第32卷 第24期 计 算 机 工 程 2006年12月
Vol.32 No.24 Computer Engineering December 2006
·人工智能及识别技术· 文章编号：1000—3428(2006)24—0162—02 文献标识码：A 中图分类号：TP18

基于商空间模型的CBR系统

赵 鹏[1,2]，蔡庆生[1]，耿焕同[1]，于 琨[1]
（1. 中国科学技术大学计算机科学与技术系，合肥 230026；2. 安徽大学计算机科学系，合肥 230039）

摘 要：传统的CBR系统采用平面结构，系统在运行过程中不断学习，范例库将变得越来越大，当范例数超过某一预设的上界时，就会出现“沼泽问题”。为了解决这个问题，该文提出了基于商空间模型的CBR系统，采用分层递阶的立体结构，在运行阶段将惰性学习算法与积极学习算法相结合。实验表明利用本方法构造的CBR系统实现E-mail分类预测时，系统的性能和有效性都得到了很大的提高。

关键词：商空间理论；信息粒度；分层递阶结构；范例推理

CBR System Based on Quotient Space Model

ZHAO Peng[1,2], CAI Qingsheng[1], GENG Huantong[1], YUN Kun[1]
(1. Department of Computer Science and Technology, USTC, Hefei 230026; 2. Department of Computer Science, Anhui University, Hefei 230039)

【Abstract】 Traditional CBR system is planar architecture. With the system's running and learning, case base will become larger and larger. When the numbers of cases surpass some boundary, there will be swarming problem. To settle this problem, this paper presents a CBR system based on quotient space model, which is hierarchical architecture. It combines active learning with lazy learning in system running phase. Experimental results show that based on this method, the system performance is greatly increased.

【Key words】 Quotient space theory; Information granularity; Hierarchy; Case-based reasoning

1 概述

范例推理(Case-based Reasoning, CBR)最早由R. Schank教授提出[1]，是根据目标范例的提示而得到历史记忆中的源范例，并由源范例来指导目标范例求解的一种策略。CBR具有信息的完全表达、增量式学习、形象思维的准确模拟、知识获取较为容易、求解效率高等特点，能够有效地克服传统知识处理系统对边界以外的知识处理十分低效、匹配冲突、组合爆炸、难以解释、自学习困难等缺陷，非常适用于那些没有很强的理论模型和领域知识不完全、难以定义、不良定义或定义不一致而经验丰富的决策环境中[2]。CBR系统所依赖的最重要的知识存储在范例中，范例的集合组成了范例库。范例的知识表示、范例库的结构以及范例的索引决定了CBR系统的推理效果。传统的CBR系统将所有的范例按照统一标准存放在一个范例库中，随着系统的运行，系统增量式学习会使范例库不断地增大，检索时间越来越长，影响系统的运行效率，这就是所谓的“沼泽问题(Swamping Problem)”[3]。传统的CBR系统所采用的学习方法属于惰性学习(Lazy Learning)，即没有真正意义的训练学习阶段，不会得到关于数据的模型，它将所有的计算工作都推延到检索阶段，当处理海量数据时，存储和检索它们的代价是很大的。

为了解决传统CBR中存在的上述问题，本文提出了基于商空间模型的CBR系统，将原始范例库划分为粒度适中的范例库商集，对范例库商集中的每个元素构造子范例库，并可以在此基础上，继续求其商集，形成分层递阶的立体结构。在测试阶段，CBR、KNN等惰性学习算法不使用原始范例，而直接访问范例库商集。这种将惰性学习与积极学习结合起来，从原始数据中得出粒度适中的直接访问范例库商集。这种将惰性学习与积极学习结合起来的若干小模型是关于部分原始数据的，所有的小模型反映了整个数据。模型是层次性的，粒度越小的模型概括的数据越少，精度越高；粒度越大的模型概括的数据越多，精度越低；粒度小与粒度大的模型之间是层次关系。根据商空间理论的同态原则[4]，通过适当的分层，可以从高层次的分析中，将一些问题无解的范围删去，缩小求解空间，加快求解速度。实验表明利用本方法构造的CBR系统实现E-mail分类预测时，系统的性能和有效性都得到了很大的提高。

