基于聚类策略的一种范例删除模型*)

耿焕同　钱　权　蔡庆生
(中国科学技术大学计算机系　合肥230026)

A Model of Case-Deletion Based on Clustering

GENG Huan-Tong　QIAN Quan　CAI Qing-Sheng
(Department of Computer Science, USTC, Hefei 230026)

Abstract　In this paper, a new Case-Deletion strategy is proposed. This method absorbs merits of clustering algorithm. It overcomes the deflection of traditional deletion strategies. The experiments show that the new algorithm can reduce cases greatly and can preserve competence of CBR system.

Keywords　CBR, Case-deletion, Clustering

1　前言

基于范例推理(Case-Based Reasoning, CBR)是由 R. Schank 教授提出的基于范例的推理方法[1]。CBR 是用先前求解问题的经验和方法，通过类比和联想来解决当前相似问题的推理方法，这种方法更好地反映专家的思维过程，兼顾了知觉、想象和经验。目前已被广泛应用于医学诊断、工业制造、法律诉讼等领域，并取得很好的效果。由于在 CBR 系统不断解决问题的过程中，越来越多的新范例被存储在范例库中，则范例库将变得越来越大，当范例数超过某一预设的上界时，则对范例检索的代价变得非常高，这种现象称为沼泽问题(swamping problem)[2]。为了解决沼泽问题，一些学者提出采用高效的并行的检索算法来解决沼泽问题；然而，这毕竟有一个极限，一旦超过了这个极限，想维持范例检索的时间为常量是不现实的。此时考虑范例删除是非常有必要的。因此范例维护(Case-Base Maintenance，简称 CRM)越来越受到研究者(如：Barry Smyth[3] & David B. Leake[4])的关注。如何在保持 CBR 解决问题的能力下，减少范例已成为当前 CBR 研究中的热点。

在 CBR 系统中，由于通过类比和范例复用来得到新问题的解；因此，相似的问题具有相似的解决方法，相似的范例解决相似的问题。而聚类分析方法[5]是对分布在 m 维空间上的一组对象 $X_1,\cdots,X_n$ 进行分析，事先给出聚类的个数 K，并任意构造 K 个初始聚类，然后通过不断的调整 K 个聚类的中心，重新修改各聚类的内容，使其相应的统计误差 $\sum_{j=1}^{k}\sum_{i=1}^{n} d(X_i,Q_j)$ 取得最小值时停止，则最终得到 K 个聚类(其中：Q_j 为聚类的中心，X_i 为 j 聚类中的元素，$d(X_i,Q_j)$ 为对象 X_i 与聚类中心对象 Q_j 之间的距离)。聚类方法使得类间的相似性尽量少，而类内的相似性尽量大。聚类学习的典型代表有 K-means 方法和 K-medoid 方法等。因此，先利用聚类方法对范例库中的范例进行聚类，使得同一聚类中的范例具有很高的相似度；在进行范例删除时，则选择同一聚类中的某些范例进行删除；基于这个思想，本文提出了一种基于聚类策略的范例删除方法。

2　范例间相似度计算

判断范例间相似程度是基于范例推理的关键，也是体现推理正确性和智能化行为的技术。下面给出基于属性-值对的范例相似度的计算公式：设 A、B 为两范例，它们的相似度 S(A,B)可用下面公式度量：

$$S(A,B)=1-\sqrt{\sum_{i=1}^{n} w_i\times d(A_i,B_i)^2}$$

其中，$\sqrt{\sum_{i=1}^{n} w_i\times d(A_i,B_i)^2}$ 为 A、B 间的欧氏距离，记 e_i 为范例的第 i 个属性字段；A_i、B_i 为范例 A、B 中属性的值，w_i 为第 i 个属性的权重，$d(A_i,B_i)$ 为两属性的规格化距离，分两种情况讨论：

1、若该属性为数值型，则 $d(A_i,B_i)=\frac{|A_i-B_i|}{ValueRange(e_i)}$；

2、若该属性为符号型，符号相同时 $d(A_i,B_i)=0$，不同时 $d(A_i,B_i)=1$。

3　范例的分类

按范例在解决新问题中的重要性，可将范例库中的范例分为4个等级[4]：核心范例(Pivotal Cases)、支持范例(Support Cases)、连接范例(Spanning Cases)、辅助范例(Auxiliary Cases)。为了更好地说明这4种等级范例，我们先定义范例的覆盖集(Coverage)和范例的可达集(Reachability)。给定一个范例库 $C=\{c_1,\cdots,c_n\}$，对于每一个范例 $c\in C$：

