doi:10.3772/j.issn.1002-0470.2018.01.006

基于增量学习和 Lasso 融合的数据可视化模式识别方法①

梁怀新② 郝连旺 宋佳霖 郑存芳 洪文学③

(燕山大学生物医学工程研究所 秦皇岛 066004)

摘 要 提出了一种基于增量学习和最小绝对值收缩和选择算子(Lasso)特征选择融合的数据可视化模式识别方法。该方法首先对归一化数据进行一级 Lasso 筛选特征降维,之后对连续数据进行基于 Gini 指数的粒化,再送入增量模式学习系统进行增量学习,针对维数大量升高的情况进行 Lasso 二级特征筛选生成一致模式决策表,生成属性偏序结构图可视化规则发现。数据采用来自 UCI 的 5 个数据库,并与分类器 KNN,SVM,Adaboost,Random Forest 进行分类准确度比较,实验表明,基于该算法的分类精度普遍高于其他分类器水平,且属性偏序结构图可视化层次清晰鲜明。通过增量学习实验设计,得到了准确率、图结构更新和不同比例增量数据的动态关系,其中 Pima Indians Diabetes 数据学习达到 40% 时准确率(77.66%)超过 Adaboost(75.32%)、SVM (77.27%)、1NN (59.74%)、3NN(75.97%)算法。结果表明该算法进行数据的可视化和模式识别是行之有效的。

关键词 增量学习, 最小绝对值收缩和选择算子(Lasso), 属性偏序结构图, 可视化, 模式识别, 粒化

0 引 言

人类生活各个领域都在每时每刻产生着多样、变化着的数据,数据规模越来越大,因此寻找一种优秀的数据可视化的方法尤为重要。根据人类的渐进式认知原理,在较大数据背景下一次性获得知识的完备模式是很困难的,增量学习(incremental learning)的出现使得机器具有了自学习能力,有助于不断进行数据学习,目前已经有很多算法与增量学习进行了结合[1-5]。形式概念分析(formal concept analysis,FCA)理论[6]是数据可视化的有效工具,自 1982 年德国 Wille 教授提出之后被广泛用于众多领域,如知识发现[7,8]、机器学习[9]、软件工程[10]、信息检索[11]等。它能反映概念间存在的泛化与例化关系,是概念的内涵和外延的统一体现。以概念为层次结构关系的概念格是形式概念分析中的核心数据结构,针对概念格的生成算法和对其结构的更新和约简,近年来一些学者进行了相关研究[12-16],针对概念格中交叉线和格结构、层次随着概念数量增多而变得复杂、层次不清等问题,洪文学教授提出了一种结合数学偏序理论的数据可视化的方法,即可以表示事物之间普遍性和特异性的属性偏序结构图,并在多个领域取得了应用[17-20]。但目前属性偏序结构图的生成一般是基于批量式构建[21],即一次性获得完备的概念前提下进行构建。基于此,本文提出一种基于增量学习和最小绝对值收缩和选择算子(least absolute shrinkage and selection operator, Lasso)(亦称“套索”)特征筛选融合的多维数据处理可视化模式识别方法,使得系统具有自学习能力。

然而系统中增量学习和粒化规则的引入会造成数据粒结构相对松散,使得数据维数升高,产生数据内部结构中对于模式分类价值相对较低的特征元

① 国家自然科学基金(61273019,81373767,61501397,61201111)和河北省自然科学基金(F2016203443)资助项目。

② 男,1992 年生,硕士生;研究方向:可视化模式识别,多维数据增量学习与特征融合,中医工程学;E-mail: tanner_k@163.com

③ 通讯作者,E-mail: hongwx@ysu.edu.cn

(收稿日期:2017-07-10)

素,不利于直接成图用于知识发现。为了得到高使用价值的特征数据,提高数据可视化效果,减少噪声数据对模式分类的准确率影响,引入 Lasso 特征筛选方法进行两级特征提取,能保留可用于规则提取的有效特征元素,提高数据模式分类规则提取的可视化效果。

此外,本文同时提出了基于基尼(Gini)粒化方法和基于 Gini 指数和覆盖对象(combination of Gini and objects, CGAO)的行列优化准则实现数据聚类。最后生成属性偏序结构图将分类规则可视化。为了验证方法的适用客观性,通过与 5 个分类器(1NN, 3NN,SVM,Random Forest,Adaboost)以及 5 个数据集进行了分类准确率对比。结果表明,本文方法不仅在准确率上达到了主流分类器水平,且规则可视化清晰明了。增量模式学习有助于挖掘完备模式临界提高数据使用效率,是一种数据可视化模式识别的有效工具。

1 概念介绍

1.1 形式概念分析相关性质

定义 1 形式概念分析反映了数据的泛化和特化的二元关系,形式背景是其具体的表现形式,$K=\{P,M,G\}$ 三元组表示一个形式背景,P 表示对象集合,M 为属性集合,二元关系表示为 $G \subseteq P \times M$。若存在对象子集 $X \subseteq P$, 属性子集 $Y \subseteq M$, 则有

$$f(X) = \{y \in M \mid \forall x \in X, xGy\} \quad (1)$$

$$g(Y) = \{x \in P \mid \forall y \in Y, xGy\} \quad (2)$$

其中,$f(X)$ 表示 X 对应属性集合, $g(Y)$ 表示 Y 的对象集合。则二元组(X,Y)就是形式背景上的一个概念,其中,X 称为概念的外延,Y 称为概念的内涵。

1.2 集合覆盖

覆盖理论是一种表示集合之间关联关系的方法之一,被定义于粗糙集和粒计算相关领域,它能通过不同的覆盖关系准确表达集合之间的交集或并集结果,可解释性强,本文引入覆盖理论说明在增量学习过程中,集合间的覆盖结果用于为指导执行不同增量学习的操作提供严谨的理论依据。为便于后文增量学习更新数据说明,完备的覆盖定义参照文献[22]。

1.3 决策表

定义 2(决策系统、决策信息表) 四元组 $S=<U,V,A,f>$ 表示的信息系统中,U 是对象集合,$A=C \cup D$ 是条件属性集合与决策属性集合之并集。$V=\cup V_a$,V_a 表示属性 a 的值域。$f: U \times A \rightarrow V$ 表示信息函数, $\forall a \in A, x \in U$, 有 $f(x,a) \in V_a$。用五元组(U,C,A,D,f)表示决策信息表,见表 1。在决策信息表中,称具有相同条件和决策属性的两个对象为相同模式,具有同样模式的数量作为模式的度来衡量其覆盖能力,例如在表 1 中,对象 $x1$ 与对象 $x5$ 属于相同模式,则保留 $x1$,模式度为 2。

