CMPE 59H - Project

Probabilistic Topic Modeling for the Analysis and Classification of Genomic Sequences

Gönül Aycı & Dilara Keküllüoğlu

Computer Engineering, Boğaziçi University
Instructor: Assoc. Dr. Arzucan Özgür

December 28, 2017

Outline

Introduction

Processing Data

Classification Methods

Results

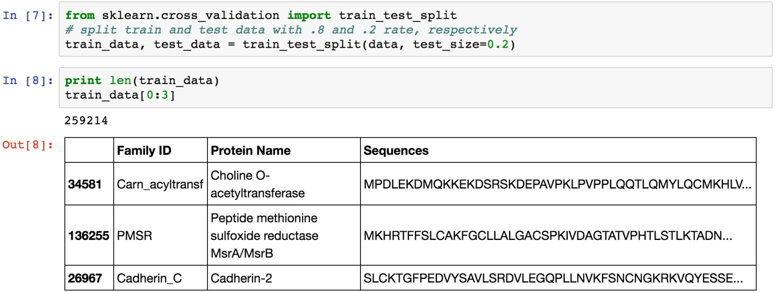
Problem Description

- To classify **DNA sequences**, we need to find similarities between them.
- This paper uses **NLP topic models** in the bioinformatics context to solve this problem.
    - For example we can find **author of a novel** by using topic models.
    - Training set is used for learning these **abstract topics** and search them in a new document to attribute an author.

Processing Data

- Downloaded **Asgari data** (from word2vec presentation),
- Reduced size from
  - 360K sequences 7K families,
  - 50K sequences and 100 families
- Split data with the **rate of 8:2** to training and test dataset,
- Create **two kinds of files**
  - Sliding window,
  - Discrete window

Sample code



In [7]:
```
from sklearn.cross_validation import train_test_split
# split train and test data with .8 and .2 rate, respectively
train_data, test_data = train_test_split(data, test_size=0.2)
```

In [8]:
```
print len(train_data)
train_data[0:3]
```

```
259214
```

Out[8]:

	Family ID	Protein Name	Sequences
34581	Carn_acyltransf	Choline O-acetyltransferase	MPDLEKDMQKKEKDSRSKDEPAVPKLPVPPLQQTLQMYLQCMKHLV...
136255	PMSR	Peptide methionine sulfoxide reductase MsrA/MsrB	MKHRTFFSLCAKFGCLLALGACSPKIVDAGTATVPHTLSTLKTADN...
26967	Cadherin_C	Cadherin-2	SLCKTGFPEDVYSAVLSRDVLEGQPLLNVKFSNCNGKRKVQYESSE...

LDA : Pair topics with families

- Create **lda model** with train data,
- Assign topic that has the **most probability** to each sequence,
- Investigate all sequences in a family,
- Find the most frequent topic in a family,
- Assign such **topics to each family**,
- Use this pairs to assign families to test data

LDA SVM: Classify with feature vectors

- Create **lda model** with train data,
- Create **feature vectors** with topic probability distribution,
- Feature vector size is the topic size (e.g. 20, 100, 200),
- Train **SVM model** with training sequences,
- Create feature vectors of test data according to lda model.

LDA W2V: Classify using Word2Vec

- Create **lda model** with train data,
- Create feature vectors with **topic probability distribution and Asgari W2V data**,
- Feature vector size is the topic size (e.g. 20, 100, 200) + 100 (W2V word embeddings),
- Train **SVM model** with training sequences,
- Create feature vectors of test data according to lda model and Asgari word embedding

Analysis of Results

- Running with ***3, 5,*** and ***8 - mers***
  - Unfortunately, **complexity problems** for 5 and 8
- We generally got **baseline results (random chance)**

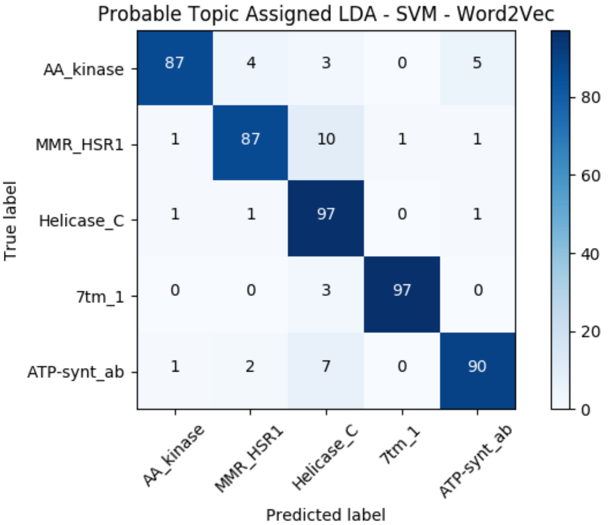
Accuracy of 5 family classification

  - **LDA** 31%,
  - **LDA** with **SVM** : 28%,
  - **Word2Vec**: 95%
  - **LDA** with **Word2Vec**: 92%,

Not get better results for

- Using **PCA**,
- Applying **sliding window**,
- Setting # of topics > # of family

LDA + W2V Best Result Confusion Matrix



Probable Topic Assigned LDA - SVM - Word2Vec

True label \ Predicted label	AA_kinase	MMR_HSR1	Helicase_C	7tm_1	ATP-synt_ab
AA_kinase	87	4	3	0	5
MMR_HSR1	1	87	10	1	1
Helicase_C	1	1	97	0	1
7tm_1	0	0	3	97	0
ATP-synt_ab	1	2	7	0	90

Numerical Results

k-mer	# of family	# of topic	lda	lda+svm
3	20	100	.07	.15
3	10	200	.11	.14
5	10	100	.12	.13
3	5	500	.31	.28

Table: Accuracy results for sliding window data

C-value	# of family	# of topic	lda	lda+svm	lda+w2v
1	5	100	.23	.28	.74
100	5	100	.23	.25	.92

Table: Accuracy results for discrete window data

References

- https://github.com/aycignl/probabilisticTopicModeling
- Asgari, Ehsaneddin, and Mohammad RK Mofrad. Continuous distributed representation of biological sequences for deep proteomics and genomics. PloS one 10.11 (2015): e0141287.
- La Rosa, Massimo and Fiannaca, Antonino and Rizzo, Riccardo and Urso, Alfonso, Probabilistic topic modeling for the analysis and classification of genomic sequences. BMC bioinformatics, BioMed Central, 16, 6, S2, 2015.

Thank you ...

Any Questions?