DOI:10.13878/j.cnki.jnuist.2019.03.012

李锋[1]　李寿梅[1]　Thierry Denoeux[1,2]

基于证据理论的聚类集成方法

摘要

单个聚类方法得到的结果会存在不稳定性等问题,为了克服这些问题,本文在证据理论(又称为信任函数理论)的基础上提出了一种新的聚类集成方法.多数情况下,聚类集成方法主要包含2个关键步骤:得到一组基划分,以及结合基划分得到最终聚类结果,本文的方法重点考虑第2步.在第1步得到基划分之后,将其转换成一种中间表示,可以称这种中间表示为关系表示.在证据理论中,我们认为得到的关系表示是不可靠的,可以用折扣过程对关系表示进行预处理,然后就可以用不同的结合法则融合关系表示.从融合后的关系表示中提取信任矩阵或似然矩阵,将其视为样本间的互相关矩阵.为了能够充分利用样本间的传递性,将得到的互相关矩阵视为一个模糊关系,对其做传递闭包处理,从而得到一个模糊等价关系.将模糊的等价关系视为新的相似性数据,用能够处理相似性数据的聚类方法得到最终的结果.通过实验,表明了该聚类集成方法的稳定性和有效性.

关键词

证据理论;聚类集成;关系表示;互相关矩阵;传递闭包

中图分类号 TP181;O213.9

文献标志码 A

收稿日期 2019-05-09

资助项目 国家自然科学基金(11571024);2018年北京工业大学研究生外培计划

作者简介

李锋,男,博士生,研究方向为证据理论下的聚类分析.lixxfeng@ emails.bjut.edu.cn

李寿梅(通信作者),女,理学博士,教授,主要研究方向为非线性随机理论及智能计算.lisma@ bjut.edu.cn

1　北京工业大学　应用数理学院,北京,100124

2　索邦大学联盟　贡比涅技术大学/国家科学研究中心,Heudiasyc(UMR 7253),法国

0　引言

聚类分析是机器学习和模式识别领域一个重要的研究课题.基于样本间的相似性,聚类分析将样本分成不同的组,使得在同一组中的样本间的相似性高,而不同组之间的样本相似度低.目前,聚类分析在数据挖掘、图像分割等领域有广泛的应用.聚类分析根据得到结果的不同可以分为硬聚类和软聚类.其中软聚类包括模糊聚类方法和证据聚类方法.

证据聚类方法最近受到越来越多的学者的关注.该方法是在证据理论的基础上提出的聚类方法.证据理论被认为是更具一般性的理论框架,概率论和模糊集合理论都可以看作是证据理论的一个特例.证据理论能够很好地处理带有不确定性的数据,也可以用于不确定性推断.在证据聚类方法[1-4]中,一般会假设样本和类之间的隶属关系并非是确定的,且可以通过一个质量函数表示.如Denoeux等[1]提出一种证据聚类方法,简称EVCLUS.该方法在给定样本间的相异性矩阵的基础上构造一个目标函数,通过梯度下降法最小化该目标函数,并给每一个样本赋予一个质量函数.这是证据理论在聚类中的一个应用,而本文就是将证据理论进一步应用到聚类集成的问题中.

聚类集成方法的提出,是为了进一步提高聚类结果的精确度和稳健性.因为目前虽然已经提出了许多种聚类方法,但是没有哪一种方法适用于所有的数据类型."集成"说的是,通过结合多个不同的聚类结果,从而得到更优的结果.目前有很多集成聚类方法,但是基于证据理论的集成聚类方法[5-8]却只有很少人涉及.在证据理论中存在多种证据结合方法,因此该方法非常适用于处理聚类集成问题.如Masson等[5]已经提出的基于证据理论的集成聚类方法,该方法在得到基划分之后,将质量函数定义在划分区间,并通过证据结合法则得到合成的质量函数,从中提取得到样本间一个新的相似性矩阵,再通过层次聚类法得到最终的结果.本文是对该方法的进一步的拓展.

不同于Masson等[5]的方法,在本文中得到基划分后,利用证据理论将其转化到另一个识别框架上,并将质量函数定义在该框架上.这样做使得该方法更容易理解和解释,本文将这一步得到的结果称为关系表示,这是互相关矩阵在证据理论下的进一步拓展.在证据理论中,如果一个证据源是不可靠的,可以通过折扣过程处理,得到新的质量函数.这里,因为在得到基划分的过程中,并不知道数据真实的类

的个数,所以得到的关系表示是不可靠的.为此,本文通过证据折扣过程处理关系表示后,用证据结合法则得到融合后的关系表示.本文从融合后的关系表示中提取信任矩阵或者似然矩阵,将其视为新的互相关矩阵.同时,为了充分利用样本间的传递性,将新得到的互相关矩阵视为模糊关系,并用传递闭包方法得到一个模糊等价关系,将其视为新的相似性数据得到最终的聚类结果.

