Published in final edited form as:

Artif Intell Med. 2010 July ; 49(3): 155–160. doi:10.1016/j.artmed.2010.04.003.

Document classification for mining host pathogen protein-protein interactions

Lanlan Yin^{a,2}, Guixian Xu^{a,b,c,2}, Manabu Torii^d, Zhendong Niu^b, Jose M. Maisog^{a,e}, Cathy Wu^f, Zhangzhi Hu^g, and Hongfang Liu^{a,1}

^aDepartment of Biostatistics, Bioinformatics, and Biomathematics, Georgetown University, Washington DC, USA

^b School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China

^c School of Information Engineering, Minzu University of China, Beijing, China

^d Imaging Science and Information Systems Center, Georgetown University Medical Center, Washington DC, USA

^e Medical Numerics, Inc., Germantown, MD, USA

^f Protein Information Resources, Georgetown University Medical Center, Washington DC, USA

^g Department of Oncology, Georgetown University Medical Center, Washington DC, USA

Abstract

Objective—Scientific findings regarding human pathogens and their host responses are buried in the growing volume of biomedical literature and there is an urgent need to mine information pertaining to pathogenesis-related proteins especially host pathogen protein-protein interactions (HP-PPIs) from literature.

Methods—In this paper, we report our exploration of developing an automated system to identify MEDLINE abstracts referring to HP-PPIs. An annotated corpus consisting of 1360 MEDLINE abstracts was generated. With this corpus, we developed and evaluated document classification systems using support vector machines (SVMs). We also investigated the effects of three feature selection methods (information gain, mutual information, and χ^2 test). The performance was measured using normalized discounted cumulative gain (NDCG) and positive predictive value (PPV) and all measures were obtained through 10-fold cross validation.

Results—NDCG measures for classification systems using all features or a subset of features selected using information gain and χ^2 range from 0.83 to 0.89 while classification systems built based on features selected using mutual information had relatively lower NDCG measures. The classification system achieved a PPV of 50.7% for the top 10% ranked documents comparing to a baseline PPV of 10.0%.

¹Corresponding Author: Building D, Room 180, Georgetown University, 4000 Reservoir Rd, NW Washington, DC 20007, USA. hl224@georgetown.edu. Phone: (202) 687-7933. Fax: (202) 687-2581.

²Equal contribution to the work

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Conclusions—Our results indicate that document classification systems can be constructed to efficiently retrieve HP-PPI related documents. Feature selection was effective in reducing the dimensionality of features to build a compact system.

Keywords

document classification; host pathogen protein-protein interaction; feature selection; literature mining

1. Introduction

The causative agents of infectious diseases consist of a great diversity of agents including bacteria, viruses, fungi, helminthes and protozoa. Because of the development of new molecular biology assays, there has been continuing progress in the study of pathogenicity mechanism. Meanwhile, due to the heightened concern about bioterrorism and emerging/reemerging infectious diseases, there have been major initiatives for large-scale genomic and proteomic projects to study the basic biology and disease-causing mechanisms of human pathogens [1,2]. As a result, a flood of molecular data is being generated, but important scientific discoveries regarding these pathogens and their host responses are often buried under the increasing volume of biomedical literature. It was reported that the growth of peer-reviewed literature in MEDLINE is exponential [3]. With this volume of publication, it is very difficult or even impossible for biologists to find or assimilate the relevant publications of pathogenicity. To effectively manage the knowledge of pathogens and to better understand the pathogens, an automated text mining system that can extract pathogen related information from the scientific literature is highly desired.

In this paper, we focus on the development of an automated text mining system to identify research articles describing host pathogen protein-protein interactions (HP-PPIs). We focus on pathogens that are bacteria. By reviewing thousands of documents in MEDLINE, we constructed a corpus consisting of 1360 abstracts where 135 abstracts are HP-PPI relevant (i.e., positive) and the remaining are not HP-PPI relevant (i.e., negative). The corpus was then used to train a machine learning classifier to identify HP-PPI related articles where samples are abstracts and features are words or phrases in the abstracts. Three feature selection methods, information gain (IG), χ^2 test, and specific mutual information (SI) were compared for reducing the high dimensionality of the feature space.

