DOI: 10.11992/tis.201905051
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20190828.1806.010.html

面向一致性样本的属性约简

高媛[1], 陈向坚[1], 王平心[2], 杨习贝[1]
(1. 江苏科技大学 计算机学院, 江苏 镇江 212003; 2. 江苏科技大学 理学院, 江苏 镇江 212003)

摘 要: 作为粗糙集理论的一个核心内容, 属性约简致力于根据给定的约束条件删除数据中的冗余属性。基于贪心策略的启发式算法是求解约简的一种有效手段, 这一手段通常使用数据中的全部样本来度量属性的重要度从而进一步得到约简子集。但实际上, 不同样本对于属性重要度计算的贡献是不同的, 有些样本对重要度贡献不高甚至几乎没有贡献, 且当数据中的样本数过大时, 利用全部样本进行约简求解会使得时间消耗过大而难以接受。为了解决这一问题, 提出了一种基于一致性样本的属性约简策略。具体算法大致由 3 个步骤组成, 首先, 将满足一致性原则的样本挑选出来; 其次, 将这些选中的样本组成新的决策系统; 最后, 利用启发式框架在新的决策系统中求解约简。实验结果表明: 与基于聚类采样的属性约简算法相比, 所提方法能够提供更高的分类精度。

关键词: 属性约简; 分类精度; 聚类; 一致性样本; 集成; 启发式算法; 邻域粗糙集; 多准则
中图分类号: TP181 **文献标志码**: A **文章编号**: 1673−4785(2019)06−1170−09

中文引用格式: 高媛, 陈向坚, 王平心, 等. 面向一致性样本的属性约简 [J]. 智能系统学报, 2019, 14(6): 1170–1178.
英文引用格式: GAO Yuan, CHEN Xiangjian, WANG Pingxin, et al. Attribute reduction over consistent samples[J]. CAAI transactions on intelligent systems, 2019, 14(6): 1170–1178.

Attribute reduction over consistent samples

GAO Yuan[1], CHEN Xiangjian[1], WANG Pingxin[2], YANG Xibei[1]
(1. School of Computer, Jiangsu University of Science and Technology, Zhenjiang 212003, China; 2. School of Science, Jiangsu University of Science and Technology, Zhenjiang 212003, China)

Abstract: As one of the key topics in rough sets theory, attribute reduction aims to remove redundant attributes in a data set according to a given constraint condition. Based on greedy strategy, the heuristic algorithm is an effective strategy in finding reductions. Traditional heuristic algorithms usually need to scan all samples in a data set to compute the significance of attributes to further obtain a reduction. However, different samples have different contributions to the process of computing significance. Some samples have little relation to the significance, and some even have no contribution to the significance. Therefore, scanning all samples to compute reductions may require too much time, and the time may be unacceptable if the number of samples is too large. To fill such a gap, we have proposed an attribute reduction algorithm with sample selection, which is based on the consistent principle. The algorithm is composed of three stages. First, the samples that satisfy the consistent principle were selected; second, a new decision system was constructed with these selected samples; finally, reductions were derived from the heuristic algorithm over the new decision system. Experimental results demonstrated that, compared with the attribute reduction algorithm with a cluster-based sample selection, our new algorithm can offer better classification accuracy.

Keywords: attribute reduction; classification accuracy; clustering; consistent samples; ensemble; heuristic algorithm; neighborhood rough set; multiple criteria

收稿日期: 2019−05−27. **网络出版日期**: 2019−08−29.
基金项目: 国家自然科学基金项目 (61572242, 61503160); 江苏省研究生科研创新计划项目 (KYCX19_1697).
通信作者: 杨习贝. E-mail: jsjxy_yxb@just.edu.cn.

粗糙集[1-2] 是 Pawlak 提出的一种用以刻画不确定性的建模方法。由于经典粗糙集所使用的等价关系仅仅适用于符号型数据, 为了弥补这一不

足, 涌现出了一批可以处理复杂类型数据的拓展粗糙集模型[3-5]。

众所周知, 无论是在经典粗糙集还是在众多拓展粗糙集研究中, 属性约简[6-10]一直扮演着核心角色。根据问题求解需求的不同, 属性约简可以使用不同的度量准则加以定义, 因此其具有丰富的解释与含义。例如近似质量可以用来度量数据中的确定性程度, 条件熵可以用来描述条件属性相对于决策属性的鉴别能力。属性约简, 就可以依据这些度量准则, 找到数据中的冗余属性并加以删除, 以达到满足度量准则所对应的约束条件。通过这一策略, 还能够使得后续的学习过程仅需在一部分属性上构建模型, 从而达到降低学习难度以及降低学习时间消耗的目的。

目前在粗糙集理论中, 穷举法与启发式方法被公认为是求解约简的两大类基本方法。然而很多穷举搜索与启发式搜索策略都将数据集中的所有样本视为同等重要, 当样本量非常巨大时, 这会带来较低的约简求解效率。为解决这一问题, 已有众多学者将样本选择的概念引入到约简求解过程中。样本选择的概念最早是由 Hart 提出, 他提出了压缩近邻规则[11], 随后亦有学者提出了很多改进策略, 如缩减最近邻[12]、有序最近邻[13]和快速压缩最近邻[14]等。当涉及约简求解的问题时, 已有学者[15-19]关注到不同的样本对属性重要度评价的贡献是不同的, 如王熙照等[15]提出的 *K*-means 样本选择算法将远离类簇中心点的样本视为重要的; 随后 Xu 等[19]将这种方法应用到多标记数据的维度压缩问题中。但他们在追求时间效率的同时忽略了约简在测试集上的分类性能。

