A New Imputation Method for Incomplete Binary Data

Munevver Mine Subasi[*]
Department of Mathematical Sciences
Florida Institute of Technology
150 W. University Blvd., Melbourne, FL 32901 USA

Ersoy Subasi[†]
RUTCOR, Rutgers Center for Operations Research
640 Bartholomew Road Piscataway, NJ 08854, USA

Martin Anthony[‡]
Department of Mathematics
London School of Economics and Political Sciences
Houghton Street, London WC2A 2AE, UK

Peter L. Hammer[§]
RUTCOR, Rutgers Center for Operations Research
640 Bartholomew Road Piscataway, NJ 08854, USA

Abstract

In data analysis problems where the data are represented by vectors of real numbers, it is often the case that some of the data points will have "missing values", meaning that one or more of the entries of the vector that describes the data point is not observed. In this paper, we propose a new approach to the imputation of missing binary values that employs a "similarity measure". We compare experimentally the performance of our technique with ones based on the usual Hamming distance measure and multiple imputation.

1 Introduction

In practical machine learning or data analysis problems in which the data to be analyzed consists of vectors of real numbers, it is often the case that some of the data points will have "missing values", meaning that one or more of the entries of the vector that describes the data point is not known. It is natural to try to "fill in" or *impute* these missing values so that one than has complete data to work from. This may be necessary, for instance, so that the data can be used to learn from using statistical or machine learning techniques. This is a classical statistical and machine learning problem and many techniques have been employed.

Since in real-life applications missing data are a nuisance rather than the primary focus, an imputation method with good properties can be preferable to one that is complicated to implement and more efficient, but problem-specific.

Some approached to handling missing data simply ignore or delete points that are incomplete. Classical approaches of this type are list-wise deletion (LD) and pairwise deletion (PD). Because of their simplicity, they are widely used (see, e.g., (Roth 1994)) and tend to be the default for most statistical packages. However, the application of these techniques may lead to a large loss of observations, which may result in data-sets that are too small if the fraction of missing values is high, and particularly if the original data-set is itself small.

One of the most challenging decisions confronting researchers is choosing the most appropriate method to handle missing data during analysis. Little and Rubin (Little and Rubin 1987) suggests that naive or unprincipled imputation methods may create more problems than they solve. The most common data imputation techniques are mean imputation also referred to as unconditional mean imputation, regression imputation (RI) also referred to as conditional mean imputation, hot-deck imputation (HDI) and multiple imputation (MI). We remark that the mean imputation and similar approaches are not proper in the sense of Rubin (Rubin 1987) and hence, are not recommended. In most situations, simple techniques for handling missing data (such as complete case analysis methods LD and PD, overall MI, and the missing-indicator method) produce biased results as documented in (Greenland 1995; Little 1992; Rubin 1987; Schafer 1997; Vach 1994). A more sophisticated technique MI gives much better results (Greenland 1995; Little 1992; Rubin 1987; Schafer 1997; Vach 1994).

MI (Rubin 1987) is a statistical technique in which each missing value is replaced by several (k) values, producing k completed data-sets for analysis. The differences between these data sets reflect the uncertainty of the missing values. Each imputed data set is analyzed by standard complete-data procedures, which ignore the distinction between real and imputed values. The k results are then combined in such a way that the variability due to imputation may be incorporated. When properly done, the results of these combined analyses not only yield unbiased estimators for parameters, but adequately incorporate the uncertainty involved because of the missing data, i.e., produce valid estimates of the variances of these parameter estimates. Rubin (Rubin 1987) gave a comprehensive treatment of MI and addressed potential uses of the technique primarily for large public-use data files from sample surveys and censuses. The technique is available in standard statistical packages such as SAS, Stata and S-Plus. It has become increasingly attractive for researchers in the biomedical, behavioral, and social sciences where missing data is a common problem. These methods are documented in the book by Schafer (Schafer 1997) on incomplete multivariate data.

