机器学习实验报告

——————病例分析与处理

姓名：林聪　　学号：16339026　　日期：2019/7/1

摘要： 病历文本的语义解析与处理是进行病历信息挖掘的基本工作。本次实验利用 jieba 模型和 Doc2Vec 模型对病历报告文本进行了文本特征提取，通过 ADASYN 样本上采样方法构建出训练数据集，并训练出了用于对肾病、酮症、心脏病、眼病、周围神经病以及足病进行辨别的逐对类别对抗 MLP 分类模型。此外还尝试借助关联规则算法 A-Priori，利用已有病历报告中的诊断结果，对关联度较高的并发症状组进行筛查。

1. 导言

a) 问题背景

病历是患者疾病发展和医生进行诊断的记录汇总，是对采集的资料进行梳理整合并按一定的格式和要求记载的患者医疗档案，对医疗、教学、科研及医院管理有着非常重要的作用。近十多年来，随着医疗信息化技术的不断发展壮大，各级医疗机构产生并积累了数目庞大的医疗病历数据，其中蕴含着宝贵的医疗信息资源。然而，大部分的病历都以不同结构或者半结构化的方式存储，并且由于缺乏统一的医学知识库和病历撰写标准，医学标准的不统一以及医学术语口语化严重影响了病历的整理、分析和共享。本次实验的问题背景就是建立在这样的大环境上的。为了能够利用医疗病历中的信息，需要对病历文本进行内容筛选和提取，并提取特征，从而进一步根据应用场景建立模型。

b) 实验分析

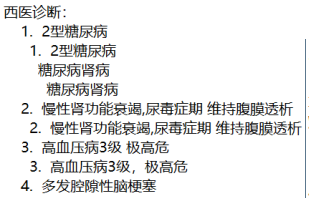
本次实验提供了 1300 多份 txt 格式的真实匿名医疗病历档案，文件大小从 4 KB 到 100 多 KB 不等。每份病历的诊断结果被明确标注为肾病、足病、酮症、眼病、心脏病和周围神经病这六种糖尿病并发症中的一种，这六种病症分别有 473、100、299、142、15、329 例。所有文件基本都遵循相同的格式：第一部分为患者的基本信息，包括职业、性别、年龄、入科时间等；第二部分为患者基本病况和初步检查结果，包括主诉内容、详细病况、体格检查、辅助检查等；第三部分为诊断结果，包括多项中医诊断结果和西医诊断结果。第四部分为患者入院期间的若干次查房记录。下图为一位匿名肾病患者的病历节选：

年龄：72岁
入科时间：2012-07-22 15:56
婚姻状况：已婚
记录时间：2012-07-22 16:25
出生地：广东省揭西县
病史陈述者：患者本人
民族：汉族
发病节气：小暑

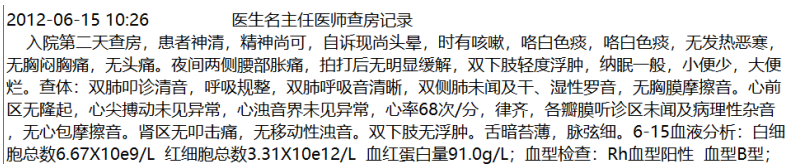
患者基本信息机节选

主 诉：反复口干多饮28年，加重伴双下肢浮肿1月余。
现病史：患者于28年前无明显诱因感口干多饮,在我院门诊诊断为"2型糖尿病"，予消渴丸及中药治疗，口干口渴症状缓解，但血糖控制欠佳。多次因血糖控制不佳在我院住院治疗，经治疗后病情好转出院。出院后皮下注射诺和灵30R（早22u晚18u）及中午口服阿卡波糖控制血糖，血糖控制一般。2012年6月4日因"咳嗽咳痰7天，双下肢紫癜1天"再次入我科住院治疗，住院期间患者出现双下肢轻度浮肿，纳差乏力等，经治疗后症状好转出院。出院后皮下注射诺和灵30R（早16u晚14u）控制血糖，血糖控制尚可，但双下肢浮肿无改善，纳差乏力症状加重。今为进一步治疗来我院就诊，门诊以"糖尿病肾病"收入院。入院时症见：患者神清，精神可，双下肢中度水肿，纳差，打嗝频繁，视物稍有模糊，偶有耳鸣，四肢末端麻木、乏力，活动后气促，无头晕头痛，无胸闷心悸，无腹痛腹泻，眠可，大便正常，小便深黄如茶色，夜尿4-5次，有少量泡沫。

