Incorporating Concept Drift in Recipient Recommendation Systems

Himangshu Ranjan Borah
hborah@ncsu.edu

Sushma Ravichandran
sravich@ncsu.edu

Sukriti Sharma
ssharm18@ncsu.edu

Dataset

The dataset used is a corpus of emails from ENRON Corporation available here. The original dataset was cleaned by a machine learning research community at Carnegie Mellon University to have a total of 200,399 messages belonging to 158 users.

Project Idea

Recipient recommendation has become very crucial in e-mail applications recently with its tremendous growth. According to [1], recipients can be recommended given either of the two sets of inputs: First, when the message contents are available, the fields predicted are TO, CC and BCC. Second, when the message contents and the TO- addresses are available, the fields predicted are CC and BCC. An Address Book is created for each user based on his activities which is later used for the prediction. We look forward to using novel features from areas of Text Mining and Natural Language Processing to effectively represent the unstructured content of emails like body and subject.

Concept Drift occurs when data evolves over time in unforeseen ways. In paper [2] the authors detect Concept Drift on the language models of emails and use it for spam filtering. As the Address Book of a user continuously varies, we propose to apply the notion of concept drift in recommending recipients and study its effects on the same with appropriate feature set and ensemble of classifiers.

Softwares

The architecture of the learning system will be designed in MATLAB. The code needed for parsing and preprocessing the raw email data has to be written from scratch. After the feature space is finalized according to the problem domain, subsequent learning algorithms will be implemented.

Work Division till Mid term

- **Himangshu :** Parsing structured features, Understanding and formatting folder structure.
- **Sushma :** Parsing un-structured features, Feature Subset Selection.
- **Sukriti :** Cleaning the data, Feature Transformations.

Midterm milestone

We plan to clean the dataset to find a subset of the same to suit our problem domain, parse the raw email data to extract structured and unstructured feature sets and preprocess. Initial testing results will be prepared for the different existing classification models.

References

[1] Vitor R. Carvalho and William Cohen. Recommending recipients in the Enron email corpus. *Technical Report CMU-LTI-07-005*, Carnegie Mellon University, 2007

[2] M. Hayat, J. Basiri, A. Shakery, Content-Based Concept Drift Detection for Email Spam Filtering, *5th International Symposim on Telecommunications*, IEEE (2010)