属性约简簇的优化选择

邓大勇[1,2],葛雅雯[2],黄厚宽[3]
(1. 浙江师范大学行知学院,浙江金华 321004;2. 浙江师范大学数理与信息工程学院,浙江金华 321004;
3. 北京交通大学计算机与信息技术学院,北京 100044)

摘 要: 属性约简是粗糙集的一个重要应用.一个数据集往往含有多个属性约简,人们一般用启发式算法找到其中的一个,再通过实验的方法验证其有效性.面对多个属性约简,人们往往难以区别,缺乏有效的手段选取最优或较优的属性约简.使用多种概念漂移的度量指标和信息损失的度量方法比较了同一个知识系统中不同 Pawlak 约简之间的区别与联系.提出了属性约简重心的概念,并研究其性质.实验结果显示,在众多的属性约简中,离重心最近的属性约简在分类准确率方面具有较大的优势.概念漂移的度量指标和信息损失的度量方法有助于区分不同的属性约简,属性约简的重心有助于在众多的属性约简中选择最优或较优的一个.

关键词: 粗糙集;属性约简;概念漂移;属性约简重心
中图分类号: TP18 **文献标识码:** A **文章编号:** 0372-2112 (2019)05-1111-10
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2019.05.019

An Optimizing Selection in a Family of Attribute Reducts

DENG Da-yong[1,2], GE Ya-wen[2], HUANG Hou-kuan[3]
(1. *Xingzhi College, Zhejiang Normal University, Jinhua, Zhejiang* 321004, *China*;
2. *College of Mathematics, Physics and Information Engineering, Zhejiang Normal University, Jinhua, Zhejiang* 321004, *China*;
3. *School of Computer and Information Technology, Beijing Jiaotong University, Beijing* 100044, *China*)

Abstract: Attribute reduction is one of important applications in rough set theory. There are more than one attribute reduct in a data set, and heuristic algorithms are always used to find one of them, which is verified with experiments. For many attribute reducts, it is hard for people to distinguish them, and lacks of valid methods of selecting the best one or a better one. Indexes of concept drift and information loss are employed to compare the same type of Pawlak attribute reducts in a knowledge system. The focus of attribute reducts is presented, and its properties are investigated in this paper. Experimental results show that the closest attribute reduct to the focus of attribute reducts is better than other attribute reducts in classification accuracy. Indexes of concept drift detection and information loss can distinguish different attribute reducts, and the focus of attribute reducts can be employed to select the best attribute reduct or a better one.

Key words: rough sets; attribute reduction; concept drift; focus of attribute reducts

1 引言

粗糙集理论[1]产生至今已经30多年,衍生了很多粗糙集模型,包括可变精度粗糙集[2]、覆盖粗糙集[3]、邻域粗糙集[4,5]、多粒度粗糙集[6]、S-粗糙集[7]、F-粗糙集[8]和全粒度粗糙集[9]等,每一种粗糙集都有很多种约简,包括:正区域约简[1]、互信息约简[10]、差别矩阵约简[11]、动态约简[12]、并行约简[8]等.

各种约简因为约简准则不同,它们之间存在着一定的区别和联系. Kryszkiewicz[13]比较了多种约简之间的关系,在不完备信息系统中提出了广义决策保持约简;邓大勇等[14]在不一致信息系统中比较了多种属性约简之间的包含关系;徐章艳等[15]研究了HU差别矩阵、信息熵、分布、最大分布、近似和正区域的属性约简的关系;李磊军等[16]从代数和信息论的观点研究了在一致和不一致信息系统中属性约简之间的等价关系和包含关系;张东娜等[17]分别用粗糙集、主成份分析和奇异值分解的方法进行属性约简,并探究了这三种约简之间的关系.所有这些研究都是基于不同的约简准则或不同的约简方法,对于同一种约简准则或约简方法下的不同属性约简却缺乏比较研究.众所周知,粗糙集每一种约简准则下都有若干个属性约简,求解同一约简准则下所有约简或最小约简

收稿日期:2018-06-10;修回日期:2018-09-06;责任编辑:蓝红杰
基金项目:国家自然科学基金(No. 61473030);浙江省自然科学基金(No. LY15F020012);浙江师范大学网络空间安全浙江省一流学科

是一个 NP-hard 问题[18]. 不同属性约简进行数据分类的分类准确率并不完全相同,求取属性约简往往使用启发式算法或智能算法,这些算法求取的属性约简的分类准确率受测试集的影响较大.

针对以上问题,本文用上下近似概念漂移[9,19]、属性约简的信息损失[20]等方法比较不同情况下多个 Pawlak 约简之间的异同,提出了属性约简重心的概念. 把每个属性约简生成的属性重要性向量看成多维空间中的一个质点,就可以求得同一决策系统或信息系统中多个属性约简的重心. 属性约简的重心符合中国古典哲学"中庸"的思想. 实验结果显示,离重心最近的属性约简在分类准确率方面具有较大的优势,属性约简的重心能够帮助人们在众多的属性约简中选择最优或较为合理的一个.

2 基础知识

假设读者对粗糙集理论较为熟悉,本节仅简单介绍粗糙集理论[1]、上下近似概念漂移[9,19]、质概念漂移和量概念漂移[21]等基础知识.

2.1 粗糙集

设 $K=(U,A)$ 是一个知识系统(根据需要,可以称为信息系统或决策系统),其中 U 为论域,A 为等价关系簇(或属性集合). A 对 U 的划分为 $U/A=\{[x]_A:x\in U\}$,其中 $[x]_A$ 是 $x\in U$ 关于 A 的等价类. 概念 $X\subseteq U$ 关于 A 的上、下近似定义为:

$$\overline{A}(X)=\cup\{[x]_A:[x]_A\cap X\neq\varnothing\};$$

$$\underline{A}(X)=\cup\{[x]_A:[x]_A\subseteq X\}.$$

序偶 $(\underline{A}(X),\overline{A}(X))$ 称为概念 $X\subseteq U$ 关于 A 的粗糙集.

