计算机科学
COMPUTER SCIENCE

动态部分标记混合数据的增量式特征选择算法

闫振超，舒文豪，谢昕

引用本文

闫振超，舒文豪，谢昕. 动态部分标记混合数据的增量式特征选择算法[J]. 计算机科学, 2022, 49(11): 98-108.

YAN Zhen-chao, SHU Wen-hao, XIE Xin. Incremental Feature Selection Algorithm for Dynamic Partially Labeled Hybrid Data[J]. Computer Science, 2022, 49(11): 98-108.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210900076

动态部分标记混合数据的增量式特征选择算法

闫振超　舒文豪　谢　昕
华东交通大学信息工程学院　南昌 330013
(zhenchao_yan@163.com)

摘　要　许多实际应用中的数据集是由符号型、数值型和缺失型特征构成的混合数据。针对混合数据的决策标记,由于获取全部数据的决策标记需要耗费大量的人工和时间成本,只能为部分数据进行决策标记,因此产生了部分标记数据。同时,现实应用领域中数据是动态产生的,即数据维度随着不同的需求动态地增加或删减。针对混合数据的高维性、部分标记和动态性,文中提出了两种面向部分标记混合数据的增量式特征选择算法。首先,利用信息粒度对部分标记混合数据的特征进行重要度分析;其次,当特征集发生动态变化时,结合增量学习的思想,给出信息粒度的增量更新机制;然后,在此基础上提出了两种面向部分标记混合数据的增量式特征选择算法;最后,通过与其他算法在UCI数据集上的实验结果进行对比,进一步验证了所提算法的可行性和有效性。

关键词:混合数据;部分标记;增量学习;信息粒度;特征选择
中图法分类号　TP391

Incremental Feature Selection Algorithm for Dynamic Partially Labeled Hybrid Data

YAN Zhen-chao,SHU Wen-hao and XIE Xin
School of Information Engineering,East China Jiaotong University,Nanchang 330013,China

Abstract　Many real-world data sets are hybrid data consisting of symbolic,numerical and missing features. For the decision labels of hybrid data,it costs much labor and it is expensive to acquire the decision labels of all data,thus the partially labeled data is generated. Meanwhile,the data in real-world applications change dynamically,i.e.,the feature set is added into and deleted from the feature sets dynamically with different requirements. In this paper,according to the characteristics of high-dimensional,partial labeled and dynamic for the hybrid data,the incremental feature selection algorithms are proposed. Firstly,the information granularity is used to analyze the feature significance for partially labeled hybrid data. Then,the incremental updating mechanisms for information granularity are proposed with the variation of a feature set. On this basis,the incremental feature selection algorithms are proposed for the partially labeled hybrid data. Finally,extensive experimental results on UCI data set demonstrate that the proposed algorithms are feasible and efficient.

Keywords　Hybrid data,Partially labeled,Incremental learning,Information granularity,Feature selection

1　引言

随着物联网、人工智能等信息技术的发展,高维数据呈现几何式增长。数据的高维特性降低了算法的运行效率并且影响其分类性能。特征选择[1-3]作为一种数据预处理的有效方法,通过删除不相关和冗余特征,有效地降低了数据维度,提高了数据的紧密度以及分类能力。粗糙集理论[4]作为粒计算的一种重要理论,已被广泛地应用于特征选择、知识发现和数据挖掘等领域[5-8]。该理论最大的优势是不需要提供数据本身以外的任何先验知识,直接处理数据,并从中挖掘潜在的有用知识。

针对现实应用中的数据集,大部分数据都是由符号型、数值型和缺失型等特征构成的混合数据[9-11]。同时,数据的标记信息获取也比较"昂贵",需要耗费昂贵的资源或者很长的实验过程进行人工标记才能得到标记数据,由此产生了部分标记的混合数据,即只能为混合数据的决策进行部分标记。如何从部分标记的混合数据中进行特征选择是当前知识发现、数据挖掘和智能信息处理等领域的一个研究热点,并得到了众多研究学者的广泛关注和研究[12-15]。Wang等[12]针对部分标记的符号型数据,提出了一种基于互补信息熵的特征选择算法。Dai等[13]针对部分标记的符号型数据,提出了基于可辨矩阵的特征选择算法。Liu等[14]针对部分标记的数值型

到稿日期:2021-09-09　返修日期:2021-12-28
基金项目:国家自然科学基金(61662023,61762037);江西省自然科学基金(20202BABL202037)
This work was supported by the National Natural Science Foundation of China(61662023,61762037) and Natural Science Foundation of Jiangxi Province(20202BABL202037).
通信作者:舒文豪(shuwenhao@126.com)

数据，提出了一种基于集成分类器的半监督特征选择算法。Xiao 等[15]针对符号型与数值型构成的部分标记混合数据，提出了一种基于属性依赖度和混合约束条件的半监督特征选择算法。综上，以上面向部分标记数据的特征选择算法主要是面向静态数据。

然而，在大数据的背景下，现实应用中的数据集往往是动态变化的[16-26]，即数据中的对象、特征或者特征值随时间不断变化。针对动态变化数据的特征选择，静态算法无法利用原有的特征选择结果进行大量的重复计算，这将耗费较多的时间。而增量学习作为处理动态数据的一种主要方法，在原特征子集的基础上通过更新局部数据结构快速得到新的特征子集，从而引起了众多研究人员的关注。Ma 等[16]针对符号型数据中对象集的动态变化，提出了一种基于压缩二元差别矩阵的增量式属性约简算法。Huang 等[18]针对由符号型和缺失型数据构成的不完备混合数据，提出了对象集增加或者删除时，正区域、边界域和负区域的增量式更新算法。

本文重点回顾了数据中特征集发生动态变化时特征选择的研究成果。Zeng 等[20]针对由布尔型、分类型、区间型和集值型特征构成的混合数据，利用混合距离与高斯核函数来度量不同对象之间的相似性，提出了单个特征发生动态变化时，基于模糊粗糙集的特征选择更新机制。Yu 等[21]针对区间值有序信息系统，提出了特征增加或者删减时下近似的增量更新机制。Cai 等[22]针对有序决策系统，提出了当特征集发生动态变化时细覆盖与粗覆盖的增量更新机制。Wang 等[23]针对有序数据，提出了当特征值与特征集同时发生变化时下近似增量更新机制。Huang 等[24]针对多源混合数据，提出了当特征集、对象集和缺失值变化时的下近似增量更新机制。上述大部分增量式特征选择算法主要针对的是决策完备的数据集。针对部分标记混合数据动态变化的特征选择研究较少，需进一步研究。因此，本文针对部分标记混合数据下特征集的动态变化，提出了基于信息粒度的增量式特征选择算法。首先，利用信息粒度对部分标记混合数据的特征进行重要度分析；然后，结合增量式策略，给出部分标记混合数据中特征集发生动态变化时，信息粒度的增量式更新机制；最后，在此基础上，设计了特征集增加和删除的两种增量式特征选择算法。实验结果表明，相比其他非增量式特征选择算法，本文提出的算法在分类精度不下降的情况下，可快速地获取特征集动态变化后的特征选择结果。

2 基础知识

2.1 部分标记混合决策系统

在粗糙集理论中，数据通常表达成一个四元组的信息系统 $IS=(U,A,V,f)$，其中 $U=\{x_1,x_2,\cdots,x_n\}$ 为对象集；$A=\{a_1,a_2,\cdots,a_m\}$ 为特征集；V 表示特征值的值域，$V=\bigcup_{a\in A}V_a$，其中 V_a 表示特征 $a\in A$ 的值域；f 是 $U\times A\to V$ 的信息函数，$\forall a\in A,x\in U,f(x,a)\in V_a$ 表示对象 x 在特征 a 下的特征值。若 $A=C\cup D$，且 $C\cap D=\varnothing$，其中 C 为条件特征集，D 为决策特征集，则该系统为决策系统。给定决策系统 $DS=(U,A=C\cup D,V,f)$，$\forall B\subseteq A$，B 上的等价关系定义为：$IND(B)=\{(x,y)\in U\times U\mid \forall b\in B,f(x,b)=f(y,b)\}$。通过该等价关系，可给出 U 在 B 下的划分为 $U/IND(B)=\{X_1,X_2,\cdots,X_j\}$，其中 $X_i(1\leqslant i\leqslant j)$ 为等价类。

针对部分标记的混合数据，将其表达为部分标记混合决策系统 $PDS=(U=U_u\cup U_l,A=C\cup D,V,f)$，其中 U_u 为无标记对象集，U_l 为有标记对象集，且 $U_u\cap U_l=\varnothing$；$C=C_s\cup C_n$，$C_s\cap C_n=\varnothing$，C_s 表示符号型特征集，C_n 表示数值型特征集；$*$ 为缺失的特征值，可以是符号型特征值也可以是数值型特征值；V 和 f 与信息系统 IS 的表达一致。以表 1 为例，该患者诊断系统为一个部分标记混合决策系统，其中 $U=\{x_1,\cdots,x_7\}$ 为 7 个患者的对象集，$C=\{c_1,\cdots,c_5\}$ 分别为血糖、胆固醇、性别、静息血压和胸痛部位 5 个条件特征，决策特征 $D=\{d\}$ 为是否患有心脏病，其中 c_1,c_2 和 c_4 为数值型特征，即 $C_n=\{c_1,c_2,c_4\}$，而 c_3 和 c_5 为符号型特征，即 $C_s=\{c_3,c_5\}$。另外，由于部分患者诊断信息缺失或者出于对患者的隐私保护等，$*$ 存在于数值型特征 c_1,c_2 和 c_4 以及符号型特征 c_5 中；同时，由于部分患者尚未得到诊断结果，$*$ 也存在于决策特征 d 中。

针对表 1 的部分标记患者诊断系统，我们将其分成两部分，一部分是无标记对象集的信息系统 $IS=(U_u,A=C,V,f)$，其中 $U_u=\{x_1,x_4,x_5,x_7\}$；而另一部分是有标记对象集的决策系统 $DS=(U_l,A=C\cup D,V,f)$，其中 $U_l=\{x_2,x_3,x_6\}$。

表 1 部分标记患者诊断系统

Table 1 Partially labeled patient diagnostic system

	c_1	c_2	c_3	c_4	c_5	d
x_1	0.1	0.3	男	0.1	其他	*
x_2	*	*	女	0.5	其他	0
x_3	0.8	0.1	女	0.8	其他	1
x_4	0.2	0.3	男	0.1	胸骨后	*
x_5	*	0.9	男	0.3	*	*
x_6	0.9	0.7	女	*	胸骨后	1
x_7	0.7	0.6	女	0.3	胸骨后	*

