

基于加权变精度容差粗糙集模型的属性约简及应用研究

余侃侃 胡孔法 王 珍

(南京中医药大学信息技术学院 南京 210023)

摘 要 对变精度粗糙集模型进行了改进,结合相容关系,提出了加权变精度容差关系模型。在此基础上,将属性重要度和信息熵相结合作为启发式信息,研究了基于属性敏感度的约简算法,克服了传统属性约简算法中仅考虑单个属性分类能力的缺陷,保证了约简的完备性。最后,以海量中医方剂数据为例,说明该模型和算法在大数据分析中的有效性。

关键词 变精度粗糙集,大数据分析,相容关系,属性约简,中医方剂
中图分类号 TP391 **文献标识码** A

Research and Application on Reduction of Weighted Variable Precision Model Based on Tolerance Theory

SHE Kan-kan HU Kong-fa WANG Zhen

(School of Information Technology, Nanjing University of Traditional Chinese Medicine, Nanjing 210023, China)

Abstract Variable precision rough set was improved in this paper. Combined with tolerance relation theory, the weighted variable precision model based on tolerance theory was proposed. Furthermore, we suggested a new heuristic algorithm using the sensitivity of attributes, which is based on the significance of attributes and entropy, can overcome the shortcoming of classical model and ensure the completeness of reduction. As a result, traditional Chinese prescriptions were used to prove the effectiveness of model and algorithm mentioned in this paper.

Keywords Variable precision rough sets, Big data analysis, Tolerance theory, Attribute reduction, Chinese prescription

1 引言

粗糙集理论^[1]是波兰学者 Pawlak 于 1982 年提出的一种数据分析工具,主要应用于模式识别、机器学习和数据挖掘等领域。它作为一种处理不完整、不一致、不精确等各种不完善信息的有效工具,与证据理论、模糊理论等其他处理不确定性问题理论的方法相比,最大的优点在于不需要先验知识,且与模糊理论有很强的互补性^[2]。中医方剂数据浩瀚如烟海,截至晚清就有近 10 万首,本文将粗糙集模型及属性约简方法进行改进,用于海量中医方剂数据分析,主要基于以下考虑:一方面,方剂数据量非常庞大,源于几千年的文化传承,以人工记载为主,存在大量的数据缺失和模糊信息,数据噪声极大,是典型的不完备信息系统。因此,本文对 Ziarko 于 1993 年提出的变精度粗糙集模型^[3,4]进行改进,在知识表示系统和决策表中引入数据对象的权值函数和属性特性函数,使之可以表示数据的不同重要性和属性的不同特性,得到加权可变精度模型,相比 Ziarko 模型,该模型能更好地克服噪声数据的干扰。另一方面,考虑到中医辨证施治的特殊性,方剂数据库中大量“可意会不可言传”的数据,并且很多数据的差别

很细微,本文将加权可变精度模型与相容关系粗糙集模型相结合,提出了加权可变精度容差关系模型,并提出了该模型下基于属性敏感度的约简算法,用以分析中医方剂配伍关系。实验结果表明,该模型能处理含有噪音的不完备方剂信息系统,得出的方证对应关系能很好地体现中医方剂配伍规律的精髓,可用于指导临床用药。

2 基本概念

粗糙集理论作为一种数据分析处理理论,其主要思想是:将数据库中的属性分为条件属性和结论属性,根据不同的属性将数据库中的元组分成相应的子集,然后根据由条件属性划分的子集与由结论属性划分的子集之间的上下近似关系生成判定规则^[5,6]。

2.1 可变精度模型^[3,4]

对于信息系统 $K=(U, R)$,经典的 Pawlak 粗糙集模型定义

$$\text{隶属度为: } \mu_X^R(a) = \frac{|X \cap [a]_R|}{|[a]_R|}$$

这是根据可用知识 R ,对象 x 隶属于概念 X 的条件概率。在可变精度粗糙集模型中,设 β 为错误分类率且 $0 \leq \beta <$

本文受江苏省自然科学基金青年项目“多数据挖掘方法集成的方剂配伍规律挖掘模式设计与系统实现”(BK20140958),江苏省高校自然科学基金项目“多数据挖掘方法集成的方剂配伍规律挖掘模式设计与系统实现”(14KJB520032),南京中医药大学青年自然科学基金项目“基于数据挖掘的方剂配伍与临床疗效相关性研究”(13XZR36)资助。

