

粗糙随机信息系统可测性的研究^{*}易树鸿¹ 张为群²(遵义师范学院计算机科学系 遵义563002)¹ (西南师范大学计算机与信息科学学院 重庆400715)²

Study of Measurability on Rough Random Information System

YI Shu-Hong¹ ZHANG Wei-Qun²(Department of Computer Science, Zhunyi Normal College, Zhunyi 563002)¹(School of Computer and Information Science, Southwest Normal University, Chongqing 400715)²

Abstract In this paper, we study measurability of rough random information system. Main results are that introduced concept of rough measurable space from the measure theory, created measurements of upper and lower approximately probability, gave strictly definition of rough probability space, final discussed its some properties.

Keywords Rough random information system, Rough measurable space, Upper approximately probability, Lower approximately probability, Rough probability space

1 引言

在知识发现、智能决策、综合评价和自动推理等应用领域,人们经常会遇到这样一类不确定性问题:某个需要处理的事件的出现带有随机性,但在寻找它出现的概率时却发现对它的描述不够精确,例如某些疑难病症的诊断、对知识产品的价值评价等。对于这类事件,由于缺乏足够的信息来精确描述这些事件,使得我们无法用通常的概率统计方法来准确地加以计算。我们称带有这样一类不确定性事件的信息系统为粗糙随机信息系统。文[1]中从概率论与粗糙集理论的结合中给出这种既带有随机性而本身又不能精确描述的不确定性现象的表达模型。但定义中描述性的成分过大,一些深层次的基础问题并没有涉及,如粗糙随机事件的上、下概率估计是否合理(即系统是否可测)等。本文将对粗糙随机信息系统的可测性做进一步的研究,以便为应用提供坚实的理论基础和一般方法。

文中涉及到的有关概率论的知识可以在文[2,3]中找到,有关粗糙集的知识请参见文[4~6]。

2 粗糙随机信息系统

为论述的需要,我们给出有关信息系统的形式化描述。

定义2.1 一个信息系统是指一个三元组 $S=(\Omega, F, \Gamma)$ 。其中 Ω 是论域, F 是 Ω 的子集簇, Γ 是 Ω 上的等价关系簇。如果 F 还是 Ω 中的 σ -代数,则称 S 是一个随机信息系统。

在随机信息系统中,论域 Ω 称为样本空间, F 叫做事件体, F 中的集合 A 称为事件。

定义2.2 设 Ω 是一个样本空间, R 是 Ω 上的等价关系。我们称商集 $\Omega/R = \{[\omega]_R | \omega \in \Omega\}$ 为 Ω 关于 R 的粗糙样本空间,简称粗糙样本空间,记为 $R(\Omega)$ 。 $R(\Omega)$ 的元素 $[\omega]_R$ 称为粗糙样本点。

定义2.3 设 Ω 是一个样本空间, R 是 Ω 上的等价关系。如果对每一集 $A \in F$, 都有 $A = \bigcup [\omega]_R$, ($[\omega]_R \in R(\Omega)$), 则称

F 是 R 可定义的, 否则称 F 是 R 不可定义的。

定义2.4 设 $S=(\Omega, F, \Gamma)$ 是一个随机信息系统, 如果对任意 $R \in P, \Gamma$ 都是 R 可定义的, 则称 S 是一个精确随机信息系统。否则称 S 是一个粗糙随机信息系统。

对于精确随机信息系统来说, 其可测性已由概率论的知识所保证。按照概率论的说法, 如果 F 是 Ω 中的 σ -代数, 则称 (Ω, F) 为可测空间, F 中的任一集合称为可测集。在可测空间 (Ω, F) 上可以定义概率测度 P , 从而得到概率空间 (Ω, F, P) 。

在粗糙随机信息系统 S 中, 其可测性需要解决两个问题, 一是 F 中事件的表示问题, 二是概率测度的可定义问题。

3 粗糙可测空间

设 S 是一个粗糙随机信息系统, 我们首先解决 F 中事件的表示问题。

定义3.1 设 F 为 Ω 中的 σ -代数, $R(\Omega)$ 是粗糙样本空间。如果对每一集 $A \in F$, 都有 $A = \bigcup [\omega]_R$, ($[\omega]_R \in R(\Omega)$), 则称 F 为 Ω 中的 R -精确 σ -代数, 否则称 F 为 Ω 中的 R -粗糙 σ -代数, 简称粗糙 σ -代数, 记为 $R(F)$ 。

