粗糙集分类算法中的近似决策规则和规则匹配方法*)

张雪英[1,2] 刘凤玉[1] Jürgen Krause[3]
(南京理工大学计算机科学与工程系 南京210094) (南京农业大学信息科技学院 南京210095)[2]
(德国社会科学信息中心)[3]

摘 要 粗糙集分类算法在应用标准决策规则进行新对象分类时,经常碰到决策规则与新对象不完全匹配的情况。因此,近似决策规则和部分匹配方法常用于提高决策规则与新对象匹配的可能性。本文在概述和比较两种近似决策规则生成算法的基础上,以一个文本分类系统为例,提出了一种综合的、更有效的近似决策规则生成算法。文章还介绍了几种通用的规则匹配方法,提出了一系列实用的完全匹配和部分匹配公式。实验表明,新提出的近似决策规则生成算法和规则匹配公式能够有效地提高决策规则与新对象的匹配可能性与准确性。

关键词 粗糙集,分类算法,近似规则,决策规则,匹配规则

Approximate Decision Rules and Matching Rules in Rough Set Based Classification Algorithms

ZHANG Xue-Ying[1,2] LIU Feng-Yu[1] Jürgen Krause[3]
(Department of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094)[1]
(Information Science and Technology College, Nanjing Agricultural University, Nanjing 210095)[2]
(Social Science Information Centre, Germany)[3]

Abstract In most cases, the decision rules inducted by rough set models are unacceptable as laws to classify new objects. Approximate decision rules and partial matching rules are proposed to overcome this problem. This paper discusses two typical algorithms for the generation of approximate rules and comparatively analyzes their performance as proven by one case study. Furthermore, one more efficient algorithm is developed based on the two algorithms. This paper also describes the general measures used for matching rules, and a set of formulae are defined for complete matching and partial matching of decision rules according to dependency coefficient in rough set theory. The experiments show that the proposed approximation algorithm and measures for matching rules can further improve the matching possibility and correctness of basic decision rules generated based on rough set theory.

Keywords Rough set, Classification algorithm, Approximate rule, Decision rule, Matching rule

1 引言

自1982年波兰学者 Zdzislaw Pawlak 首次提出粗糙集理论(rough set theory)以来,经过短短二十多年的发展,粗糙集理论已经广泛应用于知识发现、数据挖掘、智能决策和电子控制等多个领域。粗糙集的主要思想是在保持分类能力不变的前提下,通过知识约简,导出问题的决策规则(或者分类规则)[1]。大多数情况下,应用粗糙集方法推导出的标准决策规则(standard decision rules, basic decision rules)往往不能对新对象进行有效的分类。主要原因在于新对象的属性与决策规则可能完全不匹配或者不完全匹配。近似决策规则生成算法和部分匹配(partial matching rules)方法是解决这一难题的常用手段。近似决策规则(approximate decision rules)包括用各种方法从标准决策规则中导出的用于提高规则与新对象匹配可能性的所有规则[2,3]。近似决策规则生成算法可以概括为两种类型:一种是通过降低决策规则的一致性来增加规则数量,如 Banzan 的算法;另一种是运用粗糙集的上下近似基本理论,对决策规则进行扩充,如 Grzymala-Busse 的算法。规则匹配包括完全匹配和部分匹配两种类型。Bazan 在总结已有规则匹配方法的基础上提出了一系列的完全匹配公式[4]。相比较而言,人们对部分匹配的研究还较少,Grzymala-Busse 曾提出过一种部分匹配方法[3]。

本文主要研究内容包括两个部分:一是在比较研究两种典型的近似决策规则生成算法基础上,提出更有效的近似决策规则生成算法;二是比较分析现有的完全匹配方法,重新定义一系列的完全匹配和部分匹配公式。本文以一个文本自动分类系统为例,测试了各种近似决策规则生成算法和规则匹配方法的检索性能。

