文章编号:0253-2778(2016)10-0867-07

基于真实世界临床数据的失眠病判别分析

朱 威[1],颜仕星[2],张 磊[3],李国正[4]

(1. 同济大学电子与信息工程学院,上海 201804;2. 上海金灯台信息科技有限公司,上海 201801
3. 中国中医科学院中医临床基础医学研究所,北京 100700;4. 中国中医科学院中医药数据中心,北京 100700)

摘要:基于真实世界中医医疗数据集,提出了针对性的中医非结构化转结构化的数据预处理方法,并在监督分类模型和半监督分类模型上对得到的症状特征进行了实验验证.在真实医疗数据集上进行实验,发现无论是监督分类算法还是半监督分类算法在所提出的数据预处理模型上都得到了较优的分类效果,并且发现标签传播算法不仅在分类器稳定性上取得了较大的优势,在带标注数据较少时,仍能取得较好的实验结果.

关键词:结构化;半监督学习;标签传播;中医;疾病判别;失眠

中图分类号:TP391 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2016.10.011

引用格式:朱威,颜仕星,张磊,等.基于真实世界临床数据的失眠病判别分析[J].中国科学技术大学学报,2016,46(10):867-873.

ZHU Wei, YAN Shixing, ZHANG Lei, et al. Insomnia discriminant analysis based on real-world clinical data[J]. Journal of University of Science and Technology of China, 2016,46(10):867-873.

Insomnia discriminant analysis based on real-world clinical data

ZHU Wei[1], YAN Shixing[2], ZHANG Lei[3], LI Guozheng[4]

(1. *College of Electronic and Information Engineering, Tongji University, Shanghai* 201804, *China*;
2. *Shanghai Menorah Information Technology Co., Ltd, Shanghai* 201801, *China*;
3. *Institute of Basic Research in Clinical Medicine, China Academy of Chinese Medical Sciences, Beijing* 100700, *China*;
4. *National Data Center of Traditional Chinese Medicine, China Academy of Chinese Medical Sciences, Beijing* 100700, *China*)

Abstract: A new data preprocessing method based on the real-world medical database was proposed, which can change unstructured data into structured data. Supervised algorithms and semi-supervised algorithms were utilized to verify the effectiveness of the clinical features which were obtained through our data preprocessing method. From the experimental results on the real world dataset, it is found that both supervised classification and semi-supervised algorithms can get a better result based on the clinical symptom features trained from our data preprocessing method. And it is found that the label propagation algorithm not only achieves a great stability on the real Chinese medicine database when compared with classical classification algorithm, but also obtains good results when the ratio is low.

Key words: structurization; semi-supervised learning; label propagation algorithm; TCM; disease identification; insomnia

收稿日期:2016-03-01;**修回日期**:2016-09-16

基金项目:国家自然科学基金(61273305,81503680),中央级公益性科研院所基本科研业务费专项资金(ZZ0908032)资助.

作者简介:朱威,男,1992年生,硕士.研究方向:数据挖掘.E-mail:zwtj2010@163.com

通讯作者:张磊,博士/助理研究员.E-mail:leizhang@ndctcm.cn

0 引言

中医药数字化和现代化已经成为我国现阶段大力推动中医药发展前进的战略性政策. 相比于传统西医领域的医疗数据分析而言,中医领域的医疗数据具有数据特征形式混杂、非结构化现象明显、特征和类标缺失严重等特点. 因此,在对中医领域医疗数据进行疾病判别分析的过程中,不仅需要在疾病判别模型方法上进行研究,更需要构建合适的中医疾病分类症状特征. 而在构建中医疾病分类症状特征的过程中,非结构化症状描述句结构化是不可避免、也至关重要的一步.

现阶段已有的一些中医疾病分类工作一般都是在已有结构化症状数据集的基础上对疾病分类模型进行研究,针对中医非结构化症状描述句结构化的工作相对较少,并且现阶段对于中医非结构化症状描述句的处理主要是对症状描述句进行分词处理,如结巴、SnowNLP 等,然后计算症状术语之间的相似度;这一类数据预处理方法都是在通用中文术语的基础上进行操作的. 而中医领域的症状术语与通用的中文术语相比,无论是在表述方法还是在用词习惯上都有很大的差异性. 因此,由于通用中文分词方法未能结合中医领域词汇的特点,构建的症状特征较差,最终难以获得好的分类效果.

