A Bayesian Approach for Estimating and Replacing Missing Categorical Data

XIAO-BAI LI
University of Massachusetts Lowell

We propose a new approach for estimating and replacing missing categorical data. With this approach, the posterior probabilities of a missing attribute value belonging to a certain category are estimated using the simple Bayes method. Two alternative methods for replacing the missing value are proposed: The first replaces the missing value with the value having the estimated maximum probability; the second uses a value that is selected with probability proportional to the estimated posterior distribution. The effectiveness of the proposed approach is evaluated based on some important data quality measures for data warehousing and data mining. The results of the experimental study demonstrate the effectiveness of the proposed approach.

Categories and Subject Descriptors: H.2.7 [**Database Management**]: Database Administration—*Data warehouse and repository*; H.2.8 [**Database Management**]: Database Applications—*Data mining*

General Terms: Algorithms, Experimentation, Performance

Additional Key Words and Phrases: Missing data, data quality, simple Bayes

ACM Reference Format:
Li, X.-B. 2009. A Bayesian approach for estimating and replacing missing categorical data. ACM J. Data Inform. Quality 1, 1, Article 3 (June 2009), 11 pages. DOI = 10.1145/1515693.1515695. http://doi.acm.org/10.1145/1515693.1515695.

1. INTRODUCTION

Missing data treatment is an important data quality issue in data mining, data warehousing, and database management. Real-world data often has missing values. The presence of missing values can cause serious problems when the data is used for reporting, information sharing, and decision support. First, data with missing values may provide biased information. For example, a

Author's address: X.-B. Li, Department of Operations and Information Systems, College of Management, University of Massachusetts Lowell, Lowell, MA 01854, USA; email: xiaobai_li@uml.edu.

survey question that is related to personal information will more likely be left unanswered for those who are more sensitive about privacy. Second, many data modeling and analysis techniques cannot deal with missing values and have to cast out a whole record value if one of the attribute values is missing [Michie et al. 1994; SAS 1990]. Third, even though some data modeling and analysis tools can handle missing values, there are often restrictions in the domain of missing values. For example, classification systems typically do not allow missing values in the class attribute [Breiman et al. 1984; Quinlan 1993].

Various approaches to handling missing data have been proposed in the literature. The first approach is to disregard all data with any unknown attribute values (listwise deletion). This approach has been used in many statistical techniques and some early machine learning systems [Quinlan 1989; SAS 1990]. The approach is simple; but the main problem is that it will cause significant loss in the available information if the dataset contains many missing values. Empirical studies also showed that it is in general inferior to other missing value treatment approaches (see, for example, Quinlan [1989]).

The second approach is to replace missing values with a default value or a global constant. This approach is widely adopted in the database community. For example, in relational database systems it is a common practice to replace a missing attribute value with a null value (labeled, say, "Null" [Codd 1979]). While this approach is useful in resolving some database problems such as the referential integrity issue, it is not very helpful when the data is used for analysis purposes.

The third approach is to fill in the unknown values using their simple estimates. Usually, if the missing value is of a numeric type, the mean of the nonmissing values for the same attribute is used as the estimate; if it is categorical, the mode (most frequent) value is used. This is perhaps the most widely used approach in practice, and it has been discussed in numerous prior studies [Quinlan 1993; Michie et al. 1994; Clark and Niblett 1989; Pyle 1999]. This method is convenient and provides a satisfactory solution to missing data problems in many cases. A major problem is that the variability associated with missing data is biasedly represented when all missing values of an attribute are replaced with the same value. As a result, the statistical distribution of the data is altered and the quality of the data is affected.

A number of more sophisticated approaches have been proposed in order to find better solutions to missing data problems. These approaches are often geared towards a specific data analysis/mining task. For classification analysis, for instance, a surrogate split method was proposed in Breiman et al. [1984], and a probabilistic weighting method was proposed in Quinlan [1993]. For numeric data analysis, methods based on regression techniques were described in Pyle [1999], Fan et al. [2002], and Witten and Frank [2005]. In general, there are far more well-formulated methods for handling numeric missing data than for categorical data.

