Anomaly Detection with SECODA

Poster presentation of "SECODA: Segmentation- and Combination-Based Detection of Anomalies" at IEEE DSAA 2017

dr. Ralph Foorthuis, UWV, the Netherlands, ralph.foorthuis@uwv.nl

IEEE DSAA 2017
International Conference on Data Science and Advanced Analytics
IEEE / ACM / ASA
DSAA'2017
19-21, Oct 2017, Tokyo, JAPAN

Introduction

SECODA is a novel general-purpose unsupervised non-parametric anomaly detection (AD) algorithm for datasets containing continuous and categorical attributes. The method is guaranteed to identify cases with unique or sparse combinations of attribute values.

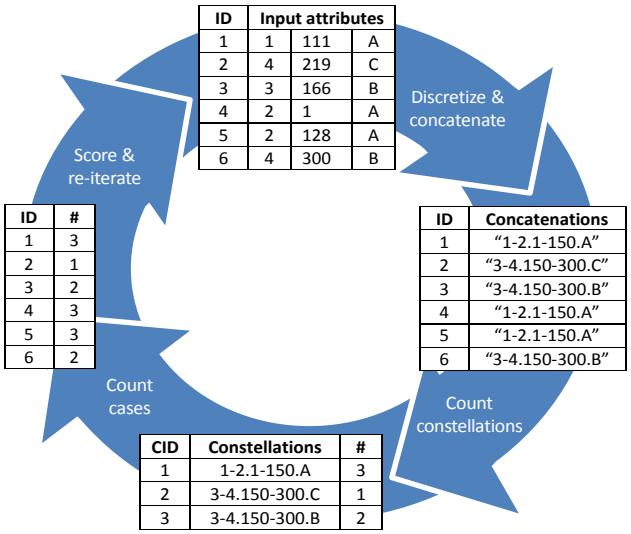
Typology of anomalies

The typology presents an overview of the types of anomalies. It provides a theoretical and tangible *understanding* of the anomaly types an analyst may encounter. It also aids in *evaluating* which types of anomalies can be detected by a given AD algorithm. The typology differentiates between the set's 'awkward cases' by means of two dimensions: The data types taken into account and the number of attributes analyzed jointly.

		Nature of the data	
		Numerical attributes	**Categorical or mixed attributes**
Number of attributes	**Univariate** Focus on individual attributes (independence)	Type I Extreme value anomaly	Type II Sparse class anomaly
	Multivariate Focus on multi-dimensionality (interactions)	Type III Multidimensional numerical anomaly	Type IV Multidimensional mixed data anomaly

The SECODA algorithm

SECODA uses the histogram-based approach to assess the density. The concatenation trick – which combines discretized continuous attributes and categorical attributes into a new variable – is used to determine the joint density distribution. In combination with recursive discretization this captures complex relationships between attributes and avoids discretization error. A pruning heuristic as well as exponentially increasing weights and arity are employed to speed up the analysis.



ID	Input attributes		
1	1	111	A
2	4	219	C
3	3	166	B
4	2	1	A
5	2	128	A
6	4	300	B

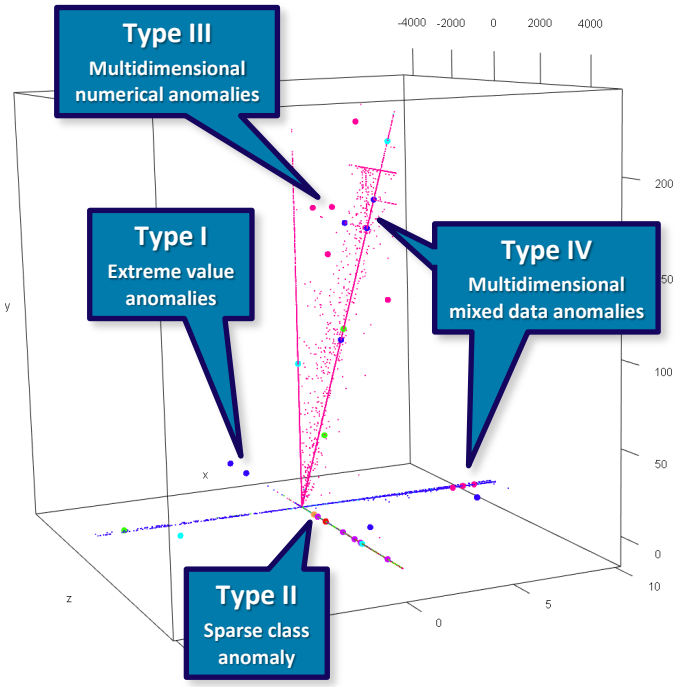
ID	Concatenations
1	"1-2.1-150.A"
2	"3-4.150-300.C"
3	"3-4.150-300.B"
4	"1-2.1-150.A"
5	"1-2.1-150.A"
6	"3-4.150-300.B"

CID	Constellations	#
1	1-2.1-150.A	3
2	3-4.150-300.C	1
3	3-4.150-300.B	2

ID	#
1	3
2	1
3	2
4	3
5	3
6	2

Evaluation

SECODA was evaluated with multiple simulated and real-world datasets. The diagram below is a 4D snapshot of income data from the Polis Administration, an official national data register in the Netherlands. As can be seen, SECODA was able to detect all four types of anomalies.



Some characteristics of SECODA:

- Simple algorithm without the need for point-to-point calculations. Only basic data operations are used, making SECODA suitable for sets with large numbers of rows as well as for in-database analytics.
- The pruning heuristic, although simple by design, is a self-regulating mechanism during runtime, dynamically deciding how many cases to discard.
- The exponentially increasing weights both speed up the analysis and prevent bias.
- The algorithm has low memory requirements and scales linearly with dataset size.
- In addition, the real-world data quality use case not only shows that all types of anomalies can be detected, but also that they can be encountered in practice.

Download R code and data examples from www.foorthuis.nl

Literature

[1] C.C. Aggarwal, "Outlier Analysis," New York: Springer, 2013.
[2] H. Liu, F. Hussain, C.L. Tan, M. Dash, "Discretization: An Enabling Technique," Data Mining and Knowledge Discovery, Vol. 6, 2002.
[3] LAK, "Loonaangifteketen," 2017, URL: www.loonaangifteketen.nl/
[4] R. Foorthuis, "SECODA: Segmentation- and Combination-Based Detection of Anomalies," IEEE DSAA 2017 Conference.

© Ralph Foorthuis 2017 DOI: 10.13140/RG.2.2.21212.08325