Quality-Aware Association Rule Mining

Laure Berti-Équille

IRISA, Campus Universitaire de Beaulieu,
35042 Rennes, France
berti@irisa.fr

Abstract. The quality of discovered association rules is commonly evaluated by interestingness measures (commonly support and confidence) with the purpose of supplying subsidies 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 whether a so-called "interesting" rule noted LHS -> RHS is meaningful when 30 % of LHS data are not up-to-date anymore, 20% of RHS data are not accurate, and 15% of LHS data come from a data source that is well-known for its bad credibility. In this paper we propose to integrate data quality measures for effective and quality-aware association rule mining and we propose a cost-based probabilistic model for selecting legitimately interesting rules. Experiments on the challenging KDD-CUP-98 datasets show for different variations of data quality indicators the corresponding cost and quality of discovered association rules that can be legitimately (or not) selected.

1. Introduction

Quality in data mining results critically depends on the preparation and on the quality of analyzed datasets [10]. Indeed data mining processes and applications require various forms of data preparation, correction and consolidation combining complex data transformation operations and cleaning techniques [11], because the data input to the mining algorithms is assumed to conform to "nice" data distributions, containing no missing, inconsistent or incorrect values [15]. 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 [**Erreur ! Source du renvoi introuvable.**], [9]. Data quality is a multidimensional, complex and morphing concept [4]. Since a 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, knowledge management), ranging from techniques in assessing information quality [13] to building large-scale data integration systems over heterogeneous data sources with different degrees of quality and trust. 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, these data are neither rigorously chosen from the various heterogeneous sources with different degrees of quality and trust, nor carefully controlled for quality [9]. 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 [5]. We illustrate this idea with the following example in the context of association rule mining. Among traditional descriptive data mining techniques, association rule mining 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 set of attributes in the same record (respectively the same transaction). The quality of discovered association rules is commonly evaluated by interestingness measures (namely support and confidence). 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 interestingness measures have been proposed in the literature for knowledge quality evaluation with the purpose of supplying subsidies to the user in the understanding and use of the new discovered knowledge [12], [7]. 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 $LHS \rightarrow RHS$ is meaningful when 30 % of *LHS* data are not up-to-date anymore, 20% of *RHS* data are not accurate, and 15% of *LHS* data come from a data source that is well-known for its bad credibility. Our assumption is that interestingness measures are not self-sufficient for representing association rule quality. Association rule quality should also integrate the measures of the quality of data the rule is computed from with considering the probability that the deficiencies in data quality may be adequately detected. The twofold contribution of this paper is to propose a method for scoring association rule quality and a probabilistic cost model that predicts the cost of low-quality data over the quality of discovered association rules. This model is used to select so-called “legitimately interesting” rules. We evaluate our approach using the KDD-Cup-98 dataset.

The rest of the paper is organized as follows. Section 2 gives a brief overview on data quality characterization and management. In Section 3, we present our decision model for estimating the cost of low-quality data on association rule mining. In Section 4, we evaluate our approach using the KDD-Cup-98 dataset. 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 can not be limited to one-shot approaches addressing simpler abstract versions the real problems of dirty or low-quality data [4]. Solving them requires highly domain- and context-dependent information and also 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 truthfully reflect the real world situation). This traditional approach of data quality management has lead to techniques such as integrity constraints, concurrency control and schema integration for distributed and heterogeneous information systems. But since a 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** and **classifications** with various definitions depending on authors and application contexts [1], [13], on the audience type or on the architecture of systems (e.g. for data warehouses [6])
- **Data quality metrics** [4],
- **Conceptual data quality models** [6], [1],
- **Frameworks** and **methodologies** for cleaning data [11], for improving or assessing data quality in databases [6] or using data mining techniques to detect anomalies [3], [5], [10], [8].

The most frequently mentioned data quality dimensions in the literature are accuracy, completeness, timeliness and consistency [1].

3. Probabilistic Cost Model for Quality-driven Selection of Interesting Association Rules

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 model that introduces data quality indicators and combines them for determining the quality of association rules.