由于中医领域病历数据症状、类标缺失严重,研究者们对中医疾病分类模型方法的研究工作大致可以分为两类:一类是将有特征、类标缺失的样本直接删除,只保留那些特征和类标完全的样本用来进行疾病分析的监督学习算法,如 SVM、KNN 等;另一类是在利用含有类标的数据的基础上,同时利用那些类标缺失样本信息作为辅助形成的半监督学习算法,如 LPA 等.

本文的主要工作是在标准中医症状库的基础上,提出了一种针对中医领域症状非结构化描述句结构化的数据预处理框架,并在监督分类算法和半监督分类算法上对结构化后的症状特征进行实验验证.

1 算法介绍

1.1 中医非结构化数据转结构化数据方法

现有常规的中医非结构化转结构化方法的主要思想是先利用分词算法对中医症状描述句进行中文分词,然后对分词后的症状描述句去掉停用词后进行相似度计算. 在这个过程中,利用分词算法对中医症状描述句进行中文分词有着至关重要的作用. 因此,本文接下来先对现阶段常用的中文分词算法进行大概介绍.

结巴分词是现阶段主流的开源中文分词框架,其主要由 3 部分算法构成:① 基于 Trie 树[1]实现的高效词图扫描模型,并根据句中汉字所有可能成词情况,构成有向无环图(DAG);② 对于未登录词,采用具有汉字成词能力的 HMM 模型[2]中的 Viterbi 算法[3-4]预测每个词出现的可能性;③ 采用了动态规划查找最大概率路径,从而确定基于词频的最大切分组合. 结巴分词的第 1 部分内容主要是在通用中文术语库的基础上,对于在中文术语库中存在的中文术语,利用 Trie 树结构实现快速查找,构建每个症状描述句的有向无环图,并将每个术语出现的频次记录在有向无环图的节点上;第 2 部分算法主要是针对在通用中文术语库中不存在中文术语词,利用隐马尔科夫模型中的 Viterbi 算法求解每种分词可能性的概率;第 3 部分算法主要是针对构建的有向无环图,将每个节点上保存的词频视作词切分的概率,并利用动态规划算法求解症状描述句的最大概率切分组合,从而得到最终的分词结果. 结巴分词是在 HMM 模型和 DAG 模型基础上实现的中文分词模型,对于通用的中文语句具有较好的分词效果,并且在工业界已经得到了一定程度的应用. 但是对于中医领域的症状描述句而言,由于中医词汇独特的构词形式和表达方式,如将其直接应用在中医症状描述句上进行中文分词,效果并不是很理想.

1.2 标签传播算法

标签传播算法(label propagation algorithm, LPA)[4]是 Zhu 等于 2003 年提出来的半监督学习算法,并在近十几年中得到了广泛推广和大量应用. 例如:Yang 等[5]于 2007 年利用 LPA 算法进行英汉双语信息检索; Kim 等[6]和贺松林等[7]提出利用 LPA 进行网页分类; Blair-Goldensohn 等[8]将 LPA 应用于情感分类,他利用少量人工标签作为引导,利用 LPA 算法、上下文以及用户提供的标签,对顾客购后评价进行情感分类;任晓娟[9]和郝建柏等[10]利用 LPA 算法进行文本分类研究.

在标签传播算法中,标签按节点与节点间的相似度大小在节点间进行传播,节点与节点间的相似度越大,标签在节点与节点间传播的影响也越大. 因此,在每次迭代中,每个节点根据其相邻节点的标签

来更新自己的标签. 同时,在每次迭代的过程中,需保持已标注数据的标签固定. 最终,当所有节点标签趋于稳定时,标签传播完成,此时相似节点之间的标签分布也会趋于一致,因此可将概率分布相似的节点划分为同一类别,这样就完成了整个标签传播算法的传播过程.