定义 1[1] 设 $DS=(U,A,d)$ 是一个决策系统,则 $DS=(U,A,d)$ 的正区域定义如下:

$$POS(DS,A)=\bigcup_{Y\in U/|d|}\underline{A}(Y)$$

定义 2[1] 设 $DS=(U,A,d)$ 是一个决策系统,$B\subseteq A$ 称为决策系统 $DS=(U,A,d)$ 的属性约简 *iff* $B\subseteq A$ 满足以下 2 个条件:

(1) $POS(DS,A)=POS(DS,B)$;

(2)对于任意 $S\subset B$,都有 $POS(DS,A)\neq POS(DS,S)$.

定义 3[1] 设 $DS=(U,A,d)$ 是一个决策系统,则 d 对 A 的依赖度定义为:

$$\gamma(DS,A,d)=\frac{|POS(DS,A)|}{|U|}$$

其中 $|\cdot|$ 表示集合的势.

定义 4[1] 设 $DS=(U,A,d)$ 是一个决策系统,则 $a\in A$ 的属性重要性(内部属性重要性)定义为:

$$\sigma(DS,A,a)=\frac{|POS(DS,A)-POS(DS,A-\{a\})|}{|U|}$$

对于条件属性集合 A 中的每一个属性都存在一个相对于 A 的属性重要性,A 中所有属性的属性重要性构成一个向量,记为 $\boldsymbol{\Pi}$.

2.2 概念漂移与信息损失

对文献[9,19]中定义的上、下近似概念漂移进行改进,并对文献[21]中的质概念漂移改进,得到下面定义:

定义 5 设 $IS=(U,A)$ 是一个知识系统,$X\subseteq U$ 是一个概念,$B_1\subseteq A$ 和 $B_2\subseteq A$ 是系统中的两个不同的知识(属性子集),则概念 $X\subseteq U$ 在 $B_1\subseteq A$ 和 $B_2\subseteq A$ 表示下的上、下近似漂移被定义为:

(1) $\overline{\Delta}(B_1,B_2,X)=\overline{B_1}(X)\oplus\overline{B_2}(X)$;

(2) $\underline{\Delta}(B_1,B_2,X)=\underline{B_1}(X)\oplus\underline{B_2}(X)$.

其中⊕表示集合的对称差.

定义 6 在决策系统 $DS=(U,A,d)$ 中,B_1,B_2 是它的 2 个不同的属性约简(或者是同一个概念的 2 个不同属性约简),则 B_1,B_2 之间的质概念漂移和量概念漂移分别定义如下:

(1)质概念漂移定义为 $B_1\oplus B_2$;

(2)量概念漂移定义为

$$\Delta(B_1,B_2)=\sqrt{\sum_{a\in B}(\sigma(DS,B_1,a)-\sigma(DS,B_2,a))^2},$$

其中 $B=B_1\cup B_2$.

定义 7[20] 设 $DS=(U,A,d)$ 是一个决策系统,$B\subseteq A$ 是它的一个约简,则 $B\subseteq A$ 约简的信息损失定义如下:

$$\delta(B)=H(DS,A)-H(DS,B)$$

$B\subseteq A$ 约简的信息损失率定义如下:

$$s(B)=\frac{\delta(B)}{H(DS,A)}\times 100\%$$

其中,$H(DS,A)$ 表示知识系统 $IS=(U,A)$ 的信息熵.

下文,我们从上、下近似概念漂移,质概念漂移,量概念漂移和信息损失等角度比较同一个知识系统中不同 Pawlak 约简之间的联系与区别,比较它们的"同"和"异".

3 Pawlak 约简之间的比较

3.1 信息系统中属性约简之间的比较

信息系统缺乏决策属性或不考虑决策属性时,信息系统的属性约简称为绝对约简. 不同绝对约简之间的区别与联系可用下面定理来表示.

定理 1 信息系统中,不同的属性约简(绝对约简)之间的质概念漂移非空,量概念漂移大于 0,信息损失都等于 0,任何概念的上、下近似概念漂移均等于∅.

证明 设 $B_1,B_2\subseteq A$ 是信息系统 $IS=(U,A)$ 的 2 个不相同的绝对约简,显然有 $B_1\oplus B_2\neq\varnothing$,从而有 $\Delta(B_1,B_2)>0$. 根据文献[20]中的命题,绝对约简的信息损失等于 0. 又由于 $U/B_1=U/B_2=U/A$,所以对于任何概念 $X\subseteq$

U,有$\underline{B_1}(X)=\underline{B_2}(X)=\underline{A}(X)$,$\overline{B_1}(X)=\overline{B_2}(X)=\overline{A}(X)$,即:任何概念的上、下近似概念漂移均等于∅. 证毕.

由定理1可知,虽然不同的绝对约简在信息损失方面都相同,对任何概念,它们的上、下近似概念漂移均等于∅,但是它们之间的质概念漂移并不等于空,量概念漂移并不等于0. 也就是说,同一个信息系统的不同绝对约简既有相同的一面,也有不同的地方.

3.2　单个概念属性约简之间的比较

对于单个概念来说,因为不太方便考虑约简的信息损失,所以,我们仅仅考虑不同属性约简之间的上、下近似概念漂移,质概念漂移和量概念漂移.

定理2　对于单个概念的不同属性约简,下面结论成立:

(1)下近似漂移等于∅;

(2)上近似漂移不一定等于∅;

(3)质概念漂移非∅;

(4)量概念漂移大于0.

证明

(1)对于单个概念来说,所有的属性约简都是保持下近似不变,所以,单个概念的不同属性约简的下近似漂移等于∅;

(2)单个概念的不同属性约简,仅仅保持下近似不变,上近似并不一定相等,所以上近似漂移不一定等于∅;

(3)因为不同属性约简里包含的属性一定不完全相同,所以它们的质概念漂移非∅;

(4)根据量概念漂移的计算公式,显然,单个概念的不同属性约简之间的量概念漂移大于0.