In fully parametric models, maximum-likelihood estimates can often be calculated directly from the incomplete data by specialized numerical methods, such as the Expectation-Maximization (EM) algorithm (Dempster and Rubin 1977; McLachlan and Krishnan 1996). The EM algo-

[*]Email: msubasi@fit.edu
[†]Email: esub@rutcor.rutgers.edu
[‡]Email: m.anthony@lse.ac.uk
[§]Deceased.

rithm is an iterative procedure in which it uses other variables to impute a value (Expectation), then checks whether that is the value most likely (Maximization). If not, it reimputes a more likely value. This goes on until it reaches the most likely value. Those procedures may be somewhat more efficient than MI because they involve no simulation. EM Imputation is available in SAS, Stata, R, and SPSS Missing Values Analysis module.

Imputation techniques have become easier to perform with the advent of several software packages. However, imputation of missing binary data is still an important practical problem. Ibrahim (Ibrahim 1990) showed that, under the assumption that the missing data are missing at random, the E step of the EM algorithm for any generalized linear model can be expressed as a weighted complete data log-likelihood when the unobserved covariates are assumed to come from a discrete distribution with finite range. Ibrahim's method of weights (Ibrahim 1990; Ibrahim and Weisberg 1992; Ibrahim and Lipsitz 1999; Ibrahim and Chen 1999; Ibrahim and Lipsitz 2001; Herring and Ibrahim 2001) can be used as a principled approach for imputation of binary data.

In this paper, we propose a new approach to the imputation of missing binary values. The technique we introduce employs a "similarity measure" introduced in (Anthony and Hammer 2006). The Boolean similarity measure has already proven to be of some application in classification problems (Subasi and Hammer 2009). Here, we use it to help indicate whether a missing value should be 0 or 1, and we compare experimentally the performance of our technique with ones based on the usual Hamming distance measure and MI technique using SAS (Inc. 2002 2004).

The framework used here requires data to be represented by binary vectors. However, in many applications, the raw data that we work with in a particular situation might be more naturally encoded as a real-valued vector. In such cases, the data may be transformed into binary data through a process known as *binarization* (see, for example, (Boros and Kogan 1997)). The transformed data-set may then be simplified or cleaned in a variety of ways, by the removal of repeated points, for instance, and the deletion of attributes (or co-ordinates) found to be statistically insignificant in determining the classification.

Section 2 provides details of the Boolean similarity measure that is at the core of our technique and describes the imputation method that derives from this measure. Section 3 describes the experiments we performed in order to test this method, and the results are reported in Section 4.

2 A measure of similarity and its application to missing values

In (Anthony and Hammer 2006), a way of measuring the similarity $s(x, A)$ of a Boolean vector x to a set A of such vectors is proposed. The measure can be described in two ways: either in terms of Boolean functions or, in a combinatorial way, in terms of substrings.

2.1 A Boolean function description

Any Boolean function $f : \{0,1\}^n \to \{0,1\}$ can be expressed by a *disjunctive normal formula* (or DNF), using *literals* $u_1, u_2, \ldots, u_n, \bar{u}_1, \ldots, \bar{u}_n$, where the $\bar{u}_i$ are known as *negated literals*. A disjunctive normal formula is one of the form

$$T_1 \vee T_2 \vee \cdots \vee T_k,$$

where each T_l is a *term* of the form

$$T_l = \left(\bigwedge_{i \in P} u_i \right) \bigwedge \left(\bigwedge_{j \in N} \bar{u}_j \right),$$

