DOI: 10.11992/tis.201706080
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20180404.1544.014.html

重要度集成的属性约简方法研究

李京政[1], 杨习贝[1,2], 窦慧莉[1], 王平心[3], 陈向坚[1]
（1. 江苏科技大学 计算机学院, 江苏 镇江 212003; 2. 南京理工大学 经济管理学院, 江苏 南京 210094; 3. 江苏科技大学 数理学院, 江苏 镇江 212003）

摘　要: 启发式算法在求解约简的过程中逐步加入重要度最高的属性, 但其忽视了数据扰动将会直接引起重要度计算的波动问题, 从而造成约简结果的不稳定。鉴于此, 提出了一种基于集成属性重要度的启发式算法框架。首先, 在原始数据上进行多重采样; 然后, 在每次循环过程中分别计算各个采样结果上的属性重要度并对这些重要度进行集成; 最后, 将集成重要度最大的属性加入到约简中去。利用邻域粗糙集方法进行的实验结果表明, 基于集成重要度的属性约简算法不仅能够获取更加稳定的约简, 而且利用所生成的约简能够得到一致性较高的分类结果。

关键词: 属性约简; 分类; 聚类; 数据扰动; 集成; 启发式算法; 邻域粗糙集; 稳定性
中图分类号: TP391　　**文献标志码:** A　　**文章编号:** 1673−4785(2018)03−0414−08

中文引用格式: 李京政, 杨习贝, 窦慧莉, 等. 重要度集成的属性约简方法研究[J]. 智能系统学报, 2018, 13(3): 414–421.
英文引用格式: LI Jingzheng, YANG Xibei, DOU Huili, et al. Research on ensemble significance based attribute reduction approach[J]. CAAI transactions on intelligent systems, 2018, 13(3): 414–421.

Research on ensemble significance based attribute reduction approach

LI Jingzheng[1], YANG Xibei[1,2], DOU Huili[1], WANG Pingxin[3], CHEN Xiangjian[1]
(1. School of Computer, Jiangsu University of Science and Technology, Zhenjiang 212003, China; 2. School of Economics and Management, Nanjing University of Science and Technology, Nanjing 210094, China; 3. School of Mathematics and Physics, Jiangsu University of Science and Technology, Zhenjiang 212003, China)

Abstract: In the process of computing reduct using a heuristic algorithm, the attribute with the highest importance is gradually added in. However, this approach neglects the fluctuation of important calculations which is directly caused by data perturbation. Notably, such fluctuation may lead to an unstable reduct result. To eliminate such an anomaly, a framework consisting of a heuristic algorithm based on the importance of the ensemble attribute was proposed. In this approach, firstly, multiple sampling is executed for raw data; secondly, in each cycle, the importance of each attribute is computed on the basis of each sampling and the importance indices are integrated; finally, the attribute with the highest importance is added into the reduct. The experimental results obtained by utilizing the neighborhood rough set method show that the new approach not only obtains a more stable reduct, but also attains the classification results with high uniformity.

Keywords: attribute reduction; classification; clustering; data perturbation; ensemble; heuristic algorithm; neighborhood rough set; stability

作为粗糙集理论[1-2]研究的核心内容, 属性约简[3-4]问题一直是众多学者关心的焦点。所谓属性约简, 是在给定某一度量标准的前提下, 期望利用较少的属性, 能够超越利用原始数据中所有的属性所得到的性能或达到与其基本相当的性能。近年来, 根据不同的需求目标以及不同类型的拓展粗糙集模型[5-8], 众多研究者提出了诸如信息熵[9]、决策代价[10-11]、分类刻画[12]等类型的度量标准作为属性约简的定义。这些不同类型的属性约简大体上可以被划分为两大类[13-14]: 1) 面向粗糙集不确定性度量的属性约简; 2) 面向粗糙集学习性能的属性约简。

为了从数据中获取约简, 在粗糙集领域的研究

收稿日期: 2017−06−24.　**网络出版日期:** 2018−04−04.
基金项目: 国家自然科学基金项目 (61572242, 61503160, 61502211); 江苏省高校哲学社会科学基金项目 (2015SJD769); 中国博士后科学基金项目 (2014M550293).
通信作者: 杨习贝. E-mail: zhenjiangyangxibei@163.com.

中有两大类方法: 穷举法与启发式方法。分辨矩阵与回溯策略是穷举法的典型算法, 虽然穷举法可以帮助我们得到所有的约简, 但由于其计算复杂度过高, 并不适用于现实世界中的大规模数据处理。启发式算法是借助贪心的搜索策略求得数据中的一个约简, 虽然启发式算法有可能陷入局部最优, 仅能得到超约简, 但因其速度优势依然得到了广大研究学者的认可。

在启发式约简求解过程中, 属性重要度扮演着重要的角色, 在向约简集合不断增加属性的过程中, 每次都加入重要度最大的属性, 直至满足所定义的约简标准。但不难发现属性重要度的计算都是基于计算所有样本的基础上, 这会带来两个问题: 1) 每次计算都需要扫描所有样本, 时间消耗过大; 2) 未考虑数据扰动带来的属性重要度变化问题。虽然王熙照等[15]已经提出了利用边界样本求解属性重要度的方法, 这一思想可以进一步降低启发式约简求解的时间消耗[16], 但他们没有考虑数据扰动问题。微小的数据扰动有可能会使约简的结果大相径庭, 这不仅表明约简本身不具备稳定性, 而且也会致使根据约简所得到的分类及预测等结果也呈现不稳定性。针对上述问题, 笔者期望利用启发式算法, 求得具有较高稳定性的约简。借助集成学习[17-18]的基本思想, 可以设计一种集成属性重要度的计算方法, 对由不同边界样本所得到的属性重要度进行集成, 其目的是使属性重要度的输出更为鲁棒。