本文在接下来的第 1 节介绍相关的理论知识和聚类集成方法.第 2 节中给出基于证据理论的聚类集成方法具体过程.最后一节通过实验表明该方法对参数选择的稳定性,以及方法的优势.

1 相关理论介绍

本节中,首先介绍证据理论[9]的基本定义,以及证据结合法则相关的概念,其次简单介绍聚类集成方法.

1.1 证据理论

假设 $\Omega=\{\omega_1,\cdots,\omega_c\}$ 是一个有限集合,在证据理论中称 Ω 为识别框架.如果函数 $m:2^{\Omega}\to[0,1]$,且满足 $\sum_{A\subseteq\Omega}m(A)=1$,则称 m 为该识别框架的质量函数.其中 2^{Ω} 表示 Ω 的幂集.在证据理论中,$m(A)$ 表示真值属于子集 A 的程度,如果 $m(A)>0$,则称集合 A 为焦元.特别的,$m(\varnothing)$ 表示真值不属于识别框架 Ω 的信任度;$m(\Omega)$ 表示只知道真值属于 Ω 的信任度,但不能具体确定是其中的哪个元素.如果一个质量函数的所有焦元都是单点集,则称其为贝叶斯质量函数.此时该质量函数等价于一个概率分布,因此证据理论可以视为广义的概率分布.

给定质量函数 m,可以得到与之对应的信任函数 b 和似然函数 p(plausibity),定义分别如下:

$$b(A)=\sum_{B\subseteq A}m(B),$$
$$p(A)=\sum_{B\cap A\neq\varnothing}m(B). \tag{1}$$

质量函数、信任函数和似然函数三者之间存在一一对应的关系.

给定 2 个质量函数 m_1 和 m_2,Dempster-Shafer 结合法则(D-S 结合法则)定义如下:

$$m_1\oplus m_2(A)=\sum_{B\cap C=A}m_1(B)\,m_2(C), \tag{2}$$

其中 $\kappa=m_1\oplus m_2(\varnothing)=\sum_{B\cap C=\varnothing}m_1(B)\,m_2(C)$ 称之为 2 个质量函数的冲突度.该结合法则满足交换律和结合律.

考虑到信息源的可靠性,有时需要对证据进行折扣.给定质量函数 m 和折扣因子 $\alpha\in[0,1]$,证据折扣定义如下:

$$m^{\alpha}(A)=\alpha\times m(A),\ \forall A\subset\Omega,$$
$$m^{\alpha}(\Omega)=\alpha\times m(\Omega)+1-\alpha. \tag{3}$$

折扣因子越大,表明原始证据越可靠.

证据理论中的另一个重要概念是细化,通过细化可以将信任程度从识别框架转移到另一识别框架.更确切地说,对于 2 个有限集合 Ω 和 Θ,如果存在映射 $f:2^{\Theta}\to2^{\Omega}$,且对于所有的 $\theta\in\Theta$,集合 $f(\{\theta\})$ 构成 Ω 一个不相交划分,而且对于 $\forall A\subset\Theta$,$f(A)=\cup_{\theta\in A}f(\{\theta\})$,则称 Ω 是 Θ 的一个细化,后者是前者的一个粗化.通过映射 f,可以将定义在 Θ 上的信任函数转移到 Ω 上,即将 $A\subset\Theta$ 上的信任度转移到 $f(A)\subset\Omega$ 上.同理,可以通过定义在 Ω 上的信任函数得到一个定义在 Θ 上的信任函数,即将 $B\subset\Omega$ 上的信任度转移到集合 $\{\theta\in\Theta\mid f(\theta)\cap B\neq\varnothing\}$.

1.2 聚类集成方法

在这一部分中,将简要介绍聚类集成方法[10-12].

对于一个数据集,不同的聚类算法会得到不同的结果,即使是同一算法在不同初始化条件下也可能得到不一样的解.没有一个聚类算法是“最好”的[13].目前,将多个聚类结果整合成一个的集成方法,能够得到具有更好的准确性和稳健性的结果.

一般来说,聚类集成方法包含 2 个关键步骤:1)得到一组不同的聚类结果(称之为基划分);2)结合多个聚类结果得到最终聚类结果.在第 1 步中,可以通过不同的方式得到想要的结果.比如可以用不同的聚类算法得到一组聚类结果,也可以对一个聚类算法,但是设置不同的初始化参数得到.集成方法最关键的步骤是第 2 步中方法的设计,不同的方法就会得到不同的集成方法.例如图划分方法[14]、投票法[15-16]、证据累积方法[17-18].

本文方法是证据累积方法的进一步拓展,因此简单介绍该方法.该方法在第 1 步得到不同的聚类结果之后,证据积累方法将其转化成一个互相关联矩阵.该矩阵是一个 $n\times n$ 的矩阵,n 为样本个数.其中的每个元素 (i,j) 表示在所有基划分中样本 o_i 和样本 o_j 属于同一类的比率,即认为 2 个样本属于同一类的基划分个数除以总划分的个数.在第 2 步将其视为一个新的聚类问题,即将互相关矩阵作为相似性数据,用能够处理相似性数据的聚类方法得到

集成的结果.该集成方法用互相关联矩阵的思想,避免了考虑集成中标签对应问题.对于证据积累方法来说,不同的基划分中类的个数可以不同.然而,该方法仅考虑了基划分中2个样本是否属于同一类这一信息,舍弃了其他的信息.目前,一些研究者已经提出利用基划分中的其他信息来构建一个相似性度量,使其能够更好地表示样本间的真实关系[19-22].

