广西科学院学报
Journal of Guangxi Academy of Sciences

2007,23(3):147~149,159
Vol. 23,No. 3 August 2007

粗糙集理论在区域降水预报中的应用研究*
Application of Rough Set Theory to Regional Mean Rainfall Forecast

孔庆燕[1],金 龙[2]
KONG Qing-yan[1],JIN Long[2]

(1.桂林航天工业高等专科学校计算机系,广西桂林 541004;2.广西气象减灾研究所,广西南宁 530022)
(1. Department of Computer,Guilin College of Aerospace Technology, Guilin,Guangxi,541004, China; 2. Guangxi Institute of Meteorological Disaster Mitigation, Nanning,Guangxi,530022, China)

摘要:用粗糙集属性约简方法选择出比较合理的预报因子组合,建立广西东南部区域日平均降水量的预报方程(预报时效为24h),并进行2006年5~6月的前汛期逐日业务预报应用试验.结果表明,采用属性约简方法建立的预报方程比传统的逐步回归预报方程有更高的预报精度,具有较好的业务应用前景.

关键词:粗糙集 属性约简 逐步回归 区域平均降水量
中图法分类号:O159;P426.6 **文献标识码**:A **文章编号**:1002-7378(2007)03-0147-03

Abstract: Using the attribute reduction method to choose reasonable combination of factors, a regional mean rainfall forecast equation is established for southeast Guangxi. An operational forecasting application test is conducted in May and June before the flood season in 2006. The results show that the new forecast equation produces higher forecast accuracy and better application prospect than traditional stepwise regression equation.

Key words: rough set, attribute reduction, stepwise regression, regional mean rainfall

粗糙集理论是一种处理模糊和不确定知识的数学工具[1],其主要思想就是在保持分类能力不变的前提下,通过知识约简,导出问题的决策或分类规则.该理论与其它处理不确定性问题理论的最显著的区别是,它无需提供问题所需处理的数据集合之外的任何先验信息,因此对问题的不确定的描述和处理是比较客观的.目前,粗糙集理论已被成功地应用于机器学习、决策分析、过程控制、模式识别与数据挖掘等领域[2~4].

在气象预报问题中,对一般的降水或气温等各种气象要素的预报都能较为方便地找到很多与之有很好相关关系的预报因子.而如何从这些众多相关预报因子中筛选出好的预报因子组合对预报量作未来状况的预测,一般来说逐步回归方法是一种便捷有效的方法.然而,逐步回归方法挑选出来的预报因子间容易存在复共线性关系,往往会产生预报方程拟合精度高而预报能力差的情况[5].为此,本文尝试采用粗糙集属性约简方法来选择预报因子,选出比较合理的预报因子组合建立回归预报模型.

收稿日期:2007-03-12
修回日期:2007-06-18

作者简介:孔庆燕(1979-),女,硕士研究生,主要从事数学建模研究.

*国家科技部社会公益性研究专项项目(2004DIB3J122),广西科学研究与技术发展计划项目(桂攻关:0592005-2A)资助.

1 粗糙集理论的基本概念

为了方便叙述,我们首先简要介绍一些粗糙集理论中的有关基本概念[6].

定义1 四元组 $S=(U,A,V,f)$ 是一个信息系统,其中 $U=\{x_1,x_2,\cdots,x_n\}$ 为论域;A 为属性的非空有限集合;$V=\bigcup_{a\in A}V_a$,V_a 是属性 a 的值域;f:$U\times A\to V$ 是一个信息函数,即 $\forall a\in A, x\in U$, $f(x,a)\in V_a$.

如果属性集 A 可以分为条件属性集 C 和决策属性集 D,即 $A=C\cup D, C\cap D=\varnothing$,则该信息系统称为决策系统或决策表,其中 D 中一般只含有一个

属性.

定义2 在信息系统 S 中，对于任何一个属性集合 $B \subseteq C$，不可分辨关系定义为

$$IND(B) = \{(x,y) \in U \times U : f(x,a) = f(y,a), \forall a \in B\}. \quad (1)$$

显然不可分辨关系就是 U 上的等价关系，为简便起见，在不产生混淆的情况下用 B 代替 $IND(B)$.

定义3 在决策系统 S 中，对于每个子集 $X \subseteq U$ 和属性子集 $B \subseteq C$，则 X 关于 B 的下近似和上近似定义为

$$\underline{B}X = \bigcup \{Y \in U/IND(B) : Y \subseteq U\},$$
$$\overline{B}X = \bigcup \{Y \in U/IND(B) : Y \cap X \neq \phi\}. \quad (2)$$

$\underline{B}X$ 表示根据关系 $IND(B)$ 判断肯定属于 X 的 U 中的集合；$\overline{B}X$ 表示根据关系 $IND(B)$ 判断可能属于 X 的 U 中的集合. 集合 $bn_R(X) = \overline{B}X - \underline{B}X$ 称为 X 的 R 边界域；集合 $pos_R(X) = \underline{B}X$ 称为 X 的 R 正域；集合 $neg_R(X) = U - \overline{B}X$ 称为 X 的 R 负域. 显然 $\overline{B}X = pos_R(X) \cup bn_R(X)$.