for some disjoint subsets P, N of $\{1, 2, \ldots, n\}$. A Boolean function is said to be a k-DNF if it has a disjunctive normal formula in which, for each term, the number of literals ($|P \cup N|$) is at most k. Such a function is said to be an l-term k-DNF if, additionally, it has a k-DNF formula in which the number of terms is at most l. For two Boolean functions f and g, we write $f \leq g$ if $f(x) \leq g(x)$ for all x; that is, if $f(x) = 1$ implies $g(x) = 1$. Similarly, for two Boolean formulae ϕ, ψ, we shall write $\phi \leq \psi$ if, when f and g are the functions represented by ϕ and ψ, then $f \leq g$. A term T of a DNF is said to *absorb* another term T' if $T' \leq T$. A term T is an *implicant* of f if $T \leq f$; in other words, if T true implies f true. The terms in any DNF representation of a function f are implicants of f. The most important type of implicants are the *prime implicants*. These are implicants with the additional property that there is no other implicant of f absorbing T. Thus, a term is a prime implicant of f if it is an implicant, and if the deletion of any literal from T results in a non-implicant T' of f (meaning that there is some x such that $T'(x) = 1$ but $f(x) = 0$). If we form the disjunction of all prime implicants of f, we have a DNF representation of f.

We now give the definition of *similarity* that is the focus of this paper.

Definition 2.1. *Suppose that A is a given subset of $\{0,1\}^n$. Let F be the Boolean function whose value is 1 in every point not in A, and 0 in every point of A, so F is the indicator function of $\bar{A}$, the complement of A. Let G_k be the disjunction of all prime implicants of length k of the function F. Then we define the similarity of $x \in \{0,1\}^n$ to A, denoted $s(x, A)$, to be the smallest k for which $G_k(x) = 1$.*

This is a slightly different description from that given in (Anthony and Hammer 2006), but it is equivalent to the formulation given there.

2.2 A combinatorial description

There is another useful way of describing the similarity measure. Suppose $x \in \{0,1\}^n$, $I \subseteq [n] = \{1, 2, \ldots, n\}$, and $|I| = k$. Then the projection of x onto I is the k-vector obtained from x by considering only the coordinates in I. For example, if $n = 5$, $I = \{2, 4\}$ and $x = 01001$ then $x|_I = 10$.

By a *positional substring* of $x \in \{0,1\}^n$, we mean a pair (z, I) where $z = x|_I$. The key point here is that the coordinates in I are specified: we will want, as part of our

later definitions, to indicate that two vectors x and y have the same entries *in exactly the same places*, as specified by some $I \subseteq [n]$. For instance, although both $x = 10101$ and $y = 01010$ have substrings equal to 00, there is no I such that $x|_I = y|_I = 00$.

We can now give an equivalent definition of similarity, from (Anthony and Hammer 2006).

Definition 2.2. *For $A \subseteq \{0,1\}^n$ and $x \in \{0,1\}^n$, the* similarity *of x to A, $s(x, A)$, is defined to be the largest s such that every positional substring (x, I) of length s appears also as a positional substring (y, I) of some $y \in A$. That is,*

$$s(x, A) = \max\{s : \forall I \subseteq [n], |I| \leq s, \exists y \in A, y|_I = x|_I\}.$$

Here $x|_I$ denotes the projection of x onto the coordinates indicated by I.

Equivalently, if r is the smallest length of a positional substring possessed by x that does *not* appear (in the same positions) anywhere in A, then $s(x, A) = r - 1$.

Notice that $s(x, A)$ is a measure of how similar x is to a *set* of vectors. It is not a metric or distance function. It can immediately be seen, indeed, that if A consists solely of one vector y, not equal to x, then $s(x, A) = 0$, since there must be some coordinate on which x and y differ (and hence a positional substring of length 1 of x that is absent from A).

Informally, then, the similarity of x to A is low if x has a short positional substring absent from A; and the similarity is high if all positional substrings of x of a fairly large length can be found in the same positions in some member y of A.