在本文中的每个基划分中,用质量函数表示每个样本和类之间的隶属关系,这种表示方法比其他方法能够更好地表现类和样本之间的关系.在此基础上,通过证据理论中粗化的概念将信息从一个识别框架转化到另一个识别框架上,即将基划分中类和样本间的关系转化成样本间的关系,这就是关系表示的由来.为了能更好地利用这种信息,使用关系表示来衡量对象之间的"相似性".该方法可以视为证据积累方法在证据理论下的拓展.

2 集成方法

在本节中,将详细阐述基于证据理论的聚类集成方法.

2.1 关系表示

假设有 n 个样本的集合为 $O=\{o_1,o_2,\cdots,o_n\}$,样本类的集合为 $\Omega=\{\omega_1,\cdots,\omega_c\}$,每个样本 i 和类之间的关系用质量函数 m_i 表示,n 个样本的质量函数就构成一个证据划分,该划分可以写成一个 $n\times f$ 的矩阵,其中 f 表示考虑的焦元的个数.

对于2个样本 i 和 j,其相对应的质量函数为 m_i 和 m_j.在识别框架 $\Omega\times\Omega$ 上定义一个粗化的识别框架 $\Theta=\{s,\neg s\}$,其中 s 表示2个样本属于同一类,$\neg s$ 表示2个样本属于不同的类.定义映射函数 $f(s)=S$ 且 $f(\neg s)=\overline{S}$,这里 $S=\{(\omega_k,\omega_k),k=1,\cdots,c\}$,$\overline{S}$ 表示它的补集.这样利用D-S结合法则,可以定义 Θ 上的质量函数 m_{ij} 为

$$\begin{aligned}
&m_{ij}(\varnothing)=m_i(\varnothing)+m_j(\varnothing)-m_i(\varnothing)\,m_j(\varnothing),\\
&m_{ij}(\{s\})=\sum_{k=1}^{c}m_i(\omega_k)m_j(\omega_k),\\
&m_{ij}(\{\neg s\})=\sum_{A\cap B=\varnothing}m_i(A)\,m_j(B)-m_{ij}(\varnothing),\\
&m_{ij}(\Theta)=\sum_{A\cap B\neq\varnothing}m_i(A)\,m_j(B)-m_{ij}(\{s\}).
\end{aligned}\tag{4}$$

在式(4)的基础上,定义在所有样本对上的质量函数就构成了 $\mathcal{R}=(m_{ij})_{1\leqslant i\leqslant j\leqslant n}$,本文称其为关系表示(relational representation).

假设在第1步中得到了 N 个基划分,将其转换成关系表示形式 $\mathcal{R}^i,i=1,\cdots,N$.因为在类个数未知的情况下,在基划分中聚类的个数也不一定相同.考虑到这些情况,由基划分得到的关系表示会存在一些不可靠的信息源.因此,用式(3)对每个关系表示做证据折扣处理,处理后的关系表示用 $\mathcal{R}_d^i$ 表示.采用Masson等[5]的算法4确定折扣因子,折扣因子的确定并非本文主要讨论的问题,故这里不做详细介绍.

2.2 证据结合

在证据结合中,可以用不同的结合法则.例如,可以用式(2)的D-S结合法则,结合后的关系表示用 $\mathcal{R}^*$ 表示.即

$$m_{ij}^*=m_{ij}^1\otimes\cdots\otimes m_{ij}^N.\tag{5}$$

这里以结合2个关系表示为例,假设目前有 $\mathcal{R}^k=(m_{ij}^k)_{1\leqslant i\leqslant j\leqslant n},k=1,2$,则

$$\begin{aligned}
m_{ij}^1\otimes m_{ij}^2(\{s\})&=m_{ij}^1(\{s\})\,m_{ij}^2(\{s\})+\\
&\quad m_{ij}^1(\{s\})\,m_{ij}^2(\Theta)+m_{ij}^1(\Theta)\,m_{ij}^2(\{s\}),\\
m_{ij}^1\otimes m_{ij}^2(\{\neg s\})&=m_{ij}^1(\{\neg s\})\,m_{ij}^2(\{\neg s\})+\\
&\quad m_{ij}^1(\{\neg s\})\,m_{ij}^2(\Theta)+m_{ij}^1(\Theta)\,m_{ij}^2(\{\neg s\}),\\
m_{ij}^1\otimes m_{ij}^2(\Theta)&=m_{ij}^1(\Theta)\,m_{ij}^2(\Theta).
\end{aligned}$$

此外,$m_{ij}^1\otimes m_{ij}^2(\varnothing)=1-m_{ij}^1\otimes m_{ij}^2(\{s\})-m_{ij}^1\otimes m_{ij}^2(\{\neg s\})-m_{ij}^1\otimes m_{ij}^2(\Theta)$.由于D-S结合法则满足交换律和结合律,结合多个关系表示可以通过依次结合2个关系来实现.

