DOI:10.13718/j.cnki.xsxb.2017.11.011

基于改进 Shapley 权力指数的特征选择算法①

巫 红 霞

镇江市高等专科学校 装备制造学院,江苏 镇江 212000

摘要:针对特征选择算法对不同类型的数据集性能不稳定的问题,提出一种基于概率模型与改进 Shapley 权力指数的通用特征选择算法.首先,计算特征对类簇表征与类簇判别的重要性值;然后,计算特征对类簇的不确定度;最终,合并特征的重要性值与不确定度,提取合适的特征.因为概率模型对数据类型、数据缺陷具有较好的鲁棒性,所以对不同的数据集获得了稳定、高性能的特征选择效果.基于人工合成数据与 benchmark 数据集的实验结果表明,本算法对不同的数据集保持了稳定的特征选择效果,优于其他算法.

关 键 词:概率模型;Shapley 权利指数;特征选择;鲁棒性;数据缺陷

中图分类号:TP391 **文献标志码**:A **文章编号**:1000-5471(2017)11-0062-10

随着互联网以及移动互联网的普及,大数据成为频繁出现的研究目标.传统的数据处理方法若针对大数据集则会导致极高的计算复杂度,因此在聚类等处理之前对大数据进行预处理,选出其中极具代表性的样本是一个可靠的思路[1-2].而特征选择则是数据挖掘领域一个重要的预处理方式,可以过滤冗余数据,并且提炼出重要的、具有代表性的少量特征用于之后的数据处理,从而提高数据分析的效率[3-5].

然而,目前主要的特征选择算法均针对某种数据集或者某些数据集具有较好的泛化效果,而无法对各种上下文数据集均保持稳定且较好的性能.文献[6-8]从不同角度提出了文档特征选择算法,此类算法对文档均取得了优异的效果,但对于其他格式的数据集性能较弱.文献[9-10]分别基于粗糙集与二阶 Hessian 矩阵提出了针对图像数据集的特征选择算法,与文档特征选择算法相似,此类算法仅对某个格式的数据集有较好的性能.

此外,有一类的特征选择方法基于特征加权,为每个特征分配一个实数权值估算对应特征的重要性,这些权值的估算方法主要有:Fisher 分数[11]、Shapley 指数[12]以及 diffusion 分数[13]等.Fisher 分数是一种有效的监督特征选择算法,它对特征独立地进行评估,因此对特征冗余的处理效果不佳;Shapley 权利指数则可估算特征的重要性,但此类方法并不稳定;diffusion 分数则通过 Markov 矩阵建立扩散距离,可获得数据结构的集合属性,此类算法对于高维数据效果较好,但并不适用于其他低维小数据集.

针对上述研究对不同类型数据集的处理性能不稳定的问题,本文开发了一种适用于不同数据类型的特征选择方法,对不同类型的数据集均可获得稳定、可靠的特征选择效果[14].概率模型可处理数据缺陷与冗余的问题,并且不受数据类型的影响,概率理论对于不同数据类型(数值、符号等)与不同的缺陷形式(模糊、不精确数据)均具有较好的建模效果.因此本文探讨特征选择的两个问题:特征重要性度量与不确定性,计算特征对类簇表征与判别的贡献.本文使用 Shapley 指数[15]选择目标特征,目标特征应当最小化类簇内距离,并且最大化类簇间距离.本文对 Shapley 指数进行扩展,实现对类簇的表征与判别,从而获得描述特征重要性的两个值,第一个值对应于特征对类簇表征的重要性,第二个值对应于类簇判别的重要性.

① 收稿日期:2016-07-10

作者简介:巫红霞(1977-),男,江苏句容人,实验师,主要从事数据挖掘研究.

1　概率理论

为了简化描述，假设 CS 表示用于类判别的 Shapley 特征重要性值，CR 表示用于类表征的 Shapley 特征重要性值，CU 表示不确定性准则. 概率理论对于不确定性信息的处理效果较好，假设 π 表示一个概率分布，在分类问题中，有限集 Ω 中的每个类 C_m 与一个概率分布 π_m 关联，假设 π_m 值的范围是[0，1].

表 1 所示是本文的一些变量定义，假设 x 是一个新元素，如果 $\pi_m(x)=1$，则说明 C_m 很可能是 x 的类，称为归一化概率分布；如果 $\pi_m(x)=0$，则说明 C_m 不可能是 x 的类.

表 1　本文预设的变量与参数

变量名	意义	变量名	意义
Ω	论域	L	X 子集
C_m，$m=1,\cdots,M$	Ω 中的第 m 个类	Var_n	第 n 个特征的概率分布方差
S_k，$k=1,\cdots,K$	第 k 个信息源	Var_L	复合特征集 L 的概率分布方差
$N(k)$	从信息源 k 中提取特征	μ	模糊隶属函数
$F_k=\{f_{k,1},\cdots,f_{k,N(k)}\}$	第 k 个信息源相关的特征集	v_n	第 n 个特征的 Shapley 值
π_{C_m}	类 C_m 的概率分布	η	概率分布 π 的均值
$U_{k,n(k)}$	S_k 源相关特征 $f_{n,k}$ 的不确定度	δ_1	阈值
$X=\{x_1,x_2,\cdots,x_B\}$	属于类 C_m 的观测量集合		

假设特征选择问题中每个数据源为 S_k，$k=1,\cdots,K$，提取的特征为 $f_{n,k}$，$n=1,\cdots,N(k)$，则可建立以下 M 个概率分布：$\pi^{C_1}_{f_{n,k}},\cdots,\pi^{C_m}_{f_{n,k}},\cdots,\pi^{C_M}_{f_{n,k}}$，该概率分布如图 1 所示. 假设 $\pi^{C_m}_{f_{n,k}}$ 是从源 S_k 提取的特征 $f_{n,k}$（属于 C_m 类）的数学形式.

