基于动态采样和迁移学习的疾病预测模型

胡满满[1),2)] 陈 旭[1),2)] 孙毓忠[1)] 沈 曦[1)] 王晓青[3)] 余天洋[4)]
梅御东[1),2)] 肖 立[1)] 程 伟[5)] 杨 杰[6)] 杨 焱[7)]

[1)](中国科学院计算技术研究所 北京 100080)
[2)](中国科学院大学 北京 100049)
[3)](首都医科大学附属北京朝阳医院 北京 100020)
[4)](南昌大学 南昌 330000)
[5)](中国中医科学院西苑医院 北京 100091)
[6)](中国中医科学院中医药数据中心 北京 100700)
[7)](中国人民解放军联勤保障部队第九八三医院信息科 天津 300142)

摘 要 基于门诊病历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.主流疾病预测模型将门诊病历转化为医学特征集合,将诊断结果作为输出标签,在此基础上利用机器学习算法训练疾病预测模型.不同疾病发病率的差异性导致医学样本具有不均衡、小样本特点,难以训练高效、准确的疾病预测模型.采样技术是目前解决样本不均衡问题的常用手段,其主要采用一定的策略生成均衡训练集,在均衡训练集上训练疾病预测模型,但是采样技术独立训练不同疾病的预测模型,没有考虑不同疾病模型之间的知识迁移性,限制了模型效果.迁移学习可以实现相似任务之间的知识迁移,如果将迁移学习运用到疾病预测模型训练过程中,在已有疾病诊断模型的基础上,训练新型疾病预测模型.受此启发,本文提出了基于动态采样和迁移学习的疾病预测模型,首先在多数类疾病上训练疾病预测模型,然后在此基础上训练少数类疾病预测模型,以实现不同疾病预测模型间的知识迁移.特别地,针对主流模型将疾病门诊病历转化为特征集合丢失文本信息的问题,本文提出了一种基于卷积神经网络的疾病预测模型,利用卷积神经网络提取语义信息;针对疾病模型知识迁移问题和小样本疾病训练问题,本文引入动态采样技术以构造均衡数据集,利用模型在不同样本上的预测结果来动态更新样本采样概率,目的是确保模型可以更多地关注错误分类样本和分类置信度不高的样本,从而提高预测模型的效果.本文在收集的门诊病历上进行了实验评估,实验结果表明,相对于目前主流疾病预测模型,本文提出的基于动态采样和迁移学习的疾病预测模型在准确率、召回率和 $F1$ 值上取得了重要的提升,尤其是召回率的提升具有十分重要的意义.

关键词 疾病预测;迁移学习;动态采样;卷积神经网络
中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2019.02339

A Disease Prediction Model Based on Dynamic Sampling and Transfer Learning

HU Man-Man[1),2)] CHEN Xu[1),2)] SUN Yu-Zhong[1)] SHEN Xi[1)] WANG Xiao-Qing[3)] YU Tian-Yang[4)]
MEI Yu-Dong[1),2)] XIAO Li[1)] CHENG Wei[5)] YANG Jie[6)] YANG Yan[7)]

[1)](*Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100080)
[2)](*University of Chinese Academy of Sciences, Beijing* 100049)
[3)](*Beijing Chao Yang Hospital Affiliate of Capital University of Medical Sciences, Beijing* 100020)
[4)](*Nanchang University, Nanchang* 330000)
[5)](*Xiyuan Hospital, China Academy of Chinese Medical Sciences, Beijing* 100091)
[6)](*TCM Data Centre, China Academy of Chinese Medical Sciences, Beijing* 100700)
[7)](*Department of Information, No. 983 Joint Logistic Support Force Hospital of PLA, Tianjin* 300142)

Abstract The disease prediction model based on clinical manifestation of outpatient records is an

收稿日期:2018-05-15;在线出版日期:2019-03-09.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)(U1611261)资助.**胡满满**,硕士研究生,主要研究方向为机器学习和数据挖掘.E-mail:humanman@ict.ac.cn.**陈 旭**,硕士研究生,主要研究方向为机器学习和数据挖掘.**孙毓忠**(通信作者),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**,本科,助理经济师,主要研究兴趣为互联网医疗、智慧医疗产业.**王晓青**,本科,主治医师,普儿内科,专业为呼吸道、消化道、新生儿.**余天洋**,硕士研究生,主要研究方向为机器学习.**梅御东**,博士研究生,主要研究方向为机器学习和数据挖掘、智能日志分析.**肖 立**(通信作者),博士,副研究员,中国计算机学会(CCF)会员,主要研究方向为人工智能、医学影像、计算生物.E-mail:xiaoli@ict.ac.cn.**程 伟**,博士,主任医师,主要研究领域为中西医结合防治老年疾病.**杨 杰**,博士,主任医师,主要研究领域为中医大数据、中医诊断学.**杨 焱**,硕士,副主任技师(高级工程师),主要研究方向为医院信息化建设.

important research content of Clinical Decision Support System (CDSS). The mainstream disease prediction models transform outpatient records into medical symptom sets, the diagnosis results into output labels, and use machine learning algorithms to train disease prediction models. Different incidences of diseases lead to the imbalance and small sample of diseases data, making it hard to train effective and accurate disease prediction model. Sampling techniques are the common methods to solve sample imbalance, which mainly use certain strategies to generate a balanced training dataset and train disease prediction models based on the new balanced training dataset. However, it independently trains different disease prediction models, which not considers the knowledge transfer between different disease models, which limits the performance. Transfer learning provides knowledge transferring between predictions on diseases with correlated knowledge. If transfer learning is applied to the training process of the disease prediction model, a new disease prediction model can be trained based on the existing disease prediction model. Inspired by this, this paper proposed a disease prediction model based on dynamic sampling and transfer learning. First, it trained prediction models for majority diseases, and then trained minority disease prediction model based on the majority disease prediction models, which can achieve knowledge transfer among different disease prediction models. In particular, in order to address the problem that transforming outpatient records into medical symptom sets leading to the loss of information, this paper proposed a disease prediction model based on convolution neural network, which used convolution neural network to extract semantic information. In view of achieve knowledge transfer on different disease prediction models and training models on minority diseases, this paper introduced the dynamic sampling technique to construct a balanced dataset, which used prediction results of different samples to update sample sampling probability dynamically, ensured that the model can learn more misclassification samples for improving the effectiveness of the prediction model. This paper has done an experimental evaluation on the collected outpatient records. Our disease prediction model based on dynamic sampling and transfer learning proposed in this paper had made important improvements in accuracy, recall rate and $F1$, especially the improvement of recall rate is of great significance.

Keywords disease prediction; transfer learning; dynamic sampling; convolutional neural network

1 引 言

医疗资源分布不均衡给社区医生带来了巨大挑战,同时很难训练充足的医生来缓解巨大的医疗压力.随着软硬件迅速发展,信息化技术广泛应用于医疗诊断过程中,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)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,为制定诊疗方案提供辅助支持信息,是一种十分有效的辅助诊断方法,其中疾病预测模型是 CDSS 的核心组成部分.目前,主流的疾病预测模型以门诊病历临床表现为训练数据集,利用机器学习和数据挖掘技术训练疾病预测模型,从而实现依据患者现有信息进行疾病预测分析.

机器学习预测模型将疾病诊断过程看作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,将已有病历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,从而可以对新病历进行预测分析.然而,稀有疾病正样本数量少,主流机器学习算法对少数类样本欠学习,难以训练高效疾病预测模型,以至于无法有效检测稀有疾病.因此,稀有疾病模型训练问题是目前疾病预测模型需要重点关注的问题.此外,疾病预测模型作为一种医疗智能辅助决策支持系统,能够将可能的疾病病历召回,即取得高召回率,具有更小的决策风险和更重要的辅助决策意义.目前,解决小样本模型训练问题的主要方法有采样技术和迁移学习技术.采样技术

采用一定的策略方法从原始训练样本集中生成均衡训练集，在均衡训练集上训练疾病预测模型，从而提高稀有疾病召回率，但是会导致模型准确率下降，并且独立训练不同疾病预测模型，没有充分考虑不同疾病模型之间的知识迁移性，限制了模型效果. 迁移学习是运用已拥有的知识对不同但相关领域问题进行求解的一种机器学习方法，首先在大规模样本集上训练有效模型，然后将模型作为稀有疾病的预训练模型，利用基于模型的迁移学习方法在稀有疾病数据集上继续训练模型，以此实现不同疾病模型之间的知识迁移. 迁移学习虽然可以提高稀有疾病预测模型效果，但是没有考虑医学小样本数据集的不均衡特点，无法充分训练高效疾病预测模型.

不同科室医生在专业学习过程中，学习的基础医学知识是类似的，之后通过学习本科室疾病知识进行疾病鉴别诊断. 医生在知识学习过程中具有很强的知识迁移能力，即医生在学习了诊断呼吸道感染的知识要点之后，可以很快将知识迁移到学习肺炎诊断中，从而很快掌握肺炎的诊断要点，这种知识迁移方式促进医生进行有效学习和诊断.

