基于概率转换函数与邓熵的关联分类融合算法

马晓剑，张家绪，王 奥，林煜华

（东北林业大学 理学院，黑龙江 哈尔滨 150040）

摘 要：在信息高度冲突情况下，基于信度转移模型的关联二分类算法会出现分类精确率低、计算量大等问题。为解决这些问题，提出一种关联分类融合算法。由类关联规则的置信度构造基本概率指派函数，给出分类规则；基于 Pignistic 概率转换函数重构邓熵，结合 PPT 散度优化分类规则；通过解决概率转换函数的高冲突，使证据的高冲突问题得到有效改善，实验结果验证了这一点。实验结果还表明，分类准确性得到有效提升，与 BJS 散度等信度转移模型相比，计算量显著降低。

关键词：数据挖掘；关联分类；证据理论；信度转移模型；邓熵；证据冲突；信度散度测度

中图法分类号：TP181 **文献标识号**：A **文章编号**：1000-7024（2023）05-1392-08

doi：10.16208/j.issn1000-7024.2023.05.015

Associative classification algorithm based on Pignistic probability transform and Deng entropy

MA Xiao-jian, ZHANG Jia-xu, WANG Ao, LIN Yu-hua

(College of Science, Northeast Forestry University, Harbin 150040, China)

Abstract: In the case of high conflicts of information, problems such as low classification accuracy and a large amount of calculation occurs when associative classification algorithm is based on transferable belief model. To solve these problems, an algorithm was proposed. BPAs were constructed by confidence, and the classification rules were given. Deng entropy was innovatively reconstructed based on Pignistic probability transformation functions, and the Pignistic functions as classification rules were optimized combined with PPT divergence. The high conflict of evidence was effectively improved by solving the high conflict of probability transformation functions. Experimental results confirm this and show that the classification accuracy is effectively improved. Compared with the transferable belief model such as BJS divergence, the calculation of the proposed algorithm is significantly reduced.

Key words: data mining; associative classification; evidence theory; transferable belief model; Deng entropy; evidential conflict; belief divergence measure

0 引 言

一些关联分类算法[1-4]在大规模挖掘类关联规则时往往需要消耗大量内存[5,6]。Guil 将关联规则挖掘与 D-S 证据理论结合，提出了基于信度转移模型的关联二分类算法[7]（以下简称 TBM），具有很好的可解释性，但精确度不具有显著优势。D-S 证据理论视多源信息为证据，在遥感检测、多智能体决策等领域中取得了较好效果[8,9]。该理论虽然具有很强的决策能力，但因证据高冲突，容易产生错误的决策[10,11]，TBM 算法正是由于这一原因导致精确度显著降低。Xiao 等提出使用 BJS 散度来自主识别证据的可靠性[12]，但推广至关联分类问题时，引入的大量计算导致耗

收稿日期：2021-12-03；**修订日期**：2023-04-28

基金项目：中央高校基本科研业务费专项基金项目（2572018BC21）；黑龙江省大学生创新创业训练计划基金项目（202110225372）；黑龙江省博士后基金项目（LBH-Z18003）

作者简介：马晓剑（1977-），女（回族），黑龙江哈尔滨人，硕士，副教授，CCF 会员，研究方向为数据挖掘和图像处理；张家绪（2001-），男，山东临沂人，本科，研究方向为数据挖掘、证据理论和机器学习；王奥（2001-），男（蒙古族），内蒙古通辽人，本科，研究方向为证据理论、数据挖掘和图像处理；林煜华（2000-），男，福建莆田人，本科，研究方向为证据理论和机器学习。

E-mail: mxjzy@nefu.edu.cn

时严重，使得改进后的算法较难应用。Xu 等基于 Pignistic 概率转换提出了一种散度——PPT 散度[13]，时间代价有所降低，但仍未能解决关联分类中计算量大的问题。

本文提出了一种基于概率转换和邓熵的加权关联规则融合算法（以下称为 PE 加权算法）。该算法利用 Pignistic 概率转换函数给出初始分类规则，将概率转换函数与邓熵结合，生成了新的熵来估计冲突信息量，再利用 PPT 散度进一步减少不确定性，实现了对 Pignistic 概率转换函数的权重再分配，优化了分类规则。该算法将以往的修正证据之间的高冲突改进为修正 Pignistic 概率转换函数的高冲突，显著降低了计算量。实验结果表明了本文算法的有效性，相较于文献［7］的算法，各项性能得到了较大的提升。

1　关联分类与证据理论的基本概念

定义 1　关联规则：假设全体项目集为 I，关联规则 R 被定义为 X 和 Y 组成的蕴含式

$$R:X\rightarrow Y;\ X,Y\subset I;$$

其中，X 称为规则的前提，Y 称为规则的结论。

定义 2　支持度：假设数据集为 Ω，关联规则的支持度为 $\{X,Y\}$ 在数据集中都出现的次数，即

$$sup(X\rightarrow Y)=\frac{Frequency(X\cup Y)}{|\Omega|}$$

其中，$Frequency(X\cup Y)$ 为 $\{X,Y\}$ 在数据集 Ω 中的出现频率。

定义 3　置信度：假设关联规则 $R:X\rightarrow Y$，关联规则 R 的置信度定义为在 X 发生的前提下，$\{X,Y\}$ 发生的概率

$$conf(R)=P(X,Y)=\frac{sup(R)}{sup(X)}=\frac{sup(X\cup Y)}{sup(X)}$$

置信度用于表征关联规则 R 的可信程度。

定义 4　类关联规则：结论为类标签的关联规则称为类关联规则。相同前提对应的类关联规则满足形式 $R^{i,*}:W_i\rightarrow C^*$，其中 $W_i\in 2^X$ 为第 i 种前提，$C^*\in Y$ 为类标签。

