Simon Fischer, Ingo Mierswa (Eds.)

Proceedings of the 2nd RapidMiner Community Meeting and Conference (RCOMM 2011)

Shaker Verlag
Aachen 2011

Preface

Welcome to the Second RapidMiner Community Meeting and Conference in Dublin, Ireland. After the first RapidMiner conference held in Dortmund in September 2010, we moved the conference to mid June in 2011 due to collisions with other events.

Despite of this resulting in a reduced preparation period, we received many excellent contributions again and selected fourteen for presentation during the conference and for publication in these proceedings. As in 2010, submissions cover a wide range of topics, including machine learning, applications and use cases, and architecture. During the conference, various new RapidMiner extensions will be presented. Even manuscripts utilizing the new data mining server RapidAnalytics were submitted, which we were particularly pleased to see since only little more than three months lay between the public release of its first version and the paper submission deadline.

Again we have been able to win excellent and renowned speakers for the keynote talks. Fionn Murtagh from the Science Foundation Ireland will talk about correspondence analysis and hierarchical clustering. His presentation is titled *"A Metric and Ultrametric Data Mining Platform for Analysis of Semantics and Change"*. Christian Kofler from the German Research Center for Artificial Intelligence (DFKI) will present his vision of how data mining will look like in the future: *"Towards Next-Generation Data Mining"*.

Training courses held on Tuesday and Friday will surround the technical programme. This year for the first time, it will be possible to participate in the certification exam to become a certified Rapid-I Analyst. Finally, we hope that you all join the new edition of "Who wants to be a data miner", the rapid process design duel. Will you challenge Rapid-I data miners to be the fastest to solve tricky data mining problems?

We hope that you enjoy the programme and the presentations and utilize the time between the presentations for exchanging ideas and establishing new collaborations!

June 2011

Simon Fischer
Head of Research and Development
Rapid-I GmbH

Contents

Part I – Keynote Talks 1

1 A Metric and Ultrametric Data Mining Platform for Analysis of Semantics and Change
Fionn Murtagh **1**

2 Towards Next-Generation Data Mining
Christian Kofler **3**

Part II – Data Transformation and Analysis 5

3 Feature Selection for High-Dimensional Data with RapidMiner
Benjamin Schowe **5**

4 Instance Selection and Prototype Based Rules. A New Extension to RapidMiner
Marcin Blachnik, Mirosław Kordos **21**

5 Radoop: Analyzing Big Data with RapidMiner and Hadoop
Zoltán Prekopcsák, Gábor Makrai, Tamás Henk, Csaba Gáspár-Papanek **31**

6 Cluster Evaluation of Highly Dimensional Data
Andrew Chisholm, Geraldine Gray **43**

Part III – Applications 63

7 An Automated Stock Price Prediction System with RapidAnalytics
Gábor Nagy, Csaba Gáspár-Papanek **63**

8 Combining RapidMiner Operators with Bioinformatics Services – A Powerful Combination
Simon Jupp, James Eales, Robert Stevens, Alan Williams, Rishi Ramgolan, Simon Fischer, Sebastian Land **75**

9 Electricity Customer Classification Using WhiBo Decision Trees
Milan Vukićević, Marin Matijaš **87**

10 Constructing Recommender Systems Workflow Templates in RapidMiner
Matko Bošnjak, Nino Antulov-Fantulin, Tomislav Šmuc, Dragan Gamberger **101**

11 Application of RapidMiner in the Medical Domain – Constructive Induction and Cellular Automata Model
Matej Mertik **113**

Part IV – Extensions **123**

12 Rapid Image Feature Extraction and Mining
Radim Burget, Vaclav Uher, Radek Cervenec **123**

13 Optimizing Performance of Decision Tree Component-based Algorithms Using Evolutionary Algorithms in RapidMiner
Miloš Jovanović, Boris Delibašić, Milan Vukićević, Milija Suknović . **135**

Part V – Text Mining and Information Extraction **150**

14 Handling Tree-Structured Values in RapidMiner
Felix Jungermann **151**

15 Supervised Learning Methods for Sentiment Classification with RapidMiner
Bruno Ohana, Brendan Tierney **163**

16 Probabilistic Topic Models in the Domain of Customer Sentiment Analysis
Clemens Forster **171**