



Universiteit
Leiden
The Netherlands

Exceptional Model Mining

Duivesteijn, W.

Citation

Duivesteijn, W. (2013, September 17). *Exceptional Model Mining*. Retrieved from <https://hdl.handle.net/1887/21760>

Version: Corrected Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/21760>

Note: To cite this publication please use the final published version (if applicable).

Cover Page



Universiteit Leiden



The handle <http://hdl.handle.net/1887/21760> holds various files of this Leiden University dissertation.

Author: Duivesteijn, Wouter

Title: Exceptional model mining

Issue Date: 2013-09-17

Exceptional Model Mining

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof.mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op dinsdag 17 september 2013
klokke 11.15 uur

door

Wouter Duivesteijn

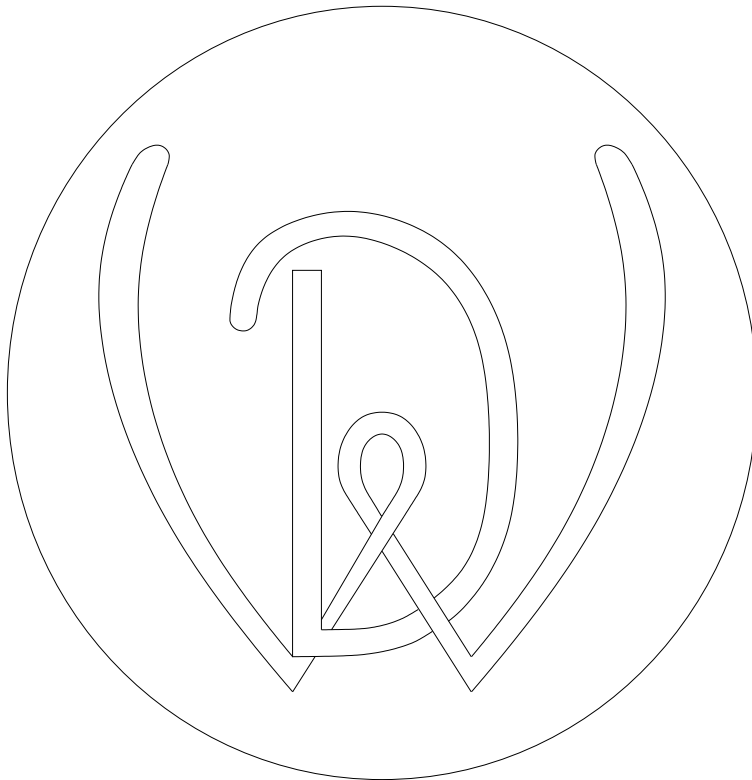
geboren te Rotterdam
in 1984

Promotiecommissie

Promotor: prof. dr. J. N. Kok
Co-promotor: dr. A. J. Knobbe
Overige leden: prof. dr. P. A. Flach (University of Bristol)
prof. dr. H. Blockeel (Katholieke Universiteit Leuven)
dr. W. A. Kusters

Cover photo: ochre sea stars (*Pisaster ochraceus*), taken at Ganges Harbour, Salt Spring Island, British Columbia, Canada. Licensed under the Creative Commons Attribution-Share Alike 3.0 Unported license by D. Gordon E. Robertson.

This research is financially supported by the Netherlands Organisation for Scientific Research (NWO) under project number 612.065.822 (Exceptional Model Mining).



Aan mijn grootouders, in liefdevolle herinnering.

Contents

1	Introduction	1
1.1	Overview	4
2	Motivation and Preliminaries	7
2.1	Preliminaries	10
3	The Exceptional Model Mining Framework	13
3.1	Search Strategy	15
3.1.1	Refinement Operator and Description Language . .	16
3.1.2	Beam Search Algorithm for Top-q EMM	18
3.1.3	Alternatives to Beam Search	21
3.2	How to Define an EMM Instance?	22
3.2.1	Quality Measure Concepts	22
3.2.2	Compared to what?	24
3.3	Related Work	26
3.3.1	Search Strategies for SD/EMM	26
3.3.2	Similar Local Pattern Mining Tasks	27
3.3.3	Similar Tasks with a Broader Scope	29
3.4	Software	31
4	Deviating Interactions – Correlation Model	33
4.1	Quality Measure φ_{scl}	33
4.2	Experiments	34
4.2.1	Datasets	34
4.2.2	Experimental Results	35
4.3	Alternatives	38
4.4	Conclusions	40

5	Deviating Predictive Performance – Classification Model	41
5.1	Quality Measure φ_{sed}	42
5.2	Experiments	42
5.2.1	Datasets	42
5.2.2	Experimental Results	42
5.3	Alternatives	43
5.3.1	BDeu Score (φ_{BDeu})	44
5.3.2	Hellinger (φ_{Hel})	44
5.3.3	Experimental Results	45
5.4	Conclusions	47
6	Unusual Conditional Interactions – Bayesian Network Model	49
6.1	Quality Measure φ_{weed}	50
6.1.1	Independence Relations in Bayesian Networks	51
6.1.2	Edit Distance for Bayesian Networks	52
6.2	Experiments	54
6.2.1	Datasets	54
6.2.2	Experimental Results	55
6.3	Alternatives	63
6.4	Conclusions	66
7	Different Slopes for Different Folks – Regression Model	69
7.1	Quality Measure φ_{Cook}	70
7.2	Experiments	73
7.2.1	Datasets	73
7.2.2	Experimental Results	76
7.3	Pruning with Bounds for Cook’s Distance	80
7.3.1	Empirical bound evaluation	83
7.4	Alternatives	86
7.5	Conclusions	87
8	Exploiting False Discoveries – Validating Found Descriptions	91
8.1	Problem Statement	92
8.2	Validation Method	93
8.2.1	Randomization Techniques	94
8.2.2	Building a Statistical Model	96
8.2.3	Comparing Quality Measures	97

8.3	Experiments	97
8.3.1	Validating Descriptions	101
8.3.2	Validating Quality Measures	102
8.3.3	Validating EMM Results	105
8.4	Discussion	107
8.4.1	Validating Descriptions	108
8.4.2	Validating Quality Measures	108
8.4.3	Validating EMM Results	110
8.5	Related Work	110
8.6	Conclusions	112
9	Multi-label LeGo – Enhancing Multi-label Classifiers with Local Patterns	115
9.1	The LeGo Framework	116
9.2	Multi-label Classification	118
9.3	LeGo Components	120
9.3.1	Local Pattern Mining Phase	120
9.3.2	Pattern Subset Discovery Phase	120
9.3.3	Global Modeling Phase	122
9.4	Experimental Setup	123
9.4.1	Evaluation Measures	124
9.4.2	Statistical Testing	125
9.5	Experimental Evaluation	126
9.5.1	Feature Selection Methods	126
9.5.2	Evaluation of the LeGo Approach	127
9.5.3	Evaluation of the Decompositive Approaches	131
9.5.4	Efficiency	133
9.6	Discussion and Related Work	134
9.7	Conclusions	136
10	Conclusions	139
	References	145
	Nederlandse Samenvatting	157
	English Summary	159
	Acknowledgments	161
	Curriculum Vitae	163