Bayesian approaches have been proposed to deal with missing categorical data. An overview of these approaches is provided in Congdon [2005, Chapter 11]. Existing approaches are mostly parametric, addressing issues in sample

surveys rather than in the context of data mining and warehousing. There is a lack of efficient methods for handling missing data spread over many dimensions, especially when the missing data is not missing at random. Chiu and Sedransk [1986] propose a Bayesian procedure for estimating and replacing missing data based on some prior knowledge about the distributions of the data. The procedure, however, primarily applies to univariate missing data and some special multivariate cases. Chen and Astebro [2003] develop a Bayesian method for estimating and replacing missing categorical data, using the uniform prior distribution and a Dirichlet posterior distribution. Their method performed very well when the missing data is missing at random, but it remains to be tested for cases where data is missing not at random.

In this article, we propose a new Bayesian approach for estimating and replacing missing categorical data. With this approach, the posterior probabilities of a missing attribute value belonging to a certain category are estimated using the simple Bayes method. Based on the estimated probabilities, two alternative methods for replacing the missing value are proposed: The first replaces the missing value with the value having the maximum probability; the second uses a value that is selected with probability proportional to the estimated posterior distribution. The proposed approach is nonparametric and does not require prior knowledge about the distributions of the data. It is computationally efficient both in time and in space. The approach is not related to any specific data analysis/mining task and thus can be applied to a wide variety of tasks. It is intended primarily for data with mostly categorical attributes, which is a problem that has not been investigated sufficiently.

The rest of the article is organized as follows. The details of the proposed approach are presented in the next section. An example is given in Section 3 to illustrate how the proposed approach works. Section 4 describes an experimental study that compares the proposed methods with two existing methods using three real-world datasets. Section 5 discusses the limitations of the proposed approach and possible extensions based on our current work.

2. THE PROPOSED BAYESIAN APPROACH

Bayesian methods are used in many different areas of data mining and database management, including classification analysis [Duda et al. 2001], information retrieval [Fung and Favero 1995], data integration [Jiang et al. 2007], and privacy-preserving data mining [Li and Sarkar 2006]. The theoretical fundamental underlying all of the Bayesian methods is the Bayes' theorem, stated next:

Let $c_1, \ldots, c_L$ be a partition of the sample space. Then for any event X in the sample space,

$$P(c_k|X) = \frac{P(c_k)P(X|c_k)}{\sum_{r=1}^{L} P(c_r)P(X|c_r)}, k = 1, \ldots, L, \quad (1)$$

where $P(c_k)$ is called the *prior probability* and $P(c_k|X)$ is called the *posterior probability*.

Our idea of using a Bayesian method to estimate missing values stems from the approach employed in a Bayesian classifier [Duda et al. 2001]. Consider

a dataset with a class attribute of two classes, c_1 and c_2, and $M-1$ nonclass attributes, $X_1, \ldots, X_{M-1}$. For a new record $\mathbf{x} = (x_1, \ldots, x_{M-1})$ to be classified, the Bayesian classifier assigns its class value to c_1 if $P(c_1|\mathbf{x}) > P(c_2|\mathbf{x})$ and otherwise to c_2. The posterior probability $P(c_k|\mathbf{x})$ can be derived from Bayes' theorem (1). The procedure involves estimating $P(c_k)$ and $P(\mathbf{x}|c_k)$ from the data. While $P(c_k)$ is easy to assess, evaluating $P(\mathbf{x}|c_k)$ is computationally very expensive for data with high dimensionality [Duda et al. 2001]. To get around this problem, it is sometimes assumed that the attributes are conditionally independent of each other, given the class value. Under this assumption, $P(\mathbf{x}|c_k)$ can be easily computed by

$$P(\mathbf{x}|c_k) = \prod_{j=1}^{M-1} P(x_j|c_k). \tag{2}$$

A classifier constructed in this fashion is called a simple (or naïve) Bayes classifier.

Our method for estimating missing values draws upon this idea of predicting class values. Instead of estimating the value of the class attribute for a given record, our task now is to estimate all missing values of any attribute in the dataset using the simple Bayes method. In the missing value replacement step, however, we propose two alternatives. The first is to fill in the missing value with the one having the maximum posterior probability (*MaxPost*), similar to the way the simple Bayes classifier assigns a class value. The second is to replace the missing value with a value that is selected with probability proportional to the estimated posterior distribution (*PropPost*).