Preliminary Definitions for Association Rule Quality

Let I be a superset of items. An association rule R is an implication of the form: $LHS \rightarrow RHS$ where $LHS \subseteq I$, $RHS \subseteq I$ and $LHS \cap RHS = \emptyset$. LHS and RHS are conjunctions of variables such as the extension of LHS is: $g(LHS)= x_1 \wedge x_2 \wedge \ldots \wedge x_n$ and the extension of Y is $g(RHS)= \mathrm{y}_1 \wedge y_2 \wedge \ldots \wedge y_{n'}$.
Let j ($j=1, 2,\ldots, k$) be the dimensions of data quality (e.g., data completeness, freshness, accuracy, consistency, completeness, credibility, etc.). Let $q_j(I_i) \in [min_{ij} , max_{ij}]$ be a scoring value for the dataset I_i on the quality dimension j ($I_i \subseteq I$). The vector, that keeps the values of all quality dimensions for each dataset I_i (normalized in [0,1]) is called quality vector and noted $q(I_i)$. The set of all possible quality vectors is called quality space Q.
Definition 1. Association Rule Quality
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 vec-

tors of the datasets constituting the extension of the right-hand and left-hand sides of the rule. The quality of the rule R is k-dimensional vector such as:

$$Quality(R) = \begin{pmatrix} q_1(R) \\ q_2(R) \\ \vdots \\ q_k(R) \end{pmatrix} = \begin{pmatrix} q_1(LHS) \circ_1 q_1(RHS) \\ q_2(LHS) \circ_2 q_2(RHS) \\ \vdots \\ q_k(LHS) \circ_k q_k(RHS) \end{pmatrix}$$
$$= \begin{pmatrix} q_1(x_1) \circ_1 q_1(x_2) \circ_1 \cdots \circ_1 q_1(x_n) \circ_1 q_1(y_1) \circ_1 q_1(y_2) \circ_1 \cdots \circ_1 q_1(y_{n'}) \\ q_2(x_1) \circ_2 q_2(x_2) \circ_2 \cdots \circ_2 q_2(x_n) \circ_2 q_2(y_1) \circ_2 q_2(y_2) \circ_2 \cdots \circ_2 q_2(y_{n'}) \\ \vdots \\ q_k(x_1) \circ_k q_k(x_2) \circ_k \cdots \circ_k q_k(x_n) \circ_k q_k(y_1) \circ_k q_k(y_2) \circ_k \cdots \circ_k q_k(y_{n'}) \end{pmatrix} \quad \textbf{(1)}$$

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

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

with w_j the weight of the quality dimension j. We assume the weights are normalized:

$$\sum_{j=1}^{k} w_j = 1 \ \ \forall j = 1,2,\ldots k \quad \textbf{(3)}$$

Definition 2. Fusion Function per Quality Dimension

Let T be the domain of values of the quality score $\overline{q(I_i)}$ for the dataset I_i on the quality dimension j. The fusion function denoted "$\circ_j$" is commutative and associative such as $\circ_j$: $T \times T$ → 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 presents several examples of definition for the fusion function allowing the combination of quality scores per quality dimension for two datasets noted x and y over the four quality dimensions; freshness, accuracy, completeness, consistency.

Table 1. Different fusion functions for merging quality scores per dimension

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

We consider that selecting an association rule 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 of the datasets composing the left-hand and right-hand sides of the rule. Consider the item $x \in LHS \cup RHS$ of a given association rule, we use $P_{CE}(x)$ to denote the probability that the item x will be classified as "erroneous" (or "polluted" and "with low-quality"), e.g., freshness, accuracy, etc. and $P_{CC}(x)$ 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, $P_{AE}(x)$ represents the probability that the item x is “actually erroneous” (AE) but detected correct, and $P_{AC}(x)$ represents the probability that it is “actually correct” (AC) but detected erroneous (see Figure 1).

