面向不均衡医学数据集的疾病预测模型研究

陈 旭[1] 刘鹏鹤[1] 孙毓忠[1] 沈 曦[1] 张 磊[4] 王晓青[3] 孙晓平[2] 程 伟[5]

[1](中国科学院计算技术研究所计算机体系结构国家重点实验室 北京 100190)

[2](中国科学院计算技术研究所智能信息处理重点实验室 北京 100190)

[3](首都医科大学附属北京朝阳医院 北京 100020)

[4](中国中医科学院中医临床基础医学研究所 北京 100700)

[5](中国中医科学院西苑医院 北京 100091)

摘 要 基于临床表现的疾病预测模型是临床决策支持系统(Clinical Decision Support System,CDSS)的一个重要研究内容.现有临床决策支持系统往往将临床病例作为训练数据集,以临床表现的描述文字为特征,采用统计机器学习方法构建疾病预测模型.然而,在医疗领域往往存在着样本数据集不均衡的问题,导致模型的预测效果降低.欠采样技术是目前解决样本不均衡问题的常用手段.其主要采用一定的方法从多数类样本中抽取部分样本,与少数类样本组成平衡数据集后再构建模型.现有的欠采样方法往往可以显著提高模型对少数类样本的召回率,然而其通常也会造成模型准确率的降低,从而限制了预测模型的整体提升效果.为此,该文提出了一种新的基于迭代提升欠采样的集成分类方法(Under-Sampling with Iteratively Boosting,USIB),该方法迭代地从多数类样本中进行欠抽样,构建多组弱分类器,并采用加权组合方式将这些弱分类器构成一个强分类器,从而提升样本不平衡条件下单种疾病预测效果.另外,医学病例样本数据集通常是多类别、多标签的,为此,该文将多个单种疾病的预测模型进行组合构成一个多标签疾病预测模型,以满足临床意义上的多病种以及并发症的诊断.为了进一步提升多标签预测模型的效果,该文设计了一种基于标签最大互信息生成树的标签选择方法(Labels Selection method based on Maximum Mutual Information Spanning Tree,LS-MMIST),该方法根据原始数据集的分布构建标签之间的最大互信息生成树,在每一次的样本预测阶段,借助树中疾病标签之间的关系确定最终的预测标签集合.实验方面,该文首先选择三种公开的不均衡二分类数据集和我们私有的四种稀有疾病的数据集,对该文提出的迭代提升欠采样方法进行性能评估.其次,分别对比了该文提出的多标签预测模型与现有的多标签预测技术在中医和西医两种多标签数据集上的预测性能.实验结果显示,相对于目前主流的八种欠采样以及两种集成采样技术,该文提出的迭代提升欠采样方法在各个不均衡二分类数据集上的 $F1$ 值平均提升 22.58%;与现有的各种多标签预测技术相比,该文提出的多标签预测方法在西医和中医数据集上正确率分别提升 6.30%和 12.43%,召回率分别提升 4.33%和 5.86%,$F1$ 值分别提升 5.48%和 11.16%.

关键词 疾病预测;不均衡数据集;欠采样;二分类;多标签分类

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2019.00596

Research on Disease Prediction Models Based on Imbalanced Medical Data Sets

CHEN Xu[1] LIU Peng-He[1] SUN Yu-Zhong[1] SHEN Xi[1] ZHANG Lei[4]
WANG Xiao-Qing[3] SUN Xiao-Ping[2] CHENG Wei[5]

[1](*State Key Laboratory of Computer Architecture, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)

[2](*Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)

[3](*Beijing Chao-Yang Hospital Affiliate of Capital University of Medical Sciences, Attending Pediatrician, Beijing* 100020)

[4](*Institute of Basic Research in Clinical Medicine, China Academy of Chinese Medical Sciences, Beijing* 100700)

[5](*Xiyuan Hospital of China Academy of Chinese Medical Sciences, Beijing* 100091)

Abstract The prediction of diseases based on clinical records is an important research topic of clinical decision support system. Existing clinical decision support systems often apply statistical

收稿日期:2017-03-23;在线出版日期:2017-11-15.本课题得到"面向云计算的网络化操作系统(2016YFB1000505)"、国家自然科学基金委员会(NSFC)-广东省人民政府联合基金超级计算科学应用研究专项计划(第二期)资助.**陈 旭**,男,1993 年生,硕士研究生,主要研究兴趣为机器学习和数据挖掘. E-mail: chenxuict@163.com. **刘鹏鹤**,男,1990 年生,硕士研究生,主要研究兴趣为机器学习和大数据挖掘. **孙毓忠**(通信作者),男,1968 年生,博士,研究员,中国计算机学会(CCF)会员,主要研究领域为大数据智能(机器学习)分析与计算. E-mail: yuzhongsun@ict.ac.cn. **沈 曦**,女,1972 年生,助理经济师,主要研究兴趣为互联网医疗、智慧医疗产业. **张 磊**,男,1981 年生,博士,助理研究员,主要研究兴趣为中医临床数据挖掘研究. **王晓青**,女,1965 年,主治医师,专业为呼吸道、消化道、新生儿疾病. **孙晓平**,男,1973 年生,博士,副研究员,中国计算机学会(CCF)会员,主要研究领域为智能计算. **程 伟**,女,1966 年生,硕士,主治医师,中西医结合内科,从事老年病专业.

machine learning methods to construct disease prediction models with the collected clinical records as training data sets and the clinical manifestation texts as feature spaces, which can help diagnose patient disease. In medical field, the common human diseases usually have much more clinical records than the rare diseases which only have few recorded samples. The imbalance of diseases samples often has a bad effect on the model's prediction effect. Under-sampling is a common method that combines the part samples extracted from majority samples and the minority samples into a balanced dataset. The existing under-sampling methods can significantly improve the recall rate of the model while it also usually leads to the decrease of the model's accuracy at the same time, which limits the overall promotion effect of the prediction model. To address this, in this paper, we propose a new ensemble classification method based on under-sampling with iteratively boosting (USIB). The method uses a boosting method to iteratively build a set of weaker classifiers by under-sampling the majority class samples, and ensemble these weaker classifiers to a strong classifier in order to improve the effect of prediction model for single disease under imbalanced data set. Besides, the medical data sets always have multiple classes and a sample has several labels. Thus, to meet the diagnosis of multiple diseases and complications in clinical significance, we combined the single disease prediction models into a multi-label disease prediction model. In order to further improve the effect of multi-label prediction model, we designed a Label Selection method based on label Maximum Mutual Information Spanning Tree (LS-MMIST). The method builds the maximum mutual information spanning tree between labels according to the distribution of the original data set and determines the final labels by the relation between the disease labels in the tree when predicting an unknown multi-label clinical sample. In experiment, we first evaluated the predictive performance of our ensemble method based on under-sampling with iteratively boosting under the imbalanced data set by using three different open imbalanced binary data sets and four private rare diseases data sets. In detail, the proposed under-sampling method with iteratively boosting improves the average of $F1$ value by 22.58% on various imbalanced binary classification data sets. Secondly, we compared the performance of the proposed multi-label prediction model with the existing multi-label prediction techniques on Traditional Chinese medicine and Western medicine data sets. The experimental results show that our method significantly outperforms the current eight mainstream under-sampling and two kinds of ensemble sampling methods in imbalanced medical data sets especially in classes of small sample size. And compared with the existing multi-label prediction technology, our multi-label disease prediction model increased the precision rate by 6.30% and 12.43%, the recall rate by 4.33% and 5.86%, the $F1$ value by 5.48% and 11.16% respectively. In addition, compared with a LSTM model, our multi-label disease prediction model is more applicable to a small-size sample set.

Keywords　disease prediction; imbalanced data set; under-sampling; binary classification; multi-label classification

1　引　言

随着软硬件迅速发展，信息化技术广泛应用于医疗过程中，为医学诊疗提供辅助支撑. 临床辅助决策支持系统(Clinical Decision Support System, CDSS)在临床诊断过程中，根据患者当前的病症信息，依据系统知识库和推理分析计算，对病情进行分析预测提示，对诊断治疗方案决策提供辅助支持信息. 智能辅助决策支持系统依据海量知识库可以帮助医生在临床诊断决策过程中更高效、更快捷地运用复杂医学知识处理各种医学问题，避免遗漏、错过重要的信息和线索，为疑难杂症寻找更多的解决方案.