也可以直接用平均方法,得到结合后的关系表示 $\mathcal{R}^*$ 如下:

$$\begin{aligned}
m_{ij}^*(\varnothing)&=\frac{1}{N}\sum_{k=1}^{N}m_{ij}^k(\varnothing),\\
m_{ij}^*(\{s\})&=\frac{1}{N}\sum_{k=1}^{N}m_{ij}^k(\{s\}),\\
m_{ij}^*(\{\neg s\})&=\frac{1}{N}\sum_{k=1}^{N}m_{ij}^k(\{\neg s\}).
\end{aligned}\tag{6}$$

由质量函数的定义,

$$m_{ij}^*(\Theta)=1-m_{ij}^*(\varnothing)-m_{ij}^*(\{s\})-m_{ij}^*(\{\neg s\}).$$

2.3 传递闭包

$\mathcal{R}^*$ 中每个样本对 (i,j) 的质量函数为

$$(m_{ij}^*(\varnothing),m_{ij}^*(\{s\}),m_{ij}^*(\{\neg s\}),m_{ij}^*(\Theta)).$$

可以通过定义得到信任函数 b 或者似然函数 p,分别为

$$\begin{aligned}
b_{ij}(\{s\})&=m_{ij}^*(\varnothing)+m_{ij}^*(\{s\}),\\
p_{ij}(\{s\})&=m_{ij}^*(\{s\})+m_{ij}^*(\Theta).
\end{aligned}\tag{7}$$

最后得到的 $(b_{ij})_{1\leqslant i,j\leqslant n}$ 和 $(p_{ij})_{1\leqslant i,j\leqslant n}$ 可以视为样本

间的互相关矩阵.

为了充分利用样本间的信息,这里引入传递闭包的概念.该概念源自等价关系,其定义为:设 R 为非空集合 X 上的二元关系,若 R 满足自反性、对称性和传递性,则 R 为 X 上的等价关系.等价关系和划分之间有一一对应的关系,如果 2 个样本是等价的,那么 2 个样本属于同一类.根据传递性的定义,如果样本 o_i 和 o_j 属于同一类,且样本 o_j 和 o_k 属于同一类,则样本 o_i 和 o_k 也属于同一类.

等价关系在模糊理论下有了进一步推广,简单来说,如果定义在集合 X 上的模糊关系 R 满足:

1)自反性:$R(x,x)=1, x\in X$;

2)对称性:$R(x,y)=R(y,x), x,y\in X$;

3)max-$*$ 传递性:$\max\limits_{y}(R(x,y)*R(y,z))\leqslant R(x,z), x,y,z\in X$;

则称 R 为模糊等价关系[23],其中 $*$ 表示 t 模(t-norm).本文考虑了 3 种常见的 t 模,包括最小(min)、乘积(prod)和 Lukasiewicz(Luk) t 模,定义分别如下:

$$\min(x,y)=\min(x,y),$$

$$\mathrm{prod}(x,y)=xy,$$

$$\mathrm{Luk}(x,y)=\max(0,x+y-1).$$

一个满足自反性和对称性的模糊关系可以通过传递闭包得到一个满足传递性的模糊等价关系.在给出传递闭包定义之前,首先需要引出合成运算的定义.假设 R 为定义在 X 上的模糊关系,通过 max-$*$ 合成得到的结果仍是一个模糊关系,且定义如下:

$$R\circ R(x,y)=\max_{z\in X}(R(x,z)*R(z,y)),\ \forall x,y\in X,$$

从上式可知合成运算满足结合律,因此可以定义一个模糊关系的幂为

$$R^1=R,$$

$$R^{n+1}=R\circ R^n,\ \forall n\in N.$$

模糊关系 R 的 max-$*$ 传递闭包可以通过下式得到:

$$\mathrm{t}(R)=\bigcup_{k=1}^{n}R^k. \tag{8}$$

对上一步得到的互相关矩阵,视其为模糊的关系矩阵,得到矩阵的传递闭包 $\mathrm{t}(b_{ij})_{1\leqslant i,j\leqslant n}$ 或 $\mathrm{t}(p_{ij})_{1\leqslant i,j\leqslant n}$.将新的互相关矩阵视为新的相似性数据,用能够处理相似性数据的聚类方法,如层次聚类法,得到最终的结果.

本文的聚类集成方法是基于证据理论提出的,由于证据理论具有一般性,概率理论和模糊理论都可以视为证据理论的特例.当所有的折扣因子都设为 0,且通过平均方法结合所有关系表示,最终得到的 $(b_{ij})_{1\leqslant i,j\leqslant n}$ 就是一般的证据累积方法中的互相关矩阵.因此一般的证据累积方法可以视为本文提出方法的一个特例.