2 近似决策规则生成算法

粗糙集理论是一种处理模糊和不确定性知识的数学工具。设 U 为一组对象组成的非空有限集合,称为论域。如果 R 为 U 上的一个等价关系,那么 $A=(U,R)$ 称为一个近似空间,R 为一个不可区分关系(indiscernibility),记为 $IND(R)$。$U/IND(R)$的等价类称为基本范畴。设 X 为 U 的一个对象子集,X 称为 U 中的一个概念和范畴。如果 $R\in IND(U)$,X 的下近似(lower approximation)和上近似(upper approximation)分别定义为:

$$\underline{R}X=\bigcup\{Y\in U/R:Y\subseteq X\} \quad (1)$$

$$\overline{R}X=\bigcup\{Y\in U/R:Y\cap X\neq\phi\} \quad (2)$$

其中下近似为 X 中的最大可定义集,上近似为含有 X 的最小可定义集[1,5~7]。

假设一个信息系统 $K=(U,R)$,$P\subseteq R$ 和 $Q\subseteq R$。如果 Q 的所有(或部分)基本范畴能由 P 来定义,那么 Q 在 $k(0\leqslant k\leqslant$

*)本文研究由德国社会科学信息中心(Social Science Information Center, Germany)提供资助。张雪英 博士研究生,副教授,主要研究领域为智能信息检索,数据挖掘和自然语言处理。

1)程度上依赖于 P。也就是说，P 和 Q 存在语义关系。

$$k=r(Q)=\frac{card\ POS_P(Q)}{cardU} \quad POS_P(Q)=\bigcup_{X\in U/Q}\underline{P}X \qquad (3)$$

当 $k=1$时，Q 完全依赖于 P；当$0<k<1$时，Q 粗糙地(roughly)依赖于 P；当 $k=0$时，Q 完全独立于 P。在下文中，K 被称之为依赖度。

假设 $A=(U,A\cup\{d\})$是一个决策表，$\mu_A(r)$表示规则 r 的一致性系数(consistency coefficient)。如果 $\mu_A(r)$等于1，说明该决策规则是一致的，否则就不一致。决策规则中每个不同的决策值对应着一个概念(concept)。假设 $Supp_A(r)$表示支持规则 r 的对象数量，$Match_A(r)$表示与规则 r 的条件属性匹配的对象数量，$\mu_A(r)$定义为：

$$\mu_A(r)=\frac{Supp_A(r)}{Match_A(r)} \qquad (4)$$

Bazan 算法的基本思想：①输入所有标准决策规则；②定义一致性系数 μ_0；③计算每条决策规则 r_0的一致性系数 $\mu_A(r_0)$；④如果其最大值大于 μ_0，去掉条件属性中一个属性后，再判断该规则的一致性系数；⑤如果此时一致性系数大于阈值，则生成一条新的近似规则；如果 $\mu_A(r_0)<\mu_0$则不产生近似规则；如果 $\mu_{max}=\mu_A(r_0)$，近似规则等于输入的决策规则；⑦以此类推，直到输入的所有规则处理完毕[4]。

Grzymala-Busse 算法的基本思想：①判断输入决策规则的一致性，如果决策规则不一致，则计算所有概念的上下近似；②将上近似规则和下近似规则放入两个不同的表中；③根据属性的优先权等特征找出最少的规则组合来描述每个概念，其实质也是尽可能减少规则中属性的个数；④如果新对象的属性与多条规则完全匹配，则选择在训练集中条件概率值最大的规则为最终匹配规则[3]。

我们以一个文本分类系统为例，分别测试了上述两种算法的性能。一个分类系统就是应用已经获得的规则对新对象进行分类。通常情况下，分类系统运用从训练集中推导出的规则来对测试集中的对象进行预测类别[3]。实验数据集(CWT)由北京大学网络实验室提供，包括15605个网页文本。实际实验用数据集包括14844个网页，原数据集中重复、分类不一致和文本内容不完整的网页被删除。训练集和测试集划分与原数据集一致。表1为数据集的分类体系结构以及训练集和测试集的划分情况。标准决策规则已经从训练集中导出。分类系统的性能用查全率和查准率两个指标进行评估。表2为标准分类规则的检索性能，其中规则匹配采用 Grzymala-Busse 算法中定义的完全匹配方法。

