DOI: 10.11992/tis.202001017

弱标记不完备决策系统的增量式属性约简算法

程龙[1,2], 钱文彬[1,2], 王映龙[1], 胡剑锋[3]

(1. 江西农业大学 计算机与信息工程学院, 江西 南昌 330045; 2. 江西农业大学 软件学院, 江西 南昌 330045; 3. 江西科技学院 信息技术研究所, 江西 南昌 330098)

摘 要: 在许多现实应用领域中, 由于数据标注代价昂贵, 且数据往往呈现动态变化, 因此存在大量弱标记的不完备数据。针对上述复杂应用场景, 本文以粒计算理论为基础, 从区分性视角给出不完备数据的区分对概念, 同时给出属性相对重要度的度量方法, 并设计面向弱标记不完备决策系统的属性约简算法。该算法能在迭代过程中不断缩减搜索空间, 提高属性约简效率; 并根据实例的动态变化情况, 分析属性约简的动态更新机制; 在此基础上, 设计了半监督条件下的增量式属性约简算法。最后, 通过实验验证了算法的可行性和有效性。

关键词: 属性约简; 粗糙集; 区分对; 混合数据; 增量学习; 半监督学习; 相对重要度; 动态数据

中图分类号: TP391 **文献标志码:** A **文章编号:** 1673−4785(2020)06−1079−12

中文引用格式: 程龙, 钱文彬, 王映龙, 等. 弱标记不完备决策系统的增量式属性约简算法 [J]. 智能系统学报, 2020, 15(6): 1079–1090.

英文引用格式: CHENG Long, QIAN Wenbin, WANG Yinglong, et al. An incremental attribute reduction algorithm for incomplete decision system with weak labeling[J]. CAAI transactions on intelligent systems, 2020, 15(6): 1079–1090.

An incremental attribute reduction algorithm for incomplete decision system with weak labeling

CHENG Long[1,2], QIAN Wenbin[1,2], WANG Yinglong[1], HU Jianfeng[3]

(1. School of Computer and Information Engineering, Jiangxi Agricultural University, Nanchang 330045, China; 2. School of Software, Jiangxi Agricultural University, Nanchang 330045, China; 3. Institute of Information Technology, Jiangxi University of Technology, Nanchang 330098, China)

Abstract: Due to the high cost of data annotation and dynamic change of data, many practical applications have a lot of incomplete data with weak labeling. In view of the above complex scenarios, based on the theory of granular computing, the concept of discernibility pairs of incomplete data is proposed and provides a measurement method for the relative importance of attributes. The attribute reduction algorithm is designed for an incomplete decision system with weak labeling, which can reduce the search space and improve the efficiency of attribute reduction. Besides, the dynamic updating mechanism of attribute reduction is analyzed based on the dynamic change of instances. In this study, an incremental attribute reduction algorithm is designed under a semi-supervised scene, and the experimental results show the feasibility and effectiveness of the proposed algorithm.

Keywords: attribute reduction; rough set; discernibility pair; mixed data; incremental learning; semi-supervised learning; relative importance; dynamic data

粗糙集理论[1-2]是一种有效的数据分析方法, 主要用于处理不确定、不一致和模糊的数据[3-4], 已被广泛地应用于知识发现、数据挖掘和机器学习等领域。属性约简是粗糙集理论的重要研究内容之一, 它旨在保持原属性集区分能力不变的情

收稿日期: 2020−01−09.

基金项目: 国家自然科学基金项目 (61966016); 江西省自然科学基金项目 (20192BAB207018); 江西省教育厅科学技术研究项目 (GJJ180200).

通信作者: 钱文彬. E-mail: qianwenbin1027@126.com.

况下，剔除不重要或冗余的属性[5-6]。由于属性约简的结果往往不是唯一的，找到数据所有约简的结果，已经被证明是一个 NP-hard 问题。因此在实际应用中通常采用启发式算法处理大规模数据，获取满足知识发现要求的属性约简结果[7-9]。

然而，在众多现实应用领域中存在大量的高维复杂数据。并且在数据采集阶段，由于采集成本和技术的限制，导致这些高维数据往往存在缺失。同时，给这些数据进行类别标注，需要耗费大量的人力资源，并且利用经典粗糙集无法直接处理不完备数据。针对上述问题研究人员引入容差关系和限制容差关系等，拓展了经典粗糙集的应用[10-12]，但这些扩展的关系模型难以直接处理含有连续型的不完备混合数据。同时，由于在实际应用中，这些高维数据通常仅包含少量已标注的数据。若仅利用带标记的数据进行属性约简，约简结果不能有效反映数据的分布，且分类性能较弱。

弱监督属性约简旨在有效利用无标记数据来增强属性约简的有效性，从而提高学习模型的分类性能。近年来，弱监督属性约简引起了许多研究人员的关注和研究。文献 [13] 针对弱标记的符号型数据，在区分对的基础上，利用有监督学习框架和无监督学习框架，构造相对应的启发式半监督属性约简算法。文献 [14] 基于粗糙集理论和信息熵的概念，在对无标记数据进行部分标注后，设计了一种基于信息熵的半监督特征选择算法。文献 [15] 将粗糙集理论和集成学习框架相结合，利用有标记的数据训练基分类器对无标记的数据进行标注，扩充有标记的数据。文献 [16] 提出了一种基于流形正则化的半监督特征选择算法，通过最大化不同类别之间的间距对特征的重要性进行度量分析。文献 [17] 提出了一种半监督特征选择算法，算法通过组合半监督散点，有效利用大量未标记的视频数据中的信息来区分目标类别。上述这些方法为弱标记数据的属性约简，提供了有效的解决方案。

另外，在大数据时代，许多数据随着时间的推移而动态变化。在这种复杂应用场景中，传统的属性约简算法在处理这些动态数据时，将会产生大量重复计算，无法快速更新属性约简结果。近年来，许多学者对动态属性约简算法进行了大量的研究。文献 [18] 提出了一种基于信息熵的组增量式属性约简算法，在动态增加一组实例后，快速更新属性约简结果。文献 [19] 采用一种复合粗糙集模型，处理动态的不完备数据，在数据动态变化后快速更新近似集合。文献 [20] 提出了一种基于知识粒度模型的动态属性约简算法，在实例变化后动态更新属性约简集。文献 [21] 提出了一种不完备动态属性约简算法，该算法在不完备决策系统中单个实例变化后动态获取新的属性约简结果集。上述研究都是针对所有实例均有标记的数据，目前对弱标记数据的增量式属性约简研究较少。为此，有效地利用无标记数据来增强属性约简结果，并在保证分类精度前提下动态更新属性约简结果，已成为了当前亟待解决的问题。

针对上述问题，本文提出了弱标记不完备数据的区分对概念，给出了属性的相对重要度的度量方法。并以此为基础，设计了启发式的半监督属性约简算法，算法在每次迭代过程中能剔除相对冗余的属性和当前属性约简集已能够区分的区分对，算法的搜索空间显著缩减。同时，根据数据中实例的动态变化情况，给出属性约简集的动态更新机制，并通过实例分析详细说明动态属性约简算法的计算过程。最后，采用来自 UCI 的真实数据集，进一步验证了算法的高效性和可行性。

1　基本知识

定义 1　四元组 $\mathrm{IS}=<U,A,V,f>$ 是一个信息系统，其中 U 表示实例的非空有限集合，称为论域；A 表示属性的非空有限集合 $V=\bigcup\limits_{a\in A}V_a$，$V_a$ 是属性 $a\in A$ 所有可能值的集合；f 表示 $U\times A\to V$，是一个信息函数，它为每个实例的每一个属性赋予一个值，即 $\forall a\in A$，$x\in U$，$f(x,a)\in V_a$。

给定信息系统，如果至少有一个属性 $a\in A$ 使得 V_a 含有缺失值，其中缺失值用 "$*$" 表示，此时该系统称为不完备信息系统，用 $\mathrm{IIS}=<U,A,V,f>$ 表示。

定义 2　在不完备信息系统 $\mathrm{IIS}=<U,A,V,f>$ 中，有 $A=A_d\cup A_r$，其中 A_d 表示离散型属性集，A_r 表示连续型属性集，对于任意 $B\subseteq A$，关于属性子集 B 的区分对定义为$\mathrm{Dis}_{\mathrm{UL}}(B,U^2)=\{<x_i,x_j>\}$。$\forall p\in B$，对 $\forall <x_i,x_j>$ 有

$$\begin{cases}\exists p\in A_r,|f(x_i,p)-f(x_j,p)|>\delta\\ \vee\exists p\in A_d,f(x_i,p)\neq f(x_j,p)\\ \wedge f(x_i,p)\neq *\wedge f(x_j,p)\neq *\end{cases}$$