1　邻域粗糙集

在粗糙集理论中, 一个决策系统可以表示为二元组$\mathrm{DS}=\langle U,\mathrm{AT}\cup D\rangle$, 其中 U 是一个非空有限的对象集合, 即论域; AT 是所有条件属性集合; D 是所有决策属性的合集且 $\mathrm{AT}\cap D=\varnothing$。

给定论域$U=\{x_1,x_2,\cdots,x_n\}$, 邻域是建立在某一种度量标准上, 通过给定半径考察样本的邻居。不妨假设$\boldsymbol{M}=(r_{ij})_{n\times n}$为论域上的相似度矩阵, r_{ij} 表示对象 x_i 与 x_j 之间的距离度量, 给定半径 $\delta\in[0,1]$, $\forall x_i\in U$, x_i 的邻域区间为

$$\mathrm{Int}(x_i)=\min_{1\leqslant j\leqslant n,j\neq i} r_{ij}+\delta\times\left(\max_{1\leqslant j\leqslant n,j\neq i} r_{ij}-\min_{1\leqslant j\leqslant n,j\neq i} r_{ij}\right) \tag{1}$$

式中: $\min\limits_{1\leqslant j\leqslant n,j\neq i} r_{ij}$表示样本 x_j 中与样本 x_i 距离的最小值, $\max\limits_{1\leqslant j\leqslant n,j\neq i} r_{ij}$表示样本 x_j 中与样本 x_i 距离的最大值。采用邻域区间的方式考察样本的邻居可以避免因半径过小而产生空邻域的情形。借助邻域区间, $\forall x_i\in U$, 其邻域为

$$\delta(x_i)=\{x_i\in U|x_j\neq x_i,r_{ij}\leqslant\mathrm{Int}(x_i)\} \tag{2}$$

定义 1[19-20]　给定一个决策系统 DS, 根据 D 可以得到所有决策类的合集形如$\{X_1,X_2,\cdots,X_n\}$。$\forall B\subseteq\mathrm{AT}$, D 关于 B 的下近似和上近似分别定义为

$$\underline{N_B}D=\bigcup_{i=1}^{N}\underline{N_B}X_i \tag{3}$$

$$\overline{N_B}D=\bigcup_{i=1}^{N}\overline{N_B}X_i \tag{4}$$

对于任一决策类 X_i 有

$$\underline{N_B}X_i=\{x_i\in U|\delta_B(x_i)\subseteq X_i\} \tag{5}$$

$$\overline{N_B}X_i=\{x_i\in U|\delta_B(x_i)\cap X_i\neq\varnothing\} \tag{6}$$

定义 2　给定一个决策系统 DS, $\forall B\subseteq\mathrm{AT}$, D 相对于对 B 的依赖度为

$$\gamma(B,D)=\frac{|\underline{N_B}D|}{|U|} \tag{7}$$

式中$|N_BD|$与$|U|$分别表示集合 N_BD 与 U 的基数。

显然 $0\leqslant\gamma(B,D)\leqslant1$ 成立。$\gamma(B,D)$ 表示属于条件属性 B 的基础上, 某种决策类的样本占总体样本的比例。若$\underline{N_B}D$越大, 则依赖度越高。

2　属性约简

2.1　属性重要度与启发式算法

定义 3　给定一决策系统 DS, $\forall B\subseteq\mathrm{AT}$, B 被称为一个约简当且仅当$\gamma(B,D)=\gamma(\mathrm{AT},D)$且$\forall B'\subseteq B$, $\gamma(B',D)\neq\gamma(\mathrm{AT},D)$。

定义 3 所示的约简是一个能够保持决策系统中依赖度不发生变化的最小属性子集。根据定义 2 所示的依赖度, 可以进一步考察属性的重要度。

给定一个决策系统 DS, $\forall B\subseteq\mathrm{AT}$, 且对于任意的 $a\in\mathrm{AT}-B$, 如果$\gamma(B\cup\{a\},D)=\gamma(B,D)$, 那么就表明属性 a 对于计算依赖度没有带来任何贡献, a 是冗余的; 如果$\gamma(B\cup\{a\},D)>\gamma(B,D)$, 那么就表示加入属性 a 后可以提高依赖度, 从而降低不确定性程度。根据这样的分析, 可以构建如式 (8) 所示的属性重要度:

$$\mathrm{Sig}(a,B,D)=\gamma(B\cup\{a\},D)-\gamma(B,D) \tag{8}$$

根据上述属性重要度, 算法 1 构建了一个启发式求解过程, 其目标是获得以式 (8) 所示重要度为依据的属性排序序列。

算法 1　启发式算法

输入　邻域决策系统$\mathrm{DS}=\langle U,\mathrm{AT}\cup D\rangle$。

输出　属性排序 seq。

1) seq←Ø, $\gamma(\mathrm{seq},D)=0$;

2) 若 $\mathrm{AT}-\mathrm{seq}=\varnothing$, 则转至 5), 否则转至 3);

3) $\forall a_i\in\mathrm{AT}-\mathrm{seq}$;

4) 选择 a_j, 满足 $\mathrm{Sig}(a_j,\mathrm{seq},D)=\max\{\mathrm{Sig}(a_i,\mathrm{seq},D):\forall a_i\in\mathrm{AT}-\mathrm{seq}\}$, 令 $\mathrm{seq}=\mathrm{seq}\cup\{a_j\}$, 返回 2),计算 $\mathrm{Sig}(a_i,\mathrm{seq},D)$;

