Knowl Inf Syst (2006)
DOI 10.1007/s10115-006-0006-x

Knowledge and Information Systems

SPECIAL ISSUE

Laure Berti-Équille

Data quality awareness: a case study for cost-optimal association rule mining

Received: 9 May 2005 / Revised: 1 November 2005 / Accepted: 14 January 2006

Abstract The quality of discovered association rules is commonly evaluated by interestingness measures (commonly support and confidence) with the purpose of supplying indicators to the user in the understanding and use of the new discovered knowledge. Low-quality datasets have a very bad impact over the quality of the discovered association rules, and one might legitimately wonder if a so-called "interesting" rule noted $LHS \rightarrow RHS$ is meaningful when 30% of the *LHS* data are not up-to-date anymore, 20% of the *RHS* data are not accurate, and 15% of the *LHS* data come from a data source that is well-known for its bad credibility. This paper presents an overview of data quality characterization and management techniques that can be advantageously employed for improving the quality awareness of the knowledge discovery and data mining processes. We propose to integrate data quality indicators for quality aware association rule mining. We propose a cost-based probabilistic model for selecting *legitimately interesting* rules. Experiments on the challenging KDD-Cup-98 datasets show that variations on data quality have a great impact on the cost and quality of discovered association rules and confirm our approach for the integrated management of data quality indicators into the KDD process that ensure the quality of data mining results.

Keywords Quality awareness mining · Data quality management · Data quality metadata · Cost model · Association rule mining

1 Introduction

The quality of data mining results and the validity of results interpretations essentially rely on the data preparation process and on the quality of the analyzed

L. Berti-Équille (✉)
IRISA, University of Rennes I, Campus Universitaire de Beaulieu,
35042 Rennes, France
E-mail: Laure.Berti-Equille@irisa.fr

datasets [48]. Indeed, data mining processes and applications require various forms of data preparation, correction, and consolidation, combining complex data transformation operations and cleaning techniques. This is because the data input to the mining algorithms is assumed to conform to "nice" data distributions, containing no missing, inconsistent, or incorrect values. This leaves a large gap between the available "dirty" data and the available machinery to process and analyze the data for discovering added-value knowledge and decision-making. Data quality is a multidimensional, complex, and morphing concept [14]. In the last decade, there has been a significant amount of work in the area of information and data quality management initiated by several research communities (database, statistics, workflow management, and knowledge engineering), ranging from the techniques that assess information quality to the design of large-scale data integration systems over heterogeneous data sources with different degrees of quality and trust. Many data quality definitions, metrics, models, and methodologies have been proposed by academics and practitioners with the aim to tackle the main classes of data quality problems:

- Duplicate detection and record matching known under various names: record linkage [21], merge/purge problem [24], object matching [12, 67], duplicate elimination [39], citation matching [7, 36], identity uncertainty [43], entity identification [33], entity resolution [5], or approximate string join [23];
- Instance conflict resolution [20] using data source selection [16, 37, 41] or data cleaning techniques [49];
- Missing values [34] and incomplete data [55];
- Staleness of data [8, 58].

In error-free data warehouses or database-backed information systems with perfectly clean data, knowledge discovery techniques (such as clustering, mining association rules, or visualization) can be relevantly used as decision-making processes to automatically derive new knowledge patterns and new concepts from data. Unfortunately, most of the time, this data is neither rigorously chosen from the various heterogeneous information sources with different degrees of quality and trust, nor carefully controlled for quality. Deficiencies in data quality still are a burning issue in many application areas, and become acute for practical applications of knowledge discovery and data mining techniques [44]. Among traditional descriptive data mining techniques, association rules discovery identifies intra-transaction patterns in a database and describes how much the presence of a set of attributes in a database's record (i.e., a transaction) implicates the presence of other distinct sets of attributes in the same record (respectively in the same transaction). The quality of association rules is commonly evaluated by the support and confidence measures. The support of a rule measures the occurrence frequency of the pattern in the rule while the confidence is the measure of the strength of implication. The problem of mining association rules is to generate all association rules that have support and confidence greater than the user-specified minimum support and confidence thresholds. Besides support and confidence, other measures for knowledge quality evaluation (called interestingness measures) have been proposed in the literature with the purpose of supplying alternative indicators to the user in the understanding and use of the new discovered knowledge [31, 57]. But, to illustrate the impact of low-quality data over discovered association rule quality, one might legitimately wonder whether a so-called "interesting" rule noted

LHS → *RHS* (with the following semantics: *L*eft-*H*and *S*ide implies *R*ight-*H*and *S*ide) is meaningful when 30% of the *LHS* data are not up-to-date anymore, 20% of the *RHS* data are not accurate, and 15% of the *LHS* data come from a data source that is well-known for its bad reputation and lack of credibility.

The main contribution of this paper is twofold: First, we give an overview of data quality characterization and management techniques that should be integrated in the KDD process for improving the quality of discovered knowledge and mining results; secondly, we propose a method for scoring association rule quality and a probabilistic cost model that predicts the cost of low-quality data on the quality of discovered association rules. This model is used to select the so-called "legitimately interesting" association rules. We evaluate our approach using the KDD-Cup-98 dataset and confirm our assumption that interestingness measures are necessary but not self-sufficient for ensuring the quality of discovered knowledge.

The rest of the chapter is organized as follows. Section 2 gives an overview on data quality characterization and management techniques. Section 3 presents a case study of quality awareness association rule mining and proposes a probabilistic decision model for estimating the cost of low-quality data on discovered association rules. In Sect. 4, we evaluate our approach using the KDD-Cup-98 dataset for our experiments. Section 5 provides concluding remarks and guidelines for future extensions of this work.

2 An overview of data quality characterization and management

Maintaining a certain level of quality of data is challenging and cannot be limited to one-shot approaches addressing simpler, more abstract versions of the problems of dirty or low-quality data [11, 14, 32]. Solving these problems requires highly domain- and context-dependent information and human expertise. Classically, the database literature refers to data quality management as ensuring: (i) syntactic correctness (e.g., constraints enforcement that prevent "garbage data" from being entered into the database) and (ii) semantic correctness (i.e., data in the database that truthfully reflects the real-world entities and situations). This traditional approach of data quality management has lead to techniques such as integrity constraints, concurrency control, and schema integration for distributed and heterogeneous systems. A broader vision of data quality management is presented in this chapter (but still with a database orientation). In the last decade, literature on data and information quality across different research communities (including databases, statistics, workflow management, and knowledge engineering) proposed a plethora of:

- Data quality dimensions with various definitions depending on authors and application contexts [4, 17, 22, 63];
- Data quality dimension classifications that are depending on the audience type: Practical and technical [51], more general [29] or depending on the system architecture type: See [42] for integrated information systems, [27] for data warehouse systems, or [53] for cooperative information systems (CIS);
- Data quality metrics [14, 25, 42, 46, 47, 68];
- Conceptual models and methodologies [38, 54, 61];

- Frameworks to improve or assess data quality in databases [1, 26, 51, 63, 64], in information systems [2, 3, 65] or in data warehouse systems [19, 27, 59, 61].

To give a detailed overview, the rest of this section will present the different paradigms for data quality characterization, modeling, measurement, and management.

2.1 Data quality dimensions

Since 1980 with Brodie's proposition [10], more than 200 dimensions have been collected to characterize data quality in the literature [4, 26, 51, 62]. The most frequently mentioned data quality dimensions in the literature are accuracy, completeness, timeliness, and consistency, with various definitions:

- Accuracy is the extent to which collected data is free of errors [35] or it can be measured by the quotient of the number of correct values in a source and the number of the overall number of values [42];
- Completeness is the percentage of the real-world information entered in the sources and/or the data warehouse [27] or it can be measured by the quotient of the number of non-null values in a source and the size of the universal relation [42];
- Timeliness is the extent to which data are sufficiently up-to-date for a task [35]; the definitions of freshness, currency, volatility are reported in [8];
- Consistency is the coherence of the same data represented in multiple copies or different data with respect of pre-defined integrity constraints and rules [4].

Figure 1 presents a classification of data quality dimensions that are divided into the following four categories:

1. Quality dimensions describing the quality of the management of data by the system based on the satisfaction of technical and physical constraints (e.g., accessibility, ease of maintenance, reliability, etc.);
2. Quality dimensions describing the quality of the representation of data based on the satisfaction of conceptual constraints on modeling and information presentation (e.g., conformance to schema, appropriate presentation, clarity, etc.);
3. Intrinsic data quality dimensions (e.g., accuracy, uniqueness, consistency, etc.);
4. Relative data quality dimensions with dependence on the user (e.g., user preferences), or on the application (e.g., criticality, conformance to business rules, etc.), or time-dependent (e.g., variability, volatility, currency, freshness, etc.) or with dependence on a given knowledge-state (e.g., data source reputation, verifiability).

