The Effects and Interactions of Data Quality and Problem Complexity on Classification

ROGER BLAKE, University of Massachusetts, Boston
PAUL MANGIAMELI, University of Rhode Island

Data quality remains a persistent problem in practice and a challenge for research. In this study we focus on the four dimensions of data quality noted as the most important to information consumers, namely accuracy, completeness, consistency, and timeliness. These dimensions are of particular concern for operational systems, and most importantly for data warehouses, which are often used as the primary data source for analyses such as classification, a general type of data mining. However, the definitions and conceptual models of these dimensions have not been collectively considered with respect to data mining in general or classification in particular. Nor have they been considered for problem complexity. Conversely, these four dimensions of data quality have only been indirectly addressed by data mining research. Using definitions and constructs of data quality dimensions, our research evaluates the effects of both data quality and problem complexity on generated data and tests the results in a real-world case. Six different classification outcomes selected from the spectrum of classification algorithms show that data quality and problem complexity have significant main and interaction effects. From the findings of significant effects, the economics of higher data quality are evaluated for a frequent application of classification and illustrated by the real-world case.

Categories and Subject Descriptors: H.1.1 [**Models and Principles**]: Systems and Information Theory—*Value of information*; H.2.8 [**Database Management**]: Database Applications—*Data mining*; H.4.2 [**Information System Applications**]: Types of Systems—*Decision support (e.g. MIS)*

General Terms: Measurement, Algorithms

Additional Key Words and Phrases: Data quality, information quality, data quality metrics and measurements, data mining

ACM Reference Format:
Blake, R. and Mangiameli, P. 2011. The effects and interactions of data quality and problem complexity on classification. ACM J. Data Inform. Quality 2, 2, Article 8 (February 2011), 28 pages.
DOI = 10.1145/1891879.1891881 http://doi.acm.org/10.1145/1891879.1891881

1. INTRODUCTION

The purpose of this research is to evaluate the impact of data quality on the outcomes of classification, and to evaluate how these impacts can be amplified when data quality is considered in combination with the structure of classification problems, as represented by problem complexity. To do so, metrics for data quality and for problem complexity are defined based on prior work and applied to analyze the effects and interactions of these two factors on the outcomes of classification, first for a series of generated datasets, and then for a real-world application. The results are used to

Authors' addresses: R. Blake, University of Massachusetts, Boston, MA; email: roger.blake@umb.edu; P. Mangiameli, University of Rhode Island, Kingston, RI 02881.

DOI 10.1145/1891879.1891881 http://doi.acm.org/10.1145/1891879.1891881

demonstrate the economic benefits that higher data quality can bring in that real-world case.

The purpose of classification is to make predictions of a categorical variable, known as a class variable, from one or more attribute variables which can be either categorical or numeric. Classification is a type of supervised machine learning in which a model to predict the class variable is built using a training dataset. Because measuring the outcomes of classification with the training dataset is known to have a positive bias, the model is usually evaluated using a separate test dataset.

Classification problems are common; typical examples include using classification for predicting which customers will respond to direct marketing promotions and which credit applications are fraudulent. Another example is the real-world case we use to illustrate the results of this study, which is a retention analysis used to predict whether customers will stay or leave the company.

Classification is one of the most important areas of data mining, and is especially significant for decision-making [Ali and Smith 2006]. Data mining research, including of classification, has increased in importance as advances in the field have enabled tighter couplings between business process systems and decision-making [Kohavi et al. 2002]. It is leading to the prospect of replacing expert analysis with fully automated decisions [Apte et al. 2002]. Data mining research has become an integral part of business intelligence as analytical capabilities and "fact-based decision making" have become a hallmark of leading companies [Davenport and Harris 2007].

While there are no universally agreed-upon definitions of data quality, there is no dispute about the importance it has and that the consequences when it is poor can be large. Inadequate data quality has had major financial consequences [Redman 2004] and led to incalculable losses [Su and Jin 2007]. In corporate databases alone the cost of poor data quality is estimated to be over $600 billion annually [Eckerson 2002]. The importance of data quality and its research is widely recognized.

Data quality is often conceptualized in terms of dimensions; our study is of four of these dimensions: accuracy, completeness, consistency, and timeliness. These particular four were chosen because "accuracy, completeness, consistency, and timeliness have been widely cited in the literature as the most important data quality dimensions to the information consumers" [Parssian 2006]. We develop a set of assessment metrics for each dimension of data quality and evaluate how varying the levels of each of these dimensions affects the outcomes of classification.

In data mining research, data quality is loosely encompassed by the general term "noise." Noise is broadly defined as anything obscuring the relation between class and attribute variables [Hickey 1996] or, alternatively, as any systematic error [Quinlan 1986]. Noise is considered to be present whenever data instances are not well-aligned with benchmark theory [Zhu and Wu 2004], and in data having different semantics, representations, perturbed values, misclassifications, or formats. Noise also encompasses data with contradictory examples; these include cases in which instances exist with the same attribute value(s) but with different values of the class variable [Zhu and Wu 2004]. To the extent one of these examples of noise may correspond to one of the dimensions of data quality, data mining research generally does not consider any linkage between the two.

Conversely, the dimensions of data quality do not generally have direct analogs to particular types of noise. For example, poor quality along Wang and Strong [1996] dimensions of objectivity, interpretability, accuracy, and relevancy could singly or together equate to noise from data with varying semantics. These dimensions could equally well relate to noise originating from data which deviates from a benchmark theory, or from noise in the form of systematic errors. Data with poor quality in the dimensions of representational consistency, concise representation, accuracy,

Table I. The PSP/IQ Model from Lee et al. [2002]

	Conforms to specifications		**Meets or exceeds consumer expectations**	
Product quality	• Free-of-error • Concise representation	• Completeness • Consistent representation	• Appropriate amount • Relevancy • Understandability	• Interpretability • Objectivity
Service quality	• Timeliness	• Completeness	• Believability • Accessibility	• Ease of operation • Reputation

or relevancy could be the result of misclassifications. The correspondence between dimensions of data quality and the types of noise considered in data mining research is not direct.

Nor do the dimensions of data quality have clear relationship to complexity. Complexity is a broad concept taking multiple forms. In this research we develop a straightforward definition and measure of complexity that relates to the structure of a classification problem. Doing so enables the results of this study to be more readily implemented in practice. We distinguish the structural complexity of a classification problem from what is known elsewhere as structural complexity, the later being related to patterns found in data. To differentiate between the two we use the term problem complexity for ours. As with noise, problem complexity and data quality dimensions have no clear analogs. A set of classification problems with varying degrees of accuracy, completeness, consistency, or timeliness may or may not correspond to classification problems with different structures or higher complexity.

The ambiguous links between data quality dimensions, noise, and problem complexity is characteristic of the differences between the two research areas of data quality and data mining; they are relatively distinct streams [Sessions and Valtorta 2006]. Our research brings together concepts from each to evaluate whether main and interaction effects exist between data quality and problem complexity on classification outcomes. Based on finding of significant effects, this study uses a real-world application to show how the economics of improving data quality can be quantified from the impact of higher data quality on classification outcomes.

In the sections that follow we provide the relevant literature used to develop our metrics, our hypotheses for effects and interactions, and then the experimental design and analysis used for generated datasets and real-world data. After presenting the analysis, the economic context of data quality in classification is explored.

2. BACKGROUND

Among the frameworks of data quality dimensions that have been put forth, perhaps the most widely accepted is Wang and Strong's framework of 15 dimensions [Wang and Strong 1996]. Lee et al.'s Product Service Performance Information Quality (PSP/IQ) model categorized Wang and Strong's dimensions for the purpose of developing benchmarks for a gap analysis of data quality [Lee et al. 2002]. Table I shows how the PSP/IQ model divides data quality by how quality is defined. On one side, quality relates to how well data meets the need of consumers, or "fitness for use" [Lee et al. 2004]; on the other quality relates to how well data meets specifications.

All four of the dimensions in this study fall within the latter category in this model; these are dimensions which are more objective because they are less based upon user expectations and can be measured by standards applied directly to a database [Kahn et al. 2002]. For three of the dimensions in our study, the PSP/IQ defines quality in terms of a product (accuracy equated to "free-of-error"); for timeliness it is defined as the quality of a service. The metrics and measurements we develop for timeliness will reflect this distinction.

The data quality metrics we develop are described in the next section, with the metrics for problem complexity and then for the outcomes of classification in following sections.

2.1 Metrics for Data Quality Assessment

A value expressed in the range of 0 to 1 has long been held as a desirable trait for data quality metrics [Dillard 1992]; data quality dimensions are normally measured using metrics defined to have values between 0 and 1 [Pipino et al. 2002]. Metrics in this form have frequently been used, especially for the intrinsic dimensions of data quality in this study. Examples include metrics for completeness and consistency developed by Ballou and Pazer [2003] and the metrics for completeness by Shankaranarayanan and Cai [2006]. We adopt the same convention as these and other authors have, and use a range of 0 to 1 to represent lowest to highest data quality.

To assess data quality we use metrics in two of the functional forms described by Pipino et al. [2002] and Lee et al. [2006]. These authors suggested simple ratios for the relatively objective dimensions. Simple ratios are based on percentages of data items meeting specific criteria, such as the percentages of data items which are complete. We follow suit by assessing data quality with simple ratios for the dimensions where applicable: accuracy, completeness, and consistency. The fourth dimension of timeliness uses a min/max metric as described later.

2.1.1 Accuracy. Of the four dimensions we study, accuracy has a particularly wide range of definitions. Many consider accuracy as meaning a correct and unambiguous correspondence with the real world. One example defines accuracy to mean "the recorded value is in conformity with the actual value" [Ballou and Pazer 1982]. Another defines accuracy as "agreement with either an attribute of a real-world entity, a value stored in another database, or the results of an arithmetic computation" [Klein et al. 1997]. Accuracy has also been defined to encompass groups of dimensions such as completeness, consistency, or timeliness. A firm definition is elusive; as Wand and Wang [1996] summarized, "there is no commonly accepted definition of what it means exactly."

When defining metrics for accuracy researchers have often represented a correct value with 1 and an incorrect value as 0, as used for the accuracy of zip codes in a customer table by Wang et al. [2000]. Effective data quality metrics can be defined at the data item, attribute, record, or database level as shown by Shankaranarayanan and Cai [2006]. Here we define the metrics for all four dimensions at the record, or tuple, level and so for accuracy the metric A is defined as $1 - (V_T/N_A)$, where $V_T$is the number of tuples in a relation having one or more incorrect values and N_A the total number of tuples.

