一种基于本体的辅助诊断治疗方法①

An assistant diagnosis and therapy approach based on ontology

裘 嵘 吴 敏 龙 军 徐德智 (中南大学信息科学与工程学院 湖南长沙 410083)

摘 要: 以医学领域本体为基础,为了在医疗领域获取知识优势,结合机器学习技术,采用基于信息熵的属性加权方法结合基于实例的方法对病例进行分类,提出了一种基于本体辅助诊断治疗的方法。对相关工作进行了分析比较,给出了一个辅助诊断治疗心肌梗塞的例子,以说明该方法的有效性。与一般的医疗诊断方法相比,具备更良好的共享能力和扩展能力,可提高辅助医疗诊断、治疗的准确性。最后给出了结论和下一步工作。

关键词: 本体 机器学习 辅助诊断治疗

1 引言

本体(ontology)最早来源于哲学,指关于存在的系统的描述[1,2],包括概念,关系,函数,公理,实例五个方面[3]。在计算机领域,引入本体的思想是为了实现知识共享和重用。知识工程学者借用这个术语及其基本思想,来建立本体知识库(简称本体, ontologies),其目的为了解决知识共享问题。

在实际应用中,本体提供某些专门知识领域的模型化概念和关系的结构性框架。本体支持生成专门领域的参考知识储存——领域知识库,供人和应用程序之间传输和共享这些知识[4,5]。

在医学领域知识的应用方面,研究方法层出不穷。

(1) 基于医学专家系统。难以表征问题域的深层知识、难以验证正确性和根据经验学习的能力较差。

(2) 基于规则的方法。不足在于可能出现把条件行为规则模式误用不当或出现任何现有规则都不适用的问题[6],启发性规则往往比较脆弱,不能处理残缺或意外数据,在领域知识的边缘附近会迅速退化。

(3) 基于案例的方法。基于案例推理的优点在于能够利用经验知识[6],但多不能包含更深层的领域知识,影响解释功能,庞大案例库可能受到存储计算平衡等问题的困扰,难以确定好的索引和匹配案例标准,单词检索和相似匹配算法需经过仔细的手工处理,抵消了基于案例推理所具有的知识获取优点。

(4) 基于本体的知识表示方法。例如统一医学语言系统 UMLS 等很多已经成为标准的医学领域内的各种基于本体的知识表示方法,为医学领域知识的应用提供了统一的共享的资源。

综上所述,根据以上在医学领域知识的应用方面国内外研究的经验和不足,我们提出一种基于本体辅助诊断治疗的方法。由于本体提供良好而统一的知识表示方法,所以利用本体构建知识库,有利于知识获取、分析和共享。

2 辅助诊断治疗的体系结构

本方法首先收集大量的电子病历原始数据,用规范词汇集来描述从电子病历中提取的重要属性,通过本体描述语言描述逻辑和推理引擎,得到统一、规范描述的知识表达,形成对本方法起重要作用的电子病历本体库。然后,经过规范表达和评估输入样例库。再使用分类器对病例的治疗效果进行分类,找出对治疗效果影响大小不同属性,给出辅助医生制定治疗方案的结果。新病例结果经过评估后形成修正后的正确结果,加入电子病历样例库,使本方法在辅助医生诊断、治疗的同时还具有良好的再学习功能。

① 国家杰出青年科学基金项目(60425310)。

3　基于本体从病例中挖掘信息辅助诊断治疗方法的构建

以下介绍具体构建本方法的步骤并举例说明：

3.1　领域知识表示

首先收集大量的的电子病历原始数据，电子病历的普及便于我们用程序自动学习，但电子病历的格式五花八门杂乱无章，所以，我们用规范词汇集来描述从电子病历中提取的重要属性，通过本体描述语言描述逻辑和推理引擎，得到统一、规范的知识表达，形成提取后的电子病历本体知识库。如果对于以前的普通手写文档病历就只能使用手工的办法提取相关知识。本体描述的是共享知识，即在医学领域内都共同接受的命题。在本方法中可以引用和补充建立疾病的症状、体征、检查、治疗和用药等诸方面和本方法相关的医学领域本体。对于已经作为标准公布的医学领域本体我们可以共享，实现统一的标准。

