Journal of Applied Logic 13 (2015) 628–642

Contents lists available at ScienceDirect

Journal of Applied Logic

www.elsevier.com/locate/jal

Learning discrete Bayesian network parameters from continuous data streams: What is the best strategy?

Parot Ratnapinda [a,b,*], Marek J. Druzdzel [a,c,**]

[a] *Decision Systems Laboratory, School of Information Sciences and Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA 15260, USA*
[b] *Faculty of Science, Information Technology Division, Maejo University, Chiang Mai 50290, Thailand*
[c] *Faculty of Computer Science, Białystok University of Technology, Wiejska 45A, 15-351 Białystok, Poland*

ARTICLE INFO

Article history:
Available online 18 March 2015

Keywords:
Bayesian networks
Parameter learning
EM algorithm

ABSTRACT

We compare three approaches to learning numerical parameters of discrete Bayesian networks from continuous data streams: (1) the EM algorithm applied to all data, (2) the EM algorithm applied to data increments, and (3) the online EM algorithm. Our results show that learning from all data at each step, whenever feasible, leads to the highest parameter accuracy and model classification accuracy. When facing computational limitations, incremental learning approaches are a reasonable alternative. While the differences in speed between incremental algorithms are not large (online EM is slightly slower), for all but small data sets online EM tends to be more accurate than incremental EM.

1. Introduction

An increasing number of domains produce continuous, massive amounts of data. World Wide Web-based systems, for example, often generate records for every user transaction. Real-time monitoring systems obtain sensor readings in fraction of a second increments. A corporate call center may deal with hundreds or even thousands of new cases daily. There exist computer programs that specialize in continuous data streams and that operate in real-time, e.g., [1,8,9]. They all need to learn from the incoming massive amounts of data and systematically update whatever they know about the system that they are monitoring.

There are two fundamental approaches to processing continuous data streams, which we will call *batch learning* and *incremental learning*. In the batch learning approach, we repeatedly add new records to the

* Corresponding author at: Decision Systems Laboratory, School of Information Sciences and Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA 15260, USA.
** Principal corresponding author at: Decision Systems Laboratory, School of Information Sciences and Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA 15260, USA.
E-mail addresses: parotr@gmail.com (P. Ratnapinda), marek@sis.pitt.edu (M.J. Druzdzel).

http://dx.doi.org/10.1016/j.jal.2015.03.007
1570-8683/

accumulated data and learn anew from the entire data set. When the number of data records becomes very large, this approach may be computationally prohibitive. In addition, it requires storing and efficiently retrieving the entire data set, which may not be feasible. In the incremental learning approach, we assume that the model learned in the previous step summarizes all the data collected up to that step and we use the newly acquired data to refine the model. Incremental learning approaches can be divided into two types: *incremental batch learning* and *online learning*. The incremental batch learning or mini-batch learning updates the model by processing the incoming data in chunks, i.e., groups of records. The online learning updates the model by processing records one at the time as they arrive.

Our work is in the context of discrete Bayesian network models [18], which are becoming increasingly popular in modeling and learning tasks. While there are other ways of updating Bayesian network parameters (e.g., [16]), the most flexible algorithm for learning discrete Bayesian network parameters is the EM (Expectation Maximization) algorithm [6,13]. While there are several variants of the EM algorithm, two are most notable: the basic EM algorithm [6] and the *online EM* algorithm [2,14,19].

The most common mode of operation of the basic EM algorithm is *batch learning*, i.e., learning from an entire data set. The basic EM algorithm can be also applied to incremental batch learning, in which case the existing set of parameters, learned previously from a database of cases, is assigned a level of reliability, captured by a number called the *equivalent sample size* (ESS). Equivalent sample size expresses the number of data records that have been used to learn the existing parameters. While updating the existing parameters, the EM algorithm weights the new cases against the existing parameters according to the relative sizes of the data sets. As each of the algorithms requires belief updating, their complexity is worst-case NP-hard [4]. Their relative complexity differs, although it is driven largely by the number of data records that they have to process. The computational complexity of the incremental batch EM depends primarily on the size of the set of additional records, i.e., the mini-batch. The *online EM* algorithm is a modification of the basic EM algorithm that allows for processing new data into the existing model one record at a time. Its complexity at each time step, both in terms of computation time and memory use, is thus the lowest of the three.

The question that we pose in this paper is which of the three approaches is best in practice when learning discrete Bayesian network parameters from continuous data streams. We assume these streams to be stationary, i.e., generated by systems whose parameters themselves are not changing over time, although we propose a way of approaching parameter learning when the system is non-stationary. We focus on the impact of choice of each of the learning schemes on (1) computational complexity of learning (speed), (2) accuracy of the learned parameters, and (3) the model's ultimate accuracy. We pose the third question in the context of classification tasks, which is a common application of Bayesian networks. While there exists literature that is related to this question, no comprehensive comparison has been made so far in the context of Bayesian networks. Some papers focus on the comparison of batch learning to incremental learning, e.g., [3,20]. They agree on the obvious truth that the online learning is computationally more efficient than batch learning and show experimentally that it also achieves accuracy that is similar to that of the batch learning. Cappe [2], who compares batch EM to online EM, suggests that the decision to select between the two algorithms depends on the size of the data set. His experiments indicate that when the size of the data is smaller than 1000 records, batch EM is preferred to online EM. Holmes et al. [10] study how mini-batch size affects the performance of incremental learning in terms of classification accuracy and speed. They demonstrate that larger chunk sizes lead to higher classification accuracy.

In this paper, we describe an experiment, in which we use several real data sets from the UCI Machine Learning Repository [7] to create gold standard Bayesian network models. We subsequently use these models to generate continuous streams of data. We learn the parameters from these streams of data with three approaches: *batch EM*, *incremental batch EM*, and *online EM*. We measure the time taken by the learning procedure, compare the accuracy of the learned parameters to the original (gold standard) parameters that have generated the data, and test the diagnostic accuracy of the learned models.