对于特征 $f_{n,k}$，通过将 $\pi^{C_m}_{f_{n,k}}$，$m=1,\cdots,M$ 曲线分离可获得类别之间的判别信息. 因为一些数据源或特征是冗余信息，此类冗余特征导致类别曲线重叠，由此增加了类簇判别程序的复杂度.

假设 $\pi_{C_1},\cdots,\pi_{C_m},\cdots,\pi_{C_M}$ 分别表示类簇 $C_1,\cdots,C_m,\cdots,C_M$ 的概率模型，数据源设为 S_k，特征设为 $f_{n,k}$，图 2 所示是相应的一个重叠实例.

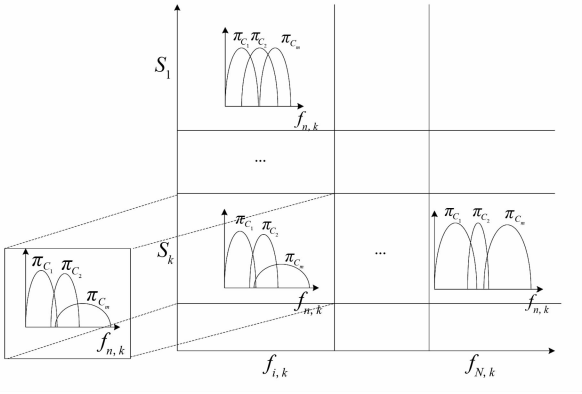

图 1　概率模型实例

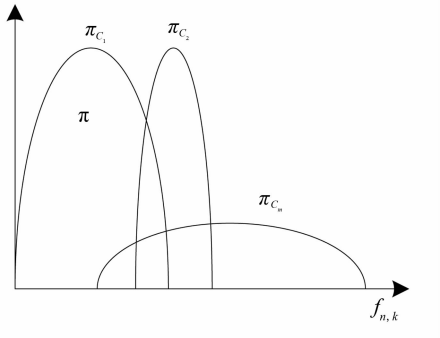

图 2　与 S_k，$f_{n,k}$ 对应的类模型重叠示例

为了评估 π_{C_m} 是否较好地表示 C_m 类，本文考虑基于概率理论的评价方法. 本文评价方法主要考虑了信息不确定度.

下文描述了概率分布“非特殊性”评估的重要性. “非特殊性”定义为：给定一个有限集 Ω 内的概率分布 $\pi(\cdot)$，定义以下 3 个知识：

1）完全的知识：对于任意给定的 x 存在 C_m，$\pi_{C_m}(x)=1$，则 $\forall C_t\neq C_m$，$\pi_{C_t}(x)=0$.

2）总忽略度：$\forall C_m\in\Omega$，$\pi_{C_m}(x)=1$（Ω 中所有类簇包含 x 的总可能性）.

3）不确定概率分布：在该情况下，概率分布在其知识中具有不确定性，可使用 U-uncertainty 度[16] 计算信息的不确定性.

图 3 所示是数据源 S_k 中特征 $f_{n,k}$ 的不确定度 $U_{k,n(k)}$.

假设一个有限集 $X=\{x_1,x_2,\cdots,x_B\}$ 属于 C_m 类，为了计算给定概率分布的不确定度，定义为：

$$\pi_{C_m}: X \to [0, 1] x \to \pi_{C_m}(x) \tag{1}$$

需要定义一个质量函数

$$\sigma: [1, \cdots, B] \longrightarrow [1, \cdots, B] b \longmapsto \sigma(b) \tag{2}$$

将 X 观测量 x_b，$b=1,\cdots,B$ 的概率值降序排列为：$\pi_{C_m}(x_{\sigma(1)}) \geqslant \pi_{C_m}(x_{\sigma(2)}) \geqslant \cdots \geqslant \pi_{C_m}(x_{\sigma(B)})$.

π_{C_m} 的 U-uncertainty 值 $U_{k,n(k)}$ 定义为：

$$U_{k,n(k)}(\pi) = \sum_{b=2}^{B}(\pi_{C_m}(x_{\sigma(b)} - \pi_{C_m}(x_{\sigma(b+1)}))\log_2(b)) + [1 - \pi_{C_m}(x_{\sigma(B)})]\log_2(B) \tag{3}$$

式中 $\pi_{C_m}(x_{\sigma(b+1)}) = 0$[16]，$U_{k,n(k)}$ 范围是$[0, \log_2(B)]$.

1) 如果 $U_{k,n(k)} = 0$，则表示完全知识的情况(不确定性为0)；

2) 如果 $U_{k,n(k)} = \log_2(B)$，则表示全部忽略的概率分布.

$U_{k,n(k)}$ 值可评价特征 $f_{n(k),k}$ 对类 C_m 的表征能力值，当特征的不确定性接近 $\log_2(B)$，则将该特征考虑为一个噪声源，并放弃该特征.