疾病预测模型是智能辅助诊断系统的核心挑战之一，可分为基于规则的专家模型、基于统计知识的

统计分析模型和基于机器学习的预测模型.20 世纪 80 年代,美国斯坦福大学的 Shortliffe[1] 开发了基于专家规则的辅助医疗诊断系统 MYCIN,用于鉴别细菌感染并提供治疗方案.MYCIN 总结了 400 多种体现专家诊断疾病的规则,利用谓词逻辑和一阶逻辑来模仿专家的推理过程.专家小组对医学专家、实习医生以及 MYCIN 系统的行为进行正式测试评价,认为 MYCIN 超过了临床医生助手的作用.随后医学专家模型的开发进入了高潮.医学专家系统的局限在于需要人工去总结大量的专家规则,维护成本过高并且拓展性不好.为了解决专家系统存在的问题,统计学习知识被运用到医学数据处理中,例如 IBM Watson 医疗辅助诊断系统[2] 通过对医学文献进行统计分析以帮助临床医生进行决策.近年,统计机器学习模型得到了飞速的发展,被广泛应用于医学病例的学习建模和预测.机器学习预测模型将疾病诊断过程看作是以疾病临床表现为特征的统计分类预测问题,根据疾病临床表现建立样本特征空间,将已有病例的样本特征和对应的标记(即诊断结果)作为训练集合,采用统计分析模型训练分类预测函数,从而可以对新病例进行预测分析.其次,深度神经网络的提出更极大提升了疾病的预测能力,例如,在文献[3]中,作者基于八万份多标签电子门诊病例构建一个 LSTM 预测模型,用于门诊病例数据的多标签疾病预测.

基于电子病例的疾病预测需要解决的一个核心问题是样本不均衡问题.某些常见疾病病例来源很多,在整体病例中比例很大;而不常见疾病则只能从书本经典教材上获取少量的病例样本.比如某三甲医院的门诊病例中的“呼吸道感染”的病例占了 50%以上,而“脑膜炎”等疾病所占比例不足 1%.甚至有很多病例只能来源于书本教材知识,只有极少量的病例.

因此,样本不均衡是基于病例医学数据构建机器学习预测模型中需要重点关注的问题.在不均衡数据集上训练疾病预测模型,主流机器学习算法会倾向于将样本预测为多数类(正样本占少数时,正样本为少数类,同时负样本为多数类),虽然可以获得较高的正确率,但是会导致模型召回率极低,以至于预测模型无法将正样本正确分类.目前,欠采样技术是解决样本不均衡问题的常用手段,然而现有方法虽然可以提升少数类样本的召回率,但是同时导致正确率下降,从而影响了预测模型的整体性能.这是由于欠采样后的数据分布与采样前的数据分布存在一定的偏差,致使很多多数类样本被预测错误,从而降低了整体的正确率.

针对这一问题,本文提出一种新的基于迭代提升欠采样的集成方法,用于提升不均衡二分类数据集中少数类样本的识别能力,并同时保证模型整体的预测性能.该方法将欠采样技术和迭代提升方法相结合,最终训练生成一个集成分类器.在每次迭代阶段,首先根据多数类样本的采样概率进行欠采样,从而生成训练数据集,进而构建基础分类器并加入到集成分类器中.随后,根据集成分类器在训练数据集上的预测效果动态更新多数类样本的采样概率以进行下一次的迭代.

另外,医学门诊病例通常是多类别、多标签数据,如图 1 所示.因此,为了满足临床意义上的多病种以及并发症的诊断,我们将该诊断问题形式化为一个多标签分类问题.首先,本文利用迭代提升欠采样方法构建出每个单种疾病的预测模型,然后将多个疾病预测模型组合为一个多标签预测模型.进一步地,本文提出了一种基于标签最大互信息生成树的标签选择方法,该方法通过疾病标签之间的互信息来更新疾病的预测概率,从而确定未知样本的最终标签集合.

患者编号:[00004041XXX] 日期:[2016-11-01]
性别:[女] 年龄:[12] 科别:[儿科]
主诉:[咳嗽2天]
现病史:[咳嗽2天,声咳,有痰,白天咳重,伴少量流涕,不发热,无发憋及喘息.发病以来饮食正常,大便、小便正常.今早呕吐]
既往史:[体健] 药物过敏史:[无]
个人及家族史:[遗传病无]
体格检查:[精神反应好,咽轻度充血,双侧扁桃体1度肿大,未见异常分泌物。口唇红润,无鼻扇及三凹征,双肺呼吸音粗重,未闻干湿罗音.心音有力.腹软,未及包块.]
辅助检查:[无]
初步诊断:[呼吸道感染,消化不良,呕吐]

图 1 医学病例示意图

首先,本文选择三种公开的不均衡二分类数据集和我们私有的四种稀有疾病的数据集,对本文提出的迭代提升欠采样方法进行性能评估.结果显示,与现有八种欠采样方法和两种集成采样方法相比,本文提出的方法效果最好,$F1$ 值平均提升了 22.58%.其次,评估本文提出的多标签预测模型在私有中西医多标签数据集上的预测效果,结果显示,与现有的主流多标签预测技术相比,本文提出的多标签预测方法在西医与中医数据集上正确率分别提升 6.30% 和 12.43%,召回率分别提升 4.33% 和 5.86%,$F1$ 值分别提升 5.48% 和 11.16%.

2　相关工作

随着信息技术的不断发展，机器学习模型被用于构建疾病辅助诊断系统，从最初的传统机器学习模型，到后面的人工神经网络，机器学习的快速发展促进了智能医疗的不断进步.

2.1　机器学习疾病预测模型

传统机器学习模型被广泛地运用在预测单标签疾病上[4-5]. 早在 1992 年，文献[4]就已经提出了一种基于视力障碍、严重疾病、认知障碍和高血尿素氮肌酐比的预测谵妄症的模型，取得了良好的效果. 贝叶斯模型由于其良好的解释性被广泛应用于疾病诊断模型中，例如，文献[6]将贝叶斯模型用于预测阿尔茨海默病. 此外，集成学习模型[7]等也被引入到疾病辅助诊断系统中. 在国内，机器学习模型也被逐渐用于中医疾病预测分析[8-9]. 但是传统的基于统计机器学习的模型大多运用在均衡数据集中，并且均为单病种疾病预测模型，很难直接适用于不均衡多标签医学数据集的预测中.

与传统的统计学和临床方法相比，人工神经网络在预测临床疾病和辅助治疗方面有更好的效果[10]. 文献[11]对比了人工神经网络和逻辑回归模型在医疗辅助诊断预测上的效果，作者分别训练急性冠状动脉综合征疾病诊断预测模型，结果表明人工神经网络在诊断疾病冠状动脉综合征上效果更好. 此外，Das 等人[12]提出了一种诊断瓣膜性心脏病的集成学习方法，其使用神经网络模型作为基础模型，并组合多个神经网络模型以建立更强的神经网络模型. 为了提高医疗诊断预测模型的效果，更多的因素被考虑进来，文献[13]提出了一种基于传统疾病特征和劳动性疼痛这样的遗传因素的人工神经网络模型，以诊断冠心病，实验表明加入更多的信息后，诊断效果具有很大的提升.

虽然大量的人工神经网络已经用于医疗诊断，但是神经网络模型需要从大量的样本中提取特征以训练诊断预测模型，很难直接用于构建少数类医学疾病诊断模型.

2.2　多标签疾病预测模型挑战

在临床诊断中，医学病例具有多标签特点和不均衡特点[14-15]，从而使得构建基于门诊病例的多标签医学预测模型面临巨大的挑战. 不均衡特点：一些疾病非常常见，具有很多临床病例记录样本，但是一些类别的疾病实例很少，甚至只能从书籍上获取少量的疾病数据；多标签特点：医学门诊诊断中，医生需要根据患者的主诉、体格检查、化验检测等信息去一一判断患者所患疾病，同时，一个门诊病例的诊断报告通常具有多个诊断结果. 因此，疾病诊断预测问题是一种不均衡多标签数据集上的预测问题.

主流的分类预测模型通常是建立在数据平衡基础上，其分类结果会趋向于多数类，从而导致少数类别的信息不能被有效地收集，从而无法保证模型对少数正样本(特别是异常类)的分类精度. 目前，学术界提出了很多解决不平衡数据集上模型训练和预测的方法[16-17]，总体可以分为基于抽样技术的算法、基于集成技术的算法以及基于样本分布以及概率密度的算法.

2.3　基于抽样技术的算法

文献[18]表明，在不平衡数据集的不同优化目标下，抽样均可以提高分类器的性能，因此可以采用合理的抽样方式来提高不平衡数据集下的样本分类精确度. 基于抽样技术的算法包括欠采样技术和过采样技术.

欠采样技术是解决样本不均衡问题的常用方法，采用从原始不均衡样本集中抽样获取均衡样本集的方法来进行模型的训练. 文献[19]提出了一种应用单边采样来解决不平衡数据问题的方法，基于采样技术来提高样本分类准确率. 虽然欠采样技术可以解决一些情况下的样本不均衡问题，但是现有欠采样技术采用随机从多数类样本集中抽取训练样本的方法，在提高模型召回率的情况下往往无法保证样本准确率，这无法满足医学多标签预测问题.

