文章编号: 1001-0920(2007)03-0357-04

相似案例自适应选择算法及其应用

程 刚, 钟秋海
(北京理工大学 自动控制系, 北京 100081)

摘 要: 为提高相似案例选择的效率和准确性,将有向无环图支持向量机(DAGSVM)多类分类器应用到相似案例选择中,提出多类分类器有效分辨阈值的概念,在保证一定案例选择准确度的前提下,对自适应构造案例集进行相似案例选择,提高相似案例选择的效率. 将该方法应用于光动力治疗(PDT)鲜红斑痣(PWS)案例推理专家系统,实验结果表明了该方法的有效性.

关键词: 基于案例推理; 案例选择; 支持向量机; 有向无环图
中图分类号: TP18 **文献标识码**: A

Adaptive case retrieval algorithm and application

CHENG Gang, ZHONG Qiu-hai
(Department of Automatic Control, Beijing Institute of Technology, Beijing 100081, China. Correspondent: CHENG Gang, E-mail: chenggang@bit.edu.cn)

Abstract: In order to improve the accuracy and efficiency of retrieving in case-based reasoning (CBR), an algorithm based on case retrieval is proposed, which uses directed acyclic graph support vector machines (DAGSVM) in retrieval. The effective differentiation threshold of DAGSVM is defined, and an adaptive subset of cases is built for each new case and the similar cases are selected in the subset. Using the approach, case retrieval is more efficient and accurate. The approach is applied in photodynamic therapy port wine stain CBR system, and the results show a dramatic increase in retrieving efficiency.

Key words: Case-based reasoning; Case retrieval; Support vector machines; Directed acyclic graph

1 引 言

随着计算机科学和信息技术的飞速发展,人类所面对的知识和信息成指数增长,这使传统的基于规则的推理(RBR)系统在知识和规则的获取上遇到难以克服的困难. 基于案例的推理(CBR)借鉴人类处理问题的方式,规避了这一瓶颈,运用以前积累的知识和经验直接解决问题,引起专家和学者的关注,逐渐成为人工智能领域的一个研究热点[1,2].

相似案例选择是 CBR 系统的一个关键步骤,也是 CBR 系统实现技术研究中的一个热点问题. 相似案例选择结果的优劣直接影响着案例的重用与修改以及整个 CBR 系统的好坏,已有大量文献研究了案例的选择方法[2]. 层次分析、神经网络等方法都已被尝试应用于案例选择[3-6]. 但这些方法都假定案例集是数据库中的全体数据,对每一个新案例均从数据库的全体案例中进行相似案例选择,浪费了大量的查找时间,降低了查询效率,同时查询出的相似案例会引入更多噪声,对决策带来困扰. 因此人们对案例集的选择问题进行了研究[7-9], Kuo 等人[9]应用 ant system clustering 算法对数据库进行了分类,相似案例选择时要先对新案例进行分类,并在相应的案例子类集合中选择相似案例. 这种对案例进行分类的方法提高了相似案例的选择效率,但存在新案例分类错误的情况. 尤其是当新案例处于多个类的边界时,错分概率较高. 对此,本文应用有向无环图支持向量机(DAGSVM)多类分类器,实现案例集的自适应调整,在保证案例选择准确度的前提下,提高了案例选择的效率.

2 基于有向无环图的多类分类支持向量机

2.1 多类分类支持向量机

统计学习理论(SLT)是目前针对小样本统计估计和预测学习的最佳理论,它从理论上系统地研究了经验风险最小化原则成立的条件,有限样本下经验风险与期望风险的关系以及如何利用这些理论找

收稿日期: 2005-11-23; **修回日期**: 2006-01-10.
作者简介: 程刚(1981—),男,河南焦作人,博士生,从事复杂系统的优化控制、人工智能等研究; 钟秋海(1941—),男,江西赣州人,教授,博士生导师,从事复杂大系统的工程控制与决策理论等研究.

到新的学习原则和方法等问题. 统计学习理论对有限样本下模式分类问题中的一些根本问题进行了系统的理论研究,在一定程度上解决了模型选择、过学习、非线性和维数灾等问题. 支持向量机(SVM)最初是为两分类问题设计的,是实现统计学习理论思想的一种方法. 目前如何将支持向量机的优良性能推广到多类分类问题中,已成为支持向量机研究的一个热点问题[10].

