Statistically Sound Pattern Discovery

Wilhelmiina Hämäläinen
University of Eastern Finland
whamalai@cs.uef.fi

Geoffrey I. Webb
Monash University, Australia
geoff.webb@monash.edu

ABSTRACT

Pattern discovery is a core data mining activity. Initial approaches were dominated by the frequent pattern discovery paradigm – only patterns that occur frequently in the data were explored. Having been thoroughly researched and its limitations now well understood, this paradigm is giving way to a new one, which can be called *statistically sound pattern discovery*. In this paradigm, the main impetus is to discover statistically significant patterns, which are unlikely to have occurred by chance and are likely to hold in future data. Thus, the new paradigm provides a strict control over false discoveries and overfitting.

This tutorial covers both classic and cutting-edge research topics on pattern discovery combined to statistical significance testing. We start with an advanced introduction to the relevant forms of statistical significance testing, including different schools and alternative models, their underlying assumptions, practical issues, and limitations. We then discuss their application to data mining specific problems, including evaluation of nested patterns, the multiple testing problem, algorithmic strategies and real-world considerations. We present the current state-of-the art solutions and explore in detail how this approach to pattern discovery can deliver efficient and effective discovery of small sets of interesting patterns.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data mining

Keywords

Pattern Discovery, Association Mining, Statistics

Who Should Attend

The tutorial is aimed at anyone interested in finding useful and reliable patterns in data. It is likely to inspire both practical data miners (how to test and improve the quality of discovered patterns) and algorithm designers (how to target the search into the most reliable patterns). The main contents are easy-to-understand to anyone who wants to get an overview of this new paradigm, but there is also deeper analysis, open problems and controversial questions which should interest experienced researchers, as well.

Prerequisites

General data mining/pattern discovery background, familiarity with basic mathematical concepts, including probability theory and statistics (undergraduate level mathematics).

Instructors

Dr. Wilhelmiina Hämäläinen is a postdoctoral researcher by the Academy of Finland, currently working in the School of Computing, University of Eastern Finland. She received a Ph.D. degree 2010 (Computer Science) from the University of Helsinki. She has worked as a teacher, lecturer, and researcher in the university since 1996, including 2 years as a university researcher of biology (applied data mining). She has often worked with interdisciplinary problems, involving computer science, statistics, and mathematics. Her main achievements in data mining are efficient algorithms for finding reliable statistical dependency patterns (a related award from The Finnish Society for Computer Science and Research Foundation of the Finnish Information Processing Association). Her research interests include statistical dependency analysis and significance testing, optimization algorithms, applied knowledge discovery (biology, educational technology) and general number crunching.

Dr. Geoff Webb is a Professor of Information Technology Research at Monash University. He has developed numerous algorithms and techniques for machine learning, data mining and computational structural biology. His commercial data mining software, Magnum Opus, incorporates many techniques from his association discovery research. Many of his learning algorithms are included in the widely-used Weka machine learning workbench. He is editor-in-chief of *Data Mining and Knowledge Discovery*, co-editor of the *Springer Encyclopedia of Machine Learning*, on the *Statistical Analysis and Data Mining* advisory board, and has served on numerous editorial boards including *Machine Learning* and *ACM Transactions on Knowledge Discovery from Data*. He is PC Co-Chair of SIGKDD 2105, was PC Co-Chair of ICDM 2010 and co-General Chair of ICDM 2012. He has received the 2013 *IEEE ICDM Service Award* and a 2014 *Australian Research Council Discovery Outstanding Researcher Award*.

KDD'14, August 24–27, 2014, New York, NY, USA.
ACM 978-1-4503-2956-9/14/08.
http://dx.doi.org/10.1145/2623330.2630812.