类比医生学习过程，如果可以有效模拟医生知识迁移能力，将迁移思想运用到疾病诊断模型训练中，在已有疾病诊断模型的基础上，训练新的疾病预测模型，则可以获取高效、准确的新型疾病预测模型. 稀有疾病（如肺炎）训练数据很难获取，难以训练学习高效、准确预测模型，但是一些疾病（如呼吸道感染）具有充足的训练数据集，可以有效学习高效、准确的预测模型，如何将常见疾病预测模型迁移到稀有疾病模型训练中，是构建高效、准确疾病预测模型要解决的问题.

此外，神经网络算法在图像识别、分类等领域取得了很好的效果，一个重要的因素就是迁移学习的运用[1-3]. Killian 等人[4]提出一种新型的隐马尔可夫决策过程（Hidden Parameter Markov Decision Process，HiP-MDP），利用贝叶斯神经网络（Bayesian Neural Network，BNN）来替换高斯过程，利用迁移学习来解决现实中存在的相似学习任务之间的迁移问题. Killian 等人指出，在最优控制理论中存在很多类似任务，如果每次都是从头训练任务，没有充分利用任务之间的相关性，无法获取高效模型. Killian 等人指出原有 HiP-MDP 采用的是高斯决策过程，很难衡量非线性交互变换，而不同任务在迁移时具有很多参数和状态之间的非线性变换，高斯过程不能很好地进行建模，因此作者采用贝叶斯神经网络来替换高斯过程，将隐含参数 w_b、状态 s 和动作 a 作为贝叶斯网络的输入，利用贝叶斯网络捕捉更加复杂的动态系统的高层次非线性交互，用于学习在给定动作 a 的基础上，状态 s 经过隐含参数 w_b 后的输出状态 s'，训练一个任务的模型之后，将构建的贝叶斯网络模型进行保存，作为相似任务的预训练模型. 在训练相似任务模型时，Killian 等人首先将预训练贝叶斯模型读取，作为本次任务训练的初始化模型，并在现有数据集上训练调整贝叶斯网络模型.

本文提出了基于动态采样和迁移学习的疾病预测模型. 主流疾病预测模型将门诊病历转化为医学特征集合，一方面导致了重要信息的丢失，另一方面，模型的好坏依赖于人工的特征设计以及特征设计的好坏. 为了更好地表征门诊病历的高层特征以解决上述问题，本文利用卷积神经网络来自动提取门诊病历的语义信息，将门诊病历转化为向量表示. 卷积神经网络通过不同的卷积核进行样本特征提取，可以提取不同长度的文本语义信息，以提高整体模型的预测性能. 同时本文提出了一种新型的动态采样技术，一方面保证采集的正负样本数量均衡，另一方面，每次迭代采样时提高分类错误的样本和分类置信度不高的样本的采样概率，让模型更加关注分类错误的样本，从而提高模型对分类错误的样本和分类置信度不高的样本的关注. 此外，本文在疾病预测模型的训练过程中引入了迁移学习，将样本充足的疾病预测模型学习到的知识迁移到样本不足的疾病预测模型中，来提高疾病预测模型的效果，同时探索研究了疾病间的共现频次（共同出现在一个病历中的次数）对迁移效果的影响. 本文提出的这种基于动态采样和迁移学习的疾病预测模型，首先采用卷积神经网络在多数类疾病上训练预测模型，然后将此模型作为稀有疾病的预训练模型，在稀有疾病训练集上继续训练疾病预测模型，同时将动态采样技术用于模型训练中，提高了模型效果. 西医门诊病历上的实验结果表明，相比于其他疾病预测模型，本文提出的采用动态采样和迁移学习技术的新型疾病预测模型在准确率、召回率和 $F1$ 值上均有较大提升.

2　相关工作

信息技术的不断发展为构建临床决策支持系统提供了更多方法，机器学习模型被逐渐用于疾病预

测模型中，大大改善了临床决策支持系统.疾病预测模型从最初的基于专家规则的预测模型，逐步发展到基于统计分析和案例分析的预测模型，再到后来的基于机器学习和深度学习的预测模型.

2.1 疾病预测模型

基于专家规则的预测模型，收集专家诊断经验，转化为诊断规则信息，进而形成疾病诊断路径，从而构建诊断模型，称为专家系统.典型的专家系统包括1976年Shortliffe开发的MYCIN专家系统[5]，用于判断细菌感染并给出相关治疗信息，和王加宽构建的颈椎病专家诊断系统[6].专家系统的核心是构建专家规则知识库，需要人工总结大量专家规则，维护成本高、不易拓展.

随着信息技术的发展和医疗数据的信息化，基于统计分析和案例分析的疾病预测模型，采用数据挖掘技术和统计分析技术从医学数据中自动总结诊断规则和获取统计知识，避免了人工的大量参与，但需要大量医学数据，不适用样本数量上的稀有疾病.包括基于对大量医学文献进行统计分析来辅助诊断的IBM Watson医疗辅助诊断系统[7].

基于机器学习的疾病预测模型将疾病预测问题形式化为分类问题[8-10]，从门诊病历数据中抽取、选择特征，训练机器学习分类模型，实现疾病的预测.如Prince提出将贝叶斯模型运用到阿尔茨海默病预测问题中，取得了很好的效果[11].此外，文献[12]提出了一种迭代提升欠采样模型(Under Sampling with Iteratively Boosting，USIB)，进行疾病预测.USIB迭代地从多数类样本中进行欠采样，构建多组弱分类器，通过加权组合方式集成为一个强分类器，最后基于标签最大互信息树对预测标签进行选择优化，实现疾病预测.

神经网络技术被逐渐应用于疾病预测模型中.Green等人[13]提出了分别基于回归模型和基于神经网络的急性冠状动脉综合征诊断预测模型，其中神经网络模型取得了更好的效果.Das等人[14]将集成学习与神经网络模型进行结合，构建了高效的疾病预测模型.Atkov等人[15]提出了基于神经网络的冠心病预测模型，充分考虑了传统疾病特征这样的遗传因素，取得了很好的效果.其次，Lipton等人[16]对多标签病历进行建模分析，构建基于长短期记忆网络的疾病预测模型，充分考虑病人医学特征变化规律，在一些疾病上取得了很好的效果.国内也将深度学习模型用于构建疾病预测模型，如蔡航[17]将深度学习模型用于诊断肺癌，利用深度学习分析医学图像，从而进行疾病诊断分析.此外，侯桂英等人[18]构建了基于深度学习的高血压诊断模型，充分考虑了不同指标之间的关系，取得了不错的效果.

虽然机器学习技术，包括深度学习技术在疾病预测模型中取得了很大的进展，但在由于机器学习模型倾向于将样本分到训练数据中样本量更大的类别中，故将其应用医疗领域时，亟需解决存在不均衡和样本量小特点的医疗数据上的模型构建问题，以得到高效的疾病预测效果.

目前，学术界提出了很多解决小样本、不均衡数据集上模型训练方法[19-20]，总体可以分为基于抽样技术的算法、基于集成技术的算法和基于迁移学习的算法.

2.2 采样技术

基于采样技术的算法通过构造均衡训练数据集来解决数据的不均衡问题[21]，主要分为欠采样技术和过采样技术.

欠采样技术从原始不均衡样本集中抽取训练样本集的子集，通过抽取部分多数类样本来降低不均衡程度，从而构造均衡训练集，如基于最近邻的采样算法[22]、基于单边采样的算法[23-24]、基于近邻清除规则的欠采样算法(neighborhood cleaning rule)[25]、基于聚类的欠采样算法[26].现欠采样技术选取部分样本作为训练集，丢失了部分多数类样本信息，虽然可以提高少数类样本召回率，但是往往导致多数类样本准确率的降低.

文献[24]提出单边采样算法从多数类样本中随机抽取部分样本来欠采样.Zhang等人[26]提出了基于聚类的欠采样技术，通过对多数类样本进行聚类分析进而按比例抽取样本来欠采样.文献[27]提出了一种基于样本权重进行欠采样的方法，采用聚类结果更新样本权重，基于权重进行采样得到均衡训练集.

过采样技术[28]采用少数类样本合成的方法提高少数类样本数量，以构造均衡训练集.Chawla等人提出了经典的SMOTE算法[29]，根据任一少数类样本的最近k个少数类样本来合成一个新样本，摈弃了随机采样容易产生过拟合的问题，但存在样本重叠问题.针对SMOTE的样本问题，Han等人[30]提出了Borderline-SMOTE，该算法更多地对处于边界上的样本进行合成，因为样本边界点更加不易区分.此外，文献[31]提出RAMOBoost算法，该算法通过自适应排序少数类样本，逐渐将决策边界移向难以分类的样本.

2.3 集成技术

基于集成技术的算法融合集成技术和采样技术，充分利用采样技术和集成技术的优点[32-34]. Liu 等人[35]提出了结合 Bagging 和欠采样技术的集成采样算法 EasyEnsemble，从多数类中有放回地采样部分样本来生成均衡训练集，多次采样训练多个基础分类器，最终加权集成为强分类器. Liang 和 Cohn[36]提出 UBagging 算法，该算法将 Bagging 算法用于不平衡数据集训练中，通过逐渐增加负样本采样数量来训练多个分类器，集成多个分类器以提高整体分类性能. 此外，Sun 等人[37]提出融合异构模型和采样技术的算法，首先利用采样技术抽取生成多个均衡数据集，将多数类样本分布在不同的训练集中，然后分别训练不同的分类器，最后利用集成技术将多个分类器组合为强分类器. 融合集成技术和采样技术的算法利用多次采样来解决单次采样样本信息缺失问题，但是每次采样均采用随机采样算法，没有充分考虑不同分类器之间的关系，限制了整体性能的提升.