患者基本病况节选



患者诊断结果节选



患者入院查房记录节选

- 问题一：并发症预测

问题一要求根据尚未确诊的病历的内容判断其患者的糖尿病并发症类型。测试数据包含 510 份 txt 文件，有效文件大小在 50 KB 以内。每份文件仅含上述格式的第一部分和第二部分，即患者的基本信息和患者的基本病况和初步检查结果，并不包含明确的诊断结果，其余内容与已有的带标记病历数据相同。

- 问题二：并发症状组筛查

问题二是一个开放性问题，要求从病历数据的丰富信息中分析提取一些有价值的信息。本次实验选择了并发症状组筛查这个任务，具体而言，试图利用带标记的病例文本中的诊断结果部分的内容，对不局限于肾病、酮症、心脏病、眼病、周围神经病以及足病的其它事先或事后并发的病症进行挖掘，从而为将来的症状预测和排查及医学研究提供指导。

c) 模型选择

- 问题一

问题一的本质是多分类问题，它包含一个从数据获取、处理到训练分类模型的完整流程。由于数据是文本数据，具体而言，解决该问题有以下几个步骤：数据准备，包含文本的读取、文段的选择、字词的分割和清洗；数据特征提取，即将文本转化为特征向量；训练数据准备，包括训练集与验证集的划分以及对本次实验中不均衡样例的采样；训练分类模型，即训练并测试分类模型。

数据准备的重点在于字词的分割，由于本次实验涉及中文病历文档，并且大量非常见的医学词语和口语化表达，比较适合进行分词任务的模型为可以基于隐马尔科夫链对未收录词进行划分的 jieba 分词模型。

文本数据特征提取的常用方法有 TF-IDF、Word2Vec、Doc2Vec 等。由于医疗文本篇幅较长，且描述性的内容很多，TF-IDF 较难控制特征维数，并且它与 Word2Vec 都较难保留长文段的词语顺序信息，因此首选的模型是基于 Word2Vec 原理的、能直接对文段进行特征化的 Doc2Vec 模型。

由于原始数据各类别间的数量高度不平衡，如肾病案例有 473 例，但心脏病案例只有 15 例，对样本进行上采样显得尤为重要。常用的借助样本特点进行样本上采样方法有 SMOTE 算法及其变种和 ADASYN 算法，它们都是基于对已有原始样本及其近邻样本进行插值的上采样算法，主要的区别在于如何选取需要在其周边进行采样的原始样本。SMOTE 算法随机选取原始样本，ADASYN 则根据 KNN 算法的结果判断原始样本点需不需要上采样，SMOTE 算法的变种也是类似地采用如 KNN、K-means 和 SVM 等辅助算法和标准来判断原始样本点需不需要上采样。

考虑到本次实验的病历文本内容复杂，不同类别之间的界限不明晰，病状及诊疗记录也有许多相似或相同之处，为了能成功进行分类，除了进行上述的上采样之外，还要求分类器有较强的表达能力，本次实验选择了这类模型中较为简单的 MLP 模型。

- 问题二

问题二的本质是关联规则挖掘问题，其关键在于寻找频繁项集，即频繁地同时出现的元素的集合。本次实验采用了关联规则算法中最有代表性的 A-Priori 算法，这是一种从较小的

