

Contents

1	An Introduction to Outlier Ensembles	1
1.1	Introduction	1
1.1.1	Motivations for Ensemble Methods in Outlier Analysis	3
1.1.2	Common Settings for Existing Ensemble Methods	4
1.1.3	Types of Ensemble Methods	6
1.1.4	Overview of Outlier Ensemble Design	7
1.2	Categorization by Component Independence	8
1.2.1	Sequential Ensembles	9
1.2.2	Independent Ensembles	11
1.3	Categorization by Constituent Components	12
1.3.1	Model-Centered Ensembles	12
1.3.2	Data-Centered Ensembles	14
1.3.3	Discussion of Categorization Schemes	16
1.4	Categorization by Theoretical Approach	17
1.4.1	Variance Reduction in Outlier Ensembles	18
1.4.2	Bias Reduction in Outlier Ensembles	18
1.5	Defining Combination Functions	19
1.5.1	Normalization Issues	19
1.5.2	Combining Scores from Different Models	20
1.6	Research Overview and Book Organization	23
1.6.1	Overview of Book	27
1.7	Conclusions and Discussion	30
	References	31
2	Theory of Outlier Ensembles	35
2.1	Introduction	35
2.2	The Bias-Variance Trade-Off for Outlier Detection	37
2.2.1	Relationship of Ensemble Analysis to Bias-Variance Trade-Off	42
2.2.2	Out-of-Sample Issues	43

2.2.3	Understanding How Ensemble Analysis Works	44
2.2.4	Data-Centric View Versus Model-Centric View	50
2.3	Examples and Applications of the Bias-Variance Tradeoff	58
2.3.1	Bagging and Subsampling.	59
2.3.2	Feature Bagging	60
2.3.3	Boosting	61
2.4	Experimental Illustration of Bias-Variance Theory	61
2.4.1	Understanding the Effects of Ensembles on Data-Centric Bias and Variance	62
2.4.2	Experimental Examples of Bias-Variance Decomposition	68
2.5	Conclusions	72
	References.	73
3	Variance Reduction in Outlier Ensembles	75
3.1	Introduction	75
3.2	Motivations for Basic Variance Reduction Framework.	78
3.3	Variance Reduction Is Not a Panacea.	83
3.3.1	When Does Data-Centric Variance Reduction Help?	84
3.3.2	When Does Model-Centric Variance Reduction Help?	91
3.3.3	The Subtle Differences Between AUCs and MSEs	93
3.4	Variance Reduction Methods	93
3.4.1	Feature Bagging (FB) for High-Dimensional Outlier Detection.	94
3.4.2	Rotated Bagging (RB).	99
3.4.3	Projected Clustering and Subspace Histograms	100
3.4.4	The Point-Wise Bagging and Subsampling Class of Methods	107
3.4.5	Wagging (WAG).	130
3.4.6	Data-Centric and Model-Centric Perturbation	131
3.4.7	Parameter-Centric Ensembles	131
3.4.8	Explicit Randomization of Base Models	132
3.5	Some New Techniques for Variance Reduction	134
3.5.1	Geometric Subsampling (GS)	134
3.5.2	Randomized Feature Weighting (RFW)	136
3.6	Forcing Stability by Reducing Impact of Abnormal Detector Executions	137
3.6.1	Performance Analysis of Trimmed Combination Methods	140
3.6.2	Discussion of Commonly Used Combination Methods	143
3.7	Performance Analysis of Methods	145
3.7.1	Data Set Descriptions	145
3.7.2	Comparison of Variance Reduction Methods	147

3.8	Conclusions	157
	References.	158
4	Bias Reduction in Outlier Ensembles: The Guessing Game	163
4.1	Introduction	163
4.2	Bias Reduction in Classification and Outlier Detection.	165
4.2.1	Boosting	166
4.2.2	Training Data Pruning	167
4.2.3	Model Pruning	168
4.2.4	Model Weighting	169
4.2.5	Differences Between Classification and Outlier Detection.	170
4.3	Training Data Pruning	171
4.3.1	Deterministic Pruning	171
4.3.2	Fixed Bias Sampling	172
4.3.3	Variable Bias Sampling.	174
4.4	Model Pruning	175
4.4.1	Implicit Model Pruning in Subspace Outlier Detection	178
4.4.2	Revisiting Pruning by Trimming	178
4.4.3	Model Weighting	180
4.5	Supervised Bias Reduction with Unsupervised Feature Engineering	181
4.6	Bias Reduction by Human Intervention	182
4.7	Conclusions	184
	References.	184
5	Model Combination Methods for Outlier Ensembles	187
5.1	Introduction	187
5.2	Impact of Outlier Evaluation Measures.	190
5.3	Score Normalization Issues.	193
5.4	Model Combination for Variance Reduction.	195
5.5	Model Combination for Bias Reduction	196
5.5.1	A Simple Example	198
5.5.2	Sequential Combination Methods	199
5.6	Combining Bias and Variance Reduction	200
5.6.1	Factorized Consensus	201
5.7	Using Mild Supervision in Model Combination	203
5.8	Conclusions and Summary	204
	References.	204
6	Which Outlier Detection Algorithm Should I Use?	207
6.1	Introduction	207
6.2	A Review of Classical Distance-Based Detectors	212
6.2.1	Exact k -Nearest Neighbor Detector.	213
6.2.2	Average k -Nearest Neighbor Detector.	214

6.2.3	An Analysis of Bagged and Subsampled 1-Nearest Neighbor Detectors	214
6.2.4	Harmonic k -Nearest Neighbor Detector.	216
6.2.5	Local Outlier Factor (LOF).	217
6.3	A Review of Clustering, Histograms, and Density-Based Methods	219
6.3.1	Histogram and Clustering Methods	219
6.3.2	Kernel Density Methods	224
6.4	A Review of Dependency-Oriented Detectors.	225
6.4.1	Soft PCA: The Mahalanobis Method	226
6.4.2	Kernel Mahalanobis Method	231
6.4.3	Decomposing Unsupervised Learning into Supervised Learning Problems	239
6.4.4	High-Dimensional Outliers Based on Group-Wise Dependencies	242
6.5	The Hidden Wildcard of Algorithm Parameters	243
6.5.1	Variable Subsampling and the Tyranny of Parameter Choice.	245
6.6	TRINITY: A Blend of Heterogeneous Base Detectors	247
6.7	Analysis of Performance.	248
6.7.1	Data Set Descriptions	249
6.7.2	Specific Details of Setting.	251
6.7.3	Summary of Findings	253
6.7.4	The Great Equalizing Power of Ensembles	260
6.7.5	The Argument for a Heterogeneous Combination	264
6.7.6	Discussion.	269
6.8	Conclusions	271
	References.	271
	Index	275

Outlier Ensembles

An Introduction

Aggarwal, C.C.; Sathe, S.

2017, XVI, 276 p. 55 illus., 9 illus. in color., Hardcover

ISBN: 978-3-319-54764-0