KNOWLEDGE MEDIA INSTITUTE

Learning Conditional Probabilities from Incomplete Data: An Experimental Comparison

Marco Ramoni and Paola Sebastiani

KMI-TR-64

July 1998

Learning Conditional Probabilities from Incomplete Data: An Experimental Comparison

Marco Ramoni
Knowledge Media Institute
The Open University

Paola Sebastiani
Department of Actuarial Science and Statistics
City University

Abstract

This paper reports some experimental results comparing three parametric methods, Gibbs Sampling, EM algorithm and Bound and Collapse, for the estimation of conditional probability distributions from incomplete databases.

Keywords: Bayesian Learning, Missing Data, Gibbs Sampling, EM Algorithm, Bound and Collapse.

Reference: KMi Technical Report KMi-TR-64, July 1998.

Address: Marco Ramoni, Knowledge Media Institute, The Open University, Milton Keynes, United Kingdom MK7 6AA. PHONE: +44 (1908) 655721, FAX: +44 (1908) 653169, EMAIL: m.ramoni@open.ac.uk, URL: http://kmi.open.ac.uk/people/marco.

1. Introduction

A Bayesian Belief Network (BBN) is defined by a set of *variables* $\mathcal{X} = \{X_1, \ldots, X_I\}$ and a directed acyclic graph defining a model $\mathcal{M}$ of conditional dependencies among the elements of $\mathcal{X}$. A conditional dependency links a *child* variable X_i to a set of *parent* variables Π_i, and it is defined by the conditional distributions of X_i given the configurations of the parent variables. We consider discrete variables only and denote by c_i the number of states of X_i and $X_i = x_{ik}$ by x_{ik}. The structure $\mathcal{M}$ yields a factorization of the joint probability of a set of values $x_k = \{x_{1k}, \ldots, x_{Ik}\}$ of the variables in $\mathcal{X}$ as $p(\mathcal{X} = x_k) = \prod_{i=1}^{I} p(x_{ik}|\pi_{ij})$, where π_{ij} denotes one of the q_i states of Π_i in x_k. Suppose we are given a database of n cases $\mathcal{D} = \{x_1, \ldots, x_n\}$ and a graphical model $\mathcal{M}$ specifying the dependencies among the variables $\mathcal{X}$. Our task is to estimate, from $\mathcal{D}$, the conditional probabilities $\theta_{ijk} = p(x_{ik}|\pi_{ij}, \theta)$ defining the dependencies in the graph, where $\theta = (\theta_{ijk})$.

When the database is complete, closed form solutions allow efficient estimation of the conditional probabilities in time proportional to the size of the database. The Maximum Likelihood estimates (MLE) of θ_{ijk} are the values $\hat{\theta}_{ijk}$ that maximize the likelihood function $l(\theta) = \prod_{ijk} \theta_{ijk}^{n(x_{ik}|\pi_{ij})}$ where $n(x_{ik}|\pi_{ij})$ is the frequency of (x_{ik}, π_{ij}) in $\mathcal{D}$. Thus, $\hat{\theta}_{ijk} = n(x_{ik}|\pi_{ij})/n(\pi_{ij})$ which is the relative frequencies of relevant cases, and $n(\pi_{ij}) = \sum_k n(x_{ik}|\pi_{ij})$ is the frequency of π_{ij}. The Bayesian approach generalizes the MLE by introducing a flattening constant $\alpha_{ijk} > 0$ for each frequency, so that the estimate is computed as

$$\hat{\theta}_{ijk} = \frac{\alpha_{ijk} + n(x_{ik}|\pi_{ij}) - 1}{\alpha_{ij} + n(\pi_{ij}) - c_i} \tag{1}$$

where $\alpha_{ij} = \sum_k \alpha_{ijk}$. The quantity $(\alpha_{ijk} - 1)/(\alpha_{ij} - c_i)$ is interpreted as the *prior probability* believed by the learning system before seeing any data, and the learning process is regarded as the updating of this prior probability by means of Bayes' theorem. When the prior distribution of $(\theta_{ij1}..., \theta_{ijc_i})$ is Dirichlet, with *hyperparameters* $(\alpha_{ij1}, ..., \alpha_{ijc_1})$, then (1) is the posterior mode, the *Maximum a Posteriori* (MAP), of θ_{ijk}. Unfortunately, simplicity and efficiency of these closed form solutions are lost when the database is incomplete, that is, some entries are reported as unknown. In this case, the exact estimate is the mixture of the estimates given by (1) for each database generated by the combination of all possible values for each missing entry, and the computational cost of this operation would grow exponentially in the number of missing data.