令数据集 $X=\{X_L,X_U\}\in R^D$,其中 X_L 表示的是类标已知数据集,X_U 表示的是类标未知数据集,且 $L\ll U$. 于是问题转化为:利用标注数据集 X_L,Y_L 和未标注数据集 X_U 的内部结构特征,为数据集 X_U 中的每个数据打标签,即求解未标注样本数据的类标 Y_U.

第 1 步:将所有样本作为节点(未标注数据和标注数据),创建一个完全连接图 $V=\{Node, W\}$. 其中,$Node$ 表示图中的节点集合,W 表示节点与节点之间的权重,计算公式如下:

$$w_{ij}=\exp(-d_{ij}^2/\sigma^2) \tag{1}$$

式中,$d_{ij}=\sum_{d=1}^{D}(x_i^d-x_j^d)^2$.

在本次实验中,w_{ij} 表示节点 i 与节点 j 之间的余弦相似度.

$$w_{ij}=\begin{cases}w_{ij}, & x_i\in K(x_j)\\ 0, & \text{else}\end{cases} \tag{2}$$

式中,$K(x_j)$ 表示除了节点 j 之外与节点 j 之间相似度最大的前 K 个节点的集合.

第 2 步:在完全连接图 V 的基础上,构建概率转移矩阵 $\boldsymbol{T}$:

$$T_{ij}=P(j\to i)=w_{ij}/\sum_{k=1}^{l+u}w_{kj} \tag{3}$$

式中,T_{ij} 表示标签从节点 j 到节点 i 的传播概率,即将节点 j 的标签传给节点 i 时的权重.

第 3 步:定义一个由 $Y_{ic}=\sigma(y_i,c)$ 组成的标注矩阵 $\boldsymbol{Y}\in R^{(l+u)\times c}$,其中 Y_{ic} 表示节点 i 的标注 y_i 为第 c 类的标注概率.

第 4 步:标签传播方式为

$$\boldsymbol{Y}=\boldsymbol{T}\cdot\boldsymbol{Y} \tag{4}$$

通过不断更新标注矩阵 $\boldsymbol{Y}$ 直到整个 $\boldsymbol{Y}$ 收敛,从而得到最终的标注结果. 但每一次迭代的过程中,需保证已标注节点的类标保持不变,即在每次迭代的过程中,都要重置 Y_L.

标签传播算法相比于经典监督分类算法最大的优势在于,标签传播算法是建立在未标注数据的内部规律和结构的基础上,以少量标注数据作为引导的一种半监督学习算法;而经典的分类算法,则是建立在带标注数据的内部规律和结构的基础上进行分类的算法. 因此,相对于监督分类算法,标签传播算法利用了未标注数据的内部结构信息.

2　数据集介绍

本次研究采用的是中国中医科学院采集的 9 000例真实病例数据集,包含医生、患者、症状和诊断疾病等信息. 由于该数据集是从医院病历数据库中直接导出的患者就诊信息数据,因此数据质量差,类标缺失严重,症状非结构化现象明显,噪音影响较严重.

在本次实验中,为了对非结构化转结构化数据预处理框架进行分析,我们主要采用该数据集的症状信息和诊断信息. 症状信息主要包括刻下症、舌脉诊、主诉、主症、兼症和四诊信息等. 在这些信息中,刻下症、主诉、主症、兼症和四诊信息缺失严重且噪音严重,故在本次实验过程中主要采用舌、脉诊信息进行分析. 经过观察发现,舌、脉诊症状信息按照属性可以分为舌诊症状和脉诊症状,其中舌诊症状又可分为舌质症状和舌苔症状.

对该 9 000 例患者就诊数据集进行统计分析发现,该数据集中同时含有舌、脉诊症状信息和诊断信息的患者就诊记录为 1 400 条,包含舌、脉诊症状信息但不含有诊断信息的就诊记录为 2 600 条.

在含有舌、脉诊症状信息和诊断信息的 1 400 条记录中,诊断结果为“失眠”的有 400 条,另外 1 000条记录的诊断结果中无“失眠”.

3　数据预处理框架

接下来主要介绍本文中提出的非结构化症状描述句结构化的预处理框架,包括中医标准症状库的构建,中医症状描述句分词模型的建立,分词后症状术语的结构化,以及结构化后的症状统计结果说明等.