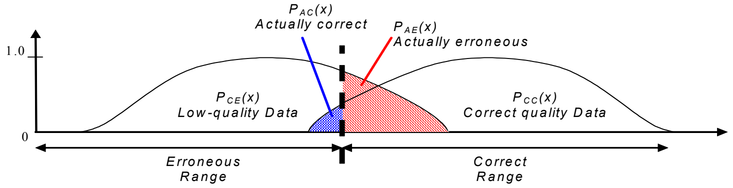

Fig. 1. Probabilities of detection of correct and low-quality data

For an arbitrary average quality vector $\overline{q} \in Q$ on the datasets in $LHS \cup RHS$ of the rule, we denote by $P(\overline{q}\in Q \mid CC)$ or $f_{CC}(\overline{q})$ the conditional probability that the average quality vector $\overline{q}$ corresponds to the datasets that are classified as correct (CC). Similarly, we denote by $P(\overline{q}\in Q \mid CE)$ or $f_{CE}(\overline{q})$ 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 $P(d=D_i \mid s=j)$ correspondingly, the joint and the conditional probability that the decision D_i is taken, when the actual status of data quality (*i.e.*, CC, CE, AE, AC) is j. 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. Based on the example presented in Table 3 where we can see how the cost of decisions could affect the result of selection among interesting association rules, we need to minimize the mean cōst c that results from making such a decision. In Table 3, c_{10} is the cost of a confident decision (D_1) for the selection of a rule based on correct-quality data (CC). c_{21} is the cost of a neutral decision (D_2) for the selection of a rule based on low-quality data (CE). c_{33} is the cost of a suspicious decision (D_3) of selecting a rule based on low-quality data but actually detected as correct (AC). The corresponding mean cost $\overline{c}$ is written as 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 $j= CC,CE,AE,AC$. Let us also assume that $\overline{q}$ is the average quality vector drawn randomly from the space of all quality vectors of the datasets of the rule. The following equality holds for the conditional probability $P(d=D_i \mid s=j)$:

$$P(d = D_i \mid s = j) = \sum_{q \in Q_i} f_j(\overline{q}). \tag{6}$$

where $i=1,2,3$ and $j=CC,CE,AE,AC$. f_j is the probability density of the quality vectors when the actual data quality status is j. We also denote the a priori probability of CC

or else *P(s=CC)* as π^0, the a priori probability of $P(s=AC)=\pi^0_{AC}$, the a priori probability of $P(s=AE)=\pi^0_{AE}$ and the a priori probability of $P(s=CE)=1-(\pi^0+\pi^0_{AE}+\pi^0_{AC})$. The mean cost $\overline{c}$ in Eq. (4) based on Eq. (5) is written as follows:

$$
\begin{aligned}
\overline{c} = {} & c_{10}.P(d=D_1|s=CC).P(s=CC)+c_{20}.P(d=D_2|s=CC).P(s=CC) \\
& +c_{30}.P(d=D_3|s=CC).P(s=CC)+c_{11}.P(d=D_1|s=CE).P(s=CE) \\
& +c_{21}.P(d=D_2|s=CE).P(s=CE)+c_{31}.P(d=D_3|s=CE).P(s=CE) \\
& +c_{12}.P(d=D_1|s=AE).P(s=AE)+c_{22}.P(d=D_2|s=AE).P(s=AE) \\
& +c_{32}.P(d=D_3|s=AE).P(s=AE)+c_{13}.P(d=D_1|s=AC).P(s=AC) \\
& +c_{23}.P(d=D_2|s=AC).P(s=AC)+c_{33}.P(d=D_3|s=AC).P(s=AC)
\end{aligned}
\qquad \textbf{(7)}
$$

