3B1-1

Classifying biomedical text abstracts using binary and multi-class Support Vector Machine

Rozilawati binti Dollah Masaki Aono

Department of Information and Computer Sciences
Toyohashi University of Technology

Text classification systems on biomedical literature aim to select relevant articles that match query keywords from large corpora. For this purpose, systems for finding relevant documents must be able to identify terms related to the search in the abstracts and also must distinguish between relevant and irrelevant results. Lately, many researchers attempt to find more applicable ways for classifying biomedical text articles in order to help users find relevant articles on the web. Due to this reason, our focus is on the problem of identifying relevant and irrelevant documents based on binary and multi-class classification in biomedical texts, especially for text biomedical abstracts. For our experiments, we have randomly downloaded and collected 400 paper abstracts of four diseases, including cancer, hepatitis, HIV/AIDS and thyroid from Medline database. Then, we have tested and compared the performance of binary classification and multi-class classification using LIBSVM. The results obtained in our experiments demonstrate that the accuracy of binary classification on the average 80.89% (with scaling) and 86.92% (without scaling), meanwhile multi-class classification on the average 75.73% (with scaling) and 85.25% (without scaling) for our biomedical text data with four categories of diseases. We observe that the choice of percentage for training and testing dataset has little influence on the classification accuracy.

1. Introduction

Literature searching is one of the most common information processing tasks in the biomedical sciences. Systems for finding relevant documents must be able to identify terms related to the search in the abstracts and also must distinguish between relevant and irrelevant results [5]. Therefore, text classification systems on biomedical literature aim to select relevant articles to a specific issue from large corpora [4] Text classification is the process of using automated techniques to assign text samples into one or more set of predefined classes [1]. However, in [4] text classification task can be defined as assigning category labels to new documents based on the knowledge gained in a classification system at the training stage.

The text classification task is called "binary" if it assigns a given documents into one of two classes either positive or negative class, meanwhile it is called "multi-class" if it assigns a given documents into one of *k* classes. Many researchers attempt to find more applicable way for classifying biomedical text articles in order to help users find relevant articles on the web. Several of statistical classification methods and machine learning techniques have been applied to text classification including techniques based on Decision Tree, Neural Network and Support Vector Machine (SVM). SVM has been prominently and widely used for binary and multi-class classification.

The goal of automatic text classification is to learn a classification scheme from training examples of previously classified documents. The learned scheme can then be used to classify test text documents automatically [3]. Due to this reason, in this paper we will focus on the problem of identifying relevant and irrelevant documents based on binary and multi-class classification in biomedical texts, especially in diseases category. Therefore, we use the approach that involves term and word frequencies to calculate a score of biomedical paper abstracts, focusing on four categories of diseases, namely cancer, hepatitis, HIV/AIDS and thyroid diseases.

We choose these diseases due to the number of patients who suffer from these critical diseases were increased lately. Other than that, the awareness among the individuals to get more information about these diseases caused them trying to find the related articles. At the same time, the increasing number of researches on these diseases also influence on this matter.

Consequently, we have conducted several experiments to compare the performance of binary classification and multi-class classification based on different percentage of training and testing of biomedical paper abstracts dataset. The result of our experiments demonstrates that the different percentage of training and testing dataset gave almost no influence to the classification accuracy. Moreover, we also have experimented with both training and testing dataset (with scaling and without scaling). We found that, the accuracy of classification using multi-class classification was much more intensively affected by the choice of "scaling" or "without scaling" than the choice of the percentage of training and testing dataset.

Author: Rozilawati binti Dollah, Toyohashi University of Technology, 1-1 Hibarigaoka, Tempaku-cho, Toyohashi, 0532-44-6764, rozeela@kde.ics.tut.ac.jp

2. Text Pre-processing

A main goal of text pre-processing is to transforms the text string representation into numeric feature vectors, where we represent documents as Vector Space Model. In our experiments, the text pre-processing step includes, stop word elimination, word stemming, word or feature selection and weighting. A set of vectors is extracted from the collected paper abstracts after text pre-processing.

Firstly, we generated a list of words for each paper abstract. Then, we eliminated words such as articles (*a, an, the*), preposition (*in, of, at*), conjunction (*and, but, or, nor*), pronouns *(I, you, them, it*) and etc. from each paper abstract using a standard stop-word list. The purpose of stop word elimination is to purge the list of words from "noise".

Secondly, we performed word stemming. Word stemming is a process of indentifying the base form of words by removing suffixes. Thus, the keywords of a query or paper abstract are represented by base forms rather than by the original words. This step was performed for each list of words to increase words or features coverage, which will increase the accuracy of classification process. For our experiments, we employed the Porter's stemming algorithm. Then, we created the vocabulary by combining a list of words that describe all paper abstracts for our experiments.

The third step in text pre-processing is word selection. The main objective of this step is to reduce the total number of words in the vocabulary for weighting process and also to remove noise from the datasets in order to optimize the classification accuracy. Learning process for a large amount of data is time consuming. Therefore, the words that appear below than three times in the vocabulary will be removed and the rest will be considered as feature vector and then, will be calculated the weight.

Finally, we calculated the weight. In this step, the weight for each word or feature in the vocabulary will be calculated using

the TFIDF formulation. TFIDF is one of the common weighting methods that can be used to describe documents in the Vector Space Model. This method was used to calculate weight for each word in vocabulary according to the frequency and the total number of paper abstracts containing that particular word. Then, a set of vectors will be divided into two sets for classifier training and testing.