2.2 信息粒度和相对信息粒度

定义 1(邻域粒度) 给定部分标记混合决策系统 $PDS=(U=U_u\cup U_l,A=C\cup D,V,f)$，对于 $\forall B=B_s\cup B_n\subseteq C$，$U$ 在 B 下的邻域关系 NR_B^ε 定义为：

$$NR_B^\varepsilon=\{(x,y)\in U\times U\mid \forall b\in B_s,f(x,b)=f(y,b)\vee f(x,b)=*\vee f(y,b)=*\}\cap\{(x,y)\in U\times U\mid \forall b\in B_n,DIS_{B_n}(x,y)\leqslant\varepsilon\}$$

其中，ε 为邻域半径，$DIS_{B_n}(x,y)=\sqrt{\sum_{i=1}^{|B_n|}|f(x,b_i)-f(y,b_i)|^2}$ 用于度量对象 x 与 y 之间的距离，若 $f(x,b_i)=*\vee f(y,b_i)=*$，则 $|f(x,b_i)-f(y,b_i)|=0$。通过邻域关系 NR_B^ε，得到 U 在 B 下的分类为 $U/NR_B^\varepsilon=\{\delta_B(x_1),\delta_B(x_2),\cdots,\delta_B(x_{|U|})\}$，其中 $\delta_B(x_i)(1\leqslant i\leqslant|U|)$ 表示对象 x_i 在 B 下的邻域粒度。

定义 2(信息粒度) 给定部分标记混合决策系统 $PDS=(U=U_u\cup U_l,A=C\cup D,V,f)$，$U_u=\{x_1,x_2,\cdots,x_{|U_u|}\}$，$\forall B\subseteq C$，若 U_u 在 B 下的分类为 $U_u/NR_B^\varepsilon=\{\delta_B(x_1),\delta_B(x_2),\cdots,\delta_B(x_{|U_u|})\}$，则 U_u 在 B 下的信息粒度为：

$$IG_u(B)=\frac{1}{|U_u|}\sum_{i=1}^{|U_u|}\frac{|\delta_B(x_i)|}{|U_u|}$$

定义 3(相对信息粒度) 给定部分标记混合决策系统

$PDS=(U=U_u\cup U_l, A=C\cup D, V, f)$, $U_l=\{x_1, x_2, \cdots, x_{|U_l|}\}$, $\forall B\subseteq C$, 若 U_l 在 B 下的分类为 $U_l/NR_B^\varepsilon=\{\delta_B(x_1), \delta_B(x_2), \cdots, \delta_B(x_{|U_l|})\}$, U_l 在 D 下的划分为 $\{D_1, D_2, \cdots, D_j\}$, 则 U_l 下 D 相对于 B 的信息粒度为:

$$IG_l(D|B)=IG_l(B)-IG_l(B\cup D)=\frac{1}{|U_l|}\sum_{i=1}^{|U_l|}\frac{|\delta_B(x_i)-D_k|}{|U_l|}$$

$(1\leqslant k\leqslant j)$

其中, $IG_l(B\cup D)=\frac{1}{|U_l|}\sum_{i=1}^{|U_l|}\frac{|\delta_B(x_i)\cap D_k|}{|U_l|}$。

实例 1(以表 1 为例) 假设邻域半径为 $\varepsilon=0.3$, $B=\{c_1, c_2, c_3\}$, 则 $U_u=\{x_1, x_4, x_5, x_7\}$ 在 B 下的分类为 $\{\delta_B(x_1), \delta_B(x_4), \delta_B(x_5), \delta_B(x_7)\}$, 其中 $\delta_B(x_1)=\{x_1, x_4\}$, $\delta_B(x_4)=\{x_1, x_4\}$, $\delta_B(x_5)=\{x_5\}$, $\delta_B(x_7)=\{x_7\}$; $U_l=\{x_2, x_3, x_6\}$ 在 B 下的分类为 $\{\delta_B(x_2), \delta_B(x_3), \delta_B(x_6)\}$, 其中 $\delta_B(x_2)=\{x_2, x_3, x_6\}$, $\delta_B(x_3)=\{x_2, x_3\}$, $\delta_B(x_6)=\{x_2, x_6\}$; U_l 在 D 下的划分为 $\{D_1, D_2\}$, 其中 $D_1=\{x_2\}$, $D_2=\{x_3, x_6\}$。由定义 2 可知, U_u 在 B 下的信息粒度为: $IG_u(B)=\frac{1}{4}*\frac{|\delta_B(x_1)|+|\delta_B(x_4)|+|\delta_B(x_5)|+|\delta_B(x_7)|}{4}=\frac{3}{8}$。由定义 3 可知, U_l 在 D 下关于 B 的相对信息粒度为: $IG_l(D|B)=\frac{1}{3}*\frac{|\delta_B(x_2)-D_1|+|\delta_B(x_3)-D_2|+|\delta_B(x_6)-D_2|}{3}=\frac{4}{9}$。

引理 1(单调性) 给定部分标记混合决策系统 $PDS=(U=U_u\cup U_l, A=C\cup D, V, f)$, 对于 $\forall B_1, B_2\subseteq C$, 且 $B_1\subseteq B_2$, 给定相同的邻域半径 ε, 则有 $IG_u(B_2)\leqslant IG_u(B_1)$, $IG_l(B_2\cup D)\leqslant IG_l(B_1\cup D)$。

证明: 假设 $U_u=\{x_1, x_2, \cdots, x_{|U_u|}\}$, 给定相同的邻域半径 ε, 因为 $B_1\subseteq B_2\subseteq C$, 根据定义 1, 容易得到 $\delta_{B_2}(x_i)\subseteq\delta_{B_1}(x_i)$ $(1\leqslant i\leqslant|U_u|)$。根据定义 2 可得, $\sum_{i=1}^{|U_u|}|\delta_{B_2}(x_i)|\leqslant\sum_{i=1}^{|U_u|}|\delta_{B_1}(x_i)|$, 则有 $IG_u(B_2)\leqslant IG_u(B_1)$。同理, 假设 $U_l=\{x_1, x_2, \cdots, x_{|U_l|}\}$, U_l 在 D 下的划分为 $U/IND(D)=\{D_1, D_2, \cdots, D_j\}$, 根据定义 3 可得, $\sum_{i=1}^{|U_l|}|\delta_{B_2}(x_i)-D_k|\leqslant\sum_{i=1}^{|U_l|}|\delta_{B_1}(x_i)-D_k|$ $(1\leqslant k\leqslant j)$, 因此, $IG_l(B_2\cup D)\leqslant IG_l(B_1\cup D)$。

定义 4(内部重要度) 给定部分标记混合决策系统 $PDS=(U=U_u\cup U_l, A=C\cup D, V, f)$, $\forall B\subseteq C, b\in B$, 则特征 b 的内部重要度为:

$$Sig_{\text{in}}(b,B,D)=\frac{Sig_{U_u}^{\text{in}}(b,B)}{IG_u(B)}+\frac{Sig_{U_l}^{\text{in}}(b,B,D)}{1+IG_l(D|B)}$$

其中, $Sig_{U_u}^{\text{in}}(b,B)=IG_u(B-\{b\})-IG_u(B)$, $Sig_{U_l}^{\text{in}}(b,B,D)=IG_l(D|B-\{b\})-IG_l(D|B)$。

由定义 4 可知, 当从特征子集 B 中删除特征 b 时, 若信息粒度的变化值越大, 说明该特征 b 越重要。当 $Sig_{\text{in}}(b,B,D)=0$ 时, 则说明特征 b 不重要。因此, 在特征选择过程中, 可利用定义 4 对选择的特征子集进行冗余特征的删除。

定义 5(外部重要度) 给定部分标记混合决策系统 $PDS=(U=U_u\cup U_l, A=C\cup D, V, f)$, $\forall B\subseteq C, b\in C-B$, 则特征 b 的外部重要度为:

$$Sig_{\text{out}}(b,B,D)=\frac{Sig_{U_u}^{\text{out}}(b,B)}{IG_u(B)}+\frac{1+Sig_{U_l}^{\text{out}}(b,B,D)}{1+IG_l(D|B)}$$

其中, $Sig_{U_u}^{\text{out}}(b,B)=IG_u(B)-IG_u(B\cup\{b\})$, $Sig_{U_l}^{\text{out}}(b,B,D)=IG_l(D|B)-IG_l(D|B\cup\{b\})$。由定义 5 可知, $Sig_{\text{out}}(b,B,D)\geqslant 0$, 当 b 增加到特征子集 B 时, 若信息粒度的变化值越大, 则特征越重要。因此, 在特征选择过程中, 可利用定义 5 对所有候选特征按照特征外部重要度进行排序, 从而加速特征子集的选择。

定义 6(特征选择) 给定部分标记混合决策系统 $PDS=(U=U_u\cup U_l, A=C\cup D, V, f)$, $\forall B\subseteq C$, 若 B 为特征选择结果, 则需要满足以下两个条件:

(1) $IG_u(B)=IG_u(C)$, $IG_l(D|B)=IG_l(D|C)$;

(2) 对于 $\forall b\in B$, 使得 $IG_u(B)<IG_u(B-\{b\})$, $IG_l(D|B)<IG_l(D|B-\{b\})$。

条件(1)保证特征子集和特征全集保持相同的分类能力; 条件(2)保证选择的特征子集没有冗余的特征。

3 动态部分标记混合数据下基于信息粒度的特征选择算法

针对部分标记混合数据下特征的动态增加或删除, 首先给出静态(非增量式)特征选择算法, 该算法将动态变化后的数据集看成一个新数据集, 重新对新数据集进行特征选择; 然后给出增量式特征选择算法, 该算法在原数据集特征选择结果的基础上, 通过更新局部数据快速地得到新数据集的特征选择结果。

3.1 动态部分标记混合数据下基于信息粒度的非增量式特征选择算法

当部分标记混合数据特征集增加或删除时, 静态(非增量式)特征选择算法将变化后的数据集看成是一个新数据集, 重新对新数据集进行特征选择, 该过程需要重复操作, 将耗费大量时间和存储空间, 无法满足时效性较强的数据集。

静态特征选择算法如算法 1(NIFS 算法)所示, 该算法采用贪心向前的启发式搜索策略, 依次计算每个特征的重要度, 将特征重要度最大的特征加入候选特征子集, 直至候选特征子集下的信息粒度与整个条件特征集下的信息粒度相同为止, 最后删除选择特征子集结果中的冗余特征。

算法 1 动态部分标记混合数据下基于信息粒度的非增量式特征选择算法(NIFS)