余侃侃(1982—),女,博士生,讲师,主要研究方向为数据挖掘、智能信息处理, E-mail: she_kk@126.com; 胡孔法(1970—),男,教授,博士生导师,主要研究方向为物联网与大数据管理、数据仓库和商务智能; 王 珍(1968—),女,副教授,硕士生导师,主要研究方向为软件体系架构、大数据分析。

0.5, 定义为:

$$R(x) \subseteq_{\beta} X \text{ iff } \mu_X^R(x) \geq 1 - \beta$$

则 X 的 R 下近似和上近似分别定义为:

$$\begin{aligned} \underline{R}_{\beta}(x) &= \bigcup_{R(x) \subseteq_{\beta} X} R(x) \\ \bar{R}_{\beta}(x) &= \bigcup_{\mu_X^R(x) > \beta} R(x) \end{aligned}$$

由 $[\underline{R}_{\beta}(x), \bar{R}_{\beta}(x)]$ 得到的粗糙集模型称为可变精度模型, 它克服了 Pawlak 经典粗糙集对于分类的处理必须是完全正确或肯定的这一局限, 解决了某种程度上的“包含”与“属于”。

2.2 容差关系^[7]

给定信息系统 $I = \langle U, C \cup D, V, f \rangle$, 令 $A = C \cup D$, 对于具有空值的属性子集 $B \subseteq A$, 记空值为“*”, Kryszkiewicz 定义了如下容差关系 T :

定义 1 $T_B(x) = \{y \in U \mid \forall b \in B, b(x) = b(y) \text{ or } b(x) = * \text{ or } b(y) = *\}$

显然 T 是自反且对称的, 但不传递。

根据定义 1, 得到上下近似的定义。

定义 2 完备信息表 $I = \langle U, C \cup D, V, f \rangle$, 令 $A = C \cup D$, 对象集合 X 关于属性集 $B \subseteq A$ 的上近似 X_B^+ 和下近似 X_B^- 分别为

$$\begin{aligned} X_B^+ &= \{x \mid T_B(x) \subseteq X (x \in U)\} \\ X_B^- &= \{x \mid T_B(x) \cap X \neq \emptyset (x \in U)\} \end{aligned}$$

3 加权变精度容差关系模型的建立

Ziarko 的可变精度模型在一定程度上克服了噪音数据的干扰, 但是抗噪能力不强^[8]。而 Slowinski 的相似关系模型对分析描述模糊的数据或存在细微差别的数据有着重要影响。考虑到海量方剂信息系统中存在上述两种情况, 本文将 Ziarko 的可变精度模型进行改进, 与相容关系模型融合, 提出粗糙集理论的加权可变精度相容关系模型。

3.1 微差关系

定义 3 设信息系统 $I = \langle U, C \cup \{d\}, V, f \rangle$, 非空集合 $B \subseteq C$, 对于论域 U 中的任意两个元素 x_i, x_j , 有

$$(MD)_{ij} = \delta_B(x_i, x_j) = \frac{\sum_{b_k \in B} \text{dist}_{b_k}(x_i, x_j)}{\text{Card}(B)}$$

称 $\delta_B(x_i, x_j)$ 或 $(MD)_{ij}$ 为元素 x_i, x_j 关于 B 的特征距离, 其中 $\text{dist}_{b_k}(x_i, x_j)$ 为:

$$\text{dist}_{b_k}(x_i, x_j) = \sum_{d_r \in U/d} \left(\frac{\text{Card}([x_i]_{b_k} \cap d_r)}{\text{Card}([x_i]_{b_k})} - \frac{\text{Card}([x_j]_{b_k} \cap d_r)}{\text{Card}([x_j]_{b_k})} \right)^2$$

式中, 所有与 x 存在微差关系的元素集合构成相容类, 记为 $B(x)$ 或 $[x]_B$, 所有相容类构成 U 的覆盖, 记为 $\text{Cover}_B(U)$, 也称为知识 B 。同一相容类中的元素只有细微差别。