目前针对多类分类问题,支持向量机的解决途径主要有两种:一种是通过构造多个 SVM 二值分类器并将它们组合起来实现多类分类,常用的方法有:"一对多"方法,"一对一"方法,有向无环图支持向量机方法(DAGSVM);另一种是直接在一个优化公式中同时考虑所有的分类器参数优化.

决策有向无环图支持向量机是目前理论上比较完善的一种多值分类算法,相对其他方法,该算法建立了对推广性界的理论分析,指出了推广误差取决于无环图的大小以及在各个决策点处的边界,与原样本空间的维数无关. 同时 DAGSVM 还解决了不可分区域问题[11-14]. 因此本文采用 DAGSVM 构建案例库的多类分类器.

2.2 有向无环图支持向量机多类分类算法

已知问题空间 X,问题空间分为 K 类,则有向无环图支持向量机多类分类器中包括了 $N = K \cdot (K-1)/2$ 个 2 值分类器. 每个 2 值分类器都是由支持向量机构成的. 一个 4 分类问题的有向无环图如图 1 所示[12].

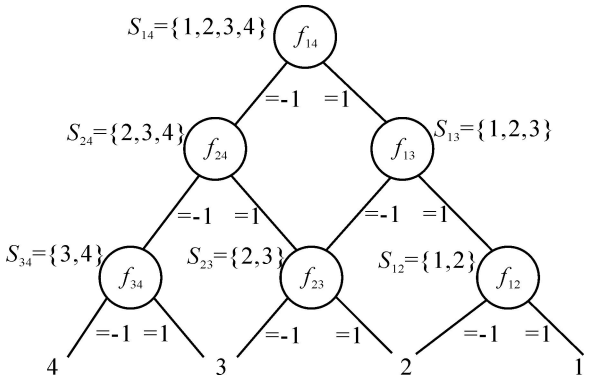

图 1 4 分类问题的有向无环图

由图 1 可知,在有向无环图中除了叶节点外,每一个节点包括两部分内容:当前类型集合 S_{ij} 和一个 2 值分类器 f_{ij},如果假定对于叶节点定义一种形式上的分类器 $f_{ij} = 0, i = j$,分类器取 0 值,则表示已经确定了样例的类型. 这样一个有向无环图便可由 $K \cdot (K+1)/2$ 个分类器对组成,记为

$$\begin{aligned} \mathrm{DAG}(F) = \{ & S_{ij}, f_{ij} \mid i = 1,2,\cdots,K, \\ & j = 1,2,\cdots,K, i \leq j \}, \end{aligned} \tag{1}$$

其中 $F = \{f_{ij}\}$ 表示分类器分类函数的集合. 用支持向量机实现上述的 2 值分类器,最优分类函数为

$$f_{ij} = \begin{cases} \mathrm{sign}\left(\sum_{k=1}^{L_{ij}} \alpha_{ijk} \cdot \mathrm{Kernel}(x, \mathrm{SV}_{ijk}) + b_{ij} \right), \\ \quad i < j; \\ 0, \ i = j. \end{cases} \tag{2}$$

其中:L_{ij} 是 f_{ij} 对应的支持向量个数,SV_{ijk} 是支持向量,x 是待分类样例,α_{ijk} 和 b_{ij} 是分类器参数.

在训练阶段,决策有向无环图需训练上述 $N = K \cdot (K-1)/2$ 个子分类器. 在测试阶段,对一个测试样例 x,首先将其输入到根节点分类器,由该分类器的输出决定测试样本下一步的走向;然后,第 2 个分类器的输出决定测试样本再下一步的走向,依此类推,直到测试样例达到某叶节点,该叶节点所代表的类就是样例 x 所属于的类型. 在测试阶段只需评价 $K-1$ 个分类器的输出.

3 相似案例自适应选择算法

用有向无环图支持向量机多类分类器对样例分类,避免了不可分区域,只需评价 $K-1$ 个分类器的输出,效率高. 但每个 2 值分类器都需有很高的准确度,例如,假定每个 2 值分类器的准确度是 95 %,经过 10 个分类器后,误分类概率为 $1 - 95\%^{10} = 40.13\%$. 在实际应用中,每个 2 值分类器难于达到很高的分类准确度,尤其是当类型之间存在重叠时. 直接将上述多类分类器应用于基于案例的推理中,对每个新案例首先选择一个子类作为其类型,误差较大. 如果在错误类型中进行相似案例选择,尽管选择速度会提高,但准确度却较低.

