Michal Krupka, Jan Laštovička, Petr Osička: Attribute dependencies in incomplete binary data

Department of Computer Science, Palacky University, Olomouc (17. listopadu 12, CZ–77146 Olomouc, Czech Republic)

Introduction

In practical applications of methods of data analysis it is often the case that the data are incomplete. It is therefore desirable for the methods to be able to cope with such data. We focus on formal concept analysis (FCA), a particular method of analysis of binary tabular data that also allows to process incomplete datasets. We study a mechanism in which the most relevant part of FCA's output can be filtered out. The mechanism utilizes so-called attribute dependencies, an additional data supplied by the user to specify what is considered as relevant output.

Formal concept analysis

For thorough material we refer the reader to [3].

A formal context is a tuple $\langle X, Y, I\rangle$ where X and Y are finite non-empty sets (of objects and attributes, respectively), and I is a binary relation in between X and Y.

A formal context $\langle X, Y, I\rangle$ induces a pair of mappings: ${}^{\uparrow}: 2^X \to 2^Y$ and ${}^{\downarrow}: 2^Y \to 2^X$ that are for $A \in X$ and $B \in Y$ given by

$$A^{\uparrow} = \{y \in Y \mid \langle x, y\rangle \in I \text{ for all } x \in A\},$$
$$B^{\downarrow} = \{x \in X \mid \langle x, y\rangle \in I \text{ for all } y \in B\}.$$

A formal concept is a pair $\langle A, B\rangle$ consisting of a set of objects A and a set of attributes B such that $A^{\uparrow} = B$ and $B^{\downarrow} = A$.

The set of all formal concepts present in formal context $\langle X, Y, I\rangle$ is called a concept lattice of $\langle X, Y, I\rangle$, denoted by $\mathcal{B}(X, Y, I)$. Under a partial order given by

$$\langle A_1, B_1\rangle \leq \langle A_2, B_2\rangle \quad \text{iff} \quad A_1 \subseteq A_2 \quad (\text{iff} \quad B_2 \subseteq B_1),$$

the concept lattice $\mathcal{B}(X, Y, I)$ is indeed a complete lattice.

Atrribute dependencies

The extension of FCA by attribute dependencies (AD) utilizes background knowledge concerning the relative importance of attributes to filter out formal concepts that are the most relevant for the user. The method has been elaborated in [1].

An AD-formula over a set Y is an expression $C \sqsubset D$, where $C, D \subseteq Y$. The set of all possible AD-formulas over a set of attributes Y is denoted by $ADF(Y)$.

An AD-formula $C \sqsubset D$ is valid in a set E of attributes, denoted by $||C \sqsubset D||_E = 1$ iff

$$C \cap E \neq \emptyset \text{ implies } D \cap E \neq \emptyset.$$

A formal concept $\langle A, B\rangle$ is compatible with $C \sqsubset D$ if $C \sqsubset D$ is valid in B. A formal concept is compatible with a set T of AD-formulas if it is compatible with every AD-formula of T. The set of all formal concepts of a formal context $\langle X, Y, I\rangle$ compatible with T is denoted by $\mathcal{B}_T(X, Y, I)$.

Incomplete contexts

We briefly describe the framework we use for dealing with formal contexts with incomplete data and their concept lattices. More details can be found in [4].

Boolean algebras with variables

A Boolean algebra with variables is a finite Boolean algebra $\mathbf{L}$ together with a distinguished set $U = \{u_1, \dots, u_k\} \subseteq L$ such that $\mathbf{L}$ is generated by U. Elements of U are called variables.

Mappings $v: U \to \mathbf{2}$ are called assignments. Since $\mathbf{L}$ is generated by U, then each assignment v can be extended in at most one way to a homomorphism $\bar{v}: \mathbf{L} \to \mathbf{2}$. If $\bar{v}$ exists, the assignment v is called admissible.

Concept lattices of incomplete contexts

We use results of formal concept analysis in fuzzy setting [2].

An incomplete $\mathbf{L}$-context is a triple $\langle X, Y, I\rangle$, where X and Y are sets and $I \in L^{X\times Y}$ is an $\mathbf{L}$-relation such that $I(X \times Y) \subseteq U \cup \{0, 1\}$. An ordinary formal context $\langle X, Y, J\rangle$ is a v-completion of $\langle X, Y, I\rangle$, if v is an admissible assignment and $J = \bar{v} \circ I$.

For any $A \in L^X$ and $B \in L^Y$ we set

$$A^{\uparrow}(y) = \bigwedge_{x\in X} \big(A(x) \to I(x, y)\big),$$
$$B^{\downarrow}(x) = \bigwedge_{y\in Y} \big(B(y) \to I(x, y)\big).$$

An incomplete formal $\mathbf{L}$-concept of the incomplete $\mathbf{L}$-context $\langle X, Y, I\rangle$ is a pair $\langle A, B\rangle \in L^X \times L^Y$ where $A^{\uparrow} = B$ and $B^{\downarrow} = A$. The set of all incomplete formal concepts of $\langle X, Y, I\rangle$ is denoted $\mathcal{B}(X, Y, I)$.

