Relations in GUHA style data mining

Petr Hájek

Institute of Computer Science AS CR,
Pod Vodárenskou věží 2, Prague 8, Czech Republic

Abstract. The formalism of GUHA style data mining is confronted with the approach of relational structures of Düntsch, Orlowska and others. A computational complexity result on tautologies with implicational quantifiers is presented.

1 Introduction

The first aim of the present paper is to compare the logic of observational calculi, presented is depth in the monograph [3] and serving as foundation for the GUHA method of automated generation of hypotheses, (see e.g. [4, 5]) with the formalism of relational structures of Düntsch, Orlowska and others as presented in [7] (see e.g. Chapter 16 by Düntsch and Orlowska). We hope to show possible mutual influence and application. The second aim is to present a result on the computational complexity of an important class of logical formulas called implicational tautologies, i.e. formulas built using a binary *quantifier* $\Rightarrow$ and logically true for each interpretation of $\Rightarrow$ as an implicational quantifier (in the sense of GUHA): the set of all such tautologies is shown to be co-NP-complete.

The two parts of the paper can be read independently of each other; the reader not familiar with GUHA finds in the first part all definitions needed for the second part. Interestingly, the second part makes non-trivial use of mathematical fuzzy logic.

Support of COST Action 274 TARSKI (Theory and Applications of Relational Structures as Knowledge Instruments) is acknowledged. The second part of the paper is also relevant for the grant project A13004/00 of the Grant Agency of the Academy of Sciences of the Czech Republic.

2 Relations – where are they from

One basic notion common to both approaches is that called *information system* **U** in Düntsch-Orlowska terminology and *data matrix* or *observational model* in GUHA. It is given by a finite non-empty set U of objects, a finite non-empty set A of attributes, each $a \in A$ having a finite domain V_a and an evaluation function f_a assigning to each $u \in U$ an element $f_a(u) \in V_a$. More generally, one may assume $f_a(u) \subseteq V_a$; less generally, one may assume $f_a(u) \in \{0,1\}$, i.e. $V_a = \{0,1\}$ for all a. Obviously, the most general notion is reducible to particular case of $\{0,1\}$-valued attributes by an appropriate coding; for simplicity we shall

restrict ourselves to this particular case. But note that this is equivalent to the assumption of having just one attribute $\hat{a}$ and $f_{\hat{a}}(u) \subseteq V_{\hat{a}}$ is the *item set* in the terminology of [1] - the set of attributes for which u has the value 1.

GUHA uses monadic predicate calculus with generalized quantifiers; open formulas are built from attributes using logical connectives and for φ, ψ open formulas $\varphi \sim \psi$ is a formula with the quantifier $\sim$ (read $\varphi \sim \psi$ "φ is associated with ψ"). For each data matrix let a, b, c, d be the number of objects $u \in U$ satisfying $\varphi\&\psi, \varphi\&\psi, \neg\varphi\&\psi, \neg\varphi\&\neg\psi$ respectively (the four-fold table of φ, ψ in **U**). The semantics of $\sim$ is given by a truth function $tr_\sim$ assigning to each (a, b, c, d) the truth value $tr_\sim(a, b, c, d) \in \{0, 1\}$. A quantifier $\sim$ is *associational* if

$$a_1 \geq a_2, b_1 \leq b_2, c_1 \leq c_2, d_1 \geq d_2 \text{ and } tr_\sim(a_1, b_1, c_1, d_1) = 1 \text{ implies } tr_\sim(a_2, b_2, c_2, d_2) = 1.$$

It is *implicational* if

$$a_1 \geq a_2, b_1 \leq b_2 \text{ and } tr_\sim(a_1, b_1, c_1, d_1) = 1 \text{ implies } tr_\sim(a_2, b_2, c_2, d_2) = 1.$$

(Thus each implicational quantifier is associational. GUHA software works with particular associational/implicational quantifiers making use of their theory. It generates formulas $\varphi \sim \psi$ true in the data.)

The approach of Düntsch-Orlowska relational structures uses information systems to define binary relations on U, particularly relations of similarity and diversity, e.g.
uRv iff u, v have the same attributes,
$uR'v$ iff u, v have no attribute in common.

For deep modal and algebraic aspect of this approach see [7]. Our question now reads: does the GUHA approach lead to some interesting relations? Let us offer three possibilities.

