第 33 卷 第 4 期 计 算 机 工 程 2007 年 2 月
Vol.33 No.4 Computer Engineering February 2007
·人工智能及识别技术· 文章编号：1000—3428(2007)04—0165—03 文献标识码：A 中图分类号：TP39

CBR 检索的线性回归模型及实现

刘缵敏[1,2]，孙 义[1]，史忠植[2]
(1. 北京科技大学计算机系，北京 100083；2. 中国科学院计算技术研究所，北京 100080)

摘 要：通过对 CBR 传统模型的分析与研究，针对传统 CBR 检索中主观确定特征权重的不足，提出了 CBR 检索的线性回归模型，该模型利用最小二乘法的线性回归性，更加科学、准确地确定各特征的权重，依据成熟的距离公式准确地求出范例的相似度，达到范例准确高效重用的目的。最后介绍了模型的实现方法，并且给出了详细的模型参数。

关键词：CBR；检索；回归；最小二乘法；权重

Linear-regression Model of CBR Retrieve and Implementation

LIU Zuanmin[1,2], SUN Yi[1], SHI Zhongzhi[2]
(1. Department of Computer Science, University of Science and Technology Beijing, Beijing 100083;
2. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080)

【Abstract】After studying the traditional CBR model, according to the deficiency that the feature weights are subjectively defined, this paper proposes a linear-regression model for CBR retrieval, which will improve the effectiveness of CBR retrieval model. The key idea is to decide the weight of each feature by the method of least square. And its property for linear regression helps to make the weights more exact. Thus in the model the similarity degree between cases is more precise than the traditional one, which facilitates the reuse of the existing cases greatly. It also gives one method to implement the model and describes the parameters.

【Key words】Case-based reasoning; Retrieval; Regression; Least square; Weight

Case-based Reasoning(CBR)是以范例为基础进行推理，把人们以往的经验存成一个个的范例，当面临新的问题时，就可以对范例库进行搜索，找到合适的范例作为参考，这其实是实现经验的重用；如果对找到的实例有不满之处，就可以进行修改以适应当前情况，修改后的实例将被再次存入范例库，以便下次使用时作为参考，这其实是实现经验的自学习。

1 CBR 的传统检索模型

1.1 CBR 的工作原理

基于范例的推理过程大致如图 1。

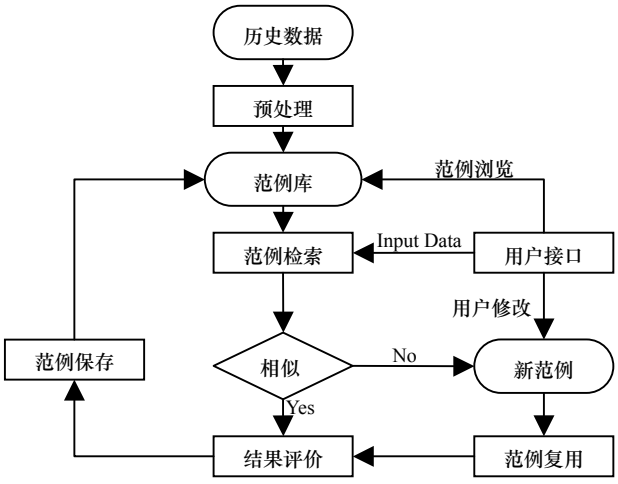

图 1 CBR 工作流程

1.2 范例的检索模型

(1)CBR 方法和人类解决问题的方式很相近。碰到一个新问题时，CBR 首先是从记忆或范例库中回忆出与当前问题相关的最佳范例。后面所有工作能否发挥出应有的作用，很大程度上依赖这一阶段得到的范例质量的高低，因此这步非常关键。它可分为 3 个子过程：特征辨识，初步匹配以及最佳选定。

(2)相似度定义。CBR 很关键的一个环节也是应用成功与否的前提是——范例检索得到的相似范例应该尽可能的好，由于范例检索是在相似比较的基础上进行的，要检索到相似的范例，完全要靠“什么程度才算相似”的定义了。如果定义得不好，检索的结果就不理想，也谈不上应用的成功。因此相似度的定义十分重要。