输入: 部分标记混合决策系统 PDS= (U=U$_u$∪U$_l$, A=C∪D, V, f), 增加特征集 C$_{ad}$, 删除特征集 C$_{de}$

输出: 特征选择结果 Red

1. 令 Red←∅, C′=C∪C$_{ad}$−C$_{de}$, 并对数值型特征值进行归一化处理, 使其值域为[0,1];
2. 分别利用定义 2 和定义 3 计算 IG$_u$(C′), IG$_l$(D|C′);
3. $\forall a_i\in C'-\text{Red}(1\leqslant i\leqslant|C'-\text{Red}|)$, 利用定义 5 计算特征的外部重要度 Sig$_{out}$(a$_i$, Red, D), 并选择重要度最大的特征, Red←Red∪argmax{Sig$_{out}$(a$_i$, Red, D): $\forall a_i\in C'-\text{Red}$};
4. 若 IG$_u$(Red)=IG$_u$(C′), IG$_l$(D|Red)=IG$_l$(D|C′), 则执行步骤 5, 否则执行步骤 3;
5. ∀b∈Red, 利用定义 4 计算特征内部重要度 Sig$_{in}$(b, Red, D), 若 Sig$_{in}$(b, Red, D)=0, 则 Red←Red−{b};
6. 返回 Red。

算法 1 的时间复杂度分析如下: 步骤 2 主要是重新计算

新数据集下的信息粒度和相对信息粒度，其时间复杂度为 $O(|U|^2|C'|)$；步骤 3、步骤 4 主要是依次选择特征重要度最大的特征到特征子集结果中，直到满足终止条件，其时间复杂度为 $O(|U|^2|C'||C'|+|U|^2|C'|(|C'|-1)+\cdots+|U|^2|C'|*1)=O(|U|^2|C'|^3)$，步骤 5 从特征子集结果中删除冗余特征，时间复杂度为 $O(|U|^2|C'|^2)$。因此，算法 1 的时间复杂度为 $O(|U|^2|C'|^3)$。

3.2 动态部分标记混合数据下基于信息粒度的增量式特征选择算法

当部分标记混合数据中的特征集增加或删除时，为进一步提高特征选择效率，下文将结合增量式策略，在原始特征选择结果的基础上对局部变化的数据进行更新，以得到新特征选择结果。下文将给出特征集动态变化时信息粒度的增量更新机制，在此基础上，提出了特征集增加和删除时的两种增量式特征选择算法。

3.2.1 增加特征集的增量式特征选择算法

定理 1 给定部分标记混合决策系统 $PDS=(U=U_u\cup U_l,A=C\cup D,V,f)$，$\forall B\subseteq C$，对于无标记对象集 $U_u=\{x_1,x_2,\cdots,x_{|U_u|}\}$，$U_u$ 在 B 下的分类为 $U_u/NR_B^\varepsilon=\{\delta_B(x_i)|x_i\in U_u\}$，假设增加特征集 C_{ad}，若 $\delta_{C_{ad}}(x_i)(1\leqslant i\leqslant|U_u|)$ 为对象 x_i 在 C_{ad} 下的邻域粒度，则 U_u 在 $B\cup C_{ad}$ 下的信息粒度为：

$$IG_u(B\cup C_{ad})=IG_u(B)-\frac{1}{|U_u|^2}\sum_{i=1}^{|U_u|}|\delta_B(x_i)-\delta_{C_{ad}}(x_i)|$$

证明：对于决策系统 PDS，当增加特征集 C_{ad} 时，U_u 在 C_{ad} 下的分类为 $U_u/NR_{C_{ad}}^\varepsilon=\{\delta_{C_{ad}}(x_i)|x_i\in U_u\}$，在 $B\cup C_{ad}$ 下的分类为 $U_u/NR_{B\cup C_{ad}}^\varepsilon=\{\delta_{B\cup C_{ad}}(x_i)|x_i\in U_u\}$。因为 $|\delta_B(x_i)|=|\delta_{B\cup C_{ad}}(x_i)|+|\delta_B(x_i)-\delta_{C_{ad}}(x_i)|$，所以 $IG_u(B)=\frac{1}{|U_u|}\sum_{i=1}^{|U_u|}\frac{|\delta_B(x_i)|}{|U_u|}=\frac{1}{|U_u|}\sum_{i=1}^{|U_u|}\frac{|\delta_{B\cup C_{ad}}(x_i)|}{|U_u|}+\frac{1}{|U_u|}\sum_{i=1}^{|U_u|}\frac{|\delta_B(x_i)-\delta_{C_{ad}}(x_i)|}{|U_u|}=IG_u(B\cup C_{ad})+\frac{1}{|U_u|^2}\sum_{i=1}^{|U_u|}|\delta_B(x_i)-\delta_{C_{ad}}(x_i)|$，故 $IG_u(B\cup C_{ad})=IG_u(B)-\frac{1}{|U_u|^2}\sum_{i=1}^{|U_u|}|\delta_B(x_i)-\delta_{C_{ad}}(x_i)|$。

实例 2(接实例 1)　若特征集 $\{c_4,c_5\}$ 增加到系统中，即 $C_{ad}=\{c_4,c_5\}$，则计算可得 U_u 在 C_{ad} 下的划分为 $\delta_{C_{ad}}(x_1)=\{x_1,x_5\}$，$\delta_{C_{ad}}(x_4)=\{x_4,x_5,x_7\}$，$\delta_{C_{ad}}(x_5)=\{x_1,x_4,x_5,x_7\}$，$\delta_{C_{ad}}(x_7)=\{x_4,x_5,x_7\}$；且 $\delta_B(x_1)-\delta_{C_{ad}}(x_1)=\{x_4\}$，$\delta_B(x_4)-\delta_{C_{ad}}(x_4)=\{x_1\}$，$\delta_B(x_5)-\delta_{C_{ad}}(x_5)=\emptyset$，$\delta_B(x_7)-\delta_{C_{ad}}(x_7)=\emptyset$。根据定理 1，可得 $IG_u(B\cup C_{ad})=IG_u(B)-\frac{1}{|U_u|^2}\sum_{i=1}^{|U_u|}|\delta_B(x_i)-\delta_{C_{ad}}(x_i)|=\frac{3}{8}-\frac{2}{16}=\frac{1}{4}$。

定理 2 给定部分标记混合决策系统 $PDS=(U=U_u\cup U_l,A=C\cup D,V,f)$，$\forall B\subseteq C$，对于有标记对象集 $U_l=\{x_1,x_2,\cdots,x_{|U_l|}\}$，$U_l$ 在 B 下的分类为 $U_l/NR_B^\varepsilon=\{\delta_B(x_i)|x_i\in U_l\}$，$U/D=\{D_1,D_2,\cdots,D_j\}$，假设增加特征集 C_{ad}，若 $\delta_{C_{ad}}(x_i)(1\leqslant i\leqslant|U_l|)$ 为有标记对象 x_i 在 C_{ad} 下的邻域粒度，且 $x_i\in D_k(1\leqslant k\leqslant j)$，则 U_l 下 D 相对于 $B\cup C_{ad}$ 的信息粒度为：

$$IG_l(D|B\cup C_{ad})=IG_l(D|B)-\frac{1}{|U_l|^2}\sum_{i=1}^{|U_l|}|\delta_B(x_i)-\delta_{C_{ad}}(x_i)-D_k|$$

证明：对于决策系统 PDS，当增加特征集 C_{ad} 时，U_l 在 C_{ad} 下的分类为 $U_l/NR_{C_{ad}}^\varepsilon=\{\delta_{C_{ad}}(x_i)|x_i\in U_l\}$，$U_l$ 在 $B\cup C_{ad}$ 下的分类为 $U_l/NR_{B\cup C_{ad}}^\varepsilon=\{\delta_{B\cup C_{ad}}(x_i)|x_i\in U_l\}$。因为 $|\delta_B(x_i)-D_k|=|\delta_{B\cup C_{ad}}(x_i)-D_k|+|\delta_B(x_i)-\delta_{C_{ad}}(x_i)-D_k|$，所以 $IG_l(D|B)=\frac{1}{|U_l|}\sum_{i=1}^{|U_l|}\frac{|\delta_B(x_i)-D_k|}{|U_l|}=\frac{1}{|U_l|}\sum_{i=1}^{|U_l|}\frac{|\delta_{B\cup C_{ad}}(x_i)-D_k|}{|U_l|}+\frac{1}{|U_l|}\sum_{i=1}^{|U_l|}\frac{|\delta_B(x_i)-\delta_{C_{ad}}(x_i)-D_k|}{|U_l|}=IG_l(D|B\cup C_{ad})+\frac{1}{|U_l|^2}\sum_{i=1}^{|U_l|}|\delta_B(x_i)-\delta_{C_{ad}}(x_i)-D_k|$，故 $IG_l(D|B\cup C_{ad})=IG_l(D|B)-\frac{1}{|U_l|^2}\sum_{i=1}^{|U_l|}|\delta_B(x_i)-\delta_{C_{ad}}(x_i)-D_k|$。

实例 3(接实例 1)　若特征集 $\{c_4,c_5\}$ 增加到系统，即 $C_{ad}=\{c_4,c_5\}$，则计算可得 U_l 在 C_{ad} 下的划分为 $\delta_{C_{ad}}(x_2)=\{x_2,x_3\}$，$\delta_{C_{ad}}(x_3)=\{x_2,x_3\}$，$\delta_{C_{ad}}(x_6)=\{x_6\}$；且 $\delta_B(x_2)-\delta_{C_{ad}}(x_2)-D_1=\{x_6\}$，$\delta_B(x_3)-\delta_{C_{ad}}(x_3)-D_2=\emptyset$，$\delta_B(x_6)-\delta_{C_{ad}}(x_6)-D_2=\{x_2\}$。根据定理 2，可得 $IG_l(D|B\cup C_{ad})=IG_l(D|B)-\frac{1}{|U_l|^2}\sum_{i=1}^{|U_l|}|\delta_B(x_i)-\delta_{C_{ad}}(x_i)-D_k|=\frac{4}{9}-\frac{2}{9}=\frac{2}{9}$。

基于以上信息粒度的增量更新机制，下面给出增加特征集的增量式特征选择算法 IFSA，如算法 2 所示。

算法 2 一种增加特征集的增量式特征选择算法(IFSA)

输入：部分标记混合决策系统 PDS=(U=$U_u\cup U_l$,A=C$\cup$D,V,f)，原特征选择结果 Red，对象在 Red，C 下的邻域粒度与信息粒度，邻域半径 ε，增加特征集 C_{ad}

