Processing Uncertainty with Interval Pattern Structures

Sergei O. Kuznetsov

National Research University Higher School of Economics,
Kochnovski pr.3, Moscow, Russia, skuznetsov@hse.ru

Pattern structures [1], an extension of Formal Concept Analysis [2], propose a direct way to knowledge discovery in nonbinary data. On the one hand, pattern structures on logical formulas, graphs and strings give a means for relational knowledge discovery. On the other hand, interval pattern structures [3, 4] propose an approach for processing uncertainty in numerical information: For a numerical attribute uncertainty is given in terms of an interval of possible values.

Let G be a set (of objects), let $(D, \sqcap)$ be a meet-semi-lattice (of all possible object descriptions) and let $\delta : G \to D$ be a mapping. Then $(G, \underline{D}, \delta)$, where $\underline{D} = (D, \sqcap)$, is called a *pattern structure*, provided that the set $\delta(G) := \{\delta(g) | g \in G\}$ generates a complete subsemilattice $(D_\delta, \sqcap)$ of $(D, \sqcap)$, i.e., every subset X of $\delta(G)$ has an infimum $\sqcap X$ in $(D, \sqcap)$. Elements of D are called *patterns* and are naturally ordered by subsumption relation $\sqsubseteq$: given $c, d \in D$ one has $c \sqsubseteq d \Leftrightarrow c \sqcap d = c$. A pattern structure $(G, \underline{D}, \delta)$ gives rise to the following derivation operators $(\cdot)^\diamond$:

$$A^\diamond = \sqcap_{g \in A} \delta(g) \qquad \text{for } A \subseteq G,$$
$$d^\diamond = \{g \in G \mid d \sqsubseteq \delta(g)\} \qquad \text{for } d \in (D, \sqcap)$$

These operators form a Galois connection between the powerset of G and $(D, \sqsubseteq)$. The pairs (A, d) satisfying $A \subseteq G$, $d \in D$, $A^\diamond = d$, and $A = d^\diamond$ are called the *pattern concepts* of $(G, \underline{D}, \delta)$, with *pattern extent* A and *pattern intent* d. Pattern concepts are ordered wrt. set inclusion on extents. The ordered set of pattern concepts makes a lattice, called *pattern concept lattice*. This lattice can be computed with a polynomial delay algorithm CbO [5]. For $a, b \in D$ the *pattern implication* $a \to b$ holds if $a^\diamond \subseteq b^\diamond$, and the *pattern association rule* $a \to_{c,s} b$ with *confidence* c and *support* s holds if $s \leq \frac{|a^\diamond \cap b^\diamond|}{|G|}$ and $c \leq \frac{|a^\diamond \cap b^\diamond|}{|a^\diamond|}$. Pattern association rules may be derived from a concise representation given by the edges of the diagram of the pattern concept lattice. Let us have a set of positive examples G_+ and a set of negative examples G_- w.r.t. a *target attribute*, $G_+ \cap G_- = \emptyset$.

A pattern $p \in D$ is a *k-weak positive premise (classifier)* iff

$$| p^\diamond \cap E_- | \leq k \text{ and } \exists A \subseteq E_+ : p \sqsubseteq A^\diamond$$

A k-weak positive premise is a *k-weak positive hypothesis* if $\exists A \subseteq E_+ : h = A^\diamond$. A weak k-hypothesis is the *least general generalization* of descriptions of positive examples, which is contained in no more than k negative examples. *Negative* premises and hypotheses are defined similarly. Various classification schemes using premises are possible, including very efficient “lazy” ones [6].

Now we come to interval pattern structures. For two intervals $[a_1,\ b_1]$ and $[a_2,\ b_2]$, with $a_1,\ b_1,\ a_2,\ b_2 \in \mathbb{R}$, their meet is defined as

$$[a_1,\ b_1] \sqcap [a_2,\ b_2] = [\min(a_1,\ a_2),\ \max(b_1,\ b_2)].$$

This operator is obviously idempotent, commutative and associative, thus defining a pattern structure on tuples (vectors) of intervals of attribute values of a numerical object-attribute table (G, M, W, J), where G is the set of objects, M is the set of attributes, $W \subseteq \mathbb{R}$ is the set of attribute values, and $J \subseteq G \times M \times W$ gives the table itself. The lattice of interval pattern structure $(G, (Int, \sqcap), \delta)$ is isomorphic to the concept lattice of the context that arises from the *interordinal scaling* of the initial many-valued numerical context [2], where for each table value a two binary attributes $\geq a$ and $\leq a$ are introduced. However, interval tuples give better understanding of results and computation with them is faster than that with the interordinal scaling [3].

A bicluster of ε-similar values is an inclusion-maximal subtable of numerical object-attribute matrix (G, M, W, J), such that any two entries of it differ no more than in ε. We show that a bicluster of similar values are closed descriptions in terms of a modified pattern structure on intervals. Namely, each object of this pattern structure corresponds to a row of the original table and has description consisting of all maximal pairs of the form $(B, [a, b])$, where $B \subseteq M,\ a, b \in \mathbb{R}$, $|a - b| \leq \varepsilon$, and maximality is taken wrt. partial order defined as $(B_1, [a_1, b_1]) \leq (B_2, [a_2, b_2]) \iff B_1 \subseteq B_2$ and $[a_1, b_1] \supseteq [a_2, b_2]$. Intersection $\sqcap$ of sets of these pairs is defined as the set of all maximal pairwise intersections on pairs wrt. $\leq$ with nonempty set of attributes and intervals not exceeding ε.

We present and discuss results of computer experiments on gene expression data justifying the use of interval patterns for mining biclusters.

References

1. Ganter, B., Kuznetsov, S.O.: *Pattern Structures and Their Projections.* In Proc. ICCS 2001. LNAI, vol. 2120, 129–142 (2001).
2. Ganter, B., Wille, R.: *Formal Concept Analysis: Mathematical Foundations.* Springer, Heidelberg (1999).
3. Kaytoue, M., Kuznetsov, S.O., Napoli, A., Duplessis, S., Mining gene expression data with pattern structures in formal concept analysis, Inf.Sci., 181(10): 1989–2001 (2011).
4. Kaytoue, M., Kuznetsov, S.O., Macko J., Napoli A., Biclustering meets triadic concept analysis. Annals of Mathematics and Artificial Intelligence, 70(1-2) (2014).
5. Kuznetsov, S.O., *Learning of Simple Conceptual Graphs from Positive and Negative Examples.* In: Proc. PKDD 1999. LNAI, vol. 1704, 384–391 (1999).
6. Kuznetsov, S.O., Fitting Pattern Structures to Knowledge Discovery in Big Data. In: Proc. ICFCA 2013, LNAI, vol. 7880, 254–266 (2013).