1) 范例 c 的覆盖集定义为：Coverage(c)$=\{c'\in C\mid solves(c,c')\}$；

2) 范例 c 的可达集定义为：Reachability(c)$=\{c'\in C\mid solves(c',c)\}$；

其中，$solves(c,c')$ 表示范例 c′能解决的问题都能由范例 c 来解决，而 $solves(c',c)$ 表示范例 c 能解决的问题都能由范例 c′来解决。简单地说，范例的覆盖集表示该范例能解决问题空间的大小；而范例的可达集表示该集合的任意范例都能取代该范例解决问题，若某范例的可达集越大，表示可替换该范例解决问题的范例就越多，本文就用这一特性来对范例进行删除。

*)本文得到国家自然科学基金项目资助(No. 60075015, No. 90104030)。耿焕同　硕士研究生，主要研究领域为人工智能、知识发现。钱　权　博士生，主要研究领域为人工智能、网络安全。蔡庆生　教授，博导，主要研究领域为人工智能、机器学习、知识发现。

下面给出4类范例的定义：

1)如果一个范例c为核心范例(Pivotal Cases)，当且仅当 $Reachability(c)-\{c\}=\phi$；

2)如果一个范例c为支持范例(Support Cases)，当且仅当

$\neg Pivotal(c) \wedge Coverage(c) \wedge \bigcup_{c' \in Reachability(c)-\{c\}} Coverage(c') \neq \phi$；

3)如果一个范例为连接范例(Spanning Cases)，当且仅当

$\exists\ c' \in Reachability(c)-\{c\}$ 且 $Coverage(c) \subset Coverage(c')$；

4)如果一个范例为辅助范例(Auxiliary Cases)，当且仅当

$\exists\ c' \in Reachability(c)-\{c\}$ 且 $Coverage(c') \subset Coverage(c)$。

从上面定义可知，在进行范例删除时，应按辅助范例、连接范例、支持范例、核心范例的次序对范例进行删除。

4 范例删除主要思想及其算法

解决问题的能力(Competence)和解决问题的效率(Efficiency)是衡量范例删除策略好坏的两个主要指标。解决问题的能力由CBR系统能解决新问题的数目和类型来衡量，若范例数越多，则系统的能力就越强；解决问题的效率由CBR系统解决目标问题的速度来决定，若范例数越少，则系统的效率越高。显然这两个性能指标是相互矛盾的；如何在尽量保持CBR系统能力的情况下，减少范例数成为当今CBM研究的热点。传统的删除策略有：随机删除策略，Minton删除策略[6]。

1)随机删除是当范例库的长度超过某一预定的界值时，就从范例库中随机挑选范例进行删除。这种方法虽简单易实现，但以牺牲CBR系统的能力为代价。

2)Minton删除策略是在对范例进行删除时，基于下面的公式：

实用性(Utility) = (ApplicationFreq * AverageSaving) −MatchCost

删除那些实用性(Utility)的值为负的范例。

在本文中，提出一种新的基于聚类集的删除策略，其基本思想是：用聚类分析方法对范例库进行分区聚类，使相似度大的范例位于同一聚类中，而使相似度小或不相似的两个范例位于不同的聚类中，同时，特别注意的是将离群范例单独作为一类，而且该类中范例往往是核心范例(Pivotal Cases)，一般该类不作删除；并且从范例的推理可知：相似的问题具有相似的解决方法，相似的范例解决相似的问题。本文基于以上思想，在给出删除策略之前，先重新定义范例的覆盖集(Coverage)和范例的可达集(Reachability)：

给定一个范例库 $C=\{c_1,\cdots,c_n\}$，对于每一个范例 $c \in C$：

1)范例c的覆盖集定义为：$Coverage(c)=\{c' \in C \mid S(c,c') \geqslant 1-R_C\}$；

2)范例c的可达集定义为：$Reachability(c)=\{c' \in C \mid S(c',c) \geqslant 1-R_R\}$；

其中，$S(c,c')$和$S(c',c)$为范例c和c'之间的相似性，R_C和R_R分别为覆盖集和可达集的聚类半径。它们的含义为：范例c的覆盖集表示范例解决问题的空间，该集越大，则该范例处理问题的能力越强，对这种范例尽量保留；而范例c的可达集表示该范例可被该集中范例集取代，若该集越大，说明该范例可被取代的范例就越多，对这种范例的删除将不会影响CBR系统解决问题的能力。

通过以上的定义，下面给出基于聚类策略的范例删除步骤：

Step1：初始化参数R_C、R_R、聚类半径R、聚类数K；

Step2：用某聚类算法(如K-means方法)对范例库中的范例进行聚类操作，将那些与所有的聚类中心都大于聚类半径R的范例单独作为一类(记为C_0类)；此操作后，范例库$C=\{C_0,C_1,\cdots,C_K\}$，其中$C_i(0 \leqslant i \leqslant K)$为一个聚类；

