基于多分类器融合的呼吸病药物推荐方法研究

NJU MIP Laboratory 刘洋

摘要

呼吸病种类繁杂且频发，在治疗过程中，药物辅助治疗极为关键，但由于耐药菌的出现，造成了开药难和治愈难的情况。为解决上述问题，本文研究利用江苏省中医院的诊疗记录及药敏试验数据，首先针对给定的呼吸病类型利用 LDA 主题模型进行聚类挖掘，之后通过主成分分析法及样本均衡化技术解决特征维度爆炸和样本失衡的问题，最后运用 stacking 集成学习框架，使用随机森林、XGBoost 和支持向量机三种模型作为基模型，逻辑回归模型作为元模型，通过计算机辅助挖掘其中蕴含的用药方案制定的规律和知识，实现了基于多分类器融合的呼吸病药物推荐模型，实验结果表明预测准确率达 86.1%，相较于随机森林、XGBoost、支持向量机和逻辑回归这四个单分类器药物推荐模型高出 1.4，1.7 和 7.9 和 13.1 个百分点。因此，该模型能够有效完成呼吸病的辅助治疗药物的智能推荐。

关键词：呼吸病、耐药菌、stacking 框架、多分类器融合、药物推荐模型

批注 [XF1]: 这句显得空，一般讲：针对 xxxx 数据集，或江苏省中医院的诊疗记录及药敏试验数据，

批注 [刘2R1]: 已更改

批注 [XF3]: 先基于 LDA 主题模型进行呼吸病聚类挖掘、独热向量编码、主成分分析法降维、**样本均衡化，这些工作也要体现在摘要里**

批注 [刘4R3]: 已更改，编码工作属于基本流程操作，可以不用写

批注 [XF5]: 框架

批注 [XF6]: 直接讲具体实验结果数据

批注 [XF7]: 框架

目录

1 引言

随着空气污染加剧和人们日常生活习惯的不断改变，呼吸病的发病几率大大提高[1]。同时，由于呼吸病种类繁杂，具体可分为哮喘病、气管炎、肺心病等，严重时可导致死亡，因此，药物辅助在整个治疗过程中显得极为关键[2]。但是，在药物治疗中，当细菌多次与药物接触后，细菌可能对治疗药物的敏感性减小甚至消失，从而致使药物对耐药菌的疗效降低甚至无效，因此，耐药菌的出现会导致呼吸病的开药难和治愈难的问题[3]。综上，科学用药对于提高呼吸病治疗效果具有重要作用，而目前的用药方案主要由医生根据自身经验和知识制定，这种传统方法受限于个人认知的局限性，用药方案的制定难免存在不足。

随着医疗信息化的不断提高，各类医院的医疗信息系统中记录了大量的历史病患诊疗数据，从这些医疗数据出发，研究可以通过计算机辅助挖掘出其中隐藏的用药方案制定的规律和知识，进而为医生制定用药方案提供科学合理的决策支持[4]。

当前，很多研究者已经从各类医疗数据出发，分析挖掘数据中存在的隐藏规律进而辅助药物推荐方法的设计。例如，刘文斌等研究者利用二维高斯分布方法筛选基因表达数据及蛋白质网络信息得出个性化网络标志物，提出基于个性化标志物的药物推荐方法，有助于实现个性化用药及推动精准医疗的发展[5]；李鹏飞等研究者则基于历史患者的用药数据对疾病的药物治疗过程建模进行挖掘，进而提出一种过程模型与用户体征数据相融合的药物推荐方法[4]；Arianna Dagliati 等研究者则对医嘱指令进行聚类，结合时序分析和相似度计算方法进行数据挖掘并提出治疗方案的推荐方法[6]；李睿易等研究者通过将医疗信息系统中的处方数据与 DrugBank 数据库中记录的药物功效融合生成患者的每日用药疗效文档，进而结合 LDA 主题模型和概率后缀树提出了药物治疗的临床路径模型，可用于辅助个性化临床路径的制定和药物的推荐[7]。