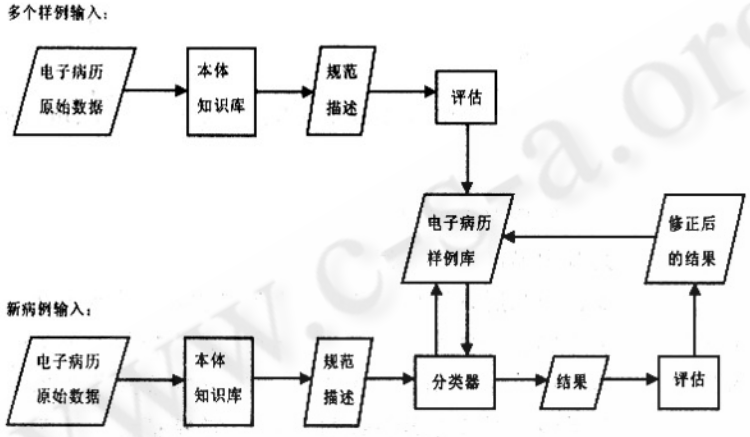

图 1　基于本体辅助诊断治疗的方法的体系结构

下面利用一个心脏病的病例对此方法详细说明。

在基于本体的领域知识表示阶段，要完成的任务在于提取某患者所患心脏病的电子病历中所涉及的基本情况、症状、体征、检查、治疗和治疗结果等方面的重要属性，引用和补充建立本体，形成提取后的电子病历本体知识库。一旦这些领域本体形成标准后，所有的应用都可以以此为依据来建立自己的知识库。我们在此摘取一个简单的子集做演示性说明。提取某患者 a 所患心脏病的电子病历，其中对本病的诊断具有决定性意义的重要属性，可以使用标记或其他方法将其用程序提取出来，将这部分知识转化为本体，建立应用于本方法的电子病历本体知识库。OWL（Web Ontology Language）是 W3C 推荐的语义互联网中本体描述语言的标准，以下表 1 采用较为简洁的描述逻辑语法来表示，所表示的知识均能转化成相应的 OWL 形式[9]。

以此类推，完成提取治疗部分重要属性的本体知识库，这就完成了领域本体知识表示阶段。有成为标准的相关医学领域本体可以直接引用，这样可以与国际统一标准作无缝的对接。

3.2　建立样例库

在建立样例库阶段，根据所建立的本体，将病人病例中的提取出的相关信息输入样例库。如上例中心脏病人病例信息在样例库中的规范表达为向量形式：< s1，s2，s3，s4，s5，t1，t2，t3，t4，r >。

其中，s1，s2，s3，s4，s5 表示症状，分别对应为：绞痛，坏死 Q 波，ST 波上抬，T 波改变；t1，t2，t3 表示治疗方法，分别对应为：尿激酶，硝酸甘油，阿斯匹林，GIK 疗法；r 表示结果，取值 true 为治愈，false 为未愈。

表 1　用描述逻辑语法表示的心脏病患者 a 电子病历中部分重要属性

Symptom	//定义症状概念
Symptom（symptom_a）	//定义患者 a 的症状为 symptom_a
character（symptom_a，angina）	//定义患者 a 的症状 symptom_a 的重要性质之一为绞痛
≥1 character ⊆ Symptom	//定义性质的 domain 为 Symptom

此过程具体完成方式可以利用现有的电子病历系统获取。通过程序自动学习从电子病历中提取的经过评估的新病例，具有良好的再学习功能。

3.3　病例分类

使用采用基于信息熵的属性加权方法结合基于实例的方法对病例进行分类，辅助医生制定治疗方案。在分类阶段，我们构建的分类器是采用基于实例的机器学习方法对病例的治疗效果进行分类，并结合采用基于信息熵的属性加权方法评价出对治疗效果影响大小不同的属性，给出辅助医生制定治疗方案的结果。

经分析，在基于实例的分类法中最近邻算法[10]是最适合本分类方法的。这种方法的基本思想是如果两个实例非常相似时，它们的分类也非常接近。采用这种方法的原因在于：第一、这种方法简单的把病例信息存储起来，知识获取过程简单而高效。第二、可以对目标函数实现良好的局部逼近。而通常的目标函数形式都非常复杂，难以实现好的全局逼近。

