Analysis of Textual Data Based on Inductive Learning Techniques

Shigeaki Sakurai, IT Research and Development Center, Toshiba Solutions Corporation, Tokyo, Japan

ABSTRACT

This paper introduces knowledge discovery methods based on inductive learning techniques from textual data. The author argues three methods extracting features of the textual data. First one activates a key concept dictionary, second one does a key phrase pattern dictionary, and third one does a named entity extractor. These features are used in order to generate rules representing relationships between the features and text classes. The rules are described in the format of a fuzzy decision tree. Also, these features are used in order to acquire a classification model based on SVM (Support Vector Machine). The model can classify new textual data into the text classes with high classification accuracy. Lastly, this paper introduces two application tasks based on these methods and verifies the effect of the methods.

Keywords: Fuzzy Decision Tree, Inductive Learning Techniques, Key Concept Dictionary, Key Phrase Pattern, Knowledge Discovery Methods, Named Entity Extractor, Support Vector Machine (SVM), Textual Data

INTRODUCTION

Large amounts of textual data, such as daily business reports, e-mail, and electronic newspapers, can be stored easily on computers, owing to the dramatic progress of computer environments and network environments. The textual data includes various kinds of knowledge. The knowledge can facilitate decision making in many situations; therefore, knowledge discovery from the textual data is significant. However, it is difficult to discover the knowledge because of the huge amounts of textual data and it is impracticable to thoroughly investigate all the textual data. Methods are needed that facilitate knowledge discovery. This paper focuses on a method of knowledge discovery described by a rule set. The rule set can reveal relationships between the features of the textual data and text classes in which the data is classified. Analysts can visually grasp the relationships. Also, the rule set can be activated in order to classify
newly collected textual data into appropriate text classes. However, some papers (Cardoso-Cachopo et al., 2003; Joachims, 1998; Yang & Lin, 1999) show that SVM (Support Vector Machine) (Vapnik, 1995) gives best performance in a viewpoint of classification accuracy. SVM inductively acquires hyperplanes which are able to separate text classes. But, they are not always visualized easily. Therefore, it is necessary to select appropriate methods according to the aim of analysis for the textual data. Thus, this paper introduces both a method based on the rule set and a method based on SVM. Firstly, it introduces the format of the textual data. Next, it introduces a fuzzy decision tree and its inductive learning method. The fuzzy decision tree is one of the methods representing a rule set. It can deal with ambiguity included in the textual data. Also, this paper introduces a method acquiring a classification model based on SVM. In addition, this paper focuses on three methods in order to extract features from the data. First one is based on a key concept dictionary, second one is based on a key phrase pattern dictionary, and the third one is based on a named entity extractor. The named entity extractor can acquire proper nouns such as person’s name or place name, and time or date expression from the textual data. Lastly, this paper introduces two application tasks in order to verify introduced methods.

BACKGROUND

Rule discovery methods have been studied since the start of research into artificial intelligence in the field of machine learning. These studies have yielded many techniques, such as decision tree, neural network, genetic algorithm, and association rules, which acquire a rule set from the structured data. A decision tree can describe a rule set in the format of a tree structure. The tree is regarded as the set of IF-THEN rules. C4.5 (Quinlan, 1992) is one example of the algorithms that acquire a compact tree with comparatively high classification accuracy from the structured data. Each item of the data is composed of attribute values and a class. The algorithm uses an information criterion to effectively acquire the tree. A neural network can describe a rule set in the format of a network structure. The network stores relationships between attributes and classes as weights of the links in the network. The weights are appropriately adjusted by the back propagation algorithm. A genetic algorithm (Holland, 1992) inspired by the concept of evolution can acquire a rule set from structured data. The algorithm describes a rule or a rule set as a solution. The algorithm repeatedly improves a solution set to acquire the optimum solution by using three operations: cross-over, mutation, and selection. Association rules (Agrawal & Srikant, 1994) can describe relationships between items. In the case of the retail field, an item is a product item in a receipt. If an item set is frequent, its subsets are frequent. This is called the Apriori property. The association rules can be discovered by expanding small item sets to big item sets including small ones based on the property.

Even if these techniques are important for the rule discovery, they cannot directly deal with the textual data because the textual data is not structured. It is necessary to deal with the textual data by extracting its structured features to acquire a rule set from the textual data. A key point of the extraction is the ambiguity of textual data. That is, the same words and phrases can represent different meanings. Also, different words and phrases can represent similar meanings. In addition, even if the same textual data is given, its interpretation depends on a human. It is necessary to overcome the ambiguity. Thus, fuzzy set theory (Zadeh, 1978) is activated, because it can represent the ambiguity by defining appropriate membership functions. Rule discovery methods based on fuzzy set theory are prominent. Analysts can directly grasp meanings buried in the textual data by checking the rules.
Comparing Different Sparse Matrix Storage Structures as Index Structure for Arabic Text Collection
[www.igi-global.com/article/comparing-different-sparse-matrix-storage/74784?camid=4v1a](www.igi-global.com/article/comparing-different-sparse-matrix-storage/74784?camid=4v1a)