3. Experiment and Result

The aim of this paper is to test and compare the accuracy of binary classification and multi-class classification for biomedical paper abstracts that involved four categories of diseases. 400 paper abstracts were randomly collected from Medline database using PubMed search engine. Each category of disease consists of 100 paper abstracts. In our experiments, we have used two categories of datasets, which are binary classification dataset and multi-class classification dataset. For each binary classification dataset, we used 100 positive and 100 negative of paper abstracts, meanwhile for multi-class classification dataset, we combined all of 400 paper abstracts. Then, we have conducted several experiments using LIBSVM [2] because it supports binary and multi-class classification.

The performance evaluation of text classifier is conducted on a testing data set which is different from the training set. For this purpose, we keep the same dataset, but change the percentage for training and testing into seven groups. In each group, we randomly selected document data vectors from dataset and put them into training and testing data respectively. The details of percentage for each group are as follows;

- Experiment I- training data (90%) and testing data (10%).
- Experiment II- training data (80%) and testing data (20%).
- Experiment III- training data (75%) and testing data (25%).
- Experiment IV- training data (70%) and testing data (30%).
- Experiment V- training data (60%) and testing data (40%).
- Experiment VI- training data (50%) and testing data (50%).
- Experiment VII- training data (40%) and testing data (60%).

For each Experiment I through VII, we have repeated runs for binary classification and multi-class classification using Radial Basic Function (RBF) kernel in LIBSVM. In the RBF kernel, there are two parameters to be determined in the SVM model, which are C (cost) and γ (gamma). Finally, we have compared the performance of classification based on the accuracy for all datasets done in these experiments. Table 1 below shows different accuracy of binary classification for each dataset (with scaling and without scaling).

Table 1: A comparison of binary classification accuracy

Dataset	Experiment	I	II	III	IV	V	VI	VII
Dataset (without scaling)	Cancer	**95**	**87**	86	**80**	83	**87**	**85**
	Hepatitis	**85**	**87**	**80**	**88**	**92.5**	**83**	64.17
	HIV/AIDS	**100**	**92**	90	**95**	**92.5**	**89**	**85**
	Thyroid	80	**90**	88	**93.34**	87.5	**83**	**85.83**
Dataset (with scaling)	Cancer	80	80	**92**	76.67	**85**	75	72
	Hepatitis	75	82.5	78	66.67	83.75	64	**75.83**
	HIV/AIDS	85	85	**92**	81.67	85	81	76.67
	Thyroid	80	85	**98.34**	80	**88.75**	75	85

From the Table 1, we found that the dataset (without scaling) outperforms the dataset (with scaling) in most of the experiments. Meanwhile, the group of experiment III which involved 75% training data and 25% testing data (with scaling) perform better than the same group category of dataset without scaling. The performance of HIV/AIDS dataset shows good accuracy in all experiments without scaling dataset and most of the experiments with scaling dataset. It is followed by cancer dataset. This might be caused by the keywords frequency in cancer and HIV/AIDS training data respectively, higher than the keywords frequency in other training data to build usable models.

Table 2 below is the results of accuracy for multi-class classification experiments based on different percentage of training and testing. Generally, the accuracy of multi-class classification (without scaling) outperforms the multi-class classification (with scaling) in all the experiments conducted. From the result of accuracy in all the experiments, we have observed that different percentage of training and testing data produce different performance of accuracy.

Table 2: A comparison of multi-class classification accuracy

Dataset	Experiment	I	II	III	IV	V	VI	VII
Dataset (without scaling)	Cancer + Hepatitis + HIV/AIDS + Thyroid	**92.5**	**83.75**	**80**	**85**	**87**	**83.5**	**85**
Dataset (with scaling)	Cancer + Hepatitis + HIV/AIDS + Thyroid	70	76.5	72	78.75	75.5	75.5	81.88

Overall, from the experiments that have been done, the results show the different performance in the binary and multi-class classification. The accuracy of binary classification on the average 80.89% (with scaling) and 86.92% (without scaling), meanwhile multi-class classification on the average 75.73% (with scaling) and 85.25% (without scaling) respectively. Moreover, we compare the different percentage of training and testing dataset. From Tables 1 and 2, we observe that the choice of percentage for training and testing dataset has little influence on the classification performance.

4. Conclusion and Discussion

In our experiments, we have employed LIBSVM to classify our datasets. From the results of our experiments, we found that LIBSVM performed well in the binary classification and multi-class classification for our biomedical text data with four categories. Even though, we conducted several experiments using different percentage of training and testing dataset, in most of the experiments done, it produced 60% and more accuracy in classification. Also we observed that the accuracy of classification in term of percentage of training and testing dataset was neither monotonically increasing nor monotonically decreasing as attached tables showed. This might be caused by the frequency of query keywords in the training datasets are higher than other keywords. This threshold must be further examined with different values. In this experiment, we employed all the words that appear at least three times in all paper abstracts.

Our future target is to improve the result of performance by reducing the dimension of dataset. Therefore, we will attempt to filter and choose only relevant keywords as a basis for constructing a set of vectors. Then, we will conduct the experiments with list of specific terminologies that gathered from biology experts. We are also interested in increasing the number of disease categories in order to analyze the performance of binary classification and multi-class classification.

References

[1] A. M. Cohen, An effective general purpose approach for automated biomedical document classification, AMIA 2006 Symposium Proceeding, pp.161-162 (2006)

[2] C.-C. Chang and C.-J. Lin, LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm (2001)

[3] C.Y. Ho and W. Lam, Automatic discovery of document classification knowledge from text databases, available at http://citeseer.ist.psu.edu/310941.html (1998)

[4] F. M. Couto, B. Martins and M. J. Silva, Classifying biological articles using web resources, Proceedings of the 2004 ACM symposium on Applied Computing, pp.111-115 (2004)

[5] J. E. Leonard, J. B. Colombe, and J. L. Levy, Finding relevant references to genes and proteins in Medline using a Bayesian approach, Bioinformatics 18(11), pp.1515-1522 (2002)