综上所述,聚类集成算法具体流程如下:

输入:通过聚类方法得到的 N 个基划分;

输出:聚类集成的结果.

步骤 1:用式(4)将 N 个基划分转化成关系表示 $\mathcal{R}^i, i=1,\cdots,N$;

步骤 2:利用 Masson 等[5]的算法得到折扣因子,用式(3)对所有关系表示做折扣处理得到 $\mathcal{R}_d^i, i=1,\cdots,N$,通过结合法则得到合成的 $\mathcal{R}^*$;

步骤 3:通过式(7)从融合后的关系表示 $\mathcal{R}^*$ 中提取得到互相关矩阵 $(b_{ij})_{1\leqslant i,j\leqslant n}$ 或 $(p_{ij})_{1\leqslant i,j\leqslant n}$;

步骤 4:将得到的互相关矩阵视为一个模糊关系,通过式(8)从而得到它的模糊等价关系,将其作为新的互相关矩阵;

步骤 5:选取能够处理相异性数据的聚类算法,对新的互相关矩阵进行再聚类,得到最终的集成结果.

3 实验分析

在这一节中,将证据理论下的聚类集成方法应用到模拟和真实数据集,证明本文所提方法的有效性.

通过调整的兰德指数[24](ARI)评价聚类结果:

$$I_{\mathrm{AR}}=\frac{2(ab-cd)}{(a+d)(d+b)+(a+c)(c+b)},$$

记 p_1 和 p_2 为 2 个基划分,上式中 a 表示在 p_1 中属于同一类在 p_2 中也属于同一类的样本对个数;b 表示在 p_1 中属于不同类在 p_2 中也属于不同类的样本对个数;c 表示在 p_1 中属于同一类而在 p_2 中属于不同类的样本对个数;d 表示在 p_1 中属于不同类在 p_2 中属于同一类的样本对个数.

3.1 数据集

表 1 给出了本文中将要考虑的数据集.其中的 Half-rings 数据(图 1)和 Moon(图 2)都是模拟数据集,数据中类集合之间是非线性可分的;Two-spirals 数据是另一种复杂数据的典型代表(图 3).8D5K 数据[14]是 8 维的模拟数据集,其在 2 维主成分上的投影如图 4 所示,从中可以看出数据是线性可分的.其他 3 个数据集 Seeds、Iris 和 Wine 都是 UCI

(University of California Irvine)数据集[25].在实验中我们对 Seeds 和 Wine 数据进行标准化预处理.

表 1　实验数据集

Table 1　Datasets

数据集	样本数	维数	类个数
Half-rings	200	2	2
Moon	400	2	2
Two-spirals	200	2	2
8D5K	1 000	8	5
Iris	150	4	3
Wine	178	13	3
Seeds	210	7	3

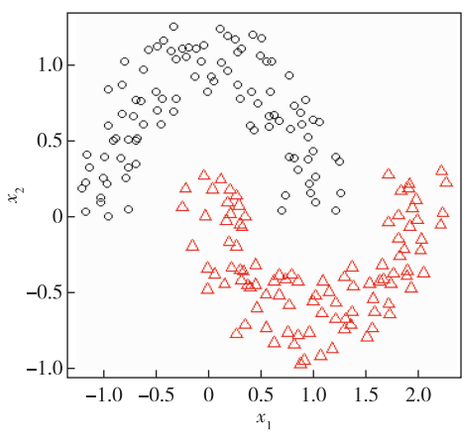

图 1　Half-rings 数据

Fig. 1　Half-rings dataset

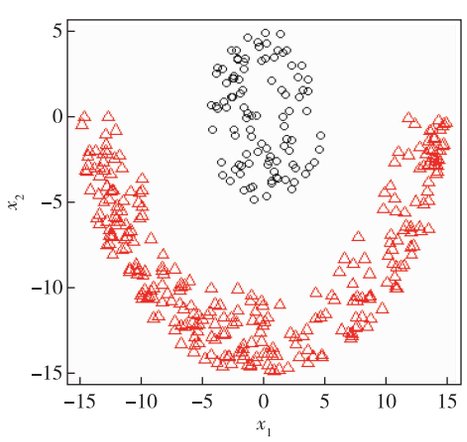

图 2　Moon 数据

Fig. 2　Moon dataset

3.2　模糊系数的影响

在实验中,通过模糊 c-means 方法[26]得到的硬划分作为基划分.模糊 c-means 方法的主要参数是模糊系数 m ,模糊系数越大,得到结果整体的模糊度越高.这一节主要研究模糊系数对于集成结果的影响,主要考虑的是 Half-rings 数据集.

