一种基于粗糙集增量式规则学习的问题分类方法研究

李 鹏 王晓龙 关 毅
(哈尔滨工业大学计算机学院 哈尔滨 150001)

摘 要：该文提出一种基于粗糙集增量式规则自动学习来实现问题分类的方法，通过深入提取问句特征并采用决策表形式构建训练语料，利用机器学习的方法自动获取分类规则。与其他方法相比优势在于，用于分类的规则自动生成，并采用粗糙集理论的简约方法获得优化的最小规则集；首次在问题分类中引入增量式学习理念，不但提高了分类精度，而且避免了繁琐的重新训练过程，大大提高了学习速度，并且提高了分类的可扩展性和适应性。对比实验表明，该方法分类精度高，适应性好。在国际 TREC2005 Q/A 实际评测中表现良好。

关键词： 粗糙集；问题分类；增量式学习；决策表；特征选择
中图分类号：TP391.6 **文献标识码**：A **文章编号**：1009-5896(2008)05-1127-04

Question Classification with Incremental Rule Learning Algorithm Based on Rough Set

Li Peng Wang Xiao-long Guan Yi
(*School of Computer Science and Technology*, *Harbin Institute of Technology*, *Harbin* 150001, *China*)

Abstract: This paper presents a method on automatic question classification through incremental rule learning based on rough set theory. The core of the method is appling the machine learning approach to gain classified rules automatically through extract the features of query sentence thoroughly, and the decision table is used to construct the training collection. Comparing with the alternative means, the superiority is that it acquires the classified rule automatically and uses the rough set method to obtain the optimized smallest rule set. Especially, the incremental learning is induced to improve the precision and avoid the tedious re-training process. The performance of the approach is promising, when tested on opposite test. Meanwhile, the method obtains a very good result in the international TREC2005 Q/A track.

Key words: Rough set; Question classification; Incremental learning; Decision table; Feature selection

1 引言

面向开放域的问答系统是近年来信息处理领域研究的热点，是下一代搜索技术的发展方向[1]。通常一个完整的问答系统包含问题分类、检索和答案抽取 3 个主要组成部分。因此，快速、高准确率的问题分类技术是问答系统实现的基础，它的准确性将直接影响到最终抽取答案的准确性[2]。问题分类的主要任务是将用户提出的问题进行语义类别上的分类。问题分类所含特征信息来源于用户问句，所以特征比较少，这样给问题分类带来了很大的难度。问题分类的分类体系也需要根据问答系统的整体需求进行变化，这使得问题分类具有很高的灵活性，需要良好的扩展性和适应性。因此，研究具有高准确率和适应性的问题分类方法对分类技术的研究具有重要的现实意义。

问题分类最早采用的是基于手工规则的方法进行分类[3]，这种分类方法的适应性和扩展性很差，而且费时费力。因此，近年来基于统计学习的方法成为人们解决这一问题的主要途径。目前，已经应用于问题分类的统计学习方法有简单贝叶斯[4]，核方法[5]，SnoW[6]以及支持向量机[7]等。这些研究结果都表明统计学习方法在问题分类应用上有十分优秀的表现。但遗憾的是，所有这些方法都把训练语料作为一种静态数据来处理，一旦改变分类体系，或补充新的训练语料，都需要重新进行训练，这极大地增加了系统的运行代价和时间，也使得系统适应性逐渐下降。因此，本文首次在利用统计学习方法解决问题分类过程中引入了增量式学习的理念，把训练语料作为一种动态数据，由此解决了重训练问题，提高了分类的适应性。

本文的第 2 节介绍问句分析与决策表知识表达；第 3 节介绍基于粗糙集理论的规则生成与属性最小约简算法；并且介绍基于粗糙集的增量式规则生成算法；第 4 节进行了对比试验及性能分析，并且介绍了参加 TREC2005 QA 评测的情况；最后为结论。

2 问句分析与决策表知识表示

2.1 问句分析与特征提取

问题分类体系通常有两种方式，即平行分类和层次分

2006-10-30 收到，2007-05-21 改回
国家自然科学基金重点项目(60435020)、国家自然科学基金项目(60504021)和国家 863 目标导向类课题(2006AA01Z197)资助课题