给定的不完备信息系统 $\mathrm{IIS}=<U,A,V,f>$，对 $\forall B\subseteq A$，$|\mathrm{Dis}_{\mathrm{UL}}(B,U^2)|$ 表示属性集 B 重要度，其物理含义为属性集 B 的区分度。属性集区分的实例对的数量越多，该属性集越重要。由定义可知 $<x_i,x_j>\in\mathrm{Dis}_{\mathrm{UL}}(B,U^2)$ 满足自反性、对称性，因此在考虑实例之间的区分对 $<x_i,x_j>$ 后，则不再重复

考虑区分对 $<x_j,x_i>$。

定义 3　对于给定信息系统 $IS=<U,A,V,f>$, 令 $A=C\cup D$, 其中子集 C 是条件属性集合, D 是决策属性集合, 又称四元组为决策系统, 用 $DS=<U,C\cup D,V,f>$ 表示。

对于给定决策系统 $DS=<U,C\cup D,V,f>$, 如果至少有一个属性 $c\in C$ 使得 V_c 含有缺失值, 其中缺失值用 "$*$" 表示。此时, 该系统称为不完备决策系统, 用 $IDS=<U,C\cup D,V,f>$ 表示。

定义 4　给定不完备决策系统 $IDS=<U,C\cup D,V,f>$, 有 $C_d,C_r\subseteq B$, 其中 C_d 表示离散型属性, C_r 表示连续型属性, 设 $<x_i,x_j>\in U^2$, 对 $\forall p\in B$、$B\subseteq C$, D 关于 B 的区分对定义为 $\mathrm{Dis}_L(B,U^2)=<x_i,x_j>$(其中 L 表示有标记), 对 $\forall <x_i,x_j>$ 有

$$\left\{\begin{array}{l}\left\{\begin{array}{l}(\exists p\in C_r,|f(x_i,p)-f(x_j,p)|>\delta)\\ \vee\exists p\in C_d,f(x_i,p)\neq f(x_j,p)\\ f(x_i,p)\neq *\wedge f(x_j,p)\neq *\end{array}\right\}\wedge\\ f(x_i,D)\neq f(x_j,D)\end{array}\right.$$

给定的不完备决策系统 $IDS=<U,C\cup D,V,f>$, 对 $\forall B\subseteq C$, $|\mathrm{Dis}_L(B,U^2)|$ 表示属性集 B 的重要度, 其物理含义为属性集 B 的区分度。属性集能区分的实例对的数量越多, 则表明该属性集越重要。

由于在现实应用中, 在大量不完备数据中只有少部分实例存在标记, 若仅利用带标记的实例获取属性约简结果, 由于无标记的实例未得到有效的利用, 使得属性约简结果较难反映数据的整体信息, 分类算法难以学习有效的知识规则, 导致分类模型的分类性能较弱。为此, 针对弱标记不完备决策系统, 设计有效的属性重性度量方法显得尤为重要。本文在文献 [8,13] 的基础上, 构造了面向弱标记不完备数据的属性重要性度量方法。

给定的不完备决策系统 $IDS=<U,C\cup D,V,f>$, 若决策属性 d 存在缺失值。此时该系统称为弱标记不完备决策系统, 用 $WIDS=<U,C\cup D,V,f>$ 表示。

定义 5　给定弱标记不完备决策系统 $WIDS=<U,C\cup D,V,f>$, 设 $U=L\cup M$, 其中 L 表示有标记实例的集合, M 表示无标记实例的集合, 属性集 $B\subseteq C$ 的重要度定义为

$$\mathrm{Dis}(B,U^2)=\left\{\begin{array}{ll}\mathrm{Dis}_L(B,L^2)+\mathrm{Dis}_M(B,UL^2), & L\neq\varnothing,M\neq\varnothing\\ \mathrm{Dis}_L(B,L^2), & L\neq\varnothing,M=\varnothing\\ \mathrm{Dis}_M(B,UL^2), & L=\varnothing,M\neq\varnothing\end{array}\right.$$

性质 1　给定弱标记不完备决策系统 $WIDS=<U,C\cup D,V,f>$, 设 $U=L\cup M$, 其中 L 表示有标记实例的集合, M 表示无标记实例的集合, 对 $\forall a\in C-B$ 满足:

$$\mathrm{Dis}(B\cup a,U^2)\geqslant\mathrm{Dis}(B,U^2)$$

证明　根据定义 2 可得 $\mathrm{Dis}_M(B\cup a,U^2)\geqslant\mathrm{Dis}_M(B,U^2)$, 根据定义 4 有 $\mathrm{Dis}_L(B\cup a,U^2)\geqslant\mathrm{Dis}_L(B,U^2)$。从而可证 $\mathrm{Dis}_L(B\cup a,U^2)+\mathrm{Dis}_M(B\cup a,U^2)\geqslant\mathrm{Dis}_L(B,U^2)+\mathrm{Dis}_M(B,U^2)$, 即 $\mathrm{Dis}(B\cup a,U^2)\geqslant\mathrm{Dis}(B,U^2)$。

定义 6　给定弱标记不完备决策系统 $WIDS=<U,C\cup D,V,f>$, 设 $U=L\cup M$, 其中 L 表示有标记实例的集合, M 表示无标记实例的集合, $R\subseteq C$ 是一个属性约简, 当且仅当 R 满足:

1) $\mathrm{Dis}(R,U^2)=\mathrm{Dis}(C,U^2)$;

2) $\forall c\in R$, $\mathrm{Dis}(R-c,U^2)\neq\mathrm{Dis}(C,U^2)$。

定义 7　给定弱标记不完备系统 $WIDS=<U,C\cup D,V,f>$, 设 $U=L\cup M$, 其中 L 表示有标记实例的集合, M 表示无标记实例的集合, 并且 $B\subseteq C$, 对 $\forall c\in C-B$, 则

1) 属性 c 相对于属性集合 B 的区分度:

$$\mathrm{RSig}(c,B,U^2)=\left\{\begin{array}{ll}|\mathrm{Dis}(B\cup c,U^2)|-|\mathrm{Dis}(c,U^2)|, & B\neq\varnothing\\ |\mathrm{Dis}(c,U^2)|, & B=\varnothing\end{array}\right.$$

2) 当 $\mathrm{RSig}(c,B,U^2)=0$ 时, 则称属性 c 相对于 B 是不必要的 (相对冗余的)。

如图 1 所示, 利用 $|\mathrm{Dis}(B\cup c,U^2)|-|\mathrm{Dis}(c,U^2)|=\mathrm{RSig}(c,B,U^2)$ 计算属性 c 相对于属性集 B 的重要度时, 仅需要在属性集 B 无法区分的区分对中, 搜索属性 c 能够区分的实例对。由于相对重要度的引入, 在算法的迭代过程中, 可不断地剔除当前属性约简集已能够区分的实例对和相对冗余的属性, 使得算法的搜索空间不断缩减, 避免了大量的重复计算, 从而有效地减少算法的计算时间。

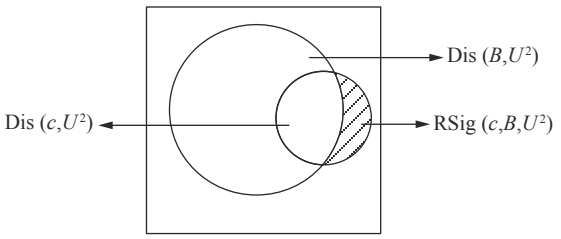


图 1　属性 c 相对于属性集 B 的重要度

Fig. 1　Significance of attribute c with respect to B

性质 2　给定弱标记不完备决策系统 $WIDS=<U,C\cup D,V,f>$, 设 $U=L\cup M$, 其中 L 表示有标记实例的集合, M 表示无标记实例的集合, 并且 $R\subseteq C$ 是 C 的一个属性约简集有

1) $\mathrm{RSig}(c,R,U^2)=0,\quad\forall c\in C-R$;