定义 5　基本概率指派（mass 函数）：假设完备集合 $\Theta=\{\theta_1,\theta_2,\cdots,\theta_n\}$ 是辨识框架，其中 θ_i，θ_j，$i\neq j$ 是互斥的元事件，如果命题集 2^Θ 到 $[0,1]$ 的映射 m，即 $m:2^\Theta\rightarrow[0,1]$，满足下列条件

$$\begin{cases}m(\varnothing)=0\\ \sum_{W\subset 2^\Theta}m(W)=1\end{cases}$$

则称该映射 m 为基本概率指派，记为 BPA。这里 $m(W)$ 被视为准确分配给 W 的信度。在证据理论中，W 被称为焦元。在本文中，辨识框架 $\Theta=\{C^1,C^2\}$，其中 C^1 和 C^2 是焦元，是二分类的两个标签。

定义 6　Dempster 组合规则：假设 $m_1,\cdots,m_n$ 是辨识框架 Θ 下的 n 组基本概率指派，$W_{11},W_{12},\cdots,W_{1s},\cdots,W_{n1},W_{n2},\cdots,W_{nt}$ 为 n 组焦元，则 Dempster 组合规则表示如下

$$\begin{cases}m(\varnothing)=0\\ m(W)=\dfrac{1}{1-k}\sum\limits_{\cap W_j=W}\prod\limits_{1\leqslant j\leqslant n}m_j(W_{ji})\end{cases}$$

其中，$k=\sum\limits_{\cap W_j=\varnothing}\prod\limits_{1\leqslant j\leqslant n}m_j(W_{ji})$ 定义为冲突系数，是经典证据理论中的冲突度量。

定义 7　Pignistic 概率转换：假设 m 为一组基本概率指派，则辨识框架 Θ 下的 Pignistic 概率分布函数 $BetP$ 定义为

$$BetP(W)=\sum_{W\subset\Theta,\omega\subset W}\frac{m(W)}{|W|}$$

式中：$|W|$ 表示焦元 W 的势，Pignistic 概率转换是将基本概率指派的证据转换为由 Pignistic 概率分布函数 $BetP$ 表示的离散概率分布。

定义 8　邓熵[14]：假设 m 为基本概率指派，$W\in 2^\Theta$，则 m 函数的邓熵可表示如下

$$E(m)=\sum\nolimits_{W\in 2^\Theta}m(W)\log_2\frac{m(W)}{2^{|W|-1}}$$

其中，$|W|$ 表示焦元 W 的势。

定义 9　PPT 散度[13]：假设基本概率指派 m_i 和 m_j 对应的 Pignistic 概率转换函数为 $BetP_{m_i}$ 和 $BetP_{m_j}$，则转换函数 $BetP_{m_i}$ 和 $BetP_{m_j}$ 之间的 PPT 散度为

$$PPT(BetPm_i,BetPm_j)=\frac{1}{2}\left[S\left(BetPm_i,\frac{BetPm_i+BetPm_j}{2}\right)+S\left(BetPm_j,\frac{BetPm_i+BetPm_j}{2}\right)\right]$$

其中，$S(BetPm_i,BetPm_j)=-\sum_k BetP_{m_i}(W_k)\log\dfrac{BetP_{m_i}(W_k)}{BetP_{m_j}(W_k)}$。

2　基于 PE 的关联分类融合算法

常见的关联分类算法[4]利用剪枝、排序等方法在类关联规则集中生成分类规则，与它们不同的是 TBM 在分类规则生成阶段引入了证据理论，使用 Pignistic 概率转换生成能覆盖数据集的分类规则。使用证据理论构建的分类器虽然具有模型简单、计算高效等优点[7]，但是在分类规则生成阶段易受到冲突信息的干扰，导致分类器的精确度下降。BJS 散度加权算法是一种冲突信息抑制算法，该算法虽然能在一定程度上克服信息高冲突的影响，提高分类性能，但未针对指数级增长的证据量进行优化，在解决关联二分类问题时，分类器消耗大量时间用于优化分类规则，降低了这种分类算法的应用价值。为克服以上缺陷，本文从提高分类器的分类精确度、减少计算量的角度出发，提出了 PE 分类器，用于在大规模高冲突的信息下进行关联二分类预测。首先，该算法使用 Apriori 算法挖掘类关联规则，利用类关联规则的置信度构造基本概率指派函数，给出初步的分类规则；其次，将 Pignistic 概率转换函数与邓熵结合，

构建了一种新的熵来描述概率转换函数的不确定性，再结合 PPT 散度生成差异度量矩阵对概率转换函数进行权重分配，达到优化分类规则的目的；最后，结合软投票方法给出了预测的分类结果。该算法提高了证据理论生成关联二分类器的实用性能。

2.1 基于 Pignistic 概率转换的分类规则

本文先利用 One-Hot 编码将特征转化为二值型，再使用 Apriori 算法挖掘类关联规则。假设经 One-Hot 编码后的稀疏特征空间为 X，其维度为 $N_{sample}\times N$，X_k 为该特征空间的第 k 个特征，Y 为类标签集。由 Apriori 算法挖掘后形成的类关联规则集记为 *CAR-Set*。将 *CAR-Set* 中类关联规则的置信度转换为基本概率指派，见式（1）

$$m^{(i)}(C^j)=\overline{conf}(R^{i,j}) \tag{1}$$

其中，$R^{i,j}$ 为第 i 个前提对应的类关联规则，由式（1）可知第 i 个类关联规则的前提对应第 i 个基本概率指派，其结论 C^j 为第 j 个类标签，由于解决的是二分类问题，所以这里 $j=1$，2。$\overline{conf}(R^{i,j})$ 为类关联规则 $R^{i,j}$ 对应的归一化置信度，满足

$$\overline{conf}(R^{i,j})=\frac{conf(R^{i,j})}{\sum_{l=1}^{|Y|}conf(R^{i,l})} \tag{2}$$