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

$$
\begin{aligned}
\overline{c} = {} & \sum_{\overline{q}\in Q_1}\left[f_{CC}.c_{10}.\pi^0+f_{CE}.c_{11}.(1-(\pi^0+\pi^0_{AC}+\pi^0_{AE}))+f_{AE}.c_{12}.\pi^0_{AE}+f_{AC}.c_{13}.\pi^0_{AC}\right] \\
& +\sum_{\overline{q}\in Q_2}\left[f_{CC}.c_{20}.\pi^0+f_{CE}.c_{21}.(1-(\pi^0+\pi^0_{AC}+\pi^0_{AE}))+f_{AE}.c_{22}.\pi^0_{AE}+f_{AC}.c_{23}.\pi^0_{AC}\right] \\
& +\sum_{\overline{q}\in Q_3}\left[f_{CC}.c_{30}.\pi^0+f_{CE}.c_{31}.(1-(\pi^0+\pi^0_{AC}+\pi^0_{AE}))+f_{AE}.c_{32}.\pi^0_{AE}+f_{AC}.c_{33}.\pi^0_{AC}\right]
\end{aligned}
\qquad \textbf{(8)}
$$

For the sake of simplicity for the following of the paper, let's now consider the case of the absence of the misclassification region (*i.e.*, f_{AC}, f_{AE} are null and $\pi^0_{AE}=\pi^0_{AC}=0$). Without misclassification region *P(s=CE)* could be simplified as $1-\pi^0$. Every point $\overline{q}$ in the quality space *Q* belongs to the partitions of quality Q_1 or Q_2 or Q_3 that correspond respectively to 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. This 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. Based on this observation, a point $\overline{q}$ that represents the quality of a rule defined in Eq. (2) is assigned to one of the three optimal areas as follows:

$$
\begin{aligned}
D^0_1 &= \left\{\overline{q}:\frac{f_{CE}}{f_{CC}}\le\frac{\pi^0}{1-\pi^0}\cdot\frac{c_{30}-c_{10}}{c_{11}-c_{31}}\text{ and, }\frac{f_{CE}}{f_{CC}}\le\frac{\pi^0}{1-\pi^0}\cdot\frac{c_{20}-c_{10}}{c_{11}-c_{21}}\right\} \\
D^0_2 &= \left\{\overline{q}:\frac{f_{CE}}{f_{CC}}\ge\frac{\pi^0}{1-\pi^0}\cdot\frac{c_{20}-c_{10}}{c_{11}-c_{21}}\text{ and, }\frac{f_{CE}}{f_{CC}}\le\frac{\pi^0}{1-\pi^0}\cdot\frac{c_{30}-c_{20}}{c_{21}-c_{31}}\right\} \\
D^0_3 &= \left\{\overline{q}:\frac{f_{CE}}{f_{CC}}\ge\frac{\pi^0}{1-\pi^0}\cdot\frac{c_{30}-c_{10}}{c_{11}-c_{31}}\text{ and, }\frac{f_{CE}}{f_{CC}}\ge\frac{\pi^0}{1-\pi^0}\cdot\frac{c_{30}-c_{20}}{c_{21}-c_{31}}\right\}
\end{aligned}
\qquad \textbf{(9)}
$$

The inequalities of Eq. (9) give rise to three different threshold values *L, P* and *N* (respectively for legitimately, potentially and not interesting rules) in the decision space as defined in Eq. (10):

$$
L=\frac{\pi^0}{1-\pi^0}\cdot\frac{c_{30}-c_{10}}{c_{11}-c_{31}},\quad P=\frac{\pi^0}{1-\pi^0}\cdot\frac{c_{20}-c_{10}}{c_{11}-c_{21}},\text{ and }\quad N=\frac{\pi^0}{1-\pi^0}\cdot\frac{c_{30}-c_{20}}{c_{21}-c_{31}}
\qquad \textbf{(10)}
$$