2.1.2 Completeness. Complete data has been defined as data having all values recorded [Gomes et al. 2007]. Whether an incomplete value represents an unknown or missing value in the real world, or it represents a value yet to be entered into a database, a value of null is used to represent an incomplete data item. Our metric for completeness, K, is the ratio of the number of tuples with null values to the total number of tuples in a relation and is defined as $1 - (M_T/N_K)$, where $M_T$is the number of tuples in a relation having a null value and $N_K$the total number of tuples.

2.1.3 Consistency. Definitions of consistency often refer to uniformity. Ballou and Pazer [2003] defined consistency as meaning "the representation of the data value is the same in all cases" and as "format and definitional uniformity within and across all comparable datasets." [Gomes et al. 2007] defined data as consistent "if it doesn't convey heterogeneity, neither in contents nor in form," and uniformity was again expressed in a definition of consistency as being related to ambiguity and the "same value repeatedly expressed for the same situation in the real world" [Wand and Wang 1996].

Consistency has been defined with respect to referential integrity [Lee et al. 2006]. Similarly, Karr et al. [2006] tied consistency to the degree to which a database is designed with proper primary and secondary keys and normalized. Both defined a metric for consistency using the ratio of tuples with violations of consistency to the total number of tuples. We do the same and define our metric for consistency C as $1-(R_T/N_C)$, where $N_C$is the number of tuples in the relation and R_T the number of tuples with violations of referential integrity.

Inconsistency in our datasets was created by switching an attribute value to the value of another attribute selected at random from the instances in a different class. This switch creates an overlap between class boundaries and an ambiguity whereby instances with the same attribute values can belong to two (or more) classes. A higher rate of switched attribute values represents higher ambiguity, lower uniformity, and a decreased level of consistency.

Ordonez and García-García [2008] developed and explored a referential integrity metric using an example of two relations, R and S. Relation S had a primary key S_k and an attribute S_f; R was a denormalized relation with primary key R_k and a foreign key R_f functionally dependent on S_f. The authors' metric was based on the number of tuples in R with nonnull values of R_k in which $R_k = S_k$ and $R_f = S_f$ taken as the percentage of the number of tuples in R with nonnull values of R_k in which $R_k = S_k$. In general terms this metric is based on the number of tuples with matches between R and S in both k and f, divided by the number of tuples with matches between R and S in only k. In more formal terms this metric can be expressed as $\Im_{Count(R.k)}$ $(\sigma_{R.k\neq null}$ $(R\bowtie_{R.k=S.k,R.f=S.f}$ $S))$ / $\Im_{Count(R.k)}$ $(\sigma_{R.k\neq null}(R\bowtie_{R.k=S.k}$ $S))$.

The datasets in our study are equivalent to the relation R with class variable k and attribute variable f. In the case of perfect consistency there is a relation S for which there are no mismatches, $\Im_{Count(R.k)}$ $(\sigma_{R.k\neq null}$ $(R\bowtie_{R.k=S.k}$ $S)) = \Im_{Count(R.k)}$. Changing a value R_f to correspond to another class in R_k by definition means that there will be a tuple in R which will not match on both R_k and R_f. For example, if the original value of a tuple's attribute was R_f = p and it was changed to R_f = q, then the denominator $\Im_{Count(R.k)}$ $(\sigma_{R.k\neq null}(R\bowtie_{R.k=S.k}$ $S))$ would be unaffected. However, this change would create an unmatched pair of k and f values in R and S, and the numerator of the metric would decrease by 1. This mismatch means that $\Im_{Count(R.k)}$ $(\sigma_{R.k\neq null},\ _{R.k=S.k},_{R.f=q,Sf=p}(R$ $\bowtie$ $S)) = 0$ and the consistency metric decreases as ambiguity is added using this method. The metric we use is the same as Ordonez and García-García's metric calculated from the conditions described before.

2.1.4 Timeliness. In general, researchers have defined timeliness using three events as points of reference. The first occurs with a change in the real world, the second when that change is recorded as data in an information system, and the third on the use of that data. Figure 1 displays a timeline to show how these three events relate.

Definitions of timeliness often refer to volatility and currency. Volatility is the length of time between real-world change and subsequent change which invalidates the original data [Pipino et al. 2002; Wand and Wang 1996]. Terms and definitions for currency vary; some consider currency as a separate dimension [Heinrich et al. 2009]. Here we define currency as a component of timeliness which corresponds to Ballou et al. [1998] definition.

In Figure 1 it can be seen that currency and volatility are independent and can either precede or succeed the other. Because a simple ratio does not suffice for a metric based on two or more component factors Pipino et al. [2002] and Lee et al. [2006]

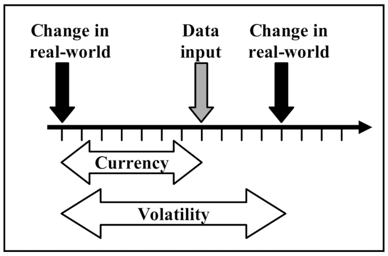


Fig. 1. Events used to define currency and volatility for timeliness.

recommend a metric in the min/max form which for timelines would be *Max(0, 1 – (Currency / Volatility)*. This metric is the same as proposed by Ballou et al. [1998] and closely corresponds to a metric for timeliness proposed by Su and Jin [2007]. However, timeliness is unlike the other dimensions in our study. In the PSP/IQ model [Kahn et al. 2002], accuracy, completeness, and consistency were categorized as dimensions where quality could be more readily quantified and measured from the data. Timeliness, on the other hand, was categorized as a dimension for which quality could not be determined directly from data but rather after data had been delivered.

Therefore, there is no direct method of manipulating data to create varying levels of timeliness. To represent different levels of timeliness we first derived a classification model and then replaced varying percentages of new instances in the learning data for reclassification by that model. We took currency to be fixed (the data in a dataset has already been inputted), but allowed volatility to vary as differing percentages of new instances were used to replace older and potentially stale data. From these assumptions, our metric for timeliness T is defined as $(N_T - R_T)/N_T$), where N_T is the number of instances, or tuples, in the training and test data and R_T the number of replacement instances introduced for reclassification.

2.2 Metric for Problem Complexity

Studies needing a measure of problem complexity have often used entropy. One example is from the model of decision strategy developed by Swait and Adamowicz [2001]. To quantify the complexity that consumers face when choosing from a number of alternatives, entropy was used. Defining higher complexity as a higher number of discrete choices is directly analogous to our definition of higher complexity as a higher number of categories in a class variable. Entropy was also used as a measure of complexity in a model of supplier-customer systems [Wu et al. 2007]. In this study, structural complexity was defined as the amount of information needed to describe the state of a planned system, and operational complexity was defined as the amount of information needed to describe the difference between a planned system and actual system. Both forms of complexity were defined using the amount of information in a discrete number of states and both measured complexity with entropy, as we do. We follow these precedents and use entropy in our metric for problem complexity.

Although neither of the two studies mentioned earlier used them, there are forms of entropy such as conditional or mutual entropy that measure complexity using two or more variables. In our study the effects of problem complexity need to be evaluated independently from the effects of data quality. The measure of problem complexity in this study needs to be based on a single variable only and cannot encompass both class and attribute variables. Therefore we represent problem complexity by the entropy

Table II. Elements of a Confusion Matrix

	Actual positive	**Actual positive**
Predicted positive	*TP*	*TP*
Predicted negative	*FN*	*TN*

of the class variable alone, and represent data quality in the attribute variable alone. This is done specifically to avoid a potential covariance.

Entropy is a measure of uncertainty contained in data and is defined using the probabilities of membership in each of the categories a variable can assume, which for classification problems is the class variable, by

$$-\sum_{i=1}^{n} p(x_1)\log(p(x_1),$$

where the n categories of the categorical variable represented by x, and $p(x_i)$ is the probability of any instance belonging to category x_i, i assuming any value from $1 \ldots n$.

Entropy increases with the number of categories, therefore using entropy ties problem complexity to the structure of a classification problem. In turn, this enables the results of this study to be more readily applied to a given application in practice. We demonstrate this in Section 6.

2.3 Metric for Classification Outcomes

Among the popular alternative metrics to measure the outcomes of classification are classification-accuracy, the Area Under the Curve (AUC), and the F-measure. For statistical analysis a metric in the form of a single figure is highly desirable, true for all three. Classification-accuracy is the percentage of correct classifications. However, this metric can be highly affected by class imbalance in which a large percentage of the learning data belongs to a small number of classes. The F-measure is not affected by this drawback. AUC is a measure of the area under a particular region of the Receiver Operating Characteristic (ROC) curve; the ROC curve is based on a plot of the rates of true and false positives. The AUC is a single figure metric, but is appropriately applied in scenarios in which the costs of correct and incorrect classifications are known and fixed. The F-measure does not have these restrictions, and being more general, is the suitable metric for this study.

The F-measure is calculated using recall and precision. Both are from the confusion matrix outputted from a classification model used on a test dataset. Recall is the percentage of instances that are classified as positive out of the total number of positive instances in the data; precision is the percentage of instances that are classified as positive that are correct and are actually positive. The elements of a confusion matrix are shown in Table II followed by the more formal definitions of recall, precision, and the F-measure. In Table II

where:
TP = Percentage of True Positives,
FP = Percentage of False Positives,
TN = Percentage of True Negatives,
FN = Percentage of False Negatives, and

$TP + FP + TN + FN = 100\%$.

Recall and precision are defined as

$$\mathit{Recall} = TP/(TP + FN) \text{ and}$$

$$\mathit{Precision} = TP/(TP + FP),$$

where *TP, FP, TN, FN* are defined as earlier.

The F-measure is calculated as the harmonic mean of precision and recall.

$$F - \mathit{measure} = (2 * \mathit{Recall} * \mathit{Precision})/(\mathit{Recall} + \mathit{Precision})$$

2.3.1 Classification Algorithms. This study involves classification algorithms in general, without the purpose of comparing specific algorithms or improving their performance. Our aim is to avoid results tied to artifacts of specific algorithms; therefore, we elected to analyze the results from a broad range of classification algorithms. Classification algorithms can be divided into three broad categories: neural-based, rules-based, and statistically-based. To cover the spectrum we used algorithms from each category, and purposefully chose algorithms with widely different characteristics from each of these categories.

