Concept lattices constrained by systems of partitions

Radim BĚLOHLÁVEK, Vladimír SKLENÁŘ, Jiří ZACPAL

Dept. Computer Science, Palacký University, Tomkova 40, CZ-779 00, Olomouc, Czech Republic, {radim.belohlavek,vladimir.sklenar,jiri.zacpal}@upol.cz

Abstract. We present a method of reduction of the number of clusters in formal concept analysis of binary object-attribute data. In our preliminary study, we use a similarity of objects computed from the object-attribute data table to reduce the number of extracted formal concepts. The method consists in constraining a concept lattice by a hierarchical clustering obtained from the similarity on objects. The level in the corresponding dendrogram can be used as parameter to control the size of the resulting constrained concept lattice.

1 Introduction and problem setting

A formal context is a triplet $\langle X, Y, I\rangle$ where I is a binary relation between X and Y, $\langle x, y\rangle \in I$ meaning that object x has attribute y. For each $A \subseteq X$ and $B \subseteq Y$, denote by $A^{\uparrow}$ and $B^{\downarrow}$ subsets of Y and X, respectivey, defined by $A^{\uparrow} = \{y \mid \text{for each } x \in X : \langle x, y\rangle \in I\}$ and $B^{\downarrow} = \{x \mid \text{for each } y \in Y : \langle x, y\rangle \in I\}$. That is, $A^{\uparrow}$ is the set of all attributes from Y shared by all objects from A (and similarly for $B^{\downarrow}$). A formal concept in $\langle X, Y, I\rangle$ is a pair $\langle A, B\rangle$ of $A \subseteq X$ and $B \subseteq Y$ satisfying $A^{\uparrow} = B$ and $B^{\downarrow} = A$, i.e. A is the set of all objects sharing all attributes from B and, conversely, B is the collection of all attributes from Y shared by all objects from A. The set $\mathcal{B}(X, Y, I) = \{\langle A, B\rangle \mid A^{\uparrow} = B, B^{\downarrow} = A\}$ of all formal concepts in $\langle X, Y, I\rangle$ can be naturally equipped with a partial order $\leq$ (modeling the subconcept-superconcept hierarchy, e.g. $\mathsf{dog} \leq \mathsf{mammal}$) defined by $\langle A_1, B_1\rangle \leq \langle A_2, B_2\rangle$ iff $A_1 \subseteq A_2$ (or, equivalently, $B_2 \subseteq B_1$). Under $\leq$, $\mathcal{B}(X, Y, I)$ happens to be a complete lattice, called a concept lattice, the basic structure of which is described by the so-called main theorem of concept lattices [3].

2 Concept lattices constrained by systems of partitions generated from data

Constraints by equivalence relations A considerabe problem of FCA is a possibly arge number of formal concepts extracted from data. In our previous papers we presented a way to reduce the number of extracted formal concepts by means of constraints. Briefly, a constraint represents an additional (to $\langle X, Y, I\rangle$) information. Only those formal concepts which satisfy a given constraint (satisfiability needs to be defined) are considered interesting and are extracted from data. We now recall the basic concepts of [1] in which we introduced a particular form of a constraint by an equivalence relation.

Definition 1. *A formal context with a binary relation (R-context, for short) is a structure $\langle\langle X, \equiv\rangle, Y, I\rangle$ where $\langle X, Y, I\rangle$ is a formal context and $\equiv$ is a binary relation on X. A formal concept $\langle A, B\rangle \in \mathcal{B}(X, Y, I)$ is called compatible with $\equiv$ if for each $x_1, x_2 \in X$, if $x_1 \in A$, and $x_1 \equiv x_2$ or $x_2 \equiv x_1$, then $x_2 \in A$.*

Remark 1. (1) We are primarily interested in case when $\equiv$ is an equivalence relation. Then $x_1 \equiv x_2$ means that objects x_1 and x_2 are equivalent from some point of view (similar, indistuinguishable). (2) Equivalence $\equiv$ may be supplied by an expert or may result from some previous analysis or external source. The latter option will be the case considered in our paper. (3) Formal concepts compatible with $\equiv$ respect the indistinguishability represented by $\equiv$. By $\mathcal{B}(\langle X, \equiv\rangle, Y, I)$, we denote the collection of all formal concepts of $\langle X, Y, I\rangle$ compatible with $\equiv$.