Simple solutions to handle this problem are either to ignore the cases including missing entries or to ascribe the missing cases to an *ad hoc* dummy state. Both solutions introduce potentially dangerous bias in the estimate, as they ignore the fact that the missing entries may be relevant to the estimation of θ. Current methods rely on more sophisticated techniques, such as the EM algorithm [2] and Gibbs Sampling (GS) [3]. More recently, Ramoni and Sebastiani [5] introduced a deterministic method, called *Bound* and *Collapse* (BC) to

perform the estimation task. This paper presents an experimental comparison of EM, GS and BC on a real-world database.

2. Methods

This paper will compare the accuracy and the efficiency of EM, GS, and BC. This section outlines the character of these three methods.

The EM algorithm is an iterative method to compute MLEs and MAP. The EM algorithm alternates an expectation step, in which unknown quantities depending on the missing entries are replaced by their expectation in the likelihood, to a maximization step, in which the likelihood completed in the expectation step is maximized with respect to the unknown parameters, and the resulting estimates are used to replace unknown quantities in the next expectation step, until the difference between successive estimates is smaller than a fixed threshold. The convergence rate of this process can be "painfully slow" [6], and several modification have been proposed to increase its speed under certain circumstances.

GS is one of the most popular Markov Chain Monte Carlo methods for Bayesian inference. The algorithm treats each missing entry as a parameter to be estimated and iterates a stochastic process that provides a sample from the posterior distribution of θ. This sample is used to compute empirical estimates of the posterior mean, the posterior mode, or any other function of the parameters. In practical applications, the algorithm iterates a number of times to reach stability (*burn-in*) and then a final sample from the joint posterior distribution of the parameters is taken.

Both EM and GS provide reliable estimates of the parameters and they are currently regarded as the most viable solutions to the problem of missing data. However, both these iterative methods can be trapped into local minima and the convergence detection can be difficult. Furthermore, they rely on the assumption that the missing data mechanism is *ignorable*: within each observed parent configuration, the available data are a representative sample of the complete database and the distribution of missing data can be therefore inferred from the available entries [4]. When this assumption fails, and the missing data mechanism is *not ignorable* (NI), the accuracy of these methods can dramatically decrease. Finally, the computational cost of these methods heavily depends on the absolute number of missing data, and this can prevent their scalability to large databases.

These limitations motivated the development of BC, a deterministic method to estimate conditional probabilities from an incomplete database. The method *bounds* the set of possible estimates consistent with the available information by computing the minimum and the maximum estimate that would be obtained from all possible completions of the database. This process returns probability intervals containing all possible estimates consistent with the available information. These bounds are then *collapsed* into a unique value via a convex combination of the extreme points with weights depending on the assumed pattern of missing data. Details and properties of BC are described in [5].

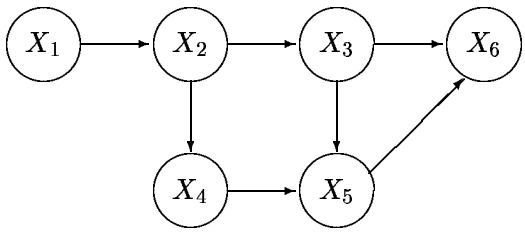

Figure 1: The BBN used for the evaluation.