根据张启明等[13]编写的《中医症状学研究》一书提取得到中医症状名称数万条(包括 399 个标准症状和这些标准症状的别名). 标准症状例如脉沉、苔黄等,其中脉沉的别名词包括脉息沉、脉来沉等,苔黄的别名词包括苔色黄、舌苔黄、舌苔罩黄等. 同时,利用爬虫工具从网络上获得公开的中医症状术语库 4 000 余条. 由上述两部分症状术语去重后构成本次数据预处理框架中的中医症状术语库,将该中医症状术语库和搜狗中文词典一起作为症状描述

句切分时的登陆词典.

根据上述构建的登陆词典对9 000例病历数据的症状描述句进行全切分,构建每个就诊记录症状描述句的有向无环图,并且统计有向无环图中每个节点词出现的次数并记录在节点上;在构建症状描述句有向无环图的过程中,对于登陆词典中未收录的词,使用隐马尔科夫模型中的viterbi算法进行分词处理,得到每一种可能的分词情况和其概率情况;至此,有向无环图构建基本完成.对构建的有向无环图,将每个术语节点上记录的频次信息视作词频,利用动态规划算法求解每个病历症状描述句的最大概率路径.根据求得的最大概率路径,即可得到最优的症状句分词结果.

然后根据前文构建的症状术语的别名信息对分词后的结果进行模糊查询和同义词匹配处理,得到结构化后的症状信息.例如症状描述句为:胖质暗有瘀斑,苔薄白.按照本文中设计的分词框架该症状描述句被划分为:胖、质暗、有、瘀斑、苔薄、白;而按照结巴分词的分词结果,该描述句被划分为:胖质,暗有,瘀斑,苔,薄白.由上面的例子可以很明显地看出,结巴分词模型没有考虑中医症状术语的表达特性,而本文提出的非结构化转结构化框架将这一特性考虑在内,因此取得了较好的结果.

接下来,我们将舌脉诊症状分为舌质症状、舌苔症状和脉诊症状.通过统计发现,舌质症状大致可分为:裂纹、齿痕、震、颤、胖、少津、红点、紫、正常、舌下静脉曲张等;舌苔症状大致可以分为:腻、厚、黄、白、少、燥、薄、花剥等;脉诊症状大致可分为:细、弦、沉、滑、数、缓、浮、弱、涩、濡、紧、平、正常等.后文将分别采用监督分类算法和半监督分类算法在本文提出的非结构化转结构化数据预处理框架和通用数据预处理框架下进行比较实验.对该数据集数据预处理的过程,如图1所示.

4 实验

本次实验采用LPA算法、SVM算法[11]和KNN算法[12]对不同的数据预处理框架下得到的症状特征进行比较实验.

4.1 实验设计

将ratio定义为训练集中带标注的样本数占本次实验中带标注的样本总数的比例.实验中设置ratio为0.15,0.25,0.50,0.75,0.85共5组参数.接下来的实验设置部分,为了方便起见,均以ratio=0.15为例进行说明.于是对LPA实验而言,ratio=0.15意味着从带标注数据集中选取85%的样本集作为测试集,剩下15%的带标注数据集和不带标注的数据集共同构成训练集.

在对LPA算法标注矩阵$\boldsymbol{Y}$的求解过程中,对于训练集的85%的标注数据,根据其诊断结果是否为"失眠"打上0,1标注,即如果该患者诊断结果有"失眠",则给该患者打上1,标记为正样本,否则打上0,标记为负样本;对于另外一部分未标注的训练集,由于其诊断结果未知,给其每个样本打上0~1之间的随机标注;对于测试集,同样给其每个样本打上0~1之间的随机标注.

对于KNN算法和SVM算法而言,根据ratio的大小从带标注数据集中选取相应比例的标注数据作为训练集,剩下的带标注数据作为测试集.当实验中的ratio参数设置为0.15时,KNN和SVM算法均随机从标注数据集中选择15%的带标注样本作为训练集,剩下85%的带标注样本则作为测试集.根据训练集中每个样本病例是否诊断为"失眠",给其打上1,0标签进行实验.