设 R 是由属性集 A 的子集诱导的论域 U 上的微差关系族, 称 R 为 U 上的一个知识库, 记为 $K = (U, R)$ 。

3.2 加权可变精度容差关系模型

粗糙集认为每个数据对象均具有同等重要程度, 但很多实际问题并非这样简单。数据对象的不同来源、领域的先验知识、决策者的主观判断或偏好等都可能对数据对象产生影

响。粗糙集通过分析属性间的依赖关系来评价属性的重要程度, 但在实际应用中属性往往还具有各自的特性, 比如, 不同属性可能具有不同的类型, 属性值的获取可能具有不同的置信度等, 粗糙集并没有提供对这种属性特性的描述手段。为解决这一问题, 我们在知识表示系统和决策表中引入数据对象的权值函数和属性特性函数, 使之可以表示数据的不同重要性和属性的不同特性。近似集合的定义采用可变精度粗糙集的原则, 以适应有噪声的数据, 这样就构成了一种新的扩展的粗糙集模型, 即加权可变精度模型。

称 $S = \langle U, A, V_a, f_a, V_w, f_w, V_c, f_c \rangle$ 为广义知识表示系统, A 为非空、有限的属性集合; 对每个属性 $a \in A$, V_a 为 a 的属性值的集合, 称为 a 的值域; $f_a: U \rightarrow V_a$ 为从论域 U 到属性 a 的值域的映射; V_w 为 U 中数据权值的值域; $f_w: U \rightarrow V_w$ 为权值函数, 使 $x \in U$ 在 V_w 中取权值 $f_w(x)$; V_c 为属性特性函数的值域; $f_c: A \times V_a \rightarrow V_c$ 为属性特性函数, 当 a 的值为 v_a 时, 取特性值 $f_c(a, v_a)$ 。

对于上述加权变精度模型, 我们定义论域中的任意对象 x 隶属于某概念 X 的隶属度为:

$$\mu_X^R(x) = \frac{\sum_{x \in I(x) \cap X} f_w(x)}{\sum_{x \in I(x)} f_w(x)}$$

设 β 为错误分类率且 $\beta \in [0, 0.5)$, 定义 $[x]_R \subseteq_{\beta} X$ iff $\mu_X^R(x) \geq 1 - \beta$, 则 X 的 R -下近似和 R -上近似分别定义为:

$$\begin{aligned} \underline{R}_{\beta, \delta_0}^w(X) &= \bigcup \{I(x) : \frac{\sum_{x \in I(x) \cap X} f_w(x)}{\sum_{x \in I(x)} f_w(x)} \geq \beta\} \\ \bar{R}_{\beta, \delta_0}^w(X) &= \bigcup \{I(x) : \frac{\sum_{x \in I(x) \cap X} f_w(x)}{\sum_{x \in I(x)} f_w(x)} > \beta\} \end{aligned}$$

4 加权变精度容差模型下基于属性敏感度的约简算法

基于属性依赖度的算法^[9]和基于信息熵^[10]的算法都只考虑了单个属性对分类能力的影响, 通常情况下, 一个属性本身不足以分类, 还需综合考虑其它属性^[11]。因此, 本文对上述算法进行改进, 将加权变精度模型与互信息概念相结合, 计算加权特征差和值差, 采用属性上下文敏感度作为属性约简的启发式信息。

定义 4 设决策表 $S = \langle U, A, V_a, f_a, V_w, f_w, V_c, f_c \rangle$, 论域中的任意两个元素 x_i, x_j 关于 C 的特征距离为 $(MD)_{ij}$, 则 x_i, x_j 的加权特征差为:

$$WFD_{ij} = \frac{1}{(MD)_{ij}}$$

设属性 c_k 对 U 的划分为 $U/c_k = \{C_1^{(k)}, C_2^{(k)}, \dots, C_n^{(k)}\}$, 决策属性集 D 对 U 的划分为 $U/D = \{D_1, D_2, \dots, D_m\}$, 则 D 相对于 c_k 的条件熵为:

$$H(D; c_k) = - \sum_{i=1}^n P(C_i^{(k)}) \sum_{j=1}^m P(D_j | C_i^{(k)}) \log P(D_j | C_i^{(k)})$$

