Note: I just got the news only two days before the end line (that's my fault), pardon me if this document is too short. **If needed, I can write a more specific version.** After reading it, if there's any question, don't be hesitated to get in touch with me.

Combined Vocabulary: A Fast Approach for Genetic Variants Classification based on Machine Learning

Wudi Fang

1 Brief Description

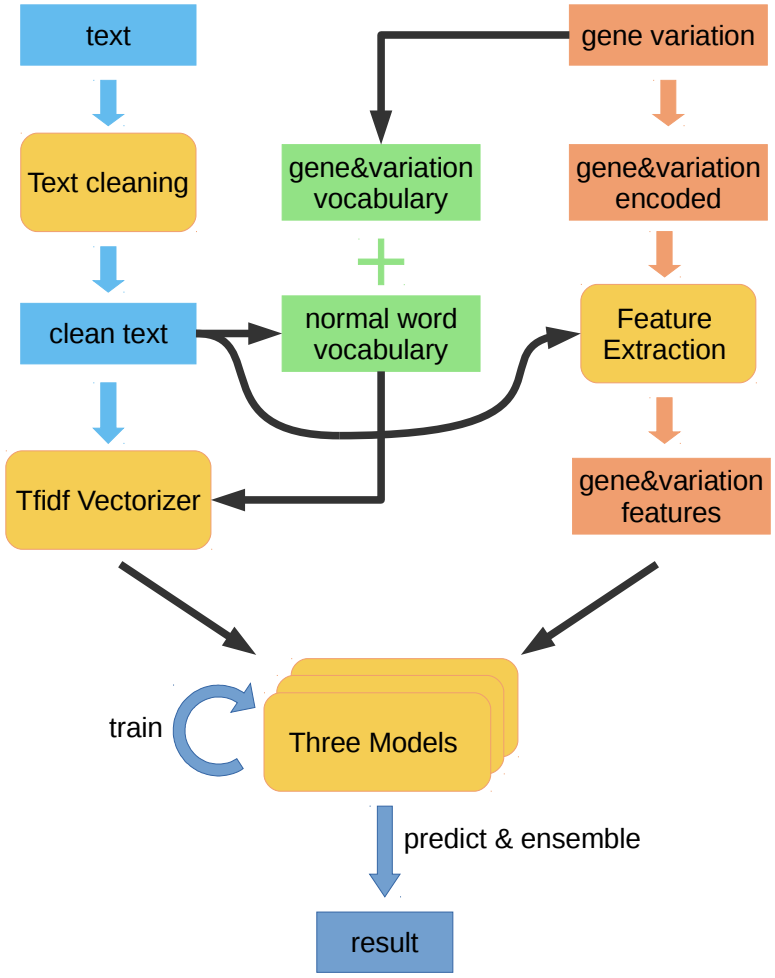

Figure 1. Solution scheme

As shown in Figure 1, this solution scheme takes text (one document for each variation) and gene variation list as inputs. After text cleaning and gene&variation encoding, gene&variation vocabulary and normal word vocabulary are generated to help doing tf-idf vectorizing. Meanwhile, gene&variation encoded information are used to assist feature extraction from text. Afterwards, all features are sent to three models for training and predicting. In the end, predictions will be averaged and output as the result.

Three main novel approaches will be introduced as below.

