[文章编号]1000-1832(2015)01-0077-06
[DOI]10.16163/j.cnki.22-1123/n.2015.01.015

基于SVM的一种医疗数据分析模型

田宇驰,胡 亮
(吉林大学计算机科学与技术学院,吉林 长春 130012)

[摘 要] 基于SVM分类算法和Web服务框架,提出了一种医疗数据分析与疾病预测模型,改进了医疗数据分析系统与医院数据库之间的数据传输协议.采用该模型与长春某三级甲等医院合作,获取了总共1 695条病人电子病历数据与病人疾病信息作为实验数据,并在医疗数据分析系统中进行数据挖掘分析.通过数据条数的变化和对属性的控制来测试设计的数据分析模型和改进的数据传输协议的传输效率.实验表明,在传输数据之前对数据进行预处理并且通过特征选择算法进行降维处理有助于提高整个系统的医疗数据传输效率和预测准确度.

[关键词] Web服务;SVM;电子病历;特征选择
[中图分类号] TP 393 [学科代码] 520·3040 [文献标志码] A

医院现在广泛采用电子病历系统,通过电子病历系统对病人的健康状况进行长期跟踪.一部分医院除了在病人发病时对病人进行治疗外,还借助病历记录的医疗数据和疾病预测系统对病人的健康状态进行分析,对疾病进行预测,从而达到提前预防和及时治疗的目的.

电子病历的研究在计算机和医学领域同时得到了很高的关注.电子病历是医疗数据的数字化信息,它包括病人的健康状况、治疗过程和影像信息等.这些数字化信息的出现催生了医疗数据信息的管理与共享[1]和医疗数据的分析与疾病预测2个领域的研究.

医疗数据的数字化使得用计算机辅助医学进行数据采集、分析与疾病预测成为可能.近几年,软计算方法、支持向量机(SVM)和人工神经网络(ANN)已经应用在疾病预测上.如SVM算法用于诊断缺血性心脏病[2];SVM和鉴别集算法的结合能诊断老年痴呆[3];ANN算法用于进行动脉粥样硬化和心血管疾病的早期预防[4]等.

本文提出了一种基于SVM分类算法和Web服务框架的医疗数据分析与疾病预测模型.该模型先对电子病历中部分医疗数据进行特征选择和降维处理,再将这些维度的所有数据按照是否感染某种疾病进行分类并作为SVM算法的训练数据集,最后使用训练得到Lagrange乘子,对病人的疾病进行分析预测.这一模型为数据挖掘中的软计算方法应用在医疗数据的分析中提供了方案.使用医疗数据对模型进行测试发现,特征选择方法对部分医疗数据进行处理后,医疗数据的传输效率明显高于原始模型.

1 医疗数据分析模型

将医院、医疗保健组织等数字化的医疗数据以特定的格式、协议发送到医疗数据分析模块进行分析与疾病预测.

医疗数据分析模型组成见图1.

[收稿日期] 2014-02-24
[基金项目] 国家自然科学基金资助项目(61103197,61073009);国家高技术研究发展计划项目(2011AA010101).
[作者简介] 田宇驰(1989—),男,硕士研究生;通讯作者:胡亮(1968—),男,教授,博士研究生导师,主要从事分布式系统和网络与信息安全研究.

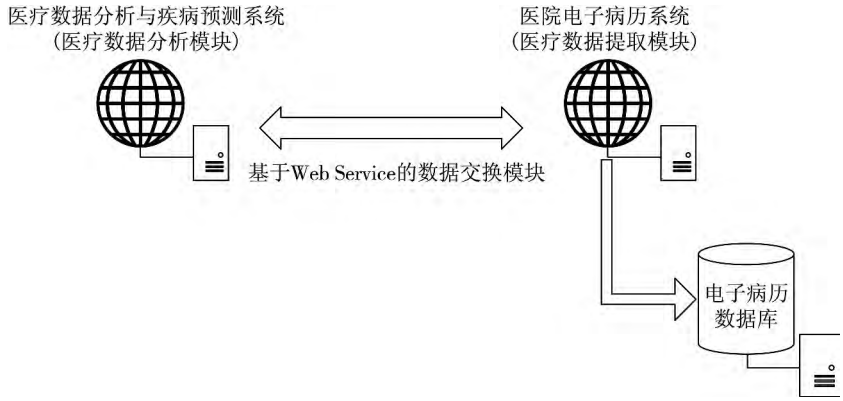

图1 医疗数据分析模型

医疗数据提取模块:该模块由医院电子病历系统负责实现,我们使用 openEHR 系统作为医院电子病历系统,并在 openEHR 中实现医疗数据的提取功能. openEHR 系统是一个开源、灵活的电子病历系统,支持 HL7 卫生信息交换标准[5]. 很多医疗健康组织、政府和学术科研单位都使用 openEHR 进行开发和科研工作. 如一种基于 openEHR 的患者病历数据管理模型、openEHR 等许多开源的电子病历平台的对比与评估和基于 openEHR 的档案建模等[6-8].

