Pythagoras: Semantic Type Detection of Numerical Data in Enterprise Data Lakes (Technical Report)

Sven Langenecker
LÄPPLE AG & DHBW Mosbach
s.langenecker@laepple.de

Christoph Sturm
DHBW Mosbach
christoph.sturm@mosbach.dhbw.de

Christian Schalles
DHBW Mosbach
christian.schalles@mosbach.dhbw.de

Carsten Binnig
Technical University of Darmstadt & DFKI
carsten.binnig@cs.tu-darmstadt.de

1 LIST OF FEATURES

Table 1 shows the complete list of the features that were extracted from a numerical column to build the representation of the graph node V_{ncf}. Except for the last feature listed, all features are also included in the model of [1].

Table 1: Features of a numerical column for graph node V_{ncf}

Feature	#
Character-level distribution (any, all, mean, variance, min, max, median, sum, kurtosis, skewness of the digits 0-9 and the symbols comma, point, plus, minus, blank)	150
Number of values	1
Column entropy	1
Fraction of values with unique content	1
Fraction of values with numerical characters	1
Fraction of values with alphabetical characters	1
Mean and std. of the number of numerical characters in cell-values	2
Mean and std. of the number of alphabetical characters in cell-values	2
Mean and std. of the number special characters in cell-values	2
Mean and std. of the number of words in values	2
Percentage, count, only/has-Boolean of the None values	4
Stats, sum, min, max, median, mode, kurtosis, skewness, any/all-Boolean of length of values	10
Column values statistics (min, max, mean, median, 8*quantile, mode, skewness, kurtosis of all column values)	15
Total	**192**

2 PERFORMANCE BREAKDOWN

Figure 1 and Figure 2 displays the F1-Score to individual semantic types of the SportsTables dataset achieved by our model *Pythagoras* in comparison to Sato. Figure 1 reports the types in which *Pythagoras* reached higher values than Sato, sorted by the largest performance difference and restricted to the top 30 types. In the same way, Figure 2 shows the top 30 types where Sato performed better. We chose to compare against Sato in this detailed analysis because Sato was the best existing model on the SportsTables dataset for predicting the types on numerical columns. As we see in Figure 1, there are many types for which *Pythagoras* performs much better than Sato. For these types, Sato can only achieve very low F1-Scores and the differences to our model are significant. In the other case, we can see in Figure 2 that for most types where Sato performs better, our model *Pythagoras* scores only slightly worse. At this point, the gaps are not that dominant.

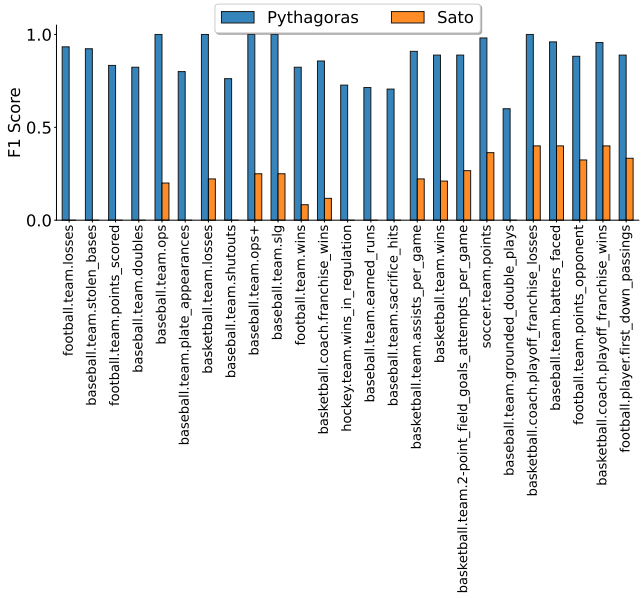

Figure 1: F1-Scores for individual semantic types assigned to numerical table columns where our model *Pythagoras* performs better than the existing model Sato. It shows the top 30 semantic types from left to right sorted by the descending value of the F1-Score difference.

EDBT/ICDT 2024, 25th March - 28th March, 2024, Paestum, Italy
2018. ACM ISBN 978-1-4503-XXXX-X/18/06.
https://doi.org/XXXXXXX.XXXXXXX

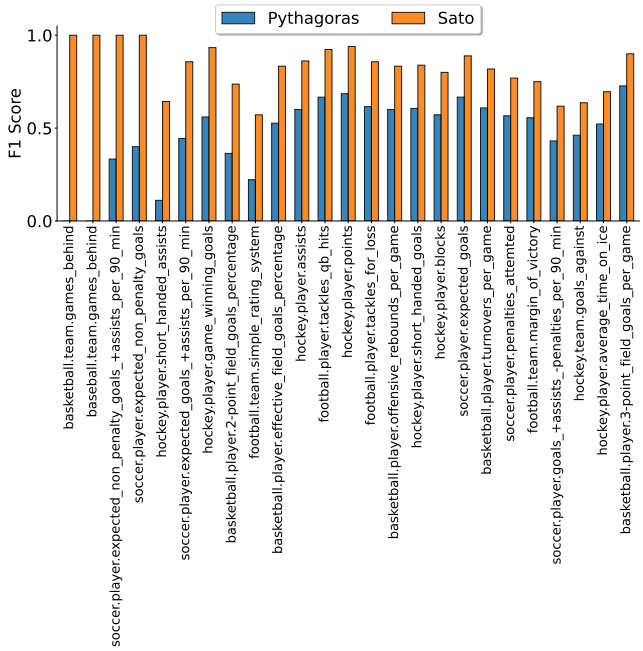

Figure 2: F1-Scores for individual semantic types assigned to numerical table columns where the existing model Sato performs better than our model *Pythagoras*. It shows the top 30 semantic types from left to right sorted by the descending value of the F1-Score difference.

REFERENCES

[1] Madelon Hulsebos, Kevin Hu, Michiel Bakker, Emanuel Zgraggen, Arvind Satyanarayan, Tim Kraska, Çagatay Demiralp, and César Hidalgo. 2019. Sherlock: A Deep Learning Approach to Semantic Data Type Detection. In *SIGKDD* (Anchorage, AK, USA) *(KDD '19)*. ACM, New York, NY, USA, 1500–1508. https://doi.org/10.1145/3292500.3330993