2) $\mathrm{RSig}(c,R-c,U^2)\neq 0,\quad\forall c\in R$。

证明　因为 $R\subseteq C$ 是 C 的一个属性约简集, 根据定义 6 有 $\mathrm{Dis}(R,U^2)=\mathrm{Dis}(C,U^2)$, 对 $\forall c\in C-R$ 有: $\mathrm{RSig}(c,R,U^2)=\left|\mathrm{Dis}(c,U^2)-\mathrm{Dis}(c,U^2)\cap\mathrm{Dis}(R,U^2)\right|=\left|\mathrm{Dis}(c,U^2)-\mathrm{Dis}(c,U^2)\cap\mathrm{Dis}(c,U^2)\right|=0$

证明了 1) 的充分性。

若 $R\subseteq C$ 是 C 的一个属性约简集, 对 $\forall c\in$

$C-R$, 假设 $\mathrm{RSig}(c,R,U^2)\neq 0$, 根据定义 6 有 $\mathrm{Dis}(R,U^2)=\mathrm{Dis}(C,U^2)$, 根据定义 7 可知:

$$\begin{aligned}\mathrm{RSig}(c,R,U^2)&=|\mathrm{Dis}(c,U^2)-\mathrm{Dis}(c,U^2)\cap\mathrm{Dis}(R,U^2)|\\&=|\mathrm{Dis}(c,U^2)-\mathrm{Dis}(c,U^2)\cap\mathrm{Dis}(C,U^2)|=0\end{aligned}$$

与假设矛盾。证明了 1) 的必要性, 同理可证 2)。

2　弱标记不完备决策系统的属性约简

基于属性重要度设计启发式属性约简算法, 被广泛地用于粗糙集理论的属性约简, 传统的启发式算法每次迭代时, 每次将最重要的属性增加到属性约简集中, 但是在迭代过程中无法剔除相对冗余的属性。本文提出一种基于区分对的向前启发式属性约简算法, 在每次迭代中, 将相对于当前属性约简集最重要的属性加入属性约简集中, 并剔除相对冗余属性和当前属性集已经能够区分的实例对, 快速缩减算法的搜索空间。

2.1　属性约简算法

算法 1　弱标记不完备决策系统属性约简算法 (WIDAR 算法)

输入　弱标记不完备决策系统 $\mathrm{WIDS}=<U,C,V,f>$;

输出　属性约简集 red。

1) $j=0$, $\mathrm{Pair}_j=\varnothing$, $C_j=C$, $C'=\varnothing$, $\mathrm{Pair}'=\varnothing$, $\mathrm{red}=\varnothing$;

2) 计算属性集 C 的区分对 $\mathrm{Pair}_j=\mathrm{Dis}(C,U^2)$;

3) while($|\mathrm{Pair}_j|\neq 0$)

{对 $c_k\in C_j$,计算 $\mathrm{RSig}(c_k,\mathrm{red},\mathrm{Pair}_j)$:

当 $\mathrm{RSig}(c_k,\mathrm{red},\mathrm{Pair}_j)=0$, c_k 相对 red 为冗余属性, $C'=C'\cup c_k$;

$c_k=\mathrm{argmax}\{\mathrm{RSig}(c_k,\mathrm{red},\mathrm{Pair}_j)|c_k\in C_j-C'\}$;

$\mathrm{red}=\mathrm{red}\cup c_k,\mathrm{Pair}'=\mathrm{Dis}(c_k,\mathrm{Pair}_j)$;

$C'=C'\cup c_k$;

$\mathrm{Pair}_{j+1}=\mathrm{Pair}_j-\mathrm{Pair}'$//剔除当前属性集已经能区分的区分对

$C_{j+1}=C_j-C'$//剔除已加入属性约简集的属性和相对冗余的属性

$j=j+1$;}

4) 对可能存在的冗余属性进行逆向剔除

计算 $\mathrm{Pair}=\mathrm{Dis}(\mathrm{red},U^2)$

对于 $\forall c_k\in\mathrm{red}$, 如果 $\mathrm{RSig}(c_k,\mathrm{red}-c_k,\mathrm{Pair})=0$, $\mathrm{red}=\mathrm{red}-c_k$;

5) 输出属性约简集 red

2.2　时间复杂度分析

算法 1 的 1) 初始化变量; 2) 根据定义 2 和定义 4 计算实例在属性集 C 上的区分对, 其时间复杂度为 $O(|U|^2|C|)$; 3) 是在性质 2 的基础上, 采用相对重要度为标准设计启发式算法, 在迭代过程中不断缩减算法的搜索空间, 其时间复杂度为 $O\left(\sum_{j=0}^{|\mathrm{red}|}|\mathrm{Pair}_j||C_j|\right)$; 4) 对可能存在的冗余属性进行剔除, 时间复杂度为 $O(|\mathrm{Pair}||\mathrm{red}|^2)$。

3　弱标记不完备决策系统的增量式属性约简

在动态数据中使用传统的属性约简方法, 往往无法快速获取约简结果, 而增量式属性约简算法能有效地利用原约简结果, 避免大量的重复计算。针对数据动态变化的情况, 本节分析了数据动态变化对实例区分对的影响, 图 2 给出了属性约简的增量式更新机制, 并针对动态数据中的实例变化情况, 算法 2 设计了属性约简增量式更新算法。

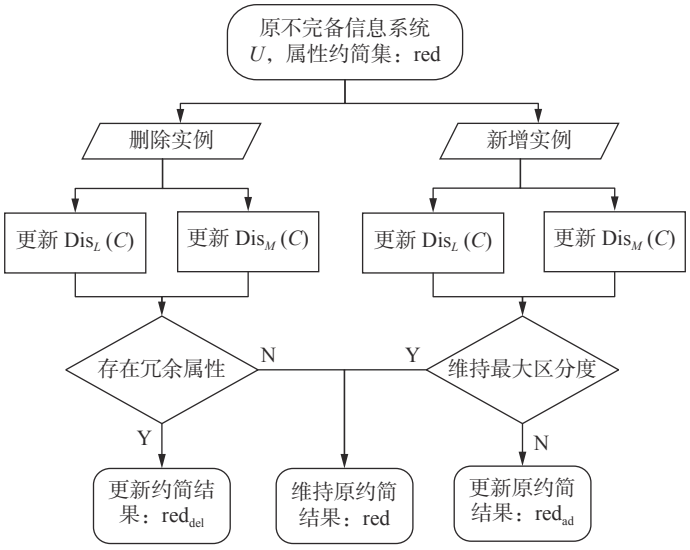


图 2　属性约简的增量式更新机制

Fig. 2　Incremental updating mechanism of attribute reduction

3.1　增量式属性约简算法

算法 2　弱标记不完备决策系统的增量式约简算法 (WIDIAR 算法)

输入　弱标记不完备决策系统 WIDS =< U, $C\cup D, V, f$ >, $U = L\cup M$, 删除的实例 $\Delta U_{de} = \Delta M_{de}\cup \Delta L_{de}$, 增加的实例 $\Delta U_{ad} = \Delta M_{ad}\cup \Delta L_{ad}$, 原属性约简结果 red;

输出　属性约简结果 red′。

1) 计算删除实例 ΔU_{de} 后存在的区分对 Pair_{de};

2) 对于 $\forall c_k \in \text{red}'$, $\text{red}' = \text{red}$, 如果满足条件: $\text{RSig}(c_k, \text{red}' - c_k, \text{Pair}_{de}) = 0$, 则根据定义 7 逆向剔除冗余属性, $\text{red}' = \text{red}' - c_k$;

3) 计算增加的区分对 ΔPair_{ad};

4) $j = 0, C' = \text{red}', C_j = C - C'$;

$\Delta\text{Pair}_{ad}^{j} = \Delta\text{Pair}_{ad}$;

while($\Delta\text{Pair}_{ad}^{j} \neq 0$) :

{ $\forall c_k \in C_j$, if($\text{RSig}(c_k, \text{red}, \text{Pair}_j) = 0$), $C' = C' \cup c_k$;

$c_k = \text{argmax}\{\text{RSig}(c_k, \text{red}', \Delta\text{Pair}_{ad}^{j}) | c_k \in C_j - C'\}$;
$\text{red}' = \text{red}' \cup c_k, C' = C' \cup c_k$;
$\text{Pair}' = \text{Dis}(c_k, \Delta\text{Pair}_{ad}^{j})$;
$C_{j+1} = C - C', \Delta\text{Pair}_{ad}^{j+1} = \Delta\text{Pair}_{ad}^{j} - \text{Pair}'$;