In practice, assessing data quality in database systems has mainly been conducted by professional assessors with more and more cost-competitive auditing practices. Well-known approaches from industrial quality management and software quality assessment have been adapted for data quality and came up with an extension of metadata management. The use of metadata for data quality evaluation and improvement have been advocated by many authors, e.g., [14, 37, 52]. Although considerable efforts have been invested in the development of metadata standard vocabularies for the exchange of information across different applications domains

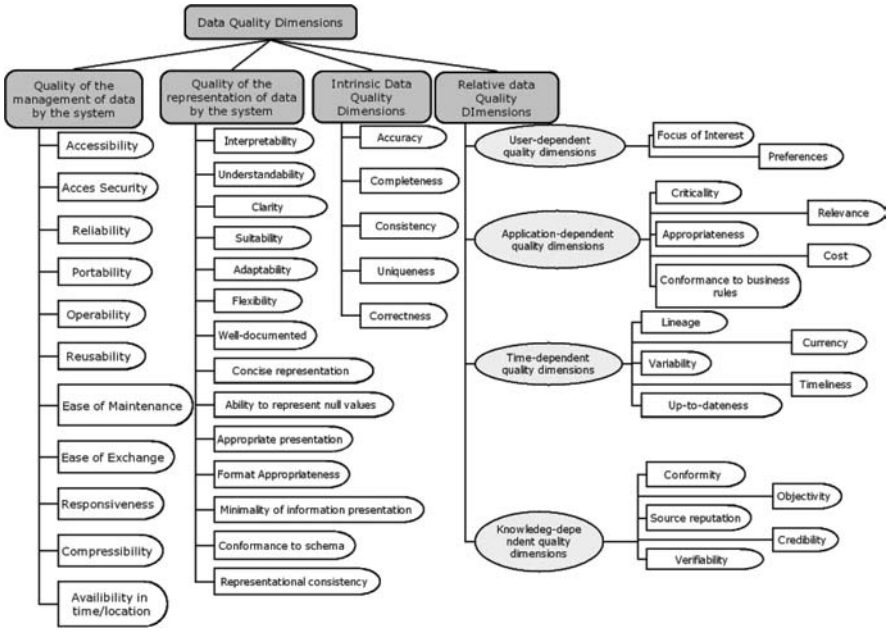

Fig. 1 Map of data quality dimensions

(e.g., for geographic information systems or digital libraries) including substantial work on data quality characterization in theses domains, the obvious fact is that in practice the quality metadata in many application domains remains a luxury. For data warehouse systems, many proposition concern the definition of quality models [27, 59, 61] with particular attention paid to data lineage and data transformation logs [13]. These metadata are very useful for the analysis and the interpretation of the probability distributions on the data, and also for debugging, implementing quality feedback loops, and analyzing the causes of data errors.

2.2 Data quality models

Several propositions fully integrate the modeling and the management of quality metadata as a whole part of the database design. Among these process-oriented approaches, the TDQM program (Total Data Quality Management) proposed by [62] provides a methodology including the modeling of data quality in the Entity-Relationship (ER) conceptual data model. It proposes also guidelines for adding step-by-step data quality metadata on each element of the model (entity, attribute, association). Classically, the first step consists in modeling the application domain. Figure 2 is adapted from [62] and illustrates quality metadata with an example of the purchase of a product by a company. The following step of modeling consists in making explicit of the subjective parameters of data quality for each entity, attribute, or association type (thus constituting the first level of quality metadata). These subjective parameters (e.g., reputation) are added as qualifiers to the attributes of the ER conceptual model. Objective quality indicators are then added

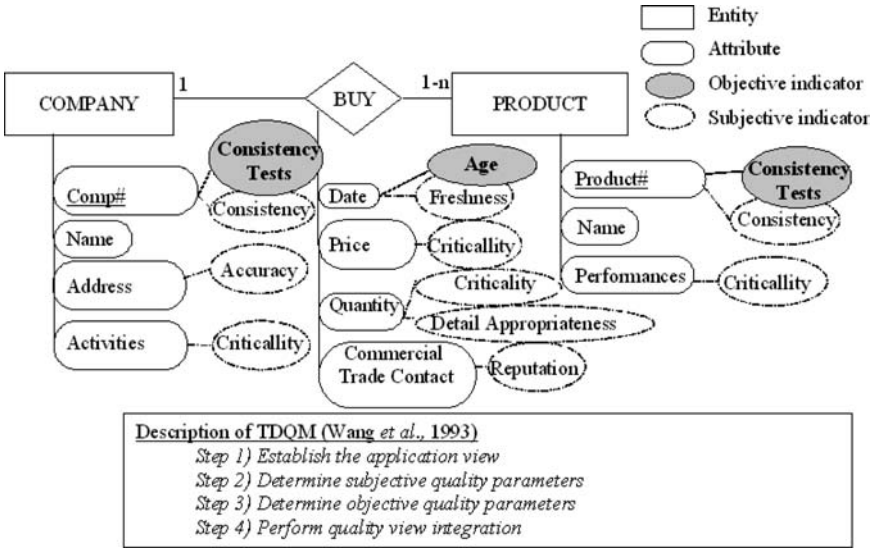

Fig. 2 Example of quality metadata modeling with TDQM

(e.g., age, consistency tests). They correspond to the measurable subjective parameters. They are automatically computed (by statistical methods or constraints verification). The last step of quality metadata modeling consists in integrating the previously defined quality views at the global level of the conceptual model.

Other works have taken other (still similar) approaches in modeling and capturing data quality [38, 53, 54]. Most of the proposed data quality models rely on data quality metadata being available, such as data source publishing such information. Unfortunately these approaches rely on precise and accurate metadata. However, such metadata are not always available with no unified standard describing data quality dimensions.

2.3 Data quality measures and quality metadata generation

Generating and managing statistical indicators of data quality have been the primary focus of methods of imputation: i.e., inferring missing data from statistical patterns of available data, predicting accuracy of the estimates based on the given data, data editing, and automating detection and handling of outliers [9, 14, 30, 68]. Utilization of statistical techniques for improving correctness of databases through introduction of new integrity constraints were first proposed in [25]. The constraints are derived from the database instances using the conventional statistical techniques (e.g., sampling and regression), and every update of the database is validated against these constraints. If an update does not comply with them, then the data administrator is alerted and prompted to check correctness of the update. Since databases model a portion of the real world which constantly evolves, the data quality estimates become outdated as time passes. Therefore, the estimation process should be repeated periodically depending on the dynamics of what is being modeled. The general trend is the use of artificial intelligence methods (machine learning, knowledge representation schemes, management of

uncertainty) for data validation [14, 15]. The use of machine learning techniques for data validation and correction was first presented by Parsaye and Chignell: rules inferred from the database instances by machine learning methods were used to identify outliers in data and facilitate the data validation process. Another similar approach was proposed by [56].

Exploratory Data Mining (EDM) [14] is a set of statistical techniques providing summaries that characterize data with typical values (e.g., medians and averages), variance, range, quantiles, and correlations. Used as a first pass, EDM methods can be advantageously employed for data pre-processing before carrying out more expensive analyses. EDM aims to be widely applicable while dealing with unfamiliar datasets. These techniques have a quick response time, and have results which are easy to interpret, to store, and to update. EDM can either be driven by the models to facilitate the use of parametric methods (model log-linear, for instance), or be driven by the data without any prior assumptions about inter-relationships between data. Well-known non-parametric techniques for exploring multivariate distributions (such as clustering, hierarchical, or neural networks) can be used. The EDM summaries (e.g., averages, standard deviations, medians, or other quantiles) can be used to characterize the data distribution, the correlations between attributes, or the center of the value distribution of a representative attribute. They can also be used to quantify and describe the dispersion of the attribute values around the center (form, density, symmetry, etc.). Other techniques are used to detect and cope with other problems on data, such as missing values, improbable outliers, and incomplete values. Concerning the techniques of analysis on missing data, the method of imputation through regression described by Little and Rubin [34] is generally used. Other methods such as Markov Chain Monte Carlo (MCMC) [55] are used to simulate data under the multivariate normal distribution assumption. Other references related to the problem of missing values are described by Pearson [44]. Concerning the outliers: The techniques of detection are mainly control charts and various techniques based: (i) on a mathematical model, (ii) on geometrical methods for distance measurement in the dataset (called geometric outliers), or (iii) on the distribution (or the density) of data population [30] extended by the concept of local exception (called local distributional outliers) [9]. The interested reader is invited to read the survey of Pyle [48], in particular for the use of entropy as a preliminary data characterization measure ([48], Sect. 11.3), and [14] for a description of these techniques.

3 Quality-aware rule mining

Our initial assumption is that the quality of an association rule depends on the quality of the data which the rule is computed from. This section will present the formal definitions of our approach that introduces data quality indicators and combines them for determining the quality of association rules.

3.1 Preliminary definitions for association rule quality

Let $\mathcal{I}$ be a set of literals, called *items*. An *association rule* R is an expression $LHS \rightarrow RHS$, where $LHS, RHS \subseteq \mathcal{I}$ and $LHS \cap RHS = \emptyset$. LHS and RHS are

conjunctions of variables such as the extension of the *L*eft-*H*and *S*ide *LHS* of the rule is: $g(LHS) = x_1 \wedge x_2 \wedge \ldots \wedge x_n$ and the extension of the *R*ight-*H*and *S*ide *RHS* of the rule is $g(RHS) = y_1 \wedge y_2 \wedge \ldots \wedge y_{n'}$.

Let $j (j = 1, 2, \ldots, k)$ be the k dimensions of data quality (e.g., data completeness, freshness, accuracy, consistency, completeness, credibility, etc.). Let $q_j(\mathcal{I}_i) \in [\mathsf{min}_{ij}, \mathsf{max}_{ij}]$ be a scoring value normalized in [0,1] for the dataset $\mathcal{I}_i$ on the quality dimension j ($\mathcal{I}_i \subseteq \mathcal{I}$). The vector, that keeps the values of all quality dimensions for each dataset $\mathcal{I}_i$ is called quality vector and noted $q(\mathcal{I}_i)$. The set of all possible quality vectors is called *quality space* $\mathcal{Q}$.

Definition 3.1 The quality of the association rule R is defined by a fusion function denoted $\circ_j$ specific for each quality dimension j that merges the components of the quality vectors of the datasets constituting the extension of the right-hand and left-hand sides of the rule. The quality of the rule R is defined as a k-dimensional vector such as:

$$q(R) = \begin{pmatrix} q_1(R) \\ q_2(R) \\ \ldots \\ q_k(R) \end{pmatrix} = \begin{pmatrix} q_1(LHS) \circ_1 q_1(RHS) \\ q_2(LHS) \circ_2 q_2(RHS) \\ \ldots \\ q_k(LHS) \circ_k q_k(RHS) \end{pmatrix}$$

$$= \begin{pmatrix} q_1(x_1) \circ_1 q_1(x_2) \circ_1 \ldots \circ_1 q_1(x_n) \circ_1 q_1(y_1) \circ_1 q_1(y_2) \circ_1 \ldots \circ_1 q_1(y_{n'}) \\ q_2(x_1) \circ_2 q_2(x_2) \circ_2 \ldots \circ_2 q_2(x_n) \circ_2 q_2(y_1) \circ_2 q_2(y_2) \circ_2 \ldots \circ_2 q_2(y_{n'}) \\ \ldots \\ q_k(x_1) \circ_k q_k(x_2) \circ_k \ldots \circ_k q_k(x_n) \circ_k q_k(y_1) \circ_k q_k(y_2) \circ_k \ldots \circ_k q_k(y_{n'}) \end{pmatrix} \quad (1)$$

The average quality of the association rule R denoted $\overline{q}(R)$ can be computed by a weighted sum of the quality vector components of the rule as follows:

$$\overline{q}(R) = \sum_{j=1}^{k} w_j \cdot q_j(R) \quad (2)$$

with w_j the weight of the quality dimension j.

