

Preface

Data mining has emerged as a new science—the exploration, algorithmically and systematically, of data in order to extract patterns that can be used as a means of supporting organizational decision making. Data mining has evolved from machine learning and pattern recognition theories and algorithms for modeling data and extracting patterns. The underlying assumption of the inductive approach is that the trained model is applicable to future, unseen examples. Data mining can be considered as a central step in the overall knowledge discovery in databases (KDD) process.

In recent years, data mining has become extremely widespread, emerging as a discipline featured by an increasing large number of publications. Although an immense number of algorithms have been published in the literature, most of these algorithms stop short of the final objective of data mining—providing possible actions to maximize utility while reducing costs. While these algorithms are essential in moving data mining results to eventual application, they nevertheless require considerable pre- and post-process guided by experts.

The gap between what is being discussed in the academic literature and real life business applications is due to three main shortcomings in traditional data mining methods. (i) Most existing classification algorithms are ‘passive’ in the sense that the induced models merely predict or explain a phenomenon, rather than help users to proactively achieve their goals by intervening with the distribution of the input data. (ii) Most methods ignore relevant environmental/domain knowledge. (iii) The traditional classification methods are mainly focused on model accuracy. There are very few, if any, data mining methods that overcome all these shortcomings altogether.

In this book we present a proactive and domain-driven method to classification tasks. This novel proactive approach to data-mining, not only induces a model for predicting or explaining a phenomenon, but also utilizes specific problem/domain knowledge to suggest specific actions to achieve optimal changes in the value of the target attribute. In particular, this work suggests a specific implementation of the domain-driven proactive approach for classification trees. The proactive method is a two-phase process. In the first phase, it trains a probabilistic classifier using a supervised learning algorithm. The resulting classification model from the first-phase is a model that is predisposed to potential interventions and oriented toward maximizing

a utility function the organization sets. In the second phase, it utilizes the induced classifier to suggest potential actions for maximizing utility while reducing costs.

This new approach involves intervening in the distribution of the input data, with the aim of maximizing an economic utility measure. This intervention requires the consideration of domain-knowledge that is exogenous to the typical classification task. The work is focused on decision trees and based on the idea of moving observations from one branch of the tree to another. This work introduces a novel splitting criterion for decision trees, termed maximal-utility, which maximizes the potential for enhancing profitability in the output tree.

This book presents two real case studies, one of a leading wireless operator and the other of a major security company. In these case studies, we utilized our new approach to solve the real world problems that these corporations faced. This book demonstrates that by applying the proactive approach to classification tasks, it becomes possible to solve business problems that cannot be approach through traditional, passive data mining methods.

Tel Aviv, Israel
July, 2013

Haim Dahan
Shahar Cohen
Lior Rokach
Oded Maimon

Proactive Data Mining with Decision Trees

Dahan, H.; Cohen, S.; Rokach, L.; Maimon, O.

2014, X, 88 p. 20 illus., Softcover

ISBN: 978-1-4939-0538-6