First, note each information system = data matrix ($\{0, 1\}$-valued) can be seen as a binary relation, subset of $U \times A$, namely $(u, a) \in R$ iff $f_a(u) = 1$. This is common in relational databases; evidently, this relation uniquely extends to a relation on $U \times Form$, where $Form$ is the set of all open formulas bilt from atoms $P_1(x), P_2(x), \ldots$. Each n-tuple $\{\varphi_1, \ldots, \varphi_n\}$ of such open formulas determines the corresponding relation on $U \times \{\varphi_1, \ldots, \varphi_n\}$ – the truth table of $\varphi_1, \ldots, \varphi_n$ given by **U.** This is trivial but useful.

Second, assume a data *matrix* **U** and a quantifier $\sim$ to be fixed; this defines a binary relation HYP on $Form$, namely of all pairs φ, ψ such that $\varphi \sim \psi$ is true in **U**. If $\sim$ is associational then HYP represents all pairs (φ, ψ) of open formulas (compared properties) such that φ is associated with ψ. GUHA generates subrelations of this relation given by syntactical conditions on φ, ψ (think e.g. on φ as a combination of symptoms and ψ a combination of diseases or so). Let us stress the importance of *logical rules* (deduction rules) for an optimal representation of true hypotheses; for example if $\sim$ is symmetric ($\varphi \sim \psi$ logically implies $\psi \sim \varphi$) one can make use of it.

Third, let $\mathbf{U}$ and $\sim$ be as before; instead of pairs of formulas let us think of pairs of *subsets* of U definable by open formulas. We get a binary relation on the Boolean algebra of definable subsets of $\mathbf{U}$; properties of the quantifier $\sim$ (for example $\sim$ being an implicational quantifier) determine properties of the relation. This leads us to the following

Definition 1. A *Boolean algebra with an implicational relation* (briefly, an i-Boolean algebra) is a structure $\mathbf{B} = (B, \cup, \cap, -, R)$ where $(B, \cup, \cap, -)$ is a Boolean algebra and $B \subseteq B \times B$ is a relation such that for each $u_1, u_2, v_1, v_2 \in B$, if $u_1 \wedge v_1 \leq u_2 \wedge v_2, u_1 \wedge -v_1 \leq u_2 \wedge -v_2$ and $(u_1, v_1) \in R$ then $(u_2, v_2) \in R$.

Clearly, for each $\mathbf{U}$ and each implicational quantifier $\sim$, the algebra of definable subsets of $\mathbf{U}$ together with the relation HYP as above is an i-Boolean algebra. This leads us to the following

Problem 2. Is each (finite) i-Boolean algebra representable as the algebra given by a data matrix $\mathbf{U}$ and an implicational quantifier $\sim$?

This may be easy to decide; but it seems at least to show that our approach may offer some possibly interesting purely algebraic problems. The methods of the next section seem to be useful for solution of the present problem.

3 Complexity of GUHA-implicational quantifiers

Implicational quantifiers[1] were defined in the preceding section; for full treatment see [3]. Just recall the following examples:

- classical implicational quantifier – $\varphi \Rightarrow \psi$ is $(\forall x)(\varphi(x) \rightarrow \psi(x))$, i.e. $b = 0$
- founded implication $\varphi \Rightarrow_{p,s} \psi$ is true iff $a \geq s$ and $a \geq p(a+b)$ where s is a positive natural number and $0 < p \leq 1$,
- lower critical implication: $\varphi \Rightarrow^{L}_{p,\alpha} \psi$ is true iff $\sum_{i=a}^{a+b} \binom{a+b}{i} p^i (1-p)^{a+b-i} \leq \alpha$

Consider the language with unary predicates $P_1, P_2, \ldots$ and a binary quantifier $\Rightarrow^*$ (as well as object variables and logical connectives). Let x be an object variable called the *designated variable.* A *pure prenex formula* has the form $\varphi \Rightarrow^* \psi$, where φ and ψ are quantifier-free open formulas containing no variable except x (i.e. Boolean conditions of formulas $P_1(x^1), P_2(x), \ldots$). It is understood that $\Rightarrow^*$ binds the variable x; the pedantical writing would be $(\Rightarrow^* x)(\varphi(x), \psi(x))$. The *normal form theorem* (see [3] 3.1.30) says that each closed formula of our language is logically equivalent to a Boolean combination of pure prenex formulas. Note the the proof of this fact is constructive and uniform: given a closed formula Φ one finds a normal form $NF(\Phi)$ which is a (possibly empty) disjunction of elementary conjunctions of pure prenex formulas and is logically equivalent to Φ for *any* semantics of the quantifier $\Rightarrow^*$. It follows