定义4 在决策系统 S 中，设 $R \subseteq C, Y \subseteq U$，则集合 Y 关于属性集 R 的近似精度为

$$\alpha_R(Y) = card(\underline{R}Y)/card(\overline{R}Y). \quad (3)$$

式中 $card(\#)$ 表示集合 # 所含元素的个数.

设 L 为由决策属性集 D 所决定 U 的划分 $\{Y_1, Y_2, \cdots, Y_k\}$，则划分 L 关于属性集 R 的近似分类精度和近似分类质量分别为

$$\alpha_R(L) = \sum_{i=1}^{k} card(\underline{R}Y_i)/card(\overline{R}Y_i),$$
$$\gamma_R(L) = \sum_{i=1}^{k} card(\underline{R}Y_i)/card(U). \quad (4)$$

近似分类的精度表示用 R 对对象分类时，可能的决策中正确决策的百分比；近似分类的质量表示用 R 对对象分类时，能确切地划入 L 类的对象的百分比.

定义5 设 $a \in R$，如果 $IND(R) = IND(R - \{a\})$，则称 a 为 R 中可约简的属性；否则称 a 为 R 中不可约简的属性. 如果每一个 $a \in R$ 都为 R 中不可约简的属性，则称 R 为独立的；否则称 R 为依赖的. 若 R 是独立的，且 $IND(R) = IND(C)$，则称 R 是 C 的一个约简. C 中所有不可约简的属性的集合称为 C 的核，记为 $core(C)$.

定理1 $core(C) = \bigcap red(C)$，其中 $red(C)$ 表示 C 的所有约简.

2 基于属性重要性的属性约简算法

2.1 属性重要性的定义

粗糙集理论的一个重要的思想是，信息是具有粒度的，根据某个等价关系可以把论域划分为正域、负域和边界域，因此可以用粗糙集的近似分类精度和质量来定义属性重要性[7].

设条件属性集 $C = \{c_1, c_2, \cdots, c_m\}$ 为有限集，决策属性集为 D，L 为由 D 所决定 U 的划分 $\{Y_1, Y_2, \cdots, Y_k\}$. 对每个条件属性 $c_i (i = 1, \cdots, m)$ 计算 $k+2$ 个参数，即 $\alpha_{c_i}(Y_j), j = 1, 2, \cdots, k$，以及 $\alpha_{c_i}(L)$ 和 $\gamma_{c_i}(L)$，其中 α_{c_i} 和 σ_{c_i} 分别为这 $k+2$ 个参数的均值和方差.

定义6 属性 $c_i (i = 1, \cdots, m)$ 的重要性为 $S_{c_i} = \alpha_{c_i}$.

定义7 当属性重要性相同时，使方差 σ_{c_i} 最小的属性 c_i 的重要性为优.

由上可知，对任一属性 c_i，当 $k+2$ 个参数均为非0时，表明该属性对划分 L 的各子集均有影响；当 $\alpha_{c_i} = \alpha_{c_j}$ 时，其方差能体现出各自的差别.

2.2 算法描述

本文基于属性重要性的属性约简算法，设计从条件属性集 C 的核开始，逐步增加不可缺少的属性，从而得到最小属性集 R.

找到条件属性集 C 的核 $core(C)$ 最简单的想法就是求出 C 的所有可能的属性约简，再求交集. 但这是一个NP-hard问题：首先，一个含有 m 个条件属性的训练集可能的属性组合有 $2^m - 1$ 种，这本身就是指数时间复杂度的问题；其次，对于给定的属性集，判断其是否为一个约简需要对样本做 $n(n-1)/2$ 次比较（设训练集中含有 n 个样本），它的计算复杂度为 $O(n^2)$. 由此可见，这种方法是行不通的. 文献[8]给出了一个利用改进的差别矩阵直接计算核的方法，其中，改进的差别矩阵 $M = \{m_{ij}\}$ 定义为

$$m_{ij} = \begin{cases} a \in C : f(x_i,a) \neq f(x_j,a), \text{当 } f(x_i,D) \neq f(x_j,D) \text{ 时}; \\ \phi, \text{其他情况时}. \end{cases} \quad (5)$$

当且仅当某个 m_{ij} 为单个属性时，该属性属于 $core(C)$. 容易验证，这种求核方法的计算复杂度为 $O(n^2m)$.

我们利用2.1中的计算单个属性重要性的方法和文献[8]中改进的差别矩阵直接计算核的方法，得到一个属性约简算法，其基本过程如下：

步骤1 对原始决策表中的连续属性进行离散化.

