A Machine Learning Approach for the Curation of Biomedical Literature--KDD Cup 2002 (Task 1)

S. Sathiya Keerthi[1], Chong Jin Ong[1],
Keng Boon Siah[1],David B.L. Lim[1],Wei Chu[1]
[1] Mechanical Engineering Department,
National University of Singapore
10 Kent Ridge Crescent, Singapore 119260
[mpessk,mpeongcj,engp2438,engp1431]@nus.edu.sg;
chuwei@gatsby.ucl.ac.uk

Min Shi[2], David S. Edwin[2], Rakesh Menon[2],
Lixiang Shen[2],Jonathan Y.K. Lim[2],Han Tong Loh[1,2]
[2]Design Technology Institute Ltd,
National University of Singapore
10 Kent Ridge Crescent, Singapore 119260
[dtishim,dtidsaej,dtirm,dtislx,dtilykj,mpelht,]@nus.edu.sg

ABSTRACT

In this paper, we present an automated text classification system for the classification of biomedical papers. This classification is based on whether there is experimental evidence for the expression of molecular gene products for specified genes within a given paper. The system performs pre-processing and data cleaning, followed by feature extraction from the raw text. It subsequently classifies the paper using the extracted features with a Naïve Bayes Classifier. Our approach has made it possible to classify (and curate) biomedical papers automatically, thus potentially saving considerable time and resources.

Keywords: Text mining, Pre-processing, Naïve Bayes Classifier, ROC curve, Paper curation

1. INTRODUCTION

Background and introduction of this KDD Cup 2002 competition task1 were described in an overview article by A. Yeh et. al. (this issue).

2. APPROACH TAKEN

In building the system, we took three steps – data pre-processing, data preparation for the classifier, and model building.

2.1 Data Pre-processing

The original articles were formatted as raw text files. The text in these files contained large amounts of noise. As such, extensive pre-processing and data cleaning is needed to be carried out so that the articles could be represented in a manner by which features could be properly extracted and relevant statistics about these features generated.

2.1.1 Noise Removal

The first step in pre-processing the data was the removal of certain control characters in the raw text that would otherwise interfere with the detection of gene symbols within the text. This output from noise removal was then passed into the module that performed the synonym replacement.

2.1.2 Synonym Replacement

Using different synonyms to reference a particular gene is very common in biomedical papers. To find out the occurrence of a gene, it is necessary to replace all the synonyms with the original gene symbol.

2.1.3 Formatting

Document formatting was included as part of the pre-processing. The pre-processed article files were used as input to generate a corresponding file that was formatted in a consistent manner. These files had special tags to indicate the location of different sections and paragraphs in the text.

2.1.4 Feature Extraction

Extraction of features from the articles involved searching for gene symbols and various evidence keywords from the text. According to the task documentation, there were tens of evidence types. We managed to use 12 types in building the model. We generated two types of keywords in which one type was from evidence files themselves (e.g. the phrase "northern blot" was picked as one keyword for an evidence of a transcript) and the other type was manually extracted from the training texts by domain experts. When the two sets of keywords were ready, the statistics generation was carried out. We were interested in the distance between a gene symbol and the keyword. For instance, if the gene and the keyword were in the same sentence, the distance was 0. If keyword is in the next sentence, the distance was 1. Within a single paragraph, the distances were calculated and recorded in an output file that was generated for each evidence type.

To find the needed yes/no (Y/N) answers, we decided to split our system into 2 levels. One level determined whether experimental evidence existed for a product (transcript (TR) or polypeptide (PP)) of a particular gene within a particular paper or document (doc-gene-level). The other level determined whether such evidence existed for any gene in a particular paper (doc-level). A 'Y' for any gene product at the doc-gene-level would lead to a 'Y' at that paper's doc-level, which meant that the paper should be curated.

2.2 Data Preparation for Classifier Building

After the final stage of text processing, a list of doc-gene examples is produced.

2.2.1 Condensation of Paragraph

For each doc-gene combination, if gene occurs several times within one paragraph, there are several repetitive representations for this doc-gene example. However, only one example is necessary. As such, we removed the redundant examples.

In essence, the doc-gene example that had the minimum distance between a particular gene under consideration and the other keywords was chosen as the representative example. All ties were kept. This was done to both the positive and negative examples.

2.2.2 *Manual Checking*

This step is only applied to the training set. Upon removing the repetitive doc-gene examples found within the same paragraph, the positive examples to be given to the classifier were manually chosen. Once the positive examples are manually selected, they are combined with the other negative examples. Because the answers are an abbreviated form of the evidence files, these files were not available for the test set.

2.2.3 *Final Input into Classifier*

The data generated from the text are transformed to the final format. For the *n*th keyword, the minimal absolute distance (KW*n*) is kept and the count (KW*n*_count) of its appearance around the gene under consideration is listed as a feature. Hence, for each keyword two features are created. The section in which the particular example occurs is also used as a feature. After this pre-processing, the data is ready for classifier building.

