3 – 20] http://archive.ics.uci.edu/ml/datasets/Iris