Summary

1 DataSet

由于尚无适合我们研究的公开数据集。因此我们从中国公众裁判文件网在线收集数据。我们获取部分有关机动车的交通事故案件，并从这些案件中抽取了案件描述，以及案件所引用的相关法条作为原始数据。在对上述案件统计分析过程发现，案件所引用的法条数量不超过10条所占比例为74.2%，大部分案件引用法条数量为4和5条，分别占比9.5%和9.6%。引用法条数的平均值和标准差分别为7.51和5.35。因总的法条种类很多，但是部分法条出现次数极少，我们从中选取了出现次数较多的204类法条并对其按照1到204的序号进行编号，最终我们保留了3万条数据。

2 Preprocess

2.1 机器学习方法

我们对上述 3 万条数据进行划分，其中训练集、验证集、测试集的比例为8:1:1。首先，我们对数据的案件描述部分进行预处理，包括去除非法字符，数值型数据类别化等。其次，我们使用jieba作为分词工具，对案情部分进行分词。由于分词结果种类太多，构建出的特征矩阵过于稀疏，并且会消耗巨大的内存。因此，我们需要在构建向量空间模型的过程中进行特征选择。我们使用sklearn中的TF-IDF对上述分词后的结果进行处理，其中的关键参数设置如表1。为了提高分类正确率，并且加速模型收敛，我们也对TF-IDF处理结果进行了标准化处理。对于分类标签部分，我们把每个样本标签构建为维度为204的向量，并把相应法条序号位置设为1，其他位置为0。

$key\ params$	
min_df	5
$max\_features$	10000
$ngram\_range$	$(1,3)$
use_idf	$true$
$smooth\_idf$	1

2.2 深度学习方法

在数据集划分、以及对案件描述部分、分词部分都和机器学习方法保持一致。我们在深度学习部分采用了 $FastText$、$TextCNN$、$RestNet$、$Transformer$、$TextRNN$ 这几种方法进行实验。 各方法参数设置如下

$key\ params$	$FastText$	$TextCNN$	$RestNet$	$Transformer$
$batch\ size$	32	32	32	32

$key\ params$	$FastText$	$TextCNN$	$RestNet$	$Transformer$
$paddding\ size$	1000	1000	1000	1000
$embedding\ size$	512	512	512	512
$num\ words$	40000	40000	40000	40000
$dropout$	0.5	0.5	0.5	0.5
$weight\ decay$	1e-4	1e-4	1e-4	1e-4
$learning\ rate$	1e-3	1e-3	1e-3	1e-3
$optim\ methods$	$Adam$	$Adam$	$Adam$	$Adam$
$filter\ size$	$None$	[2,3,4,5,6]	$None$	$None$
$num\ head$	$None$	$None$	$None$	5

上述方法均采用 $Batch\ Norm$ 对数据进行归一化。

2 Metrics

常用的评估方法例如精确率，召回率，F1值常常被用于通常的分类模型评估方法。但是由于实验数据的标签类别非平衡且每个样本对应的平均标签长度较长，上述评估方法并非理想的评估方法。 本文中，我们自己定义了一种新的评估方法。对每个样本 x_i，其对应的真实标签（可能有多个）记为 y_i 。我们选取模型样本 x_i 预测的概率最大的 k 个类别标签，记为 k_i。统计 k_i 中正确标签个数 m_i 。其中对样本 x_i 正确预测标签集合 m_i 定义如下:

$$m_i = k_i \cap y_i \tag{1}$$

则样本 x_i 预测的准确率如下：

$$acc(x_i) = \frac{L(m_i)}{L(k_i)} \tag{2}$$

注: $L(x)$ 表示 x 集合的基。

最后对所有的样本预测准确率加权求均值得到模型准确率如下：

$$model\ acc = \sum_{i=1}^{N} acc(x_i) \tag{3}$$

注: N 表示样本个数。

3 Result

k	*LR*	*SVM*	*Bayes*	*MLkNN*	*FastText*	*TextCNN*	*ResNet*	*T*
1	66.87	63.53	42.67	58.07	66.47	65.47	66.93	57
2	62.00	58.43	30.73	52.07	60.07	59.77	61.37	51
3	57.07	52.36	25.73	46.93	53.71	54.49	55.91	45
4	52.38	47.15	22.10	41.12	48.98	49.98	51.42	40
5	48.49	42.52	20.17	37.59	44.88	46.21	47.64	36
6	45.30	39.34	19.08	34.66	42.18	43.10	44.69	33
7	42.81	36.88	18.34	32.59	39.49	40.54	42.16	31
8	40.67	34.59	17.78	31.02	37.42	38.42	39.98	30
9	38.56	32.77	17.46	29.62	35.57	36.56	37.78	28
10	36.84	31.28	17.27	28.43	34.12	35.06	36.05	27