对于已经学习好的分类器,误分样例一般都分布在紧邻分类面的两侧,针对上述问题,在有向无环图多类分类器中引入一个停止继续分类的阈值 $\theta > 0$,有向无环图中每个节点对应于如下 3 元组:

$$\begin{aligned} \mathrm{DAG}(F) = \{ & S_{ij}, f_{ij}, \theta \mid i = 1,2,\cdots,K, \\ & j = 1,2,\cdots,K, i \leq j \}, \end{aligned} \tag{3}$$

最优分类函数修正为

$$f_{ij} = \begin{cases} \mathrm{sign}(\Omega), \ |\Omega| > \theta \text{且} \ i < j; \\ 0, \ |\Omega| \leq \theta \text{或} \ i = j, \end{cases} \tag{4}$$

其中

$$\Omega = \sum_{k=1}^{L_{ij}} \alpha_{ijk} \cdot \mathrm{Kernel}(x, \mathrm{SV}_{ijk}) + b_{ij}.$$

当 $f_{ij} = 0$ 时,表示当前案例已紧邻分类面或到达有向无环图的叶节点,停止进一步分类. 当前案例属于类型集合 S_{ij} 中的某一个类型.

θ 代表分类器进行分类的有效分辨能力,这里定义为分类器的"有效分辨阈值". θ 值越大多类分类器对类型的分辨就越差,当 θ 足够大时,整个案例

库将重新成为一个整体,多类分类器不起任何分类作用;当 θ 为零时,多类分类器分辨程度最高,将为任何一个输入样例指定一个确定的类型. 相应的,θ 越大误分概率就越小,当 θ 足够大时,不存在误分情况,误分概率最小;当 θ 为零时,误分概率最大.

由于阈值 θ 和多类分类器所要求的准确率之间有紧密联系,学习时首先赋予阈值初值 θ_0,然后以给定步长 $\Delta\theta$ 递增,直到多类分类器达到要求的准确率. 将上述方法用在基于案例推理时,可将集合 S_{ij} 中类型对应的案例集作为新案例所在类型,在该案例集中进行相似案例选择. 这样可根据新案例的具体情况,自适应地选择可用于进行相似案例选择的案例集,在保证精度的前提下,提高案例选择效率. 该方法避免了当新案例类型误分时,相似案例选择无效的问题. 算法流程如图 2 所示.

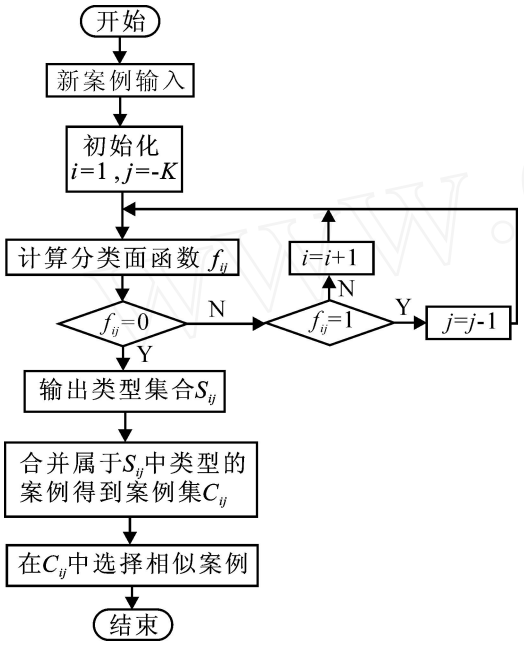

图 2　自适应案例选择算法流程

4　实验和结果

4.1　实验数据

光动力治疗鲜红斑痣是一项新型的诊疗技术,通过注射光敏剂后用激光照射皮肤表面,使得血液中的光敏剂受光激发与血液中的氧作用产生可杀伤病变组织和血管的活性物质,最终治愈病变. 该治疗方法可通过光敏剂和激光的定位,达到组织的选择性损伤,即杀伤病变组织的同时保护正常组织. 该方法有很好的临床疗效,但由于同时涉及到生物、化学、物理等多个过程,因素多且难于检测,很难提取出简单易用的模型或规则. 为了更好地利用病例库中已有病例信息,建立了基于案例推理的专家系统,应用上述方法进行案例的选择,得到了较好的效果.