2 Novel Approaches

2.1 combined vocabulary

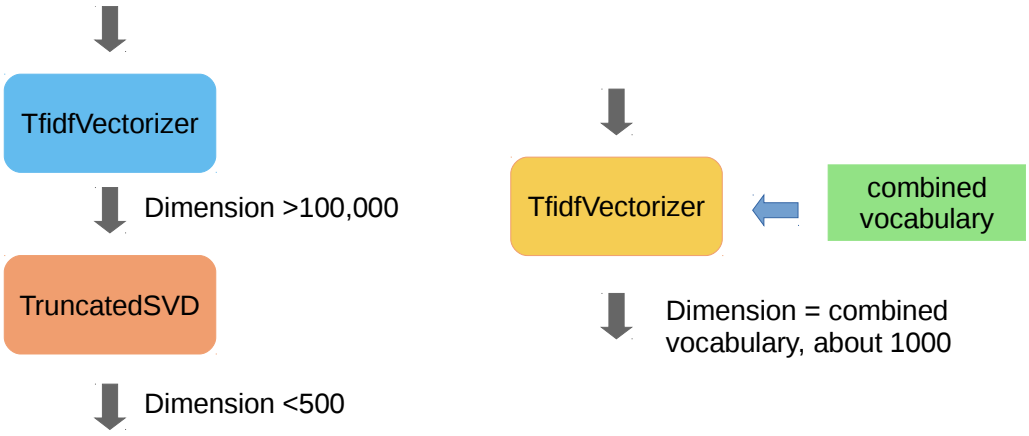

Figure 2. Comparison in text feature engineering

For text feature engineering, traditional way has two major drawbacks: very slow and space consuming. Because it need to use all the words in the corpus as feature dimension and then use SVD to reduce dimension.

In this article, combined vocabulary method was proposed. This method use medium-scale (e.g. 1000) vocabulary. to process TfidfVectorizer. (1) It can be at least dozens of times faster than traditional way. (2) It can also preserve more dimensions than traditional way. (3) Combined vocabulary. mentioned here is very feasible to customize, which means, experts with domain knowledge (e.g. researcher in medical field) can expend the vocabulary.

Combined vocabulary. has two sources: gene variation and train corpus. (1) gene&variation vocabulary. is composed of gene words and variation words. (2) normal word vocabulary. comes from train corpus in a tricky way as follows.

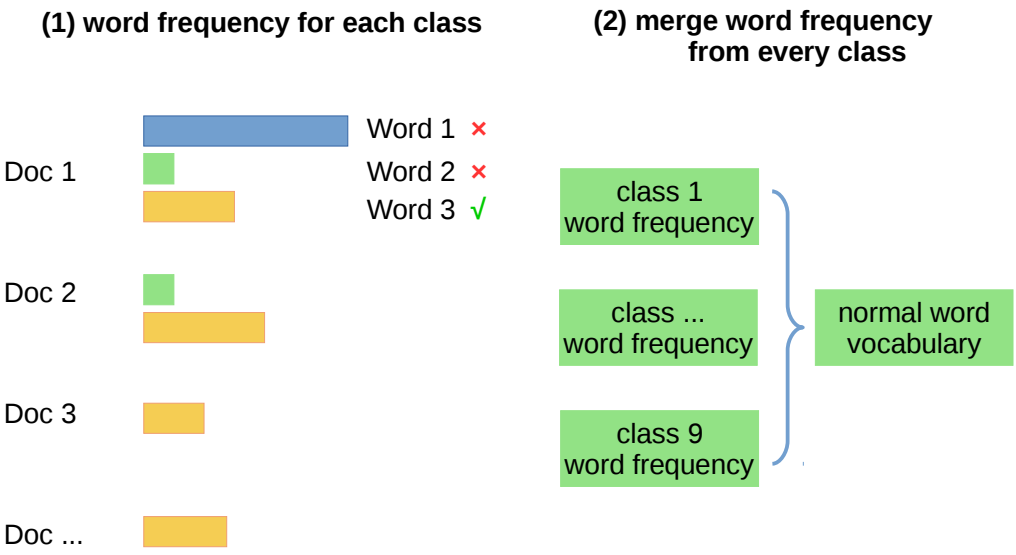

Figure 3. Building normal word vocabulary.

As shown in Figure 3, first step is counting word frequency in documents and then build a word frequency table for each class. During the process, words in these two cases will be deleted: (1) word frequency is high but only crowded in limited several documents; (2) word frequency is lower than 20% of document number in the class, which means at most 20% of the class's documents have this word. These two deletions are based on the main idea that documents belong to the same class should share some words which are meaningful for classification. The second step is merging those word frequency tables from every class into one word vocabulary.