因此, D 相对于对象 x_i (不妨设 $x_i \in C_i^{(k)}$) 的属性值 $c_k(x_i)$ 的条件熵为:

$$H(D | c_k(x_i)) = H(D | C_i^{(k)}) = - p(C_i^{(k)}) \sum_{j=1}^m p(D_j | C_i^{(k)}) \log p(D_j | C_i^{(k)})$$

定义 5 设决策表 $S = \langle U, A, V_a, f_a, V_w, f_w, V_c, f_c \rangle$, 对于条件属性 $c_k \in C$, 论域中的任意两个元素 x_i, x_j 的值差定义

为:

$$VD(c_k(x_i), c_k(x_j)) = \frac{|H(D|c_k(x_i)) - H(D|c_k(x_j))|}{H_{\max} - H_{\min}}$$

其中,

$$H_{\max} = \max(H(D|C_1^{(k)}), H(D|C_2^{(k)}), \dots, H(D|C_n^{(k)}))$$

$$H_{\min} = \min(H(D|C_1^{(k)}), H(D|C_2^{(k)}), \dots, H(D|C_n^{(k)}))$$

这样,我们可基于加权特征差和值差给出属性 c_k 的上下文敏感的价值度量:

$$CS(c_k) = \sum_{x_i \in D_l} \sum_{x_j \notin D_l} WFD_{ij} \cdot VD(c_k(x_i), c_k(x_j))$$

其中, D_l 是 x_i 所属的类。 $CS(c_k)$ 反映了 c_k 对对象差别的影响程度。 $CS(c_k)$ 越大, c_k 的重要性越高。因此 $CS(c_k)$ 可作为寻找最小属性约简的启发式信息。加入回溯思想,我们得到改进的基于上下文敏感度的属性约简算法。

算法 1 基于上下文敏感度的属性约简算法(带回溯思想)

输入:决策表 $S = \langle U, A, V_a, f_a, V_w, f_w, V_c, f_c \rangle, \delta_0, \beta$

输出:决策表的一个相对约简

步骤:

- (1) 计算决策中决策属性对条件属性的依赖度 $\gamma_C(D)$
- (2) 计算 C 相对于 D 的核 $Core_D(C)$
- (3) 令 $RED = C_0$, 重复
 - ① 对每个属性 $p \in C - RED$, 计算 p 的上下文敏感度 $CS(p)$
 - ② 选择使 $\gamma_p(D|RED)$ 最大的属性记作 p , 且 $RED \leftarrow RED \cup \{p\}$
 - ③ 若 $\gamma_{RED}(D) = \gamma_C(D)$, 则终止, 否则转①
- (4) 从 RED 的尾部开始, 从后往前对每个属性 a 进行判断是否可省。

若 $a \in Core_D(C)$, 则从 a 开始往前的属性都是不可省的, 算法结束, RED 就是所求结果;

否则, 若 $|POS_C(D)| = |POS_{RED}(D)|$, 则说明 a 是可省的, 因此把它从 RED 中删除。

由于计算 $CS(p)$ 和 $\gamma_C(D)$ 的时间复杂度都是 $O(|U|^2)$, 故该算法的时间复杂度是 $O(|C|^2|U|^2)$ 。

5 算法分析与应用

下面将本文模型和算法应用到中医方剂数据分析, 主要基于两点考虑: 1) 方剂数据库中数据量极其庞大, 且由于人工传承, 存在着大量信息缺失, 是典型的海量不完备信息系统, 可用于检验加权变精度容差关系模型及约简算法的有效性与完备性; 2) 由于每条方剂中涉及药物和症状很多, 给药症对应分析带来一定的困难, 因此可将药物与症状映射为粗集的属性, 对这两个属性集进行约简, 从约简后的属性中寻找药症对应关系, 探讨方证相应联系。

采用属性约简进行方证对应分析的具体步骤如下:

