Final Project Report

Paper implementation of:" Protein classification using modified n-gram and skip-gram models"

Student: Zhen Liu
NID: zh116855

1. Paper Background

Using machine learning method to do classification of protein characteristics from their primary sequence has make great contributions under the condition of the primary sequence provides the blueprint which encodes the purpose of the protein, ultimately determining the proteins characteristics, functions, subcellular localization and interactions. Also, there are variety of classification models based on machine learning method have been proposed and bring huge benefits to many fields. According to this background, the step of feature generation become more and more important in the whole process. However, lack of professional knowledge may risk selection of irrelevant features. So, researchers proposed lots of generalized and data-specific feature extraction methods based on natural language processing to this problem. But, the classification machine learning model depends heavily on training data, feature extraction, classifier algorithm and optimization. As we know, n-gram method combined with skip-gram model is a powerful feature extraction method to many fields and it's also very valuable in biomedical area.

Under this condition, this paper proposed a modified k-skip-n-gram model which is fully automated and agnostic to peptide function or chain size.

2. Method of Paper

2.1 Dataset

In the meta-comparison step, the n-gram and skip-gram model(m-NGSG) has been compared with several models based on the availability of benchmark data reported by those models. Which means, when m-NGSG is compared with different models which may aim to handle specific classification problem, the dataset will be different too as the *figure 1* shown.

iAMP-2L	One pair of training and test dataset of two classes of protein sequences based on antimicrobial activity
Cypred	One pair of training and test dataset consist of two classes of sequences based on cyclic and noncyclic structure.
PredSTP	One training dataset divided into two classes based the cysteine bonding pattern in the 3D structure of the proteins.

Figure 1 Example of datasets used in this paper

The dataset format is ‘fasta’ which is a text-based format for representing nucleotide sequence or peptide sequence. It also allows sequence names and comments to precede sequences. *Figure 2* is the example of ‘fasta’ format file.

```
>VIR|P82889|CAPSD_BP
MDFNPSEVASQVTNYIQAIAAAGVGVLALAIGLSAAWKYAKRFLKG
>VIR|P03619|CAPSD_BP
MKKSVVAKIIAGSTLVIGSSAFAADDATSQAKAAFDSLTAQATEMSGYAWALVVLVVGAT
>VIR|P03620|CAPSD_BP
MRVLSTVLAAKNKIALGAATMLVSAGSFAAEPNAATNYATEAMDSLKTQAIDLISQTWP
>VIR|P03621|CAPSD_BP
MKAMKQRIAKFSPVASFRNLCIAGSVTAATSLPAFAGVIDTSAVESAITDGQGDMKAIGG
>VIR|P03623|CAPSD_BP
MQSVITDVTGQLTAVQADITTIGGAIIVLAAVVLGIRWIKAQFF
>VIR|P15794|CAPSD_BP
MRSFLNLNSIPNVAAGNSCSIKLPIGQTYEVIDLRYSGVTPSQIKNVRVELDGRLLSTYKT
>VIR|P22535|CAPSD_BP
MAQVQQLTPAQQAALRNQQAMAANLQARQIVLQQSYPVIQQVETQTFDPANRSVFDVT
>VIR|P85987|CAPSD_BP
MAAYQTYTMAGIKEDFADWVSNISPEYTPLISMIRKFPVHNTMFQWQWDVLKDVDTE
>VIR|P85989|CAPSD_BP
MSKKLVTEEMRTQWLPVLEKKSEQIQPLTAENVSVRLLQNQAEWNAKNLGESEGPSSV
>VIR|P03622|CAPSD_BP
SGVGDGVDVVSAIEGAAGPIAAIGGAVLTVMVGIKVYKWVRRAM
>VIR|P03614|COAT_BPF
MASNFEEFVLVDNGGTGDVKVAPSNFANGVAEWISSNSRSQAYKVTCSVRQSSANNRK
>VIR|P49861|COAT_BPH
MSELALIQKAIEESQQKMTQLFDAQKAEIESTGQVSKQLQSDLMKVQEELTKSGTRLFD
>VIR|Q04754|COAT_BPL
MTVVLDSKDLARIDEEYKADSQVWSYLTGGNGVTQRFRGHNEVRINKLSGFVDATAYK
>VIR|P85500|COAT_BPP
MKTNRAYSTLEVKALDDEKRVITGIASTPSPDRMQDVVEPKGAQFKLPIPFLWQHNHDEP
>VIR|P03630|COAT_BPP
SKTIVLSVGEATRTLTEIQSTADRQIFEEKVGPLVGRLRLTASLRQNGAKTAYRVNLKLDQ
```

Figure 2 'fasta' format

The data preprocessing step will include extracting different class tags and building feature index. The class type is different between datasets, like protein structure, function, protein name and etc. And the class type information will be recorded in the preceding comments before every sequences. For building feature index, we can use NLP method, such as one-hot vector. For example, we can create a 400-length feature vector to record the 2-gram features.

2.2 Basic Method of n-gram & skip-gram

As we know, n-gram and skip gram model are very basic feature extraction methods and have been used in many fields. So, the concept of these two models are easy to understand by the given example: we have a protein sequence MISHW. All 2-gram will be MI, IS, SH, HW and 1-skip-bi-gram will be MI, IS, SH, HW, MXS, IXH, SXW.
N-gram motifs provide information in protein functionality which can be represented as GM_p^s. s is a positive integer not longer than the length L, p is the permutation index. In order to avoid duplicating features extracted with m-gram, it exclude the motifs produced: SM_p^b. b is the number of skips between two amino acids.

