Readmission Prediction

张少魁 计算机系 2018310862

胡旭强 计算机系 2018310800

概述

本次实验，我们分别选用了两个模型：神经网络与决策树，来对 10 年的医疗数据进行建模和预测。对于神经网络，我们分别实现了三种不同的网络结构，并对其进行实验和比较。

数据预处理

数据预处理我们根据不同的模型，有相应不同的预处理方法。由于我们操作数据的属性均为离散取值，因此我们会应用专门针对类别属性的预处理方法。首先，对于神经网络，我们对数据集做了 one-hot 编码，即用一个向量表示每一个属性，对应的属性位取 1，其余位均取 0。对于决策树模型，我们选用了整型（integer）的 label 编码，用每一个无比较意义的整数来代表每一个属性。

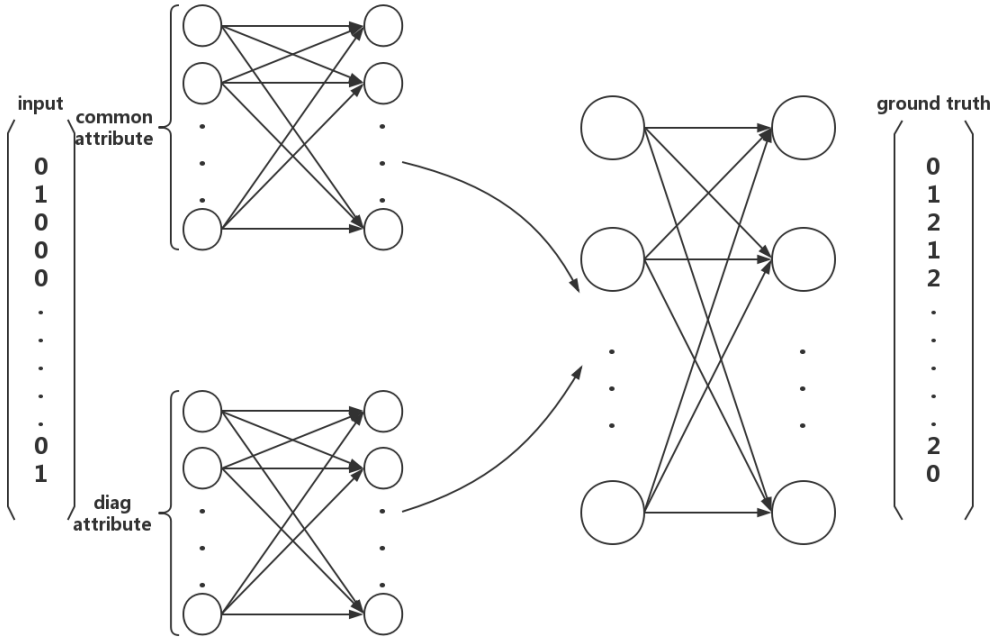
数据集中有大量的缺失值，对于不同的模型，我们也同样使用了不同的应对策略。对于二者，我们完全删掉了 ID、NUMBER 两个属性，因为这些属性都是属于每一个特定实例的属性，并没有很强的泛化性，很容易导致过拟合。并且这些属性会导致 ont-hot 编码过长，试神经网络难以训练。剩余的属性，对于神经网络而言，我们并没有显示地删掉任何属性，对于缺失值，我们使用所有位均为 0 的向量作为输入。直觉上讲，这样的输入并不会对神经网络的权值调节造成任何影响（ReLU 激活函数，0 点导数为 0）。另外，三个 diag 属性由于 ont-hot 编码后过长，我们也会单独处理。其余属性 ont-hot 后为 573 维，而三个 diag 属性已经达到了 2253 维。因此专门提取出的 diag 属性，会在模型设计上时做特别处理。

对于决策树的预处理，除了两个 ID、NUMBER 属性外，缺失值使用了最频繁的值作为填充，同时由于 weights 属性缺失值过多，我们也同样删掉了这个属性。最终，完全整型编码的属性将用于决策树分类。

模型

如上文所说，神经网络我们采用了三种不同的模型。下图为带分支（branch）的多层神经网络。其中上部的分支包含了除了 diag 之外的所有属性，而下方的分支包含了三个 diag 属性。定性来讲，这样首先可以加快模型的训练速度和防止过拟合，因为完全连接的网络，中间会产生大量的权值，使网络难以训练。其次，这样可以专门针对 diag 数据，提取更细化的特

征来表达三个特征，将不同的属性分而治之。最后的实验结果可以证明，使用带分支的网络确实在速度上和 F1 分数上有强于全连接的网络。除了带分支的网络以外，我们还实现了两个全连接的神经网络（多层感知机）来做对比实验，一个为包括 diag 属性的网络，一个只包含 diag 之外的属性。



神经网络的实现上，我们尝试了 leaky_relu、sigmoid、反正切函数、relu 作为激活函数。Sigmoid 和反正切函数得到的准确率均比较低，relu 对于训练来说比较慢，存在冷启动的情况，即初期训练让准确率来回震荡。因此最终选用了 leaky_relu 作为最终的激活函数。权值的初始化均使用均值为 0、方差为 0.1 的高斯分布。我们先后尝试了 ADAM 和单纯的梯度下降，ADAM 的收敛速度和精确率都明显高于后者。关于防止过拟合的策略，我们尝试了 L2 正则化、学习率指数衰减、Dropout、归一化、Residual Block 等策略。前两者能为模型的效果带来泛化性的提高。Dropout、归一化让网络难以收敛，所以并没有被我们最终采用。我们虽然尝试了一个脑洞，即 ResNet 中的残差模块（Residual Block），但是最终的效果没有任何提升（实现的源代码也包含在了最终提交的代码中）。（ResNet：He et al. Deep residual learning for image recognition）

我们选用 CART 作为我们决策树的模型，并且限制了决策树的最大深度来放置过拟合。同时我们尝试了不同的分割指标，最后选用了基尼系数。

实验

实现上，我们用 Tensor-Flow Core 来实现神经网络，并且用 Sklearn 来实现决策树，Numpy 被用于数据预处理。最终的实验结果使用了 Ten fold cross validation，数据集被**随机等分**成了 10 分，每个模型均被训练十次，每次使用不同的训练集和验证集。实验的最终结果如下表所示：

	Accuracy	Precision	Recall	MACRO-F1
Decision Tree	0.576332	0.492892	0.400452	0.441653
ANN	0.586621	**0.647774**	0.409252	0.498674
ANN Dense	**0.591126**	0.619101	**0.421538**	0.498526
ANN with Branches	0.588568	0.624058	0.416974	**0.500300**
Chou *et al.*	**0.650**		**0.552**	

从表中可知，稠密的神经网络（多层感知机）拥有最高的准确率，但相比其他模型，并没有很明显的提升。使用神经网络的精确率（precision），要明显高于决策树的精确率，但神经网络之间的精确率并没有很大的区分，只是使用简化的神经网络会有比较好一些的精确率。四个模型之间，召回率不相上下，稠密的网络会有略微的提高。对于最终的 F1 分数，神经网络的结果仍然明显高于决策树的结果，神经网络之间仍然在 F1 分数上不相上下，带分支优化的网络会有略微的提高。最后一行的结果参考了这篇论文：Chou et al. A Fully Private Pipeline for Deep Learning on Electronic Health Records。他们也是使用了神经网络，对这组数据进行分类。