2.4 迁移学习技术

迁移学习是运用已存有的知识对不同但相关领域问题进行求解的一种新的机器学习方法[3]. 迁移学习可以分为基于特征选择的迁移学习、基于特征映射的方法以及基于权重的方法[3]. 基于特征选择的方法首先识别出源领域和目标领域中共同特征，然后在源领域内基于这些特征训练分类器，最终通过目标领域特有的特征再进行分类器的优化，从而使分类器适用于目标领域的分类任务[3,38-39]，例如，文献[38]提出了一种基于联合聚类的迁移学习方法，用于解决领域外文档分类问题. 基于特征映射方法将源领域和目标领域的数据从高维特征空间映射到低维特征空间，从而使源领域和目标领域具有相同的分布，然后，利用新空间内源领域数据集训练分类器，从而对目标领域数据进行分类预测[40-42]. 基于权重的迁移方法主要包括戴文渊提出的 TrAdaBoost 算法[43]，TrAdaBoost 借助 Boosting 技术用源领域的数据知识辅助目标领域的分类任务，通过迭代减小误分类的源领域样本的权值，增加误分类的目标领域样本的权值，使得分类器逐渐将目标领域样本准确分类[3].

迁移学习在传统机器学习算法中被大量而广泛地研究应用，如 TrAdaBoost 算法等，在神经网络算法中，则主要应用于图像、视频领域，在图像识别、分类等领域，并取得了很好的效果[1-2]，如目标定位[44]、图像排序[45]、图像检索[46]等领域. 但在自然语言领域尤其是不均衡医学数据上的疾病预测问题上缺乏探索和研究，这就是我们引入迁移学习到疾病预测领域的动机和原因之一. 如果将样本充足的疾病模型学习到的知识迁移到样本不足的疾病模型的构建中，来辅助训练稀有疾病的预测模型，则可以有效解决医学数据不均衡的问题和部分稀有疾病数据不足的问题.

3 CNN 组合型疾病预测模型

本文提出了基于卷积神经网络的深度学习疾病预测模型，将门诊病历转化为词语序列，利用卷积神经网络对医学门诊病历中患者体征文本信息进行语义特征提取，训练单种疾病预测模型，实现对疾病的辅助诊断. 其中，将医院门诊病历中主诉、现病史和体格检查等文本信息作为输入，病历中初步诊断结果的疾病标签作为输出，将基于门诊病历的疾病预测问题抽象为文本二分类问题. 同时，本文提出了将迁移学习和动态采样技术融入到疾病预测模型的训练中的方法，解决医疗病例样本不均衡的问题给疾病预测模型的训练和预测性能带来的影响，来提高疾病预测模型的预测性能.

本文选择基于卷积神经网络(Convolutional Neural Network，CNN)的疾病预测模型，具有对文本自动提取语义特征进而进行分类的优点. 首先，卷积神经网络中能够对文本从语义角度进行表示进而提取特征，它将文本中的字词表达为多维空间的连续稠密向量，语义相似的词对应的词向量的也相近. 其次，卷积神经网络是一种局部连接的网络，其特征提取是自动学习实现的，卷积与池化操作可以看成是一种局部特征提取过程，相比于传统机器学习模型，避免了人工提取特征的成本和模型实现的效果对人工提取特征的好坏的依赖. 此外，主流模型从门诊病历中抽取医学特征集合，导致重要信息丢失，如将“咳嗽三天”转化为“咳嗽”，丢失了重要信息“三天”，而本文中的卷积神经网络疾病预测模型则不存在该问题. 最后，CNN 解决问题的能力比较强大，当需要解决的问题较复杂、难度较大时，需要建模较复杂的 CNN 模型，此时问题的复杂度使得算法需要足够大的样本去训练和测试模型，且复杂的 CNN 模型也需要足够大的训练样本，防止模型记住样本和防止模型的过拟合问题；当需要解决的问题较简单时，设计简单的 CNN 模型即可解决问题，而

简单的 CNN 模型不需要大量的样本去训练，在本文中，疾病的诊断通构建 CNN 二分类模型实现对单个疾病的预测，且每个疾病包含的症状比较集中，任务复杂度不是很大，故使用 CNN 模型实现疾病预测是可行、必要且优越的.

其次，疾病发病率的差异性导致疾病数据集具有不均衡特点，不同疾病的病例样本数量差异较大，如“呼吸道感染”的样本数量有 1219 例，而“支气管炎”的病例样本量只有 433 例. 而主流模型单独训练每一个疾病的预测模型，没有考虑不同疾病预测模型之间的知识迁移性，限制了模型性能的提升. 针对小样本疾病的病例样本量不足的问题和疾病间存在的共享知识的现象，本文提出了将迁移学习技术融合到疾病预测模型的训练中，改善不同疾病样本量不均衡问题带来的影响，将大样本疾病的预测模型学习的知识迁移到小样本疾病的预测模型中，进而提高疾病预测模型的收敛速度和预测性能. 其中，我们提出了以疾病共现频次作为疾病预测模型迁移学习的依据，当两个疾病对间的共现频次较高时，即同时被诊断为该两个疾病对的病例较多，也即该疾病对的病例的相似性较大，共有的知识较多，病例的相似特征较多，可在这两个疾病对之间进行模型的迁移.

此外，病例样本的极度不均衡特点，导致难以直接训练基于深度学习的多标签分类模型，而本文中病例的疾病预测通过对其进行每种疾病的二分类实现的. 训练每种疾病的疾病预测模型时，该疾病的负样本是其他所有疾病的病例，致使疾病训练集中正样本比例明显低于负样本，而机器学习模型偏向将样本分到样本量更多的类别中，如果直接训练基于深度学习的分类模型，会导致疾病的召回率低，无法满足临床诊断需求. 针对不同疾病的正负样本的不均衡问题，本文引入文献[12]中的动态采样技术，将动态采样技术融入到疾病模型的训练中，提出基于动态采样和迁移学习的深度学习疾病预测模型.

首先利用 One-Vs-Rest 方式将多标签疾病训练集转化为多种疾病的二分类训练集，以训练不同疾病的诊断模型；然后，在样本数量多的疾病上训练深度学习疾病预测模型，并将训练产生的模型参数进行保存；随后在训练小样本疾病预测模型时，选取与小样本疾病共现次数最多的多数类疾病，将此多数类疾病模型参数知识迁移到少数类疾病模型训练中，即将多数类疾病模型参数作为少数类模型的初始化值，同时采用动态采样技术获取均衡数据集来训练疾病预测模型，提高整体疾病预测模型的性能. 最后，本文将多个单种疾病预测模型组合为多标签疾病预测模型，整体疾病预测模型示意图如图 1 所示.

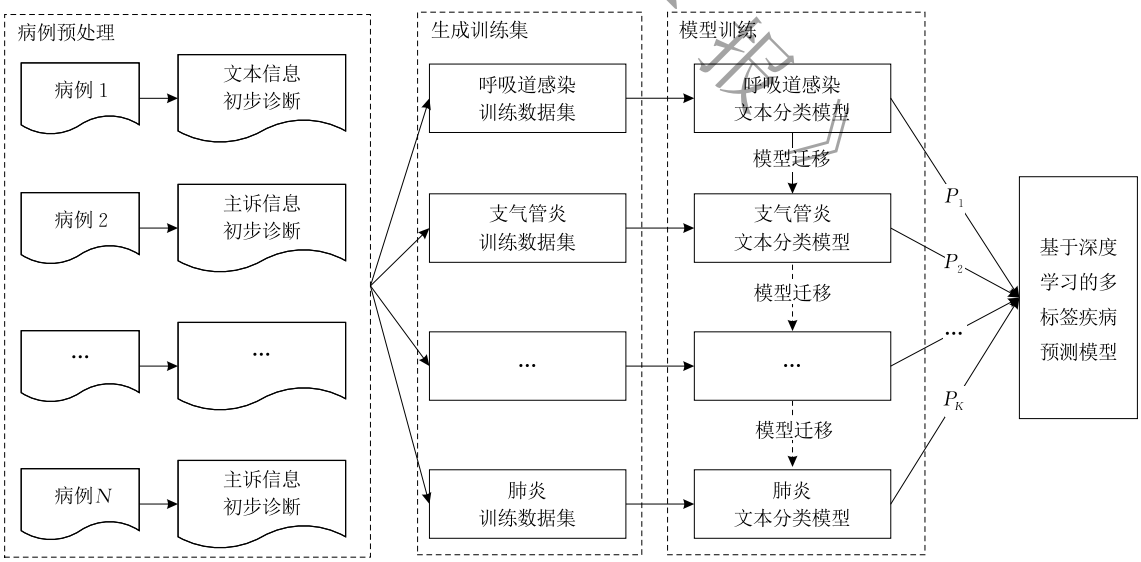


图 1 结合迁移学习和深度学习的多标签疾病预测模型框架图

3.1 基于卷积神经网络的疾病预测模型:Single CNN

在对病例进行疾病预测时，分别输入各种疾病的 CNN 疾病预测模型中，进行是否为这些疾病的二分类判断来进行病例的疾病预测. 在基于 CNN 的二分类疾病预测模型中，首先对病历中文本进行分词处理，转化为词语序列；然后，利用 Word2Vector 的 Skip-gram 模型在医疗相关文本数据上预训练词语的词向量，将离散的词语符号表示为低维连续空间的语义向量；接下来，将病例的各词语表达为其词向量后，得到表达病例的二维

词向量矩阵；最后，使用 CNN 的卷积、池化操作对病例的词向量矩阵提取特征并进行疾病的二分类. 图 2 给出了基于卷积神经网络的疾病预测模型结构图.