上述光动力治疗鲜红斑痣案例推理专家系统,是通过输入病人的检查信息,在当前病例库中查询相似案例,根据以往案例的治疗剂量确定当前病人的治疗剂量. 输入的病人检查信息包括病人的性别、年龄、病变程度和病变部位. 治疗剂量信息包括光敏剂剂量、激光功率密度和照光时间长度. 其中光敏剂浓度和激光功率密度是治疗中最重要的两个因素.

根据光敏剂剂量和激光功率密度将数据库中的病例分为 8 个类型. 分类原则兼顾各类型间的相关性和类型中包含的病例数目. 保证类型内部的治疗剂量相关度大,类型之间的治疗剂量相关度小,每个类型中的病例个数适中. 实验中各类型参数和病例分布如表 1 所示.

表 1　各类型参数及病例库中病例分布情况

类型	光敏剂剂量范围 mg/kg	激光功率密度 mW/cm^2	照光时间长度范围/min	病例数目
1	[3.0,3.5)	80	[40,70]	38
2	[3.5,4.0)	80	[40,70]	23
3	[4.0,4.5)	80	[40,70]	33
4	[4.5,6.0]	80	[40,70]	17
5	[3.0,4.0)	100	[40,70]	12
6	[4.0,4.5)	100	[40,70]	42
7	[4.5,5.0)	100	[40,70]	15
8	[5.0,6.0]	100	[40,70]	31

4.2　实验结果与讨论

对实验数据进行归一化处理,具体公式如下:

$$\rho_{\text{sex}} = \begin{cases} 0.9, & \text{male}, \\ 0.1, & \text{female}; \end{cases}$$

$$\rho_{\text{age}} = 1 - e^{-\text{age}/25};$$

$$\rho_{\text{level}} = 0.1 + (0.9 - 0.1) \times \frac{\text{level} - 1}{5},$$

$$\text{level} = 1, 2, \cdots, 6;$$

$$\rho_{\text{location}} = 0.1 + (0.9 - 0.1) \times \frac{\text{location} - 1}{5},$$

$$\text{location} = 1, 2, \cdots, 6.$$

其中:ρ_{sex} 是性别属性,ρ_{age} 是年龄属性,ρ_{level} 是病变程度属性,ρ_{location} 是病变位置属性. 病变分级分为 1 ～ 6 级,病变位置主要有 6 个特征部位. 事例的相似度计算公式为

$$\begin{aligned} \text{Sim}(C_1, C_2) = {} & \omega_{\text{sex}} \cdot | \rho_{\text{sex},c_1} - \rho_{\text{sex},c_2} | + \\ & \omega_{\text{age}} \cdot | \rho_{\text{age},c_1} - \rho_{\text{age},c_2} | + \\ & \omega_{\text{level}} \cdot | \rho_{\text{level},c_1} - \rho_{\text{level},c_2} | + \\ & \omega_{\text{location}} \cdot | \rho_{\text{location},c_1} - \rho_{\text{location},c_2} |, \end{aligned}$$

其中:ω_{sex}, ω_{age}, ω_{level}, ω_{location} 是相应属性的权值. 构造神经网络,通过使每个事例中病人特征的相似度和

治疗方案的相似度保持一致,从事例数据中直接学习得到上述各个属性的权值,这些权值说明了相应属性的重要程度[6].

事例库中共有 8 个类型,211 条数据,取 180 条数据作为训练集学习 28 个支持向量机参数,余下的 31 条数据作为测试集. 8 个类型存在交叉区域,因此学习后无法使全体 2 值分类支持向量机的分类准确率达到 100 %. 构建好有向无环图支持向量机后,可学习确定阈值. 学习时首先赋予阈值 $\theta = 0$,然后以给定步长 $\Delta\theta = 0.01$ 递增,直到多类分类器达到要求的准确率. 实验中按不同的多类分类器分类准确率计算出相应的阈值 θ,结果如表 2 所示,其中实验 1 ~ 4 取不同的分类器准确率进行实验,实验 5 是直接在全体数据集中进行案例选择.