5) 输出 seq。

2.2 集成属性重要度

算法 1 在迭代过程中, 求解属性重要度是利用全体样本所得到的依赖度差异, 如式 (8)。但这种重要度计算方法忽视了数据扰动对重要度计算产生的影响, 当样本集发生变化时, 属性重要度势必也会发生相应的变化, 从而导致约简变化。如何降低样本集变化所引起的约简变化程度, 其本质是期望所求约简应尽可能稳定、鲁棒, 因此需要重新考察属性重要度的计算方法。

从分类学习的角度来看, 不同样本对学习性能的贡献程度是不相同的。一般来说, 那些对于学习性能影响比较重要的样本大都分布在边界区域上[15-16]。从这一考虑出发, 可以将边界区域的样本挑选出来, 作为计算属性重要度的依据。一个可行且直观的办法是采用聚类算法对原始样本集进行聚类, 在各类簇中挑选出距离类簇中心较远的样本, 将这些样本组合成一个新的决策系统, 这实际上是一个采样的过程 (具体描述如算法 2 所示)。又因为传统的聚类算法, 如 k-means 聚类的结果并不稳定, 其初始类簇中心是随机选取的, 故可以在原始样本集上通过多次聚类后得到多个决策系统, 分别在这多个决策系统上求得各个属性的重要度并进行融合, 最后选择融合重要度较高的属性, 其具体描述如算法 3 所示。

算法 2　基于 k-means 聚类的采样

输入　邻域决策系统DS = $\langle U, \mathrm{AT}\cup D\rangle$。

输出　采样后的决策系统DS′。

1) $U' = \varnothing$;

2) 利用 k-means 聚类获得 U 上的类簇 $C = \{C_1, C_2, \cdots, C_N\}$, 其中 N 为决策类的个数;

3) for j = 1 to N

①计算类簇 C_j 中每个样本到类簇中心的平均距离 $\overline{d}_j$;

②将 C_j 中到类簇中心的距离大于平均距离 $\overline{d}_j$ 的样本挑选出来加入U';

end for

4) 输出DS′。

算法 3　重要度集成的启发式算法

输入　邻域决策系统DS = $\langle U, \mathrm{AT}\cup D\rangle$, 采样次数 k;

输出　属性排序 seq。

1) seq←Ø, γ(seq, D)=0;

2) for r = 1 to k

利用算法 2 进行采样得到决策系统 DS$_r$;

end for

3) 若 AT–seq = Ø, 则转至 7), 否则转至 4);

4) for r = 1 to k

$\forall a_i \in$ AT–seq, 计算属性 a_i 在决策系统 DS$_r$ 上的重要度 Sig$_r$(a_i, AT, D);

end for

5) $\forall a_i \in$ AT– seq, 融合属性 a_i 在各个决策系统上的重要度:

$$\mathrm{Sig}(a_i, \mathrm{AT}, D) = \frac{\sum_{r=1}^{K} \mathrm{Sig}_r(a_i, \mathrm{AT}, D)}{k};$$

6) 选择 a_j, 满足 Sig(a_j, seq, D)=max{Sig(a_i, seq, D): $\forall a_i \in$ AT–seq}, 令 seq=seq$\cup${a_j}, 返回 3);

7) 输出 seq。

在邻域粗糙集上求解属性重要度的时间复杂度为 $O(U^2\times n)$, 其中 U 表示论域中对象个数, n 代表条件属性个数, 算法 1 的时间复杂度为 $O(U^2\times n^3)$。k-means 聚类的时间复杂度为 $O(N\times T\times U)$, N 为类簇个数, T 为迭代次数。算法 3 的时间消耗为$k\times N\times T\times U + k\times[U]^2\times n^3$, 其中$[U]$是一个不确定集, 表示每次 k-means 聚类采样的对象个数, 聚类次数 k 为常数, 故算法 3 的时间复杂度为 $O([U]^2\times n^3)$。通过大量实验表明$[U]<U$, 所以在聚类次数较少的前提下, 算法 3 的时间消耗一般小于算法 1。

3 实验分析

为了验证所提算法的有效性, 本文从 UCI 数据集中选择了 9 组数据, 数据的基本描述如表 1 所示。实验中取 k=5, 即进行 5 次聚类采样, 由算法 3 得到的属性序列不仅和传统的启发式算法比较, 而且和王熙照等[15]提出的基于样例选取的求解属性重要度算法对比分析。

表 1　实验数据的基本信息

Table 1　Data sets description

数据集编号	数据集名称	样本数	属性数	决策类数
1	Dermatology	366	35	6
2	Diabetic Retinopathy Debrecen	1 151	20	2
3	Ecoli	336	8	8
4	Ionosphere	351	35	2
5	Iris	150	4	3
6	Parkinson Multiple Sound Recordings	1 208	27	2
7	Pima Indians Diabetes	768	9	2
8	Tic-Tac-Toe Endgame	958	9	2
9	Yeast	1 484	9	10

为了比较 3 种约简算法在样本扰动情况下属性的排序结果, 采用了 5 折交叉验证来实现。具体过程为: 在每个数据集上, 将数据集随机地平均分成 5 份, 即 $U_1, U_2, \cdots, U_5$。第一次使用 $U_2 \cup U_3 \cup \cdots \cup U_5$ 求得属性排序结果 seq_1; 第二次使用 $U_1 \cup U_3 \cup \cdots \cup U_5$ 求得属性排序结果 seq_2; 依次类推, 第 5 次使用 $U_1 \cup U_2 \cup \cdots \cup U_4$ 求得属性排序结果 seq_5。

