文章编号: 1001-0920(2011)04-0637-04

基于微粒群的案例推理方法研究

韩 敏, 沈力华

(大连理工大学 电子信息与电气工程学院，辽宁 大连 116024)

摘 要: 距离测度是案例检索的关键问题, 它直接影响案例检索精度. 针对距离测度进行研究, 提出一种基于微粒群方法的自学习距离测度, 并将该自学习距离测度引入案例推理中, 使案例推理在处理由相关属性表述的案例时有了合理的解决方法, 从而扩展了案例推理的应用范围. 最后, 利用实际数据与 UCI 数据对基于新距离测度的案例推理技术进行了仿真实验, 实验结果表明, 与其他方法相比,该方法可以提高案例检索的准确性.

关键词: 案例推理；案例检索；微粒群；自学习距离测度

中图分类号: TP182 **文献标识码:** A

Research of CBR based on particle swarm optimization

HAN Min, *SHEN Li-hua*

(Faculty of Electronic Information and Electrical Engineering，Dalian University of Technology，Dalian 116024，China. Correspondent：HAN Min，E-mail：minhan@dlut.edu.cn)

Abstract: Distance measure is the key issue in case-based reasoning(CBR), which influences the accuracy of case retrieval directly. For distance measure, a learning distance measure based on particle swarm optimization is proposed. The application range of CBR is extended by introducing leaning distance measure into CBR technology for the first time, which makes CBR technology have reasonable method to deal with the cases with correlative attributes. Finally, the simulation is conducted with real data and UCI data. The result shows that, compared with the other methods, this distance measure improves the accuracy of case retrieval.

Key words: case-based reasoning；case retrieval；particle swarm optimization；learning distance measure

1 引 言

案例推理 (CBR) 是一门人工智能技术, 它可通过结合各种方法来提高预测精度. 案例推理过程类似于人类解决问题的过程, 通过采用以往的成功经验来解决新问题. 其工作原理是通过类比当前案例与案例库中案例, 找到与当前案例最相似的历史案例, 利用历史案例的解决方案来处理当前问题[1]. 案例检索是案例推理中最关键的问题, 其检索结果的优劣直接影响到问题方案的求解.

CBR 中经常采用欧氏距离作为检索过程中相似度的计算, 但是欧氏距离通常假设各属性权重相等且属性间是相互独立的, 这一假设在实际应用中往往不能满足[2]. 自学习距离测度是利用已知特征数据学习得到, 不必事先假设数据符合何种分布, 且不必事先假设属性间相互独立, 考虑到了各个属性间的关系. 在过去的几年中, 已经出现了许多关于自学习的距离测度的研究, 这种基于自学习的距离测度在许多问题上相比于欧氏距离都表现出很大的优势[3]. 目前比较典型且有效的监督自学习距离测度方法有 DCA (discriminative component analysis) 方法[4] 和 RCA (relevant component analysis) 方法[5]. 文献 [6] 通过同时考虑类内类间信息, 提出了改进的 RCA 方法. 本文在已有的自学习距离测度基础上, 提出一种基于微粒群 (PSO) 的自学习距离测度, 并将此自学习距离测度用于案例检索中相似案例的选取. 最后利用实际数据与 UCI 数据进行了仿真实验, 实验结果表明, PSO 自学习距离测度相比于传统欧氏距离测度和几种主要的监督自学习距离测度都具有一定的优势.

2 基于微粒群方法的自学习距离测度

距离测度是案例推理以及许多机器学习算法的

收稿日期: 2010-01-31；**修回日期:** 2010-04-15.

基金项目: 国家科技支撑计划项目(2006BAB14B05)；国家973计划项目(2006CB403405).

作者简介: 韩敏(1959−), 女, 教授, 博士生导师, 从事神经网络、专家系统等研究；沈力华(1984−), 女, 硕士生, 从事案例推理相关算法的研究.

关键, 如 K 均值算法, K-NN 算法等都需要有一个好的距离测度以更好地实现学习过程. 目前最常采用的距离测度是欧氏距离测度和改进的欧氏距离[7]. 欧氏距离及其改进方法的计算公式均可由下式表示:

$$d_{ij}^2 = (\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}}\boldsymbol{W}(\boldsymbol{x}_i - \boldsymbol{x}_j).$$

其中: d_{ij} 为第 i 个案例与第 j 个案例的距离; $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 分别为第 i 个案例与第 j 个案例的各个属性所组成的列向量; $\boldsymbol{W}$ 为权值矩阵, 是一个对角阵, 其维数表示各案例的属性个数, 各个元素为各属性的权值. 显然, 若 $\boldsymbol{W}$ 为对角阵, 则必须先假定属性间相互独立, 这在现实中往往不能满足. 为提高检索精度, 改变上述方法对于各个属性相互独立的不合理假设, 本文提出一种基于微粒群方法的自学习距离测度方法.

