Higher order feature selection for text classification

Abstract In this paper, we present the MIFS-C variant of the mutual information feature-selection algorithms. We present an algorithm to find the optimal value of the redundancy parameter, which is a key parameter in the MIFS-type algorithms. Furthermore, we present an algorithm that speeds up the execution time of all the MIFS variants. Overall, the presented MIFS-C has comparable classification accuracy (in some cases even better) compared with other MIFS algorithms, while its running time is faster. We compared this feature selector with other feature selectors, and found that it performs better in most cases. The MIFS-C performed especially well for the breakeven and $F$-measure because the algorithm can be tuned to optimise these evaluation measures.

Keywords Feature selection · Text classification

1 Introduction

Growing numbers of text information available on the Internet is raising interest in automatic analysis of text data. One such technique is text classification, which is a problem of automatically assigning free text documents to one predefined category from a set of possible categories [30]. This is accomplished using machine learning techniques.

A successful solution is based on the bag-of-words approach, where each document is represented as a list of words present in the document. Each word is then
considered as a feature, and the document is represented as a feature vector. Then a feature-selection filters out the useless features, leaving the useful ones that are used by the classifier to make a decision on the document class.

A number of popular approaches are successful in reducing the feature space. Some of these are chi-squared \[29\], information gain \[18\], mutual information \[7\] and odds ratio \[23\]. All of these techniques measure the correlation between each feature and the class, and the feature selection is performed by selecting the required number of features with the highest correlation values, and a comparison of these techniques can be found in \[34\]. However, all of these techniques are simple filters, which ignore the correlation among the features.

In this paper, we investigate some more advanced feature-selection approaches that use higher order decisions and take the feature-to-feature correlation into account when selecting the feature set. We investigate the correlation feature selection presented by Hall in \[11\] and Markov blanket feature selection presented by Koller and Sahami in \[16\]. In addition, we consider the mutual information feature-selection (MIFS) class of feature selector.

We investigate the MIFS presented by Battiti in \[3\] and its extension by Kwak in \[17\] as the MIFS-U. In particular, we focus on the MIFS-C variant presented by Bakus and Kamel in \[2\] and present a number of improvements. As part of all the MIFS-type algorithms, a redundancy parameter is required, and we present a search technique to find an optimal value for this parameter. We also present an optimised algorithm that skips the unnecessary calculations, which results in a training time improvement.

This paper is organised as follows. Section 2 gives some background on feature selection and text classification. Section 3 presents the MIFS feature-selection algorithms. Numerical results are presented in Sect. 4, and Sect. 5 offers some concluding remarks.

2 Background

2.1 Text classification

The objective of the text classification is to infer a classification rule from a sample of labelled training documents, such that it is able to assign an unknown document to one of \(L\) possible classes \(C = \{C_1, \ldots, C_L\}\). Each document, \(d\), is represented by a vector of features \(\{F_1, \ldots, F_N\}\) that describe the document. The most popular representation is called the bag-of-words approach, where each feature corresponds to one individual word \[19, 30\].

In this paper, we focus on the binary classification problem, where each document is classified into one of two categories. The document either belongs to a given topic or does not. We test the effectiveness of the feature selection using the following classifiers: naive Bayes, Rocchio, K-nearest neighbor, C4.5 decision tree and support vector machine classifiers.

2.1.1 Naive Bayes (NB)

The naive Bayes classifier \[12, 14, 22, 35\] is a simple, yet very effective, classifier. The assumption behind naive Bayes classifier is that the features are independent