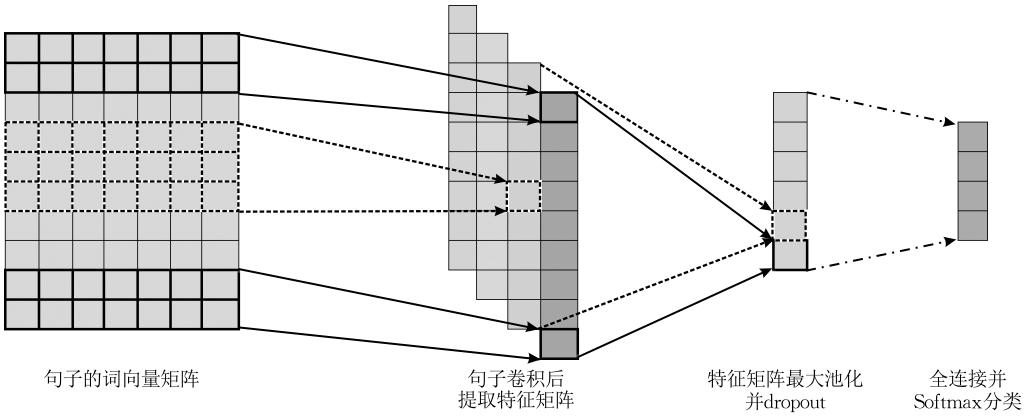


图 2　基于卷积神经网络的疾病预测模型

(1) Skip-gram 词向量模型

Skip-gram 模型是一种利用单词 W_t 来预测 n 窗口内的上下文单词 Context(W_t) = $\{W_{t-n}, W_{t-n+1}, \cdots, W_{t-1+n}, W_{t+n}\}$ 的模型，通过最大化对数似然函数来训练模型从而获得每个词的向量表示，式(1)给出了最大化的对数似然函数. 模型的结构图如图 3 所示.

$$L=\sum_{c\in \text{Context}(W_t)} \log(p(c|W_t)) \tag{1}$$

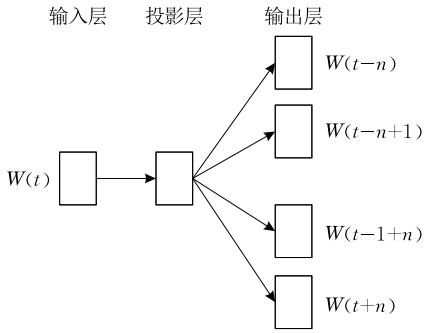


图 3　Skip-gram 模型示意图

(2) 基于卷积神经网络的预测模型

本文提出的基于 CNN 的疾病预测模型如图 2 所示. 模型包含一个卷积层和一个池化层，首先对二维特征矩阵进行卷积操作，其中卷积核的长度与词向量的长度一致，每一个卷积核产生一个列向量表示；针对每一个列向量采用最大池化方法选择其中最大值作为输出；将所有列向量的最大值按照顺序组成一个固定维度的向量，向量长度与卷积核数量一致，称为特征向量；将池化后的节点进行全连接分类.

假设采用 k 维向量来表示词向量，$\boldsymbol{x}_i \in \mathbb{R}^k$ 表示第 i 个单词的词向量表示. 门诊病历 Case 包含 n 个词语，可以将门诊病历表示为 $X_{1:n}$，其中⊕表示向量连接操作，如式(2)所示.

$$X_{1:n}=\boldsymbol{x}_1 \oplus \boldsymbol{x}_2 \oplus \cdots \oplus \boldsymbol{x}_n \tag{2}$$

生成文本二维矩阵表示之后，本文将结果输入到卷积层，利用卷积核来抽取训练数据中的一些语义特征. 给定(1) $X_{i:i+m-1}$ 表示词序列中第 i 个词到第 $i+m-1$ 个词的窗口向量矩阵；(2) 一个卷积核矩阵 $\boldsymbol{w} \in \mathbb{R}^{hk}$，目的是将 $\boldsymbol{w}$ 作用到 m 个连续的词向量产生一个输出结果.

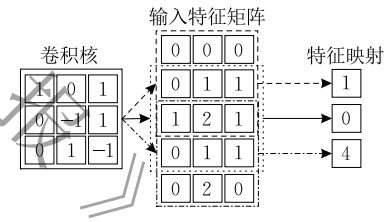


图 4　卷积操作示意图

如图 4 所示为卷积操作示意图，卷积核 $\boldsymbol{w}$ 作用在 $X_{i:i+m-1}$ 上产生的结果 c_i 可以如式(3)计算.

$$c_i=f(\boldsymbol{w}\cdot X_{i:i+m-1}+b) \tag{3}$$

其中 $f(*)$ 通常为一个非线性函数，可以为 $ReLU$ 函数、tanh 函数等，$\boldsymbol{w}$ 是上述卷积核，b 是偏置项.

为了提取文本更加丰富的数据表示，模型采用多个不同窗口的卷积核来获得更多的语义信息. 通过卷积层之后，生成了维度随句子长度变化的特征图，通常维度很大，难以直接训练合适的分类模型，将这些特征图作为池化层的输入，进行维度降低，同时捕获最重要的信息. 本文模型采用最大池化方法，最大池化将特征图中的最大值作为结果输出. 通过池化层之后，产生了固定长度的特征向量(长度与卷积核数量一致)，特征向量输入到全连接分类层进行样本分类.

3.2 基于动态采样和迁移学习的小样本疾病预测模型:CNN+TL+DS

本文在训练基于卷积神经网络的疾病预测二分类模型时,提出了一种新的动态采样技术,来降低训练数据的不均衡对疾病预测性能的影响,提出了使用迁移学习技术来提高疾病预测的性能和收敛速度.

当训练数据中各类别的数据不均衡时,卷积神经网络倾向于将样本分到训练数据中样本量更大的标签类别中.而医学门诊病历样本是不均衡的,在每个疾病的二分类训练数据中,由于正样本集是该疾病的病例样本,负样本集是其他所有疾病的病例样本,负样本数量远大于正样本数量,导致疾病预测时召回率低;此外,各个疾病类别的病例样本也是不均衡的,一些常见疾病的病历样本数量较多,而部分稀有疾病的病历样本数量较少,各疾病类别的训练数据是不均衡的,而且不同疾病类别的训练数据的不均衡程度相差较大,小样本疾病的不均衡性尤其严重,被选取进入训练集的概率远小于大样本疾病,因此导致模型对小样本疾病的欠学习,导致少数类疾病预测模型召回率低,影响疾病的多标签预测性能,无法满足临床使用.

深度学习中模型迭代训练时包含三个部分:训练样本选择、误差计算和参数更新.通常将训练数据集进行分块,每一块数据具有固定大小样本(本文中采用每块 64 个样本),按照顺序依次选择块数据进行模型训练.如图 5 所示,将训练数据集分为 n 块,依次选择块数据进行模型训练.

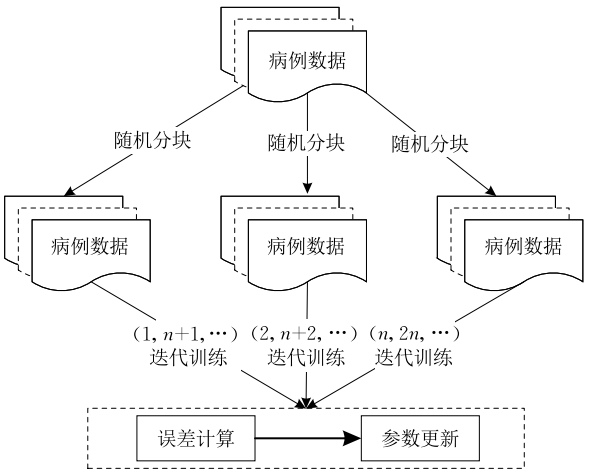


图 5 卷积神经网络训练过程

为了在小样本疾病数据集上训练有效模型,本文将迁移学习技术和动态概率采样技术运用到小样本疾病预测模型上,首先选择与本疾病共现频次高的、具有充足数据集的大样本疾病上训练高效疾病模型,将大样本疾病预测模型作为小样本模型的初始化值,重新在小样本数据集上训练疾病预测模型,同时在训练时通过动态采样选择训练样本集合.在每次迭代训练之后,根据模型在样本集上的预测结果,更新样本采样概率,增加分类错误的和分类置信度不高的样本的采样概率,进而通过动态采样构建一个均衡的训练数据集训练模型.为了保证选择的数据集中正样本个数与负样本数量的平衡,本文将正样本和负样本分开采样,最后将采样后的数据集进行合并,作为下一轮的训练数据.为了提高迁移学习的效果,本文计算两两标签之间的共现频次,选取共现次数最多的多数类疾病模型作为初始化模型.本文也提出了分别融合迁移学习和动态采样的 2 个小样本疾病预测模型(前者简称:CNN+TL,后者简称:CNN+DS).其中 CNN+DS 模型流程如图 6 所示.