2. Background and related work

The task considered in this study is a special case of identifying papers that describe protein-protein interactions (PPIs). There are several components in developing an automated literature mining system, including the construction of an annotated corpus, the selection of features and their representations, and the choice of machine learning algorithms. In the following, we present the research background and related work of each component.

2.1. Construction of annotated corpora from MEDLINE

One step towards constructing annotated corpora from MEDLINE is to select a subset of MEDLINE abstracts. There are different ways to obtain such subset. One approach is to use keyword search. For example, abstracts selected for the GENIA corpus were retrieved from MEDLINE using three MeSH terms, “human”, “blood cell” and “transcription factor” [4]. An alternative way to obtain a subset is to exploit the use of existing biomedical databases. For example, in order to construct an annotated corpus for the Interaction Article Subtask at the second BioCreative workshop, contents of two existing interaction databases, namely IntAct

and MINT, have been exploited [5]. After deriving such subset, domain experts can manually annotate them.

2.2. Feature representation/selection

In order to use machine learning methods, usually each document needs to be transformed into a feature vector. Commonly, features are based on words appearing in the document. Various feature selection techniques have been explored to overcome the high-dimensionality of word-based features. In this paper, three widely used feature selection methods, information gain (IG), specific mutual information (SI), and χ^2 test, were applied and compared.

IG represents the quantity of information in a feature with regard to class prediction on the basis of presence/absence of the feature in a document. Let $\{c_i\}_{i=1}^m$ be a set of categories to be predicted. Then the IG value of feature t in a document collection $IG(t)$ is defined as follows:

$$IG(t) = E - E_1 - E_2, \quad (1)$$

$$E = - \sum_{i=1}^m P(c_i) \log_2 P(c_i), \quad (2)$$

$$E_1 = - P(t) \sum_{i=1}^m P(c_i|t) \log_2 P(c_i|t), \quad (3)$$

$$E_2 = - P(\bar{t}) \sum_{i=1}^m P(c_i|\bar{t}) \log_2 P(c_i|\bar{t}), \quad (4)$$

where E is the entropy of the document collection; $P(c)$ is the occurrence probability of category c ; $P(t)$ and $P(\bar{t})$ are the occurrence probabilities of presence and absence of t ; and finally $P(c|t)$ and $P(c|\bar{t})$ are the conditional probabilities of the occurrence of category c with or without feature t . The larger $IG(t)$ is, the more important t is. By calculating IG value for each variable appearing in the abstracts, a rank list for all the variables can be obtained. Given a threshold value, features with IG values ranked high are selected to build classifiers.

In information theory, the SI of two random variables has been used to describe the mutual dependence of the two variables. In text mining, the SI of feature t in category c , $SI(t, c)$, can be defined as:

$$SI(t, c) = \log \frac{p(t, c)}{p(t)p(c)}, \quad (5)$$

where $p(t, c)$ is the joint occurrence probability of t and c ; and $p(t)$ and $p(c)$ are occurrence probabilities of t and c , respectively. Then the mutual information of t , $MI(t)$, can be defined as [6]:

$$MI(t) = \sum_{i=1}^m p(t, c_i) S I(t, c_i) + \sum_{i=1}^m p(\bar{t}, c_i) S I(\bar{t}, c_i) \quad (6)$$

The definition here yields the equivalence of $MI(t)$ and $IG(t)$ [7]. To distinguish from IG, Yang and Pedersen computed SI only based on the presence of a specific term. The feature value of feature t was defined in two alternative ways [7]:

$$MI_MAX(t) = \max_{i=1}^m \{S I(t, c_i)\}, \quad (7)$$

$$MI_AVG(t) = \sum_{i=1}^m p(c_i) S I(t, c_i). \quad (8)$$

Feature words were ranked accordingly and only the top-ranked features were used to build classifiers.

