Information Sciences 425 (2018) 62–75

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Mining massive hierarchical data using a scalable probabilistic graphical model

Khalifeh AlJadda[a,*], Mohammed Korayem[a], Camilo Ortiz[d], Trey Grainger[a], John A. Miller[b], Khaled M. Rasheed[b], Krys J. Kochut[b], Hao Peng[b], William S. York[c], Rene Ranzinger[c], Melody Porterfield[c]

[a] *CareerBuilder, Norcross, GA, USA*
[b] *Computer Science Department, University of Georgia, Athens, GA, USA*
[c] *Complex Carbohydrate Research Center, University of Georgia, Athens, GA, USA*
[d] *AI Research and Development, Bloomberg, NYC, NY*

ARTICLE INFO

Article history:
Received 8 February 2016
Revised 4 October 2017
Accepted 6 October 2017
Available online 10 October 2017

MSC:
00-01
99-00

Keywords:
Probabilistic model
Mass spectrometry annotation
Big data
Large scale machine learning
Smantic discovery

ABSTRACT

Probabilistic Graphical Models (PGM) are very useful in the fields of machine learning and data mining. The crucial limitation of those models, however, is their scalability. The Bayesian Network, which is one of the most common PGMs used in machine learning and data mining, demonstrates this limitation when the training data consists of random variables, in which each of them has a large set of possible values. In the big data era, one could expect new extensions to the existing PGMs to handle the massive amount of data produced these days by computers, sensors and other electronic devices. With hierarchical data - data that is arranged in a treelike structure with several levels - one may see hundreds of thousands or millions of values distributed over even just a small number of levels. When modeling this kind of hierarchical data across large data sets, unrestricted Bayesian Networks may become infeasible for representing the probability distributions. In this paper, we introduce an extension to Bayesian Networks that can handle massive sets of hierarchical data in a reasonable amount of time and space. The proposed model achieves high precision and high recall when used as a multi-label classifier for the annotation of mass spectrometry data. On another data set of 1.5 billion search logs provided by CareerBuilder.com, the model was able to predict latent semantic relationships among search keywords with high accuracy.

1. Introduction

A Probabilistic Graphical Model (PGM) consists of a structural model and a set of conditional probabilities [21,40]. PGMs are widely used in machine learning and data mining in many fields, e.g., speech recognition [25], bioinformatics [14,41], Natural Language Processing (NLP) [9,27]. In the big data era, the complexity of data and scalability can be major challenges for PGMs. To overcome these challenges, one could expect extensions to the existing PGMs. One such extension is the Hierarchical Probabilistic Graphical Model (HPGM), which aims to extend the PGM to work with more structured domains

[*] Corresponding author.
E-mail address: khalifeh.aljadda@careerbuilder.com (K. AlJadda).

https://doi.org/10.1016/j.ins.2017.10.014
0020-0255/© 2017 Published by Elsevier Inc.

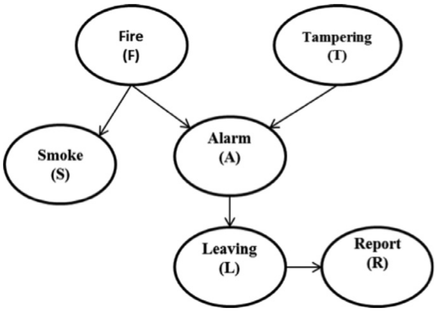

Fig. 1. Example Bayesian Network.

[16,18]. However, this extension focuses more on the complex data problem, but less so on scalability. For hierarchical data that can be divided into several levels and arranged in tree-like structures, data items in each level depend on or are influenced only by the data items in the upper levels, so a Bayesian Network is an appropriate type of PGM to represent such a probability distribution, since the dependencies in this kind of data are not bidirectional. For big data, an unrestricted Bayesian Network may be infeasible, though, since it may not be able provide a concise representation or approximation of a huge joint probability distribution. In massive hierarchical data, which is becoming increasingly common in the big data era, each level represents a random variable, while each node in that level represents an outcome (possible value) of that random variable. The infeasibility of an unrestricted Bayesian Network is evident because the data can grow horizontally (number of values) at a much faster rate than vertically (number of random variables). Moreover, since the dependencies among the random variables are predefined in the hierarchical data, the structure of the network is also predefined. Hence, the first phase of building a Bayesian Network, finding the optimal structure, is substantially restricted.

For example, consider the glycan ontology “GlycO” [43] which describes 1300 glycan structures (see section d), for which the theoretical tandem mass spectra (MS) can be predicted by GlycoWorkbench [8]. If the maximum of cleavages is set to two and the number of cross-ring cleavages is set to one, then the theoretical MS^2 spectrum contains 2,979,334 ions, which themselves can be fragmented to form tens of millions of ions in MS^3. Representing two levels of MS data using a Bayesian Network (BN) involves a network composed of nodes at two levels, one for MS^1 and one for MS^2. The Conditional Probability Tables for the MS^2 could contain 3,873,134,200 (2,979,334 × 1300) entries. For this kind of data, we propose a simple Probabilistic Graphical Model for massive Hierarchical Data (PGMHD). Currently, we focus on using a Bayesian Network, a common Probabilistic Graphical Model, to represent massive hierarchical data in a more efficient way. We successfully apply the PGMHD in two different domains: bioinformatics, for multi-label classification, and search log analytics, for latent semantic discovery of related terms.

The main contributions of this paper are as follows: We propose a simple, efficient and scalable probabilistic graphical model for massive hierarchical data. We successfully apply this model to the bioinformatics domain in which we automatically classify and annotate high-throughput mass spectrometry data. We also apply this model to large-scale latent semantic discovery using 1.6 billion search log entries provided by CareerBuilder.com, using the Hadoop MapReduce framework.

Graphical models can be classified into two major categories: (1) directed graphical models (the focus of this paper), which are often referred to as Bayesian Networks, or belief networks, and (2) undirected graphical models which are often referred to as Markov Random Fields, Markov networks, Boltzmann machines, or log-linear models [24]. Probabilistic graphical models consist of both the graph structure and parameters. The graph structure represents a set of conditionally independent relations for the probability model, while the parameters consist of the joint probability distributions [40].

PGMs are used in many domains. For example, Hidden Markov Models (HMM) are considered crucial components for most speech recognition systems [25]. In bioinformatics, probabilistic graphical models are used in RNA sequence analysis [14]. In natural language processing (NLP), HMM and Bayesian models are used for part of speech (POS) tagging [9]. A problem with PGMs in general, and Bayesian Networks in particular, is that they may not be suitable for representing massive data due to the time complexity of learning the structure of the network and the space complexity of storing a network with thousands of random variables or random variables with large discrete domains.

Bayesian Networks provide a relatively concise way to represent a large joint probability distribution [11]. The structure of a Bayesian Network is a Directed Acyclic Graph (DAG) [21]. A Bayesian Network consists of two components: a DAG representing the structure (as shown in Fig. 1), and a set of Conditional Probability Tables (CPTs). Each node in a Bayesian Network typically has a CPT which quantifies the relationship between the random variable represented by that node and its parents (other random variables it is dependent upon) in the network. Completeness and consistency are guaranteed in a Bayesian Network since there is only one probability distribution that satisfies the Bayesian Network constraints [11]. The constraints that guarantee a unique probability distribution are the numerical constraints represented by CPT and the independence constraints represented by the structure itself. The independence constraints are shown in Fig. 1: once the information about A is known, for example, the probability of L will not be affected by any new information about F or T,