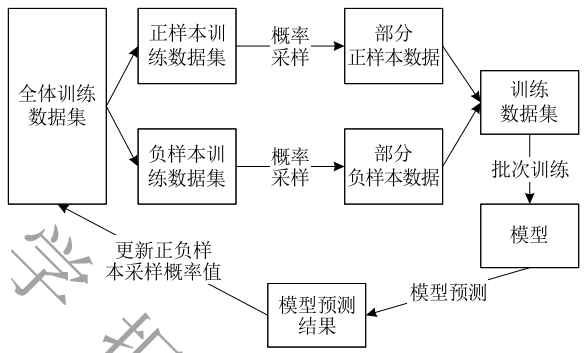


图 6 融合动态采样的疾病预测模型

算法 1. 结合迁移学习和动态采样的小样本疾病预测深度学习模型训练算法.

输入:多标签疾病数据集 $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$,$x_i\in X\subseteq R^n$,$y_i\in\{c_1,c_2,\cdots,c_K\}$,$K$ 为疾病标签总个数;待训练疾病标签 c_i;迭代次数 T;每次迭代训练数据块的大小 num

输出:c_i 预测模型 $F_i(x)$

(1) 源模型训练

(1.a) 对于任意标签 c_j,计算疾病标签 c_i 与 c_j 的共现频次,如式(4)和式(5)所示.

$$P(c_i,c_j)=\sum_{(x_n,y_n)\in D} I(\{c_i,c_j\}\subseteq y_n) \tag{4}$$

$$I(\{c_i,c_j\}\subseteq y_n)=\begin{cases}1, & c_i\in y_n \text{ and } c_j\in y_n\\ 0, & c_i\notin y_n \text{ or } c_j\notin y_n\end{cases} \tag{5}$$

(1.b) 利用 One-Vs-Rest 方式将 D 拆分为多个疾病的二分类数据集$\{D_1,D_2,\cdots,D_K\}$,其中 D_i 为疾病标签 c_i 的训练集合,选择与疾病标签 c_i 共现频次最多的多数疾病标签 c_k,在 c_k 的训练数据集 D_k 上训练疾病预测模型 $F_k(x)$(见式(6)),并将 $F_k(x)$ 参数进行保存.

$$F_k(x)=Train(D_k) \tag{6}$$

(2) 小样本疾病训练阶段

(2.a) 将深度学习模型 $F_k(x)$ 参数读取，作为疾病标签 c_i 的初始化模型 $F_{i,1}(x)$，c_i 的多数类别样本集为 $D_{k,neg}$，少数类别样本集为 $D_{k,pos}$，样本量分别是 N_{neg} 和 N_{pos}，总样本量为 N，初始化少数类样本采样概率 $\beta_{i,1}=\{\beta_{i,1}(1),\beta_{i,1}(2),\cdots,\beta_{i,1}(N)\}$，如式(7)所示. 由于正、负样本的采样概率之和均为 $\frac{num}{2}$，故针对每个正、负样本按照以下采样方法(2.b)采样后，采样得到的正、负样本数量的均值均为 $\frac{num}{2}$，故采样构建的样本是均衡的.

$$\beta_{i,1}(j)=\begin{cases}\dfrac{num}{2\times N_{pos}}, & c_i\in y_j\\ \dfrac{num}{2\times N_{neg}}, & c_i\notin y_j\end{cases} \tag{7}$$

(2.b) 基于样本抽样概率 $\beta_{i,t}$ 进行样本采样，分别对正样本集和负样本集进行采样. 对于任意样本 (x_j,y_j)，其抽样概率为 $\beta_{i,t}(j)$，利用 $Random(*)$ 随机产生一个 0 到 1 之间均匀分布的值 $Random(x_j)$，如果 $Random(x_j)<\beta_{i,t}(j)$，则将样本 (x_j,y_j) 加入到新的均衡样本集 $D_{i,train}$ 中，此时，若 $c_i\notin y_j$，则将样本 (x_j,y_j) 加入到部分多数类样本集 $D^{sel}_{i,neg}$ 中，若 $c_i\in y_j$，则将样本 (x_j,y_j) 加入到少数类样本集 $D_{i,pos}$ 中，如式(8)和式(9)所示. 针对每个样本 (x_j,y_j)，其采样概率是 $\beta_{i,t}(j)$，等于随机生成的数 $Random(x_j)$ 小于 $\beta_{i,t}(j)$ 的概率，故 $Random(x_j)$ 小于 $\beta_{i,t}(j)$ 时，将样本 (x_j,y_j) 加入该均衡的样本集中，故采用该算法更新采样概率是合理的.

$$D^{sel}_{i,neg}=\{(x_j,y_j)\,|\,Random(x_j)\le\beta_{i,t}(j),(x_j,y_j)\in D_{i,neg}\} \tag{8}$$

$$D^{sel}_{i,pos}=\{(x_j,y_j)\,|\,Random(x_j)\le\beta_{i,t}(j),(x_j,y_j)\in D_{j,pos}\} \tag{9}$$

最终将 $D^{sel}_{i,neg}$ 与 $D^{sel}_{i,pos}$ 组成训练集 $D_{i,train}$：

$$D_{i,train}=D^{sel}_{i,neg}\bigcup D^{sel}_{i,pos} \tag{10}$$

(2.c) $F_{i,t-1}(x)$ 基于集数据 $D_{i,train}$ 进行训练，生成新模型 $F_{i,t}(x)$，如式(11)所示；

$$F_{i,t}(x)=Train(F_{i,t-1}(x);D_{i,train}) \tag{11}$$

(2.d) 计算模型 $F_{i,t}(x)$ 在整体训练样本上的预测样本为正样本的概率结果是 $P_{i,t}$，$P_{i,t}(j)\in[0,1]$，表示分类器预测样本属于正样本的概率值. 对于正样本来说，$P_{i,t}(j)$ 越大越好，对于负样本来说，$P_{i,t}(j)$ 越小越好. 利用 $P_{i,t}$ 来更新抽样概率 $\beta_{i,t+1}=\{\beta_{i,t+1}(1),\beta_{i,t+1}(2),\cdots,\beta_{i,t+1}(N)\}$，如式(12)所示.

$$\beta_{i,t+1}(j)=\begin{cases}\beta_{i,t}(j)\exp(1-P_{i,t}(j)), & c_i\in y_j\\ \beta_{i,t}(j)\exp(P_{i,t}(j)), & c_i\notin y_j\end{cases} \tag{12}$$

当模型 $F_{i,t}(x)$ 预测训练样本错误或预测正确但预测置信度不高时，增加该样本的采样概率，从而提高模型对该样本的关注度；当模型预测样本正确且置信度高时，相对降低该样本的采样概率，从而减小模型对该样本的关注度. 从而增加模型对正负样本的可区分性，提高模型的预测准确度和置信度. 故当样本 (x_j,y_j) 为正样本时，$P_{i,t}(j)$ 越接近 0，分类错误或分类正确但置信度不高，其更新后的采样概率提高；当为负样本时，$P_{i,t}(j)$ 越接近 1，分类错误或分类正确但置信度不高时，其更新后的采样概率提高.

对正样本采样概率进行正则化，其中 $Z_{t,pos}$ 为所有正样本采样概率之和，分别如式(13)和式(14)所示.

$$\beta_{i,t+1}(j)=\frac{num\times\beta_{i,t+1}(j)}{2\times Z_{t,pos}} \tag{13}$$

$$Z_{t,pos}=\sum_{(x_n,y_n)\in D_{i,pos}}\beta_{i,t+1}(n) \tag{14}$$

对负样本采样概率进行正则化，其中 $Z_{t,neg}$ 为所有负样本采样概率之和，分别如式(15)和式(16)所示.

$$\beta_{i,t+1}(j)=\frac{num\times\beta_{i,t+1}(j)}{2\times Z_{t,neg}} \tag{15}$$

$$Z_{t,neg}=\sum_{(x_n,y_n)\in D_{i,neg}}\beta_{i,t+1}(n) \tag{16}$$

(2.e) 判断是否达到指定的迭代次数，如果满足则返回最终分类器；否则，利用新的采样概率继续进行步骤(2.b)～(2.e).

4　实验与分析

本节首先说明了我们使用的疾病数据集、公开数据集和采用的机器学习模型，然后阐述了实验的设置和评估方法. 我们在不同疾病类别的医院门诊病历集(见表 1 和表 2)上做了传统机器学习疾病预测方法[9-11]与我们提出的 4 种改进方法(CNN＋TL＋DS、CNN＋DS、CNN＋TL、Single CNN)的疾病预测效果对比实验.

