# Neurosymbolic Association Rule Mining from Tabular Data


Erkan Karabulut (e.karabulut@uva.nl), Paul Groth, Victoria Degeler

University of Amsterdam


Presented @ NeSy 2025

INDE lab

University of Amsterdam

## 1 Learning Rules?

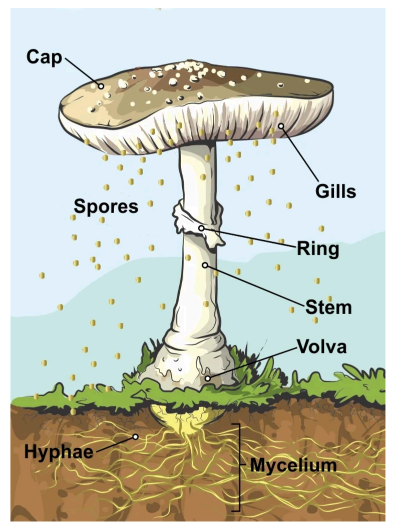
**Knowledge discovery**: Reveal associations between data features, e.g., columns of a given table.
**Interpretable inference**: Draw conclusions using learned rules instead of black box models, such as classification rules.

**Formalization**: Table with $k$ features $F = \{f_1, ..., f_k\}$, each with categories $f_i^1, ..., f_i^{c_i}$. Define the item universe $I = \left\{ f_i^j \mid 1 \leq i \leq k, 1 \leq j \leq c_i \right\}$.
Each row (transaction, $n$) $T \subset I$ satisfies $\forall i \in \{1, ..., k\}, \exists! j \in \{1, ..., c_i\}, f_i^j \in T$
An association rule is $X \to Y$ with $X, Y \subset I$, $X \cap Y = \emptyset$, $|Y| = 1$.
Logical form: $X \to Y \equiv (\neg \wedge_{x \in X}\, x) \vee y$ (Horn clause in CNF).

**Mushroom example:**

| cap-shape | cap-surface | odor | ... | poisonous |
|---|---|---|---|---|
| b | y | l | | e |
| x | y | p | ... | p |
| b | s | l | | e |

https://archive.ics.uci.edu/dataset/73/mushroom




https://grocycle.com/parts-of-a-mushroom/

cap-shape(b) ∧ cap-color(w) → odor(l)
cap-shape(b) ∧ cap-surface(y) → poisonous(e)

## 2 Research Question
How to address Combinatorial Explosion in Rule Mining?

**Intuition**: Even a small dataset can generate an overwhelming number of rules, most of which are redundant or trivial. Long execution times, harder to interpret. Existing methods are algorithmic, which rely on 'counting' co-occurrences.

**Formal**: For itemset universe $I$, each disjoint $X, Y \subset I, Y \neq \emptyset$ defines a rule $X \to Y$, with $|X| + |Y| \leq a$.