频繁项集逐步构建并筛选较大频繁项集的算法。

2．实验过程

a) 问题一

- 数据准备

首先对病历文本文段进行选择。根据观察并借助常识，应当认为第一部分的患者基本情况，包括职业、年龄、性别等信息，对后续预测并发症是没有信息贡献的，因此将第一部分文段完全除去。由于第三第四部分并不会出现在测试数据当中，且第三部分内容为具体的诊断结果，而第四部分为非常繁杂、难以处理的住院信息，并且这部分信息是建立在已经知道诊断结果的基础之上的，为了避免数据泄露并确保本次实验的顺利进行，在此将这两部分文段内容也除去。剩下的第二部分包含主诉内容、现病史、既往史、过敏史、个人史、婚育史、家族史、体格检查、辅助检查及一些专科检查。根据观察并借助常理，病历中的既往史多为否定性的，可反映的问题不多，过敏史、个人史、婚育史、家族史则千篇一律且与并发症关系不大。体格检查内容繁杂但也是千篇一律地以否定性内容为主，而辅助检查和专科检查内容非常多变，并且涉及大量出现与否不确定的项目及数值。这些项目都不利于文本特征的提取，而主诉及现病史则能比较精炼且有倾向性地反映患者显现出来的病症，并且这部分确诊前的内容是测试数据中也有的，故选取这部分内容为病历的核心，以进行后续的步骤。

接下用 python3.6.4 中的 jieba 0.39 对病历文段进行分词，并去除包括数字和标点在内的符号以及其它与本次实验关系不大的停用词。

- 数据特征提取

利用 gensim 3.7.2 中的 Doc2Vec 方法提取病历文段的特征向量，综合性能和速度的考量，特征向量的维数设为 256，窗口大小根据对病历文本的观察设置为 3。

- 训练数据准备

可选择的方法有 SMOTE、ADASYN、考虑类边界的 BorderlineSMOTE 和借助支持向量的 SVMSMOTE，这几种方法尤其各自的特点，由于病历文本特征向量的分布难以预料，经过测试，本次实验最终选择了能产生最好效果的 ADASYN，所使用的是基于 imblearn 0.0 的实现。

- 训练分类模型

本次实验采取了多种策略来训练 MLP 分类器，它们分别是直接多分类、OneVsRest 分类和 OneVsOne 分类策略。之所以尝试了三种不同的方法，是因为实际测试过程中发现采用直接多分类，即直接训练 MLP 输出六个类别的概率分布向量，效果并不理想，实验结果表明可能是数据分布过于诡异使得 MLP 无法将六类数据很好地区分开来。由此希望将 MLP 的表达能力集中到具体的类别上去，于是便尝试了 OneVsRest 分类的策略，即训练多个分类器，每个分类器仅仅负责判断数据是否属于特定的某个类。结果表明采用 OneVsRest 策略相比直接多分类的策略所得到的结果略有提升，最终采用 OneVsOne 分类策略，即训练多个分类器，每个分类器仅仅负责判断数据属于特定的某两个类中的哪一个，获得了目前最好的结果。

b) 问题二

- 数据准备

由于病历文本的诊断结果部分撰写并不规范，症状总量太大，不同类别并发症所确诊的病症相差较大，鉴于 A-Priori 需要枚举型的数据以及一定的支持度，同时为了后续验证的方便，本次实验仅选择样本数量最多达 473 的肾病病历样本来进行并发症状组筛查。

- 模型训练

由于病状总量大，表述不规范，清洗难度较大，本次实验主要关注规则的置信度和提升

度，支持度、置信度和提升度的阈值分别设置为 0.01、0.75、1.0。

3．结果分析

a) 问题一

经过多次实验，所获得的较好的、合理的结果的具体情况如下：

并发症类	肾病	酮症	心脏病	眼病	周围神经病	足病
原始训练样例数	373	230	8	110	254	77
采样后训练样例数	373	366	376	371	367	399
验证样例数	94	69	7	32	75	23

实验样例情况

	肾病	酮症	心脏病	眼病	周围神经病	足病	总体
查准率	0.9032	0.8173	0.9921	0.9437	0.6980	0.8839	0.8739
查全率	0.6756	0.9044	1.0000	0.7224	0.8692	0.9925	0.8623
F1	0.7730	0.8586	0.9960	0.8183	0.7743	0.9351	0.8605