显然有 $\sum_{l=1}^{|Y|}m^{(i)}(C^l)=\sum_{j=1}^{|Y|}\overline{conf}(R^{i,l})=1$。由式（2）可以看出，基本概率指派获得的信息为第 i 个前提与第 j 个类标签同时发生的概率。

有了基本概率指派函数后，由 Pignistic 概率转换函数将基本概率指派转换为一个概率测度 $BetP$[15]，以此来表征特征 X_k 发生条件下的待预测类别 C^l 发生的条件概率（见式（3）），此时分类器的分类规则初步形成

$$BetP_{(X_k^{(C^l)})}=P(y=C^l\mid X_k)=\sum_{X_k\in W_i}\frac{m^{(i)}(C^l)}{|W_i|},\ X_k\in X,\ C^l\in Y \tag{3}$$

2.2 基于 PE 加权的分类规则优化

在二分类问题中，由类关联规则产生的基本概率指派往往具有高度冲突性，这种冲突会对分类器的决策产生显著影响，导致分类器的判别精度下降，因此需要对初步形成的分类规则进行修正。

以特征 X_m 和 X_n 为例：当训练集中仅存在类关联规则 $X_m\rightarrow C^1$ 和 $X_m\rightarrow C^2$，且两条规则具有相同的归一化置信度时，分类器将学习到以下信息：$P(y=C^1\mid X_m)=1$，$P(y=C^2\mid X_m)=1$，$P(y=C^1\mid X_n)=0$，$P(y=C^2\mid X_n)=1$。这表示测试样本 y 有特征 X_m 和 X_n 发生，在这种情况下，该样本既不能被判定为第 C^1 类样本，也不能被判定为第 C^2 类样本，原因在于类关联规则 $X_m\rightarrow C^1$ 和 $X_m\rightarrow C^2$ 提供给分类器的证据信息是完全相反的、高度冲突的，分类器在这种相互矛盾的分类规则下不能做出正确的判别，因此，本文引入 PPT 散度对高冲突的信息进行修正。

事实上，Xu 提出的 PPT 散度是利用最大熵思想将 BJS 散度推广为 Pignistic 概率转换之间的距离测度，PPT 散度的优势之一是可以为每种概率分布分配不同的权重，它代表了证据间可靠性的差异大小，PPT 散度越大，说明两个证据间越相似。在多于两个证据的情形时，它对分类结果会产生重要的作用，并且它的算法复杂度要低于 BJS 散度，但是在关联分类问题中，由于 Apriori 算法最多可以挖掘到 $2^{|2^x|\times|Y|}$ 条关联规则，如果直接利用 PPT 散度加权，有可能计算多达 $2^{|2x|\times|Y|}$ 条证据的权重，产生巨大的计算开销。

考虑到上述情况，本文提出了一种信息熵加权方法，该方法尽管利用 PPT 散度，却不是对概率指派函数进行权重再分配，而是利用 Pignistic 概率转换函数给出新的熵，通过这个新的熵来构造新的权重，实现对 $BetP$ 的再分配，从而达到优化分类规则的目的，抑制高冲突的同时又加速了计算。具体优化分类规则的算法如下：

（1）给出 PPT 散度的差异度量矩阵 *DMM*。

令 $PPT_{mn}=PPT(BetP(X_m),BetP(X_n))$，它定量地描述了第 m 个特征 X_m 和第 n 个特征 X_n 的 Pignistic 概率转换函数之间的冲突性，由它建立的差异度量矩阵 *DMM* 形式如下

$$DMM=\begin{pmatrix}0 & PPT_{12} & \cdots & PPT_{1(N-1)} & PPT_{1N}\\ PPT_{21} & 0 & \cdots & PPT_{2(N-1)} & PPT_{2N}\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ PPT_{(N-1)1} & PPT_{(N-1)2} & \cdots & 0 & PPT_{(N-1)N}\\ PPT_{N1} & PPT_{N2} & \cdots & PPT_{N(N-1)} & 0\end{pmatrix} \tag{4}$$

其中，N 为 One-Hot 特征空间的维数，PPT_{mn} 也可以利用 Shannon 熵，将其改写为

$$PPT(BetP(X_m),BetP(X_n))=H\left(\frac{BetP(X_m)+BetP(X_n)}{2}\right)-\frac{1}{2}H(BetP(X_m))-\frac{1}{2}H(BetP(X_n))=\frac{1}{2}\left[\sum_{l=1}^{2}BetP(X_m^{(C^l)})log\left(\frac{2BetP(X_m^{(C^l)})}{BetP(X_m^{(C^l)})+BetP(X_n^{(C^l)})}\right)+\sum_{l=1}^{2}BetP(X_n^{(C^l)})log\left(\frac{2BetP(X_n^{(C^l)})}{BetP(X_m^{(C^l)})+BetP(X_n^{(C^l)})}\right)\right] \tag{5}$$