同时,对于每1个不同的ratio参数,实验中均

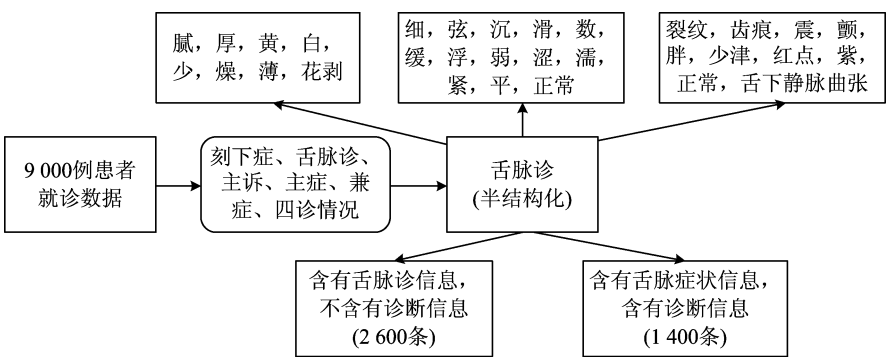

图1 数据预处理过程

Fig. 1 Data preprocessing

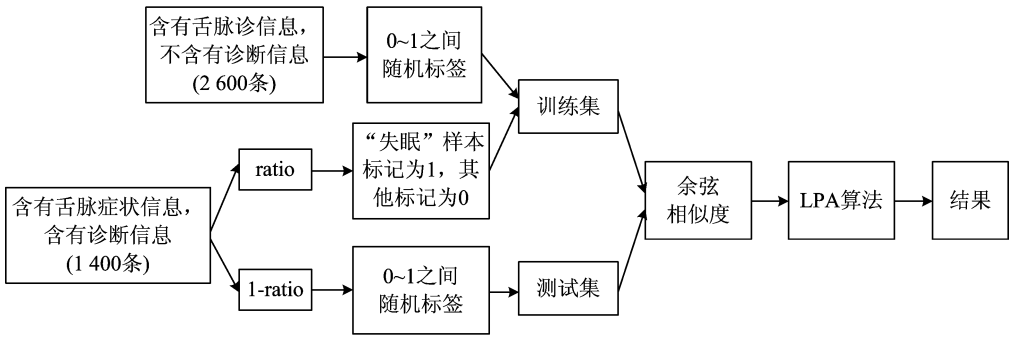

图 2　LPA 算法实验过程设置

Fig. 2　The experiment setting of LPA algorithm

采用十次随机进行相应结果的求解.

4.2　实验评价指标

本次实验过程中采用 Accuracy, Recall, Precision 和 F1 等指标进行评价.

其中,准确率(Accuracy)表示的是分类正确样本数占总样本数的比例:

$$\text{Accuracy}=\frac{(N_{\text{TP}}+N_{\text{TN}})}{(N_{\text{TP}}+N_{\text{FP}}+N_{\text{FN}}+N_{\text{TN}})}$$

召回率(Recall)表示的是分类正确的正样本数占所有正样本数的比例,其定义如下所示:

$$\text{Recall}=\frac{N_{\text{TP}}}{N_{\text{TP}}+N_{\text{FN}}}$$

在实验中 Recall 越低表示预测为“失眠”但是实际上却不是“失眠”的人越多,即误诊率越高.

精确率(Precision)表示分类正确的正样本数占该分类器所有分类为正样本数的比例:

$$\text{Precision}=\frac{N_{\text{TP}}}{N_{\text{TP}}+N_{\text{FP}}}$$

实验中 Precision 越低表示实际上是“失眠”但是经过 LPA 算法却没有将其预测为“失眠”的患者越多,即漏诊率越高.

F1 表示精确率和召回率的调和平均值:

$$\text{F1}=\frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}}$$

图 2 给出了对 LPA 算法实验过程设置的具体描述.