Constraints by nested systems of partitions generated from data by hierarchical clustering In the following, we use hierarchical clustering to generate a series of nested equivalence relations. For an overview on clustering techniques, theory, and algorithms, we refer to [2]. Recall that methods of hierarchical clustering assume (as an input) a set X of objects and a dissimilarity function $D : X \times X \to [0, \infty)$. The output of hierarchical clustering can be seen as a system $\Pi = \{\Pi_t \mid t = 0, 1, \ldots, n\}$ of nested partitions on X together with a corresponding vector $h = \langle h_0, h_1, \ldots, h_n\rangle$ of dissimilarity indices. h_t can be seen as an index of heterogenity of Π_t. The nesting condition says that for $t < t'$, Π_t is a subpartition of $\Pi_{t'}$, i.e. each class/cluster of Π_t is contained (is a subset) in some class of $\Pi_{t'}$. In terms of the corresponding eqivalence relations, if $t < t'$ then $\theta_{\Pi_t} \subseteq \theta_{\Pi_{t'}}$. A formal context $\langle X, Y, I\rangle$ plus a suitable dissimilarity function D on X can be seen as input to hierarchical clustering. For D, we use a weighted Hamming distance D_w defined by $D_w(x_i, x_j) = \sum_{y\in Y} w_y \cdot |I(x_i, y) - I(x_j, y)|$, where w_y ane nonnegative weights. For instance, for $w_y = 1$ for each $y \in Y$, D_w becomes the simple matching coefficient.

If Π is a partition corresponding to an equivalence relation $\equiv$, i.e. $\Pi = \{[x]_\equiv\}$, we denote $\mathcal{B}(\langle X, \equiv\rangle, Y, I)$ also by $\mathcal{B}(\langle X, \Pi\rangle, Y, I)$. In what follows we list basic properties of the approach described above. They are mainly consequences of those established in [1]. We asusme that we are given a formal context $\langle X, Y, I\rangle$, and a system $\Pi = \{\Pi_t \mid t = 0, 1, \ldots, n\}$ of nested partitions obtained by hierarchical clustering. That is, $\Pi_0 = \{\{x\} \mid x \in X\}$, $\Pi_n = \{X\}$, and for each $t = 0, 1, \ldots, n-1$, Π_{t+1} results from Π_t by joining two classes/clusters of Π_t. The following are the main properties:

- $\mathcal{B}(\langle X, \Pi_t\rangle, Y, I)$'s form a dually nested system of subsets of the concept lattice $\mathcal{B}(X, Y, I)$, i.e. for $t < t'$ we have $\mathcal{B}(\langle X, \Pi_t\rangle, Y, I) \supseteq \mathcal{B}(\langle X, \Pi_{t'}\rangle, Y, I)$. Moreover, $\mathcal{B}(\langle X, \Pi_0\rangle, Y, I) = \mathcal{B}(X, Y, I)$ and $\mathcal{B}(\langle X, \Pi_n\rangle, Y, I) = \{\langle X, X^\uparrow\rangle\}$.