3.1　属性序列的稳定性比较

度量属性序列的稳定性, 就是在样本扰动时度量不同属性序列之间的相似性, 相似性越高, 说明所得到的属性序列越稳定, 可使用式 (9)[21]计算属性序列的相似性:

$$\text{Sta} = \frac{2\sum_{i=1}^{d-1}\sum_{j=i+1}^{d}\text{Sim}(\text{seq}_i, \text{seq}_j)}{d \times (d-1)} \tag{9}$$

式中: seq_i 与 seq_j 表示第 i 和第 j 个属性排序结果; d=5 表示 5 折交叉验证; Sim 表示序列之间的相似性, 可采用 Spearman 排序关联系数进行计算:

$$\text{Sim}(\text{seq}_i, \text{seq}_j) = 1 - 6 \times \sum_{l=1}^{n} \frac{(\text{seq}_i^l - \text{seq}_j^l)^2}{n \times (n^2 - 1)} \tag{10}$$

式中: n 表示属性个数, seq_i^l 表示第 l 个属性在第 i 个序列中的排序值, 本文将排在最前端的属性排序值定为 n, 往后依次减 1。

表 2 列出了 4 个不同的邻域半径下 3 种约简算法所求得的属性序列的稳定性结果。

观察表 2 可以发现, 在大多数的半径参数 δ 下, 利用算法 3 所求得的属性序列相似度都比利用算法 1 及文献[15]算法所求得的属性序列相似度高, 这说明算法 3 在增加属性的过程中所得到的属性序列是比较稳定的。

表 2　属性序列的稳定性对比

Table 2　Comparisons of stabilities of attribute sequences

数据集编号	δ=0.1			δ=0.2			δ=0.3			δ=0.4			平均值		
	算法 1	算法 3	文献[15]算法	算法 1	算法 3	文献[15]算法	算法 1	算法 3	文献[15]算法	算法 1	算法 3	文献[15]算法	算法 1	算法 3	文献[15]算法
1	0.312 4	0.491 7	0.325 1	0.343 4	0.505 6	0.213 2	0.323 6	0.572 0	0.238 5	0.456 6	0.531 2	0.392 9	0.359 0	0.525 1	0.292 4
2	0.297 0	0.507 9	0.347 0	0.263 3	0.381 6	0.309 5	0.385 4	0.476 7	0.246 8	0.473 3	0.474 6	0.547 0	0.354 8	0.460 2	0.362 6
3	0.803 6	0.764 3	0.528 6	0.414 3	0.628 6	0.443 2	0.157 1	0.607 1	0.389 3	0.260 7	0.792 9	0.521 4	0.408 9	0.698 2	0.470 6
4	0.229 3	0.491 1	0.152 4	0.291 3	0.248 7	0.398 1	0.385 3	0.404 7	0.126 5	0.293 4	0.307 7	0.388 0	0.299 8	0.363 0	0.266 2
5	0.200 0	0.680 0	0.200 0	0.260 0	0.200 0	0.100 0	0.260 0	0.360 0	0.320 0	0.880 0	0.560 0	0.280 0	0.400 0	0.450 0	0.225 0
6	0.353 7	0.407 9	0.364 7	0.267 4	0.262 7	0.240 1	0.120 9	0.377 3	0.217 1	0.325 3	0.467 7	0.211 5	0.266 8	0.378 9	0.258 4
7	0.402 4	0.609 5	0.209 5	0.211 9	0.483 3	0.254 8	0.190 5	0.261 9	0.270 6	0.383 3	0.519 0	0.366 7	0.297 0	0.468 4	0.275 4
8	0.285 7	0.463 5	0.369 0	0.235 7	0.547 6	0.410 7	0.404 8	0.531 0	0.269 0	0.325 3	0.208 4	0.388 1	0.312 9	0.437 6	0.359 2
9	0.419 0	0.616 7	0.268 0	0.178 6	0.309 5	0.214 8	0.319 0	0.421 4	0.164 7	0.407 1	0.540 5	0.267 4	0.330 9	0.472 0	0.228 7

此外, 为了检验新算法约简结果稳定性在统计学上是否具有显著性差异, 对各算法的属性序列稳定性的值, 采用 Friedman 检验[22]分别计算它们的秩及 APV(adjusted p-value), 判断其是否拒绝原假设。其中, 显著性水平 α 设为 0.05。统计分析结果如表 3 所示。

表 3　各个算法的统计结果

Table 3　Statistical results of various algorithms

算法	秩	APV
算法 1	2.33	8.14×10^{-4}
算法 3	1.00	—
文献[15]算法	2.67	0.4×10^{-2}

从表 3 可以看出, 算法 3 在各个算法里的秩最小, 这表明算法 3 性能最好。此外, 算法 1 与文献[15]的算法的 APV 值均小于显著性水平 α=0.05, 这意味着算法 3 与其余两种算法有着显著性的差异。

3.2　分类结果的一致性比较

在求解属性排序序列的过程中, 将重要度较大的属性逐个添加到约简结果中。在属性序列逐步增长的过程中, 不同序列在同一分类器上也会产生不同的分类结果。借助交叉验证, 由属性序列 $\text{seq}_u^{\text{num}}{}_{1\leqslant\text{num}\leqslant\text{AT}}$ 与 $\text{seq}_v^{\text{num}}{}_{1\leqslant\text{num}\leqslant\text{AT}}$ 可构造联合分布矩阵, 如表 4 所示。