4.1　实验数据集

本文采用的数据集是我们论文合作医生(王晓青大夫)的 1990 份医院门诊病历私有数据集. 我们从门诊病历中抽取主诉、现病史、体格检查等作为样本文本输入，初步诊断结果作为样本标签结果，按照初步诊断结果进行不同疾病数据集的划分，获得各个疾病二分类数据集. 为了进一步泛化验证我们所提新模型在通用文本类数据集上的有效性，我们选用了搜狗实验室情感分类数据集中的酒店评论、笔记本评论、图书评论等三种公开的不均衡二分类数据集.

根据病例数量多寡，本文选择呼吸道感染和支气管炎作为多数类疾病实验数据集；选择支气管肺炎、胃肠功能紊乱和细菌性感染作为少数类疾病实验数据集. 本文采用不均衡率(Imbalance Ratio, IR)[47]来表征数据集的不均衡程度，IR 是数据集中

多数类样本与少数类样本的数量比值.表 1 给出了数据集的具体情况.

表 1 实验数据集基本情况表

	疾病名称	正例数	负例数	IR
公开数据集	图书评论	500	1500	3
	酒店评论	500	1500	3
	笔记本评论	500	1500	3
私有数据集(多数类)	呼吸道感染	1210	750	1.61
	发热	631	1359	2.15
	支气管炎	433	1527	3.53
私有数据集(少数类)	支气管肺炎	131	1829	13.96
	胃肠功能紊乱	113	1847	16.35
	细菌性感染	66	1894	28.70

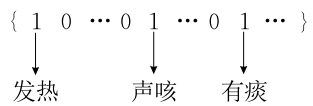
表 2 标签共现频次信息

疾病名称 A	疾病名称 B	共现频次
支气管炎	呼吸道感染	163
支气管肺炎	支气管炎	56
支气管肺炎	呼吸道感染	39
发热	呼吸道感染	456
发热	支气管炎	92
发热	胃肠功能紊乱	30
胃肠功能紊乱	呼吸道感染	40
胃肠功能紊乱	发热	30
胃肠功能紊乱	支气管炎	9
细菌性感染	呼吸道感染	33
细菌性感染	发热	26
细菌性感染	支气管炎	12
细菌性感染	胃肠功能紊乱	2

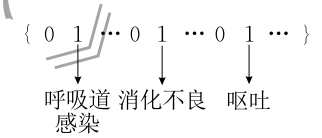
在迁移学习中为了选取更好的预训练模型,本文统计了不同疾病标签共现频次,对于标签组合(c_i,c_j),采用算法 1 步骤(1.a)统计标签共现频次,得到了表 2 所示的标签共现频次信息.Single CNN 首先将输入文本分词后得到词序列,利用 skip-gram 模型训练得到的词语高维连续稠密词向量,进而将词序列转化为词向量矩阵,从而得到输入文本的语义特征表示,在卷积、池化、全连接后进行疾病的分类预测.本文用于对比的基于传统机器学习的模型有:逻辑回归(Logistic Regression,LR)、支持向量机(Support Vector Machine,SVM)、决策树(Decision Tree,DT)、朴素贝叶斯模型(Naive Bayes,NB)、迭代提升欠采样模型(Under Sampling with Iteratively Boosting,USIB)[12].

USIB 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,USIB 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,USIB 采用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.这些模型需要将门诊病历进行抽取转化为特征集合,因此本文将门诊病历中涉及到的所有医学特征进行组合,构建医学特征集,记为$\{S_1,S_2,\cdots,S_n\}$,共有 n 个医学特征;将所有初步诊断中的疾病进行合并,构造医学疾病标签集,记为$\{c_1,c_2,\cdots,c_K\}$,共有 K 个医学标签.本文采用位编码方式将每一份医学病历表示为 n 维的特征向量和 K 维标签向量的组合.如对于某一病历,病历特征集合表示为{“发热、声咳、有痰,…”},病历标签集合表示为{“呼吸道感染、消化不良、呕吐”},其特征向量和标签向量可以如下表示:

特征向量 $\boldsymbol{X}$:其中第 i 维表示特征 S_i 是否包含在这一病历中,若病历特征集合包含特征 S_i,则 x_i 为 1,否则为 0.



标签向量 $\boldsymbol{Y}$:其中第 i 维的值表示门诊病历初步诊断结果中是否包含此疾病 c_i,若疾病标签集合包含疾病标签 c_i,则 y_i 为 1,否则为 0.



4.2 实验设置及评估

本文首先在呼吸道感染和支气管炎二分类数据集上分别训练疾病预测模型.在训练支气管肺炎模型时,将支气管炎模型参数作为初始模型,训练疾病二分类器;呼吸道感染模型作为细菌性感染预训练模型;呼吸道感染模型作为急性扁桃体炎预训练模型.本文采用 *Precision*(准确率),*Recall*(召回率)和 $F1$ 值三个指标评价模型结果[48].定义如下:设 D 是一个某疾病的二标签数据集,其中包含了 $|D|$ 个样本$(\boldsymbol{x}_i,\boldsymbol{Y}_i)$,$i=1,\cdots,|D|$,$\boldsymbol{Y}_i\in L$,$L=\{0,1\}$是标签集合.设 H 是一个二标签分类器,而 $Z_i=H(\boldsymbol{x}_i)$是由 H 对样本 $\boldsymbol{x}_i$ 预测的标签.*Precision*、*Recall* 和 $F1$ 值分别由式(17)、(18)和式(19)计算得出.

$$Precision(H,D)=\frac{\sum_{i=1}^{|D|}|\boldsymbol{Y}_i\times Z_i|}{\sum_{i=1}^{|D|}|Z_i|} \tag{17}$$

$$Recall(H,D)=\frac{\sum_{i=1}^{|D|}|\boldsymbol{Y}_i\times Z_i|}{\sum_{i=1}^{|D|}|\boldsymbol{Y}_i|} \tag{18}$$

$$F1(H,D)=\frac{2\times Precision(H,D)\times Recall(H,D)}{Precision(H,D)+Recall(H,D)} \tag{19}$$

4.3　实验结果与分析

本文实验中用于疾病预测的卷积神经网络模型采用了 50 维度的词向量，多个不同卷积大小的卷积核，包括 2、3、4、6、8、10 大小的卷积核；全连接层采用 0.75 大小的 dropout rate；二范式的正则化比值为 3；批量梯度下降算法的批量为 64 个样本.

通过设计不同卷积核大小、卷积核数量等超参数结构的卷积神经网络分别进行实验，我们发现最终选取以上超参数时，疾病预测的性能最好，故本文中采用以上结构的卷积神经网络进行卷积神经网络疾病预测模型的对比实验、融合动态采样的效果提升实验和迁移学习的效果提升研究实验. 我们提出的 4 种模型在准确率、召回率和 $F1$ 值方面的实验结果分别如表 3、表 4 和表 5 所示.

表 3　不同疾病数据集上疾病预测模型的准确率

	呼吸道感染	支气管炎	支气管肺炎	胃肠功能紊乱	细菌性感染
LR	0.7321	0.7705	0.8194	0.7035	/
NB	0.6982	0.4915	0.6328	0.3136	0.0875
DT	0.7834	0.8356	0.8242	0.7199	0.7399
SVM	0.7261	0.6716	/	/	/
USIB	0.8373	0.8527	0.8264	0.6631	0.6427
Single CNN	0.7512	0.657	0.5882	0.4286	0.628
CNN+DS	**0.8115**	**0.7167**	**0.6428**	**0.5429**	**0.6992**
CNN+TL	/	/	/	**0.6154**	**0.6317**
CNN+TL+DS	/	/	/	**0.7167**	**0.7454**

表 4　不同疾病数据集上疾病预测模型的召回率

	呼吸道感染	支气管炎	支气管肺炎	胃肠功能紊乱	细菌性感染
LR	0.8493	0.4660	0.3191	0.1203	/
NB	**0.9306**	0.6066	0.3511	0.6354	0.0788
DT	0.7495	0.5637	0.5237	0.5026	0.3515
SVM	0.8375	0.3610	/	/	/
USIB	0.8552	0.7668	0.7027	0.7948	0.8024
Single CNN	**0.8267**	**0.9441**	**0.9524**	**0.9231**	**0.6465**
CNN+DS	**0.8440**	**0.9461**	**0.9575**	**0.9265**	**0.6473**
CNN+TL	/	/	/	**0.9271**	**0.8571**
CNN+TL+DS	/	/	/	**0.9293**	**0.9286**

表 5　不同疾病数据集上疾病预测模型的 $F1$ 值

	呼吸道感染	支气管炎	支气管肺炎	胃肠功能紊乱	细菌性感染
LR	0.7848	0.5807	0.4575	0.1977	/
NB	0.7976	0.5427	0.4511	0.4198	0.0771
DT	0.7656	0.6679	0.6387	0.5917	0.4758
SVM	0.7716	0.4694	/	/	/
USIB	0.8439	0.8024	0.7435	0.7104	0.6362
Single CNN	0.7871	0.7389	0.6961	0.5637	0.6371
CNN+DS	**0.8509**	**0.8118**	**0.7593**	**0.6791**	**0.7782**
CNN+TL	/	/	/	**0.7398**	**0.7273**
CNN+TL+DS	/	/	/	**0.8093**	**0.8189**

表 3、表 4 和表 5 表明所提出的 4 种模型在门诊病历数据集上分别优于对比模型，其中在预测效果最好的多数类疾病呼吸道感染上 $F1=0.8509$，在预测效果最好的少数类疾病细菌性感染疾病上 $F1=0.8189$. 其次，对比分析多数类疾病类别和少数类疾病类别的实验结果可知：在少数类疾病类别上，对比的传统机器学习模型的预测效果不理想，而本文 CNN+TL+DS 目标模型取得更好的预测准确率、召回率和 $F1$ 值，例如在胃肠功能紊乱和细菌性感染上因为同时采用动态采样和迁移学习技术分别取得了 0.8093 和 0.8189 的高 $F1$ 值，从而验证了本文所提目标模型的有效性. 再者，在疾病的辅助预诊中，高召回率具有更小的辅助诊断风险，对疾病辅助决策具有更加重要的意义. 本文所有 CNN 组合模型在召回率上均取得了显著提升，尤其是结合动态采样技术和迁移学习技术可对召回率进一步地提升，例如我们模型的召回率在胃肠功能紊乱上较传统方法提升达 18%～672.5%.

