一、本周研究内容

1. 整理了一份 ‘Twitter 预警数据集’ 文档，里面总结了项目背景，任务描述，数据集说明以及评价指标。

2. 对数据集分类类别不平衡的问题(imbalanced class dataset) 进行了处理，使用了重采样（resampling）方法，具体来说，直接对训练集里的反类样例进行欠采样 （undersample majority class）以及对训练集的正类样例进行过采样 (oversample minority class)。

3. 使用 Baseline 方法 KNN 和 SVM，对不使用任何采样的原始数据，使用了 oversampling 的数据以及使用了 undersampling 的数据这三种情况下进行了实验。实验结果表明使用 oversampling 处理后的模型表现最好。

二、项目实施当前状态

已经完成了对数据集类别不平衡问题做了 resampling 处理，使用 oversampling 和 undersampling 并取得了不错的结果。

三、本周成果

1. Twitter 预警数据集

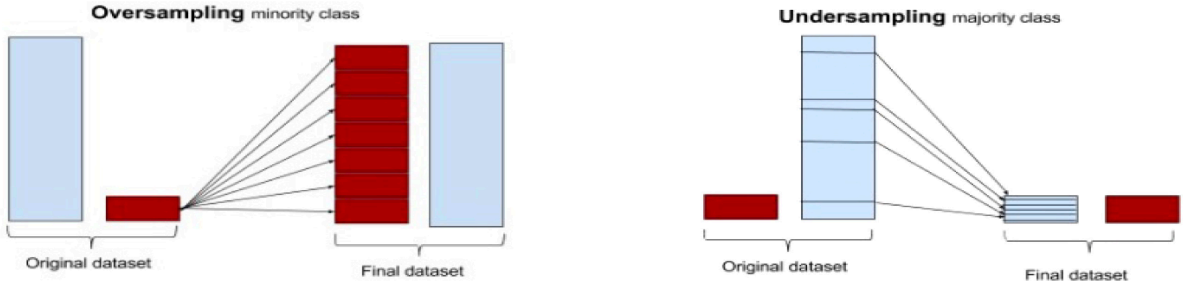
在 ‘Twitter 预警数据集’ 文档里，总结了项目背景，任务描述，数据集说明以及评价指标。

Annotation Guidelines	Examples in Dataset
1. Twitter 文本中有与 COVID19 相关症状的描述, 则被认为该数据为有关。 COVID-19 Symptoms including fever, dry cough, tiredness, aches, pains, nasal congestion, sore throat, diarrhea. Hence the first guideline is to label as positive if it describes such symptoms in the tweet. The descriptions can be formal (include keywords) or very colloquial.	I think I've caught a cold, unless it's Coronavirus...
2. Twitter 文本中涉及 COVID19 在本地区/国家确诊的。 If the tweet indicates the user or someone in his community/city/state has been tested positive for either viral test or antibody test, then this tweet should be considered positive.	"Not a hoax as #Trump said in SC rally. Illinois officials say patient has tested positive for #coronavirus https://t.co/lbgsYxqRo8"
3. Twitter 文本中涉及发现本地区有疑似 COVID19 的患者。 Someone may have COVID-19(suspect) or someone was infected or there is community spread, but not confirmed.	" A letter sent to the @PlymouthSch community warns that a student who just got back from Italy last month was hospitalized with flu-like symptoms. We're tracking this potential case of #coronavirus on @boston25 at 5 and 6:30 https://t.co/yZAswzS2dd"

图一： Twitter 预警数据集中的部分截图

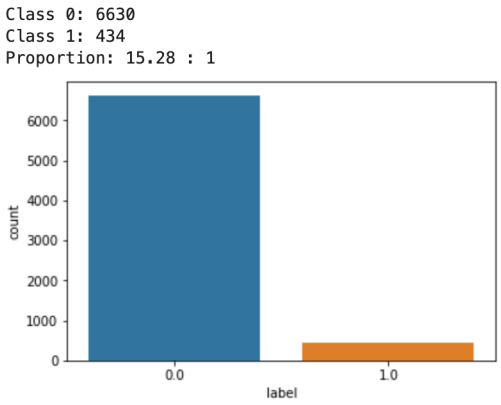
2. 重采样 (resampling)