Step3：在各聚类中(除C_0类外)求出具有最大可达集的范例；

Step4：若有多个范例的可达集相同，则考虑它们的覆盖集的大小，并将具有小覆盖集的范例进行删除；否则将具有最大可达集的范例删除。

最后，给出基于聚类的范例删除算法的主要处理流程：

```
1  参数 R_C、R_R、聚类半径 R、聚类数 K 的设定；
2  调用聚类算法(如 K-means 方法)后生成各聚类集，记范例库 C={C_0,C_1,…,C_K}；
3  For 每一个聚类 C_i∈{C_1,C_2,…,C_K} Do // 对除 C_0 外的聚类进行范例删除操作
4     For 每一范例 c∈C_i Do // 计算 C_i 聚类中每个范例的可达集
5        计算 Reachability(c)；
6     将具有 |Reachability(c*)|=max_{c∈C_i} |Reachability(c)| 的 c*∈C_i 插入到删除集
      DelSet 中；// |Reachability(c*)|表示集合的元素个数
7     If 集合 DelSet 的元素个数>1 Then
8        For 每一个范例 c*∈DelSet Do // 计算 DelSet 中每个范例的覆盖集
9           计算 Coverage(c*)；
10    找到具有 |Coverage(c**)|=min_{c*∈DelSet} |Coverage(c*)| 的 c**∈DelSet；
11       DelSet={c**}；// 对删除集 DelSet 重新赋值；
12    在 C_i 聚类中，对删除集 DelSet 中的范例进行删除。
```

与传统的范例删除策略相比，该策略能在保持CBR系统解决问题能力的前提下，最大限度地减少范例库中的范例，从而大大地提高范例提取的效率，改进CBR系统的性能。

5 实验结果及说明

为了进一步说明基于聚类策略的范例删除方法的有效性，本文利用Jeffrey C. Schlimmer提供的一批用于CBR领域的实验数据[7]，这批数据是关于汽车规格方面的，共有26个属性，有1个整型、10个符号型，15个实数型属性；库中有205个范例。实验准备阶段，先从205个范例中选取150个作为范例库，并利用K-means聚类算法对范例库进行聚类分析，产生范例分类的结果；再选取50个范例作为测试集。实验运行环境为：前台为VC开发的测试程序，后台用MSSQL Server数据库服务器来存放范例库、分类结果等数据，800MHZ的CPU，128M内存和Windows 2000操作系统。

本实验通过范例删除前、后对测试范例集的平均相似度改变来评价删除策略性能，在对测试集进行范例提取时采用最近邻提取算法，并将测试结果与随机删除策略进行比较，比

(下转第149页)

Kaufmann,1996. 134～145
9 Hipp J,Güntzer U,Nakhaeizadeh G. Algorithms for Association Rule Mining-A General Surey and Comparison. ACM-SIGKDD, July 2000
10 Hipp J,Güntzer U,Nakhaeizadeh G. Mining Association Rules: Derving A Superior Algorithm by Analysing Today's Approaches. In:Proc. of the 4th European Conf. on Principles and Practice of Knowledge Discovery,Lyon,France,Sep. 2000
11 Gouda K,Zaki M J. Efficiently Mining Maximal Frequent Itemsets
12 Webb G I. Efficient Search for Association Rules. In KDD-2000 Boston,MA,Aug. 2000
13 Webb G I. Discovering Association with Numeric Variables. In KDD-2001 San Francisco,CA,Aug. 2001.
14 Han J,Pei J,Yin Y. Mining Frequent Patterns without Candidate Generation. In SIGMOD'00,Dallas,TX,May 2000. 1～12
15 Pei J,Han J,Mao R. CLOSET:An Efficient Algorithm for Mining Frequent Closed Itemsets. In DMKD'00,Dallas,TX,May 2000. 11～20
16 Han J,Pei J. Mining Frequent Patterns by Pattern-Growth: Methodology and Implications. 2001
17 Pei J,Han J,Lakshmanan L V S. Mining Frequent Itemsets with Convertible Constraints. In ICDE'01,Heidelberg,Germany,April 2001
18 Özel S A,Güvenir H A. An Algorithm for Mining Association Rules Using Perfect Hashing and Database Pruning
19 Brin S,Motwani R,Ullman J D,Tsur S. Dynamic Itemset Counting and Implication Rules for Market Basket Data. In:Proc. of the ACM-SIGMOD'97,1997
20 Mannila H. Theoretical Frameworks for Data Mining. 2000 ACM-SIGKDD',Jan. 2000
21 Kleinberg J,Papadimitriou C,Raghavan P. A Microeconomic View of Data Mining. Data Mining and Knowledge Discovery,1998,2(4):311～324
22 Boulicaut J-F,et al. Modeling KDD Processes within the Inductive Database Framework. Data Warehousing and Knowledge Discovery(DaWak 1999),M. K. Mohania and A. M. Tjoa(eds), PP. 293～302
23 Clofton C. Privacy Preserving Distributed Data Mining
24 Agrawal R,Srikant R. Privacy-Preserving Data Mining. In:Proc. of the 1997 ACM-SIGMOD Conf. on Management of Data,Dallas, TX,May 2000. 14～19
25 Lindel Y,Pinkas B. Privacy Preserving Data Mining. In Advances in Cryptology-CRYPTO,Springer-Verlag,Aug. 2000. 36～54
26 Han J,Kammbr M. Data Mining Concepts and Techniques. Morgan Kaufmann Publishers,2000
27 Park J S,Chen M S,Yu P S. Using a Hash-Based Method with Transaction Trimming for Mining Association Rules. IEEE Transactions on Knowledge and Data Engineering,1997,9(5)
28 Gehrke J. New Research Directions in KDD. Report on the SIGKDD 2001 Conference Panel
29 Smyth P. Data Mining at the Interface of Computer Science and Statistics. Data Mining for Scientific and Engineering Applications,to appear,2002
30 Bayardo R J Jr.,Agrawal R. Mining the Most Interesting rules. In:Proc. of the Fifth ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining,1999. 145～154
31 Orlando S,Palmerini P,Perego R. Enhancing the Apriori Algorithm for Frequent Set Counting. DaWak 2001,LNCS 2114, 2001. 71～82

