支持向量机基础上的银屑病辅助诊断方法研究*

李 鹏

（1 湖南中医药大学信息科学与工程学院 长沙 410208 2 中南大学湘雅三医院 长沙 410006
3 医学信息研究湖南省普通高等学校重点实验室（中南大学） 长沙 410006）

闵 慧 罗爱静

（湖南信息职业技术学院 （1 中南大学湘雅三医院 长沙 410006
软件学院 长沙 410200） 2 医学信息研究湖南省普通高等学校重点实验室（中南大学） 长沙 410006）

〔摘要〕 介绍基于支持向量机的银屑病辅助诊断方法实现流程，包括数据收集和预处理、构建数据库群、特征提取、建立基于支持向量机的辅助诊断模型。通过实验验证该方法的有效性，其诊断精度较高，可以为银屑病数据分析、疾病预防提供技术支持。

〔关键词〕 银屑病；数据清洗；特征提取；支持向量机；序列最小优化算法；精度
〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10. 3969/j. issn. 1673－6036. 2020. 09. 007

Study on the Assistant Diagnosis Method of Psoriasis Based on Support Vector Machine *LI Peng, 1School of Informatics, Hunan University of Chinese Medicine, Changsha 410208, 2The Third Xiangya Hospital of Central South University, Changsha 410006, 3Key Laboratory of Medical Information Research (CSU), College of Hunan Province, Changsha 410006, China; MIN Hui, Software Institute, Hunan College of Information, Changsha 410200, China; LUO Aijing, 1The Third Xiangya Hospital of Central South University, Changsha 410006, 2Key Laboratory of Medical Information Research (CSU), College of Hunan Province, Changsha 410006, China*

〔**Abstract**〕 The paper introduces the realization process of the assistant diagnosis method of Psoriasis (PS) based on Support Vector Machine (SVM), including data collection and preprocessing, data base construction, feature extraction, and the building of assistant diagnosis model based on SVM. The effectiveness of the method is verified by experiments. The diagnosis accuracy of the method is rela-

〔修回日期〕 2020－05－25
〔作者简介〕 李鹏，博士，讲师，发表论文25 篇；通讯作者：闵慧，硕士，讲师。
〔基金项目〕 国家社会科学基金重点项目“网络健康信息资源聚合与精准信息服务研究”（项目编号：17AZD037）；国家重点研发计划“中医智能舌诊系统及数据平台研发与应用”（项目编号：2017YFC1703306）；湖南省自然科学基金青年项目“无线传感网中基于压缩感知的数据收集关键技术研究”（项目编号：2019JJ50453）；湖南省自然科学基金面上项目“基于‘法－方－药’网络机器学习的中医治疗银屑病复方功效预测研究”（项目编号：2018JJ2301）；湖南省科技厅重点项目“基于大数据的中西医结合防治脑梗死创新技术研究与推广应用”（项目编号：2017SK2111）；湖南中医药大学开放基金项目" 面向动态蛋白质网络的功能模块挖掘方法研究"（项目编号：2018JK02）。

tively high, which can provide technical support for data analysis and disease prevention of psoriasis.

〔**Keywords**〕 Psoriasis (PS); data cleaning; feature extraction; Support Vector Machine (SVM); Sequential Minimal Optimization (SMO) algorithm; accuracy

1 引言

银屑病（Psoriasis，PS）[1]又称“牛皮癣”，是一种以表皮细胞过度增殖及免疫性炎症为特征的慢性反复发作性疾病，被世界卫生组织列入世界 10 大顽症之一[2]。在 2018 年举办的“世界银屑病日科普活动”上，专家表示人工智能技术（Artificial Intelligence，AI）结合大数据可以为患者提供更为直观的银屑病辅助诊断。大量实证研究表明[3]当前银屑病诊治方法已经无法满足患者多样化、个性化需求，将机器学习技术与银屑病诊治结合起来，提出新的诊治手段是未来的必然趋势。本文提出一种基于支持向量机（Support Vector Machine，SVM）[4]的银屑病辅助诊断方法，通过构建诊断模型对银屑病发病或复发情况进行判断，为医生治疗工作提供决策支持。