The third feature selection method we applied is χ^2 test which is commonly used to test the independence of two variables. Here, the two variables are feature t and document class c . The null hypothesis is that the occurrence of t and the occurrence of c are independent. The statistics of χ^2 is defined as:

$$\chi^2(c, t) = \sum_{ec \in \{0,1\}} \sum_{et \in \{0,1\}} \frac{(O_{et,ec} - E_{et,ec})^2}{E_{et,ec}}, \quad (9)$$

where χ^2 is the test statistic that asymptotically approaches a χ^2 distribution. O is an observed frequency; E is an expected (theoretical) frequency, asserted by the null hypothesis. Similarly, features are ranked with respect to their χ^2 scores, and the top-ranked features are selected to train the classifier, since a high χ^2 score indicates that the hypothesis of independence between the feature and the class is incorrect.

3. Method and experimental design

Figure 1 illustrates the overall data flow of constructing the classification system. It consists of several steps:

- Generating an annotated MEDLINE corpus: each abstract was annotated either positive or negative based on its relevance to HP-PPI;
- Reducing the high dimensional feature space: three feature selection methods (IG, MI, and χ^2 test), and the resulting features were applied to train classifiers;
- Evaluating the performance: ten-fold cross-validation was used to evaluate the performance.

3.1. Annotated data generation

In order to gather HP-PPI related abstracts, two biomedical databases were investigated. First, a data file was downloaded from UniProtKB, where the HP-PPI information is annotated for the protein entries, and the relevant MEDLINE abstracts are cited. If a cited abstract contains HP-PPI information, it is considered as positive, while unrelated abstracts are labeled as negative. Second, a set of MEDLINE abstracts obtained by keyword searching were reviewed by two domain experts. The pathogen related and unrelated abstracts were tagged manually. Mining these two databases resulted in 135 positive abstracts and 1225 negative abstracts, with a total of 1360 samples.

3.2. Feature representation and selection

Each document was normalized by changing lexical variants to their base forms and replacing nouns and adjectives by their corresponding verbs based on the SPECIALIST lexicon, a component of the Unified Medical Language System (UMLS) [8]. We also replaced punctuation marks with spaces, and changed uppercase letters to lowercase letters. After normalization, we used uni-grams and bi-grams as features. An n -gram is a subsequence of n items from a given sequence. Accordingly, our features included every single normalized word (uni-gram) in the corpus and every two neighboring normalized words (bi-gram) present in the corpus. The frequency of a uni- and bi-gram in each abstract was used as the feature value. Three feature selection methods introduced previously were applied. Additionally, for mutual information, we experimented with different document frequency thresholds where features with frequency lower than the given threshold were removed.

3.3. Document classification

A growing number of statistical and probabilistic machine learning algorithms have been applied to document classification, including K nearest neighbor, Bayesian approaches, decision trees, symbolic rule learning, and neural networks [9]. Here, we chose Support Vector Machines (SVMs), a supervised learning algorithm proposed by Vapnik and his co-workers [10,11]. It has been widely used for text mining and achieved promising results. The purpose of training SVM classifiers is to find a hyperplane to separate the two classes with the maximum margin [10,11]. SVMlight, by Joachims, is one of the most widely used SVM classification and regression packages. The algorithms used in SVMlight has scalable memory requirements and can handle problems with many thousands of support vectors efficiently [12,13]. In the present project, we used the SVMlight package and chose the linear kernel. We also experimented with other types of kernels such as polynomial or radial basis function (RBF), but observed no performance improvement.

3.4. Performance evaluation

The performance was evaluated through 10-fold cross validation. In 10-fold cross validation, an annotated corpus is partitioned into 10 portions, and each portion is used to evaluate a classifier trained with the remaining 9 portions. Instead of traditional binary classification, for each run, we generated a rank list based on the classification scores.