4. Experiments and Results

In order to validate and evaluate our decision model, 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 on datasets and on discovered association rules which are assigned to the decision areas D_1, D_2 or D_3. In the set of experiments that we present, we make use the KDD-CUP-98[1] dataset from the UCI repository. 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" 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 Σ(*actual donation* - $0.68) over the validation records with predicted donation greater than the mailing cost $0.68 (see [14] for details). Because we ignored the quality of the data collected during this campaign, we generated synthetic data quality indicators with different distributions representative of 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 ten discovered "respond" rules) and we use different assumptions on data quality indicators that do not affect the top ten list of discovered association rules but that significantly change the reliability (and quality) of this mining result and also the cost of the decisions relying on these rules. The variable names, definitions, estimated probabilities and average quality score per attribute are given in Table 2. For the sake of simplicity, we suppose that the quality dimension scores are uniformly representative of the quality of the attribute value domain. The average quality per attribute in Table 2 is computed from the equi-weighted function given in Eq. (2). $f_{CC}\ (q(I_i))$ (also noted f_{CC} in Table 2) is the probability density that the dataset I_i is "correct" when the average quality score of I_i is $q(I_i)$. $f_{CE}\ (q(I_i))$ is the probability density that the dataset I_i is "erroneous" when the average quality score of I_i is $q(I_i)$. Table 3 shows tentative unit costs developed by the staff of the direct marketing department on the basis of consideration of the consequences of the decisions on selecting and using the discovered association rules. Without misclassification problem, the costs c_{12}, c_{13}, c_{22}, c_{23}, c_{32}, and c_{33} are null; the cost c_{30} of a suspicious decision for rule selection based on correct data is $500. Based on the values assigned to the various costs in Table 2, we also assume that the a priori probability that a certain quality vector belongs to *CC* equals the a priori probability that the same vector belongs to *CE*. For this reason, the ratio $\frac{\pi^0}{1-\pi^0}$ in Eq. (9) and (10) equals 1. By using Eq. (10) and Table 3, we compute the values of the three decision thresholds for rule selection for the a priori probability π^0= 0.200 without misclassification and we obtain: L=0.125, P = 0.0131579 and N = 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(L)*=-0.9031; *log(P)* = -1.8808 and *log(N)* = 0.3522. Based on the values for these thresholds, we can assign the rules to one of the three decision areas. The top 10 a priori association rules discovered by Wang *et al.* [14] are given in Table 4 with the confidence, the support (in number of records), and the

[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

profit. Table 4 also shows the score per quality dimension, the average quality and the cost of selecting the association rule. The scores are computed from the definitions of the quality dimensions given in Table 1. The costs are computed from Eq. (8). It's very interesting to notice that the predicted profit per rule may be considerably affected by the cost of the rule computed from low-quality data (e.g., the second best rule R2 whose predicted profit is $61.73 has a cost of $109.5 and thus is classified as "not interesting" due to the bad quality of its datasets). Let us now introduce different variations on the average quality of the datasets composing the rules. Based on the cost Table 3, Figure 2 shows the behavior of the decision cost of rule selection when data quality varies from the initial average quality down to -10%, -30%, and -50% and up to +10%, +30% and +50% for a priori probability π^0=0.200 and without misclassification. In Figure 2 we observe that the quality degradation of the datasets composing the rules increases the cost of these rules with variable amplitudes.

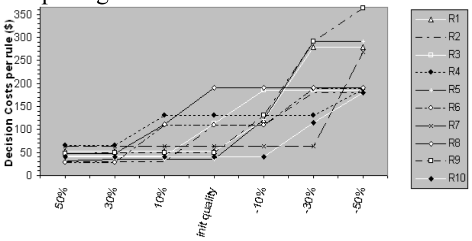

Fig. 2. Decision cost for rule selection with different data quality variations without misclassification for $\pi^0 = 0.200$

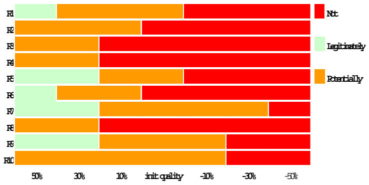

Fig. 3. Decision status on rule selection for data quality variations for π^0=0.200

Data quality amelioration implies a stabilization trend of the decision cost for legitimately interesting rule selection. Another interesting result is shown in Figure 3 where the decisions for rule selection change simultaneously with the data quality variations. Among the top 10 interesting rule discovered by Wang *et al.* [14] with the initial data quality (noted `Init Qual`), 5 rules (R1, R5, R7, R9 and R10) are potentially worth being selected based on their average data quality. Increasing data quality up to +30%, 3 rules were legitimately interesting (R5, R7 and R9). This observation offers two (among others) 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 decision costs for rule selection and secondly, for the optimal scheduling of data quality improvement activities (e.g., cleaning) driven and tuned by the rule pruning step. Additionally to the interestingness measures the three thresholds can be used as a predictive technique for quality awareness in association rule mining for the appropriate selection of legitimately interesting rules based on the data quality indicators.