定义3.2 包含有 Ω 、粗糙样本空间 $R(\Omega)$ 和粗糙 σ -代数 $R(F)$ 的三元组 $(\Omega, R(\Omega), R(F))$ 称为粗糙可测空间。 $R(F)$ 中的任一集合称为粗糙可测集。

注: $R(F)$ 也叫做粗糙事件体, $R(F)$ 中的集合 A 称为粗糙事件。

由于 $R(F)$ 中的集 A 不能被 $R(\Omega)$ 的元素精确表示, 只能通过逼近的方式来表达, 因此我们引进下面的定义。

定义3.3 设 $(\Omega, R(\Omega), R(F))$ 是粗糙可测空间。对每一集 $A \in R(F)$, A 的上近似可测集和 A 的下近似可测集分别如下:

$$A^+ = \bigcup_{[\omega]_R \cap A \neq \emptyset} [\omega]_R \quad A^- = \bigcup_{[\omega]_R \subseteq A} [\omega]_R$$

4 粗糙概率空间

下面解决粗糙随机信息系统 S 中概率测度的可定义问

^{*}贵州省科学技术基金项目(20023088)资助, 易树鸿 副教授, 主要从事人工智能和 Rough 集理论的研究。

题。我们要在粗糙可测空间 $(\Omega, R(\Omega), R(F))$ 上引进概率测度,由于 $R(F)$ 中的集合不是都能用 $R(\Omega)$ 中的元素来精确表示,因此不能按通常可测空间的做法直接引入概率。为此,我们先做如下的准备工作。

4.1 $(\Omega, R(\Omega), R(F))$ 的近似可测空间

考虑 Ω 的子集所成的集类:

$$F_R = \{A^* | A^* = \bigcup_{[\omega]_R \in R(\Omega)} [\omega]_R, A_- \subseteq A^* \subseteq B^-(A, B \in F)\}$$

F_R 中的集合具有这样的特征,首先它是由 Ω 的一些 R -等价类组成的,其次它包含 F 的某个集合的下近似,同时含于 F 的某个集合的上近似中。集类 F_R 具有如下性质。

定理4.1.1 设 $A \in R(F)$,则 $A_- \in F_R, A^+ \in F_R$ 。

证明:由 F_R 的构造,显然。

定理4.1.2 F_R 是 Ω 中的一个 σ -代数。

证明:(1) $\because \Omega \in R(F)$,而 $\Omega = \Omega_- = \Omega^+$, $\therefore$ 由定理4.1.1, $\Omega \in F_R$ 。

(2) $\forall A^* \in F_R$,则由它是由 Ω 的一些 R -等价类组成的可知, $\overline{A^*}$ 也是由 Ω 的一些 R -等价类组成的。再由 $\exists A, B \in F, A_- \subseteq A^* \subseteq B^-$ 知, $\overline{(B^-)} \subseteq \overline{A^*} \subseteq \overline{(A_-)}$,即 $\overline{(B^-)} \subseteq \overline{A^*} \subseteq \overline{(A_-)}$,而 $\overline{A}, \overline{B} \in F, \therefore \overline{A^*} \in F_R$ 。

(3)只要证明 F_R 中两个集合的并还在 F_R 中,则利用数学归纳法就有 F_R 中可数个集合的并仍在 F_R 中。 $\forall A_1^*, A_2^* \in F_R$,则显然有 $A_1^* \cup A_2^*$ 是由 Ω 的一些 R -等价类组成的。

$$\text{设 } A_1_- \subseteq A_1^* \subseteq B_1^-, A_2_- \subseteq A_2^* \subseteq B_2^-$$

$(A_1, A_2, B_1, B_2 \in F)$,于是有:

$$A_1_- \subseteq A_1_- \cup A_2_- \subseteq A_1^* \cup A_2^* \subseteq B_1^- \cup B_2^- = (B_1 \cup B_2)^-$$

$$\therefore A_1^* \cup A_2^* \in F_R.$$

综合(1)(2)(3)可知 F_R 是 Ω 中的一个 σ -代数。证毕。

由于 F_R 是 Ω 中的一个 σ -代数,并且包含了 $R(F)$ 中所有集合的上、下近似集,因此我们称 F_R 是 $R(F)$ 的近似 σ -代数,而称可测空间 (Ω, F_R) 是 $(\Omega, R(\Omega), R(F))$ 的近似可测空间。