（2）构造可信度 *Crd*。

PPT 散度是 Pignistic 概率转换函数的距离测度，可用于测量转换函数之间的冲突性大小，如果一个特征与其它特征之间的冲突越大，该特征对应的 PPT 散度就越大，可以认为该函数具有的可信度就越低，其对应的信息量测度也越小，为了提高决策的可信度，PPT 散度赋予该函数的权重就要越小。根据这一原理，由 PPT 散度构造可信度 Crd_m，即：$Crd_m=\frac{Sup_m}{\sum_{s=1}^{N}Sup_s}$，其中 $Sup_m=\frac{1}{\overline{PPT}_m}$，

$\overline{PPT}_m = \dfrac{\sum_{m=1,n\neq m}^{N} PPT_{mn}}{N-1}$，$1 \leqslant m \leqslant N$，$1 \leqslant n \leqslant N$。

（3）基于 $BetP$ 函数构造新的熵。

当分类的不确定性越高时，邓熵[14]就越大，证据信息的支持度就要越小，于是可以利用邓熵来进一步描述分类的不确定性。与文献［13］不同的是，本文是基于 Pignisitc 概率转换函数 $BetP$ 构造邓熵，并没有利用概率指派函数。

不妨令 $BetP$ 重构特征 X_m 的邓熵是 S_{X_m}，即

$$S_{X_m} = \sum\nolimits_l BetP(X_m)\log\frac{BetP(X_m^{(C^l)})}{2^{|X_m|}-1} \tag{6}$$

容易证明，该定义满足熵的性质。于是 Pignistic 概率转换函数的不确定信息量为：$IV_m = e^{-S_{X_m}}$；进一步将 IV_m 归一化为：$\tilde{I}V_m = \dfrac{IV_m}{\sum_{s=1}^{N} IV_s}$。

（4）给 Pignistic 概率转换函数（即 $BetP$）重新分配权重，实现分类规则的优化。

令第 m 个 Pignistic 概率转换函数的信息权重为 $ACrd_m = Crd_m \times \tilde{I}V_m$，$1 \leqslant m \leqslant N$，并给所有 Pignistic 概率转换函数重新分配权重，则将加权修正后的 Pignistic 概率转换函数作为分类规则，即

$$NBetP(m) = \tilde{A}Crd_m \times BetP(X_m) \tag{7}$$

这里 $\tilde{A}Crd_m = \dfrac{ACrd_m}{\sum_{s=1}^{N} ACrd_s}$ 为新的权重，$1 \leqslant m \leqslant N$。

在 Pignistic 概率转换函数生成后，一方面 PPT 散度作为距离的度量给出置信度的估计，另一方面邓熵作为信息量的度量给出不确定信息量的估计，二者作为权重的分支需要分别计算（如图 1 所示），需要指出的是本文算法中的 PPT 散度、邓熵均由 Pignistic 概率转换函数重新表述，这一点与以往算法不同，其中使用 Pignistic 概率转换函数生成权重的计算流程如图 1 所示。

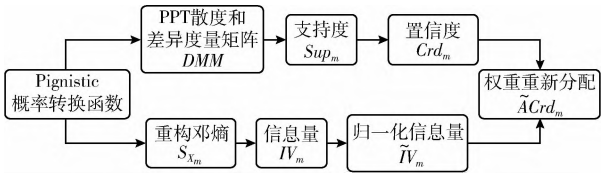

图 1　Pignistic 概率转换函数优化权重的流程

至此，分类规则优化结束，分类器构建完成。伪代码给出了 PE 算法优化分类规则和生成分类器的过程。

伪代码：PE 优化分类规则与分类器构建

输入：Pignistic 概率函数集 $BetPSet$

for $BetP(X_m)$ in $BetPSet$：

　　for $BetP(X_n)$ in $BetPSet$：

　　$DMM(i,j) = PPT(BetP(X_m), BetP(X_n))$

　　end

　　输出 DMM，计算 $\tilde{A}Crd_m$

　　for m in 1：N：

　　计算 $NBetP(X_m) = \tilde{A}Crd_m \times BetP(X_m)$

　　end

end

输出：优化后的分类规则集 $\{BetP(\cdot)\}$ 和第 i 个分类器，$i = 1,\cdots,n$

2.3　PE 加权关联规则分类的实现

在分类器学习完成后，优化后的分类规则集 $\{BetP(\cdot)\}$ 得以生成，在分类阶段将优化后的 $BetP(\cdot)$ 信息融合，向分类器输入类别未知的待预测样本即可预测类别。假设待测样本 $x_{1\times N}$ 为一个 One-Hot 编码后的特征向量，其第 s 个特征值为 x_s，计算待测样本 x 属于类别 C^l 的概率如下

$$SBetP(x,C^l) = \sum_{s=1}^{N} NBetP(X_s^{(C^l)}) \times I_{(x_s=1)} \tag{8}$$

其中，$I_{(x_s=1)}$ 为示性函数，当 $x_s=1$ 时，值为 1，否则值为 0。待测样本的类别由条件概率和取最大值时的类别标签决定，即

$$\hat{y} = \operatorname{argmax} \sum_i SBetP^i(x,C^l) \tag{9}$$

本文算法的流程如图 2 所示。

由上述分类规则的优化过程可知，本文提出的 PE 加权算法与 BJS 加权算法[12]、PPT 散度加权[13]有很大的不同，虽然三者都将 Pignistic 概率转换函数作为分类规则的组成部分，但是由于它们对高冲突的抑制机理不同，导致在关联分类场景中，其构造的分类器就产生了差异性。BJS 散度加权法是通过基本概率指派来实现 Pignistic 概率转换，属于间接优化分类器参数的优化算法；PPT 散度加权法虽然考虑了 Pignistic 概率函数之间的冲突，但其最终的优化目标仍然是缩小基本概率指派之间的冲突，这与 BJS 散度加权法的思想类似；只有 PE 加权算法具有直接对 Pignistic 概率函数的冲突进行抑制的特性，若某个 Pignistic 概率函数与其它概率函数之间存在的冲突越大，则表明该函数具有的可信度越小，其对应的归一化信息量测度也就越低，PE 加权法赋予该转换函数的权重就越小，分类的可行性就越高。此外，PE 加权算法的更大优势还体现在可以避免处理海量证据、缩短计算时间、提高计算效率等方面，这一内容的讨论将在 3.2 节中详细给出。