We should point out that although the problem of categorical missing value replacement can be considered a classification problem, most classification techniques do not naturally translate to a missing value replacement method, because they themselves require a solution for missing values in the nonclass attributes. The proposed simple Bayes method, however, can estimate the probabilities of multiple missing attribute values based directly on nonmissing attribute values.

Given a dataset with N records and M categorical attributes, $X_1, \ldots, X_M$, let L_i be the number of categories in X_i, N_i be the number of records with X_i values known, and N_{ik} be the number of records where X_i equals its kth category c_{ik}. Further, let $N_{jr|ik}$ be the number of records where X_j equals its rth category c_{jr}, given $X_i = c_{ik}$, $j \neq i$. The proposed algorithm for estimating and replacing missing data is given in Table I.

Let L be the average number of categories in each attribute. The time complexity for each step in the algorithm (and thus for the entire algorithm) is of order $O(NML)$. In large datasets, N is typically much larger than M and L. If $N >> ML$, then the time complexity is linear in N. In terms of space complexity, prior and conditional probabilities require $O(ML)$ and $O(M^2L^2)$ space, respectively. Posterior probabilities would need $O(NML)$ space. However, it is not required to store posteriors, since each of them can be computed for each missing value and immediately disregarded once the missing value is filled (estimation and substitution can be done in one pass). Clearly, the algorithm is very efficient for large data both in time and in space.

Table I. The Missing Value Estimation and Replacement Algorithm

1. Compute the prior probabilities for each attribute:
$$P(X_i = c_{ik}) = N_{ik}/N_i, \quad i = 1, \ldots, M; \; k = 1, \ldots, L_i.$$
2. Compute the conditional probabilities of X_j, given $X_i = c_{ik}$:
$$P(X_j = c_{jr}|X_i = c_{ik}) = N_{jr|ik}/N_{ik}, \quad j = 1, \ldots, M; \; j \neq i; \; r = 1, \ldots, L_j.$$
3. For a record $\mathbf{x}$ having a missing value in X_i, let J be the index set for all attributes with nonmissing values in $\mathbf{x}$, and $\mathbf{x}_J$ be the corresponding part of $\mathbf{x}$. Compute the posterior probabilities, based on Eqs. (1) and (2), as follows:
$$P(X_i = c_{ik}|\mathbf{x}_J) = \frac{1}{P(\mathbf{x}_J)} P(X_i = c_{ik}) \prod_{j \in J} P(X_j = c_{jr}|X_i = c_{ik}), \; k = 1, \ldots, L_i,$$
where it is not necessary to compute $P(\mathbf{x}_J)$ as it will be cancelled out when the posteriors are normalized.
4. Replace the missing X_i value in $\mathbf{x}$ based on probabilities computed in Step 3 with one of the two alternative methods below:
(a) MaxPost: using the value with the maximum posterior probability;
(b) PropPost: using a value that is selected with probability proportional to the estimated posterior distribution [Law and Kelton 1991, p. 497].

Table II. An Example Data

No.	Income	Age	Gender	HomeOwner
1	low	<30	female	no
2	low	<30	male	no
3	low	30-55	female	yes
4	low	30-55	female	no
5	low	>55	female	no
6	high	<30	male	yes
7	high	30-55	female	yes
8	high	30-55	male	yes
9	high	30-55	male	yes
10	high	30-55	male	no
11	high	>55	male	yes
12	?	30-55	female	yes
13	?	30-55	female	yes
14	?	<30	female	?
15	?	?	male	no
16	?	?	male	no

3. AN ILLUSTRATIVE EXAMPLE

To illustrate how to use the proposed method, consider the example dataset in Table II, which contains 16 records with 4 attributes: Income, Age, Gender, and HomeOwner (whether or not the person owns a home). There are eight missing values: 5 in Income, 2 in Age, and 1 in HomeOwner. A missing value is indicated by a question mark.