基于以上分析, 笔者提出了一种基于样本一致性原则的样本选择算法 (以下简称为一致性采样), 一致性采样的主要思想为: 1) 给定一个样本, 找到距离自己最近的邻居; 2) 判断这一邻居是否与自身属于同一类别, 若属于同一类别, 则将该样本选中; 3) 最后将所有选中的样本组成一个新的数据集。

1　基础知识

在粗糙集理论中, 一个决策系统可表示为一个二元组 DS=<*U*, AT∪*D*>, *U* 是所有样本构成的非空有限集合, 即论域; AT 是所有条件属性的集合; *D* 是决策属性的集合且 AT∩*D*=*Ø*。当 *D* 的取值都为离散型时, 可得 *U*/IND (*D*)={$X_1, X_2, \cdots, X_q$}, 其表示根据决策属性 *D* 所诱导出的论域上的划分, 对于$\forall X_p \in U/\mathrm{IND}(D)$, X_p 表示第 *p* 个决策类, 其中 $[x]_D$ 表示与 *x* 属于同一个决策类的样本的集合。

定义 1　给定一个决策系统 DS, $\forall A \subseteq \mathrm{AT}$, 则邻域关系定义为

$$N_A = \{(x,y) \in U \times U : r(x,y) \leqslant \delta\} \tag{1}$$

其中, $\forall x, y \in U$, $r(x, y)$ 表示样本 *x* 与 *y* 之间的欧氏距离, $\delta>0$ 称为邻域半径。

则由式 (1), 容易得到关于 *A* 样本 *x* 的邻域:

$$\delta_A(x) = \{y | r(x,y) \leqslant \delta\} \tag{2}$$

定义 2　给定一个决策系统 DS, *U*/IND(*D*)={$X_1, X_2, \cdots, X_q$}, $\forall A \subseteq \mathrm{AT}$, $\forall X_p \in U/\mathrm{IND}(D)$, X_p 关于 *A* 的下近似集和上近似集分别定义为

$$\underline{N_A} X_p = \{x \in U | \delta_A(x) \subseteq X_p\} \tag{3}$$

$$\overline{N_A} X_p = \{x \in U | \delta_A(x) \cap X_p \neq \phi\} \tag{4}$$

定义 3[20]　给定一个决策系统 DS, *U*/IND(*D*)={$X_1, X_2, \cdots, X_q$}, $\forall A \subseteq \mathrm{AT}$, *D* 关于 *A* 的近似质量定义如下:

$$\gamma(A,D) = \frac{\left| \bigcup_{p=1}^{q} \underline{N_A} X_p \right|}{|U|} \tag{5}$$

其中|*U*|表示集合 *X* 的基数。

显然 $0 \leqslant \gamma(A, D) \leqslant 1$ 成立。$\gamma(A, D)$ 表示根据条件属性 *A*, 那些确定属于某一决策类别的样本占总体样本的比例。

条件熵是属性约简中另外一种常用的度量准则, 它能反映条件属性相对于决策属性的鉴别能力。根据不同的需求, 很多学者设计并定义了多种形式的条件熵[21-25], 一种经典的形式可定义为:

定义 4[25]　给定一个决策系统 DS, $\forall A \subseteq \mathrm{AT}$, *D* 关于 *A* 的条件熵定义如下:

$$\mathrm{ENT}(A,D) = -\frac{1}{|U|} \sum_{x \in U} \log \frac{|\delta_A(x) \cap [x]_D|}{|U|} \tag{6}$$

显然, 条件熵的值越小, 条件属性相对于决策属性鉴别能力越大。

2　属性约简

属性约简作为粗糙集领域的一个核心内容, 其主要目的是根据某一给定的约束条件来去除全部属性中的冗余、不相关的属性。目前求解约简的常用策略包括穷举式算法和启发式算法。虽然前者可以得到一个数据中的所有约简, 但当数据维数急剧升高时, 它的时间消耗随之增大, 过大的时间消耗导致算法并不适用于处理实际问题。与穷举式算法不同, 启发式算法因其较高的时间效率得到了众多学者的青睐, 它运用贪心策略, 每次迭代过程中其将属性重要度最大的属性加入到

潜在约简集合中，直至满足约束条件则终止算法。因此，接下来需要给出属性重要度的表达式。

$$\mathrm{Sig}_{\gamma}(a_i,A,D)=\gamma(A\cup\{a_i\},D)-\gamma(A,D) \tag{7}$$

$$\mathrm{Sig}_{\mathrm{ENT}}(a_i,A,D)=\mathrm{ENT}(A,D)-\mathrm{ENT}(A\cup\{a_i\},D) \tag{8}$$

对于近似质量 (利用式 (7) 来计算属性重要度)，a_i 的重要度越大，表示 a_i 对其值的提升效果越明显。而对于条件熵而言 (利用式 (8) 来计算属性重要度)，a_i 的重要度越大，则表示 a_i 对其值的降低效果越明显。