The six algorithms used in this study are the MultiLayer Perceptron (MLP) algorithm (a neural-based learner), J48 (a decision tree algorithm), Sequential Minimal Optimization (SMO) which is a type of Support Vector Machine (SVM), IBk (a nearest-neighbor algorithm similar to *k*-means clustering), and two statistically-based algorithms, Bayesian networks and logistic regression. These six algorithms were used as implemented in Weka, a popular open-source data mining package.

3. HYPOTHESES

Each of our hypotheses posits significant effects on the outcomes of classification from two factors: data quality and problem complexity. In our experimental design, a continuous attribute variable is used for the representation of varying levels of data quality, and a categorical class variable for varying levels of problem complexity, again to avoid a potential confound.

There are four major "umbrella" hypotheses, each consisting of a group of analogous individual hypotheses. Each individual hypothesis involves a single dimension of data quality and the effects of either data quality, problem complexity, or both, on classification outcomes. The null hypothesis for each is that there will be no significant effects; each hypothesis is stated in the positive form as an alternate hypothesis that there will be significant effects. None of the "umbrella" hypotheses is intended to be collective; rather, analogous individual hypotheses have been grouped together to avoid a lengthy list of highly similar hypotheses.

> *Hypothesis I: Taken individually, each level of data quality or level of problem complexity will have a significant effect on the outcomes of classification as indicated by the F-measure.*

"Umbrella" Hypothesis I is comprised of five individual hypotheses, one each for accuracy, consistency, completeness, timeliness, and problem complexity as diagrammed in Figure 2. It is a straightforward hypothesis that subsequent hypotheses build upon.

> *Hypothesis II: For any given level of problem complexity, as the level of data quality in any given dimension decreases there will be a corresponding negative and significant effect on classification outcomes as indicated by the F-measure.*

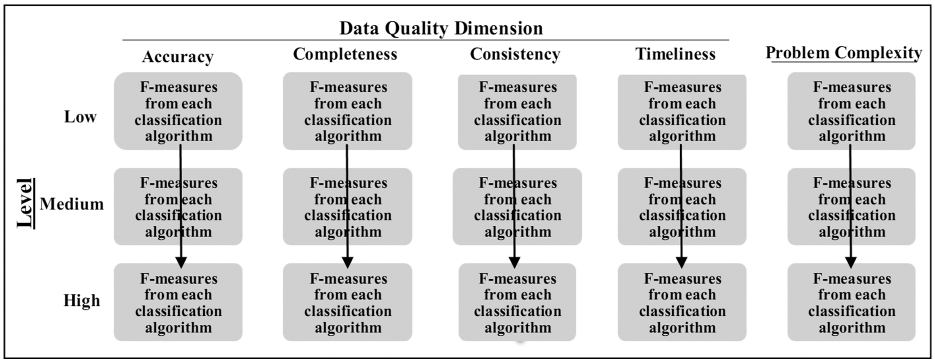


Fig. 2. Hypothesized significant effects in Hypothesis I.

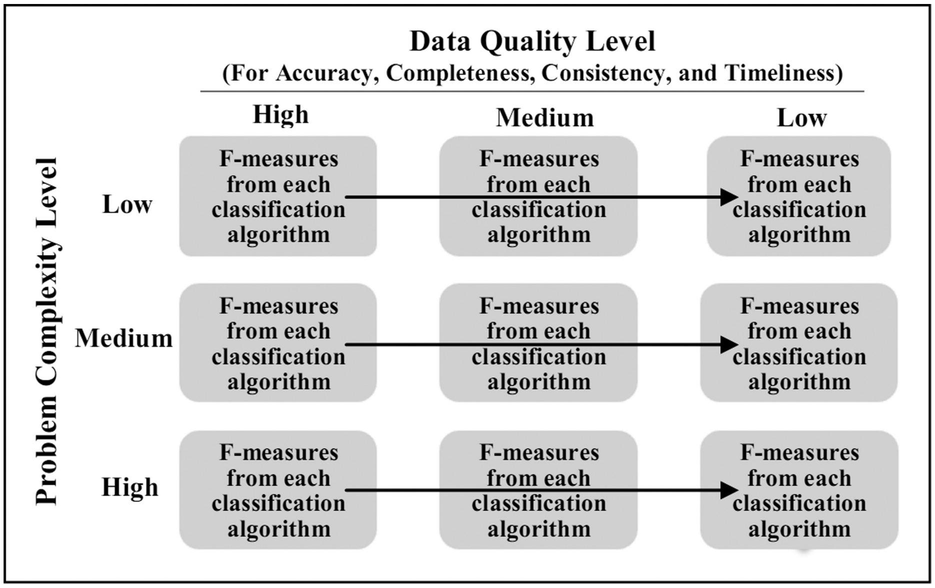
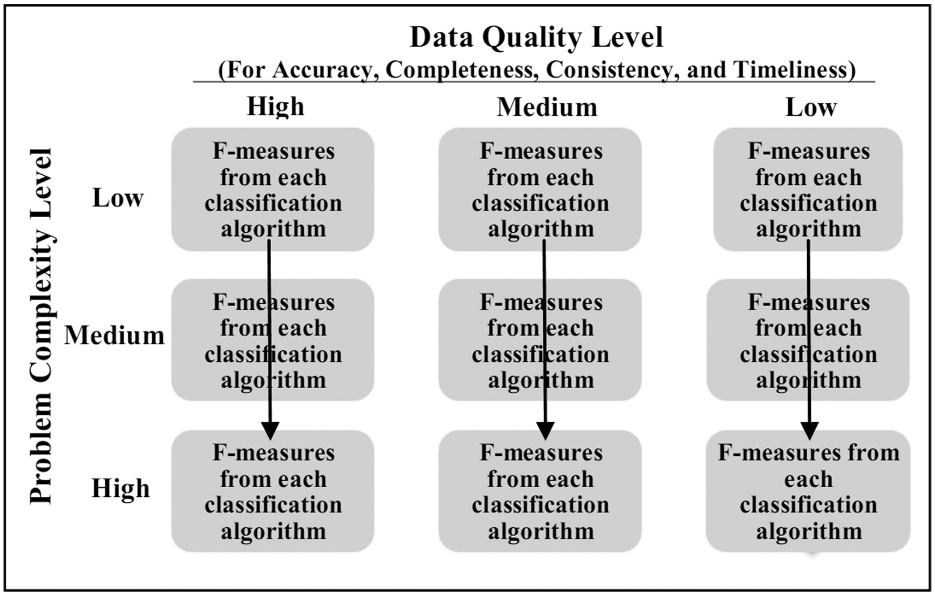

Data Quality Level
(For Accuracy, Completeness, Consistency, and Timeliness)
High
Medium
Low
Problem Complexity Level
Low
Medium
High
F-measures from each classification algorithm

Fig. 3. Hypothesized significant effects and their direction in Hypothesis II.

Hypothesis II is concerned with the impact of data quality on classification outcomes. Hypothesis II posits that significant and negative effects of decreasing data quality on F-measures will be found while holding the level of problem complexity constant. With four dimensions and three levels of problem complexity, “umbrella” Hypothesis II consists of 12 individual hypotheses. Figure 3 diagrams the hypothesized effects and direction of decreasing F-measures expected for each dimension.

> *Hypothesis III: For any given level of a stated data quality dimension, as the level of problem complexity increases there will be a corresponding negative and significant effect on classification outcomes as indicated by the F-measure.*

Hypothesis III is concerned with the effects of problem complexity on classification outcomes. Hypothesis III posits that significant and negative effects of higher problem complexity will be found while holding the level of data quality constant. With



Problem Complexity Level	Data Quality Level (For Accuracy, Completeness, Consistency, and Timeliness): High	Medium	Low
Low	F-measures from each classification algorithm	F-measures from each classification algorithm	F-measures from each classification algorithm
Medium	F-measures from each classification algorithm	F-measures from each classification algorithm	F-measures from each classification algorithm
High	F-measures from each classification algorithm	F-measures from each classification algorithm	F-measures from each classification algorithm

Fig. 4. Significant effects and their direction in Hypothesis III.

three levels of data quality and four dimensions there are 12 individual hypotheses in “umbrella” Hypothesis III. Figure 4 illustrates the hypothesized effects and the direction of decreasing F-measures expected for each dimension of data quality.

> *Hypothesis IV: Taken together, as the level of a given data quality dimension decreases and the level of problem complexity increases, their interactions will have a corresponding significant and negative effect on classification outcomes as indicated by the F-measure.*

“Umbrella” Hypothesis IV has one individual hypothesis for each dimension of data quality, totaling four analogous hypotheses. Each is based on a statement that the effects of either decreasing data quality or increasing problem complexity will have significant and negative effects. Hypothesis IV further states there will be interaction effects as data quality decreases and problem complexity increases together, and that these interaction effects will be both significant and negative. These hypothesized effects applicable to each dimension of data quality are shown in Figure 5.

The main effects hypothesized in Hypotheses I through III are intended to build towards the evaluation of the main and interaction effects posited in Hypothesis IV. Presented next is the experimental design used for testing Hypotheses I through IV, first with a description of how data quality and problem complexity are used for the factors, then for the methods used to generate datasets.

4. EXPERIMENTAL DESIGN

Please note that in the experimental design we employ the common research standard of low, medium, and high levels for the four dimensions of data quality and for problem complexity.