2 Shapley 权力指数

Shapley[15] 基于合作博弈论提出了 Shapley 指数，它可以较好地描述不同特征的重要性，Shapley 指数表示了每个特征的全局重要性.

原 Shapley 权利指数基于模糊指数来表征类簇，如(4)式所示. 假设 X 是属于 C_m 类的观测量集合，L 是 X 的一个子集，文献[15]中第 n 个特征的 Shapley 指数如下定义：

$$\upsilon_n = \sum_{i=1}^{N-1}\gamma_i \sum_{L\subset X\backslash n,\ |L|=i}(\mu_{uL} - \mu_L) \tag{4}$$

$$\gamma_i = \frac{(N-i-1)!\,i!}{N!} = \frac{1}{C_i^{N-1}N} \tag{5}$$

其中 $|L|$ 是 L 的基数.

Shapley 指数值应当满足 $\sum_{n=1}^{N}\upsilon_n = 1$. Shapley 指数值接近1的特征重要性高于接近0的特征，将此值乘以一个因子 N，最终 $\sum_{n=1}^{N}N * \upsilon_n = N$，如果重要性指标高于1，则表示该特征的重要性高于平均值.

图4所示是计算 S_k 数据源中特征 $f_{n,k}$ 的 Shapley 指数 $\upsilon(S_k, f_{n,k})$. 在本文的特征选择策略中，为 Shapley 指数引入概率分布的散度参数来评估特征的重要性值，通过分布的方差获得概率分布的散度，定义如下：

设 $\pi = \{\pi_1, \pi_2, \cdots, \pi_B\}$ 是概率分布，特征 n 的分布方差 Var_n 定义为：

$$Var_n = \frac{\sum_{b=1}^{B}(\pi_b - \eta)^2}{B} \tag{6}$$

其中 η 表示分布 π 的均值.

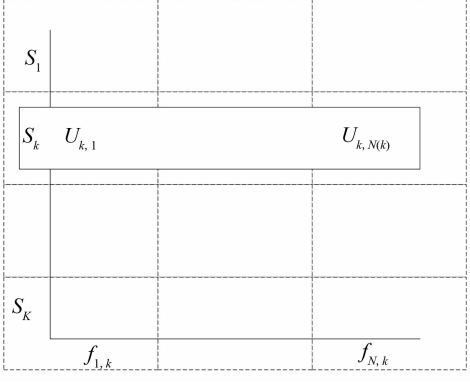

图3 数据源 S_k 中特征 $f_{n,k}$ 的不确定度 $U_{k,n(k)}$

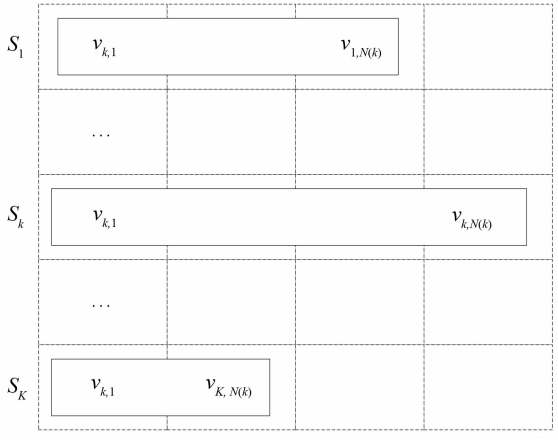

图4 Shapley 权利指标形式

3　本文方法

3.1　本文算法

本文特征选择策略不仅考虑了特征对于类簇判别的重要性，也考虑了特征对于类簇表征的重要性，本方法包括基于 Shapley 权利指数与分布不确定度的特征选择. 本方法使用归一化概率分布方差 Var 计算 Shapley 指数.

假设 π 是概率分布，特征 n 的 Shapley 指数采用归一化方差 Var 作为模糊度量，定义如下：

$$\upsilon_n = \sum_{i=0}^{N-1} \gamma_i \sum_{L \subset X \backslash n,\ |L|=i} (Var_{nL} - Var_L) \tag{7}$$

式中 Var_L 表示特征集 L 的概率分布方差，且满足下式：

$$Var_{nL} = Var(Var_n, Var_L) \tag{8}$$

Shapley 方法一般用于选择表征能力强的特征，而本文采用 Shapley 方法选择判别能力强的特征. 使用两个值代表特征重要性，第一个值表示特征对类簇表征的重要性，第二个值表示特征对类簇判别的重要性，分别如图 5 与图 6 所示.

