Universal latent structure mining

Carlos Cotrini
Department of Computer Science
ETH Zürich
Switzerland

Abstract

We propose a new method for designing miners of latent structures. Our new method is accessible to anyone with knowledge on programming and elementary probability theory, thus removing the expertise on machine learning or combinatorial algorithms. Moreover, we show how our method attains a competitive performance for a wide variety of learning problems, including mining access control policies, business processes, king-and-rook endgame configurations, and decision trees. Our method makes then machine learning more accessible to non-experts.

1 Introduction

Learning algorithms have been proposed for a wide variety of tasks: autonomous cars, diagnosis of human diseases, face recognition, translation, etc... However, developing such learning algorithms require expertise in machine learning, mathematics, and computer science. The breakthrough that machine learning has made across many disciplines shows that there are fields where non-experts would benefit from applying machine learning to their domain-specific problems. However, the lack of expertise is a major obstacle preventing the application of machine learning to those problems.

We propose a new method for making machine learning more accessible to non-experts. This method is based on deterministic annealing with mean-field approximations. This method facilitates the design of miners of latent structures for a wide variety of learning tasks, including learning access control policies, business processes, king-and-rook endgame configurations, and elementary mathematical functions.

2 Methodology

To design a miner of latent structures for a specific domain, we must perform three tasks:

1. Specify a *template for latent structures*. This is a data structure that has the ability to encode a hypothesis class $\mathcal{C}$ of latent structures.
2. Specify a *cost function* $R(\cdot, \cdot)$ that receives as input a sample X of instances and a latent structure c and outputs a positive real number $R(c, X)$. The lower $R(c, X)$, the best c fits X.
3. Approximate $\arg\min_{c \in \mathcal{C}} R(c, X)$ using deterministic annealing and mean-field approximation.

3 Tasks

The main task is to implement the approach above and evaluate its effectiveness and versatility with the following problems:

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

Mining business processes. A business process is a directed acyclic graph (DAG) whose vertices denote operations within an organization and whose edges denote dependence relations between the operations. Business process mining helps organizations to optimize their processes. The goal is to observe sequences of operations originating from the organization's current business process. Then, from these sequences, we mine a DAG that fits as good as possible the observed sequences. The mined DAG should be simple in the sense that there are no unnecessary repetitions or dependencies. We use publicly available datasets used for business process mining competitions[1].

Mining access control policies. An access control policy is a set of logical rules that define who can access what in an organization. These rules usually depend on the employees' and the organizational resources' attribute values. The idea is to observe how employees access resources and then mine from that a simple and precise access control policy. We use available datasets from Amazon[2].

To illustrate the versatility of our approach, we also show how our approach mines discrete data structures for various types of tasks.

Mining king-rook chess configurations from features like the distance from the black king to the rook and whether the rook checks the black king[3].

Mining decision trees for wine classification. For this we use the wine data set[4].

Mining elementary mathematical functions. We show how our approach mines, given pairs $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R} \times \mathbb{R}$, a formula for a function $f : \mathbb{R} \to \mathbb{R}$ that minimizes $\sum_{i \leq n} \mathcal{L}(y_i, f(x_i))$.

[1] `https://icpmconference.org/2019/process-discovery-contest`

[2] `https://archive.ics.uci.edu/ml/datasets/Amazon+Access+Samples`

[3] `https://archive.ics.uci.edu/ml/datasets/Chess+%28King-Rook+vs.+King-Knight%29`

[4] `https://archive.ics.uci.edu/ml/datasets/wine`