Asian Journal of Information Technology 15 (22): 4551-4553, 2016
ISSN: 1682-3915

Classification of Chemical Documents Using Multi Class Support Vector Machines

R. Hema and T.V. Geetha
Department of Computer Science and Engineering, Anna University, Chennai, India

Abstract: As most information are unstructured text on the web, text mining plays a vital role in information retrieval field. Text classification is the process of classification of text documents into predefined categories. This type of organizing text documents is useful for fast retrieval of documents and it fetches smaller space in which the users may find their documents. In organic chemistry, organic reactions have been used for preparing the huge amount of organic compounds and they are further used for the synthesis of new organic molecules. This study explores the application of Support Vector Machines (SVMs) forclassifying the chemical text documents based on the types of chemical reactions. SVMs achieve the better results over the existing methods and eliminates the needfor manual parameter tuning.

Key words: Chemical documents, classification, one vs. all SVM, organic reactions, India

INTRODUCTION

Text classification is the process of sorting text documents into one or more classes of similar documents. This classification is based on the set of feature sets within the documents that can be used to differentiate among the text documents and assign the documents to the different classes. These classes may be determined either by humans or algorithmically or may be determined dynamically as needed.Many algorithms have been developed to deal with automatic text classification (Kruengkrai and Jaruskulchai, 2002).

Recently, most researchers use the machine learning techniques to automatically classify the text documents into predefined classes by first using a training set and then adapting the classifier to the future set of text documents or test documents (Joachims, 1998). The machine learning process is initiated by an examination of sample documents to identify the feature sets and to train the classifier to classify the text documents. This training phase may be supervised or unsupervised. Here the set of classes has been defined in prior. Machine learning techniques will automatically classify and discover patterns from the different types of the documents. The task is to determine a classification model which is able to assign the correctclass to a new document *d*.

Finding relevant information is one of the most important challenges on the web. Now a days in chemical domain, a huge amount of new publications, research reports and patents are produced. Organic reactions have been used for preparing the huge amount of organic compounds and will be used for the synthesis of new organic molecules in the future. A detailed knowledge of organic reactions and their mechanisms is therefore an essential tool for any scientist or technician involved with the development of organic molecules in any scientific and technological field. The production of many man-made chemicals such as drugs, plastics, food additives, fabrics depend on organic reactions.

In this study, we explore the idea of using Support Vector Machines (SVMs) for the classification of chemical literature based on the organic reactions. It analyzes the feature set selection for each class and the classification process according to the feature sets. SVMs achieve better results over the existing methods and behave robustly among different learningtasks. Furthermore, they are fully automatic, eliminating the needfor manual parameter tuning.

Text classification process: Generally the stages of Text Classification is shown in the following Fig. 1. The first step in the classification process is to collect the documents in various forms like pdf, doc, html and from websites. After the collection, the documents have to be represented in a standard format. Some preprocessing techniques like tokenization, stopword removal and stemming has to be done to reduce the number of words in the documents. The document representation is one of the pre-processing technique that is used to reduce the complexity of the documents and make them easier to

Corresponding Author: R. Hema, Department of Computer Science and Engineering, Anna University, Chennai, India

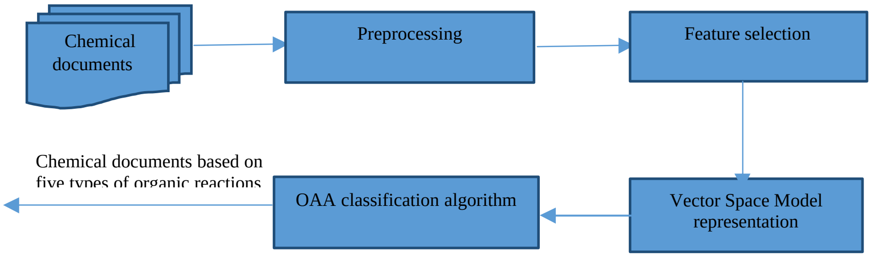
Chemical documents
Preprocessing
Feature selection
Vector Space Model representation
OAA classification algorithm
Chemical documents based on five types of organic reactions

Fig. 1: Stages of text classification process

handle for the learning algorithm and the classification task. This leads to feature representation of text. Each distinct word w_i corresponds to a feature and the number of times the word $w_i$occurs in the document is taken as its value. To avoidunnecessarily large feature vectors, words are considered as features only if they occur in the training data at least five times and if they are not stop-words like and or, the etc.

The most commonly used document representation is Vector space model which is an algebraic model and in this the documents are represented as vectors of words.In this representation scheme, the feature space is very high-dimensional containing >10000 dimensions. Here, the stopwords (Salton and McGill, 1983) are removed to reduce the dimensions and the information gain isused to select asubset of features.

The term weight will control the precision and recall of the accuracy of document classification. The term weighting for the vector space model is entirely based on single term statistics. There are three main factors for term weighting: term frequency factor, collection frequency factor and length normalization factor. These three factors are multiplied together to make the resulting term weight. The frequency of occurrence is used as a common weighting method for various terms within a document (Luhn, 1958). The term frequency will describe the crux of the documents and is basically used as a weighted document vector (Salton and McGill, 1983).