其次，过采样技术[20]也是解决样本不均衡问题的常用技术. SMOTE 算法[21]是主流的过采样算法，该算法摒弃了随机过采样简单复制样本的做法，可以防止随机过采样易过拟合的问题，实践证明此方法可以提高分类器的性能. 但是由于对每个少数类样本都生成新样本，SMOTE 算法容易发生样本重叠问题，并且没有考虑近邻样本的分布特点，合成样本具有一定的盲目性.

大多数现有的过采样方法不能考虑序列的时间结构，文献[22]提出了一种新的过采样算法，使用循环神经网络生成序列，取得了良好的结果. 此外，文献[23]提出了带权少数过采样 Boosting 方法(RAMOBoost)，其基于自适应合成数据，简而言之，根据基于数据分布的采样概率，RAMOBoost 在每个学习迭代过程中自适应地排序少数类别实例，并且可以自适应地将决策边界移向难以学习的少数

类样本.虽然过采样技术可以解决样本不均衡问题,但是通常针对的是连续型数值,不适用于离散型数据集.本文将医学门诊病例抽象为多个特征的组合,是典型的离散型数据集,因此不适合直接将过采样技术运用于基于门诊病例的多标签疾病预测模型中.

2.4 基于集成技术的算法

集成方法也是解决样本不均衡问题的常用方法[24-25].文献[26]提出了一种采用 AdaBoost 算法生成边界集群数据来处理不平衡数据集的方法,该方法考虑了样本的分布情况,同样是一种样本合成技术.另外,文献[27]通过 bagging 方法来提高不平衡样本下二分类模型的预测性能,多次采样训练多个基础分类器,最终组合为强分类器,虽然能够提高一定的性能,但是各个分类器之间的训练相互分离,限制了整体性能的提升.

在文献[15]中,作者将三种学习技术:(a)集合学习;(b)人工样本生成及(c)通过将数据重新标记为新集合,组合为一个新的学习框架即多样化集成分类器以用于解决样本不均衡条件下的学习问题,有效地解决了现实世界中的蛋白质甲基化预测问题.在文献[28]中,作者提出了一种新颖的集成方法,其首先将不平衡数据集转换为多个平衡数据集,然后用特定分类算法在这些数据集上构建多个基础分类器,最终将这些基础分类器组合为集成分类器.现有集成学习方法考虑了单次采样、单个算法的局限性,融合多种算法、多次采样过程,从一定程度上提升了预测模型在少数类上的召回率,但是其通常也会造成模型准确率的降低,从而限制了预测模型的整体提升效果.

2.5 基于样本分布以及概率密度的算法

此外,还有很多基于样本分布以及概率密度的方法.文献[14]提出了一种基于实例的分类方法,其考虑了集群区域内的局部点密度,提高了小样本数据集上的分类效果.文献[29]提出了一种不平衡进化自组织图(IESOMS)的混合学习模型,其减少竞争学习阶段中的局部误差以搜索获胜神经元,以便从有限和不足的少数类别中获取有用的知识,从而提高不均衡数据集上的分类效果.

目前,上述解决不均衡问题的方法也开始运用在医学领域,从而解决医学样本不均衡问题.例如,文献[30]提出了一种将人工对象引入到数据集中来解决乳腺癌恶性肿瘤分级中恶性类别的不平衡数量问题.其次,文献[31]通过研究阿尔茨海默病神经影像学数据集(ADNI),对不平衡数据的采样技术进行了分析,作者通过测试不同比例和类型的欠采样、过采样以及过采样和欠采样方法的组合效果,对特征选择和数据采样的集成系统进行研究,并得出基于多个欠采样数据集的集成模型可以产生稳定和更好结果的结论.

综上,医学门诊病例数据集具有多标签、不均衡特点,是典型的不均衡医学数据集,现有针对不均衡数据集的解决方案无法同时保证预测模型召回率和准确率,研究新的针对不均衡医学数据集的多标签分类模型具有重要意义.

3 疾病预测模型

针对医学门诊病例的样本不均衡特性,本文首先提出了一个面向单种疾病标签的基于迭代提升欠采样的集成分类方法,旨在提高预测模型对稀有疾病的检测与识别能力;同时疾病诊断具有多标签特点,因此为了满足临床上的多病种以及并发症的诊断,我们将各个单种疾病的二分类预测模型组合成一个多标签疾病预测模型,并借助标签最大互信息生成树来确定未知样本的最终预测标签集合.整体的疾病预测模型示意图如图 2 所示.

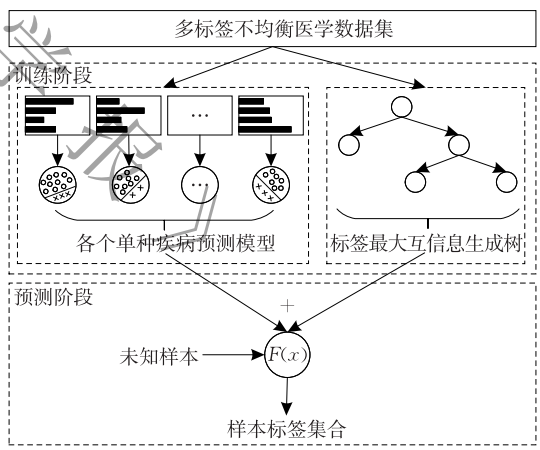

图 2 面向不均衡医学数据集的疾病预测模型

3.1 单标签疾病预测模型构建

为了解决单标签疾病数据集中样本不均衡问题,本文提出了一种基于迭代提升欠采样的集成分类方法(USIB).如图 3 所示,该方法通过多次迭代学习生成最终的疾病预测分类器.在每次迭代过程中,依据数据集中多数类样本的抽样概率对多数类样本进行多次有放回抽样,将每次抽样的结果加入到少数类样本中形成多个二分类训练数据集,使用各个训练数据集学习得到多个基础分类器;随后,分别将各个基础分类器加入到上一次迭代生成的集成

模型中形成多个新的集成模型，计算每个新的集成模型在采样前的训练数据集上的性能，选取最好的一个集成模型作为当前迭代轮次的输出模型. 与上次迭代生成的集成模型相比，如果本次迭代生成的集成模型的性能不再提高，则停止迭代；否则基于本次集成模型的预测效果来更新训练数据集中多数类样本的抽样概率，并继续迭代生成新的集成模型. 算法 1 具体地描述了基于迭代提升欠采样的二分类集成方法.

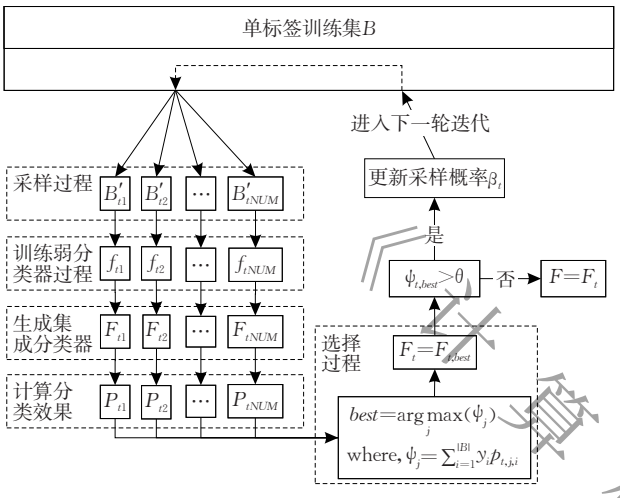

图 3　基于迭代提升欠采样的集成分类方法

算法 1.　基于迭代提升欠采样的集成分类方法.