范例的表示表明，范例的情境是由许多属性组成，要谈范例间的相似度，就必须先定义范例属性(或变量)的相似度。

1)数值属性的相似度

$$Sim(V_i,V_j)=1-d(V_i,V_j)=1-d_{ij}$$ 或

$$Sim(V_i,V_j)=\frac{1}{1+d(V_i,V_j)}=\frac{1}{1+d_{ij}}$$

$$d_{ij}=|V_i-V_j| \quad 或 \quad d_{ij}=\frac{|V_i-V_j|}{\max(V_i,V_j)-\min(V_i-V_j)}$$

2)名称属性的相似度：名称属性相似度一般有两种，一种是一刀切，它比较简单而武断，只要两个属性值不同，就认为两者之间的相似度为 0，否则为 1。另一种则依据具体情况而定，不是如上述那样进行简单的非此即彼划分，而是针对不同的属性值间不同的关系给以具体的定义。前者其实是

作者简介：刘缵敏(1981－)，男，硕士生，主研方向：CBR，Grid Computing，知识管理等；孙 义，副教授；史忠植，研究员、博导

收稿日期：2006-02-21 **E-mail：**liuzm@ics.ict.ac.cn

质上的，即非此即彼的二值分割；而后者则是量上的，进一步细化值间的区别。一般来讲，前者通用，适用于各种情况；而后者则要由人来预定义，与领域知识相关的，从而专用性强。两种方法各有自己的适用范围。

3)有序属性的相似度：有序属性介于数值和名称属性之间，也介于定性和定量之间。由于属性值有序，因此可以赋予不同等级值间有不同的相似度。和名称属性相比，有序属性规整性强。假设属性值分为 n 个等级，则等级 i 和 j(1≤i，j≤n)之间的相似度可以定义为 1-|i-j|/n。

其实，数值属性、有序属性和名称属性之间可以相互转化，依具体问题、对象的性质和研究对象的刻画程度不同，有时一个属性可以由数值属性来刻画，也可以由有序属性来刻画，比如学生成绩可以用 0~100 的分数来反映也可以用 A、B、C 来反映，只不过刻画的程度不同而已。

4)范例相似度的定义：要计算范例之间的相似度，必须考虑组成一个范例的各个属性相似度综合在一起形成的效应。

范例的相似度也常常是通过距离来定义的。常用的典型距离定义有

①绝对值距离(Manhattan)

$$d_{ij}=\sum_{i=1}^{N}\left|V_{ik}-V_{jk}\right|$$

其中 v_{ik} 和 v_{jk} 分别表示范例 i 和范例 j 的第 k 个属性值。

②欧氏距离(Euclidean)

$$d_{ij}=\sqrt{\sum_{k=1}^{N}(V_{ik}-V_{jk})^2}$$

③麦考斯基距离

$$d_{ij}=\left[\sum_{k=1}^{N}\left|V_{ik}-V_{jk}\right|^{q}\right]^{1/q},q>0$$

绝对值距离和欧氏距离就是分别当 q=1 和 q=2 时的距离。

上面的距离定义还只是属于平凡的定义，把各属性所起的作用一视同仁。由于事实上各属性对一个范例整体上的相似度有不同的贡献，因此还需要加上权值。即上面的公式可以写成

$$d_{ij}=\sum_{k=1}^{N}w_k d(V_{ik}-V_{jk})$$

其中，w_k为第k个属性权值大小，一般要求$\sum_{k=1}^{N}w_k=1$，d(v_{ik}，v_{jk})表示第i个范例和第j个范例在第k个属性上的距离，它可以为前面定义的英武距离，也可以是其它距离的定义。

有了距离的定义，就可以类似地得到两个范例间相似度的定义：

$SIM_{ij}=1-d_{ik}$(当$d_{ij}\in[0,1]$)

或为

$$SIM_{ij}=\frac{1}{1+d_{ij}}$$

2 CBR 回归检索模型

2.1 回归模型的提出

回归模型主要是针对 CBR 的检索而提出的改进方法，使检索精确度相比传统模型有很大提高。

传统方法中，特征的权重全是靠主观设定，缺乏科学依据，既而在范例检索时精确度不高，影响了 CBR 在实际中的应用。根据这一需求，在总结传统 CBR 理论的基础上，提出了一种科学确定范例特征权重的模型与算法。