2.3 Example

Suppose the set A consists of the following 10 points of $\{0,1\}^5$.

$$\begin{array}{ccccc} 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \end{array}$$

Note, first, that no x can have $s(x, A) = 0$, since this could only happen if, on one of the five coordinates, all elements of A had a fixed value, either 0 or 1. Consider any x of the form $x = 01x_3x_4x_5$. Since there is no $y \in A$ with $y|_{\{1,2\}} = x|_{\{1,2\}} = 01$, we have $s(x, A) = 1$. Consider, however, $x = 10101$. For this x, we have $s(x, A) = 3$, because all (positional) substrings of x of length 3 belong to A, but there is no $y \in A$ such that $y|_{\{1,2,4,5\}} = x|_{\{1,2,4,5\}} = 1001$. Suppose now that $x = 00001$. Then, since all (positional) substrings of x of length 2 appear in A, $s(x, A) \geq 2$. However, there are substrings of length 3 missing from A: for example, there is no $y \in A$ with $y|_{\{1,3,4\}} = x|_{\{1,3,4\}} = 000$. So $s(x, A) = 2$.

2.4 A new technique for imputing missing values

In general terms, a natural approach to determining a missing value in a data-point is to find the value which will make the data-point appear to be most in line with the rest of the data-set. For example, suppose the first value of a data-point is missing. Then we may think it sensible to replace it by the average of all the known first values of all the other points in the data set. Note that when the data consists of binary vectors, then the counterpart to this technique is the MAJORITY method, in which the missing entry is replaced by the value, 0 or 1, that is most commonly taken in that same position in the rest of the data.

The general approach we propose here applies to binary data, and is as follows: suppose $x \in \{0,1\}^n$ belongs to a data-set A and that x has one or more missing values. Then the values we impute to those missing values are those that maximize the similarity of the resulting vector to the subset A^* of the data-set consisting of all the data-points with no missing values. That is, we impute the missing values in such a way that the resulting vector $\hat{x}$ minimizes $s(\hat{x}, A^*)$. If there is more than one assignment of values achieving the maximum similarity, the values can be assigned randomly among these.

For example, suppose that $x \in \{0,1\}^n$ is missing one value, in its first entry. Let us indicate this by writing $x =$ **?**x', where $x' \in \{0,1\}^{n-1}$. (The '**?**' symbol denotes the missing value.) If $s(1x', A^*) > s(0x', A^*)$ then we will take **?** $= 1$, so that x, once the missing value is imputed, becomes $1x'$. If $s(1x', A^*) < s(0x', A^*)$, then we will take **?** $= 0$. If $s(1x', A^*) = s(0x', A^*)$, then we will take **?** $= 1$ or 0, randomly (with equal probability).

3 Experiments

3.1 An overview of the experiments

To test the method, we conducted experiments based on real data-sets (binarized if necessary) in which single entries of some data-points are assumed to be missing. We compared the performance of the method with that of other methods. For each data-set D, and each of the imputation methods considered, we randomly selected 10% of the data-set, giving a subset D' of D.

Then, for each $x \in D'$ and each $i \in \{1, 2, \ldots, n\}$, supposing the ith entry of x (but no others) to be missing, we used the imputation technique to determine what that value should be, taking the data-set to be $(D \setminus D') \cup \{x\}$. This involves mn applications of the imputation technique, where n is the number of attributes of the data (that is, its dimension) and m the size of the selected set D'. In each case, we knew what the true value should have been, so we can know whether the method has worked correctly or not. So, explicitly, for the similarity-based method, for $x \in D'$, let us denote by $x^{(i)}$ observation x with its ith entry assumed to be missing. Let $x_1^{(i)}$ be x with ith entry taking value 1 (so this is the binary vector that agrees with x in all entries except possibly entry i, which is 1) and define $x_0^{(i)}$ similarly. Then, we determine the similarities $s_1 = s\left(x_1^{(i)}, D \setminus D'\right)$

and $s_0 = s\left(x_0^{(i)}, D \setminus D'\right)$. If $s_1 > s_0$, then we impute value 1 for the ith entry of x; if $s_1 < s_0$, then we impute value 0; and if $s_1 = s_0$, then the method fails to determine uniquely a value to impute (and in this case, one could instead simply choose randomly between 0 and 1).