- (1) 按照症状检索出符合条件的方剂, 如主治中包含“恶寒、发热”症状的方剂;
- (2) 建立决策表, 令决策属性 $D = \{\text{恶寒、发热}\}$, 条件属性 $C = \{\text{主治“恶寒、发热”症状的方剂所有组成药物}\}$;
- (3) 建立加权变精度容差关系模型, 以属性敏感度作为启发式条件, 进行属性约简, 找出“恶寒、发热”方中所有药物的约简属性, 即可得到在保持方剂功效不变的条件下最小的药物组合。

在实验中, 我们以《中医方剂大辞典》^[12] 为数据来源, 共十万多条方剂数据。针对包含“恶寒、发热”症状的所有方剂

进行约简, 为说明本文算法的有效性, 特与基于属性依赖度算法和基于互信息的算法进行比较, 结果如表 1 所列。

表 1 不同算法结果比较

β	δ_0	基于属性 依赖度算法	基于信息 熵算法	基于属性 敏感度算法
0	0	{麻黄, 桂枝, 防风, 甘草, 羌活, 细辛}	{麻黄, 桂枝, 防风, 甘草, 羌活, 细辛}	{麻黄, 桂枝, 防 风, 羌活, 细辛}
0	0.3	{麻黄, 桂枝, 防风, 羌活, 细辛}	{麻黄, 桂枝, 防风, 甘草, 细辛}	{麻黄, 桂枝, 防风, 甘草}
0.25	0	{麻黄, 桂枝, 防风, 细辛}	{麻黄, 桂枝, 防风, 甘草}	{麻黄, 桂枝, 防风}
0.2	0.15	{桂枝, 防风, 羌活, 细辛}	{桂枝, 防风, 羌活}	{桂枝, 防风}

从表 1 可以看出, 本文算法所获得的约简最小, 即为治疗“恶寒、发热”症状的君药, 与中医理论相符, 防止了非核心药物的干扰, 杜绝了方剂配伍中“过于频繁反而平凡”的现象, 可用于指导临床用药。

结束语 本文对不完备信息系统下的变精度粗糙集进行了拓展, 得到了加权变精度容差关系模型, 此模型相比传统的粗糙集模型拥有更大的下近似集和更小的上近似集, 提高了近似精度。本文提出基于属性敏感度的约简算法能得到最小的约简结果, 在海量方剂配伍规律分析中得到了验证, 可用于发现药症对应关系, 指导临床用药。今后的工作仍将围绕变精度粗糙集模型的拓展而展开, 探讨更高效完备的属性约简算法以用于大数据分析。

参考文献

- [1] Pawlak Z. Rough sets[J]. International Journal of Computer and Information Sciences, 1984, 11: 341-356
- [2] Kryszkiewicz M. Rough set approach to incomplete information systems[J]. Information Sciences, 1998, 112(1-4): 39-49
- [3] Ziarko W. Analysis of uncertain information in the framework of variable precision rough sets[J]. Foundations of Computing and Decision Science, 1993, 18(3-4): 381-396
- [4] Ziarko W. Variable precision rough set model[J]. Journal of Computer and System Sciences, 1993, 46: 39-59
- [5] 张文东, 李明壮, 石小艳. 基于粗糙集理论的属性约简算法[J]. 计算机工程与设计, 2008, 11(29): 95-97
- [6] He X. coefficient of variation and its application to strength prediction of adhesively bonded Joints[C]//Proceedings of International Conference on Measuring Technology and Mechatronics Automation. 2009: 602-605
- [7] Kryszkiewicz M. Rough set approach to incomplete information systems[J]. Information Science, 1998, 112: 39-49
- [8] 胡军, 王国胤. 粗糙集不确定性度量准则[J]. 模式识别与人工智能, 2010, 23(5): 606-614
- [9] 张长胜. 基于决策表的区分矩阵增量属性约简算法[J]. 计算机工程与应用, 2012, 48(35): 110-113
- [10] 陈媛, 杨栋. 基于信息熵的属性约简算法及应用[J]. 重庆理工大学学报: 自然科学版, 2013, 27(1): 42-46
- [11] 滕书华, 孙即祥, 周石琳, 等. 基于信息观点的约简算法比较[J]. 计算机科学, 2011, 38(1): 259-263
- [12] 彭怀仁. 中医方剂大辞典[M]. 北京: 人民卫生出版社, 2005