[1] An alternative name is *multitudinal quantifiers*, cf. [2].

that if you have only finitely many predicates (atributes) $P_1, \ldots, P_n$ then there is a finite set **NF** of normal form formulas such that for each Φ, NF(Φ) can be taken from **NF.**

Call Φ an *implicational tautology* (or a tautology with implicational quantifiers) if Φ is true in each interpretation of the language, interpreting $\Rightarrow^*$ as an implicational quantifier.

Examples (easy to verify):

$$((\varphi\&\psi) \Rightarrow^* x) \rightarrow (\varphi \Rightarrow^* (\chi \vee \neg\psi)),$$
$$(\varphi \Rightarrow^* (\psi\&\chi)) \rightarrow ((\varphi\&\psi) \Rightarrow^* \chi).$$

(Compute the corresponding four-fold tables.)

The normal form theorem gives a cheap decidability result:
For each fixed n, the set of all implicational tautologies containing no predicate except $P_1, \ldots, P_n$, is decidable (since the set of the corresponding normal forms is finite, cf. [3] Chapter III, Problem 9). We are going to show decidability and determine the computational complexity of the set of implicational tautologies having arbitrary many predicates. Since a formula is a tautology iff its negation is not satisfiable, everything is solved by the following

Theorem 3. *The problem of implicational satisfiability of a formula in normal form is NP-complete.*

Proof. Assume a disjunction Φ of elementary conjunctions of pure prenex formulas given. Nondeterministically choose one such elementary conjunction K and test satisfiability as follows: let $\varphi_i \Rightarrow^* \psi_i$ be the positive members of K and $\neg(\alpha_j \Rightarrow^* \beta_j)$ the negative ones. The *critical formulas* will be

$$\varphi_i, \varphi_i\&\psi_i, \varphi_i\&\neg\psi_i, \alpha_j, \alpha_j\&\beta_j, \alpha_j\&\neg\beta_j.$$

We look for an interpretation giving these formulas frequences compatible with an implicational quantifier, making $\varphi_i \Rightarrow^* \psi_i$ true and $\alpha_j \Rightarrow^* \beta_j$ false.

Guess a linear preorder $\leq$ of the critical formulas obeying logical consequence (if γ is a subformulas of δ then $\gamma \leq \delta$) and obeying K : there is no i, j such that

$$\varphi_i\&\psi_i \leq \alpha_j\&\beta_j \text{ and } \varphi_i\&\neg\psi_i \geq \alpha_j\&\neg\beta_j.$$

(If there is no such preorder then K is not satisfiable.)

To test realizability of $\leq$ by frequencies of formulas in a model we use the result of [6] on the complexity of fuzzy probabilistic logic over Lukasiewicz propositional calculus. (This calculus is also described in [2].) For each critical formula γ, let $\mathcal{P}(5)$ be the formula saying "γ is probable". Let $\gamma_1, \ldots, \gamma_n$ be a sequence of all critical formulas non-decreasing with respect to $\leq$. Let ε be a new unary predicate. For each $1 \leq i < k$, if $\gamma_i < \gamma_k$ let Λ_i be the formula $(\mathcal{P}(\gamma_i) \oplus \mathcal{P}(\varepsilon) \rightarrow \mathcal{P}(\gamma_{i+1})$ (where $\oplus$ is Lukasiewicz strong disjunction, also denoted by $\underline{\vee}$); if $\gamma_i \leq \gamma_{i+1}$ and $\gamma_{i+1} \leq \gamma_i$ then Λ_i is $\mathcal{P}(\gamma_i) \equiv \mathcal{P}(\gamma_{i+1})$. Let T be the finite theory over FPL whose axioms are all the $\Lambda_i (i = 1, \ldots k-1)$ and also $P(\gamma_{i_0}) \rightarrow P(\neg\varepsilon)$ where i_0 is the

largest index such that $\gamma_{i_o} < \gamma_k$. (The singular case that all γ_i are $\leq$-equivalent is left to the reader as an exercise.) Observe that there is a probability on open formulas built from predicates occuring the critical formulas and coherent with $\leq$ iff the theory T has a model over FPL in which $\mathcal{P}(\varepsilon)$ has a non-extremal value, i.e. $\mathcal{P}(\varepsilon) \vee \neg\mathcal{P}(\varepsilon)$ has not value 1. Using the method of [6] and [2] one easily reduces this problem to a Mixed Integer Programming problem, showing that the last problem is in NP.