表 1 决策信息表

U	C			D
	$c1$	$c2$	$c3$	
$x1$	1	0	1	$d1$
$x2$	1	0	1	$d2$
$x3$	1	1	0	$d2$
$x4$	0	0	1	$d1$
$x5$	1	1	1	$d1$

定义 3(决策模式信息表) 定义六元组(U,C,I',D,K',De)为决策模式信息表,新增了 I' 和 K' 两种映射关系,其中前者表示 U 与 C 之间的映射关系,后者为 U 与 D 之间的映射关系,De 是模式的度。表 2 为将表 1 转换后的决策模式信息表。因对象 $x5$ 与 $x1$ 属性相同,合并属性度 De。为了得到一致决策模式信息表,需要将条件属性相同但决策属性差异的对象合并,规定中只保留属性度较大的对象。如 $x1$ 与 $x2$ 的 C 相同,D 不同,因此在一致决策模式信息表中,保留 $x1$,删除 $x2$,记 $x1$ 的属性度为 3。

表 2 决策模式信息表

U	C			D	De
	$c1$	$c2$	$c3$		
$x1$	1	0	1	$d1$	2
$x2$	1	0	1	$d2$	1
$x3$	1	1	0	$d2$	1
$x4$	0	0	1	$d1$	1

2 增量学习算法

将增量学习算法引入多维数据的可视化和模式识别过程中,使得机器学习具有自学习能力,并根据对象具有的模式实现概念调整,完整的增量学习数据模式识别过程见图 1。

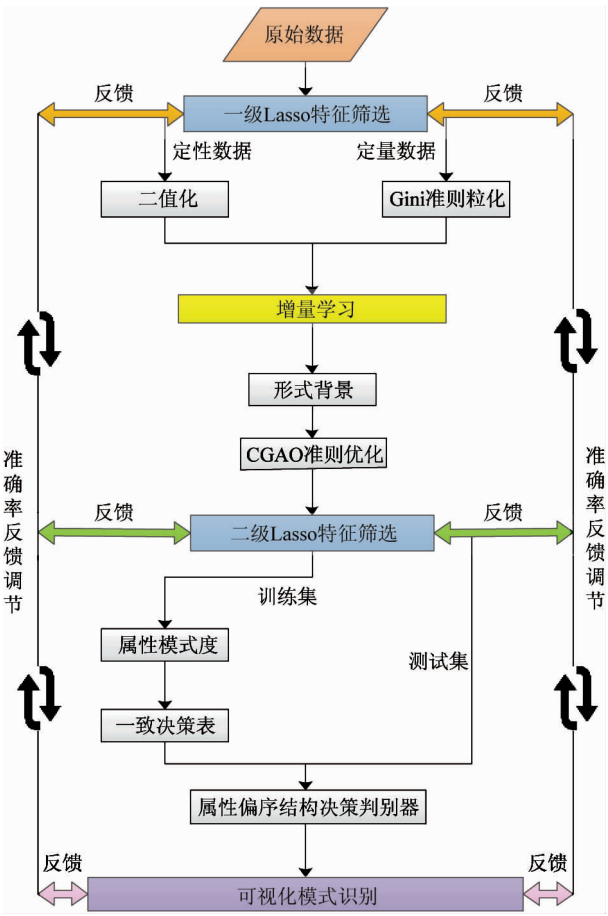


图 1 增量学习混合数据模式识别说明图

在形式背景 $K=\{P,M,G\}$ 中,假设 $\forall p_1 \in P$,对象 p_1 对应的属性集合 M_1,特征模式的学习基本思想就是通过不同的特征集合描述表征不同的对象。根据模式之间求交集得到覆盖结果进行形式背景的完备操作。依据现有的属性偏序结构图的优化过程[23],需要一次性获得完备的形式背景才可以进行优化,本文通过建立模式库、属性库等特殊数据集合,可以实现在增量学习的同时,保证存在模式的最简性。此过程总结见图 2。

2.1 初始模式的增量学习

初始的形式背景 $K=\{P,M,G\}$ 为空,即论域为 $\varnothing$,当存在新增对象 $X*$,设存在新增概念为 $(X*, f(X*))$,其中,根据定义 1,$f(X*)$ 表示新增的属性集合。此时,$f(X*)\cap M=\varnothing$ 一定成立,因此可以省略覆盖交集运算直接添加到形式背景中。这里,将属性集合 $f(X*)$ 中的属性按原始顺序保存到属性库 L 中作为下一次学习的覆盖判定集合 $L=\{l_1,l_2,\cdots,l_n\}$,n 为属性个数。自动生成由一个概念组成的形式背景保存在新形式背景 $K*=\{X*, f(X*),G\}$ 中。

2.2 增量属性学习

大规模的增量学习可以认为是很多单次增量学习的叠加,这里以每次学习一个概念为例进行说明。

原始形式背景 $K=\{P, M, G\}$,假设 $(X*, f(X*))$ 仍为新增的概念,将属性集合 $f(X*)$ 与属性库集合 L 做覆盖运算 $f(X*)\cap L$,根据覆盖结果进行下一步操作。为说明覆盖情况,设定标志集合 $ind=\{ind_1,ind_2,\cdots,ind_n\}$ (n 为属性库中属性个数),其中,$ind_i(i=1,2,\cdots,n)$ 表示新增单个属性与属性集合 L 的覆盖结果,单次比较中某属性已经存在则置 ind_i 为 1,否则 ind_i 为 0。

若 ind 为全零集,则判定该属性为新增属性,$f(X*)$ 与 L 之间是互斥覆盖关系,固然此模式也一定是新增模式,需更新模式库 H,将新增属性默认追加到属性库 L 末尾形成 $L*=(L\cup f_{add}(X*))$,其中,$f_{add}(X*)$ 表示属性集合 $f(X*)$ 中新增的属性。同时,生成单行形式背景添加到原形式背景中,即做更新 $K*=\{P\cup X*, M\cup f_{add}(X*), G\}$。

若 ind 为非全零集,即新增属性集合与原集合之间可能存在:子域覆盖关系、全覆盖关系、互不包含覆盖关系,超覆盖关系。根据不同覆盖关系对属性库和模式库做相关操作:

对于属性库 L 更新有以下几种情况:(1)若为全覆盖关系,说明新增属性集合 $f(X*)=L$,则属性库 L 保持不变,不做新增;(2)若为子域覆盖关系,属性库已覆盖全部新增对象,保持属性库 L 不变;(3)若为互不包含覆盖关系,则只更新新增属性 $L\cup f_{add}(X*)$;(4)若为超覆盖,则属性集合全部覆盖原属性库且有新属性加入,此时将新属性 $f_{add}(X*)$ 追加属性库 L 末尾,保留原有属性集合不变。

