支持向量机算法在中医证候信息分类中的应用*

□杨小波** 梁兆晖 罗云坚 (广东省中医院 广州 510120)
陈 玲 (中山大学信息科学与技术学院 510275)

摘 要:目的:探讨带先验知识的支持向量机(P- SVM)数据挖掘算法在中医证候信息自动分类中的应用。方法:以中医证候数据库收集的30余万条中医证候文献信息作为训练和测试数据集,以中医专业知识作为先验知识,将样本集置信度通过带权分类间隔导入SVM模型中进行分类,计算其分类置信度。结果:在有中医专业知识的情况下,中医证候信息分类的正确率得到了很大的提高,正确率约为95%。结论:P- SVM算法是统计学习理论在小样本数据集中较成功的应用,能对中医证候信息进行有效的分类,实现了数据挖掘技术在中医证候信息研究中的应用。实验表明P- SVM算法能把先验知识与训练样本中的信息量很好地结合起来,对一种对中医证候信息进行正确分类的有效算法。

关键词:中医证候 数据挖掘 信息技术 支持向量机

一、研究背景

当代科技发展主要的趋势,是以信息技术革命为中心的当代科技革命正在全球蓬蓬勃勃的兴起,它标志着人类从工业社会向信息社会历史性的跨越[1]。证候研究作为中医药学的核心问题之一,其信息化进程备受关注。自20世纪50年代,证候研究经历了从宏观到微观、从整体到局部的多层次研究,相关的研究数据逐年增多,然而由于中医药术语的规范化程度不高,加上资料数量庞大,造成文献研究效率低下,严重制约了研究人员对相关信息的利用。对此,中医药事业"十五"计划明确提出:"要积极推进中医药信息化,引进信息技术和设备,加强中医药信息化基础建设,加快信息技术在中医药领域的广泛应用,提高信息网络的使用效益;积极开发利用信息资源,为中医药事业发展提供支撑和条件"的要求。

目前文本挖掘、人工智能等技术已逐渐广泛地应用于情报科学、图书馆管理等信息管理领域,并取得了显著成果,而中医药领域的信息管理仍以人工标引等手段为主,证候信息等的自动分类技术的研究和应用基本上仍处于空白状态,限制了中医药信息管理和利用的效率。本文以广东省中医药管理局课题-"中医证候专题信息数据库的开发与应用"为基础,利用中医证候专题数据库(内含约30万条数据)作为训练和测试数据集,以中医专业知识为先验知识,探讨带先验

收稿日期:2006-03-13
修回日期:2006-07-24

* 广东省中医药局资助课题(1040014):中医证候专题数据库的开发与利用,负责人:黄燕;国家科技部"十五"攻关计划课题(2004BA721A02)急性缺血中风辨证规范和疗效评价的示范研究,负责人:杨小波。

** 联系人:杨小波,副主任医师,主要研究方向:中医证候研究,中医科研方法学研究,Email yangxiaobomd@163.com

知识的支持向量机(P- SVM)数据挖掘算法在中医证候信息分类中的应用。

二、相关理论背景

统计学习理论(Statistical Learning Theory或 SLT)是一种专门研究小样本情况下机器学习规律的理论,。该理论针对小样本统计问题,其统计推理规则不仅考虑了对渐近性能的要求,而且追求在现有有限信息条件下得到最优结果[3]。而且,它能将很多现有方法纳入其中,有望帮助解决许多原来难以解决的问题。支持向量机(Support Vector Machine或 SVM)方法是近年来兴起的基于统计学习理论的分类和预测算法[3][4],它建立在统计学习理论的 VC维理论和结构风险最小原理基础上的,根据有限的样本信息在模型的复杂性和学习能力之间寻求最佳折衷,以期获得最好的推广能力(generalization ability)。随着人们对统计学习理论研究的深入和 SVM 模型的改良,把应用背景知识引入到 SVM 模型中以及把 SVM 与各种应用相结合产生的各种解决方案。然而,尽管人们在先验知识引入 SVM 模型方面做了很多工作,但仍有一些问题尚待解决,有时难以取得令人满意的效果。

本文将中医证候背景知识引入 SVM 模型并对之进行训练方面进行了一些尝试,通过改进现有算法,对中医证候数据进行分类。基本思路是从中医证候的先验知识出发,定义对于中医证候样本的置信度,也就是样本在实际应用可能属于某一类的可能性,进而提出带置信度的 SVM 模型 P- SVM,然后应用带先验知识的间隔最小优化算法(P- SMO)对带置信度属性的数据集进行训练,得出 P- SVM 分类器。