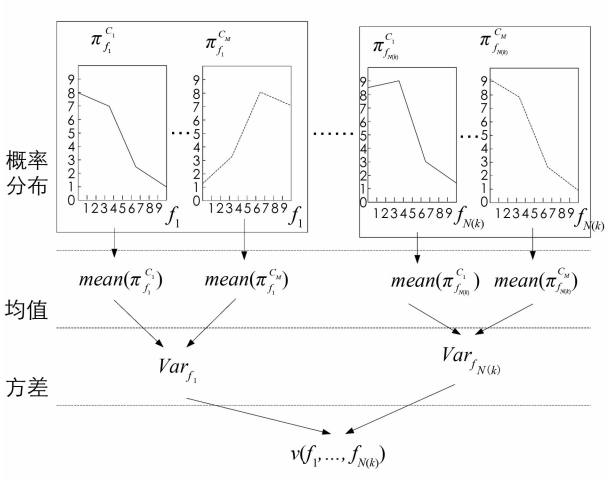

图 5　本文的特征类簇分离方法
（表示特征对类簇表征的重要性）

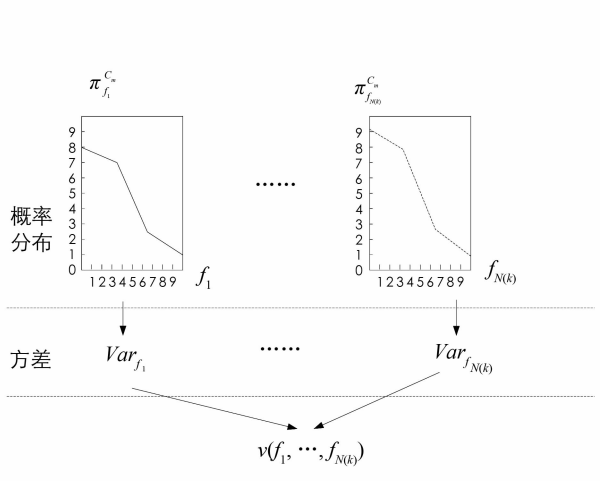

图 6　本文的特征类表征方法
（表示特征对类簇判别的重要性）

某个特征对类簇表征与类簇判别的重要性评价方法分别如下所示：

1）类簇表征的重要性评价：首先，计算每个特征 $f_{n,k}$ 的概率分布方差 Var_n；然后，使用 Shapley 方法（基于(7)式）定义特征重要性；最终，选择 M 个类中最小的一个特征，选择方法如下所示，算法 1 所示是具体算法实现：

$$\upsilon_n \geqslant 1,\ \upsilon_n \geqslant \delta_1 \tag{9}$$

式中 δ_1 是一个阈值，其计算方法如下：

$$\delta_1 = \frac{\max\limits_{n=1,\cdots,N}(\upsilon_n)}{\sum\limits_{n=1}^{N}\upsilon_n} \tag{10}$$

式中 N 是特征的数量.

2）类簇判别的重要性评价：计算信息源 S_k 中每个特征 f_n 的概率分布方差，首先计算每个概率分布的均值，根据 M 个均值计算其方差；最终，选出满足(9)，(10) 两式的特征.

例 1　给定一个论域包含 M 个类 $\Omega = \{C_1, C_2, \cdots, C_M\}$，假设 K 个信息源集合为 $S = \{S_1, S_2, \cdots, S_k\}$ 用于观察 Ω 中不同的类别. 从第 k 个信息源 S_k 提取 $N(k)$ 个特征 $F_k = \{f_{1,k}, f_{2,k}, \cdots, f_{N(k),k}\}$，$\pi_{f_{n,k}}^{C_m}$ 是类 C_m 拥有特征 $f_{n,k}$ 的概率分布.

算法 1　从 S_k 数据源中对特征采用类表征方法（假设类簇为 C_m）

输入：从 S_k 中提取的所有特征的概率分布集合(对于 C_m)：$\pi=\{\pi_{f_1}^{C_m},\pi_{f_2}^{C_m},\cdots,\pi_{f_{N(k)}}^{C_m}\}$.

输出：布尔型值，如果特征被类表征方法选择，则 $TR=1$；否则 $TR=0$.

步骤1：根据(6)式计算每个特征 $f_{n,k}$ 的概率分布.

步骤2：基于(7)式计算每个特征 $f_{n,k}$ 的重要性 v_n.

步骤3：假设从(10)式获得的阈值为 δ_1，如果对于某个类，满足 $v_n\geqslant 1$ 并且 $v_n\geqslant\delta_1$，则返回1；否则返回0.

算法2 从源 S_k 对特征进行类簇分离.

输入：从 S_k 提取的所有特征概率分布集合(对于 C_m)：$\pi=\{\pi_{f_1}^{C_m},\pi_{f_2}^{C_m},\cdots,\pi_{f_{N(k)}}^{C_m}\}$.

输出：布尔型值，如果特征被类表征方法选择，则 $TS=1$；否则 $TS=0$.

步骤1：计算信息源 S_k 中每个类(共 M 个类)所提取特征 f_n 的所有概率分布均值.

步骤2：计算 S_k 中所提取每个特征 f_n 均值向量的方差.

步骤3：基于(7)式(Shapley方法)计算每个特征 f_n 的重要性 v_n.

步骤4：假设(10)式获得的阈值为 δ_1，如果对于某个类，满足 $v_n\geqslant 1$ 并且 $v_n\geqslant\delta_1$，则返回1；否则返回0.

算法3 合并 Shapley 重要性值与概率分布不确定性值，选择特征 f_n.

输入：TR 值与 TS 值

输出：布尔型值，如果特征被选择，则 $TS=1$；否则 $TS=0$.

步骤1：验证特征 f_n 是否被类表征方法或类簇分离方法选择：

1) 如果 $TR=1$ 与 $TS=0$，则转至步骤2；

2) 如果 $TR=0$ 与 $TS=1$，则转至步骤5；

3) 如果 $TR=0$ 与 $TS=0$，结束迭代并返回0；

4) 如果 $TR=1$ 与 $TS=1$，结束迭代并返回1；

步骤2：计算类表征方法选择特征的概率分布Shapley值$\{v_{f_n}^{C_{m_1}},v_{f_n}^{C_{m_2}},\cdots,v_{f_n}^{C_{m_i}}\}$，其中$\{m_1,m_2,\cdots,m_i\}\subset\{1,\cdots,M\}$. 因此选出满足 $v_{f_n}^{C_{m_i}}\geqslant\delta_1$ 与 $v_{f_n}^{C_{m_i}}\geqslant 1$ 的所有特征 f_n.

