A New and Useful Syntactic Restriction on Rule Semantics for Tabular Datasets

Marie Agier[1,2] and Jean-Marc Petit[3]

[1] DIAGNOGENE, Aurillac
[2] LIMOS, UMR 6158 CNRS, Univ. Clermont-Ferrand II
[3] LIRIS, UMR 5205 CNRS, INSA Lyon France

Abstract. Different rule semantics have been successively defined in many contexts such as implications in artificial intelligence, functional dependencies in databases or association rules in data mining. We are interested in defining on tabular datasets a class of rule semantics for which Armstrong's axioms are sound and complete, so-called *well-formed semantics*. The main contribution of this paper is to show that an *equivalence* does exist between some syntactic restrictions on the natural definition of a given semantics and the fact that this semantics is well-formed. From a practical point of view, this equivalence allows to prove easily whether or not a new semantics is well-formed. We also point out the relationship between our generic definition of rule satisfaction and the underlying data mining problem, i.e. given a well-formed semantics and a tabular dataset, discover a cover of rules satisfied in this dataset. This work takes its roots from a bioinformatics application, the discovery of gene regulatory networks from gene expression data.

1 Introduction

The notion of *rules* or *implications* is very popular and appears in different flavors in databases, data mining or artificial intelligence communities. Famous examples of rules are functional dependencies [1], implications [2] or association rules [3]. As such, a simple remark can be done on such rules: their syntax is the same but their semantics widely differs.

In this paper, we consider rules to be defined on *tabular datasets*. Basically, a tabular dataset is equivalent to a *relation* over a set U of distinguished attributes in databases terminology [4]. In this setting, a *rule* is an expression of the shape $X \rightarrow Y$ i.e. "X implies Y" with $X, Y \subseteq U$. The *semantics* of a rule $X \rightarrow Y$ over U is the *meaning*, the *sense* one wants to give to this rule: Given a relation r, a rule $X \rightarrow Y$ is said to be *satisfied* in r with the semantics s, noted $r \models_s X \rightarrow Y$ if the semantics of the rule is true (or valid) in r.

From our analysis of existing rule semantics, we identify two main components to specify a rule semantics: the subsets of the relation on which the rule applies and the predicates occurring in the "if... then..." part of the rule. By the way, a natural and "generic" definition of rule semantics can be elaborated in order to be able not only to capture most of existing semantics already known on tabular datasets, but also to devise new semantics specific to some application domains.

S.O. Kuznetsov and S. Schmidt (Eds.): ICFCA 2007, LNAI 4390, pp. 39–50, 2007.

We also chose to focus on those semantics verifying Armstrong's axioms, i.e. semantics for rules on which Armstrong's axioms are sound and complete, so-called "well-formed semantics". For functional dependencies and implications, this result is known for a long time but more surprisingly, many other semantics also fit into this framework [2]. Roughly speaking, our aim is to define syntactical boundaries of well-formed semantics. Practical interests are for instance that some form of *reasoning* can be done on rules (e.g. implication problem in linear time [5]). Moreover, it is also possible to work on "small" *covers* of rules [6, 7, 8] and to use data mining techniques specific to the considered cover, but applicable to *every* well-formed semantics.

Paper contribution. The contribution of this paper is to show that an *equivalence* does exist between some syntactic restrictions on the natural definition of a given semantics and the fact that this semantics is *well-formed*.

From a practical point of view, this equivalence allows to prove easily whether or not a new semantics is well-formed: So far, for a given semantics, a proof of the soundness and the completeness of the Armstrong's axioms for this semantics should be given. Now, it is just enough to show that this semantics complies with the proposed syntactic restrictions.

We also point out a relationship between our generic definition of rule satisfaction and the underlying data mining problem, i.e. given a well-formed semantics and a relation, discover a cover of rules satisfied in this relation. More precisely, we show how a base of the closure system for any well-formed semantics can be computed from a dataset.

Application. This work takes its roots from a bioinformatics application, the discovery of gene regulatory networks from gene expression data. The challenge is to find out relationships between genes that reflect observations of how expression level of each gene affects those of others. The conjecture that association rules could be a model for the discovery of gene regulatory networks has been partially validated, see for example [9, 10, 11]. Nevertheless, we believe that many different kinds of rules between genes could be useful with respect to some biological objectives and the restricted setting of association rules could be not enough to cope with this diversity. Therefore, the main application of this paper is to offer a framework in which biologists may define their "own customized semantics" for rules with regard to their requirements. It is worth noting that other application domains could benefit from the propositions made in this paper.

Paper organization. We give in Section 2 the motivations of our proposition with examples of rule semantics. In Section 3, we propose a natural definition of a semantics using some syntactic restrictions. In Section 4, we further restrict the syntax and give the main result of this paper then we point out the relationships between our proposition and the underlying data mining problem. In Section 5, we give the related contributions of this work and finally, we conclude and give some perspectives in Section 6.

2 Motivating Examples

We give in the sequel three examples of semantics for tabular datasets, some of them in the context of gene expression data.