定义 5　给定一个决策系统 DS，$\forall A\subseteq \mathrm{AT}$，$A$ 是 DS 中的一个关于 φ 的约简当且仅当：

(1) $\varphi(A, D)$ 满足约束条件；

(2) $\forall A'\subseteq A$，$\varphi(A', D)$ 不满足约束条件。

在定义 5 中，“φ”可以是“γ”也可以是“ENT”。当 $\varphi=\gamma$ 时，约束条件可以定义为 $\gamma(A, D)\geqslant\gamma(\mathrm{AT}, D)$，它表示利用约简 A 中的属性计算的近似质量值应不低于利用全部属性 (AT) 计算的近似质量值；而当 $\varphi=\mathrm{ENT}$ 时，约束条件则定义为 $\mathrm{ENT}(A, D)\leqslant \mathrm{ENT}(\mathrm{AT}, D)$，它表示利用约简 A 中的属性计算的条件熵值应不高于利用全部属性 (AT) 计算的条件熵值。

算法 1 给出了一个求解定义 5 所示 φ 约简的启发式框架型描述。

算法 1　启发式算法

输入　决策系统 DS=<U, AT∪D>，半径 δ

输出　一个关于 φ 的约简：A

1) 计算 $\varphi(\mathrm{AT}, D)$；

2) $A\leftarrow\phi$；

3)：

(1) $\forall a_i\in \mathrm{AT}-A$，计算属性 a_i 的重要度 $\mathrm{Sig}_{\varphi}(a_i, A, D)$；

(2) 选出一个重要度最大的属性 b，令 $A=A\cup\{b\}$；

(3) 计算 $\varphi(A, D)$；

4) 如果 $\varphi(A, D)$ 满足约束条件，则直接转至 5)

5) 返回 A

算法 1 的时间复杂度为 $O(|U|^2\cdot|\mathrm{AT}|^2)$，其中 $|U|$为论域中样本数目，|AT|为条件属性数目。

算法 1 是基于单准则设计的，而运用基于单准则的算法得到的约简往往仅能满足一个约束条件，而此约简结果可能无法满足其他约束条件。例如：仅使用近似质量得到的约简满足自身的约束条件，但它往往无法同时提高分类能力，这主要是因为近似质量是用来描述数据中的确定性程度，而与数据的分类关系不大。为了弥补这一局限，近年来，根据多准则设计的约简也开始受到学者的重视。如 Yang 与 Yao[26] 提出的集成选择器极大地丰富了约简的求解策略；随后，Li 等[27] 利用这一思想设计了基于调和平均的多准则属性约简。Liu 等[21] 进一步利用集成思想，将其扩展应用到半监督领域中。一般来说，多准则启发式算法可由算法 2 实现。

算法 2　多准则启发式算法

输入　决策系统 DS=<U, AT∪D>，半径 δ

输出　一个多准则约简：A

1) 计算 $\varphi_1(\mathrm{AT}, D)$, $\varphi_2(\mathrm{AT}, D)$, …, $\varphi_m(\mathrm{AT}, D)$；

2) $A\leftarrow\phi$；

3)

(1) **For** $1\leqslant k\leqslant m$

$\forall a_i\in \mathrm{AT}-A$，计算属性 a_i 的重要度 $\mathrm{Sig}_{\varphi_k}(a_i, A, D)$；选出重要度最大的属性 a_k^j；

End For

(2) 在 $\{a_1^j, a_2^j, \cdots, a_m^j\}$中选择一个出现频次最高的属性 b；

(3) $A=A\cup\{b\}$；

(4) 计算 $\varphi_1(A, D)$, $\varphi_2(A, D)$, …, $\varphi_m(A, D)$；

4) 如果对于任意的 $k(1\leqslant k\leqslant m)$，$\varphi_k(A, D)$ 满足约束条件，则直接转至步骤 5)；否则转至步骤 3)；

5) 返回 A。

算法 2 的时间复杂度为 $O(m\cdot|U|^2\cdot|\mathrm{AT}|^2)$。在每次迭代过程，3) 将 m 个准则下重要度最大的属性分别选择出来并记录每个属性出现的频次，选取频次最高的属性加入到潜在约简集合中：1) 如果出现频次最高的属性是唯一的，则直接将其加入到潜在约简集合中；2) 否则，出现频次最高的属性发生冲突情况，即两个或多个属性的频次同时达到最高，则需要进行选择，为了保证算法的稳定性，将位置靠前的属性加入到潜在约简集合中。

3　一致性采样约简

显然，第 2 节所示的两个算法都是基于扫描数据中的全部样本来实现的。但当数据体量较大时，这种求解策略的时间消耗较高。为了进一步压缩算法的时间消耗，可以将样本选择的方法引入到约简求解过程中。不同的样本选择方法会选取不同的样本，进而产生不同的分类效果。本文将一致性从样本间距离与样本的决策属性值角度出发，目的是使得算法可以利用选择出的样本获取更高的分类精度。算法大致分为两个主要部分：1) 要全部样本组成的决策系统上进行采样处

理以构建含有较少样本个数的新决策系统; 2) 随后, 将一致性采样的思想应用到属性约简的求解过程中。具体算法流程如下所示。

算法 3　一致性采样约简算法

输入　决策系统 DS=<U, AT∪D>, 半径 δ;

输出　一个约简 A。

1)

