In *Intelligent Information Systems 2002: Proceedings of the IIS'2002 Symposium,* M. Klopotek, S.T. Wierzchon, M. Michalewicz (eds.), pages 351-360, *Advances in Soft Computing Series,* Physica-Verlag (A Springer-Verlag Company), Heidelberg, 2002

# An experimental comparison of methods for handling incomplete data in learning parameters of Bayesian networks

Agnieszka Oniśko[1], Marek J. Druzdzel[2], and Hanna Wasyluk[3]

[1] Faculty of Computer Science, Białystok University of Technology, ul. Wiejska 45-A, 15–351 Białystok, Poland, aonisko@ii.pb.bialystok.pl
[2] Decision Systems Laboratory, School of Information Sciences, Intelligent Systems Program, and Center for Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA 15260, USA, marek@sis.pitt.edu
[3] The Medical Center of Postgraduate Education, and Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Sciences, Marymoncka 99, 01–813 Warsaw, Poland, hwasyluk@cmkp.edu.pl

**Abstract.** Missing values of attributes in data sets, also referred to as incomplete data, pose difficulties in learning tasks, such as classification, data mining, or learning Bayesian network structure and its numerical parameters. Because of the predominance of incomplete data in practice, many methods have been proposed to deal with them while there are few studies that compare their performance. The HEPAR II project presents an excellent opportunity to test experimentally how these methods perform on a real data set. We briefly review several popular methods for handling incomplete data and then compare them on the task of learning conditional probability distributions of a Bayesian network model, where the comparison criterion is the resulting diagnostic accuracy. While substitution of "normal" values of missing attributes seemed to perform best, we observed only a small difference in performance among the studied methods.

## 1 Introduction

It is a fact of life that most practical databases of measurements or cases contain missing values of some of their attributes. There are many reasons for missing data. Sometimes they result from human errors of omission (e.g., a nurse forgetting to record the result of a measurement) sometimes the value of the attribute in question was not known (e.g., a patient forgetting whether or not she had chicken pox as a child). At other times, the value might have not made sense (e.g., presence or absence of pregnancy in a male patient). While the causes of missing values may be of interest in choosing how to handle them, the fact that a measurement is missing is uniformly a complication in any algorithm that analyzes the data.

Cowell *et. al* [3] define a database to be complete when all cases that it contains are complete. In turn, a case is complete if every random variable has a state or a value assigned to it. A database is incomplete, if it contains at

least one incomplete case. A case is incomplete, if one or more of the random variables has no value associated with it.

The data in an incomplete case can be missing, unobserved, or censored at random, but there may also be some structure, known or unknown, in why some values are missing. Little and Rubin [11,17] define three kinds of possible mechanisms that account for missing data. The first account is referred to as the missing at random (MAR) property. One way to formulate the MAR property is that while cases with incomplete data differ from cases with complete data, the pattern of data missingness is predictable from other variables in the database rather than being due to the specific variable on which the data is missing. The second mechanism is related to a situation when the data are missing completely at random (MCAR), i.e., when cases with complete data are indistinguishable from cases with incomplete data. The third type of missing data mechanism involves non-ignorable (NI) property, i.e., when the pattern of data missingness is not random and it is not predictable from other variables in the database. In case of medical data sets, both the MAR and the MCAR assumptions seem invalid. There are typically identifiable reasons why a measurement is missing.

Little and Rubin [11] offer an extensive review of various statistical approaches to handle missing data. The first group of methods involves listwise or casewise data deletion, pairwise data deletion, mean substitution, or hot deck imputation. There are also more sophisticated approaches involving regression methods, Expectation Maximization (EM) approach, raw maximum likelihood methods, or multiple imputation. All these methods require that the data meet the MAR assumption. For cases with non-ignorable mechanisms for missing data, a pattern-mixture model was developed [9,10,12].

Various approaches have been developed for learning parameters in probabilistic systems from incomplete data. These techniques include iterative methods like stochastic Gibbs Sampling [8], EM algorithm [5], and methods based on probability intervals, for example, deterministic method Bound and Collapse [16], or methods presented in [1,4]. Most of these methods assume usually the MAR property for all incomplete cases, however, Bound and Collapse algorithm proved to be robust also for NI data.