根据定理2,对于单个概念来说,不同的属性约简仅仅可以保持下近似概念漂移等于∅,而上近似概念漂移有可能存在,质概念漂移一定非∅,量概念漂移一定大于0. 也就是说,单个概念的不同属性约简虽然存在相同的部分,但是在一些指标上也存在着很大的不同.

3.3　决策系统中属性约简之间的比较

决策系统可以看成是决策属性划分的集合,由若干个概念组成[21]. 同一个决策系统中不同属性约简(Pawlak 约简)之间的关系如下.

定理3　对于决策系统不同的属性约简,下面结论成立:

(1)下近似漂移等于∅;

(2)上近似漂移不一定等于∅;

(3)质概念漂移非∅;

(4)量概念漂移大于0;

(5)信息损失不一定相等.

证明

(1)决策属性的每个概念对于 Pawlak 约简都能保持下近似不变,根据下近似漂移的定义,决策系统中的不同约简下近似漂移等于∅;

(2)根据 Pawlak 约简的定义和上近似漂移的定义,易证此结论成立.

根据相关定义,显然,结论(3)(4)(5)成立.

在一致的决策系统中,不同属性约简之间的关系也满足定理3. 只不过定理3中的结论(2)可以更强,如下所述.

定理4　在一致的决策系统中,不同约简之间的上近似漂移等于∅.

证明　根据相关定义,易证本定理.

4　属性约简的重心

通过前面的讨论发现,不同的属性约简既有相同的一面,又有不同的一面,如何在众多的属性约简中寻找到最优或较优的一个,这是人们一直在探寻的问题. 人们往往使用启发式算法或智能算法找到一个属性约简,再通过实验的方法验证其有效性. 面对众多的属性约简,通过实验的方法验证其有效性,往往工作量大,时间复杂性高. 如何在众多的属性约简中找到最优或较优的一个是以下2节要探讨的问题.

定义8　设 $B_1,B_2,\cdots,B_k$ 是决策系统 $DS=(U,A,d)$(或信息系统 $IS=(U,A)$)的 k 个属性约简,$\boldsymbol{\Pi}_1,\boldsymbol{\Pi}_2,\cdots,\boldsymbol{\Pi}_k$ 分别是 $B_1,B_2,\cdots,B_k$ 对应的属性重要性向量,则 $B_1,B_2,\cdots,B_k$ 的重心定义为 $\dfrac{\sum_{i=1}^{k}\boldsymbol{\Pi}_i}{k}$.

注:在求属性约简的重心时,不同属性约简所含的属性并不完全相同,当属性约简中不含有某个属性时,在该属性约简中这个属性的重要性为0.

属性集的重心是属性约简生成的属性重要性向量的均值,反映了不同属性约简内部各属性重要性的中心,它并不对应着一个具体的属性约简,只是一个虚拟的重心(或中心). 它具有如下性质.

命题1　2个属性约简的重心与这2个属性约简的距离(量概念漂移)相等.

命题2　$k(k>2)$个属性约简与它们的重心之间的距离并不完全相等.

特殊情况下,k 个属性约简与重心之间的距离可能均相等.

命题3　求解决策系统 $DS=(U,A,d)$ 所有属性约简的重心是一个 NP-hard 问题.

证明　因为求解决策系统 $DS=(U,A,d)$ 的所有约简是一个 NP-hard 问题[18],对于每个属性约简求解属性重要性向量是一个多项式时间复杂度问题,求解每个属性约简与属性约简的重心之间的距离也是一个多项式时间复杂度问题,所以本命题成立.

关于属性约简的重心,我们有如下猜想:对多个(大于等于3)属性约简来说,离重心最近的属性约简的分类准确率较大.

这个猜想提供了一种在若干个属性约简中选择一个最优或较优属性约简的方法. 它满足中国古典哲学"中庸"的思想,也符合人们的直觉:性格是大家的平均数的人比较容易被大家所接受,外貌是大家平均数的人被人们认为最美,男女生的平均身高被人们认为是标准身高,等等.

下一节,我们将通过实验来验证这个猜想,并比较离重心最近的属性约简与基于属性重要度的属性约简在分类准确率上的优劣.

5 实验结果

实验选用了 UCI 机器学习数据库中的 6 个数据集. 其中,kddcup99 随机选取了 10,000 条数据(注:实验中需要求出若干个属性约简,由于计算机的计算能力限制,我们仅取 kddcup99 中的一部分数据)记录,如表 1 所示. 每个数据集随机选取一组 90% 的数据作为训练集,每组数据集再随机选取 10 组 10% 的数据作为测试集.

表 1 数据集

序号	数据名称	Instances	Condition attributes number
1	Horse Colic	300	27
2	Credit Approval	690	15
3	Solar Flare	1066	12
4	Chess	3196	36
5	Mushroom	8124	22
6	Kddcup99	10000	41

本文采用重庆邮电大学开发的 RIDAS 系统进行属性约简. 使用差别矩阵约简方法求取属性约简集合,实验通过 3 ~ 6 个属性约简组成不同的形状(如,三个属性约简组成一个三边形),找出每组属性约简的重心,并算出重心与每个属性约简的欧式距离. 同时算出每个属性约简的分类准确率(注:使用决策树进行分类). 每组属性约简用于分类时的分类准确率的关系如表 2 所示.

表 2 中平均比例的计算公式为:百分比 $= \frac{\sum |\{\alpha:\beta \geq \alpha\}|}{n} \times 100\%$,其中 α 表示离重心距离较远的属性约简的分类准确率,β 表示离重心距离最近的属性约简的分类准确率,$|\{\alpha:\beta \geq \alpha\}|$ 表示离重心距离最近的属性约简的分类准确率大于等于离重心距离较远的属性约简的分类准确率的属性约简个数,$n = 10$(组测试集)× 多边形的个数. N/A 表示该数据集没有五个或六个以上的属性约简.

由表 2 可知,6 组数据集满足属性约简离重心距离越近分类准确率越高的平均比例均在 50% 以上,最高百分比达到 100%,即离重心最近的属性约简分类准确率最高.

下面比较 6 个数据集不同情况下,离重心最近的属性约简与基于属性重要度的属性约简在分类准确率方面的优劣.