4.2 粗糙概率空间

注意到 (Ω, F_R) 是可测空间,因而可以直接在 (Ω, F_R) 上引入概率 P ,从而得到概率空间 (Ω, F_R, P) 。

下面我们将利用概率空间 (Ω, F_R, P) 来寻求 $R(F)$ 中粗糙可测集的概率。对任一集 $A \in R(F)$,由于 A 的粗糙性,我们无法从 (Ω, F_R, P) 来计算 A 的精确概率,只能给出它的逼近值。

对任一集 $A \in R(F)$,定义实值集函数 $P_-(A) = P(A_-)$,这里 A_- 是 A 的下近似集,而 $P(A_-)$ 是概率空间 (Ω, F_R, P) 中事件 A_- 的概率,因此这样定义的实值集函数 $P_-(A)$ 是有意义的,下面看它的性质。

定理4.2.1 由上定义的实值集函数 $P_-(A)$ 满足下面三条性质:

$$(1) \text{对每一 } A \in R(F), \text{有 } 0 \leq P_-(A) \leq 1;$$

$$(2) P_-(\Omega) = 1;$$

$$(3) \text{对任意 } A_i \in R(F) (i=1, 2, \dots),$$

$$A_i \cap A_j = \Phi (i \neq j), \text{恒有:}$$

$$P_-(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P_-(A_i).$$

证明:(1)对每一 $A \in R(F)$,有 $A_- \in F_R$,由于 P 是可测空间 (Ω, F_R) 上的概率,因此有 $0 \leq P(A_-) \leq 1$,从而有 $0 \leq P_-(A) \leq 1$ 。

$$(2) \because \Omega \in F_R, \Omega = \Omega_- , \therefore P_-(\Omega) = P(\Omega_-) = P(\Omega) = 1.$$

(3)只需考虑对两个 $A, B \in F_R, A \cap B = \Phi$,时结论成立,即可用数学归纳法得出所要的结论。

对此,我们首先证明当 $A \cap B = \Phi$ 时,有 $(A \cup B)_- = A_- \cup B_-$ 。事实上,由粗糙集的性质有 $(A \cup B)_- \supseteq A_- \cup B_-$,因此只要证明相反的包含也成立即可。

$$\forall x \in (A \cup B)_-, \text{则 } \exists [\omega] \subseteq A \cup B, x \in [\omega].$$

由于 $A \cap B = \Phi$,故 $[\omega]$ 只含于其中一个,不妨设为 A 。

则有 $[\omega] \subseteq A, x \in [\omega]$,即 $x \in A_-$,从而 $x \in A_- \cup B_-$,因此 $(A \cup B)_- \subseteq A_- \cup B_-$, $\therefore (A \cup B)_- = A_- \cup B_-$ 。

再由 $A \cap B = \Phi$ 知 $A_- \cap B_- = \Phi$,于是我们有:

$$P_-(A \cup B) = P((A \cup B)_-) = P(A_- \cup B_-) = P(A_-) + P(B_-) = P_-(A) + P_-(B).$$

证毕。

由于 P_- 具有概率函数的三条性质,因此下面的定义是有意义的。

定义4.2.1 设 $(\Omega, R(\Omega), F_R)$ 是粗糙可测空间,对任一集 $A \in R(F)$,定义实值集函数 $P_-(A) = P(A_-)$ 。称 P_- 为 $(\Omega, R(\Omega), R(F))$ 上的下近似概率, $P_-(A)$ 为事件 A 的下近似概率。

类似地,对任一集 $A \in R(F)$,定义实值集函数 $P^-(A) = P(A^+)$,可以证明 P^- 满足下面的性质。

定理4.2.2 由上定义的实值集函数 $P^-(A)$ 满足下面三条性质:

$$(1) \text{对每一 } A \in R(F), \text{有 } 0 \leq P^-(A) \leq 1;$$

$$(2) P^-(\Omega) = 1;$$

$$(3) \text{对任意 } A_i \in R(F) (i=1, 2, \dots), A_i \cap A_j = \Phi (i \neq j), \text{恒}$$

$$\text{有: } P^-(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P^-(A_i).$$

由此可以引进下面的定义。

定义4.2.2 设 $(\Omega, R(\Omega), R(F))$ 是粗糙可测空间,对任一集 $A \in R(F)$,定义实值集函数 $P^-(A) = P(A^+)$ 。称 P^- 为 $(\Omega, R(\Omega), R(F))$ 上的上近似概率, $P^-(A)$ 为事件 A 的上近似概率。

