University of Stuttgart
Institute for Parallel and Distributed Systems (IPVS)
Applications of Parallel and Distributed Systems

Towards Automated Data Preprocessing for Clustering: Challenges and Future Directions

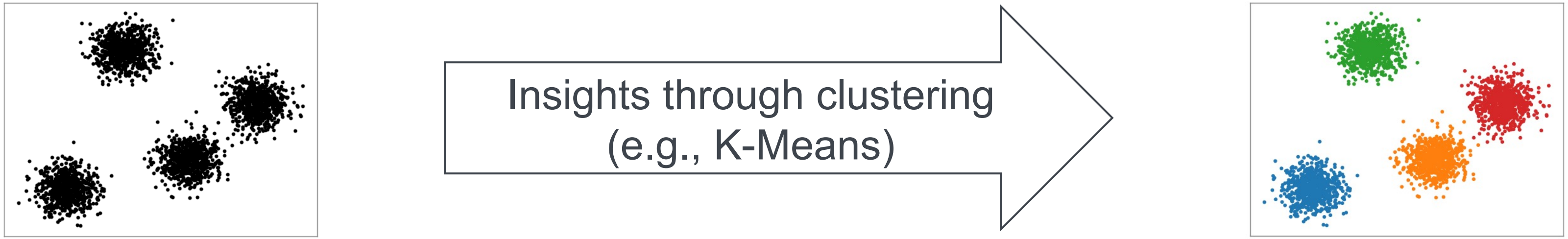
Leonard Labes
36th GI-Workshop on Foundations of Databases

IPVS

Anwendungssoftware

Motivation

- Various organizations rely on data to gain insights and knowledge
  - Customer segmentation in E-Commerce
  - Social network recommendations
  - Fraud detection in finance
  - …
- One option to gain insights from data is clustering (unsupervised machine learning task)



Motivation

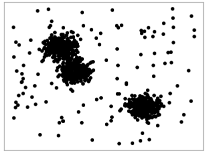
- Which clustering algorithm should I use?
- What are the parameter values?
- How should I evaluate it?

Solution: AutoML for Clustering:

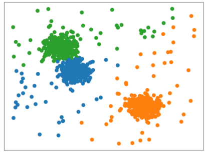
- Automatically find good algorithms, parameters, and evaluation scores.
- ML2DAC [7], AutoClust [9], TPE-AutoClust [17],…

Motivation

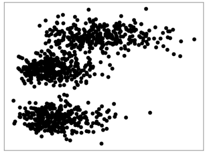
- However, data often have quality problems like:
- Outlier



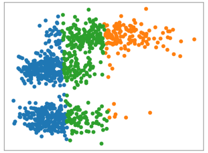
Clustering (K-Means)



- Skewed Data



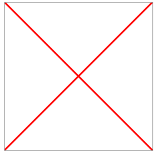
Clustering (K-Means)



- Missing Values

A	B	C
1	-	8
3	0	-

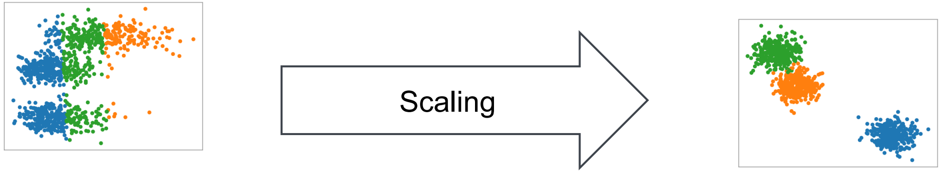
Clustering (K-Means)



Motivation

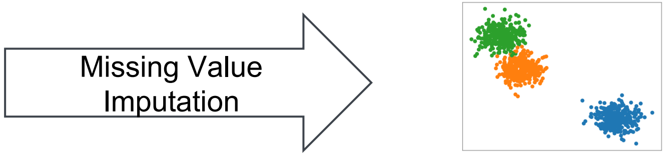
- One possibility to use is **preprocessing**
- Outlier

- Skewed Data



- Missing Values

A	B	C
1	-	8
3	0	-



Motivation

- Identifying data characteristics can range from easy to very difficult

Column A	Column B	Column C	Column D
45	120000	Red	4.5
10	1		-100
	450000	Green	
13	790000	Yellow	1.3