步骤3：使用(3)式计算 f_n 概率分布的不确定度(对于 C_{m_i} 个类).

步骤4：假设(10)式获得的阈值为 δ_1，如果对于某个类，满足 $v_n^{C_{m_i}}<\delta_1$，则返回1；否则返回0.

步骤5：基于(3)式(所有分类的概率分布)计算 f_n 的不确定度 U_n；

步骤6：使用步骤4中的阈值 δ_1，如果所有分类满足 $U_n<\delta_1$，则返回1，否则返回0.

3.2 算法复杂度分析

本算法的复杂度依赖 Shapley 指数与概率分布不确定性度量的复杂度：概率分布不确定度的处理时间较低，其复杂度是 $O(\log_2(n))$；Shapley 算法的处理时间与特征数量成比例，其复杂度是 $O(2^n)$，但是本文使用文献[17]的母函数计算 Shapley 权利指数，从而降低了复杂度.

4 实验结果与分析

4.1 人工合成数据实验

首先，使用合成数据来评估本文方法的性能. 考虑的数据形式为概率分布，每个类假设为一个相关、非相关特征的混合. 给定2个类组成的论域，用3个信息源 $S=S_1,S_2,S_3$ 描述 Ω，假设从每个源 S_k 提取5个特征 f_1,f_2,f_3,f_4,f_5，则：f_4 较好地代表 C_2；S_1 中 f_3 特征对类的判别能力较强；f_2 对所有数据源均具有较好的表征与判别能力；另外两个特征 f_1 与 f_5 对于所有数据源均没有表征与判别能力.

从 S_k 中提取的每个特征包含4个样本，根据(10)式获得一个特征概率分布不确定度，计算阈值 δ_1：$\delta_1=\dfrac{\log_2(4)}{1.6}=1.25$.

考虑表2所示的概率分布：估算不同特征的重要性值以实现类簇表征与类簇分离，表3－5分别列出了不同概率分布的不确定性值，根据这3个表的结果评价本文算法.

表 2　类与特征的条件概率分布

类		f_1	f_2	f_3	f_4	f_5
C_1	S_1	1	0.69	0.05	0.91	0.83
		1	0.60	0.73	0.94	0.72
		1	0.21	0.76	0.09	0.42
		1	0.11	0.02	0.05	0.30
	S_2	1	0.70	0.10	0.88	0.78
		1	0.75	0.52	0.92	0.69
		1	0.09	0.68	0.07	0.35
		1	0.14	0.09	0.06	0.27
	S_3	1	0.69	0.20	0.90	0.74
		1	0.71	0.60	0.89	0.66
		1	0.18	0.66	0.10	0.37
		1	0.16	0.13	0.03	0.25
C_2	S_1	1	0.02	0.09	0.20	0.82
		1	0.20	0.31	0.31	0.77
		1	0.90	0.73	0.52	0.42
		1	0.81	0.69	0.68	0.24
	S_2	1	0.10	0.15	0.19	0.79
		1	0.22	0.35	0.33	0.65
		1	0.96	0.70	0.56	0.38
		1	0.87	0.60	0.60	0.16
	S_3	1	0.08	0.01	0.21	0.73
		1	0.20	0.29	0.29	0.70
		1	0.87	0.68	0.54	0.40
		1	0.91	0.65	0.68	0.22

特征 1：从表 2 中可看出，f_1 将忽略所有的条件概率分布，导致重要性值与不确定度较低. 详细分析：表 3 中 f_1 对类分离的重要性值等于 0；表 4 显示特征 1 对类表征的重要性值等于 0；特征 1 的不确定度等于最大值，即 $\log_2(B)(\log_2(B)=2)$，其中 B 是每个类中观测量的次数($B=4$)，因此，该特征并未被本方法选择.

特征 2：图 7 为特征 f_2(从每个源 S_k 中提取) 的两个类概率分布曲线，从图 7 可看出，f_2 具有区分不同类簇的能力以及特征 C_1 类的能力，对 S_2 与 S_3 两个源效果尤佳. 如果可以清晰地区分两个类所属概率的两个区域，则认为该概率分布可较好地代表一个类，如图 8 所示：弱区域的类簇所属概率接近 0；强区域的所属概率接近 1.

表 3 是特征 2 对于类分离的重要性值，其结果远大于 δ_1(对于所有的信息源)，因此该特征可以分离类簇. 表 4 是特征 2 对于表征 C_2 类的重要性值(对于所有信息源)，从中可看出，该特征仅能代表 C_2 类. 表 5 所示是特征 2 对于 C_1 类的不确定度(对于所有的信息源)，其值高于 δ_1，因此，该特征对 C_2 类有重要的类簇区别能力. 因此，所有信息源均应当保留该特征.

特征 3：从图 9 可看出，特征 3 可区分不同的类簇. 表 3 表明特征 3 对类分离的重要性值高于 δ_1(对于 S_1 源)；表 4 表明特征 3 对 C_2 类表征的重要性值较低(对于所有的信息源)；表 5 表明特征 3 对 C_2 类的不确定度较高(对于所有的信息源). 因此，仅有 S_1 源应当保留特征 3.