- For each $\langle A, B\rangle \in \mathcal{B}(\langle X, \Pi_t\rangle, Y, I)$, A is a union of some classes from Π_t. For each $t' < t$, A is a union of disjoint extents A_j of some $\langle A_j, B_j\rangle \in \mathcal{B}(\langle X, \Pi_{t'}\rangle, Y, I)$.
- For a formal context $\langle X, Y, I\rangle$ denote by $\cong_X$ the binary relation defined on X by $x_1 \cong_X x_2$ if and only if for each $y \in Y$ we have $\langle x_1, y\rangle \in I$ iff $\langle x_2, y\rangle \in I$. Then, $\mathcal{B}(\langle X, \Pi_t\rangle, Y, I) = \mathcal{B}(X, Y, I)$ if and only if for each $x_1, x_2 \in X$, $x_1, x_2 \in C$ for some $C \in \Pi_t$ implies $x_1 \cong_X x_2$. In such a case, $\mathcal{B}(\langle X, \Pi'_t\rangle, Y, I) = \mathcal{B}(X, Y, I)$ for each $t' < t$.
- Each $\mathcal{B}(\langle X, \Pi_t\rangle, Y, I)$ equipped with $\leq$ inherited from $\mathcal{B}(X, Y, I)$ is a complete lattice in which arbitrary infima coincide with infima in $\mathcal{B}(X, Y, I)$. Therefore, for $t < t'$, $\mathcal{B}(X, \langle Y, \Pi_{t'}\rangle, I)$ complete $\bigwedge$-sublattice of $\mathcal{B}(X, \langle Y, \Pi_t\rangle, I)$.

3 Illustrative examples

Consider the formal context $\langle X, Y, I\rangle$ in Tab. 1 . $\mathcal{B}(X, Y, I)$ is depicted in

Table 1. Money funds data

	fund	type	1 2 3	4 5 6	7 8 9
1	CPI Penezniho trhu	money	1 0 0	0 1 0	0 1 0
2	CSOB Akciovy	stock	1 0 0	0 0 1	0 0 1
3	CSOB Bond mix	bond	0 1 0	1 0 0	0 1 0
4	IKS Dluhopisovy	bond	0 1 0	1 0 0	1 0 0
5	IKS Globalni	mixed	0 1 0	0 1 0	0 1 0
6	IKS Penezni trh	money	1 0 0	0 1 0	0 1 0
7	ISCS Sporoinvest	money	1 0 0	0 1 0	0 1 0
8	ISCS Sporotrend	stock	0 0 1	0 0 1	0 0 1
9	ISCS Trendbond	bond	0 0 1	1 0 0	1 0 0
10	ISCS Vynosovy	mixed	0 0 1	0 1 0	0 1 0

attributes: 1 - rating for 1 week <= $0,5$, 2 - rating for 1 week > $0,5$ and <= 1, 3 - rating for 1 week > 1, 4 - rating for 26 weeks <= $0,5$, 5 - rating for 26 weeks > $0,5$ and <= 4, 6 - rating for 26 weeks > 4, 7 - rating for 52 weeks <= $0,5$, 8 - rating for 56 weeks > $0,5$ and <= 10, 9 - rating for 56 weeks > 10

Fig. 1. First, consider the equivalence relation $\equiv$ given by the type of fund (four $\equiv$-classes corresponding to stock funds, bond funds, mixed funds and money funds). The constrained $\mathcal{B}(\langle X, \equiv\rangle, Y, I)$ contains 6 formal concepts $(\{\langle\{1,2,3,4,5,6,7,8,9,10\},\{\}\rangle, \langle\{2,8\},\{6,9\}\rangle, \langle\{1,5,6,7,10\},\{5,8\}\rangle,$ $\langle\{3,4,9\},\{5\}\rangle, \langle\{1,6,7\},\{1,5,8\}\rangle, \langle\{\},\{1,2,3,4,5,6,7,8,9\}\rangle\})$.

Now, by hierarchical clustering (single-linkage version) using weighted Hamming distance with weights for attributes 1–9 equal to 0.4, 0.3, 0.3, 0.3, 0.5, 0.2, 0.2, 0.6, 0.2, respectively, we get a collection of nested partitions $\Pi_1, \ldots, \Pi_5$ and