输出：新的特征选择结果 Red′

1. 令 Red′←Red，C′=C$\cup C_{ad}$，并对新增数值型特征值进行归一化处理；
2. 利用定理 1 和定理 2 分别计算 $IG_u(C')$ 和 $IG_l(D|C')$；
3. 若 $IG_u(C')=IG_u(Red)$，$IG_l(D|C')=IG_l(D|Red)$，则转向步骤 7；否则转向步骤 4；
4. $\forall a_i\in C_{ad}(1\leqslant i\leqslant|C_{ad}|)$，分别利用定理 1 和定理 2 计算特征外部重要度 $Sig_{out}(a_i,Red,D)$，并对特征重要度由大到小排序为 $\{a_1,a_2,\cdots,a_{|C_{ad}|}\}$；
5. While $IG_u(Red')\neq IG_u(C')$，$IG_l(D|Red')\neq IG_l(D|C')$ do
   for j=1 to $|C_{ad}|$ do
   Red′=Red′$\cup\{a_j\}$；
6. $\forall b\in Red$，利用定义 4 计算其特征内部重要度 $Sig_{in}(b,Red',D)$，若 $Sig_{in}(b,Red',D)=0$，则 Red′←Red′−{b}；
7. 返回 Red′。

算法 2 的时间复杂度分析如下：步骤 2 使用增量式方法计算增加特征集后 $C\cup C_{ad}$ 下的信息粒度，时间复杂度为 $O(|U|^2|C_{ad}|)$；步骤 4 对特征集 C_{ad} 中的特征进行重要度排序，时间复杂度为 $O(|U|^2|C_{ad}|)$；步骤 5 使用迭代的方法将最重要的特征添加到候选特征选择子集中，直至满足终止条件为止，时间复杂度最坏为 $O(|U|^2|C_{ad}|)$；步骤 6 从候选特征子集中删除冗余特征，时间复杂度为 $O(|U|^2|Red|)$。因此，算法 2 的时间复杂度为 $O(|U|^2|C|)$，相比非增量式算法 NIFS 的时间复杂度 $O(|U|^2|C'|^3)$，算法 2 的时间复杂度得到了有效的降低。

3.2.2 删除特征集的增量式特征选择算法

定理 3 给定部分标记混合决策系统 $PDS=(U=U_u\cup$

$U_l, A=C\cup D, V, f)$，$\forall B\subseteq C$，对于无标记对象集 $U_u=\{x_1, x_2, \cdots, x_{|U_u|}\}$，$U_u$ 在 B 下的分类为 $U_u/NR_B^\varepsilon=\{\delta_B(x_i)\mid x_i\in U_u\}$，假设删除特征集 C_{de}，则 U_u 在 $B-C_{de}$ 下的信息粒度为：

$$IG_u(B-C_{de})=IG_u(B)+\frac{1}{|U_u|^2}\sum_{i=1}^{|U_u|}|\{y\in U-\delta_B(x_i)\mid (x_i,y)\in NR_{B-C_{de}}^\varepsilon\}|$$

证明：对于决策系统 PDS，当删除特征集 C_{de} 时，因为 $\delta_{B-C_{de}}(x_i)=\delta_B(x_i)\bigcup\{y\in U_u-\delta_B(x_i)\mid (x_i,y)\in NR_{B-C_{de}}^\varepsilon\}$，所以 $IG_u(B-C_{de})=\frac{1}{|U_u|}*\sum_{i=1}^{|U_u|}\frac{|\delta_B(x_i)\bigcup\{y\in U_u-\delta_B(x_i)\mid (x_i,y)\in NR_{B-C_{de}}^\varepsilon\}|}{|U_u|}=IG_u(B)+\frac{1}{|U_u|^2}\sum_{i=1}^{|U_u|}|\{y\in U_u-\delta_B(x_i)\mid (x_i,y)\in NR_{B-C_{de}}^\varepsilon\}|$。

实例 4(接实例 2) 若特征集 $\{c_4,c_5\}$ 从系统中删除，即 $C_{de}=\{c_4,c_5\}$，则计算可得 $U_u-\delta_C(x_1)=\{x_4,x_5,x_7\}$，$U_u-\delta_C(x_4)=\{x_1,x_5,x_7\}$，$U_u-\delta_C(x_5)=\{x_1,x_4,x_7\}$，$U_u-\delta_C(x_7)=\{x_1,x_4,x_5\}$，且 $\{(x_1,x_4),(x_4,x_1)\}\in NR_{C-C_{de}}^\varepsilon$。根据定理 3，可得 $IG_u(C-C_{de})=IG_u(C)+\frac{1}{|U_u|^2}\sum_{i=1}^{|U_u|}|\{y\in U_u-\delta_C(x_i)\mid (x_i,y)\in NR_{C-C_{de}}^\varepsilon\}|=\frac{1}{4}+\frac{1}{16}*2=\frac{3}{8}$。

定理 4 给定部分标记混合决策系统 $PDS=(U=U_u\cup U_l, A=C\cup D, V, f)$，$\forall B\subseteq C$，对于有标记对象集 $U_l=\{x_1, x_2, \cdots, x_{|U_l|}\}$，$U_l$ 在 B 下的分类为 $U_l/NR_B^\varepsilon=\{\delta_B(x_i)\mid x_i\in U_l\}$，假设删除特征集 C_{de}，则 U_l 下 $B-C_{de}$ 相对于 D 的信息粒度为：

$$IG_l(D|B-C_{de})=IG_l(D|B)+\frac{1}{|U_l|^2}\sum_{i=1}^{|U_l|}|\{y\in U_l-\delta_B(x_i)-D_k\mid (x_i,y)\in NR_{B-C_{de}}^\varepsilon\}|$$

证明：对于决策系统 PDS，当删除特征集 C_{de} 时，因为 $\delta_{B-C_{de}}(x_i)-D_k=\{\delta_B(x_i)-D_k\}\bigcup\{y\in U_l-\delta_B(x_i)-D_k\mid (x_i,y)\in NR_{B-C_{de}}^\varepsilon\}$，所以 $IG_l(D\mid B-C_{de})=\frac{1}{|U_l|}\sum_{i=1}^{|U_l|}\frac{|\delta_B(x_i)-D_k|+|y\in U_l-\delta_B(x_i)-D_k\mid (x_i,y)\in NR_{B-C_{de}}^\varepsilon|}{|U_l|}=IG_l(D|B)+\frac{1}{|U_l|^2}*\sum_{i=1}^{|U_l|}|\{y\in U_l-\delta_B(x_i)-D_k\mid (x_i,y)\in NR_{B-C_{de}}^\varepsilon\}|$。

实例 5(接实例 3) 若特征集 $\{c_4,c_5\}$ 从系统中删除，即 $c_{de}=\{c_4,c_5\}$，则计算可得 $U_l-\delta_C(x_2)-D_1=\{x_6\}$，$U_l-\delta_C(x_3)-D_2=\varnothing$，$U_l-\delta_C(x_6)-D_1=\{x_2\}$，且 $\{(x_2,x_6),(x_6,x_2)\}\in NR_{C-C_{de}}^\varepsilon$。根据定理 4，可得 $IG_l(D|C-C_{de})=IG_l(D|C)+\frac{1}{|U_l|^2}\sum_{i=1}^{|U_l|}|\{y\in U_l-\delta_C(x_i)-D_k\mid (x_i,y)\in NR_{C-C_{de}}^\varepsilon\}|=\frac{2}{9}+\frac{1}{9}*2=\frac{4}{9}$。

基于以上信息粒度的增量更新机制，下面给出删除特征集的增量式特征选择算法 IFSD，如算法 3 所示。

算法 3 一种删除特征集的增量式特征选择算法(IFSD)

输入：部分标记混合决策系统 PDS=(U=U$_u$∪U$_l$, A=C∪D, V, f)，原特征选择结果 Red，对象在 Red，C 下的邻域粒度和信息粒度，邻域半径 ε，删除特征集 C$_{de}$

输出：新的特征选择结果 Red′

1. 令 Red′=Red−C$_{de}$，C′=C−C$_{de}$；
2. 利用定理 3 分别计算 IG$_u$(C′)和 IG$_u$(Red′)，利用定理 4 分别计算 IG$_l$(D|C′)和 IG$_l$(D|Red′)；
3. 若 IG$_u$(C′)=IG$_u$(Red′)，IG$_l$(D|C′)=IG$_l$(D|Red′)，则转向步骤 7，否则转向步骤 4；
4. ∀a$_i$∈C′−Red′(1≤i≤|C′−Red′|)，计算特征外部重要度 Sig$_{out}$(a$_i$, Red′, D)，并对外部重要度由大到小排序为{a$_1$, a$_2$, …, a$_{|C'-Red'|}$}；
5. While IG$_u$(Red′)≠IG$_u$(C′)，IG$_l$(D|Red′)≠IG$_l$(D|C′) do
   for j=1 to |C′−Red′| do
   Red′=Red′∪{a$_j$}
6. ∀b∈Red′，利用定义 4 计算其特征内部重要度 Sig$_{in}$(b, Red′, D)，若 Sig$_{in}$(b, Red′, D)=0，则 Red′←Red′−{b}；
7. 返回 Red′。

算法 3 的时间复杂度分析如下：步骤 2 使用增量式方法计算删除特征集后 $C-C_{de}$ 和 $Red-C_{de}$ 下的信息粒度，时间复杂度最坏为 $O(|U|^2|C-C_{de}|)$；步骤 4 对剩余特征进行重要度排序，时间复杂度为 $O(|U|^2|C'-Red'|)$；步骤 5 使用迭代的方法将最重要的特征添加到候选特征子集中，直至满足终止条件，时间复杂度最坏为 $O(|U|^2|C'-Red'|)$；步骤 6 从候选特征子集中删除冗余特征，时间复杂度为 $O(|U|^2|C-C_{de}|)$。因此，算法 IFSD 的时间复杂度为 $O(|U|^2|C-C_{de}|)$。

4 实验分析

为了验证本文算法的可行性和有效性，从 UCI[27] 中选取 8 个数据集进行测试，具体的数据集信息如表 2 所列。本实验的测试环境为：CPU Intel(R) Core(TM) i5-4210 (2.40GHz)，内存 8.0GB，算法编程语言为 Python，使用的开发工具是 JetBrains PyCharm Community Edition 2017。针对数据集的数值型特征，先对其使用 Rossta[28] 进行归一化处理。使用文献[13]的实验设置将数据集分成决策标记数据和无决策标记数据两部分，其中无决策标记数据占每个数据集的 80%。