2 粒度和商空间模型

粒度(granularity)本是一个物理概念，被借用来做“信息粗细的平均度量”，也称信息粒度。所谓信息粒度，是指人类在解决和处理大量复杂信息问题时，由于能力有限，需要把大量复杂信息按各自的特征和性能将其划分成数个较简单的信息块，以方便处理，每个如此划分的信息块就被认为是一个粒度[5]。

商空间模型将不同的粒度世界与数学上的商集概念统一起来，表示对象模型的方法，即以商集作为不同粒度世界的数学模型的方法。

设用三元组(X，f，T)描述一个问题[4,6,7]，其中X表示问题的论域，f(.)表示论域的属性，T是论域的结构(即论域X中各元素的相互关系)。分析或求解问题(X，f，T)，是指对论域X及其有关的结构、属性进行分析、研究。对论域X，在其上给定一个等价关系R，对应R的商集为[X]。然后将[X]当作新的论域，对它进行分析、研究。商集是将等价类看作新元

基金项目：国家自然科学基金资助项目(70171052)；安徽省自然科学基金资助项目(2004kj011)；安徽省高等学校青年教师基金资助项目(2006jq1040)
作者简介：赵 鹏(1976－)，女，博士生、讲师，主研方向：人工智能，机器学习；蔡庆生，教授、博导；耿焕同、于 琨，博士
收稿日期：2006-01-04 **E-mail：**zhp2004@mail.ustc.edu.cn

素而构成的新空间，这就得到一个较粗的粒度世界[X]。问题的不同粒度表示对应于不同的等价关系R。可以看出不同的粒度是对论域的不同的划分。

文献[4]中论述了论域 X 与[X]的关系，证明了不同粒度的论域形成完备半序格。

命题 1 p：(X,T)→([X],[T])是自然投影，p 是连续的。若 A⊂X 且 A 是 X 中的连通集，则 p(A)是[X]中的连通集。

命题 1 表明，如果一个问题在原论域 X 中有解(是连通的)，在适当的粗粒度论域[X]上也有解。反之，如果粗粒度论域上无解，则原问题必无解(不连通)。这个性质说明，商空间变换的保假的特点。

命题 2 设(X,T)是半序空间(或是拟半序空间)，R 是相容的，若 x,y∈(X,T)且 x<y，则[x]<[y]，其中[x]，[y] ∈([X],[T])。

命题 2 表明，如果原论域 X 本身很复杂，可在 X 上引入一个分类 R，得[X]。若 R 与 X 的结构 T 是相容的，则在[X]上诱导出一个商半序[T]。于是就将原来求 x 到 y 的问题转化为在[X]中求[x]到[y]的问题。由于 R 是相容的，则 p：(X,T)→([X],[T])是保序的。也就是，当利用适当的分类技术在粗粒度世界讨论问题时，如果问题无解，那么在细粒度的原问题上也无解，由于粗粒度世界通常比原粒度世界简单，这样就缩小了求解的范围，加快了求解的进度。

3 基于商空间模型的 CBR 系统

传统的 CBR 系统将所有的范例按照统一格式存放在一个范例库中，内存开销很大，并且存在大量冗余信息和无用信息。当对新范例预测类别时，经典的 KNN 算法必须访问所有的范例，算法的复杂性高。为了构造更为合理、高效的 CBR 系统，我们提出了基于商空间模型的 CBR 系统。

仿照商空间模型，用三元组(CB,f,T)描述一个范例库，其中 CB 表示原始范例集，即问题的论域，f(.)表示范例的属性，T 是范例集的结构(即范例集 CB 中各范例的相互关系)。对于论域 CB，按照一定的划分标准，即给定一个等价关系 R，对应 R 的商集为[CB]，即为子范例库集。然后将[CB]当作新的论域，对它进行分析、研究。依此下去，直到得到合适的粒度。这样就将 CBR 系统由原先平面结构转换为基于商空间模型的分层递阶的立体结构。

此外，要使得分层递阶的结构和方法带来系统的高效率，还必须从f构造出适当的[f]。对于商集[CB]中的每一个元素，即每一个子范例库i，对其属性进行分析，按照一定的标准(同领域的具体知识有关)，选择其中最有代表性的n个属性a_1，a_2，…，a_n，构成属性集合A_i(i=1,2,…,m)，m为类别数，所有的A_i构成集合空间A，由此得到[f]，记为[f]：[X]→A。将每一个类别i上的范例在A_i上的投影作为子范例存入子范例库中。然后再对商集[CB]中的每一个元素进行抽象形成超范例，存入超范例库中。当对新范例检索近似范例、预测类别时，可以先扫描超范例库，初步确定新范例可能隶属的类别，然后再到相应的子范例库中做进一步的检索，这样就大大减少了范例检索的空间，从而提高了检索的效率。