There seems to be little in terms of comparative studies that would test the proposed approaches in practical settings. Many approaches are typically tested on artificial data (or artificially introduced missing values to real world data, e.g., [16]). The Hepar II project and its underlying Hepar data set have provided us with an opportunity to test various approaches to handle missing data on a real data set. It has given us also a natural and fairly objective criterion for such a comparison — the quality of the resulting model. We test the diagnostic accuracy of the Hepar II model for various methods and present the results of experimental comparison.

The remainder of this paper is structured as follows. Section 2 describes briefly the Hepar data set and the Hepar II model. Section 3 reviews several

methods for handling incomplete data. Section 4 reports the results of an experimental comparison of selected methods that we tested on the HEPAR data set and the HEPAR II model. Finally, Section 5 discusses general issues related to the performed study and directions for further work.

## 2 The HEPAR data set and the HEPAR II model

Our work on the HEPAR II system is a continuation of the HEPAR project [2,18], conducted in the Institute of Biocybernetics and Biomedical Engineering of the Polish Academy of Sciences in collaboration with physicians at the Medical Center of Postgraduate Education in Warsaw. The HEPAR system was designed for gathering and processing of clinical data of patients with liver disorders and aimed at reducing the need for hepatic biopsy by modern computer-based diagnostic tools. An integral part of the HEPAR system is its database, created in 1990 and thoroughly maintained since then at the Gastroentorogical Clinic of the Institute of Food and Feeding in Warsaw. The current database contains over 800 patient records and its size is steadily growing. Each hepatological case is described by over 160 different medical findings, such as patient self-reported data, results of physical examination, laboratory tests, and finally a histopathologically verified diagnosis.

The version of the HEPAR data set, available to us, consisted of 699 patient records. The HEPAR data set contains many missing values. While there may be some randomly missing values that can be attributed to errors of omission, these are not very likely, as the data set is well maintained and utmost care is exercised in keeping it complete and correct. One of the main reasons for missing values is sheer economics. There are more than 40 variables that represent laboratory tests. It is obvious that not every patient will undergo all the possible tests since not all of them are relevant to a particular diagnostic situation. Also, performing a laboratory test is often expensive.

The HEPAR II project [13,14] aims at applying decision-theoretic techniques to diagnosis of liver disorders. Its main component is a Bayesian network model involving a subset of over 70 variables included in the HEPAR database. The model covers 11 different liver diseases and 61 feature nodes encoding medical findings such as patient self-reported data, signs, symptoms and laboratory tests results. The structure of the model, (i.e., the nodes of the graph along with arcs among them) was built based on medical literature and conversations with our domain expert, a hepatologist Dr. Hanna Wasyluk (third author of the current paper) and two American experts, a pathologist, Dr. Daniel Schwartz, and a specialist in infectious diseases, Dr. John N. Dowling, from the University of Pittsburgh. The elicitation of the structure took approximately 50 hours of interviews with the experts, of which roughly 40 hours were spent with Dr. Wasyluk and roughly 10 hours spent with Drs. Schwartz and Dowling. This includes model refinement sessions, where previously elicited structure was reevaluated in a group setting. The numerical

parameters of the model, i.e., the prior and conditional probability distributions, were learned from the HEPAR database. All continuous variables in the database were discretized by our expert.

Missing values in the HEPAR database have been a major problem in our work on the HEPAR II project. We counted that there were 7,792 missing values (15.9% of all entries!) in the learning data set. Figure 1 presents the cumulative distribution of the number of cases in the HEPAR data set as a function of the number of missing values per patient case. For example, there were 200 records in the HEPAR data set where each case had at most nine missing values. Please, note that there were no records that are complete.

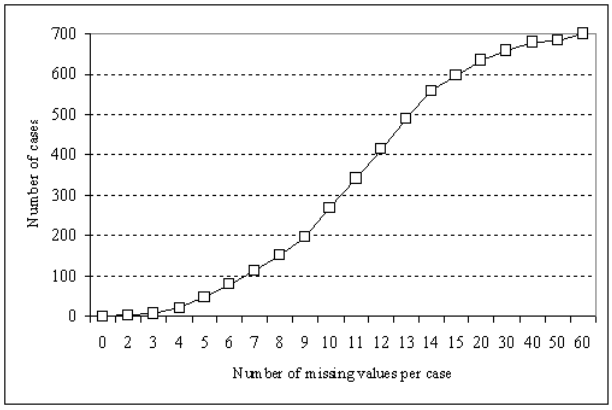


**Fig. 1.** Number of cases as a function of the number of missing values per case

We have tried several approaches to handle incomplete data when learning conditional probability distribution of the HEPAR II model, the choice of which was based on conversations with our expert and understanding that resulted from these.