数据交换模块:基于 Web 服务的数据交换模块使用医疗数据通信协议实现医疗数据分析模块与医疗数据提取模块的数据交换. Web 服务[9]是一个平台独立、松耦合的 Web 应用程序. 由于 Web 服务的跨平台特性,许多模型与框架是基于 Web 服务构建的,如基于 Web 服务集成分布式资源[10]和数据流分析测试[11]等. 在本文提出的医疗数据分析模型中,使用 Web 服务来连接医疗数据分析模块和医疗数据提取模块. 医疗数据提取模块作为 Web 服务的服务端,实现的方法包括存取数据、数据预处理、序列化等,改进后的模型要求实现指定维度,指定属性数据的读取. 本文提出的医疗数据分析模块作为 Web 服务的客户端,通过 HTTP 服务向数据提取模块请求获取数据,并对数据进行预处理.

医疗数据分析模块:我们使用 Caisis 开源平台作为医疗数据分析与疾病预测系统实现这一模块. Caisis 是基于 Web 的开源癌症数据管理系统,一些临床医学研究使用 Caisis 系统管理和归档数字显微图像[12],通过向 Caisis 系统中添加特征选择和 SVM 算法,使用 SVM 算法对医疗数据进行分析和疾病预测,因此使用的特征选择算法需要基于 SVM,可以提高数据分析和疾病预测过程的效率和准确度.

2 数据分析模块与算法

2.1 SVM 算法

SVM 算法最初是由 Vapnik 等人在 1995 年提出的一种可训练的机器学习算法[13]. 依据统计学习理论、VC 维理论[14-15]和结构风险最小化理论[16],从一定数目的样本信息在学习能力和复杂度(对训练样本的学习程度)中找到最佳折中,以期望获得最好的推广能力(或称泛化能力).

SVM 一般用在二分类问题上,二分类问题的形式化定义:对于给定的训练集$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\in \boldsymbol{R}^n\times\mathbf{R}$,其中 $x_i\in\boldsymbol{R}^n$,$y_n\in\{-1,1\}$,$i=1,\cdots,n$,根据训练集在 $\boldsymbol{R}^n$ 空间上找出一个实值函数 $g(x)$,使得指示函数(或称决策函数、分类函数)

$$f(x)=\mathrm{sgn}(g(x)). \tag{1}$$

在对 x_i 进行分类的时候,取一个阈值 ε,当 $g(x_i)>\varepsilon$ 时 $\mathrm{sgn}(g(x_i))$选择一个类别;当 $g(x_i)<\varepsilon$ 时 $\mathrm{sgn}(g(x_i))$选择另一个类别[17].

二分类问题的本质就是获得一个可以将 $\boldsymbol{R}^n$ 空间分成两部分的实值函数 $g(x)$. 如果 $g(x)$为线性函数,则分类器就是线性分类器;如果 $g(x)$为非线性函数,则分类器就是非线性分类器. 对于线性分类问题,$g(x)=(\boldsymbol{w}\cdot\boldsymbol{x})-b$(其中$(\boldsymbol{w}\cdot\boldsymbol{x})$是向量 w 与向量 x 的内积),可以将 2 个类别无错误的分割开来,所表示的分隔函数被称为超平面[18-19].

对于线性分类问题，2 个类别中间的那条超平面可能并不是唯一的，我们需要一个指标来评价分类函数的优劣. 其中分类间隔(margin)是 SVM 分类中的一个指标，这个指标通过分类间隔大小来评价一个超平面是否是最终的最优超平面. 下面给出最优超平面的定义：

假设训练集数据$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\in \boldsymbol{R}^n\times\mathbf{R}, y\in\{-1,+1\}$可以被超平面，有

$$(\boldsymbol{w}\cdot\boldsymbol{x})-b=0. \tag{2}$$

且超平面与每个分类中最近的样本点之间的距离(分类间隔)最大，因此这个超平面为最大间隔超平面，即最优超平面.

