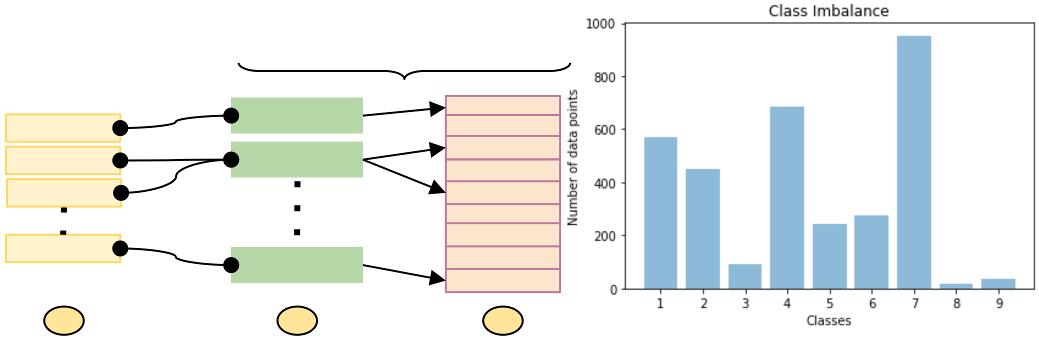
Personalized Medicine: Redefining Cancer Treatment

Problem Description:



(a) (b)

Figure 1 (a). Experts classify each mutation by going through corresponding research text (b) Class imbalance in training data

- Experts classify mutations based on clinical evidence (research texts)
- A particular research text has been used to classify multiple mutations as shown in (2)
- Training data has huge class imbalance

Solution Setup:

Data Transformation:

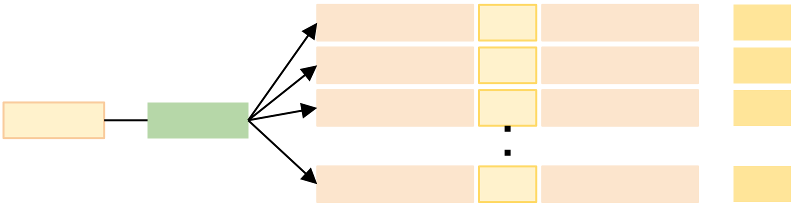


Figure 2. As each research text can be used to classify many mutations, only the relevant parts of these texts are important. We therefore extract a 15 word context window around each occurence of a gene variation and store it as a new document

- Search for all *instances* of “gene variations” in corresponding research text
- Define a context window of 15 words around each *instance* and store these as new documents
- This will serve as the new training data

Vector Space Representation using Doc2Vec:

We need some way of representing these documents in a vector space, for which we choose Doc2Vec technique, which is an extension of the popular Word2Vec. The model we chose is PV-DM as it learns the word vectors along with document vectors and takes into account the order of words in a small window. The resulting 100 dimensional document vectors are used to train a SVM classifier.