特征 4：从图 10 可看出，特征 4 可较好代表 C_1 类，并且该特征可较好地区分类簇. 表 3 显示特征 4 对类簇判别的重要性高于 δ_1(对于 S_1 源)；表 4 显示特征 4 对 C_2 类表征的重要性较低；表 5 显示特征 4 对 C_1 类的类簇判别能力较差. 因此，所有信息可以保留特征 4.

特征 5：从图 11 可看出，特征 5 不能较好地代表 C_2 类. 因此本方法过滤掉特征 5.

表3 区分类簇的特征重要性值(大于 δ_1)

特征	S_1	S_2	S_3
f_1	0	0	0
f_2	2.534 2	4.228 2	3.902 6
f_3	1.098 6	0.056 6	0.230 7
f_4	1.362 3	0.642 8	0.855 7
f_5	0.004 9	0.072 4	0.011 0
δ_1	0.506 8	0.845 6	0.780 5

表4 代表类簇的特征重要性值(大于 δ_1)

特征	S_1		S_2		S_3	
	C_1	C_2	C_1	C_2	C_1	C_2
f_1	0	0	0	0	0	0
f_2	0.450 4	3.322 7	0.990 3	3.636 6	0.580 9	3.295 9
f_3	1.340 6	0.800 6	0.366 1	0.449 4	0.947 4	0.967 9
f_4	2.892 4	0.299 6	3.307 2	0.243 0	3.178 6	0.312 4
f_5	0.316 6	0.577 1	0.336 4	0.670 9	0.293 8	0.423 7
δ_1	0.578 5	0.666 5	0.661 4	0.727 3	0.635 7	0.659 2

表5 概率分布不确定性值(低于 δ_2)

特征	S_1		S_2		S_3	
	C_1	C_2	C_1	C_2	C_1	C_2
f_1	2	2	2	2	2	2
f_2	1.388 5	1.135 3	1.319 2	1.120 2	1.441 7	1.200 2
f_3	1.247 5	1.448 7	1.585 7	1.467 0	1.162 6	1.463 8
f_4	1.103 4	1.424 3	1.105 8	1.631 9	1.160 9	1.436 8
f_5	1.430 2	1.475 3	1.446 8	1.358 7	1.500 2	1.565 3

