Formal Concept Analysis Based Clustering

Jamil M. Saquer
Southwest Missouri State University, USA

INTRODUCTION

Formal concept analysis (FCA) is a branch of applied mathematics with roots in lattice theory (Wille, 1982; Ganter & Wille, 1999). It deals with the notion of a concept in a given universe, which it calls context. For example, consider the context of transactions at a grocery store where each transaction consists of the items bought together. A concept here is a pair of two sets (A, B). A is the set of transactions that contain all the items in B and B is the set of items common to all the transactions in A. A successful area of application for FCA has been data mining. In particular, techniques from FCA have been successfully used in the association mining problem and in clustering (Kryszkiewicz, 1998; Saquer, 2003; Zaki & Hsiao, 2002). In this article, we review the basic notions of FCA and show how they can be used in clustering.

BACKGROUND

A fundamental notion in FCA is that of a context, which is defined as a triple (G, M, I), where G is a set of objects, M is a set of features (or attributes), and I is a binary relation between G and M. For object g and feature m, gIm if and only if g possesses the feature m. An example of a context is given in *Table 1*, where an "X" is placed in the i^{th} row and j^{th} column to indicate that the object in row i possesses the feature in column j.

The set of features common to a set of objects A is denoted by $\beta(A)$ and is defined as $\{m \in M \mid gIm\ \forall g \in A\}$.

Table 1. A context excerpted from (Ganter & Wille, 1999, p. 18) a = needs water to live; b = lives in water; c = lives on land; d = needs chlorophyll; e = two seeds leaf; f = one seed leaf; g = can move around; h = has limbs; i = suckles its offsprings

		a	b	c	d	e	f	g	h	i
1	Leech	X	X					X		
2	Bream	X	X					X	X	
3	Frog	X	X	X				X	X	
4	Dog	X		X				X	X	X
5	Spike-weed	X	X		X		X			
6	Reed	X	X	X	X		X			
7	Bean	X		X	X	X				
8	Maize	X		X	X		X			

Similarly, the set of objects possessing all the features in a set of features B is denoted by $\alpha(B)$ and is given by $\{g \in G \mid gIm\ \forall m \in B\}$. The operators α and β satisfy the assertions given in the following lemma.

Lemma 1 (Wille, 1982): Let (G, M, I) be a context. Then the following assertions hold:

1. $A_1 \subseteq A_2$ implies $\beta(A_2) \subseteq \beta(A_1)$ for every $A_1, A_2 \subseteq G$, and $B_1 \subseteq B_2$ implies $\alpha(B_2) \subseteq \alpha(B_1)$ for every $B_1, B_2 \subseteq M$.
2. $A \subseteq \alpha(\beta(A)$ and $A = \beta(\alpha(\beta(A))$ for all $A \subseteq G$, and $B \subseteq \beta(\alpha((B))$ and $B = \alpha(\beta(\alpha(B)))$ for all $B \subseteq M$.

A formal concept in the context (G, M, I) is defined as a pair (A, B) where $A \subseteq G, B \subseteq M, \beta(A) = B$, and $\alpha(B) = A$. A is called the extent of the formal concept and B is called its intent. For example, the pair (A, B) where $A = \{2, 3, 4\}$ and $B = \{a, g, h\}$ is a formal concept in the context given in *Table 1*. A subconcept/superconcept order relation on concepts is as follows: $(A_1, B_1) \leq (A_2, B_2)$ iff $A_1 \subseteq A_2$ (or equivalently, iff $B_2 \subseteq B_1$). The fundamental theorem of FCA states that the set of all concepts on a given context is a complete lattice, called the concept lattice (Ganter & Wille, 1999). Concept lattices are drawn using Hasse diagrams, where concepts are represented as nodes. An edge is drawn between concepts C_1 and C_2 iff $C_1 \leq C_2$ and there is no concept C_3 such that $C_1 \leq C_3 \leq C_2$. The concept lattice for the context in *Table* 1 is given in *Figure 1*.

A less condensed representation of a concept lattice is possible using reduced labeling (Ganter & Wille, 1999). *Figure 2* shows the concept lattice in *Figure 1* with reduced labeling. It is easier to see the relationships and similarities among objects when reduced labeling is used. The extent of a concept C in *Figure 2* consists of the objects at C and the objects at the concepts that can be reached from C going downward following descending paths towards the bottom concept. Similarly, the intent of C consists of the features at C and the features at the concepts that can be reached from C going upwards following ascending paths to the top concept.