For each of the methods we consider, and on each data-set, we determined three numbers that describe the performance of the method:

- the percentage of times the method determined (uniquely and unambiguously) the correct value;
- the percentage of times it did not determine uniquely a value to impute;
- the percentage of times the method did uniquely determine a value, but the value determined was incorrect.

We call the third of these three statistics the *error rate* of the method in this experiment.

In the second of the cases above, in which the value is not determined, one could chosen a value randomly. Of course, the resulting imputation will be incorrect some of the time, so it could be argued that the true error rate should take this into account. But what we want to assess is the extent to which the method fails when it *definitively* (unambiguously) makes a decision as to what the missing value should be. When the decision is ambiguous, we could, instead of choosing randomly, simply declare that the method has not determined the missing value, or we could invoke some other method. In applications where it is acceptable to have some points remaining incomplete, or where the penalty for imputing wrong values is severe, the appropriate error measure is then the one we propose.

In addition to the similarity-based method, which we'll denote SIM, we considered several others. One, which we call the Hamming method, denoted HAMM, uses Hamming distance as a guide to how similar a point is to the others. This method acts as follows: suppose $x \in \{0,1\}^n$ belongs to a data-set D and that x has one or more missing values. Then the values we impute to those missing values are those such that the resulting vector $\hat{x}$ will *minimize* the Hamming distance

$$d_H(\hat{x}, D^*) = \min\{\sum_{i=1}^{n} |\hat{x}_i - y_i| : y \in D^*\}$$

of $\hat{x}$ to the subset D^* of the data-set consisting of all the data-points with no missing values. (If there is more than one such possible $\hat{x}$ then one is chosen randomly from among these.)

So, in the context of our single-missing-value experiments, and with the notation as above, the Hamming method acts as follows: we determine the Hamming distances $d_1 = d_H\left(x_1^{(i)}, D \setminus D'\right)$ and $d_0 = d_H\left(x_0^{(i)}, D \setminus D'\right)$. If $d_1 < d_0$, then we impute value 1 for the ith entry of x; if $d_1 > d_0$, then we impute value 0; and if $d_1 = d_0$, then we choose randomly between 0 and 1.

Two methods can be derived by using the similarity measure and Hamming distance together.

First, SIM & HAMM acts as follows: first, SIM is applied. If that does not return a uniquely determined (that is, non-random) value for the missing value, then HAMM is then applied to determine the missing value. This subsequent application of HAMM may, of course, still not uniquely determine a value, in which case HAMM chooses randomly as its definition requires.

Secondly, HAMM & SIM first applies HAMM and, if that does not uniquely determine the missing value, then SIM is employed.

Another method is to apply MI technique (Rubin 1987) implemented in PROC MI (Yuan 2000) procedure in SAS (Inc. 2002 2004). This method always determines uniquely a value for the missing value, however the actual value is stochastic. In order to convert a multivariate normal imputed value into a binary imputed value we adapt a common approach that treats a binary variable as continuous in the imputation process and subsequently round each imputed value to the nearest category, i.e., to a "0" or a "1" (see e.g., Schafer (Schafer 1997), p. 148).

Finally, we also used the MAJORITY method in which the missing value imputed into $x^{(i)}$ is 1 (respectively, 0) if a majority of the elements of $D \setminus D'$ have a 1 (respectively, 0) is position i, and is chosen to be 0 or 1 at random if each occurs equally often.

3.2 The data-sets

In our experiments we used the following nine data-sets, taken from the UCI Machine Learning Repository (`http://www.ics.uci.edu/~mlearn/MLRepository.html`). The data-sets were pre-processed in several ways before we ran our experiments. First, any observations in the data-set that had any missing attribute values were deleted. Next, the data-sets were binarized, according to the method described in (Boros and Kogan 1997), so that any numerical or nominal attribute values were changed to binary values. Next, techniques from (Boros and Muchnik 2000) were used to determine that some attributes (of the binarized data) could be deemed irrelevant and therefore deleted. (Set covering was used to find a small "support set".) The binarized data was then projected onto the remaining binary attributes. If this process resulted in any repetitions, these were deleted, and if any of the processed observations appeared once with each class label, all its occurrences were deleted. After pre-processing in this manner, the data-sets consisted of binary vectors, generally in a higher-dimensional space than the original data. Table 1 describes the characteristics of the data-sets before and after this pre-processing.

