肿瘤电子病历数据挖掘技术的应用研究

童 刚,姜 宁,刘 焕
(青岛科技大学 信息科学技术学院,山东 青岛 266061)

摘 要:旨在研究肿瘤电子病历数据挖掘技术,重点探究数据抽取及挖掘分析实验。数据抽取是对文本信息进行针对性抽取,以结构化的形式将结果储存起来,从而为分类算法的研究奠定数据基础。重点研究了肿瘤电子病历的中文分词及分类挖掘算法的选取,对于中文分词的研究,提出了改进后的逆向最大匹配算法,提高了分词准确度和分词效率。对于分类挖掘算法的研究,采用分类效果较好的C4.5算法和BP神经网络算法分别进行分类挖掘实验,通过对分类算法的性能对比,在研究肿瘤电子病历的分类挖掘上,C4.5算法更有利于辅助医生进行肿瘤疾病诊断,提高疾病诊断的精确率及效率进而提高肿瘤患者的治愈率。

关键词:肿瘤电子病历;辅助诊断;逆向最大匹配分词;C4.5;神经网络
中图分类号:TP399 **文献标识码**:A **文章编号**:1673-629X(2020)08-0152-05
doi:10.3969/j.issn.1673-629X.2020.08.026

Application of Data Mining Technology in Electronic Medical Record of Cancer

TONG Gang, JIANG Ning, LIU Huan
(School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, China)

Abstract: The aim is to study the data mining technology of electronic medical records of tumor, especially the data extraction and mining analysis experiments. The data extraction carries out the targeted extraction of the text information and stores the results in a structured form, so as to lay a data foundation for the research of classification algorithms. The Chinese word segmentation of tumor electronic medical records and the selection of classification mining algorithms are studied. For the Chinese word segmentation, an improved inverse maximum matching algorithm is proposed to improve the segmentation accuracy and word segmentation efficiency. For the classification mining algorithm, the classification mining experiment is carried out by C4.5 algorithm and BP neural network algorithm with better classification effect. Through the comparison of the performance of the classification algorithm, in the classification mining of tumor electronic medical records, the C4.5 algorithm is more conducive to assisting doctors in the diagnosis of tumor diseases, improving the accuracy and efficiency of disease diagnosis and improving the cure rate of cancer patients.

Key words: electronic medical record of cancer; assisted diagnosis; reverse maximum matching segmentation; C4.5; neural network

0 引 言

数据挖掘是指从大量的数据中通过算法搜索其中重要信息的过程。在医学中,医疗诊断的方法及选择模式尤其重要,将数据挖掘技术应用在此便于医生对疾病进行诊断,从而在医疗科研方面提供了科学依据[1]。随着医疗信息系统的发展,医院的数据库信息在医疗分类诊断上变得更加重要,如何有效利用这些信息进行分类挖掘是很多研究者的工作重心。冠心病是目前威胁人类身体健康的一项重大疾病,利用当今流行的数据挖掘技术提炼出冠心病积累的临床信息资料中的有用信息,并通过神经网络算法进行分类诊断,诊断的精确率已经高达90%[2]。除此之外,在其他相关疾病诊断中,此类技术的应用也达到了预期效果。Chen等[3]在提取规则方面,运用了决策树算法,然后采用CBR技术修改过往问题的解决流程,并应用到肿瘤疾病的新情况中进行诊断。Murate等[4]将神经网络算法及支持向量机算法应用在早期前列腺疾病的诊断中。Anand等[5]在疾病的诊断分类中,将病人的医

收稿日期:2019-09-12 **修回日期**:2020-01-15
基金项目:国家自然科学基金(61572268)
作者简介:童 刚(1962-),男,教授,研究方向为工业自动化、信息管理、智能交通;姜 宁(1992-),男,硕士,CCF会员(85456G),研究方向为计算机通信网技术。