2.1 目标函数和初始位置的确定

两个案例之间的距离可表示为

$$d_{ij}^2 = |(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}}\boldsymbol{M}(\boldsymbol{x}_i - \boldsymbol{x}_j)|. \tag{1}$$

其中: d_{ij} 为第 i 个案例与第 j 个案例的距离; $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 为两个列向量, 分别表示第 i 个案例与第 j 个案例; 在传统马氏距离中, $\boldsymbol{M}$ 为数据集协方差矩阵的逆, 在此, 令 $\boldsymbol{M} = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}$, 这里 $\boldsymbol{B}$ 是一个距离变换矩阵, 也是一个通过渐近学习而得到的矩阵. 定义如下两个矩阵:

$$\boldsymbol{A}_{\mathrm{dif}} = \frac{1}{n_d}\sum |(\boldsymbol{x}_j - \boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}(\boldsymbol{x}_j - \boldsymbol{x}_i)|, \tag{2}$$

$$\boldsymbol{A}_{\mathrm{sam}} = \frac{1}{n_s}\sum |(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}(\boldsymbol{x}_i - \boldsymbol{x}_j)|. \tag{3}$$

其中: 式 (3) 中的 $\boldsymbol{x}_j$ 与 $\boldsymbol{x}_i$ 表示不同类之间的两个案例, n_d 为不同类之间的案例对的个数, $\boldsymbol{A}_{\mathrm{dif}}$ 描述了各不同类之间的距离平方之和; 式 (4) 中的 $\boldsymbol{x}_j$ 与 $\boldsymbol{x}_i$ 为同一类中的两个案例, n_s 为各类类内案例对的个数, $\boldsymbol{A}_{\mathrm{sam}}$ 描述了类内各个案例之间的距离平方之和. 下面的问题是如何学习得到一个矩阵 $\boldsymbol{B}$ 使得类间距离最大, 同时类内距离最小, 即转化为使目标函数

$$J(\boldsymbol{B}) = \arg\max_{\boldsymbol{B}} \frac{\boldsymbol{A}_{\mathrm{dif}}}{\boldsymbol{A}_{\mathrm{sam}}} \tag{4}$$

最大化的优化问题. 为使分母中矩阵 $\boldsymbol{A}_{\mathrm{sam}}$ 非奇异, 约束条件为 $\boldsymbol{A}_{\mathrm{sam}} \neq 0$. 最终目标函数如下式所示:

$$J(\boldsymbol{B}) = \arg\max_{\boldsymbol{B}} \frac{\boldsymbol{A}_{\mathrm{dif}}}{\boldsymbol{A}_{\mathrm{sam}}},\ \boldsymbol{A}_{\mathrm{sam}} \neq 0. \tag{5}$$

优化的最终目标是使类内距离比类间距离最小化.

PSO 方法在迭代过程中通过调整矩阵 $\boldsymbol{B}$ 的各个元素, 达到合理分配各个方向上的权重. 因为 $\boldsymbol{B}$ 不是一个对角阵, 所以不用假设各个属性间相互独立, 这在处理具有相关属性案例时具有重要意义. 为使微粒群在搜索最优解的过程中保证一定的精度和搜索速度, 将微粒初始位置设置为一个固定的值, 使其在此基础上进一步优化. 根据文献 [6], 分别定义类间与类内方差矩阵 $\boldsymbol{C}_d$ 和 $\boldsymbol{C}_s$, 令变换矩阵 $\boldsymbol{W}$ 为 $\boldsymbol{C}_d^{0.5}\boldsymbol{C}_s^{-0.5}$, 并设 $\boldsymbol{B}$ 的初始值也为 $\boldsymbol{C}_d^{0.5}\boldsymbol{C}_s^{-0.5}$.