(1) 令 U'=Ø;

(2) $\forall y \in U$, 计算样本之间距离 $r(x, y)$;

(3) 对 $r(x, y)$ 进行排序;

(4) 取距离测试样本 y 最近的样本, 若二者的决策值相同, 则选中该测试样本并将其加入到 U'中;

(5) 将新选中的样本组成新的决策系统 DS′=<U', AT∪D>;

2) 在新的决策系统 DS′中计算 $\varphi_1(\mathrm{AT}, D)$, $\varphi_2(\mathrm{AT}, D), \cdots, \varphi_m(\mathrm{AT}, D)$;

3) $A \leftarrow$Ø;

4)

(1) **For** $1 \leqslant k \leqslant m$

$\forall a_i \in \mathrm{AT}-A$, 计算属性 a_i 的重要度 $\mathrm{Sig}_{\varphi_k}(a_i, A, D)$; 选出重要度最大的属性 a_k^j;

End For

(2) 在 $\{a_1^j, \cdots, a_m^j\}$中选择一个出现频次最高的属性 b;

(3) $A=A \cup \{b\}$;

(4) 计算 $\varphi_1(A, D), \varphi_2(A, D), \cdots, \varphi_m(A, D)$;

5) 如果对于任意的 $k(1 \leqslant k \leqslant m)$, $\varphi_k(A, D)$ 满足约束条件, 则直接转至步骤 6); 否则转至步骤 4);

6) 返回 A。

算法 3 的时间复杂度为 $O(|U|^2+m \cdot |U'|^2 \cdot |\mathrm{AT}|^2)$。其中, $|U'|$为新的决策系统中样本的个数。步骤 1 为样本选择的过程, 将一致性样本选择出来的时间复杂度为 $O(|U|^2)$。而之后的步骤则是用启发式算法求解约简, 由于使用的是新的决策系统, 则时间复杂度为 $O(m \cdot |U'|^2 \cdot |\mathrm{AT}|^2)$。这里的启发式算法可以为单准则属性约简算法也可以为多准则属性约简算法, 当 m=1 时则简化为单准则属性约简算法。换言之, 无论单准则还是多准则约简算法, 都可先经过采样后再根据具体需求设计相应的属性约简算法。

4　实验分析

为了验证算法 3 的有效性, 笔者从 UCI 数据集中选取了 8 组数据, 数据的基本描述如表 1 所列。实验环境为 PC 机, 双核 2.60 GHz CPU, 8 GB 内存, windows 10 操作系统, Matlab R2016a 实验平台。

表 1　数据描述

Table 1　Data sets description

ID	数据集	样本数	属性数	决策类数
1	Gesture Phase Segmentation	9 901	19	5
2	MAGIC Gamma Telescope	19 020	11	2
3	QSAR Biodegradation	1 055	41	2
4	Sonar	208	61	2
5	Statlog (German Credit Data)	1 000	25	2
6	Ultrasonic Flowmeter Diagnostics	180	44	4
7	Wall-Following Robot Navigation Data	5 466	25	4
8	Wine	178	13	3

实验采用了 5 折交叉验证的方法, 并且选取了 10 个不同的半径 δ, 其值分别为 0.03, 0.06, …, 0.3。5 折交叉验证的具体实施过程是将实验数据中的样本平均分成 5 份, 即 $U_1, U_2, \cup \cdots, U_5$。第一次使用 $U_2 \cup U_3 \cup \cdots \cup U_5$ 作为训练集求得约简 A_1, 使用 U_1 作为测试集并在其中利用 A_1 中的属性计算分类精度; 第 2 次使用 $U_1 \cup U_3 \cup \ldots \cup U_5$ 作为训练集求得约简 A_2, 使用 U_2 作为测试集并在其中利用 A_2 中的属性计算分类精度; 依次类推, 第 5 次使用 $U_1 \cup U_2 \cup \cdots \cup U_4$ 作为训练集求得约简 A_5, 使用 U_5 作为测试集并在其中利用 A_5 中的属性计算分类精度。

本组实验选取了近似质量、条件熵以及多准则 (同时满足近似质量和条件熵的约束条件) 作为约简的度量准则。实验将一致性采样属性约简算法与基于 K-means 采样[15] 的属性约简算法 (这

里 K 的取值等于数据的决策类数目) 进行对比分析。在上述 8 组数据集上分别计算并比较了基于这 3 种约简的分类精度。其中, 在计算分类精度时, 分别采用了邻域分类器 (NEC)[28] 和 SVM 分类器[29], 实验结果如图 1、图 2 所示。

