Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 33 (2007) 1–5
Expert Systems with Applications
www.elsevier.com/locate/eswa

A novel feature selection algorithm for text categorization

Wenqian Shang [a,*], Houkuan Huang [a], Haibin Zhu [b], Yongmin Lin [a], Youli Qu [a], Zhihai Wang [a]

[a] School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, PR China
[b] Department of Computer Science, Nipissing University, North Bay, Ont., Canada P1B 8L7

Abstract

With the development of the web, large numbers of documents are available on the Internet. Digital libraries, news sources and inner data of companies surge more and more. Automatic text categorization becomes more and more important for dealing with massive data. However the major problem of text categorization is the high dimensionality of the feature space. At present there are many methods to deal with text feature selection. To improve the performance of text categorization, we present another method of dealing with text feature selection. Our study is based on Gini index theory and we design a novel Gini index algorithm to reduce the high dimensionality of the feature space. A new measure function of Gini index is constructed and made to fit text categorization. The results of experiments show that our improvements of Gini index behave better than other methods of feature selection.

Keywords: Text feature selection; Text categorization; Gini index; kNN classifier; Text preprocessing

1. Introduction

With the advance of WWW (world wide web), text categorization becomes a key technology to deal with and organize large numbers of documents. More and more methods based on statistical theory and machine learning has been applied to text categorization in recent years. For example, k-nearest neighbor (kNN) (Cover & Hart, 1967; Yang, 1997; Yang & Lin, 1999; Tan, 2005), Naive Bayes (Lewis, 1998), decision tree (Lewis & Ringuette, 1994), support vector machines (SVM) (Joachims, 1998), linear least squares fit, neural network, SWAP-1, and Rocchio are all such kinds of methods.

A major problem of text categorization is the high dimensionality of the feature space. For many learning algorithms, such high dimensionality is not permitted. Moreover most of these dimensions are not relative to text categorization; even some noise data hurt the precision of the classifier. Hence, we need to select some representative features from the original feature space (i.e., feature selection) to reduce the dimensionality of feature space and improve the efficiency and precision of classifier. At present the feature selection method is based on statistical theory and machine learning. Some well-known methods are information gain, expected cross entropy, the weight of evidence of text, odds ratio, term frequency, mutual information, CHI (Yang & Pedersen, 1997; Mladenic & Grobelnik, 2003; Mladenic & Grobelnik, 1999) and so on.

In this paper, we do not discuss these methods in detail. We present another new text feature selection method—Gini index. Gini index was early used in decision tree for splitting attributes and got better categorization precision. However, it is rarely used for feature selection in text categorization. Shankar and Karypis discuss how to use Gini index for text feature selection and weight-adjustment. They mainly pay attention on weight-adjustment. Their method only limits to centroid based classifier and their iterative method is time-consuming. Our method is very different from theirs. Through deeply analyzing the principles of Gini index and text feature, we construct a new

* Corresponding author.
E-mail addresses: shangwenqian@hotmail.com (W. Shang), haibinz@npissingu.ca (H. Zhu).

0957-4174/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2006.04.001

measure function of Gini index and use it to select features in the original feature space. It not only fit centroid classifiers but also fit other classifiers. The experiments show that its quality is comparable with other text feature selection methods. However, its complexity of computing is lower and its speed is higher.

The rest of this paper is organized as follows. Section 2 describes the classical Gini index algorithm. Section 3 gives the improved Gini index algorithm. Section 4 discusses the classifiers using in the experiments to compare Gini index with the other text feature selection methods. Section 5 presents the experiments' results and their analysis. In the last section, we give the conclusion.

