Expert Systems with Applications 39 (2012) 3127–3134

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

A multi-class SVM classification system based on learning methods from indistinguishable chinese official documents ☆

JuiHsi Fu *, SingLing Lee

Department of Computer Science and Information Engineering, National Chung Cheng University, 168 University Road, Minhsiung Township, 62162 Chiayi, Taiwan, ROC

ARTICLE INFO

Keywords:
Support Vector Machines (SVM)
Multi-class classification
Chinese official document classification
Indistinguishability identification
Incremental learning

ABSTRACT

Support Vector Machines (SVM) has been developed for Chinese official document classification in One-against-All (OAA) multi-class scheme. Several data retrieving techniques including sentence segmentation, term weighting, and feature extraction are used in preprocess. We observe that most documents of which contents are indistinguishable make poor classification results. The traditional solution is to add misclassified documents to the training set in order to adjust classification rules. In this paper, indistinguishable documents are observed to be informative for strengthening prediction performance since their labels are predicted by the current model in low confidence. A general approach is proposed to utilize decision values in SVM to identify indistinguishable documents. Based on verified classification results and distinguishability of documents, four learning strategies that select certain documents to training sets are proposed to improve classification performance. Experiments report that indistinguishable documents are able to be identified in a high probability and are informative for learning strategies. Furthermore, LMID that adds both of misclassified documents and indistinguishable documents to training sets is the most effective learning strategy in SVM classification for large set of Chinese official documents in terms of computing efficiency and classification accuracy.

1. Introduction

In government departments and companies, some kinds of articles and documents are still handled by human labors. Among these data, Chinese official documents have been used very often to inform and communicate officially with companies/corporations. Each employee is responsible for dispatching official documents to all related departments. Hence, designing an accurate classification system to handle Chinese official documents will improve government employee's working efficiency. However, in Chinese documents, it is very often that segmented terms could not represent the original meaning of this content completely, since no delimiter exits in the content. Moreover, no formal stop-word list is defined like English to remove those meaningless words. Due to lack of complete content representation and stop-word lists, distinct features are difficult to be extracted from the document content. Additionally, in Chinese official documents which are well-formed and textual, some special characteristics are listed below:

1. Short and brief content
   The abstract of an official document usually is used to represent this document for classification. The abstract needs to be short and brief to describe government affairs.
2. Fewer distinct features
   In official affairs, Chinese official documents belonging to different units (classes) tend to be similar due to most terms in a document are not discriminative. Classifying Chinese official documents is difficult when depending on only few distinct terms.

Our objective is to classify Chinese official documents more precisely. One department in the company is corresponding to a class label, so the problem of automatically dispatching an official document is reduced to a multi-class classification problem. However, special characteristics of Chinese official documents usually cause poor classification results. The traditional learning strategy is to add misclassified documents to the training set in order to adjust classification rules. In this paper, some correctly classified documents are observed to be indistinguishable since their labels are predicted in low confidence. It is worth noting that, indistinguishable documents should be informative for strengthening prediction rules since following similar ones could be correctly classified in a higher probability. Hence, a distinguishing method, Identifying Possibly Misclassification Documents (IPMD), is proposed to

☆ This work is supported by NSC, Taiwan, ROC under Grant No. NSC 97-2221-E-194-029-MY2.

* Corresponding author.
E-mail addresses: fjh95p@cs.ccu.edu.tw (J. Fu), singling@cs.ccu.edu.tw (S. Lee).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.08.176

distinguish whether verified documents tend to be misclassified or not. Based on verified classification results and document distinguishability, four learning strategies that select certain documents to training sets are proposed to enhance classification accuracy and reduce the size of training sets. They are introduced in more details in Section 3.

Fig. 1 is an overview of our document classification system. Initially, training documents are processed by modules of Text Preprocessing and Classifier Training to build a prediction model. Feature extraction eliminates terms with weights lower than a predefined threshold, and term weighting methods (Combarro, Montanes, e Diaz, Ranilla, & Mones, 2005; Quinlan, 1986; Salton & Buckley, 1988; Salton & McGill, 1983) are used to represent document vectors. Classifier Training is kernel in classification systems. Some well-known classification methods, K-Nearest Neighbors (KNN) (Yuan, Yang, & Yu, 2005), Support Vector Machines (SVM) (Cortes & Vapnik, 1995; Cristianini & Taylor, 2000; Liang, 2004), Naive Bayes (Lewis, 1998; Lewis & Ringuette, 1994), and neural network (Wiener, 1995), have been well studied recently. Notably, SVM is adopted for solving our document classification problems since it has been proven to perform very effectively in many research results (Deng & Peng, 2006; Diaz, e Ranilla, Montanes, Fernandez, & Combarro, 2004; Dumais, Platt, Heckerman, & Sahami, 1998; Joachims, 1998; Kecman, 2001; Lee & Lee, 2005; Ramirez, Durdle, Raso, & Hill, 2006; Özgür & Güngör, 2006; Wang, 2005; Wang, Sun, Zhang, & Li, 2006; Wang & Fu, 2005) and is able to deal with high dimensional feature spaces. Geometrically speaking, SVM (Cortes & Vapnik, 1995) generates a hyperplane to separate positive instances from negative ones. The objective function is to maximize the distance from the nearest training instance to the separating hyperplane.

