Journal of Computer and Communications, **2013, 1, 9-13**
http://dx.doi.org/10.4236/jcc.2013.15002 Published Online October 2013 (http://www.scirp.org/journal/jcc)

Protein-Protein Interaction Extraction Based on Convex Combination Kernel Function

Peng Chen, Jianyi Guo, Zhengtao Yu, Sichao Wei, Feng Zhou, Xin Yan

[1]The School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China; [2]Key Laboratory of Intelligent Information Processing, Kunming University of Science and Technology, Kunming, China.
Email: gjade86@hotmail.com

Received August 2013

ABSTRACT

Owing to the effect of classified models was different in Protein-Protein Interaction (PPI) extraction, which was made by different single kernel functions, and only using single kernel function hardly trained the optimal classified model to extract PPI, this paper presents a strategy to find the optimal kernel function from a kernel function set. The strategy is that in the kernel function set which consists of different single kernel functions, endlessly finding the last two kernel functions on the performance in PPI extraction, using their optimal kernel function to replace them, until there is only one kernel function and it's the final optimal kernel function. Finally, extracting PPI using the classified model made by this kernel function. This paper conducted the PPI extraction experiment on AIMed corpus, the experimental result shows that the optimal convex combination kernel function this paper presents can effectively improve the extraction performance than single kernel function, and it gets the best precision which reaches 65.0 among the similar PPI extraction systems.

Keywords: Protein-Protein Interaction; Support Vector Machine; Convex Combination Kernel Function

1. Introduction

That the presented protein entities exist interaction relation of the article in biomedical field is called Protein-Protein Interaction. PPI is vital for knowing the single proteins and the organization of the entire biological process. Recently, PPI becomes an important and hot task in biomedical field. There is a surge of research interest in PPI.

Early studies on PPI extraction mostly employ the machine learning method, it includes the convolution kernel based method [1-3] and feature based method [4-7]. The main difference of these two methods is the way they get their high dimension matrixes. The convolution kernel-based method expresses the sentence as string [2], tree [1], graph [3] or other structured ways, and determine the high dimension matrix by counting the number of same substructure in two sentences. The feature based method needs select features firstly, Miyao, Y. *et al*. [5] employed three syntactic analysis methods and made syntactic information as the feature, Liu B *et al*. [6] used dependent information as features, Bui Q-C *et al*. [7] used keywords as the feature. When features have been determined, the method expresses the sentence as feature vector, then gets the high dimension matrix mapping by common kernel functions. For the high computation complexity, the convolution kernel based method or the composite method which includes the convolution kernel-based method is not adaptive in the practical PPI extraction system. The feature based method which can get the high dimension matrix at a high rate of speed becomes the mainstream. What calls for special attention is that in the feature-based method, the high dimension matrix is the only information in getting the classified model by training, so the selection of the kernel function is crucial in PPI extraction. Niu Y *et al*. [4] and Bui Q-C *et al*. [7] respectively used a linear kernel function and radial basis function (RBF Kernel) as the kernel function. The effect of classified models was different in PPI extraction, which was made by different single different kernel functions, and only using single kernel function hardly trained the optimal classified model to extract PPI.

For the above problem, this paper presents a strategy that finds the optimal kernel function from a single kernel function set consists of some single kernel functions. On the basis of corpus pretreatment, selection features and getting high dimension matrixes by mapping from different single kernel functions, then achieving their classified models by training. The next step is that endlessly finding the last two kernel functions on the performance in PPI extraction, using their optimal kernel

function to replace them, until there is only one kernel function and it's the final optimal kernel function. Finally, extracting PPI using the classified model made by this kernel function. This paper conducted the PPI extraction experiment on AIMed corpus, the experimental result shows that the optimal convex combination kernel function this paper presents can effectively improve the extraction performance than single kernel function.

2. The PPI Extraction Method Based on Convex Combination Kernel Function

2.1. Preprocessing

Protein entity recognition is the basis of the PPI extraction work. This paper uses AIMed corpus which has been labeled protein entities, and therefore don't consider issues related to protein entity recognition. We have to pretreat AIMed corpus before select features.

- In order to avoid the interference of PPI extraction by the protein text information, we use PROT1, PROT2 respectively to represent the first protein and the second protein in the sentence, the rest of the protein are expressed as PROT.
- Using OAOD (One Answer per Occurrence in a Document) principle processes the relationship of protein, that is, each occurrence of protein relations from the corpus are considered to be unique.

2.2. Feature Selection

- Location Information

Location information expresses proteins position in a sentence, using the location of "P" of protein PROT1 or PROT2.

- Local context information

Local context information includes left word of the first protein entity and number of other protein on the left side of the entity, and right word of the second protein entity and number of other protein on the right side of the entity, as well as words between the first protein entity and the second protein entity and number of other protein entity. Because multiple words provide similar information, the number of words will be divided into four levels: no word, one word, two words, greater than or equal to three words. In addition protein context of speech may also contain information on the interactions between the protein, we extract three words of speech before and after each protein in the protein pair.