学数据通过混合人工神经网络进行分析,在分类精度上有所提高。Huang Z 等[6]提出了增强迭代次数的分类算法,对处理急性冠脉综合征患者心脏不良事件失衡问题有显著效果。肖勤[7]在建立乳腺 X 线分类模型上选用决策树算法,在分类诊断上取得了很好的效果。Feng 等[8]在慢性胃炎中的分类诊断中应用了信息熵决策树算法。刘绿[9]将一些分类算法进行了性能对比,结果显示决策树的综合性能最佳。许腾[10]在甲状腺疾病的分析研究中,将纹理及超声图像进行了融合运用。于霄[11]创建了基于分类算法的医疗服务系统,并弥补了决策树本身存在的过拟合问题。

1　肿瘤电子病历的分类挖掘实验

电子病历中包含的医疗信息十分丰富。对其数据的有效处理和利用,是一项非常有意义的工作。通过数据预处理等[12]可部分消除数据中的噪声和不完整性,实现数据的规范化和有效压缩,从而使数据的再处理更加有效。最终将非结构化的电子病历文本数据转换成可直接挖掘利用的结构化数据。在电子病历中,病程记录是其重要组成部分,病程记录中包含了大量可供挖掘的患者就诊信息及过往病史信息,因此病程记录可以作为数据抽取的关键。首次病程记录的内容结构如图 1 所示。

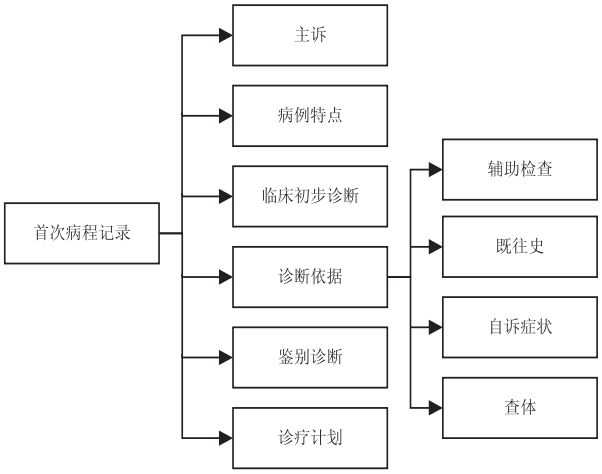

图 1　首次病程记录的内容结构

数据抽取又叫信息抽取,是数据预处理技术中的关键。基于目前的实体抽取模型的优劣性并结合研究数据的特点,文中采用了基于条件随机场的多特征融合的医疗实体识别方法[13-14]。具体识别流程如图 2 所示。

如图 2 所示,首先将原始语料库进行相应的中文分词和标注处理后,变为训练语料,再将训练语料分词进行同样的处理形成训练模型。其次将测试语料输入到训练模型中进行实体识别。最后将识别后形成的结果按照一定的方法规则进行评测,得到评测结果来检验整体模型的科学性。

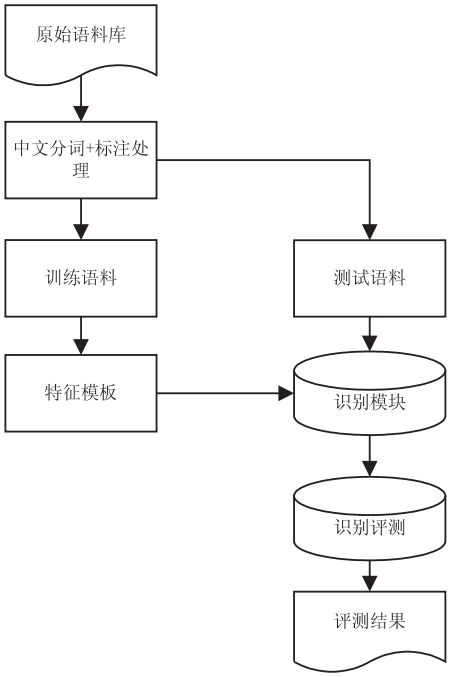

图 2　基于条件随机场的实体识别方法框架

2　挖掘实验重点探究

中文分词被视为最基础的问题,中文分词常用的方法有三种:基于词典的分词、基于统计的分词、基于理解的分词。根据电子病历中医疗术语较多的特点,采用基于词典的分词算法[15],即将分字符串中的词,按照一定的标准和规则与词典中的词进行比对,若可以在词典中找到该字符串,则匹配成功。若找不到,则按照一定的算法策略继续匹配。基于词典的分词算法中逆向最大匹配法的分词精确率较高,缺点是分词速度较慢[16]。为解决这个问题,结合电子病历数据的表达特点提出了改进后的逆向最大匹配算法,在分词速度上有明显提高。