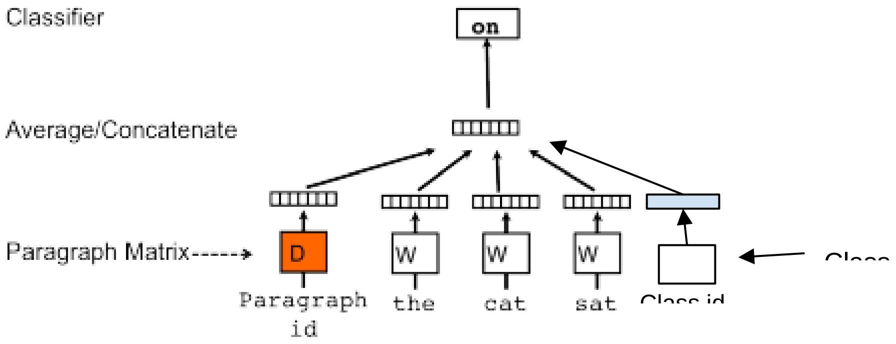
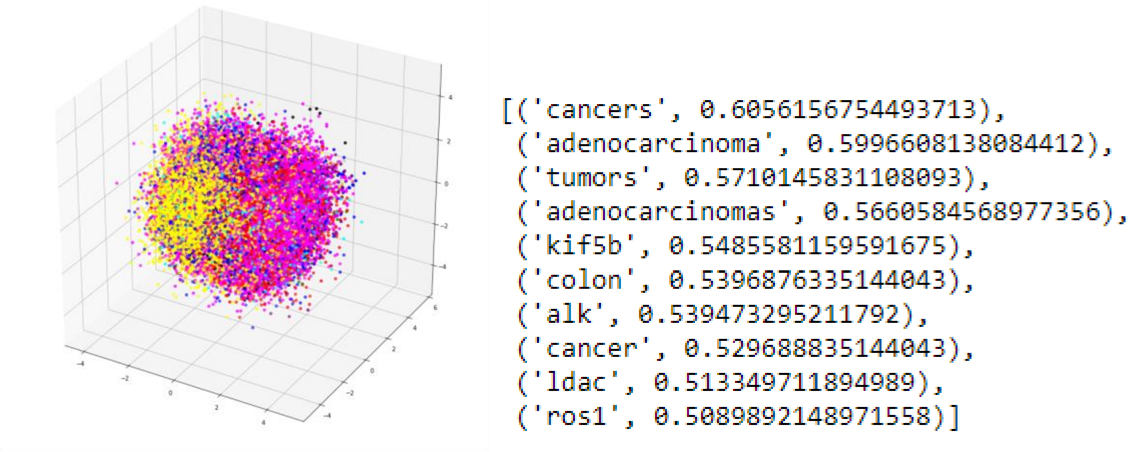


Figure 3. Doc2vec PV-DM model with a slight modification to include a class id along with paragraph ID



```
[('cancers', 0.6056156754493713),
 ('adenocarcinoma', 0.5996608138084412),
 ('tumors', 0.5710145831108093),
 ('adenocarcinomas', 0.5660584568977356),
 ('kif5b', 0.5485581159591675),
 ('colon', 0.5396876335144043),
 ('alk', 0.539473295211792),
 ('cancer', 0.529688835144043),
 ('ldac', 0.513349711894989),
 ('ros1', 0.5089892148971558)]
```

(a) (b)

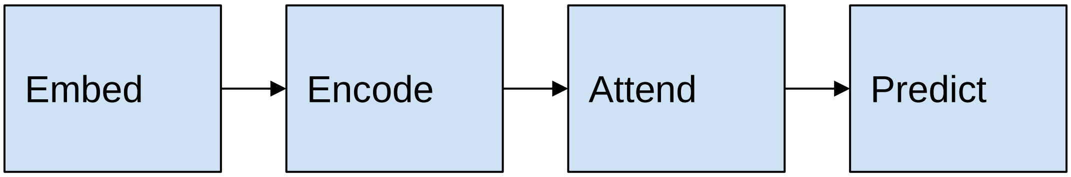
Figure 4. (a) Document vectors in 3D space (dimensionality reduced with PCA) (b) Words most similar to ‘lung’

Sr. No	Tasks Accomplished	Description
1	Document Cleaning	Tokenization, lemmatization, stemming, and stop word removal from texts
2	Region of interest extraction	Extract the context around mutations mentioned in a research document
3	Represent documents in vector space	Used gensim’s implementation of doc2vec
4	Visualize vector space and word embeddings	
5	Initial Baseline Model	SVM + doc2vec for initial baseline model

Going Forward:

Intuitively it seems that the context in which a mutation is mentioned in a research text would be of primary importance in determining the class of that particular mutation. RNN model sequences quite well and their variants have achieved state-of-the art performance on various

NLP tasks. We will implement a Bi-directional LSTM with attention model with the well known architecture.



The top challenges would be:

- How to select relevant pieces of text from research articles
- All sentences that are extracted as context do not have the same amount of predictive power, we would need a mechanism that attends relevant pieces of text
- In some research articles the mutation in question is not even mentioned once. It would be very challenging to correctly classify such records.
- There is a paper titled “Hierarchical Attention Networks for Document Classification” that could be key to solving this problem. It uses a hierarchical architecture mirroring the structure of documents while using attention mechanism first at the word level and then at the sentence level[1].

[1] "Hierarchical Attention Networks for Document Classification." http://www.cs.cmu.edu/~./hovy/papers/16HLT-hierarchical-attention-networks.pdf. Accessed 11 May. 2018.