- Keywords information

In the extraction of protein interaction relations, some special keyword provides important information. In this paper, use a list of key words which provided by [10] to structure keyword dictionary. We take the number of keywords as features, and the same, divided into no keyword, a keyword, greater than or equal to two keyword as feature.

- Interactions sentence information

Interactions sentence information indicates whether there are other protein pairs in the sentence with protein pair.

- Phrase syntax information

Phrase syntactic tree reflects the grammatical structure of sentences and can express semantic information over long distances. For example"PROT1, PROT and PROT2 levels were statistically greater in patients."Phrase syntactic tree is shown in **Figure 1**. The phrase syntactic features selected specifically:

Syntactic tree feature 1: the common root class of two proteins

Syntactic tree feature 2: the number of nodes forms the first protein to the root protein

Syntactic tree feature 3: the number of nodes forms the second protein to the root protein

- Dependent information

Dependent information can be revealed of the relation of long-distance dependencies in the sentence, and can avoid noise arising from the unstructured characteristics. In [6] it proves that the dependent information as characteristics can effectively improve the effect of PPI extraction. For example "Collectively, PROT1 is yet another functional ligand for PROT2." Dependent information is shown in **Figure 2**. Observed dependency information, whether there is dependence between protein entity and in [10] it presented the keyword may provide information for PPI extraction, therefore we add two features on the basis of Liu *et al*. [6] raised the dependency information feature, respectively:

Dependent information feature 1: Whether there is a direct dependency PROT1 and keyword.

Dependent information feature 2: Whether there is a direct dependency PROT2 and keyword.

2.3. The Optimal Convex Combination Kernel Function

Before looking for optimal combination kernel function

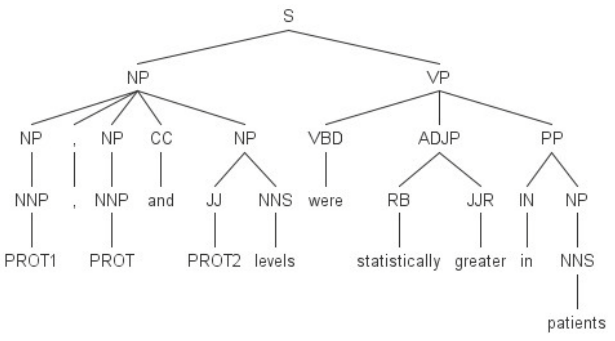

Figure 1. An instance of the phrase syntactic tree.


```
advmod(ligand-8, Collectively-1)
nsubj(ligand-8, PROT1-3)
cop(ligand-8, is-4)
advmod(ligand-8, yet-5)
dep(ligand-8, another-6)
amod(ligand-8, functional-7)
root(ROOT-0, ligand-8)
prep_for(ligand-8, PROT2-10)
```

Figure 2. An instance of the dependent information.

is convex, the first thing you need is to get the PPI performance of each single kernel function. Under the above feature selection strategy, then we can convert all instances to feature matrixes. In order to obtain different PPI extraction performance of the single kernel function, the high dimension matrixes are obtained by different feature matrixes mapping from single kernel function, the high dimensional matrix after training can obtain the different classification models. Finally using the test corpus and we can achieve the PPI extraction of each single kernel function.

For convex combination kernel function, the determination of convex combination ratio parameter is a very important problem. Suppose there are n kinds of kernel function, such as k_1, …, k_n. The convex combination kernel (CCK, Convex Combination Kernel) function consists by the n kinds of single kernel functions is shown in type (1). a_i is called the convex combination ratio parameter, $\sum_{i=1}^{n} a_i = 1$.

$$CCK = \sum_{i=1}^{n} a_i k_i \quad (1)$$

If the convex combination parameter selection may have m, the computational complexity is O (m^n). If the parameter selection of the range is 0 to 1, step length is 0.1, and there are four kinds of kernel functions, that is to say, m is 11, n is 4, and there are 14641 times you will need to experiment to determine the optimal parameters of kernel function, Obviously, O (m^n), such a huge computational complexity is unacceptable in practical application. Aiming at this problem, this paper presents some principles constantly looking for two kinds of kernel function is the optimal convex combination kernel function instead of these two kinds of kernel function strategy, for a variety of kernel function is the optimal kernel function is convex combination. This strategy can be decreased from computational complexity O (m * (n − 1)), under the condition of just to find the optimal convex combination kernel function only needs 33 experiments. As for kernel function with the worst performance for the principle, to obtain the optimal algorithm of convex combination kernel function is shown in **Figure 3**.