三、实验设计与结果分析

SVM 的核心为核函数,反映了特征空间的性质,尤其对小样本学习十分有效,但存在运算量较大,算法时间复杂度会较高等问题[5]。针对这一缺点,可以通过引入先验知识加以解决。在先验知识与 SVM 模型结合方面,有学者提出 WM SVM 模型及一种基于间隔最小优化算法(SMO)的训练算法,通过先验知识计算分类标识的置信度,以确定样本离分类面间隔的大小,提出一种带权距离的思想,也是本实验设计的理论基础。所谓带权分类间隔,即将样本置信度引入分类器模型中,在原始的训练数据集中,除了数据属性之外增加一个类标签,用于表明该样本所属类别。SVM 算法对分类面有以下要求:如果某一样本的置信度越大,则它和分类面的距离就应相对于没有考虑置信度时的距离要大,反之就要小。也就是说,在实验中,要考虑的距离是一种带权的距离,而这个权是样本的置信度。图 1中表达了这种思想,正方形和圆形分别表示了两类样本,而它们的面积则代表样本属于该类的置信度的大小。从图 1中可以看出,对于置信度较大的样本,它离分类面的距离也应该相对较大。

在广东省中医药管理局资助课题“中医证候信息数据库系统的设计与开发”的研究中,我们尝试将 SVM 算法模型运用于中医证候数据的分析,以中医证候数据库作为训练和测试数据集对算法的性能进行测试。具体步骤为首先建立中医证候专业知识库,然后此作为先验知识规则用于训练数据集,生成置信度属性。这些带置信工属性的数据作为 P- SMO算法的输入,生成 P- SVM 分类器模型。最后我们在不同训练样本数的情况下对分类的正确率进行比较。

所使用的数据来自中医证候专题数据库,其中收集了国内自 1978年至 2004年的大部分中医文献,总数约 30万条。数据库中以单条文献为单位进行数据组织,每条文献记录具有若干属性,根据经验,我们认为标题、主题词、副主题词、关键词、中文摘要这几个字段能够标识出文献所属的主题分类。其中每条记录的主题词和副主题词属性是匹配使用的,最能反映该条记录的主题信息。一条数据可以同时标引数个主题

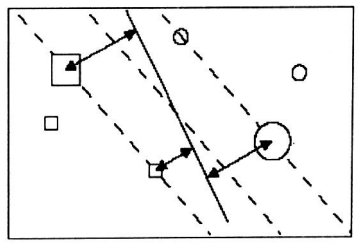

图 1 带权分类间隔的支持向量

词,每个主题词又可与一定数量的副主题词相匹配,共同组成了该条记录的主题信息。利用 P- SVM 来对数据进行分类,使分类器学习到特定主题特征信息,以便迅速从海量数据中找出相关的数据信息。

实验数据的具体处理流程如下:以“脾气虚”为例,它在主题词树中的位置为“证候→脏腑证候→脾和胃证候→脾系证候→脾虚→脾气虚”,可匹配副主题词包括:血液、脑脊髓液、化学诱导、分类、先天性、并发症、膳食疗法、诊断、药物疗法、按摩疗法、气功疗法、穴位疗法、针灸疗法、中西医结合疗法、中医病机、中医药疗法、中药疗法、中医疗法等。现使用 50条与该主题相关的文献记录作为训练数据,加入了使用规则表示的先验知识(Prior Knowledge)计算出来的置信度,表 1所示:

每个样本给予初置信度 80%,若经过先验知识处理后置信度 > 100%,则按 100% 计算,若置信度 < 0,则舍弃该样本。

对于不同的类别,可以有不同的先验知识。针对每一种分类可以提取出不同的先验知识,于是对于每一种分类都可以有一个带先验知识的 SVM 对之进行识别。50条数据的训练样本集中有 21条被标记为“属于此类”的文章,其余则标记为“不属于此类”的文章。通过先验知识的处理,增加其分类置信度的属性。再对 SVM 进行训练。

在使用训练之后的 SVM 对其他 2000条测试数据进行识别,发现其中大约有 95% 的测试数据能够被准确识别,可以预测,如果能引入更多更全面的专业知识作为先验知识,SVM 模型的分类准确率会得到大幅度的提升。图 2和图 3显示了没有先验知识和有先验知识的情况下,不同大小的训练样本集对于分类器推广能力的影响:

图 2和图 3提示,先验知识对于分类器的推广能力有重要影响。图 3呈一条折线,是因为每一条先验知识对分类正确性的贡献均不相同,但可以肯定的是,使用越多越全面的专业知识作为先验知识,将会得出越高的分类正确率。