Consider records 12 and 13, both having missing Income value, and the same values for Age, Gender, and HomeOwner: {30-55, female, yes}. We first compute the prior probabilities for Income (step 1 of the algorithm):

$$P(\text{low}) = 5/11, \text{ and } P(\text{high}) = 6/11,$$

Table III. Posterior Probabilities for Missing Items

No.	Income		Age			Gender	HomeOwner	
	low	high	<30	30-55	>55		yes	no
12	0.3655	0.6345	-	-	-	-	-	-
13	0.3655	0.6345	-	-	-	-	-	-
14	0.9057	0.0943	-	-	-	-	0.2941	0.7059
15	0.4898	0.5102	0.5161	0.2903	0.1935	-	-	-
16	0.4898	0.5102	0.5161	0.2903	0.1935	-	-	-

where P(low) means P(Income = low), and so on. We then compute the conditional probabilities for {Age = 30-55}, given a certain Income level (step 2):

$$P(30-55|\text{low}) = 2/5, \text{ and } P(30-55|\text{high}) = 4/6.$$

Similarly, we can obtain the conditionals for {Gender = female} and {HomeOwner = yes}, given an Income level, as follows.

$$P(\text{female}|\text{low}) = 4/5, \text{ and } P(\text{female}|\text{high}) = 1/6$$
$$P(\text{yes}|\text{low}) = 1/5, \text{ and } P(\text{yes}|\text{high}) = 5/6$$

Finally, we compute the posterior probabilities (step 3):

$$P(\text{low}|30-55, \text{female}, \text{yes}) = \frac{1}{P}(5/11)(2/5)(4/5)(1/5) = 0.0291/P,$$

$$P(\text{high}|30-55, \text{female,yes}) = \frac{1}{P}(6/11)(4/6)(1/6)(5/6) = 0.0505/P,$$

where $P = P(30-55, \textit{female}, \textit{yes})$; this is not calculated as it will be cancelled out when we normalize the posteriors as follows.

$$P(\text{low}|30-55, \text{female,yes}) = \frac{0.0291P}{0.0291P + 0.0505P} = 0.3655$$

$$P(\text{high}|30-55, \text{female,yes}) = \frac{0.0505P}{0.0291P + 0.0505P} = 0.6345$$

Posteriors for the other missing values can be computed similarly. The results of the computation are shown in Table III.

There are two alternatives for replacing missing values in Step 4 of the algorithm. If MaxPost is used, the missing HomeOwner value will be replaced with "no"; the two missing Age values will both be replaced with "<30". The missing Income value for record 14 will be replaced with "low" and all remaining missing Income values will be replaced with "high". If PropPost is used, the replacing values will be selected with probability proportional to the posterior distribution. For records 15 and 16, for instance, the posterior probabilities for the Income values are 0.4898 and 0.5102 for "low" and "high" respectively. So it is more likely that one record will have a "low" value and the other a "high"

Table IV. Replaced Missing Values

	Income		Age		HomeOwner	
No.	PropPost	MaxPost	PropPost	MaxPost	PropPost	MaxPost
12	low	high	-	-	-	-
13	high	high	-	-	-	-
14	low	low	-	-	no	no
15	low	high	<30	<30	-	-
16	high	high	30-55	<30	-	-

value. The replaced missing values based on these two methods are provided in Table IV, where the values for PropPost represent just one possible senario, and they may vary for different random numbers.

4. EXPERIMENTAL EVALUATION

In order to evaluate the effectiveness of the proposed approach, we conducted an experimental study that compares it with two existing methods: The first is the mode substitution method, which is widely used in practice for handling missing categorical data; the second is the Bayesian method proposed by Chen and Astebro [2003]. Like the proposed approach, both methods are intended for use in a general setting (i.e., not related to a specific task).

Three datasets were used in the experiment; all of them were taken from the UCI Machine Learning Repository [Asuncion and Newman 2007]. The first dataset, called Adult, was originally extracted from the U.S. Census Bureau databases. It contains 48,842 individual records, each with 15 attributes (9 categorical and 6 numeric). These attributes provide an individual's demographic information such as age, gender, race, education, occupation, marital status, income, and so on. The second dataset, Credit, consists of 1,000 customer records and 21 attributes (14 categorical and 7 numeric), representing credit rating, age, gender, marital status, length of employment, housing status, and a number of attributes related to the customer's account. The third dataset, Cancer, contains 286 patient records and 10 attributes (all categorical), including diagnostic results, age, and a set of physical condition and medical test-related attributes.