4.1 Data Quality Measurements

The metric for accuracy is defined from the rate (percentage) of the instances with correct attribute values. In our datasets, we introduce inaccuracy changing an attribute

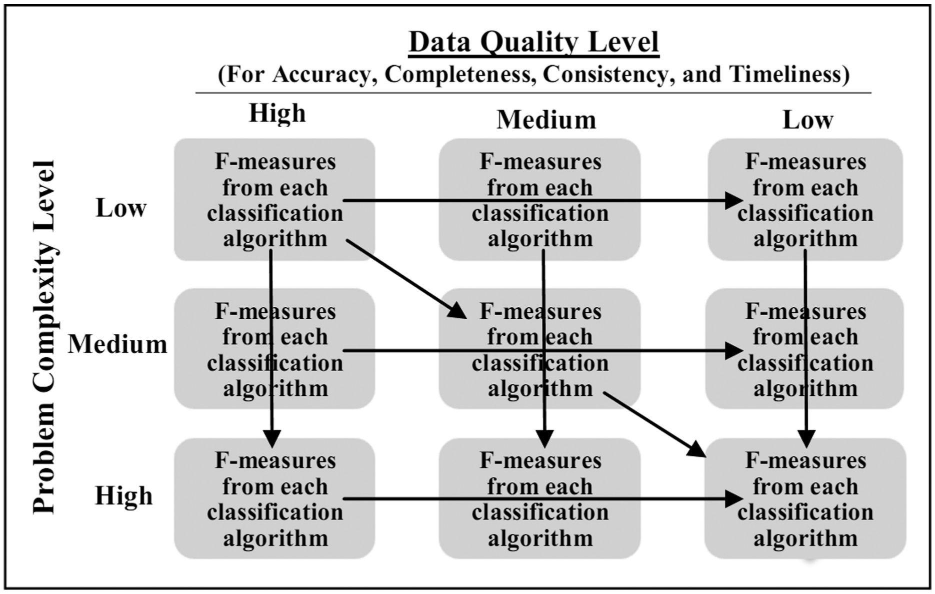

Data Quality Level
(For Accuracy, Completeness, Consistency, and Timeliness)
High
Medium
Low
Problem Complexity Level
Low
Medium
High
F-measures from each classification algorithm

Fig. 5. Significant effects in Hypothesis IV and their direction.

value to that of another instance selected at random. Prior research provides several guidelines for appropriate rates of incorrect values. Parssian et al. [2004] discussed accuracy levels from three sources ranging from 80% to 90%. Fisher et al. [2007] explored the effects of random on accuracy metrics by simulating accuracy rates in the range of 90%. Klein et al. [1997] found that even the most critical databases have estimated inaccuracy rates of 1% to 10%. Given that they are critical, presumably the quality would be as high as possible. As the research suggests accuracy levels between 80% and 100%, we chose a high level of accuracy as having a range of 92–100% correct values (mean of 96%), a medium level as 88–92% (mean of 90%), and a low level as 80–88% (mean of 84%).

The metric for completeness is derived from the rate of null instances in an attribute variable. Starting with base datasets generated as described later, different rates of null values were inserted into the attribute variable to represent different levels of completeness. Guidelines from past research indicate a wide range of null value rates have been used to represent completeness. Parssian et al. [2004] discuss examples of completeness rates ranging from 75% to 95%. Others have experimented with 50% null attribute values [Ge and Helfert 2006], and some have reported databases with missing value rates of 50% and more [Lakshminarayan et al. 1999]. Based on the literature we use the same mean rates and ranges of completeness as for accuracy.

The consistency metric by [Ordonez and García-García 2008] was well developed but the authors did not explore particular values to represent levels of data quality. Referential integrity metrics developed elsewhere also provide little guidance. We therefore chose the same mean and range of values for consistency that we used for accuracy and completeness.

Prior work does not suggest values for the low, medium, or high levels of timeliness. Varying levels of timeliness were represented by replacing varying percentages of instances in the learning data with new data in our test datasets. We ran our data-mining algorithms with 10-fold cross-validation to avoid spurious effects. Based on preliminary investigations with similar datasets, we established a high level of

Table III. Descriptive Statistics for Entropy as a Function of Class Variable Categories

Number of categories	Entropy			
	Mean	Std Dev	Min	Max
2	0.253	0.071	0.004	0.301
3	0.405	0.066	0.139	0.476
4	0.520	0.058	0.378	0.599
5	0.614	0.057	0.330	0.693
6	0.698	0.044	0.571	0.774
7	0.774	0.046	0.623	0.833
8	0.824	0.042	0.689	0.896

timeliness by retaining an average of 82% of the instances, a medium level by retaining 50%, and a low level by 18%.

4.2 Problem Complexity Measurement

In classification studies paralleling ours, entropy is used to define complexity. Entropy increases with the number of categories comprising a class variable. At one extreme is a class variable with one category, and therefore no complexity; at the other extreme is a class variable with many categories and high complexity. To develop measurement levels, Monte-Carlo simulations were employed to calculate the entropies associated with a varying number of class variable categories. Table III shows the descriptive statistics for entropy from those simulations.

A *k*-means cluster analysis was conducted to determine the association between mean entropies and the three problem complexity levels. These results showed that low complexity was best represented by a class variable having two categories, medium complexity by a class variable having three to four categories, and high complexity as a class variable having five through eight categories.

4.3 Dataset Generation

The data mining classification problems we analyzed each had a dataset with a categorical class variable and one continuous attribute variable. Data with this structure corresponds to a relation extracted from a data warehouse, such as those having one identifier and one nonidentifier in the schema used for data quality analysis by Parssian et al. [2004]. This structure also equates to the denormalized relations Ordonez and García-García [2008] used for their data quality metrics.

Each dataset was generated with 300 instances. This number of instances is an approximate performance threshold beyond which improvements in the performance of classification algorithms may not be appreciable [Oates and Jensen 1997]. A base dataset with perfect data quality was first created for each dimension and level of problem complexity. Varying levels of data quality were next induced to create additional datasets having lower levels of data quality.

Figure 6 is a graphical representation of the 100 datasets that were generated for the three data quality and three problem complexity levels for each dimension of data quality. The 900 generated datasets for each dimension were analyzed by all six classification algorithms (i.e., Bayesian networks, IBk, J48, logistic regression, MLP, and SMO). F-measures were collected from the results of analysis by each of the six algorithms individually.

In order to make the analysis of a large number of combinations of dimensions, data quality levels, and problem complexity levels feasible, an application was written to automate the entire procedure from dataset generation through statistical analysis

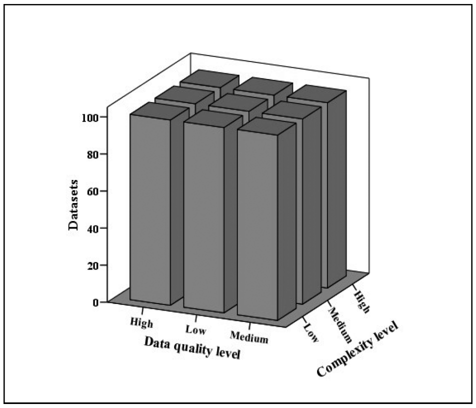


Fig. 6. Representation of the datasets used for statistical analysis.

Table IV. Significant Main Effects Found from ANOVAs to Evaluate Hypothesis I

Dimension	**Bayes Net**	**IBk**	**J48**	**Logistic**	**MLP**	**SMO**
Accuracy	YES	YES	YES	YES	YES	YES
Completeness	YES	YES	YES	YES	YES	YES
Consistency	YES	YES	YES	YES	YES	YES
Timeliness	YES	YES	YES	YES	*NO*	*NO*
Problem complexity	YES	YES	YES	YES	YES	YES

and hypothesis testing. This application was written in C# and architected by utilizing Weka 3.5.6 using the IKVM 0.36.0.5 bridge and using SPSS version 15.0.1.

The default parameters from Weka and ten-fold cross-validation were used for the analyses by each of the six algorithms. Our purpose is not to compare algorithms, but to evaluate the results from a wide range of classification algorithms to find whether general agreement exists.

5. RESULTS

An ANOVA was used to evaluate the effects for each of the six classification algorithms for all individual hypotheses. Presented next are the results summarized by the four major “umbrella” hypotheses.

5.1 Hypothesis I

For Hypothesis I, one-way ANOVAs were used to evaluate the effects of problem complexity and the four dimensions on the outcomes produced by each of the six classification algorithms. The ANOVAs showed significant effects for all dimensions and problem complexity as predicted by Hypothesis I ($p < .05$, the criteria used throughout) except for two cases of timeliness, which appear as the italicized and shaded cells in Table IV. We therefore accept Hypothesis I, which was expected and a necessary step before proceeding to the next hypotheses.

5.2 Hypothesis II

Hypothesis II theorizes the existence of significant and negative effects as levels of data quality decrease while controlling for the effects of problem complexity. One-way ANOVAs were conducted to test the 12 individual hypotheses comprising

Table V. Mean F-Measures for Hypothesis II with Significant Effects Indicated by Nonshaded Cells

Dimension	Data Quality Level	Bayes Net	IBk	J48	Logistic	MLP	SMO	Average
Accuracy	High	0.974	0.975	0.974	0.960	0.957	0.953	0.965
	Medium	0.934	0.940	0.931	0.898	0.914	0.904	0.920
	Low	0.899	0.912	0.895	0.852	0.888	0.850	0.883
Completeness	High	0.971	0.979	0.975	0.963	0.964	0.958	0.968
	Medium	0.932	0.949	0.939	0.913	0.934	0.918	0.931
	Low	0.901	0.922	0.904	0.866	0.905	0.875	0.895
Consistency	High	0.958	0.965	0.962	0.945	0.945	0.941	0.953
	Medium	0.906	0.913	0.904	0.864	0.891	0.873	0.892
	Low	0.858	0.882	0.853	0.802	0.839	0.793	0.838
Timeliness	High	0.998	0.999	0.998	0.998	*0.984*	*0.981*	0.993
	Medium	0.997	0.997	0.996	0.996	*0.983*	*0.981*	0.991
	Low	0.994	0.994	0.991	0.994	*0.980*	*0.981*	0.989

Hypothesis II. With six algorithms used to test the individual hypotheses, a total of 72 ANOVAs were conducted.

With the exception of five ANOVAs for timeliness, all indicated significant effects. Those five were for the MLP algorithm while any of the three levels of complexity was held constant, and two were for the SMO algorithm with problem complexity held at either the medium or high levels. These cases are highlighted by the shaded cells with italicized text in Table V, which summarizes the F-measures produced by the six algorithms for each dimension and data quality level.

A comparison of the F-measures in Table V shows the decreases in overall F-measures from decreasing data quality in all but one case for timeliness and the SMO algorithm. More detailed comparisons at each of the three levels of problem complexity showed similar results.

The magnitude of decreases for timeliness were small, especially when compared with the decreases for other dimensions, averaging .002 across all six algorithms and problem complexity levels as can be seen in Table V. However, while small, the decreases for four of the six algorithms, the majority, were statistically significant. Therefore we conclude that there are significant effects of timeliness as well as accuracy, completeness, and consistency on classification outcomes.

To evaluate whether these results varied by classification algorithm, the range of variability across the six algorithms in each row of Table V was calculated. For example, in the first row, for low accuracy, the range of variability as a percent of the overall average was (.975-.953)/.965, or 2.35%. In ten of 12 evaluations in Hypothesis II, this percentage was less than 5%. Given the spectrum of classification algorithms used in this study, this low degree of variability gives us reason to believe that the results are not algorithm dependent. Therefore we accept Hypothesis II.