H 是分类面，而 $H1$ 和 $H2$ 是平行于 H，且过离 H 最近的两类样本的直线，$H1$ 与 H，$H2$ 与 H 之间的距离就是几何间隔.

这里的分类间隔指的是 $H1$ 与 $H2$ 之间的几何间距.

下面给出几何间距的定义：

对于给定训练集$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\in \boldsymbol{R}^n\times\mathbf{R}$，其中 $x_i\in\boldsymbol{R}^n, y_i\in\{-1,1\}, i=1,\cdots,n$. 我们定义样本点$(x_i,y_i)$到超平面的间距

$$\delta_i=y_i(\boldsymbol{w}\cdot x_i+b)=|g(x_i)|. \tag{3}$$

在归一化处理的过程中，公式中的 w 和 b 被替换成$\frac{w}{\|w\|_p}$和$\frac{b}{\|w\|_p}$，其中

$$\|w\|_p=\sqrt[p]{w_1^p+w_2^p+\cdots+w_n^p}.$$

几何间距表达式

$$\Delta_i'=\frac{1}{\|w\|_p}y_i(\boldsymbol{w}\cdot x_i+b)=\frac{1}{\|w\|_p}g(x_i)=\frac{1}{\|w\|_p}\delta_i. \tag{4}$$

为了构造最优超平面，需要在约束条件 $y_i(\boldsymbol{w}\cdot x_i+b)\geqslant 1, i=1,2,\cdots,n$ 的情况下，最小化函数

$$\Phi(x)=\frac{1}{2}(\boldsymbol{w}\cdot\boldsymbol{w}). \tag{5}$$

这是一个在线性约束条件下凸二次规划问题，根据 Lagrange 求解方法[20]，通过构造 Lagrange 乘子 $\alpha=(\alpha_1,\alpha_2,\cdots,\alpha_n)$，得到最后的判定函数

$$f(x)=\mathrm{sgn}[\sum y_i\alpha_i(x_i,\boldsymbol{x})-b]. \tag{6}$$

2.2　基于 SVM 的医疗数据分析模块

将 SVM 分类算法应用到医疗数据分析模块中，进行疾病预测. 如图 2 是基于 SVM 的医疗数据分析模块，通过数据交换模块获取原始组数据(患病病人医疗数据和对照组病人数据). 通过特征选择过程输入到 SVM 分类器中进行训练，训练后可以对新的医疗数据进行分析预测.

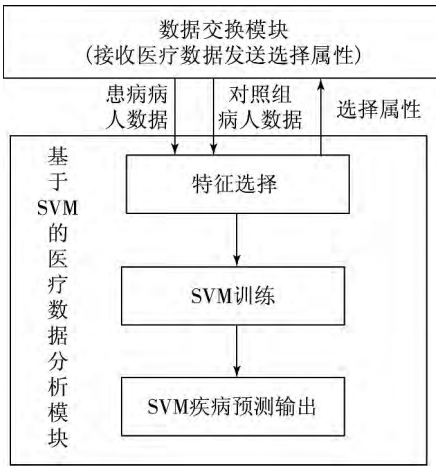

图 2　基于 SVM 的医疗数据分析模块系统架构

3　改进的医疗数据交换模块

3.1　医疗数据交换模块

在原始的医疗数据交换模块中，数据请求原语只由 4 条通信原语组成(如图 3 所示). 图 3 由原始医疗数据分析模型的 3 个模块构建，其中在医疗数据分析模块与医疗数据提取模块之间的 4 条通信原语包括 2 条请求和 2 条应答. 由于医疗数据的维度极大，属性很多，但是在预测某个疾病时，只有很少的一部分属性会对分类预测产生影响. 这样的全部维度的数据都需要传输，浪费了时间，降低了数据传输效率，影响了医疗数据分析模块的算法效率.