表 2 离重心最近的属性约简分类准确率高的百分比

数据名称	三边形平均比例	四边形平均比例	五边形平均比例	六边形平均比例
Horse Colic	100%	100%	N/A	N/A
Credit Approval	90%	97.78%	100%	100%
Solar Flare	86.25%	73.23%	N/A	N/A
Chess	56.67%	65.56%	77.08%	64%
Mushroom	100%	100%	N/A	N/A
Kddcup99	80%	81.67%	59.17%	100%

表 3 中"三边形平均分类准确率"表示几组由同类型属性约简组成的三边形中,离重心距离最近的属性约简分类准确率的均值,三边形的分类准确率比较也可用图 1 表示. 四边形的分类准确率比较可用图 2 表示.

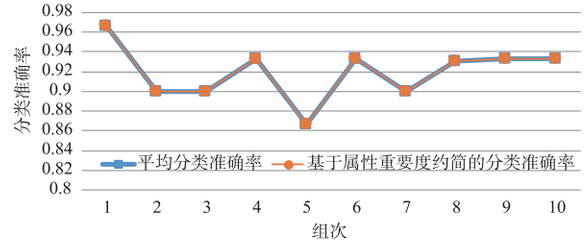


图1 分类准确率比较(Horse Colic 三边形)

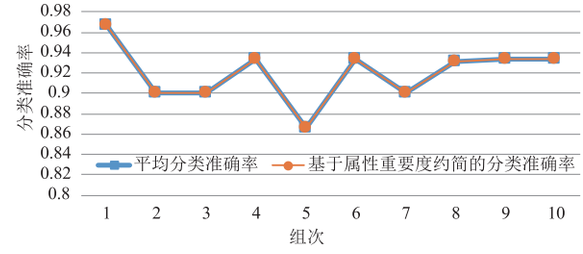


图2 分类准确率比较(Horse Colic 四边形)

类似的,表 4 ~ 表 8 均同表 3,图 3 ~ 图 18 与图 1、图 2 意义相同.

从表 3 可得,离重心最近的属性约简分类准确率与基于属性重要度约简的分类准确率均相等.

从图 1 和图 2 可以看出,离重心距离最近的属性约简的分类准确率等于基于属性重要度约简的分类准确率.

从图 3、图 4、图 5 和图 6 中可以看出,离重心距离最近的属性约简的分类准确率高于基于属性重要度约简的分类准确率,离重心距离最近的属性约简更优.

将表 4、表 5 中的数据进行比较,可以发现离重心距离最近的属性约简分类准确率均大于等于基于属性重要度约简的分类准确率.

表 3　Horse Colic 分类准确率比较

序号	三边形平均分类准确率	四边形平均分类准确率	五边形平均分类准确率	六边形平均分类准确率	基于属性重要度约简的平均分类准确率
1	0.966667 (±0.0)	0.966667 (±0.0)	N/A	N/A	0.966667
2	0.9 (±0.0)	0.9 (±0.0)	N/A	N/A	0.9
3	0.9 (±0.0)	0.9 (±0.0)	N/A	N/A	0.9
4	0.933333 (±0.0)	0.933333 (±0.0)	N/A	N/A	0.933333
5	0.866667 (±0.0)	0.866667 (±0.0)	N/A	N/A	0.866667
6	0.933333 (±0.0)	0.933333 (±0.0)	N/A	N/A	0.933333
7	0.9 (±0.0)	0.9 (±0.0)	N/A	N/A	0.9
8	0.931034 (±0.0)	0.931034 (±0.0)	N/A	N/A	0.931034
9	0.933333 (±0.0)	0.933333 (±0.0)	N/A	N/A	0.933333
10	0.933333 (±0.0)	0.933333 (±0.0)	N/A	N/A	0.933333

表 4　Credit Approval 分类准确率比较

序号	三边形平均分类准确率	四边形平均分类准确率	五边形平均分类准确率	六边形平均分类准确率	基于属性重要度约简的平均分类准确率
1	0.915459 (±0.005917)	0.922705 (±0.007484)	0.925121 (±0.005917)	0927536 (±0.005917)	0.898551
2	0.942029 (±0.0)	0.942029 (±0.0)	0.942029 (±0.0)	0.942029 (±0.0)	0.942029
3	0.913043 (±0.0)	0.913043 (±0.0)	0.913043 (±0.0)	0.913043 (±0.0)	0.913043
4	0.913043 (±0.0)	0.913043 (±0.0)	0.913043 (±0.0)	0.913043 (±0.0)	0.913043
5	0.898551 (±0.0)	0.898551 (±0.0)	0.898551 (±0.0)	0.898551 (±0.0)	0.898551
6	0.929952 (±0.005917)	0.937198 (±0.007484)	0.939613 (±0.005917)	0.942029 (±0.005917)	0.927536
7	0.9009663 (±0.005916)	0.908212 (±0.007484)	0.908212 (±0.005916)	0.913043 (±0.005916)	0.898551
8	0.9009663 (±0.005916)	0.908212 (±0.007484)	0.908212 (±0.005916)	0.913043 (±0.005916)	0.898551
9	0.867647 (±0.0)	0.867647 (±0.0)	0.867647 (±0.0)	0.867647 (±0.0)	0.867647
10	0.898551 (±0.0)	0.898551 (±0.0)	0.898551 (±0.0)	0.898551 (±0.0)	0.898551

由图 7、图 8 可以看出,离重心距离最近的属性约简的分类准确率相对于基于属性重要度约简的分类准确率更有优势.

将表 6 中的数据进行比较,可以发现绝大多数离重心距离最近的属性约简分类准确率大于等于基于属性重要度约简的分类准确率. 其中,四、五、六边形离最近的属性约简分类准确率大于等于基于属性重要度约简的分类准确率的情况优于三边形.

由图 9、图 10、图 11 和图 12 可以看出,绝大多数离重心距离最近的属性约简的分类准确率大于等于基于属性重要度约简的分类准确率.