In general, there are various weighting schemes to differentiate one document from the other.Most of themassume that the importance of a term is proportional with the number of document in which the term appears (Salton and McGill, 1983) .These document discrimination factors lead to a more effective retrieval, i.e., an improvement in precision and recall (Salton and McGill, 1983). Experimentally, it is shown that the best results are obtained by using term frequency with inverse document frequency (Salton and Buckley, 1988). So in this research, TF-IDF (Term Frequency-Inverse Document Frequency) is used as the term weighting scheme and each document feature vector is normalized to unit length.

MATERIALS AND METHODS

Data Set: For this research, two datasets have been collected for classification. First, we have taken 12000 chemical research articles from beilstein Journal of organic chemistry. Of them, 9500 articles are used for training and the remaining articles are used testing. After the preprocessing steps, the training data contains 10,113 different terms.

Second, totally 8350 research articles have taken from the chemical sciences Journal. In the second set, 7000 documents are training data and the remaining 1350 documents are test data. After the preprocessing steps, this corpus contains 7654 different terms.

RESULTS AND DISCUSSION

Support vector machines: A Support Vector Machine (SVM) is a supervised classification machine learning technique that has been extensively used for text classification tasks. SVM are based on statistical learning theory and have the aim of determining the location of decision boundaries that produce the optimal separation of classes (Vapnik, 1995). Initially, a SVM is a binary classifier, that is the classes will take the values of + 1 or -1. Thus, the selected decision boundary will be the greatest margin between the two classes.

SVM for multiclass classification: Originally, SVMs were developed to perform binary classification.In machine learning, multiclass classification is the problem of classifying instances into more than two classes. However, for multiclass classification, a number of methods to generate multiclass SVMs from binary SVMs have been proposed by researchers. For this research, “One Against All (OAA)” approach is chosen for classifying the chemical documents based on five basic types of organic reactions. This multiclass method has an advantage of constructing the equal number of binary classifiers as the number of classes and this method is represented in the Fig. 2. The following section describes

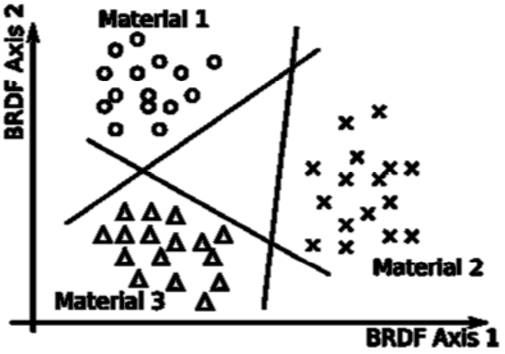

Fig. 2: One vs all multiclass SVM (Adopted from web)

Table 1: Classification accuracy of chemical documents based on organic reactions

Classifiers	Accuracy (%)	Organic reactions
1	82.0	Addition
2	79.0	Substitution
3	76.0	Elimination
4	86.0	Rearrangement
5	71.0	Organic redox
Average accuracy	78.8	

the classification of chemical documents into five classes in detail.

One against all approach: This method is also called one-against-rest classification. In this research, we want to classify the chemical documents into five classes based on the five types of organic reactions. Therefore, five binary SVM classifiers may be created where each classifier is trained to distinguish one class from the remaining four classes. Other SVM classifiers are constructed in the same manner. During the testing phase, data vectors are classified by finding margin from the linear separating hyperplane. The final output is the class that corresponds to the SVM with the largest margin. The above OAA approach is described in the following Pseudocode:

```
Inputs:
•  L, a learner (training algorithm for binary classifiers)
•  Samples X
•  Labels y where y? {addition, substitution, elimination,
   rearrangement, organic redox} is the label for the
   sample X

Output:
•  A list of classifiers $f_k$ for k∈{addition, substitution,
   elimination, rearrangement, organic redox}

Procedure:
•  For each k in {addition, substitution,…… organic
   redox}
•  Construct a new label vector y' = 1 where y = k, 0 (or
   -1) elsewhere
•  Apply L to X, y' to obtain $f_k$
```

This OAA classification model is tested by applying it to test data with known target values and comparing the predicted values with the known values. Metrics are used to assess the accuracy of theclassification model. If the model performs well it can then be applied to make predictions on new data. Accuracy refers to the percentage of correct predictions made by the model when compared with the actual classifications in the test data. Table 1 shows the accuracy of our multiclass classification model for chemical documents.

CONCLUSION

This study introduces the multiclass classification of chemical research articles based on the five basic types of chemical organic reactions. Text classification process and one vs. All SVMs are discussed. The experimental results shows that the multiclass classification technique achieve the classification accuracy of 78.8%. In future, we are in a plan to classify the chemical documents based on more types of chemical organic reactions.

REFERENCES

Joachims, T., 1998. Text categorization with support vector machines: Learning with many relevant features. Proceedings of the 10th European Conference on Machine Learning, Chemnitz, Germany, April 21-23, 1998, Springer, Berlin, Heidelberg, pp: 137-142.

Kruengkrai, C. and C. Jaruskulchai, 2002. A parallel learning algorithm for text classification. Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-25, 2002, ACM, Edmonton, Alberta, ISBN:1-58113-567-X, pp: 201-206.

Luhn, H.P., 1958. The automatic creation of literature abstracts. IBM J. Res. Dev., 2: 159-165.

Salton, G. and C. Buckley, 1988. Term-weighting approaches in automatic text retrieval. Inform. Process. Manage., 24: 513-523.

Salton, G. and M.J. McGill, 1983. An Introduction to Modern Information Retrieval. 2nd Edn., Mcgraw Hill, USA, Pages: 448.

Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. 1st Edn., Springer-Verlag, New York, USA.