Chapter XII
Bayesian Belief Networks for Data Cleaning

Enrico Fagiuoli
Università degli Studi di Milano-Bicocca, Italy

Sara Omerino
ETNOTEAM S.p.A., Italy

Fabio Stella
Università degli Studi di Milano-Bicocca, Italy

ABSTRACT

The importance of data cleaning and data quality is becoming increasingly clear, as evidenced by the surge in software, tools, consulting companies, and seminars addressing data quality issues. In this contribution, the authors present and describe how Bayesian computational techniques can be exploited for data-cleaning purposes to the extent of reducing the time to clean and understand the data. The proposed approach relies on the computational device named Bayesian belief network, which is a general statistical model that allows the efficient description and treatment of joint probability distributions. This work describes the conceptual framework that maps the Bayesian belief network computational device to some of the most difficult tasks in data cleaning, namely imputing missing values, completing truncated datasets, and outliers detection. The proposed framework is described and supported by a set of numerical experiments performed by exploiting the Bayesian belief network programming suite named HUGIN.

INTRODUCTION

Every data analysis task starts by gathering, characterizing, and cleaning a new, unfamiliar dataset (Dasu & Johnson, 2003). After this process, the data can be analyzed and the results delivered. It is well established that the first step is far more difficult and time consuming than the second.

Indeed, data gathering is complicated by sociological (turf sensitivity) and technological problems (different software and hardware platforms make transferring and sharing data very difficult). Once the data are in place, acquiring the metadata (data description and business rules) is another challenge. Indeed, very often the metadata are poorly documented, and when we are ready to analyze the data, its quality is suspect. Fortunately, automated techniques can be applied to understand the data through *exploratory data mining,* and to ensure *data quality* through *data cleaning*.

Data cleaning and quality monitoring is an incessant and continuous activity starting right from data gathering stage to the ultimate choice of analysis and interpretation of the results. It is needed to update the static conventional definitions and metrics of data quality to reflect the continuous and flexible nature of data quality process and metrics required to effectively measure and monitor data quality (Scannapieco, Missier, & Batini, 2005).

According to a study conducted by The Data Warehouse Institute, commissioned by DataFlux (The Data Warehouse Institute, 2003), current data quality problems cost U.S. business more than 600 billion dollars a year. Furthermore, a survey from conversation to practitioners of *data mining* leads to assert that between 30% to 80% of the data-analysis task is spent in cleaning and understanding the data. Therefore, the importance of data cleaning and data quality is becoming increasingly clear as evidenced by the surge in software, tools, consulting companies, and seminars addressing data quality issues.

A taxonomy of data-quality problems, addressed by data cleaning, together with an overview of the main solution approaches has been proposed by Rahm and Hai Do (2000).

Several contributions, devoted to cleaning databases containing corrupted data, have been proposed in the specialized literature.

Guyon, Matic, and Vapnik (1996) emphasize the link between informative patterns and data cleaning, and describe how machine learning approaches can be exploited to remove noise from a database.

Schwarm and Wolfman (2000) pointed out that many techniques have attempted to use learners to predict problems with class values in datasets, with the same approach being extended to correct errors in data. However, these techniques suffer several problems: they can only actually correct noise in the class attribute, do not fully leverage dependencies among attributes, and are inappropriate for datasets with no distinguished class attribute.

As far as the authors know, Schwarm and Wolfman were the first to propose the use of Bayesian belief networks for data cleaning. However, the approach described through this work significantly differs from their approach in the sense that it is unsupervised, and therefore it does not require the availability of any subset of precleaned instances from the database to be cleaned, as required by Schwarm and Wolfman (2000).

Arning, Agrawal, and Raghavan (1996) described a linear time method for detecting deviations in a database. The authors assume that all records should be similar, which may not be true in unsupervised learning tasks, and that an entire record is either noisy or clean.

The assumption that entire records are noisy or clean is also common in outlier and novelty detection (Hampel, Rousseeuw, Ronchetti, & Stahel, 1986; Huber, 1981).

However, as pointed out in Kubika and Moore (2003), a significant downside to looking at noise on the scale of records is that entire records are thrown out, and useful, uncorrupted data may be lost. In datasets where almost all records have at least a few corrupted cells, this may prove disastrous. The approach described in Kubika and Moore (2003) is particularly interesting,