A Comparative Study on Feature Selection of Text Categorization for Hidden Markov Models

Abstract

Background & Purpose

Text categorization (TC) is the task of automatically assigning pre-set categories to textual digital documents. A number of statistical TC models with machine learning techniques have been proposed and reported to tackle the TC problem. One major difficulty with TC models is a high-dimensional feature space. A feature, normally a word or a phrase, is an informative attribute conveying the subject content of a document. Text categorization tasks with a large number of training documents can easily lead to a few thousand features. Since a large number of features, referred to as high dimensionality of feature space, may significantly increase computational times for the TC models, it is highly desirable to reduce the feature space without the loss of model performance.

This paper is a comparative study of feature selection methods in a statistical learning model of text categorization. The dual purposes of this study are to investigate popular dimensionality reduction methods in general TC domain and to probe task-dependent dimensionality reduction methods, for the Hidden Markov Models (HMMs) of the text categorization problem. Our aim is to explore how features should be selected for this model, based on the statistical properties of the related TC task. This study seeks empirical evidence for the following questions: (1) how much do the selected dimensionality reduction methods improve the classification accuracy of the HMM for text categorization? (2) How can the statistical properties of the task be used for reduction methods?

Conceptual framework

Since the early 1990’s, research on TC has been shifted towards the development of statistical learning models using machine learning techniques, such as decision trees, Bayesian models, and Support Vector Machines. A large number of the dimensionality of feature space for a TC task can not be accommodated in learning models due to the formidable computational times. The reduction of high dimensionality can be done in two different ways. First, a feature-selection approach may reduce the size of the features to be considered by selecting a subset of all available features. Second, a feature-extraction approach may use synthetic features, which do not occur in original documents. Our study focuses on the first approach. Various dimensionality reduction methods have been proposed and tested on different learning models in the TC domain, including Document frequency, Information gain, Mutual information, Chi-square, Odds ratio, and Relevancy score. Comparison studies of reduction methods have been carried out, and the ranking of reduction techniques on the effectiveness of the method.

* It refers to the number of distinct features
has been reported\textsuperscript{15, 16}. However, as Sebastiani (2002) points out, more comparative studies on diverse experimental settings, such as different classifiers, and different tasks, need to be conducted.

**Methodology**

Several popular reduction methods - Information gain, Mutual information, and Chi-Square - and other task-related methods are tested in this study. The research is conducted in three phases: (1) Data collection, training data\textsuperscript{17} and test data\textsuperscript{18} (2) Classifier creation, and (3) Classifier Testing.

(1) A set of cataloguing records from the OCLC\textsuperscript{19} WorldCat database is collected to constitute a training set for the statistical model. A subset of information in cataloguing records containing topical subjects and their descriptors is used as the training data set in the proposed model. A database for test data sets is created containing digital documents previously classified by professional librarians. In selecting these documents, the availability of content and the type of content is considered. The dissertation abstracts from Proquest Digital Dissertations database (PQDD)\textsuperscript{20} has been selected for test set of this system. Library of Congress Classification (LCC) and Library of Congress Subject Headings (LCSH) for the selected abstracts are found in OCLC FirstSearch Database-WorldCat database.

(2) The statistical learning HMMs using different reduction methods are designed and trained (using the sample data from OCLC) as a text classifier. The basic model of this comparison study was designed and built earlier.

(3) Experimental results from different methods are compared to that of the basic classifier. The performance of each classifier will be measured in classification accuracy by comparing the result to the manual classifications by professionals.

* This paper is directly relevant to the theme of *technologies*, as it discusses the improvement of the performance of automatic machine classification system.
Notes

17 A set of data required for building the statistical model.
18 A set of data for measuring the system performance.
19 Online Computer Library Center, Inc (OCLC) is a nonprofit membership organization offering services for libraries and their users. The primary service of this organization is to provide online shared cataloging system for its members.