总结上面的讨论,我们有如下的结果。

定义4.2.3 设 Ω 是一样本空间, $R(\Omega)$ 是由 Ω 通过等价关系 R 生成的粗糙样本空间, $R(F)$ 是 Ω 中的粗糙 σ -代数, P_- 为 $R(F)$ 上的下近似概率, P^- 为 $R(F)$ 上的上近似概率。我们称具有上述结构的样本空间为粗糙概率空间,记为 $(\Omega, R(\Omega), R(F), P_-, P^-)$ 。

至此,我们可以用粗糙概率空间 $(\Omega, R(\Omega), R(F), P_-, P^-)$ 来处理粗糙随机信息系统的可测性问题。在不能直接求得粗糙事件 A 出现的概率的情况下,可以通过它的下近似概率和上近似概率的值进行估计。

5 上、下近似概率的性质

在用粗糙概率空间 $(\Omega, R(\Omega), R(F), P_-, P^-)$ 来处理粗糙随机信息系统的可测性问题中,上、下近似概率在其中扮演着主要的角色。为了今后应用的需要,下面给出上、下近似概率 P^- 和 P_- 的若干性质。

定理5.1 对任一 $A \in R(\Phi)$,有 $P_-(A) \leq P(A) \leq P^-(A)$ 。

证明: $\because$ 对任一 $A \in R(\Phi)$,有 $A_- \subseteq A \subseteq A^+$, $\therefore$ 由概率的性质有

$$P(A_-) \leq P(A) \leq P(A^+), \text{即 } P_-(A) \leq P(A) \leq P^-(A).$$

证毕。

结合定理5.1,我们给出下列说法。

定义5.1 对任一 $A \in R(\Phi)$, 称 $P_-(A)$ 为事件 A 的概率下估计值, $P^-(A)$ 为 A 的概率上估计值, 而称 $[P_-(A), P^-(A)]$ 为 A 的粗糙概率区间估计。

定理5.2 $P_-(\Phi) = P^-(\Phi) = 0$ 。

证明: $\because \Phi_- = \Phi = \Phi^+$, $\therefore P_-(\Phi) = P(\Phi_-) = P(\Phi) = 0, P^-(\Phi) = P(\Phi^+) = P(\Phi) = 0$ 。证毕。

定理5.3 若 $A_i \in R(F) (i=1, 2, \dots, n), A_i \cap A_j = \Phi (i \neq j)$, 则:

$$P_-(\bigcup_{i=1}^n A_i) = \sum_{i=1}^n P_-(A_i), P^-(\bigcup_{i=1}^n A_i) = \sum_{i=1}^n P^-(A_i)。$$

证明: 由定理4.2.1之(3), 令 $A_{n+1} = A_{n+2} = \dots = \Phi$, 应用 $P_-(\Phi) = 0$ 这一结果, 得:

$$P_-(\bigcup_{i=1}^n A_i) = P_-(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P_-(A_i) = \sum_{i=1}^n P_-(A_i)。$$

同理可得 $P^-(\bigcup_{i=1}^n A_i) = \sum_{i=1}^n P^-(A_i)$ 。证毕。

定理5.4 对任一 $A \in R(F)$, 有 $P_-(\bar{A}) = 1 - P_-(A), P^-(\bar{A}) = 1 - P^-(A)$ 。

证明: $\because A \cup \bar{A} = \Omega, A \cap \bar{A} = \Phi$, $\therefore$ 由定理5.3得 $P_-(A) + P_-(\bar{A}) = 1$, 即 $P_-(\bar{A}) = 1 - P_-(A)$ 。

同理可证 $P^-(\bar{A}) = 1 - P^-(A)$ 。证毕。

定理5.5 对 $A, B \in R(F)$, 若 $A \supseteq B$, 则:

(1) $P_-(A-B) = P_-(A) - P_-(B)$, 且 $P_-(A) \geq P_-(B)$;