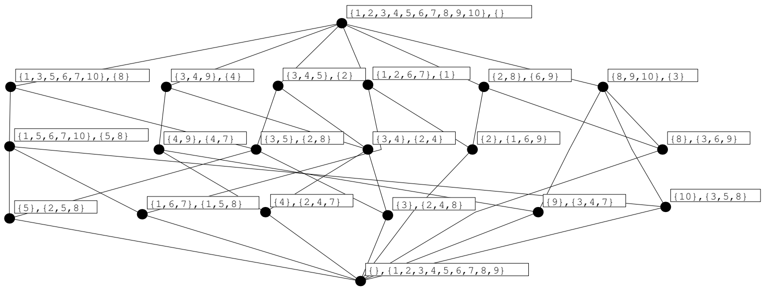

Fig. 1. Concept lattice corresponding to data from Tab. 1

the corresponding equivalence relations$\equiv_1, \ldots, \equiv_5$:

$$
\begin{aligned}
\Pi_1 &= \{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\},\{7\},\{8\},\{9\},\{10\}\},\\
\Pi_2 &= \{\{1\},\{2\},\{3\},\{4\},\{5\},\{8\},\{9\},\{10\},\{6,7\}\},\\
\Pi_3 &= \{\{1\},\{2\},\{3\},\{8\},\{4,9\},\{5,10\},\{6,7\}\},\\
\Pi_4 &= \{\{4,9\},\{1,3\},\{2,8\},\{5,6,7,10\}\},\\
\Pi_5 &= \{\{2,8\},\{1,3,4,5,6,7,9,10\}\},\\
\Pi_6 &= \{\{1,2,3,4,5,6,7,8,9,10\}\}.
\end{aligned}
$$

We obtain the following constrained concept lattices: $\mathcal{B}(\langle X,\Pi_1\rangle,Y,I) = \mathcal{B}(X,Y,I)$; $\mathcal{B}(\langle X,\Pi_2\rangle,Y,I) = \mathcal{B}(X,Y,I)$; $\mathcal{B}(\langle X,\Pi_3\rangle,Y,I) = \{\langle\{1,2,3,4,5,6,7,8,9,10\},\{\}\rangle, \langle\{1,2,6,7\},\{1\}\rangle, \langle\{2,8\},\{6,9\}\rangle, \langle\{3,4,9\},\{4\}\rangle,$ $\langle\{1,3,5,6,7,10\},\{8\}\rangle, \langle\{2\},\{1,6,9\}\rangle, \langle\{8\},\{3,6,9\}\rangle, \langle\{3\},\{2,4,8\}\rangle,$ $\langle\{1,5,6,7,10\},\{5,8\}\rangle, \langle\{4,9\},\{4,7\}\rangle, \langle\{1,6,7\},\{1,5,8\}\rangle,$ $\langle\{\},\{1,2,3,4,5,6,7,8,9\}\rangle\}$; $\mathcal{B}(\langle X,\Pi_4\rangle,Y,I) = \{\langle\{1,2,3,4,5,6,7,8,9,10\},\{\}\rangle,$ $\langle\{2,8\},\{6,9\}\rangle, \langle\{4,9\},\{4,7\}\rangle, \langle\{1,3,5,6,7,10\},\{8\}\rangle, \langle\{\},\{1,2,3,4,5,6,7,8,9\}\rangle\}$; $\mathcal{B}(\langle X,\Pi_5\rangle,Y,I) = \{\langle\{1,2,3,4,5,6,7,8,9,10\},\{\}\rangle, \langle\{2,8\},\{6,9\}\rangle,$ $\langle\{\},\{1,2,3,4,5,6,7,8,9\}\rangle\}$; $\mathcal{B}(\langle X,\Pi_6\rangle,Y,I) = \{\langle X,\emptyset\rangle\}$.

More detailed results will be available in full version of this paper.

Acknowledgement Supported by grant No. 1ET101370417 of the GA AV CR. The research of the first author was supported by grant No. 201/02/P076 of the GAČR.

References

1. Bělohlávek R., Sklenář V., Zacpal J.: Concept lattices constrained by equivalence relations. In: Proc. CLA 2004, pp. 58–66. [ISBN 80-248-0597-9]
2. Everitt, Brian S.: *Cluster Analysis, 4th ed.* Edward Arnold, 2001.
3. Ganter B., Wille R.: *Formal concept analysis. Mathematical Foundations.* Springer-Verlag, Berlin, 1999.