逆向最大匹配算法的思想如下:事先设置一个 n 值,然后从最后一个字开始向前截取 n 个字,先把这 n 个字与词典进行匹配,看能否找到匹配的词语,若匹配成功,即识别出一个词。若不能,则删除这 n 个字最左边的字,然后再把这 $n-1$ 个字与词典继续匹配直到匹配成功,或者前 $n-1$ 个字都没匹配成功,那就把第 n 个字当成一个独立的词,然后再向前移动分出来的词的长度,再截取 n 个字直到全部分好词为止。

改进后的算法思想:

(1)将分字符串中的词 A 与词典中的词 B 进行对比,如果词典 B 中没有 A,则选择逆向最大匹配法进行分词。

(2)如果词典 B 中有 A,将 A 前后位置的词分别与 A 进行组词,将新组成的分词与词典 B 进行比对:

若有一个存在于词典 B 中,将 A 和新匹配的词一起作为一个分词,并在此处将字符串分为两段,最后再利用逆向最大匹配算法将这两段进行分词;若两个词都存在于词典 B 中,采用最大概率分词法进行确定;若在词典 B 中两个词均无法找到,那么以 A 为切点将字符串分成前后两段,再采用逆向最大匹配算法进行分词。

改进后的算法流程如图 3 所示。

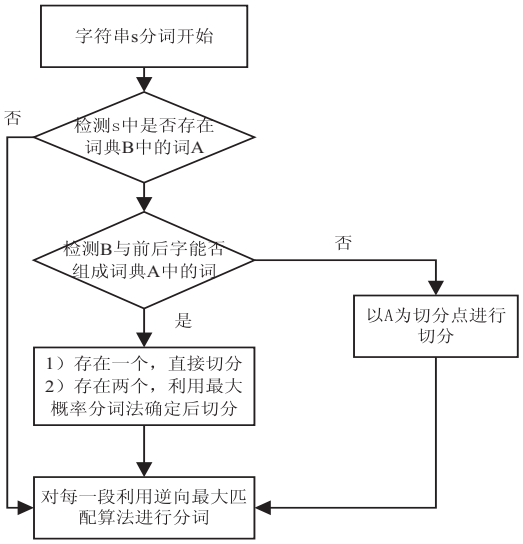

图 3 改进后的逆向最大匹配算法流程

电子病历的语言表达中会有很多单用词如“和”“到”“若”“及”等,对这些单用词进行切分,不仅提高了分词效率和准确率,还不影响最终结果。因此将类似的这类单用词组建成一个新的词典,同时找出一些症状专有名词和疾病判断词也放入新词典中。最后判断待分字符串中是否有新词典中的词,若有则在此处分词,对切分后的每个词,再继续分词。利用传统的逆向最大匹配算法及改进后逆向最大匹配算法分别对电子病历部分内容进行分词的对比结果如表 1 所示。

表 1 逆向最大匹配算法改进前后分词结果比对

传统逆向最大匹配法分词结果	改进后的逆向最大匹配算法
右/f 肺/n	右肺/n
上/f 叶/ng	上叶/n
左/a 肾上腺/n	左肾上腺/n
颈/ng 部/q	颈部/q
增/v 厚	增厚/v