The optimal convex combination kernel function generation algorithm
Algorithm input: N kinds of kernel function is a collection of PPI extraction performance F, F = {F_i, i = 1,2….,n.}
Algorithm output: The optimal kernel function is convex combination(Optimal Convex Combination Kernel, OCCk)
Algorithm steps:

```
While(number(F)>1)
    candidate1←min(F)
     F=F-min(F)
     candidate2←min(F)
     F=F-min(F)
     F=F+Optimal (candidate1,candidate2)
     if number(F)=1
       then Return Optimal (candidate1,candidate2)
```

△Number (F) is the number of F_i in F
△min (F) is the minimum value in F
△Optimal (k1, k2) is the Optimal kernel function from k1,k2

Figure 3. The optimal convex combination kernel function generation algorithm.

3. Experiment and Results

This Paper uses AIMed corpus as the experimental data. AIMed is the corpus of PPI extraction used most frequently. It contains 225 MEDLINE abstracts, there are 177 abstracts contain PPI instance, and 48 abstracts does not contain the PPI instance, referring to 4084 proteins entities. After preprocessing, this paper removes 59 autocorrelation PPI instances, retains 154 nested instances. Finally, there are 1000 positive instances and 4084 negative instances. In [11,12], they verified that the SVM classification model have better effect in PPI extraction, therefore, this paper also uses the SVM as a machine learning method. In addition, in order to facilitate comparison with other experiments, the paper uses 10 times the cross validation method. This paper uses precision, recall and F1 Value to evaluate the result.

3.1. Experimental Method

The paper sets up three groups of experiments:

- Experiment 1 will test different kernel functions for the same characteristics in terms of PPI extraction performances are different. Experiment will get all the PPI extraction performance of the single kernel function, and sort them, then search for the optimal preparation convex combination kernel function. This paper uses three kinds of single kernel function including: Radial Basis Function, Polynomial Kernel Function, Linear kernel. Respectively, such as type (2), (3), (4), where *x* and *y* are two arbitrary dimension vector.

$$K_{RBF}(x, y) = \exp\left(-|x - y|^2\right) \quad (2)$$

$$K_{polynomial}(x, y) = \left(x \cdot y^T\right)^3 \quad (3)$$

$$K_{linear}(x, y) = x \cdot y^T \quad (4)$$

- Experiment 2 will compare three optimal convex

combination principle of kernel function to find the best, and verify that the proposed model of optimal classification of convex combination kernel function training is better than single kernel function training out of classification model. Principle 1: Find two worst performances for the optimal kernel function is convex combination to replace the original kernel function. Principle 2: look for two performances best optimal convex combination of nuclear function to replace the original kernel function. Principle 3: look for a better performance and a worst performance optimal kernel function is convex combination to replace the original kernel function.

- Experiment 3 will compare other systems in PPI extraction.

3.2. Experimental Results and Analysis

- Extraction performance of different single kernel functions

Table 1 shows that between the RBF kernel, polynomial kernel function and linear kernel function, RBF kernel function gets the best performance in PPI extraction. The results of this sort will prepare future experiments. At the same time, **Table 1** also shows that different kernel functions on the same characteristics have different PPI extraction performance.

- Compared with different principles for finding the optimal kernel

Table 2 shows that, in accordance with the strategy that finding the optimal convex combination kernel from two worst kernels and using this optimal kernel instead of them can get the best PPI extraction performance. But in accordance with the strategy that finding the optimal convex combination kernel from two best kernels and using this optimal kernel instead of them can get the worst PPI extraction performance. In addition, PPI extraction performance of each principle is higher than RBF kernel, which gets the best PPI extraction performance in all single kernels.

- Compared with other systems

Table 3 shows that, firstly, the proposed method gets the best precision at 65% in all feature based systems, and its F1 Value is close to 60. Niu *et al*. [4] is using a linear kernel, Bui *et al*. [7] using the RBF kernel function. They both use single kernel function. Compared to the single kernel function, the convex combination kernel function can increase the PPI extraction precision. Secondly, the F1 value on the feature based method has surpassed it on convolution kernel based method, furthermore, the feature based method is much faster than the convolution kernel based method, so the feature-based method will become the mainstream in the future research in PPI extraction.

Table 1. PPI Extraction performance on different single kernel functions.

Kernel method	P/%	R/%	F1/%
KRBF	**56.8**	**51.3**	**53.9**
Kpolynomial	53.2	49.6	51.3
Klinear	51.2	45.0	50.6

Table 2. PPI Extraction performance on different principles.

Principle method	P/%	R/%	F1/%
Principle 1 (two best performance)	58.2	52.4	55.1
Principle 2 (two worst performance)	**65.0**	**55.3**	**59.8**
Principle 3 (one best one worst)	57.3	53.9	55.5
KRBF	56.8	51.3	53.9

Table 3. Compared with other systems.