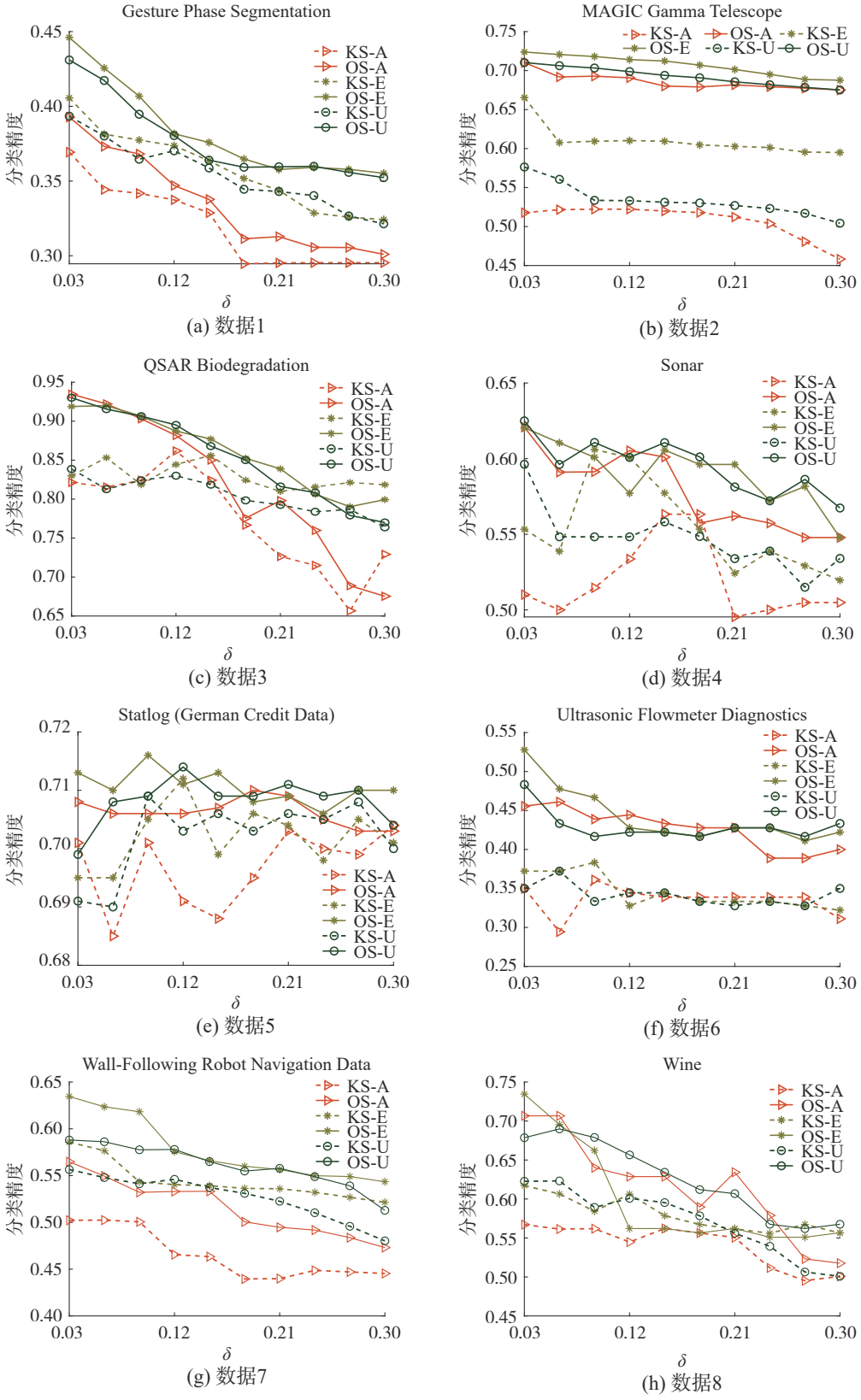

图 1　邻域分类器下分类精度的对比

Fig. 1　Comparisons among classification accuracies with using NEC

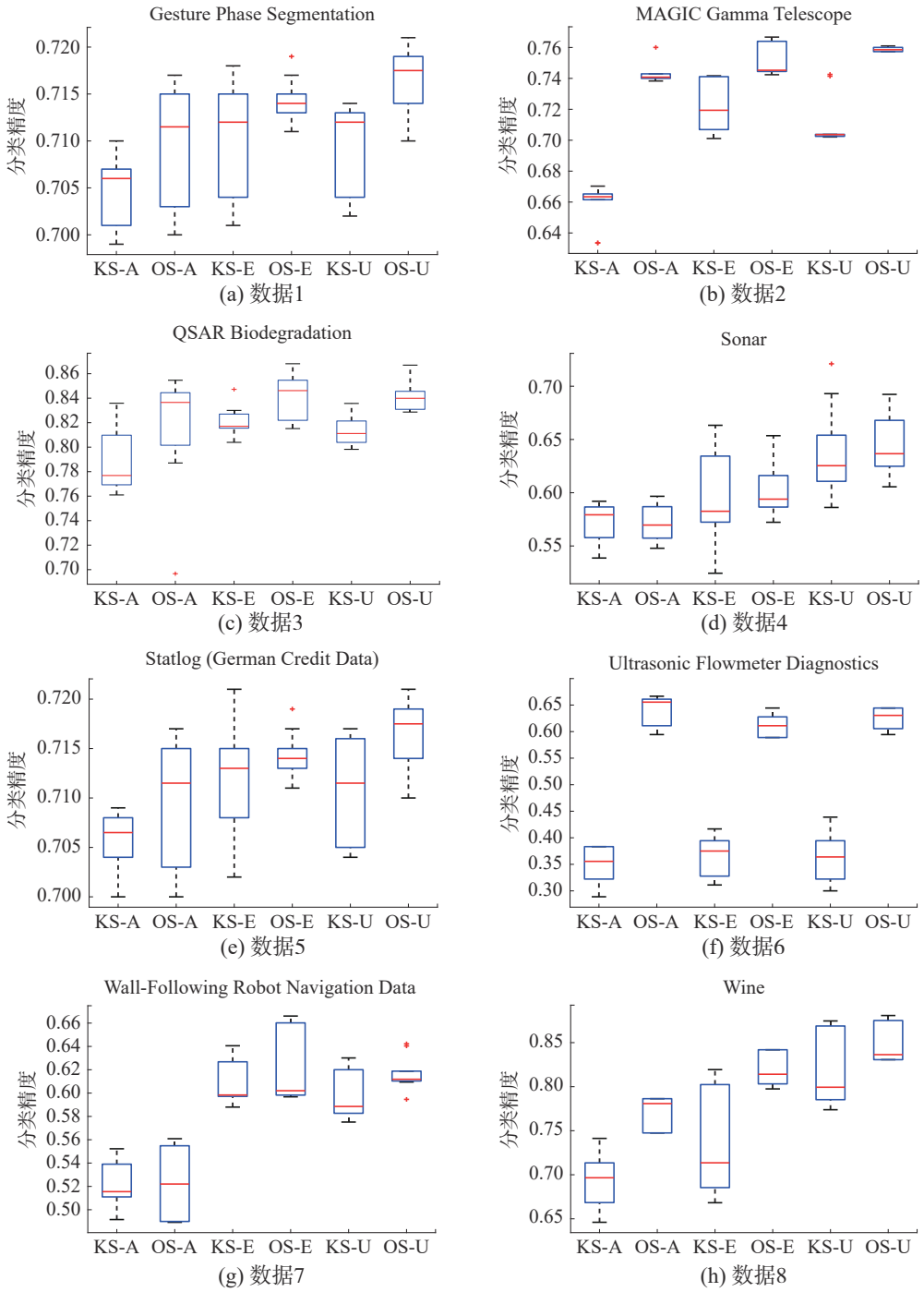

图 2　SVM 分类器下分类精度的对比

Fig. 2　Comparisons among classification accuracies with using SVM

在以下的结果图中, 用 KS-A、KS-E、KS-U 分别表示基于 *K*-means 采样的近似质量约简、条件熵约简、多准则约简, OS-A、OS-E、OS-U 分别表示基于一致性采样的近似质量约简、条件熵约简、多准则约简。

观察图 1 可以发现, 在 10 个半径下, 不难得出如下结论:

1) 相较于基于 *K*-means 采样的约简, 利用基于一致性采样的约简在测试样本上可以获得更好的分类效果;