$j = j + 1$;}

5) 对可能存在的冗余属性进行逆向剔除, 计算 $\text{Pair} = \text{Dis}(\text{red}', ((U - \Delta U_{de}) \cup \Delta U_{ad})^2)$; $\forall c_k \in \text{red}'$, 如果 $\text{RSig}(c_k, \text{red}' - c_k, \text{Pair}) = 0$, $\text{red}' = \text{red}' - c_k$;

6) 输出新的属性约简结果 red′。

3.2　时间复杂度分析

算法 2 中: 1) 计算删除实例后属性集 red 的区分对, 其时间复杂度为 $O((U - \Delta U_{de})^2|\text{red}|)$; 2) 对可能存在的冗余属性逆向删除, 其时间复杂度为 $O((U - \Delta U_{de})^2|\text{red}|^2)$; 3) 计算由于增加实例 ΔU_{de} 而增加的区分对, 其时间复杂度为 $O(|U - \Delta U_{de}||\Delta U_{ad}||C|)$; 4) 若原属性约简集 red′ 无法维持最大区分度, 则需要在属性集 $C - C'$ 中筛选属性加入属性约简集, 其时间复杂度为 $O\left(\sum_{i=0}^{j}|\Delta\text{Pair}_{ad}^{i}||C_i|\right)$; 5) 对可能存在的冗余属性进行剔除, 时间复杂度为 $O(|\text{Pair}||\text{red}'|^2)$。

4　实例分析

为进一步详细说明算法的流程。以表 1 中的弱标记不完备数据为例, 进行分析说明。其中共有 6 个实例和 4 个属性, $U = \{x_1, x_2, x_3, x_4 x_5, x_6, x_7\}$ 为实例集, $C = \{c_1, c_2, c_3, c_4\}$ 为条件属性集, D 为决策属性, "*"表示缺失值, 阈值 $\delta = 0.1$。

表 1　弱标记不完备决策系统

Table 1　Incomplete decision system with weak labeling

U	c_1	c_2	c_3	c_4	D
x_1	a	b	*	0.71	1
x_2	a	*	b	0.58	0
x_3	a	b	a	0.80	1
x_4	b	a	*	0.65	*
x_5	b	a	b	0.72	*
x_6	b	*	b	0.72	*
x_7	a	b	a	0.77	*

根据表 1 详细描述算法 1 属性约简的具体步骤: 算法 1 中的 1) 变量初始化; 2) 计算关于属性集 C 的区分对, 确定算法的搜索空间 $\text{Pair}_0 = \{< x_1, x_2 >, < x_2, x_3 >, < x_4, x_7 >, < x_5, x_7 >, < x_6, x_7 >\}$; 在算法 1 的 3) 中第一次迭代时 $\text{red} = \varnothing$、$\text{RSig}(c_1, \text{red}, \text{Pair}_0) = 3$ 、$\text{RSig}(c_2, \text{red}, \text{Pair}_0) = 2$ 、$\text{RSig}(c_3, \text{red}, \text{Pair}_0) = 2$、$\text{RSig}(c_4, \text{red}, \text{Pair}_0) = 2$, 第一次迭代结束后, $\text{Pair}_1 = \{< x_1, x_2 >, < x_2, x_3 >\}$, $\text{red} = \{c_1\}$。由于此时 $|\text{Pair}_1| \neq 0$, 因此需进行第二次迭代, $\text{red} = \{c_1\}$, $\text{RSig}(c_2, \text{red}, \text{Pair}_1) = 0$、$\text{RSig}(c_3, \text{red}, \text{Pair}_1) = 0$、$\text{RSig}(c_4, \text{red}, \text{Pair}_1) = 2$, 第 2 轮迭代结束后, $\text{Pair}_2 = \varnothing$、$\text{red} = \{c_1, c_4\}$ 循环结束; 算法 1 在 4) 中未发现冗余属性, 最后输出属性约简结果 $\text{red} = \{c_1, c_4\}$。

为进一步详细说明算法 2 针对不同实例变化后对原属性约简结果产生的影响, 下面首先分析算法 2 的 2) 中实例动态删除情况。在数据中删除实例将存在 6 种情况: 1) 只删除有标记的实例 $\{x_1\}$, 对于 $\forall c_k \in \text{red} = \{c_1, c_4\}$ 有 $\text{RSig}(c_k, \text{red} - c_k, \text{Pair}) \neq 0$, 属性约简集保持不变 $\text{red}' = \{c_1, c_4\}$; 2) 只删除有标记的实例 $\{x_2\}$, 有 $\text{RSig}(c_4, \text{red} - c_4, \text{Pair}) = 0$, 此时 c_4 相对于 c_1 为冗余属性, 属性约简集更新为 $\text{red}' = \{c_1\}$; 3) 只删除无标记的实例 $\{x_4\}$, 对 $\forall c_k \in \text{red} = \{c_1, c_4\}$, $\text{RSig}(c_k, \text{red} - c_k, \text{Pair}) \neq 0$, 属性约简集保持不变 $\text{red}' = \{c_1, c_4\}$; 4) 只删除无标记的实例 $\{x_7\}$, 有 $\text{RSig}(c_4, \text{red} - c_4, \text{Pair}) = 0$, 此时 c_4 相对于 c_1 为冗余属性, 属性约简集更新为 $\text{red}' = \{c_1\}$; 5) 同时删除有标记的实例 $\{x_1\}$ 和无标记的实例 $\{x_4\}$, 对 $\forall c_k \in \text{red} = \{c_1, c_4\}$, $\text{RSig}(c_k, \text{red} - c_k, \text{Pair}) \neq 0$, 属性约简集维持不变 $\text{red}' = \{c_1, c_4\}$; 6) 同时删除有标记的实例 $\{x_1\}$ 和无标记的实例 $\{x_7\}$, 有 $\text{RSig}(c_1, \text{red} - c_1, \text{Pair}) = 0$, 此时 c_1 相对于 c_4 为冗余属性, 属性约简集更新为 $\text{red}' = \{c_4\}$。

由于在许多现实动态场景中不仅存在数据删除还有数据动态增加的情况, 为了进一步分析算法 2 的 3) 和 4) 的计算流程, 在上述实例中若在原始数据集中删除实例 $U_{de} = \{x_1, x_7\}$, 删除实例后属性约简集为 $\text{red}' = \{c_4\}$, 在此基础上, 在数据中增

加实例将存在如下 6 种情况：1) 只增加有标记实例 $\{x_7\}=\{[a,b,*,0.62,0]\}$，原属性约简能够维持最大区分度，根据性质 2 属性约简集维持不变，$\mathrm{red}'=\{c_4\}$；2) 只增加有标记实例 $\{x_7\}=\{[a,a,*,0.62,0]\}$，原属性约简无法维持最大区分度，$\exists c\in C-\mathrm{red}'$ 使得 $\mathrm{RSig}(c,C-c,\Delta\mathrm{Pair}_{\mathrm{ad}})\neq 0$，根据性质 2 可知 $\mathrm{red}'=\{c_2\}$ 不满足约简条件，在算法 2 的 4) 中属性约简更新为 $\mathrm{red}'=\{c_4,c_2\}$；3) 只增加无标记的实例 $\{x_7\}=\{[b,a,b,0.68,*]\}$，同理根据性质 2 属性约简集维持不变，$\mathrm{red}'=\{c_4\}$；4) 只增加无标记的实例 $\{x_7\}=\{[b,a,*,0.62,0]\}$，原属性约简集无法维持最大区分度，同理属性约简更新为 $\mathrm{red}'=\{c_4,c_1\}$；5) 同时增加有标记的实例 $\{x_7\}=\{[a,b,*,0.77,1]\}$ 和无标记的实例 $\{x_8\}=\{[b,a,a,0.65,*]\}$，同理属性约简集维持不变，$\mathrm{red}'=\{c_4\}$；6) 同时增加有标记的实例 $\{x_7\}=\{[a,b,*,0.77,1]\}$ 和无标记的实例 $\{x_8\}=\{[b,a,a,0.65,*]\}$，原属性约简无法维持最大区分度，同理可得属性约简集更新为 $\mathrm{red}'=\{c_4,c_3\}$。

