Fitness Quality of the Corpus' Projects to their Respective Time-Related Patterns of Schema Evolution

P. Vassiliadis, A. Karakasidis [2024-08-28]

Given the extraction of Time-Related Patters of Schema Evolution and the respective classification of the 151 projects of the corpus with respect to them, we need to validate that the classification schema is of adequate quality. The present report addresses the following validation question:

VQ2: Can we claim that the classification of projects into different patterns is producing patterns that are (a) pairwise disjoint and (b) internally cohesive?

1 Disjointness

We claim that our patterns are both (i) disjoint and (ii) cover significantly (not fully) the space of possible behaviors.

Disjointedness requires that the different patterns are essentially different with one another. It is straightforward to verify that our classification scheme attains disjointedness, as the formal definitions cover disjoint areas in the space produced by the Cartesian Product of values for the defining attributes. However, apart from formally, the pattern set is also hiding an inherent, essential disjointedness.

The essential disjointedness is depicted in Figure 1.1 reporting the placement of patterns in the active domain space formed by the Cartesian Product of the domains of the involved class-based metrics. We report only the active domain that contains only the combinations which are populated by projects, and, for each point of this multidimensional space, we report the number of projects that pertain to it. So, for each part of the active domain, we show how many projects "live" there, and to which pattern they belong. With the exception of (a) a couple of Siesta projects being born early that is overlapping with Regularly Curated projects of similar definition, and, (b) the Quantum Steps and Regularly Cu-rated patterns that span a large area of the domain space of values (understandably, as they are produced by the union of similar sub-families) which they would share had they not being discriminated by change rate, the rest of the patterns are focused in a specific area of the domain space and disjoint from the others. Quantum Steps and Regularly Curated are still disjoint; we just point out that they are the only patterns practically separated by change rate in their growth period.

As a side-effect of the validation of pattern disjointedness, we have extracted a simple decision tree from the labeled values of their properties, after the manual annotation had been completed (Figure 1.2). The simple classification tree shows that the patterns can be fairly well (although not 100%) separated automatically, with only 4 out of 151 projects that would have been erroneously classified under this classification scheme.

PointBirthClass	PointTopBandClass	IntervalBirth-To-TopBand Class	GrowthMonthsWithChange	11_FlatLiner	12_RadicalSign	13_Sigmoid	14_LateRiser	21_QuantumSteps	22_RegularlyCurated	31_SmokingFunnel	32_Siesta	Σ
0_V0	**0_V0**	0_Zero	0-3	23								23
	1_early	1_soon	0-3		8							8
		2_fair	0-3		7							7
			> 3		1							1
	2_middle	2_fair	0-3					2				2
		3_long	0-3					2				2
			> 3						2			2
	3_late	4_vlong	0-3								6	6
			> 3						1			1
1_earl	**1_early**	0_Zero	0-3		15							15
		1_soon	0-3		9							9
		2_fair	0-3		1							1
	2_middle	1_soon	0-3			2						2
		2_fair	0-3					5				5
			> 3						1			1
		3_long	0-3					7				7
			> 3						1			1
	3_late	3_long	0-3					1			1	2
			> 3						3			3
		4_vlong	0-3								1	1
			> 3						3		2	5
2_mid	**2_middle**	0_Zero	0-3			12						12
		1_soon	0-3			5						5
		2_fair	> 3							7		7
		3_long	0-3					1				1
	3_late	1_soon	0-3				1					1
		2_fair	0-3					2				2
			> 3						1			1
		3_long	0-3					3				3
			> 3						2			2
3_late	**3_late**	0_Zero	0-3				12					12
		1_soon	> 3				1					1
Σ				**23**	**41**	**19**	**14**	**23**	**14**	**7**	**10**	**151**

Figure 1.1 Coverage of the space of possible values by the time-related patterns of schema evolution

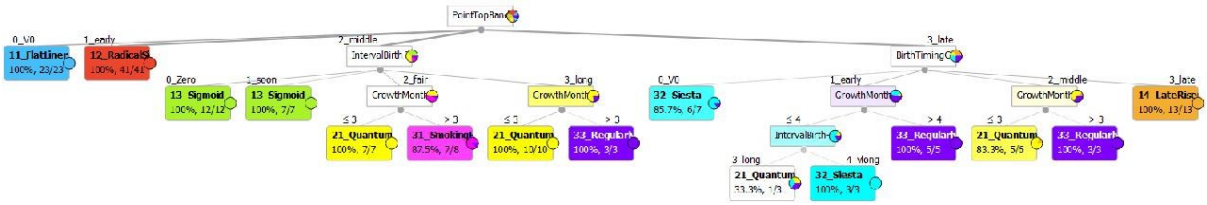

Figure 1.2 Schema evolution time-related patterns classified via a decision tree

2 Cohesion

Pattern cohesion refers to the internal homogeneity of the projects that pertain to each pattern. We investigate cohesion with respect to three different ways to assess it: (a) nominally, i.e., the extent to which pattern definitions have exceptions in the corpus, (b) visually, i.e., to the extent that the visual depiction of the patterns demonstrates discrepancies, and, (c) quantitatively, by assessing the magnitude of discrepancies in the corpus of each pattern.