2.3 Model Building

The Naïve Bayes Classifier (NBC) has been used as an effective classifier for many years [1]. The use of NBC facilitated a quick running time. This was essential for our system, which required building classifiers, each with a large number of input records. The ROC curve [2] and the F-measure were used as scoring mechanisms.

2.3.1 *Building Classifiers*

There are two-stages for model building (each having a classifier). In the first stage, the doc-gene examples (statistics of KWn, KWn_count) are given as input to the first classifier. An initial classifier model is built. The output of the classifier was probabilistic estimates of whether a 'YES' class (for Ev*i*) is obtained for a given doc-gene example which was done by 5-fold cross validation. As such, there was a probability estimate for all doc-gene examples. The doc-gene examples are not unique because some ties were kept. There could be more than one example for a particular doc-gene. Since the final scores are computed for distinct doc-gene examples, it was necessary to finally compute a F-measure based on distinct doc-gene examples. This was carried out by picking the example with the highest probability for similar doc-genes so that only one of them was used as the representative example. At the end of the first stage, a set of distinct doc-gene examples was available for input into a second classifier.

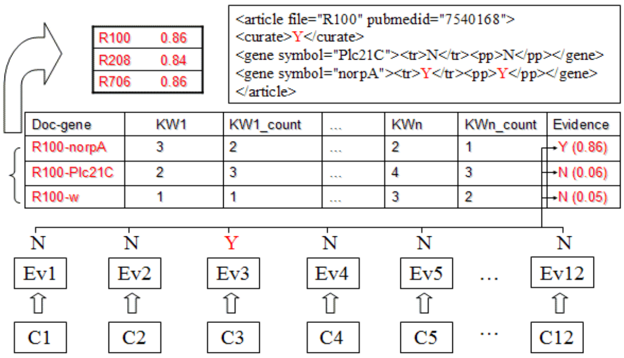
In the second stage, distinct doc-gene patterns are used to train the classifier. The probabilistic output is used to compute the ROC curve and F-measure at the doc-gene level. In plotting the ROC, the number of true positives was taken to be the actual number of positively classified documents in the entire set of training documents. In this way the model was built for each evidence type.

Figure 1 shows the model.

2.3.2 *Evidence at Doc-Gene Level and Doc- Level*

For each gene listed with each paper, steps were carried out to answer the doc-gene-level Y/N questions by determining whether experimental evidence existed for that gene's TR products and similarly for PP products. For example, let the scores for the three genes in paper R100 be as follows: gene R100-norpA is 0.86, R100-Plc21C is 0.06 and R100-w is 0.05. Then 0.86 will be chosen as paper R100's representative score.

A threshold was set by maximizing [ROC + F-measure].The score of a paper was not only critical for curation of a single paper. It also determined the position of the paper in the ranked list. The higher the score is, the higher the paper is ranked.



R100	0.86
R208	0.84
R706	0.86

```
<article file="R100" pubmedid="7540168">
<curate>Y</curate>
<gene symbol="Plc21C"><tr>N</tr><pp>N</pp></gene>
<gene symbol="norpA"><tr>Y</tr><pp>Y</pp></gene>
</article>
```

Doc-gene	KW1	KW1_count	...	KWn	KWn_count	Evidence
R100-norpA	3	2	...	2	1	Y (0.86)
R100-Plc21C	2	3	...	4	3	N (0.06)
R100-w	1	1	...	3	2	N (0.05)

C*i* is the *i*th part of the first classifier; Ev*i* is the *i*th Evidence type

Figure 1 Model

3. RESULTS

In this issue's task 1 overview article (A. Yeh et. al.) are the results from this task's submissions. Each of our scores were within the corresponding first quartile. Our approach performed quite well on the "ranked-list" (81%) and "yes/no curate paper" (73%) subtasks.

4. CONCLUSION

In this work, an approach has been provided for detecting evidence of gene-product formation in biomedical papers. In particular, a document collection on Drosophila was studied. The biomedical articles were initially pre-processed to remove noise and to provide for standardization between articles. Important features were then extracted and used as input into a Naïve Bayes Classifier. We found that domain knowledge was essential for the feature extraction task. A classifier was built for each evidence type, the results from which were finally combined for evidence detection of gene-product formation.

5. ACKNOWLEDGEMENTS

This work was supported by the Design Technology Institute Ltd. management. We would like to thank P. Long and P. Kolatkar from the Genome Institute of Singapore for their contributions to this work.

6. REFERENCES

[1] CHENG, J., AND GREINER, R. (2001) Learning Bayesian Belief Network Classifiers: Algorithms and System, (*Proceedings of the fourteenth Canadian conference on artificial intelligence*) AI'2001

[2] BRADLEY, A.P. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7), 1145—1159