Consider the context presented in *Table 1*. Let $B = \{a, f\}$. Then, $\alpha(B) = \{5, 6, 8\}$, and $\beta(\alpha(B)) = b(\{5, 6, 8\}) = \{a, d, f\} \neq \{a, f\}$; therefore, in general, $\beta(\alpha(B)) \neq B$. A set of features B that satisfies the condition $b(\alpha(B)) = B$ is called a closed feature set. Intuitively, a closed feature set is a

Figure 1. Concept lattice for the context in Table 1

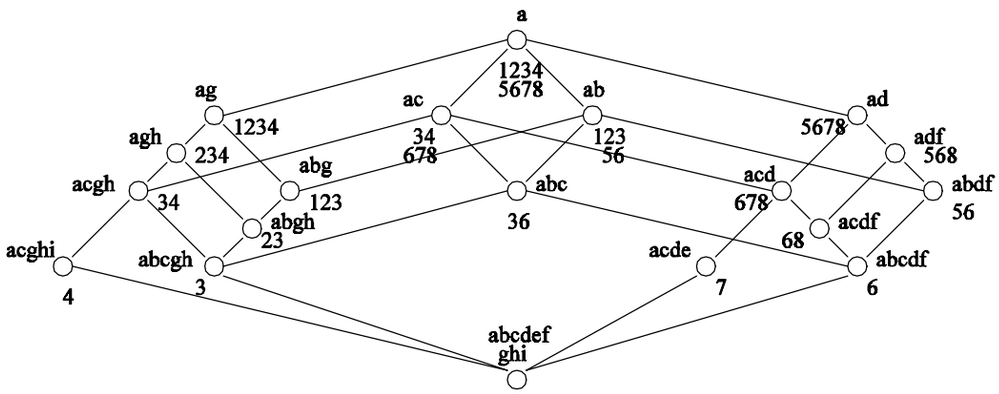
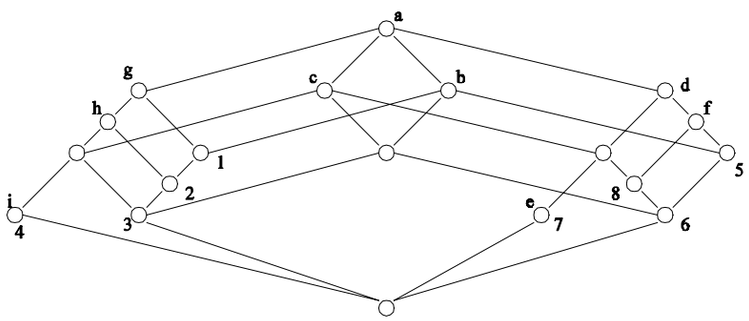


Figure 2. Concept lattice for the context in Table 1 with reduced labeling



maximal set of features shared by a set of objects. It is easy to show that intents of the concepts of a concept lattice are all closed feature sets.

The support of a set of features B is defined as the percentage of objects that possess every feature in B. That is, support(B) = $|\alpha(B)|/|G|$, where |B| is the cardinality of B. Let minSupport be a user-specified threshold value for minimum support. A feature set B is frequent iff support(B) ≥ minSupport. A frequent closed feature set is a closed feature set, which is also frequent. For example, for minSupport = 0.3, {a, f} is frequent, {a, d, f} is frequent closed, while {a, c, d, f} is closed but not frequent.

CLUSTERING BASED ON FCA

It is believed that the method described below is the first for using FCA for disjoint clustering. Using FCA for conceptual clustering to gain more information about data is discussed in Carpineto & Romano (1999) and Mineau & Godin (1995). In the remainder of this article we show how FCA can be used for clustering.

Traditionally, most clustering algorithms do not allow clusters to overlap. However, this is not a valid assumption for many applications. For example, in Web documents clustering, many documents have more than one topic and need to reside in more than one cluster (Beil, Ester, & Xu, 2002; Hearst, 1999; Zamir & Etzioni, 1998). Similarly, in the market basket data, items purchased in a transaction may belong to more than one category of items.

The concept lattice structure provides a hierarchical clustering of objects, where the extent of each node could be a cluster and the intent provides a description of that cluster. There are two main problems, though, that make it difficult to recognize the clusters to be used. First, not all objects are present at all levels of the lattice. Second, the presence of overlapping clusters at different levels is not acceptable for disjoint clustering. The techniques described in this chapter solve these problems. For example, for a node to be a cluster candidate, its intent must be frequent (meaning a minimum percentage of objects must possess all the features of the intent). The intuition is that the objects within a cluster must contain many