4 Experimental results

In this section we present experimental results obtained by the use of imputation techniques described in Section 3.1. In each table we show the percentages of times the method determined (uniquely and unambiguously) the correct value (*correct determination*); the percentage of times it did not determine uniquely what the value should be (*non-determination*) and the percentage of times the method did

Dataset	# of observations		# of attributes		After preprocess		
					# of observations		# of binary
	Positive	Negative	Numeric	Nominal	Positive	Negative	attributes
Cleveland Heart Disease	139	164	10	3	137	158	63
Pima Indian Diabetes	130	262	8	0	130	262	47
German credit	700	300	7	13	697	300	66
Hepatitis	123	32	6	13	92	19	28
Ionosphere	225	126	34	0	216	125	49
Mushroom	3916	4208	0	22	2188	2047	50
Tic-Tac-Toe	626	332	0	9	626	332	27
Voting	267	168	16	0	96	64	16
Wisconsin Breast Cancer	458	241	9	0	203	182	48

Table 1: Nine data-sets from UCI Machine Learning Repository

	Determination rate	Correct determination rate
SIM	75%	97%
SIM & HAMM	95%	91%
SAS	100%	89%
HAMM & SIM	95%	85%
HAMM	86%	84%
MAJORITY	100%	69%

Table 2: Overall Results

uniquely determine a value, but the value determined was incorrect (*incorrect determination; that is, error*).

For all data-sets Figures 1–6 show the rates of correct determination, non-determination and errors (incorrect determination) obtained by different imputation methods. Table 2 shows the average *determination rate* (by which we mean the percentage of times the missing value was determined, either correctly or incorrectly) over all nine data-sets, and the *correct determination rate* (by which we mean the proportion of determinations that are correct) for all imputation methods discussed in Section 3.1.

It appears from the results of the experiments that the similarity-based method achieves, on average, a lower error rate than the other methods. What this means is that when it is used to determine a missing value, if it does so uniquely, then it appears to perform relatively well. However, as Table 2 makes clear, the method also generally has a lower determination rate than the other methods. So there is, in a sense, a trade-off between determination rate and error rate. If it is important to be sure of imputed values, and also acceptable to have some missing values as yet undetermined, then the similarity-based approach looks like a useful one. In the experiments, the HAMM method achieved a higher determination rate, but a lower rate of correct determination, than the SIM method. When the SIM and HAMM methods are used in combination, a higher still determination rate is achieved, and the rate of correct determination appears to be as good as (and possibly better) than other methods (and better than the HAMM method used on its own).

Acknowledgements

M. Anthony's work is supported in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence. P.L. Hammer's work was partially supported by NSF Grant NSF-IIS-0312953 and NIH Grants NIH-002748-001 and NIH-HL-072771-01.