Table 2. KDD-Cup-98 dataset with quality measures and estimatedprobabilities

Attribute	Definition	Quality					f_{CC}	f_{CE}
		Fresh.	Accur.	Compl.	Consi.	Average		
AGE904	Average Age of Population	0,50	0,21	0,39	0,73	0,46	0,9	0,05
CHIL2	Percent Children Age 7 - 13	0,16	0,99	0,75	0,71	0,65	0,95	0,1
DMA	DMA Code	0,49	0,58	0,16	0,95	0,55	0,95	0,01
EIC16	Percent Employed in Public Administration	0,03	0,56	0,33	0,61	0,38	0,98	0,01
EIC4	Percent Employed in Manufacturing	0,17	0,37	0,87	0,15	0,39	0,9	0,2
ETH1	Percent White	0,21	0,76	0,50	0,53	0,50	0,55	0,15
ETH13	Percent Mexican	0,52	0,77	0,87	0,79	0,74	0,9	0,6
ETHC4	Percent Black < Age 15	0,84	0,52	0,32	0,35	0,51	0,95	0,45
HC6	Percent Owner Occupied Structures Built Since 1970	0,47	0,96	0,74	0,11	0,57	0,98	0,03
HHD1	Percent Households w/ Related Children	0,61	0,95	0,27	0,08	0,48	0,96	0,41
HU3	Percent Occupied Housing Units	0,07	0,40	0,18	0,57	0,30	0,94	0,53
HUPA1	Percent Housing Units w/ 2 thru 9 Units at the Address	0,76	0,85	0,96	0,93	0,88	0,95	0,52
HVP5	Percent Home Value >= $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	0,2
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

Table 3. Costs of various decisions for classifying association rules

Decision for Rule Selection	Cost#	Data Quality Status	Cost without misclassification
D_1	c_{10}	CC	$0.00
	c_{11}	CE	$1 000.00
	c_{12}	AE	$0.00
	c_{13}	AC	$0.00
D_2	c_{20}	CC	$50.00
	c_{21}	CE	$50.00
	c_{22}	AE	$0.00
	c_{23}	AC	$0.00
D_3	c_{30}	CC	$500.00
	c_{31}	CE	$0.00
	c_{32}	AE	$0.00
	c_{33}	AC	$0.00

Table 4. The top 10 "respond" rules by Wang et al. [14] with quality, cost, and decision area

#	Association Rule	(Conf. ; Supp.)	Profit (Wang et al., 2005)	Quality					Cost	Decision Area
				Fresh.	Accur.	Compl.	Consi.	Average		
1	ETHC4=[2.5,4.5], ETH1=[22.84,29.76], HC6=[60.91,68.53]	(0.11; 13)	$81.11	0,21	0,38	0,79	0,53	0,48	$ 53	potentially
2	RFA_14=f1d, ETH1=[29.76,36.69]	(0.17; 8)	$61.73	0,21	0,52	0,62	0,53	0,47	$109.5	not
3	HHD1=[24.33,28.91], EIC4=[33.72,37.36]	(0.12;12)	$47.07	0,17	0,35	0,90	0,15	0,39	$113	not
4	RFA_23=s2g, ETH13=[27.34,31.23]	(0.12;16)	$40.82	0,34	0,01	0,90	0,79	0,51	$130	not
5	EIC16=[11.25,13.12], CHIL2=[33,35.33], HC6=[45.69,53.30]	(0.16;11)	$35.17	0,03	0,53	0,77	0,71	0,51	$ 34.7	potentially
6	RHP2=[36.72,40.45], AGE904=[42.2,44.9]	(0.16;7)	$28.71	0,50	0,15	0,44	0,73	0,46	$109	not
7	HVP5=[56.07,63.23], ETH13=[31.23,35.61], RAMNT_22=[7.90,10.36]	(0.14;10)	$24.32	0,37	0,65	0,68	0,95	0,66	$ 62.8	potentially
8	NUMCHLD=[2.5,3.25], HU3=[66.27,70.36]	(0.08;31)	$19.32	0,07	0,09	0,61	0,57	0,34	$190	not
9	RFA_11=f1g, DMA=[743,766.8], POP903=[4088.208,4391.917], WEALTH2=[6.428571,7.714286]	(0.25;8)	$17.59	0,24	0,08	0,72	0,95	0,50	$ 49.6	potentially
10	HUPA1=[41.81+,], TPE11=[27,64,31.58]	(0.23;9)	$9.46	0,20	0,22	0,99	0,93	0,59	$ 40.8	potentially