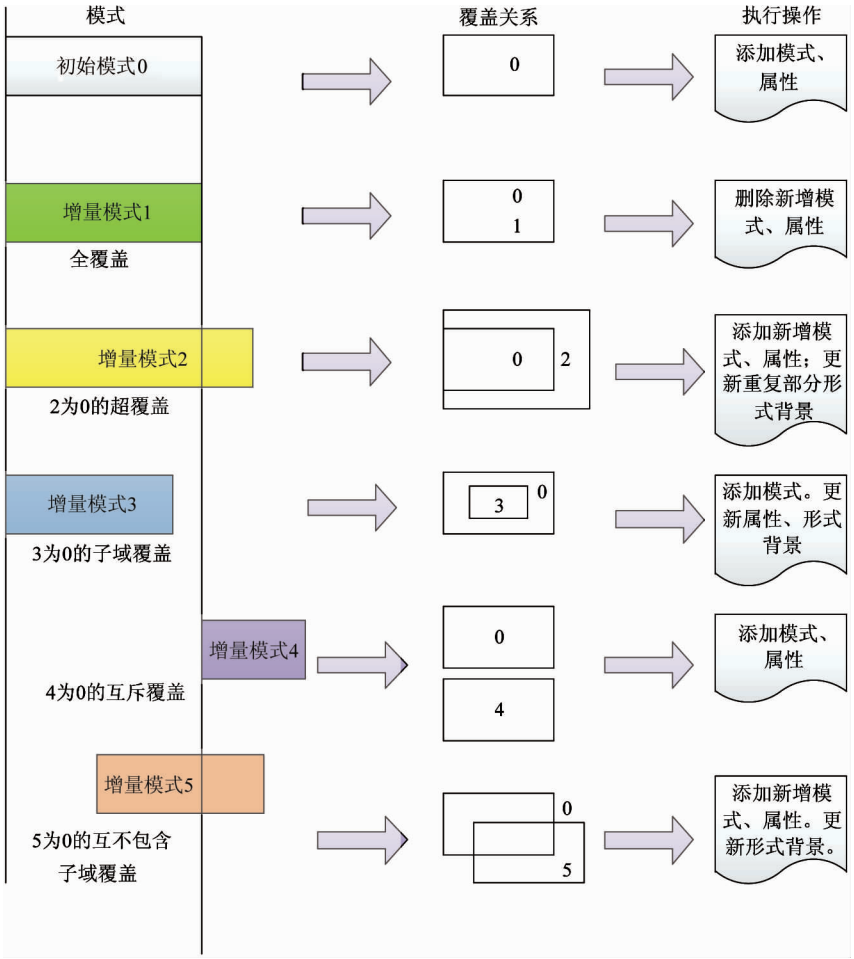


图 2　不同覆盖关系对应操作图

2.3　相同模式处理

当数据规模较大，概念的外延很多，内涵数量有限时，更容易出现重复的模式，根据定义重复模式即为既存模式库中模式的全覆盖关系。因此为了得到约简的模式，减少后续优化操作和比较，提高生成形式背景的时间效率，建立模式库 H 保存最简模式，H 中每个模式按照增量学习顺序添加。

当有新的概念$(X*, f(X*))$加入到形式背景中形成的一行概念的二值背景 F 时，设已存在形式背景为 $K_{old} = \{K_1, K_2, \cdots, K_i, \cdots, K_k\}$（$k$ 为已形成子背景的个数），$Class_i$ 表示概念对应标签，进行模式检测：

令 $ind' = \{ind_1', ind_2', ind_3', \cdots, ind_k'\} = \{\{F \cap k_1\}, \{F \cap k_2\}, \{F \cap k_3\}, \cdots, \{F \cap k_k\}\}$，若存在某两个有序的形式背景和标签不同，将此模式移到模式库 H 中追加到末尾 $H* = (H \cup (M \cup f_{\text{add}}(X*))$，否则，删除相同冗余模式。

3　Lasso

在很多情况下，并不是维数越多分类效果越好，噪声和偏差会提高维数并给结果带来误差，因此在多维数据中选择有价值、区分能力强的特征尤为重要，Lasso 算法有效起到了降维的作用[24]，其通过构造惩罚函数，限定各自变量绝对值之和小于特定值，进而实现对变量系数的压缩，当某些自变量回归系数为零时，非零系数特征便被筛选出来。Lasso 算法具有子集收缩的特点，是一种有偏估计方法。

最小角回归算法（least angle regression，LARS）的出现提高了 Lasso 算法的计算效率[25]，是一个计算残差逐渐减小的过程，其基本思想为：通过计算残差作为特征选择的参数指标，只有当某一变量的残

差关联系数和当前相同时,该变量才会被选入到回归路径中,LARS 算法的基本过程为:

设数据(X,Y), $X=(x_1,x_2,\cdots,x_j,\cdots,x_p)^{\mathrm{T}}$, $Y=(y_1,y_2,\cdots,y_i,\cdots,_n)^{\mathrm{T}}$是回归量,$x_j=(x_{1j},\ x_{2j},\cdots,x_{nj})$为预测变量属性,将数据进行标准化、中心化:

$$\sum_{i=1}^{n} y_i = 0,\ \sum_{i=1}^{n} x_{ij} = 0,\ \sum_{i=1}^{n} x_{ij^2} = 1,\quad j = 1,2,\cdots,p \tag{3}$$

计算当前残差:

$$r = y - \hat{y} \tag{4}$$

置回归系数估计$\beta=(\beta_1,\beta_2,\cdots,\beta_p)$为 0,首先找到某变量$x_j$和类标签相关系数最高,将路径设定为沿着$x_j$与$r$内积方向调整系数$\beta_j$,另一变量$x_k$加入回归模型的条件是其与当前残差的相关性与上一变量的残差相关系数相同。此时β_j已确定,此后,沿着变量x_j和x_k角分线方向继续调整系数选择下一特征变量。

4 属性偏序结构图生成

对于连续数据,需要将数据根据有效方式粒化,即将连续的属性特征值映射到不同的离散区间将其划分为一个粒,以进行之后的数据挖掘过程。

形式背景的优化对于属性偏序图的生成十分重要,现采用的形式背景的优化方法未考虑类别信息,即完全根据对象包含的属性模式进行聚类。然而类别信息对于模式发现起着至关重要的作用,是之后进行规则提取、分类器设计等数据挖掘的重要基本信息。基于此,本文提出形式背景的生成方法,分为划分和优化两个步骤。具体而言,提出了基于有监督的新的连续数据粒化方法,同时提出基于基尼指数和属性覆盖对象综合指标作为行列优化中重要属性选择的指标,不仅保留了基于属性覆盖最大原理的优化方法,同时引入基于类别纯度表征的基尼指数以突出类别信息,有利于发现类独有属性和类独有复合属性以便实现规则提取、知识发现等数据挖掘方法。根据前文定义,其中基尼指数越小保证了类别信息越"纯",属性覆盖对象越大说明该属性普遍性越大,算法需要的信息正是信息最纯的最大量概念信息。

