Artificial Intelligence and Knowledge Engineering Laboratory

Task 4. Document classification

Authors: Maciej Piasecki

Task Objectives

Getting familiar with the representation of text documents as vectors of word frequencies. Learning basic methods for feature selection and Machine Learning algorithms for classification. Obtaining basic skills in using Weka environment for Machine Learning.
The main task is to train a classifier based on Machine Learning for recognising documents that belong to one of the selected categories on the basis of the content of the documents.

Subtasks

The minimal part:

1. Read chapter 13 from "Introduction to Information Retrieval" – described in the bibliography.
2. Download Weka environment for Data Mining and Machine Learning: http://www.cs.waikato.ac.nz/ml/weka/
3. Download the collection of news group documents (an newsgroups archive) called 20k Newsgroups: http://qwone.com/~jason/20Newsgroups/
4. Select for the experiments 5 different categories (if they have very different topics, it will be easier to obtain better results during the experiments).
5. Write a program for converting the newsgroups documents into the vectors of the word frequencies:
   doc_id, category, word1, frequency_of _w1, …, word1, frequency_of _w1
6. Select k=10 000 words that are most frequent and occur in not too small number of documents.
7. Convert document vectors to the arff format (all vectors in one file), in which the attributes are words selected in the step 6 and the decision class is the document category.
8. Upload the arff file into Weka system.
9. Select a method for feature selection.
10. Choose one classifier – you can start with a version of Naïve Bayes.
11. Evaluate the performance of the classifier in 10-fold cross validation scheme for different parameter setting and different methods for feature selection.

12. Prepare a report: describe the classifier and the feature selection method in your own words, analyse the results of the evaluation, draw conclusions.

The rest:

13. Test one more classifier that is based on a different Machine Learning algorithm than the first one.
14. Add to the report information about this classifier and the obtained results

Additional information

Task rating

2 points – building vector representation
2 points – preparing the data for Weka and uploading the data to Weka
2 points – running feature selection and classification
2 points – writing report
2 point – testing the second classifier and extending the report
Extra 3 points: implementation of a feature weighting algorithm, e.g. tf.idf and testing the classifier on the weighted values of features instead of the raw frequencies – this step must be described in the report too

Bibliography

1. Christopher D. Manning. Prabhakar Raghavan. Hinrich Schütze. *Introduction* to. *Information*. *Retrieval*. Cambridge University Press, 2008. (there will be also a copy in the Board):
http://www-nlp.stanford.edu/IR-book
or
https://archive.org/details/AnIntroductionToInformationRetrieval
or
http://www-connex.lip6.fr/~gallinar/livres%20-%20fichiers/2007-%20Manning-irbookonlinereading.pdf
2. Weka documentation: http://www.cs.waikato.ac.nz/ml/weka/documentation.html
3. Papers suggested in Weka for the selected classifier(s).