The condition

$$\langle A_1, B_1\rangle \leq \langle A_2, B_2\rangle \quad \text{iff} \quad A_1 \subseteq A_2 \quad (\text{iff} \quad B_2 \subseteq B_1)$$

defines a partial ordering on $\mathcal{B}(X, Y, I)$. Together with this ordering, $\mathcal{B}(X, Y, I)$ is a complete lattice.

In addition to the partial ordering, we have a binary $\mathbf{L}$-relation $\preceq$ on $\mathcal{B}(X, Y, I)$, defined by

$$(\langle A_1, B_1\rangle \preceq \langle A_2, B_2\rangle) = S(A_1, A_2) \quad (= S(B_2, B_1)).$$

Consider an admissible assignment v and the v-completion $\langle X, Y, \bar{v} \circ I\rangle$ of $\langle X, Y, I\rangle$. The assignment v induces a homomorphism $\bar{v}^{\mathcal{B}(X,Y,I)}: \mathcal{B}(X, Y, I) \to \mathcal{B}(X, Y, \bar{v} \circ I)$ of complete lattices given by

$$\bar{v}^{\mathcal{B}(X,Y,I)}(\langle A, B\rangle) = \langle \bar{v} \circ A, \bar{v} \circ B\rangle.$$

AD in incomplete contexts

Let $\mathbf{L}$ be a Boolean algebra, $C, D \subseteq Y$ and $B \in \mathbf{L}^Y$.
The validity $||C \sqsubset D||_B$ of the formula $C \sqsubset D$ in B is defined as

$$||C \sqsubset D||_B = \bigvee_{y\in C} B(y) \to \bigvee_{y\in D} B(y)$$

$\mathbf{L}$-theory is an $\mathbf{L}$-set of AD-formulas.
The validity $||T||_B$ of $\mathbf{L}$-theory T in B is defined as

$$||T||_B = \bigwedge_{C\sqsubset D\in \mathrm{ADF(Y)}} \big(T(C \sqsubset D) \to ||C \sqsubset D||_B\big),$$

Let h be a homomorphism of $\mathbf{L}$ to the two element Boolean algebra $\mathbf{2}$.

Lemma 1 *For every AD-formula $C \sqsubset D$ and $B \in L^Y$ it holds*

$$h(||C \sqsubset D||_B) = ||C \sqsubset D||_{h\circ B}.$$

Lemma 2 *For every $\mathbf{L}$-theory T and $B \in L^Y$ it holds*

$$h(||T||_B) = ||h \circ T||_{h\circ B}.$$

Let $\langle X, Y, I\rangle$ be an incomplete $\mathbf{L}$-context, F be a subset of L, and T be an $\mathbf{L}$-theory. An incomplete formal $\mathbf{L}$-concept $\langle A, B\rangle \in \mathcal{B}(X, Y, I)$ is said to be F-compatible with T if $||T||_B \in F$. We denote the set of all formal concepts of $\langle X, Y, I\rangle$ that are F-compatible with T by $\mathcal{B}_{F,T}(X, Y, I)$.

Theorem 1 *For any admissible assignment v, the corresponding homomorphism $\bar{v}$, and the $\bar{v}$'s preimage of 1 denoted by F it holds*

$$\bar{v}^{\mathcal{B}(X,Y,I)}(\mathcal{B}_{F,T}(X, Y, I)) = \mathcal{B}_{\bar{v}\circ T}(X, Y, \bar{v} \circ I).$$

We show that certain structural properties of the set of formal concepts compatible with a set T of AD-formulas that hold true in the case of complete context also hold true for formal concepts of incomplete context that are compatible with T.

To do so, we utilize a specific instance of first-order fuzzy logic (see [2], chapter 3.2).

The language of our logic contains unary relation symbol r_T, and binary relation symbol $\leq$; binary function symbols $\inf$, and $\sup$; constants c_0 and c_1.

Let $\langle X, Y, I\rangle$ be an $\mathbf{L}$-context and T be an $\mathbf{L}$-theory. The structure corresponding to $\langle X, Y, I\rangle$ and T is an $\mathbf{L}$-structure $\mathbf{M}(\langle X, Y, I\rangle, T)$ with $M = \mathcal{B}(X, Y, I)$, $r_T^{\mathbf{M}}(\langle A, B\rangle) = ||T||_B$, $\leq^{\mathbf{M}} = \preceq$, $c_0^{\mathbf{M}} = \langle Y^{\downarrow}, Y\rangle$, $c_1^{\mathbf{M}} = \langle X, X^{\uparrow}\rangle$, $\inf^{\mathbf{M}} = \wedge$, $\sup^{\mathbf{M}} = \vee$.