表 4　联合分布矩阵

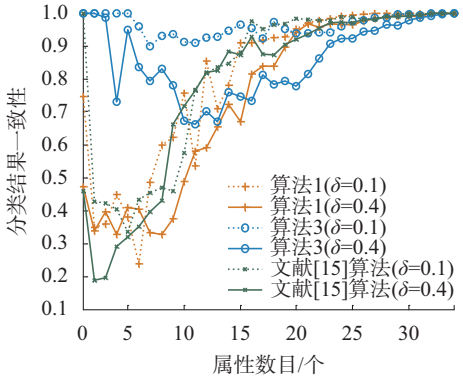
Table 4　Joint distribution matrix

真实情况	$\text{seq}_u^{\text{num}}(x) = d(x)$ $_{1\leqslant\text{num}\leqslant\text{AT}}$	$\text{seq}_u^{\text{num}}(x) \neq d(x)$ $_{1\leqslant\text{num}\leqslant\text{AT}}$
$\text{seq}_v^{\text{num}}(x) = d(x)$ $_{1\leqslant\text{num}\leqslant\text{AT}}$	a_{uv}	b_{uv}
$\text{seq}_v^{\text{num}}(x) \neq d(x)$ $_{1\leqslant\text{num}\leqslant\text{AT}}$	c_{uv}	d_{uv}

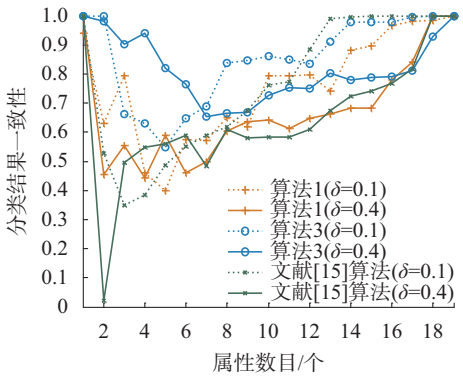
在表 4 中，当前 $\mathrm{seq}_u^{\mathrm{num}}{}_{1\leqslant \mathrm{num}\leqslant AT}$ 包含 num 个条件属性，$\mathrm{seq}_u^{\mathrm{num}}{}_{1\leqslant \mathrm{num}\leqslant AT}(x)$ 表示利用交叉验证第 u 轮由属性序列 $\mathrm{seq}_u^{\mathrm{num}}{}_{1\leqslant \mathrm{num}\leqslant AT}$ 在某一种分类器上对样本 x 做出的预测结果。$\forall x \in U$，若 $\mathrm{seq}_u^{\mathrm{num}}{}_{1\leqslant \mathrm{num}\leqslant AT}(x) = d(x)$，表示利用当前的属性序列，可以做出正确的分类结果；反之，则表示做出的分类结果是错误的。基于联合分布矩阵，采用 Yule 提出的 Q-统计量方法来度量两种算法的约简在分类器上分类结果的一致性，一致性的度量是反映分类性能稳定性的指标，其计算式为

$$Q = \frac{a_{uv}d_{uv} - b_{uv}c_{uv}}{a_{uv}d_{uv} + b_{uv}c_{uv}} \tag{11}$$

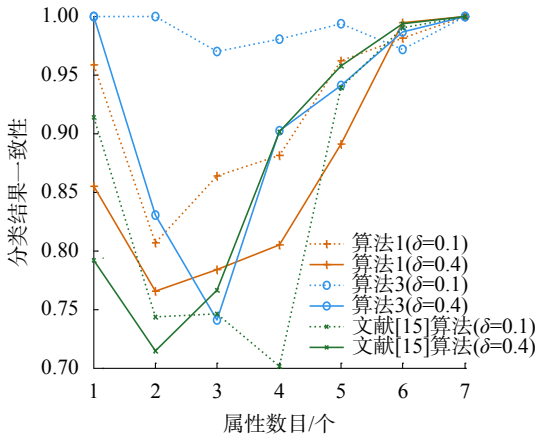
式中 Q 的取值范围为[−1, 1]。Q 值为 0 时，表示两个排序序列在同一分类器上的预测结果毫不相关；Q 值越大，表示当前两个排序结果在同一分类器上的预测结果的一致性越高。整体的一致性可取平均值 $\overline{Q}$ 作为分类结果的稳定性指标。实验中采用 KNN 分类器去分类，因为不同的数据集对 K 的敏感程度不一样，为了降低 K 的取值对分类结果影响，每个数据集对 K 寻优，在最佳的 K 值情况下，再比较各个算法下的分类性能。按照表 1 顺序，K 分别取值为 3、5、9、3、5、9、7、3、5。为了能直观比较 3 种约简算法分类结果的一致性，以及不同邻域半径参数下对分类结果一致性的影响，分别在邻域参数 δ=0.1 与 δ=0.4 时完成本组实验。实验结果如图 1 所示。