输入：不均衡训练数据集 $B=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，其中 $x_i\in X\subseteq R^n$，$y_i\in\Upsilon=\{-1,+1\}$；基础学习算法；用户设定的收敛阈值

输出：最终分类器 $F(x)$

(1) 初始化训练数据集中多数类样本的抽样概率分布

$$D_1=(\beta_{1,1},\cdots,\beta_{1,i},\cdots,\beta_{1,|B_-|}),$$
$$\beta_{1i}=\frac{|B_+|}{|B_-|},i=1,2,\cdots,|B_-| \tag{1}$$

(2) 对 $t=1,2,\cdots,T$

(a) 使用抽样概率分布 D_{t-1} 对训练数据集中的多数类样本进行 NUM 次有放回的采样，分别加入到少数类样本中得到 NUM 个采样后的数据集：

$$\{B'_{t1},\cdots,B'_{tj},\cdots,B'_{tNUM}\},\ j=1,2,\cdots,NUM.$$

(b) 使用基础学习算法在 NUM 个数据集上进行学习，得到 NUM 个基础分类器：

$$\{f_{t1}(x),\cdots,f_{tj}(x),\cdots,f_{tNUM}(x)\}$$
$$f_{tj}(x):X\to\{-1,+1\},j=1,2,\cdots,NUM.$$

(c) 计算每个基础分类器在训练数据集 B 上的效果：

$$\{\alpha_{t1},\cdots,\alpha_{tj},\cdots,\alpha_{tNUM}\}$$
$$\alpha_{tj}=F_{\text{score}}(f_{tj}(x),B),j=1,2,\cdots,NUM \tag{2}$$

(d) 将 NUM 个基础分类器加入到之前的集成分类器中形成 NUM 个新的集成分类器：

$$\{F_{t1}(x),\cdots,F_{tj}(x),\cdots,F_{tNUM}(x)\}$$
$$F_{tj}(x)=F_{t-1}(x)+\omega_{tj}f_{tj}(x),\ j=1,2,\cdots,NUM \tag{3}$$
$$\omega_{tj}=\left(\frac{\alpha_{tj}}{1-\alpha_{tj}}\right)^2 \tag{4}$$

(e) 计算每个新的集成分类器在训练数据集 B 的提升效果：

$$\{\Delta\psi_{t1},\cdots,\Delta\psi_{tj},\cdots,\Delta\psi_{tNUM}\},\ \psi_{tj}=\sum_{i=1}^{|B|}y_i p_{tj,i} \tag{5}$$
$$\Delta\psi_{tj}=\sum_i^{|B|}y_i(p_{tj,i}-p_{t-1,i})=\psi_{tj}-\psi_{t-1} \tag{6}$$
$$p_{tj,i}=F_{tj}(x_i),\ p_{t-1,i}=F_{t-1}(x_i) \tag{7}$$

(f) 选取提升效果最好的集成分类器作为本轮迭代的输出

$$F_t(x)=\text{argmax}_j(\{\Delta\psi_{t1},\cdots,\Delta\psi_{tj},\cdots,\Delta\psi_{tNUM}\}),$$
$$j=1,2,\cdots,NUM \tag{8}$$

(g) 若本轮迭代的提升效果小于阈值(由用户自行设定)，即模型已经收敛，则停止迭代，将本轮生成的集成分类器作为最终分类器输出：

$$F(x)=F_t(x) \tag{9}$$

否则，更新多数类样本的抽样概率分布，继续下一轮的迭代计算过程：

$$D_t=(\beta_{t+1,1},\cdots,\beta_{t+1,i},\cdots,\beta_{t+1,|B_-|}),\ \beta_{t+1,i}=\frac{\beta_{t,i}e^{(p_{t-1,i}-p_{t,i})}}{Z_t} \tag{10}$$

$$Z_t=\frac{1}{|B_+|}\sum_i^{|B_-|}\beta_{t-1,i}e^{(p_{t-1,i}-p_{t,i})} \tag{11}$$

对算法 1 作如下说明：

步骤 1. 当多数类样本抽样概率总和为 $|B_+|$ 时(式(12)所示)，可以保证多数类样本抽样期望值为 $|B_+|$，因为对于每一个样本 B_i 来说，其被抽到的期望值即为 $E(\beta_i)$，则多数类样本抽样期望为：

$$\sum_i\beta_i=|B_+| \tag{12}$$
$$\sum_i E(\beta_i)=E(\sum_i\beta_i)=|B_+| \tag{13}$$

假设初始时多数类样本具有均匀的抽样概率分布，为了使抽取的多数类样本与少数类样本数量相当，对每一个多数类样本赋予抽样概率为 $\frac{|B_+|}{|B_-|}$.

步骤 2. 进行最多 T 次的迭代训练过程，每次迭代过程生成一个新的集成二分类器，如果这个新生成的集成二分类器较上次迭代生成的集成二分类器提升效果小于设定的阈值(阈值根据实验自行设定，通常为一个很小的数，例如 0.05，是迭代算法中判断收敛的一种方式)，此时表明集成模型已经收敛，退出迭代，并将这个新的集成分类器作为最终分类器 $F(x)$.

每次迭代过程中，步骤 a 首先根据抽样概率分布 D_{t-1} 对训练数据集中的多数类样本进行 NUM 次有放回的欠采样，分别加入到少数类样本中从而得到 NUM 个互不相同的训练数据集 $\{B'_{t1},\cdots,B'_{tj},$

$\cdots,B'_{tNUM}\}$,$j=1,2,\cdots,NUM$.

进而,步骤 b 使用基础学习算法在 NUM 个数据集上进行学习,从而得到 NUM 个基础分类器:

$$\{f_{t1}(x),\cdots,f_{tj}(x),\cdots,f_{tNUM}(x)\}$$

步骤 c 计算每个基础分类器在整体训练数据集 B 上的 F 值(模型评价指标),用以衡量每个基础分类器的分类效果,得到基础分类器的权重值:$\{\alpha_{t1},\cdots,\alpha_{tj},\cdots,\alpha_{tNUM}\}$,步骤 d 将步骤 b 生成的 NUM 个基础分类器分别集成到上次迭代过程中生成的集成分类器 $F_{t-1}(x)$ 中,从而形成 NUM 个新的集成分类器

$$\{F_{t1}(x),\cdots,F_{tj}(x),\cdots,F_{tNUM}(x)\}.$$

每一个集成分类器 $F_{tj}(x)$ 由集成分类器 $F_{t-1}(x)$ 和基础分类器 $f_{tj}(x)$ 加权组成,其中 $f_{tj}(x)$ 的权重 ω_{tj} 由式(4)计算,α_{tj} 是步骤 c 计算的分类器 $f_{tj}(x)$ 的分类效果.

步骤 e 到步骤 g 中,计算每一个集成分类器的分类效果,从中选择分类效果最好的分类器作为本次迭代生成的最终的分类器 $F_t(x)$,并根据最终分类器的结果来更新样本抽样概率.

本文采用预测概率来衡量分类器的效果,具体而言,假设 y_i 为样本 B_i 的真实标签值(正样本为+1,负样本为−1),对于每一个样本 B_i,集成分类器 $F_{tj}(x)$ 均输出一个概率值 $p_{tj,i}$ 来预测样本 B_i 属于正样本的概率.因此,对于正样本,预测概率 $p_{tj,i}$ 越大(分类器预测这个样本是正样本的概率越大),表明预测模型越准确;对于负样本,预测概率 $p_{tj,i}$ 越小,表明预测模型越准确.

本文采用 ϕ_{tj}($\phi_{tj}=\sum_{i=1}^{|B|}y_i p_{tj,i}$)来衡量分类器 $F_{tj}(x)$ 的性能.ϕ_{tj} 最大值为 $|B_+|$,此时代表预测模型 $F_{tj}(x)$ 将所有样本完全正确分类(正样本预测概率全为 1,负样本全为 0).分类器 $F_{tj}(x)$ 相对于上次迭代生成的分类器 $F_{t-1}(x)$ 提升的效果可以采用如下公式计算,其中 ϕ_{t-1} 为 $F_{t-1}(x)$ 在训练集 B 上的预测效果.

$$\Delta\phi_{tj}=\sum_{i}^{|B|}y_i(p_{tj,i}-p_{t-1,i})=\phi_{tj}-\phi_{t-1}.$$

据此计算本次迭代生成的每一个集成分类器 $F_{tj}(x)$ 较 $F_{t-1}(x)$ 的提升效果 $\Delta\phi_{tj}$,并选取提升效果最大的分类器作为本次迭代的最终集成分类器 $F_t(x)$.

$$F_t(x)=\mathrm{argmax}_j(\{\phi_{t1},\cdots,\phi_{tj},\cdots,\phi_{tNUM}\})\leftrightarrow \mathrm{argmax}_j(\{\Delta\phi_{t1},\cdots,\Delta\phi_{tj},\cdots,\Delta\phi_{tNUM}\}),\quad j=1,2,\cdots,NUM.$$

当 $\phi_t-\phi_{t-1}\geq\theta(\Delta\phi_t\geq\theta)$,说明本次迭代生成的最终分类器 $F_t(x)$ 较 $F_{t-1}(x)$ 来说,预测效果有所提升,将 $F(x)$ 更新为 $F_t(x)$,并继续下一轮迭代训练;当 ϕ_t 相对于 ϕ_{t-1} 不增加或者增加的效果小于一定阈值($\Delta\phi_t<\theta$)时,表明算法收敛,则停止迭代循环,输出最终分类器 $F(x)$.

在进入下一次迭代训练过程之前,算法更新多数类样本的抽样概率.具体而言,算法根据本次迭代选择的最终分类器 $F_t(x)$ 的预测概率 p_t 和上次迭代生成的分类器 $F_{t-1}(x)$ 的预测概率 p_{t-1} 来更新多数类样本的抽样概率,从而形成多数类样本的新的样本抽样分布 $D_t=(\beta_{t+1,1},\cdots,\beta_{t+1,i},\cdots,\beta_{t+1,|B_-|})$,

$$\beta_{t+1,i}=\frac{\beta_{t,i}e^{(p_{t,i}-p_{t-1,i})}}{Z_t},$$

$$Z_t=\frac{1}{|B_+|}\sum_{i}^{|B_-|}\beta_{t,i}e^{(p_{t,i}-p_{t-1,i})}.$$

抽样概率更新原则:对于任一多数类样本 B_i,若分类器 $F_t(x)$ 较 $F_{t-1}(x)$ 预测更加准确,则降低 B_i 的抽样概率,否则增加 B_i 的抽样概率.