We assume the weights are normalized:

$$\sum_{j=1}^{k} w_j = 1 \quad (3)$$

Definition 3.2 Let T be the domain of values of the quality score $q_j(\mathcal{I}_i)$ for the dataset $\mathcal{I}_i$ on the quality dimension j. The fusion function denoted $\circ_j$ is commutative and associative such as: $\circ_j : T \times T \rightarrow T$. The fusion function may have different definitions depending on the considered quality dimension j in order to suit the properties of each quality criterion.

Table 1 Examples of fusion functions for merging quality scores per dimension

j	Quality dimension	Function fusion $\circ_j$	Quality dimension of the rule $x \rightarrow y$
1	Freshness	$\min[q_1(x), q_1(y)]$	The freshness of the association rule $x \rightarrow y$ is estimated pessimistically as the lower score of freshness of the two datasets composing the rule.
2	Accuracy	$q_2(x) \cdot q_2(y)$	The accuracy of the association rule $x \rightarrow y$ is estimated as the probability of accuracy of the two datasets x and y of the rule.
3	Completeness	$q_3(x) + q_3(y) - q_3(x) \cdot q_3(y)$	The completeness of the association rule $x \rightarrow y$ is estimated as the probability that one of the two datasets of the rule is complete.
4	Consistency	$\max[q_4(x), q_4(y)]$	The consistency of the association rule $x \rightarrow y$ is estimated optimistically as the higher score of consistency of the two datasets composing the rule.

Based on the main data quality dimensions (e.g., freshness, accuracy, completeness, and consistency) defined in Sect. 2.1, Table 1 presents several definition examples of the fusion function (per quality dimension) based on the combination of quality scores of the two datasets x and y of the rule $x \rightarrow y$.

We consider that selecting an association rule (among the top N) is a decision that designates the rule as *legitimately interesting* (noted D_1), *potentially interesting* (D_2), or *not interesting* (D_3) based both on good interestingness measures and on the actual quality status of the datasets composing the left-hand and right-hand sides of the rule.

Consider the data item $x \in LHS \cup RHS$ of a given association rule, we use $P_{\mathrm{CE}}(x)$ to denote the probability that the item x will be classified as "erroneous" (i.e., "with low-quality") with reference to one or more quality dimensions relevant to the application (e.g., freshness, accuracy, etc.), and $P_{\mathrm{CC}}(x)$ denotes the probability that the item x will be classified as "correct" (i.e., "with correct quality" in the range of acceptable values for each pre-selected quality dimension). Also, $P_{\mathrm{AE}}(x)$ represents the probability that the item x is "actually erroneous" (*AE*) but detected correct, and $P_{\mathrm{AC}}(x)$ represents the probability that it is "actually correct" (*AC*) but detected erroneous (see Fig. 3).

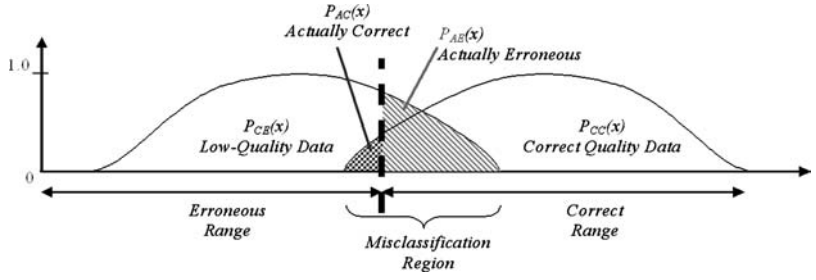

Fig. 3 Probabilities of detection of correct vs. low-quality data

3.2 Probabilistic decision model for quality-driven and cost-optimal association rule mining

For an arbitrary average quality vector $\overline{q}(R) \in \mathcal{Q}$ on the datasets in $LHS \cup RHS$ of the rule R, we denote by $P(\overline{q} \in \mathcal{Q}|CC)$ or $f_{\mathrm{CC}}(\overline{q})$ the conditional probability that the average quality vector $\overline{q}$ corresponds to the datasets that are classified as correct (CC). Similarly, we denote by $P(\overline{q} \in \mathcal{Q}|CE)$ or $f_{\mathrm{CE}}(\overline{q})$ the conditional probability that the average quality vector $\overline{q}$ corresponds to the datasets that are classified erroneous (CE). We denote by d the decision of the predicted class of the rule, i.e., *legitimately interesting* (D_1), *potentially interesting* (D_2), or *not interesting* (D_3), and by s the actual status of quality of the datasets upon which the rule has been computed.

Let us also denote by $P(d = D_i, s = j)$ and $P(d = D_i|s = j)$ correspondingly, the joint and the conditional probability that the decision D_i is taken, when the actual status of data quality is j (i.e., CC, CE, AE, AC).

We also denote by c_{ij} the cost of making a decision D_i for classifying an association rule with the actual data quality status j of the datasets composing the two parts of the rule. As an illustrative example, Table 2 shows tentative unit costs developed by the staff of a direct marketing department on the basis of consideration of the consequences of the decisions on selecting and using the discovered association rules in both cases: With and without misclassification. In Table 2, c_{10} is the cost of a confident decision (D_1) for the selection of a *legitimately interesting* rule based on correct-quality data (CC). c_{21} is the cost of a neutral decision (D_2) for the selection of a *potentially interesting* rule based on low-quality data (CE). c_{33} is the cost of a suspicious decision (D_3) for the selection of a *not interesting* rule based on low-quality data but actually detected as correct (AC). Based on the example presented in Table 2 where we can see how the cost of decisions could affect the result of the selection among interesting association rules (i.e., in the Top N list), we need to minimize the mean cost $\overline{c}$ that results from making such a decision. The corresponding mean cost $\overline{c}$ is written as

Table 2 Costs of various decisions for classifying association rules based on data quality

Decision for rule selection	Cost #	Data quality status	Cost ($) without misclassification	Cost ($) with misclassification
D_1	c_{10}	CC	0	0
	c_{11}	CE	1000	1000
	c_{12}	AE	0	1000
	c_{13}	AC	0	500
D_2	c_{20}	CC	50	50
	c_{21}	CE	50	50
	c_{22}	AE	0	500
	c_{23}	AC	0	500
D_3	c_{30}	CC	500	500
	c_{31}	CE	0	0
	c_{32}	AE	0	500
	c_{33}	AC	0	1000

follows:

$$\begin{aligned}\overline{c} = {} & c_{10}.P(d = D_1, s = CC) + c_{20}.P(d = D_2, s = CC) \\ & + c_{30}.P(d = D_3, s = CC) \\ & + c_{11}.P(d = D_1, s = CE) + c_{21}.P(d = D_2, s = CE) \\ & + c_{31}.P(d = D_3, s = CE) \\ & + c_{12}.P(d = D_1, s = AE) + c_{22}.P(d = D_2, s = AE) \\ & + c_{32}.P(d = D_3, s = AE) \\ & + c_{13}.P(d = D_1, s = AC) + c_{23}.P(d = D_2, s = AC) \\ & + c_{33}.P(d = D_3, s = AC) \end{aligned} \tag{4}$$

From the Bayes theorem, the following is true:

$$P(d = D_i, s = j) = P(d = D_i \mid s = j).P(s = j) \tag{5}$$

where $i = 1, 2, 3$ and $j = CC, CE, AE, AC$. The mean cost $\overline{c}$ in Eq. (4) based on Eq. (5) is written as follows:

$$\begin{aligned}\overline{c} = {} & c_{10} \cdot P(d = D_1 \mid s = CC) \cdot P(s = CC) \\ & + c_{20} \cdot P(d = D_2 \mid s = CC) \cdot P(s = CC) \\ & + c_{30} \cdot P(d = D_3 \mid s = CC) \cdot P(s = CC) \\ & + c_{11} \cdot P(d = D_1 \mid s = CE) \cdot P(s = CE) \\ & + c_{21} \cdot P(d = D_2 \mid s = CE) \cdot P(s = CE) \\ & + c_{31} \cdot P(d = D_3 \mid s = CE) \cdot P(s = CE) \\ & + c_{12} \cdot P(d = D_1 \mid s = AE) \cdot P(s = AE) \\ & + c_{22} \cdot P(d = D_2 \mid s = AE) \cdot P(s = AE) \\ & + c_{32} \cdot P(d = D_3 \mid s = AE) \cdot P(s = AE) \\ & + c_{13} \cdot P(d = D_1 \mid s = AC) \cdot P(s = AC) \\ & + c_{23} \cdot P(d = D_2 \mid s = AC) \cdot P(s = AC) \\ & + c_{33} \cdot P(d = D_3 \mid s = AC) \cdot P(s = AC) \end{aligned} \tag{6}$$

Let us also assume that $\overline{q}$ is the average quality vector drawn randomly from the space of all quality vectors of the item sets of the rule. The following equality holds for the conditional probability $P(d = D_i \mid s = j)$:

$$P(d = D_i \mid s = j) = \sum_{\overline{q} \in \mathcal{Q}_i} f_j(\overline{q}) \tag{7}$$

where $i = 1, 2, 3$ and $j = CC, CE, AE, AC$.

f_j is the probability density of the quality vectors when the actual quality status is j. Every point $\overline{q}$ in the quality space $\mathcal{Q}$ belongs to the partitions of quality $\mathcal{Q}_1$ or $\mathcal{Q}_2$ or $\mathcal{Q}_3$ that correspond respectively to the partitions of the decision space: D_1, or D_2 or D_3 in such a way that its contribution to the mean cost is minimum.