Missing Values

Motivation

- Identifying data characteristics can range from easy to very difficult

Column A	Column B	Column C	Column D
45	120000	Red	4.5
10	1		-100
	450000	Green	
13	790000	Yellow	1.3

Missing Values

Categorical Features

Motivation

- Identifying data characteristics can range from easy to very difficult

Column A	Column B	Column C	Column D
45	120000	Red	4.5
10	1		-100 ?
	450000	Green	
13	790000	Yellow	1.3

Missing Values

Categorical Features

Outlier

Motivation

- Identifying data characteristics can range from easy to very difficult

Column A	Column B	Column C	Column D
45	120000	Red	4.5
10	1		-100
	450000	Green	
13	790000	Yellow	1.3

Missing Values

Categorical Features

Outlier

Duplicate Features

Motivation

- Identifying data characteristics can range from easy to very difficult

Column A	Column B	Column C	Column D
45	120000	Red	4.5
10	1		-100
	450000	Green	
13	790000	Yellow	1.3

Missing Values

Categorical Features

Outlier

Duplicate Features

Different Scales

Motivation

- Identifying data characteristics can range from easy to very difficult

Column A	Column B	Column C	Column D

Identifying all relevant characteristics is **<u>hard</u>**

Fixing them requires not only a single preprocessing method but a whole **<u>pipeline</u>**

Missing Values

Categorical Features

Outlier

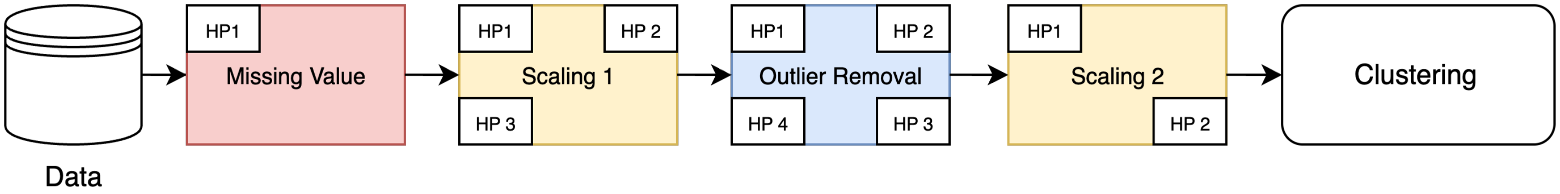
Duplicate Features

Different Scales

Motivation

- A pipeline raises more questions

Which Methods? Interdependency? Order of Methods? Evaluation?



Which Parameter Values? Runtime and Scalability? Repetition of Methods?

Contribution

- To guide and support data analysts, an automated approach is needed

---

To make the described problems more explicit, this work contributes:

- A five-step conceptual process
- Five derived challenges from this process
- A systematic comparison of related work
- Hints to possible future directions

Process

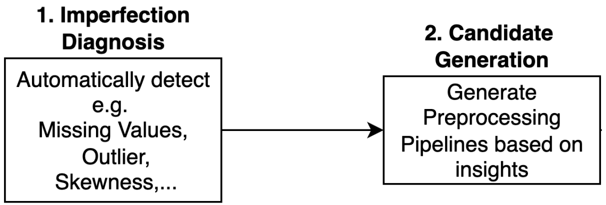
1. Imperfection Diagnosis

Automatically detect
e.g.
Missing Values,
Outlier,
Skewness,...