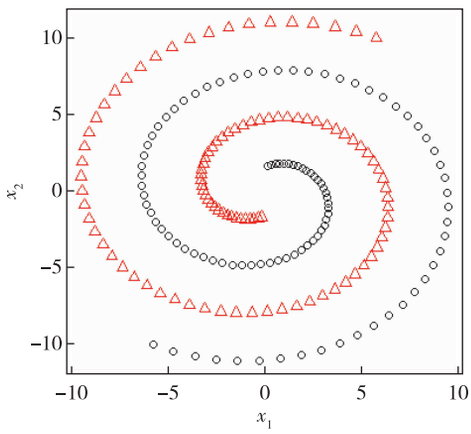

图 3　Two-spirals 数据

Fig. 3　Two-spirals dataset

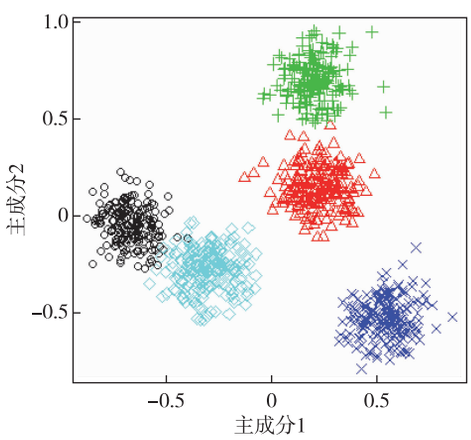

图 4　8D5K 数据

Fig. 4　8D5K dataset

本文方法不会涉及标签对应的问题,因此在集成的第 1 步时,基划分的类个数可以不同.这里考虑从 10 到 20 中随机选取类个数,最后得到 20 个基划分.在第 2 步中,用基于离差平方和的层次聚类方法得到最终的聚类结果,在这一步中考虑类个数已知.

考虑 5 个模糊系数(1.1、1.5、2、2.5、3)对结果的影响.对于每个系数,本文研究了平均结合法则和 D-S 结合法则的结果.在实验中考虑了将信任矩阵作为互相关矩阵.对于每种结合法则得到的互相关矩阵,又考虑了 4 种情况:不考虑传递性时原始的互相关矩阵,以及分别用 min、prod 和 Luk 3 种 t 模传递闭包后得到新的互相关矩阵.每次实验重复 10 次,最终结果分别如表 2 和表 3 所示,在表中给出的是 10 次结果的平均值,括号中的值为 10 次实验的标

准差.

从表 2 可以看出,当使用平均结合法则,不考虑传递性时,得到的结果精确度不高且波动性很大.经过 min t 模处理后,效果变好,主要表现是聚类结果的精确度变高,方差变小.ARI 值为 1 表明,经过传递闭包处理后的结果能够完全识别数据的非线性结构.由 prod t 模得到的结果,优于原始互相关矩阵的结果.经过 Luk t 模得到的结果同原始结果一致.在平均法则下,min t 模得到的结果优于 prod t 模,也优于 Luk t 模得到的结果.

从表 3 可以看出,在使用 D-S 结合法则时,当模糊系数很低时,如 $m = 1.1$,不考虑传递性得到的结果很好,但当模糊系数变大时,得到的结果的精确度和稳定性都变差.但是经过 3 种 t 模处理后的结果,精确度和稳定性都达到最好.同表 2 中平均法则得到的结果相比,D-S 法则经过传递闭包处理后的结果,在稳定性和精确度上表现更好,整体上优于平均法则得到的结果.

总而言之,本文所提的方法对参数选取具有稳健性,而原有方法依赖模糊系数的选取,这是本文方法的优势之一.在使用模糊 c-means 方法时,人们普遍认为模糊系数为 2 时,得到的结果最好.在接下来的实验中,将只考虑一个模糊系数下的结果.

表 2 Half-rings 数据平均结合法则的结果

Table 2 Results from average combination rule

模糊度	原始	min	prod	Luk
1.1	0.23(0.27)	1(0)	0.61(0.42)	0.23(0.27)
1.5	0.32(0.29)	1(0)	1(0)	0.32(0.29)
2.0	0.74(0.36)	1(0)	0.92(0.24)	0.74(0.36)
2.5	0.53(0.41)	1(0)	1(0)	0.53(0.41)
3.0	0.29(0.43)	1(0)	0.87(0.17)	0.29(0.43)

注:表中数值为 10 次的平均值,括号中为 10 次的标准差.

表 3 Half-rings 数据 D-S 结合法则的结果

Table 3 Results from D-S combination rule

模糊度	原始	min	prod	Luk
1.1	1(0)	1(0)	1(0)	1(0)
1.5	0.85(0.31)	1(0)	1(0)	1(0)
2.0	0.83(0.35)	1(0)	1(0)	1(0)
2.5	0.33(0.29)	1(0)	1(0)	1(0)
3.0	0.77(0.37)	1(0)	1(0)	1(0)

注:表中数值为 10 次的平均值,括号中为 10 次的标准差.