The following metrics were used to measure the performance:

- Simplified normalized discounted cumulative gain (NDCG).

$$NDCG = \frac{1}{Z_k} \sum_{m=1}^k \frac{2^{R_m} - 1}{\log(1+m)} \quad (10)$$

where Z_k is a normalization factor calculated to make it so that the NDCG of a perfect ranking at k is 1. R_m is the relevance of an abstract to HP-PPI, either 1 (relevant) or 0 (irrelevant), m is the rank of the abstract in the final list, and k is the total number of the abstract [6]. The advantage of NDCG is that among the classifiers with same accuracy, the classifier which can rank the true positive literature higher will be awarded more.

- Receiver operating characteristic curve (ROC curve). This is a graphical plot of the true positive rate against the false positive rate for the different possible cut-points of a binary classifier system [14].
- Another measure used is the positive predictive value (PPV) [15] which is the same as precision (i.e., the probability of predicted positives to be true positives) given a cut-point of a binary classifier system.

3.5. System implementation

As we have discussed, the machine learning task considered here is to rank a set of documents according to their PH-PPI relevance. In order to judge the PH-PPI relevance for any given abstract, we used the following method:

- obtain N score lists by executing N runs of 10-fold cross validation using the corpus where scores were ones assigned by SVM classifiers,
- build an SVM classifier C with all documents in the corpus,
- for a new abstract d , use classifier C to obtain a score $S(d)$ for d ,
- for each score list that was obtained, compute the percentage of documents that are positive among the documents with scores larger than $S(d)$, and
- average the above percentage over N score lists and display the percentage as the relevance score. The higher the score, the more relevant the abstract.

To test the effectiveness of the proposed method, we used one run of 10-fold cross validation and measured PPVs for a given relevance score threshold.

4. Results and discussion

4.1. Document frequency for specific mutual information

Figures 2 and 3 display the performance of SVM classifiers on our corpus after using MI_MAX and MI_AVG as the feature selection method with different document frequency thresholds. In general, MI_MAX has better performance than MI_AVG. The classification results showed that the NDCG (normalized discounted cumulative gain) value of both MI_MAX and MI_AVG generally decreases as the number of features decreased, which can be explained by the smaller amount of information (fewer features) recruited by the classifier. However, the performance of MI_MAX was improved as the document frequency threshold increased. By setting the threshold of document frequency, low frequency terms with document frequency less than the threshold can be removed from the feature space. In our case, the NDCG of the classifier based on MI_MAX remains above 0.83 even with only 1000 feature terms if the document frequency threshold was no less than 3, while the NDCG of other classifiers with threshold of 1 and 2 was less than 0.82 with 3000 feature terms. Therefore, setting document frequency threshold is a crucial step for applying MI_MAX. But for MI_AVG, the performance was not improved by increasing the threshold of document frequency. To calculate the average mutual information for each term, a weight was assigned to each term for each class. Here, we use the occurrence probability of each class as the weight. Due to the imbalanced distribution of classes (only 10% documents are positive), the weight for the terms in positive abstracts would be 0.1,

much lower than the terms in negative abstracts. Consequently, the informative terms in positive documents were swamped by the terms in negatives. In our project, the features of positive documents are more helpful in recognizing the pattern. Together, the poor performance of MI_AVG and the distinct characteristics from MI_MAX were caused by the bias towards low-frequency words and the bias towards words in negative abstracts.

4.2. Comparison of feature selection methods

Table 1 and Figure 4 show the comparison results of three feature selection methods. When there were more than 4000 feature terms, MI_MAX, IG, and CHI had similar performance. But as the number of features used in the classifier decreases to less than 4000, the performance of MI_MAX declines much faster than IG and CHI. The classifier curve goes to the minimum 0.769 if the classifier used 100 terms selected based on MI_MAX, while using the same number of features selected from information gain or χ^2 test the classifier's NDCG is still above the line of 0.831, indicating MI_MAX does not have comparable performance to the other two methods: information gain and χ^2 test.