2. Classical Gini index algorithm

Gini index is a non-purity split method. It fits sorting, binary systems, continuous numerical values, etc. It was put forward by Breiman, Friedman, and Olshen (1984) and was widely used in decision tree algorithms of CART, SLIQ, SPRINT and intelligent miner. The main idea of Gini index algorithm is as follows:

Suppose S is the set of s samples. These samples have m different classes ($C_i, i = 1,\ldots,m$). According to the differences of classes, we can divide S into m subset ($S_i, i = 1,\ldots,m$). Suppose S_i is the sample set which belongs to class C_i, s_i is the sample number of set S_i, then the Gini index of set S is:

$$\mathrm{Gini}(S) = 1 - \sum_{i=1}^{m} P_i^2, \tag{1}$$

where P_i is the probability that any sample belongs to C_i and estimating with s_i/s. Gini(S)'s minimum is 0, that is, all the members in the set belong to the same class; this denotes it can get the maximum useful information. When all the samples in the set distribute equably for the class field, Gini(S) is maximum; this denotes it can get the minimum useful information. If the set is divided into n subset, then the Gini after splitting is:

$$\mathrm{Gini}_{\mathrm{split}}(S) = \sum_{j=1}^{n} \frac{s_j}{s} \mathrm{Gini}(S_j). \tag{2}$$

The minimum $\mathrm{Gini}_{\mathrm{split}}$ is selected for splitting attribute.

The main idea of Gini index is: for every attribute, after it traverses all possible segmentation methods, if it can provide the minimum Gini index then it is selected as the divisive criterion of this node no matter it is the root node or a sub node.

3. The improved Gini index algorithm

To apply the Gini index theory described above directly to the text feature selection, we can construct the new formula:

$$\mathrm{Gini}(W) = P(W)\left(1 - \sum_i P(C_i|W)^2\right) + P(\overline{W})\left(1 - \sum_i P(C_i|\overline{W})^2\right) \tag{3}$$

After we analyze and compare the merits and demerits of the existing text feature selection measure functions, we improve formula (3) to:

$$\mathrm{Gini\,Text}(W) = \sum P(W|C_i)^2 P(C_i|W)^2. \tag{4}$$

Why we amend formula (3) to formula (4)? The reasons include three aspects as follows:

(1) The original form of Gini index is used to measure the impurity of attributes towards categorization. Smaller the impurity is, better the attribute is. If we adopt the form $\mathrm{Gini}(S) = \sum_{i=1}^{m} P_i^2$, it is to measure the purity of attributes towards categorization. Bigger the value of purity is, better the attribute is. In this paper, we adopt the measure form of purity. This form is more adapt to text feature selection. In paper (Gupta, Somayajulu, Arora, & Vasudha, 1998; Shankar & Karypis), they all adopt the measure form of purity.

(2) In other authors' papers, they all emphasize that text feature selection inclines to high frequency words, namely, including the $P(W)$ factor in the formula. Experiments show that some words that do not appear have contributions to judge the class of text, but this contribution is far less significant than the effort to consider the words that do not appear, especially when the distribution of the class and feature values is highly unbalanced. Yang and Pedersen (1997) and Mladenic and Grobelnik (1999) compare and analyze synthetically the merits and demerits of many feature measure functions in their papers. Their experiments show that the demerits of information gain are to consider the word that does not appear. The demerits of mutual information are not to consider the affect of the $P(W)$ factor leading to select rare words. Expected cross entropy and weight of evidence of text overcome these demerits, hence their results are better. Therefore, when we construct the new measure function of Gini index, we get ride of the affection factor expressing words that do not appear.

(3) Iff W_1 appears in the documents of class C_1 and W_1 appears in every document of class C_1; Iff W_2 appears in the documents of class C_2 and W_2 appears in every documents of class C_2, then W_1 and W_2 is the same important feature. But due to $P(C_i) \neq P(C_j)$, from $\mathrm{Gini\,Text}(W) = P(W)\sum_i P(C_i|W)^2$ to compute out $\mathrm{Gini\,Text}(W_1) \neq \mathrm{Gini\,Text}(W_2)$, this is not consistent with domain knowledge. So we adopt $P(W|C_i)^2$ to replace $P(W)$, for considering the unbalanced class