利用数据抽取中常用的 P 值、R 值、F 值三个评价指标加上分词速度对实验结果进行对比评测[17],评测结果如表 2 所示。

表 2 逆向最大匹配算法改进前后性能比对

算法	P	R	F	速度(KB/s)
逆向最大匹配	0.871	0.814	0.842	159
改进后的逆向最大匹配	0.912	0.897	0.904	323

按照改进之后的逆向最大匹配分词法对电子病历分词后,经过标注处理及相应的特征选择后,得到初步的数据抽取结果,再对其进行数据清理、数据变换、数据归约等操作完成整个的数据预处理,图 4 为预处理之后的部分截图。

age	sex	desease ag	smoke	weight	shit	nausea	temperatur	rhythm of	blood	kexie	furhang	chest tight	painlocati	aiqi	sleep
4	1	4	1	2	0	0	0	0	0	0	0	0	2	0	0
5	2	3	2	0	0	0	0	0	1	0	0	0	1	0	0
1	1	1	1	0	0	0	1	1	0	0	0	0	2	0	1
6	2	2	2	1	0	1	0	0	0	0	0	0	2	1	0
4	1	1	1	0	0	0	0	0	0	0	0	0	2	0	0
6	1	3	1	1	0	0	0	0	0	0	0	1	1	0	1
5	1	1	2	1	1	0	0	0	1	0	0	0	1	0	0
4	2	2	2	0	0	0	1	0	0	0	0	0	2	0	1
5	1	1	1	1	0	1	0	0	0	0	0	0	1	0	0
6	1	1	1	0	0	0	1	1	1	0	1	0	1	0	1
6	2	6	2	1	0	1	0	0	0	0	0	0	2	0	0
6	1	1	1	2	0	0	0	0	1	0	0	0	2	1	0
7	2	2	2	0	0	0	0	0	0	0	0	1	2	0	1
6	1	1	1	0	0	0	1	0	0	0	0	0	2	0	0
6	2	4	2	1	0	1	0	0	0	1	0	0	2	0	0
6	1	1	2	0	0	0	0	1	0	0	0	0	1	0	0
5	2	2	2	1	0	0	0	0	1	0	0	0	1	0	0
6	2	3	2	1	0	0	0	1	0	0	1	0	2	0	0
5	1	1	1	1	0	0	0	0	0	0	0	1	1	0	1
6	2	2	2	0	0	0	0	0	0	0	0	0	2	0	0
5	1	1	1	1	0	0	0	0	1	0	0	0	1	0	0
6	1	6	1	0	0	0	0	1	0	0	0	0	1	0	1
6	1	1	2	1	0	0	0	0	0	0	0	0	2	0	0
6	2	2	2	0	2	0	0	0	0	0	0	0	2	0	0
5	1	5	1	0	0	0	0	0	0	0	0	0	2	0	1
8	2	2	2	2	0	0	0	0	1	0	0	0	2	1	0
8	1	7	1	0	0	0	0	0	0	0	0	0	2	0	0
6	2	2	2	0	0	1	1	0	0	0	0	0	2	0	1
5	1	1	1	1	0	0	0	0	1	1	0	0	2	0	0
7	1	1	1	0	0	0	0	0	0	0	0	0	2	0	1

图 4 预处理之后的肿瘤疾病数据集部分截图

3 分类算法的选取

挖掘实验过程中的关键问题在于挖掘算法的选取,针对医疗数据自身的独特性,筛选出合适的算法进而实现辅助诊断变得更加重要。然而不同的数据挖掘算法具有不同的特性[18],通过其特性对比发现,在分类选取方面,C4.5 算法和 BP 神经网络效果最佳[19]。C4.5 算法的基点是 ID3 算法,具备 ID3 算法的优点,在属性选择上用信息增益率进行选择,由于属性选择时会优先选择取值多的属性,C4.5 算法有效解决了这

类问题。不仅可以将连续属性离散化处理,还能够处理一些不完整数据。BP 神经网络的主要特点是信号和误差按照相反方向进行传播。信号传播过程中,信号从输入层进入隐藏层,最后到达输出层,下一层的信号状态只由上一层影响。如果最后输出的信号并不是期望信号,则进入误差的反向传播过程。再根据误差进行调整权值和偏向,最后使得输出信号不断逼近期望输出。因此 BP 神经网络具有高度自学习和自适应的能力。下面对这两种算法进行分类挖掘实验。