Bazan 算法必须人为设定一致性系数的阈值。实际操作中，我们很难事先知道哪一个阈值可以获得最优的检索性能。因此，必须采用不同的阈值进行实验，然后再确定最佳的阈值及近似决策规则。我们分别用0.9、0.8和0.7三个阈值进行了实验，实验结果见表3和表4。表5为应用 Grzymala-Busse 算法的实验结果。规则匹配同样采用 Grzymala-Busse 算法中定义的完全匹配方法。

表1 CWT 数据集的类别分布及划分

大类	大类代码	二级类类目数	三级类类目数	总计类目数	训练文献数	平均训练文献数	测试文献数	平均测试文献数
人文与艺术	01	4	19	24	399	16.6	103	4.3
新闻与媒体	02	6	0	7	121	17.3	18	2.6
商业与经济	03	10	37	48	778	16.2	206	4.3
娱乐与休闲	04	16	71	88	1470	16.7	362	4.1
计算机与因特网	05	11	45	58	904	15.6	232	4.0
教育	07	13	4	18	272	15.1	80	4.4
区域	08	2	50	53	867	16.4	225	4.2
自然科学	10	14	96	113	1772	15.7	492	4.4
政府与政治	11	6	10	17	264	15.5	73	4.3
社会科学	12	15	87	104	1676	16.1	450	4.3
医疗与健康	13	12	123	136	2216	16.3	590	4.3
社会与文化	14	14	51	66	997	15.1	277	4.2

表2 标准决策规则的检索性能评估

类目	01	02	03	04	05	07	08	10	11	12	13	14
查准率	0.49	0.39	0.58	0.69	0.65	0.42	0.58	0.66	0.50	0.61	0.64	0.49
查全率	0.32	0.44	0.31	0.33	0.43	0.35	0.28	0.20	0.30	0.22	0.20	0.33

表3 Bazan 算法的查准率

阈值＼类目	01	02	03	04	05	07	08	10	11	12	13	14
阈值=0.9	0.50	0.42	0.61	0.69	0.66	0.45	0.58	0.67	0.63	0.62	0.64	0.50
阈值=0.8	0.62	0.49	0.71	0.77	0.76	0.61	0.74	0.78	0.63	0.72	0.76	0.65
阈值=0.7	0.64	0.50	0.64	0.75	0.74	0.62	0.70	0.74	0.61	0.69	0.74	0.62

表4 Bazan 算法的查全率

阈值＼类目	01	02	03	04	05	07	08	10	11	12	13	14
阈值=0.9	0.25	0.21	0.28	0.27	0.32	0.30	0.21	0.20	0.26	0.15	0.12	0.24
阈值=0.8	0.35	0.47	0.35	0.38	0.48	0.39	0.36	0.44	0.39	0.28	0.32	0.41
阈值=0.7	0.28	0.24	0.29	0.29	0.33	0.32	0.24	0.20	0.26	0.17	0.17	0.25

表5 Grzymala-Busse 算法的检索性能评估

类目	01	02	03	04	05	07	08	10	11	12	13	14
查准率	0.45	0.39	0.57	0.64	0.62	0.40	0.55	0.64	0.49	0.57	0.61	0.47
查全率	0.35	0.54	0.41	0.37	0.44	0.45	0.34	0.24	0.36	0.32	0.28	0.37

从表2、表3和表4中可以看出，Bazan 算法明显比标准分类规则获得的检索性能高。因为通过该算法增加了大量的近似规则，提高了规则对新对象的匹配能力。实验发现，当一致性系数阈值定义为0.8，系统可以获得最佳的检索性能。因此，针对不同的分类系统必须经过多次试验，才能确定最佳的近似规则。

很显然，Grzymala-Busse 算法虽然可以大大提高规则的匹配能力，但准确性却相应降低。该算法能够比 Bazan 算法生成更多的近似规则，而且规则的平均属性个数明显少于 Bazan 算法生成的近似规则。因此，规则属性数量的减少导致了对新对象分类预测准确性的降低。