Feasible itemsets: $\prod_{i=1}^{a}(c_i + 1) - 1$

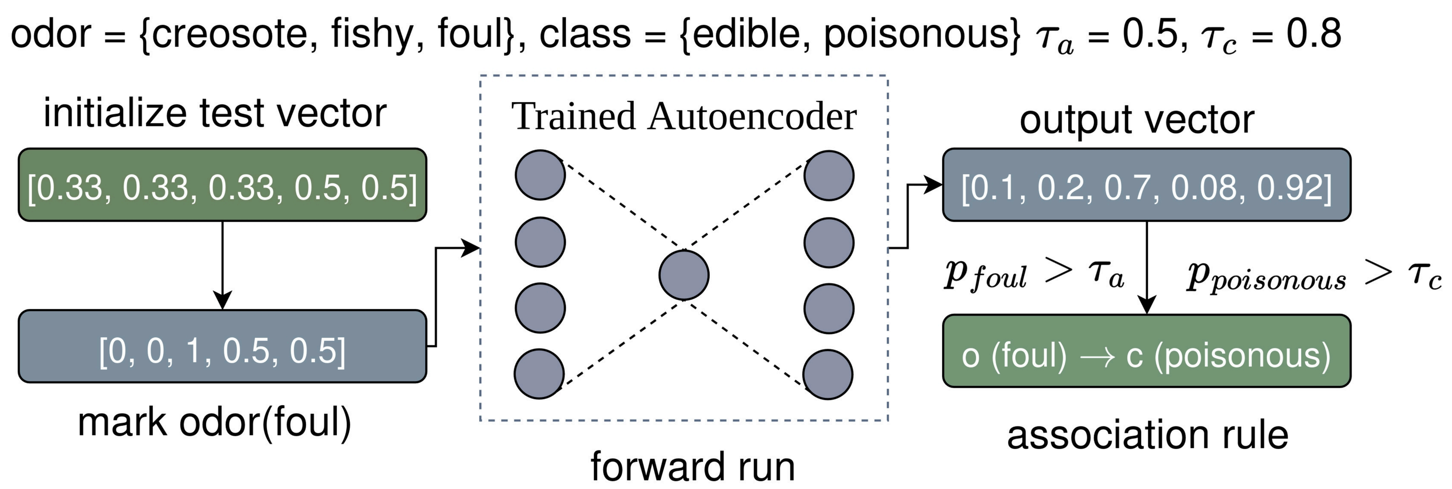
Number of rules: $\Sigma_{p=1}^{a} c_i \left( \prod_{i \neq p}(c_i + 1) - 1 \right)$

**Example**:
Table: 20 ($k$) columns $(f_1, ..., f_{20})$, 4 ($c_i$) values each $(q, r, t, y)$, and $a = 4$.
→ 5,186,240 rules!

Which rules to use? Hard to interpret, and unscalable on high-dimensional data.

## 3 Aerial+: Addressing Rule Explosion

**Intuition:** Autoencoders capture feature associations via reconstruction. If, after training, a forward pass with marked categories $A$ reconstructs categories $C$ with high probability, then $A \to C \setminus A$ (no self-implication).

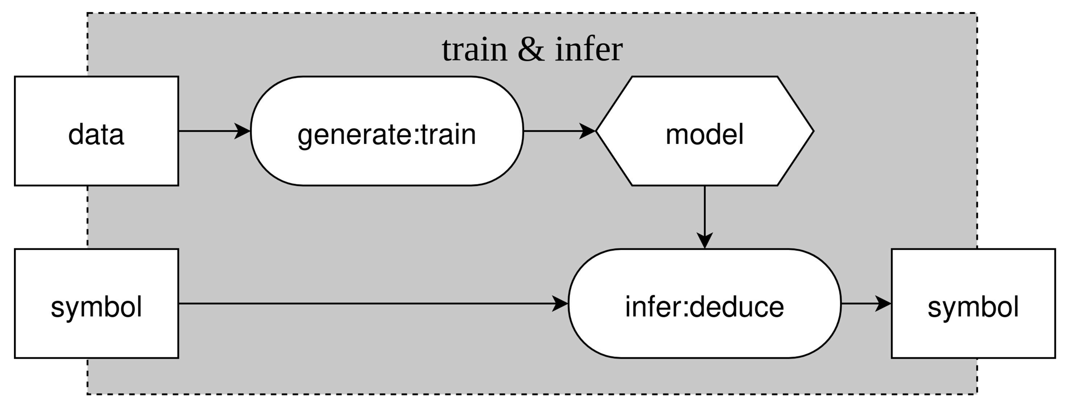



**Train to learn associations: shallow under-complete denoising Autoencoder**
Autoencoder Input: vectors of dim $\Sigma_{i=1}^{k} c_i$.
Noise: $N \sim [-0.5, 0.5]$ added per feature category $f_i^j$, clipped to $[0, 1]$.
Output: softmax per feature, values sum to 1 across categories.
Loss: per-feature BCE, aggregated as

$$BCE(F) = \Sigma_{i=1}^{k} \left( \frac{1}{c_i} \right) \Sigma_{j=1}^{c_i} - (y_{i,j} \log(p_{i,j}) + (1 - y_{i,j}) \log(1 - p_{i,j})),$$

with $p_{i,j} = \sigma\left(f_i^j\right), y_{i,j} =$ original (noise-free).