## 3 Methods for handling missing data

The following section describes briefly several simple approaches to handling missing values in databases. We applied some of these methods in the course of the HEPAR II project.

### 3.1 Discarding records with missing data

The simplest way of dealing with missing data involves *listwise or casewise data deletion* approach. The method simply omits entire records if they have missing values for any of the variables. In those cases where only a small fraction of records contain missing values, this is a simple method that works

well. An underlying assumption in this approach is that the values are missing at random and, thus, discarding records with missing data will not bias the remaining data set. When many records contain missing values, this method becomes unreliable. For example, in the HEPAR data set, no single record contained all values. Application of this method would thus result in an empty data set.

### 3.2 Missing as an additional state

A simple approach to handling incomplete data is treating a missing value of an attribute as an additional state of the attribute, i.e., the missing measurements are interpreted as possible values of the variables in question. This interpretation requires some care when using the system. It is assumed namely that the fact that a measurement was not taken is meaningful — for example, in case of a medical database, the physician did not find taking the measurement appropriate. The meaning of the thus construed outcome *unmeasured* can be in this way equivalent to a measured value of the variable. This approach does not assume that data are missing at random.

### 3.3 Replacement by "normal" values

The third approach for handling missing values is based on the suggestions of Peot and Shachter [15] on the interpretation of missing values in medical data sets. They argued convincingly that data in medical data sets is not missing at random and that there are two important factors influencing the probability of reporting a finding. The first factor is a preference for reporting present symptoms over absent symptoms. The second factor is a preference for reporting more severe symptoms before those that are less severe. In other words, if a symptom is absent, there is a high chance that it is not reported, i.e., it is missing from the patient record. And conversely, a missing value suggests that the symptom was absent. Then, in learning the model parameters, missing values for discrete variables are assigned to state *absent* (e.g., a missing value for *Jaundice* is interpreted as *absent*). In case of continuous variables, a missing value is assigned as a typical value for a healthy patient elicited from the expert (e.g., a missing value for *Bilirubin* is interpreted as being in the range of 0–1 *mg/dl*). Similarly to the previous approach, this method assumes that the pattern of data missingness is not random.

### 3.4 Replacement by mean values

*Replacement by mean values* approach relates to filling in missing data values with a variable's mean that is computed from available cases. In case of discrete binary variables, a missing value is substituted by the outcome that occurs most frequently in the data.

### 3.5 Hot deck imputation

*Hot deck imputation* approach [6] examines the cases with complete records and identifies the most similar case to the case with a missing value. Then, it substitutes a missing value with the most similar case's variable value. More sophisticated hot deck algorithms identify more than one similar record and then randomly select one of those available donor records to impute the missing value or use an average value if that were appropriate.

### 3.6 K-NN techniques

The distance weighted k-Nearest Neighbor techniques (k-NN) [7] are widely used in many practical research problems. The k-NN techniques involve searching for k nearest neighbors of a given data point, i.e., in case of a medical data set, it would consist in looking for neighboring patient cases. One practical issue in applying k-NN approach is that the distance between instances is calculated based on all attributes of the instance. The k-NN approach can be also used in the imputation of incomplete data, where it involves imputation of missing values based on the neighboring patient records.

## 4 Experimental comparison of the approaches

Our experiments involved learning conditional probability distributions of the HEPAR II model from the HEPAR data set. We compared the diagnostic accuracy of HEPAR II for each of the methods dealing with missing data. In each case, the model had the same graphical structure elicited from the experts. In other words, the various approaches to deal with missing data had impact only on the numerical parameters of the model and not on its structure. The diagnostic accuracy was defined as the percentage of correct diagnoses and was determined by cross-validation using the leave-one-out method. When testing the diagnostic accuracy of HEPAR II, we were interested in both (1) whether the most probable diagnosis indicated by the model is indeed the correct diagnosis, and (2) whether the set of $w$ most probable diagnoses contains the correct diagnosis for small values of $w$ (we chose a "window" of $w$=1, 2, 3, and 4). The latter focus is of interest in diagnostic settings, where a decision support system only suggest possible diagnoses to a physician. The physician, who is the ultimate decision maker, may want to see several alternative diagnoses before focusing on one.