类。平行分类体系是一种粗分类方法，这种分类方法将要处理的问题分成平等的若干类别，问题分类如果采用平行分类往往会有比较高的分类精度，但这种分类体系会为后续答案抽取部分增加技术难度。层次分类体系要求在粗分类的基础上对每一个类别进一步逐层细分类，随着层次的增加难度也不断加大。层次分类也是问题分类研究的发展方向以及分类技术研究的一个难点。

特征提取是采用统计机器学习方法解决分类问题中至关重要的一个部分。问题分类所面向的处理对象是相对简短的问句，问句中通常包含较少的词，因此所含特征信息也就相对较少。如果要使分类的准确性得以提高就需要从简短的问句中尽可能多地提取对分类有帮助的信息。本文将每一种特征都用一种属性标记来表示，如疑问词[Q-QW]、关键词[Q-KW]、命名实体[Q-NE]、基本名词短语[Q-BNP] 等等。针对每一个具体的问句，每一种属性会对应具体的值。如下面例子中表现的一些属性特征：

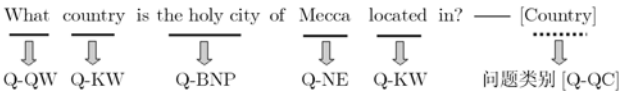


图 1 问句特征提取

特别指出的是，还有一种缺省属性 NULL，即当某一属性的值没有时就用它来补齐，这主要是为了使后面的决策表成为一个完备信息表，减少处理上的复杂程度。

2.2 决策表知识表达

规则获取就是要从大量原始数据中分析发现有用的规律信息，即是将知识从一种原来的表达形式(原始数据表达形式)转换为一种新的目标表达形式(人类或者计算机便于处理的形式，如逻辑规则等)[8]。基于粗糙集理论的知识发现，主要是借助于信息表这样一种有效的数据表知识表达方式。决策表是一种特殊的信息表，可以用来生成决策规则并用于解决分类问题。决策表的定义如下：

定义 1 一个决策表是一个信息表知识表达系统 $S=\langle U,\ R,\ V,\ f\rangle$，这里，$U$ 是对象的集合，也称为论域，$R=C\cup D$ 是属性集合，子集 C 和 D 分别称为条件属性集和结果属性集，$D\neq\varnothing$。$V=\bigcup_{r\in R}V_r$ 是属性值的集合。V_r 表示属性 $r\in R$ 的属性值范围。$f:U\times R\to V$ 是一个信息函数，它指定 U 中每一个对象 x 的属性值。

可以通过表 1 这个典型的决策表说明图，来进一步认识决策表的构成以及在获取决策规则中所起的作用。论域 U 其实就是所要处理对象的样本集合；条件属性 C 就是能决定分类的分类属性，也就是特征；结果属性 D 就是最终要分类的类别集合，也就是决策属性。本文的训练样本都可以转化成这种决策表信息表达形式。这里，为了适应粗糙集方法处理的要求并使程序上能执行的更加方便，一些原始属性值已经进行量化，被离散值所取代了；决策属性也同样被量化为离散数，有多少种分类类别就有多少个数，一旦有所变化，只需增加或删减数字即可。

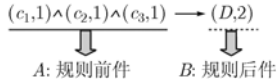
表 1 问题分类样本的决策信息表

U (论域)	C(条件属性)				D (决策属性)
	C_1(Q-QW)	C_2(Q-NE)	C_3(Q-BNP)	…	
x_1	1	2	1	…	2
x_2	5	4	0	…	5
x_3	3	1	1	…	3
⋮	⋮	⋮	⋮	…	⋮

决策表中的每一行就代表一条决策规则，下面我们给出有关决策规则的一些定义，表明如何从决策表中生成决策规则。

定义 2 公式 $A\to B$ 的逻辑含义称为决策规则，A 称为规则前件，B 称为规则后件，它们表达一种因果关系。其中，公式 A 中所包含的原子公式中只有决策表中的条件属性，B 中所包含的原子公式中只有决策表中的决策属性。

如表 1 所示的决策表，条件属性集合为 $\{c_1,c_2,c_3,\cdots\}$，决策属性为 D，则可以生成如下的决策规则：



3 基于粗糙集理论的增量式最小规则集自动获取

决策表中的一个样本就代表一条基本决策规则，如果把所有这样的决策规则罗列出来，就可以得到一个决策规则集合。但是，这样的决策规则集合只是机械地记录了一个个样本的情况，没有得到优化，不能适应情况的变化。为了从决策表中抽取得到适应度大的规则，需要对决策表进行约简，使得经过约简优化处理的决策表中的一个记录就代表一类具有相同规律特性的样本，这样得到的决策规则就具有较高的适应性。每一条决策规则的不确定性可以用可信度来度量，它用来表示利用该规则得到正确结论的概率估计。可信度的定义和计算方法如下：