5. Conclusion

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 (or not) interesting" association rules based on correct- or low-quality data. The model defines the thresholds of three decision areas for the predicted class of the discovered rules (*i.e.*, legitimately interesting, potentially interesting, 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 ten association rules discovered by Wang *et al.* [14], *iii)* computing the cost of selecting low-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 also considered for effective and 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 biomedical datasets (see [2]) with on-line collecting and computing operational data quality indicators with the aim to select high-quality and interesting association rules.

References

1. Batini C., Catarci T. and Scannapiceco M., A Survey of Data Quality Issues in Cooperative Information Systems, *Tutorial, Intl. Conf. on Conceptual Modeling (ER)*, 2004.
2. Berti-Equille L., Moussouni F., Quality-Aware Integration and Warehousing of Genomic Data., *Proc. of the Intl. Conf. on Information Quality*, M.I.T., Cambridge, U.S.A., 2005.
3. Dasu T. and Johnson T., Hunting of the Snark: Finding Data Glitches with Data Mining Methods, *Intl. Conf. on Information Quality*, M.I.T., Cambridge, M.A., U.S.A., 1999.
4. Dasu T., Johnson T., *Exploratory Data Mining and Data Cleaning*, Wiley, 2003.
5. Hipp J., Guntzer U., and Grimmer U., Data Quality Mining - Making a Virtue of Necessity. *Proc. of the Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD2001)*, Santa Barbara, CA, U.S.A, May 20th, 2001.
6. Jeusfeld, M. A., Quix C., Jarke M., Design and Analysis of Quality Information for Data Warehouses, *17th Intl. Conf. on Conceptual Modeling (ER'98)*, Singapore, 1998.
7. Lavrač N., Flach P.A., Zupan B., Rule Evaluation Measures: A Unifying View, *ILP*, p. 174-185, 1999.
8. Lübbers D., Grimmer U. and Jarke M., Systematic Development of Data Mining-Based Data Quality Tools, *Proc. of the Intl. VLDB Conf.*, p. 548-559, 2003.
9. Pearson R.K., Data Mining in Face of Contaminated and Incomplete Records, *Proc. of SIAM Intl. Conf. Data Mining*, 2002.
10. Pyle D., Data Preparation for Data Mining, Morgan Kaufmann, 1999.
11. Rahm E., Do H., Data Cleaning: Problems and Current Approaches, *IEEE Data Eng. Bull.* 23(4): 3-13, 2000.
12. Tan P-N., Kumar V. and Srivastava J., Selecting the Right Interestingness Measure for Association Patterns, *Proc. of Intl. KDD Conf.*, p. 32-41, 2002.
13. Wang R., Storey V., Firth C., A Framework for Analysis of Data Quality Research, *IEEE TKDE*, 7(4): 670-677, 1995.
14. Wang K., Zhou S., Yang Q. and, Yeung J.M.S., Mining Customer Value: from Association Rules to Direct Marketing, *J. of Data Mining and Knowledge Discovery*, 2005.
15. Zhang C., Yang Q. and Liu B. (Eds). Introduction: Special Section on Intelligent Data Preparation, *IEEE Transactions on Knowledge and Data Engineering,* 17(9), 2005.