批注 [XF8]: 这一段要补充国外的研究现状

之前的研究工作针对临床数据出发，从多角度利用不同方法进行分析挖掘，进而设计相应的药物推荐方法。本文研究从前人的成果出发，结合呼吸病相关数据并利用创新算法设计了基于多分类器融合的呼吸病药物推荐方法，与之前相比具有如下特色：

1、基于 LDA 主题模型的呼吸病聚类挖掘。由于呼吸病种类复杂繁多，且部分呼吸病在成因、症状及治疗药物选择等方面具有极大相似性和关联性，因此，本文研究借助 LDA 主题模型对呼吸病类型进行聚类挖掘，进而完成药物推荐方法的设计。

2、基于多分类器融合的药物推荐分类器。由于单一算法训练的分类器效果不稳定，且不同分类器之间具有差异性，因此，本研究中利用 stacking 集成学习的思想将若干个单分类器进行组合，形成效果稳定的药物推荐强分类器。

3、基于多方面数据训练的药物推荐方法。本文主要解决当引发不同呼吸病的微生物对某些抗生素产生耐药性后的药物推荐问题，本着多角度充分考虑的原则，研究将呼吸病类型、微生物检测结果和药敏试验多方面数据结合进行药物推荐方法的设计。

2 研究路线

本文旨在解决面对不同种类的耐药菌引发的呼吸病的最佳药物推荐问题，而研究就如何设计出更稳定准确的呼吸病药物推荐模型提出了相应研究路线，如图 1 所示。

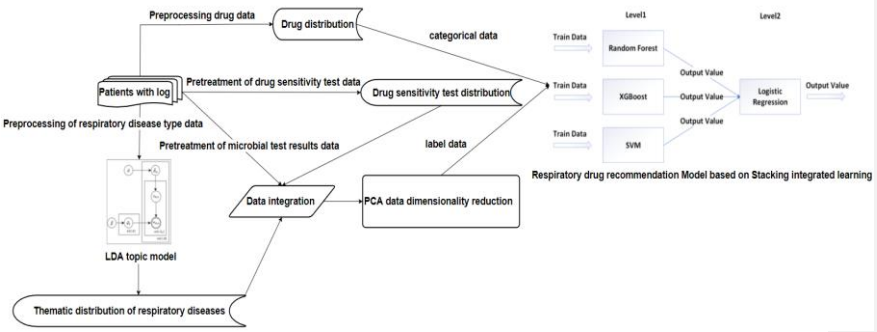

图 1 研究路线

结合图 1，医院中的医疗信息系统中记录了大量患者的患病记录日志和药敏试验数据，首先，研究针对种类繁杂的呼吸病种类进行了 LDA 主题聚类挖掘，得到了呼吸病种类的主题分布；之后，研究将上述聚类得到的呼吸病主题、微生物检测结果和药敏试验数据进行融合编码，通过主成分分析对融合指标进行提取降维操作；然后，研究利用上述降维处理完成的融合编码数据作为特征体系，同时对药物种类数据进行编码作为分类标签，进行随机森林、XGBoost 和支持向量机三种单分类器的训练；最后，研究基于 stacking 集成学习的思路对上述单分类器的输出结果进行特征融合，所得结果作为输入进行逻辑回归模型的训练，进而输出最终的药物推荐清单，至此，研究完成了基于多分类器融合的呼吸病药物推荐模型的设计实现。

3 基于 LDA 主题模型的呼吸病聚类挖掘

本文研究基于江苏省中医院的医疗信息系统中的实际数据进行实验，经过数据预处理，研究整理呼吸病类型共 889 种，种类的繁多导致后续模型的建立有一定难度。同时，由于呼

批注 [XF9]: 江苏省中

吸病在成因、症状及治疗药物选择等方面具有极大相似性和关联性，因此，研究依据呼吸病名称中隐含的发病部位、症状描述以及状态信息等针对种类繁杂的呼吸病进行 LDA 主题建模，最终得出相应聚类结果。

