Information-theoretic partition-based models for interpretable machine learning

Yang, L.

Summary

Data is recorded wherever digital systems are. For instance, websites and mobile phone applications record how people interact with them. Sensors measure and record parameters (such as temperature) of the manufacturing process of industrial products. Financial computer software systems record transactions of financial activities. Similarly, healthcare information management systems record conditions of patients.

Such data can be used to reveal information about the physical process that 'generates' them. Research in the field of *data mining* and *machine learning* concerns developing models and algorithms that can extract and leverage information contained in the dataset effectively and efficiently.

For instance, the information revealed from the medical records of a large number of patients in a hospital can be used to understand why a certain condition is very risky for one type of patient but not for the rest. Additionally, monitoring systems can be developed to automatically alert medical staff to dangerous situations for hospitalized patients

This dissertation focuses on developing new methods that can construct *partition-based models* from data. By partitioning a dataset into subsets where each subset is homogeneous from certain perspectives, a partition-based model extracts data regularities, such as patterns indicating which types of patients in hospitals are at risk for certain diseases. It also makes data-driven predictions, such as raising alarms for potentially dangerous situations. Thus, the goal is to develop data mining and machine learning methods that partition the data at hand 'properly'—the concept of *properness* in this dissertation is defined based on an information-theoretic approach.

Specifically, we focus on studying partition-based models for different tasks,

including interpretable multi-class classification, data discretization and summarization, and dependency structure learning. First of all, for interpretable multi-class classification, we develop an interpretable machine learning algorithm that can extract *probabilistic rules* from data. As an example, a probabilistic rule could be *If weather is foggy and flight time is before 9 am, Then the probability of a flight delay is 10%*. Moreover, our developed method is applied to a case study to predict the risk of patients in the ICU of a hospital being readmitted to the ICU after they are discharged, in which we showcase that the model can improve itself by incorporating feedback from medical experts—a pilot study towards human-AI collaboration.

Further, we consider the task of discretization and summarization of two-dimensional datasets, with the potential use case of summarizing destinations of trajectories recorded by GPS devices. The corresponding research question is how to partition datasets that record the locations with a self-adaptive granularity, neither too coarse nor too refined, as the former means a large amount of information is lost while the latter means noise instead of regularities in the data is captured.

Lastly, we study the problem of understanding '(conditional) dependency' structure in data, which is about developing algorithms that can automatically draw conclusions like 'the risk of forest fire is independent of which month it is, conditioned on the temperature and humidity' from data (i.e., once the temperature and humidity are known, the risk of forest fire is the same no matter what month it is). Both theoretical and methodological research is conducted specifically for a challenging data type, i.e., the mixture of discrete and continuous values. For instance, forest fire historic data are often recorded in this type, which contains 'no fire' (a discrete label) and 'area of fires' (a quantitative, continuous value).

To sum up, partition-based models are studied for various tasks, with the goal of making them more interpretable and transparent to the end-user.