2) 在 3 个度量准则的比较中, 利用条件熵约简能够大体上使得分类精度达到最高。此外, 一致性采样相较于 *K*-means 采样来说, 当利用近似质量作为约简的度量准则时, 约简在测试样本上分类效果的提升最为明显。这主要是因为相较于 *K*-means 采样来说, 一致性采样能够使得较多的样本落入下近似集中, 从而较大幅度地提升近似质量的值, 使得在约简迭代过程中, 需要更多

的属性被加入到约简集合中。

观察图 2, 不难得出如下结论:

1) 由于 SVM 分类器在计算分类精度时没有使用半径这一参数, 所以本文主要比较两者的分类精度的平均值, 可以发现相较于基于 *K*-means 采样的约简, 基于一致性采样的约简在测试样本上能够提供较高的分类精度;

2) 在 3 个度量准则的比较中, 利用多准则策略大体上可以使得分类精度达到最高, 这主要是因为多准则约简同时满足近似质量与条件熵的约束条件, 较多的约束条件需要较多的属性才能完成目标。

观察图 3 可以发现, 相较于 *K*-means 采样, 利用一致性采样进行约简求解, 大体上需要更多的时间消耗, 这主要是因为利用一致性采样得到的样本数量往往比利用 *K*-means 采样所得到的样本数量多, 这一事实可以参照表 2。