所谓 LDA 主题模型[8]，是 David M. Blei 等人提出的“文档—主题—词”的三层贝叶斯模型，该模型的假设在于每个文档是多类主题的混合分布，而主题则是词组成的概率分布，即一篇文档中蕴藏的主题按照多项式分布生成，而每个主题中的词按照多项式分布生成，具体生成原理如图 2，其中假设超参数为α的 Dirchlet 分布为文档—主题的先验分布，模型根据该分布取样生成文档 m 的主题多项式分布θ_m;然后，每次从θ_m中取样生成关于文档 m 的主题 n 的主题$z_{m,n}$；同时，模型的另一边假设存在一个超参数为β的 Dirchlet 分布为主题—词的先验分布，接下来从该分布中可以取样生成关于主题$z_{m,n}$的词分布φ_k，最终可通过词分布φ_k生成词语$\omega_{m,n}$，因此，当文档集合确定时，文档—主题—词的联合概率分布$p(\omega,z)$如公式（1）：

$$p(\omega,z|\alpha,\beta)=p(\omega|z,\beta)p(z|\alpha) \quad (1).$$

然后，LDA 主题模型则是采用 Gibbs 抽样可以推出文档 m 出现主题 k 和主题φ_k出现词语 t 的分布期望，分别为公式（2）和（3）：

$$\theta_{m,k}=\frac{n_m^{(k)}+\alpha_k}{\sum_{k-1}^{K}n_m^{(k)}+\alpha_k} \quad (2),$$

$$\varphi_{k,t}=\frac{n_k^{(t)}+\beta_t}{\sum_{t-1}^{V}n_k^{(t)}+\beta_t} \quad (3).$$

公式中$n_m^{(k)}$代表第 m 篇文档中第 k 个主题的个数，$n_k^{(t)}$代表第 k 个主题中第 t 个词的个数。

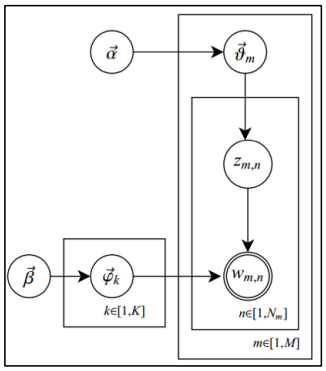

图 2 LDA 主题模型原理图

本文研究通过将整理好的呼吸病种类文档集合作为输入进行 LDA 主题模型的训练，通过调参优化，最终将主题数设置为 5 时效果最佳，其中呼吸病主题下概率值最大的前 5 个项目及其名称如表 1 所示。

表 1 呼吸病主题数为 5 时概率最大的前 5 个项目及其名称

主题	概率最大的前 5 个项目及其名称
1	输尿管、膀胱、泌尿道、前列腺、结石
2	急性、状态、肺炎、白血病、功能
3	肿物、贫血、创伤性、下肢、肿瘤
4	糖尿病、高血压、肾结石、积水、泌尿道
5	胆管、骨折、综合征、结石、占位性

4 基于多分类器融合的呼吸病药物推荐模型

4.1 实验数据准备

本文研究数据来自于江苏省中医院的医疗信息系统，其中包含了大量患者的个人基本信息、包括临床诊断结果、患病类型和微生物检测结果等在内的诊疗记录以及关于 Amikacin（AMK）、Amoxicillin（AMC）和 Moxifloxacin（MFX）等 50 种抗生素药物的药敏试验数据等。

批注 [XF10]: 更具体讲，列举 5-8 种诊疗记录及药敏试验主要数据种类

批注 [刘11R10]: 诊疗记录模

文章第三节内容已经对 889 种呼吸病类型进行了处理，聚类得到 5 类呼吸病主题。之后，研究通过对数据进行缺失值处理、离群点处理、噪声处理和重复值清洗等工作，整理结果共包括 1918 条药敏试验类数据，内含 20 种微生物检测结果和 39 类抗生素的药敏试验结果；5625 条患者的用药记录，共涉及到 18 类药物。

4.2 特征构建与样本选择

为验证研究路线方案的有效性，本文模型需利用样本数据进行实验验证，但在特征构建和样本选择的过程中，数据本身出现的特征维度爆炸和样本不平衡的问题极容易影响模型的训练效果，本小节就上述问题提供了解决方案。