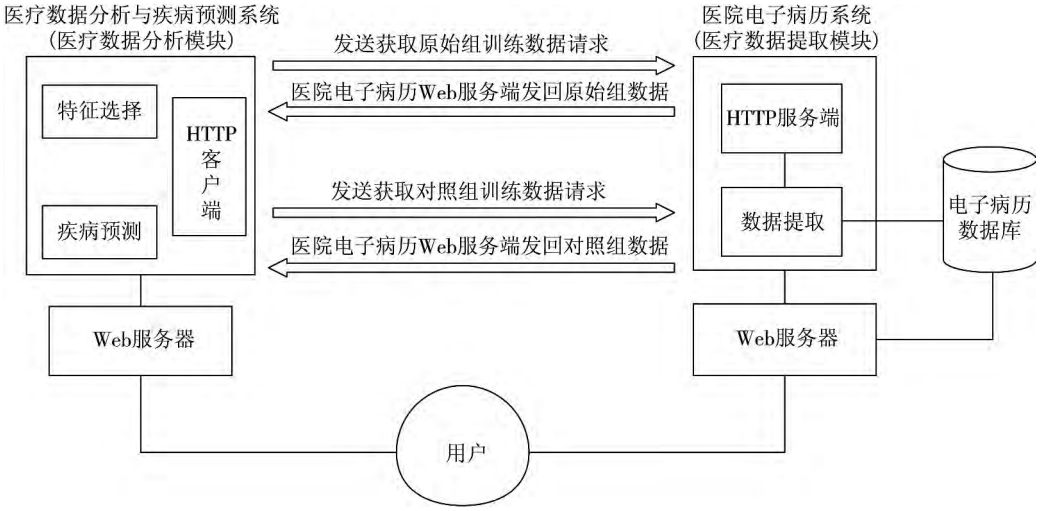

图3 医疗数据分析模型3个模块的构建

3.2 改进的医疗数据交换模块

在改进的医疗数据交换模块中,在数据传输协议中增加了4条原语(见图4).在每条原语中不仅有医疗记录条数的要求,还包括对所请求医疗数据维度和属性的具体说明.医疗数据分析模块先请求一小部分全部维度的数据,对这小部分数据进行特征选择.然后医疗数据分析模块只请求特征选择出来的对预测相关的属性的剩余所有医疗数据.最后通过SVM分类算法进行训练和预测.在新的医疗数据交换模块中,大部分数据中只有小部分相关属性被传输到数据分析模块,极大地减少了数据传输总量,也同时增加了分析模块预测算法的效率.

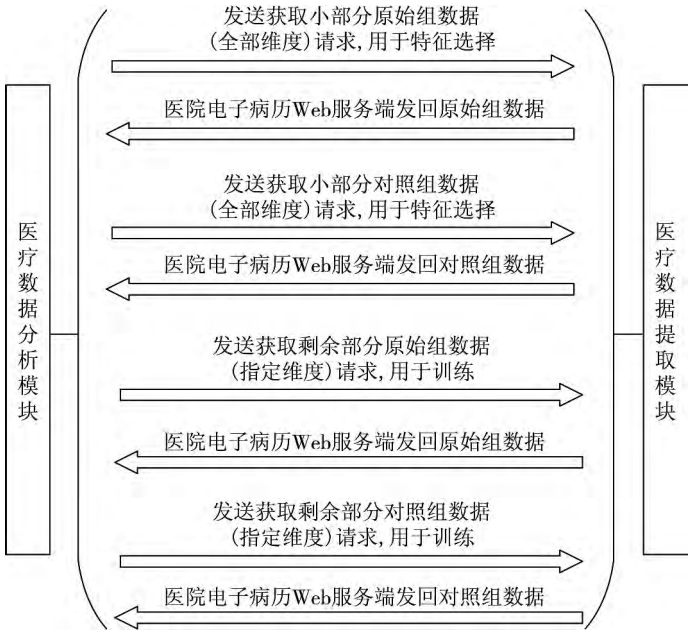

图4 改进医疗数据传输协议后的模型

4 医疗数据模型的对比评估

4.1 数据传输效率的计算

在实验中假设特征选择出的结果为141个属性中的22个属性,这22个属性是运行特征选择算法得到的结果,假设Caisis中训练需要的原始组数据和对照组数据总条数取相同的N值.

在原始的设计模型中,在Caisis和openEHR分别发送请求和接收数据的过程中,分别针对原始组

和对照组病人数据记录了 8 次系统时间 t_1-t_8. 原始设计模型的总数据传输时间可以表示为

$$T=[(t_2-t_1)-(t_4-t_3)]+[(t_6-t_5)-(t_8-t_7)]. \tag{7}$$

改进的设计模型中，假设 Caisis 在特征选择过程中需要的原始组和对照组数据条数相等，设为 n，训练过程需要的数据总数为 $N-n$. 在实验中分别针对原始组和对照组数据 16 次记录了系统时间 g_1-g_{16}. 改进后的设计模型的总数据传输时间可以表示为

$$G=\{[(g_2-g_1)-(g_3-g_4)]+[(g_6-g_5)-(g_8-g_7)]\}+$$
$$\{[(g_{10}-g_9)-(g_{12}-g_{11})]+[(g_{14}-g_{13})-(g_{16}-g_{15})]\}. \tag{8}$$