4.1 数据粒化

基尼指数是一种基于不纯度分裂方法的参数指标,通常用于选择分裂的属性,进而进行决策树的构建。典型的运用基尼指数的决策树算法有:分类回归树算法(CART 算法)、SLIQ 算法、SPRINT 算法等。假设集合S中包含s个数据,m个不同类别,将m个不同类定义为$C_i(i=1,2,\cdots,m)$。根据属性值将集合S划分为m个子集,集合S_i包含的样本数目为s_i,则集合S的 Gini 指数为

$$Gini(S) = 1 - \sum_{i=1}^{m} p_i^2 \tag{5}$$

其中,p_i表示的是某一样本属于类别C_i的概率。在构建决策树选择分裂属性时,假设根据某个属性将集合S划分为N个子集$S_j(j=1,2,\cdots,N)$,则分裂后的$Gini_{\mathrm{split}}$指标表示为

$$Gini_{\mathrm{split}}(S) = \sum_{j=1}^{N} \frac{s_j}{s} Gini(S_j) \tag{6}$$

其中,s_j为属于某一个类别的样本数,s为所有类别数目。

4.2 连续数据粒化

连续数据粒化方法说明如下:

输入:连续数据集合S、类别标签$Class$

输出:形式背景K

(1) 对每列属性值求两两数据中点P为潜在分割点。

(2) 计算每一个中点P的$Gini_{\mathrm{split}}$。

(3) 选择$\min(Gini_{\mathrm{split}}(P))$对应的分割点,当前行数为$w_i$,若$w_i+1$行到末尾行$Class$一致,停止本列计算,否则重复步骤(2)和(3)。

(4) 得到全部分割区间,生成形式背景K。

4.3 形式背景优化方法

为了实现数据聚类生成属性偏序结构图,进行基于行列变换的形式背景优化,本文提出新的参数指标,即以基尼指数和覆盖对象(CGAO)为依据,定义形式背景K中每一列属性值$a_{i1},a_{i2},\cdots,a_{ij},\cdots,a_{im}$的 CGAO:

$$CGAO = Gini_{\mathrm{split}(m_i)} + \frac{1}{\sum m_i} \tag{7}$$

其中,$m_i \in M(i=1,2,\cdots,n)$表示某个属性。

将最大值对应的列与第一列交换,从 1 到q_1行

使得 $\sum_{i=1}^{q_1} a_{i1} = q_1$ 且 $\sum_{i=q_1+1}^{m} a_{i1} = 0$。再对第 q_1+1 行到第 m 行进行计算 CGAO 值。将最小值对应列与第 2 列交换，对 q_1 到 q_2 行使得 $\sum_{i=q_1+1}^{q_2} a_{i2} = q_2 - q_1$，且 $\sum_{i=q_2+1}^{m} a_{i2} = 0$。直到覆盖全部对象，得到属性偏序结构图的第一层。对每一个子形式背景 $K_1, K_2, \cdots, K_k$（k 为子背景个数）都进行如上操作，直到全部形式背景 K 分割完成。优化之后的形式背景中，属性出现的次序是根据其在相应子域内覆盖对象最大和类别信息最纯为依据，经过 CGAO 指标优化的对象簇集聚类效果也会优于之前的形式背景。

4.4　属性偏序结构图的生成

经过粒化和行列优化形成了形式背景，也具备了生成属性偏序结构图的条件，之后采用文献[18]的生成方法，可以通过人工或计算机编程方式生成新的属性偏序结构图，具体生成过程不再赘述。

5　实验过程

为对应实验过程，本文采用来自 UCI 数据库中的 Pima Indians Diabetes 数据进行数据处理过程说明，该数据来自国家糖尿病与消化和肾脏疾病研究所，由 Research Center 教授等提供。其中涉及对象 768 例，属性 8 个，为实现模式识别过程，随机抽取 80% 数据作为训练集，在剩余 20% 数据中进行测试。

步骤 1　一级 Lasso 特征筛选：将原始连续数据进行归一化去除量纲影响后输入 R 语言的 lars 工具中，生成回归路径图见图 3。从图中可以看到按照顺序筛选的特征为 $X2 > X6 > X1 > X7 > X8 > X3 > X5 > X4$。综合考虑分类准确率（后续给出特征数优选比较实验）和可视化效果这里选取前 4 个属性 $X2$、$X6$、$X1$、$X7$ 为一级 Lasso 属性。

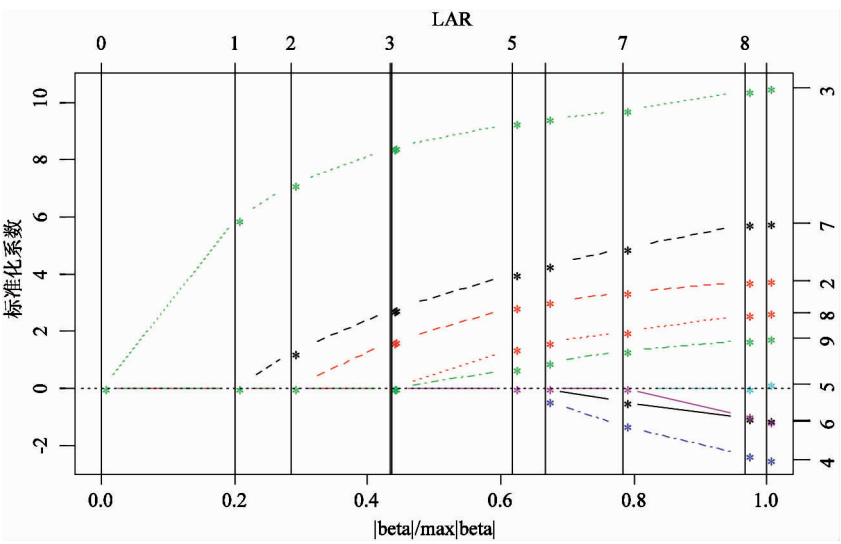


图 3　一级 Lasso 筛选回归路线图

步骤 2　数据粒化：针对选择出的连续属性进行基于 Gini 指数的粒化，使用字母表示顺序属性，数字表示映射顺序区间，得到的分割点和原数据映射到对应的区间表见表 3 和表 4。