4.3　实验结果与分析

为了选择 LPA,KNN 算法在实验过程中最合适的近邻数,本文接下来首先给出在 ratio=0.15 的情况下,LPA,KNN 算法随着 K 值的变化趋势图. 由图 3 可以看出,当 $K<70$ 时,LPA 算法的 F1 值随着 ratio 的增加逐步增加,KNN 算法的 F1 值随着 ratio 的增加虽然有所减低,但是减低速度较慢,依然维持在较高的值. 而当 $K>70$ 时,LPA 算法的 F1 值随着 K 值的增加而下降,KNN 算法分类器的 F1 值随着 K 值的增加下降增快. 当 $K=70$ 时,LPA,KNN 都取得较好的分类效果. 因此,在 LPA 算法中计算节点与节点之间的相关性矩阵和 KNN 算法选取近邻时,均取 $K=70$.

对于 SVM 算法,利用网格算法寻优发现,当 $C=1.0$, $g=0.0$ 时,SVM 分类在该数据集上取得最优的分类结果.

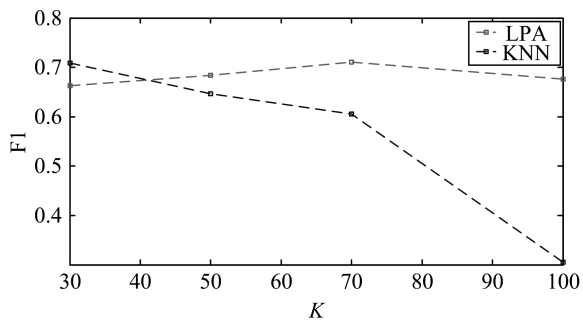

图 3　ratio=0.15 时,LPA、KNN 算法 F1 值随着 K 值的变化趋势

Fig. 3　F1 changed with the tendency of K of LPA and KNN algorithm

接下来,为了验证本文提出的非结构化转结构化数据预处理框架的有效性和通用性,分别利用监督疾病判别算法和半监督疾病判别算法对本文提出的预处理框架下和通用数据预处理框架下得到的症状特征进行对比实验,并根据不同的评价指标 F1, Recall, Precision 和 Accuracy 等画出了 LPA, KNN,SVM 3 种分类算法的分类结果比较图.

在图 4 中,pri-KNN,pri-SVM,pri-LPA 分别表示的是在本文提出的非结构化数据转结构化预处理框架下进行 LPA,KNN,SVM 的实验结果,jieba-