为了使分类系统既能提高对新对象的匹配能力，又能提高分类预测的准确性，我们提出了一种综合算法来生成近似决策规则（见算法1）。

算法1：

input standard decision rules (D_rules)
generate approximate rules (B_rules) of D_rules with Bazan's algorithm
Integrate D_rules and B_rules into C_rules
generate approximate rules (G_rules) of C_rules with Grzymala-Busse's algorithm
Integrate C_rules and G_rules into F_rules
Compute the consistency coefficient of F_rules
Delete the rules from F_rules for consistency coefficient < defined threshold
Output F_rules

实验表明，算法1不仅可以生成比 Bazan 算法更多的决策规则，而且能够明显提高系统的分类性能。实验结果见表6。

表6 算法1的检索性能评估

类目	01	02	03	04	05	07	08	10	11	12	13	14
查准率	0.65	0.59	0.74	0.84	0.81	0.65	0.74	0.79	0.71	0.77	0.83	0.69
查全率	0.45	0.55	0.42	0.39	0.47	0.47	0.36	0.29	0.40	0.42	0.38	0.45

3 规则匹配(matching rules)

近似决策规则生成算法是从增加规则的角度来提高决策规则对新对象的匹配能力。因此，近似决策规则生成算法会影响分类规则本身的质量。而且，即使生成较多的近似决策规则，在分类系统的实际应用中，仍然会经常发生新对象不能找到匹配规则的现象。由于新对象千变万化，要在决策规则中找到完全匹配的规则会有一定的难度。于是，人们提出了部分匹配方法。部分匹配实质上就是通过缩减新对象中属性的个数，以尽可能在决策规则中找到合适的完全匹配规则。与近似决策规则生成算法相比，部分匹配方法有较强的灵活性，但是容易造成分类准确率降低。在详细讨论部分匹配方法之前，我们先简单介绍一下完全匹配方法。

3.1 完全匹配(complete matching)

通常情况下，完全匹配按照以下步骤操作：

·在一个粗糙集分类器中对新对象进行表达，抽取出表达新对象的属性值对(attribute-value pairs)。

·在规则集中查找匹配的规则。如果有只有一条规则与之完全匹配，则将新对象归至匹配规则决策值对应的类别。

·如果没有规则与之相匹配，则将训练集最常出现的分类结果赋给该对象。

·如果有多个规则与之相匹配，必须对所有匹配规则进行相关排序，然后将新对象归至相关性值最高的规则所定义的类别[8]。事实上，这也是完全匹配要解决的主要问题。本文将这个过程称之为完全匹配规则的优选。

Grzymala-Busse，Holland 和 Booker 通过研究发现，完全匹配规则的优选需要考虑三个因素[9~11]：

·规则强度(Strength)：训练集中支持该规则的对象个数。

·规则专指度(Specificity)：规则中条件属性的个数。

·规则支持度(Support)：所有匹配规则的相关性值总和。规则支持度定义为：

$$\delta=\sum_{\text{matching rules } R \text{ describing } C} Strength(R)\times Specificity(R) \tag{5}$$

δ 值越大，表示新对象与 C 类越相关。因此，新对象应归至 δ 值最大的类别。

Bazan 在总结已有完全匹配方法的基础上，提出了一系列通用的完全匹配规则的优选方法。下文主要讨论 Bazan 的方法。

假设 $W=(W,A\cup\{d\})$ 是一个决策表，包括训练集和测试集对象。$A=(U,A\cup\{d\})$ 是 W 的一个子表，$u_i\in W$ 是一个新对象，$Rul(X_j,u_i)$ 是一组决策规则，v_d^j 表示决策值，$Pred(r)$ 是规则 r 的条件属性，$MRul(X_j,u_i)\subseteq Rul(X_j)$ 表示类 X_j 中匹配新对象 u_i 的所有规则[4]。

$$SimpleStrength(X_j,u_i)=\frac{card(MRul(X_j,u_i))}{card(Rul(X_j))} \tag{6}$$

$$BasicStrength(X_j,u_i)=\frac{\sum_{r\in MRul(X_j,u_i)} Supp_A(r)}{\sum_{r\in Rul(X_j)} Supp_A(r)} \tag{7}$$