5.3 Hypothesis III

Hypothesis III theorizes significant and negative effects from increasing problem complexity levels when controlling for data quality. Analogously to Hypothesis II, 72 ANOVAs were used to evaluate the effects of problem complexity for the six algorithms. Of these ANOVAs all but seven indicated significant effects; these are indicated by the shaded and italicized cells in Table VI. For the dimension of consistency they were for the Bayes Net algorithm with data quality held at the high level and the J48 algorithm

Table VI. Mean F-Measures for Hypothesis III with Significant Effects Indicated by Nonshaded Cells

Dimension	**Problem Complexity Level**	**Bayes Net**	**IBk**	**J48**	**Logistic**	**MLP**	**SMO**	**Average**
Accuracy	Low	0.951	0.958	0.952	0.943	0.950	0.941	0.949
	Medium	0.934	0.940	0.929	0.914	0.913	0.914	0.924
	High	0.923	0.929	0.917	0.854	0.895	0.852	0.895
Completeness	Low	0.952	0.954	0.952	0.947	0.951	0.946	0.950
	Medium	0.932	0.947	0.939	0.911	0.932	0.927	0.931
	High	0.921	0.948	0.928	0.885	0.919	0.878	0.913
Consistency	Low	*0.908*	0.927	*0.908*	0.898	0.906	0.898	0.907
	Medium	*0.905*	0.919	*0.906*	0.877	0.889	0.884	0.897
	High	*0.908*	0.915	*0.906*	0.837	0.879	0.824	0.878
Timeliness	Low	0.998	*0.998*	0.998	0.998	0.996	0.990	0.997
	Medium	0.997	*0.997*	0.996	0.996	0.983	0.983	0.992
	High	0.994	*0.994*	0.992	0.994	0.967	0.969	0.985

with data quality held at any of the three data quality levels, and for timeliness they were for the IBk algorithm, also while data quality was held at any of its three levels.

Table VI also shows decreases in F-measures with increasing problem complexity, consistent with the behavior predicted by Hypothesis III. The only three exceptions to this were for comparisons of F-measures in the same cases for which no significant effects were found.

As with Hypothesis II, to evaluate the variability of results by algorithm, the range of values from lowest to highest F-measure, taken as a percentage of the overall average F-measure for all six algorithms across each row of Table VI, is less than 5% in 11 of the 12 cases. This again gives us reason to believe the results are not algorithm dependent. From the high degree of agreement among algorithms for the hypothesized effects, the decreasing F-measures with increasing complexity, and the lack of a pattern among the six classification algorithms, we accept Hypothesis III.

5.4 Hypothesis IV

Hypothesis IV posits significant interaction effects from both data quality and problem complexity together. To test Hypothesis IV for each of the four data quality dimensions an ANOVA was conducted for each of the six classification algorithms. Table VII enumerates the main and interaction effects which were evaluated, and shows that 64 of these 72 effects were significant. The eight cases with no significant effects are indicated in Table VII by the shaded and italicized cells. These few cases, distributed among the four dimensions, indicate that both the main and interaction effects are significant. Table VII also shows that these few cases without significant effects are distributed among three of the six algorithms, indicating that these results are not algorithm dependent.

Confirmation of Hypothesis IV requires that F-measures decrease with lower data quality, higher complexity, and their interactions. There are 84 comparisons of F-measures to be made for each dimension: two for data quality (high to medium, medium to low) at each problem complexity level, two for problem complexity (low to medium, medium to high) at each data quality level, and two comparisons for interaction effects; all of these to be made for six algorithms. For accuracy, all 84 showed decreases as expected, for completeness, 82 of the 84 comparisons were as expected, for consistency 77 of 84 comparisons showed the expected decreases, three from Bayesian

Table VII. Mean F-Measures and Significant Main and Interaction Effects for Hypothesis IV

Dimension	F-measures / Effect	Bayes Net	IBk	J48	Logistic	MLP	SMO	Average
Accuracy	Mean F-measure	0.936	0.942	0.933	0.904	0.919	0.902	0.923
	Data Quality	YES	YES	YES	YES	YES	YES	YES
	Problem Complexity	YES	YES	YES	YES	YES	*NO*	YES
	Interaction	YES	YES	YES	YES	YES	*NO*	YES
Completeness	Mean F-measure	0.935	0.950	0.940	0.914	0.934	0.917	0.931
	Data Quality	YES	YES	YES	YES	YES	YES	YES
	Problem Complexity	YES	YES	YES	YES	YES	YES	YES
	Interaction	YES	YES	YES	YES	*NO*	*NO*	YES
Consistency	Mean F-measure	0.907	0.920	0.906	0.870	0.891	0.869	0.894
	Data Quality	YES	YES	YES	YES	YES	YES	YES
	Problem Complexity	YES	YES	*NO*	YES	YES	YES	YES
	Interaction	YES	YES	YES	YES	YES	*NO*	YES
Timeliness	Mean F-measure	0.996	0.996	0.995	0.996	0.982	0.981	0.991
	Data Quality	YES	YES	YES	YES	YES	YES	YES
	Problem Complexity	YES	YES	YES	YES	YES	YES	YES
	Interaction	YES	YES	YES	YES	*NO*	*NO*	YES

networks, one from IBk, and three from J48. One comparison for timeliness involving Bayesian networks did not show decreasing F-measures.

Because significant main and interaction effects were found for the large majority of cases, the F-measures decrease with lower data quality and higher problem complexity as posited by Hypothesis IV, and there was no apparent pattern among classification algorithms, we accept Hypothesis IV.

We conclude that the effects of lower data quality can be amplified by the particular structure of a classification problem, a result of importance to practitioners as illustrated in the following real-world case.

6. REAL-WORLD APPLICATION

This section describes how we applied the results found for data quality and problem complexity to a real-world case. Our goal was to evaluate the extent of support for the conclusions made using generated data.

6.1 Introduction

Even for the more objective dimensions of data quality such as the four in our study, quality needs to be evaluated within the context of usage [Shankaranarayanan and Cai 2006], and both procedures and metrics are needed to be able to quantitatively assess data quality in an economic context [Heinrich et al. 2009]. However, assessments made using only impartial measures focus on defects present in data, and are most helpful for determining the costs associated with data quality. To evaluate the utility of data quality, including the benefits of improving it, specific contexts of use must be incorporated into assessment [Even and Shankaranarayanan 2009]. These authors demonstrated how both impartial and contextual factors can be used to assess data quality in an economic context.

This prior research motivates us to apply our results in the economic context of a real-world example. This example is from the case of Tel-Fast, a New England-based telecommunications company that resells voice, Internet, and other services to individuals and small businesses. The strategy of the company is to compete on price while meeting industry standards for customer service, a strategy the company has pursued with success. Tel-Fast is able to compete on price by offering customers packages of bundled services it purchases from major telecommunications vendors at wholesale rates.

Developing and maintaining a flow of recurring revenue from steady customers is critical for companies in this industry to maintain profitability, and Tel-Fast is no exception. A customer defection means not only the loss of an established revenue stream, but the costly prospect of acquiring a new customer to balance that loss. The company attempts to minimize customer defections by offering discounted rate plans to customers who might leave. Although the loss of profits from giving a customer a discounted plan has less consequence than losing a customer, the discount still affects profitability negatively. Therefore it is important that Tel-Fast avoid offering discounts to customers who would be staying with the company whether or not they are offered a discount.

Tel-Fast's database has a log into which customer service representatives record each contact they have with a customer. Representatives assign a type code to each contact from one of the following five categories.

- *Regular business* – These are contacts for routine activities such as customer requests to add a new line, change features, recover PIN codes, question an item on an invoice, or move to ebilling. Entries in this category are most often initiated by the customer.
- *Service problems* – These include notifications by a customer of a problem, notes to and from the company's internal trouble ticket system, and events recorded by service technicians concerning the problem and resolution.
- *Complaints* – These include contacts first initiated by a customer to register a complaint. In addition, contacts can be placed in this category for a customer contact regarding a service or invoice problem that turns into what is judged as a complaint. That judgment is at the discretion of the customer service representative; categorizations by individual representatives for similar situations can vary.
- *Company initiated* – Log entries placed in this category are predominately from contacts made by the company to prompt customers to make payments on overdue invoices, or to establish payment plans for delinquent balances. This category is also used for phone calls to maintain contact with a select group of high volume customers.
- *Others N.E.C.* – This category, for others Not Elsewhere Classified, is used for contacts not fitting into the other categories such as internal notes from service providers and vendors that affect a customer's account.

For this study Tel-Fast provided data for 1,000 randomly selected customers from a two-and-a-half year period. For the customer retention analysis the class variable has two categories, whether a customer will stay or leave, and there are five attribute variables which are from the number of entries in each of the contact log categories given before. Although not identical, we posit that the data structures for Tel-Fast and the generated datasets are reasonably close and that this allows us to determine how well a real-world example supports the conclusions made using generated data.

Table VIII. Sample of Tel-Fast's Log Entries and Customer Retention Data

Customer ID	Regular business	Service problems	Complaints	Company initiated	Others NEC	Outcome
109	6	0	2	1	0	Lost
114	23	0	5	1	6	Retained
237	12	6	0	0	2	Lost
389	9	8	0	1	2	Retained

Table IX. Customer Counts for Each Scenario of Problem Complexity

Low problem complexity		Medium problem complexity	
Outcome	Number of customers in sample	Outcome	Number of customers in sample
Retained	661	Retained – no payment problems	480
		Retained – payment problems	181
Lost	339	Lost– no payment problems	191
		Lost – payment problems	148
Total	1,000	Total	1,000

The sample of customer contact data in Table VIII exemplifies Tel-Fast's difficulty predicting customer retention. Customer 109 meets the perhaps intuitive expectation that a customer with complaints will leave the company. But customer 114, who had more complaints than did customer 109, stayed with the company and is a contrary example. Unlike customer 114, customer 237 had no complaints but registered six service problems and then left the company. If service problems have the same significance as complaints in predicting retention, customer 389 had more service problems than did customer 237, but is another contradictory case because customer 389 stayed with the company while customer 237 left.

6.2 Levels of Problem Complexity

Two levels of problem complexity were used in Tel-Fast's classification problem. The first is a low problem complexity level in which two class categories represent customer retention or loss. We could determinewhether a customer was retained using the date the most recent invoice was sent to a customer; customers still receiving invoices at the end of the sample period were known to be retained; customers receiving their last invoice prior to the end of that period were no longer customers.