2.2 PSO 方法求优化距离测度的过程

微粒群算法将每个个体看作是在 n 维搜索空间中的一个没有重量和体积的微粒, 并在搜索空间中以一定的速度飞行. 该飞行速度由个体的飞行经验和群体的飞行经验进行动态调整[8]. 设第 i 个微粒当前位置为 $\boldsymbol{X}_i = (x_{i1}, x_{i2}, \cdots, x_{in})$, 它经历过的最好位置记为 $\boldsymbol{P}_i = (p_{i1}, p_{i2}, \cdots, p_{in})$, 也称 $\boldsymbol{p}_{\mathrm{best}}$. 在群体中, 所有微粒经历过的最好位置的索引号用 g 表示, 即 $\boldsymbol{P}_g$, 也称 $\boldsymbol{g}_{\mathrm{best}}$. 微粒 i 的当前速度为 $\boldsymbol{V}_i = (v_{i1}, v_{i2}, \cdots, v_{in})$. 对于第 i 个粒子, 其第 j 维 $(1 \leqslant j \leqslant n)$ 根据如下方程变化:

$$v_{ij}(t+1) = wv_{ij}(t) + c_1r_{1j}(t)(p_{ij}(t) - x_{ij}(t)) + c_2r_{2j}(t)(p_{gj}(t) - x_{ij}(t)), \tag{6}$$

$$x_{ij}(t+1) = x_{ij}(t) + v_{ij}(t+1). \tag{7}$$

基于 PSO 算法求取距离测度的具体流程如下:

1) 在约束条件 $\boldsymbol{A}_{\mathrm{sam}} \neq 0$ 下, 初始化一群微粒的位置和随机速度, 初始位置为 $\boldsymbol{C}_d^{0.5}\boldsymbol{C}_s^{-0.5}$. 其中, 每个微粒对应于矩阵 $\boldsymbol{B}$ 的一个解, 每个微粒的各个元素即为矩阵 $\boldsymbol{B}$ 的各个元素. 以 4 个条件属性的案例推理为例, 则 $\boldsymbol{B}$ 矩阵是一个 4×4 的方阵, 每个微粒是一个 16 维的向量. 将 16 维的向量从第 1 个元素到最后一个元素按顺序拆分成 4 组, 每组为 4 个元素, 并构成矩阵 $\boldsymbol{B}$ 的每一行.

2) 根据目标函数, 如式 (5) 所示, 计算各个微粒的适应值.

其他步骤与传统 PSO 方法基本相同.

经过上述步骤可计算出使适应值最大化的矩阵 $\boldsymbol{B}$, 再根据公式 $\boldsymbol{M} = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}$ 即可求得矩阵 $\boldsymbol{M}$.

3 基于自学习距离测度的 CBR 模型

基于本文方法提出的优化距离测度检索公式如下:

$$d_{ij}^2 = |(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}(\boldsymbol{x}_i - \boldsymbol{x}_j)|, \tag{8}$$

其中 $\boldsymbol{B}$ 即为上述基于微粒群方法学习得到的一个变换矩阵. 利用式 (8) 可计算待处理案例与案例库中的各个案例的距离.

基于上述优化距离测度的 CBR 模型如图 1 所示. 首先, 搜集并整理之前的成功案例, 存入案例库中; 然后, 根据上述学习过程得到新的自学习距离测度, 并利用该距离测度计算案例间的距离, 找到案例库中与目标案例最相似的 K 个案例; 最后通过案例调整确定目标案例的最终解. 在分类问题上, 解决方案即为案例所属类别.

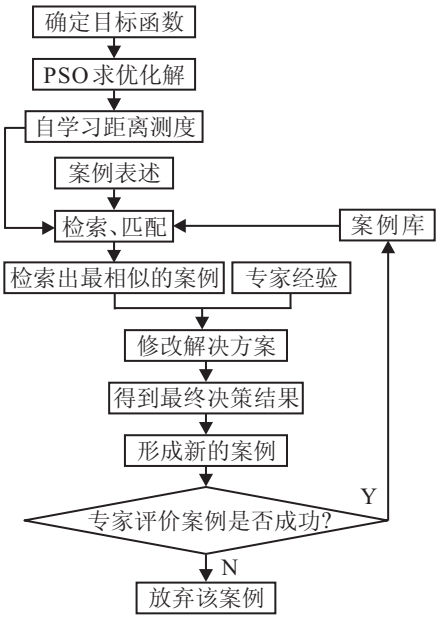

图 1　基于优化的距离测度的案例推理流程

将检索结果作为参考, 结合专家经验, 对检索得到的解决方案进行修改. 结合专家经验对案例进行调整, 实际是人工经验与自动识别的结合过程. 案例推理只能给出一般化的解决方案, 专家还要根据自己的经验判断案例检索结果是否符合现实情况, 并根据实际情况对检索到的解决方案的相关参数进行调整, 以获得最终的适合于具体案例的解决方案, 并用其处理当前待解决的问题. 最后, 专家根据决策所产生的效果对处理该问题的决策结果进行评价, 并判断是否将该案例存入案例库.