$$MaximalStrength(X_j,u_i)=\max_{r\in MRul(X_j,u_i)}\left\{\frac{Supp_A(r)}{card(d=v_d^j|_A)}\right\} \tag{8}$$

$$GlobaslStrength(X_j,u_i)=\frac{card\left(\bigcup_{r\in MRul(X_j,u_i)}|Pred(r)|_A\cap|d=v_d^j|_A\right)}{card(|d=v_d^j|_A)} \tag{9}$$

Bazan 在上述公式基础上，进一步提出了考虑动态系数的完全匹配规则的优选方法[4]。由于动态系数的计算必须与其规则推导方法相配套，非常复杂，不具有通用性，本文不予以讨论。实验发现，式(6)和式(9)的效果明显不如式(6)和式(7)，式(6)和公(7)效果比较接近。因此，本文只讨论如何进一步优化式(7)和式(8)。

完全匹配分为两种类型：第一种，新对象所有属性与一个或多个决策规则完全匹配，也就是决策规则和新对象的属性及属性值完全相同。在这种情况下，新对象可以确定地归入预

定义的类别中。如果系统规定是一元分类，还必须确定最佳的类别，匹配规则的优选可用式(5)～(9)。第二种，决策规则与新对象完全匹配，但是包含一个或多个新对象没有的属性。如果仍然应用式(5)～(9)进行新对象类别预测，分类准确率会比较低。在这种情况下，有两个因素影响分类准确性：匹配规则中与新对象不匹配的属性数量和匹配规则中与新对象不匹配的属性相对决策属性的依赖性度的大小。如果匹配规则与新对象不匹配的属性越多或者匹配规则中与新对象不匹配的属性相对决策属性的依赖度越大，那么新对象与该类的相关性就越低。

假设 k 是决策规则中某个条件属性相对于决策属性的依赖度(参见第二部分)，n 是决策规则与新对象不相匹配的条件属性的个数，$k_i(r)$表示匹配规则中第 i 个与对象不匹配的属性的依赖性系数，那么式(7)和式(8)可以重新定义为：

$$BasicStren\ gth(X_j,u_t)=\frac{\sum_{r\in M\,Rul(X_j,u_t)} supp\ A^{(r)}\times\prod_{i=1}^{n}[1-K_i(r)]}{\sum_{r\in Rul(x_j)}Supp_A(r)} \tag{10}$$

$$MaximalStrength(X_j,u_t)=\max_{r\in MRul(X_j,u_t)}\left\{\frac{Supp_A(r)\times\prod_{i=1}^{n}[1-K_i(r)]}{card(d=v_d^t|_A)}\right\} \tag{11}$$

3.2 部分匹配(partial matching)

Grzymala-Busse 认为部分匹配应该是找出与新对象匹配属性数量最多的规则。其基本原理是逐渐减少新对象的属性个数，直到有一条规则与之完全相匹配为止。换句话说，就是将部分匹配转换为完全匹配[12]。实际上，新对象的属性可能与一个类里的多条规则部分匹配。因此，我们认为，部分匹配必须从两方面考虑：一是新对象与规则匹配的属性数量；二是匹配规则与新对象不相匹配的属性数量。

假设 L 为一条决策规则和新对象匹配的属性个数，n 为规则中与该对象不相匹配的属性个数，部分匹配公式定义为：

$$BasicStrength(X_j,u_t)=\frac{\sum_{r\in M\,Rul(x_j,u_t)}\left\{\sum_{i=1}^{L}K_i(r)\times Supp_A(r)\times\prod_{i=1}^{n}[1-K_i(r)]\right\}}{\sum_{r\in Rul(X_j)}Supp_A(r)} \tag{12}$$

$$MaximalStrength(X_j,u_t)=\max_{r\in M\,Rul(x_j,u_t)}\left\{\frac{Supp_A(r)\times\sum_{i=1}^{L}K_i(r)\times\prod_{i=1}^{n}[1-K_i(r)]}{card(|d=v_d^t|_A)}\right\} \tag{13}$$