The second level of problem complexity was for Tel-Fast's classification models with two additional class categories: one for customers who had given them problems with payments (payments persistently late or incomplete) and one for those who had not. By knowing which customers will give them problems with payments, Tel-Fast can involuntarily terminate customer contracts and thereby avoid potential future losses. This practice of Tel-Fast is common in the telecommunications industry [Hadden et al. 2007]. The classification problem for retention and problem payments combined has four class categories: for customers that are retained or lost, and either have or do not have payment problems. This is a problem of medium complexity. Customer counts from Tel-Fast's sample data for the low and medium complexity problems are displayed in Table IX.

6.3 Levels of Data Quality

Tel-Fast's management perceived the quality of their data to be very good. We therefore defined the "as-is" quality of data in their sample as a high level of data quality. The medium and low levels were defined as reduced levels of data quality relative to the high level. The metrics, methods, and parameters used to represent the three levels of accuracy, completeness, and consistency for Tel-Fast were the same as for the generated data. Inaccuracies were created by switching one attribute value for that in another instance; incomplete values by inserting null values; inconsistencies by switching attribute values to one from an instance in a different class. The high level of data quality was reduced by 8% to represent medium quality and 15% for low quality.

Tel-Fast's data and the nature of their classification problems did not enable us to use an identical representation of timeliness as we did for the generated data. Recall that in Tel-Fast's data we could not determine whether a customer was retained or lost until after the company stopped sending invoices. We therefore defined decreased timeliness from the delays which could occur in recording invoice data in the company's database. Using this definition the level of timeliness could be varied by changing volatility (i.e. "how fast" changes in the real world occur), the same component of timeliness manipulated in the generated data. Representing timeliness in this way matched Tel-Fast's billing operations which were outsourced to an application service provider. Tel-Fast experienced occasional delays receiving the processed invoice data back from this vendor because of the vendor's capacity constraints and commitments to other customers. In our analysis a one-month delay was used for a medium level of timeliness and a two-month delay for a low level.

6.4 Evaluating Data Quality and Problem Complexity

The six classification algorithms were used to analyze each combination of the data quality levels for each of the data quality dimensions (accuracy, completeness, complexity, and timeliness), and the two levels of problem complexity for Tel-Fast.

A general sense of how much support the real-world application might hold for the four hypotheses is shown in Figure 7, which provides a visualization of F-measures averaged across all six algorithms. Except for one case of consistency and one for timeliness, both at low complexity, the F-measures show general support as can be seen by comparing the F-measures at varying levels of data quality and problem complexity.

Table X shows the F-measures for Tel-Fast from all six algorithms and all combinations of dimensions, data quality levels, and problem complexity levels. Of the 144 F-measures from the six algorithms, 118, or 82%, decrease and are consistent with Hypotheses I through IV. This shows support for the real-world application from the results of the four hypotheses evaluated with generated data.

However, in 26 cases F-measures increased where decreases were anticipated, as highlighted in Table X by italicized figures in shaded cells. Of those, 10 instances are from Bayesian networks and one each from J48 and MLP. Although on the surface there does not seem to be an obvious pattern, this imbalance leads to the question of whether or not the classification algorithms have produced similar results. To compare the F-measures from all six algorithms, coefficients of determination, as adjusted R^2values, were calculated and shown in Table XI. These coefficients indicate that a high degree of variance in F-measures of any algorithm, above .50, can be explained by those from any other algorithm except in one case. That single instance is for Bayesian networks compared with SMO, although that coefficient is still above .3. We therefore conclude that classification algorithms in this real-world case produce similar results, and that these results are analogous to those for generated data.

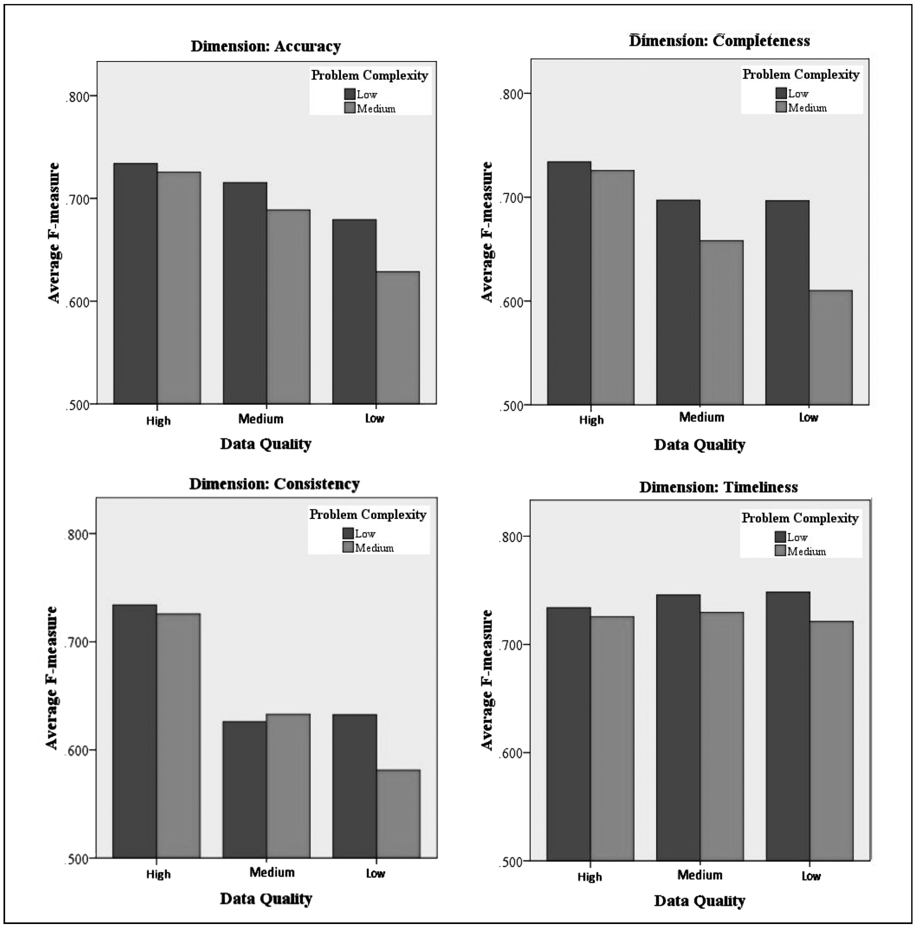


Fig. 7. F-measures averaged for the six algorithms and displayed by dimension of data quality.

To further how well the real-world application matches our results from generated data, we conducted a series of linear regressions to determine if a statistically significant relationship might exist between the F-measures from the real-world application and those from the generated data. These regressions modeled this potential relationship as $F(RW_{q,p,a}) = \beta_0 + \beta_1 F(G_{q,p,a})$, where $F(RW_{q,p,a})$ is the set of F-measures from the real-world application for each data quality level q, problem complexity level p, and algorithm a, and $F(G_{q,p,a})$ is the corresponding set of F-measures from the generated data. Summary statistics from the regressions for each dimension are shown in Table XII. Each regression shows a statistically significant relationship, $p_{df=34} < .05$, between the F-measures from the generated data results and the real-world application. This lends more support to our findings.

This real-world case was used to see if it would lend support to the findings from generated data. The real-world datasets were analyzed for the four data quality dimensions and all levels of data quality and problem complexity by the same six algorithms as for the generated data. Comparisons of F-measures from the six classification algorithms showed that all six algorithms produced consistent results for all four dimensions.

Table X. F-Measures by Dimension, Data Quality Level, Problem Complexity Level, and Algorithm

Dimension	Problem Complexity	Data Quality	Bayes Net	IBk	J48	Logistic	MLP	SMO	Avg
Accuracy	Low	High	0.757	0.762	0.790	0.729	0.792	0.574	0.734
		Medium	0.749	0.752	0.781	0.684	0.755	0.572	0.715
		Low	0.726	0.683	0.746	0.627	0.727	0.567	0.679
	Medium	High	*0.787*	0.741	0.775	0.757	0.781	0.512	0.726
		Medium	*0.750*	0.714	0.745	0.693	0.731	0.500	0.689
		Low	0.707	0.632	0.702	0.576	0.695	0.461	0.629
Completeness	Low	High	0.757	0.762	0.790	0.729	0.792	0.574	0.734
		Medium	0.744	0.691	0.766	0.692	0.715	0.574	0.697
		Low	0.743	0.687	0.780	*0.699*	0.697	0.573	0.696
	Medium	High	*0.787*	0.741	0.775	*0.757*	0.781	0.512	0.726
		Medium	*0.751*	0.561	0.756	*0.717*	0.692	0.471	0.658
		Low	*0.714*	0.440	0.704	0.674	0.661	0.467	0.610
Consistency	Low	High	0.757	0.762	0.790	0.729	0.792	0.574	0.734
		Medium	0.655	0.647	0.675	0.604	0.659	0.515	0.626
		Low	0.677	*0.666*	0.689	0.576	0.659	0.526	0.632
	Medium	High	*0.787*	0.741	0.775	0.757	0.781	0.512	0.726
		Medium	*0.712*	0.643	0.706	0.597	*0.718*	0.421	*0.633*
		Low	0.653	0.605	0.650	0.547	0.635	0.398	0.581
Timeliness	Low	High	0.757	0.762	0.790	0.729	0.792	0.574	0.734
		Medium	0.757	0.775	0.789	*0.768*	0.791	*0.594*	0.746
		Low	0.751	0.757	0.765	*0.770*	0.775	*0.672*	*0.748*
	Medium	High	*0.787*	0.741	0.775	*0.757*	0.781	0.512	0.726
		Medium	*0.782*	*0.751*	*0.786*	*0.792*	0.768	0.498	0.730
		Low	*0.766*	0.731	0.749	*0.778*	0.765	0.538	0.721

Table XI. Coefficients of Determination (adjusted R^2) for F-Measures between the Six Classification Algorithms

	Bayes Net	IBk	J48	Logistic	MLP	SMO
Bayes Net	1.000	0.587	0.892	0.905	0.853	0.380
IBk	0.587	1.000	0.700	0.569	0.826	0.620
J48	0.892	0.700	1.000	0.850	0.871	0.647
Logistic	0.905	0.569	0.850	1.000	0.813	0.559
MLP	0.853	0.826	0.871	0.813	1.000	0.516
SMO	0.380	0.620	0.647	0.559	0.516	1.000

Hypotheses I through IV each posited decreases in F-measures as data quality decreases or problem complexity increases. In the real-world case the F-measures from all six algorithms were in general agreement with those from the generated datasets for accuracy, completeness, and consistency. Although the majority of F-measures agreed for timeliness, the support was less strong. This may be due to Tel-Fast use of two billing cycles per month which made invoice dates "lumpy." Depending on the specific metric values used for timeliness, a quality level might encompass all invoices in a billing cycle or it might encompass none. Nonetheless, a regression analysis showed a statistically significant relationship between the F-measure from generated datasets and those found from the real-world datasets.