3　实验与分析

3.1　分类结果分析

本文选取 UCI 数据库和 Wolfram 数据库中的 11 个二分类数据集用于测试算法性能，数据集描述见表 1。

对于具有连续特征的数据集，本文在 One-Hot 编码预处理阶段选取 25%、50%、75%作为自适应分箱法的区间

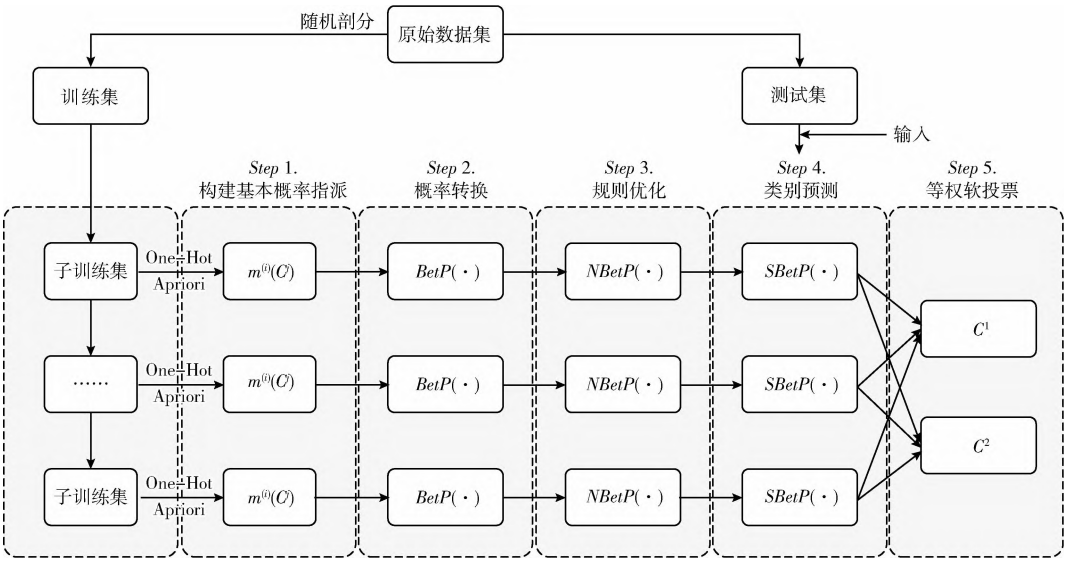

图 2 PE 加权算法流程

表 1 测试数据集描述

数据集	来源	类别数	样本数	正样本数	负样本数	特征数
adult	UCI	2	32 561	7841	24 720	14
blood	UCI	2	748	178	570	5
blogger	UCI	2	100	32	68	6
breast	UCI	2	699	241	458	10
cervical	UCI	2	72	21	51	19
covid	Wolfram	2	218	79	139	58
dishonest	UCI	2	322	97	255	5
HCV	UCI	2	615	82	533	14
inflammation	UCI	2	120	59	61	6
liver	UCI	2	583	167	416	10
vote	UCI	2	435	168	267	16

剖分点，将连续特征转换为离散特征。此外，由于单一评价指标只能反映出分类器在某一假设下的分类效果，为了全面评价算法的分类性能，本文选取了 Accuracy（精确度）、Precision（准确率）、Recall（召回率）、F1 得分和 AUC 值作为评价指标，基于 WEKA 平台选取 CBA、C4.5、Ripper、PART 以及 Python3.7 编写的 TBM 作为对比算法进行比较。以 33% 作为测试集尺度，设置超参数 $minsup=0.01$、$minconf=0.05$ 进行了 10 次随机实验，记录 10 次重复实验下各个评价指标的平均值。表 2 和图 3 给出了 PE 加权融合算法与对比算法在 11 个测试数据库下的分类结果。

由表 2 可以看出，本文提出的 PE 加权融合算法与经典的 C4.5 算法在精确度方面几乎具有一样的表现力。图 3 描述了 11 个数据集在 4 个指标下的表现，PE 算法相对于其它算法在 Precision、F1-score 和 AUC 3 个指标方面呈现了较高的优势；不难看出与 TBM 算法相比，本文算法的各指标均有所提升，特别是 PE 算法的 AUC 表现更突出，说明本文所提出的分类器在不同判别阈值下的泛化能力更强，分类器在不平衡数据集下的表现性能更突出。图 2 说明 PE 算法对于正负样本分布有偏性较不敏感。为进一步说明本文算法在 AUC 指标下的性能具有较强的鲁棒性，本文又对 covid、cervical、HCV、liver4 个数据集进行了 10 次重复实验，给出了 Ripper、CBA、PART 这 3 个算法与 PE 算法的 AUC 对比（如图 4 所示）。对于 4 个数据库，CBA 和 PART 两种算法的 AUC 曲线均有较大的波动性，说明分类的正确性存在严重的不稳定性，特别是 CBA 算法对于 HCV 和 liver 库 AUC 表现过低，说明该种算法不能有效识别正样本；Ripper 算法对于 cervical、HCV 和 liver 数据库