2 基于支持向量机的银屑病辅助诊断

2.1 诊断流程（图 1）

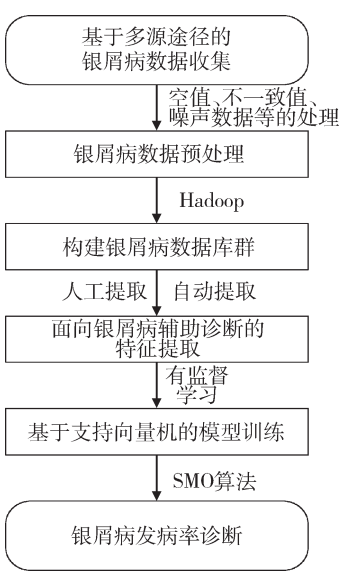

图 1 基于支持向量机的银屑病辅助诊断流程

银屑病是一种遗传与环境共同作用诱发、免疫介导的慢性、复发性、炎症性、系统性疾病，针对银屑病诊断困难这一现状，本研究通过对数据进行预处理、构建银屑病数据库群、面向银屑病辅助诊断的特征提取和基于支持向量机的模型训练等步骤对个体是否患有银屑病进行诊断，帮助医生判断病情，为患者做出诊断决定，还可辅助医生根据个体病情的不同，给出个性化治疗方案建议。

2.2 数据收集和预处理

首先利用数据接入及导入工具对分散在基层医疗机构、社区卫生中心、保健机构、体检机构和三甲医院等不同银屑病数据源的患者信息进行采集和集成，最终形成银屑病患者病历信息库。由于数据来源异构性，收集得到的银屑病相关数据在大多数情况下是不完整的，含有噪声以及不一致，采用这些数据进行机器学习往往得到的结果不准确。为此通过清洗和规约对数据进行预处理[5]。

2.3 构建数据库群

在数据收集基础上，采用 Hadoop[6]作为基本的分布式执行架构，使用 Ascential 公司的 Datastage 作为提取 - 转换 - 加载（Extract - Transform - Load，ETL）[7]工具，通过数据抽取、转换和加工、装载等步骤构建出数据库群，见图 2。

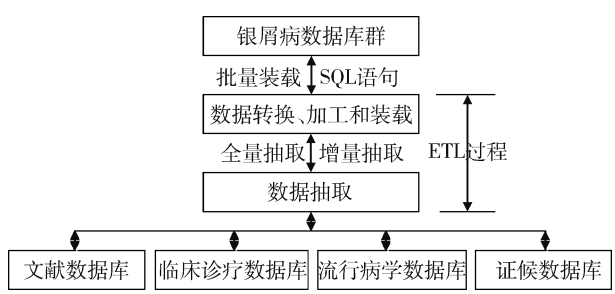

图 2 银屑病数据库群逻辑结构

利用可扩展标记语言（Extensive Markup Language，XML）技术构建元数据库和词语表。其中元数据库以都柏林核心元数据集为基础，位于这些异构数据库的上层，记录各个异构数据库对象位置，

负责整合并传递可以理解的描述信息。词语表被用来规范用户提交的词语，解决命名差异问题。引入数据更新中间件和缓存表以实现原有存储银屑病数据子系统的互联互通、信息同步与共享。系统互联同步框架，见图3。

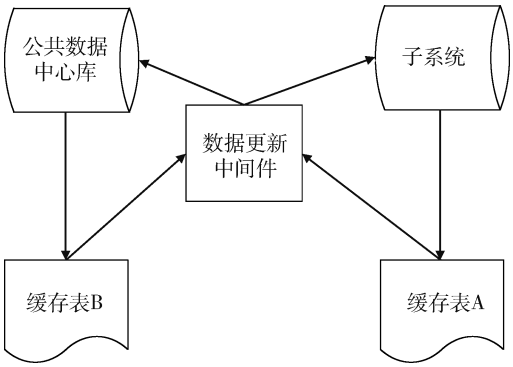

图3 系统互联同步框架

2.4 特征提取

银屑病发病诱因主要包括病毒、细菌感染、压力、外伤手术、妊娠、肥胖、吸烟、酗酒、某些药物作用和免疫因素等。从预处理后的银屑病数据集中提取出上述10种因素作为特征进行模型训练。考虑到基于SVM模型的输入一定是数值类型，而提取的10个特征中大部分是字符串类型，因此需要将字符串类型转换成数值类型，如向量、矩阵或张量形式。使用独热编码（One-Hot Encoding）[8]技术将其转换为数值类型后再作为模型的输入。编码过程，见图4。

