Email Categorization with learning rationales

Domain

Email categorization can automatically organizes emails into different tabs, like primary, promotion and social. In other words, users don't need to label each mail manually, but they can still check a specific category directly.

Goal

In order to improve the learning efficiency in Email categorization, there are two approach I use. First, learning with rationale, which I checked domain knowledge first then provide rationales for each labeled email. Second, using active learning, which I adapt uncertainty sampling to pick up the most uncertain email. I will explain more detail in next section.

Experiment

In this section, I describe the datasets, classifiers and approach, used for the experiments and explain the approach for providing rationales more detail.

Dataset

I used **imaplib** library to collect emails from my own two mailbox, my private mailbox and school mailbox. The interval date I choose are from 8/27 to 10/27 for private mailbox and 1/27 to 10/27 for school mailbox, then filter non-English emails. And I checked my gmail category and labeled each email in 3 categories, primary, promotion and social, which represents 0,1, and 2 respectively (Figure 1.)

```
print(New_dataset.groupby('Label').count())
```

Label	ID	From	Date	Subject	Email
0	751	751	751	751	751
1	571	571	571	571	571
2	456	456	456	456	456

Fig.1 Class distribution

According to Figure 1, my original dataset structure contains 5 columns. “Email” is “From” appending “Date”, which is my main text analysis goal. Features are constructed by CountVectorizer in binary. I also choose **nltk.stem import WordNetLemmatizer** to be stemmer analyzer.

Approach

I used multinomial naive bayes to be my classifier , then chose mean accuracy to be my performance measure. To evaluate my rationale method, I compare the performance between traditional learning and active learning.

Regarding rationales sources, I construct rationale dataset for each categorization. Then I checked similar rationales in each categorization to put in rationale dataset.

According to active learning with rationales, I applied uncertain sampling. To explain further, I used predict_prob to find out 5 most

uncertain emails. Then I emphasize rationale feature weight(r =10) and de-emphasize un-rationale feature weight(O =1) for these emails

Result

Regarding Figure 2, there is the huge difference between Random sampling with rationale and without rationale. It proves that providing rationales can get good performance more efficiency. However, in active learning case, both of learning models get the same accuracy, which means it can't increase the efficiency in active learning. Interestingly, we can compare the difference between two groups, AL_Lwo_R, RS_Lwo_R and AL_Lw_R, RS_Lw_R. And we can find out that using active learning can get better efficiency than using rationales.

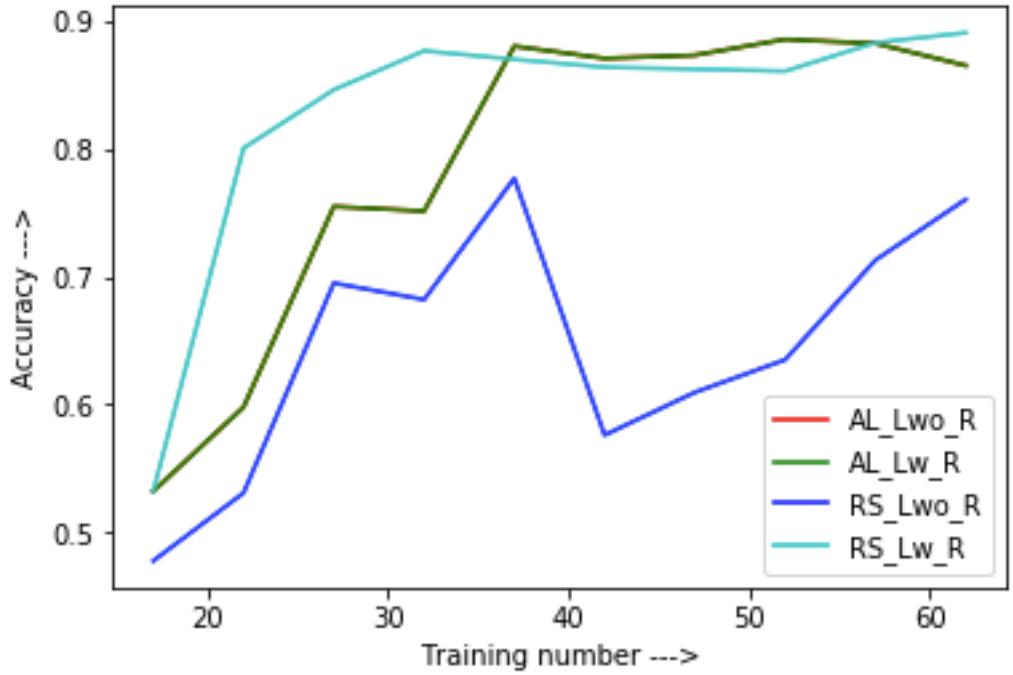

Fig.2 Accuracy with CountVectorizer

Conclusion

I successfully implement rationales in email categorization in active learning and traditional learning. It is really clear that learning rationales can improve the efficiency when training learning models. Regarding the future work, I can put more different mailbox and keep finding more interesting rationales.