表 2 数据集

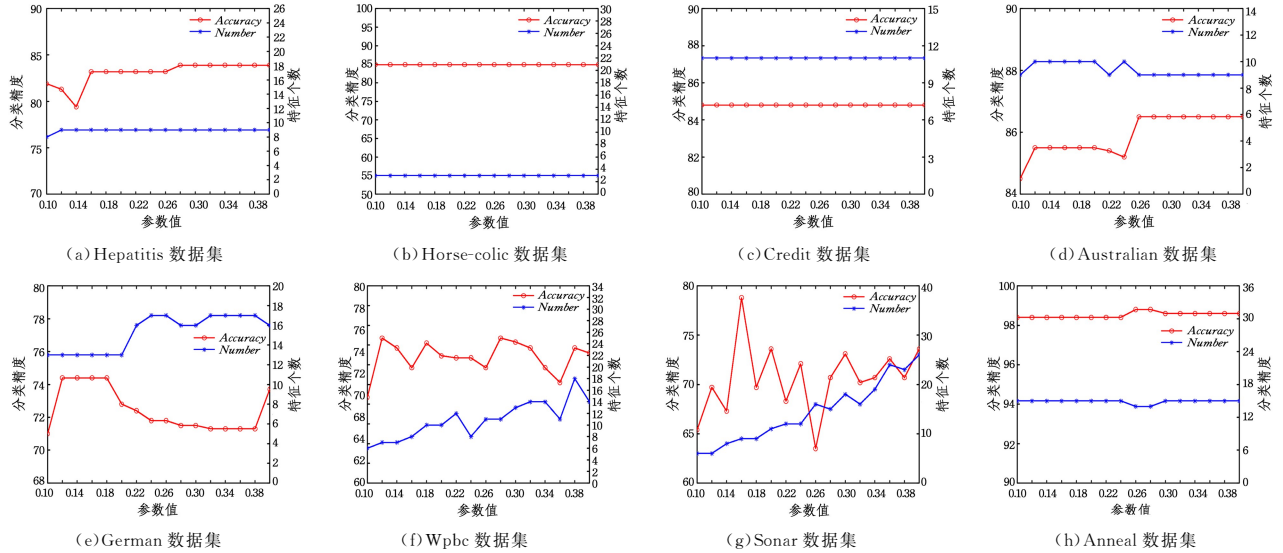
Table 2 Data sets

数据集	样本数	特征数	类别数	特征值是否缺失
Hepatitis	155	19	3	是
Horse-colic	301	27	2	是
Credit	680	15	2	是
Australian	691	14	2	否
German	1000	20	2	否
Wpbc	198	34	2	是
Sonar	208	60	2	否
Anneal	798	38	6	是

4.1 ε 特征参数的设置

在基于信息粒度的特征选择过程中，计算信息粒度时需要使用邻域半径 ε，该参数决定了邻域粒度的大小，影响着特征子集的选择结果，对分类精度起着重要的作用。因此，使用文献[8]的方法，将 ε 以步长 0.02 从 0.1 增加到 0.4 进行实验，分析参数对部分标记混合数据下特征选择的影响，为每个数据集选取最佳的邻域参数。图 1 给出了表 1 中各数据集在不同参数 ε 下的特征选择结果分类精度，图 1 中的纵坐标 *Accuracy*

表示特征选择结果在 C4.5 分类器下的分类精度，*Number* 表示特征选择结果的个数，横坐标为不同的邻域参数值。



(a) Hepatitis 数据集 (b) Horse-colic 数据集 (c) Credit 数据集 (d) Australian 数据集

(e) German 数据集 (f) Wpbc 数据集 (g) Sonar 数据集 (h) Anneal 数据集

图 1 8 个数据集在不同参数 ε 下的分类精度

Fig. 1 Classification accuracies of eight datasets with different parameters ε

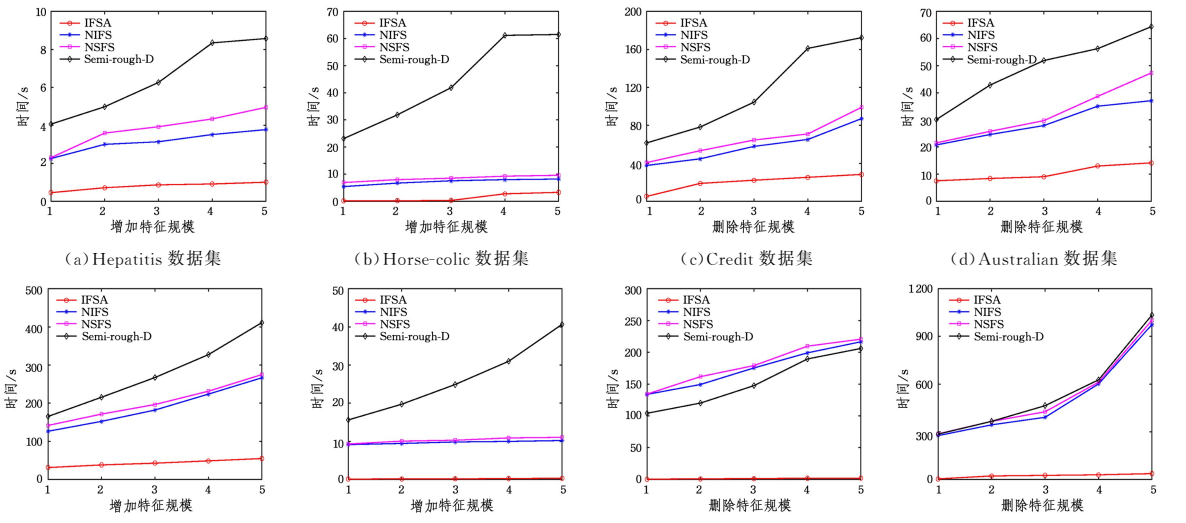
由图 1 可知，随着邻域参数 ε 的增加，分类精度和特征选择个数在不断地变化。在 Hepatitis 数据集中，当 ε = 0.28 时，该数据集特征选择结果的分类精度为 84.3%，个数为 9，此时特征选择结果拥有最高的分类精度，因此 Hepatitis 数据集的最优特征参数设置为 0.28。Australian 数据集在 ε = 0.26 时可得到最优的分类精度 85.9%以及最少的特征个数 9。由于 Horse-colic 数据集的特征选择结果均为符号性特征，分类精度不受邻域参数的影响，因此将参数 ε 设置为 0.10。而针对剩余 5 个数据集 Credit，German，Wpbc，Sonar，Anneal，当 ε 的值分别设置为 0.10，0.12，0.12，0.16，0.26 时，数据集获得最优的分类性能。

4.2 不同算法的性能比较

为了验证所提算法的可行性与高效性，下文将本文提出的两种增量式特征选择算法(IFSA 和 IFSD)与基于信息粒度的非增量式特征选择算法(NIFS)、基于可辨矩阵的半监督特征选择算法 Semi-rough-D[13] 以及基于邻域熵的特征选择算法 NSFS[29] 进行实验比较。由于算法 Semi-rough-D 只能处理符号型数据，因此针对表 2 所列数据集中的数值型特征，采用等频方法[28]进行离散化，并对缺失特征值采用特征平均值进行补齐等数据预处理操作。

4.2.1 增加特征集时不同算法的性能比较

针对表 2 中的每个数据集，令 C 为每个数据集的特征集，选取 50%的特征作为原始特征集，即 $0.5*|C|$，将剩余 50%的特征平均分为 5 部分，即 $|c_{ad}^{i}|=\dfrac{0.5*|C|}{5}$，$i=1,2,\cdots,5$，令 $C_{ad}^{i}=\bigcup_{j=1}^{i}c_{ad}^{j}$，$i=1,2,\cdots,5$ 为 5 个增加特征集。利用 5 个不同规模的增加特征集，使用算法 NIFS，Semi-rough-D，NSFS 和 IFSA 进行实验。图 2 给出了算法 NIFS，Semi-rough-D，NSFS 和 IFSA 随着增加特征集规模变化的运行时间。



(a) Hepatitis 数据集 (b) Horse-colic 数据集 (c) Credit 数据集 (d) Australian 数据集

(e) German 数据集 (f) Wpbc 数据集 (g) Sonar 数据集 (h) Anneal 数据集

图 2 算法 NIFS，Semi-rough-D，NSFS 和 IFSA 的运行时间

Fig. 2 Running time of algorithms NIFS，Semi-rough-D，NSFS and IFSA

由图 2 可知，随着特征集不断增加，特征选择算法 NIFS，Semi-rough-D，NSFS 和 IFSA 的运行时间也不断增加。相比算法 NIFS，NSFS 和 Semi-rough-D，本文算法 IFSA 的运行时间明显短于其他 3 种算法，如在数据集 German 增加第三个特征集时，算法 NIFS，NSFS 和 Semi-rough-D 的运行时间分别为 181.93 s，196.01 s 和 267.17 s，而算法 IFSA 的运行时间为 42.85s，相比算法 NIFS，NSFS 和 Semi-rough-D 分别缩短了 76.44%，78.13%，83.96%。再比如，在向数据集 Wpbc 增加第三个特征集时，算法 NIFS，NSFS 和 Semi-rough-D 的运行时间分别为 9.82 s，10.30 s 和 24.86 s，而算法 IFSA 的运行时间为 0.17 s，相比算法 NIFS，NSFS 和 Semi-rough-D 分别缩短了 98.26%，98.34%，99.31%。很明显，算法 IFSA 在运行时间上短于其他 3 种算法，主要原因是算法 IFSA 利用增量式策略，在原始特征选择结果的基础上，只需更新计算局部数据，减少了大量重复计算。随着增加特征集规模的增加，算法 IFSA 所需的运行时间与算法 NIFS，NSFS，Semi-rough-D 所需的运行时间差值增大。以 Hepatitis 数据集为例，在增加第三个特征集时，算法 IFSA 所需的运行时间相比算法 NIFS，NSFS，Semi-rough-D 缩短了 2.27 s，3.06 s，5.40 s；增加第四个特征集时，算法 IFSA 所需的运行时间与算法 NIFS，NSFS，Semi-rough-D 所需的运行时间的差值为 2.60 s，3.43 s，7.44 s，运行时间差值增加。这是因为增加特征集的规模越大，静态特征选择算法需要重复计算的工作量相应增加。