2.2 回归模型的思想

在范例特征相似检索算法中，这样定义相似度 S。

$S=s_1*w_1+s_2*w_2+\ldots+s_n*w_n$

其中，s_i为各特征之间的相似度，并假定所有权重（w_i）之和为 1。这里包括以下特征：s_1、$s_2\cdots s_n$。这样定义特征间的相似度：

$S_i=1-|d_i|$，（$0<|d_i|<1$）

其中，$|d_i|$是两个特征值的广义距离。如果比较不同范例的相同的特征值，则在具体环境下需要量化这个特征值，最终转化为几何距离，或者数值差。(无论是几何距离还是数值差，应将结果转换成在区域（0,1）之间。）有了各个特征相似度的定义，如果再知道各个特征所占的权重，就可以确定整个范例间的相似度，从而容易地进行范例相似检索了。但问题的关键和难点是，很难准确地定义权重。即使是领域专家，也只能指出哪个特征比较重要，哪个特征比较不重要，但如果要定量地给出各特征的权重，是非常困难的。

在这里利用最小二乘法，动态调整各个特征权值，以使结果达到最佳。

权值调整的思想是基于有专家指导的学习（supervised learning)。这里有两种方法定义专家策略：(1)由专家担任评判；(2)根据预测结果与实际结果的符合度。第 1 种策略实际一些且准确些；第 2 种则还要定义什么叫符合度，如果是分类问题则好解决，分类结果相同则符合，而对于范例这种复杂的预测，结果则难以给出。这样，本文考虑采取专家评判的方案。

首先给出一组权重的初值，为了表示一个特征的重要性，设这一特征的权重为 0.5，而其它特征的权重之和为 0.5（如开始时可以平均地分配使另外 x 个特征每个特征的权重都为$\frac{0.5}{x}$)。这样可以求每个相似范例与样本的相似度。然后再由专家对每个范例的相似度进行评估，给出该相似度的真值。接着，使用最小二乘法调整权值。在一般使用最小二乘法时，通过实验测得真值。但在这里，函数值是范例相似度，是无法通过实验求得的。所以，只能由领域专家通过对该相似范例进行评估，给出它的相似度作为真值。在系统实现时，可以设计一个子模块，它能将样本范例的特征和相似范例的特征同时用可视化的方式展示给用户，并详细提供相关的数据，以供专家对每个范例特征的相似度更好地进行评估，得到真值。

任取一范例集，选中它的一个范例作为样本，然后再选取 m 个范例的样本进行相似比较。这里设各特征相似度为S_i，权值为W_i，共有 n 个特征，计算得出的范例相似度为 y，即 y=$\sum_{i=1}^{n}S_i*W_i$（设$\sum W_i=1$)。将 m 个范例逐个与样本范例进行比较计算，可以记录积累下如下的数据（上标表示第几个相似范例)：

$(S_1^1;S_2^1,\ldots,S_n^1\ ;\ y^1)$

$(S_1^2;S_2^2,\ldots,S_n^2\ ;\ y^2)$

…

$(S_1^m, S_2^m, \ldots, S_n^m; y^m)$

每一次范例检索测试得到相似度y后，接着根据y值可以确定相似范例。对于这些可能的相似范例，专家能够判别哪些是对的，哪些是错的，甚至指出哪些很好、较好、一般、较坏和很坏。有了这些判断，可以假设y的值应该做一调整。比如对某个y^i，它能确定的范例较坏，则表明y^i值过高，应该予以降低。专家给出每个范例与样本范例的相似度估计值：y_i，允许误差范围为$(-\varepsilon, \varepsilon)$。以上这些参数值只是一种参数，可根据情况而定。设置允许误差范围的目的是，由于专家给出的判断只是一个模糊语言值而非精确值，因此当y值落入允许误差范围内，便可认为该值可靠性已满足。如果满足可靠性的各个y的所占比率达到一定量如90%，则可以认为权值调整已经满意了，停止。

调整权值的方法是最小二乘法。设m组数据（S_{ij}即前面的S_j^i）

S_{11}、S_{12}、…、S_{1n}

S_{21}、S_{22}、…、S_{2n}

…

S_{m1}、S_{m2}、…、S_{mn}

的最小二乘法误差总和记为

$$E = \sum[(\sum_{j=i}^{n} S_{ij} * W_j) - y_i]^2$$ (y_i是专家估计值))

则根据梯度下降原理知：当取

$$\Delta W_j = -\delta \Big/ \frac{\partial \mathrm{E}}{\partial \mathrm{W}_j}$$

时，E能以最快速度收敛到最小。这里δ是步长，应取为一较小的正实数，如δ=0.02。把 $\frac{\Delta E}{\Delta W_j}$展开得

$$\Delta W_j = -\delta / \ 2\sum_{i=1}^{m} [W_j S_{ij}^2 + S_{ij} \sum_{k \neq j} S_{ik} W_k - S_{ij} y_i]$$

