How to Learn Enough Data Mining to be Dangerous in 60 Minutes

Abraham Bernstein
Dynamic and Distributed Information Systems Group, Department of Informatics
University of Zurich
Binzmühlestrasse 14
8050 Zürich, Switzerland
bernstein@ifi.uzh.ch

ABSTRACT

The field of data mining provides some methods highly relevant to researchers when mining software repositories. Whether one predicts bug locations, discovers hidden architectural structures and software patterns, or identifies experts of modules, data mining algorithms are usually the working horses for these studies. The goal of this tutorial is to convey some of the most relevant theoretical foundations and practical issues when using data mining algorithms.

The tutorial will first discuss the usual data mining tasks (prediction, filtering, smoothing, and elucidation of the most likely explanation or structure). Then, it will introduce a general framework for data mining paving the way to explain the functionality of some of the most used data mining algorithms. The tutorial will close with an overview over the typical evaluation methods for induced results and a number of pointers for further study. Where possible, it will use examples from software engineering.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: [Database applications: Data mining]; D.2.8 [**Software Engineering**]: Software/Program Verification—*Statistical methods*; D.2.8 [**Software Engineering**]: Testing and Debugging—*Diagnostics*; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Measurement, Performance

Keywords

Mining Software Repositories

MSR'08, May 10-11, 2008, Leipzig, Germany.
ACM 978-1-60558-024-1/08/05.