2018年12月17日

按照上周的电话沟通内容，做了以下工作：

1. 构造多分类标签，发现每个角度有如下情况：

脏腑 13个类别 气血津液 11个类别 八纲10个类别 卫气营血 4个 三焦 3个，通过这种方式，确实比上次用sklearn构造的多标签效果好（比如上次我发送的不匹配数据，有125条左右，现在不匹配有75条左右）

1. 第二部分的预测严格按照表格匹配（不再使用SVM），并计算准确率，得到如下结果：

Cidian\_data15\_size3.pkl 准确率89.83

Cidian\_data15\_size5 准确率 89.54

Cidian\_data15\_size7 准确率 89.26

1. 使用用不同的窗口大小，在训练词向量的时候。最开始我用的window size是5，上面展示了不同size的结果（差千分之三，如果考虑误差的情况下，可以认为基本没啥区别）

问题：

1. 苗总说的排序还没做？我想问下排序是啥意思？是指在训练词向量之前做排序，还是输入神经网络之前做排序？
2. 神经网络结构暂时用的是全连接，这个我觉得咱们可以把整个的解决方案都敲定，再来调整神经网络的结构。

思考：

关于怎么把这个效果再提升，我想法是如下这样的，如果苗总有啥其他的想法我们电话再沟通！

我们想提升这个最终效果其实有三个方向可以做。A 选定一种更好的词向量表示（比如有相应的医学语料，或者能不能用到迁移学习什么的） 。B 调出最合适的神经网络模型（包括层数，激活函数，dropout等）。C对于得到的五个角度的辨证，转化到四十分类的这部分，如何优化！

我觉的感觉C可以放到较高的优先级，首先说A：之前也试验过，我们目前用Word2vec训练出来的词向量根随机赋值也没有什么区别（因为word2vec对语料的质量和数量都有要求，）再说B:我们目前来看其实最大的难度就是数据集太小，在小的数据集上可能模型M1准确率95% 模型M2准确率90%，可能也不见得真实使用的时候M1真的比较好。可能就是花费大量时间去做，提高了一个百分点或者两个百分点，也不一定是真正意义的提高。说到C：我倒是觉得这部分可改进的空间比较大，如果将整个称之为一个系统的话，我觉得系统的瓶颈主要被限制到根据八纲 脏腑预测最终四十分类的环节上了，所以咱们能不能有啥好的办法去提升这一部分（比如肝胃不和 真实的应该是 脏腑：“肝胃”，八纲：“实”，模型预测成了“肝胃， 虚实”，虽然错了，没有匹配上，但我们知道这个“肝胃 和 虚实” 不会错的太离谱，我们预测最终是肝胃不和！就是我们让根据八纲预测最终辨证有较强的鲁棒性！这是我的一点比较朦胧的解决思路！） 我的目的是能不能把 根据中间辨证预测四十分类这一环节的鲁棒性提高！