数据的离散化问题属于粗糙集理论的预处理

问题之一[9].为了简便起见,在本文中采用的是把条件属性的值域按等频率划分为几个离散空间,而决策属性的分割点由具体问题而定.

步骤2 删除决策表中重复的实例.

步骤3 利用改进的差别矩阵直接计算条件属性集 C 的核 $core(C)$.

步骤4 令 $R=core(C)$,计算 $\gamma_R(L)$,若 $\gamma_R(L)<\gamma_C(L)$,令 $A=C-R$,下转步骤5,否则下转步骤8.

步骤5 对每个属性 $a\in A$,计算属性重要性 S_a.

步骤6 选择 A 中使 S_a 最大的属性 a(出现多个属性重要性相同时,选择方差小的属性),令 $R=R\cup\{a\}$,同时在集合 A 中去掉属性 a.

步骤7 如果 $\gamma_R(L)<\gamma_C(L)$,上转步骤6,否则下转步骤8.

步骤8 输出 R,即 R 为条件属性 C 的一个相对约简.

3 应用实例

3.1 预报因子及预报对象

以广西东南部23个气象观测站2002年至2005年的5～6月的逐日平均降水量作为预报量,并以同期的中国气象局T213和日本数值天气预报模式的数值预报产品作为预报因子的基本资料.所选用的数值预报产品包括:T213模式各标准层17个常规气象要素及物理量要素场(100－120°E,15－30°N,1°*1°,共336个格点)和日本细网格模式降水预报场(100－120°E,15－30°N,1.25*1.25,共221个格点).通过对2002年至2005年的5～6月逐日数值预报产品场与预报对象进行场相关普查,将成片稳定(置信水平高于0.05)的高相关格点作为预报因子的选择区,在区内选2个相邻格点的最大平均值作为待选因子.并以达到或超过0.01置信度水平作为选择预报因子的标准,最终得到40个预报因子(39个T213模式预报场因子和1个日本细网格降水预报因子).各预报因子与预报量的相关系数在0.35～0.50之间.以下方法的讨论均基于这些预选的因子.

3.2 基于属性约简的预报方程

根据3.1中对预报因子选取结果可知,初步选定的40个预报因子中,有39个T213模式预报场因子和1个日本细网格模式降水预报因子.把每个预报因子按等频分为4类,预报量(日平均降水量)以10mm为分界点(24小时降水总量超过10mm为中雨以上),采用2.2节基于属性重要性的属性约简算法,最后选出的7个预报因子,依次为 $X_2,X_{22},X_{23},X_{30},X_{38},X_1,X_3$.利用这7个预报因子,剔除数值预报产品资料不齐的日期,以2002年至2005年5～6月资料为建模样本(217天),建立广西东南部区域日平均降水量的逐步回归预报方程为

$$Y=2.329046+0.085983X_1+0.083185X_2+0.029432X_3+0.003503X_{22}+0.117030X_{23}+0.670733X_{30}-0.154785X_{38}. \quad (6)$$

采用方程(6)对2006年5～6月广西东南部区域日平均降水量进行业务应用预报试验,剔除数值预报产品资料不齐,不能进行预报试验的日期外,共进行了54天的试预报运行.试验的预报平均绝对误差值为6.3929mm.

3.3 区域平均降水量的逐步回归预报方程

为了考察用粗糙集属性约简方法选取预报因子所建立的预报方程的预报效果,再进一步用建立方程(6)时相同的建模样本(217天)和资料,采用较易实现且预报效果较为客观的传统的逐步回归法建立回归预报方程.为了便于对比分析,新建立的回归方程因子数和方程(6)一样,也是7个。当 F 值取6.0时,入选回归因子有7个,依次是 $X_1,X_{34},X_{18},X_{38},X_2,X_{39},X_{22}$.所建立的传统逐步回归预报方程为

$$Y=4.469287+0.567104X_1+0.063359X_2+0.186434X_{18}+0.047998X_{22}+0.810476X_{34}+0.130244X_{38}-0.109549X_{39}. \quad (7)$$

采用与方程(6)相同的检验方法,用方程(7)对2006年5～6月广西东南部区域日平均降水量进行业务应用预报试验,同样进行了54天的业务预报试运行.方程(7)试验的预报平均绝对误差为7.8152mm.

3.4 两种预报方程的预报效果对比分析

由于预报方程(6)和(7)的建模样本长度相同,因子数相同,预报对象相同,试验预报的时段也相同,因此预报效果具有较大的可比性.从两个回归方程所进行的54天业务预报平均绝对误差对比来看,方程(6)的误差6.3929mm小于方程(7)的误差7.8152mm,采用粗糙集属性约简方法选取预报因子的方法比传统的回归方法预报绝对误差减小了1.4223 mm,预报精度提高了18.2%,预报效果比传统的逐步回归方法有了改进和提高.在基本相同的条件下所建立的两个回归预报方程预报效果不同,这可能是因为方程所选因子间的复相关系数不同,影

(下转第159页)

0.99948.

