Chapter XXV
Rule Discovery from Textual Data

Shigeaki Sakurai
Toshiba Corporation, Japan

ABSTRACT

This chapter introduces knowledge discovery methods based on a fuzzy decision tree from textual data. The author argues that the methods extract features of the textual data based on a key concept dictionary, which is a hierarchical thesaurus, and a key phrase pattern dictionary, which stores characteristic rows of both words and parts of speech, and generate knowledge in the format of a fuzzy decision tree. The author also discusses two application tasks. One is an analysis system for daily business reports and the other is an e-mail analysis system. The author hopes that the methods will provide new knowledge for researchers engaged in text mining studies, facilitating their understanding of the importance of the fuzzy decision tree in processing 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 tex-

tual 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. Thus, this chapter focuses on a method of knowledge discovery described by a rule set, that is, a rule discovery method. The rule set can classify the textual data based on viewpoints of the analysis. Also, it can reveal relationships between the features of the textual data, which constitute knowledge.

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 the 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 high classification efficiency 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 the relationships between attributes and classes as weights of the arcs in the network. The weights are appropriately adjusted by the back propagation algorithm. A genetic algorithm 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 can describe relationships between items. 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.

These techniques are important for the rule discovery, but 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, we employ fuzzy set theory, because fuzzy set theory can describe ambiguity by defining appropriate membership functions. We introduce rule discovery methods based on fuzzy set theory.

On the other hand, we need to grasp the meaning of discovered rules in order to check their validity and to gain new knowledge from them. Rules described in a visible format are required. Thus, we employ a decision tree, because the tree is an IF-THEN rule set and we intuitively grasp the meaning of rules by looking through attribute values in the IF-part and classes in the THEN-part. We introduce rule discovery methods based on the decision tree.

As anticipated in the above introduction, this chapter focuses on rule discovery methods from textual data based on a fuzzy decision tree. The tree expands the concept of the previous decision tree by incorporating the concept of fuzzy set theory. In this chapter, first, we introduce the format of the textual data. Next, we introduce the format of the fuzzy decision tree, its inductive learning method, and the inference method using it. In addition, we introduce two methods of extracting features included in the textual data and the rule discovery methods based on the features. One method is based on a key concept dictionary and the other method is based on a key phrase