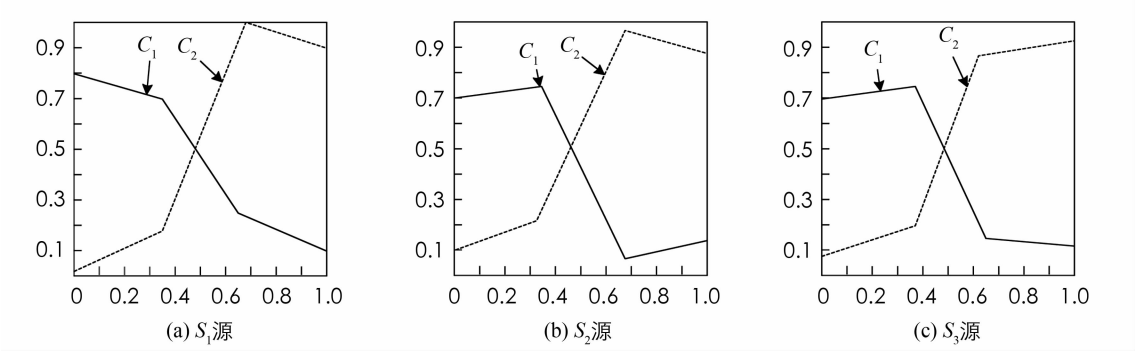

图7 特征2的概率分布曲线

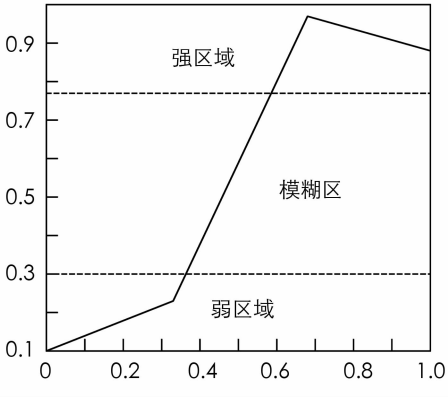

图8 类簇所属概率分布的区域

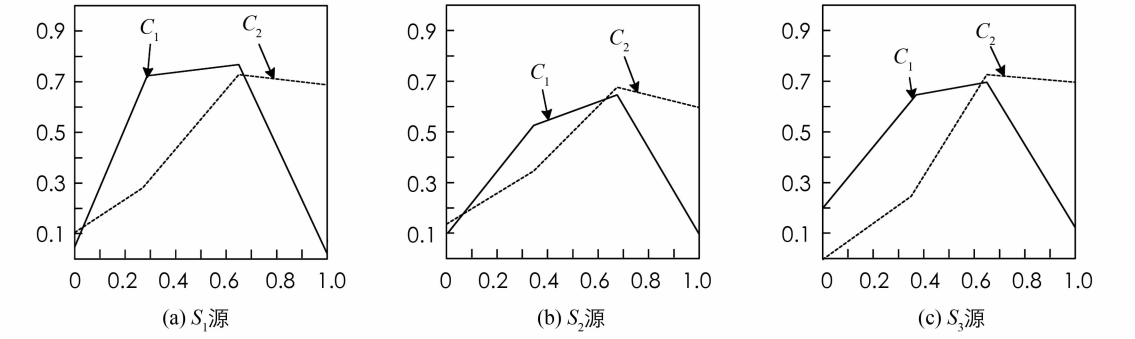

图 9　特征 3 的概率分布曲线

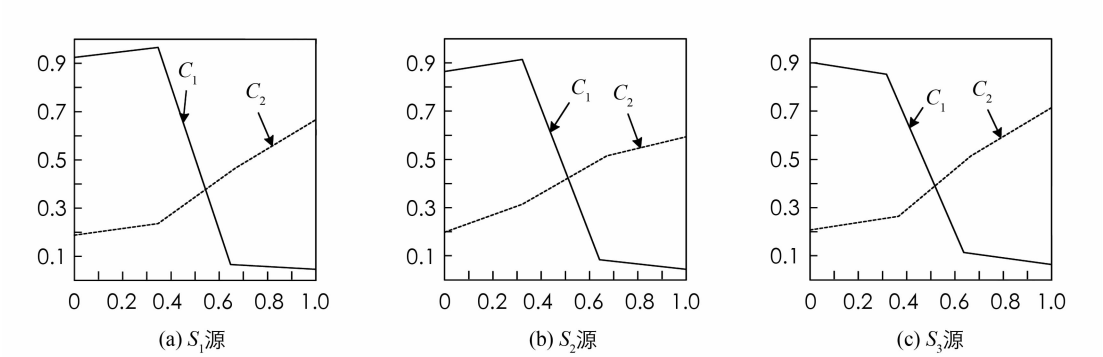

图 10　特征 4 的概率分布曲线

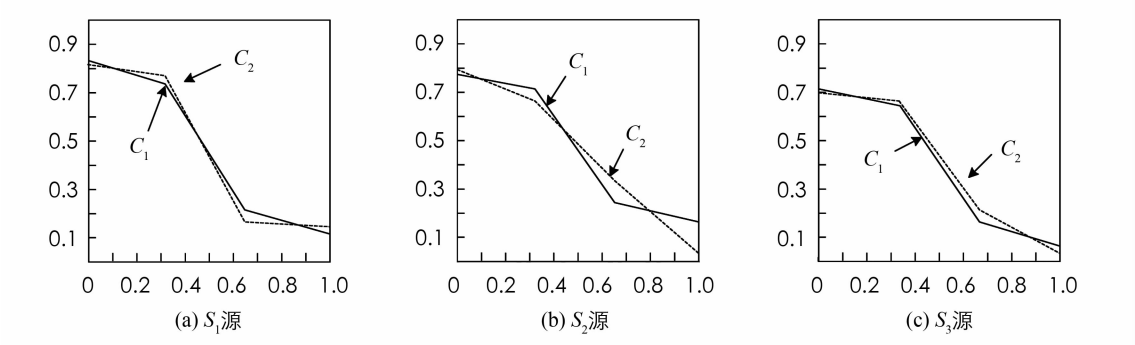

图 11　特征 5 的概率分布曲线

4.2　Benchmark 数据集

本文采用 6 个 benchmark 数据集[18]：Glass，Liver-disorders，IRIS，SVMguide3，Pima Indians Diabetes 与 Vehicle. 实验将特征选择算法与 SVM 分类器结合，测试特征选择算法的效果. 本文实验选择 4 个性能较好的特征选择算法作为对比实验：文献[19]是一种基于类簇判别信息最大化的特征选择算法 FSDMED；文献[20]是一种基于依赖度最大化的特征选择算法 FSDM；文献[21]是一种基于改进最大相关最小冗余判据的暂态稳定评估特征选择方法 MRMR；文献[22]是一种基于随机森林的封装式特征选择算法 RFFS.

表 6 所示是本文特征选择算法的处理结果，可见本方法对于 IRIS，Glass，SVMguide3，Diabetes，Vehicle 5 个数据集的分类正确率均排名第二，仅对 Liver-disorders 算法的分类正确率略低于 FSDMED 与 RFFS 算法，综合可看出，本算法的分类正确率较优.

表6 使用不同方法获得的特征选择结果统计

算法	IRIS		Glass		Liver-disorders		SVMguide3		Diabetes		Vehicle	
	特征数量/个	分类正确率/%	特征数量/个	分类正确率/%	特征数量/个	分类正确率/%	特征数量/个	分类正确率/%	特征数量/个	分类正确率/%	特征数量/个	分类正确率/%
所有特征	4	96.2	9	95.3	6	67.3	21	83.2	8	78.1	18	82.8
FSDMED	2	81.3	6	93.9	5	68.4	13	78.3	4	75.6	7	74.7
FSDM	2	95.3	7	93.9	3	60.5	4	79.6	4	77.8	10	70.6
MRMR	2	81.3	4	91.5	3	64.8	11	80.4	6	76.4	10	74.2
RFFS	2	81.3	4	93.4	4	69.3	4	76.2	4	77.2	10	72.8
本文方法	2	95.3	7	94.4	5	66.3	17	82.3	4	77.3	8	75.7

5 结束语

本文针对特征选择算法对不同类型的数据集性能不稳定的问题，提出一种基于概率模型与改进Shapley权力指数的通用特征选择算法. 因为概率模型对数据类型、数据缺陷具有较好的鲁棒性，所以对不同的数据集获得了稳定、高性能的特征选择效果. 数据实验结果表明，本算法对于不同类型的数据集均可保持稳定、高效的特征选择效果，不受数据格式或者数据缺陷的限制.