4.3. Comparison of systems with and without feature selection

The overall performance of no feature selection, information gain, and χ^2 test is shown in Figure 5. We found that as the feature number became greater than 500, the performance of the IG and χ^2 test were comparable to that of no feature selection. Both IG and χ^2 test outperformed no feature selection by virtue of the much lower dimensional feature spaces they used. These two feature selection methods selected a small number of variables and then generated compact models.

In implementation, each abstract in the test dataset was assigned a score by an SVM classifier, and the abstracts were ordered by those scores. The higher the score, the more likely the abstract to be positive. Therefore, given a rank threshold N , the abstracts with rank above N were classified as positive abstracts, while the abstracts with lower rank were categorized as negative. Given a series of rank thresholds, ROC curves of different classifiers built upon no feature selection, information gain, and χ^2 test were shown in Figure 6. All three curves approach the left-hand border and then the top border of the ROC space, located far from the no-discrimination line, indicating competent classification capability. The χ^2 curve lies closer to the 45-degree diagonal of the ROC space, suggesting poor performance. Figure 7 shows the positive predictive value of three models at rank thresholds of 25, 50, 75, 100, and 135. A classifier with no feature selection gave the best performance among the three at each threshold because it utilized all uni-grams and bi-grams as features, thereby using as much information as possible from the samples. However, classifiers built upon information gain and χ^2 achieved comparable results with much lower cost. The number of term used was reduced to 2000 (a 98.3% reduction).

4.4. Evaluation of implementation

Table 2 shows the performance of PPVs when implementing the system using information gain (IG) threshold of 0.002 and the number of runs parameter is set to 5. Given a relevance score threshold 0.5, PPV is 50.7% which indicates that if an abstract receives relevance score higher than 0.5, the probability of the abstract to be positive is 50.7%, which is much higher than the random chance to select positive abstracts (9.9%; 135 out of 1360).

5. Conclusions

In summary, we built a text mining system that retrieves MEDLINE abstracts pertaining to host-pathogen protein-protein interaction. We manually constructed a literature corpus consisting of 1360 Medline abstracts, where 135 are HP-PPI related and the remaining ones

are HP-PPI unrelated. This corpus was used to build automated text categorization system that classifies MEDLINE abstracts as HP-PPI related or not. As a classification algorithm, SVM was used. In addition, three feature selection methods (IG, MI, and χ^2 test) were considered to reduce the high dimensionality of the feature space. Among them, IG and χ^2 test were found effective in reducing the dimensionality and, thus, in building a compact system. Our results indicate that an automated document classification system can help curators search and retrieve HP-PPI related biomedical literature.

Acknowledgments

This work was supported by grants USAMRMC-W81XWH0720112, NSF-IIS-0639092, NSF-DBI-0845523, and NIH-R01LM009959.