当 $p_{t,i}-p_{t-1,i}>0$ 时,此时分类器 $F_t(x)$ 较 $F_{t-1}(x)$ 预测更加不准确,考虑到 $e^{(p_{t,i}-p_{t-1,i})}>1\leftrightarrow\beta_{t,i}e^{(p_{t,i}-p_{t-1,i})}>\beta_{t,i}$,从而提高 B_i 的抽样概率;当 $p_{t,i}-p_{t-1,i}<0$ 时,此时分类器 $F_t(x)$ 较 $F_{t-1}(x)$ 预测更加准确,考虑到 $e^{(p_{t,i}-p_{t-1,i})}<1\leftrightarrow\beta_{t,i}e^{(p_{t,i}-p_{t-1,i})}<\beta_{t,i}$,故降低了 B_i 的抽样概率.为了保证更新后的多数类样本抽样概率的总和等于少数类样本个数,采用 Z_t 来对更新后的抽样概率进行放缩,式(14)说明了更新后的多数类样本的抽样概率期望仍为 $|B_+|$.

$$\begin{aligned}\sum_{i}^{|B_-|}\beta_{t+1,i}&=\sum_{i}^{|B_-|}\frac{\beta_{t,i}e^{(p_{t,i}-p_{t-1,i})}}{Z_t}\\&=\sum_{i}^{|B_-|}\frac{\beta_{t,i}e^{(p_{t,i}-p_{t-1,i})}}{\frac{1}{|B_+|}\sum_{i}^{|B_-|}\beta_{t,i}e^{(p_{t,i}-p_{t-1,i})}}\\&=|B_+|\sum_{i}^{|B_-|}\frac{\beta_{t,i}e^{(p_{t,i}-p_{t-1,i})}}{\sum_{i}^{|B_-|}\beta_{t,i}e^{(p_{t,i}-p_{t-1,i})}}\\&=|B_+|\frac{\sum_{i}^{|B_-|}\beta_{t,i}e^{(p_{t,i}-p_{t-1,i})}}{\sum_{i}^{|B_-|}\beta_{t,i}e^{(p_{t,i}-p_{t-1,i})}}\\&=|B_+|\leftrightarrow\sum_{i}^{|B_-|}\beta_{t+1,i}=|B_+|\end{aligned}\quad(14)$$

综上,本算法每次迭代过程中根据抽样概率有放回地采样 NUM 次,生成 NUM 个基础训练集 $\{B'_{t1},\cdots,B'_{tj},\cdots,B'_{tNUM}\}$,然后独立训练出 NUM 个基础分类器 $\{f_{t1}(x),\cdots,f_{tj}(x),\cdots,f_{tNUM}(x)\}$;对于每一个基础分类器 $f_{tj}(x)$,计算权重 ω_{tj},与上次迭代生成的最终分类器 $F_{t-1}(x)$ 一起组成新的集成分类器

$F_{tj}(x)$，形成 NUM 个集成分类器$\{F_{t1}(x),\cdots,F_{tj}(x),\cdots,F_{tNUM}(x)\}$；计算每一个集成分类器 $F_{tj}(x)$ 相对于 $F_{t-1}(x)$ 的提升效果 $\Delta\psi_{tj}$，从 NUM 个集成分类器中选择提升效果最大的分类器作为本次迭代的最终分类器 $F_t(x)$；若 $F_t(x)$ 较 $F_{t-1}(x)$ 提升效果小于设定的阈值 θ（模型收敛），则退出循环并输出最终分类器 $F(x)$，否则根据 $F_t(x)$ 预测概率 p_t 来更新每一个多数类样本的抽样概率，继续进入下一次迭代训练过程.

3.2　多标签疾病预测模型的构建

基于 One-Vs-All 的转换策略是目前学术界广泛使用的一种多标签分类方法，其基本思想是首先将多标签数据集转换为多个单标签数据集，分别构建多个单标签的二分类器，然后依次使用每一个标签的二分类模型对未知样本进行预测，如果预测结果为正，则将该标签加入到预测标签集合中. 其原理如式(15)所示.

$$H(x)=\bigcup_{l\in L}\{l\}:H_l(x)\geq\theta \tag{15}$$

其中 θ 是某一个阈值（通常为 0.5），H_l 是标签 l 的二分类预测模型，$H_l(x)$ 是预测样本 x 具有标签 l 的概率，当 $H_l(x)$ 大于阈值 θ 时，模型认为此样本具有标签 l. $H(x)$ 表示集成模型对样本 x 的最终分类结果集合，对应于样本 x 的多标签属性.

本文使用基于 One-Vs-All 的转换策略构建多标签疾病预测模型. 我们将多标签疾病数据集拆分为各个疾病的二分类数据集，并使用 3.1 节中的基于迭代提升欠采样的集成分类方法依次构建每个疾病的二分类模型. 与现有方法相比，在预测阶段，我们没有将每个标签看成是相互独立的随机变量，而是充分考虑了疾病与疾病之间的相关性. 具体而言，本文采用了一种基于标签最大互信息生成树的标签选择方法（LS-MMIST）用于确定最终的预测标签集合，如算法 2 所示.

算法 2.　多标签疾病预测方法.

输入：多标签疾病数据集 $B\_ML=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，其中 $x_i\in X\subseteq R^n$，$y_i\in\Upsilon=\{l_1,l_2,\cdots,l_K\}$，$K$ 为标签个数；kruskal 最小生成树算法；基于迭代提升欠采样框架的集成分类算法；未知样本 x

输出：样本 x 的标签集合

(1) 训练阶段

(a) 计算多标签疾病数据集中各个疾病标签之间的互信息

$$\mathrm{I}(l_i;l_j)=\sum_{u\in L_i}\sum_{v\in L_j}p(u,v)\log\frac{p(u,v)}{p(u)p(v)},$$
$$L_i=L_j=\{0,1\},\ 0<i<K,\ 0<j<K,\ i\neq j \tag{16}$$

(b) 将每个疾病标签当作一个结点，使用 kruskal 最小生成树算法，生成最大互信息生成树，即向树中添加边时优先选择互信息最大的边

$$T_{l_1,l_2,\cdots,l_K}=\mathrm{kruskal}(\{\mathrm{I}(l_1;l_2),\cdots,\mathrm{I}(l_i;l_j),\cdots,\mathrm{I}(l_{K-1};l_K)\}) \tag{17}$$

(c) 将多标签疾病数据集按照疾病标签个数拆分为 K 个单标签数据集

$$\{B_{l_1},B_{l_2},\cdots,B_{l_K}\}.$$

(d) 使用基于迭代提升欠采样的集成分类算法构建 K 个疾病标签的二分类预测模型

$$\{F_{l_1}(x),F_{l_2}(x),\cdots,F_{l_K}(x)\}.$$

(2) 预测阶段

(a) 使用训练好的 K 个疾病的二分类模型分别预测样本 x 具有相应疾病标签的概率

$$\{p_{l_1}(x),p_{l_2}(x),\cdots,p_{l_K}(x)\}.$$

(b) 选取具有最大预测概率的疾病标签作为生成树的根结点，开始遍历生成树，更新当前结点的预测概率

$$p_{cur}=\begin{cases}p_{cur}, & cur=root\\ \max(p_{cur},p_{father}\cdot\mathrm{I}(father;cur)), & cur\neq root\end{cases} \tag{18}$$

如果当前结点的预测概率大于阈值 θ，则将相应的疾病标签加入到预测标签集合中；否则，不考虑，直到遍历结束.

对算法 2 作如下说明：

步骤 1. 该步骤主要用于训练模型

(a) 互信息是变量间相互依赖性的量度，我们这里用其对两个不同疾病之间的相关性进行表达，因而这里将某个疾病标签 l_i 看作为一个随机变量，那么其在数据集中的取值范围为 $L_i=\{0,1\}$，即表示该疾病是否出现. 因而，该随机变量取值为 1 或 0 的概率可通过训练数据集统计得到：

$$p(u=1)=\frac{\sum_m^{|B\_ML|}I(l_i\in y_m)}{|B\_ML|},\ 0<m<|B\_ML| \tag{19}$$

$$p(u=0)=1-p(u=1) \tag{20}$$

同样地，可以计算出其它疾病标签在训练数据集中出现或不出现的概率以及疾病标签两两之间的联合概率，从而求出疾病标签之间的互信息.

