AutoFeat:
Transitive Feature Discovery over Join Paths

Andra Ionescu, Kiril Vasilev, Florena Buse, Rihan Hai, Asterios Katsifodimos

TU Delft Netherlands

AutoFeat Pipeline

Base Table, Dataset Repository

1 **Input data**

Data repository + Base table

Base Table

T0	Applicants
X1	Applicant_ID
X2	Name
X3	Social_security_number
X4	Age
Y	Loan approval (yes/no)

Dataset Repository

T2	Personal_information
X9	Social_security_number
X10	Nationality
X11	Date of birth
X12	Place of birth
X13	Zipcode

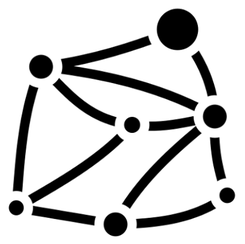
T4	Property value
X18	Property_ID
X19	Zipcode
X20	Street_name
X21	Neighbourhood
X22	Housing_value

T1	Credit_profile
X5	Credit_ID
X6	SSN
X7	Income
X8	Credit_score

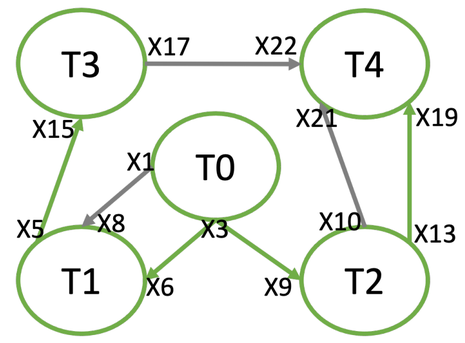
T3	Loan history
X14	Loan_ID
X15	Credit_ID
X16	Loan_type
X17	Loan_value

Dataset Discovery (e.g. Jaccard Similarity)

2 **Find relationships**



Dataset Relation Graph



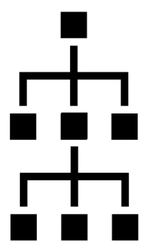
- Find relationships with Valentine [1].
- Create a weighted graph:
  - nodes -> tables
  - edges -> relations

BFS Traversal, Left Join & Prune Paths

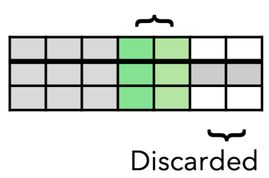
Relevance & Redundancy feature selection

3 **Streaming feature selection**

Create join trees



Select features



Join Trees

- BFS Traversal
- Left join
- Prune paths:
  - Similarity score
  - Null value ratio

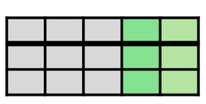
Feature Selection

- Relevance - Spearman
- Redundancy - MRMR
- Ranking - Linear function

Evaluate top-k ranked join trees

4 **Evaluate**

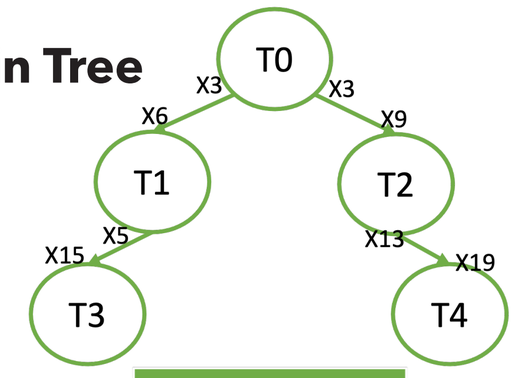
Augmented table



Evaluation

- Top-k join trees
  - Based on ranking
- Augment base table
- Train ML models

Join Tree



Applicants_augmented
Applicant_ID
Name
Social_security_number
Age
Income
Credit_score
Loan_type
Loan_value
Neighbourhood
Housing_value
Loan approval (yes/no)

AutoFeat Evaluation

8 Datasets: 7 OpenML, 1 SOTA

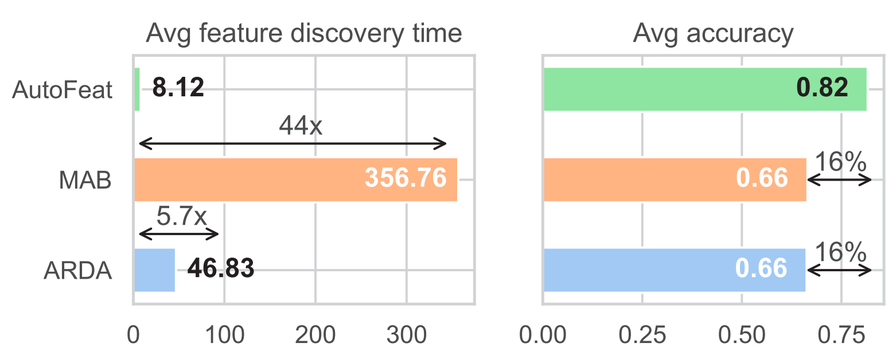
5 Baselines: Base, JoinAll, JoinAll + FS, ARDA [2], MAB [3]

4 ML models: AutoGluon decision trees

2 Scenarios:

- Benchmark - Snowflake schemata, known PK-FK
- Data Lake - Dense multi-graph, no PK-FK relations

2 Metrics: runtime, accuracy



Same accuracy as Join All(+FS) at a fraction of time
AutoFeat explores the join space in depth
10x faster than MAB, 3x faster than ARDA

16% average increase in accuracy across all datasets and models

[1] Christos Koutras, et al. "Valentine: Evaluating matching techniques for dataset discovery." 2021 ICDE
[2] Nadiia Chepurko, et al. "ARDA: Automatic Relational Data Augmentation for Machine Learning." 2020 VLDB
[3] Jiabin Liu, et al. "Feature augmentation with reinforcement learning." 2022 ICDE

https://github.com/delftdata/autofeat