重采样方法是使用 oversampling 和 undersampling 来使不平衡的数据集变得平衡的一种方法。示意图如图二所示。



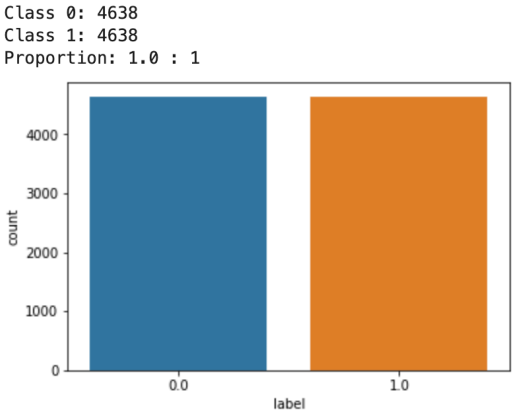
图二： 过采样和欠采样来解决数据集分类不平衡

Twitter 预警数据集中，class 0 有 6630，class 1 有 434, 比例为 15.28：1。示意图如图三所示。



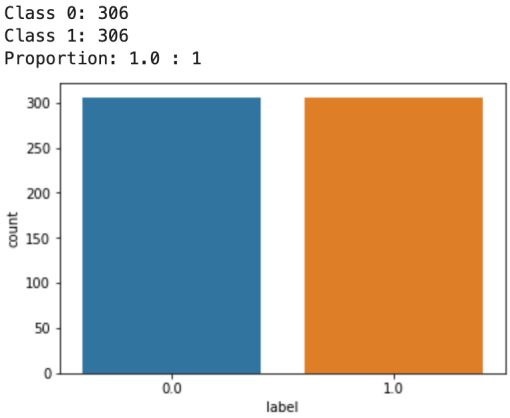
图三：Twitter 预警数据集类别分类情况

使用 Oversampling 方法，对训练集中的数据进行了过采样，采样后 class 0 和 class 1 都为 4638，类别比例为 1：1。示意图如图四所示。



图四：训练集中使用 oversampling

使用 Undersampling 方法，对训练集中的数据进行了欠采样，采样后 class 0 和 class 1 都为 306，类别比例为 1：1。示意图如图四所示。



图五：训练集中使用 undersampling

3. 实验结果

模型使用了 baseline 模型 KNN 和 SVM, 文本进行了预处理，并使用了 TFIDF 作为文本向量化方法。

3.1 在没有 resampling 处理过的数据集上表现：

	precision	recall	f1-score	support
0.0	0.95	0.99	0.97	1992
1.0	0.68	0.27	0.38	128
accuracy			0.95	2120
macro avg	0.82	0.63	0.68	2120
weighted avg	0.94	0.95	0.94	2120

图六：KNN Classification Report without Resampling

	precision	recall	f1-score	support
0.0	0.96	0.99	0.97	1992
1.0	0.64	0.37	0.47	128
accuracy			0.95	2120
macro avg	0.80	0.68	0.72	2120
weighted avg	0.94	0.95	0.94	2120

图七：SVM Classification Report without Resampling

3.2 在 over-sampling 处理过的数据集上表现：

	precision	recall	f1-score	support
0.0	0.99	0.93	0.96	1986
1.0	0.46	0.90	0.61	134
accuracy			0.93	2120
macro avg	0.73	0.91	0.79	2120
weighted avg	0.96	0.93	0.94	2120

图八：KNN Classification Report with over-sampling

	precision	recall	f1-score	support
0.0	0.99	0.99	0.99	1986
1.0	0.85	0.84	0.85	134
accuracy			0.98	2120
macro avg	0.92	0.91	0.92	2120
weighted avg	0.98	0.98	0.98	2120

图九：SVM Classification Report with over-sampling

3.3 在 under-sampling 处理过的数据集上表现：

	precision	recall	f1-score	support
0.0	1.00	0.58	0.74	1986
1.0	0.14	0.96	0.24	134
accuracy			0.61	2120
macro avg	0.57	0.77	0.49	2120
weighted avg	0.94	0.61	0.70	2120

图十：KNN Classification Report with under-sampling

	precision	recall	f1-score	support
0.0	1.00	0.80	0.89	1986
1.0	0.24	0.96	0.38	134
accuracy			0.81	2120
macro avg	0.62	0.88	0.63	2120
weighted avg	0.95	0.81	0.85	2120

图十一：SVM Classification Report with under-sampling

四、下周计划

下周可以是继续探究 resampling 方法如基于 Clustering 的 under sampling 算法 centroid k-means, 以及基于距离的 over sampling 算法 Synthetic Minority Oversampling Technique （SMOTE)。或者是使用目前已经取得较好效果的 oversampling 方法，并开始开展 RNN, CNN 模型的部分。