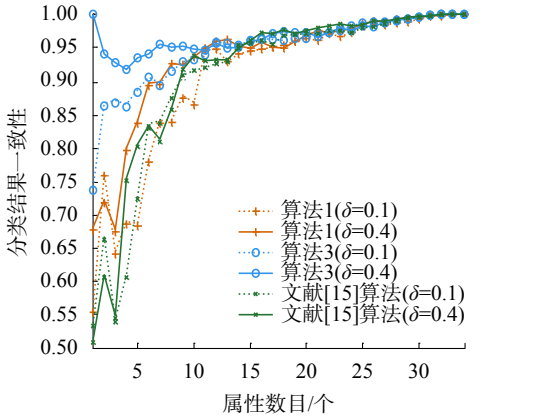
(a) Dermatology



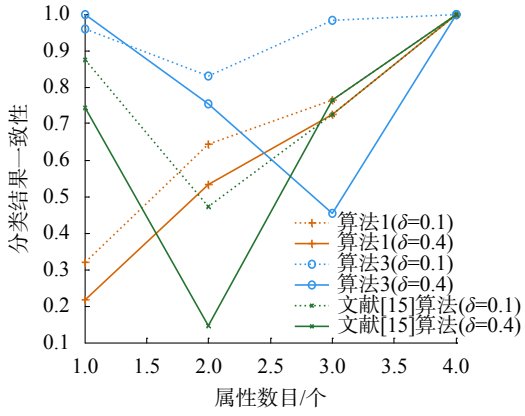
(b) Diabetic Retinopathy Debrecen



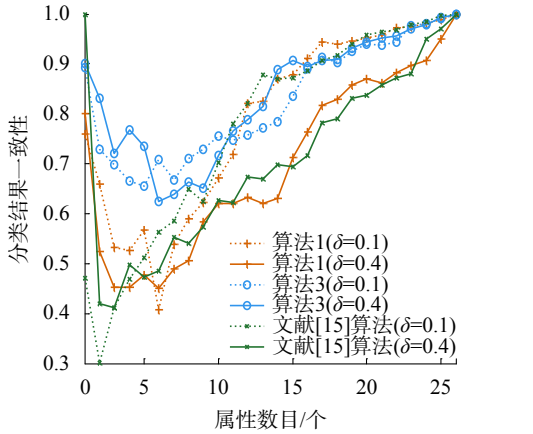
(c) Ecoli



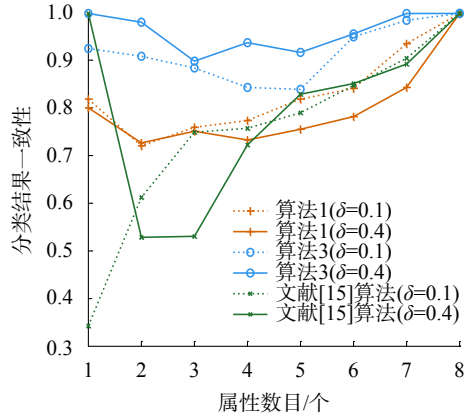
(d) Ionosphere



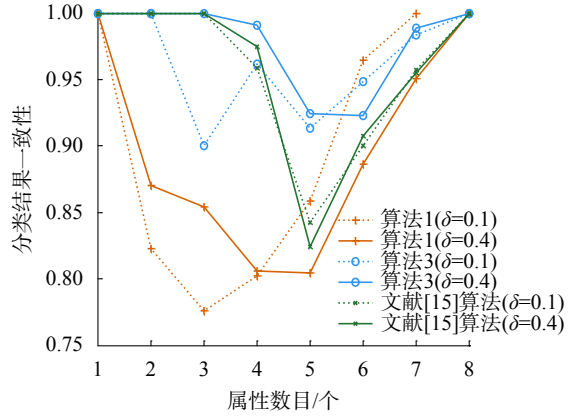
(e) Iris



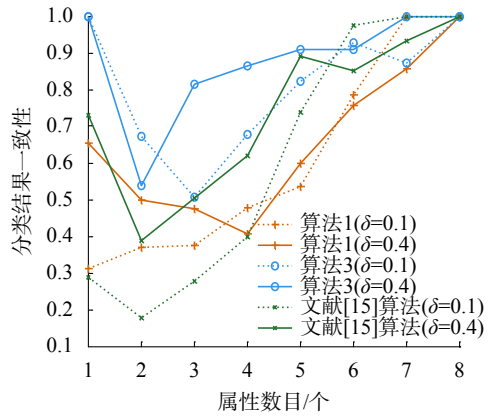
(f) Parkinson Multiple Sound Recording



(g) Pima Indians Diabetes



(h) Tic-Tac-Toe Endgame



(i) Yeast

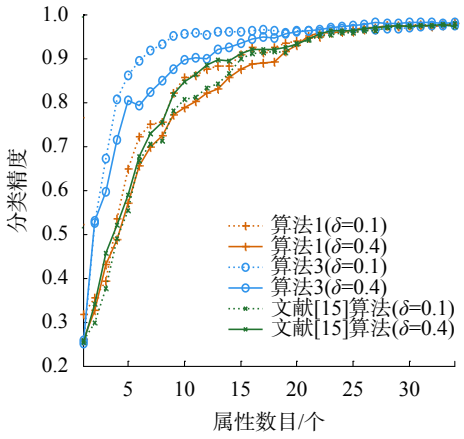
图 1　分类结果一致性的对比

Fig. 1　Comparisons of classification agreements

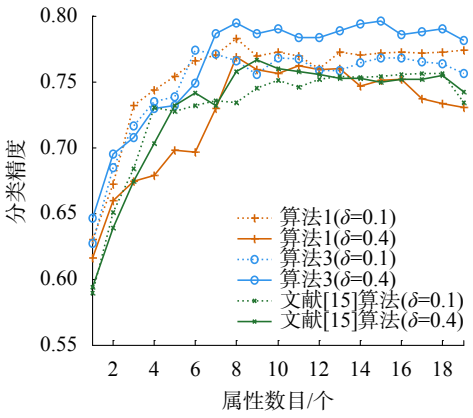
由图 1 可知, 随着属性逐个加入排序序列中在相同的邻域半径参数下, 算法 3 在依次增加属性时做出分类结果的一致性总体比算法 1 以及文献[15]算法要高, 验证了由算法 3 求得属性序列做出分类结果的稳定性要高于算法 1 以及文献[15]算法。