In our comparison we have taken into account the methods described in Sections 3.2 through 3.6. We could not include approaches that are based on discarding records with missing data (Section 3.1) because there was not even one complete record in our data set (see Figure 1). In case of the *replacement be mean values* approach, we noticed that most of mean values that were calculated for the variables representing laboratory tests were significantly

higher/lower than normal values. For example, a mean value for $AST$ was equal to $111U/L$[1] while the normal value for this finding is between 5 and $35U/L$. When analyzing the HEPAR data set, we found that there were only three binary variables, for which "present" value was the most frequent occurring value. In case of the *hot deck imputation* approach, we defined the most similar case as a case that has the highest number of similar or equal values for corresponding variables. We also employed the k-nearest neighbor method for $k = 1, 5$. We chose the Euclidean distance as a metric. For $k = 1$, we substituted missing values with the values of the nearest neighbor, in case of the five nearest neighbors, we have calculated mean values based on the neighboring cases and replaced with them missing values. Because the results were similar for both values of $k$, we present the results only for $k = 5$.

In addition, we have included the following three methods that played the function of the baseline, i.e., we expected that they would perform poorly.

#### Replacement by "abnormal" values

This method is the opposite of replacement by "normal" values described in Section 3.3 and involves replacing missing values with values that are considered "abnormal." Missing values for discrete variables are replaced by "present" value and for continuous variables are replaced by the values indicating most abnormal result (elicited from the expert).

#### Proportional random replacement

In this method we replaced missing values by a random drawing from the set of possible states of the variable. The probability of drawing a state was proportional to the probability of that state.

#### Replacement at random

In this method, we replaced missing values by a random drawing from the set of possible states of the variable. The probability of drawing a state was uniform, i.e., each state was equally likely to be drawn.

Table 1 captures the results of the diagnostic accuracy of HEPAR II for different approaches to handle missing values. The methods marked by an asterisk are also presented graphically in Figure 2.

## 5 Discussion

We tested the diagnostic accuracy of HEPAR II for several methods dealing with incomplete data. In each case, the model had the same graphical structure elicited from the experts. The accuracy for most of the methods that

[1] An abbreviation U/L stands for units/liter.

**Table 1.** The diagnostic accuracy of HEPAR II for different approaches to handle incomplete data for window size equal $w = 1, 2, 3, 4$

| **Approach** | **w=1** | **w=2** | **w=3** | **w=4** |
|---|---|---|---|---|
| Replacement by "normal" values* | 0.57 | 0.69 | 0.75 | 0.79 |
| Missing as an additional state* | 0.54 | 0.67 | 0.75 | 0.82 |
| Replacement by mean values* | 0.49 | 0.63 | 0.72 | 0.77 |
| Hot deck imputation | 0.51 | 0.64 | 0.72 | 0.77 |
| k-NN | 0.51 | 0.63 | 0.71 | 0.77 |
| Replacement by "abnormal" values* | 0.51 | 0.65 | 0.72 | 0.78 |
| Proportional random replacement | 0.52 | 0.66 | 0.74 | 0.79 |
| Replacement at random* | 0.49 | 0.64 | 0.72 | 0.78 |

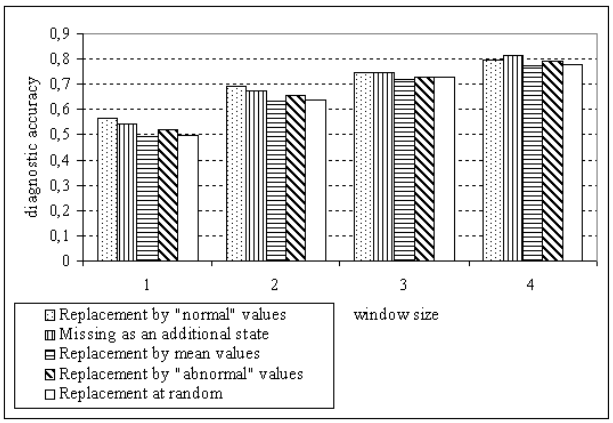


**Fig. 2.** The diagnostic accuracy of HEPAR II as a function of the window size for selected approaches to handling missing data.

we have tested was similar, with *replacement by "normal" values* and *missing as an additional state* performing slightly better than other approaches. It is interesting that while there are some performance differences between the methods, they are minimal. Even though the data set contained many incomplete values and one would expect even small performance differences to be amplified, this did not happen. It will be interesting to probe this issue further by performing tests on another real medical data set.