1. Definition of imperfection

2. Many detection techniques

Process



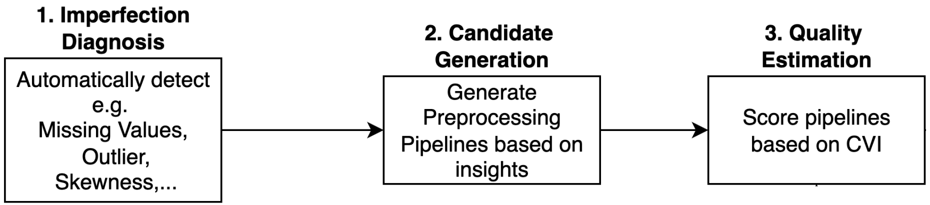
1. Definition of imperfection

2. Many detection techniques

3. Pipeline constraints

4. Search space constraints

Process



1. Definition of imperfection

2. Many detection techniques

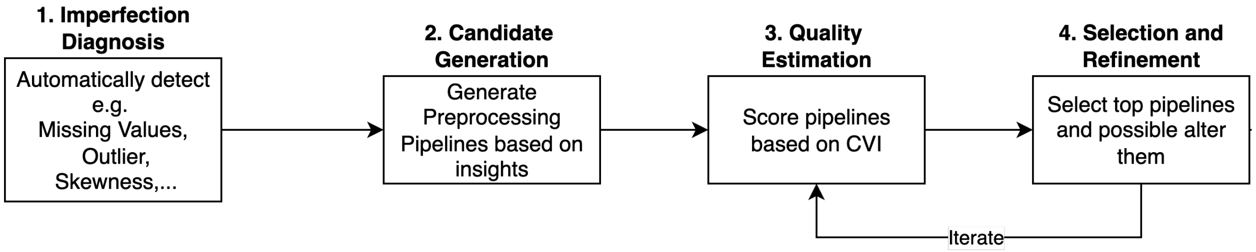
3. Pipeline constraints

4. Search space constraints

5. No ground truth

6. Which CVI should be considered

Process



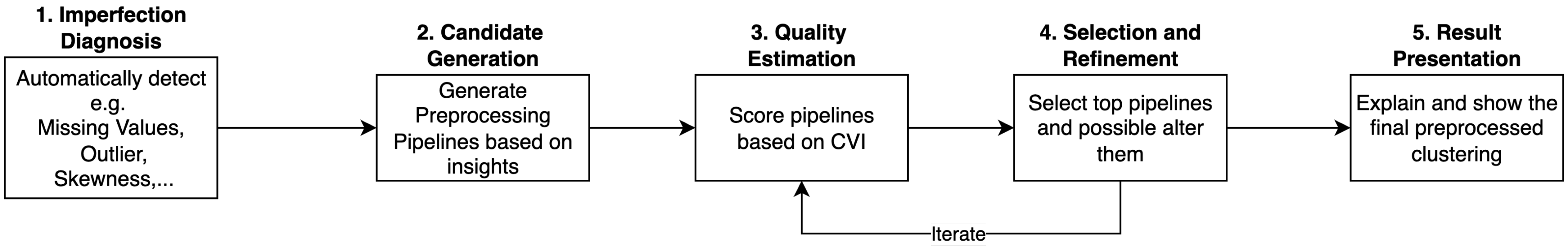
1. Definition of imperfection
2. Many detection techniques

3. Pipeline constraints
4. Search space constraints

5. No ground truth
6. Which CVI should be considered

7. Refinement strategy
8. Huge search space

Process



1. Definition of imperfection

2. Many detection techniques

3. Pipeline constraints

4. Search space constraints

5. No ground truth

6. Which CVI should be considered

7. Refinement strategy

8. Huge search space

9. How to visualize

10. Explainability

Process

1. Imperfection Diagnosis

Automatically detect e.g.
Missing Values,
Outlier,
Skewness,...

. Result
esentation

and show the
preprocessed
lustering

Every step has its own challenges

To make approaches comparable, a clear definition of the challenges is needed

1. Definition of imperfection

2. Many detection techniques

constraints

4. Search space constraints

truth

6. Which CVI should be considered

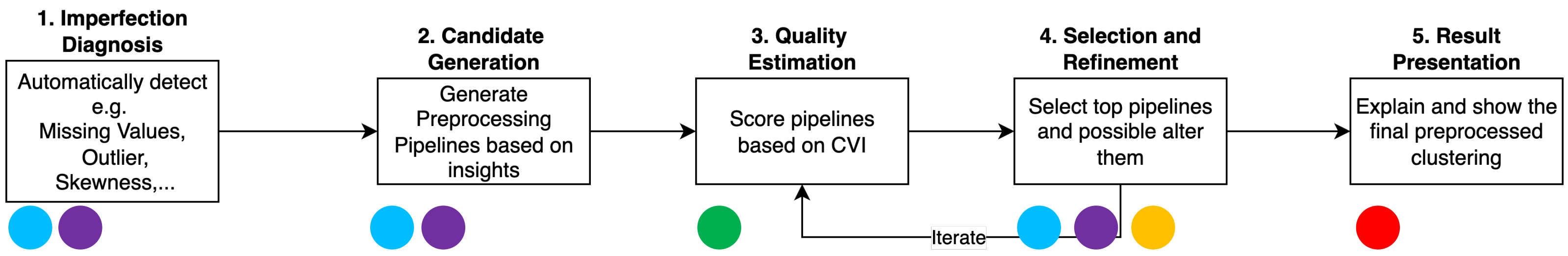
strategy

8. Huge search space

visualize

10. Explainability

Process



- **C1: Diverse Techniques:** Many different detection mechanisms and preprocessing techniques are needed
- **C2: Search Space Explosion:** Availability of techniques, their parameters, etc., leads to a vast search space
- **C3: Evaluation Metric:** In unsupervised machine learning, the evaluation technique is not clear
- **C4: Refinement Strategy:** A robust and scalable strategy for improving the pipelines is needed
- **C5: Visualization & Explainability:** A data analyst must understand why and how the data is preprocessed