(b) 为了更好地描述疾病之间的依赖关系，我们将各个疾病标签结点映射到一棵树上，该树通过 kruskal 最小生成树算法建立，将疾病之间的互信息看作是结点之间边的权重，为了突出不同疾病之间的相关性，在建树的过程中，每次取权重最大的边加入树中，称为标签最大互信息生成树.

(c) 由于多标签数据集无法直接用于二分类模型的构建，因此，对其进行预处理，拆分为各个疾病标签相对应的二分类数据集，具体拆分策略可以表

示为针对某种疾病标签,将具有该疾病标签的样本标记为正样本,其余则为负样本.

(d)使用拆分后的各个疾病标签的二分类数据集按照算法 1 描述的基于迭代提升欠采样的集成分类方法构建各个疾病标签的二分类模型.

步骤 2. 主要用于对未知样本进行预测以确定其可能具有的标签.

(a)根据步骤 1 中构建好的各个疾病标签的二分类模型依次对未知样本进行预测其具有相应标签的概率.

(b)在步骤 1(b)中生成的标签最大互信息生成树并没有确定根结点,为了遍历树,将具有最大预测概率的疾病标签作为根结点,从根结点开始遍历树,对当前结点的预测概率进行更新,如果当前结点为根结点,则无需更新;否则,计算当前结点的父亲结点与它们之间的互信息的乘积,取该乘积与当前结点预测概率中的较大值作为当前结点的新的预测概率.判断当前结点的预测概率,如果预测概率大于阈值 θ,则将这个标签加入到标签集合中;否则,继续遍历,直至结束.

4 实验结果与分析

我们采用了两组不同的实验用于评估本文提出的两种疾病预测模型.

首先评估了本文提出的基于迭代提升欠采样的集成分类方法在不均衡二分类数据集上的性能.我们使用现有的八种欠采样技术(Random Under Sampler (RUS)[32]、Tomek Links (TL)[33]、Near Miss (NM)[34]、Cluster Centroids (CC)[35]、One Sided Selection (OSS)[36]、Edited Nearest Neighbours (ENN)[37]、Neighbourhood Cleaning Rule (NCR)[36,38] 和 Repeated Edited Nearest Neighbours (RENN)[39])以及两种集成采样技术(Easy Ensemble (EE)[40]、Balance Cascade (BC)[40])作为我们的对比方法.使用 3 种公开的不均衡二分类数据集[41]与我们私有的四种单种疾病数据集对各个方法进行评估,如表 1 所示,其中 IR[41] 指的是多数类样本与少数类样本的比例,主要用于衡量二分类数据集的不平衡度.由于部分数据集名称较长,为了显示方便,后续的结果表中的数据集名称使用表 1 中各个数据集英文名称的简写或缩写代替.另外,由于我们的集成分类方法的基础分类器为决策树模型,因此对于其它的采样技术同样采用以决策树为基础分类器的 Adaboost 模型作为预测模型.

表 1 本文使用的不均衡二分类数据集

	数据集	样本数	正例数	IR
公开数据集	abalone 9~18 (aba)	731	42	16.4
	yeast 6 (yeast)	1484	35	41.4
	poker-8-9_vs_6 (poker)	1485	25	58.4
私有数据集	哮喘性支气管炎 Asthmatic Bronchitis (AB)	1990	12	164.8
	胃炎 Gastritis (Ga)	1990	51	38.0
	肝木乘脾 Liver Subjugating Spleen (LSS)	1454	36	39.4
	血瘀 Blood Stasis (BS)	1454	75	18.4

在实验之前,我们首先选择不同阈值 0.1、0.05、0.01、0.001 进行预实验,最终我们选择 0.01 作为收敛阈值,既保证了收敛速度又保证了分类器性能,其次,为了避免随机性对实验结果造成的影响,本文做了 5 次实验,每次实验随机选取 80% 的数据作为训练数据集,余下的 20% 作为测试数据集,取 5 次实验的平均值作为最终结果.本文采用二分类最常用的正确率(*Precision*),召回率(*Recall*)和 *F*1 值(*F*1)三种指标对实验结果进行评估.

实验结果如表2~4所示.在正确率方面,本文

表 2 各种欠采样方法在不均衡数据集上的正确率 /%
(USIB 在除 poker 外的数据集外表现均好于其它方法)

	yeast	aba	poker	AB	Ga	LSS	BS
RUS	10.26	9.71	3.65	8.12	17.93	8.64	14.57
NM	1.62	6.10	6.10	5.94	6.97	5.53	12.85
TL	48.00	25.59	60.00	56.67	68.34	18.61	30.95
CC	8.08	12.62	2.61	1.92	15.08	3.57	4.24
OSS	43.71	23.75	60.00	58.57	71.06	15.11	28.35
ENN	45.24	15.16	60.00	44.64	52.21	15.27	40.02
NCR	46.64	16.39	**70.00**	42.14	45.84	15.29	23.11
RENN	31.66	18.50	60.00	41.07	22.00	13.59	41.52
EE	19.47	19.47	7.01	41.07	18.52	8.99	26.27
BC	21.19	21.19	8.02	13.05	23.91	10.33	25.89
USIB	**51.79**	**34.62**	55.46	**83.33**	**84.06**	**49.34**	**48.14**

表 3 各种欠采样方法在不均衡数据集上的召回率 /%
(USIB 表现稍弱于其它最好的欠采样)

	yeast	aba	poker	AB	Ga	LSS	BS
RUS	89.99	45.00	70.00	93.33	95.56	70.00	76.00
NM	66.67	62.50	62.50	86.67	**97.78**	62.50	80.00
TL	33.33	37.50	10.00	66.67	66.67	25.00	42.00
CC	**93.33**	**70.00**	**96.67**	93.33	95.56	**87.50**	**82.00**
OSS	33.33	32.50	10.00	73.34	68.89	20.00	44.00
ENN	50.00	15.00	10.00	93.33	88.89	25.00	50.00
NCR	50.00	17.50	13.34	93.33	**97.78**	25.00	42.00
RENN	60.00	22.50	10.00	93.33	88.89	25.00	58.00
EE	50.00	50.00	53.33	93.33	**97.78**	65.00	76.00
BC	40.00	40.00	20.00	**100.0**	95.56	60.00	64.00
USIB	83.33	67.50	86.66	80.00	88.89	77.50	**82.00**

表 4　各种欠采样方法在不均衡数据集上的 *F*1 值 /%
（USIB 在各个数据集上表现均好于其它方法）

	yeast	aba	poker	AB	Ga	LSS	BS
RUS	18.21	15.94	6.90	14.67	30.18	15.34	24.30
NM	3.17	11.11	11.11	10.99	12.43	10.14	22.08
TL	38.83	30.38	17.14	60.95	67.45	21.31	35.60
CC	14.52	21.34	5.09	3.63	25.93	6.84	8.04
OSS	37.43	27.36	17.14	63.43	69.69	17.15	34.37
ENN	46.37	15.04	17.14	59.67	63.75	18.92	44.33
NCR	47.94	16.91	22.14	57.25	62.40	18.93	29.74
RENN	41.44	20.28	17.14	56.16	35.27	17.59	48.37
EE	28.01	28.01	12.38	56.16	31.13	15.79	39.03
BC	27.66	27.66	11.42	23.09	38.25	17.62	36.81
USIB	**63.65**	**45.72**	**66.75**	**79.62**	**86.01**	**59.67**	**59.89**

提出的迭代提升欠采样方法在除 poker 数据集外的其它 6 份数据集上表现均优于其它欠采样方法；而在召回率方面，本文方法稍弱于其它采样方法中最好的结果，但也优于大多数欠采样方法. 单纯的看正确率和召回率并不能说明一个方法的优劣，在 *F*1 值方面，我们的方法在各个数据集上普遍优于其它欠采样方法.

其它欠采样方法正确率低而召回率高的原因在于采样出的负样本的分布和整体负样本分布可能存在偏差，导致模型对负样本学习不够，从而将更多的负样本错误预测. 本文方法通过每次迭代抽取不同的负样本，将抽样出的样本分布逼近整体负样本的分布，确保了模型学习到更多的不同负样本的特征，从而在保证预测结果的召回率的同时提高预测结果的正确率.

其次，我们对本文提出的多标签疾病预测模型（实验结果表中使用 LS-MMIST 表示）进行评估，采用的数据集为中西医两种多标签数据集，数据集的属性如表 5 所示，其中标签势和标签密度计算公式见文献[42].