参考文献：

[1] 李先锋，朱伟兴，纪 滨，等. 基于图像处理和蚁群优化的形状特征选择与杂草识别[J]. 农业工程学报，2010，26(10)：178－182.

[2] 王 璐，邱桃荣，何 妞，等. 基于粗糙集和蚁群优化算法的特征选择方法[J]. 南京大学学报(自然科学版)，2010，46(5)：487－493.

[3] 徐峻岭，周毓明，陈 林，等. 基于互信息的无监督特征选择[J]. 计算机研究与发展，2012(02)：372－382.

[4] 朱鹏飞，胡清华，于达仁. 基于大间隔粒计算的特征选择[J]. 重庆邮电大学学报(自然科学版)，2010，22(5)：641－647.

[5] 朱颢东，周 姝，钟 勇. 结合ODF和辨识集的特征选择[J]. 重庆邮电大学学报(自然科学版)，2010，22(1)：94－98.

[6] BASU T，MURTHY C A. A Feature Selection Method for Improved Document Classification[M]// Advanced Data Mining and Applications. Berlin：Springer，2012：296－305.

[7] 阿力木江·艾沙，吐尔根·依布拉音，库尔班·吾布力，等. 基于类别分布差异和特征熵的维吾尔语文本特征选择[J]. 计算机应用研究，2013，30(10)：2958－2961.

[8] 张延祥，潘海侠. 一种基于区分能力的多类不平衡文本分类特征选择方法[J]. 中文信息学报，2015，29(4)：111－119.

[9] 胡 燕，王慧琴，秦薇薇，等. 基于粗糙集的火灾图像特征选择与识别[J]. 计算机应用，2013，33(3)：704－707.

[10] 史彩娟，阮秋琦，刘 健，等. 基于Hessian半监督特征选择的网络图像标注[J]. 计算机应用研究，2015，32(2)：606－608.

[11] SINGH B，SANKHWAR J S，VYAS O P. Optimization of Feature Selection Method for High Dimensional Data Using Fisher Score and Minimum Spanning Tree[C]// India Conference 2014. New York：IEEE Computer Society Press，2014：1－6.

[12] SASIKALA S，APPAVU A B S，GEETHA S. An Efficient Feature Selection Paradigm Using PCA－CFS－Shapley Values Ensemble Applied to Small Medical Data Sets[C]// Fourth International Conference on Computing，Communications and Networking Technologies (ICCCNT). New York：IEEE Computer Society Press，2013：1－5.

[13] YU J. Cartoon Features Selection Using Diffusion Score[J]. Signal Processing，2013，93(6)：1510 - 1520.

[14] 崔 超，吴 双，张宪忠，等. 基于贝叶斯概率理论的防火墙技术研究[J]. 北京理工大学学报，2012，32(8)：801－804.

[15] GRABISCH M. The Representation of Importance and Interaction of Features by Fuzzy Measures[J]. Pattern Recognition Letters，1996，17(6)：567－575.

[16] HIGASHI M，KLIR G J. Measures of Uncertainty and Information Based on Possibility Distributions[J]. International

Journal of General Systems, 1982, 9(1): 43－58.
[17] BILBAO J M, FERNÁNDEZ J R, LOSADA A J, et al. Generating Functions for Computing Power Indices Efficiently[J]. Top, 2000, 8(2): 191－213.
[18] AMARNATH B, BALAMURUGAN S A A. Review on Feature Selection Techniques and Its Impact for Effective Data Classification Using UCI Machine Learning Repository Dataset[J]. Journal of Engineering Science & Technology, 2016, 11(11): 1639－1646.
[19] MURATA M, UCHIMOTO K, UTIYAMa M, et al. Using the Maximum Entropy Method for Natural Language Processing: Category Estimation, Feature Extraction, and Error Correction[J]. Cognitive Computation, 2010, 2(4): 272－279.
[20] Gretton A. Feature Selection via Dependence Maximization[J]. Journal of Machine Learning Research, 2012, 13(1): 1393－1434.
[21] 李　扬，顾雪平. 基于改进最大相关最小冗余判据的暂态稳定评估特征选择[J]. 中国电机工程学报，2013(34): 179－185.
[22] 姚登举，杨　静，詹晓娟. 基于随机森林的特征选择算法[J]. 吉林大学学报(工学版)，2014，44(1): 137－141.

Probabilistic Model and Improved Shapley Power Index Based General Feature Selection Algorithm

WU Hong-xia

School of Equipment Manufacturing, Zhenjiang College, Zhenjiang Jiangsu 212000, China

Abstract: Aimed at the problem that feature selection algorithms show unstable performance to different types of datasets, a probabilistic model and improved Shapley power index based general feature selection algorithm. Firstly, the importance values of features to class representation and discrimination are computed. Secondly, the uncertainness values of features to classes are computed. Lastly, the importance values and uncertainness values are merged to abstract suitable feature. Because probabilistic model performs good robustness to data types and data imperfection, stable feature selection results to different datasets are got with high performance. Synthetic datasets and benchmark datasets based experiments results show that the proposed algorithm show stable feature selection effect to different datasets and is better than the other algorithms.

Key words: probabilistic model; Shapley power index; feature selection; robustness; data imperfection

责任编辑　张　枸