We denote the a priori probability of CC or else $P(s = CC)$ as π^0, the a priori probability of $P(s = AC) = \pi^0_{\mathrm{AC}}$, the a priori probability of $P(s = AE) = \pi^0_{\mathrm{AE}}$ and the a priori probability of $P(s = CE) = 1 - (\pi^0 + \pi^0_{\mathrm{AE}} + \pi^0_{\mathrm{AC}})$.

For the sake of clarity, we define three variables denoted κ_{AE}, κ_{AC}, and κ_{CE} as follows:

$$\kappa_{\mathrm{AE}} = \frac{f_{\mathrm{AE}}}{f_{\mathrm{CC}}} \cdot \frac{\pi^0_{\mathrm{AE}}}{\pi^0} \tag{8}$$

$$\kappa_{\mathrm{AC}} = \frac{f_{\mathrm{AC}}}{f_{\mathrm{CC}}} \cdot \frac{\pi^0_{\mathrm{AC}}}{\pi^0} \tag{9}$$

$$\kappa_{\mathrm{CE}} = \frac{f_{\mathrm{CE}}}{f_{\mathrm{CC}}} \cdot \frac{1 - (\pi^0 + \pi^0_{\mathrm{AC}} + \pi^0_{\mathrm{AE}})}{\pi^0} \tag{10}$$

By using Eq. (7) and by dropping the dependent vector variable $\overline{q}$, Eq. (6) becomes:

$$\begin{aligned} \overline{c} = \pi^0 \cdot f_{\mathrm{CC}} \cdot \Bigg(& \sum_{\overline{q} \in \mathcal{Q}_1} [c_{10} + c_{11} \cdot \kappa_{\mathrm{CE}} + c_{12} \cdot \kappa_{\mathrm{AE}} + c_{13} \cdot \kappa_{\mathrm{AC}}] \\ & + \sum_{\overline{q} \in \mathcal{Q}_2} [c_{20} + c_{21} \cdot \kappa_{\mathrm{CE}} + c_{22} \cdot \kappa_{\mathrm{AE}} + c_{23} \cdot \kappa_{\mathrm{AC}}] \\ & + \sum_{\overline{q} \in \mathcal{Q}_3} [c_{30} + c_{31} \cdot \kappa_{\mathrm{CE}} + c_{32} \cdot \kappa_{\mathrm{AE}} + c_{33} \cdot \kappa_{\mathrm{AC}}] \Bigg) \end{aligned} \tag{11}$$

3.2.1 Cost-optimal selection of rule with misclassification

In the case of misclassification, minimizing the mean cost $\bar{c}$ in Eq. (11) will lead to the optimal selection for the three sets of rules which we denote by D^0_1, D^0_2, and D^0_3. To minimize the cost $\bar{c}$, a point $\overline{q}$ is assigned to one of the three optimal areas as follows:
To D^0_1 iff:

$$\begin{aligned} & c_{10} + c_{11} \cdot \kappa_{\mathrm{CE}} + c_{12} \cdot \kappa_{\mathrm{AE}} + c_{13} \cdot \kappa_{\mathrm{AC}} \\ & \quad \leq c_{30} + c_{31} \cdot \kappa_{\mathrm{CE}} + c_{32} \cdot \kappa_{\mathrm{AE}} + c_{33} \cdot \kappa_{\mathrm{AC}} \\ \text{and,} \quad & c_{10} + c_{11} \cdot \kappa_{\mathrm{CE}} + c_{12} \cdot \kappa_{\mathrm{AE}} + c_{13} \cdot \kappa_{\mathrm{AC}} \\ & \quad \leq c_{20} + c_{21} \cdot \kappa_{\mathrm{CE}} + c_{22} \cdot \kappa_{\mathrm{AE}} + c_{23} \cdot \kappa_{\mathrm{AC}} \end{aligned}$$

To D^0_2 iff:

$$\begin{aligned} & c_{20} + c_{21} \cdot \kappa_{\mathrm{CE}} + c_{22} \cdot \kappa_{\mathrm{AE}} + c_{23} \cdot \kappa_{\mathrm{AC}} \\ & \quad \leq c_{30} + c_{31} \cdot \kappa_{\mathrm{CE}} + c_{32} \cdot \kappa_{\mathrm{AE}} + c_{33} \cdot \kappa_{\mathrm{AC}} \\ \text{and,} \quad & c_{20} + c_{21} \cdot \kappa_{\mathrm{CE}} + c_{22} \cdot \kappa_{\mathrm{AE}} + c_{23} \cdot \kappa_{\mathrm{AC}} \\ & \quad \leq c_{10} + c_{11} \cdot \kappa_{\mathrm{CE}} + c_{12} \cdot \kappa_{\mathrm{AE}} + c_{13} \cdot \kappa_{\mathrm{AC}} \end{aligned}$$

To D_3^0 iff:

$$\begin{aligned} & c_{30} + c_{31} \cdot \kappa_{\mathrm{CE}} + c_{32} \cdot \kappa_{\mathrm{AE}} + c_{33} \cdot \kappa_{\mathrm{AC}} \\ & \quad \leq c_{10} + c_{11} \cdot \kappa_{\mathrm{CE}} + c_{12} \cdot \kappa_{\mathrm{AE}} + c_{13} \cdot \kappa_{\mathrm{AC}} \\ \text{and,} \quad & c_{30} + c_{31} \cdot \kappa_{\mathrm{CE}} + c_{32} \cdot \kappa_{\mathrm{AE}} + c_{33} \cdot \kappa_{\mathrm{AC}} \\ & \quad \leq c_{20} + c_{21} \cdot \kappa_{\mathrm{CE}} + c_{22} \cdot \kappa_{\mathrm{AE}} + c_{23} \cdot \kappa_{\mathrm{AC}} \end{aligned}$$

The three decision areas for rule selection are then defined as follows:

$$D_1^0 = \begin{cases} \bar{q} : \kappa_{\mathrm{CE}} \leq \dfrac{(c_{30} - c_{10})}{(c_{11} - c_{31})} + \kappa_{\mathrm{AE}} \cdot \dfrac{(c_{32} - c_{12})}{(c_{11} - c_{31})} + \kappa_{\mathrm{AC}} \cdot \dfrac{(c_{33} - c_{13})}{(c_{11} - c_{31})} \\ \text{and, } \kappa_{\mathrm{CE}} \leq \dfrac{(c_{20} - c_{10})}{(c_{11} - c_{21})} + \kappa_{\mathrm{AE}} \cdot \dfrac{(c_{22} - c_{12})}{(c_{11} - c_{21})} + \kappa_{\mathrm{AC}} \cdot \dfrac{(c_{23} - c_{13})}{(c_{11} - c_{21})} \end{cases}$$

$$D_2^0 = \begin{cases} \bar{q} : \kappa_{\mathrm{CE}} \leq \dfrac{(c_{30} - c_{20})}{(c_{21} - c_{31})} + \kappa_{\mathrm{AE}} \cdot \dfrac{(c_{32} - c_{22})}{(c_{21} - c_{31})} + \kappa_{\mathrm{AC}} \cdot \dfrac{(c_{33} - c_{23})}{(c_{21} - c_{31})} \\ \text{and, } \kappa_{\mathrm{CE}} \geq \dfrac{(c_{20} - c_{10})}{(c_{11} - c_{21})} + \kappa_{\mathrm{AE}} \cdot \dfrac{(c_{22} - c_{12})}{(c_{11} - c_{21})} + \kappa_{\mathrm{AC}} \cdot \dfrac{(c_{23} - c_{13})}{(c_{11} - c_{21})} \end{cases}$$

$$D_3^0 = \begin{cases} \bar{q} : \kappa_{\mathrm{CE}} \geq \dfrac{(c_{30} - c_{10})}{(c_{11} - c_{31})} + \kappa_{\mathrm{AE}} \cdot \dfrac{(c_{32} - c_{12})}{(c_{11} - c_{31})} + \kappa_{\mathrm{AC}} \cdot \dfrac{(c_{33} - c_{13})}{(c_{11} - c_{31})} \\ \text{and, } \kappa_{\mathrm{CE}} \geq \dfrac{(c_{30} - c_{20})}{(c_{21} - c_{31})} + \kappa_{\mathrm{AE}} \cdot \dfrac{(c_{32} - c_{22})}{(c_{21} - c_{31})} + \kappa_{\mathrm{AC}} \cdot \dfrac{(c_{33} - c_{23})}{(c_{21} - c_{31})} \end{cases}$$

In the case of misclassification these inequalities give rise to three different threshold values λ, ρ, and ν (respectively for *legitimately*, *potentially*, and *not interesting* rules) in the decision space (see Fig. 4). D_2^0 can be seen as the union of two subareas representing the *potentially legitimately interesting* rules and the *potentially not interesting* rules.

$$\lambda = \frac{1}{(c_{11} - c_{21})} \cdot (c_{20} - c_{10} + \kappa_{\mathrm{AE}} \cdot (c_{22} - c_{12}) + \kappa_{\mathrm{AC}} \cdot (c_{23} - c_{13})) \tag{12}$$

$$\rho = \frac{1}{(c_{11} - c_{31})} \cdot (c_{30} - c_{10} + \kappa_{\mathrm{AE}} \cdot (c_{32} - c_{12}) + \kappa_{\mathrm{AC}} \cdot (c_{33} - c_{13})) \tag{13}$$

$$\nu = \frac{1}{(c_{21} - c_{31})} \cdot (c_{30} - c_{20} + \kappa_{\mathrm{AE}} \cdot (c_{33} - c_{23}) + \kappa_{\mathrm{AC}} \cdot (c_{32} - c_{22})) \tag{14}$$