References

1. Zhang CG, Chromy BA, McCutchen-Maloney SL. Host-pathogen interactions: a proteomic view. *Expert Rev Proteomics* 2005;2(2):187–202. [PubMed: 15892564]
2. Nomura K, DebRoy S, Lee YH, Pumphlin N, Jones J, He SY. A bacterial virulence protein suppresses host innate immunity to cause plant disease. *Science* 2006;313(5784):220–223. [PubMed: 16840699]
3. Hunter L, Cohen KB. Biomedical Language Processing: What's Beyond PubMed? *Molecular Cell* 2006;21(5):589–594. [PubMed: 16507357]
4. Kim JD, Ohta T, Tateisi Y, Tsujii J. GENIA corpus: a semantically annotated corpus for bio-textmining. *Bioinformatics* 2003;19(1):180–182.
5. Krallinger M, Leitner F, Rodriguez-Penagos C, Valencia A. Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome Biology* 2008;9(Suppl 2):S4. [PubMed: 18834495]
6. Manning, CD.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*. Cambridge University Press; New York, NY, USA: 2008.
7. Yang, Y.; Pedersen, JO. A comparative study on feature selection in text categorization. In: Fisher, DH., editor. *The Fourteenth International Conference on Machine Learning*; San Francisco, CA, USA: Morgan Kaufmann Publishers, Inc.; 1997. p. 412-420.
8. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research* 2004;32(Database):D267. [PubMed: 14681409]
9. Sebastiani F. Machine learning in automated text categorization. *ACM computing surveys (CSUR)* 2002;34(1):1–47.
10. Joachims, T. Text categorization with support vector machines: learning with many relevant features. In: Nédellec, C.; Robveiol, C., editors. *Proceedings of ECML-98, 10th European Conference on Machine Learning*; London, UK: Springer Verlag; 1998. p. 137-142.
11. Vapnik, V. *Statistical Learning Theory*. Wiley-Interscience; New York, NY, USA: 1998.
12. Joachims, T. *Learning to classify text using support vector machines: methods, theory and algorithms*. Kluwer Academic Publishers; Norwell, MA, USA: 2002.
13. Joachims, T. *Making large-scale support vector machine learning practical, advances in kernel methods: support vector learning*. MIT Press; Cambridge, MA, USA: 1999.
14. Fawcett T, Flach PA. A response to webb and ting's on the application of roc analysis to predict classification performance under varying class distributions. *Machine Learning* 2005;58(1):33–38.
15. Altman D, Bland J. Diagnostic tests 2: predictive values. *BMJ. British medical journal (International)* 1994;309(6947)

Figure 1.
Experimental design

Figure 2.
Performance of maximum mutual information (MI_MAX)

Figure 3.
Performance of average mutual information (MI_AVG)

Figure 4.
Comparison results of three feature selection methods

Figure 5.
Performance of no feature selection vs. feature selection

Figure 6.
ROC curve of different classifiers

Figure 7.
PPV of different classifiers

Table 1

Average NDCG of classifiers with feature selection

FEATURE NUMBER	IG		CHI		ML_MAX	
	AVG	SD	AVG	SD	AVG	SD
14,000	0.881	0.011	0.883	0.006	0.881	0.008
13,000	0.885	0.013	0.875	0.013	0.881	0.007
12,000	0.888	0.007	0.880	0.009	0.879	0.009
11,000	0.890	0.006	0.885	0.005	0.882	0.004
10,000	0.888	0.005	0.884	0.006	0.880	0.007
9,000	0.886	0.008	0.886	0.005	0.881	0.009
8,000	0.882	0.006	0.883	0.006	0.876	0.006
7,000	0.882	0.007	0.880	0.006	0.874	0.008
6,000	0.880	0.006	0.878	0.008	0.876	0.005
5,000	0.882	0.007	0.880	0.007	0.871	0.007
4,000	0.881	0.006	0.874	0.007	0.873	0.007
3,000	0.883	0.006	0.881	0.006	0.862	0.005
2,000	0.881	0.009	0.880	0.006	0.846	0.011
1,000	0.874	0.010	0.872	0.015	0.838	0.015
900	0.876	0.006	0.878	0.012	0.843	0.018
800	0.871	0.011	0.877	0.008	0.831	0.015
700	0.872	0.008	0.872	0.007	0.833	0.013
600	0.873	0.008	0.872	0.011	0.842	0.015
500	0.874	0.005	0.863	0.012	0.839	0.014
400	0.874	0.009	0.860	0.013	0.831	0.015
300	0.868	0.009	0.866	0.007	0.817	0.013
200	0.862	0.006	0.862	0.007	0.792	0.013
100	0.850	0.009	0.831	0.008	0.769	0.010

Table 2

The performance of implementation

Threshold	TOTAL	TP	PPV
0	1360	135	0.099
0.1	1185	118	0.099
0.2	519	106	0.204
0.3	304	93	0.306
0.4	207	82	0.396
0.5	136	69	0.507
0.6	96	63	0.656
0.7	69	52	0.754
0.8	41	30	0.732
0.9	8	7	0.875