Three things remain: First, to show that if such probability exists then we may assume it has rational values on all open formulas from our predicates – this is done as in [2] 8.4.16. Thus multiplying by the common denominator we may get a finite model alias data matrix alias information system **U** such that frequences of object satisfying critical formulas order than in accordence with $\leq$. Now let (a_i, b_i, c_i, d_i) be four-fold tables of pairs (φ_i, ψ_i) of formulas occuring in the sentences $\varphi_i \Rightarrow^* \psi_i$ in K; define a quantifier $\Rightarrow^*$ by letting $tr_{\Rightarrow^*}(a, b, c, d) = 1$ iff for some i, $a \geq a_i, b \leq b_i, c \leq c_i$ and $d \geq d_i$. This is an implicational quantifier and makes **U** to a model of K. This shows that our problem of satisfiabilibty of K (and of Φ) is in NP.

The last (third) thing is to show NP-hardness. To this we reduce the satisfiability problem of propositional logic (equivalently, satisfiability of open formulas built from atom $P_1(x), \ldots, P_n(x), \ldots$ to our problem. Let ε be as before; for each open formula φ as above let φ^* be $(\varepsilon \Rightarrow^* \varepsilon) \rightarrow (\varepsilon \Rightarrow^* \varphi)$. Let us show that φ is a Boolean tautology logic iff φ^* is an implicational tautology; thus φ is satisfiable in Boolean logic iff $\neg(\neg\varphi)^*$ is satisfiable in our logic with an implicational quantifier.

Indeed, if φ is a Boolean tautology and (a, b, c, d) is a four-fold table of (ε, φ) given by an **U**; then $b = d = 0$; if in the same **U** the formula $P \Rightarrow^* P$ is true for an implicational quantifier $\Rightarrow^*$ then $tr_{\Rightarrow^*}(a, 0, 0, c) = 1$ and hence also $tr_{\Rightarrow^*}(a, 0, c, 0) = 1$. Then φ^* is true in each **U**. Conversely, if φ is not a Boolean tautology then take an **U** in which all objects satisfy $\neg\varphi$ and all satisfy ε, thus $\varepsilon \Rightarrow^* \varepsilon$ is true and $\varepsilon \Rightarrow^* \varphi$ is false for the classical implicational quantifier $\Rightarrow^*$ ($\gamma \Rightarrow^* \delta$ being $(\forall x)(\gamma \rightarrow \delta)$). This completes the proof.

Corollary 4. *The set of all implicational tautologies is co-NP-complete.*

Remark. The same idea can be used to show that the problem of associational satisfiability (more precisely, the problem of showing that a formula Φ of the language will one binary quantifier, Φ in normal form, is satisfiable in a **U** by an associational quantifier) is in NP. The reader may try to show NP-completeness (possibly easy).

Conclusion. We hope to have shown that the logic of observational calculi as calculi speaking on data (alias information systems) is interesting and relevant for the study of relational structures and to have contributed to the study of its computational complexity.

References

1. Agrawal R., Mannila H., Srikant R., Toivonen H., Verkamo A. I.: Fast discovery of association rules. In: (Fayyad V. et. al, ed.) Advances in knowledge discovery and data mining, AAA Press/MIT Press 1996, pp. 307–328.
2. Hájek P.: Metamathematics of fuzzy logic, Kluwer 1998.
3. Hájek P., Havránek T.: Mechanizing hypothesis formation – Mathematical foundations for a general theory, Springer Verlag 1978.
4. Hájek P., Holeňa M.: Formal logics of discovery and hypothesis formation by machine. In: (Arikawa et al, ed.) Discovery Science, Springer Verlag 1998, 291-302
5. Hájek P., Sochorová A., Zvárováá J.: GUHA for personal computers. Comp. Statistics and Data Analysis 19(1995), 149–153.
6. Hájek P., Tulipani S.: Complexity of fuzzy probability logic. Fundamenta Informaticae 45 (2001), 207–213.
7. Orlowska E., Szalas A. (ed.): Relational methods in computer science applications, Springer-Physica Verlag 2001, 263–285.

This article was processed using the LaTeX macro package with LLNCS style