# 基于 Python 数据挖掘算法对中国罕见病医疗保障体系的多维度研究

刘祥帅

（携程集团，北京 310000）

**摘　要：**罕见病是由近七千种不同疾病组成的多样化集合，其中大多数疾病只影响少数人。由于罕见病患者在数千种不同疾病中的罕见性和分散性，医疗系统没有很好地识别或量化罕见病患者的医疗需求，从而未能形成合适的医疗保障体系。为了有效地利用医疗数据给罕见病防治和辅助诊断提供科学依据，本课题将使用 Python 数据挖掘算法作为核心算法，开展对中国罕见病医疗保障体系的多维度研究。基于 python 的数据挖掘算法的采用将患者的数据进行整合，通过实现每个患者的 360 度视角，提供以患者为中心的医疗保障体系，从而在改变医疗行业的结果方面发挥重要作用。

**关键词：**数据挖掘；罕见病；医疗保障体系

根据世界卫生组织对一种罕见疾病的定义，中国至少有 1000 万人患有罕见疾病。对于我国超过 13 亿的人口来说，这一估计似乎是保守的。2 国的罕见疾病包括成骨不全、神经肌肉疾病、法布里病、高切氏病、苯丙酮尿症、血友病 A 和 B、淋巴管平滑肌瘤病、白化病和肢端肥大症。不幸的是，由于现有医疗保障体系对罕见病的识别的不完善，患有此类疾病的中国患者通常很难获得最好的保证。尽管公众对罕见疾病的认识正在提高，罕见病患者及其家属、患者权益团体、医疗保健专业人士、律师和人大代表正在共同努力制定罕见病医疗保障体系而努力，在推动罕见病在医疗保障体系中自动识别和定级，人为的努力往往是不够的，还需要利用先进的数据挖掘算法来辅助。本文将整合多个罕见病数据，构建罕见病医疗保障体系的研究，为罕见病病防治和管理提供科学的指导。

## 一、罕见病及其挑战

在 30000 种已知疾病中，约 6000–7000 种被定义为罕见疾病，也称为孤儿疾病。尽管每种罕见疾病指征的患者人数较少，但患有罕见疾病的患者总数却很大。他们中的许多人不知道自己受到一种罕见疾病的影响，仍在寻求诊断或治疗。总的来说，所有罕见疾病的患病率约为 5%，这相当于糖尿病患病率的一半。许多罕见疾病会对预期寿命和生活质量产生重大负面影响。由于这些疾病中有相当一部分是由基因突变引起的，因此许多患者都是患有遗传性疾病的儿童。对医生来说，治疗罕见疾病并不少见。然而，在许多情况下，护理人员并不知道他们正在治疗一种罕见疾病。

对于医疗保健系统，我们从经济角度来看，医疗保健系统的三大任务是：医疗保健系统应为国内患者提供稳定、适当和高质量的护理（稳定）。人们应该享有平等的医疗保健设施。应该优先考虑那些需求最大的人。卫生保健系统的融资方案应符合社会的公平价值（分配）。卫生服务应有效地产生健康结果，医疗保健中使用的资源应用于最大限度地提高患者的福利（分配）。这些正是世界各国所有卫生政策面临的经济挑战，尤其适用于罕见疾病的治疗。

从经济和组织的双重角度来看，罕见疾病是卫生保健系统面临诸多挑战的原因：首先，在许多情况下，罕见疾病的诊断既困难又耗时，因为大多数医生对这些疾病几乎没有或根本没有经验。因此，需要教育努力和更好的信息系统来支持医生和告知患者。在进行诊断的过程中，一些患者在医疗系统中经历了一段漫长的旅程，通常既得不到正确的治疗，也得不到疾病的名称。加速诊断可以减少与健康相关的痛苦以及医疗资源的使用不足和滥用。在未来，基于基因组测序方法的新型诊断技术将改善诊断。目前，基因组测序仍然很昂贵。其次 . 由于患者人数较少，问题是如何为这些患者组织适当的护理。特别是医疗中心之外，由社会医疗保险系统资助，为患者提供治疗。许多这些专科护理单位资金不足，医院为罕见疾病患者提供高质量护理的经济激励也较少。这些例子说明了为什么罕见疾病患者的医疗费用昂贵，而且在某些情况下治疗资源太少。最为重要的是，在许多情况下，罕见疾病的治疗费用极高。特别是，由于接受治疗的患者数量少，以及提供罕见病的制药公司的独特地位，支付者的药品成本相当高。社会和付款人必须共同努力，决定他们愿意为罕见疾病治疗付费的程度。由于资源有限，存在宏观经济配置困境：如果社会在治疗罕见疾病上花费更多，那么用于治疗更常见疾病的资源就会更少。

罕见病通常很难诊断，许多患者都经历了漫长的诊断旅程，称为诊断之旅，以获得准确的诊断。即使在准确诊断的情况下，只有不到一半的罕见病映射到国际疾病分类（ICD）10 代码，具有特定 ICD 10 代码的罕见病更少（<20%），这导致大多数罕见病在数据库（如付款人 / 保险数据库）中被低估和计数不足，以及无数下游效应，例如罕见病患者的不精确编码以及对罕见病患者和疾病本身的不良跟踪和理解。此外，在没有诊断的情况下，通常无法可靠或一致地使用一组实验室、笔记和其他特征（如可计算表型）来识别罕见病