(上接第144页)

较结果如图1所示。另外从范例删除前、后对CBR系统范例提取效率(仍采用最近邻提取算法)的影响来评价删除策略性能,实验结果如图2所示。

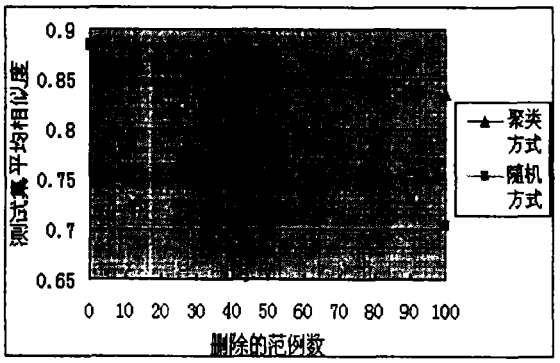

图1 两种删除策略下测试集平均相似度随删除范例数增加的变化情况

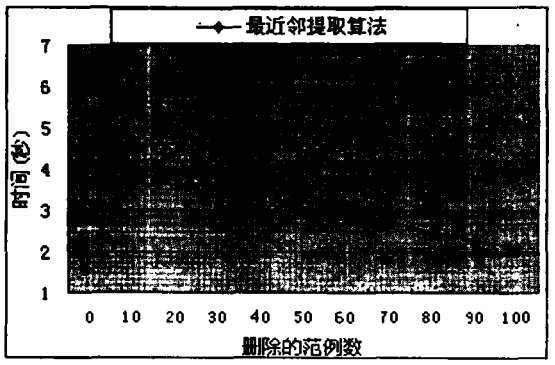

图2 提取时间随删除范例数增加的变化情况

从图1、图2可以看出,基于聚类策略的范例删除方法比随机范例删除方法具有更高的平均相似度;而且在删除范例数增加时,该方法下的平均相似度减少缓慢,而范例提取时间明显下降,从而该方法能够在保持CBR系统解决问题能力的前提下,最大限度地减少范例库,改进CBR系统的性能。

结束语 本文提出了一种有效的基于聚类策略的范例删除方法,从实验结果可知,该方法能在保持CBR系统解决问题能力的前提下,最大限度地减少范例库,提高范例提取的效率,改进CBR系统的性能。与此同时,算法中经验参数的选取有待进一步的研究。

参考文献

1 Schank R. Dynamic Memory: A Theory of Reminding and Learning in Computers and People. Cambridge University Press, Cambridge,UK,1982
2 Francis AG,Ram A. The utility problem in case-based reasoning. In:Proc. AAAI-93 Case-Based Reasoning Workshop, 1993
3 Smyth B,Keane M T. Remembering To Forget: A Competence-Preserving Case Deletion Policy for Case-Based Reasoning Systems. IJCAI,1995. 377～383
4 Leake D B,Wilson D C. Remembering Why to Remember: Performance-Guided Case-Base Maintenance. EWCBR 2000, Springer Verlag,pp 161～172
5 范明,孟小峰等译. 数据挖掘概念与技术. 北京:机械工业出版社, 2001
6 Minton S. Qualitative Results Concerning the Utility of Explanation-Based Learning. Artificial Intelligence, 1990,42:363～391
7 http: // www-old. ics. uci. edu/pub/machine-learning-databases/autos/