**Aerial+ is a Neurosymbolic approach:**




(Boxology), van Bekkum et al. 2021

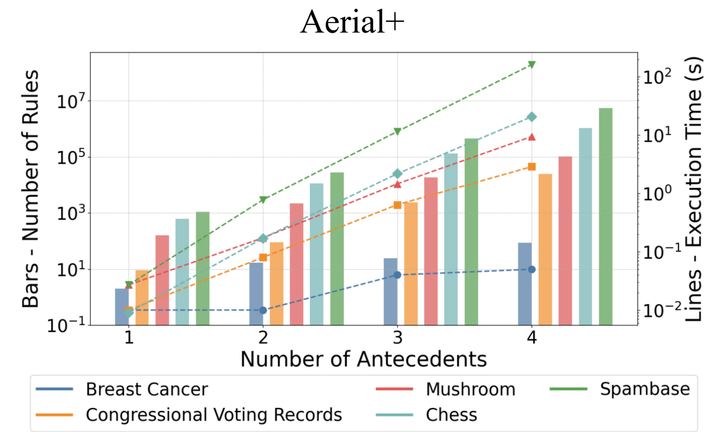
**Algorithm 1:** Aerial+'s rule extraction algorithm from a trained autoencoder

**Input:** Trained autoencoder: $AE$, max antecedents: $a$, similarity thresholds $\tau_a, \tau_c$
**Output:** Extracted rules $\mathcal{R}$

$\mathcal{R} \leftarrow \emptyset$, $\mathcal{F} \leftarrow AE.input\_feature\_categories$;
**for** $i \leftarrow 1$ **to** $a$ **do**
  $\mathcal{C} \leftarrow \binom{\mathcal{F}}{i}$;
  **foreach** $S \in \mathcal{C}$ **do**
    $\mathbf{v}_0 \leftarrow$ UniformProbabilityVectorPerFeature($\mathcal{F}$);
    $\mathcal{V} \leftarrow$ MarkFeatures($S, \mathbf{v}_0$)
    **foreach** $\mathbf{v} \in \mathcal{V}$ **do**
      $\mathbf{p} \leftarrow AE(\mathbf{v})$;
      **if** $\min_{f \in S} p_f < \tau_a$ **then**
        $S$.low_support ← **True**;
        **continue** with the next $\mathbf{v}$;
      **foreach** $f \in \mathcal{F} \setminus S$ **do**
        **if** $p_f > \tau_c$ **then** $\mathcal{R} \leftarrow \mathcal{R} \cup \{(S \to f)\}$
  $\mathcal{F} \leftarrow \{f \in \mathcal{F} \mid f.\text{low\_support} = \textbf{False}\}$;
**return** $\mathcal{R}$;