类别特征		类别特征
病毒		1000000000
细菌感染		0100000000
压力		0010000000
外伤手术		0001000000
妊娠	独热编码	0000100000
肥胖		0000010000
吸烟		0000001000
酗酒		0000000100
药物作用		0000000010
免疫因素		0000000001

图4 独热编码

2.5 构建基于支持向量机的辅助诊断模型

支持向量机属于有监督学习模型，主要用于解决数据分类问题。SVM将每个样本数据表示为空间中的点，使不同类别的样本点尽可能明显地区分开。对于一个任意给定的银屑病训练样本集 $D=(x_1,y_1),(x_2,y_2),\ldots,(x_m,y_m)$，其中 $y_i\in\{-1,+1\}$ 分别表示健康人群和银屑病患者。本研究基本思路是：基于训练集 D 在样本空间中找到一个划分超平面，将不同类别样本分开，见图5。这样的超平面有很多，从直观判断红色线代表的超平面抗“扰动”性最好。找到红色线代表的超平面后，在进行分类时，遇到一个新的数据点 x，将 x 代入 y 中，如果 $y<0$ 则将 x 类别赋为 -1，如果 $y>0$ 则将 x 类别赋为1。本文研究的重点是如何确定超平面，以下将详细阐述。

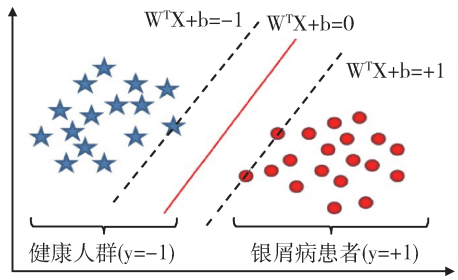

图5 基于SVM的样本分类

用函数 $y=w^Tx+b$ 表示该超平面。当 $y=0$ 时，x 是位于超平面上的点，而 $y>0$ 点对应 $y=1$ 的数据点，$y<0$ 的点对应 $y=-1$ 的点。令：

$$\begin{cases} w^Tx_i+b\geqslant+1, y_i=+1 \\ w^Tx_i+b\leqslant-1, y_i=-1 \end{cases} \quad (1)$$

根据式（1）可知任意样本点 x 到超平面的距离为：$r=\frac{|w^Tx+b|}{||w||}$，又因为 $y_i\in\{-1,+1\}$，两个异类支持向量到超平面的距离之和（也称为间隔）可表示为：$r=\frac{2}{||w||}$。要找到具有最大间隔的划分超平面，即要找到满足式（1）中约束的参数 w 和 b，使 r 最大，即：

$$\max\frac{2}{||w||}$$
$$s.t.\ y_i(w^Tx_i+b)\geqslant1, i=1,2,\ldots,m \quad (2)$$

最大化 $2/\|w\|$ 相当于最小化 $\|w\|$，为了计算方便，转化成以下公式，即为支持向量机的基本型：

$$\min_{w,b} \frac{1}{2}\|w\|^2$$
$$s.t.\ y_i(w^T x_i + b) \geqslant 1, i = 1,2,\ldots,m \quad (3)$$

上式所示的 SVM 基本型是一个凸二次规划（Convex Quadratic Programming）[9] 问题，其拉格朗日函数可表示为：

$$L(w,b,\alpha) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{m}\alpha_i(1 - y_i(w^T x_i + b)) \quad (4)$$

其对偶形式为：

$$\max(\min L(w,b,\alpha)) \quad (5)$$

为计算拉格朗日函数极小值，分别对 w,b 求偏导，使其为 0 可得：

$$\begin{cases} \frac{\partial L}{\partial w} = w - \sum_{i=1}^{m}\alpha_i y_i x_i \\ \frac{\partial L}{\partial b} = \sum_{i=1}^{m}\alpha_i y_i \end{cases} \quad (6)$$

令其分别为 0，可得：

$$\begin{cases} w = \sum_{i=1}^{m}\alpha_i y_i x_i \\ \sum_{i=1}^{m}\alpha_i y_i = 0 \end{cases} \quad (7)$$