Since the purpose of the experiment is to compare the methods for handling categorical missing data, we removed the numeric attributes in the Adult and Credit data. Originally, there are some missing values in the Adult (1.47%) and Cancer (0.31%) data and no missing value in the Credit data. It is difficult to use the original missing values to assess the performance of a missing value treatment method, since their true values are unknown. For the evaluation purposes, we randomly set additional 25% of the total values in each dataset to missing values. With their true values known, it is then easier to examine the effectiveness of the proposed approach. Given that each dataset in the experiment has at least nine attributes, setting 25% of the values to missing implies that, on average, each record has at least two missing attribute values. If a listwise deletion method is used for this setting, then most, if not all, of the records will be deleted.

In many real-world applications, data is not missing at random. For example, individuals with very high income are less likely to provide their income information than those with normal income. Likewise, a survey question related to a person's criminal history will more likely be left unanswered for those whose criminal record is not clean. To examine the effectiveness of the proposed method in these circumstances, we employed a different mechanism in setting missing values. In this setup, the random selection of missing values was applied only to those categories with an odd-numbered index. Here the index number of a category is determined by the order of its first appearance in the dataset (e.g., the odd-numbered categories for Age in Table II will be "<30" and ">55"). The amount of data set to missing remained the same (25% of the total). This setup is called "missing not at random," while the first setup is called "missing at random."

Data quality can be measured in a number of different dimensions, as described in Pipino et al. [2002], and Zhu and Wang [2008]. A straightforward measure is to calculate the number (or percentage) of replaced values that correctly match the original values. This measure is suitable for evaluating methods that assign a fixed value to replace a missing value (e.g., Mode and MaxPost). However, the measure is not appropriate for methods like PropPost or Dirichlet, where the replacement value is assigned based on a probability distribution. For example, we demonstrated in Section 3 that for records 15 and 16, there is an almost equal chance that the Income value will be "low" or "high." Assume that PropPost assigns a "low" to record 15 and a "high" to record 16. The true values of records 15 and 16 may be the opposite, namely, a "high" for record 15 and a "low" for record 16. In this situation, PropPost is "wrong" in both cases. However, when the dataset is used for statistical analysis and data mining tasks such as classification and association rules mining, which are the purposes of this study, the results will be the same no matter which of record 15 or 16 gets a "low" or "high". Due to this consideration, distribution-based measures are normally used in the literature (see, for example, Chiu and Sedransk [1986], and Chen and Astebro [2003]).

We selected two important tasks for performance evaluation: (i) univariate summary statistics, which are used in most data-warehousing applications; (ii) association rules, which comprise a popular data-mining task that deals with categorical data. The most relevant univariate summary statistics for categorical data is the categorical frequency (count) distributions of the individual attributes. In this aspect, we use a measure called Root Mean Square Error (RMSE), based on a similar measure used in Chen and Astebro [2003], defined as

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{M}\sum_{k=1}^{L_i}(\tilde{m}_{ik} - m_{ik})^2}, \tag{3}$$

where m_{ik} is the number of values being set to missing that originally belong to the kth category of the ith attribute, $\tilde{m}_{ik}$ is the corresponding number calculated using the replaced values, and m is the total number of values set to missing. The other symbols have been described in Section 2. This statistic

Table V. Experimental Results

Dataset	Method	Missing At Random		Missing Not At Random	
		RMSE[1]	Error in Large Itemset (%)[2]	RMSE[1]	Error in Large Itemset (%)[2]
Adult	Mode	60.57	39.09	73.55	45.79
	Dirichlet[3,4]	1.22	10.76	43.43	28.07
	PropPost[4]	4.51	3.98	28.30	15.87
	MaxPost	16.48	15.94	26.94	21.01
Credit	Mode	8.77	44.99	9.53	33.92
	Dirichlet[3,4]	1.12	5.87	6.50	23.51
	PropPost[4]	1.09	5.38	5.69	20.19
	MaxPost	3.33	12.00	4.93	15.75
Cancer	Mode	5.42	39.72	8.52	55.46
	Dirichlet[3,4]	1.10	12.47	5.91	35.39
	PropPost[4]	1.13	9.81	4.99	30.79
	MaxPost	1.42	18.41	5.52	33.89