在最近邻算法中，假定所有的实例对应于 n 维空间中的点。即把实例表示为特征向量，即 $<a_1(x), a_2(x), \cdots, a_n(x)>$，其中 $a_i(x)$ 表示实例 x 的第 i 个属性值。本文中所提到的心脏病的实例，则为：<s1, s2, s3, s4, s5, t1, t2, t3, t4, r>。

最近邻算法如下

训练算法：

对于每个病例信息 $<x, f(x)>$，x 为实例，y 为实例所对应的值，f(x) 函数是求 x 与 y 的对应关系的目标函数，把它加入列表 training_examples

分类算法：

给定一个要分类的查询实例 x_q

在 training_examples 选出离 x_q 距离最近的 k 个实例，并用 $x_1, \cdots, x_k$ 表示

返回 $\tilde{f}(x_q) = \arg\max_{v \in V} \sum_{i=1}^{k} \delta(v, f(x_i))$，$\sum_{i=1}^{k} \delta(v, f(x_i))$ 为关于 v 的函数，$\arg\max_{v \in V} \sum_{i=1}^{k} \delta(v, f(x_i))$ 函数意义为在 v 遍历 V 中所有值的情况下，返回使 $\sum_{i=1}^{k} \delta(v, f(x_i))$ 最大的那个 v 的取值。

在该算法中，V 是分类的可能值的集合，对应本例即为治疗效果为治愈和未愈两个元素组成的集合。

最近邻算法的关键在于距离函数的定义，如前所述，两个点的距离一般定义为两个点的欧氏距离，但这种方法存在一个问题，即它把所有的属性对结论的影响都考虑作同样的重要，如果有少量重要的属性比其他属性对实例的分类结果影响更大，则近邻间的距离被大量的不重要属性所支配，导致分类被误导。本文中考虑的方案为对每种属性加权值，即定义该距离为 $d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} w_r (a_r(x_i) - a_r(x_j))^2}$。r 为实例的第 r 个属性，具体权值 $w_r, r = 1, \cdots, n$ 取什么值呢。本文结合决策树算法提出一种启发式方法[10]。即根据各个属性的信息增益来做 $w_r, r = 1, \cdots, n$ 的启发式估计。

在信息论中，设有一个实例集 S 和一个可能的分类的集合 V，对于要把实例集分成几个 V 中的类时，该实例集相对于这个分类的熵为：

$$Entropy(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

其中 p_i 是 S 中属于某个类别的比例。当使用某个属性 A 来将样例集分成几个小的实例集后，定义信息增益为：

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

其中 Values(A) 是属性 A 所有可能值的集合，S_v 是 S 中属性 A 的值为 v 的子集。则 Gain(S,A) 的意思是由于知道 A 而导致的期望熵的减少。这个量反映了各个特征分类实例的能力，也就是这个特征对分类的重要性。当 Gain(S,A) 越大，表明该特征对分类的贡献越大。对应距离函数的定义上，该特征细微的变化可能导致分类的不同，所以该属性的权值也就越大。

取 $w_r = (Gain(S, a_r))^2$。可以满足上述说明。当 $Gain(S, a_r)$ 越大，w_r 越大，对距离的影响也就越大。

4 实验结果与分析

为了验证这种方法在本体系结构中的有效性，需要将它的实验结果与其它经典的分类算法的实验结果进行比较分析。

验证中针对关于心脏病的辅助诊断问题。问题的背景是基于这两种疾病的电子病历样例库，来验证上文提到的几种算法的性能表现。样例库的来源都是标准公开的 UCI 数据库，这样可以在验证时保证样例结果的正确性。

在 Weka 软件中，可以调用和基于熵的属性加权最近邻算法进行比较的其他机器学习算法。本测试采用 10 – fold 交叉验证的方法，即把样例库随机均分为 10 个集合，每次取一个集合作为测试集，其它的所有样例都作为训练集，分别使用最近邻算法 IBK、朴素贝叶斯算法、贝叶斯信念网算法、J48 决策树算法、AdaBoostM1 算法对测试集的目标属性进行预测，再与测试集原有的正确的目标属性值进行比较，得出分类正