同时，我们利用决策树(C4.5)和支持向量机(SVM)两种分类器，对 4 种不同特征选择算法的分类性能进行比较。采用十字交叉验证法，针对每个数据集，我们将数据集等分为 10 份，轮流将其中 9 份作为训练数据，将剩余 1 份作为测试数据进行实验。表 3 列出了特征集增加时算法 NIFS，Semi-rough-D，NSFS 和 IFSA 的最终特征选择个数。表 4 和表 5 分别列出了算法 NIFS，Semi-rough-D，NSFS 和 IFSA 在 C4.5 和 SVM 分类器下的分类精度，其中 Raw 表示在特征全集下的分类精度结果，“Average”表示不同算法下的平均分类精度结果。表中的“+”“=”“−”表示本文提出的算法 IFSA 相比算法 NIFS，Semi-rough-D，NSFS 拥有“更好”“一样”“更差”的分类精度。同时，表中“Win/Tie/Lose”这行用于统计所提算法 IFSA 在所有数据集中拥有“更好/一样好/更差”的数据集个数。

表 3 算法 NIFS，Semi-rough-D，NSFS 和 IFSA 的特征选择个数

Table 3 Feature selection numbers of algorithms NIFS, Semi-rough-D, NSFS and IFSA

数据集	特征选择个数			
	NIFS	Semi-rough-D	NSFS	IFSA
Hepatitis	9	7	8	9
Horse-colic	3	8	3	3
Credit	11	13	14	14
Australian	9	8	10	11
German	13	13	12	13
Wpbc	7	13	7	8
Sonar	9	9	8	10
Anneal	16	12	16	16

表 4 算法 NIFS，Semi-rough-D，NSFS 和 IFSA 在 C4.5 分类器下的分类精度

Table 4 Classification accuracies of algorithms NIFS, Semi-rough-D, NSFS and IFSA with classifier C4.5

数据集	Raw	NIFS	Semi-rough-D	NSFS	IFSA
Hepatitis	81.50±0.42	84.63±0.24$^{=}$	84.59±0.28^{+}	84.52±0.19^{+}	**84.63±0.24**
Horse-colic	84.46±0.92	84.70±1.97$^{=}$	81.76±1.43^{+}	84.70±1.97$^{=}$	**84.70±1.97**
Credit	84.75±0.07	84.85±0.16^{+}	85.45±0.62^{+}	85.63±0.37$^{=}$	**85.63±0.37**
Australian	85.13±0.87	85.71±0.28^{+}	85.41±0.47^{+}	85.15±0.36^{+}	**85.92±0.51**
German	73.03±0.17	71.46±0.55^{-}	70.62±0.67^{+}	**74.53±0.26**$^{-}$	71.42±0.06
Wpbc	73.83±1.31	74.36±0.32^{+}	73.75±0.55^{+}	74.36±0.32^{+}	**75.33±0.47**
Sonar	71.93±3.28	74.16±1.31^{+}	60.46±0.65^{+}	67.70±0.64^{+}	**74.83±0.47**
Anneal	98.56±0.04	98.66±0.04$^{=}$	98.33±0.23^{+}	98.66±0.04$^{=}$	**98.66±0.04**
Average	81.64	82.31	80.04	81.90	**82.64**
Win\Tie\Lose	−	4/3/1	8/0/0	4/3/1	−

表 5 算法 NIFS，Semi-rough-D，NSFS 和 IFSA 在 SVM 分类器下的分类精度

Table 5 Classification accuracies of algorithms NIFS, Semi-rough-D, NSFS and IFSA with classifier SVM

数据集	Raw	NIFS	Semi-rough-D	NSFS	IFSA
Hepatitis	81.29±0.84	83.90±0.49$^{=}$	81.33±0.19^{+}	82.63±0.88^{+}	**83.90±0.49**
Horse-colic	64.15±0.08	80.13±0.10$^{=}$	66.22±1.09^{+}	80.13±0.10$^{=}$	**80.13±0.10**
Credit	85.60±0.19	**85.92±0.10**$^{-}$	85.80±0.12^{-}	85.70±0.04$^{=}$	85.70±0.04
Australian	85.52±0.19	**86.42±0.10**$^{-}$	85.55±0.02^{+}	85.70±0.10^{+}	86.10±0.08
German	76.70±0.13	74.54±0.07^{+}	72.46±0.03^{+}	**76.81±0.18**$^{-}$	75.81±0.76
Wpbc	79.68±0.89	79.80±0.28$^{=}$	76.30±0.11^{+}	79.80±0.28$^{=}$	**79.80±0.12**
Sonar	79.66±2.49	81.42±2.01^{+}	62.57±1.74^{+}	73.56±0.36^{+}	**82.36±0.47**
Anneal	98.83±0.09	99.40±0.14$^{=}$	98.40±0.16^{+}	99.40±0.14$^{=}$	**99.40±0.14**
Average	81.42	84.06	78.57	82.96	**84.15**
Win\Tie\Lose	−	2/4/2	7/0/1	3/4/1	−

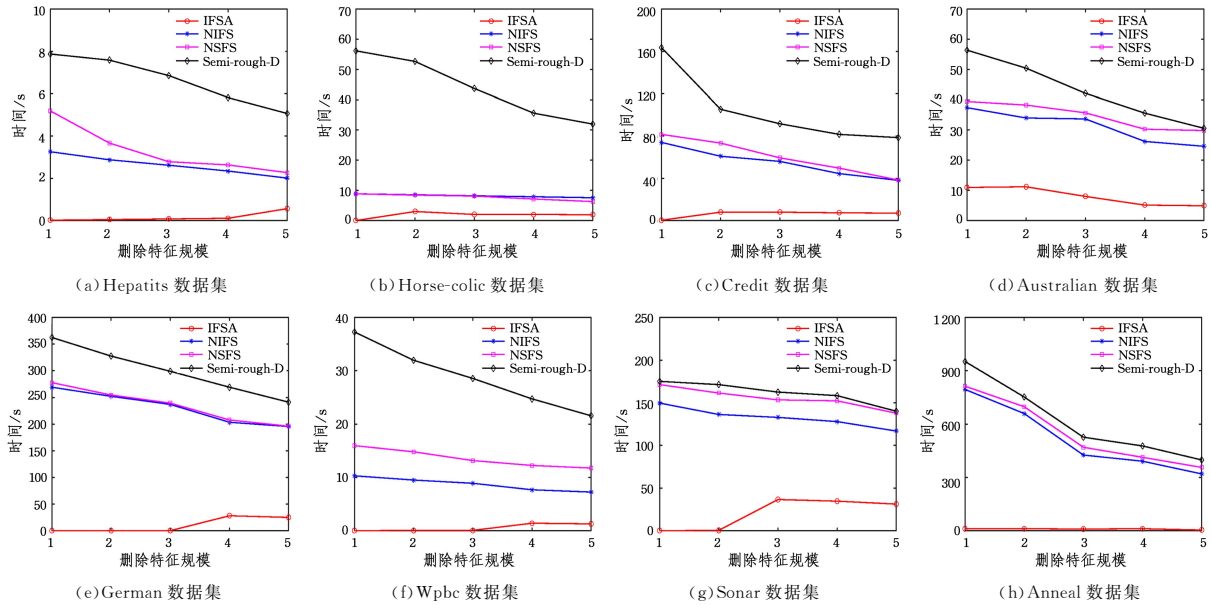
从表 3 可以看出，在大部分数据集中增加特征集时，使用算法 IFSA 得到的特征子集个数相比数据集原有的特征降幅明显。算法 NIFS 和 IFSA 在 Hepatitis，Horse-colic，German，Anneal 数据集下的特征选择结果是相同的，在 Credit，Australian，Wpbc，Sonar 数据集下的特征选择结果相似，因为这两种算法都是基于信息粒度的方法进行特征选择，所使用的特征度量方法一致，因此得到的特征选择结果相近。使用算法 NSFS 得到的特征选择结果与算法 IFSA 相近，例如在 Australian 数据集下，算法 NSFS 和 IFSA 得到的特征选择个数分别为 10，11。但是算法 Semi-rough-D 由于在进行特征

选择前对数值型数据进行离散化以及对缺失值进行预处理，可能造成部分信息的丢失，导致在 Horse-colic，Wpbc 数据集下与其他 3 种算法得到的结果有明显的不同。从表 4 和表 5 的平均分类精度结果可以看出，在 C4.5 分类器和 SVM 分类器下，算法 NIFS，NSFS 和 IFSA 得到的平均分类精度结果均优于没有进行特征选择时特征全集的平均分类精度。同时，算法 NIFS，NSFS 和 IFSA 拥有较为相似的平均分类精度结果。通过 Win\Tie\Lose 的统计结果可以看出，相比算法 NIFS，Semi-rough-D 和 NSFS，算法 IFSA 在大部分数据集下可以得到“更好”或者“同样”的分类精度结果。

综上，实验结果表明，当特征集增加到部分标记混合数据时，使用算法 IFSA 进行特征选择是可行的。

4.2.2 删除特征集时不同算法的性能比较

为测试算法 IFSD 的性能，针对表 2 中的每个数据集，令 C 为每个数据集的特征集，选取 30% 的特征作为删除的特征，即 $0.3*|C|$，并将其平均分为 5 部分，$|c_{de}^{i}|=\frac{0.3*|C|}{5}$，$i=1,2,\cdots 5$，令 $C_{de}^{i}=\bigcup_{j=1}^{i}c_{de}^{j}$，$i=1,2,\cdots,5$ 为 5 个删除特征集。图 3 给出了算法 NIFS，Semi-rough-D，NSFS 和 IFSD 随着删除特征集规模变化的运行时间。



(a) Hepatits 数据集 (b) Horse-colic 数据集 (c) Credit 数据集 (d) Australian 数据集

(e) German 数据集 (f) Wpbc 数据集 (g) Sonar 数据集 (h) Anneal 数据集

图 3 算法 NIFS，Semi-rough-D，NSFS 和 IFSD 的运行时间

Fig. 3 Running time of algorithms NIFS，Semi-rough-D，NSFS and IFSD

由图 3 可知，当删除特征集时，随着删除特征集规模的增大，算法 NIFS，NSFS 和 Semi-rough-D 的运行时间不断缩短，而 IFSD 算法的运行时间并不是完全单调的。当从数据集 Credit 中删除第三个、第四个和第五个特征集时，算法 IFSD 的运行时间分别为 7.94 s，7.37 s 和 7.01 s，这是因为删除特征集时，候选特征子集的变化具有随机性。但是，算法 IFSD 的运行时间明显短于其他 3 种算法，如当数据集 Australian 删除第二个特征集时，算法 NIFS，NSFS 和 Semi-rough-D 的运行时间分别为 34.04 s，38.27 s 和 50.39 s，而算法 IFSA 的运行时间为 11.17 s，相比算法 NIFS，NSFS，Semi- rough-D 分别缩短了 67.18%，70.81%，77.83%；当数据集 Horse-colic 删除第二个特征集时，算法 NIFS，NSFS 和 Semi-rough-D 的运行时间分别为 8.51 s，8.45 s，52.62 s，而算法 IFSA 的运行时间为 3.01 s，相比算法 NIFS，NSFS，Semi- rough-D 分别缩短了 64.62%，64.37%，94.27%。