表 2　Accuracy（精确度）实验结果

数据集	C4.5	Ripper	CBA	PART	TBM	PE
adult	0.821(±0.000)	0.819(±0.000)	0.759(±0.000)	**0.823(±0.000)**	0.692(±0.000)	0.712(±0.000)
blood	0.749(±0.000)	**0.750(±0.000)**	0.748(±0.000)	0.739(±0.001)	0.637(±0.000)	0.645(±0.000)
blogger	0.750(±0.003)	0.749(±0.005)	0.803(±0.006)	0.818(±0.001)	**0.833(±0.000)**	0.780(±0.002)
breast	0.967(±0.001)	0.965(±0.000)	0.954(±0.002)	0.957(±0.000)	0.961(±0.000)	**0.969(±0.000)**
cervical	0.804(±0.013)	0.804(±0.004)	0.704(±0.006)	0.758(±0.023)	0.817(±0.005)	**0.821(±0.006)**
covid	0.752(±0.006)	0.744(±0.001)	0.719(±0.002)	0.775(±0.005)	0.779(±0.002)	**0.809(±0.001)**
dishonest	**0.995(±0.000)**	0.990(±0.000)	0.936(±0.013)	0.990(±0.000)	0.911(±0.000)	0.911(±0.000)
HCV	**0.937(±0.000)**	0.934(±0.000)	0.874(±0.000)	0.926(±0.000)	0.744(±0.001)	0.869(±0.001)
inflammation	**1.000(±0.000)**	**1.000(±0.000)**	0.881(±0.002)	0.943(±0.012)	0.982(±0.000)	**1.000(±0.000)**
liver	0.674(±0.001)	0.687(±0.002)	**0.711(±0.001)**	0.667(±0.001)	0.658(±0.001)	0.700(±0.001)
vote	**0.954(±0.001)**	0.951(±0.000)	0.935(±0.001)	0.952(±0.000)	0.888(±0.000)	0.897(±0.000)

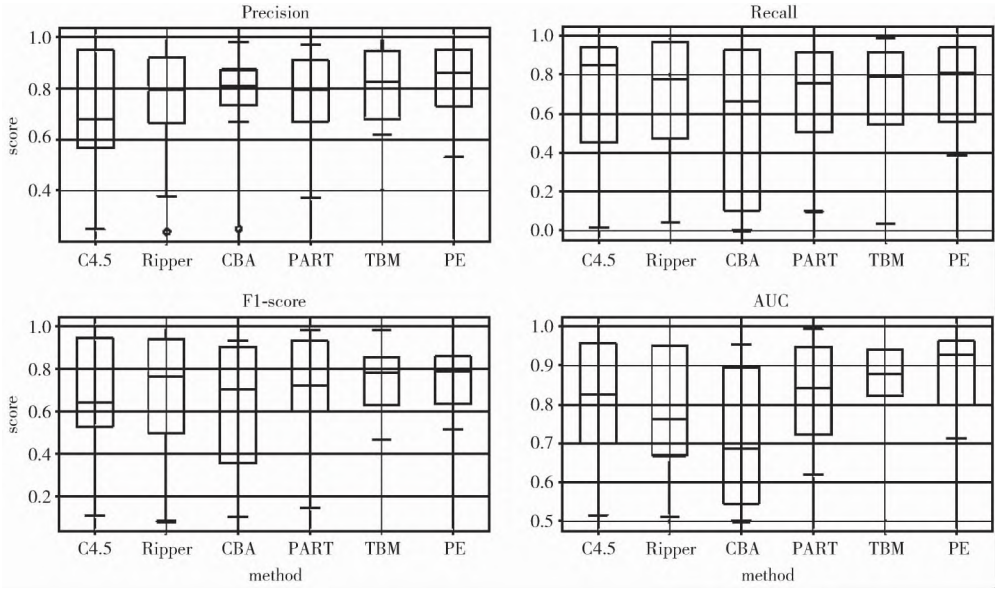

图 3　测试数据集 4 类评价指标的箱线

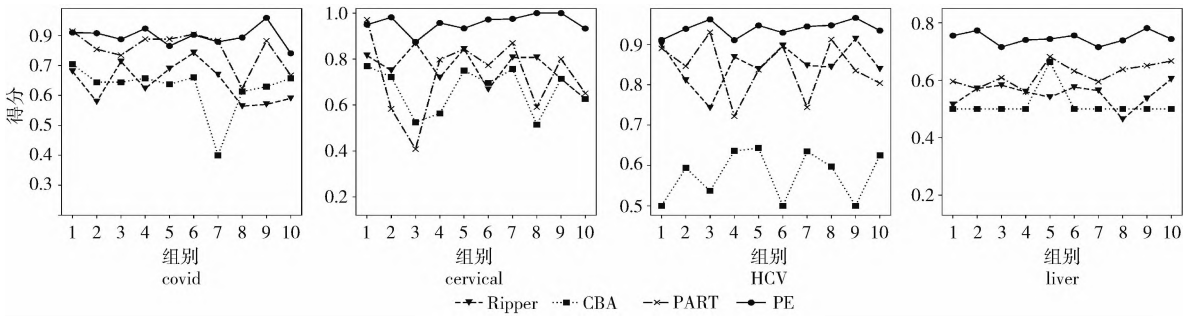

图 4　4 个数据集在 AUC 指标下 10 次重复实验结果对比

的分类正确性基本上在 75%到 90%之间振荡，但是对于 covid 库的分类结果较差，10 组实验基本都是在 85%以下；而本文提出的 PE 加权融合算法对于不同数据库的分类性能稳定，AUC 指标基本优于其它算法，且正确性基