确率,从而判断出算法性能的优良程度。

4.1 心脏病辅助诊断问题

全世界三分之一的人口死亡是因心脏病引起的,而我国每年耗费在心脏病的诊断上的资源相当大,为了实现充分的资源共享和智能诊断,使用基于语义 Web 的电子病历本体结合辅助诊断的分类方法对心脏病的正确诊断和节约医疗成本具有非常重要的现实意义。

测试采用 10 – fold 交叉验证的方法。即把样例库随机均分为 10 个集合,每次取一个集合作为测试集,其它的所有样例都作为训练集,一共做 10 次。取平均的测试结果作为最终结果。

实验过程如下:

首先,Weka 软件中打开 heart – statlog 样例库,左侧窗口显示样例库中各样例的所有属性特征,右侧窗口显示某一属性特征的统计特性,然后,进入分类界面,选择朴素贝叶斯分类算法,测试方法使用 10 – fold 交叉验证,运行结果如下所示。

用朴素贝叶斯分类算法分析 heart – statlog 样例库的统计信息项目:正确分类的样例为 226 、错误分类的样例为 44、样例总数为 270、正确分类率为 83.7037% 、错误分类率为 16.2693% 。用朴素贝叶斯分类算法分析 heart – statlog 样例库两种分类的详细精确度,如:分类的真实正确率、真实错误率、精确率和召回率等,这样可以全面了解分类器在样例集上的分类效果。比较用朴素贝叶斯分类算法分析 heart – statlog 样例库,得出对于两种分类判断出的真值与其分类的真值。

以上验证结果中,正确分类率是最重要的因素,因此,比较算法性能时,只取正确分类率作为比较因素。

依照上例,分别调用 IBK 算法、贝叶斯信念网算法、J48 决策树算法和 AdaBoostM1 算法,分别得出他们的正确分类率。

为了将本文提出的基于熵的属性加权最近邻算法与朴素贝叶斯分类算法、IBK 算法、贝叶斯信念网算法、J48 决策树算法和 AdaBoostM1 算法进行比较,使用 java 语言对本算法进行开发,写成程序名为 medicinelearning. java 的文件,开发过程中调用了与 Weka 中一致的统计信息项目,得出对应这些统计信息项目的相应结果。将本算法生成的程序运行在 Eclipse 控制台上,对 heart – statlog 样例库进行分析的结果如下所示。

用基于熵的属性加权最近邻算法分析 heart – statlog 样例库信息项目:正确分类的样例为 228 、错误分类的样例为 42、样例总数为 270、正确分类率为 84.4444% 、错误分类率为 15.5556% 。

在 Weka 中使用基于熵的属性加权最近邻算法分析 heart – statlog 样例库,对于两种分类,比较用该算法判断出的真值与其分类的真值,得到分类矩阵的数据。

我们通过在 Eclipse 的控制台窗口运行基于熵的属性加权最近邻算法,给出了基于熵的属性加权最近邻算法的性能评估,在输出界面中选取了与 Weka 软件输出相同的项目,以便进行全面的比较,当然,我们考察的最重要的属性仍然是正确分类率。

4.2 结果比较

整理比较以上各种机器学习方法对关于心脏病电子病历的样例库进行分析的效果,如图 2 所示。

从上图可以看出,本文提出的基于熵的属性加权最近邻算法分类正确率最高,其余算法的实验效果在不同程度上低于它。它比同为最近邻算法的 IBK 算法分类正确率高,是因为它用了基于熵的属性加权的距离度量,同时也说明最近邻算法中的距离度量对于算法的分类正确率影响很大。

5 结论

本文中提出的基于领域本体知识库的辅助诊断、治疗的方法的意义在于:提出了基于信息增益的最近邻算法中的启发式距离度量定义,有效地解决了维度灾难,使近邻间的距离不会被大量的不相关属性所支配。本方法从使用医学领域本体作知识库的表示方式,获取知识优势出发,结合机器学习技术给出有效的医疗辅助诊断方案,比一般的医疗诊断系统具备更良好的共享能力和扩展能力,提高辅助医疗诊断、治疗的准确性。可广泛地应用于基于语义 Web 的医疗诊断、医学教学及分析药物的有效性等方面,并能实现不同医疗应用系统间的良好对接。为进一步构建可扩展、可重用、更广泛的语义 Web 智能医学系统工程提供有益参考。