	Estimation Errors											
	Ignorable						Non Ignorable					
	9%			37%			3%			9%		
	GS	EM	BC	GS	EM	BC	GS	EM	BC	GS	EM	BC
Min	0.0000	0.0000	0.0000	0.0002	0.0001	0.0001	0.0000	0.0000	0.0000	0.0003	0.0003	0.0003
1st Qu.	0.0001	0.0001	0.0001	0.0008	0.0008	0.0009	0.0003	0.0003	0.0003	0.0017	0.0017	0.0016
Median	0.0004	0.0004	0.0003	0.0018	0.0019	0.0021	0.0008	0.0008	0.0008	0.0057	0.0056	0.0057
3rd Qu.	0.0007	0.0007	0.0010	0.0031	0.0030	0.0041	0.0021	0.0020	0.0020	0.0116	0.0116	0.0116
Max	0.0035	0.0034	0.0037	0.0119	0.0117	0.0229	0.0112	0.0111	0.0112	0.0917	0.0907	0.0918
X Ent.	0.0012	0.0011	0.0018	0.0336	0.0319	0.0473	0.0078	0.0078	0.0078	0.3583	0.3501	0.3587
	Prediction Errors											
Min	0.0000	0.0001	0.0001	0.0066	0.0049	0.0018	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
1st Qu.	0.0021	0.0028	0.0013	0.0088	0.0097	0.0175	0.0272	0.0275	0.0273	0.1186	0.1131	0.1186
Median	0.0046	0.0051	0.0021	0.0220	0.0224	0.0199	0.0314	0.0315	0.0314	0.1436	0.1440	0.1438
3rd Qu.	0.0059	0.0062	0.0031	0.0404	0.0396	0.0414	0.0355	0.0359	0.0356	0.1696	0.1756	0.1697
Max	0.0087	0.0094	0.0089	0.0719	0.0699	0.0858	0.0669	0.0678	0.0672	0.4786	0.4750	0.4789

Table 1: Summary statistics of estimation and prediction errors.

3. Experimental Evaluation

Aim of these experiments is to evaluate the estimation accuracy of BC, EM algorithm and GS as the available information in the database decreases. We used a database reporting the values of six risk factors of coronary diseases in 1841 employees of a Czechoslovakian car factory during a follow up study [8, page 261]. All variables are binary and they represent *Anamnesis* (X_1), *Strenuous Mental Work* (X_2), *Ratio of Beta and Alpha Lipoproteins* (X_3), *Strenuous Physical Work* (X_4) *Smoking* (X_5) and *Systolic Blood Pressure* (X_6). The database is complete and we extracted the most probable directed graphical model, depicted in Figure 1, using the K2 algorithm [1]. The 15 conditional dependencies specified by this BBN were then quantified from the complete database. Uniform prior distributions were assumed for the parameters θ. We then used the BBN in Figure 1 to run two different learning tests in order to evaluate the accuracy of BC relative to GS and EM, using the implementation of accelerated EM in GAMES [6], the implementation of GS in BUGS [7] and the implementation of BC in BKD. The threshold for the EM was 10^{-4}. Results of GS are based on a first burn-in of 5,000 iterations, sufficient to reach stability, and a successive sample of 5,000 cases.

The aim of the first test was to compare the estimation accuracy of the three methods when the missing data mechanism is ignorable. For this purpose, four incomplete databases

were created by incrementally deleting data from the complete database. A vector ψ of 15 numbers between 0 and 1 was randomly generated, and elements of ψ were taken as the probability of deleting the occurrences of each variable X_i, independently of its value, given the parent configuration, in the the 10%, 20%, 30% and 40% of the database. This process generated four databases with 9% (1004), 18% (2035), 28% (3041) and 37% (4092) missing entries, and the observed entries within parent configurations are representative of the complete samples.

The rationale of the second test was to compare the robustness of these methods with a not ignorable missing data mechanism. Four samples were generated from the complete database by incrementally deleting respectively 25%, 50%, 75% and 100% of the entries $(X_5 = 2, X_6 = 1)$ with probability 0.9, and $(X_5 = 2, X_6 = 2)$ with probability 0.1. This process generated 4 databases with 3% (278), 5% (532), 7% (790) and 9% (1030) missing entries. The estimation accuracy is measured by comparing the exact joint probability distribution of $(X_1, ..., X_6)$ to those learned from the incomplete data using GS, EM and BC. Table 1 reports some summary statistics of the results obtained on the two extreme databases generated in the two tests. The first five rows of Table 1 report summary statistics, in terms of minimum, maximum, first, third quartile and median of the absolute difference between the 64 exact joint probabilities and those obtained from the BBNs learned with GS, EM and BC. X Ent. labels the cross entropy between joint probability distributions.