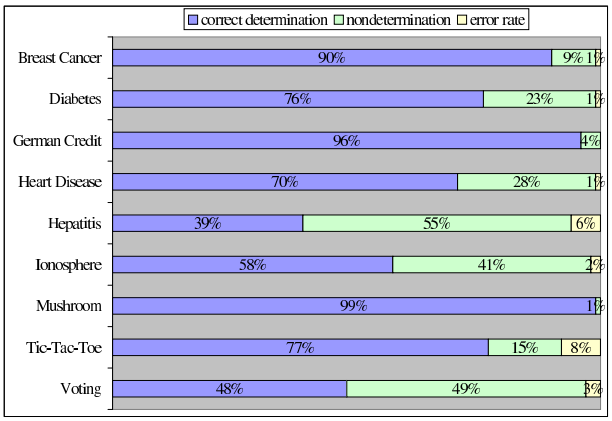

Figure 1: Imputation using similarity with SIM

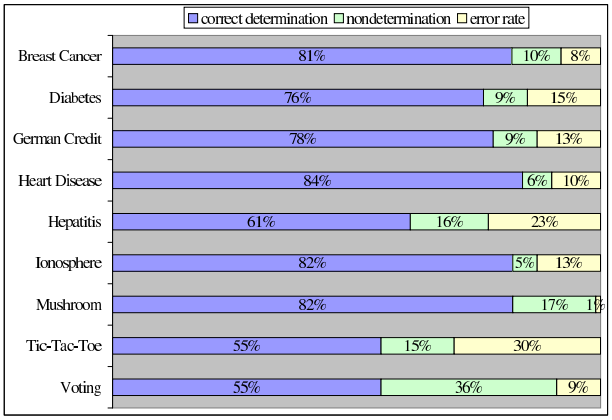

Figure 2: Imputation using Hamming distance with HAMM

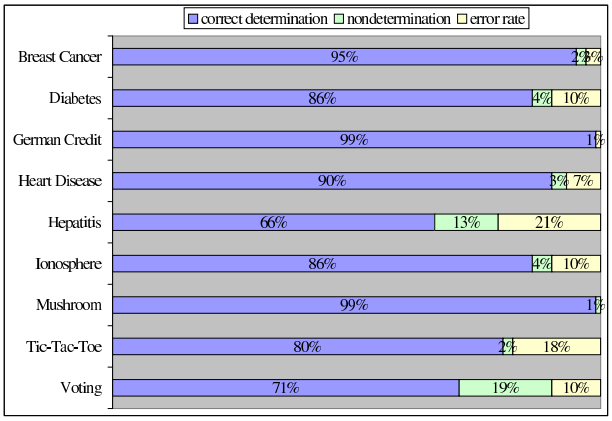

Figure 3: Imputation Combining similarity and Hamming distance: SIM & HAMM

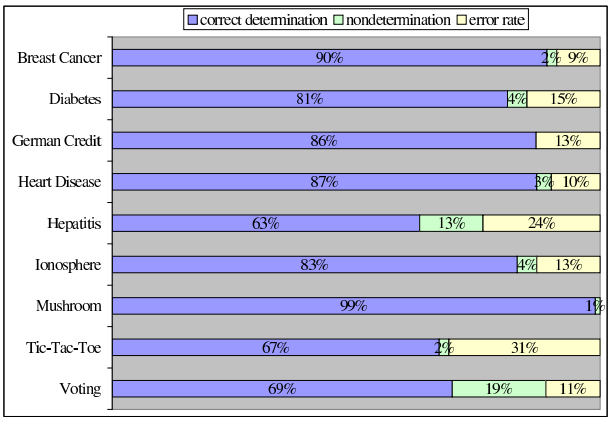

Figure 4: Imputation combining similarity and Hamming distance: HAMM & SIM

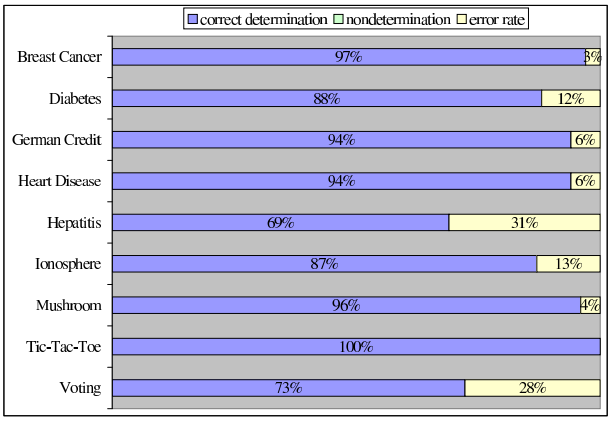

Figure 5: Imputation using MI

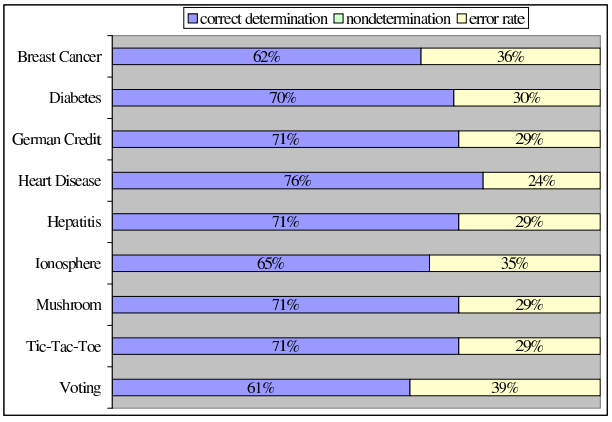

Figure 6: Imputation using MAJORITY

References

Anthony, M., and Hammer, P. 2006. A boolean measure of similarity. *Discrete Applied Mathematics* 154:2242–2246.

Boros, E., H. P. I. T., and Kogan, A. 1997. Logical analysis of numerical data. *Mathematical Programming* 79:163–190.

Boros, E., H. P. I. T. K. A. M. E., and Muchnik, I. 2000. Implementation of logical analysis of data. *IEEE Trans. on Knowledge and Data Engineering* 12:292–306.

Dempster, A.P., L. N., and Rubin, D. 1977. Maximum likelihood from incomplete data via the em algorithm. *J. of the Royal Stat. Society: Series B* 39:1–38.

Greenland, S.and Finkle, W. 1995. A critical look at methods for handling missing covariates in epidemiologic regression analyses. *Am J Epidemiol* 142:1255–64.

Herring, A., and Ibrahim, J. 2001. Likelihood-based methods for missing covariates in the cox proportional hazards model. *J. of the Am. Stat. Association* 96:292–302.

Ibrahim, J.G., L. S., and Chen, M. 1999. Missing covariates in generalized linear models when the missing data mechanism is nonignorable. *J. of the Royal Stat. Society, Series B* 61:173–190.