从图 1 可以看出, 相比于传统 CBR 模型, 该模型在案例检索、匹配过程中引进了自学习距离测度, 该距离测度在考虑各属性相关性的同时, 根据最大化类间距离同时最小化类内距离的思想得到了更合理的距离测度, 从而实现了对整个模型检索精度的提高. 因为该距离测度是根据数据集本身学习得到的, 所以不同的数据集均会得到适合于自身的距离测度.

4　仿真实例

松花江流域每年降水情况多变, 缺乏规律性, 丰水年份经常洪水泛滥, 枯水年份则引起严重缺水. 因此, 灾情的判断对于如何有效利用洪水资源, 使枯水年份有充足的水资源, 同时在丰水年份避免洪涝灾害具有指导性的作用.

本文选取松花江流域 8 个分区的 43 组数据作为实验数据, 其中部分数据见表 1. 决定灾情等级的属性主要有 4 个: 年降水量 (mm), 水库年末蓄水量 ($10^8 m^3$), 年流量 ($10^8 m^3$) 和总用水量 ($10^8 m^3$). 以年流量和总用水量为例, 这两个属性的变化趋势如图 2 所示.

表 1　部分学习样本

D	年降水量/mm	水库年末蓄水量/$10^8 m^3$	年流量/$10^8 m^3$	总用水量/$10^8 m^3$	灾情
1	713.7	128.02	247.75	62.03	洪涝严重
2	635.3	110.84	167.48	59.19	正常
3	536.2	81.79	101.19	57.3	干旱较轻
4	746.6	114.58	159.18	52.44	洪涝较轻
⋮	⋮	⋮	⋮	⋮	⋮

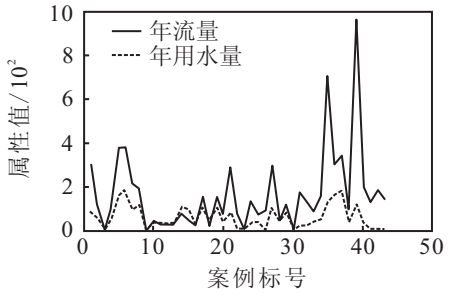

图 2　属性值变化趋势曲线

从图 2 可以明显看出, 这两个属性并不相互独立, 而且具有较强的正相关性, 因此, 假设属性间相互独立的加权欧氏距离并不合理. 本文所提出的基于微粒群的距离测度是在不假设属性间相互独立的情况下得到的, 因此更具合理性. 通过适当选择, 在 K-NN 中, 当 K 取 1 时, 各方法均得到更好的检索结果.

在实验中, 为验证在不同案例库的情况下, 本文方法仍具有一定的优势, 通过如下方法来改变案例库中的案例: 分别按照 1:1, 3:2, 2:3 和 2:1 分配案例库中样本与测试样本的比例. 案例库中的样本也是确定目标函数中 $\boldsymbol{A}_{\mathrm{dif}}$ 和 $\boldsymbol{A}_{\mathrm{sam}}$ 时所用到的数据.

分别将 RCA 方法, 改进的 RCA 方法, 经典的加权欧氏距离方法, DCA 方法以及本文方法应用于 CBR 检索过程的距离计算, 得到各个方法的准确率, 如表 2 所示. 从表 2 可以看出, 在不同案例库的情况下, 本文方法相对于其他方法都具有一定的优势, 说明了所提出方法的有效性. 由表 2 还可看出, 旱涝等级评定不能单纯根据案例检索结果, 而应将检索结果作为一个参考依据, 辅助解决问题.

表 2　各案例推理方法实验结果　　%

分配比例	加权欧氏	RCA	改进 RCA	DCA	本文方法
1:1	65.22	39.13	69.57	56.52	78.26
3:2	60.00	30.00	70.00	50.00	75.00
2:1	75.00	43.75	68.75	60.00	81.25
2:3	56.00	40.00	64.00	60.00	64.00

为进一步验证所提出方法的有效性, 利用UCI中Iris数据对本文方法进行仿真实验. 将数据按1:1分配案例库样本与测试样本, 在案例库不变的情况下, 改变选取K-NN中的K值, 分别取K为$1, 2, \cdots, 9$. 在案例调整过程中利用式(11)[9]确定案例最终所属类别. 实验结果如图3所示. 从图3可以看出, 除K等于3和9以外, K取其他7个值时, 基于本文方法的案例检索精度均好于其他方法.