表 3　各列属性分割点

	分割点							
	1	2	3	4	5	6	7	8
属性 1	0.641	0.776	0.837	0.867	0.947	0.977	0.987	0.992
属性 2	0.150	0.205	0.242	0.257	0.264	0.282	0.293	/
属性 3	0.033	0.068	/	/	/	/	/	/
属性 4	0.003	0.003	0.004	0.007	0.009	0.012	0.012	/

表 4　部分属性映射区间表

	*A*2	*B*2	*C*1	*D*3	…
1	1	1	1	1	
2	0	0	1	0	…
3	0	0	0	1	
4	0	0	1	0	
		…			

步骤 3　增量学习和形式背景优化:将步骤 2 中的数据输入到增量学习系统中进行数据的顺序学习,之后对生成的形式背景进行行列优化处理,得到部分优化的形式背景。

步骤 4　二级 Lasso 特征筛选:经过粒化和增量学习过程的数据由 4 维变为了 18 维,直接生成属性偏序结构图不利于数据的可视化规则发现,针对此种情况,本文引入二次 Lasso 特征筛选,实现数据的二次降维,图 4 为增量形式背景的回归路线图,这里选择前五个特征 *X*5、*X*6、*X*3、*X*8、*X*19 作为 Lasso 二级筛选特征。

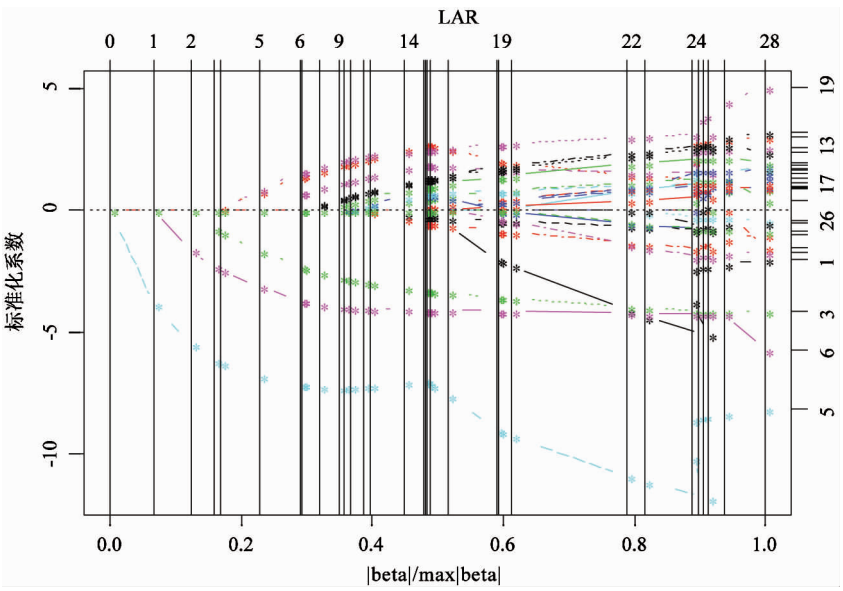


图 4　二级 Lasso 特征筛选回归路径图

步骤 5　生成一致决策模式信息表:经过降维后的决策信息表中存在大量冗余模式,为了得到最简模式的一致决策模式信息表,只保留不一致决策模式中属性度大的模式,并取多模式属性度之和作为最后的属性度。考虑篇幅影响,部分一致决策信息表见表 5。

表 5　一致决策模式信息表

	*X*7 (0.003,0.004]	*X*2 (0,0.641]	*X*6 (0.15,0.205]	*X*7 (0,0.003]	*X*2 (0.977,0.987]	标签
1	0	0	0	0	0	1
2	0	0	0	0	1	1
3	0	0	0	1	0	1
4	0	0	0	1	1	1
5	0	0	1	0	0	1
…			…			
22	1	1	1	0	1	-1

步骤 6　生成属性偏序结构图:从一致决策模式信息表整合出形式背景,生成属性偏序结构图见图 5。从图中可以看出:条件属性总共分为 5 层,对应二级 Lasso 筛选的五个特征,在第二层中包含属性 a3,a2,a5,a1,a4,其中属性 a3 对应的簇集最大,分支最多,体现了该属性的普遍性。图中用方框标注出的对象分支属于一类,可以清晰地看出:相同类别的对象实例被分在同一簇集中,可见属性偏序结构图实现了数据的良好聚类效果。

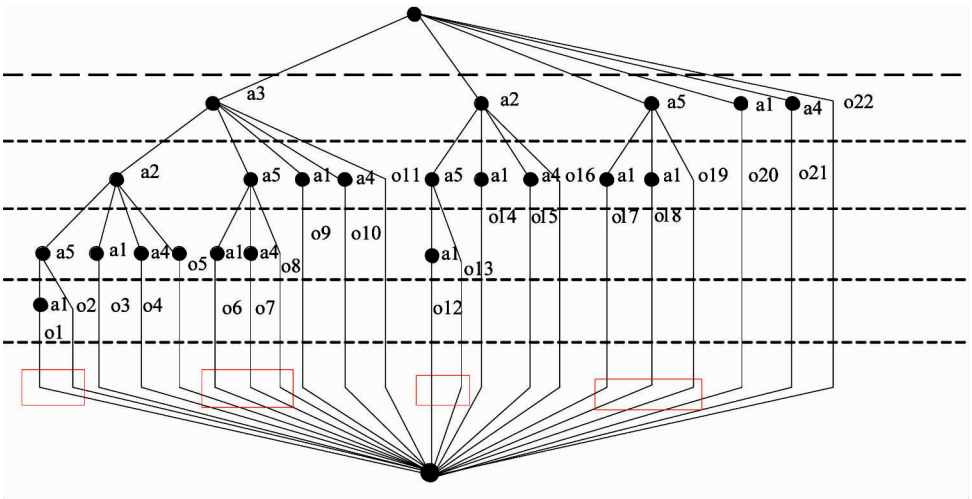


图 5　属性偏序结构图

步骤 7　规则提取:利用属性偏序结构图实现了规则的可视化,同时可以进行规则的约简,比如对象 o1 对应的规则为:a3,a2,a5,a1→第一类,对象 o2 支路对应规则为:a3,a2,a5→第一类,因二者在相同的以 a5 为顶点的小簇集中且为同类,因此将其规则合并约简为 a3,a2,a5→第一类。采用同样方法对其余支路规则进行约简,经过对测试集数据测试得到平均分类准确率为78.57%。

