Text classification and keyword extraction by learning decision trees

Yasubumi Sakakibara  Kazuo Misue  Takeshi Koshiba
IIAS-SIS, Fujitsu Laboratories Ltd.
140, Miyamoto, Numazu, Shizuoka 410-03, Japan

Abstract

In this paper, we propose a completely new approach to the problem of text classification and automatic keyword extraction by using machine learning techniques. We introduce a class of representations for classifying text data based on decision trees, and present an algorithm for learning it inductively. Our algorithm has the following features: it does not need any natural language processing technique, and it is robust for noisy data. We show that our learning algorithm can be used for automatic extraction of keywords for text retrieval and automatic text categorization. We also demonstrate some experimental results using our algorithm.

1 Decision trees on strings

We introduce a class of representations for classifying text data based on decision trees. We use simple attributes on strings, denoted key(w), defined as follows: For a string w in \(\Sigma^*\),

\[
\text{key}(w)(z) = \begin{cases} 
1 & \text{if a string } z \text{ contains } w, \\
0 & \text{otherwise.} 
\end{cases}
\]

Let \(CT = \{\text{key}(w) \mid w \in \Sigma^*\}\). We call decision trees defined over \(CT\) on the domain \(\Sigma^*\) text-classification trees. Thus a text-classification tree is a binary tree where each internal node is labeled with a string and each leaf is labeled with a class name. Each text-classification tree classifies an input string as follows. An input string determines a unique path from the root to a leaf: at each internal node the right (respectively left) edge to a child is taken if the input string contains the string labeled at that internal node as a substring (respectively does not contain the labeled string). The class that the input string is classified into is the class at the leaf reached.

2 Learning text-classification trees

We propose an algorithm for learning text-classification trees. The specific features of our learning algorithm are as follows:

1. The algorithm constructs a text-classification tree in a top-down manner started from the root node inductively from the given sample (Top-Down Induction of Decision Tree).
2. The algorithm uses the Quinlan's (ID3) entropy function as the attribute selection function.
3. The algorithm is robust for classification noise contained in the sample.
4. The algorithm does not need any natural language processing technique.

3 Experimental results

Our experiment has been done on the problem of learning text-classification trees for classifying books according to their titles.

Our data consist of a set of pairs of the title of a book and its category in some library. By feeding the data to our learning algorithm, we get a text-classification tree as the output of the learning algorithm. Figure 1 shows a part of the learned text-classification tree output by our learning algorithm, and shows the ratios of test data correctly classified by learned text-classification trees.

![Figure 1: A part of the learned text-classification tree and the ratios of test data correctly classified by it.](image-url)