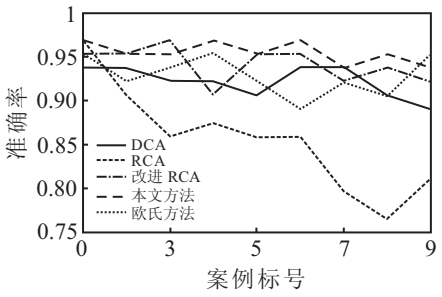

图 3 K取不同值时各方法检索精度

$$C(x_q) \leftarrow \arg\max_{v\in \boldsymbol{V}} \sum_{i=1}^{k} w_i \delta(v, f(\boldsymbol{x}_i)). \tag{9}$$

其中: $w_i \equiv 1/d(\boldsymbol{x}_q, \boldsymbol{x}_i)^2$; $C(x_q)$为输入案例$\boldsymbol{x}_q$的预测类别; $\boldsymbol{V}$为类别$(v_1, \cdots, v_s)$的有限集合; $f(\boldsymbol{x}_i)$为检索到的第i个相似案例$\boldsymbol{x}_i$的所属类别; 当$a = b$时, $\delta(a, b) = 1$; 否则, $\delta(a, b) = 0$.

相对于其他几种方法, 本文方法提高了检索精度. 在时间复杂性方面, 由于引入了智能优化算法, 本文方法需要相对长的时间来寻找最优解. 但在实际应用中, 当要对新案例进行预测时, 距离测度已学习完毕, 只需利用学习好的距离测度直接计算案例间的距离即可, 并不影响案例检索的速度.

5 结 论

本文对自学习距离测度算法进行了研究, 根据最大化类间距离, 同时最小化类内距离的思想确定新的目标函数, 利用微粒群方法求取最优解, 学习得到距离测度, 并将该距离测度引入案例推理检索过程中. 最后, 将基于PSO的自学习距离测度的案例推理技术成功地用于松花江流域8个分区灾情数据和UCI中Iris数据的分类, 实验结果表明了本文方法的有效性.

参考文献(References)

[1] 韩雪, 冯玉强. 基于案例推理的谈判支持系统的研究[J]. 控制与决策, 2008, 23(7): 791-794.
(Han X, Feng Y Q. Negotiation support system based on case-based reasoning[J]. Control and Decision, 2008, 23(7): 791-794.)

[2] Kwang Hyuk Im, Sang Chan Park. Case-based reasoning and neural network based expert system for personalization[J]. Expert Systems with Applications, 2007, 32(1): 77-85.

[3] Xiang Shiming, Nie Feiping, Zhang Changshui. Learning a mahalanobis distance metric for data clustering and classification[J]. Pattern Recognition, 2008, 41(12): 3600-3612.

[4] Hoi S C H, Liu W, Lyu M R, et al. Learning distance metrics with contextual constraints for image retrieval[C]. Proc of the IEEE Computer Society Conf on Computer Vision and Pattern Recognition. New York: IEEE Press, 2006: 2072-2078.

[5] Aharon Bar-hillel, Tomer Hertz, Noam Shental, et al. Learning distance functions using equivalence relations[C]. Proc of the 20th Int Conf on Machine Learning. Washington DC, 2003: 11-18.

[6] Yeung Dit-yan, Chang Hong. Extending the relevant component analysis algorithm for metric learning using both positive and negative equivalence constraints[J]. Pattern Recognition, 2006, 39(5): 1007-1010.

[7] Li Hui, Sun Jie, Sun Bo-liang. Financial distress prediction based on OR-CBR in the principle of k-nearest neighbors[J]. Expert Systems with Applications, 2009, 40(93): 643-659.

[8] 李军军, 肖健梅, 王锡淮. 一种精英退火微粒群算法[J]. 控制与决策, 2008, 23(7): 756-761.
(Li J J, Xiao J M, Wang X H. Elitist annealing particle swarm optimization algorithm[J]. Control and Decision, 2008, 23(7): 756-761.)

[9] Niloofar Arshadi, Igor Jurisica. Data mining for case-based reasoning in high-dimensional biological domains[J]. IEEE Trans on Knowledge and Data Engineering, 2005, 17(8): 1127-1137.