其中,在步骤 1、3 的多属性 Lasso 分级特征筛选过程中,为了说明特征数量选择的客观性和组合的有效性,本文对不同的 Lasso 特征筛选数目组合进行了研究,结果见表 6。

表 6　Pima Indians Diabetes 数据集不同特征筛选数目组合比较

一级\二级	3	4	5	6
3	75.32	77.27	77.27	77.27
4	75.32	77.27	**78.57**	78.57
5	74.68	77.27	77.27	77.92

通过表 4 可知,当一级 Lasso 筛选属性为前 4 个,二级 Lasso 特征筛选数目为前 5 个时,在测试集上分类准确率相对较高,特征组合记为 4#5,与组合 4#6 准确率相同,综合考量可视化效果,因此选择组合 4#5 作为最终的分类可视化的筛选特征组合。

6　模式识别

为了验证本文方法的有效性,采用来自标准数据库 UCI 中的共 5 个数据集(Liver-disorders、Heart、Breast cancer、Ionosphere、Pima Indians Diabetes)进行测试。按照随机抽取 80% 数据作为训练集,剩余 20% 数据作为测试集实验 5 次求取准确率,实验中为保证分类准确率可比性,在同一数据集上的不同 Lasso 特征筛选组合中保持训练集测试集不变。表 7 ~ 表 10 表示其余 4 个数据集的不同 Lasso 特征筛选数目组合的分类准确率比较。

通过分析表 5 ~ 表 8 可以得出不同级特征筛选数目组合对模式识别分类准确率的影响,从中可以在有限范围内选择最佳的Lasso特征筛选数目组合

表 7　Liver-disorders 数据特征筛选数目组合准确率

一级\二级	3	4	5	6
3	65.58	65.58	61.96	62.32
4	66.30	67.39	65.58	61.59
5	65.58	64.86	65.94	63.41

表 8　Heart 数据特征筛选数目组合准确率

一级\二级	3	4	5	6
3	81.48	81.48	81.48	77.78
4	74.07	87.04	87.04	87.04
5	74.07	87.04	87.04	85.19

表 9　Breast cancer 数据特征筛选数目组合准确率

一级\二级	3	4	5	6
3	94.89	92.7	91.24	86.86
4	96.52	94.89	95.62	91.97
5	92.7	89.78	85.4	81.75

表 10　Ionosphere 数据特征筛选数目组合准确率

一级\二级	3	4	5	6
3	81.69	87.32	87.32	87.32
4	76.06	78.87	80.28	84.51
5	76.06	78.87	80.28	85.92

已达到分类相对最优,同时,通过设计反馈控制系统(见图 1),根据不同的特征组合进而对一级、二级 Lasso 特征数目进行反馈调节,在保证分类准确性的前提下达到可视化结构最优,得到有效的分类器算法。

同时,为了说明本文基于增量学习和 Lasso 特征筛选融合的算法可靠性,与不同分类器如 KNN(K=1、3)、SVM、Adaboost、Random Forest(RF)5 个主流分类器进行准确率比较,其中,均采用经过本文增量学习和 Lasso 特征选择后生成的形式背景为数据源。比较结果见表 11。

表 11　不同分类器准确率比较

	1NN(%)	3NN(%)	SVM(%)	Adaboost(%)	RF(%)	本文(%)
Liver-disorders	57.97	57.97	60.87	60.87	60.87	67.39
Heart	85.19	85.19	87.04	87.04	79.63	87.04
Breast cancer	88.32	89.78	92.00	92.70	92.70	96.52
Ionosphere	87.32	87.32	88.73	88.73	87.32	87.32
Pima Indians Diabetes	59.74	75.97	77.27	75.32	75.32	78.57

经过分析表 11 可以得出,本文算法分类准确率普遍高于其他主流分类器水平,只有在 IONO 数据集中低于 SVM 和 Adaboost 分类算法。结合表 7 - 10 中二级 Lasso 特征筛选数目可以得到在维数相对较低的情况下,本文方法仍然可以保持较高的准确率,提高了可视化的简洁性。

7　增量学习验证

为了更加突出增量学习的特征,将 Pima Indians Diabetes 数据集合打乱并按照顺序学习方式按比例提取数据,以寻找模式完备的临界,提高数据利用效率,将数据按照 10% ~90% 增量顺序输入增量学习系统中作为训练集,用剩下的数据进行测试,保持最佳 Lasso 特征筛选数目组合 4#5,得到的不同比例分类准确率见表 12。同时根据本文算法生成属性偏序结构图见图 6 中(a)至(f),分别代表增量数据比例为 10%、20%、30%、40%、50% 和 60%。

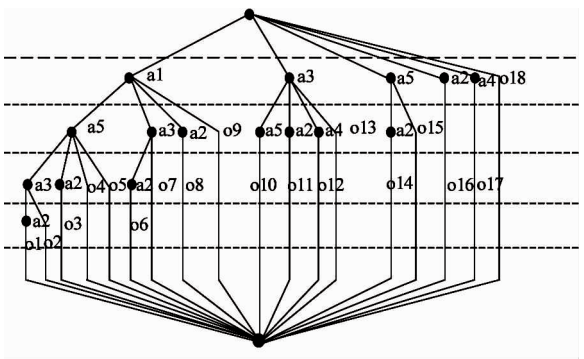
通过分析表 12 可以发现,随着增量学习数据量逐渐增大,分类准确率基本持逐渐上升状态,但在个别数据增量比例情况下准确率有所下降,可能是与难分样本分布以及测试集合基数有关。

由图 6 可以看出:图(b)相比图(a)有新的规则分支 o19 加入,图(c)中新加入规则分支为 o20,图(e)的最右也有新分支 o21 加入,图(f)和图(e)规则分支、结构完全相同,最后在图(g)中加入最后

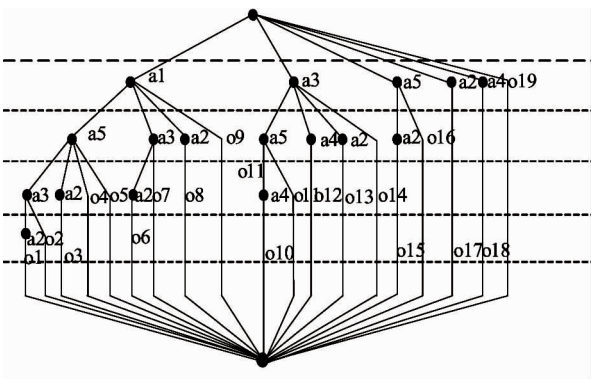
一条规则分支 o22。可见,随着增量数据比例逐渐增加,模式逐渐趋于完备,在前 10% ~60% 的数据比例中,增量学习每次均有不同的模式加入,其中,数据增量为 40% 和 50% 情况下模式相同,没有变化,且自数据量达到 60% 起,数据集模式完备保持不变,且结构不再进行更新,因此 70% 到 90% 比例增量数据属性偏序结构图在此没有给出。