通过上述实例分析可知，本文算法 1 采用相对重要度为属性重要度的度量标准，在迭代过程中不断剔除当前属性集已能够区分的区分对和相对冗余的属性，使得每次迭代的搜索空间不断缩减，避免了大量的重复计算。算法 2 通过分析实例动态变化对原属性约简集的影响，在实例变化后动态获取属性约简集，无需重新计算属性约简集。在删除实例后，对可能存在的冗余属性逆向剔除；增加实例后，通过搜索原属性约简集无法辨识的区分对，确定算法的搜索空间。为弱标记混合数据的属性约简提供了一种可借鉴的处理方法。

5　实验分析

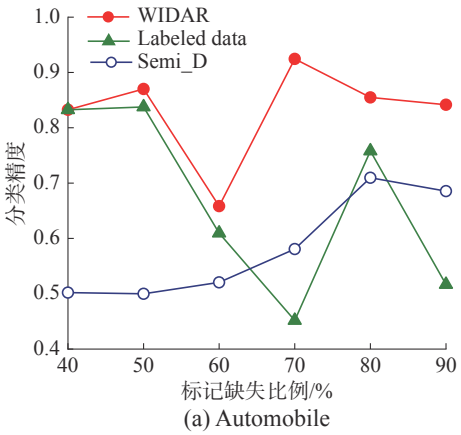
为进一步验证本文算法的有效性，从 UCI 数据集中选取了 6 个真实数据集进行测试和分析，数据的详细信息如表 2 所示。实验的运行环境为：CPU Intel(R)Core(TM)i5-6500(3.20 Hz),内存 8.0 GB，操作系统为 Windows 10，采用 Python 编程语言，开发工具为 Pycharm 2018.2.4。

表 2　数据集描述

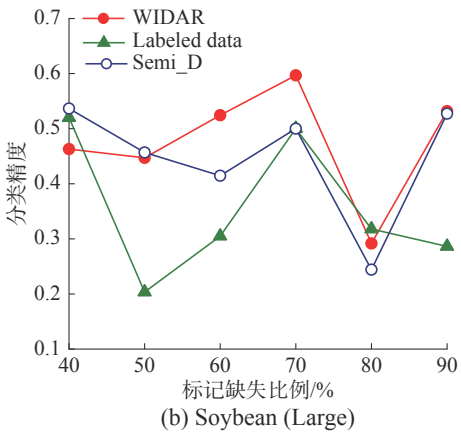
Table 2　Description of UCI data sets

数据集	实例个数	属性个数	属性值缺失
Automobile	205	26	是
Soybean(Large)	307	35	是
Dermatology	366	34	是
Cylinder Bands	512	40	否
shroom	8 124	22	是
Letter Recognition	20 000	17	否

本节详细讨论标记缺失对算法 1 的影响，首先以数据集的 40% 为基础数据，数据集大小的 10% 为梯度递增，对数据集的标记进行随机缺失处理。然后分别对弱标记的数据 (weak labeled data) 采用算法 1(WIDAR)、Semi_D 算法[13]、Semi_P 算法[10] 进行属性约简，并和算法 1 对有标记数据 (Labeled data) 的约简结果进行比较分析。对属性约简结果的分类性能评估，将采用 KNN、CART、Naive Bayes 三个分类器的精度作为约简结果的评价指标，将 Automobile、Soybean、Dermatology、Cylinder Bands 数据集随机分为两部分，一部分作为训练集，另一部分作为测试集，获取分类精度；Mushroom 和 Letter Recognition 数据集采用 10 倍交叉验证，获取分类精度。针对数据中的连续型属性，本文的 δ 计算方式为 $\delta=(S_i/n)/\lambda$，其中，S_i 为每个连续型属性的标准差，S_i/n 为连续型属性标准差的平均值，由于每个数据集的连续型属性的平均标准差为固定值，δ 的取值由 λ 决定[22]。在本文中先对连续型的属性采用 Min-Max Normalization 归一化方法处理，λ 取 0.6。由于 Semi_P 和 Semi_D 算法的属性约简结果的分类精度基本相同，本节以 Semi_D 算法为例进行比较分析，实验结果如图 3~5 所示。



(a) Automobile



(b) Soybean (Large)

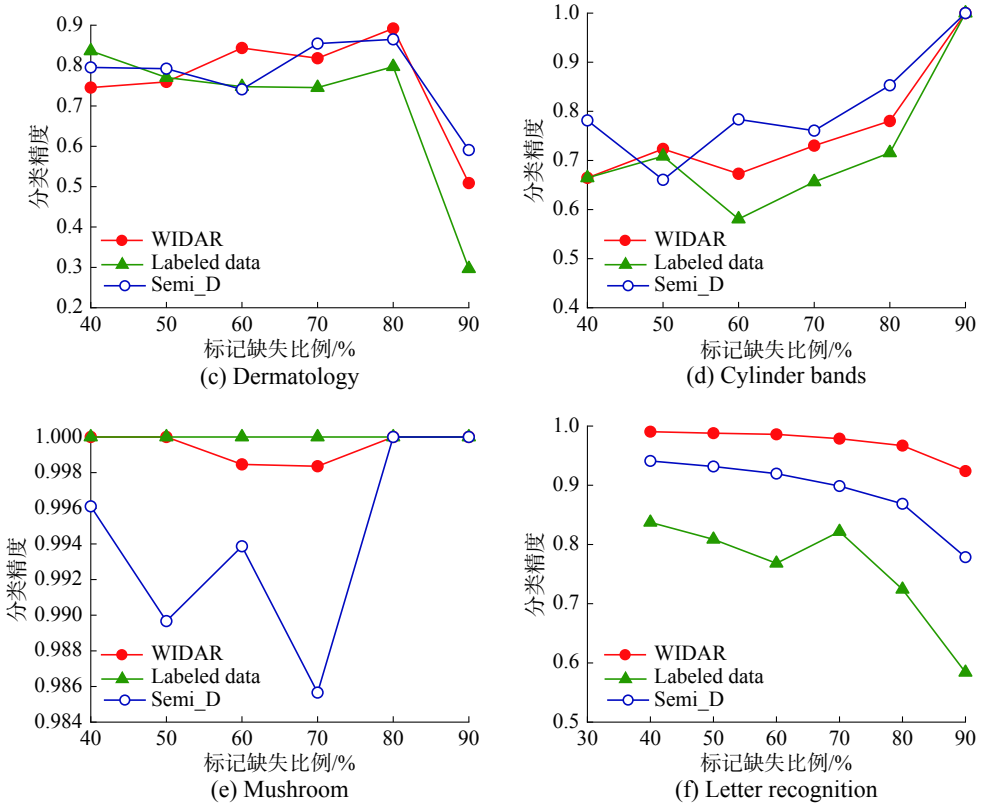
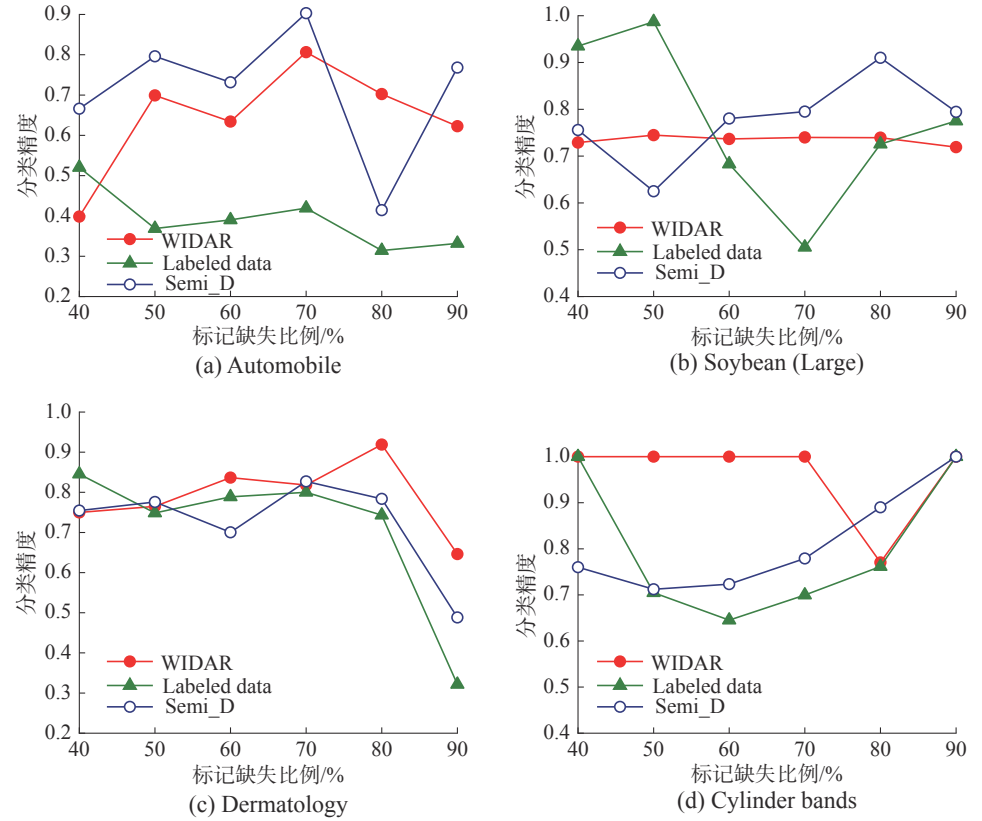