将式（7）代入式（4）中的拉格朗日函数可得：

$$\begin{aligned} \min L(w,b,\alpha) &= \frac{1}{2}\|w\|^2 + \sum_{i=1}^{m}\alpha_i(1 - y_i(w^T x_i + b)) \\ &= \frac{1}{2}\|w\|^2 - \sum_{i=1}^{m}\alpha_i(y_i(w^T x_i + b) - 1) \\ &= \frac{1}{2}\|w\|^2 - \sum_{i=1}^{m}\alpha_i y_i w^T x_i - \sum_{i=1}^{m}\alpha_i y_i b + \sum_{i=1}^{m}\alpha_i \\ &= \frac{1}{2}w^T \cdot w - \sum_{i=1}^{m}\alpha_i y_i w^T x_i - \sum_{i=1}^{m}\alpha_i y_i b + \sum_{i=1}^{m}\alpha_i \\ &= \frac{1}{2}w^T \cdot \sum_{i=1}^{m}\alpha_i y_i x_i - \sum_{i=1}^{m}\alpha_i y_i w^T x_i - \sum_{i=1}^{m}\alpha_i y_i b + \sum_{i=1}^{m}\alpha_i \\ &= -\frac{1}{2}w^T \cdot \sum_{i=1}^{m}\alpha_i y_i x_i - \sum_{i=1}^{m}\alpha_i y_i b + \sum_{i=1}^{m}\alpha_i \\ &= \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i x_j \end{aligned}$$
$$s.t.\ \sum_{i=1}^{m}\alpha_i y_i = 0, \alpha_i \geqslant 0, i = 1,2,\ldots,m \quad (8)$$

根据式（5）可得最终的优化表达式为：

$$\max\left\{\sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^T x_j\right\}$$
$$s.t.\ \sum_{i=1}^{m}\alpha_i y_i = 0, \alpha_i \geqslant 0, i = 1,2,\ldots,m \quad (9)$$

由于基本型中存在不等式约束，上式满足卡罗需－库恩－塔克条件（Karush－Kuhn－Tucker Conditions，KKT）[10]，即有：

$$\begin{cases} \alpha_i \geqslant 0 \\ y_i(w^T x_i + b) - 1 \geqslant 0 \\ \alpha_i(y_i(w^T x_i + b) - 1) = 0 \end{cases} \Rightarrow \begin{cases} \alpha_i \geqslant 0 \\ y_i f(x_i) - 1 \geqslant 0 \\ \alpha_i(y_i f(x_i) - 1) = 0 \end{cases} \quad (10)$$

式（9）是一个二次规划问题，虽可采用二次规划算法来求解，但算法开销太大。为此采用经典的序列最小优化（Sequential Minimal Optimization，SMO）算法[11] 解出 α 之后，根据公式（7）求得 w，进而求得 b，得到 SVM 模型为：

$$f(x) = w^T x + b = \sum_{i=1}^{m}\alpha_i y_i x_i x + b \quad (11)$$

基于 SVM 的银屑病辅助诊断模型建立后，对于任何一名疑似患者而言，只需将其 10 大特征数据（$x_1, x_2, \ldots, x_{10}$）输入公式（11）中就可以准确诊断是否患有银屑病，这一结果为临床医生提供决策依据。

3 实验

以 2017 年 1 月 1 日－2019 年 12 月 31 日为一个时间周期，从湖南中医药大学第一和第二附属医院的皮肤科获取有关银屑病门诊数据，将其作为构建诊断模型的数据集，共筛选得到数据样本 20 000 余份。随机抽取 70% 样本数据集作为训练集来训练模型，余下 30% 数据作为测试集用以评估模型性能。在一台 8 核 16 线程的计算机上进行实验。其中 CPU 型号为 Intel Core i9－9960X @ 3.10GHz，内存为 16G，操作系统为 Ubuntu 16.04 LTS 64 位。使用包含 Python 内核的 Anaconda 平台和 Scikit－learn 库实现基于支持向量机的银屑病辅助诊断方法，采用目前常用的精度来评价方法性能，见图 6。可以看出随着数据规模的增加，本文方法的诊断精度呈线性上升，但当样本规模超过 10 000 后，诊断精度逐渐趋于稳定。分析其原因可知：本文方法在建模过程中采用多种技术对银屑病原始大数据进行清洗，对影响银屑病高危因素进行分析和特征提取，将噪声数据对模型