1. Errors calculated based on Eq. (3).
2. Error rates calculated based on Eq. (4).
3. The method proposed by Chen and Astebro [2003], which uses a Dirichlet distribution.
4. The results of the Dirichlet and PropPost methods vary slightly with different random number seeds. Therefore, these two algorithms were run five times, each run using a different seed. The average results are reported in the table.

measures the closeness in univariate frequency distribution between the sets of original and replaced values. A smaller RMSE indicates better data quality.

For association rules mining, we use a performance measure related to large itemsets. An itemset is a set of attribute values that appear together in a record. An itemset is said to be large if its support (percentage of the records in the dataset containing the itemset) is greater than a prespecified value. The measure we use is based on Evfimievski et al. [2002], and Rizvi and Haritsa [2002], defined as

$$\text{Error rate in large itemset count} = \frac{1}{|T|}\sum_{t=1}^{|T|}\frac{|\tilde{Q}_t - Q_t|}{Q_t}, \tag{4}$$

where T represents the set of all large itemsets with support count larger than a specified value, Q_t is the frequency count of the tth large itemset from the original data, and $\tilde{Q}_t$ is the count for the same itemset from the data with 25% of the values set to missing and replaced. This measure indicates the closeness in large itemset count between the two sets. Again, a small value in this measure is desirable. A key difference between the measures in Eqs. (3) and (4) is that the former is univariate while the latter is multivariate. For each dataset, a sufficient number of large itemsets ($|T|$ ranged between 400 and 1,250) were identified to compute the error rate.

The experimental results are shown in Table V. When data is missing at random, Dirichlet (the method proposed by Chen and Astebro [2003]) leads in the univariate RMSE measure for the Adult and Cancer data, while PropPost leads for the Credit data. In terms of the multivariate itemset measure, however, PropPost is the best for all three datasets. MaxPost, although much

better than Mode, is not as good as Dirichlet and PropPost in either measure. This is understandable because MaxPost always uses the value with the maximum posterior probability for replacement, which will cause this value to be overweighted when categorical values are missing completely at random.

When data is missing not at random, both of the proposed methods perform favorably over the existing methods. MaxPost is the best (and PropPost is the second best) in the RMSE measure for the Adult and Credit data, while PropPost is the best (and MaxPost is the second best) for the Cancer data. In terms of the itemset measure, PropPost is the best (and MaxPost is the second best) for the Adult and Cancer data, while MaxPost is the best (and PropPost is the second best) for the Credit data. In general, both PropPost and MaxPost perform relatively better than the two existing methods when data is missing not at random.

5. LIMITATIONS AND EXTENSIONS

We have presented a new Bayesian approach for estimating and replacing missing data. The proposed method requires nonmissing values used for estimation to be categorical. One way to extend the method to numeric domain is to group the numeric data into categories. For example, the Age attribute in Table II might be the result of such grouping. Numerous grouping/discretization methods exist [Witten and Frank 2005], which attempt to preserve the dependence between attributes after grouping. When the attributes in data are overwhelmingly numeric, however, grouping of numeric values into categories may cause considerable loss of information. A more appropriate approach is then to estimate probabilities using a density function obtained by making an assumption about the underlying distribution (e.g., Gaussian distribution), or by nonparametric approach such as kernel density estimation [Witten and Frank 2005].

The proposed approach may also be limited by the conditional independence assumption made in the simple Bayes method. The issue has been investigated in a number of studies [Witten and Frank 2005; Clark and Niblett 1989; Li and Sarkar 2006]. The results of the studies indicated that, despite seemingly very restrictive assumption, the simple Bayes method performed surprisingly well. A possible explanation is that the assumption about conditional independence is not as restrictive as it appears to be. In general, for instance, there may be a strong dependency between age and home owner status. Given a certain income level, however, the dependency between age and home owner status will likely be much weaker. As a result, the simple Bayes approach can work well in many situations.