实验也表明：只采用 Single CNN 模型在整体性能上比传统机器学习模型取得相当或甚至更好的效果，在准确率上不及我们之前提出的 USIB 模型，但大幅提升了召回率；同时，疾病预测效果在小样本疾病类别上也有所提升，虽然 Single CNN 模型准确率低于决策树模型，但是决策树模型的召回率明显低于 Single CNN 模型，整体 $F1$ 值也得到了提升. 分析可知，卷积神经网络是一种重要且有效的疾病预测技术. 分析对比 Single CNN 模型和采用动态采样后的 CNN+DS 模型，采用动态采样技术后，疾病预测的准确率、召回率和 $F1$ 值均得到了有效的提升，这说明本文提出的动态采样技术有效缓解了训练样本不均衡的影响. 同时采用动态采样技术和迁移学习技术的 CNN+TL+DS 模型展现了更好的疾病预测效果，证明本文提出的迁移学习技术和动态采样技术显著提升疾病预测效果.

另一方面,病例中疾病的共现频次体现了疾病间的知识共享程度,本文针对大样本疾病中的“发热”的疾病预测模型,分别从高共现频次“呼吸道感染”、中共现频次“支气管炎”和低共现频次“胃肠功能紊乱”的疾病预测模型中迁移知识,针对小样本疾病中的“胃肠功能紊乱”和“细菌性感染”的疾病预测模型,分别从高共现频次“呼吸道感染”、中共现频次“发热”和低共现频次“支气管炎”的疾病预测模型中迁移知识,研究迁移学习对基于卷积神经网络疾病预测模型的作用以及共现频次与迁移效果的关系.实验结果中准确率、召回率和 $F1$ 值分别如表 6、表 7 和表 8 所示.

表 6 CNN 模型在不同迁移源下的准确率

	发热	胃肠功能紊乱	细菌性感染
呼吸道感染(高)	0.8113	0.6154	0.6317
发热(中)	/	0.5714	0.5217
支气管炎(中)	0.8104	0.6364	0.5455
胃肠功能紊乱(低)	0.8232	/	/
无迁移	**0.8279**	**0.6667**	**0.6454**

表 7 CNN 模型在不同迁移源下的召回率

	发热	胃肠功能紊乱	细菌性感染
呼吸道感染(高)	0.9070	0.9271	0.8571
发热(中)	/	0.9231	0.8487
支气管炎(中)	0.8839	0.9304	0.8571
胃肠功能紊乱(低)	0.8715	/	/
无迁移	**0.8413**	**0.8231**	**0.6465**

表 8 CNN 模型在不同迁移源下的 $F1$ 值

	发热	胃肠功能紊乱	细菌性感染
呼吸道感染(高)	0.8565	0.7398	0.7273
发热(中)	/	0.7059	0.6462
支气管炎(中)	0.8456	0.7558	0.6667
胃肠功能紊乱(低)	0.8467	/	/
无迁移	**0.8676**	**0.7764**	**0.7615**

表 6、表 7 和表 8 对比了从高、中、低共现频次疾病模型进行迁移学习后的结果,与不进行迁移的 Single CNN 模型相比,采用迁移学习技术后,CNN+TL 模型召回率得到了提升(尤其是提升了少数类疾病的召回率),迁移的效果与迁移源疾病和迁移目标疾病的共现频次非正相关关系.我们在后续研究中将继续探索迁移学习的影响因素,从而进一步提升疾病的预测效果.

此外,为了进一步评估本文提出的新模型以及动态采样的效果,我们在 3 种公开的不均衡二分类搜狗数据集上分别训练 LR、NB、DT、SVM 和 USIB 等对比模型和文本提出的新模型,实验结果如表 9、表 10 和表 11 所示.

表 9 公开数据集上各模型的准确率

	图书评论	酒店评论	笔记本评论
LR	0.7672	0.6192	0.7957
NB	0.6405	0.7761	0.6468
DT	0.6979	0.6968	0.6945
SVM	0.6679	0.5259	0.7289
Single CNN	**0.6827**	**0.7481**	**0.7549**
CNN+DS	**0.7886**	**0.8214**	**0.7977**

表 10 公开数据集上各模型的召回率

	图书评论	酒店评论	笔记本评论
LR	0.7952	0.7715	0.6832
NB	0.7435	0.7325	0.8400
DT	0.7629	0.5426	0.7280
SVM	0.5565	1.0000	0.2640
Single CNN	**0.8024**	**0.7523**	**0.7977**
CNN+DS	**0.8586**	**0.7977**	**0.8321**

表 11 公开数据集上各模型的 $F1$ 值

	图书评论	酒店评论	笔记本评论
LR	0.7809	0.6870	0.7352
NB	0.6882	0.7537	0.7308
DT	0.7290	0.6101	0.7109
SVM	0.6071	0.6893	0.3876
Single CNN	**0.7377**	**0.7502**	**0.7757**
CNN+DS	**0.8221**	**0.8094**	**0.8145**

表 9、表 10 和表 11 表明,本文的融合动态采样的 CNN+DS 模型在公开不均衡数据集上在准确率、召回率和 $F1$ 值方面取得了更好的效果,进一步证明了融合动态采样技术的卷积神经网络技术在基于不均衡、小样本数据集任务上的有效性,是一种重要的面向不均衡数据集提升模型召回率以及整体效果的有效技术.

进一步分析本文提出的新模型的性能提升,本文模型在准确率、召回率和 $F1$ 值上优于其他模型,主要原因在于以下几方面:

(1)本文采用将整体文本信息作为模型的输入,保留了原始的文本信息,信息丢失更少,同时利用 skip-gram 模型训练分布式词向量,保留了不同词之间的语义相关性.

(2)传统机器学习的疾病预测模型将门诊病历转化为医学特征集合时丢失了重要信息,且预测效果对病历特征处理的依赖大,而基于 CNN 组合疾病预测模型是端到端的学习,通过词语的分布式语义词向量表示、利用卷积核自动学习提取门诊病历中与疾病预测相关的语义特征来实现疾病的预测,获得了更好的预测效果.

(3)利用卷积神经网络抽取病历语义信息,可以获取不同症状之间的关系,抽取了更高层次的语义信息,同时采用多个卷积核进行语义提取,更多地

保留了门诊病历中不同层次的信息.

(4)采用动态采样技术,根据样本的预测结果动态调整样本的采样概率,使模型更多地关注、学习预测困难的样本数据;其次,正样本和负样本数据上的分开采样,保证了每批次训练数据的均衡性,提升了样本的召回率.即动态采样技术能够保证模型训练数据的均衡性,从而避免模型对样本更多的负样本的过学习和预测偏向.

(5)采用迁移学习技术将多数类疾病模型作为预训练模型,在此基础上训练少数类疾病模型,实现了不同疾病模型间的知识共享和迁移,在一定程度上避免了样本不足情况下无法获取足够信息的缺点.此外,迁移学习提升少数类疾病模型的效果,原因在于不同疾病间存在共享知识,迁移学习能够将多数类疾病模型中学习到的共享知识迁移到少数类疾病模型中.

5　总　结

本文提出了一种融合动态采样和迁移学习的卷积神经网络疾病预测模型,用在面向小样本、不均衡数据集的医学病历分类预测任务中.同时进行基于传统机器学习的疾病预测模型以及 USIB 疾病预测模型实验的对比,本文疾病预测模型在准确率、召回率和 $F1$ 值上均得到了提升,验证了本文提出的基于卷积神经网络的动态采样和迁移技术的有效性.在医学病历分类预测任务中,一方面,因为疾病的发病率不同和稀有疾病发病率低导致疾病病历具有小样本、不均衡的特点,这极大限制了疾病预测模型的性能提升,另一方面,具有高的召回率的疾病预测模型具有更小的决策风险和更为重要的实际意义.本文提出的融合动态采样和迁移学习技术的疾病预测模型,直接从病历文本利用卷积神经网络自动抽取病历的语义的、高层次的、不同层面的信息,采用动态采样方法以构造均衡数据集,利用模型的预测结果动态更新样本采样概率,更关注易错样本,从而提高模型效果,避免数据不均衡的影响,同时采用迁移学习技术实现从多数类疾病到少数类疾病模型的知识迁移,在一定程度上避免了样本和信息不足的缺点.实验结果表明,本文提出的融合动态采样和迁移学习的疾病预测模型,能够取得最好的疾病预测效果,还可以在保证预测准确性的情况下提升模型的召回率,尤其是能够提升少数类疾病的召回率,具有更小的决策风险和更重要的意义.