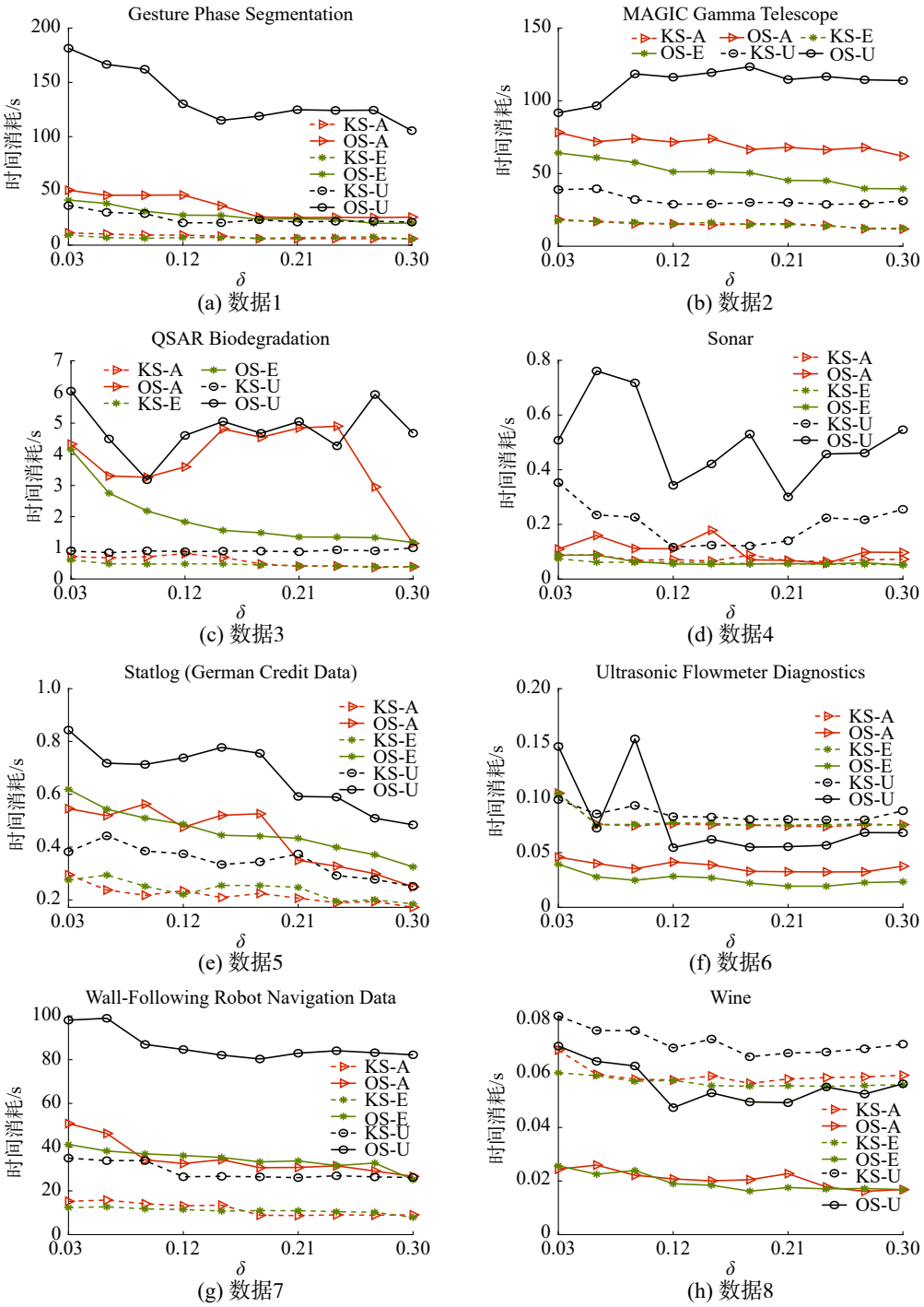

图 3　约简求解的时间消耗对比

Fig. 3　Comparisons among elapsed time for computing reducts

表 2　采样后数目

Table 2　Number of data after sample selection

ID	样本数	*K*-means采样 (占总样本比例/%)	一致性采样 (占总样本比例/%)
1	9 901	3 511(35.46)	7 400(74.27)
2	19 020	6 144(32.30)	12 310(64.72)
3	1 055	347(32.89)	887(84.08)
4	208	66(31.73)	143(68.75)
5	1 000	382(38.20)	537(53.70)
6	180	45(25.00)	133(73.88)
7	5 466	2 154(39.41)	3 835(70.16)
8	178	62(34.83)	136(76.41)

5　结束语

为了提高约简的求解效率, 本文提出一种基于一致性原则的采样方法。进一步地, 将基于一致性采样与基于聚类采样所求得的约简结果进行对比分析, 实验结果表明, 相较于聚类采样, 一致性采样的约简结果可以有效地提升分类器的分类性能。在这一工作的基础上, 本文将就以下问题展开进一步探讨:

1) 本文仅仅从整体角度考虑度量准则, 在之后的研究中可以进一步引入一些局部度量准则[30]如: 局部近似质量、局部条件熵等。

2) 本文算法及所使用的对比算法都仅仅是建立在一种采样技术的基础上的, 今后可以尝试使用混合采样的方法[31]以进一步地提升约简的性能。

参考文献:

[1] PAWLAK Z. Rough sets: Theoretical aspects of reasoning about data[M]. Dordrecht: Kluwer Academic Publishers, 1991.

[2] PAWLAK Z, GRZYMALA-BUSSE J, SLOWINSKI R, et al. Rough sets[J]. Communications of the ACM, 1995, 38(11): 88–95.

[3] CHEN Hongmei, LI Tianrui, LUO Chuan, et al. A decision-theoretic rough set approach for dynamic data mining[J]. IEEE transactions on fuzzy systems, 2015, 23(6): 1958–1970.

[4] WANG Changzhong, HU Qinghua, WANG Xizhao, et al. Feature selection based on neighborhood discrimination index[J]. IEEE transactions on neural networks and learning systems, 2018, 29(7): 2986–2999.

[5] HU Qinghua, YU Daren, XIE Zongxia, et al. EROS: Ensemble rough subspaces[J]. Pattern recognition, 2007, 40(12): 3728–3739.

[6] JIA Xiuyi, SHANG Lin, ZHOU Bing, et al. Generalized attribute reduct in rough set theory[J]. Knowledge-based systems, 2016, 91: 204–218.

[7] YANG Xibei, QI Yunsong, SONG Xiaoning, et al. Test cost sensitive multigranulation rough set: model and minimal cost selection[J]. Information sciences, 2013, 250: 184–199.

[8] CHEN Degang, YANG Yanyan, DONG Ze. An incremental algorithm for attribute reduction with variable precision rough sets[J]. Applied soft computing, 2016, 45: 129–149.