表 5　本文使用的多标签中西医数据集

数据集	样本数	标签数	特征数	标签势	标签密度
西医	1990	77	161	1.744	0.023
中医	1454	34	438	1.378	0.041

对比模型有 ML-kNN[43]，基于 One-Vs-All 的 Adaboost[44] 多标签预测模型（ML-Ada）. 另外，我们将在不均衡二分类数据集上表现较好的 5 种欠采样技术（Tomek Links、One Sided Selection、Edited Nearest Neighbours、Neighbourhood Cleaning Rule 和 Repeated Edited Nearest Neighbour）加入到基于 One-Vs-All 的 Adaboost[44] 多标签预测模型构成新的预测模型（ML-TL-Ada、ML-OSS-Ada、ML-ENN-Ada、ML-NCR-Ada、ML-RENN-Ada）以扩充我们的对比模型. 每一个 Adaboost 集成模型也均采用决策树模型作为基础分类器. 此外，我们采用隐藏层为 512 个神经元的多标签神经网络（ML-NN）模型进行了实验评估.

本文采用适用于多标签分类的正确率 Precision，召回率 Recall 和 *F*1 值[42] 对多标签分类效果进行评估. 定义如下：设 D 是一个多标签数据集，其中包含了 $|D|$ 个多标签样本 (x_i, Y_i)，$i=1,\cdots,|D|$，$Y_i \subseteq L$，L 是标签集合. 设 H 是一个多标签分类器，而 $Z_i = H(x_i)$ 是由 H 对样本 x_i 预测的标签集合. *Precision*，*Recall* 和 *F*1 值分别由式（21）～（23）计算得出：

$$Precision(H,D)=\frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i \cap Z_i|}{|Z_i|} \tag{21}$$

$$Recall(H,D)=\frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i \cap Z_i|}{|Y_i|} \tag{22}$$

$$F1(H,D)=\frac{2\times Precision(H,D)\times Recall\ (H,D)}{Precision(H,D)+Recall(H,D)} \tag{23}$$

同样，对每个模型我们均做了 5 次实验，每次实验随机选取 80% 的数据作为训练集，剩余的 20% 作为测试集，最后取多次实验结果的平均值作为最终结果.

实验结果如表 6 所示. 从表 6 中可以看出，在西医数据集上，本文提出的多标签预测模型明显好于其它对比模型，其中正确率为 82.25%，相对对比模型中表现最好的 ML-*k*NN 模型提高 6.3%；召回率为 76.54%，提高 4.33%；*F* 值为 77.49%，提高 5.48%. 另外，在中医数据集上，各个方法的表现均有所下降，这与数据集自身的属性有关. 然而，本文方法仍然要好于其它对比模型，其中正确率为 56.41%，相对对比模型中正确率最高的 ML-NCR-Ada 模型提高 12.43%，召回率为 66.03%，相对对比模型中正确率最高的 ML-RENN-Ada 模型提高 5.86%，*F*1 值为 58.03%，相对对比模型中正确率最高的 ML-ENN-Ada 模型提高 11.16%.

另外，我们也评估了文献[3]中提出的基于 LSTM 的 RNN 分类方法在我们的西医数据集上的效果，结果并不理想，如表 7 所示.

这是由于原文方法使用的美国电子病例，这与中国门诊病例有很大不同. 美国电子病例具有强烈的时序特征，每隔一个小时会记录患者的检测指标信息. 我国的门诊病例主要是文本形式，记录着患者

表 6 不同多标签分类方法在中西医数据集上的效果

	西医数据集			中医数据集		
	precision	recall	*F*-score	precision	recall	*F*-score
ML-kNN	0.7595	0.7221	0.7201	0.4052	0.4439	0.4103
ML-Ada	0.7560	0.7201	0.7176	0.3879	0.4325	0.3922
ML-TL-Ada	0.749	0.7175	0.7115	0.3974	0.4439	0.4040
ML-OSS-Ada	0.7152	0.7021	0.6828	0.3577	0.4457	0.3776
ML-NCR-Ada	0.5742	0.7162	0.6108	0.4398	0.5464	0.4641
ML-ENN-Ada	0.5420	0.7153	0.5917	0.4388	0.5683	0.4687
ML-RENN-Ada	0.3617	0.9038	0.4857	0.4140	0.6017	0.4663
ML-NN	0.6175	0.5718	0.5717	0.3748	0.4413	0.3961
LS-MMIST	**0.8225**	**0.7654**	**0.7749**	**0.5641**	**0.6603**	**0.5803**

表 7 LSTM 在西医数据集上的效果

	Precision	Recall	*F*1-score
LSTM	0.3442	0.3153	0.3171

主诉和现病史信息，时间特征很少，并且粒度较粗，往往以天为单位，如“患者 3 天前出现发热，2 天后好转”. 在实验上，我们将西医病例中的时序信息抽取并标注出来以适用于基于 LSTM 的 RNN 分类模型. 然而由于时序信息较为稀疏，文献[3]中提出的方法在我们的西医数据集上并没有取得较好的效果. 而中医数据集并没有时序信息，因此并未做相关实验.

5 总 结

医疗数据集往往是不均衡的，不同来源的疾病具有不同的样本数，有的疾病类别会有上千甚至上万个样本案例，而有的疾病类别却只有少量甚至 1 个样本，这大大增加了现在主流预测分类模型应用的难度. 基于大数据的主流算法，以总体精度为优化目标，往往会偏向于比例较大的类，从而淹没了少数类样本信息. 在医学领域中，少数类样本也极其重要，疑难杂症的预测可以极大地帮助医生对患者进行提前治疗.

本文提出了一种新的基于迭代提升欠采样的集成分类方法，该方法每次根据样本采样概率对多数类样本进行多次欠采样，训练多个分类器，从中选择最好的分类器并根据分类器预测结果去更新多数类样本的采样概率直至模型的性能达到一个稳定阶段. 另外，为了满足医学临床诊断中的多病种以及并发症的诊断，我们将各个基于迭代提升欠采样的单病种预测模型组合为一个多标签预测模型，并利用标签最大互信息生成树来确定待预测样本的标签集合.

首先我们通过实验对本文提出的基于迭代提升欠采样的集成分类方法在 3 种公开的不均衡二分类数据集与我们私有的 4 种疾病数据集上进行实验评估，结果显示我们的方法比现有 8 种欠采样方法以及两种集成采样方法在这 7 种不均衡数据集上 *F*1 值平均提升 22.58%. 其次，我们使用私有的中西医多标签数据集对本文的多标签预测模型进行评估，结果显示，与现有的各种多标签预测技术相比，本文提出的多标签预测方法正确率分别提升 6.30% 和 12.43%，召回率分别提升 4.33% 和 5.86%，*F*1 值分别提升 5.48% 和 11.16%.

尽管本文所提基于迭代提升欠采样集成分类方法与多标签预测模型较其它模型在不平衡数据集上的效果有了较好的提升，但与真实医生的医学诊断思维过程还是有非常大的差距. 后续我们将进一步重点研究医学诊疗思维的理解、学习和模拟方法，将思维模型加入到预测模型中以期获得更为真实、可解读的医学病例预测效果.

致 谢 感谢《基于天河二号的生物医学健康大数据应用支撑平台》(U1611261)项目与《智能化数据中心管理、编程规范与应用生态》(2016YFB1000505)课题对本工作的支持；感谢审稿专家们给出的宝贵修改意见，让我们对我们的工作认识更加深刻；同样感谢《计算机学报》编辑部老师们的辛勤工作.

参 考 文 献

[1] Shortliffe E H. Computer-based medical consultations: MYCIN. Elsevier, 1976, 85(6): iii

[2] Kohn M S, Sun J, Knoop S, et al. IBM's health analytics and clinical decision support. Yearbook of Medical Informatics, 2014, 9(1): 154-162

[3] Lipton Z C, Kale D C, Elkan C, et al. Learning to diagnose with LSTM recurrent neural networks. arXiv preprint arXiv: 1511.03677, 2015

[4] Inouye S K, Viscoli C M, Horwitz R I, et al. A predictive model for delirium in hospitalized elderly medical patients based on admission characteristics. Annals of Internal Medicine, 1993, 119(6): 474-481

[5] Lin D, Vasilakos A V, Tang Y, et al. Neural networks for computer-aided diagnosis in medicine: A review. Neurocomputing, 2016, 216: 700-708

[6] Prince M J. Predicting the onset of alzheimer's disease using Bayes' theorem. American Journal of Epidemiology, 1996, 143(3): 301-308

[7] Chu N, Ma L, Chen X, et al. Ensemble learning for synthesis of the four diagnostics of TCM//Proceedings of the 2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW). Atlanta, USA, 2011: 843-847

[8] Zhang N L, Yuan S, Chen T, et al. Latent tree models and diagnosis in traditional Chinese medicine. Artificial Intelligence in Medicine, 2008, 42(3): 229-245

[9] Wang Y, Ma L, Liu P. Feature selection and syndrome prediction for liver cirrhosis in traditional Chinese medicine. Computer Methods and Programs in Biomedicine, 2009, 95(3): 249-257