3.1　C4.5 分类实验

预处理后得到的数据集使用 C4.5 算法进行挖掘实验,采用十折交叉验证法测试算法的准确性[20]。

运行结果如图 5 所示。

```
Time taken to build model:  129 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances                86.17 %
Incorrectly Classified Instances              13.83 %
```

图 5　C4.5 算法疾病分类效果

实验结果表明,C4.5 算法分类结果性能:分类正确率约为 86%,错误率约为 14%,建模时间为 129 s。

3.2　BP 神经网络分类实验

BP 神经网络算法具有实现任何复杂非线性映射的功能且可以进行复杂的数学运算[21]。它还具有一定的推广、概括、自学习等能力。在实际应用中,多数神经网络模型都采用 BP 神经网络的变化形式,在分类挖掘应用方面有较好的实验效果。运行结果如图 6 所示。

```
Time taken to build model:  398 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances                87.50 %
Incorrectly Classified Instances              12.50 %
```

图 6　BP 神经网络算法疾病分类效果

由以上效果图可以看到 BP 神经网络算法的分类精确率约为 88%,错误率约为 12%,建模时间为 398 s。

3.3　实验结果对比

C4.5 与 BP 神经网络在肿瘤病历数据上的实验对比如表 3 所示。

表 3　分类实验精度性能对比

算法名称	测试集准确率/%	建模时间/s
C4.5	86	129
BP 神经网络	88	398

通过以上分析可以得出结论,BP 神经网络算法在分类的精确率上略高于 C4.5 算法,但是其运行时间效率要比 C4.5 算法慢 3 倍。综合来看,两种算法的精确率相差较小,但是 C4.5 算法的运算效率却远远超过 BP 神经网络算法,因此 C4.5 算法具有较高的综合性能,更适用于肿瘤电子病历的分类挖掘。

4　结束语

肿瘤电子病历挖掘过程中包含两个重要环节:中文分词及算法选取,针对中文分词,文中结合肿瘤电子病历的表达特点,采用了一种基于特定字词切分的方法对最大逆向匹配分词算法进行改进。实验结果表明,改进后的算法不仅提高了分词效率同时在分词精确度上也有明显提高。在算法选取阶段,对比了分类领域中性能较高的两种算法:C4.5 和 BP 神经网络算法,经对比之后发现 C4.5 算法的综合性能要高于 BP 神经网络,因此选用 C4.5 算法作为肿瘤电子病历的分类挖掘算法。通过以上研究,可以实现利用数据挖掘技术辅助医生进行疾病诊断的目的,能够提高肿瘤疾病诊断的精确率及效率,进而提高肿瘤疾病的治愈率。

参考文献：

[1] KAUR H, WASAN S K. Empirical study on applications of data mining techniques in healthcare[J]. Journal of Computer Science, 2006, 2(2): 194-200.

[2] DAS R S, TURKOGLU I, SENGUR A. Effective diagnosis of heart disease through neural networks ensembles[J]. Expert Systems with Applications, 2009, 36(4): 7675-7680.

[3] HUANG M J, CHEN M Y, LEE S C. Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis[J]. Expert Systems with Applications, 2007, 32(3): 856-867.

[4] ÇINAR M, ENGIN M, ZEKIENGIN E, et al. Early prostate cancer diagnosis by using artificial neural networks and support vector machines[J]. Expert Systems with Applications, 2009, 36(3): 6357-6361.

[5] ANAND L, IBRAHIM S P S. HANN: a hybrid model for liver syndrome classification by feature assortment optimization [J]. Journal of Medical Systems, 2018, 42(11): 211-222.

[6] HUANG Z, CHAN T M, DONG W. MACE prediction of acute coronary syndrome via boosted resampling classification using electronic medical records[J]. Journal of Biomedical Informatics, 2017, 66: 161-170.

[7] 肖 勤. 数据挖掘技术在乳腺 X 线诊断中的应用[D]. 上海：复旦大学，2009.