表 12　顺序增量比例数据分类准确率

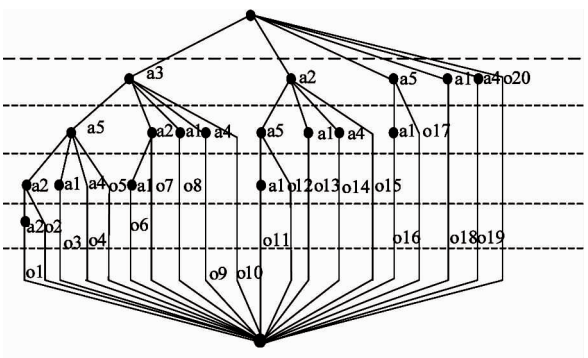
	10%	20%	30%	40%	50%	60%	70%	80%	90%
准确率(%)	72.25	74.8	75.65	77.66	77.34	76.62	77.06	77.92	80.52



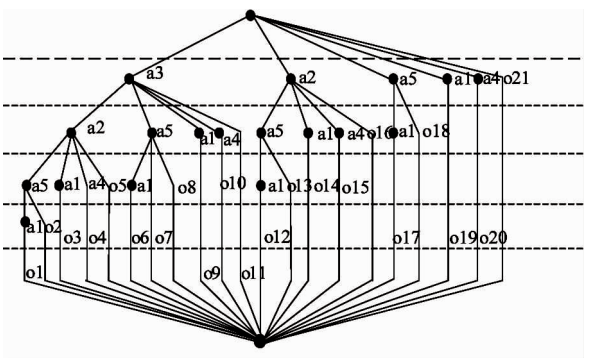
(a) 增量数据比例 10%



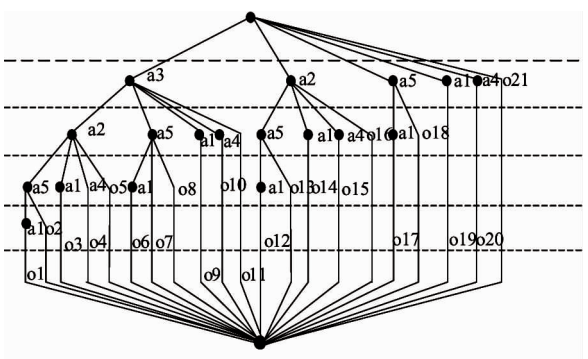
(b) 增量数据比例 20%



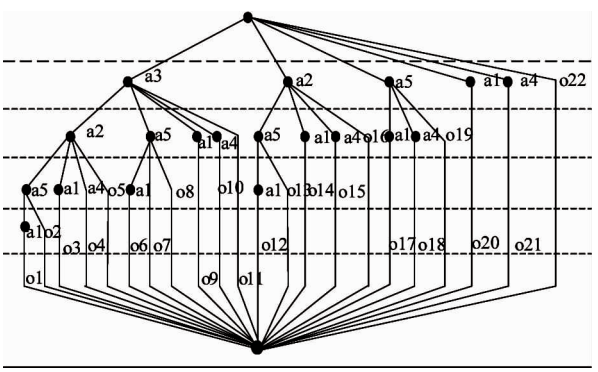
(c) 增量数据比例 30%



(d) 增量数据比例 40%



(e) 增量数据比例 50%



(f) 增量数据比例 60%

图 6　顺序增量比例模式属性偏序结构图

结合表 11、表 12 和图 6 进行分析可得,按照本文增量顺序学习方式进行模式挖掘过程中,当训练集数据量达到总量的 30% 时已经超过 Adaboost 算法(75.32%)采用 80% 训练及数据量的分类水平。当数据量达到 40% 时,超过 SVM 算法(77.27%)以及 1NN(59.74%)和 3NN(75.97%)算法的分类准确率。

为了进一步验证本文算法的可行性,针对 5 个数据集按照打乱顺序后的顺序增量比例学习方式进行学习并进行准确率验证,并与其它文献[26-30]中针对同数据集的分类准确率进行比较,比较结果见表 13。

表 13　与不同文献准确率对比

数据集	最优增量学习比例(%)	本文准确率(%)	文献	对比准确率(%)
Heart	50	84.44	[25]	83.20
			[26]	75.10
Liver-disorders	90	88.57	[27]	70.14
Pima Indians Diabetes	90	80.52	[28]	80.21
			[29]	92.00
Breast cancer	80	99.27	[18]	94.16
Ionosphere	40	91.94	[30]	90 .57
			[25]	94.00

通过对表 13 的分析可得,本文算法中有 3 个数据集在不同增量学习比例中的最优准确率高于其它文献,在 Pima Indians Diabetes、Ionosphere 数据集上准确率欠优,但在 Ionosphere 数据集中,顺序学习只采用了 40% 数据量进行训练,数据使用效率较高且准确率已经超过文献[26]。同时考虑到样本分布对数据分类的影响,本文还将对挖掘数据分布、指定学习顺序、寻找最优 Lasso 特征筛选组合方面进行进一步研究。

8　结　论

本文提出了一种基于增量学习和 Lasso 特征多级筛选结合的处理数据的可视化模式识别方法,通过对原始连续数据进行一级 Lasso 特征筛选实现初步降维,粒化,对于增量学习后数据维数提高的情况,引入二级 Lasso 特征筛选实现降维,筛选作用相对较大的特征。其中,通过设计特征数目识别率反馈环节得到最优的分级特征筛选数目组合,并基于此进行后续工作。属性偏序结构图自动聚类效果明显,使得决策规则可视化,结构层次清晰,结合 Lasso 特征筛选方法可用于高维数据动态结构可视化。最后,为了验证增量学习对挖掘完备模式临界的作用,设计顺序增量比例学习实验,得到了不同比例增量数据的学习与模式更新的关系,本文方法可以清晰看出准确率和结构更新的动态联系,有效地挖掘了完备模式的临界数据比例,有利于在保证准确率的前提下提高数据的使用效率。此外,本研究在寻找最佳 Lasso 特征筛选数目组合、探寻样本分布对准确率的影响、更改学习顺序以提高准确率等方面还有待更进一步的研究以取得更优的分类结果。下一步研究将采用更多数据集,扩大 Lasso 特征筛选数目组合研究分类准确率变化,同时寻找顺序增量学习分类最优的学习方式以提高模式识别效果。