[10] Grossi E, Mancini A, Buscema M. International experience on the use of artificial neural networks in gastroenterology. Digestive and Liver Disease, 2007, 39(3): 278-285

[11] Green M, Björk J., Forberg J, et al. Comparison between neural networks and multiple logistic regression to predict acute coronary syndrome in the emergency room. Artificial Intelligence in Medicine, 2006, 38(3): 305-318

[12] Das R, Turkoglu I, Sengur A. Diagnosis of valvular heart disease through neural networks ensembles. Computer methods and programs in biomedicine, 2009, 93(2): 185-191

[13] Atkov O Y, Gorokhova S G, Sboev A G, et al. Coronary heart disease diagnosis by artificial neural networks including genetic polymorphisms and clinical parameters. Journal of Cardiology, 2012, 59(2): 190-194

[14] Plant C, Böhm C, Tilg B, et al. Enhancing instance-based classification with local density: A new algorithm for classifying unbalanced biomedical data. Bioinformatics, 2006, 22(8): 981-988

[15] Ding Z. Diversified ensemble classifiers for highly imbalanced data learning and its application in bioinformatics [Ph. D. dissertation]. Georgia State University, Atlanta, USA, 2011

[16] He H, Garcia E A. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263-1284

[17] Branco P, Torgo L, Ribeiro R P. A survey of predictive modeling on imbalanced domains. ACM Computing Surveys (CSUR), 2016, 49(2): 1-50

[18] Van Hulse J, Khoshgoftaar T M, Napolitano A. Experimental perspectives on learning from imbalanced data//Proceedings of the 24th International Conference on Machine Learning. Corvallis, USA, 2007: 935-942

[19] Kermanidis K, Maragoudakis M, Fakotakis N, et al. Learning Greek verb complements: Addressing the class imbalance//Proceedings of the 20th International Conference on Computational Linguistics. Geneva, Switzerland, 2004: 1065

[20] Zhang X, Ma D, Gan L, et al. CGMOS: Certainty guided minority OverSampling//Proceedings of the 25th ACM International Conference on Information and Knowledge Management. Indianapolis, USA, 2016: 1623-1631

[21] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 2002, 16(1): 321-357

[22] Gong Z, Chen H. Model-based oversampling for imbalanced sequence classification//Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. Indianapolis, USA, 2016: 1009-1018

[23] Chen S, He H, Garcia E A. RAMOBoost: Ranked minority oversampling in boosting. IEEE Transactions on Neural Networks, 2010, 21(10): 1624-1642

[24] Rodríuez J J, Díez-Pastor J F, García-Osorio C, et al. Using model trees and their ensembles for imbalanced data//Proceedings of the 14th Spanish Association for Artificial Intelligence. Puebla, Mexico, 2011: 94-103

[25] Yu H, Ni J. An improved ensemble learning method for classifying high-dimensional and imbalanced biomedicine data. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2014, 11(4): 657-666

[26] Thanathamathee P, Lursinsap C. Handling imbalanced data sets with synthetic boundary data generation using bootstrap re-sampling and AdaBoost techniques. Pattern Recognition Letters, 2013, 34(12): 1339-1347

[27] Liang G, Cohn A G. An effective approach for imbalanced classification: Unevenly balanced bagging//Proceedings of the 27th AAAI Conference on Artificial Intelligence. Bellevue, USA, 2013: 1633-1634

[28] Sun Z, Song Q, Zhu X, et al. A novel ensemble method for classifying imbalanced data. Pattern Recognition, 2015, 48(5): 1623-1637

[29] Cai Q, He H, Man H. Imbalanced evolving self-organizing learning. Neurocomputing, 2014, 133: 258-270

[30] Krawczyk B, Jelen L, Krzyzak A, et al. Oversampling methods for classification of imbalanced breast cancer malignancy data//Proceedings of the Computer Vision and Graphics: International Conference. Warsaw, Poland, 2012: 483-490

[31] Dubey R, Zhou J, Wang Y, et al. Analysis of sampling techniques for imbalanced data: An $n=648$ ADNI study. NeuroImage, 2014, 87(3): 220-241

[32] Tahir M A, Kittler J, Yan F. Inverse random under sampling for class imbalance problem and its application to multi-label classification. Pattern Recognition, 2012, 45(10): 3738-3750

[33] Tomek I. Two modifications of CNN. IEEE Transactions on Systems, Man, and Cybernetics, 1976, 6(11): 769-772

[34] Zhang J, Mani I. kNN approach to unbalanced data distributions: A case study involving information extraction//Proceedings of the ICML 2003 Workshop on Learning from Imbalanced Datasets. Washington, USA, 2003: 42-48

[35] Zhang Y P, Zhang L N, Wang Y C. Cluster-based majority under-sampling approaches for class imbalance learning//Proceedings of the 2nd IEEE International Conference on Information and Financial Engineering. Chongqing, China, 2010: 400-404

[36] Kubat M, Matwin S. Addressing the curse of imbalanced training sets: One-sided selection//Proceedings of the 14th International Conference on Machine Learning. Nashville, USA, 1997: 179-186

[37] Wilson D L. Asymptotic properties of nearest neighbor rules using edited data. IEEE Transactions on Systems, Man, and Cybernetics, 1972, 2(3): 408-421

[38] Laurikkala J. Improving identification of difficult small classes by balancing class distribution//Proceedings of the 8th Conference on Artificial Intelligence in Medicine in Europe. Cascais, Portugal, 2001: 63-66

[39] Tomek I. An experiment with the edited nearest-neighbor rule. IEEE Transactions on Systems, Man, and Cybernetics, 1976(6): 448-452

[40] Liu X-Y, Wu J, Zhou Z-H. Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2009, 39(2): 539-550

[41] Alcalákdez J., Fernández A, Luengo J, et al. KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. Journal of Multiple-Valued Logic & Soft Computing, 2011, 17(2-3): 255-287

[42] Tsoumakas G, Katakis I. Multi-label classification: An overview. International Journal of Data Warehousing and Mining, 2007, 3(3): 1-13

[43] Zhang M-L, Zhou Z-H. ML-KNN: A lazy learning approach to multi-label learning. Pattern recognition, 2007, 40(7): 2038-2048

[44] Margineantu D D, Dietterich T G. Pruning adaptive boosting//Proceedings of the 14th International Conference on Machine Learning. San Francisco, USA, 1997: 211-218

CHEN Xu, born in 1993, M. S. candidate. His current research interests include medical diagnosis and data mining.

LIU Peng-He, born in 1990, M. S. candidate. His main research interests include machine learning and data mining.

SUN Yu-Zhong, born in 1968, Ph. D., professor. His main research interests include big data intelligence analysis (machine learning) and calculation.

SHEN Xi, born in 1972, assistant economist. Her main research interests include internet medical, wisdom medical.

ZHANG Lei, born in 1981, Ph. D. His main research interest is clinical data mining of Traditional Chinese Medicine.

WANG Xiao-Qing, born in 1965, chief physician. Her medical profession is the respiratory tract, the digestive tract, and the newborn.

SUN Xiao-Ping, born in 1973, Ph. D., associate professor. His main research interest is Intelligent computing.

CHENG Wei, born in 1966, M. S., chief physician. Her medical profession is geriatrics and integrative medicine.

Background

Disease prediction based on medical records is a classic problem in clinical decision support systems (CDSS). Medical data sets are often heterogeneous and imbalanced, which impose challenges on current statistical prediction models whose targets are to optimize the overall performance. Under-sampling and over-sampling based ensemble methods are two main techniques to handle imbalanced data. However, when predicting rare classes in an extremely imbalanced data set, these major techniques have no obvious improvement.

In medical diagnosis process, detecting majority class is important, but the detection of rare diseases is also very helpful for doctors to improve the diagnosis quality.

This paper proposes a new under-sampling framework. It uses a boosting method to build a set of weaker classifiers by iteratively under-sampling the majority class and ensemble these weaker classifiers to form a strong classifier. This

method can learn from the small number of samples and gain a better performance by an iterative boosting process. We also introduced a label selection method based on the maximum mutual information spanning tree to consider the relationship between the disease labels.

We evaluate the proposed method on a modern medical record sample set and a traditional Chinese medical sample set. The results show that our method performs better than other major models using the mainstream over-sampling and under-sampling techniques. Our models perform significantly better than other models on the detection of minority class.

The proposed work is under the support of the Networked Operating System for Cloud Computing (Grant No. 2016YFB1000505) and the Special Program for Applied Research on Super Computation of the NSFC-Guangdong Joint Fund (U1611261). The project is to investigate advanced statistical models for the intelligent log records analysis for predicting log case labels in a large multi-label, heterogeneous and imbalanced data set. The authors have proposed a Bayesian model for the multi-label prediction in medical records analysis. This work is a significant extension and improvement over previous models in that it can effectively process very extremely imbalanced data samples.