图 3　3NN 分类器的分类精度

Fig. 3　Classification accuracy with the 3NN classifier



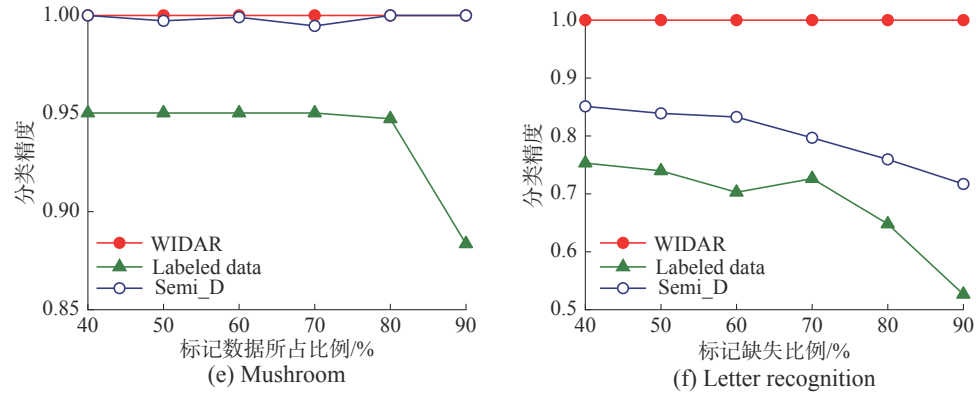


图 4　CART 分类器的分类精

Fig. 4　Classification accuracy with the CART classifier

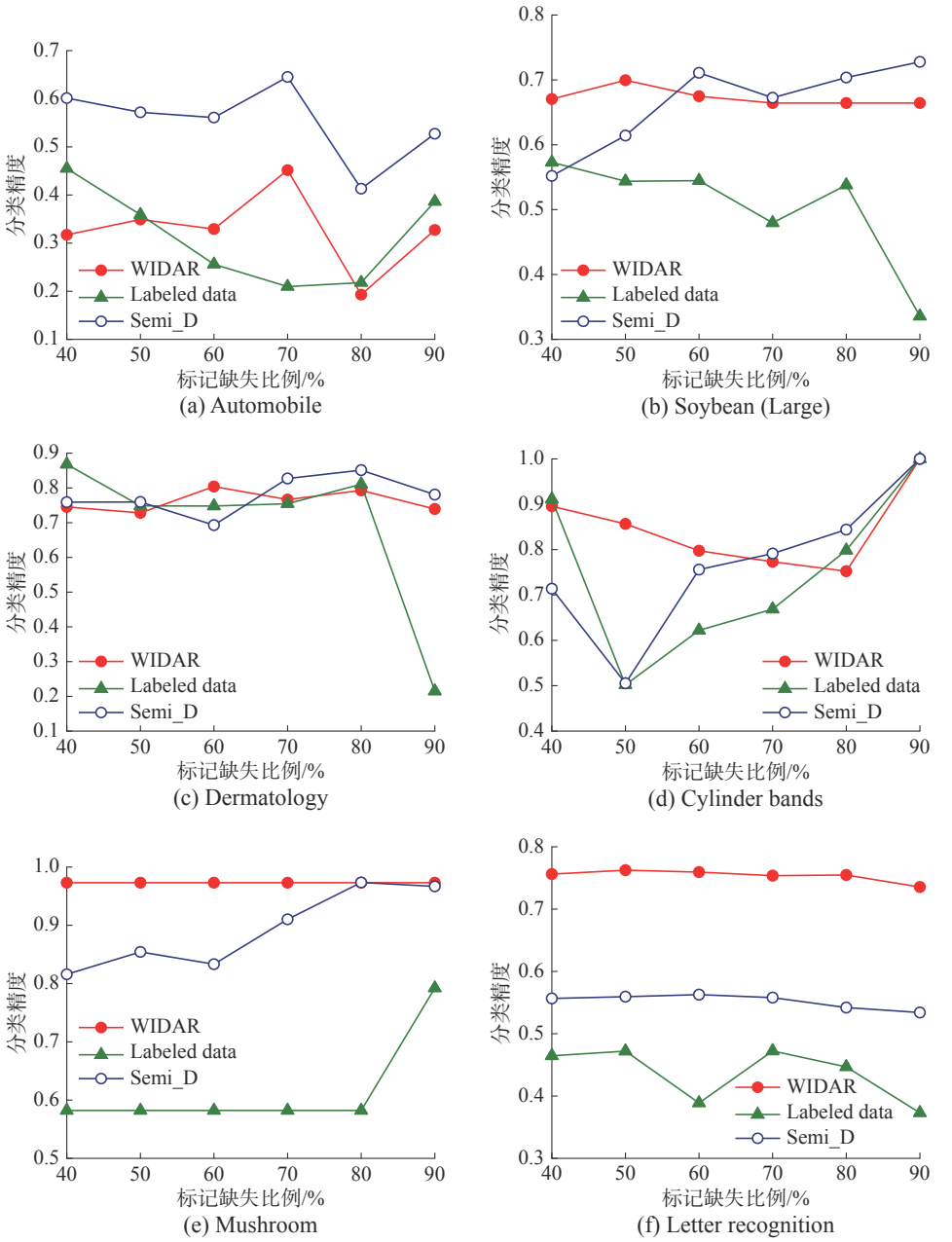


图 5　Naive Bayes 分类器分类精度

Fig. 5　Classification accuracy with the Naive Bayes classifier

由图 3~5 可知, WIDAR 算法能够有效利用有标记和无标记的数据, 获取分类性能较优的属性约简集。特别地, 随着标记缺失比例的增加, 若采用 WIDAR 算法仅处理有标记数据 (Labeled data) 获取的属性约简结果, 由于信息利用率的下降, 分类器难以学习到较好的分类模型, 在 3 个分类器中分类性能总体偏低, 分类精度较低并且不稳定。相反, 充分利用弱标记数据获取的属性约简结果, 能够较好地利用原数据集的信息。分类模型的分类精度较高且稳定, 分类性能较优。在前 4 个数据集中, 由于数据的规模较小, 使得分类器较难准确学习到其内在规则或模式, 且对小数据集的标记进行随机缺失, 对分类效果产生了一定影响。因此不同分类器在同一属性约简结果上的分类表现差异较大, 但仅利用有标记的数据获取属性约简结果的分类性能显著偏弱。随着数据规模增大, 在 Mushroom、Letter Recognition 数据集中, 随着标记缺失比例的增加, WIDAR 算法在有标记的数据中和 Semi_D 算法在弱标记数据中获取的属性约简结果分类精度出现较大的波动, 但采用 WIDAR 算法利用弱标记的数据获取的属性约简结果分类精度稳定且对比 Semi_D 算法有较明显的优势。以数据集 Letter Recognition 为例, 在图 (3) 的 (f) 中随着标记缺失比例的增加, WIDAR 算法使用弱标记的数据获取的属性约简结果在 KNN 分类器上的分类精度仅出现了较小的波动, 在标记缺失比例为 40% 时分类精度为 99.0%, 在标记缺失比例为 90% 时, 分类精度仍有 92.4%, 与 Semi_D 算法相比分类精度从 83.7% 下降到 58.4%。而在 CART 和 Naive Bayes 分类器中, WIDAR 算法的表现更加稳定, CART 分类器的分类精度稳定在 100%, Naive Bayes 分类器的分类精度则在 73.6%~76.3%, 与 Semi_D 算法相比具有较明显的优势。