表 2 阈值 θ,多类分类器精度和案例选择效率关系表

	实验 1	实验 2	实验 3	实验 4	实验 5
多类分类器分类准确率 / %	85	90	95	100	—
θ值	0.71	0.88	1.08	2.39	—
案例选择时间 ms/ 1 000 次	136	150	155	164	182

实验结果表明,采用多类分类器后相似案例选择时间大大缩短. 而且由于多类分类器对原始数据进行了一次筛选,选择出的相似案例集合与不采用多类分类器的相似案例集合比较,相似案例显得更加均一. 案例选择效率与阈值 θ 之间有密切联系,θ 值越大,多类分类器对类型的分辨就越差,当 θ 足够大时,整个案例库将重新成为一个整体,任何新案例都在整个案例库中进行相似案例选择. 当 θ 为零时,多类分类器将为任何一个新案例指定一个确定的子类型,只在该类型的案例集合中进行相似案例选择.

上述实验的案例数据库规模比较小,相似案例选择效率已明显提高,当案例数据库规模增大时,选择效率的提高将更加明显.

总之,上述算法实现了相似案例自适应选择,在保证一定精度的前提下,提高了相似案例选择的效率. 阈值 θ 的选取可根据问题的重要性,依据分类器的准确率进行学习得到.

5 结 语

提高案例选择准确度和效率是基于案例推理系统中需要解决的重要问题. 本文应用有向无环图多类分类器对案例数据库中的数据进行分类. 进行相似案例选择时,首先确定与新案例可能相似的案例子类,并将这些子类包含的案例组成新的案例集合,从中选择相似案例. 该方法在保证案例选择准确度的前提下,提高了案例选择的效率. 最后通过实验证明该方法是有效的.

参考文献(References)

[1] Xu L D. Case-based reasoning[J]. Potentials of IEEE, 1993, 13(5): 10-13.

[2] Ramon Lopez de Mantaras. Case-based reasoning[J]. Lecture Notes in Computer Science, 2001, 2049: 127-145.

[3] Park Cheol-Soo, Han Ingoo. A case-based reasoning with feature weights derived by analytic hierarchy process for bankruptcy prediction[J]. Expert Systems with Applications, 2002, 23(3): 255-264.

[4] Policastro C A, Carvalho A C P L F, Delbem A C B. Hybrid approaches for case retrieval and adaptation[J]. Lecture Notes in Computer Science, 2003, 2821: 297-311.

[5] Chen D Q, Burrell Phillip. Case-based reasoning system and artificial neural networks: A review[J]. Neural Computer and Application, 2001, 10(3): 264-276.

[6] Juell P, Paulson P. Case-based systems[J]. Intelligent Systems IEEE, 2003, 18(4): 60-67.

[7] Finnie Gavin, Sun Z H. R5 model for case-based reasoning[J]. Knowledge-based Systems, 2003, 16(1): 59-65.

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

[9] Kuo R J, Kuo Y P, Chen K Y. Developing a diagnostic system through integration of fuzzy case-based reasoning and fuzzy ant colony system[J]. Expert Systems with Applications, 2005, 28(4): 783-797.

[10] 安金龙, 王正欧, 马振平. 一种新的支持向量机多类分类方法[J]. 信息与控制, 2004, 33(3): 262-267.
(An J L, Wang Z O, Ma Z P. A new SVM multiclass classification method[J]. Information and Control, 2004, 33(3): 262-267.)

[11] 唐发明, 王仲东, 陈绵云. 支持向量机多类分类算法研究[J]. 控制与决策, 2005, 20(7): 746-749.
(Tang F M, Wang Z D, Chen M Y. On multiclass classification methods for support vector machines[J]. Control and Decision, 2005, 20(7): 746-749.)

[12] Platt J, Cristianini N, Shawe-Taylor J. Large Margin DAG's for multiclass classification[C]. Advances in Neural Information Processing Systems 12. Cambridge: MIT Press, 2000: 547-553.

[13] Hsu Chih-Wei, Lin Chih-Jen. A comparison of methods for multicalss support vector machines[J]. IEEE Trans on Neural Networks, 2002, 13(2): 415-425.

[14] Boonserm Kijsirkul, Nitiwut Ussivakul. Multiclass support vector machines using adaptive directed acyclic graph[C]. IEEE/ INNS Int Joint Conf on Neural Networks. 2002: 980-985.