Table XII. Summary of Regressions Evaluating the Relationship between Results for Generated and Real-World Data

Dimension	**Model**	**Unstandardized coefficients**		**Standardized coefficients**	$t_{(df=34)}$
		B	**Std**	**Beta**	
Accuracy	Intercept	.805	.038		21.22
	F-measure from generated data results	.189	054	.513	3.49
Completeness	Intercept	.864	.034		25.22
	F-measure from generated data results	.112	.049	.363	2.27
Consistency	Intercept	.720	.040		18.00
	F-measure from generated data results	.278	.060	.621	4.62
Timeliness	Intercept	.963	.007		137.94
	F-measure from generated data results	.043	.009	.611	4.50

We conclude that the real-world application indicates broad support for the conclusions of significant effects found for each of the four hypotheses.

6.5 Economic Impacts of Data Quality Improvement

Understanding data quality in an economic context is an important research challenge and one that has yet to be fully addressed [Even and Shankaranarayanan 2007]. An economic context is especially important in our case because customer retention is a key driver of profitability; even a modest 5% increase in customer retention can increase the profits of a firm from 25% to 85% [Reichheld and Sasser 1990].

From the results of this study, we know that improving data quality has significant positive effects on the outcomes of classification. Improved classification outcomes result from an increased number of correct predictions, increases which can come from a higher rate of true positives, a higher rate of true negatives, or both together.

The Appendix shows the derivation of the gain in profits that will accrue when improved data quality improves the outcomes of classification. Eq. (A.5) in the Appendix shows the gain in profits to be

$$D\Delta_{12} + W\Delta_{24}(P - D),$$

where P equals the profit lost from losing a customer, D is the reduction in profit due to a discounted rate plan offered and accepted by a customer, and W is the win-back percentage of customers who take the discount and decide to stay. In addition, Δ_{13} is the increase in the rate of true positives and Δ_{24} is the increase in the rate of true negatives that result from improving classification.

A review of Tel-Fast's billing data showed that the average monthly customer invoice totaled $277.10. The company's net profit before taxes was 9%. A customer who leaves the company reduces monthly profits by $24.94, the value of P. If accepted, a typical discounted rate plan used to induce customers to stay would result in an overall 10% discount on their bills. Given that cost of offering a discounted plan to a customer is negligible, the monthly reduction in profits is $2.49, the value of D. The company's win-back rates varied by retention program; here we will use a win-back rate of 35% for W.

From the analysis of low complexity problems in the previous sections, the average rate of improvement in true positives, Δ_{13}, across all quality dimensions was 3.3%, and for true negatives, Δ_{24}, the average improvement was 2.2%. Given the aforesaid values for P, D, and W, Eq. (A.5) is used to calculate Tel-Fast's incremental gains in profit accrued from improved data quality.

The resulting gain per month per customer is \$0.256; across Tel-Fast's 60,000 customers the annual increase in profits is \$184,320. When considered in conjunction with the costs of improving data quality, this analysis can be used for targeting the approaches that would reap the highest economic benefit from those improvements.

The economic analysis of data quality presented here might be integrated with the Total Data Quality Management (TDQM) cycle first introduced by Madnick and Wang [1992]. In the first step of the cycle, data quality is defined for a set of dimensions to be considered for improvement and in the second step data quality is measured for those dimensions. In this study, we defined four important dimensions of data quality and developed metrics to measure each. The third step in TDQM is to determine whether improvements are justified, and if so, which dimensions warrant improving. The knowledge of significant effects and interactions, combined with an economic analysis such as in this study, could help target specific actions for improvement.

Note that this section is not intended as a generalization of which dimensions of data quality are most important. It shows the economic implications of improving data quality for a real-world case. The process for doing so as outlined in this example (and based upon the development in the Appendix) might be generalized for customer retention analyses in companies similar to Tel-Fast.

7. CONCLUSIONS

Our research takes a comprehensive view of data quality and demonstrates that specific metrics for accuracy, completeness, consistency, and timeliness can be formulated and used to measure data quality. It demonstrates that each of the four dimensions of data quality has a significant and negative effect on the outcomes of classification using both generated data and in the real-world case of Tel-Fast, a telecommunications company.

Although it still had significant effects, there was somewhat less support for timeliness when evaluated in the real-world case. A set of probabilistic nonlinear metrics for this dimension were recently developed by Heinrich et al. [2009], who applied these metrics to a company in the same industry as Tel-Fast. The results for timeliness could be revisited in light of these new metrics.

Our results show that independently of data quality, problem complexity has significant effects on classification outcomes. Problem complexity is related to the structure of a classification problem and to concepts from data mining, not often considered directly alongside data quality.

This research confirms that there are significant interaction effects between problem complexity and each of the four dimensions of data quality. It shows how the structure of a classification problem can amplify the negative effects of poor data quality, finding these significant effects for both the generated data and the real-world case. All conclusions in this study were based on having consistent classification outcomes produced by six divergent classification algorithms, indicating that the results are not algorithm dependent.

Studying data quality in an economic context is an important research challenge. This research evaluates the effects and interactions of data quality and problem complexity in the economic context of a real-world case and a widely used application of classification. A method is presented to compare the profit implications of data quality

Table AI. Table of Classifications by Customer State

	Customer state – prior to retention program	
	Will stay	**Will leave**
Classified as staying	p_1	p_2
Classified as leaving	p_3	p_4

improvements in each dimension. The data quality metrics and method for analyzing data quality in an economic context could be useful if integrated with the Total Data Quality Management (TDQM) cycle for measuring data quality and evaluating dimensions for improvement.

A preliminary analysis of the real-world case has suggested there may also be significant interaction effects between data quality dimensions. The results from this study could be applied to existing utility models of the trade-offs between quality dimensions. These include Ballou and Pazer's [1998] model of the completeness and consistency trade-off and Even and Shankaranarayanan's [2007] model of the trade-offs between completeness and accuracy.

Finally, the datasets generated for this study had one class and one attribute variable; this schema could be expanded to include multiple attribute variables, possibly with statistical distributions as a factor. Our findings could be tested for additional real-world applications. Lastly, this research could be extended to evaluate additional classes of data mining algorithms, such as those for unsupervised learning. Each is a potentially productive area for new research.

APPENDIX: QUANTIFYING THE ECONOMICS OF DATA QUALITY

Consider a classification model which classifies whether a customer will stay or leave a company; this model is built and customers are classified prior to any consideration of a retention program. A customer will or will not leave the company at some point in the future. Table AI illustrates the four combinations of classification (classified to stay or to leave) and customer state (will stay or will leave):

In Table AI, customer state designates whether a customer actually will or will not leave the company at some point in the future. Note that statistics generated for a classification model such as the F-measure, the rate of true positives, the rate of true negatives, etc., are derived from training and test data used to build that model. This data is past data from a sample of the company's customer database. Whether a customer stays or leaves at some future point is an unknown state at the time the model is built and customers are classified.

Table AI shows the division of customer classifications into $p_1 + p_2$ as the proportion of customers classified as staying with the company, and $p_3 + p_4$ the proportion classified to leave. Without a retention program, the proportion of customers whose state is to stay will be $p_1 + p_3$, and proportion in the state to leave will be $p_2 + p_4$; when summed $p_1 + p_2 + p_3 + p_4 = 1$. If a customer leaves then the profit from that customer leaves with them. Let P equal the profit lost from losing a customer. The impact on the company's profits of customers leaving and the company does not have a retention program inducing them to stay is

$$(p_2 + p_4)P. \tag{A.1}$$

Next consider a retention program the company offers to induce customers to stay who would otherwise leave the company. It is reasonable to assume that the loss of a profit stream from a customer who leaves is significantly greater than a financial incentive that might induce a customer to stay, even without considering the costs of new customer acquisition. In Tel-Fast's case retention programs involve a discounted rate plan; we consider a discount as the financial incentive here. We assume that the offer of the discount plan is economically appropriate (i.e., it is better economically to offer the plan to all customers classified as leaving in order to retain some of the customers in the "will leave" state than not offer the retention program and lose all the customers in the "will leave" state).

The retention program is implemented using the classification analysis in the following way.

(1) Any customer classified to stay is not offered a discount. The financial results to the company will be such that:

(1.1) A customer who was staying with the company, and is not targeted with an offer of a discount, will continue to stay. These are customers correctly classified as in the state of "will stay" prior to a consideration of the retention program, and the fact that they were not targeted for the program will not affect their state of retention. For this group of customers, profits are unchanged.

(1.2) A customer who was leaving the company, and is not targeted with an offer of a discount, will leave. These customers are incorrectly classified as staying will not be offered an incentive to stay, and so the state of "will leave" will be unchanged. Gone with the customer are their profits, P (note that P is a negative number indicating lost profits). The incorrect identification of customers who will leave and the subsequent failure to target retention efforts to them has the greatest economic consequence for the company.

(2) All customers classified to leave are offered a discount to stay with the company. It should be noted that Tel-Fast is in routine contact with their customers and has the capacity to make additional contacts when needed. The cost of offering a customer the discount is minor or even negligible. This differentiates the use of classification for retention analysis here from other applications such as direct marketing in which the costs of reaching customers with an offer (such as a catalog) are significant. In such cases, the area under a ROC curve (AUC) is commonly used and appropriate metric to select a subset of customers that will maximize profitability. In this appendix, as with Tel-Fast, all customers predicted to leave will be offered the discount.

(2.1) A customer staying with the company, and is offered a discount, will accept the discount and continue to stay with the company. These customers are incorrectly classified as leaving. As these customers take the discount when, in fact, none need have been offered, the profit stream is reduced by the amount of the discount D (note that D is a negative number indicating a reduction in profits due to the discount).

(2.2) A customer who was leaving, but is offered a discount to stay with the company, might be induced to stay. A discount will be incentive enough to keep some customers who otherwise intended to leave. The rate of success for these win-backs is designated W. Based on experience management is able to provide an estimate of W for a particular discount plan. Because the discount was required for retention, profits will be decreased by the discount amount D.

