Expert Systems with Applications 41 (2014) 2259–2268

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Assessment of data quality in accounting data with association rules

Paul Alpar *, Sven Winkelsträter

School of Business and Economics, University at Marburg, Universitätsstr. 24, 35032 Marburg, Germany

ARTICLE INFO

Keywords:
Data quality
Data mining
Association rules
Business rules
Accounting data

ABSTRACT

Business rules are an effective way to control data quality. Business experts can directly enter the rules into appropriate software without error prone communication with programmers. However, not all business situations and possible data quality problems can be considered in advance. In situations where business rules have not been defined yet, patterns of data handling may arise in practice. We employ data mining to accounting transactions in order to discover such patterns. The discovered patterns are represented in form of association rules. Then, deviations from discovered patterns can be marked as potential data quality violations that need to be examined by humans. Data quality breaches can be expensive but manual examination of many transactions is also expensive. Therefore, the goal is to find a balance between marking too many and too few transactions as being potentially erroneous. We apply appropriate procedures to evaluate the classification accuracy of developed association rules and support the decision on the number of deviations to be manually examined based on economic principles.

1. Introduction

Data quality has been a concern since the beginning of automated data processing. Users of information systems quickly learned the consequences of bad data and coined the adage "Garbage in, garbage out". Data quality encompasses many aspects which makes it difficult to operationalize the concept. It is often decomposed into a number of characteristics or dimensions. For example, Wang and Strong (1996) identified 16 dimensions through interviews with information technology (IT) users. The same number of characteristics has been suggested by English (1999). The dimensions are usually grouped into categories, often four of them. A lack of data quality does not necessarily mean that data are wrong. They just may lack some desired properties, e.g. they may not be as precise or as current as needed. Data quality assessment depends on the needs and perceptions of its users (Strong, Lee, & Wang, 1997). This does not mean that data quality measurement is mainly a subjective exercise. The judgment of usefulness of data with a given quality is subjective because it depends on the user and his concrete task. The same level of quality may be sufficient for one task but inacceptable for another task. For example, the birth year of a consumer can be sufficient to judge whether she is an adult but it is not sufficient if a company wants to send her birthday wishes. Thus, data quality is satisfactory if the relevant dimensions are satisfying for the intended use (Olson, 2003).

Once the dimensions of data quality are determined, metrics need to be defined to be able to measure and control data quality. The object of quality examination can be an individual attribute, a tuple, a relation, the whole data base, or a transaction that transforms data. A data quality requirement can be expressed in the form of a consistency rule. Each object can violate or fulfill more than one rule. An appropriate quality measure should take this into account. A question arises as to where the consistency rules come from (Hinrichs, 2002). They can be defined in advance by business experts based on known business practices and regulations. Nowadays, software packages exist which allow experts to codify their knowledge in form of business rules in almost natural language. A business rule can be defined as "a statement that defines or constrains some aspect of the business. It is intended to assert business structure, or to control or influence the behavior of the business (Business Rule Group, 2012)".

Specification of rules by business experts has, first, the advantage that misunderstandings with programmers can be avoided. Second, rather than being scattered all over the program and perhaps not being well documented, they are easy to locate and fairly explicit. This makes them more readable and easier to maintain. However, the following problems may limit the possibility of specification of business rules (Gebauer & Windheuser, 2011, pp. 93):

- Business rules may not be documented well enough.
- Business rules may not be regularly applied so that employees are unaware of them.
- Business rules may be fuzzy, outdated, or too generic so that their use requires further implicit knowledge.

* Corresponding author. Tel.: +49 6421 2823703; fax: +49 6421 2826554.
E-mail addresses: alpar@wiwi.uni-marburg.de (P. Alpar), sven.winkelstraeter@gmx.de (S. Winkelsträter).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.09.024

Therefore, it may be helpful, in addition to recorded business rules, to recover further rules from practice which is mirrored in actual business transactions and corresponding data. Deviations from these patterns may signal data quality problems. Objects containing such deviations are often referred to as outliers (see, for example, Maervoet, Vens, Vanden Berghe, Blockeel & De Causmaecker, 2012). This approach to outlier detection in the context of data quality was proposed in previous work (Hipp, Güntzer, & Grimmer, 2001). However, the reported experience with the application of the approach to real-world data is limited (Hipp, Müller, Hohendorff, & Naumann, 2007). We apply the approach to accounting data and show that simple accounting-specific knowledge can help to improve the results. Second, the referenced work examines the strength of the procedure without any cost considerations. On one hand, the examination of outliers also creates costs. On the other hand, the costs of non-detected data errors can be much higher than costs of suspected problems that turn out to be false alarms. Therefore, we extend the approach by adding cost considerations. Third, we present result visualizations that can support decision makers in choosing parameters. Finally, we suggest some heuristics for the practical application of the approach.

