初步结果

|  |  |
| --- | --- |
| 任务 | 细节 |
| Jigsaw | 英文网络有攻击性语言分析 |
| Quora 类似知乎 | 英文评测问题是否真诚 |
| 20 Newsgroups | 英文新闻分类 |
| SMS | 中文短句情感分析 |
| SOUGOU | 中文网购评语分析 |

|  |  |
| --- | --- |
| 模型 |  |
| Multinomial Naive Bayes |  |
| Logistic Regression |  |
| QDA | QuadraticDiscriminantAnalysis |
| Decision Tree |  |
| Random Forest |  |
| Gaussian Naive Bayes |  |

All initial label rate = 0.9

Accuracy quoted when proportion of training size at 0.9 (showing theoretical limitation of the algorithm on the challenge for comparison with other algorithm doing the same task)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 有用吗 | Jigsaw | Quora | 20 Newsgroups | SMS | SOUGOU |
| Multinomial Naive Bayes | X (0.72) | X (0.80) | X (0.70) | X (0.61) | X (0.76) |
| Logistic Regression | X (0.84) | X (0.82) | X (0.86) | X (0.80) | X (0.88) |
| QDA | Yes (0.85) | X (0.80) | X (0.80) | Yes (0.77) | X (0.86) |
| Decision Tree | X (0.72) | X (0.71) | X (0.50) | X (0.75) | X (0.76) |
| Random Forest | X (0.74) | X (0.72) | X (0.60) | X (0.78) | X (0.80) |
| Gaussian Naive Bayes | Yes (0.66) | Yes (0.66) | Yes (0.60) | Yes (0.70) | X (0.74) |

**Uneven initial group size [ratio of target==1 << ratio of target==0]**

**Model = Logistical Regression because it is much better than most other models on a wide variety of tasks.**

**Metric: f1-score of target==1** 🡪 chosen because both recall and precision are important in this application.

**\* Initial label rate 0.9**

|  |  |  |  |
| --- | --- | --- | --- |
| Ratio of target==1 /有用吗 | Jigsaw | Quora | Sougou |
| **0.05** | **yes** | **yes when few training samples are used** | **yes** |
| **0.1** | **x** | **yes** | **x** |
| **0.5** | **x** | **x** | **x** |

**\* Initial label rate 0.5**

|  |  |  |  |
| --- | --- | --- | --- |
| Ratio of target==1 /有用吗 | Jigsaw | Quora | Sougou |
| **0.05** | **X** | **x** | **x** |
| **0.1** | **X** | **x** | **x** |
| **0.5** | **x** | **x** | **x** |

**\* Initial label rate 0.1**

|  |  |  |  |
| --- | --- | --- | --- |
| Ratio of target==1 /有用吗 | Jigsaw | Quora | Sougou |
| **0.05** | **X** | **x** | **x** |
| **0.1** | **X** | **x** | **x** |
| **0.5** | **x** | **x** | **x** |

**Conclusion**

Active learning improves f1\_score of target==1 significantly on all three binary classification tasks when using highly unbalanced datasets with high initial label rate of 0.9 (This demonstrates the value of good initialization for active learning, which AL can then improve upon). The different strategies do not perform too differently compared to one another.

**Other observations**

For the logistic regression model, the three-strategies of QueryInstanceUncertainty (least\_confident, margin and entropy) produce identical results

The Query Instance Random selection criteria, has default batch\_size of 1, which means that the model is retrained after selecting one more sample. This makes it perform much worse than ‘random\_sampling’ when selecting num\_sample \* train\_size \* initial\_label\_rate samples at one go and training the model together.