定义 3 对于决策表 $S=\langle U,\ R,\ V,\ f\rangle$，$R=C\cup D$ 是属性集合，子集 C 和 D 分别称为条件属性集和结果属性集，$D\neq\varnothing$。决策规则 $A\to B$ 的可信度 $CF(A\to B)$ 定义为

$$CF(A\to B)=\frac{|X\cap Y|}{X}$$

其中集合 X 是条件属性值满足公式 A 的实例集合，集合 Y 是决策属性值满足公式 B 的实例集合。

基于粗糙集理论的增量式最小规则集的获取过程主要是在保持决策表条件属性和决策属性之间的依赖关系不变的前提下对决策表进行约简，也就是对从决策表中生成的原始规则集进行优化，再经过设置决策规则可信度的阈值对低

可信度的规则进行过滤，如果有新的训练数据加入则采用增量式学习策略，最终获得最小规则集合。

3.1 属性约简算法

原始的决策表中的条件属性并不是同等重要的，甚至其中某些条件属性是冗余的。这些冗余属性的存在，一方面是对资源的浪费；另一方面，也干扰人们做出正确简洁的决策。因此，决策表的属性约简，就是要在保持条件属性相对于决策属性的分类能力不变的条件下，删除其中不必要的或不重要的属性。属性约简问题是一个 NP 难题，有一些针对属性约简的专题性研究[9]，本文采用比较经典的基于特征选择的属性约简算法，其步骤如下：

输入　决策表 $S=\langle U,\ C\bigcup D,\ V,\ f\rangle$

输出　属性约简后的条件属性集

第 1 步　计算条件属性集 C 和决策属性 D 之间的相关程度 $K_\beta(C,D)$；

第 2 步　$\mathrm{REDU}=C$；

第 3 步　While $K_\beta(C,D)=K_\beta(\mathrm{REDU},D)$ Do

(1)计算 REDU 中所有属性的上、下文价值 CM；

(2)根据上、下文价值对 REDU 中的属性进行排序；

(3)选择属性 α_j，α_j 具有最小是上、下文价值，且 $K_\beta(\mathrm{REDU},D)=K_\beta(\mathrm{REDU}\setminus\{\alpha_j\},D)$；

(4) $\mathrm{REDU}\leftarrow\mathrm{REDU}\setminus\{\alpha_j\}$；

(5)End while;

第 4 步　输出　REDU。

3.2 增量式规则学习算法

通常在统计学习方法中，当有新的数据到来时，每次都需要把新数据和原始数据放在一起，重新学习来得到新的规则集。当训练数据不断增加而变的十分庞大时，每次新增加数据都需要重新进行训练，这无疑大量地浪费了训练时间和资源，使得系统的适应性变得越来越差。为了解决这个问题，引入了增量式学习思想。增量式学习是模拟人脑学习过程的一种方法，当新的信息到来时，只对新的信息进行学习，然后用得到的新知识来修改原有的知识，它并不是把新的信息和原来的信息合并起来重新进行学习。针对增量式学习方法的专题性研究一直以来是研究者所关注的热点，本文采用一种简单的增量式学习方法获取分类规则[10]。

利用先前得到的知识来对新样本进行分类是一种动态学习[11]。Pawlak 认为加入一个新的样本有 3 种情况：(1)新样本与原有知识相同(新样本已经出现过)；(2)新样本与原有知识矛盾(新样本的规则前件在已有样本中出现，但它们的规则后件不同)；(3)新样本完全是新情况(新样本不属于已知的任何决策类)；针对新样本的不同情况，我们采用的增量式学习策略如图 2 所示：