四、结 论

本文中应用带先验知识的支持向量机 P- SVM 对中医证候信息进行自动分类研究,采用一种改进的 SMO 算法对 P- SVM 模型进行训练,再通过由专业人员提供的中医证候知识,对训练样本集进行置信度的

表 1 由专家知识计算置信度

规则	置信度变化
主题词包含“脾气虚”	+ 20%
主题词包含“脾虚”	+ 15%
主题词包含“脾系证候”	+ 10%
标题包含“脾气虚”	+ 5%
标题含“脾虚”	+ 3%
主题词含“肾气虚”	- 10%

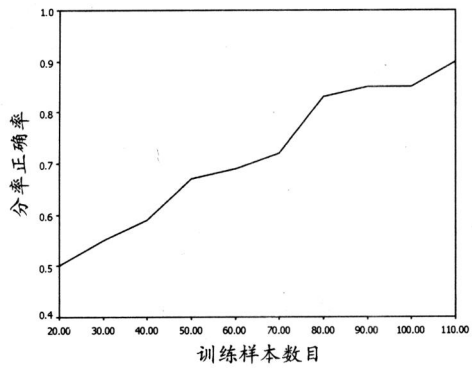

图 2 训练样本数据训练样本数分类正确率的影响

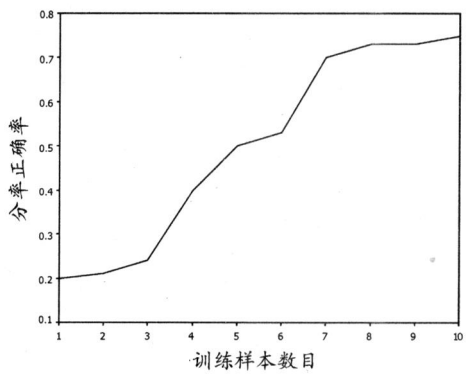

图 3 专家知识条目数分类正确率的影响

计算,然后使用 P-SMO 算法确定 P-SVM 模型的参数,得出最终分类器并应用到中医证候数据的分类中。实验结果表明,在有先验知识的情况下,分类的正确率有比较大的提高,可视为 SMV 算法在中医证候信息研究中的一个较为成功的应用,提示统计学习理论在中医证候信息的管理和分类中有良好的应用前景。

对于先验知识表达的研究是当前学术界的研究热点之一,在本研究中仅使用了较简单的方法引入先验专业知识,表现为样本到分类面距离权值的单一度量。在今后的研究中,我们会研究一种复合度量,以向量形式表示的度量,以表达更复杂的先验知识,进一步提高算法的推广能力。

参考文献

1 徐冠华. 当代科技发展趋势和我国的对策. 中国软科学, 2002, (5): 1~12

2 关彤. 浅述中医药信息管理. 中国中医药近代远程教育, 2003, 4: 20~22.

3 V. Vapnik. The Nature of Statistical Learning Theory. Springer, N. Y., 1995.

4 V. Vapnik著, 许建华, 张学工译. 统计学习理论. 北京: 电子工业出版社, 2004.6

5 吴涛, 贺汉根, 贺明科. 基于插值的核函数构造. 计算机学报, 2003, 26(8): 990~996.

6 印鉴, 张钢, 陈忆群, 等. Multi-events Analysis for Anomaly Intrusion Detection. International Conference of Machine Learning and Cybernetics, 2004.

P-SVM Applications in TCM Syndrome Classifications

Yang Xiaobo, Liang Zhaohui, Luo Yunjian

Guangdong Provincial Hospital of TCM, Guangzhou 510120, *China*

The paper explores possible applications of Prior knowledge Support Vector Machine (P-SVM) based data mining algorithm in an automatic TCM syndrome classification system. In the study, a TCM syndrome database containing some 300,000 medical records is used as a sample set for algorithm training and test. In addition, a range of TCM syndrome theories are incorporated into a prior knowledge set. The sample set is made part of the SVM model, with weighted sequence for classification. The confidence value for each result is also calculated on an individualized basis. It is proved that with prior TCM knowledge, the accuracy of the automatic TCM syndrome classification system can be raised to a level as high as 95%. It is concluded that P-SVM has made a successful application of statistical learning theory (SLT) to the given samples, though limited in number, which heralds an effective approach for improving automatic TCM syndrome classification, and proves the applicable features of data mining in TCM syndrome researches. The results show that P-SVM marries prior knowledge with the trained samples, and it is an effective algorithm for TCM syndrome classification.

Keyword: TCM Syndrome; data mining; information technology; support vector machine

(责任编辑:张述庆,责任编审:张志华,责任译审:邹春申)