Systematic Comparison of Related Work

- Existing work is grouped into:
  - Unsupervised AutoML Approaches
  - Data Preprocessing for Supervised Machine Learning
  - Libraries
  - Semi Automated Visualization Tools

Paper	C1 Diverse Techniques	C2 Search Space Explosion	C3 Evaluation Metric	C4 Refinement Strategy	C5 Visualization & Explainability

○ Not addressed

◐ Partially addressed

● Completely addressed

Systematic Comparison of Related Work

- Unsupervised AutoML Approaches
  - Focus on the machine learning part
    - Finding the best clustering algorithm
    - Finding the best evaluation metric
  - If preprocessing is addressed, only partially

Paper	C1 Diverse Techniques	C2 Search Space Explosion	C3 Evaluation Metric	C4 Refinement Strategy	C5 Visualization & Explainability
ML2DAC [7]	○	○	●	○	○
AutoClust [9]	○	○	◑	○	○
TPE-AutoClust [17]	◑	◑	◑	◑	○

Systematic Comparison of Related Work

- Data Preprocessing for Supervised Machine Learning
  - Solutions with more extensive preprocessing exist
    - By design, all require class labels
    - Various strategies and focuses exist

Paper	C1 Diverse Techniques	C2 Search Space Explosion	C3 Evaluation Metric	C4 Refinement Strategy	C5 Visualization & Explainability
Saga [18]	◑	●	○	●	○
CtxPipe [19]	●	◑	○	●	○
LLMClean [20]	◑	○	○	○	◑

Systematic Comparison of Related Work

- Libraries
  - Research has matured into widely used libraries
    - Most of them only do static preprocessing
    - TPOT optimizes preprocessing pipelines
      - Only supports supervised learning

Paper	C1 Diverse Techniques	C2 Search Space Explosion	C3 Evaluation Metric	C4 Refinement Strategy	C5 Visualization & Explainability
Auto-Weka [21, 22]	○	◑	○	○	○
Auto-Sklearn [23, 24]	○	◑	○	○	○
AutoGluon [25]	○	◑	○	○	○
TPOT [26]	◑	◑	○	◑	○

Systematic Comparison of Related Work

- Semi Automated Visualization Tools
  - Most of the focus is on visualization and explainability of the data
  - Some suggest preprocessing steps
  - Some highlight the difference after a preprocessing operation has been applied

Paper	C1 Diverse Techniques	C2 Search Space Explosion	C3 Evaluation Metric	C4 Refinement Strategy	C5 Visualization & Explainability
Wrangler [27]	○	○	○	○	◑
Vizier [28]	○	○	○	○	◑

Systematic Comparison of Related Work

- Much research with a focus on preprocessing exists
- None can fulfill all the defined challenges
- Combining approaches is not easily achievable.

Paper	C1 Diverse Techniques	C2 Search Space Explosion	C3 Evaluation Metric	C4 Refinement Strategy	C5 Visualization & Explainability
ML2DAC [7]	○	○	●	○	○
AutoClust [9]	○	○	◑	○	○
TPE-AutoClust [17]	◑	◑	◑	◑	○
Saga [18]	◑	●	○	●	○
CtxPipe [19]	●	◑	○	●	○
LLMClean [20]	◑	○	○	○	◑
Auto-Weka [21, 22]	○	◑	○	○	○
Auto-Sklearn [23, 24]	○	◑	○	○	○
AutoGluon [25]	○	◑	○	○	○
TPOT [26]	◑	◑	○	◑	○
Wrangler [27]	○	○	○	○	◑
Vizier [28]	○	○	○	○	◑

Future Directions

- Detecting data characteristics
  - Rule-based, Meta Learning, utilizing embeddings
- Creating and refining pipelines
  - Optimization techniques are needed: Bayesian Optimization, Genetic Optimization, Reinforcement Learning.
  - LLMs: Still a focus on small datasets and domain knowledge
    - A good representation of the data is needed
- Visualization and Explainability
  - Depends on the approach
  - Many tools provide good visualization,
    - Often require human input or are user-focused

Summary & Conclusion

- This work
  - Defines a conceptual process
  - Identifies challenges
  - Compares existing work
  - Shows what potential solutions may consider
- Preprocessing is essential, but automated preprocessing for clustering does not exist.
- Existing solutions show bits and pieces that could be adapted or point in a specific direction.

