Towards Automatic Feature Construction for Supervised Classification

Marc Boullé

Orange Labs,
2 avenue Pierre Marzin, 22300 Lannion, France
marc.boulle@orange.com
http://perso.rd.francetelecom.fr/boulle

Abstract. We suggest an approach to automate variable construction for supervised learning, especially in the multi-relational setting. Domain knowledge is specified by describing the structure of data by the means of variables, tables and links across tables, and choosing construction rules. The space of variables that can be constructed is virtually infinite, which raises both combinatorial and over-fitting problems. We introduce a prior distribution over all the constructed variables, as well as an effective algorithm to draw samples of constructed variables from this distribution. Experiments show that the approach is robust and efficient.

Keywords: supervised learning, relational learning, feature construction, feature selection, regularization.

1 Introduction

In a data mining project, the data preparation phase aims at constructing a data table for the modeling phase [19,6]. The data preparation is both time consuming and critical for the quality of the mining results. It mainly consists in a search of an effective data representation, based on variable construction and selection. Variable selection has been extensively studied in the literature [12]. Two main approaches, filter and wrapper, have been proposed. Filter methods consider the correlation between the input variables and the output variable as a pre-processing step, independently of the chosen classifier. Wrapper methods search the best subset of variables for a given classification technique, used as a black box. Wrapper methods, which are time consuming, are restricted to the modeling phase of data mining, as a post-optimization of a classifier. Filter methods are better suited to the data preparation phase, since they can be combined with any data modeling approach and can deal with large numbers of input variables. In this paper, we focus on the filter approach, in the context of supervised classification.

Variable construction [18] has been less studied than variable selection in the literature. It implies a large amount of work for the data analyst and heavily relies on domain knowledge to construct new potentially informative variables. In practice, the initial raw data usually originate from relational databases. As most
classification techniques need a flat input data table with instances \( \times \) variables tabular format, such relational data cannot be directly analyzed.

Learning from relational data has recently received an increasing attention in the literature. The term Multi-Relational Data Mining (MRDM) was initially introduced in [13] to address novel knowledge discovery techniques from multiple relational tables. The common point between these techniques is that they need to transform the relational representation. In Inductive Logic Programming (ILP) [9], data is recoded as logic formulas. In 1BC method [16] and its successor 1BC2 [17], simple predicates are used together with a naive Bayes classifier. More expressive approaches cause scalability problems especially with large-scale data. Other methods named by propositionalisation [14] try to flatten the relational data by constructing new variables. These variables aggregate the information contained in non target tables in order to obtain a classical tabular format. For example, the RELAGGS method [15] uses functions such as mean, median, min, max to summarize numerical variables from secondary tables in zero to many relationship, or counts per value for the categorical variables. The TILDE method [2,24] aims at constructing complex variables based on conjunctions of selection conditions of records in secondary tables. However, the expressiveness of such methods faces the following problems: complex parameter setting, combinatorial explosion of the number of potentially constructed variables and risk of over-fitting that increases with the number of constructed variables.

In this paper, we suggest an approach aiming at the automation of variable construction, with the three-fold following objective: simplicity of parameters, efficient control of the combinatorial search in the space of variable construction and robustness w.r.t. over-fitting. Section 2 presents a formal description of a variable construction domain. Section 3 introduces an evaluation criterion of the constructed variables exploiting a Bayesian approach, by suggesting a prior distribution over the space of variables that can be constructed. Section 4 studies the problem of drawing a sample from this space and describes an efficient and computable algorithm for drawing samples of constructed variables of given size. Section 5 evaluate the approach on several datasets. Finally, Section 6 gives a summary and discusses future work.

2 Specification of Variable Construction Domain

We suggest a formal specification of a variable construction domain in order to efficiently drive the construction of variables for supervised classification. The objective is not to propose a new expressive and general formalism for describing domain knowledge, but simply to clarify the framework exploited by the variable construction algorithms presented in Section 4. This framework consists in two parts: description of the data structure and choice of the construction rules.

2.1 Data Structure

The simplest data structure is the tabular one. Data instances are represented by a list of variables, each defined by its name and type. The standard types,