(2.3) Some customers who were leaving will not be swayed by the offer of a discount. The profits of the company will decrease by P for the $1 - W$ attempts at a win-back that are unsuccessful.

The changes in profit that will accrue from each of these scenarios are shown in Table AII.

Table AII. Implications of a Retention Program for Changes in Profit

	Customer state – after offering a retention program	
	Will stay	**Will leave**
Classified as staying	0	*P*
Classified as leaving	*D*	*WD + (1 − W)P*

Table AIII. Changes in Customer Groupings from Improved Data Quality

	Customer state – after offering a retention program	
	Will stay	**Will leave**
Classified as staying	$p_1 + \Delta_{13}$	$p_2 - \Delta_{24}$
Classified as leaving	$p_3 - \Delta_{13}$	$p_4 + \Delta_{24}$

D, P,and$W$ are defined as before.

The impact on the company's profit stream of offering the discount program is obtained by multiplying the values in Tables AI and AII yielding

$$0p_1 + Pp_2 + Dp_3 + WDp_4 + (1 - W)Pp_4. \tag{A.2}$$

We reiterate the assumption that it is better economically to offer the plan to all customers classified as leaving in order to retain some of the customers in the "will leave" state than not offer the retention program and lose all the customers in the "will leave" state. We can subtract Eq. (A.2) from (A.1) and find the net effect on profit by offering the discount.

$$WPp_4 - Dp_3 - WDp_4 \tag{A.3}$$

Note that D and P are negative numbers. Given our stated assumptions, the preceding equation will be positive. This is an improved profit situation due to offering the discount and retaining customers who would have left without this incentive.

As the results of this study show, improving data quality improves the outcomes from classification. Improved classification outcomes result in increases in the number of true positives (with a corresponding decrease in the number of false negatives), the number of true negatives (with a corresponding decrease in the number of false positives), or both. Assume the increase in the rate of true positives is Δ_{13} and the increase in the rate of true negatives is Δ_{24}. By definition, $\Delta_{13} + \Delta_{24} > 0$.

Taking customers as they are grouped in Table AI, when classification is improved p_1 will increase by Δ_{13} and p_3 will decrease by the same amount, and p_4 will increase by Δ_{24} and p_2 will decrease by the same amount. The classification improvements Δ_{13} and Δ_{24} each affect a single state, and the resulting changes from those improvements are applied to the proportions in Table AI and shown in Table AIII.

We can now quantify the change in profits from a retention program when data quality and classification outcomes are improved by multiplying the values in Tables AII and AIII. This yields

$$0(p_1{+}\Delta_{13}) + P(p_2 - \Delta_{24}) + D(p_3 - \Delta_{13}) + (WD + (1 - W)P)(p_4 + \Delta_{24}). \tag{A.4}$$

Subtracting (A.4) from (A.2) results in

$$D\Delta_{13} + W\Delta_{24}(P - D). \tag{A.5}$$

Eq. (A.5) is the incremental and positive change in profits that will accrue when improved classification due to better data quality increases true positives by Δ_{13} and true negatives by Δ_{24} as compared to not improving data quality and classification.

REFERENCES

ALI, S. AND SMITH, K. A. 2006. On learning algorithm selection for classification. *Appl. Soft Comput. J. 6*, 119–138.

APTE, C., LIU, B., PEDNAULT, E. P. D., AND SMYTH, P. 2002. Business applications of data mining. *Comm. ACM 45*, 49–53.

BALLOU, D., WANG, R., PAZER, H., AND KUMAR, T. G. 1998. Modeling information manufacturing systems to determine information product quality. *Manag. Sci. 44*, 462–484.

BALLOU, D. P. AND PAZER, H. L. 1985. Modeling data and process quality in multi-input, multi-output information systems. *Manag. Sci. 31*, 150–162.

BALLOU, D. P. AND PAZER, H. L. 2003. Modeling completeness versus consistency tradeoffs in information decision contexts. *IEEE Trans. Knowl. Data Engin. 15*, 240–243.

DAVENPORT, T. H. AND HARRIS, J. G. 2007. Competing on Analytics: The New Science of Winning. Harvard Business School Publishing Company, Boston, MA.

DILLARD, R. A. 1992. Using data quality measures in decision-making algorithms. *IEEE Intell. Syst. Appl. 7*, 63–72.

ECKERSON, W. W. 2002. Data warehousing special report: Data quality and the bottom line. In *Applications Development Trends.*

EVEN, A. AND SHANKARANARAYANAN, G. 2007. Utility-driven configuration of data quality in data repositories. *Int. J. Inf. Quality 1*, 22–40.

EVEN, A. AND SHANKARANARAYANAN, G. 2009. Dual assessment of data quality in customer databases. *ACM J. Inf. Data Quality 1*, 3.

FISHER, C., LAURIA, E., AND MATHEUS, C. 2007. In search of an accuracy metric. In *Proceedings of the 12th International Conference on Information Quality.*

GE, M. AND HELFERT, M. 2006. A framework to assess decision quality using information quality dimensions. In *Proceedings of the International Conference on Information Quality.*

GOMES, P., FARINHA, J., AND TRIGUEIROS, M. J. 2007. A data quality metamodel extension to CWM. In *Proceedings of the 4th Asia-Pacific Conference on Conceptual Modeling.* 17–26.

HADDEN, J., TIWARI, A., ROY, R., AND RUTA, D. 2007. Computer assisted customer churn management: State-of-the-Art and future trends. *Comput. Oper. Res. 34*, 2902–2917.

HEINRICH, B., KLIER, M., AND KAISER, M. 2009. A procedure to develop metrics for currency and its application in CRM. *ACM J. Inf. Data Quality 1*, 3.

HICKEY, R. 1996. Noise modelling and evaluating learning from examples. *Artif. Intell. 82*, 157–179.

KAHN, B. K., STRONG, D. M., AND WANG, R. Y. 2002. Information quality benchmarks: Product and service performance. *Comm. ACM 45*, 185–192.

KARR, A. F., SANIL, A. P., AND BANKS, D. L. 2006. Data quality: A statistical perspective. *Statist. Method. 3*, 137–173.

KLEIN, B. D., GOODHUE, D. L., AND DAVIS, G. B. 1997. Can humans detect errors in data? Impact of base rates, incentives, and goals. *MIS Quart. 21*, 169–194.

KOHAVI, R., ROTHLEDER, N. J., AND SIMOUDIS, E. 2002. Emerging trends in business analytics. *Comm. ACM 45*, 45–48.

LAKSHMINARAYAN, K., HARP, S. A., AND SAMAD, T. 1999. Imputation of missing data in industrial databases. *Appl. Intell. 11*, 259–275.

LEE, Y. W., PIPINO, L., STRONG, D. M., AND WANG, R. Y. 2004. Process-embedded data integrity. *J. Datab. Manag. 15*, 87–103.

LEE, Y. W., PIPINO, L. L., FUNK, J. D., AND WANG, R. Y. 2006. *Journey to Data Quality*. The MIT Press.

LEE, Y. W., STRONG, D. M., KAHN, B. K., AND WANG, R. Y. 2002. AIMQ: A methodology for information quality assessment. *Inf. Manag. 40*, 133–146.

MADNICK, S. AND WANG, R. Y. 1992. Introduction to total data quality management (TDQM). Research Program TDQM-92-01, Total Data Quality Management Program, MIT Sloan School of Management.

MARCH, S. T. AND HEVNER, A. R. 2007. Integrated decision support systems: A data warehousing perspective. *Decis. Support Syst. 43*, 1031–1043.

OATES, T. AND JENSEN, D. 1997. The effects of training set size on decision tree complexity. In *Proceedings of the 14th International Conference on Machine Learning.* Morgan Kaufmann Publishers, 254–262.

ORDONEZ, C. AND GARCÍA-GARCÍA, J. 2008. Referential integrity quality metrics. *Decis. Support Syst. 44*, 495–508.

PARSSIAN, A. 2006. Managerial decision support with knowledge of accuracy and completeness of the relational aggregate functions. *Decis. Support Syst. 42*, 1494–1502.

PARSSIAN, A., SARKAR, S., AND JACOB, V. S. 2004. Assessing data quality for information products: Impact of selection, projection, and cartesian product. *Manag. Sci. 50*, 967–982.

PIPINO, L. L., LEE, Y. W., AND WANG, R. Y. 2002. Data quality assessment. *Comm. ACM 45*, 211–218.

QUINLAN, J. R. 1986. Induction of decision trees. *Mach. Learn. 1*, 81–106.

REDMAN, T. C. 2004. Data: An unfolding quality disaster. *DM Rev. 6.*

REICHHELD, F. F. AND SASSER, W. E. 1990. Zero defections. *Harvard Bus. Rev. 68*, 105–111.

SESSIONS, V. AND VALTORTA, M. 2006. Learning Bayesian networks from inaccurate data. In *Proceedings of the 11th International Conference on Information Quality.*

SHANKARANARAYANAN, G. AND CAI, Y. 2006. Supporting data quality management in decision-making. *Decis. Support Syst. 42*, 302–317.

SU, Y. AND JIN, Z. 2007. Assessment and improvement of data and information quality. In *Information Quality Management: Theory and Applications*. Idea Group, Inc.

SWAIT, J. AND ADAMOWICZ, W. 2001. Choice environment, market complexity, and consumer behavior: A theoretical and empirical approach for incorporating decision complexity into models of consumer choice. *Organiz. Behav. Hum. Decis. Process. 86*, 141–167.

WAND, Y. AND WANG, R. Y. 1996. Anchoring data quality dimensions in ontological foundations. *Comm. ACM 39,* 86–95.

WANG, R. Y. AND STRONG, D. M. 1996. Beyond accuracy: What data quality means to data consumers. *J. Manag. Inf. Syst. 12,* 5–33.

WANG, R. Y., ZIAD, M., AND LEE, Y. W. 2000. *Data Quality.* Kluwer Academic Publishers.

WU, Y., FRIZELLE, G., AND EFSTATHIOU, J. 2007. A study on the cost of operational complexity in customersupplier systems. *Int. J. Product. Econom. 106*, 217–229.

ZHU, X. AND WU, X. 2004. Class noise vs. attribute noise: A quantitative study. *Artif. Intell. Rev. 22*, 177–210.

Received December 2008; revised September 2010; accepted November 2010