4.2　实验结果分析与评价

原始模型与改进模型的对比结果见图 5. 由图 5 可知，在对改进后的模型进行实验评估时，当 Caisis 系统请求的训练数据总数从 100～600 条变化时，特征选择请求的数据条数均取 100 条. 当 Caisis 系统请求的训练数据总数为 100 条时，改进模型与原始模型的总数据传输时间是相同的，这是因为当 Caisis 系统请求的训练数据总数与改进模型的特征选择请求的数据条数都是 100 条. 无论是原始模型还是改进模型，openEHR 系统发送回来的数据都是 100 条的全部属性. 所以，当 Caisis 请求的训练数据总数与进行特征选择的条数相同时，改进模型降级为与原始模型具有相同传输效率. 但是随着 Caisis 系统请求的训练数据总数的不断增加，改进模型则具有非常明显的优势（见图 5）.

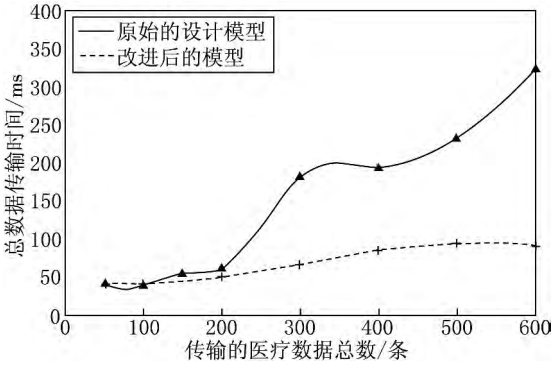

图 5　总数据传输时间随传输的医疗数据总条数的增加时的变化曲线

图 6 是对改进模型的评估. 当 Caisis 系统请求的医疗数据总数一定时，随着 Caisis 系统进行特征选择所请求的数据条数变化，总数据传输时间不断增加，传输效率逐渐降低. Caisis 系统第一步进行特征选择所请求的数据条数越小，就会有更少的数据以全部属性传输，也就是说，更多的数据会以更少的属性传输. 这样，总的数据量变小，医疗数据的传输效率就增大. 相反，随着 Caisis 系统第一步特征选择所请求的数据条数的增加，总的传输数据量变大，医疗数据的总传输时间也会变长. 当 Caisis 系统第一步特征选择所请求的数据条数等于需要训练的总数时，全部数据的全部属性都需要传输给 Caisis，这就使得改进模型降级为与原始模型具有相同的低效率模型系统.

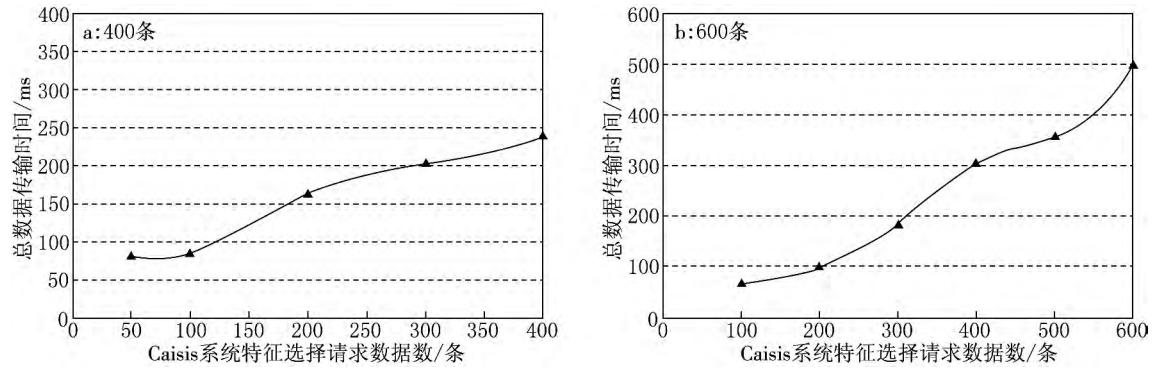

图 6　总数据传输时间随着用于特征选择医疗记录条数的增加时的变化曲线

实验结果表明，改进后的医疗数据交换协议具有更高的数据交换效率，并且医疗数据分析与预测系统进行特征选择时使用的数据量越小，医疗数据的交换效率越高；基于 SVM 的特征选择算法提高了医疗数据分析系统的效率，并提高了使用 SVM 算法进行预测分析的效率和准确度.