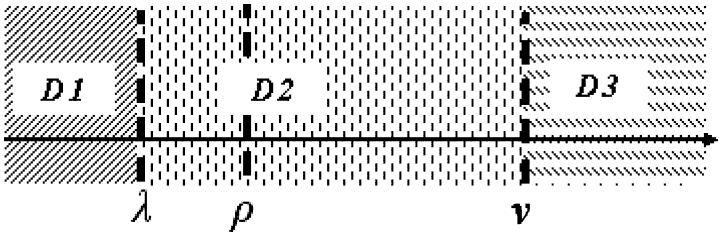

Fig. 4 Decision areas for rule selection

3.2.2 Cost-optimal selection of rule without misclassification

For the sake of simplicity, let us now consider the case of the absence of the misclassification region. f_{AC}, f_{AE}, κ_{AC}, κ_{AE}, π^0_{AE}, and π^0_{AC} are null. Without misclassification region $P(s = CE)$ could be simplified as $1 - \pi^0$. κ_{CE} is equal to $\frac{f_{CE}}{f_{CC}}$ and we can thus simplify the inequalities above:

$$D_1^0 = \left\{ \overline{q} : \frac{f_{CE}}{f_{CC}} \leq \frac{\pi^0}{1-\pi^0} \cdot \frac{c_{30} - c_{10}}{c_{11} - c_{31}} \quad \text{and,} \quad \frac{f_{CE}}{f_{CC}} \leq \frac{\pi^0}{1-\pi^0} \cdot \frac{c_{20} - c_{10}}{c_{11} - c_{21}} \right\} \tag{15}$$

$$D_2^0 = \left\{ \overline{q} : \frac{f_{CE}}{f_{CC}} \leq \frac{\pi^0}{1-\pi^0} \cdot \frac{c_{30} - c_{20}}{c_{21} - c_{31}} \quad \text{and,} \quad \frac{f_{CE}}{f_{CC}} \geq \frac{\pi^0}{1-\pi^0} \cdot \frac{c_{20} - c_{10}}{c_{11} - c_{21}} \right\} \tag{16}$$

$$D_3^0 = \left\{ \overline{q} : \frac{f_{CE}}{f_{CC}} \geq \frac{\pi^0}{1-\pi^0} \cdot \frac{c_{30} - c_{10}}{c_{11} - c_{31}} \quad \text{and,} \quad \frac{f_{CE}}{f_{CC}} \geq \frac{\pi^0}{1-\pi^0} \cdot \frac{c_{30} - c_{20}}{c_{21} - c_{31}} \right\} \tag{17}$$

Without misclassification the inequalities (15)–(17) give rise to three different threshold values λ, ρ, and ν (respectively for *legitimately, potentially, and not interesting* rules) in the decision space that define concretely the decision regions based on the cost of rule selection decision with the following relationship:

$$\lambda = \frac{\pi^0}{1-\pi^0} \cdot \frac{c_{20} - c_{10}}{c_{11} - c_{21}} \leq \rho = \frac{\pi^0}{1-\pi^0} \cdot \frac{c_{30} - c_{10}}{c_{11} - c_{31}} \leq \nu = \frac{\pi^0}{1-\pi^0} \cdot \frac{c_{30} - c_{20}}{c_{21} - c_{31}} \tag{18}$$

Additionally to the interestingness measures these thresholds can be used for quality awareness in association rule mining for a predictive selection of *legitimately interesting* rules based on the data quality indicators.

4 Experiments and results

In order to evaluate our decision model (in both cases with and without misclassification), we built an experimental system. The system relies on a data generator that automatically generates data quality metadata with a priori known characteristics. This system also allows us to perform controlled studies so as to establish data quality indicators and quality variations both on datasets and on discovered association rules. In this section, we present a set of experiments using the KDD-Cup-98 dataset from the UCI repository.[1] The KDD-Cup-98 dataset contains 191,779 records about individuals contacted in the 1997 mailing campaign. Each record is described by 479 non-target variables and two target variables indicating the "respond" / "not respond" classes and the actual donation in dollars. About 5% of records are "respond" records and the rest are "not respond"

[1] http://kdd.ics.uci.edu/databases/kddcup98/kddcup98.html for the dataset and http://www.kdnuggets.com/meetings/kdd98/kdd-cup-98.html for the results.

records. The KDD-Cup-98 competition task was to build a prediction model of the donation amount. The participants were contested on the sum of actual profit $\sum(\textit{actual donation} - \$0.68)$ over the validation records with predicted donation greater than the mailing cost \$0.68 (see [66] for details). Because we ignored the quality of the data collected during this campaign, we generated synthetic data quality indicators with different assumptions representing common data pollutions. In this experiment, our goal is to demonstrate that data quality variations may have a great impact on the significance of KDD-Cup-98 results (i.e., the top 10 discovered "respond" rules and profit predictions). Although data quality indicators do not affect the top 10 list of discovered association rules, they significantly change the reliability (and the quality) of this mining result and also the cost of the decisions relying on these rules.

The names, definitions, generated quality indicators for four data quality dimensions (i.e., freshness, accuracy, completeness, and consistency), average quality scores, and estimated probabilities per variable of the KDD-Cup-98 dataset are given in Table 3. For the sake of simplicity, we generated the quality dimension scores such as they are uniformly representative of the quality dimension of the values in the attribute domain. The average quality $\overline{q}$ per variable in Table 3 is computed from the equi-weighted function given in Eq. (2). f_{CC} in Table 3 (also noted $f_{\text{CC}}(\overline{q}(\mathcal{I}_i))$ in our formalism) is the probability density that the dataset $\mathcal{I}_i$ is considered "correct" (i.e., with "correct quality") when the average quality score of the variable $\mathcal{I}_i$ is $\overline{q}(\mathcal{I}_i)$. f_{CE} (also noted $f_{\text{CE}}(\overline{q}(\mathcal{I}_i))$) is the probability density that the dataset $\mathcal{I}_i$ is considered "erroneous" (i.e., with "low quality") when the average quality score of $\mathcal{I}_i$ is $\overline{q}(\mathcal{I}_i)$.

The top 10 a priori association rules discovered by [66] are given in Table 4 with the confidence, the support (in number of records), and the quality scores. Table 4 shows the score per quality dimension and the average quality score for each association rule. The scores are computed from the definitions of the quality dimensions given in Table 1 and the data quality scores previously given per attribute in Table 3.

In the next subsections, we study the impact of data quality variations on the decision cost for rule selection respectively in the two cases: With and without misclassification. We use the decision costs previously mentioned in Table 2 for classifying the rules based on the quality of their data.

4.1 Quality and cost of association rules without misclassification

First, we identify the value of the a priori probability that implies the largest amplitude of decision costs for the rule selection based on Table 2 and Eq. (11) for the top 10 rules discovered by [66]. Figure 5 shows this case for the a priori probability $\pi^0 = 0.200$ in the absence of misclassification region (i.e., $\pi^0_{\text{AE}} = \pi^0_{\text{AC}} = 0$). By using Eq. (18), we compute the values of the three decision thresholds for the rule selection with the a priori probability $\pi^0 = 0.200$. We obtain the following thresholds: $\lambda = 0.0131579$, $\rho = 0.125$, and $\nu = 2.25$. In order to be consistent with the conditional independency of the quality vector components we also need to take the logarithms of the thresholds values. By doing this we obtain: $\log(\lambda) = -1.8808$, $\log(\rho) = -0.9031$, and $\log(\nu) = 0.3522$. Based

Table 3 Quality measures and estimated probabilities of selected attributes of the KDD-Cup-98 dataset

		Quality						
Variable	Definition	Fresh.	Accur.	Compl.	Cons.	$\overline{q}$	f_{CC}	f_{CE}
AGE904	Average age of population	0.50	0.21	0.39	0.73	0.46	0.90	0.05
CHIL2	% Children age 7–13	0.16	0.99	0.75	0.71	0.65	0.95	0.10
DMA	DMA Code	0.49	0.58	0.16	0.95	0.55	0.95	0.01
EIC16	% Employed in public administration	0.03	0.56	0.33	0.61	0.38	0.98	0.01
EIC4	% Employed in manufacturing	0.17	0.37	0.87	0.15	0.39	0.90	0.20
ETH1	% White	0.21	0.76	0.50	0.53	0.50	0.55	0.15
ETH13	% Mexican	0.52	0.77	0.87	0.79	0.74	0.90	0.60
ETHC4	% Black $\leq$Age 15	0.84	0.52	0.32	0.35	0.51	0.95	0.45
HC6	% Owner occupied structures built since 1970	0.47	0.96	0.74	0.11	0.57	0.98	0.03
HHD1	% Households w/ related children	0.61	0.95	0.27	0.08	0.48	0.96	0.41
HU3	% Occupied housing units	0.07	0.40	0.18	0.57	0.30	0.94	0.53
HUPA1	% Housing units w/2 through 9 at the address	0.76	0.85	0.96	0.93	0.88	0.95	0.52
HVP5	% Home value $\geq$ \$50.000	0.99	0.88	0.38	0.95	0.80	0.94	0.05
NUMCHLD	Number of children	0.44	0.23	0.53	0.50	0.42	0.96	0.17
POP903	Number of households	0.77	0.52	0.74	0.61	0.66	0.87	0.15
RAMNT_22	Dollar amount of the gift for 95XK	0.37	0.95	0.95	0.75	0.76	0.84	0.25
RFA_11	Donor's RFA status as of 96X1 promotion date	0.59	0.34	0.34	0.76	0.51	0.95	0.12
RFA_14	Donor's RFA status as of 95NK promotion date	0.60	0.69	0.24	0.10	0.41	0.95	0.13
RFA_23	Donor's RFA status as of 94FS promotion date	0.34	0.01	0.23	0.63	0.30	0.97	0.55
RHP2	Average number of rooms per housing unit	0.66	0.72	0.08	0.26	0.43	0.98	0.20
TPE11	Mean travel time to work in minutes	0.20	0.26	0.78	0.32	0.39	0.85	0.05
WEALTH2	Wealth rating	0.24	0.82	0.41	0.58	0.51	0.87	0.05

on the values for these thresholds, we can assign each rule to one of the three decision areas. Table 5 shows the profit per rule predicted by [66], the decision cost of the rule selection computed from Table 2 and the decision area per rule.

