Let the Data Speak for Themselves - Simplistic Approaches to Difficult Data Mining Problems

Wei Fan

IBM T. J. Watson Research, USA

Abstract:

Inductive learning is to construct an accurate model from labeled training examples to match the true model that generates the data. One major difficulty is that the actual true model is never known for many practical problems. Any assumption about the exact form of the true model could be wrong. The validity of an assumption is difficult to verify since labeled examples are non-exhaustive for most applications, and there may be little known truth about the exact generating mechanism. The main stream research of machine learning has been focusing on rather sophisticated and well-thought approaches to approximating the true models in classification, regression and probability estimation problems. Examples of well-known algorithms belonging to this family include Boosting, Bagging, SVM, Mixture models, Logistic Regression, and GUIDE, among many others.

In this talk, we will discuss a family of Randomized Decision Tree algorithms or RDT that can be used efficiently and accurately for classification, regression and probability estimation problems. The training procedure of RDT incorporates some surprisingly simple and unconventional random factors that "encode" the training data into multiple decision trees. However, its accuracy in all three major problems is either higher or significantly higher than many well-known sophisticated approaches.

In summary, this talk offers the following insights:

1. Introduction of Randomized Decision Trees and its application in classification, regression, and probability estimation.
2. Selected applications of RDT under many difficult situations.
   a) sample selection bias where the training set is not representative.
   b) extremely skewed distribution.
   c) very large number of categorical features.
   e) true answer is non-deterministic given the feature vector.
   Example applications include:
   a) equity trading fraud detection.
   b) customer default payment prediction.
   c) information retrieval.
   d) storage component latency modeling.
   e) ground ozone level estimation.
   f) chip failure prediction.
   g. parametric query optimization.
3. A fresh and unconventional look at accurate machine learning and data mining without making strong assumptions.