的影响降到最低；对每个特征进行编码，提高特征对于模型的吻合度；采用 SMO 算法进行模型训练，在减少训练时间的同时进一步保证诊断准确性。

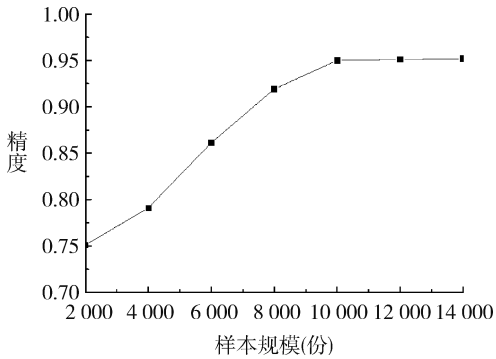

图6 本文方法精度比较

4 结语

银屑病是一种慢性炎症性皮肤病，病程较长，有易复发倾向，有的病例几乎终生不愈。如果能在潜伏期对该病进行准确诊断对于后续治疗具有重要意义，为此本文提出一种基于支持向量机的银屑病辅助诊断方法，通过实验验证其有效性。下一步将采用深度学习技术自动提取影响银屑病发病的重要因素，设计一种基于图卷积神经网络的银屑病发病率诊断方法，为医生智能诊疗提供更好的技术支持。

参考文献

1 Blome C, Costanzo A, Dauden E, et al. Patient – relevant Needs and Treatment Goals in Nail Psoriasis [J]. Quality of Life Research, 2016, 25 (5): 1179 – 1188.

2 Reich K, Hartl C, Gambichler T, et al. Retrospective Data Collection of Psoriasis Treatment with Fumaric Acid Esters in Children and Adolescents in Germany (KIDS FUTURE Study) [J]. JDDG: Journal der Deutschen Dermatologischen Gesellschaft, 2016, 14 (1): 50 – 57.

3 胡满满, 陈旭, 孙毓忠, 等. 基于动态采样和迁移学习的疾病预测模型 [J]. 计算机学报, 2019, 42 (10): 2339 – 2354.

4 Gu B, Quan X, Gu Y, et al. Chunk Incremental Learning for Cost – sensitive Hinge Loss Support Vector Machine [J]. Pattern Recognition, 2018 (83): 196 – 208.

5 姜慧勇, 娄冬华. 临床试验既往病史、不良事件和合并用药数据清理的方法 [J]. 南京医科大学学报 (自然科学版), 2018, 38 (10): 1463 – 1466.

6 Tariq H, Al – Sahaf H, Welch I. Modelling and Prediction of Resource Utilization of Hadoop Clusters: a machine learning approach [C]. Paris: Proceedings of the 12th IEEE/ACM International Conference on Utility and Cloud Computing, 2019: 93 – 100.

7 Nath R P D, Hose K, Pedersen T B, et al. SETL: a programmable semantic extract – transform – load framework for semantic data warehouses [J]. Information Systems, 2017 (68): 17 – 43.

8 Rodríguez P, Bautista M A, Gonzalez J, et al. Beyond One – hot Encoding: lower dimensional target embedding [J]. Image and Vision Computing, 2018 (75): 21 – 31.

9 Tam H H M, Tuan H D, Ngo D T. Successive Convex Quadratic Programming for Quality – of – service Management in Full – duplex MU – MIMO Multicell Networks [J]. IEEE Transactions on Communications, 2016, 64 (6): 2340 – 2353.

10 Van Tuyen N, Yao J C, Wen C F. A Note on Approximate Karush – Kuhn – Tucker Conditions in Locally Lipschitz Multiobjective Optimization [J]. Optimization Letters, 2019, 13 (1): 163 – 174.

11 Chen L, Chu C, Feng K. Predicting the Types of Metabolic Pathway of Compounds Using Molecular Fragments and Sequential Minimal Optimization [J]. Combinatorial Chemistry & High Throughput Screening, 2016, 19 (2): 136 – 143.