为了进一步详细分析属性约简结果, 本文以数据标记缺失比率为 50% 的情况为例, 在表 3 中列举出了详细的属性约简结果,属性 C_i 简写为 i。在同一数据集上, 不同的算法属性结果存在一定的差异, 结合图 5 和表 3 中的信息可知, 仅利用有标记的数据获取的属性约简结果会丢失部分有效的分类信息。Semi_D 算法和 Semi_P 算法中 Automobile 和 Cylinder Bands 数据集的约简结果完全相同, 而这两个算法在其他数据集中属性约简结果大致相同, 分类精度相近, 为此在讨论属性约简结果的分类性能时, 本文以 Semi_D 算法为例。与 Semi_D 算法和 WIDAR 算法仅利用有标记数据 (Labeled data) 相比较, 本文提出的 WIDAR 算法能够获取一个分类性相对较优的属性约简结果。在实验过程中发现, 本文的 WIDAR 算法在规模较小的数据集上对比 Semi_D 算法, 其分类性能有时存在效果偏弱的情况, 但随着数据规模的增大, WIDAR 算法的性能表现趋于稳定, 并且对比 Semi_D 算法存在较明显的优势。综上可得, 本文的算法在大数据集中能够有效利用无标记的数据, 增强属性约简结果的分类性能, 显著提升了算法的鲁棒性。

表 3　属性约简结果的对比

Table 3　Comparison of attribute reduction results

数据集	WIDAR算法	Labeled data	Semi_D算法	Semi_P算法
Automobile	3,24,7,6,8,5	3,24,7,6	1,22,5,4,6,3,8	1,22,5,4,6,3,8
Soybean(Large)	1,7,6,15,10,17,35,4,12,8,21,3,30	1,7,6,22,10,35,3,2	1,7,6,10,4,22,3,8,9,30	1,29,3,4,6,7,10,8,9,30
Dermatology	16,4,3,19,2,32,17,26,18,5	16,4,3,19,2,32,17	34,16,4,19,3,2,17,5,13	34,17,29,28,3,17,1,32,16
Cylinder Bands	2,35,25,3	2,4,14	1,24,2	1,24,2
Mushroom	9,3,22,1,2,15,5,21,14,13,20,12,17,7,6	5,20,22,21	5,9,3,21,13,22,1,15,2,14,20,12,17,7,6	9,3,22,1,2,15,5,21,14,13,20,12,17,7,6
Letter Recognition	2,15,8,9,11,3,6,4,7,1,12,10,5,16,13,14	2,15,8,9,11,12,13,10,6,1	2,15,8,9,11,3,6,4,7,1,12,10,5,16,13,14	2,10,7,8,15,9,11,3,12,13,6,4,1,5,16,14

综上可知, 无标记数据也内含丰富的分类信息, 仅利用有标记的数据获取属性约简集, 往往会丢失部分信息, 导致分类器的分类性能偏低。充分利用有标记数据和无标记数据, 获取的属性约简结果, 在分类器中的表现较优。WIDAR 算法在对大数据进行属性约简时, 能够快速获取分类性能较优的属性约简结果, 且算法具有良好的鲁棒性。另外, 为了进一步说明算法 2 (WIDIAR 算法) 的有效性, 在数据集的标记随机缺失 50% 后, 将 6 组数据集划分为基准数据集和候选数据集两部分, 取原数据集的前 50% 作为基准数据集。为应对现实应用领域中常见的复杂应用场景, 每次

实验剔除基准数据集中 5% 的数据后增加数据，并且增加的数据以候选数据集的 10% 为梯度增加，与 WIDAR 算法、Semi_P 算法和 Semi_D 算法进行对比分析，实验结果如图 6 所示。

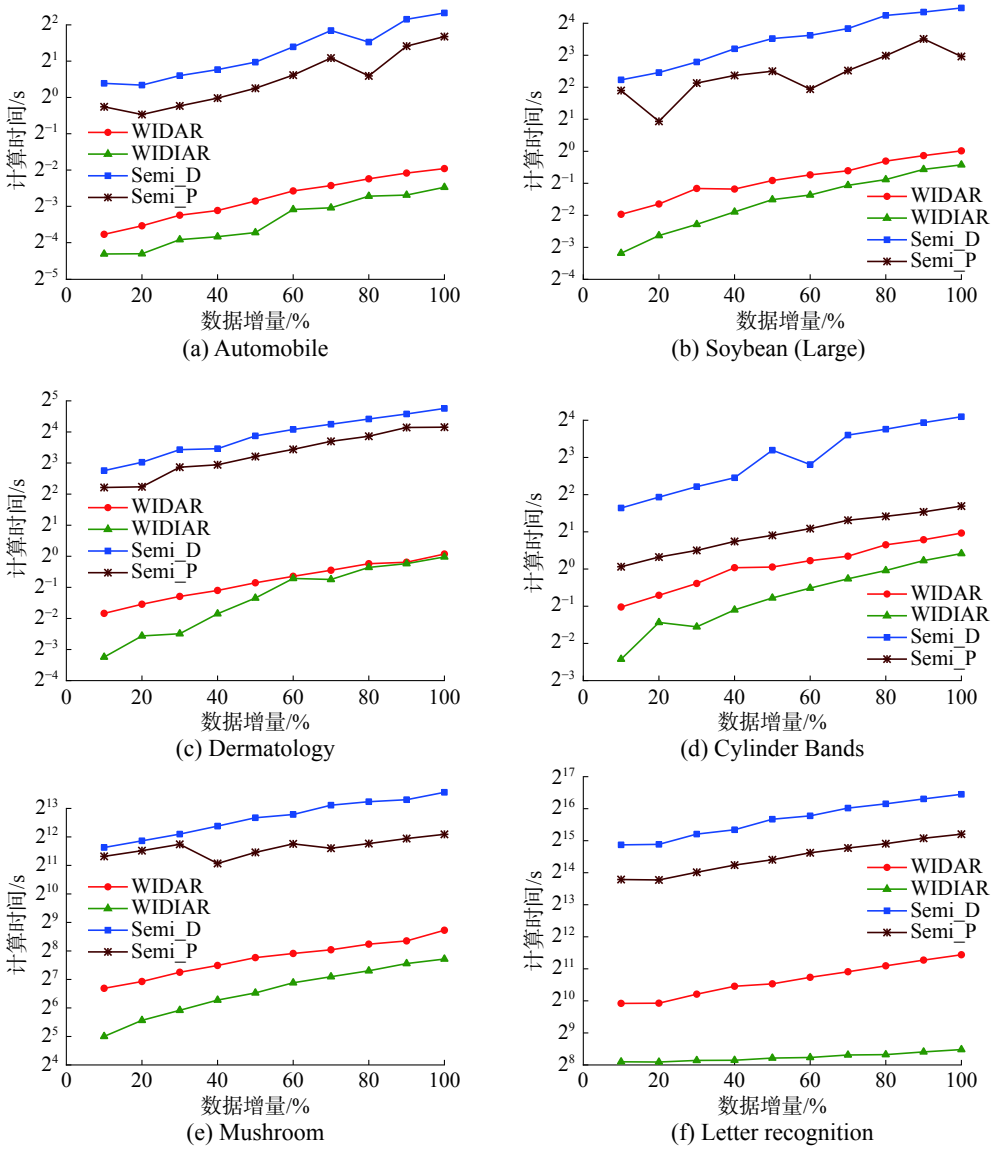


图 6　4 种算法的计算时间对比

Fig. 6　Comparison of time consumption of four algorithms

由图 6(a)~(c) 可知，WIDAR 算法引入了相对重要度作为属性重要度的度量标准，在迭代中不断缩减算法搜索空间。因此在处理数据规模较小的数据集时，WIDIAR 算法与 WIDAR 算法计算效率相近，但相比 Semi_P 算法和 Semi_D 算法有较明显的优势。但随着数据规模的增大，WIDAR 算法在动态数据中更新属性约简结果需要进行大量的重复计算。采用 WIDIAR 算法，对属性约简集进行增量式更新，能够有效减少重复的计算，相比 WIDAR 算法，能够节约大量的时间。当增加的数据集的大小为 100% 时，在 Mushroom 数据集中，采用动态属性约简算法动态更新属性约简结果需要的时间仅为 210.156 s，而采用静态属性约简算法，获取属性约简需要花费 423.521 s，与静态属性约简算法相比较能够节约 50.38% 的时间。在 Letter Recognition 数据集中，采用动态属性约简算法动态更新属性约简结果需要的时间仅为 356.823 s，而采用 WIDAR 算法获取属性约简需要花费 2 776.922 s，与 WIDAR 算法相比较能够节约 87.15% 的时间。