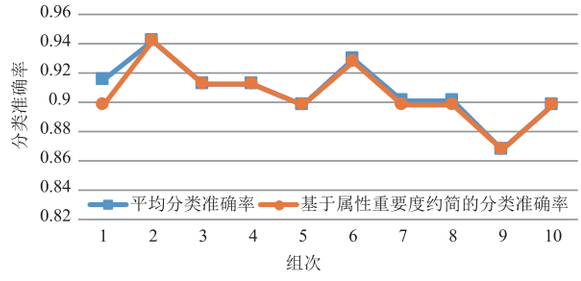


图3 分类准确率比较（Credit Approval 三边形）

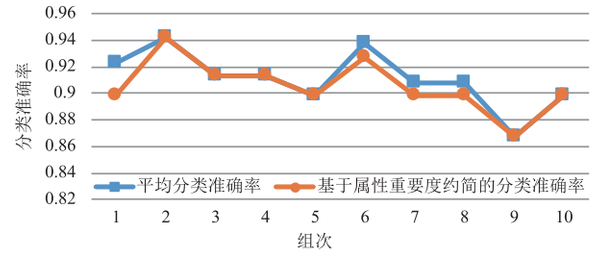


图4 分类准确率比较（Credit Approval 四边形）

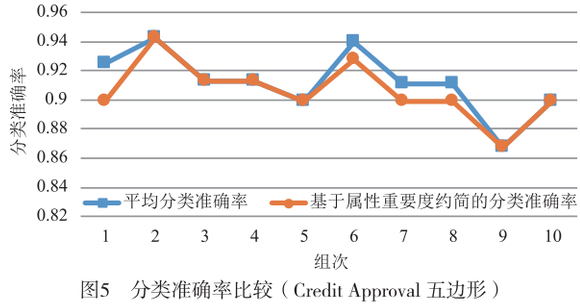


图5 分类准确率比较（Credit Approval 五边形）

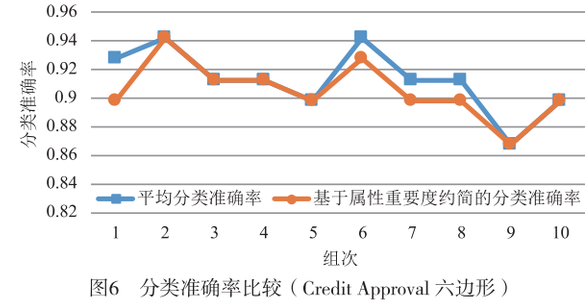


图6 分类准确率比较（Credit Approval 六边形）

表 5 Solar Flare 分类准确率比较

序号	三边形平均分类准确率	四边形平均分类准确率	五边形平均分类准确率	六边形平均分类准确率	基于属性重要度约简的平均分类准确率
1	0.958904 (±0.0)	0.958904 (±0.0)	N/A	N/A	0.958904
2	0.967742 (±0.0)	0.967742 (±0.0)	N/A	N/A	0.967742
3	0.985714 (±0.0)	0.985714 (±0.0)	N/A	N/A	0.985714
4	0.954545 (±0.0)	0.954545 (±0.0)	N/A	N/A	0.954545
5	0.992308 (±0.007693)	0.984615 (±0.0)	N/A	N/A	0.969231
6	0.960145 (±0.013876)	0.956522 (±0.0)	N/A	N/A	0.942029
7	1.0 (±0.007813)	1.0 (±0.0)	N/A	N/A	1.0
8	0.988281 (±0.006329)	0.984375 (±0.0)	N/A	N/A	0.96875
9	0.955696 (±0.007693)	0.949367 (±0.0)	N/A	N/A	0.936709
10	0.996154 (±0.0)	1.0 (±0.0)	N/A	N/A	0.984615

表 6 Chess 分类准确率比较

序号	三边形平均分类准确率	四边形平均分类准确率	五边形平均分类准确率	六边形平均分类准确率	基于属性重要度约简的平均分类准确率
1	0.955208 (±0.015052)	0.945833 (±0.004705)	0.942188 (±0.001712)	0.940625 (±0.016268)	0.912382
2	0.957813 (±0.023854)	0.96875 (±0.004841)	0.970313 (±0.005135)	0.975 (±0.022045)	0.953125
3	0.968229 (±0.010029)	0.965104 (±0.009358)	0.960938 (±0.008558)	0.953125 (±0.012565)	0.96875

续表

序号	三边形平均分类准确率	四边形平均分类准确率	五边形平均分类准确率	六边形平均分类准确率	基于属性重要度约简的平均分类准确率
4	0.958854 (±0.012250)	0.974479 (±0.003653)	0.978125 (±0.003423)	0.975 (±0.013172)	0.965625
5	0.959375 (±0.014034)	0.966667 (±0.002552)	0.978125 (±0.003423)	0.975 (±0.014389)	0.965625
6	0.952083 (±0.024606)	0.972917 (±0.005103)	0.978125 (±0.003423)	0.975 (±0.024240)	0.9625
7	0.958333 (±0.006455)	0.954688 (±0.003827)	0.957813 (±0.005135)	0.953125 (±0.007126)	0.953125
8	0.921875 (±0.015138)	0.969271 (±0.003072)	0.967188 (±0.001712)	0.965625 (±0.016467)	0.971875
9	0.946875 ±(0.014114)	0.926042 (±0.008307)	0.934375 (±0.003423)	0.93125 (±0.013752)	0.9125
10	0.945486 (±0.019862)	0.957292 (±0.005103)	0.9625 (±0.003423)	0.959375 (±0.019230)	0.953125

表 7　Mushroom 分类准确率比较

序号	三边形平均分类准确率	四边形平均分类准确率	五边形平均分类准确率	六边形平均分类准确率	基于属性重要度约简的平均分类准确率
1	1.0 (±0.0)	1.0 (±0.0)	N/A	N/A	0.94528
2	1.0 (±0.0)	1.0 (±0.0)	N/A	N/A	0.964432
3	1.0 (±0.0)	1.0 (±0.0)	N/A	N/A	0.963064
4	1.0 (±0.0)	1.0 (±0.0)	N/A	N/A	0.963064
5	1.0 (±0.0)	1.0 (±0.0)	N/A	N/A	0.957592
6	1.0 (±0.0)	1.0 (±0.0)	N/A	N/A	0.94528
7	1.0 (±0.0)	1.0 (±0.0)	N/A	N/A	0.956224
8	1.0 (±0.0)	1.0 (±0.0)	N/A	N/A	0.963064
9	1.0 (±0.0)	1.0 (±0.0)	N/A	N/A	0.956224
10	1.0 (±0.0)	1.0 (±0.0)	N/A	N/A	0.942544