[8] YANG F, MAO K Z. Robust feature selection for microarray data based on multicriterion fusion[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2011, 8(4): 1080-1092.

[9] 刘 绿. Logistic 回归模型、神经网络模型和决策树模型在乳腺癌的彩超影像诊断中的比较研究[D]. 衡阳：南华大学，2013.

[10] 许 腾. 基于甲状腺疾病的临床数据挖掘与分析研究[D]. 上海：东华大学，2016.

[11] 于 霄. 基于分类算法的智慧医疗服务系统的设计与实现[D]. 成都：电子科技大学，2018.

[12] 庄 军，郭 平，周 杨，等. 电子病历数据预处理技术[J]. 计算机科学，2007，34(3)：141-144.

[13] 何 彬，关 毅. 基于字级别条件随机场的医学实体识别[J]. 智能计算机与应用，2019，9(2)：130-134.

[14] 翟菊叶，陈春燕，张 钰，等. 基于 CRF 与规则相结合的中文电子病历命名实体识别研究[J]. 包头医学院学报，2017，33(11)：124-125.

[15] WANG Xu, YANG Chen, GUAN Renchu, et al. A comparative study for biomedical named entity recognition[J]. International Journal of Machine Learning and Cybernetics, 2018, 9(3): 373-382.

[16] ABSA A H A, DERICHE M, ELSHAFEI-AHMED M, et al. A hybrid unsupervised segmentation algorithm for arabic speech using feature fusion and a genetic algorithm[J]. IEEE Access, 2018, 6: 157-162.

[17] DE CLERCQ E, VAN CASTEREN V, BOSSUYT N, et al. Belgian primary care EPR: assessment of nationwide routine data extraction[J]. Studies in Health Technology & Informatics, 2014, 197(18): 85-89.

[18] 黄 雯. 数据挖掘算法及其应用研究[D]. 南京：南京邮电大学，2013.

[19] 方金城. 分类挖掘算法综述[J]. 沈阳工程学院学报：自然科学版，2006，2(1)：73-76.

[20] LI Min. A study on the influence of non-intelligence factors on college students' english learning achievement based on C4.5 algorithm of decision tree[J]. Wireless Personal Communications, 2018, 102(2): 1213-1222.

[21] 郭红霞，师义民. 中医脉象的 BP 神经网络分类方法研究[J]. 计算机工程与应用，2005，41(32)：187-189.

(上接第 151 页)

[5] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector[C]//European conference on computer vision. Amsterdam, The Netherlands: Springer, 2016: 21-37.

[6] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Washington, DC: IEEE, 2016: 779-788.

[7] CHEN L C, ZHU Y, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//European conference on computer vision. Amsterdam, The Netherlands: Springer, 2018: 833-851.

[8] CHOLLET F. Xception: deep learning with depthwise separable convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Washington, DC: IEEE, 2017: 1800-1807.

[9] HE K, ZHANG X, REN S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.

[10] 陈鸿翔. 基于卷积神经网络的图像语义分割[D]. 杭州：浙江大学，2016.

[11] 刘婷婷，张惊雷. 基于 ORB 特征的无人机遥感图像拼接改进算法[J]. 计算机工程与应用，2018，54(2)：193-197.

[12] RUBLEE E, RABAUD V, KONOLIGE K. ORB: an efficient alternative to SIFT or SURF[C]//IEEE international conference on computer vision. Barcelona, Spain: IEEE, 2011: 2564-2571.

[13] 李小红，谢成明，贾易臻，等. 基于 ORB 特征的快速目标检测算法[J]. 电子测量与仪器学报，2013，27(5)：455-460.

[14] 单 欣，王耀明，董建萍. 基于 RANSAC 算法的基本矩阵估计的匹配方法[J]. 上海电机学院学报，2006，9(4)：66-69.

[15] 周剑军，欧阳宁，张 彤，等. 基于 RANSAC 的图像拼接方法[J]. 计算机工程与设计，2009，30(24)：5692-5694.