从上述实验结果可以看出，本文提出的增量式特征选择算法 IFSD 在删除特征集时可以有效缩短运行时间，主要原因是算法 IFSD 在原有特征选择结果的基础上进行特征选择，只需计算部分局部对象的粒度结构，进行特征结果的增量式更新，缩短了计算时间。

表 6 列出了删除特征集时 4 种算法 NIFS，Semi-rough-D，NSFS 和 IFSD 的最终特征选择个数。表 7 和表 8 列出了 4 种算法在 C4.5 和 SVM 分类器下的平均分类精度结果。

表 6 算法 NIFS，Semi-rough-D，NSFS 和 IFSD 的特征选择个数

Table 6 Feature selection numbers of algorithms NIFS，Semi-rough-D，NSFS and IFSD

数据集	特征选择个数			
	NIFS	Semi-rough-D	NSFS	IFSA
Hepatitis	9	8	9	9
Horse-colic	3	2	3	3
Credit	9	8	9	9
Australian	9	8	9	9
German	12	12	12	12
Wpbc	7	7	13	9
Sonar	10	7	10	9
Anneal	10	7	10	10

从表 6 可以看出，当从数据集中删除特征集时，算法 IFSD，NIFS，Semi-rough-D 和 NSFS 均可以有效降低数据维度。同时，针对不同数据集，使用算法 NIFS，Semi-rough-D，NSFS，IFSD 得到的特征选择结果是相近的，例如，在 Credit 数据集下，算法 NIFS，Semi-rough-D，NSFS，IFSD 得到的特征子集个数分别为 9，8，9，9。从表 7 和表 8 的分类精度结果可以看出，与算法 NIFS，Semi-rough-D，NSFS 相比，算法 IFSD 得到的特征子集结果的分类精度是相近的。例如，在 Australian 数据集下，在 SVM 分类器下算法 NIFS，Semi-rough-D，

NSFS,FSD 的分类精度分别为 85.70%,85.55%,85.70%,85.70%。从 Win\Tie\Lose 的统计结果可以看出,与算法 NIFS,Semi-rough-D,NSFS 相比,算法 IFSD 在大部分数据集下可以得到"更好"或者"相同"的分类精度。例如,在 C4.5 分类器下,与算法 Semi-rough-D 相比,算法 IFSD 在 8 个数据集上均可以得到更好的分类精度结果。

表 7 算法 NIFS,Semi-rough-D,NSFS 和 IFSD 在 C4.5 分类器下的分类精度

Table 7 Classification accuracies of algorithms NIFS,Semi-rough-D,NSFS and IFSD with classifier C4.5

数据集	Raw	NIFS	Semi-rough-D	NSFS	IFSD
Hepatitis	**83.95±0.15**	81.93±0.21$^{=}$	81.91±0.63^{+}	81.93±0.21$^{=}$	81.93±0.21
Horse-colic	**79.71±1.08**	73.78±1.58$^{=}$	63.73±0.35^{+}	73.78±1.58$^{=}$	73.78±1.58
Credit	84.94±0.64	85.78±0.48$^{=}$	85.76±0.33^{+}	85.78±0.48$^{=}$	**85.78±0.48**
Australian	85.51±1.16	86.86±0.06$^{=}$	84.83±0.08^{+}	86.86±0.06$^{=}$	**86.86±0.06**
German	73.36±0.56	74.44±0.54$^{=}$	72.62±0.08^{+}	74.44±0.54$^{=}$	**74.44±0.54**
Wpbc	72.22±0.48	75.37±0.22^{+}	76.32±0.17^{+}	74.20±0.15^{+}	**76.33±0.14**
Sonar	69.86±1.62	74.36±0.98^{+}	66.58±0.13^{+}	71.63±1.36^{+}	**75.03±1.35**
Anneal	98.56±0.09	98.93±0.16$^{=}$	98.13±0.11^{+}	98.93±0.16$^{=}$	**98.93±0.16**
Average	81.01	81.43	78.73	80.94	**81.63**
Win\Tie\Lose	—	2/6/0	8/0/0	3/5/0	—

表 8 算法 NIFS,Semi-rough-D,NSFS 和 IFSD 在 SVM 分类器下的分类精度

Table 8 Classification accuracies of algorithms NIFS,Semi-rough-D,NSFS and IFSD with classifier SVM

数据集	Raw	NIFS	Semi-rough-D	NSFS	IFSD
Hepatitis	**83.24±0.13**	81.90±0.08$^{=}$	79.40±0.12^{+}	81.90±0.08$^{=}$	81.90±0.08
Horse-colic	68.70±0.56	69.72±0.52$^{=}$	63.70±0.05^{+}	69.72±0.52$^{=}$	**69.72±0.52**
Credit	85.92±0.08	85.80±0.09$=$	**86.12±0.16**$^{-}$	85.80±0.09$^{=}$	85.80±0.09
Australian	85.70±0.13	85.70±0.04$^{=}$	85.55±0.05^{+}	85.70±0.04$^{=}$	**85.70±0.04**
German	76.37±1.08	**76.58±0.38**$^{=}$	75.58±0.30^{+}	76.58±0.38$^{=}$	**76.58±0.38**
Wpbc	**79.83±0.61**	78.30±0.07^{+}	76.38±0.12^{+}	77.84±0.07^{+}	79.36±0.08
Sonar	79.83±4.76	**81.33±0.51**$^{-}$	70.53±0.23^{+}	80.73±0.44^{+}	80.93±0.63
Anneal	99.11±0.14	99.46±0.04$^{=}$	98.53±0.12^{+}	99.46±0.04$^{=}$	99.46±0.04
Average	82.33	82.34	79.47	82.21	**82.43**
Win\Tie\Lose	—	1/6/1	7/0/1	2/6/0	—

综上,实验结果表明,当特征集从部分标记混合数据中删除时,算法 IFSD 在保证得到特征子集高分类精度的同时可以显著缩短运行时间。

4.3 算法的稳定性分析

为了进一步验证本文算法的稳定性,本文使用谷元距离度量[30]对算法的稳定性进行分析,计算式为:

$$D_T(s,s')=1-\frac{|s|+|s'|-2|s\bigcap s'|}{|s|+|s'|-|s\bigcap s'|}$$

其中,s 和 s' 表示两个特征子集。

表 9 和表 10 分别列出了增加特征集和删除特征集时算法 NIFS,Semi-rough-D,NSFS 和 IFSA(IFSD)的稳定性,其中"Average"这行表示不同算法下的平均稳定性结果。

表 9 算法 NIFS,Semi-rough-D,NSFS 和 IFSA 的稳定性分析

Table 9 Stability of algorithms NIFS,Semi-rough-D,NSFS and IFSA

数据集	NIFS	Semi-rough-D	NSFS	IFSA
Hepatitis	0.63	0.50	0.70	0.63
Horse-colic	0.50	0.91	0.50	0.50
Credit	0.57	0.57	0.57	0.57
Australian	0.53	0.66	0.53	0.53
German	0.60	0.52	0.60	0.60
Wpbc	0.62	0.41	0.26	0.93
Sonar	0.20	0.50	0.50	0.66
Anneal	0.68	0.63	0.76	0.78
Average	0.54	0.58	0.55	0.65

从表 9 中最后一行的平均稳定性结果可以看出,当增加特征集时,算法 NIFS,Semi-rough-D,NSFS 和 IFSA 在 8 个数据集上的平均稳定性结果为 0.54,0.58,0.55 和 0.65,因此算法 IFSA 拥有最高的稳定性结果。同理,从表 10 的结果可以看出,当删除数据集时,相比算法 NIFS,Semi-rough-D 和 NSFS,算法 IFSD 的平均稳定性提高了 0.08,0.22 和 0.14。

表 10 算法 NIFS,Semi-rough-D,NSFS 和 IFSD 的稳定性

Table 10 Stability of algorithms NIFS,Semi-rough-D,NSFS and IFSD

数据集	NIFS	Semi-rough-D	NSFS	IFSD
Hepatitis	0.63	0.60	0.63	0.63
Horse-colic	0.66	0.33	0.25	0.66
Credit	0.80	0.80	0.72	0.80
Australian	0.80	0.80	0.80	0.80
German	0.91	0.78	0.80	0.91
Wpbc	0.60	0.50	0.73	0.68
Sonar	0.27	0.33	0.35	0.63
Anneal	0.71	0.16	0.66	0.90
Average	0.67	0.53	0.61	0.75

综上,实验结果表明,本文算法 IFSA 和 IFSD 具有较高的稳定性。

4.4 算法的统计检验分析

为进一步对不同算法的实验结果进行统计比较,选取 Friedman Test[31]及 Nemenyi Test 两种统计检验方法验证算法对比的有效性。

Friedman Test 作为一种非参数统计检验方法,它的零假设为所有实验算法的分类性能相当,表达式定义为:

$$F_F=\frac{(T-1)\chi_F^2}{T(s-1)-\chi^2{}_F}$$

$$\chi^2{}_F=\frac{12T}{s(s+1)}\left(\sum_{i=1}^{s}R_i^2-\frac{s(s+1)^2}{4}\right)$$

其中，T 和 s 分别为实验数据集和实验算法的数量，R_i 代表算法 i 在不同分类器上分类精度结果的平均排名值。

表 11 和表 12 分别列出了算法 NIFS，Semi-rough-D，NSFS 和 IFSA(IFSD)在分类器 C4.5 和 SVM 上分类精度结果的平均排名。对比实验中，$T=8$，$s=4$，由表 11 和表 12 的平均排名结果可得到算法分类性能的 Friedman 值，如表 13 和表 14 所列。本次检验中一次性检验值为 $\alpha=0.05$，最终可得置信度值为 3.027。

表 11 算法 NIFS，Semi-rough-D，NSFS 和 IFSA 分类精度的平均排名

Table 11 Average ranking of classification accuracies of algorithms NIFS，Semi-rough-D，NSFS and IFSA

性能指标	NIFS	Semi-rough-D	NSFS	IFSA
分类精度 C4.5	2.250	3.625	2.500	1.625
分类精度 SVM	1.812	3.750	2.437	2.000