The next section briefly presents previous work. The third section describes our procedural approach to rules development and outlier detection and the techniques we use. The evaluation of classification accuracy of the developed rules is discussed in Section 4. The application of our approach to three sets of independent real-world data is described in Section 5. The last section summarizes the results, gives suggestions for practical use and extension to neighboring application areas, and indicates possibilities for further research.

2. Previous research

Our approach analyzes actual data rather than processes. Therefore, we review only data-driven approaches in this section. In data-driven approaches, data quality is assessed on the basis of actual data rather than on the basis of processes that generate them (see, e.g., Kaplan, Krishnan, Padman, & Peters, 1998). Both approaches should be used together in practice. A comprehensive management of data quality needs to start with appropriate process and data design methods. On one hand, proper process design may prevent many data problems. On the other hand, problems discovered during data quality assessment may give rise to process redesign. See Batini, Cappiello, Francalanci, and Maurino (2009) for a comprehensive review of methodologies for data quality assessment.

We define three classes of techniques for data-driven quality assessment, partly in analogy to Batini and Scannapieca (2006) who distinguish probabilistic, knowledge-based, and empirical techniques. We prefer to label the third group "other" because the other two groups also work with empirical data and the authors' "empirical techniques" include methods that are assigned to knowledge-based techniques by other authors. Our class "other" contains all other approaches that do not belong to the first two classes.

2.1. Probabilistic techniques

These approaches are based on relative frequencies of attribute values in the observed data bases, on data volatilities known from other samples or populations, or some other empirical probabilities. They mainly observe one attribute at a time.

Jiang, Sarkar, De, and Dey (2007) analyze the integration of data from two or more sources where attribute values that should be the same differ. They define probability-based measures to determine which of the values has the highest probability of being the true value. The probability of a value is calculated from prior probabilities of values (which can be determined from relative frequencies in the database) and attribute-specific reliabilities of involved data sources (which could have been determined beforehand through sampling). In addition, they also develop a procedure to determine the cost-minimizing best value assuming that costs from type I and type II errors and of data misrepresentation are known. This value is assigned to the attribute with conflicting values in different data sources. This means that the procedure can automatically resolve such conflicts.

The quality dimension of data currency has been addressed in Heinrich and Klier (2011) and Heinrich, Kaiser, and Klier (2009). The authors develop a probability-based measure for attributes which quality may decline over time (e.g., last name, address, occupation). They define currency as the probability that the value of an attribute in a database still corresponds to its real-world version at the time of data quality assessment. The probability is calculated based on the average currency decline rate for the given attribute and the age of the attribute value. Average decline rates can be obtained from publicly available statistics (e.g., the yearly rate of relocation of private households can be used as the likelihood of address change of customers) or from sampling of internal data. The authors show that the metric fulfills requirements for data quality metrics they developed from literature. If the likelihood that values are outdated is high then actions to increase their currency may be worthwhile. The economic usefulness of the approach is demonstrated through an application to customer data of a mobile services provider.

2.2. Knowledge-based techniques

The techniques in this group usually analyze relations among different attributes or groups of data. In this approach, an outlier value may be a value that is within the range of values of that attribute but unusual in combination with other attributes of the same object or in relation to other objects.

Hipp et al. (2001) suggest creating association rules from existing data. These association rules can be used to identify data deviations from established patterns which may stem from data quality problems. The feasibility of this approach has been first tested with synthetic data (Kübart, Grimmer, & Hipp, 2005). The authors generated 10,000 records with eight attributes per record and a specific distribution of values for each attribute. Then, some of the records and some of the attributes were randomly chosen and their values were randomly altered. The "errors" were known and the performance of the generated association rules could be tested. Real data from the automotive industry was analyzed in Hipp, Müller, Hohendorf and Naumann (2007). The accuracy of only one attribute was examined in that work. Domain experts evaluated a sample of about 1,300 pre-selected records from a large data base and added the correct value for the observed attribute if it could be determined. This way, erroneous records could be identified. As indicated above, costs were not considered in this work. As the authors state, generalizations on rule structure and content or parameter recommendations are not possible because the experiment was limited to one very specific application.

Low, Lee, and Ling (2001) develop an expert system to detect duplicate records. In their approach, rules are defined by experts while certainty factors, also assigned by experts, serve as weights. A system that learns mapping rules for identifying identical objects in two different data sources with little input from users was developed in Tejada, Knoblock, and Minton (2001). The identification problem may arise, e.g., due to different spellings of names, abbreviations, or typing errors. Mapping rules are constructed with decision trees by attempting to maximize their accuracy.