3.3 实验

实验仍以上述分类系统为例，表7和表8分别为在标准决策规则和近似决策规则上采用完全匹配和部分匹配方法的效果，规则匹配选用式(7)、式(10)和式(12)。

表7 标准决策规则＋完全匹配＋部分匹配的检索性能评估

类目	01	02	03	04	05	07	08	10	11	12	13	14
查准率	0.49	0.41	0.59	0.70	0.73	0.48	0.63	0.66	0.52	0.67	0.65	0.52
查全率	0.33	0.42	0.33	0.33	0.44	0.37	0.31	0.22	0.31	0.29	0.25	0.36

表8 近似决策规则＋完全匹配＋部分匹配的检索性能评估

类目	01	02	03	04	05	07	08	10	11	12	13	14
查准率	0.71	0.65	0.79	0.86	0.83	0.77	0.79	0.87	0.77	0.81	0.86	0.78
查全率	0.50	0.57	0.49	0.54	0.60	0.57	0.40	0.39	0.42	0.45	0.44	0.55

以上数据表明，部分匹配可以进一步提高规则对新对象的分类能力。因此，在实际分类系统中，建议将近似规则算法和部分匹配方法同时使用。

结论 对于基于粗糙集理论的自动分类系统，人们普遍关心的问题只是如何推导出标准决策规则。事实上，如何运用规则预测新对象也是影响分类系统性能的关键因素。本文在详细讨论现有近似决策规则生成算法和规则匹配方法的基础上，提出了更有效的近似决策规则生成算法和规则匹配方法。从理论上讲，本文提出的近似决策规则算法和规则匹配方法具有普遍适用性，可以应用于任何粗糙集分类算法中。今后的研究工作将集中在两个方面：一是应用多种类型数据集进行实验，进一步验证它们的通用性；二是进一步优化部分匹配方法，降低分类错误率。

致谢 感谢北京大学李晓明教授、冯是聪博士和北京大学网络实验室免费提供 CWT 实验数据集。

参考文献

1 张文修，等. 粗糙集理论与方法. 北京：科学出版社，2001

2 Kusiak A. Rough Set Theory. A data mining tool for semiconductor manufacturing. IEEE Transactions on Electronics Packaging Manufacturing，2001，42(1)：44～50

3 Grzymala-Busse J W. A new version fo the rule induction system LERS. Fundamenta Informaticae，1997，31：27～39

4 Bazan J，Nguyen H S，Synak P，Wróblewski J. Rough set algorithms in classification problems. In：Rough set methods and applications. New Developments in Knowledge Discovery in Information Systems，Studies in Fuzziness and Soft Computing. Heidelber，Physica verlag，2000. 49～88

5 Pawlak Z. Rough sets. International Journal of Computer Information Sciences，1982，11(5)：341～356

6 Pawlak Z. Rough sets theoretical aspects of reasoning about data. Netherlands：Kluwer academic publishers，1991

7 Polkowski L. Rough sets Mathematical foundations. Heidelberg：Physica-Verlag，2002

8 Komorowski J，Pawlak Z，Polkowski L，Skowron A. Rough Sets：A Tutorial，Rough Fuzzy Hybridization. Springer-Verlagpage，1999. 3～98

9 Booker L B，Goldberg D E，Holland J F. Classifier systems and genetic algorithms，Machine Learning，Paradigms and Methods. MIT Press，1990. 235～282

10 Holland J H，Holyoak K J，Nisbett R E. Induction，Processes of Inference，Learning and Discovery. MIT Press，1986

11 Gyzymala-Busse J W. LERS- a system for learning from examples based on rough sets. Intelligent Decision Support ， Kluwer Academic Publishers，1992. 3～18

12 Grzymala-Busse J W. Preterm birth risk assessed by a new method of classification using selective partila matching. In：Proc. of the Eleventh Intl. Symposium on Methodologies for Intelligent System Springer Verlag，1999. 612～620