2.2 deep feature engineering for variation

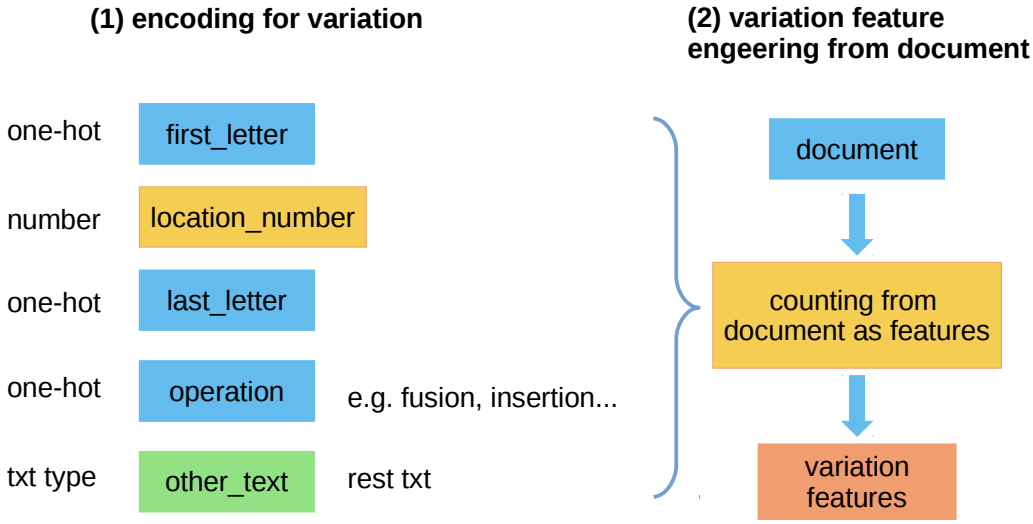

Figure 4. Feature engineering for variation

In feature engineering work, variation has been paid enough attention to. As shown in Figure 4, variation is split into several parts which is encoded accordingly. And then, encoded information is used to counting from document and generating features.

2.3 class imbalance compensation

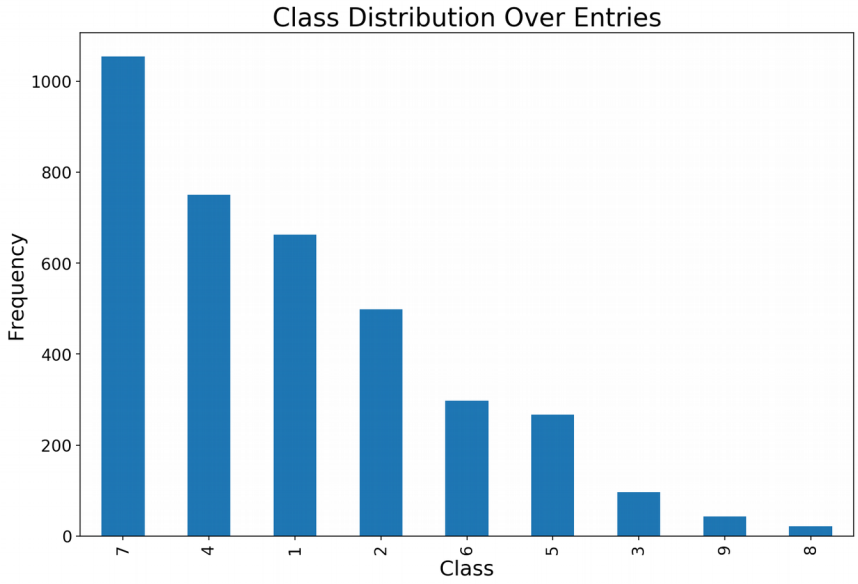

Figure 5. Class imbalance graph

As shown in Figure 5, class imbalance problem is very obvious, which can lead to model bias because of class probability distribution. To relieve this problem, SVM model was added into ensemble model selections (original ones are xgboost and lightgbm model). That's because SVM model can set class weight to balance in multi-class classification task while xgboost or lightgbm can't. Model ensemble selection seen in Figure 6.