表 4 为增加的数据达到 100% 时，WIDIAR 算法和 WIDAR 算法的属性约简结果的对比，属性 C_i 简写为 i。从表 4 中可以看到，WIDIAR 算法的属性约简结果与 WIDAR 算法相比，在较小的数据集中存在一定差异，但随着数据规模的增加，算法的属性约简结果差异逐步缩减，在 Mushroom

和 Letter Recognition 数据集中两者的属性约简结果完全相同。由此可知, 本文提出的 WIDIAR 算法能够有效节约大量计算时间, 同时能获取分类性能较优的属性约简结果。为大规模复杂数据的属性约简问题, 提供了一个可行的增量式属性约简方法。

表 4　属性约简结果对比

Table 4　Comparison of attribute reduction results

数据集	WIDIAR算法	WIDAR算法
Automobile	13,25,2	24,7,6,25,1
Soybean(Large)	1,7,6,15,10,17,4,35,12,8,3,21,30	1,7,6,15,10,35,4,21,3,22,2,28,9,13,8,17,11,16,19,24,30,5,12,14,31
Dermatology	34,4,32,3,16,1,2	4,3,19,2,32,17,34
Cylinder Bands	2,35,25,3	2,29,22,3
Mushroom	9,3,22,1,2,15,5,21,14,13,20,12,17,7,6	9,3,22,1,2,15,5,21,14,13,20,12,7,17,6
Letter Recognition	2,15,8,9,11,3,6,4,7,1,12,10,5,16,13,14	2,15,8,9,11,3,6,4,7,1,12,10,5,16,13,14

6　结束语

在众多的现实应用领域中, 由于数据采集技术和成本的限制, 存在大量的不完备高维数据, 且通常只有少量的数据存在标记信息。若仅利用有标记的数据获取属性约简结果, 将会丢失部分有效的信息。为了有效地利用弱标记不完备的高维数据, 从中获取分类效果更优的属性约简集, 本文基于实例的区分对, 提出属性相对重要度的概念, 设计了启发式属性约简算法。针对动态变化场景, 详细分析了数据的动态变化对属性约简的影响和更新机制。并在此基础上, 提出了增量式的属性约简算法。实验结果表明, 该算法在处理大规模数据时, 相比静态属性约简算法, 在保证分类性能的同时, 能够高效地获取属性约简结果。下一步工作将拓展增量式属性约简算法应用场景, 考虑属性集变化后, 属性约简集的更新问题。

参考文献:

[1] PAWLAK Z, SKOWRON A. Rough sets: some extensions[J]. Information sciences, 2007, 177(1): 28–40.

[2] 王国胤, 姚一豫, 于洪. 粗糙集理论与应用研究综述 [J]. 计算机学报, 2009, 32(7): 1229–1246.
WANG Guoyin, YAO Yiyu, YU Hong. A survey on rough set theory and applications[J]. Chinese journal of computers, 2009, 32(7): 1229–1246.

[3] HU Qinghua, LIU Jinfu, YU Daren. Mixed feature selection based on granulation and approximation[J]. Knowledge-based systems, 2008, 21(4): 294–304.

[4] 王映龙, 曾淇, 钱文彬, 等. 变精度下不完备混合数据的增量式属性约简方法 [J]. 计算机应用, 2018, 38(10): 2764–2771.
WANG Yinglong, ZENG Qi, QIAN Wenbin, et al. Incremental attribute reduction method for incomplete hybrid data with variable precision[J]. Journal of computer applications, 2018, 38(10): 2764–2771.

[5] MA Fumin, DING Mianwei, ZHANG Tengfei, et al. Compressed binary discernibility matrix based incremental attribute reduction algorithm for group dynamic data[J]. Neurocomputing, 2019, 344: 20–27.

[6] QIAN Yuhua, LIANG Jiye, PEDRYCZ W, et al. Positive approximation: an accelerator for attribute reduction in rough set theory[J]. Artificial intelligence, 2010, 174(9/10): 597–618.

[7] LIANG Jiye, MI Junrong, WEI Wei, et al. An accelerator for attribute reduction based on perspective of objects and attributes[J]. Knowledge-based systems, 2013, 44: 90–100.

[8] DAI Jianhua, HU Qinghua, HU Hu, et al. Neighbor inconsistent pair selection for attribute reduction by rough set approach[J]. IEEE transactions on fuzzy systems, 2018, 26(2): 937–950.

[9] TENG Shuhua, LU Min, YANG AFeng, et al. Efficient attribute reduction from the viewpoint of discernibility[J]. Information sciences, 2016, 326: 297–314.

[10] QIAN Yuhua, LIANG Jiye, LI Deyu, et al. Approximation reduction in inconsistent incomplete decision tables[J]. Knowledge-based systems, 2010, 23(5): 427–433.

[11] MENG Zuqiang, SHI Zhongzhi. A fast approach to attribute reduction in incomplete decision systems with tolerance relation-based rough sets[J]. Information sciences, 2009, 179(16): 2774–2793.

[12] XIE Xiaojun, QIN Xiaolin. A novel incremental attribute reduction approach for dynamic incomplete decision systems[J]. International journal of approximate reasoning, 2018, 93: 443–462.

[13] DAI Jianhua, HU Qinghua, ZHANG Jinghong, et al. Attribute selection for partially labeled categorical data by rough set approach[J]. IEEE transactions on cybernetics, 2017, 47(9): 2460–2471.

[14] 王锋, 刘吉超, 魏巍. 基于信息熵的半监督特征选择算法 [J]. 计算机科学, 2018, 45(S2): 427–430.
WANG Feng, LIU Jichao, WEI Wei. Semi-supervised feature selection algorithm based on information entropy[J]. Computer science, 2018, 45(S2): 427–430.

[15] 张维, 苗夺谦, 高灿, 等. 基于粗糙集成学习的半监督属性约简 [J]. 小型微型计算机系统, 2016, 37(12): 2727–2732.
ZHANG Wei, MIAO Duoqian, GAO Can, et al. Semi-supervised data attribute reduction based on rough-subspace ensemble learning[J]. Journal of Chinese computer systems, 2016, 37(12): 2727–2732.

[16] XU Zenglin, KING I, LYU M R T, et al. Discriminative semi-supervised feature selection via manifold regularization[J]. IEEE transactions on neural networks, 2010, 21(7): 1033–1047.

[17] HAN Yahong, YANG Yi, YAN Yan, et al. Semisupervised feature selection via spline regression for video semantic recognition[J]. IEEE transactions on neural networks and learning systems, 2015, 26(2): 252–264.

[18] LINAG Jiye, WANG Feng, DANG Chuangyin, et al. A group incremental approach to feature selection applying rough set technique[J]. IEEE transactions on knowledge and data engineering, 2014, 26(2): 294–308.

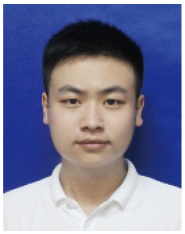
[19] LI Shaoyong, LI Tianrui, HU Jie. Update of approximations in composite information systems[J]. Knowledge-based systems, 2015, 83: 138–148.

[20] JING Yunge, LI Tianrui, LUO Chuan, et al. An incremental approach for attribute reduction based on knowledge granularity[J]. Knowledge-based systems, 2016, 104: 24–38.

[21] SHU Wenhao, QIAN Wenbin. An incremental approach to attribute reduction from dynamic incomplete decision systems in rough set theory[J]. Data & knowledge engineering, 2015, 100: 116–132.

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

作者简介：



程龙，硕士研究生，主要研究方向为粒计算与知识发现。

钱文彬，副教授，博士，主要研究方向为粒计算、知识发现与机器学习。主持国家自然科学基金项目 2 项，江西省自然科学基金项目 2 项。发表学术论文 30 余篇。

王映龙，教授，博士，主要研究方向为知识发现与数据挖掘。参与国家自然科学基金项目 2 项，主持江西省自然科学基金项目 3 项。发表学术论文 20 余篇。