下面给出基于商空间模型的 CBR 系统的构造算法和检索算法。

算法 1 基于商空间模型的 CBR 系统的构造算法。

输入 原始范例库(CB，f，T)，等价关系 R，属性选择标准 S。

输出 范例库商空间([CB]，[f]，[T])。

Step1 扫描 CB，按照等价关系 R 划分，形成[CB]；

Step2 对于[CB]中的每一个元素：

(1)按照属性选择标准S，选择最有代表性的n个属性a_1, a_2,…, a_n，将所属类别上的每一个范例在其上的投影作为子范例存入子范例库中；

(2)对于子范例库进行抽象，形成超范例，存入超范例库 SCB。

算法 2 基于商空间模型的 CBR 系统的检索算法。

输入 范例库商空间([CB]，[f]，[T])，SCB，新范例，参数 N。

输出 最近似的 N 个范例集合 RESULT。

Step1 i=0，RESULT=Φ；

Step2 扫描 SCB，使用 KNN 算法检索到最近似的 N 个超范例，将其按相似度由大到小依次存入 TEMP 中；

Step3 对于 TEMP 中的每一个超范例依次做如下操作：

(1)扫描当前超范例所指向的子范例库，使用 KNN 算法检索到最近似的 N 个范例，将其中相似度大于阈值θ的最大的 k(k<=N-i)个范例并入 RESULT 中，i←i+k；

(2)如果 i=N，则停止，否则取下一个超范例，转 step3。

4 实验结果与分析

我们构造了一个CBR原型系统，其中包含了 1 800 封来源于DELL技术服务论坛的E-mail。每一封E-mail作为一条范例，实验采用论坛已有的分类标准，即根据E-mail中所涉及到的硬件类型划分为 12 类，形成范例库商集。实验首先对每一封E-mail做预处理(包括将标题和内容中文字做去停用词和低频词处理)，通过对训练集中E-mail的统计，得到 3 420 个单词作为文本特征，然后将每一封E-mail表示为文本特征向量，作为原始范例；对于范例库商集中的每一个元素，即子类i，计算该类所有词的可分辨度$D_i(w)$(式 1)，$D_i(w)$表示词w对于将范例预测为i类的贡献大小。取$D_i(w)$值最大的前 50 个词作为类别i的特征属性。将类i的每一个原始范例在特征属性上进行投影，并将相应的词频TF(式 2)作为属性值，形成子范例存入子范例库i中。而这 50 个特征属性D_i值及附加类别信息构成超范例存入超范例库中。

$$D_i(w)=\frac{N_i}{\sum_{j\neq i}N_j} \tag{1}$$

其中，N_i为i类中出现词w的E-mail个数。

$$TF(w)=\frac{T_w}{T_E} \tag{2}$$

其中，T_w为词w在E-mail中出现的次数，T_E为E-mai l中的总词数。

为了测试基于商空间模型的CBR系统的运行效果，分别采用基于词频特征表示方法的平面结构和基于商空间模型的立体结构两种方法来构造CBR系统，并且给出了对比实验的结果。实验使用F-score(式(3))作为基本测试指标测试分类综合性能，并使用宏平均(Macro-averaging)[8]计算所有类别的平均F-score(式(4))，以及范例检索所需要的时间T。

$$F-score=\frac{2\times 精确率\times 召回率}{精确率+召回率} \tag{3}$$

其中，精确率=(分类正确的 E-mail 个数/分为该类别的所有 E-mail 个数)×100%；召回率=(分类正确的 E-mail 个数/属于该类别所有 E-mail 个数)×100%。

$$平均\ F-score=\frac{\sum_{i=1}^{N}第i类F-score}{N} \tag{4}$$

实验分别使用了经典的 KNN 算法和第 3 节中算法 2 提出的检索算法作为范例的检索算法，并根据检索到最近似范

(下转第 166 页)