参考文献

[1] 邱天宇,申富饶,赵金熙. 自组织增量学习神经网络综述[J]. 软件学报,2016,27(09):2230-2247

[2] 徐敏政,何宗宜,刘亚虹,等. 双向渐进式概念格生成算法[J]. 小型微型计算机系统,2014,35(01):172-176

[3] 王爱平,万国伟,程志全,等. 支持在线学习的增量式极端随机森林分类器[J]. 软件学报,2011,22(09):2059-2074

[4] 茅嫣蕾,魏赟,贾佳. 一种基于 KKT 条件和壳向量的 SVM 增量学习算法[J]. 电子科技,2016,29(02):38-

40 +44

[5] 曾舒如. 基于多模态增量学习模型的目标物体检测方法研究:[硕士学位论文]. 南昌:南昌大学,2016. 8-14

[6] Wille R. Restructuring lattice theory: an approach based on hierarchies of concepts. In: Proceedings of the International Conference on Formal Concept Analysis, 2009. 314-339

[7] Jonas P, Dmitry II, Ser-gei OK, et al. Formal concept analysis in knowledge processing: A survey on applications [J]. *Expert Systems with Applications*, 2013, 40(16): 6538-6560

[8] Simon A, Constantinos O. Discovering knowledge in data using formal concept analysis [J]. *International Journal of Distributed Systems and Technologies* (*IJDST*), 2013, 4(2): 31-50

[9] Nida M, Hela K, Mondher M. Parallel learning and classification for rules based on formal concepts [J]. *Procedia Computer Science*, 2014, 35: 358-367

[10] Zheng P, Da R, Dan M, et al. Formal concept analysis based on the topology for attributes of a formal context [J]. *Information Sciences*, 2013, 236(1): 66-82

[11] Li J H, Mei C L, Lv Y J. Incomplete decision contexts: Approximate concept construction, rule acquisition and knowledge reduction [J]. *International Journal of Approximate Reasoning*, 2013, 54 (1): 149-165

[12] 姜琴. 基于多属性消减的概念格构造算法研究:[硕士学位论文]. 郑州:郑州大学,2016. 37-53

[13] 刘贝玲,齐华,沈富强,等. 基于概念格的关联规则挖掘算法改进[J]. 地理信息世界,2016,23(01):64-70

[14] 康向平,苗夺谦. 一种基于概念格的集值信息系统中的知识获取方法[J]. 智能系统学报,2016,11(03): 287-293

[15] 吴杰,梁妍,马垣. 基于内涵亏值的概念格渐进式构建[J]. 计算机应用,2017,37(01):222-227

[16] 黄艳,任苗苗,魏玲. 区间值决策形式背景的属性值向量约简[J]. 计算机科学,2012,39(01):193-197

[17] 洪文学,李少雄,张涛,等. 大数据偏序结构生成原理[J]. 燕山大学学报, 2014,(5):388-393

[18] 郑存芳,洪文学,李少雄,等. 数据偏序结构关系中的知识发现可视化方法[J]. 智能系统学报,2016,11(04):475-480

[19] 郑存芳,洪文学,王金甲. 基于偏序结构图的乳腺癌诊断规则提取方法[J]. 计算机工程与设计,2016,37(06):1599-1603

[20] 张仲鹏. 基于属性偏序原理的脑功能近红外光谱分析方法研究:[博士学位论文]. 秦皇岛:燕山大学电气工程学院,2016. 12-30

[21] 李少雄,闫恩亮,宋佳霖,等. 偏序结构图的一种计算机生成算法[J]. 燕山大学学报,2014,(05):403-408

[22] Hong W X, Luan J M, Li S X. The complete definitions of covering and properties description based on partial ordered theory [C]. In: Proceedings of the 10th International Conference on Innovative Computing, Information and Control, Dalian, China, 2015. 1055-1060

[23] 樊凤杰,洪文学. 基于属性偏序结构图表示原理的中药方剂配伍规律研究[J]. 生物医学工程学杂志,2013,(04):719-723

[24] Robert T. Regression shrinkage and selection via the lasso [J]. *Journal of the Royal Statistical Society. Series B* (*Methodological*). 1996, 58(1): 267-288

[25] Bradley E, Trevor H, Iain J, et al. Least angle regression [J]. *The Annals of Statistics*, 2004, 32(2): 407-499

[26] 高建国,崔业勤. 基于信息熵理论的连续属性离散化方法[J]. 微电子学与计算机,2011,28(7):187-189

[27] 郭躬德,黄杰,陈黎飞. 基于 KNN 模型的增量学习算法[J]. 模式识别与人工智能,2010,23(5):701-707

[28] Sug H. Generating CART decision trees for health data sets [C]. In: Proceedings of the Recent Advances in Electrical and Computer Engineering, Korea, 2012. 72-76

[29] 刘晋胜. 基于熵降噪优化相似性距离的 KNN 算法研究[J]. 计算机应用与软件,2015(9):254-256

[30] 谢宏,程浩忠,牛东晓,等. 基于信息熵的粗糙集连续属性离散化算法[J]. 计算机学报,2005,28(9):1570-1574

The data visualization and pattern recognition method based on the fusion of incremental learning and Lasso

Liang Huaixin, Hao Lianwang, Song Jialin, Zheng Cunfang, Hong Wenxue
(Institute of Biomedical Engineering, Yanshan University, Qinhuangdao 066004)

Abstract

A data visualization and pattern recognition method based on the fusion of incremental learning and least absolute shrinkage and selection operator (Lasso) feature selection is proposed. The method selects the features of the normalized data by the first-order Lasso to deduce the dimensions. When the granular computing of the continuous data is completed by using the Gini index, the data is then sent to the incremental learning system. The second-order Lasso feature selection is used to deal with the increasing dimensions, and the attribute partial order structure diagram is generated to visualize the rules concerned. Five databases from UCI and five classifiers (1NN, 3NN, SVM, Adaboost, and Random Forest) are selected to make comparison with the precision result of the proposed method. The result shows that the precision of the method is higher than that of other algorithms generally, and the attribute partial order structure diagram has clear layers and structures. The incremental learning experiment is designed to testify the relationships of the precision and update of the structures of the diagram with different incremental learning proportions. When the proportion reaches 40%, the precision of the Pima Indians Diabetes database (77.66%) can exceed over the Adaboost (75.32%), SVM (77.27%), 1NN (59.74%) and 3NN (75.97%) algorithm with learning process of all of data. The result shows that the method proposed is an effective tool for the visualization and pattern recognition.

Key words: incremental learning, least absolute shrinkage and selection operator (Lasso), attribute partial order structure diagram, visualization, pattern recognition, granulation