表 8　Kddcup**99** 分类准确率比较

序号	三边形平均分类准确率	四边形平均分类准确率	五边形平均分类准确率	六边形平均分类准确率	基于属性重要度约简的平均分类准确率
1	0.999177 (±0.001006)	0.997742 (±0.001637)	0.996921 (±0.002023)	0.998768 (±0.001508)	0.992611
2	0.997883 (±0.001275)	0.999177 (±0.002016)	0.997531 (±0.002705)	1.0 (±0.002066)	0.993827
3	0.999392 (±0.001037)	0.998518 (±0.002594)	0.996824 (±0.003480)	1.0 (±0.002331)	0.993647
4	0.997691 (±0.000667)	0.999392 (±0.001018)	0.998784 (±0.001333)	1.0 (±0.000993)	0.990268
5	0.99694 (±0.000515)	0.9981111 (±0.001054)	0.997482 (±0.001380)	0.998741 (±0.000948)	0.993703

续表

序号	三边形平均分类准确率	四边形平均分类准确率	五边形平均分类准确率	六边形平均分类准确率	基于属性重要度约简的平均分类准确率
6	0.998761 (±0.002011)	0.997756 (±0.001627)	0.997552 (±0.001341)	0.998776 (±0.001843)	0.991432
7	0.997890 (±0.001919)	0.999587 (±0.001012)	0.998761 (±0.001357)	1.0 (±0.001647)	0.993804
8	0.9994 (±0.001307)	0.998312 (±0.001033)	0.997470 (±0.001386)	0.998734 (±0.001307)	0.996203
9	0.997274 (±0.000657)	0.9996 (±0.000620)	0.9994 (±0.000657)	1.0 (±0.000657)	0.995198
10	0.997694 (±0.001471)	0.998952 (±0.001672)	0.998742 (±0.001378)	1.0 (±0.001719)	0.991195

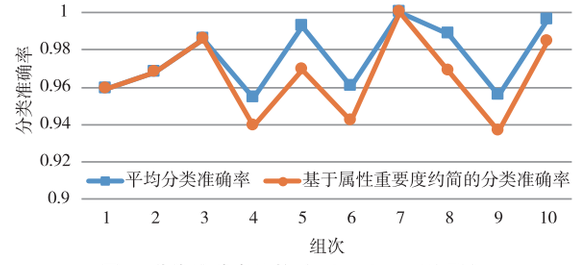


图7 分类准确率比较（Solar Flare 三边形）

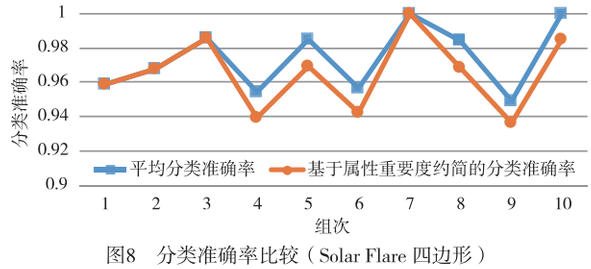


图8 分类准确率比较（Solar Flare 四边形）

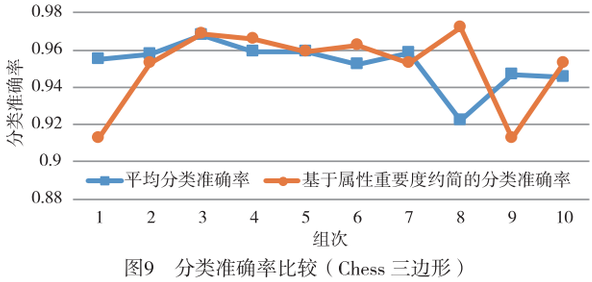


图9 分类准确率比较（Chess 三边形）

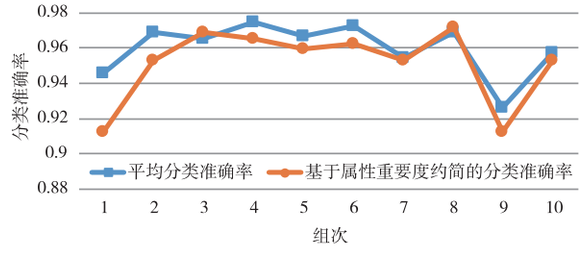


图10 分类准确率比较（Chess 四边形）

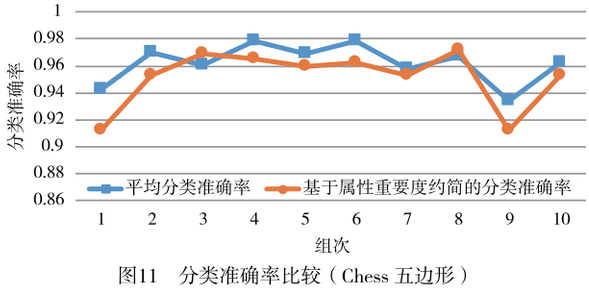


图11 分类准确率比较（Chess 五边形）

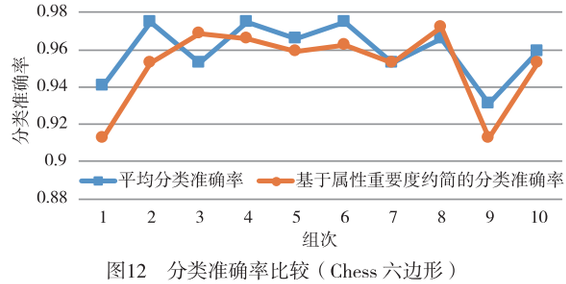


图12 分类准确率比较（Chess 六边形）

从表7和表8中，我们可以发现离重心距离最近的属性约简分类准确率均大于基于属性重要度约简的分类准确率.