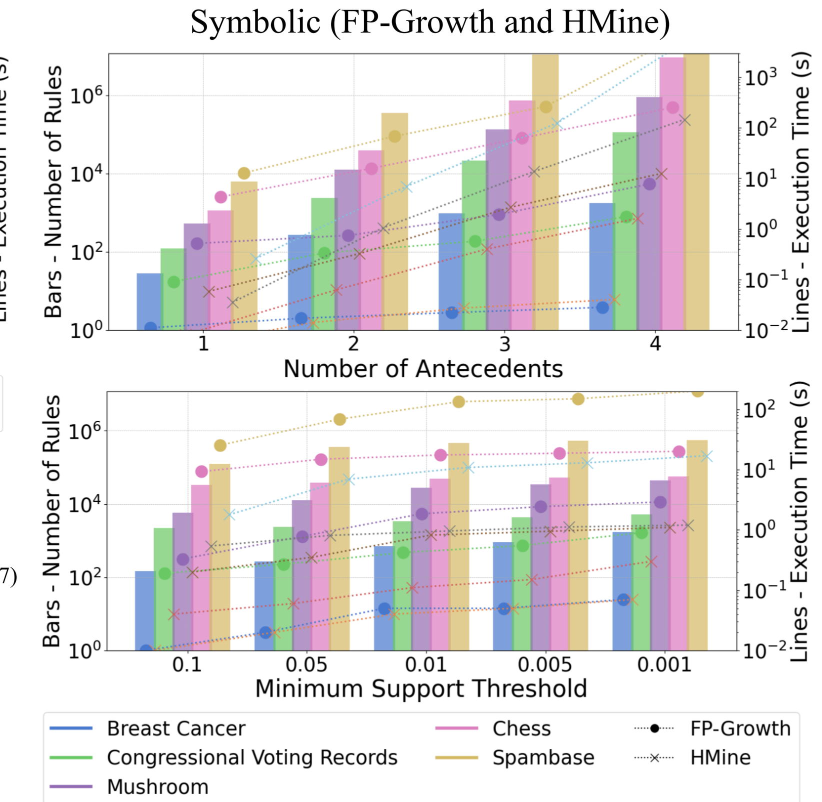
## 4 Validation

### Neurosymbolic rule learning is scalable




**Tabular datasets** from UCI ML repository (Kelly et al., 2023)

**NeSy baseline**: ARM-AE (Berteloot et al., 2024)
**Symbolic baselines**: FP-Growth, HMine (Han et al., 2000, Pei et al., 2007)
**Optimization-based baselines**: BAT, GW, SC, FSS (Heraguemi et al., 2015, Yildirim and Alatas, 2021, Altay and Alatas, 2021, Bharathi and Krishnakumari, 2014)




### Concise high-quality rule sets with full data coverage

| Algorithm | #Rules | Time (s) | Cov. | Support | Conf. |
|---|---|---|---|---|---|
| **Congressional Voting Records** | | | | | |
| BAT | 1913 | 208 | 1 | 0.06 | 0.45 |
| GW | 2542 | 186 | 1 | 0.05 | 0.48 |
| SC | 7 | 186 | 0.46 | 0.01 | 0.43 |
| FSS | 10087 | 272 | 1 | 0.01 | 0.71 |
| FP-G \| HMine | 1764 | 0.09 \| 0.04 | 1 | 0.29 | 0.88 |
| ARM-AE | 347 | 0.21 | 0.03 | 0.23 | 0.45 |
| **Aerial+** | 149 | 0.25 | 1 | 0.32 | **0.95** |
| **Mushroom** | | | | | |
| BAT | 1377.2 | 225.57 | 1 | 0.1 | 0.62 |
| GW | 1924.1 | 184.56 | 1 | 0.11 | 0.63 |
| SC | 1.33 | 281.84 | 0.07 | 0.02 | 0.48 |
| FSS | 794.9 | 352.99 | 1 | 0.04 | 0.38 |
| FP-G \| HMine | 1180 | 0.1 \| 0.07 | 1 | 0.43 | 0.95 |
| ARM-AE | 390 | 0.33 | 0 | 0.22 | 0.23 |
| **Aerial+** | 321 | 0.38 | 1 | 0.44 | **0.96** |
| **Spambase** | | | | | |
| BAT | 0 | 424 | No rules found | | |
| GW | 0 | 508 | No rules found | | |
| SC | 0 | 643 | No rules found | | |
| FSS | 0 | 677 | No rules found | | |
| FP-G \| HMine | 125223 | 21.4 \| 2.14 | 1 | 0.64 | 0.92 |
| ARM-AE | 85327 | 254 | 0.03 | 0.31 | 0.38 |
| **Aerial+** | 43996 | 1.92 | 1 | 0.62 | **0.97** |