训练情况

	肾病	酮症	心脏病	眼病	周围神经病	足病	总体
查准率	0.6061	0.5060	0.5000	0.8750	0.4455	0.3333	0.5482
查全率	0.4255	0.6087	0.1429	0.6563	0.6000	0.3478	0.5233
F1	0.5000	0.5526	0.2222	0.7500	0.5114	0.3404	0.5229

验证情况

从以上结果可以看出分类器在六类并发症的判断任务上都取得了良好的结果，其中除了肾病的查全率和周围神经病的查准率略低之外，六类并发症的查准率、查全率和 F1 度量均取得了不错的数值，说明分类器能够较好地诊断出这六类并发症。

然而验证结果相比测试结果有明显的下降，仅有对眼病的排查保持在较好的水平，对肾病、酮症和周围神经病的排查处于平均水平，对心脏病和足病的排查则处于较低水平。进行过多次修改模型、调整参数和重复试验之后，这已经是我在验证集上所能得到的最好结果。其原因很有可能在于病历数据的分布过于分散，即使经过上采样，训练集数据也很难反映验证集的分布。尤其像心脏病样本集这样的总共只有 15 个样本的数据集，如果只有其中一小部分数据则很难反映其余数据的分布。

要克服这样的问题，扩大样本容量是一个最自然的选择。在这次实验中，由于样本受限，我不得不采用不同的采样方法、分类模型和训练策略，试图解决这一问题。最终结果虽然有所好转，但仍然不容乐观，这也有待我的进一步学习和研究。

鉴于以上分析，最终用于测试集分类的分类器是利用所有训练数据的上采样数据训练的，其训练情况如下，分类结果见附件：

	肾病	酮症	心脏病	眼病	周围神经病	足病	总体
查准率	0.9000	0.7282	0.1312	0.8012	0.7107	0.6463	0.7783
查全率	0.6553	0.6990	0.5333	0.9085	0.7690	0.9500	0.7396
F1	0.7584	0.7133	0.2105	0.8515	0.7387	0.7692	0.7481

训练情况

b) 问题二

所挖掘到的关联规则，即并发症状组，见附件 syndrome_detected.txt，在这里仅展示一

些值得注意规则：

并发症状组	置信度	提升度
糖尿病性肾病、泌尿系感染→高血压病	0.8333	1.3333
三支血管病变、冠心病→糖尿病性肾病	1.0000	2.2136
脑梗塞后遗症、前列腺增生→高血压病	1.0000	1.6000
尿毒症、糖尿病性肾病→肺部感染	0.7857	1.7393
湿疹、高血压病→糖尿病性肾病	1.0000	2.2136

并发症状组规则节选

4．结论

本次实验相比上次实验更接近实际生产，也是三次实验中最具有实战性的实验。如上文结果所述，我在这次实验中获得的效果并不好，并且在尝试了许多改进方法后，所取得的结果依旧提升不大。这次实验作为本学期机器学习课程的最后一次实验，警示我在机器学习方面还有很多的知识、能力和技巧需要学习和掌握。在病例分析与处理这个接地气的课题中遇到的如样例不平衡和样例分布复杂等问题给我带来了很大的困难和挑战，也激发了我探索求知的热情，在结课后的未来学习中，我会继续了解和探索实践这方面的解决方案，希望在未来如有机会能做出一个更成功的病历确诊分类器。

主要参考文献：

[1] 病历智能处理引擎的设计、实现和应用
https://blog.csdn.net/dev_csdn/article/details/78889043
[2] imbalanced-learn Documentation
https://imbalanced-learn.readthedocs.io/en/stable/
[3] Shickel, Benjamin, et al. "Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis." IEEE journal of biomedical and health informatics 22.5 (2017): 1589-1604.
[4] Kononenko, Igor. "Machine learning for medical diagnosis: history, state of the art and perspective." Artificial Intelligence in medicine 23.1 (2001): 89-109.