由图13～图18可以看出离重心距离最近的属性约简的分类准确率均大于基于属性重要度约简的分类准确率，且离重心距离最近的属性约简的分类准确率的变化趋势比基于属性重要度约简的分类准确率更稳定.

以上实验结果显示，绝大多数情况下，离重心距离最近的属性约简的分类准确率大于等于其他离重心距离较远的属性约简的分类准确率；离重心距离最近的属性约简的分类准确率大于等于基于属性重要度约简的分类准确率. 也就是说，离重心最近的属性约简的平均分类准确率比一般情况下的属性约简（包括基于属性重要度的属性约简）的分类准确率有较大的优势，属性约简重心可以作为一种属性约简选优的方法.

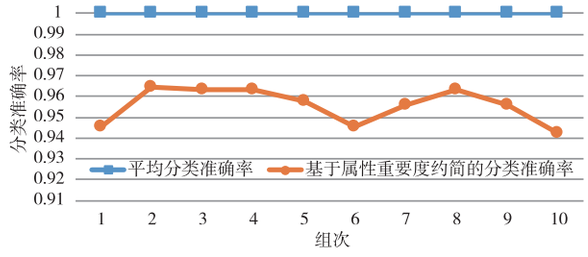


图13　分类准确率比较（Mushroom 三边形）

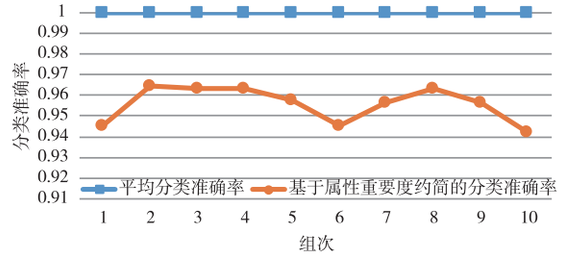


图14　分类准确率比较（Mushroom 四边形）

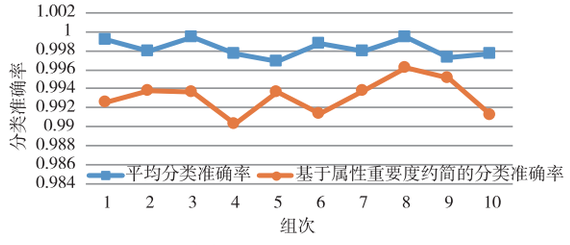


图15　分类准确率比较（Kddcup99 三边形）

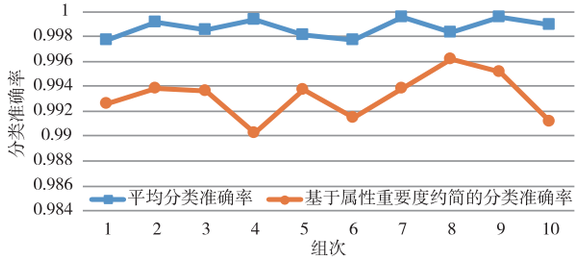


图16　分类准确率比较（Kddcup99 四边形）

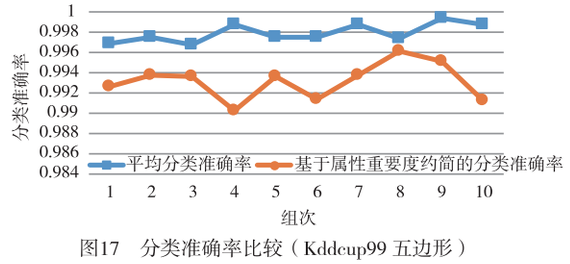


图17　分类准确率比较（Kddcup99 五边形）

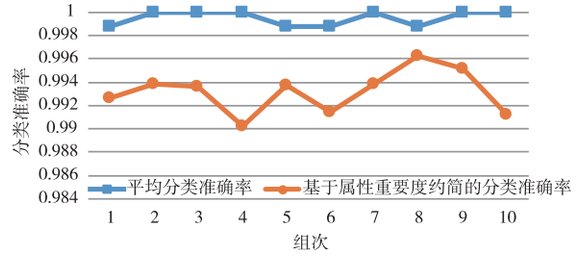


图18　分类准确率比较（Kddcup99 六边形）

离重心最近的属性约简的分类准确率比其它属性约简的分类准确率高,是因为相对于其它属性约简,离重心最近的属性约简与测试集相应属性子集之间的概念漂移要小(可用质概念漂移和量概念漂移等不确定性指标进行测量);不同数据集的训练结果对测试集进行分类时的分类准确率不同,也是因为概念漂移不同.

6　结论与展望

概念漂移探测是数据流挖掘的重要研究方向. 用概念漂移的指标和属性约简信息损失的度量指标讨论了同种类型的属性约简之间的区别与联系,提出了属性约简重心的概念,提出了一个属性约简选优的猜想:与其他属性约简相比(包括基于属性重要性的属性约简),大多数情况下离重心最近的属性约简具有较高的分类准确率. 实验结果验证了这个猜想的有效性. 属性约简的重心符合中国古典哲学“中庸”的思想. 因此,属性约简的重心可作为一种属性约简(或特征选择)的选优方法.

进一步研究为,属性约简重心的性质和进一步应用,使用重心、质概念漂移、量概念漂移等指标量化数据之间的差别与联系,并与其他概念漂移探测方法进行比较.

参考文献

[1] Pawlak Z. Rough Sets-Theoretical Aspect of Reasoning About Data[M]. Dordrecht:Kluwer Academic Publishers,1991.

[2] Ziarko W, Variable precision rough set model[J]. Journal of Computer and System Sciences,1993,46(1):39 – 59.

[3] Zhu W. Wang F. Reduction and axiomization of covering generalized rough sets[J]. Information Sciences,2003,152(1):217 – 230.

