Classical and Symbolic metadata setting for biological datasets

Haralambos Papageorgiou*, Maria Vardaki*

*Department of Mathematics University of Athens Panepistemiopolis, 15784, Athens Greece
hpapageo@uoa.gr, mvardaki@uoa.gr

Abstract. We consider an extension of classical data analysis into symbolic data analysis to describe the management process of biological datasets produced by multi-source clinical studies. Such extension leads to more complex data types and tables and the metadata under consideration hold information both on classical (original) and the symbolic data. In this paper we model these metadata items in a structured object-oriented schema for symbolic data revealing their relations. A number of transformations are also discussed both for classical and symbolic classes of our model in order to illustrate how the applied transformations on symbolic data depend on the related classical data setting.

1 Introduction

Symbolic data serve not only to summarize large datasets, but they also lead to more complex data tables, thus enabling the manipulation of huge datasets (Bock and Diday, 2000; Billard and Diday, 2006). Using the symbolic data techniques, data are aggregated into macrodata, forming Symbolic Objects (SO) and Symbolic Data Tables (SDT) (Bock and Diday, 2000; Diday and Noirhomme-Fraiture, 2008; Noirhomme-Fraiture, 1997).

A symbolic data table constitutes the main input for symbolic data analysis (Diday, 2002). It looks like a classical data table where each cell represents symbolic data, since each row corresponds to a symbolic description of a group of individuals and each column corresponds to a symbolic-valued variable (Noirhomme-Fraiture and Brito, 2011).

Consider a modern, state-of-the-art information system. As expected, it stores a considerable amount of microdata, macrodata and related metadata for each piece of information imported. In the case of an information system that manages biological datasets collected from multi-source clinical studies, due to confidentiality reasons (as emphasized by UNESCO, the Nuremberg Code, the Helsinki Declaration, etc.) all data are imported, randomized and further used in the form of macrodata. Symbolic analysis techniques are especially useful for managing large datasets from multiple sources; therefore they can adequately manage, among others, biological macrodata resulting from various clinical studies.

Since, even in the classical data setting, aggregate data can be of little value to any data consumer if explanatory information (metadata) does not accompany them (definitions, patients' eligibility criteria, the study parameters, the risk factors, how data were collected and