3.3 数据集的结果

接下来,考虑证据理论下的聚类集成方法在其他数据集上的表现.这里的参数设置与 3.2 节中的参数选择基本一致,但此时不考虑模糊系数的影响:即对于这些数据集只考虑 $m = 2$ 时的结果.此外,对于 Two-spirals 数据集,基划分中类个数从 30 到 40 之间随机选取,对于其他数据集,类个数从 6 到 10 中随机选取.这里仍考虑不同结合法则下的结果,其中由平均法则结合得到的结果如表 4 所示,表 5 中给出的是由 D-S 结合法则得到的结果.

对于模拟数据集,Moon 和 Two-spirals 数据集上的结果相似,即在平均结合法则下由 min t 模得到的互相关矩阵效果最好,prod t 模的结果优于原始结果,Luk t 模的结果同原始结果一致.在 D-S 结合法则下得到结果的稳定性和精确度都优于平均结合法则下的结果,且 3 种 t 模下的结果都能很好地识别数据的非线性结构.对于 8D5K 数据集,在不考虑传递性以及考虑 3 种 t 模下的传递性得到的结果一致,D-S 结合法则得到的结果同平均法则得到的结果一致,且每种情况都能完全识别数据的内部结构.这是因为该数据的数据结构比较简单,其在 2 个主成分上的投影是线性可分的(从图 4 中可以看出),此时不经过传递性处理时得到的结果就已经能够识别数据结构.

表 4 平均结合法则的结果

Table 4 Results from average combination rule

数据集	原始	min	prod	Luk
Moon	0.22(0.43)	1(0)	0.51(0.43)	0.22(0.43)
Two-spirals	0.01(0)	0.98(0.04)	0.02(0.02)	0.01(0)
8D5K	1(0)	1(0)	1(0)	1(0)
Iris	0.58(0.06)	0.83(0.22)	0.76(0.01)	0.58(0.06)
Wine	0.90(0.01)	0.90(0.01)	0.90(0.01)	0.90(0.01)
Seeds	0.56(0.18)	0.35(0.12)	0.75(0.06)	0.56(0.18)

注:表中数值为 10 次的平均值,括号中为 10 次的标准差.

表 5 D-S 结合法则的结果

Table 5 Results from D-S combination rule

数据集	原始	min	prod	Luk
Moon	0.82(0.39)	1(0)	1(0)	1(0)
Two-spirals	0.09(0.02)	1(0)	1(0)	1(0)
8D5K	1(0)	1(0)	1(0)	1(0)
Iris	0.54(0.06)	0.92(0)	0.92(0)	0.92(0)
Wine	0.90(0.01)	0.89(0.01)	0.90(0.01)	0.90(0.01)
Seeds	0.53(0.18)	0.34(0.16)	0.34(0.16)	0.34(0.16)

注:表中数值为 10 次的平均值,括号中为 10 次的标准差.

对于真实数据集 Iris,在平均结合法则下,经过 min 和 prod t 模传递闭包处理后的结果,优于原始以

及 Luk t 模处理后的结果,且 D-S 结合法则的结果优于平均法则.Wine 数据集上得到结果同 8D5K 数据得到的结果相似,说明 Wine 数据集同 8D5K 数据的结构相似,都是线性可分的结构.对于 Seeds 数据,在平均结合法则以及 prod t 模传递闭包后得到的结果最好,其他情况时的表现都很差.这是因为如果 2 个类之间存在重合区域时,传递闭包后的这 2 个类之间的相似度会变大,从而导致最终的结果变差.

4 结论

本文在证据理论的基础上提出一种新的聚类集成方法.证据理论是更一般性的理论,能更好地表示数据中的内在结构.在证据理论的基础上,首先将所有信息从一个框架转到另一个框架上,即从基划分转成关系表示.考虑到信息源的不可靠性,本文对关系表示进行证据折扣处理.结合处理后的关系表示,得到融合后的关系表示,从中提取互相关矩阵.为了充分利用样本间的信息,对得到的互相关矩阵进行传递闭包处理,从而得到新的互相关矩阵.将该矩阵视为新的样本数据进行再聚类,得到最终的结果.通过将该方法应用到模拟和真实数据集,证明了该方法对于模糊系数的选取具有稳健性.在调整的兰德指数评价标准下,本文所提方法大多优于已有方法.本文中折扣因子是通过 Masson 等[5] 中的算法得到的,如何更好地选取折扣因子将是本文进一步的研究方向.

参考文献

References

[1] Denoeux T, Masson M H. EVCLUS: evidential clustering of proximity data[J]. IEEE Transactions on Systems, Man and Cybernetics, Part b (Cybernetics), 2004, 34(1): 95-109

[2] Masson M H, Denœux T. ECM: an evidential version of the fuzzy c-means algorithm[J]. Pattern Recognition, 2008, 41(4): 1384-1397

[3] Masson M H, Denœux T. RECM: relational evidential c-means algorithm[J]. Pattern Recognition Letters, 2009, 30(11): 1015-1026

[4] Antoine V, Quost B, Masson M H, et al. CEVCLUS: evidential clustering with instance-level constraints for relational data[J]. Soft Computing, 2014, 18(7): 1321-1335