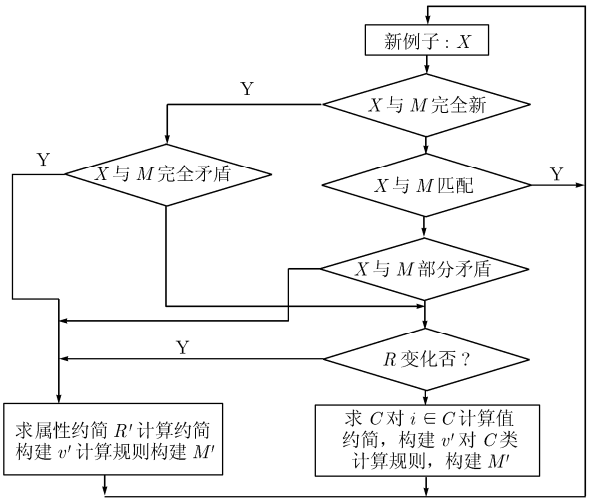


图 2 增量式规则学习算法

4　实验与性能分析

4.1 对比实验与性能分析

为了验证和分析本方法在问题分类上应用的效果，与文献[6]和文献[7]的方法进行了对比实验。在对比实验中采用了与它们相同的训练数据、测试集和分类体系。训练集中包括 5500 个标注好的问题实例，测试集为 TREC10 Q/A 评测的 500 个问题实例，分类采用两层分类体系，其中第 1 层包括 6 个粗类，第 2 层包括 50 个细类。具体结果如表 2 所示。

表 2 对比实验结果

	文献[7]		文献[6]	本文方法
	Word 特征	Ngram 特征		
粗类准确率 (%)	85.8	87.4	84.20	86.80
细类准确率 (%)	80.2	79.2	84.00	79.60

从上面的表中可以看到，本文方法粗分类的准确率达到了 86.80 %，细分类的准确率达到了 79.60 %。这样的结果与对比参照的两种方法取得的最好结果基本相当，但在这两篇对比文献中，研究者都在人为不断地考查选取各种不同的特征时分类的效果，表 2 所示的是经过反复试验后最好的结果；本文的特征是通过属性约简自动优化得来的，并不存在人工的试验性筛选。文献[7]同时也公布了其他统计学习算法在同一实验条件下的分类结果。图 3 和图 4 显示了两篇文章中所涉及到的其他方法或不同特征下所取得的结果，相比之下可以看到本文方法得到了比较优异的结果。

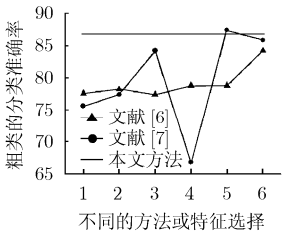


图 3 粗分类准确率对比图

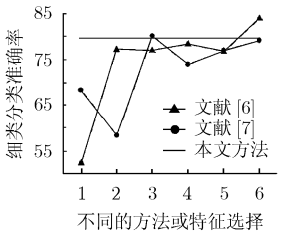


图 4 细分类准确率对比图

4.2 实际参加 TREC 评测结果

文本检索会议(Text REtrieval Conference, TREC)是由美国国家标准技术局(NIST)和国防部高级研究计划局(DARPA)组织召开的全球信息检索领域最具有权威性的评测会议。应用本文方法的问答系统参加了 TREC2005 Q/A 任务的评测(图 5)，在全世界总共 71 个参评的系统中，获得了 FACTOID 单项第五、LIST 单项第七，以及总分第八的好成绩[12]。其中，这两个项目的任务都需要进行问题分类，本文所阐述的方法正式用于这次的评测工作，为了服务于整个问答系统的实际需要，我们采用了自己制定的分类体系，问题分类的准确率达到了 92.54 %。

Run Tag	Submitter	Accuracy	NIL Prec	NIL Recall
lcc05	Language Computer Corp.	0.713	0.643	0.529
NUSCHUA1	National Univ. of Singapore	0.666	0.148	0.529
IBM05L3P	IBM T.J. Watson Research	0.326	0.200	0.118
ILQUA2	Univ. of Albany	0.309	0.075	0.235
Insun05QA1	Harbin Inst. of Technology	0.293	0.057	0.176
csail2	MIT	0.273	0.098	0.294
FDUQA14B	Fudan University	0.260	0.082	0.412
QACTIS05v2	National Security Agency (NSA)	0.257	0.045	0.176
mk2005qar2	Saarland University	0.235	0.071	0.353
Edin2005b	Univ. of Edinburgh	0.215	0.068	0.176