3.3　分类精度比较

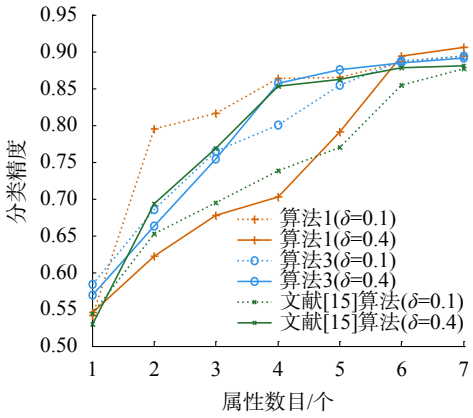
随着当前重要度最大的属性逐渐加入到属性序列中去, 进一步考虑当前属性序列的分类精度, 对此, 分别在邻域参数 δ=0.1 与 δ=0.4 时比较了 3 种约简算法的分类精度, 实验结果如图 2 所示。



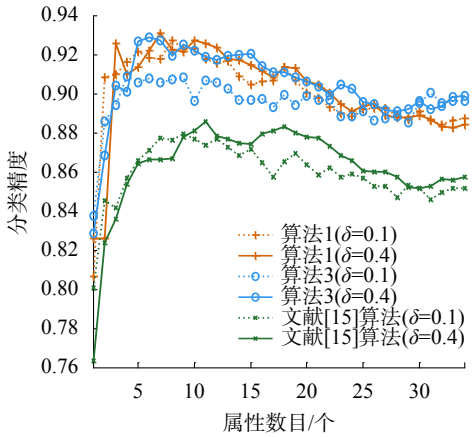
(a) Dermatology



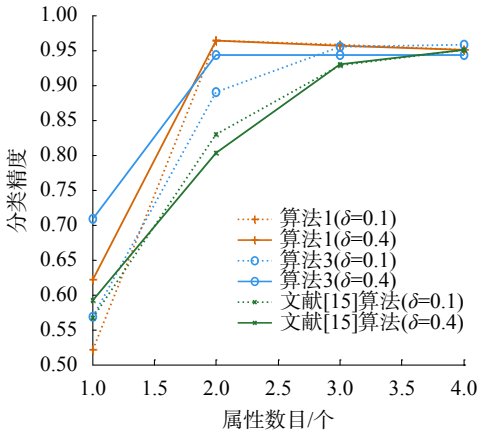
(b) Diabetic Retinopathy Debrecen



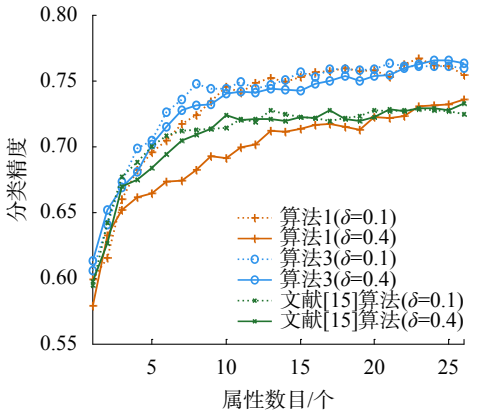
(c) Ecoli



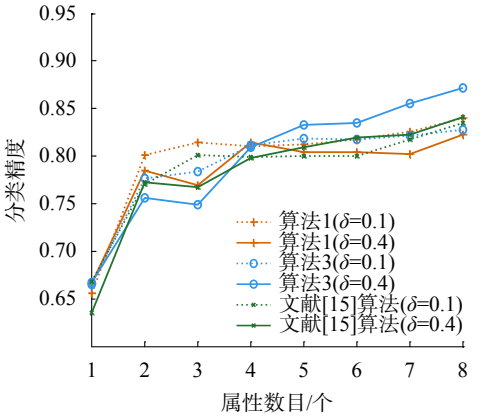
(d) Ionosphere



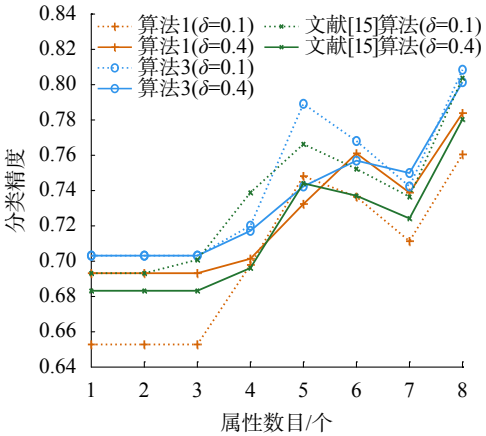
(e) Iris



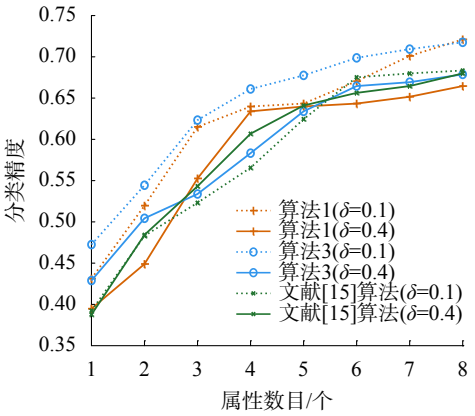
(f) Parkinson Multiple Sound Recording



(g) Pima Indians Diabetes



(h) Tic-Tac-Toe Endgame



(i) Yeast

图 2　分类精度的对比

Fig. 2　Comparisons of classification accuracies

通过图 2 可知，在相同的邻域半径参数下，随着属性逐个加入到排序序列中，其分类精度也在不断提高，当属性达到一定个数时，分类精度也趋于平稳。总体来说，由算法 1、算法 3 和文献[15]的算法得到的属性序列的分类精度是差不多的。尽管算法 3 得到的属性序列在样本扰动下的稳定性以及分类结果的一致性较算法 1 和文献[15]的算法能够得到提升，但是未能有效提升属性序列的分类精度。