We observe that only five rules (i.e., R1, R5, R7, R9, R10) are *potentially interesting* among the top 10 rules considering the quality of data they are computed from. With data quality awareness, the other rules (R2, R3, R4, R6, R8) are *not interesting* despite a good rank in the top 10 list. It is also interesting to notice that

Table 4 The top 10 "Respond" rules with confidence, support, and quality scores

Rule #	Association rule	(Conf.; Supp.)	Quality				
			Fresh.	Accur.	Compl.	Cons.	$\overline{q}$
R1	ETHC4 = [2.5,4.5], ETH1 = [22.84,29.76], HC6 = [60.91,68.53]	(0.11; 13)	0.21	0.38	0.79	0.53	0.48
R2	RFA_14 = f1d, ETH1 = [29.76,36.69]	(0.17; 8)	0.21	0.52	0.62	0.53	0.47
R3	HHD1 = [24.33,28.91], EIC4 = [33.72,37.36]	(0.12; 12)	0.17	0.35	0.90	0.15	0.39
R4	RFA_23 = s2g, ETH13 = [27.34,31.23]	(0.12; 16)	0.34	0.01	0.90	0.79	0.51
R5	EIC16 = [11.25,13.12], CHIL2 = [33,35.33], HC6 = [45.69,53.30]	(0.16; 11)	0.03	0.53	0.77	0.71	0.51
R6	RHP2 = [36.72,40.45], AGE904 = [42.2,44.9]	(0.16; 7)	0.50	0.15	0.44	0.73	0.46
R7	HVP5 = [56.07,63.23], ETH13 = [31.23,35.61], RAMNT_22 = [7.90,10.36]	(0.14; 10)	0.37	0.65	0.68	0.95	0.66
R8	NUMCHLD = [2.5,3.25], HU3 = [66.27,70.36]	(0.08; 31)	0.07	0.09	0.61	0.57	0.34
R9	RFA_11 = f1g, DMA = [743,766.8], POP903 = [4088.208,4391.917], WEALTH2 = [6.428571,7.714286]	(0.25; 8)	0.24	0.08	0.72	0.95	0.50
R10	HUPA1 = [41.81+,], TPE11 = [27,64,31.58]	(0.23; 9)	0.20	0.22	0.99	0.93	0.59

Table 5 The top 10 "Respond" rules with profit, cost, and decision area for $\pi^0 = 0.200$ without misclassification

Rule #	Profit ($)	Cost ($)	Decision area
R1	81.11	53	Potentially
R2	61.73	109.5	Not
R3	47.07	113	Not
R4	40.82	130	Not
R5	35.17	34.7	Potentially
R6	28.71	109	Not
R7	24.32	62.8	Potentially
R8	19.32	190	Not
R9	17.59	49.6	Potentially
R10	9.46	40.8	Potentially

the profit per rule predicted by [66] may be considerably counterbalanced by the cost of the rule computed from low-quality data (although it depends from initial costs defined in Table 2). The second best rule R2 whose predicted profit is $61.73 has a cost of $109.5 and is classified as *not interesting* due to the low quality of its datasets.

Let us now introduce different variations on the average quality of the datasets composing the rules. Based on the costs in Table 2, Fig. 6 shows the behavior of the decision cost for the rule selection when data quality varies from the initial

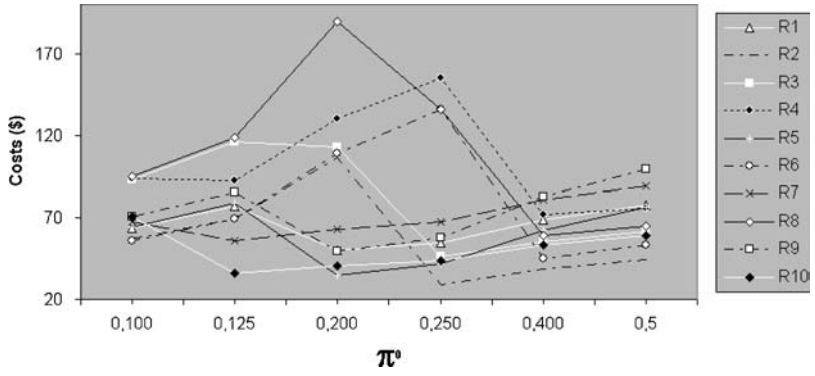

Fig. 5 Decision costs for rule selection with a priori probability in [0.1, 0.5] without misclassification

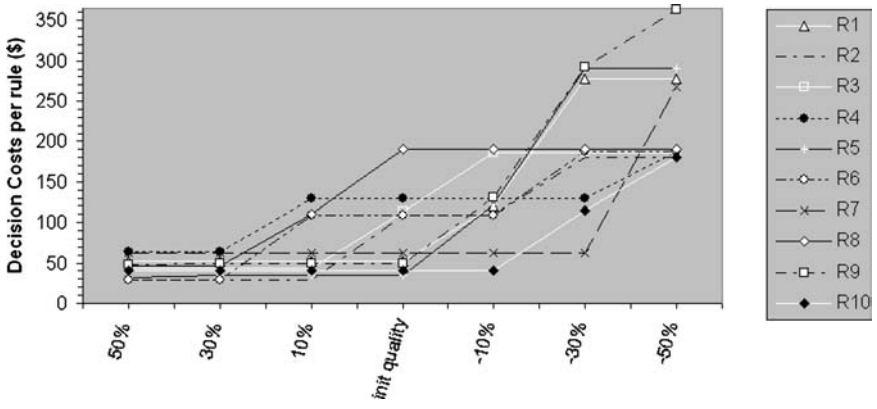

Fig. 6 Decision cost for rule selection with different data quality variations without misclassification for the a priori probability $\pi^0 = 0.200$

average quality (Init Quality) down to -10, -30, and -50% and up to $+10$, $+30$, and $+50\%$ for the a priori probability $\pi^0 = 0.200$ in the absence of misclassification.

In Fig. 6, we observe that the quality degradation of the datasets composing the rules increases the cost of these rules with various amplitudes shown in Fig. 7 (with a maximal quality degradation noted qual-50% and a maximal quality amelioration noted qual$+50\%$). Data quality amelioration (from $+30$ to $+50\%$) implies a stabilization trend of the decision cost for the rule selection (between \$20 and \$70).

Another interesting result is shown in Fig. 8 where the decisions for rule selection change simultaneously with the data quality variations. Among the top 10 interesting rules discovered by [66] with the initial data quality (noted Init Qual), five rules (R1, R5, R7, R9, and R10) were *potentially* worth being selected based on their average data quality and five rules were *not interesting* (R2, R3, R4, R6, and R8). While increasing data quality up to $+30\%$, three rules become *legitimately interesting* (R5, R7, and R9). The others become *potentially interesting*.

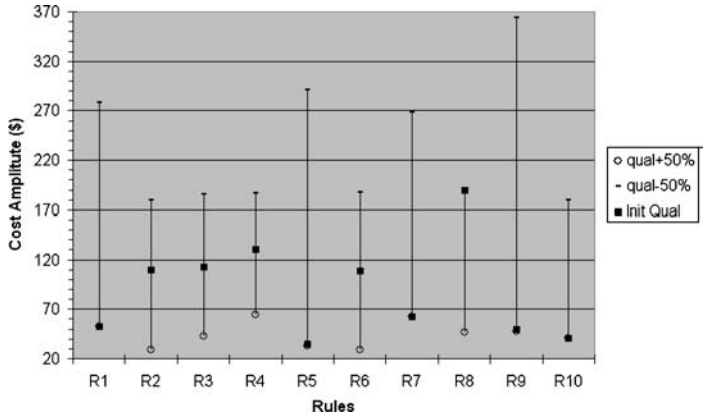

Fig. 7 Amplitude of cost variations depending on data quality variations without misclassification for the a priori probability $\pi^0 = 0.200$

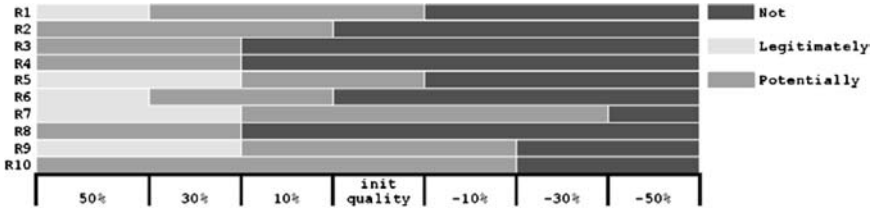

Fig. 8 Decision status on rule selection for data quality variations without misclassification for $\pi^0 = 0.200$

4.2 Quality and cost of association rules with misclassification

In the case of misclassification (with $f_{AC} = f_{AE} = f_{CC}$) we observe that the cost is high (between \$248.40 and \$594.30 compared to the case without misclassification between \$34.7 and \$190). The amplitude of the decision cost per rule depending on the a priori probability is stable (see Fig. 9) and the rule costs are stratified per rule.

While keeping the a priori probability $\pi^0 = 0.200$ and using Eqs. (12)– (14), we compute the values of the three decision thresholds for rule selection with misclassification and we obtain: $\lambda = 0.1053$, $\rho = 0.1667$, and $\nu = 4.6667$. Based on the values for these thresholds, we can assign the rules to one of the three decision areas (see Table 6). In the case of misclassification with the a priori probability $\pi^0 = 0.200$ it is interesting to notice that the cost per rule may be increased from 1.7 to 13.5 times (respectively for R8 and for R5) compared to the case of correct classification. This is mainly due to the cost of: (i) confident decisions for the selection of the rules computed from low-quality data that are incorrectly classified, and (ii) suspicious decisions for the selection of the rules computed from correct-quality that are incorrectly classified. With different variations on the average quality of the datasets composing the rules (with -10, -30, -50% and $+10$, $+30$, $+50\%$ from Init Quality) and based on the costs given in Table 2 in the case of misclassification, we study the behavior of the decision cost for the rule selection.