| Algorithm | #Rules | Time (s) | Cov. | Support | Conf. |
|---|---|---|---|---|---|
| **Breast Cancer** | | | | | |
| BAT | 787.1 | 162.18 | 1 | 0.07 | 0.41 |
| GW | 1584 | 129.18 | 1 | 0.08 | 0.42 |
| SC | 33.6 | 137.66 | 1 | 0.03 | 0.27 |
| FSS | 6451.6 | 225.71 | 1 | 0.02 | 0.36 |
| FP-G \| HMine | 94 | 0.01 \| 0.01 | 1 | 0.34 | **0.87** |
| ARM-AE | 131 | 0.09 | 0.01 | 0.19 | 0.27 |
| **Aerial+** | 50 | 0.19 | 1 | 0.39 | 0.86 |
| **Chess** | | | | | |
| BAT | 2905.9 | 235.34 | 1 | 0.17 | 0.64 |
| GW | 5605.25 | 255.56 | 1 | 0.31 | 0.65 |
| SC | 1 | 545.71 | 0 | 0 | 0.7 |
| FSS | 32.75 | 380.73 | 0.4 | 0 | 0.36 |
| FP-G \| HMine | 30087 | 12.43 \| 0.7 | 1 | 0.46 | 0.93 |
| ARM-AE | 22052 | 26.98 | 0.02 | 0.39 | 0.54 |
| **Aerial+** | 16522 | 0.22 | 1 | 0.45 | **0.95** |

**Metrics:**
$\text{Supp}(X \to Y) = |\{T : X \cup Y \subseteq T\}| \,/\, n$
$\text{Conf.}(X \to Y) = |\{T : X \cup Y \subseteq T\}| \,/\, |\{T : X \subseteq T\}|$
$\text{Cov.}(X \to Y) = |\{T : X \subseteq T\}| \,/\, n$

### Concise rule sets improves downstream task performance

| Dataset | Algorithm | # Rules or Items<br>Exhaustive \| Aerial+ | Accuracy<br>Exhaustive \| Aerial+ | Exec. Time (s)<br>Exhaustive \| Aerial+ |
|---|---|---|---|---|
| Congressional Voting Records | CBA | 3437 \| **1495** | 91.91 \| **92.66** | 0.34 \| **0.14** |
| | BRL | 2547 \| **57** | **96.97** \| **96.97** | 15.37 \| **9.69** |
| | CORELS | 4553 \| **61** | **96.97** \| **96.97** | 3.04 \| **0.17** |
| Mushroom | CBA | 27800 \| **2785** | **99.82** \| **99.82** | 1.75 \| **1.30** |
| | BRL | 5093 \| **493** | **99.87** \| 99.82 | 244 \| **167** |
| | CORELS | 23271 \| **335** | 90.14 \| **99.04** | 61 \| **2** |
| Breast Cancer | CBA | 695 \| **601** | 66.42 \| **71.13** | **0.08** \| 0.28 |
| | BRL | 2047 \| **290** | 71.13 \| **71.46** | 16.82 \| **14.5** |
| | CORELS | 2047 \| **369** | 73.69 \| **75.82** | 1.42 \| **0.40** |
| Chess | CBA | 49775 \| **34490** | **94.02** \| 93.86 | 24.31 \| **6.24** |
| | BRL | 19312 \| **1518** | **96.21** \| 95.93 | 321 \| **119** |
| | CORELS | 37104 \| **837** | 81.1 \| **93.71** | 106 \| **3.87** |
| Spambase | CBA | 125223 \| **33418** | 84.5 \| **85.42** | 23.87 \| **7.56** |
| | BRL | 37626 \| **5190** | 72.78 \| **84.93** | 1169 \| **431** |
| | CORELS | 275003 \| **1409** | 85.37 \| **87.28** | 1258 \| **5.23** |

**Library**

**Paper**


Funded by the Dutch Research Council (NWO), in the scope of Digital Twin for Evolutionary Changes in water networks (DiTEC) project, file number 19454.