Let φ be a formula, $\langle X, Y, I\rangle$ be an incomplete context, and $\bar{v}$ be a homomorphism corresponding to an admissible assignment. Furthemore, consider structures $\mathbf{M}_1 = \mathbf{M}(\langle X, Y, I\rangle, T)$ and $\mathbf{M}_v = \mathbf{M}(\langle X, Y, \bar{v} \circ I\rangle, \bar{v} \circ T)$.

Lemma 3 *For any $\mathbf{M}_1$-valuation e it holds*

$$\bar{v}(||\varphi||^{\mathbf{L}}_{\mathbf{M}_1,e}) = ||\varphi||^{\mathbf{2}}_{\mathbf{M}_v,e_v},$$

where e_v is the composition of e and $\bar{v}^{\mathcal{B}(X,Y,I)}$.

Theorem 2 *For any $\mathbf{M}_1$-valuation e it holds*

$$||\varphi||^{\mathbf{L}}_{\mathbf{M}_1,e} = 1 \textit{ iff } ||\varphi||^{\mathbf{2}}_{\mathbf{M}_v,e_v} = 1$$

for each admissible assignment v.

The following results are applications of Theorem 2. Using results known for the ordinary case, we prove their analogies for the case of incomplete data.

Corollary 1 *Let T be an $\mathbf{L}$-theory. Then, $\mathcal{B}_{F,T}(X, Y, I)$ where $F \subseteq \mathbf{L}$ with $1 \in F$ has the least element.*

Corollary 2 *Let T be an $\mathbf{L}$-theory consisting of AD-formulas of the form $\{a\} \sqsubset \{b\}$. Then, $\mathcal{B}_{\{1\},T}(X, Y, I)$ is a complete lattice which is a $\bigvee$-sublattice of $\mathcal{B}(X, Y, I)$.*

We call attributes $y_1, y_2 \in Y$ disjoint w.r.t. an $\mathbf{L}$-context $\langle X, Y, I\rangle$ if $\{y_1\}^{\downarrow_I} \cap \{y_2\}^{\downarrow_I} = \emptyset$. We call an $\mathbf{L}$-theory T tree theory w.r.t. an $\mathbf{L}$-context $\langle X, Y, I\rangle$, if for each $y_1, y_2 \in Y$ that are not disjoint w.r.t. $\langle X, Y, I\rangle$, T contains to the degree 1 an AD-formula $\{y_1\} \sqsubset Y_2$ or $\{y_2\} \sqsubset Y_1$ where $y_1 \in Y_1$ and $y_2 \in Y_2$ and attributes in both Y_1 and Y_2 are pairwise disjoint w.r.t. $\langle X, Y, I\rangle$.

Corollary 3 *Let T be a tree theory w.r.t. $\langle X, Y, I\rangle$. Then for each $\langle A_1, B_1\rangle, \langle A_2, B_2\rangle \in \mathcal{B}_{\{1\},T}(X, Y, I)$ it holds*
if $\langle A_1, B_1\rangle \preceq \langle A_2, B_2\rangle = 0$ and $\langle A_2, B_2\rangle \preceq \langle A_1, B_1\rangle = 0$, then $\langle A_1, B_1\rangle \wedge_{\mathcal{B}_{\{1\},T}(X,Y,I)} \langle A_2, B_2\rangle = \langle Y^{\downarrow}, Y\rangle$.

Conclusions

We presented foundations of a method of filtering out the interesting formal concepts present in a incomplete datasets. The information specifying which concepts are considered as interesting, is supplied as attribute dependencies, particular formulas that capture relative importance of attributes. We presented basic notions and studied their properties. Furthermore, we proved a theorem that allows us to carry results known for the case of complete data to the case of incomplete data. Such results include mainly the ones concerning a structure of the set of filtered formal concepts.

References

1. Belohlavek, R., Vychodil, V.: Formal concept analysis with background knowledge: attribute priorities. IEEE Transactions on Systems, Man, and Cybernetics, Part C 39(4), 399–409 (2009)
2. Belohlavek, R.: Fuzzy Relational Systems: Foundations and Principles. Kluwer Academic Publishers, Norwell, USA (2002)
3. Ganter, B., Wille, R.: Formal Concept Analysis – Mathematical Foundations. Springer (1999)
4. Krupka, M., Lastovicka, J.: Concept lattices of incomplete data. In: Lecture Notes in Computer Science. vol. 7278, pp. 180–194 (2012)

International Center for Information and Uncertainty
SSIU 2013: International Summer School "Information and Uncertainty"
http://mcin.upol.cz/SSWIU-2013/
INVESTMENTS IN EDUCATION DEVELOPMENT