A thorough assessment of methods, techniques, and unresolved options is necessary to develop a solution that addresses all challenges.

Thank you!

Leonard Labes

E-Mail leonard.labes@ipvs.uni-stuttgart.de

Phone +49 711 685 88298

www.ipvs.uni-stuttgart.de/de/institut/team/Labes/

University of Stuttgart

Institute for Parallel and Distributed Systems (IPVS)

Applications of Parallel and Distributed Systems

References

[7] D. Treder-Tschechlov, M. Fritz, H. Schwarz, B. Mitschang, Ml2dac: Meta-learning to democratize automl for clustering analysis, Proceedings of the ACM on Management of Data 1 (2023) 1–26.

[9] Y. Poulakis, C. Doulkeridis, D. Kyriazis, Autoclust: A framework for automated clustering based on cluster validity indices, in: 2020 IEEE International Conference on Data Mining (ICDM), IEEE, 2020, pp. 1220–1225.

[17] R. ElShawi, S. Sakr, Tpe-autoclust: A tree-based pipline ensemble framework for automated clustering, in: 2022 IEEE International Conference on Data Mining Workshops (ICDMW), IEEE, 2022, pp. 1144–1153.

[18] S. Siddiqi, R. Kern, M. Boehm, Saga: a scalable framework for optimizing data cleaning pipelines for machine learning applications, Proceedings of the ACM on Management of Data 1 (2023) 1–26.

[19] H. Gao, S. Cai, T. T. A. Dinh, Z. Huang, B. C. Ooi, Ctxpipe: Context-aware data preparation pipeline construction for machine learning, Proceedings of the ACM on Management of Data 2 (2024) 1–27.

[20] F. Biester, M. Abdelaal, D. Del Gaudio, Llmclean: Context-aware tabular data cleaning via llm-generated ofds, in: European Conference on Advances in Databases and Information Systems, Springer, 2024, pp. 68–78.

[21] C. Thornton, F. Hutter, H. H. Hoos, K. Leyton-Brown, Auto-weka: Combined selection and hyperparameter optimization of classification algorithms, in: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, 2013, pp. 847–855.

[22] L. Kotthoff, C. Thornton, H. H. Hoos, F. Hutter, K. Leyton-Brown, Auto-weka 2.0: Automatic model selection and hyperparameter optimization in weka, Journal of Machine Learning Research 18 (2017) 1–5.

[23] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, F. Hutter, Efficient and robust automated machine learning, in: Advances in Neural Information Processing Systems 28 (2015), 2015, pp. 2962–2970.

[24] M. Feurer, K. Eggensperger, S. Falkner, M. Lindauer, F. Hutter, Auto-sklearn 2.0: Hands-free automl via meta-learning, arXiv:2007.04074 [cs.LG] (2020).

[25] N. Erickson, J. Mueller, A. Shirkov, H. Zhang, P. Larroy, M. Li, A. Smola, Autogluon-tabular: Robust and accurate automl for structured data, arXiv preprint arXiv:2003.06505 (2020).

[26] R. S. Olson, J. H. Moore, Tpot: A tree-based pipeline optimization tool for automating machine learning, in: Workshop on automatic machine learning, PMLR, 2016, pp. 66–74.

[27] S. Kandel, A. Paepcke, J. Hellerstein, J. Heer, Wrangler: Interactive visual specification of data transformation scripts, in: Proceedings of the sigchi conference on human factors in computing systems, 2011, pp. 3363–3372.

[28] M. Brachmann, C. Bautista, S. Castelo, S. Feng, J. Freire, B. Glavic, O. Kennedy, H. Müeller, R. Rampin, W. Spoth, et al., Data debugging and exploration with vizier, in: Proceedings of the 2019 International Conference on Management of Data, 2019, pp. 1877–1880.