患者。自2018年以来，中国从国家战略层面启动了罕见病特别安全项目。我国政府制定了中国第一份罕见病名单（包括121种疾病），成立了罕见病诊断、治疗和安全专家委员会和中国罕见病联盟。政府还鼓励各省将罕见病安全规划和战略纳入重点工作，推动建立罕见病临床研究中心，制定诊断和治疗指南。就我国首个罕见病名单中的121种疾病的治疗药物而言，其中83种已在国内上市，50种已纳入国家医疗保险，同时还实施了罕见病药物减税政策，并开发了多个慈善基金会和企业捐赠计划，以增加患者的药物可用性和可获得性。然而，对于罕见疾病没有明确的定义和相应的编码；由于诊断困难和分类错误，大多数罕见疾病的经济负担无法准确计算。在服务提供体系方面，由于筛查涉及的罕见病种类较少、诊断能力水平较低、康复计划较少以及缺乏社会关怀和知识培训，中国的罕见病社会保障体系需要进一步完善。中国有必要为罕见病患者提供全面、全面的医疗保健和社会关怀。由于不同地区的政策制定和实施条件不同，因此在中国制定罕见病政策时，应首先在具备所需资质的地区进行试点，然后制定成熟的国家计划。

大数据在银行业、零售业等多个行业中的作用已经得到了见证，但在最近几年，即使是医疗保健系统也显示出准备好实现转型，并利用大数据解决方案为医疗保健系统提供的好处。医疗行业已经变得非常数据密集，数据来自多个来源。跨异构数据源的数据集成是目前确定的最大挑战，否则将有助于更好地了解可用数据。有一些可行的解决方案可用于处理医疗保健中的异构数据，许多研究表明Python数据挖掘算法具有容易理解、鲁棒性好、准确性高、模型可重用的特点。因此在研究选择其作为罕见病医疗保障体系多维度研究的核心算法。并以此展开详尽的文献调研，掌握已有的改进策略，为下一部分算法的设计和改进奠定基础。

## 二、罕见病数据的挖掘应用

为了为分类提供数据基础，我们使用了119例结节病和130例囊性纤维化，这是一种发病率相对较高的罕见肺部疾病。此外，还发现了35例类似症状的病例。对于通过Python（版本3.7.3）设计的自身应用程序模型，对四种常见的ML分类器支持向量机（SVM）、朴素贝叶斯（NB）、K近邻（KNN）和多层感知器（MLP）进行了训练和优化。将现有系统的分类性能与自优化应用程序进行比较，以测试现成系统对罕见病自动编码的适用性。列车测试分割法用于比较预测建模中评估算法的性能。在此过程中，数据集被划分为两个子集。第一个子集用于拟合模型，称为训练数据集。第二个数据集称为测试数据集，用于评估模型在新数据上的性能。训练数据集将再次拆分为训练集和验证集（80\20拆分）。训练集用于建立具有多个模型参数的模型。然后将每个训练模型与验证集进行比较。验证集包含具有已知标签的数据，但模型不知道这些分类。因此，基于使用验证集的预测，可以评估模型的质量。基于使用验证集时发生的错误分类，可以使用验证误差最小的参数集优化模型参数。应用程序在NLP过程中使用文本规范化、柠檬化、部分语音标记、标记化、词包、词频逆文档频率（TF–IDF）和停止词，并进行优化设计和参数设置。MM护理的详细NLP方法设计为专有。分类和模型质量使用性能指标F1分数、精确度和召回率（所有分数的最高值为1.0）进行评估。还考虑了不同训练量的结果变化。为了验证，提取了在自然语言处理过程中识别的语料库特定术语。

MM–care软件的三个分类器的性能刚好低于目标值0.8，测试数据集的平均F1得分约为0.78。在自主开发的应用程序中，分类器——SVM、NB、MLP——表现最好，F1得分高于0.81，略高于目标值。KNN分类器提供的分类精度较低（F1得分<0.73）。这里的平均F1分数为0.82。为自身应用程序的交叉验证而额外实现的学习曲线的开发（见图1）表明，对于少量训练案例，训练值显著高于所有模型的验证值。这表明模型与训练数据“拟合过度”。在所有模型中，都有一个明显的趋势，即增加训练案例的数量可以提高泛化能力。在刚刚超过50个案例的训练集中，分类器SVM、NB和MLP的F1得分已超过0.9。决策相关术语的分析表明，并非模型识别的所有术语都明确指向两个罕见病中的一个。但自己的应用程序和MM care软件也发现了越来越多地出现在肺部疾病领域的术语，偶尔明确提及本研究中考虑的呼吸窘迫综合征。在本文中，我们研究了支持罕见病编码的不同方法。商业系统MM–care的分类器在测试集上的平均F1得分为0.78，而自主开发的应用程序的平均F1分数为0.82。为了评估所开发模型在日常临床实践中的性能，需要对更为多样的数据集进行分析，这些数据集不仅包含两个代码（或疾病），而是分类系统的多种不同代码。正如我们的结果所示，较大的数据集（>200）可能会导致更好的模型性能。在本研究中，使用ICD–10分类进行编码。然而，在ICD–10中，大约6954份RD（截至2015年）中只有355份是唯一编码的。整合孤儿网命名法（其中每个RD都有一个唯一的代码）将是实现所有RD自动编码的逻辑步骤。此外，将预先训练好的模型应用于新模型的转移学习可以获得更多的训练数据，并启动医疗编码系统的长期改进。概念证明表明，即使是低发病率的疾病和相关的小训练数据集，也可以自信地识别和自动编码。通过该模型对添加到测试数据集中的具有类似症状的其他疾病进行了区分。现成的应用程序和自己的应用程序都能够提供较高的分类正确率。未来需要更全面的研究来评估模型在临床实践中的性能。

## 四、结论

为了改善患者护理并通过利用最新技术使其具有成本效益，医疗保障行业已经做好了转型准备。数字化带来了大量的数字数据，尤其是在医疗行业。本文提出了一种使用正确的技术和架构处理异构医疗数据的方法，该方法有可能以最佳成本改变医疗保障体系对于罕见病患者的保障。