Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics and Data Analysis 52 (2008) 3583–3602
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Maximum entropy and least square error minimizing procedures for estimating missing conditional probabilities in Bayesian networks

Parag C. Pendharkar*

Information Systems, School of Business Administration, Pennsylvania State University at Harrisburg, 777 West Harrisburg Pike, Middletown, PA 17057, United States

Received 15 March 2006; received in revised form 11 March 2007; accepted 22 November 2007
Available online 4 December 2007

Abstract

Conditional probability tables (CPT) in many Bayesian networks often contain missing values. The problem of missing values in CPT is a very common problem and occurs due to the lack of data on certain scenarios that are observed in the real world but are missing in the training data. The current approaches of addressing the problem of missing values in CPT are very restrictive in that they assume certain probability distributions for estimating missing values. Recently, maximum entropy (ME) approaches have been used to learn features of probability distribution functions from the observed data. The ME approaches do not require any data distribution assumptions and are shown to work well for several non-parametric distributions. The ME and least square (LS) error minimizing approaches can be used for estimating missing values in CPT for Bayesian networks. The applications of ME and LS approaches for estimating missing CPT require researchers to solve difficult constrained non-linear optimization problems. These difficult constrained non-linear optimization problems can be solved using genetic algorithms.

1. Introduction

Missing conditional probability values in conditional probability tables (CPT) are a very common problem in a Bayesian network design (Pearl, 1988). The reasons for missing values are high cost of data collection, data punching errors, longitudinal high dimensional data (Formann, 2007), and unavailable data (Demirtas et al., 2007). The current literature on missing values in CPT can be divided into two streams. The first stream is related to missing values in CPT due to incomplete data. The second stream is related to missing values in CPT with complete data.

Incomplete data sets contain some unreported entries in the data set, which makes conditional probability estimation a difficult task. Earlier approaches of handling missing values included the Monte Carlo simulation approach suggested by Henrion (1987) and described in Pearl (1988), and other approaches summarized in Heckerman (1997). All of the approaches suggested by Heckerman (1997) are either computationally intractable, or require knowledge of the expectation function or probability density/distribution function. Recent approaches of estimating conditional probabilities from incomplete data sets include the use of the expectation maximization procedure, Gibbs Sampling, Bound and Collapse approach (Ramoni and Sebastiani, 1998) and robust Bayesian estimator

* Tel.: +1 717 948 6028; fax: +1 717 948 6456.
E-mail address: pxp19@psu.edu.

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2007.11.013

(RBE) (Ramoni and Sebastiani, 2001). Some of the recent methods assume that missing data follows a certain pattern. However, the RBE approach does not assume any particular missing data pattern and has a polynomial computational complexity (Ramoni and Sebastiani, 2001). Two approaches, expectation maximization and Gibb Sampling, are known to get trapped into local minima and are not guaranteed to converge (Ramoni and Sebastiani, 1999). Since the RBE is a relatively new approach, we do not have any information on the convergence of the RBE.

Missing conditional probabilities can occur in complete data sets. This is likely to happen when the size of a data set is small. For example, assume a data set containing three variables A, B, and C with each taking three discrete values. There will be a total of $3^3 = 27$ unique combinations of their values. Assuming that all unique combinations can legitimately occur in the real world, a complete data set of size 25 will contain a minimum of two missing combinations. While the data set might appear complete, when a Bayesian network is constructed using the data, the missing combinations will lead to missing values in conditional probability tables for the Bayesian network. Since the data set is complete, approaches used to estimate missing conditional probabilities with incomplete data set cannot be used here. In the Bayesian network literature, this problem is sometimes called "filling in" missing conditional probabilities (Paris, 2005) or estimating conditional probability with incomplete information (Holmes et al., 1998). The problem of estimating missing conditional probabilities in a complete data set is known to be computational complex with multiple solutions (Paris, 2005). The solutions proposed for solving this problem assume a unique solution (Rhodes and Garside, 1996) or assume a uniform distribution for unknown conditional probabilities (Paris, 2005). A procedure proposed by Paris (2005), called center of mass inference (CMI) procedure, provides a fast computationally efficient trivial solution. All the procedures proposed to estimate missing conditional probabilities with complete data sets suffer from two drawbacks, (1) they assume a unique solution, and (2) they use a local search procedure to obtain the unique solution. In addition to these two common limitations, the procedures suffer from additional individual limitations. For example, the CMI procedure focuses only on estimating missing conditional probabilities and does not consider the values of known conditional probabilities. Additionally, the CMI procedure does not satisfy Language Invariance (Paris, 2005). The CMI procedure, however, is the most computationally efficient procedure.

In this paper, we focus on the problem of estimating missing conditional probabilities with a complete data set. We pose the problem of estimating missing conditional probabilities as non-linear mathematical programming problems with different objective functions. We show that the mathematical programming problems have multiple solutions. Unlike previous studies on estimating the values of missing conditional probabilities with complete data set, we use a global search genetic algorithm (GA) procedure to search for a best possible solution among different multiple solutions. GAs have gained popularity in computational statistics literature and recent research illustrates several successful applications of GAs for bank rating (Krink et al., 2007), variable selection in regression models (Kapetanios, 2007) and maximum likelihood estimation for the threshold vector error correction model (Yang et al., 2007).

We provide a general description of estimating the missing values in CPT for complete data sets. Assume that $\boldsymbol{p} = (p_1, \ldots, p_n)$ is a vector of actual conditional probabilities for a Bayesian network, where not all components of the vector $\boldsymbol{p}$ are known. Further, assume that a technique estimates the missing values of the CPT in the Bayesian network, and provides a vector $p^* = (p_1^*, \ldots, p_n^*)$ which is an approximation of $\boldsymbol{p}$ where all components of $\boldsymbol{p}^*$ are known. Theoretically, the magnitude of the error of approximation, e, is given by $e = \sum_{i=1}^{n} \left|\left(\boldsymbol{p_i} - \boldsymbol{p_i^*}\right)\right|$. Since not all components of vector $\boldsymbol{p}$ are known, the error e is considered only for known components. In our research, a technique that estimates all values of $\boldsymbol{p}^* = (p_1^*, \ldots, p_n^*)$, and minimizes e for known values of $\boldsymbol{p}$, and satisfies all the laws of probabilities is considered to be a good estimator of $\boldsymbol{p}$.

To design a technique that best minimizes e, certain assumptions about the distribution of e are required. If e is assumed to be normal, independent, and identically distributed then the technique minimizing least square (LS) errors may be the best estimator. However, if the distribution of e has fatter tails than the normal distribution then the LS estimator will be inefficient (Wu and Stengos, 2005) and adaptive estimators are necessary for establishing the best estimator.

Maximum entropy (ME) is a set of general purpose partially adaptive quasi-maximum likelihood estimators that nest most commonly used mathematical distributions (Wu and Stengos, 2005) including normal and t distributions. In a recent study, Wu and Stengos (2005) found that the ME estimators show considerable degree of adaptiveness to different shapes of error distributions and work well when compared to other methods.

In this paper, we propose two approaches to estimate the vector $\boldsymbol{p}^*$. The approaches primarily differ in the assumption of distribution of error e. The first approach uses least square minimization of e and the other approach uses the ME method. Both the LS and the ME approaches require a solution of a difficult constrained optimization problem.