System	P/%	R/%	F1/%
Feature-based method			
This paper proposed	**65.0**	55.3	59.8
Miyao *et al*. [5]	-	-	59.5
Liu *et al*. [6]	63.4	44.1	52.0
Niu *et al*. [4]	43.2	**70.2**	53.5
Bui *et al*. [7]	55.3	68.5	61.2
Convolution kernel-based method			
Qian *et al*. [1]	59.1	57.6	58.1
Kim *et al*. [2]	61.4	53.3	56.6
Airola *et al*. [3]	52.9	61.8	56.4
Composite kernel method			
Sætre *et al*. [8]	64.3	44.1	52.0
Sætre *et al*. [9]	-	-	**64.2**

4. Conclusion

In this paper, in accordance with the principle which is searching for the optimal kernel function for two single kernel functions with the worst performance instead of the two of kernel functions. We find the optimal concentration of convex combination kernel function in the kernel function set which is composed of RBF kernel, polynomial kernel function and linear kernel function three single-core kernel function. Experiments on AIMed corpora, with this optimal convex combination of kernel function trained classifier model can improve performance of the PPI extraction compared to single kernel function classification model, and obtained the highest accuracy rate of 65.0% in the same system. In the next step's work, we will look for more effective features, and find out other work to solve a single kernel function limitation in the PPI extraction method based on the cha-

racteristic method.

5. Acknowledgements

This research was supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. 61175068, No.6126041.

REFERENCES

[1] L. H. Qian and G. D. Zhou, "Tree Kernel-Based Protein-Protein Interaction Extraction from Biomedical Literature [J]," *Journal of Biomedical Informatics*, Vol. 45, No. 3, 2012, pp. 535-543. http://dx.doi.org/10.1016/j.jbi.2012.02.004

[2] S. Kim, J. Yoon, J. Yang and S. Park, "Walk-Weighted Subsequence Kernels for Protein-Protein Interaction Extraction [J]," *MBC Bioinformatics*, Vol. 11, 2010, p. 107.

[3] A. Airola, S. Pyysalo, J. Bjrne, T. Pahikkala, F. Ginter and T. Salakoski, "All-Paths Graphkernel for Protein-Protein Interaction Extraction with Evaluation of Crosscorpus Learning [J]," *BMC Bioinformatics*, Vol. 9, No. S1, 2008.

[4] Y. Niu, D. Otasek and I. Jurisica, "Evaluation of Linguistic Features Useful in Extraction of Interactions from PubMed; Application to Annotating Known, High-Throughput and Predicted Interactions in I2D [J]," *Bioinformatics*, Vol. 26, No. 1, 2020, pp. 111-119.

[5] Y. Miyao, *et al.*, "Evaluating Contributions of Natural Language Parsers to Protein-Protein Interaction Extraction [J]," *Bioinformatics*, Vol. 25, 2009, pp. 394-400. http://dx.doi.org/10.1093/bioinformatics/btn631

[6] B. Liu, L. H. Qian, H. L. Wang and G. D. Zhou, "Dependency-Driven Feature-Based Learning for Extracting Protein-Protein Interactions from Biomedical Text," *Proceedings of COLING*', Poster, 2010, pp. 757-765.

[7] Q.-C. Bui, S. Katrenko and P. M. A. Sloot, "A Hybrid Approach to Extract Protein-Protein Interactions [J]," *Bioinformatics*, Vol. 27, No. 2, 2011, pp. 259-265. http://dx.doi.org/10.1093/bioinformatics/btq620

[8] R. Sætre, K. Sagae and J. Tsujii, "Syntactic Features for Protein-Protein Interaction Extraction," *Proceedings of LBM*'07, Vol. 319, 2007, pp. 6.1-6.14.

[9] R. Sætre, K. Yoshida, M. Miwa, T. Matsuzaki, Y. Kano and J. Tsujii, "Extracting Protein Interactions from Text with the Unified AkaneRE Event Extraction System [J]," *IEEE/ACM Transactions on Computing Biological Bioinformatics*, Vol. 7, No. 3, 2010, pp. 442-453. http://dx.doi.org/10.1109/TCBB.2010.46

[10] C. Plake, *et al*., "Optimizing Syntax Patterns for Discovering Protein-Protein Interactions," *Proceedings of the ACM Symposium on Applied Computing*, 2005, ACM Press, New York, pp. 195-201.

[11] S. Kim, *et al*., "Walk-Weighted Subsequence Kernels for Protein-Protein Interaction Extraction [J]," *BMC Bioinformatics*, Vol. 11, 2010, p. 107. http://dx.doi.org/10.1186/1471-2105-11-107

[12] R. Sætre, *et al*., "Extracting Protein-Interactions from Text with the Unified AkaneRE Event Extraction System [J]," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 99, 2010, pp. 442-453. http://dx.doi.org/10.1109/TCBB.2010.46