4.2.1 数据降维

本文模型的原始指标体系包括 LDA 处理得到的 5 种患病类型、20 种微生物鉴定结果和针对 39 种微生物的药敏试验结果（包括敏感、中介、耐药和不确定四类取值），研究首先对指标体系数据进行独热向量编码工作，但由于指标本身规模和取值范围均较大，直接导致了编码后的特征维度爆炸的问题。数据降维工作可以有效解决上述问题，通过将高维度的数据保留下最重要的一些特征，去除噪声和不重要的特征，从而实现提升数据处理速度和防止模型过拟合的目的。

批注 [XF12]: 不是实验吧？

批注 [刘13R12]: 已更改为研究，实验用词不够严谨

数据降维工作采用算法为主成分分析法[9]，PCA 的主要思想是将 n 维特征映射到 k 维上，这 k 维是全新的正交特征也被称为主成分，是在原有 n 维特征的基础上重新构造出来的 k 维特征，本质上来讲，PCA 算法相当于只保留包含绝大部分方差的维度特征，而忽略包含方差几乎为 0 的特征维度，实现对数据特征的降维处理。最终，研究以 90%的方差保留比例将指标特征降维到 28 种指标，结果如图 5 所示。

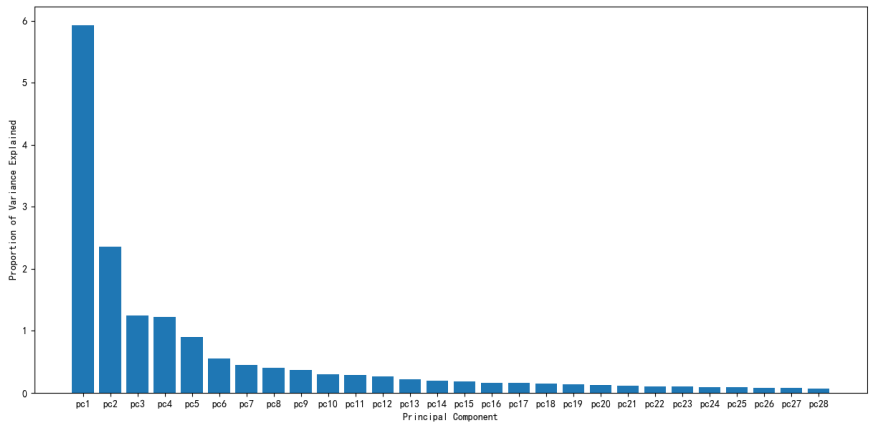

图 5 PCA 降维结果

4.2.2 样本均衡化

在机器学习的分类任务中，保证实验数据的样本均衡极为重要。在本实验中，某些药物的使用记录存在着单一类别样本大比例存在的现象，而由于实验数据的样本不平衡问题，导致了训练后的药物推荐分类器出现分类不准确的问题。为解决上述问题，本研究结合过采样和欠采样两种方法的优点，实验通过采用 SMOTE 算法[10]针对少数类进行过采样操作，配合 Tomek Links 算法[11]剔除近邻点实现欠采样，最终实现了 18 类药物类别的样本数量均衡化，结果如表 3 所示，有利于后续呼吸病药物推荐模型的训练。

表 2 样本均衡结果

药物名称	正样本	负样本	药物名称	正样本	负样本
M1:β-内酰胺酶抑制剂	972	972	M2:三唑类抗真菌药	1689	1689
M3:喹诺酮类	1350	1350	M4:噁唑烷酮类	1738	1738
M5:四环素类	1807	1807	M6:大环内酯类	1843	1843
M7:头孢菌素	1003	1003	M8:头霉素类	1852	1852
M9:恶唑酮类	1844	1844	M10:棘白菌素类抗真菌药	1763	1763
M11:氧头孢类	1858	1858	M12:氨基糖苷类	1791	1791
M13:甘氨酰四环素类	1780	1780	M14:硝基咪唑类	1828	1828
M15:碳青霉烯类	1149	1149	M16:磺胺类	1835	1835
M17:糖肽类	1715	1715	M18:青霉素类	852	852