此外,从实验结果可以看出,本文模型虽然取得了很好的召回率,但是疾病预测的准确率有待进一步提升,后续工作将对网络模型和训练算法进行深度优化,以达到更好的疾病预测效果.此外,后续也将进一步研究基于专业文献医学知识图谱的新启发方法,利用知识图谱中的不同诊断路径来模拟医学诊疗思维,以获得具有可解释性、高准确率的疾病预测模型.

致　谢　感谢《宁夏自治区重点研发计划(引才专项)》(课题编号:2018BEB04002)、《智能化数据中心管理、编程规范与应用生态》(课题编号:2016YFB-1000505)、《基于天河二号的生物医学健康大数据应用支撑平台》(项目基金号:U1611261)、中央级公益性科研院所基本科研业务费专项资金资助课题《中医药大数据与疗效评价技术研究》(课题编号 ZZ10-005)对本工作的支持.

参　考　文　献

[1] Pan S J, Yang Q. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345-1359

[2] Weiss K, Khoshgoftaar T M, Wang D D. A survey of transfer learning. Journal of Big Data, 2016, 3(1): 9

[3] Zhuang Fu-Zhen, Luo Ping, He Qing, Shi Zhong-Zhi. Survey on transfer learning research. Journal of Software, 2015, 26(1): 26-39(in Chinese)
(庄福振,罗平,何清,史忠植.迁移学习研究进展.软件学报,2015,26(1):26-39)

[4] Killian T, Daulton S, Konidaris G, et al. Robust and efficient transfer learning with hidden-parameter Markov decision processes//Proceedings of the Advances in Neural Information Processing Systems. California, USA, 2017: 6250-6261

[5] Shortliffe E H. Computer-based medical consultations: MYCIN. Elsevier, 1976, 85(6): 243-260

[6] Wang Jia-Kuan, Yu Li-Ping, Qiao Chuang. Development of expert diagnostic system for cervical and lumbar diseases. Acta Academiae Medicinae Xuzhou, 1998, 18(1): 51-53(in Chinese)
(王加宽,俞立平,乔闯.颈腰疾病专家诊断系统的研制.徐州医学院学报,1998,18(1):51-53)

[7] Kohn M S, Sun J, Knoop S, et al. IBM's health analytics and clinical decision support. Yearbook of Medical Informatics, 2014, 9(1): 154-62

[8] Chattopadhyay S, Davis R M, Menezes D D, et al. Application of Bayesian classifier for the diagnosis of dental pain. Journal of Medical Systems, 2012, 36(3): 1425-1439

[9] Patil A P, Bhosale A P, Ambre G. Intelligent heart disease prediction system using naive Bayes classifier. International Journal of Advanced and Innovative Research, 2013, 2(4): 1061-1066

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

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

[12] Chen Xu, Liu Peng-He, Sun Yu-Zhong, et al. Research on disease prediction models based on imbalanced medical data sets. Chinese Journal of Computers, 2019, 42(3): 596-609 (in Chinese)
(陈旭, 刘鹏鹤, 孙毓忠等. 面向不均衡医学数据集的疾病预测模型研究. 计算机学报, 2019, 42(3): 596-609)

[13] 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

[14] 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

[15] 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

[16] Lipton Z C, Kale D C, Elkan C, Wetzell R. Learning to diagnose with LSTM recurrent neural networks//Proceedings of the International Conference on Learning Representations (ICLR). San Juan, Puerto Rico, 2016: 1-18

[17] Cai Hang. Medical diagnosis expert system based on neural network. Journal of Mathematical Medicine, 2002, 15(4): 294-295(in Chinese)
(蔡航. 基于神经网络的医疗诊断专家系统. 数理医药学杂志, 2002, 15(4): 294-295)

[18] Hou Gui-Ying, Sun Bai-Qing, Guan Zhen-Zhong, et al. Research on diagnosis of hypertension by intelligent medical diagnostic system. Journal of Harbin Medical University, 2003, 37(3): 223-225(in Chinese)
(侯桂英, 孙佰清, 关振中等. 应用智能医疗诊断系统诊断高血压的研究. 哈尔滨医科大学学报, 2003, 37(3): 223-225)

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

[20] Branco P, Torgo L, Ribeiro R P. A survey of predictive modeling on imbalanced domains. ACM Computing Surveys, 2016, 49(2): 31

[21] Hulse J V, 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

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

[23] 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

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

[25] 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

[26] 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

[27] Xiong Bingyan, Wang Guoyin, Deng Weibin. Under-sampling method based on sample weight for imbalanced data. Journal of Computer Research and Development, 2016, 53(11): 2613-2622

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

[29] 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

[30] Han H, Wang W Y, Mao B H. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning//Proceedings of the International Conference on Advances in Intelligent Computing. Berlin, Germany, 2005: 878-887

[31] Chen S, He H, Garcia E A. RAMOBoost: Ranked minority oversampling in boosting. IEEE Transactions on Neural Networks, 2010, 21(10): 1624-42

[32] Rodríguez 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

[33] 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

[34] 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

[35] 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

[36] 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

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

[38] Dai W, Xue G R, Yang Q, et al. Co-clustering based classification for out-of-domain documents//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Jose, USA, 2007: 210-219

[39] Jiang J, Zhai C X. A two-stage approach to domain adaptation for statistical classifiers//Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management. Lisbon, Portugal, 2007: 401-410

[40] Pan S J, Kwok J T, Yang Q. Transfer learning via dimensionality reduction//Proceedings of the 23d AAAI Conference on Artificial Intelligence. Chicago, USA, 2008: 677-682

[41] Si S, Tao D, Geng B. Bregman divergence-based regularization for transfer subspace learning. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(7): 929-942

[42] Tian X, Tao D, Rui Y. Sparse transfer learning for interactive video search reranking. ACM Transactions on Multimedia Computing Communications and Applications, 2011, 8(3): 1-19

[43] Dai W, Yang Q, Xue G R, et al. Boosting for transfer learning //Proceedings of the 24th International Conference on Machine Learning. Corvallis, USA, 2007: 193-200

[44] Zhou B, Khosla A, Lapedriza A, et al. Object detectors emerge in deep scene CNNs//Proceedings of the International Conference on Learning Representations (ICLR). San Diego, USA, 2015: 1-12

[45] Wang J, Song Y, Leung T, et al. Learning fine-grained image similarity with deep ranking//Proceedings of the Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 1386-1393

[46] Babenko A, Lempitsky V. Aggregating deep convolutional features for image retrieval//Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile, 2015: 1269-1277

[47] Alcalá-Fdez 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 and Soft Computing, 2011, 17(2-3): 255-287

[48] Tsoumakas G, Katakis I, Taniar D. Multi-label classification: An overview. International Journal of Data Warehousing and Mining, 2008, 3(3): 1-13

HU Man-Man, M. S. candidate. Her current research interests include machine learning and data mining.

CHEN Xu, M. S. candidate. His current research interests include machine learning and data mining.

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

SHEN Xi, assistant economist. Her main research interests include Internet medical, wisdom medical.

WANG Xiao-Qing, bachelor, attending pediatrician. Her medical profession is respiratory tract, digestive tract, and newborn.

YU Tian-Yang, M. S. candidate. His current research interest is machine learning.

MEI Yu-Dong, Ph. D. candidate. His current research interests include machine learning, intelligent log analysis and data mining.

XIAO Li, Ph. D., associate professor. His main research interests include artificial intelligence, medical image, biological computation.

CHENG Wei, Ph. D., chief physician. Her current research interests include combination of traditional Chinese medicine and western medicine to prevent and cure senile diseases.

YANG Jie, Ph. D., chief physician. Her current research interests include big data of traditional Chinese medicine and diagnostics of traditional Chinese medicine.

YANG Yan, M. S., associate-director technician (senior engineer). Her current research interests include hospital informatization construction.

Background

Disease prediction based on medical records is a classic problem in clinical decision support systems (CDSS). Different incidences of diseases lead to the imbalance and small sample of diseases data, making it hard to train effective and accurate

disease prediction model. Sampling techniques are the common methods to solve the sample imbalance, but it independently train different disease prediction models, which not considers the knowledge transfer between different disease models, which limits the performance. Transfer learning can achieve knowledge transfer between similar tasks. In addition, Mainstream models need to transform outpatient records into medical symptom sets for training disease prediction models.

To address the above problems, this paper proposed a model based on transferring learning and dynamic sampling. It firstly trained disease models on majority diseases, and used the majority disease models as the initialization model of the minority diseases. It introduced the dynamic sampling to generate balanced training datasets for rare diseases, and dynamically updated the sample sampling probabilities by model prediction results, so that the model could learn more error samples to improve the performance of prediction models.

We evaluate the proposed method on a modern medical record sample set. The results showed that our method performed better than other mainstream models.

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 dataset. The authors have proposed a new under-sampling framework, which used 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 work is a significant extension and improvement over previous models.