[5] Masson M H, Denoeux T. Ensemble clustering in the belief functions framework[J]. International Journal of Approximate Reasoning, 2011, 52(1): 92-109

[6] Denœux T, Masson M H. Evidential reasoning in large partially ordered sets[J]. Annals of Operations Research, 2012, 195(1): 135-161

[7] Li F J, Qian Y H, Wang J T, et al. Multigranulation information fusion: a dempster-shafer evidence theory based clustering ensemble method[C]// 2015 International Conference on Machine Learning and Cybernetics (ICMLC), Guangzhou, China, 2015: 58-63

[8] Denoux T, Li S M, Sriboonchitta S. Evaluating and comparing soft partitions: an approach based on Dempster - Shafer theory[J]. IEEE Transactions on Fuzzy Systems, 2018, 26(3): 1231-1244

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

[10] Ghaemi R, Sulaiman M N, Ibrahim H. A survey: clustering ensembles techniques[C]// Proceedings of World Academy of Science Engineering and Technology, 2009: 109-114

[11] Vega-Pons S, Ruiz-Shulcloper J. A survey of clustering ensemble algorithms[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2011, 25(3): 337-372

[12] Zhan J M, Chen J T, Xing J Q. Research advance of clustering ensemble algorithm[C]// 2017 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR), Ningbo, China, 2017: 109-114

[13] Akbari E, Mohamed Dahlan H, Ibrahim R, et al. Hierarchical cluster ensemble selection[J]. Engineering Applications of Artificial Intelligence, 2015, 39: 146-156

[14] Strehl A, Ghosh J. Cluster ensembles: a knowledge reuse framework for combining multiple partitions[J]. Journal of Machine Learning Research, 2002: 583-617

[15] Dudoit S, Fridlyand J. Bagging to improve the accuracy of a clustering procedure[J]. Bioinformatics, 2003, 19(9): 1090-1099

[16] Fischer B, Buhmann J M. Bagging for path-based clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003, 25(11): 1411-1415

[17] Fred A L N, Jain A K. Data clustering using evidence accumulation[C]//. Object Recognition Supported by User Interaction for Service Robots, Quebec City, Quebec, Canada, 2002: 276-280

[18] Fred A L N, Jain A K. Combining multiple clusterings using evidence accumulation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(6): 835-850

[19] Iam-On N, Boongoen T, Garrett S. Refining pairwise similarity matrix for cluster ensemble problem with cluster relations[M]// Iam-On N, Boongoen T, Garrett S. Discovery Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008: 222-233

[20] Vega-Pons S, Ruiz-Shulcloper J. Clustering ensemble method for heterogeneous partitions[M]// Vega-Pons S, Ruiz-Shulcloper J. Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009: 481-488

[21] Wang X, Yang C N, Zhou J. Clustering aggregation by probability accumulation[J]. Pattern Recognition, 2009, 42(5): 668-675

[22] Yang L, LV H, Wang W. Soft cluster ensemble based on fuzzy similarity measure[C]// The Proceedings of the

Multiconference on Computational Engineering in Systems Applications,2006:1994-1997
[23] Dubois D,Prade H.Fundamentals of fuzzy sets[M].Boston,MA:Springer US,2000.DOI:10.1007/978-1-4615-4429-6
[24] Hubert L,Arabie P.Comparing partitions[J].Journal of Classification,1985,2(1):193-218
[25] Dua D,Graff C.UCI machine learning repository[D].Irvine,CA:University of California,School of Information and Computer Science,2019
[26] Bezdek J C,Ehrlich R,Full W.FCM:the fuzzy c-means clustering algorithm [J]. Computers & Geosciences,1984,10(2/3):191-203

Clustering ensemble method based on belief function theory

LI Feng[1] LI Shoumei[1] Denoeux Thierry[1,2]
1 College of Applied Sciences,BeijingUniversity of Technology,Beijing 100124
2 Centre National de la Recherche Scientifique,Sorbonne Universités,
Universitéde Technologie de Compiègne,Heudiasyc (UMR 7253),France

Abstract To overcome the instability of one single clustering result,we propose a new clustering ensemble method based on Dempster-Shafer theory (also known as belief function theory).In general,ensemble methods consist of two principal steps:generating base partitions and combining them into a single one;our method mainly focuses on the second step.After obtaining the base partitions in the first step,we convert them into an intermediate interpretation,which can be called a relational representation.We believe that the evidence source from the relational representations may be doubtful,which can be fixed by using the discounting process in belief function theory.After discounting the relational representations,we can combine them in the evidential level by different combination rules.Then,we can obtain the belief matrix or plausibility matrix from the fused relational representation,which can be seen as a co-association matrix between objects.To make full use of the transitive property between objects,we treat this co-association matrix as a fuzzy relation and make it the transitive closure to yield a fuzzy equivalence relation.The final partition is obtained by applying some clustering algorithms to the new co-association matrix.The experimental results show the stability and efficiency of our method.

Key words belief function;clustering ensemble;relational representation;co-association matrix;transitive closure