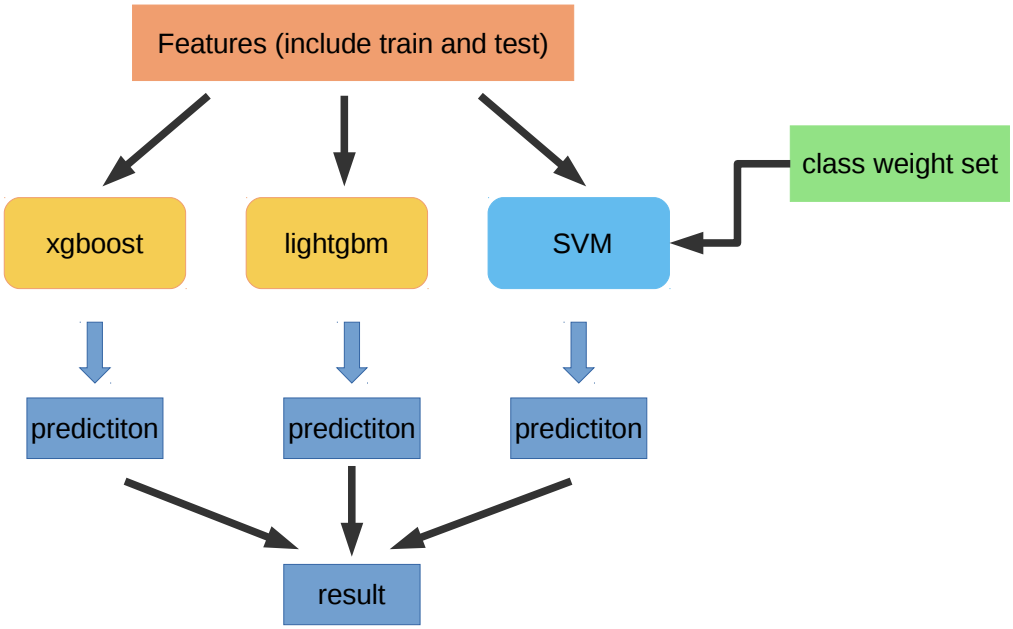

Figure 6. Model ensemble selection

3 Solution Details

3.1 carefully clean for text

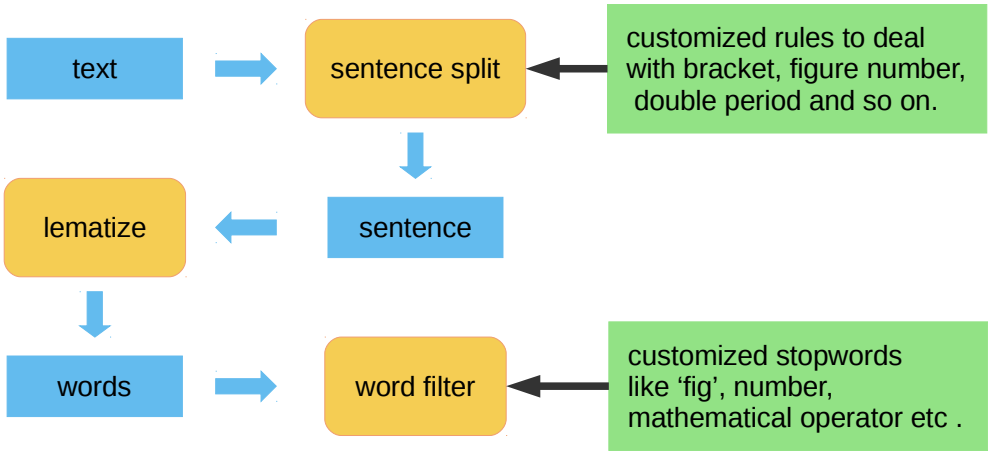

Figure 7. Text clean

As shown in Figure 7, during the process of sentence split, beside normal tokenizer, customized rules should be added to deal with those sentences unable to split correctly. Also, beside normal English stop words, customized stop words are added into the word filter.

3.2 fine tune model hyper parameters

During the process of model training, 5-fold cross validation was used to fine tune hyper parameters of the models, like ‘eta’, ‘max_depth’, 'max_delta_step', ‘num_leaves’, ‘kernel’ etc. In the end, these parameters are adjusted to a proper range or value. Details can be seen in the code.

4 Solution Result

In use of this solution, validation loss reached about 0.8 both in stage 1 and stage 2. Public score of test loss in stage 1 reached about 0.5. This method maintains the high speed in training, and also keeps the generalization ability.

A link to my code: **https://github.com/fangwudi/solution-for-kaggle-Redefining-Cancer-Treatment.git**