When the prediction model is generated, testing documents are also processed by Text Preprocess and their class labels are properly predicted by Classifier Training. Verified module judges the prediction results (supervised learning Alpaydin, 2004). Then, IPMD module utilizes the decision values of verified documents in SVM classification to determine whether they are distinguishable or not. Next, a revised Learning Strategy module that utilizes verified classification results and distinguishability of documents to select new training instances is developed in order to update prediction models and improve classification performance.

The objective of semi-supervised learning (SSL) (Joachims, 1999) is to utilize unlabeled samples for decreasing the use of labeled samples. In this paper, the proposed methods identify

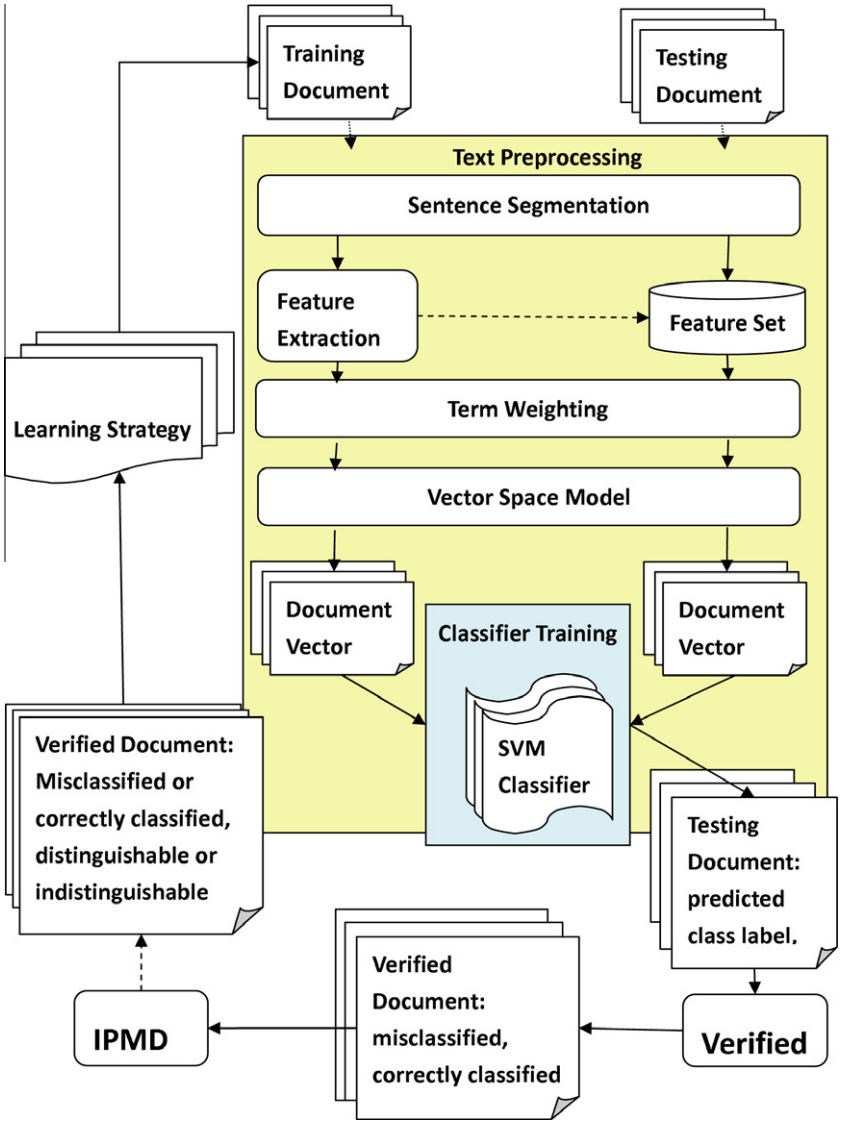

Fig. 1. Text Preprocessing, Classifier Training, and classifier testing of our SVM classification system.