4　结束语

利用邻域粗糙集求解约简时，提出了一种可以得到稳定约简的启发式算法框架。这种新的算法在多次采样基础上利用集成的思想求解属性重要度，从而可以用来提高约简的稳定性。实验结果表明，新算法在有效地提升约简稳定性的同时，亦能提高由约简所做出分类结果的稳定性。在本文工作的基础上，下一步工作主要有：1) 针对稳定的属性约简与分类度量指标之间的关系进行深入讨论，以期能够在获得稳定约简的基础上，提升分类精度等相应的学习性能；2) 文中聚类采样方法未能考虑到原始样本分布情况，对于某些非凸型分布的样本，或许不能有效地抽取到边界样本。进一步考虑数据的分布情况，寻求更有效的方法抽取到边界样本也是笔者的下一步工作。

参考文献：

[1] PAWLAK Z. Rough sets: theoretical aspects of reasoning about data[M]. Boston, Mass, USA: Kluwer Academic Publishers, 1991.

[2] PAWLAK Z. Rough sets[J]. International journal of computer & information sciences, 1982, 11(5): 341–356.

[3] JU Hengrong, LI Huaxiong, YANG Xibei, et al. Cost-sensitive rough set: a multi-granulation approach[J]. Knowledge-

based systems, 2017, 123: 137–153, doi: 10.1016/j.knosys.2017.02.019.

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

[5] DUBOIS D, PRADE H. Rough fuzzy sets and fuzzy rough sets[J]. International journal of general systems, 1990, 17(2/3): 191–209.

[6] 胡清华, 于达仁, 谢宗霞. 基于邻域粒化和粗糙逼近的数值属性约简[J]. 软件学报, 2008, 19(3): 640–649.
HU Qinghua, YU Daren, XIE Zongxia. Numerical attribute reduction based on neighborhood granulation and rough approximation[J]. Journal of software, 2008, 19(3): 640–649.

[7] YANG Xibei, CHEN Zehua, DOU Huili, et al. Neighborhood system based rough set: models and attribute reductions[J]. International journal of uncertainty, fuzziness and knowledge-based systems, 2012, 20(3): 399–419.

[8] LIANG Jiye, WANG Feng, DANG Chuangyin, et al. An efficient rough feature selection algorithm with a multi-granulation view[J]. International journal of approximate reasoning, 2012, 53(6): 912–926.

[9] 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.

[10] JU Hengrong, YANG Xibei, YU Hualong, et al. Cost-sensitive rough set approach[J]. Information sciences, 2016, 355-356: 282–298.

[11] JIA Xiuyi, LIAO Wenhe, TANG Zhenmin, et al. Minimum cost attribute reduction in decision-theoretic rough set models[J]. Information sciences, 2013, 219: 151–167.

[12] 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.

[13] MIN Fan, HE Huaping, QIAN Yuhua, et al. Test-cost-sensitive attribute reduction[J]. Information sciences, 2011, 181(22): 4928–4942.

[14] SONG Jingjing, TSANG E C C, CHEN Degang, et al. Minimal decision cost reduct in fuzzy decision-theoretic rough set model[J]. Knowledge-based systems, 2017, 126: 104–112, doi: 10.1016/j.knosys.2017.03.013.

[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] 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.

[17] LI Yun, SI J, ZHOU Guojing, et al. FREL: a stable feature selection algorithm[J]. IEEE transactions on neural networks and learning systems, 2014, 26(7): 1388–1402.

[18] 周林, 平西建, 徐森, 等. 基于谱聚类的聚类集成算法[J]. 自动化学报, 2012, 38(8): 1335–1342.
ZHOU Lin, PING Xijian, XU Sen, et al. Cluster ensemble based on spectral clustering[J]. Acta automatica sinica, 2012, 38(8): 1335–1342.

[19] YANG Xibei, ZHANG Ming, DOU Huili, et al. Neighborhood systems-based rough sets in incomplete information system[J]. Knowledge-based systems, 2011, 24(6): 858–867.

[20] QIAN Yuhua, WANG Qi, CHENG Honghong, et al. Fuzzy-rough feature selection accelerator[J]. Fuzzy sets and systems, 2014, 258: 61–78.

[21] LI Jingzheng, YANG Xibei, SONG Xiaoning, et al. Neighborhood attribute reduction: a multi-criterion approach[J]. International journal of machine learning and cybernetics, 2017, doi: 10.1007/s13042-017-0758-5.

[22] DEMŠAR J. Statistical comparisons of classifiers over multiple data sets[J]. Journal of machine learning research, 2006, 7: 1–30.

作者简介:

李京政, 男, 1993 年生, 硕士研究生, 主要研究方向为粗糙集理论、机器学习。

杨习贝, 男, 1980 年生, 副教授, 博士后, 主要研究方向为粗糙集理论、粒计算、机器学习。发表学术论文 100 余篇, 被 SCI 检索 50 余篇, 出版英文专著一部。

窦慧莉, 女, 1980 年生, 助理研究员, 主要研究方向为粒计算、智能信息处理。