[参 考 文 献]

[1] 吴信东,叶明全,胡东辉,等. 普适医疗信息管理与服务关键技术与挑战[J]. 计算机学报,2012,13(5):41-42

[2] CIECHOLEWSKI M. Ischemic heart disease detection using selected machine learning methods[J]. International Journal of Computer Mathematics,2013,90(8):1734-1759.

[3] RAMIREZ J,GORRIZ J M,SALAS-GONZALEZ D,et al. Computer-aided diagnosis of alzheimer's type dementia combining support vector machines and discriminant set of features[J]. Information Sciences,2013,237:59-72.

[4] KUPUSINAC A,DOROSLOVACKI R,MALBASKI D,et al. A primary estimation of the cardiometabolic risk by using artificial neural networks [J]. Computers in Biology and Medicine,2013,43(6):751-757.

[5] 俞汝龙. HL7 组织与 HL7 标准简介[J]. 中国数字医学,2007,2(7):41-43.

[6] SANTOS C,PEDROS T,COSTA C,el al. On the use of openehr in a portable phr[C] . Rome: Scitepress,2011: 351-356.

[7] MAGLOGIANNIS I. Towards the adoption of open source and open access electronic health record systems[J]. Journal of Healthcare Engineering,2012,3(1):141-161

[8] 张旭峰,姚志洪. 基于 openEHR 的个人健康档案建模[J]. 计算机应用与软件,2013,30(5):71-72,111.

[9] 岳昆,王晓玲,周傲英. Web 服务核心支撑技术[J]. 软件学报,2004,15(3):428-442.

[10] 何清林,杨森,徐泽同. 基于元数据和 Web Service 中间件的分布式资源库集成[J]. 计算机工程与设计,2009,30(9):2202-2204.

[11] DONG WL,HU JH. Test method for BEPL-Based Web service composition based on data flow analysis[J]. Journal of Software,2009,20(8):2102-2112

[12] KHUSHI M,CARPENTER JE,BALLEINE RL,et al. Electronic biorepository application system: web-based software to manage receipt,peer review,and approval of researcher applications to a biobank[J]. Biopreserv Biobank,2012,10(1):37-44.

[13] 魏振. 基于 GPU 的 SVM 算法在入侵检测系统中的应用[D]. 长春:吉林大学,2013.

[14] VAPNIK V. The nature of statistical learning theory[M]. New York: Springer-Verlag,1995:65-85.

[15] VLADIMIR N VAPNIK. 统计学习理论的本质[M]. 张学工,译. 北京:清华大学出版社,2000:12-13.

[16] VAPNIK V,CHERVOKNENKIS A Y. The necessary and sufficient conditions for the uniform convergence of averages to their expected value[J]. Teoriya Veroyatnostei I ee Primeneniya,1981,26(3): 543-564.

[17] 邓乃扬,田英杰. 数据挖掘中的新方法——支持向量机[M]. 北京: 科学出版社,2004:164-190.

[18] 张浩然,韩正之. 支持向量机算法及应用研究[D]. 上海:上海交通大学,2003.

[19] 席少霖. 非线性最优化方法[M]. 北京: 高等教育出版社,1992:470.

[20] BARZILAY O,BRAILOVSKY V L. On domain knowledge and feature selection using a support vector machine[J]. Pattern Recognition Letters,1999,20: 475-484.

A medical data analysis model based on SVM

TIAN Yu-chi,HU Liang

(College of Computer Science and Technology,Jilin University,Changchun 130012,China)

Abstract: This paper proposes a medical data analysis and disease prediction model based on SVM classification algorithm and Web Service framework and the medical data communication protocol has been improved. This model includes the extraction of medical data from the electronic medical records database in a hospital and the transmission of medical data to a system to be analyzed by methods of data mining. A total of 1 695 patients' electronic medical records derived from a third-grade class-A hospital in Changchun are used to examine the communication efficiency of the model. Improved medical data communication protocol is also tested by means of the variation of number of requested medical records and the control of the attributes. Experimental results show that the communication efficiency of the protocol in the system based on the model is much higher when the medical data go through a dimensionally reducing process like feature selection before transmission.

Keywords: Web Service; SVM; electronic medical record; feature selection

(责任编辑:石绍庆)