图 5 TREC2005 Q/A Factoid 项目评测结果

5 结束语

本文提出的问题分类方法，通过基于粗糙集理论所支持的多个知识获取步骤(如：数据预处理、属性约简、规则生成、数据依赖关系获取等)，实现了问题分类规则的自动生成与优化，避免了大量手工整理规则的劳动以及人为选择特征的主观干扰，具有分类精度和自动化程度高的特点。在采用统计机器学习解决问题分类的过程中，创造性地引入了增量式学习的理念，解决了新增样本后重训练问题。在同一实验条件下与其他方法的对比实验显示，本文方法在粗分类和细分类的分类准确率分别达到了 86.80 % 和 79.60 %。该方法实际应用于国际 TREC Q/A 评测比赛中效果良好。在本文分类体系下，分类准确率达到了 92.54 %，为问答系统取得优异成绩奠定了坚实的基础。

在今后的研究工作中，针对问题分类的研究还应该在以下几个方面进一步深入探索。首先，特征的选取应该考虑可以引入最新的语义信息，如组块信息、语义角色信息等；其次，需要研究高质量的增量式规则学习算法；最后，要加强对多层次分类体系问题分类的研究。

参 考 文 献

[1] Marius A Pasca. High-performance, open-domain question answering from large text collections. [Ph. D. dissertation], University of Southern Methodist, 2001.

[2] Cody Kwok, Oren Etzioni, and Daniel. Scaling question answering to the web [J]. *ACM Trans. on Information Systems*, 2001, 9(3): 242–262.

[3] Shaw M L G and Gaines B R. Question classification in rule-based systems [C]. Proceedings of Expert Systems'86, The 6Th Annual Technical Conference on Research and development in expert systems, Brighton, 1987: 123–131.

[4] 张宇, 刘挺. 基于改进贝叶斯模型的问题分类 [J]. 中文信息学报, 2005, 19(2): 100–105.
Zhang Yu, Liu Ting, and Wen Xu. Modified Bayesian model based question classification [J]. *Journal of Chinese Information Processing*, 2005, 19(2): 100–105.

[5] Taira Jun Suzuki, Sasaki Yutaka, and Maeda Eisaku. Question classification using HDAG kernel [C]. ACL Workshop on Mulitilingual Summarization and Question Answering, Sapporo, 2003: 61–68.

[6] Li Xin and Roth Dan. Learning question classifier [C]. In Proceedings of the 19th International Conference on Computational Linguistics (COLING'02). Taipei, 2002: 556–562.

[7] Zhang Dell and Lee Wee Sun. Question classification using support vector machines [C]. Proceedings of the 26th annual international ACM SIGIR Conference on Research and Development in Information Retrieval, New York, ACM Press, 2003: 26–32.

[8] 王国胤. Rough 集理论与知识获取[M]. 西安：西安交通大学出版社, 2001.
Wang Guo-yin. Rough Sets Theory and Knowledge Discovery[M]. Xi'an Jiaotong University Press, 2001.

[9] 王国胤, 于洪等. 基于条件信息熵的决策表约简 [J]. 计算机学报, 2002, 25(7): 759–766.
Wang Guo-yin and Yu Hong. Decision table reduction based on conditional information entropy [J]. *Chinese J Computer*, 2002, 25(7): 759–766.

[10] 于洪，杨大春，吴中福. 基于 Rough set 理论的增量式规则获取算法[J]. 小型微型计算机系统, 2005, 26(1): 36–41.
Yu Hong, Yang Da-chun, and Wu Zhong-fu. Incremental rule acquisition algorithm based on rough set [J]. *Mini-Micro Systems*, 2005, 26(1): 36–41.

[11] Pawlak Z. Rough set: theoretical aspects and reasoning about data [M]. Dordrecht, Kluwer Academic Publishers, 1991.

[12] Ellen M. Voorhees, and Hoa Trang Dang. Overview of the TREC 2005 question answering Track [C]. The Fourteenth Text REtrieval Conference (TREC 2005) Proceedings. New York, 2005: 1–15.

李 鹏： 男，1978 年生，博士生，研究方向为机器学习、自然语言处理、问答系统、网络信息处理.

王晓龙： 男，1955 年生，教授，博士生导师，研究方向为人工智能、机器学习、计算语言学和中文信息处理.

关 毅： 男，1970 年生，教授，研究方向为问答系统、统计语言处理、机器学.