Ibrahim, J.G., C. M., and Lipsitz, S. 1999. Monte carlo em for missing covariates in parametric regression models. *Biometrics* 55:591–596.

Ibrahim, J.G., C. M., and Lipsitz, S. 2001. Missing responses in generalized linear mixed models when the missing data mechanism is nonignorable. *Biometrika* 88:551–564.

Ibrahim, J., and Weisberg, S. 1992. Incomplete data in generalized linear models with continuous covariates. *The Australian J. of Statistics* 34:461–470.

Ibrahim, J. 1990. Incomplete data in generalized linear models. *J. of the Am. Stat. Association* 85:765–769.

Inc., S. I. 2002–2004. *SAS 9.1.3 Help and Documentation*. Cary, NC: SAS Institute Inc.

Little, R., and Rubin, D. 1987. *Statistical Analysis with Missing Data*. New York: J. Wiley & Sons.

Little, R. 1992. Regression with missing xs; a review. *J. American Stat. Assoc.* 87:1227–37.

McLachlan, G., and Krishnan, T. 1996. *The EM Algorithm and Extensions*. New York: J. Wiley & Sons.

Roth, P. 1994. Missing data: a conceptual review for applied psychologists. *Psychology* 47:537–560.

Rubin, D. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: J. Wiley & Sons.

Schafer, J. 1997. *Analysis of incomplete multivariate data*. London: Chapman & Hall/CRC Press.

Subasi, M., S. E. A. M., and Hammer, P. 2009. Using a similarity measure for credible classification. *Discrete Applied Mathematics* 157:1104–1112.

Vach, W. 1994. *Logistic regression with missing values in the covariates*. New York: Springer.

Yuan, Y. 2000. Multiple imputation for missing data: Concepts and new development. *SUGI Proceedings* 267–25.