Table 6 The top 10 "Respond" rules with profit, cost, and decision area for $\pi^0 = 0.200$ with misclassification

Rule #	Profit ($)	Cost ($)	Decision area
R1	81.11	415.70	Potentially
R2	61.73	248.40	Potentially
R3	47.07	315.90	Not
R4	40.82	356.00	Not
R5	35.17	469.80	Potentially
R6	28.71	308.30	Potentially
R7	24.32	455.80	Legitimately
R8	19.32	325.00	Not
R9	17.59	592.30	Potentially
R10	9.46	305.10	Potentially

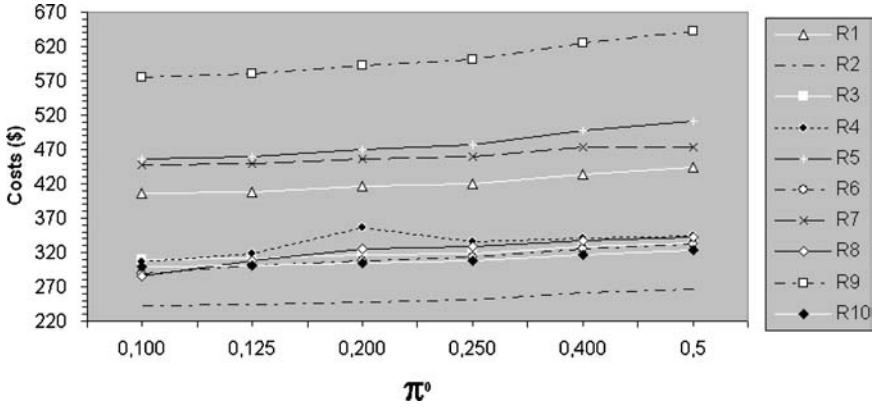

Fig. 9 Decision costs for rule selection with a priori probability in [0.1, 0.5] with misclassification

Figure 10 shows that the costs are relatively stable with smaller amplitudes and more distinct and staggered cost ranges than in the case without misclassification (see Figs. 10 and 11 compared to Figs. 6 and 7) with the exceptions of the maxima of data quality variations (i.e., $\pm 50\%$) when the misclassification has more impact on decision costs. In Fig. 12, only R7 is *legitimately interesting* among the top 10 rules discovered by [66] with the initial data quality (Init Qual). Three rules are *not interesting* (R8, R3, and R4) and the other six rules are *potentially interesting*. Misclassification globally attenuates the "verdict" that classifies each rule correspondingly to one of the decision areas for the *legitimately*, *potentially*, or *not interesting* rules. Some rules (e.g., R8) keep the same behavior with or without misclassification when data quality varies.

The observations made on this set of experiments offer two main interesting research perspectives for both association rule mining and data quality management: First, for proposing a post-filtering rule process based on data quality indicators and optimal decision costs for rule selection, and secondly, for the optimal scheduling of data quality improvement activities (with data cleaning techniques and Extraction-Transformation-Loading tools for instance) that could

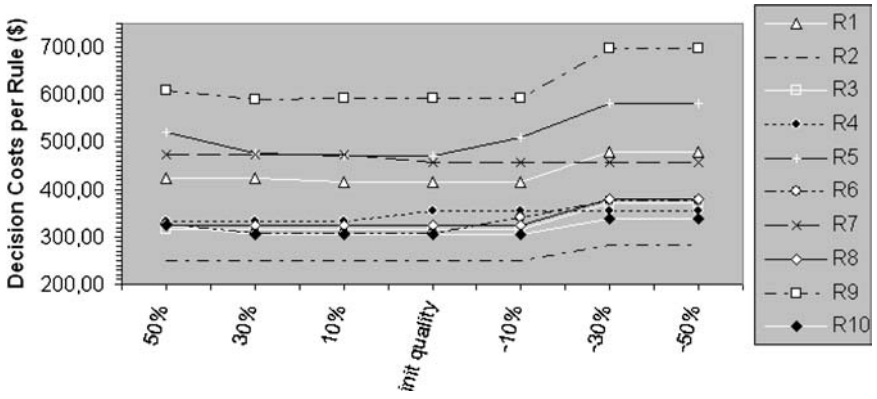

Fig. 10 Decision cost for rule selection with different data quality variations with misclassification for the a priori probability $\pi^0 = 0.200$

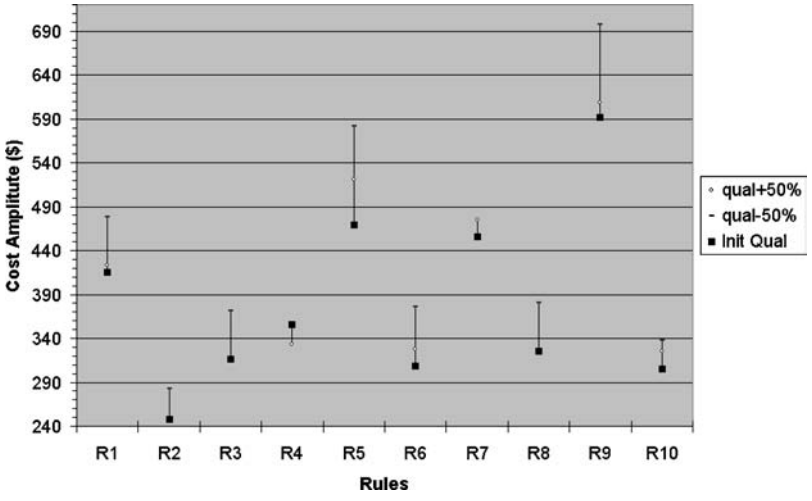

Fig. 11 Amplitude of cost variations depending on data quality variations with misclassification for the a priori probability $\pi^0 = 0.200$

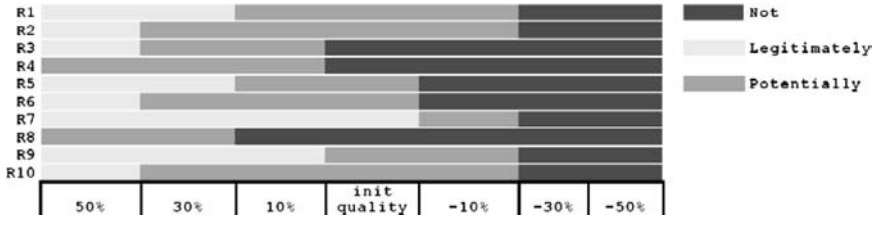

Fig. 12 Decision status on rule selection for data quality variations with misclassification for $\pi^0 = 0.200$

be relevantly used on targeted datasets for reaching the quality level expected for the mining results and for setting up quality improvement strategies for the KDD process.

5 Conclusion

This paper gives an overview of data quality characterization and management techniques that can be employed for improving quality awareness of knowledge discovery and data mining processes. The original contribution of this paper is twofold: first, we propose a method for scoring the quality of association rules that combines and integrates measures of data quality; secondly, we propose a probabilistic cost model for estimating the cost of selecting *legitimately interesting* association rules based on correct-quality data. The model defines the thresholds of three decision areas for the predicted class of the discovered rules (i.e., *legitimately*, *potentially*, or *not interesting*). To validate our approach, our experiments on the KDD-Cup-98 dataset consisted of: (i) generating synthetic data quality indicators, (ii) computing the average quality of the top 10 "respond" association rules discovered by [66], (iii) computing the cost of selecting low- versus correct-quality rules and the decision areas they belong to, (iv) examining the cost and the decision status for rule selection when the quality of underlying data varies.

Our experiments confirm our original assumption that is: Interestingness measures are not self-sufficient and the quality of association rules depends on the quality of the data which the rules are computed from. Data quality includes various dimensions (such as data freshness, accuracy, completeness, etc.) which should be combined and exploited for effective data quality-aware mining.

Our future plans regarding this work are to study the optimality of our decision model, to propose error estimation and to validate the model with experiments on large real biomedical datasets. This will extend our work in [6] with exploiting operational bio-data quality indicators for improving data quality awareness in biomedical association rule discovery.

References

1. Avenali A, Batini C, Bertolazzi P, Missier P (2004) A formulation of the data quality optimization problem. In: Proceedings of the international CAiSE workhop on data and information quality (DIQ), Riga, Latvia, pp 49–63
2. Ballou DP, Pazer H (1995) Designing information systems to optimize the accuracy-timeliness trade-off. Inf Syst Res 6(1)
3. Ballou DP, Pazer H (2002) Modeling completeness versus consistency trade-offs in information decision contexts. IEEE Trans Knowl Data Eng (TDKE) 15(1):240–243
4. Batini C, Catarci T, Scannapiceco M (2004) A survey of data quality issues in cooperative information systems. In: Tutorial presented at the 23rd international conference on conceptual modeling (ER), Shanghai, China
5. Benjelloun O, Garcia-Molina H, Su Q, Widom J (2005) Swoosh: A generic approach to entity resolution. Technical Report, Stanford Database Group
6. Berti-Équille L, Moussouni F (2005) Quality-aware integration and warehousing of genomic data. In: Proceedings of the 10th international conference on information quality (IQ'05), MIT, Cambridge, USA
7. Bilenko M, Mooney RJ (2003) Adaptive duplicate detection using learnable string similarity measures. In: Proceedings of the 9th ACM SIGKDD conference on knowledge discovery and data mining (KDD), Washington, DC, USA, pp 39–48