本在 90%以上，特别是对 cervical 库的多组分类结果达到 96%以上。

仅仅以分类精确度来评价算法性能是不充分的，比如在处理不平衡数据集时，传统分类算法容易出现分类整体精确度较高，却对少数类别关注不足的问题[16,17]。本文以 liver 数据集为例，绘制了 10 次重复实验下的平均混淆矩阵（如图 5 所示）。由图 5 可知，PE 算法与 TBM 的性能接近，都具备正确识别少数类样本的性能，但 PE 算法的正确率要略优于 TBM，且错分率也低于 TBM；Ripper、C4.5、PART 这 3 种算法正好相反，它们正确识别少数类的能力明显低于 PE 和 TBM，且在总体精确度指标下弱于 PE；对于 CBA 算法而言，虽然表 2 显示对于 liver 数据集，CBA 的精确度要略高于本文的 PE 算法，但是由图 5 可知，CBA 算法几乎完全将样本判别为多数类，说明该算法生成的分类规则对少数类的关注严重不足，CBA 算法在该数据集下已经失效。综上，PE 算法识别少数类的性能要优于其它对比算法，它继承了 TBM 识别少数类的能力，同时还具有更高的识别精确度，并且这一结论与图 4 的结果是相一致的。由此可见，当数据集正负样本分布有偏时，本文算法在 AUC 指标表现突出的前提下，对少数类识别精确度也能保持较高的水平。

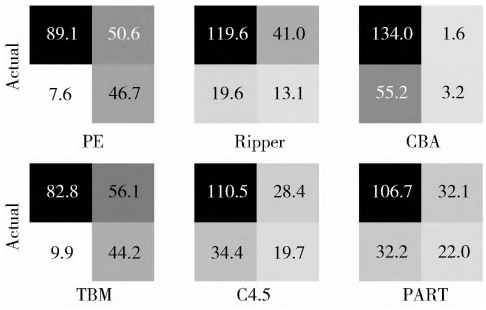

图 5 基于 liver 数据集的平均混淆矩阵

3.2 运行时间分析

为了说明利用 PE 加权优化参数在计算效率方面比直接使用 BJS 散度加权更有效，在这一部分将对 PE 加权方法加速计算的机制进行分析。本文采用 BJS 散度和 PE 两种加权方法，在 11 个测试数据集上各进行 10 次随机实验，给出了这两种方法在高冲突抑制阶段的平均运算时间，为了避免不同时间的量级影响，文中对结果进行了规范化处理，实验结果如图 6 所示。不难看出本文算法运行的时间明显低于使用 BJS 散度加权的时间，其中处理 HCV 数据集时，具有最大的平均运算时间压缩比。

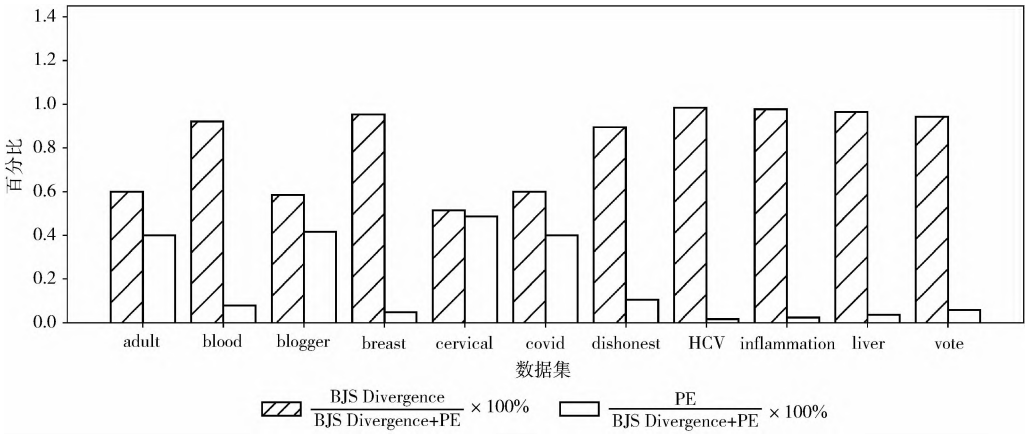

图 6 两种加权方式在高冲突抑制阶段的运算时间对比

由图 1 可知，本文权重 *ACrd* 是根据置信度和不确定信息量两部分生成的，在度量差异性和估计置信度过程中，虽然两种加权方法的时间复杂度均为 $O(n^2)$，n 为 *DMM* 矩阵维数，但在本文的 PE 加权算法中，矩阵 *DMM* 维数为 $N\times N$，N 为稀疏特征空间的特征数，而最差情况下 BJS 散度需计算的矩阵维数为 $2^{|2^{\mathcal{X}}|\times|Y|}\times 2^{|2^{\mathcal{X}}|\times|Y|}$，考虑实际情形，在一般情况下总有 $N\ll 2^{|2^{\mathcal{X}}|\times|Y|}$，因此 $O(N\times N)\ll O(2^{|2^{\mathcal{X}}|\times|Y|}\times 2^{|2^{\mathcal{X}}|\times|Y|})$ 成立，所以在生成置信度阶段，本文算法相较于原始 BJS 加权法消耗更少计算时间；而在计算不确定信息量部分中，两种加权方法的时间复杂度均为 $O(n)$，而由于一般情况下总有 $N\ll 2^{|2^{\mathcal{X}}|\times|Y|}$，即 $O(N)\ll O(2^{|2^{\mathcal{X}}|\times|Y|})$，所以本文算法又在生成不确定信息量阶段相较于原始 BJS 散度再一次降低了计算的时间代价。综上，相较于使用 BJS 散度优化参数而言，采用本文提出的 PE 加权算法能够大幅降低计算量、显著缩短分类规则的优化时间，更具有应用前景，图 6 也再次说明了这一点。