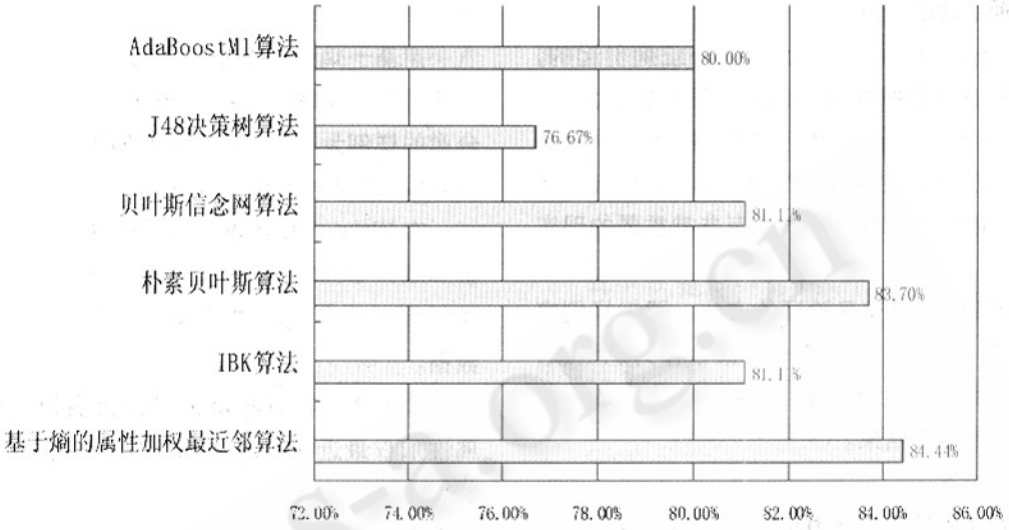

图 2　在心脏病辅助诊断问题中各种算法的效果比较

Reference

1 T. R. Gruber, A translation approach to portable ontologies. Knowledge Acquisition[J], 1993, 5(2):199 ~ 220.

2 T. R. Gruber, Towards principles for the design of ontologies used for knowledge sharing. International Journal of Human - Computer Studies [J], 1995: 907 ~928.

3 Michael K. Smith, Chris Welty and Deborah McGuinness, OWL Web Ontology Language Guide, W3C Recommendation[DB/OL], http://www.w3.org/TR/owl-guide/, 2004.

4 Nils J. Nilsson, Artificial Intelligence: A New Synthesis, China Machine Press & Morgan Kaufmann Publishers[J], 1999, 215 ~316.

5 Stuart Russell and Peter Norvig, Artificial Intelligence: A Modern Approach, PEARSON EDUCATION NORTH ASIA LIMITED and PEOPLE'S POSTS & TELECOMMUNICATIONS PRESS[J], 2002, 221 ~226.

6 Soumeya L. Achour, Michel Dojat, Claire Rieux, Philippe Bierling, and Eric Lepage: A UMLS - based Knowledge Acquisition Tool for Rule - based Clinical Decision Support System Development. J Am Med Inform Assoc[J], 2001,8(4):351 ~360.

7 Althoff KD, Bergmann R and Wess S, Case - based reasoning for medical decision support tasks: the Inreca approach., Artif Intell Med[J], 1998 Jan,12(1):25 ~41.

8 (美)卢格尔著, 史忠植译,《人工智能:复杂问题求解的结构和策略(原书第4版)》, 北京, 机械工业出版社, 2004.

9 Horrocks, I., Patel - Schneider, P. F., van Harmelen, F., From SHIQ and RDF to OWL: The making of a web ontology language, Journal of Web Semantics[J], 2003, issue 1, volume 1.

10 (美)米歇尔著, 曾华军译,《机器学习》, 北京, 机械工业出版社, 2003.