The number of feature will be affected by those parameters in two models and the skip rule.

2.3 Modified Method

About the modification for the k-skip-n-gram model, there are two main parts. First, this model allows buffering on the number of skip: SM_p^c. $c = b + ((a - b)\%a)$. Second

is adding buffering relative position of motifs with respect to C-terminus which is the end of protein transcription: $SM_p^c, (x; y)$ which $x = z + ((y - z)\%y)$. y determines the positional buffering parameter, z is the relative position to C-terminus (end of the protein transcription).

As an example of the relative position of motifs with respect to C-terminus, if we consider NTerm-AYHGFTVCKY-CTerm as a protein sequence, then two tyrosine will be members of the set of uni-gram motifs and should be considered as identical. However, if we choose to account for position, each will be assigned position identity information as defined by equation (5). If the initial buffer value y 0 equals 5 then the positional identity of the first Y and the last Y will be x = 9 + ((5 − 9) %5) = 10 and x = 1 + ((5 − 1) %5) = 5, respectively. Here the distance of first Y is 9 and the second Y is 1 from the C-terminus. In this way, rather being identical, the tyrosine will be recorded as Y10 and Y5 in the feature set. This approach can be generalized to n-grams. The bi-gram AY has a positional identity of 12, because its onset is 10 residues away from the C-terminus, and the buffer value will be 6 because y 0 is 5 and the length of the motif is 2. All of these parameters are shown in *figure 3* and will be optimized during validation step.

n	determines the maximum length of an *n-gram* motif
k	determines the maximum number of skips in a *k-skip-bi-gram* motif
np	determines the maximum length of an *n-gram* motif that gets a positional value
kp	determines the maximum skips in a *k-skip-bi-gram* motif that gets a positional value
y	determines the positional buffering parameter in both *n-gram* and *k-skip-bi-gram* motifs
c	determines the skip buffering parameter in *k-skip-bi-gram* motifs

Figure 3 Model parameters

After those above processing, feature vectors will be generated from protein sequence. However, before inputting these features into classifier, we need reduce noise by removing those words that make up more than 30% of the corpus and appears less than 3 times.

2.4 Classification

This paper using different classifiers when compared with different models. Two main classification model it used are logistic regression and SVM. Also, it did Jack-knife, 5-fold and 10-fold cross validation and used grid search method to optimize parameters.

3. Implementation

3.1 Structure

The process of implementation can be summarized as these main steps:

(1) Find benchmark data which is used in this paper.
(2) Data preprocessing.
(3) Implement basic n-gram & skip-gram feature extraction method.
(4) Implement modified n-gram & skip gram method.

(5) Train classification model and do 10-fold cross validation.
(6) Implement grid search method to optimize parameters.
(7) Try some improvements.

3.2 Method and Insight from Implementation

The dataset I used comes from the model called PvPred which is mentioned in this paper. The dataset is in 'fasta' format which I have introduced above and it contains two class: 'VIR' and 'non-VIR'. The description of every protein sequence precedes corresponding sequence. So, the first step is extracting the class of every sequence.

Then I implemented the basic 2-gram and 1-skip-bi-gram model and using grid search method to get the best parameter of SVM classifiers during validation.

Next step is implementing buffer on the number of skip, doing denoise process to feature vectors and also using grid search method to get the best parameter of SVM classifiers during validation.

3.3 Results

After implementing all these steps, I got some results to compare with the original and modified NGSG model. We can observe that the performance of skip-buffering model has been already better than the original model. Also, the buffering model with relative position to C-terminus is 0.7833 which is same with non-pos version. And it is still lower than the modified NGSG model about 7%.

I think the main reason for this result is that the dataset is not big enough to test this model. When I test my model using independent dataset, the accuracy can reach more than 0.95.

The first table shows the results of several models and the second table shows the results of the optimization of parameters.

Method	Accuracy
Basic 2-gram	0.7167
k-skip-bi-gram	0.7667
Skip-buffering NGSG	0.7833
Buffering NGSG with pos	0.7833
Original method(PvPred)	0.7719
m-NGSG	0.8502

Parameter	Optimized number
Buffer	1
Denoise threshold	2
Cost penalty	5
Kernel	Poly
Split size	0.2
gamma	0.001

4. Conclusion

Though the result is still a little bit lower than the model in this paper, through the implementation I got fully understand the process and details of this model. Also, during the implementation, I tried to combine other feature extraction method for peptide sequence, such as Information Theory, mixed-gram. However, I didn't get a higher score than the model I introduced.

This method improves the performance of traditional n-gram and skip-gram method with the concept of buffer and motif are tagged with relative position to the C-terminus. This method could reduce the work at feature extraction step even regardless of sequence size. It will benefit the machine learning –based protein classification community, especially those focus on protein classification based on primary protein sequence.

5. Further Work

According to the main limitation of my model, I need to find more dataset to test the performance of my model and compare it with the paper.

During the process of implementation, we find that fragment-based models provide a slightly better accuracy than the m-NGSG model. Because the dataset is beyond the scope of demonstrating a generalized feature extraction method on whole sequence.

According to the limitation of m-SGNG method under the condition when we need to generate feature in the whole sequence, we can modify this model in some directions, such as modify parameter, add more parameters.