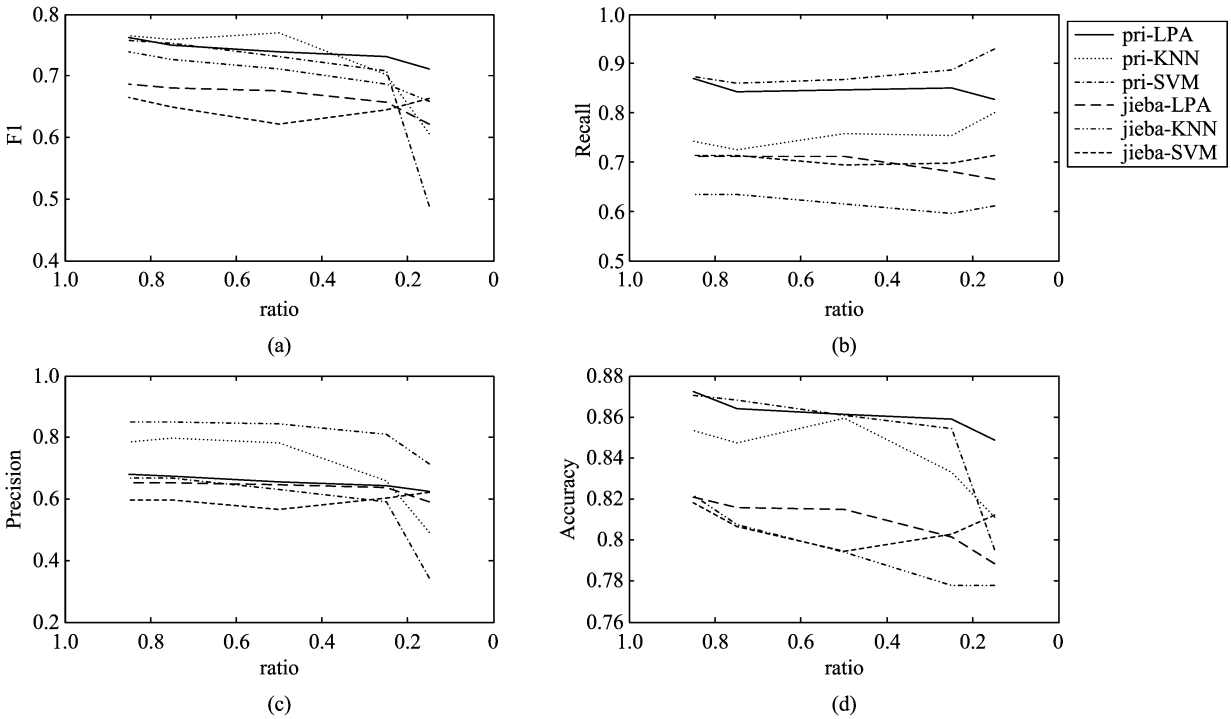

图 4 F1,Recall,Precision,Accuracy 指标随着 ratio 值的变化趋势图

Fig. 4 The change trend of values of F1, Recall, Precision and Accuracy changed with the ratio of algorithms

KNN,jieba-SVM,jieba-LPA 分别表示的是在结巴分词框架下进行 LPA,KNN,SVM 的实验结果. 从图 4 中可以很明显地看出,对同一种疾病判别算法,基本上在本文所提框架下的结果曲线都在结巴分词框架的结果曲线之上,表明本文提出的非结构化数据结构化框架比通用的非结构化数据结构化框架更适合中医领域的疾病分类模型. 至于图 4(c)中 KNN 算法在通用数据预处理框架下的效果比本文提出的非结构化转结构化框架下得到的效果较好,可能主要是由于样本的不均衡现象导致的. 由于本文研究的数据集中,含有类标的样本有 400 个是正样本,另外 1 000 个是负样本,这就致使负样本对最终的判别结果起到较大的影响.

另外,从图 4 中可以看出,在 ratio 比较大,即训练集中标注样本比较多时,KNN,SVM,LPA 3 种算法的性能相差不大;但是当 ratio 下降到 50%以下时,KNN,SVM 算法的分类效果急剧下降,而 LPA 算法还是保持在较高的水平. 可见,LPA 算法的稳定性比 KNN 和 SVM 算法有一定优势.

当 ratio=0.15 时,即随机采样 85%的标注数据作为测试集、15%的标注数据作为训练集中的标注源时进行实验,实验结果如表 1 所示.

表 1 K =70,ratio=0.15 时 LPA,KNN,SVM 的实验结果

Tab. 1 Results of LPA, KNN and SVM when K =70 and ratio=0.15

	Accuracy	F1	Recall	Precision
pri-KNN	0.81	0.61	0.80	0.49
pri-SVM	0.80	0.49	0.93	0.34
pri-LPA	0.85	0.71	0.83	0.62
jieba-KNN	0.78	0.66	0.61	0.71
jieba-SVM	0.81	0.65	0.71	0.62
jieba-LPA	0.78	0.62	0.66	0.59

由表 1 的数据可以看出,pri-KNN,pri-SVM 和 pri-LPA 在大多数指标上都比 jieba-KNN,jieba-SVM,jieba-LPA 要好,说明无论是对监督疾病判别方法还是在半监督疾病判别方法,本文提出的非结构化症状描述句结构化框架均比通用的结构化框架的效果好. 因此从一定程度可以说明本文提出的非结构化转结构化框架在中医领域比一般的非结构化转结构化框架要好.

以 15%的样本为训练集有标记部分时,半监督学习的 LPA 算法的 Recall 结果比 SVM 差,但是 Precision 结果却好很多. 如前所述,Recall 表示预测为"失眠",实际上也是"失眠"的人的比例,即 Recall

越低，误诊率越高；Precision 表示实际上是“失眠”，经过 LPA 算法也将其预测为“失眠”的患者的比例，即 Precision 越低，漏诊率越高. 对于疾病诊断来说，我们觉得漏诊率相对于误诊率而言更加重要. 我们认为，宁愿误诊而不是漏诊. 并且从 F1 上看，LPA 相比于 SVM 和 KNN 也具有一定优势. 我们认为 LPA 算法在标注样本较少时相比于 SVM 和 KNN 具有较好的预测结果.