4.3 模型架构设计

本文模型为多分类器融合的呼吸病药物推荐模型，其主要利用 Stacking 方法进行构建，该方法是一种分层模型集成框架[12]。具体思路如图 3 所示，首先将数据集分成训练集和测试集，利用训练集训练得到多个基模型，然后利用基模型对测试集进行预测，并将输出值作为下一阶段训练的输入值，最终的标签作为输出值，用于训练元模型。由于该集成学习算法可以兼顾多个基模型和元模型的学习能力，同时基模型和元模型所使用的训练数据不同，因此可以很大程度地提高模型训练的准确性。

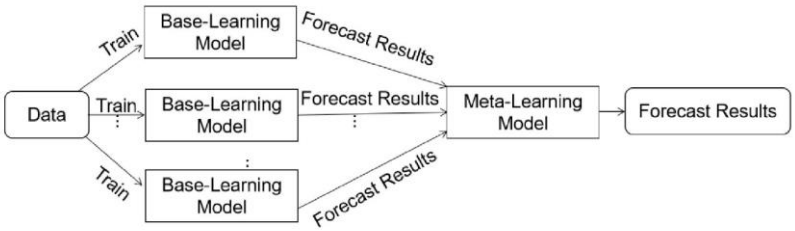

图 3 Stacking 集成学习算法

在 Stacking 集成学习框架的基础上，本文选择随机森林模型、XGBoost 模型和支持向量机模型作为第一层的基模型。其中，随机森林模型[13]和 XGBoost 模型[14]均利用决策树模型作为弱分类器，分别通过 Bagging 和 Boosting 两种集成学习方法进行强分类器的训练，使

得预测模型具有更好的预测效果和更强的泛化能力，支持向量机模型[15]是一种寻找特征空间上的间隔最大的超平面的线性分类器，除分类性能优于大多数弱分类器的优点外，尤其在样本数据高维度而导致线性不可分的情况下，支持向量机模型可以通过核函数进行维度映射实现线性可分，而本文核函数采用高斯核函数。为纠正第一层各基模型的预测结果偏差，同时防止过拟合情况的出现，本文选择逻辑回归模型[16]作为第二层的元模型，最终，该模型能够有效实现离散型预测结果的输出。

依托以上研究现状，本文设计了如图 4 所示的算法架构图，即第一层的基模型分别采用 7 折交叉验证的方法对随机森林、XGBoost 和支持向量机三种模型进行训练，之后通过将特征融合的结果作为第二层元模型的输入，采用逻辑回归算法完成多分类器融合呼吸病药物推荐模型的训练。