表 3 给出了两种加权方式下测试数据集中 *DMM* 矩阵的维数，由该表可以更直观看出 PE 加权融合算法有效利用了关联分类问题中数据集的特征数远小于 m 函数个数的性质，实现了对差异度量矩阵的降维，正是这种降维成功地实现了对 BJS 散度加权法的计算加速，因此本文所提出的 PE 加权在计算量和运行时间上具有更大的优势。而文献 [13] 给出

表 3　两种加权方式下测试数据集 *DMM* 矩阵维数

数据集	adult	blood	blogger	breast	cervical	covid	dishonest	HCV	inflammation	liver	vote
BJS 散度	1024	900	16 384	5776	7569	10 404	5776	14 641	37 249	17 956	22 801
PE	196	25	36	100	361	3364	25	196	36	100	256

的 PPT 散度加权算法，其 *DMM* 矩阵维数仍与 BJS 散度相同，该加权算法并未实现对 *DMM* 矩阵的降维优化。

4　结束语

为了解决关联分类问题中信度转移模型的分类精确率低等问题，本文利用 PPT 散度、概率转换函数和邓熵，结合关联规则挖掘构造了一种新的关联分类器，并提出了能够优化分类规则的冲突抑制算法——PE 加权融合算法。该算法通过直接优化 Pignistic 概率转换函数，实现了对差异度量矩阵的降维，与 BJS 散度和 PPT 散度加权法相比，PE 加权融合算法不但提高了信度转移模型的分类精确度，还大幅降低了计算量。此外，该算法在保持 AUC 指标占优的前提下，在正负样本分布有偏时也具有很好的判别稳定性。

参考文献：

[1] Hee-Young Park，Dong-Joon Lim. A design failure pre-alarming system using score-and vote-based associative classification [J]. Expert Systems with Applications，2020，164 (1)：113950.

[2] Geng Xiaojiao，Liang Yan，Jiao Lianmeng. ARC-SL：Association rule-based classification with soft labels [J]. Knowledge-Based Systems，2021，225 (15)：107116.

[3] Zou Yuchun，Chou Chun-An. A combinatorial optimization approach for multi-label associative classification [J]. Knowledge-Based Systems，2021，240 (6)：108088.

[4] Kumi S，Lim C，Lee S-G. Malicious URL detection based on associative classification [J]. Entropy，2021，23 (2)：182-193.

[5] Thanajiranthorn C，Songram P. Efficient rule generation for associative classification [J]. Algorithms，2020，13 (11)：299-314.

[6] QIN Chenpu，ZHANG Yunhua. Improvement of association classification algorithm based on classification pruning [J]. Computer Systems & Applications，2019，28 (4)：194-198 (in Chinese). [秦晨普，张云华. 基于分类修剪的关联分类算法改进 [J]. 计算机系统应用，2019，28 (4)：194-198.]

[7] Guil F. Associative classification based on the transferable belief model [J]. Knowledge-Based Systems，2019，182 (20)：104800.

[8] Zhao Jixiang，Liu Shanwei，Wan Jianhua，et al. Change detection method of high resolution remote sensing image based on D-S evidence theory feature fusion [J]. IEEE Access，2020，9：4673-4687.

[9] Liu Jinyu，Tang Yongchuan. Conflict data fusion in a multi-agent system premised on the base basic probability assignment and evidence distance [J]. Entropy，2021，23 (7)：820-833.

[10] JIANG Wen，DENG Xinyang. Information modeling and application of D-S evidence theory [M]. Beijing：Science Press，2018：16-17 (in Chinese). [蒋雯，邓鑫洋. D-S 证据理论信息建模与应用 [M]. 北京：科学出版社，2018：16-17.]

[11] Li Shanshan，Xiao Fuyuan，Abawajy J. Conflict management of evidence theory based on belief entropy and negation [J]. IEEE Access，2020，8：37766-37774.

[12] Xiao Fuyuan. Multi-sensor data fusion based on the belief divergence measure of evidences and the belief entropy [J]. Information Fusion，2019，46：23-32.

[13] Xu Shijun，Hou Yi，Deng Xinpu，et al. A novel divergence measure in Dempster-Shafer evidence theory based on pignistic probability transform and its application in multi-sensor data fusion [J]. International Journal of Distributed Sensor Networks，2021，17 (7)：155014772110314.

[14] Zhu Ruonan，Chen Jiaqi，Kang Bingyi. Power law and dimension of the maximum value for belief distribution with the maximum deng entropy [J]. IEEE Access，2020，8：47713-47719.

[15] Yang Yuexiang，Pan Xing，Cui Qingde. An evidence combination rule based on transferable belief model and application in reliability assessment with multi-source data [J]. IEEE Access，2020，8，69096-69104.

[16] CUI Wei，JIA Xiaolin，FAN Shuaishuai，et al. A new imbalanced association classification algorithm [J]. Computer Science，2020，47 (S1)：488-493 (in Chinese). [崔巍，贾晓琳，樊帅帅，等. 一种新的不均衡关联分类算法 [J]. 计算机科学，2020，47 (S1)：488-493.]

[17] XU Lingling，CHI Dongxiang. Machine learning classification strategy for unbalanced data sets [J]. Computer Engineering and Applications，2020，56 (24)：12-27 (in Chinese). [徐玲玲，迟冬祥. 面向不平衡数据集的机器学习分类策略 [J]. 计算机工程与应用，2020，56 (24)：12-27.]