5　结论

本文针对中医真实临床数据特征形式混杂、非结构化显著、缺失严重等特点，提出了针对中医临床数据的非结构化转结构化的数据预处理方法. 并通过实验验证，无论是在监督判别模型还是在半监督判别模型下，本文提出的非结构化转结构化框架都取得了较好的效果. 因此可以说明，本文提出的非结构化转结构化数据预处理方法在中医领域比通用的结构化数据处理方法得到的症状特征要好. 另外，在仅仅利用患者舌、脉诊信息的情况下，标签传播算法在标记数据较少时仍能取得较好的分类结果. 在后续的工作中，我们将尝试利用更多的症状特征改进现有的非结构化转结构化数据预处理框架，使其更好地服务于中医疾病分类模型.

参考文献(References)

[1] GOUDA K, RASHAD M. PreJoin: An efficient Trie-based string similarity join algorithm[C]// 2012 8th International Conference on Informatics and Systems (INFOS). Piscataway: IEEE Press, 2012: DE-37-DE-43.

[2] DENG Y, BYRNE W. HMM word and phrase alignment for statistical machine translation[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2008, 16(3): 494-507.

[3] TOSELLI A H, PUIGCERVER J, VIDAL E. Context-aware lattice based filler approach for key word spotting in handwritten documents[C]// 2015 13th International Conference on Document Analysis and Recognition (ICDAR). Piscataway: IEEE Press, 2015: 736-740.

[4] ZHU X J, GHAHRAMANI Z, LAFFERTY J D. Semi-supervised learning using gaussian fields and harmonic functions[C]// Proceedings of the Twentieth International Conference of Machine Learning. [S. l.: s. n.], 2003: 912-919.

[5] YANG Lingpeng, JI Donghong, NIE Yu. Information retrieval using label propagation based ranking[C]// Proceedings of NTCIR-6 Workshop Meeting. [S. l.: s. n.], 2007: 140-144.

[6] KIM S M, PANTEL P, DUAN L, et al. Improving Web page classification by label propagation over click graphs[C]// Proceedings of the 18th ACM Conference on Information and Knowledge Management. New York: ACM, 2009: 1 077-1 086.

[7] 贺松林，张晖. 基于 K-means 和 label propagation 的半监督网页分类[J]. 软件导刊, 2011, 10(2): 49-51.
HE Songlin, ZHANG Hui. Semi-supervised Web classification based on K-means and label propagation [J]. Software Guide, 2011, 10(2): 49-51.

[8] SPERIOSU M, SUDAN N, UPADHYAY S, et al. Twitter polarity classification with label propagation over lexical links and the follower graph [C]// Proceedings of the First Workshop on Unsupervised Learning in NLP. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011: 53-63.

[9] 任晓娟. 基于改进标注传播算法的半监督资源分类[D]. 长春：吉林大学, 2008: 1-64.

[10] 郝建柏，陈贤富，黄双福，等. 一种基于模糊近邻标签传递的半监督分类算法[J]. 微电子学与计算机, 2010, 27(2): 30-33.
HAO Jianbai, CHEN Xianfu, HUANG Shuangfu, et al. Semi-supervised classification algorithm using fuzzy nearest neighborhood label propagation [J]. Microelectronic & Computer, 2010, 27(2): 30-33.

[11] CORTES C, VAPNIK V. Support-vector networks [J]. Machine Learning, 1995, 20 (3): 273-297.

[12] NIRMALADEVI M, APPAVU S, SWATHI U V. An amalgam KNN to predict diabetes mellitus[C]// 2013 International Conference on Emerging Trends in Computing, Communication and Nanotechnology (ICE-CCN). Piscataway: IEEE Press, 2013: 691- 695.

[13] 张启明，刘保延，王永炎. 中医症状学研究[J]. 北京：中医古籍出版, 2013.