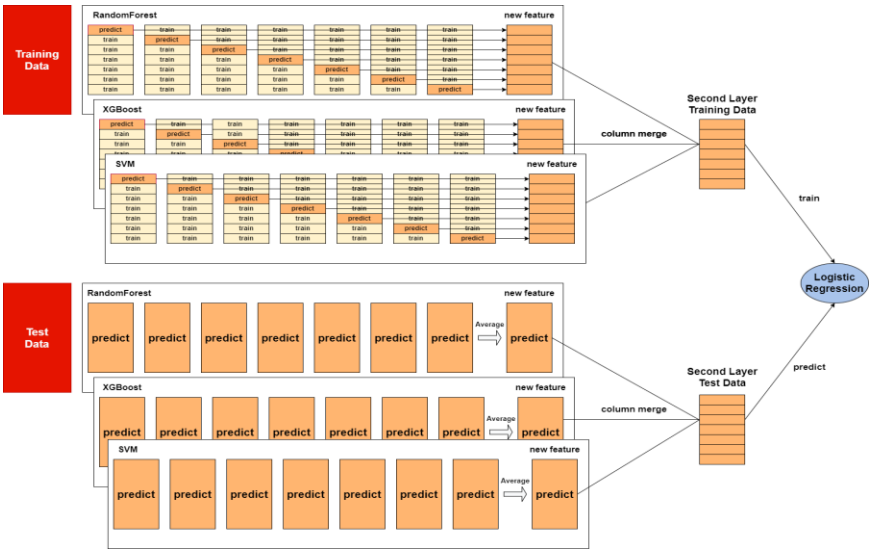

图 4 算法架构图

批注 [XF14]: 简要补充一些：基于 xxx 他人相关的研究工作，3-4 行字足够

批注 [XF15]: 这图应该是你画的，是吧？

批注 [刘16R15]: 自己画的图

5、模型评估

由于本实验中共需要解决涉及到 18 类药物的推荐预测问题，最终，本文模型采用 Scikit-learn 机器学习包，针对 18 类药物分别完成上述设计的基于多分类器融合的呼吸病药物推荐模型的训练。同时，实验分别针对其中使用的四种算法进行单分类器呼吸病药物推荐模型的训练，通过对测试结果进行对比，展示不同模型在药物推荐问题上的效果优劣。

本文针对不同药物推荐模型采用的评价标准主要为准确率和 F1 值。准确率指正确分类的样本数占总样本数的比值；F1 值与精准率与召回率有关，前者指预测为正的样本中，实际为正的比例，后者指实际为正的样本中，预测为正的样本所占的比例，而 F1 值是精准率与召回率的加权调和平均值。具体评价指标的数学公式如下：

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \quad (4),$$

$$Precision = \frac{TP}{FP + TP} \quad (5),$$

$$recall = \frac{TP}{FN + TP} \quad (6),$$

$$F1 = \frac{2 * Precision * recall}{Precision + recall} \quad (7).$$

其中，P 为正样本总数，TP 为预测为正，实际为正的数量，TN 是预测为负，实际为负的数量，FP,FN 同理。

首先，实验分别使用 18 类药物使用数据对基于 stacking 集成学习方法的多分类器融合呼吸病药物推荐模型进行训练及测试，最终模型准确率和 F1 值如图 6 所示（横坐标 M1—M18 指代具体药物见表 3），通过平均操作得到总体模型准确率为 86.1%，F1 值为 86.3%，说明该模型起到了良好的药物推荐效果。

批注 [XF17]: 与表 3 对应吗?

批注 [刘18R17]: 18 类药物完全对应

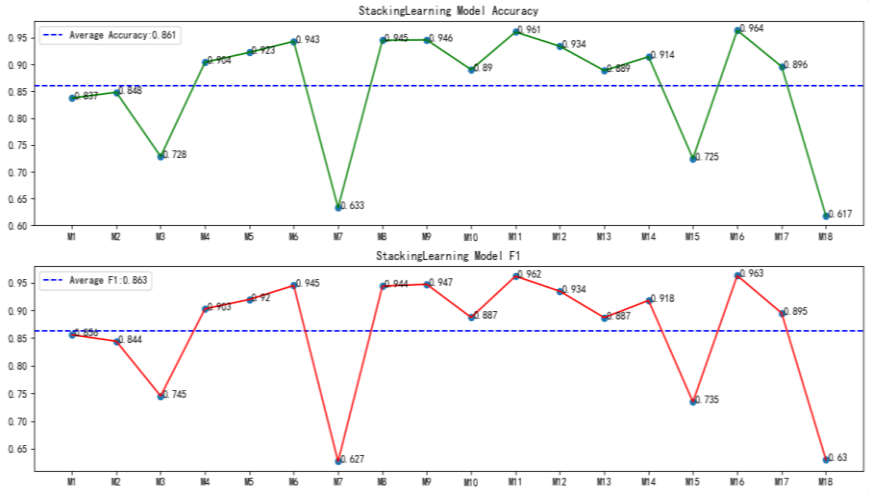

图 6 基于 stacking 集成学习的药物推荐模型准确率和 F1 值

同时，实验又分别针对 RandomFroest、XGBoost、支持向量机和逻辑回归四种算法进行了单分类器呼吸病药物推荐模型的训练，并与基于 stacking 集成学习方法的多分类器融合

呼吸病药物推荐模型进行准确率和 F1 值两类评价指标的对比，结果如表 3 所示。

表 3 不同药物推荐模型评价指标对比表

模型	准确率/%	F1 值/%
RandomForest	84.7	84.9
XGBoost	84.4	84.6
SVM	78.2	79.0
Logistic Regression	73.0	73.2
Stacking	86.1	86.3