2.1 Exceptions

Table 2.1 presents the exceptions to the patterns with respect to their definition for our 151 projects. The 7 exceptions we found are too few and not particularly significant to question the cohesion of the patterns:

- Two projects classified as Sigmoid violate the "middle-born" part of the definition by being born early.
- A Late Riser project reaches the top band in middle life, violating the requirement of late attainment of top-band.
- Siesta has 2 projects exceeding the 0-3 months growth activity in the end, and another project that reaches growth just 'long' after schema birth (and not 'very long').
- A Quantum Step project reaches top late rather than middle.

Pattern	#prjs	Exceptions	Overlaps
Flatliner	23	–	–
Radical Sign	41	–	–
Sigmoid	19	2	–
Late Riser	14	1	–
Quantum Steps	23	2	–
Regularly Curated	14	–	–
Smoking Funnel	7	–	–
Siesta	10	3	–

Table 2.1 Exceptions and Overlaps of the Definitions of Schema Evolution time-related patterns

2.2 Visual Inspection

Pattern cohesion refers to the internal homogeneity of the projects that pertain to each pattern. Our first attempt towards producing cohesive patterns was via visual inspection.

The visual inspection of the schema evolution progress of the different patterns is quite revealing. Although the production of the patterns was performed in an iterative way, we have also created a collective visual representation for each pattern to demonstrate the similarity of its members.

We believe the visualization clearly shows the validity of the derived patterns. We will complement this visual-based classification with concrete measures in the sequel.