(2) $P^-(A-B) = P^-(A) - P^-(B)$, 且 $P^-(A) \geq P^-(B)$ 。

证明: 我们只给出(1)的证明, 对(2)可做类似讨论。

$\because A = (A-B) \cup B, (A-B) \cap B = \Phi$, $\therefore$ 由定理5.3得

$$P_-(A) = P_-(A-B) + P_-(B),$$

$$\text{即 } P_-(A-B) = P_-(A) - P_-(B)。$$

$$\text{又 } \because P_-(A-B) \geq 0,$$

$$\therefore P_-(A) \geq P_-(B)。$$
 证毕。

小结 我们以粗糙随机信息系统 S 的论域 Ω 上的等价关系 R 为基础, 逐步引进了粗糙样本空间 $R(\Omega)$ 和粗糙 σ -代数 $R(F)$, 从而形成了粗糙可测空间 $(\Omega, R(\Omega), R(F))$ 。为了在 $(\Omega, R(\Omega), R(F))$ 上引进概率测度, 我们构造了 $(\Omega, R(\Omega), R(F))$ 的近似可测空间 (Ω, F_R) 。然后利用 (Ω, F_R) 上的概率测度 P 为 $(\Omega, R(\Omega), R(F))$ 引进了下近似概率 P_- 和上近似概率 P^+ , 这样形成了粗糙概率空间 $(\Omega, R(\Omega), R(F), P_-, P^+)$, 从而解决了粗糙随机信息系统 S 的可测性问题。最后我们还讨论了上、下近似概率的若干性质。有关粗糙随机信息系统的进一步研究, 我们还在进行中。

参考文献

- 1 易树鸿. 粗糙随机事件及其概率估计. 计算机科学, 2002, 29(9. 专刊): 126~134
- 2 中山大学数学力学系. 概率论及数理统计. 人民教育出版社, 1980
- 3 袁震东. 近代概率引论——测度、鞅和随机微分方程. 科学出版社, 1991
- 4 Pawlak Z. Rough set. International Journal of Information and Computer Science, 1982(11): 341~356
- 5 Pawlak Z. Rough set — theoretical aspects of reasoning about data. Dordrecht: Kluwer Academic Publishers, 1991
- 6 刘清. Rough 集及 Rough 推理. 科学出版社, 2001

(上接第152页)

表1 第一个实验两种算法性能的比较

实验方法	样本 点数	迭代 次数	实验 次数	正确聚 类次数	聚类 正确率
K 均值方法	500	20	1000	723	72.3%
混合聚类方法	500	20	1000	989	98.9%

表2 第二个实验两种算法性能的比较

实验方法	样本 点数	迭代 次数	实验 次数	正确聚 类次数	聚类 正确率
K 均值算法	500	30	1000	714	71.4%
混合聚类算法	500	30	1000	978	97.8%

结论 HCA 是在传统的聚类算法中结合了免疫进化的机理, 将全局寻优与局部搜索能力相集成, 并引入了接种疫苗和免疫检测, 从而提高了聚类算法的性能。理论分析和实验结果表明, 本文提出的 HCA 算法明显地克服了 K 均值算法中的局部极值问题和对初始值敏感性的缺点, 并且具有较快的收敛速度, 同时也减轻了进化后期的波动现象。

参考文献

- 1 Jain A K, Murty M N, Flynn P J. Data clustering: A survey. ACM Compute. Surv., 1999, 31: 264~323

- 2 Jain A K, Dubes R C. Algorithms for clustering data. Englewood Cliffs, NJ: prentice Hall, 1988
- 3 高新波, 谢维新. 模糊聚类理论发展及应用的研究进展. 科学通报, 1999, 44(21): 2241~2251
- 4 MacQueen J. Some methods for classification and analysis of multivariate observations. Proc. 5th Berkeley Symp. Math. Statist, Prob, 1967, 1: 281~297
- 5 Huang Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery, 1998, 2: 283~304
- 6 许雷. 一种聚类新算法: 模拟退火[J]. 模式识别与人工智能, 1989, 1: 1~16
- 7 Babu G P, Murty M N. Clustering with evolution strategies [J]. Pattern Recognition, 1994, 2(27): 321~329
- 8 Krishma K, Murty M N. Genetic K-Means Algorithm. IEEE Trans on System, Man, and Cybernetics, Part B, 1999, 29(3): 433~439
- 9 Rudolph G. Convergence of Canonical Genetic Algorithm. IEEE Trans on Neural Networks, 1994, 5(1): 96~100
- 10 Mali U, Bandyopadhyay S. Genetic Algorithm-Based Clustering Technique. Pattern Recognition, 2000, 33(9): 1455~1465
- 11 胡玉锁, 陈宗海. 基于混合遗传算法的聚类分析. 模式识别与人工智能, 2001, 14(3): 352~335
- 12 王磊, 潘进, 焦李成. 免疫规划. 计算机学报, 2000, 23(8): 806~812