Our expert was able to predict a-priori which method would perform best on the data or, in other words, which of the assumptions was the most reasonable, even though the performance difference turned out to be minimal. Our advice to those knowledge engineers who encounter data sets with missing values is to reflect on the data and find out what the reasons are for missing values. In case of medical data sets, the assumption postulated by Peot and

Shachter [15] seems very reasonable. Even in this case, however, we advise to run it through the expert.

## Acknowledgments

Marek Druzdzel was supported by the Air Force Office of Scientific Research grant F49620–00–1–0112, Hanna Wasyluk was supported by Medical Center of Postgraduate Education grant 501–2–1–02–18/02, and by Institute of Biocybernetics and Biomedical Engineering PAS grant 16/ST/02. Our collaboration was enhanced by travel funds from the NATO Collaborative Linkage Grant PST.CLG.976167.

The HEPAR II model was created and tested using SMILE, an inference engine, and GeNIe, a development environment for reasoning in graphical probabilistic models, both developed at the Decision Systems Laboratory, University of Pittsburgh and available at `http://www2.sis.pitt.edu/~genie`.

## References

1. Silvia Acid, Luis M. de Campos, and Juan F. Huete. Estimating probability values from an incomplete dataset. *International Journal of Approximate Reasoning*, 27(2):183–204, 2001.
2. Leon Bobrowski. HEPAR: Computer system for diagnosis support and data analysis. Prace IBIB 31, Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Sciences, Warsaw, Poland, 1992.
3. Robert G. Cowell, A. Philip Dawid, Steffen L. Lauritzen, and David J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer Verlag, New York, 1999.
4. Luis M. de Campos, Juan F. Huete, and Serafin Moral. Probability Intervals: A tool for uncertain reasoning. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2:167–196, 1994.
5. A. Dempster, D. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39:1–38, 1977.
6. B. L. Ford. An overview of hot-deck procedures. In Rubin D. B. Madow W. G., Olkin I., editor, *Incomplete data in sample surveys*, pages 185–207. Academic Press, New York, 1983.
7. K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, New York, 1972.
8. S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.
9. D. Hedeker and R.D. Gibbons. Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods*, 2(1):64–78, 1997.
10. R.J.A. Little. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88:125–134, 1993.

11. R.J.A. Little and D. B. Rubin. *Statistical analysis with missing data*. John Wiley and Sons, New York, 1987.
12. R.J.A. Little and N. Schenker. Missing data. In C.C. Clogg G. Arminger and M.E. Sobel, editors, *Handbook for Statistical Modeling in the Social and Behavioral Sciences*, pages 39–75. New York Plenum, 1994.
13. Agnieszka Oniśko, Marek J. Druzdzel, and Hanna Wasyluk. Extension of the Hepar II model to multiple-disorder diagnosis. In S.T. Wierzchoń M. Kłopotek, M. Michalewicz, editor, *Intelligent Information Systems,* Advances in Soft Computing *Series*, pages 303–313, Heidelberg, 2000. Physica-Verlag (A Springer-Verlag Company).
14. Agnieszka Oniśko, Marek J. Druzdzel, and Hanna Wasyluk. Learning Bayesian network parameters from small data sets: Application of Noisy-OR gates. *International Journal of Approximate Reasoning*, 27(2):165–182, 2001.
15. Mark Peot and Ross Shachter. Learning from what you don't observe. In *Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI–98)*, pages 439–446, San Francisco, CA, 1998. Morgan Kaufmann Publishers.
16. Marco Ramoni and Paola Sebastiani. Learning conditional probabilities from incomplete data: An experimental comparison. In *Proceedings of the The Seventh International Workshop on Artificial Intelligence and Statistics*, pages 260–265, San Francisco, CA, 1999. Morgan Kaufmann Publishers, Inc.
17. D.B. Rubin. Inference and missing data. *Biometrika*, 63:581–592, 1976.
18. Hanna Wasyluk. The four year's experience with HEPAR-computer assisted diagnostic program. In *Proceedings of the Eighth World Congress on Medical Informatics (MEDINFO–95)*, pages 1033–1034, Vancouver, BC, July 23–27 1995.