Data Mining Tools:
Formal Concept Analysis and Rough Sets

D

Sanjiv K. Bhatia
University of Missouri, USA

Jitender S. Deogun
University of Nebraska, USA

INTRODUCTION

In the previous chapter, we examined some fundamental techniques for data mining. We'll continue in this chapter by looking at other techniques, such as formal concept analysis, Bayesian classification, and rough set theory.

As the storage costs drop, the new trend is to store all the pieces of data and extract what is needed at the time of need. This is especially true with entities that have the wherewithal to perform such a task, such as government and large corporations. We have seen an example of this in the case of the bombings in Boston Marathon on April 15, 2013 where FBI analyzed the video footage from multiple sources in three days to identify the perpetrators.

BACKGROUND

We have already seen an example of dimensionality reduction in the previous chapter where the time of crime is discretely divided into four hour intervals. It is natural to think whether some of those intervals may overlap. For example, a crime occurring at 11:45pm may be classified as night or late night. Humans do not think in absolute discrete terms but more in fuzzy terms which can make the analysis somewhat harder. In this chapter, we'll look at ways to handle such classifications using formal concept analysis, Bayesian classification, and rough set theory.

MAIN FOCUS OF THE CHAPTER

Formal Concept Analysis and Data Mining

Formal concept analysis (FCA) can be used to derive conceptual structures, analyze complex structures, and discover data dependencies (Wille, 1989; Wille, 2005). FCA is useful in data mining in two ways. First, it provides tools for formal representation of knowledge in an efficient manner. Second, it helps to formalize the conceptual knowledge discovery for different data mining tasks. FCA is increasingly applied in conceptual clustering, data analysis, information retrieval, knowledge discovery, and ontology engineering. Though different from first order logic, FCA emphasizes inter-subjective communication and argumentation. FCA also facilitates importation of the notion of a concept into the modeling of knowledge discovery in databases (KDD).

Formal concept analysis is based on the notions of *formal context* and *formal concept*. A *formal context* is a binary relation between a set of objects and a set of attributes. A formal context provides logic representation of a data set and is used to extract formal concepts.

A *formal concept* is a pair of *intent* and *extent* (Saquer & Deogun, 1999). Intent is a set of features possessed by each object. The extent represents the set of all objects that belong to the concept. These objects share the features from intent. Given a set of features in intent, we can find objects that share the set or subset of features that are shared

DOI: 10.4018/978-1-4666-5202-6.ch060

by the candidates in the extent. There may exist some indiscernible objects in the extent; such objects can be classified using concept learning from formal concept analysis.

Formal Context

A formal context is defined by a triplet (*O,A,R*), where *O* and *A* are two finite and nonempty sets, namely the object set and the attribute set. The relationships between objects and attributes are described by a binary relation *R* between *O* and *A*, which is a subset of the Cartesian product $O \times A$. If an object O_x possesses an attribute A_y, we denote it as $(O_x,A_y) \in R$, or $O_x R A_y$.

Based on the definition of formal context, we know that an object $O_x \in O$ has a set of attributes:

$$O_x R = \{A_y \in A \mid O_x R A_y\} \subseteq A$$

and an attribute A_y is possessed by the set of objects:

$$RA_y = \{O_x \in O \mid O_x R A_y\} \subseteq O$$

To perform FCA, we first define a set-theoretic operator "*" to associate the subset of objects and attributes mutually in a formal context (*O,A,R*).

$$\begin{aligned} X^* &= \{A_y \in A \mid \forall O_x \in O(O_x \in X \Rightarrow O_x R A_y)\} \\ &= \{A_y \in A \mid X \subseteq RA_y\} \\ &= \bigcap_{A_y \in Y} RA_y \end{aligned}$$

$$\begin{aligned} X^* &= \{A_y \in A \mid \forall O_x \in O(O_x \in X \Rightarrow O_x R A_y)\} \\ &= \{A_y \in A \mid X \subseteq RA_y\} \\ &= \bigcap_{A_y \in Y} RA_y \end{aligned}$$

This shows that the "*" operator associates a subset of attributes *X** to the subset of objects *X*. Similarly, for any subset of attributes $Y \subseteq A$, we can associate a subset of objects $Y^* \subseteq O$ as follows:

$$\begin{aligned} Y^* &= \{O_x \in O \mid \forall A_y \in A(A_y \in Y \Rightarrow O_x R A_y)\} \\ &= \{O_x \in O \mid Y \subseteq O_x R\} \\ &= \bigcap_{A_y \in Y} RA_y \end{aligned}$$

The "*" operation induces the following attributes: for $X, X_1, X_2 \subseteq O$ and, $Y_1, Y_2 \subseteq A$,

1. $X_1 \subseteq X_2 \Rightarrow X_1^* \supseteq X_2^*$
   $Y_1 \subseteq Y_2 \Rightarrow Y_1^* \supseteq Y_2^*$
2. $X \subseteq X^{**}$
   $Y \subseteq Y^{**}$
3. $X^{***} = X^*$
   $Y^{***} = Y^*$
4. $(X_1 \cup X_2)^* = X_1^* \cap X_2^*$
   $(Y_1 \cup Y_2)^* = Y_1^* \cap Y_2^*$

A pair of mappings is called a Galois connection if it satisfies (1) and (2), and hence (3). By definition, $O_x^* = O_x R$ is the set of attributes possessed by O_x, and $A_y^* = RA_y$ is the set of objects having attributes A_y. For a set of objects *X*, *X** is the maximal set of attributes shared by all objects in *X*. Similarly, for a set of attributes *Y*, *Y** is the maximal set of objects that have all attributes in *Y* (Yao & Chen, 2006).

Formal Concept

A pair (O_x,A_y), $O_x \subseteq O$, $A_y \subseteq A$, is a formal concept if $O_x = A_y^*$ and $A_y = O_x^*$. O_x is called the *extent* of the concept and A_y is called the *intent* of the concept.

By attribute (3), for any subset $X \subseteq O$, we have a formal concept (*X***,*X**), and for any subset $Y \subseteq A$, we have a formal concept (*Y**,*Y***). With formal concepts, given either a set of attributes or objects, we can directly know all the objects that share the set of attributes or the common attributes that are possessed by a set of objects. All of the formal concepts form a complete lattice, known as a *concept lattice* (Ganter & Wille, 1997). The *meet* (*infima* or *greatest lower bound*) and