表 12 算法 NIFS，Semi-rough-D，NSFS 和 IFSD 分类精度结果的平均排名

Table 12 Average ranking of classification accuracies of algorithms NIFS，Semi-rough-D，NSFS and IFSD

性能指标	NIFS	Semi-rough-D	NSFS	IFSD
分类精度 C4.5	2.062	3.625	2.625	1.687
分类精度 SVM	2.062	3.625	2.250	2.062

表 13 算法 NIFS，Semi-rough-D，NSFS 和 IFSA 分类性能的 Friedman 值

Table 13 Friedman statistics of classification performance of algorithms NIFS，Semi-rough-D，NSFS，and IFSA

性能指标	Friedman 值	置信度值($\alpha=0.05$)
分类精度 C4.5	5.873	3.072
分类精度 SVM	5.910	

表 14 算法 NIFS，Semi-rough-D，NSFS 和 IFSD 分类性能的 Friedman 值

Table 14 Friedman statistics of classification performance of algorithms NIFS，Semi-rough-D，NSFS and IFSD

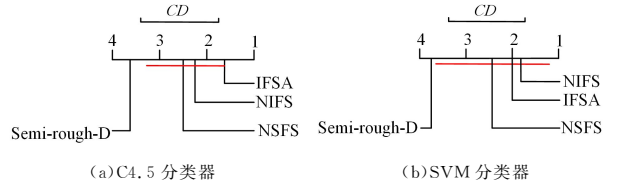
性能指标	Friedman 值	置信度值($\alpha=0.05$)
分类精度 C4.5	5.206	3.072
分类精度 SVM	3.873	

由表 13 和表 14 可知，分类性能的 Fireman 值均大于置信度值 3.027，可以拒绝零假设。这表明本文提出的算法 IFSA 和 IFSD 在分类精度(C4.5，SVM)上与其他算法存在差异。

此外，Nemenyi Test 可进一步分析所有比较算法的相对性能及差异，其临界差公式定义为：

$$CD_\alpha = q_\alpha \sqrt{\frac{s(s+1)}{6T}}$$

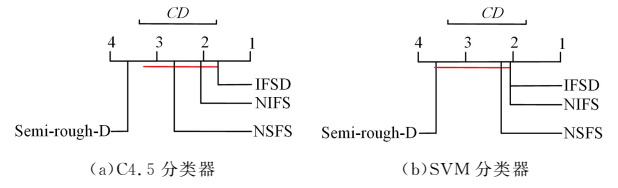
在算法对比实验中，$T=8$，$s=4$，$\alpha=0.05$。根据临界值表可得 $q_{(0.05)}=2.569$，进而可得 $CD_{(0.05)}=1.658$。由表 13 和表 14 的平均排名及 $CD_{(0.05)}$ 可绘制算法 IFSA(IFSD)与其他对比算法的 Nemenyi 检验结果图，如图 4 和图 5 所示。由图 4 可知，在 C4.5 和 SVM 分类器下，算法 IFSA 显著优于算法 Semi-rough-D。然而，对于算法 NIFS 和 NSFS，它们在分类精度值上并不存在着显著的差异。由图 5 可知，在 C4.5 分类器下，算法 IFSD 显著优于算法 Semi-rough-D。而在 SVM 分类器下，算法 NIFS，Semi-rough-D，NSFS 和 IFSD 在分类精度值上并不存在着显著的差异。



(a)C4.5 分类器 (b)SVM 分类器

图 4 算法 IFSA 与其他对比算法的 Nemenyi 检验结果

Fig. 4 Comparisons between IFSA and other comparison algorithms for Nemenyi test



(a)C4.5 分类器 (b)SVM 分类器

图 5 算法 IFSD 与其他对比算法的 Nemenyi 检验结果

Fig. 5 Comparisons between IFSD and other comparison algorithms for Nemenyi test

结束语 针对特征集动态变化的部分标记混合数据，静态(非增量式)特征选择算法在进行特征选择时将特征集动态变化后的数据看作新数据集，重新对新数据集进行特征选择，这样浪费了大量的时间，无法实时地更新特征选择结果，进而无法高效快速地得到特征集动态变化后的特征选择结果。为此，针对部分标记混合数据的特征集的动态变化，本文提出了基于信息粒度的增量式特征选择算法。首先，结合增量学习策略，给出了信息粒度的增量式更新机制；然后，在此基础上，提出了部分标记混合数据中特征集增加和删除的增量式特征选择算法；最后，实验结果表明，相比非增量式特征选择算法，所提算法在不降低分类精度的情况下，大大缩短了特征选择的运行时间。同时，我们下一步的研究工作将考虑针对部分标记混合数据下对象与特征同时发生变化时，如何高效地进行特征选择。

参 考 文 献

[1] WANG C Z，HUANG Y，SHAO M W，et al. Feature selection based on neighborhood self-information[J]. IEEE Transactions on Cybernetics，2019，99(7)：1-12.

[2] WANG Q，QIAN Y H，LIANG X Y，et al. Local neighborhood rough set[J]. Knowledge-Based Systems，2018，153(8)：53-64.

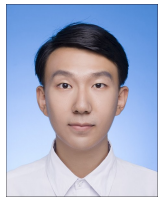
[3] WANG D，CHEN H M，LI T R，et al. A novel quantum grasshopper optimization algorithm for feature selection[J]. International Journal of Approximate Reasoning，2020，127(12)：122-150.

[4] PAWLAK Z. Rough sets[J]. International Journal of Computer and Information Sciences，1982，11(5)：341-356.

[5] ZHENG N，WANG J Y. Evidence characteristics and attribute reduction of incomplete ordered information system[J]. Computer Engineering and Applications，2018，54(21)：43-47.

[6] JIANG Z H，LIU K Y，YANG X B，et al. Accelerator for supervised neighborhood based attribute reduction[J]. International

Journal of Approximate Reasoning, 2020, 119(4): 122-150.
[7] WAN Y, CHEN X L, ZHANG J H, et al. Semi-supervised feature selection based on low-rank sparse graph embedding[J]. Journal of Image and Graphics, 2018, 23(9): 1316-1325.
[8] LIU K Y, YANG X B, YU H L, et al. Supervised information granulation strategy for attribute reduction[J]. International Journal of Machine Learning and Cybernetics, 2020, 11(3): 2149-2163.
[9] HU Q H, XIE Z X, YU D R. Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation[J]. Pattern Recognition, 2007, 40(12): 3509-3521.
[10] JING Y G, LI T R, FUJITA H, et al. An incremental attribute reduction method for dynamic data mining[J]. Information Sciences, 2018, 465(7): 202-218.
[11] WEI W, LIANG J Y, QIAN Y H. A comparative study of rough sets for hybrid data[J]. Information Sciences, 2012, 190(6): 1-16.
[12] WANG F, LIU J C, WEI W. Semi-supervised feature selection algorithm based on information entropy[J]. Computer Science, 2018, 45(11): 427-430.
[13] DAI J H, HU Q H, ZHANG J H, et al. Attribute selection for partially labeled categorical data by rough set approach[J]. IEEE Transactions on Cybernetics, 2017, 47(9): 2460-2471.
[14] LIU K Y, YANG X B, YU H L, et al. Rough set based semi-supervised feature selection via ensemble selector[J]. Knowledge-Based Systems, 2019, 165(1): 282-296.
[15] XIAO L S, WANG H J, YANG Y. Semi-supervised feature selection based on attribute dependency and hybrid constraint[J]. Journal of Computer Applications, 2015, 35(12): 80-84.
[16] MA F M, DING M W, ZHANG T F, et al. Compressed binary discernibility matrix based incremental attribute reduction algorithm for group dynamic data[J]. Neurocomputing, 2019, 334(6): 20-27.
[17] SHU W H, QIAN W B, XIE Y H. Incremental approaches for feature selection from dynamic data with the variation of multiple objects[J]. Knowledge-Based System, 2019, 163(1): 320-331.
[18] HUANG Q Q, LI T R, HUANG Y Y, et al. Incremental three-way neighborhood approach for dynamic incomplete hybrid data[J]. Information Sciences, 2020, 541(12): 98-122.
[19] LIU Y, ZHENG L D, XIU Y L, et al. Discernibility matrix based incremental feature selection on fused decision tables[J]. International Journal of Approximate Reasoning, 2020, 118(3): 1-26.
[20] ZENG A P, LI T R, LIU D, et al. A fuzzy rough set approach for incremental feature selection on hybrid information systems[J]. Fuzzy Sets and Systems, 2015, 258(6): 39-60.
[21] YU J H, CHEN M H, XU W H. Dynamic computing rough approximations approach to time-evolving information granule interval-valued ordered information system[J]. Applied Soft Computing, 2017, 60(6): 18-29.
[22] CAI M J, LANG G M, FUJITA H, et al. Incremental approaches to updating reducts under dynamic covering granularity[J]. Knowledge-Based Systems, 2019, 172(1): 130-140.
[23] WANG S, LI T R, LUO C, et al. A novel approach for efficient updating approximations in dynamic ordered information systems[J]. Information Sciences, 2020, 507(8): 197-219.
[24] HUANG Y Y, LI T R, LUO C, et al. Dynamic maintenance of rough approximations in multi-source hybrid information systems[J]. Information Sciences, 2020, 530(8): 108-127.
[25] LIU D, LI T R, ZHANG J B. Incremental updating approximations in probabilistic rough sets under the variation of attributes[J]. Knowledge-Based System, 2015, 73(1): 81-96.
[26] ZHANG Y Y, LI T R, LUO C, et al. Incremental updating of rough approximations in interval-valued information systems under attribute generalization[J]. Information Sciences, 2016, 373(12): 461-475.
[27] UCI Machine Learning Repository [OL]. http://archive.ics.uci.edu/ml/datasets.html.
[28] Rosetta: A rough set toolkit for analysis of data[OL]. http://www.lcb.uu.se/tools/rosetta/index.php.
[29] MARIELLO A, BATTITI R. Feature selection based on the neighborhood entropy[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(12): 6313-6322.
[30] LIU Y, CAO J J, DIAO X C, et al. Survey on stability of feature selection[J]. Journal of Software, 2018, 29(9): 2559-2579.
[31] FRIEDMAN M. A comparison of alternative tests of significance for the problem of m rankings[J]. The Annals of Mathematical Statistics, 1940, 11(1): 86-92.



YAN Zhen-chao, born in 1997, postgraduate. His main research interests include granular computing, knowledge discovery, data mining, etc.

SHU Wen-hao, born in 1985, Ph.D, associate professor, master supervisor. Her main research interests include data mining, knowledge discovery, etc.

(责任编辑:喻藜)