The predictive accuracy was evaluated by comparing the predictive probabilities of X_6 obtained by the three methods to those calculated by the BBN extracted from the complete database. Given the conditional independence assumptions embedded in the BBN there are 43 possible evidences that can be considered: 10 are given by observing one of the variables $X_1, \cdots, X_5$ alone, 32 by observing one of the pairs (X_1, X_3), (X_1, X_4), (X_1, X_5), (X_2, X_3), (X_2, X_4), (X_2, X_5), (X_3, X_4), (X_3, X_5), and one in which the evidence is the empty set, so that the marginal probability of X_6 is returned. Bounds on the predictive probabilities were also computed by propagating BC upper bounds for each conditional dependency. Summary statistics of the absolute errors are given in the second half of Table 1.

When 9% of data is missing and the missing data mechanism is ignorable, the distribution of the errors incurred by the three methods differs at the fourth decimal point and a slightly higher maximum value of BC is balanced by a lower median point. The difference becomes a little more evident when 37% of data are missing. Nonetheless, the median difference still remains around 0.0002 and, most of all, these differences do not affect the predictive power of the BBN extracted by BC, which, if anything, scores a slightly better median performance, larger than any difference among the estimates.

This overall equivalence among the three methods is confirmed when missing data are NI, where error distributions are almost identical, and the impact of a wrong assumption about the pattern of missing data equally affect all methods. However, BC provides bounds on the predicted values which reflect their reliability and can be taken into account during the reasoning process. It must be also remarked that BC does not rely *per se* on the ignorability

Missing	GS	EM	BC
278	00:15:22	00:00:30	00:00:14
532	00:23:00	00:01:10	00:00:13
790	00:29:42	00:01:50	00:00:14
1004	00:37:47	00:01:50	00;00:13
1030	00:29:12	00:02:00	00:00:13
2035	01:09:47	00:03:12	00:00:13
3041	01:56:02	00:04:18	00:00:14
4056	04:26:07	00:07:26	00:00:13

Table 2: Execution time for each method.

assumption and that the available information about the missing data could have been exploited by BC to achieve a better performance. The main difference among the three methods highlighted by the experiments is the execution time and, most of all, the shape of its growth curve, showing that the execution time of BC is independent of the number of missing data [5]. Table 2 shows the execution time for all eight databases generated during the experiments. The first column reports the absolute number of missing entries from the original entries of the complete database. Results are in *hours*:*minutes*:*seconds* format.

4. Conclusions

The first character of BC is its ability to explicitly and separately represent the information available in the database and the assumed pattern of missing data, so that BC does not need to rely on the ignorability assumption, although it can be easily represented as any other. Furthermore, BC reduces the cost of estimating each conditional distribution of X_i to the cost of one exact Bayesian updating and one convex combination for each state of X_i in each parent configuration. This deterministic process does not pose any problem of convergence rate and detection and its computational complexity is independent of the number of missing data. The experimental comparison with EM and GS showed a substantial equivalence of the estimates provided by these three methods and a dramatic gain in efficiency using BC.

References

[1] G.F. Cooper and E. Herskovitz. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

[2] A. Dempster, D. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

[3] 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.

[4] R.J.A. Little and D.B. Rubin. *Statistical Analysis with Missing Data.* Wiley, New York, NY, 1987.

[5] M. Ramoni and P. Sebastiani. Parameter estimation in bayesian networks from incomplete databases. *Intelligent Data Analysis Journal*, 2(1), 1998.

[6] B. Thiesson. Accelerated quantification of Bayesian networks with incomplete data. In *Proceedings of first international conference on knowledge discovery and data mining*, pages 306–11. AAAI press, 1995.

[7] A. Thomas, D.J. Spiegelhalter, and W.R. Gilks. Bugs: A program to perform Bayesian inference using Gibbs Sampling. In *Bayesian Statistics 4*, pages 837–42. Clarendon Press, Oxford, 1992.

[8] J. Whittaker. *Graphical Models in Applied Multivariate Statistics.* Wiley, New York, NY, 1990.