Definability of Association Rules in Predicate Calculus

Jan Rauch

Department of Information and Knowledge Engineering
University of Economics, Prague
nám. W. Churchilla 4, 130 67 Praha 3, Czech Republic
rauch@vse.cz

Summary. Observational calculi are special logical calculi in which statements concerning observed data can be formulated. Their special case is predicate observational calculus. It can be obtained by modifications of classical predicate calculus - only finite models are allowed and generalised quantifiers are added. Association rules can be understood as special formulas of predicate observational calculi. Such association rules correspond to general relations of two Boolean attributes. A problem of the possibility to express association rule by the means of classical predicate calculus is investigated. A reasonable criterion of classical definability of association rules is presented.

Key words: Data mining, association rules, mathematical logic, observational calculi

1 Introduction

The goal of this chapter is to contribute to the theoretical foundations of data mining. We are interested in association rules of the form $\varphi \sim \psi$ where φ and ψ are derived Boolean attributes. Meaning of association rule $\varphi \sim \psi$ is that Boolean attributes φ and ψ are associated in a way corresponding to the symbol $\sim$ that is called 4ft-quantifier. The 4ft-quantifier makes possible to express various types of associations e.g. several types of implication or equivalency and also associations corresponding to statistical hypothesis tests.

Association rules of this form are introduced in [2]. Some more examples are e.g. in [7, 8]. To keep this chapter self-contained we will overview basic related notions in the next section.

Logical calculi formulae of which correspond to such association rules were defined and studied e.g. in [2, 4, 5, 6, 7]. It was shown that there are practically important theoretical properties of these calculi. Deduction rules of the form

$\frac{\varphi \sim \psi}{\varphi' \sim \psi'}$ where $\varphi \sim \psi$ and $\varphi' \sim \psi'$ are association rules are examples of such results [7].

Logical calculus of association rules can be understood as a special case of the monadic observational predicate calculus [2]. It can be obtained by modifications of classical predicate calculus such that *only finite models are allowed* and *4ft quantifiers are added*. We call this calculus *observational calculus of association rules*.

Observational calculus is a language formulae of which are statements concerning observed data. Various types of observational calculi are defined and studied [2]. The observational calculi are introduced in Sect. 3. Association rules as formulas of observational calculus are defined in Sect. 4.

The natural question is what association rules are classically definable. We say that the association rule is classically definable if it can be expressed by means of classical predicate calculus (i.e. predicates, variables, classical quantifiers $\forall$, $\exists$, Boolean connectives and the predicate of equality). The formal definition is in Sect. 5. The problem of definability in general monadic observational predicate calculi is solved by the Tharp's theorem, see Sect. 5.

Tharp's theorem is but too general from the point of view of association rules. We show, that there is a more intuitive criterion of classical definability of association rules. This criterion concerns 4ft-quantifiers. We need some theoretical results achieved in [2], see Sect. 6. The criterion of classical definability of association rules is proved in Sect. 7.

2 Association Rules

The association rule is an expression $\varphi \sim \psi$ where φ and ψ are Boolean attributes. The association rule $\varphi \sim \psi$ means that the Boolean attributes φ and ψ are associated in the way given by the symbol $\sim$. The symbol $\sim$ is called *4ft-quantifier*. Boolean attributes φ and ψ are derived from columns of an analysed data matrix $\mathcal{M}$. An example of the association rule is the expression

$$A(\alpha) \wedge D(\delta) \sim B(\beta) \wedge C(\gamma) \ .$$

The expression $A(\alpha)$ is a *basic Boolean attribute* The symbol α denotes a subset of all possible values of the attribute A (i.e. column of the data matrix $\mathcal{M}$). The basic Boolean attribute $A(\alpha)$ is true in row o of $\mathcal{M}$ if it is $a \in \alpha$ where a is the value of the attribute A in row o. Boolean attributes φ and ψ are derived from basic Boolean attributes using propositional connectives $\vee$, $\wedge$ and $\neg$ in the usual way.

The association rule $\varphi \sim \psi$ can be true or false in the analysed data matrix $\mathcal{M}$. It is verified on the basis of the four-fold contingency table of φ and ψ in $\mathcal{M}$, see Table 1. This table is denoted 4ft(φ, ψ, $\mathcal{M}$).

Here a is the number of the objects (i.e. the rows of $\mathcal{M}$) satisfying both φ and ψ, b is the number of the objects satisfying φ and not satisfying ψ, c is