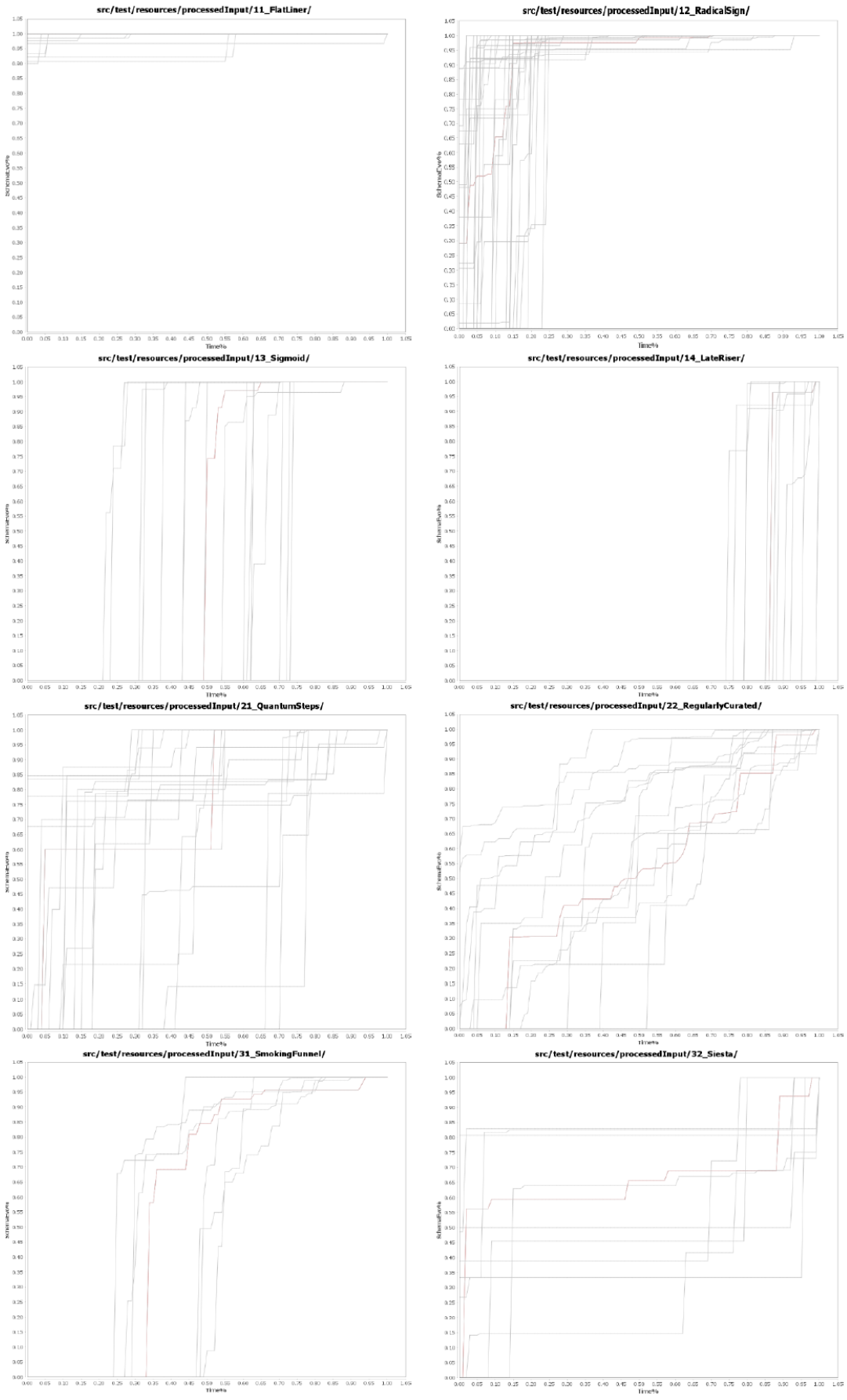

Figure 2.1Visualization of all the patterns

2.3 Metrics for Inter-Pattern Cohesion

Beyond visual inspection, we have employed quantitative measures to illustrate in a quantitative way the validity of the proposed patterns and the consequent categorization. As such, we have employed measures of cohesion, in its clustering pattern-based definition, which refers to the internal homogeneity of the projects that pertain to each pattern. Cohesion indicates how dense are a cluster's points and closer to its center. As such, for a cluster, it would be desired to illustrate lower values.

To apply these measures, we have quantized each project's time series to a vector of 20 measurements, one for each interval of 5% of time (i.e., at 0%, 5%, 10%, . . . of time), and computed the centroid for the corpus of each pattern. In what follows, we will use the following terminology:

- C_i refers to cluster *i*, with a center c_i and m_i data points
- *x* is a data point assigned to C_i
- *dist(a,b)* is a distance function between two data points, which, unless stated otherwise, we assume to be the Euclidean distance

The first measure to employ is the *Sum of Square Errors* (*SSE*), a cluster validation metric that is most frequently used in the data mining literature [Agga15], [TaSK05]:

$$SSE(C_i) = \sum_{x \in C_i} dist(x, c_i)^2$$

Table 4.1 depicts the values of SSE for the different patterns. The values range from 0.18 to 36.20. An issue here is that we are dealing with points in the 20-dimensional space and patterns of different size, therefore, the intuition behind these numbers is not clear.

Pattern	#prjs	SSE	MDC
11_FlatLiner	23	0.18	0.06
12_RadicalSign	41	23.25	0.70
13_Sigmoid	19	31.76	1.26
14_LateRiser	14	9.33	0.78
21_QuantumSteps	23	36.20	1.16
22_RegularlyCurated	14	12.64	0.90
31_SmokingFunnel	7	5.00	0.83
32_Siesta	10	9.15	0.91

Table 2.2 Cluster properties.

To come up with an intuitive measure of cohesion, we introduce the *Mean Distance to Centroid* (*MDC*) measure, formally defined as follows:

$$MDC(C_i) = \frac{1}{m_i} \sum_{x \in C_i} dist(x, c_i)$$

The MDC is an intuitive measure that represents the average distance (and not *squared* distance as in most definitions of intra-cluster variance, like SSE) of an arbitrary point in the cluster to the cluster's centroid. The MDC for the different projects ranges from 0.06 to 1.25, which is reasonable for vectors of 20 measurements in [0. .1]. These results are illustrated in Table 4.1 too.

Overall, we judge that the cohesion of the different patterns, as MDC computes, is reasonable enough to justify the validity of the clusters.

3 Conclusions

The answer to the original validation question is as follows:

- *Yes, the patterns are indeed disjoint, both in terms of definitions and essential separation.*
- *Yes, the patterns are internally cohesive in various ways: (a) in terms of few exceptions, (b) in terms of visual inspection, and, (c) in terms of their Mean Distance to Centroid measure*

References

[Agga15] Charu C. Aggarwal. Data Mining - The Textbook. Springer 2015, ISBN 978-3-319-14141-1

[TaSK05] Pang-Ning Tan, Michael S. Steinbach, Vipin Kumar. Introduction to Data Mining. Addison-Wesley 2005, ISBN 0-321-32136-7