[9] YANG Xibei, LIANG Shaochen, YU Hualong, et al. Pseudo-label neighborhood rough set: measures and attribute reductions[J]. International journal of approximate reasoning, 2019, 105: 112–129.

[10] FERONE A. Feature selection based on composition of rough sets induced by feature granulation[J]. International journal of approximate reasoning, 2018, 101: 276–292.

[11] HART P. The condensed nearest neighbor rule (Corresp.)[J]. IEEE transactions on information theor, 1968, 14(3): 515–516.

[12] GATES G. The reduced nearest neighbor rule (Corresp.)[J]. IEEE transactions on information theory, 1972, 18(3): 431–433.

[13] TOMEK I. Two modifications of CNN[J]. IEEE transactions on systems, man, and cybernetics, 1976, SMC-6(11): 769–772.

[14] ANGIULLI F. Fast nearest neighbor condensation for large data sets classification[J]. IEEE transactions on knowledge and data engineering, 2007, 19(11): 1450–1464.

[15] 王熙照, 王婷婷, 翟俊海. 基于样例选取的属性约简算法 [J]. 计算机研究与发展, 2012, 49(11): 2305–2310.
WANG Xizhao, WANG Tingting, ZHAI Junhai. An attribute reduction algorithm based on instance selection[J]. Journal of computer research and development, 2012, 49(11): 2305–2310.

[16] ZHAI Junhai, WANG Xizhao, PANG Xiaohe. Voting-based instance selection from large data sets with mapreduce and random weight networks[J]. Information sciences, 2016, 367-368: 1066–1077.

[17] ZHAI Junhai, LI Ta, WANG Xizhao. A cross-selection instance algorithm[J]. Journal of intelligent and fuzzy systems, 2016, 30(2): 717–728.

[18] 杨习贝, 颜旭, 徐苏平, 等. 基于样本选择的启发式属性约简方法研究 [J]. 计算机科学, 2016, 43(1): 40–43.
YANG Xibei, YAN Xu, XU Suping, et al. New heuristic attribute reduction algorithm based on sample

selection[J]. Computer science, 2016, 43(1): 40–43.

[19] XU Suping, YANG Xibei, YU Hualong, et al. Multi-label learning with label-specific feature reduction[J]. Knowledge-based systems, 2016, 104: 52–61.

[20] GAO Yuan, CHEN Xiangjian, YANG Xibei, et al. Neighborhood attribute reduction: a multicriterion strategy based on sample selection[J]. Information, 2018, 9(11): 282.

[21] LIU Keyu, YANG Xibei, YU Hualong, et al. Rough set based semi-supervised feature selection via ensemble selector[J]. Knowledge-based systems, 2019, 165: 282–296.

[22] DAI Jianhua, WANG Wentao, XU Qing, et al. Uncertainty measurement for interval-valued decision systems based on extended conditional entropy[J]. Knowledge-based systems, 2012, 27: 443–450.

[23] ZHANG Xiao, MEI Changlin, CHEN Degang, et al. Feature selection in mixed data: a method using a novel fuzzy rough set-based information entropy[J]. Pattern recognition, 2016, 56: 1–15.

[24] LIN Jianhua. Divergence measures based on the Shannon entropy[J]. IEEE transactions on information theory, 1991, 37(1): 145–151.

[25] HU Qinghua, CHE Xunjian, ZHANG Lei, et al. Rank entropy-based decision trees for monotonic classification[J]. IEEE transactions on knowledge and data engineering, 2012, 24(11): 2052–2064.

[26] YANG Xibei, YAO Yiyu. Ensemble selector for attribute reduction[J]. Applied soft computing, 2018, 70: 1–11.

[27] LI Jingzheng, YANG Xibei, SONG Xiaoning, et al. Neighborhood attribute reduction: a multi-criterion approach[J]. International journal of machine learning and cybernetics, 2019, 10(4): 731–742.

[28] HU Qinghua, YU Daren, XIE Zongxia. Neighborhood classifiers[J]. Expert systems with applications, 2008, 34(2): 866–876.

[29] WANG Rui, LI Wei, LI Rui, et al. Automatic blur type classification via ensemble SVM[J]. Signal processing: image communication, 2019, 71: 24–35.

[30] CHEN Degang, ZHAO Suyun. Local reduction of decision system with fuzzy rough sets[J]. Fuzzy sets and systems, 2010, 161(13): 1871–1883.

[31] 孟军, 张晶, 姜丁菱, 等. 结合近邻传播聚类的选择性集成分类方法 [J]. 计算机研究与发展, 2018, 55(5): 986–993.
MENG Jun, ZHANG Jing, JIANG Dingling, et al. Selective ensemble classification integrated with affinity propagation clustering[J]. Journal of computer research and development, 2018, 55(5): 986–993.

作者简介：

高媛，女，1994 年生，硕士研究生，主要研究方向为粗糙集理论、机器学习。

陈向坚，女，1983 年生，副教授，博士，主要研究方向为模糊神经网络与智能控制。主持国家自然科学基金项目 1 项，发表学术论文 20 余篇。

王平心，男，1980 年生，副教授，博士，主要研究方向为矩阵分析与粒计算。主持国家自然科学基金项目 1 项，发表学术论文 30 余篇。