[4] Lin T Y. Neighborhood Systems:A Qualitative Theory for Fuzzy and Rough Sets[M]. Duke University, Durham:Advances in Machine Intelligence and Soft Computing, Department of Electrical Engineering,1997. 132 – 155.

[5] Zhu P, Hu Q, Zuo W, et al. Multi-granularity distance metric learning via neighborhood granule margin maximization [J]. Information Sciences,2014,282(282):321 – 331.

[6] Qian Y H, Liang J Y, Yao Y Y, et al. MGRS:A multi-granulation rough set[J]. Information Sciences,2010,180(6):949 – 970.

[7] 史开泉,崔玉泉. S-粗集和它的一般结构[J]. 山东大学学报(理学版),2002,37(6):471 – 474.

Shi K Q, Cui Y Q. S-rough set and its general structures [J]. Journal of Shandong University (Nat Sci), 2002, 37 (6): 471 – 474. (in Chinese)

[8] 邓大勇,陈林. 并行约简与 F-粗糙集. 云模型与粒计算[M]. 北京:科学出版社,2012. 210 – 228.
Deng D Y, Chen L. Parallel Reducts and F-rough Sets. Cloud Model and Granular Computing[M]. Beijing: Science Press, 2012. 210 – 228. (in Chinese)

[9] 邓大勇,卢克文,苗夺谦,黄厚宽. 知识系统中全粒度粗糙集及概念漂移的研究[J]. 计算机学报,2016. 39:在线出版号 No. 177.
Deng Dayong, Lu kewen, Miao Duoqian, Huang Houkuan. Study on entire-granulation rough sets and concept drifting in a knowledge system[J]. Chinese Journal of Computers, 2016, 39: Online Publishing No. 177. (in Chinese)

[10] 苗夺谦,胡桂荣. 知识约简的一种启发式算法[J]. 计算机研究与发展,1999,36(6):681 – 684.
Miao D Q, Hu G R. A heuristic algorithm for reduction of knowledge[J]. Journal of Computer Research & Development, 1999, 36(6): 681 – 684. (in Chinese)

[11] Skowron A, Rauszer C. The Discernibility Matrices and Functions in Information Systems [M]. Netherlands Springer, 1992.

[12] Bazan G J. A comparison of dynamic non-dynamic rough set methods for extracting laws from decision tables[A]. Rough Sets in Knowledge Discovery 1: Methodology and Applications, Physica-Verlag[C]. Heidelberg, 1998. 321 – 365.

[13] Kryszkiewicz M. Comparative study of alternative types of knowledge reduction in inconsistent systems[J]. International Journal of Intelligent Systems, 2001, 16(1): 105 – 120.

[14] 邓大勇,黄厚宽,李向军. 不一致决策系统中约简之间的比较[J]. 电子学报,2007,35(2):252 – 255.
DENG D Y, HUANG H K, LI X J. Comparison of various types of reductions in inconsistent systems[J]. Acta Electronica Sinica, 2007, 35(2): 252 – 255. (in Chinese)

[15] 徐章艳,杨炳儒,宋威,等. 几种不同属性约简的比较研究[J]. 小型微型计算机系统,2008,29(5):848 – 853.
Xu Z Y, Yang B R, Song W, et al. Comparative research of different attribute reduction definitions[J]. Journal of Chinese Computer Systems, 2008, 29(5): 848 – 853. (in Chinese)

[16] 李磊军,米据生. 信息系统属性约简的比较研究[J]. 计算机科学与探索,2009,3(6):577 – 584.
Li L J, Mi J S. A comparative study of attribute reduction in information systems[J]. Journal of Frontiers of Computer Science and Technology, 2009, 3(6): 577 – 584. (in Chinese)

[17] 张东娜,刘博. 三种属性约简方法的比较研究[J]. 电脑知识与技术:学术交流,2008,1(2):613 – 614.
Zhang D N, Liu B. The comparison research of three reduction methods[J]. Computer Knowledge and Technology, 2008, 1(2): 613 – 614. (in Chinese)

[18] Wong S K M, Ziarko W. On optimal decision rules in decision tables[J]. Bulletin of the Polish Academy of Sciences Mathematics, 1985, 33(11 – 12): 693 – 696.

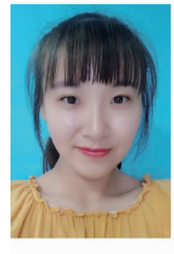
[19] 邓大勇,裴明华,黄厚宽. F-粗糙集方法对概念漂移的度量[J]. 浙江师范大学学报(自然科学版),2013,36(3):303 – 308.
Deng D Y, Pei M H, Huang H K. The F-rough sets approaches to the measures of concept drift[J]. Journal of Zhejiang Normal University (Nat Sci), 2013, 36(3): 303 – 308. (in Chinese)

[20] 邓大勇,薛欢欢,苗夺谦,等. 属性约简准则与约简信息损失的研究[J]. 电子学报,2017,45(2):401 – 407.
Deng D Y, Xue H H, Miao D Q, et al. Study on criteria of attribute reduction and information loss of attribute reduction[J]. Acta Electronica Sinica, 2017, 45(2): 401 – 407. (in Chinese)

[21] 邓大勇,卢克文,黄厚宽,等. 概念的属性约简及异构数据概念漂移探测[J]. 电子学报,2018,46(5):1234 – 1239.
Deng D Y, Lu K W, Huang H K, et al. Attribute reduction for concepts and concept drifting detection in heterogeneous data[J]. Acta Electronica Sinica, 2018, 46(5): 1234 – 1239. (in Chinese)

作者简介

邓大勇 男,1968 年出生,副教授,博士,现为浙江师范大学行知学院教师,主要研究方向为粗糙集、粒计算、数据挖掘等.
E-mail: dayongd@163.com



葛雅雯 女,1992 年生,硕士研究生. 主要研究方向为粗糙集、数据挖掘.
E-mial: 541865527@qq.com

黄厚宽 男,1940 年生,教授,博士生导师. 主要方向为计算智能、数据挖掘和粗糙集等.
E-mial: hkhuang@bjtu.edu.cn