对于表2的数据，令 $y=\ln\lambda, x=\ln f$，则 y 与 x 成线性关系，从而(4)式问题可用最小二乘法来处理. 令 $y=bx+a$，则 $b=-0.97349$，为 -1 左右，从而验证 $\lambda \propto f^{-1}$ 的关系成立. 计算过程如下：

$$b=\frac{\overline{xy}-\bar{x}\cdot\bar{y}}{\overline{x^2}-(\bar{x})^2}=$$

$$[-3.24118-4.55599\times(-0.66688)]/(20.96545-20.7570)=-0.97349,$$

$$a=\bar{y}-b\bar{x}=-0.66688-(-0.97349\times 4.55599)=3.76833,$$

相关系数

$$r=\frac{\overline{xy}-\bar{x}\cdot\bar{y}}{\sqrt{(\overline{x^2}-\bar{x}^2)(\overline{y^2}-\bar{y}^2)}}=\frac{-3.24118-4.55599\times(-0.66688)}{\sqrt{(20.96545-20.757)(0.64238-0.44473)}}=-0.99963.$$

3 结束语

从上面分析可以看出，用最小二乘法线性拟合的结果与用 Origin 软件拟合的结果完全吻合，但是利用 Origin 软件进行拟合可以避免一系列的人为计算过程造成的误差，整个计算过程及作图完全由计算机进行，既简洁、快速、直观，而且更精确. 在教学过程中运用此方法，学生更容易接受.

参考文献：

[1] 刘东红. 弦振动驻波分析[J]. 大学物理实验，2002，15(1)：13-15.

[2] 王武廷. 对弦振动实验中振源的改进[J]. 大学物理，2004，23(5)：30-32.

[3] 郑连琴，溪海英. 定滑轮对弦振动实验的影响[J]. 物理实验，2001，21(5)：42.

[4] 李莞，冯正南，卢海燕. 弦振动实验装置的改进[J]. 大学物理实验，2004，17(4)：56-57.

[5] 徐志东，陈世涛. 大学物理实验[M]. 成都：西南交通大学出版社，1999：129-130.

（责任编辑：韦廷宗）

（上接第149页）

响了方程的预报效果.

进一步深入分析这两种预报结果的绝对误差，统计两种方法的预报绝对误差落在不同误差范围内的次数，预报绝对误差大于10mm的次数，粗糙集属性约简方法选取预报因子法的次数为11次，传统逐步回归法的次数为16次；而预报绝对误差小于5mm的次数，粗糙集属性约简方法选取预报因子法的次数为30次，传统逐步回归法的次数为18次. 表明方程(6)预报误差大的次数明显少于方程(7)，而预报误差小的次数又明显多于方程(7)，方程(6)可以为降水预报提供一种更为可靠的预报参考.

4 结束语

本文将粗糙集理论中基于属性重要性的属性约简方法用于区域日平均降水量预报，通过分析属性的重要性，剔除不必要的因素，在不改变决策结果的前提下，选出相对较好预报因子组合，建立回归预报方程，通过对2006年5～6月广西东南部区域日平均降水量的业务预报试运行，结果表明，利用粗糙集属性约简方法选择预报因子方法建立的预报方程比传统的逐步回归方法的预报误差小，是降水预报服务可靠的参考依据，具有较好的业务应用前景.

参考文献：

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

[2] CUI YUQUAN，SHI KAIQUAN. Function S-Rough sets and its applications [J]. Journal of Systems Engineering and Electronics，2006，17(2)：331-338.

[3] ZHOU LEI，SHU LAN. Rough set model based on new set pair analysis [J]. Fuzzy Systems and Mathematics，2006，20(4)：111-116.

[4] 阎维明，乔亚玲，何浩祥. 粗糙集理论在震害预测中的应用[J]. 自然灾害学报，2006，15(3)：147-151.

[5] 金龙. 神经网络气象预报建模理论方法与应用[M]. 北京：气象出版社，2004.

[6] 张文修，吴伟志，梁吉业，等. 粗糙集理论与方法[M]. 北京：科学出版社，2001.

[7] 瞿彬彬，卢炎生. 基于粗糙集的属性约简算法研究[J]. 华中科技大学学报：自然科学版，2005，33(8)：30-33.

[8] HU XIAOHUA，NICK CERCONE. Learning in relational databases：a rough set approach [J]. Computational intelligence，1995，11(2)：323-338.

[9] 侯利娟，王国胤，聂能，等. 粗糙集理论中的离散化问题[J]. 计算机科学，2000，27(12)：89-94.

（责任编辑：韦廷宗）