This article focuses on the issue of estimating and replacing missing values to maintain the relationships between attributes and the statistical properties of the data. We do not consider problems that involve examination of relationships between individual records. These problems involve different objectives and assumptions, which likely result in different approaches and preferences. One typical example of such problems is record linkage. When the proposed methods are applied to the record linkage problem, we would expect MaxPost

to perform better than PropPost because the primary objective for record linkage is to correctly match the individual values, rather than maintaining statistical distributions. It is interesting to investigate how to extend our approach to such individual-level problems in future study.

REFERENCES

ASUNCION, A. AND NEWMAN, D. J. 2007. *UCI Machine Learning Repository*. School of Information and Computer Science, University of California, Irvine, CA. http://www.ics.uci.edu/~mlearn/MLRepository.html.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A., AND STONE, C. J. 1984. *Classification and Regression Trees*. Wadsworth, Belmont, CA.

CHEN, G. AND ASTEBRO, T. 2003. How to deal with missing categorical data: Test of a simple Bayesian method. *Organ. Res. Methods 6*, 3, 309–327.

CHIU, H. Y. AND SEDRANSK, J. 1986. A Bayesian procedure for imputing missing values in sample surveys. *J. Amer. Statist. Assoc. 81*, 3905, 5667–5676.

CLARK, P. AND NIBLETT, T. 1989. The CN2 induction algorithm. *Mach. Learn. 3*, 4, 261–283.

CODD, E. F. 1979. Extending the database relational model to capture more meaning. *ACM Trans. Database Syst. 4*, 4, 397–434.

CONGDON, P. 2005. *Bayesian Models for Categorical Data*. John Wiley & Sons, New York.

DUDA, R. O., HART, P. E., AND STORK, D. G. 2001. *Pattern Classification*. John Wiley & Sons, New York.

EVFIMIEVSKI, A., SRIKANT, R., AGRAWAL, R., AND GEHRKE, J. 2002. Privacy preserving mining of association rules. In *Proceedings of 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 217–228.

FAN, W., LU, H., MADNICK, S. E., AND CHEUNG, D. 2002. DIRECT: A system for mining data value conversion rules from disparate data sources. *Decis. Support Syst. 34*, 1, 19–39.

FUNG, R. AND DEL FAVERO, B. 1995. Applying Bayesian networks to information retrieval. *Commun. ACM 38*, 5, 42–57.

JIANG, Z., SARKAR, S., DE, P., AND DEY, D. 2007. A framework for reconciling attribute values from multiple data sources. *Manag. Sci. 53*, 12, 1946–1963.

LAW, A. M. AND KELTON, W. D. 1991. *Simulation Modeling and Analysis*. McGraw-Hill, New York.

LI, X.-B. AND SARKAR, S. 2006. Privacy protection in data mining: A perturbation approach for categorical data. *Inf. Syst. Res. 17*, 3, 254–270.

MICHIE, D., SPIEGELHALTER, D. J., AND TAYLOR, C. C., Eds. 1994. *Machine Learning, Neural, and Statistical Classification*. Ellis Horwood, New York.

PIPINO, L. L., LEE, Y. W., AND WANG, R. Y. 2002. Data quality assessment. *Commun. ACM 45*, 4, 211–218.

PYLE, D. 1999. *Data Preparation for Data Mining*. Morgan Kaufmann, San Mateo, CA.

QUINLAN, J. R. 1989. Unknown attribute values in induction. In *Proceedings of the 6th International Workshop on Machine Learning*. Morgan Kaufmann, San Mateo, CA, 164–168.

QUINLAN, J. R. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.

RIZVI, S. J. AND HARITSA, J. R. 2002. Maintaining data privacy in association rule mining. In *Proceedings of the 28th Very Large Data Base Conference*.

SAS INSTITUTE, INC. 1990. *SAS Procedure Guide*. SAS Institute Inc., Cary, NC.

WITTEN, I. H. AND FRANK, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann of Elsevier, San Francisco, CA.

ZHU, H. AND WANG, R. 2008. An information quality framework for verifiable intelligence products. In *Data Engineering: Mining, Information, and Intelligence*. Y. Chan et al., Eds. Springer, New York. to appear.