8. Bouzeghoub M, Peralta V (2004) A framework for analysis of data freshness. In: Proceedings of the 1st ACM SIGMOD workshop on information quality in information systems (IQIS), Paris, France, pp 59–67
9. Breunig M, Kriegel H, Ng R, Sander J (2000) LOF: Identifying density-based local outliers. In: Proceedings of 2000 ACM SIGMOD conference, Dallas, TX, USA, pp 93–104
10. Brodie ML (1980) Data quality in information systems. Inform Manage 3:245–258
11. Celko J, McDonald J (1995) Don't warehouse dirty data. Datamation 41(18)
12. Chaudhuri S, Ganjam K, Ganti V, Motwani R (2003) Robust and efficient fuzzy match for online data cleaning. In: Proceedings of the 2003 ACM SIGMOD international conference on management of data, San Diego, CA, USA, pp 313–324
13. Cui Y, Widom J (2001) Lineage tracing for general data warehouse transformation. In: Proceedings of the 27th international conference on very large data bases (VLDB), Roma, Italy, September 11–14, pp 471–480
14. Dasu T, Johnson T (2003) Exploratory data mining and data cleaning. Wiley, New York
15. Dasu T, Johnson T, Muthukrishnan S, Shkapenyuk V (2002) Mining database structure or, how to build a data quality browser. In: Proceedings of the 2002 ACM SIGMOD international conference on management of data, Madison, WI, USA, pp 240–251
16. De Giacomo G, Lembo D, Lenzerini M, Rosati R (2004) Tackling inconsistencies in data integration through source preferences. In: Proceedings of the 1st ACM SIGMOD workshop on information quality in information systems (IQIS), Paris, France, pp 27–34
17. Delen G, Rijsenbrij D (1992) The specification, engineering and measurement of information systems quality. J Softw Syst 17:205–217
18. Elfeky MG, Verykios VS, Elmagarmid AK (2002) Tailor: A record linkage toolbox. In: Proceedings of the 19th international conference on data engineering (ICDE), San Jose, CA, USA, pp 1–28
19. English L (1998) Improving data warehouse and business information quality. Wiley, New York
20. Fan K, Lu H, Madnick S, Cheung D (2001) Discovering and reconciling value conflicts for numerical data integration. Inform Syst 26(8):235–656
21. Fellegi IP, Sunter AB (1969) A theory for record linkage. J Am Stat Assoc 64:1183–1210
22. Fox C, Levitin A, Redman T (1994) The notion of data and its quality dimensions. Information Processing and Management 30(1)
23. Gravano L, Ipeirotis PG, Koudas N, Srivastava D (2003) Text joins in an RDBMS for web data integration. In: Proceedings of the 12th international world wide web conference (WWW), Budapest, Hungary, pp 90–101
24. Hernandez M, Stolfo S (1998) Real-world data is dirty: Data cleansing and the merge/purge problem. Data Min Knowl Discov 2(1):9–37
25. Hou WC, Zhang Z (1995) Enhancing database correctness: A statistical approach. In: Proceedings of the 1995 ACM SIGMOD international conference on management of data, San Jose, CA, USA
26. Huang K, Lee Y, Wang R (1999) Quality information and knowledge management. Prentice Hall, New Jersey
27. Jarke M, Jeusfeld MA, Quix C, Vassiliadis P (1998) Architecture and quality in data warehouses. In: Proceedings of the 10th international conference on advanced information systems engineering (CAiSE), Pisa, Italy, pp 93–113
28. Johnson T, Dasu T (1998) Comparing massive high-dimensional data sets. In: Proceedings of the 4th international conference KDD, New York City, New York, USA, pp 229–233
29. Kahn B, Strong D, Wang R (2002) Information quality benchmark: Product and service performance. Com. ACM 45(4):184–192
30. Knorr E, Ng R (1998) Algorithms for mining distance-based outliers in large datasets. In: Proceedings of the 24th international conference on very large data bases (VLDB), New York City, USA, pp 392–403
31. Lavrač N, Flach PA, Zupan B (1999) Rule evaluation measures: A unifying view. In: Proceedings of the international workshop on inductive logic programming (ILP), Bled, Slovenia, pp 174–185
32. Liepins G, Uppuluri V (1990) Data quality control: Theory and pragmatics. Marcel Dekker, New York

33. Lim L, Srivastava J, Prabhakar S, Richardson J (1993) Entity identification in database integration. In: Proceedings of the 9th international conference on data engineering (ICDE), Vienna, Austria, pp 294–301
34. Little RJ, Rubin DB (1987) Statistical analysis with missing data. Wiley, New York
35. Liu L, Chi L (2002) Evolutionary data quality. In: Proceedings of the 7th international conference on information quality (IQ), MIT, Cambridge, USA
36. McCallum A, Nigam K, Ungar LH (2000) Efficient clustering of high-dimensional data sets with application to reference matching. In: Proceedings of the 6th ACM SIGKDD conference on knowledge discovery and data mining (KDD), Boston, MA, USA, pp 169–178
37. Mihaila GA, Raschid L, Vidal M (2000) Using quality of data metadata for source selection and ranking. In: Proceedings of the 3rd international WebDB workshop, Dallas, TX, USA, pp 93–98
38. Missier P, Batini C (2003) A multidimensional model for information quality in CIS. In: Proceedings of the 8th international conference on information quality (IQ), MIT, Cambridge, MA, USA
39. Monge A (2000) Matching algorithms within a duplicate detection system. IEEE Data Eng Bull 23(4):14–20
40. Müller H, Leser U, Freytag JC (2004) Mining for patterns in contradictory data. In: Proceedings of the 1st ACM SIGMOD workshop on information quality in information systems (IQIS) in conjunction with ACM PODS/SIGMOD, Paris, France, pp 51–58
41. Naumann F, Leser U, Freytag J (1999) Quality-driven integration of heterogeneous information systems. In: Proceedings of the 25th international conference on very large data bases (VLDB), Edinburgh, Scotland, pp 447–458
42. Naumann F (2002) Quality-driven query answering for integrated information systems. LNCS 2261, Springer, Berlin Heidelberg New York
43. Pasula H, Marthi B, Milch B, Russell S, Shpitser I (2003) Identity uncertainty and citation matching. In: Proceedings of the international conference advances in neural information processing systems (NIPS), Vancouver, British Colombia, pp 1401–1408
44. Pearson RK (2002) Data mining in face of contaminated and incomplete records. In: Proceedings of SIAM international conference on data mining
45. Perner P (2002) Data mining on multimedia. LNCS 2558, Springer, Berlin Heidelberg New York
46. Piattini M, Genero M, Calero C, Polo C, Ruiz F (2000) Database quality. Chapter 14: Advanced database technology and design. Artech House, Norwood, MA, pp 485–509
47. Piattini, M, Calero C, Genero M (eds)(2002) Information and database quality. The Kluwer International Series on Advances in Database Systems, 25
48. Pyle D (1999) Data preparation for data mining. Morgan Kaufmann, San Mateo, CA
49. Rahm E, Do H (2000) Data cleaning: Problems and current approaches. IEEE Data Eng Bull 23(4):3–13
50. Raman V, Hellerstein JM (2001) Potter's wheel: An interactive data cleaning system. In: Proceedings of the 26th international conference on very large data bases (VLDB), Roma, Italy, pp 381–390
51. Redman T (2001) Data quality: The field guide. Digital Press, Elsevier
52. Rothenberg J (1996) Metadata to support data quality and longevity. In: Proceedings of the 1st IEEE metadata conference, Silver Spring, MD
53. Santis LD, Scannapieco M, Catarci T (2003) Trusting data quality in cooperative information systems. In: Proceedings of the international conference on cooperative information systems (CoopIS), Catania, Sicily, Italy, pp 354–369
54. Scannapieco M, Pernici B, Pierce E (2004) IP-UML: A methodology for quality improvement based on IP-MAP and UML. Advances in Management Information Systems-Information Quality Monograph (AMIS-IQ), Sharpe
55. Schafer JL (1997) Analysis of incomplete multivariate data. Chapman & Hall, London
56. Schlimmer J (1991) Learning determinations and checking databases. In: Proceedings of AAAI workshop on knowledge discovery in databases, AAAI–1991 Anaheim California
57. Tan P-N, Kumar V, Srivastava J (2002) Selecting the right interestingness measure for association patterns. In: Proceedings of the 8th ACM SIGKDD conference on knowledge discovery and data mining (KDD), Edmonton, Canada, pp 32–41

58. Theodoratos D, Bouzeghoub M (2001) Data currency quality satisfaction in the design of a data warehouse. Special Issue on design and management of data warehouses. Int J Coop Inf Syst 10(3):299–326
59. Vassiliadis P, Bouzeghoub M, Quix C (1999) Towards quality-oriented data warehouse usage and evolution. In: Proceedings of the 11th international conference on advanced information systems engineering (CAiSE), Heidelberg, Germany, pp 164–179
60. Vassiliadis P, Simitsis A, Georgantas P, Terrovitis M (2003) A framework for the design of ETL scenarios. In: Proceedings of the 15th international conference on advanced information systems engineering (CAiSE), Klagenfurt, Austria, pp 520–535
61. Vassiliadis P (2000) Data warehouse modeling and quality issues. PhD thesis, Technical University of Athens, Greece
62. Wang R, Kon HB, Madnick SE (1993) Data quality requirements analysis and modeling. In: Proceedings of the 9th international conference on data engineering (ICDE), Vienna, Austria, pp 670–677
63. Wang R, Storey V, Firth C (1995) A framework for analysis of data quality research. IEEE Trans Knowl Data Eng (TDKE) 7(4):670–677
64. Wang R (1998) A product perspective on total data quality management. Com. ACM 41(2):58–65
65. Wang R (2002) Journey to data quality, vol 23 of Advances in database systems. Kluwer, Boston, MA, USA
66. Wang K, Zhou S, Yang Q, Yeung JMS (2005) Mining customer value: From association rules to direct marketing. J Data Min Knowl Discov
67. Weis M, Naumann F (2004) Detecting duplicate objects in XML documents. In: Proceedings of the 1st international ACM SIGMOD workshop on information quality in information systems (IQIS) in conjunction with ACM PODS/SIGMOD, Paris, France, pp 10–19
68. Winkler WE (2004) Methods for evaluating and creating data quality. Inf Syst 29(7)

Laure Berti-Équille is currently an Associate Professor at the Computer Science Department (IFSIC) of the University of Rennes (France). Her research interests at IRISA lab (CNRS-INRIA-University of Rennes, France) are multisource data quality, quality-aware data integration, data cleaning techniques, recommender system, and multimedia data mining.