这样，给定一组W_j初值后，可依据上式得到的W_j的下一组调整值，即

$$W_j^{(n+1)} = W_j^{(n)} - \Delta W_j^{(n)}$$

经过不断调整权值，当计算所得的各个y值落入允许误差范围的达到90%左右，就认为是满意的，可以停止调整了。

权值确定以后，整个相似度的公式就确定了。接下来，就可以从范例库中选择同类的范例，进行相似检索，从而检索得到相似度最好的一组相似范例，提供给后面的处理。

3 实现方法

根据上述模型的描述，确定问题的输入(见表1)和问题的输出（见表2）。

表1 输入参数

Hit	FEATURE1	FEATURE2	…	FEATUREN
Hit_1_value	Feature_11_value	Feature_12_value	…	Feature_1n_value
Hit_2_value	Feature_21_value	Feature_22_value	…	Feature_2n_value
⋮	⋮	⋮	…	⋮
Hit_m_value	Feature_m1_value	Feature_m2_value	…	Feature_mn_value

表2 输出结果

FEATURE_W1	FEATURE_W2	FEATURE_W3	…	FEATURE_Wm
FEATURE_W1_VALUE	FEATURE_W2_VALUE	FEATURE_W3_VALUE	…	FEATURE_Wm_VALUE

问题的求解：根据模型的思想及问题的输入输出，问题的求解最终转化为方程组的求解。这里采用成熟的迭代算法，迅速求解方程组。

4 结论

传统检索模型定义了距离公式，利用专家的经验来确定特征权重，最后求得范例的相似度；因为没有科学准确地确定特征权重，相应利用距离公式求得的范例相似度准确度不高。

针对传统检索模型的这一缺点，本文提出的新的回归检索模型对这一缺点做了很大修改，利用最小二乘法的回归性，科学准确地确定权重，最终准确检索到匹配范例，然后再考虑范例的复用等后续工作，真正准确地实现记忆与经验的回忆。

参考文献

1 史忠植. 高级人工智能[M]. 北京: 科学出版社, 1997.

2 史忠植. 智能主体及其应用[M]. 北京: 科学出版社, 2000.

3 史忠植. 知识发现[M]. 北京: 清华大学出版社, 2002.

4 周　涵. 基于范例学习的内燃机产品设计系统[D]. 北京: 石油大学, 1993.

5 王　军. 基于范例推理的淮河王家坝洪水预报系统技术报告[R]. 中国科学院计算技术研究所, 1997.

6 赵　钢. 基于范例推理技术及其在降水过程预测中的应用[D]. 北京: 中国科学院计算技术研究所, 1995.

7 Klaus-Dieter A, Wess S, Bergmann R. Induction and Case-based Reasoning for Classification Tasks[C]//Proc. of the 17th Annual Conference on the GfKl. 1994: 3-16.

8 Aamodt A, Plaza E. Case-based Reasoning: Foundational Issues, Methodological Variations, and System Approaches[J]. AI-Communications, 1994, 7(1).

(上接第164页)

参考文献

1 Figueiredo M, Jain A K. Unsupervised Learning of Finite Mixture Models[J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2002, 24(3): 381-396.

2 Celeux G, Forbes F, Peyrard N. EM Procedures Using Mean Field-like Approximations for Markov Model-based Image Segmentation[J]. Pattern Recognition, 2003, 36(1): 131-144.

3 Blekas K, Likas A, Galatsanos N P, et al. A Spatially Constrained Mixture Model for Image Segmentation[J]. IEEE Trans. on Neural Networks, 2005, 16(2): 494-498.

4 Akaike H. A New Look at the Statistical Model Identification[J]. IEEE Trans. on Automatic Control, 1974, 19(6): 716-723.

5 Richardson S, Green P. On Bayesian Analysis of Mixture Models with Unknown Number of Components[J]. Journal of Royal Statistical Society, Series B, 1997, 59(4): 731-792.

6 Dempster A P, Laird N M, Rubin D B. Maximum-likelihood from Incomplete Data via the EM Algorithm[J]. Journal of Royal Statistical Society, Series B, 1977, 39(1): 1-38.