由表中数据分析得出，相较于分别基于 RandomFroest、XGBoost、支持向量机和逻辑回归四种算法的单分类器呼吸病药物推荐模型，运用 stacking 集成学习的多分类器融合呼吸病药物推荐模型在模型评价指标上均有了一定的提升，具体在准确率方面分别提升了 1.4，1.7 和 7.9 和 13.1 个百分点。实验证明，该模型较单分类器呼吸病药物推荐模型更能有效利用呼吸病类型、药敏试验等医疗数据进行该 18 类药物的智能推荐，从而为耐药菌导致的呼吸病开药难问题提供了有效解决思路。

6、结语

本文设计了一种基于 stacking 集成学习的多分类器融合呼吸病药物推荐模型，具体在 RandomForest、XGBoost 和 SVM 的基础上，采用逻辑回归对以上三种模型的训练结果构建的新数据集进行二次训练，最终得出的呼吸病药物推荐模型在准确性和 F1 值两类评价指标均有了一定提升。该呼吸病药物推荐模型可利用一系列病患数据进行呼吸病药物的智能推荐，可有效针对耐药菌的出现提出正确的药物推荐清单,对医院的呼吸病智能辅助治疗有参考价值。

参考文献

[1] Xue Y， Chu J， Li Y，et al. The influence of air pollution on respiratory microbiome: A link to respiratory disease[J]. Toxicology Letters, 2020, 334.

[2] 周云春,董文斌,李世福,陈黎跃,张继华.呼吸系统疾病住院患者医院感染的相关因素[J].昆明医科大学学报,2020,41(09):86-92.

[3] 蔡晶娟,余春丽,卢亚陵,罗艳,孙照华,刘新颜,杨力夫.病原菌检测在急性呼吸系统感染性疾病患儿中的应用价值[J].检验医学与临床,2021,18(17):2508-2511+2516.

[4] 李鹏飞,鲁法明,包云霞,曾庆田,朱冠烨.基于医疗过程挖掘与患者体征的药物推荐方法[J].计算机集成制造系统,2020,26(06):1668-1678.DOI:10.13196/j.cims.2020.06.023.

[5] 刘文斌,吴倩,杜玉改,方刚,石晓龙,许鹏.基于个性化网络标志物的药物推荐方法研究[J].电子与信息学报,2020,42(06):1340-1347.

[6] Dagliati A， Sacchi L， Cerra C，et al. Temporal data mining and process mining techniques to identify cardiovascular risk-associated clinical pathways in Type 2 diabetes patients[C]// IEEE-EMBS International Conference on Biomedical & Health Informatics. IEEE, 2014.

[7] 李睿易,鲁法明,包云霞,曾庆田,朱冠烨.基于药物疗效日志的临床路径挖掘方法[J].计算机集成制造系统,2019,25(04):1017-1025.DOI:10.13196/j.cims.2019.04.026.

[8] Blei D M， Ng A Y， Jordan M I . Latent Dirichlet Allocation[J]. The Annals of Applied Statistics, 2001.

[9] Shlens J . A Tutorial on Principal Component Analysis[J]. International Journal of Remote Sensing, 2014, 51(2).

[10] Chawla N V， Bowyer K W， Hall L O，et al. SMOTE: Synthetic Minority Over-sampling Technique[J]. 2011.

[11] Tomek I . Two Modifications of CNN[J]. IEEE Transactions on Systems Man & Cybernetics, 1976, SMC-6(11):769-772.

[12] ZHOU Z H . ENSEMBLE METHODS: FOUNDATIONS AND ALGORITHMS[M]. TAYLOR & FRANCIS, 2012.

[13] Breiman. Random forests[J]. MACH LEARN, 2001, 2001,45(1)(-):5-32.

[14] Chen T , Guestrin C . XGBoost: A Scalable Tree Boosting System[C]// the 22nd ACM SIGKDD International Conference. ACM, 2016.

[15] Platt J C . Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. 1998.

[16] N BACAËR. THE DIFFUSION OF GENES (1937)[J]. A SHORT HISTORY OF MATHEMATICAL POPULATION DYNAMICS, 2011.