Institute for Application-Oriented Knowledge Processing

Student Topics

JOHANNES KEPLER
UNIVERSITÄT LINZ

https://www.jku.at/en/faw/

Our Faculty

Data – Information – Knowledge

Johannes Fürnkranz

- Computational Data Analytics
- Data Mining and Knowledge Discovery
- Rule Learning and Interpretability
- Machine Learning in Games
- Preference Learning and Multi-Label Classification

Josef Küng

- Knowledge-based Systems and Knowledge Representation
- Security and Trust in Information Systems
- Process-aware Information Systems
- Similarity Queries

Birgit Pröll

- Information Retrieval & Extraction
- Natural Language Processing
- Web Search and Mining
- Web Engineering und Web Science

Wolfram Wöß

- Information Integration (Semantic-based, Ontologies)
- Data Modeling
- Knowledge Representation and Knowledge Graphs
- Data Quality, Data Profiling, Data Catalogs

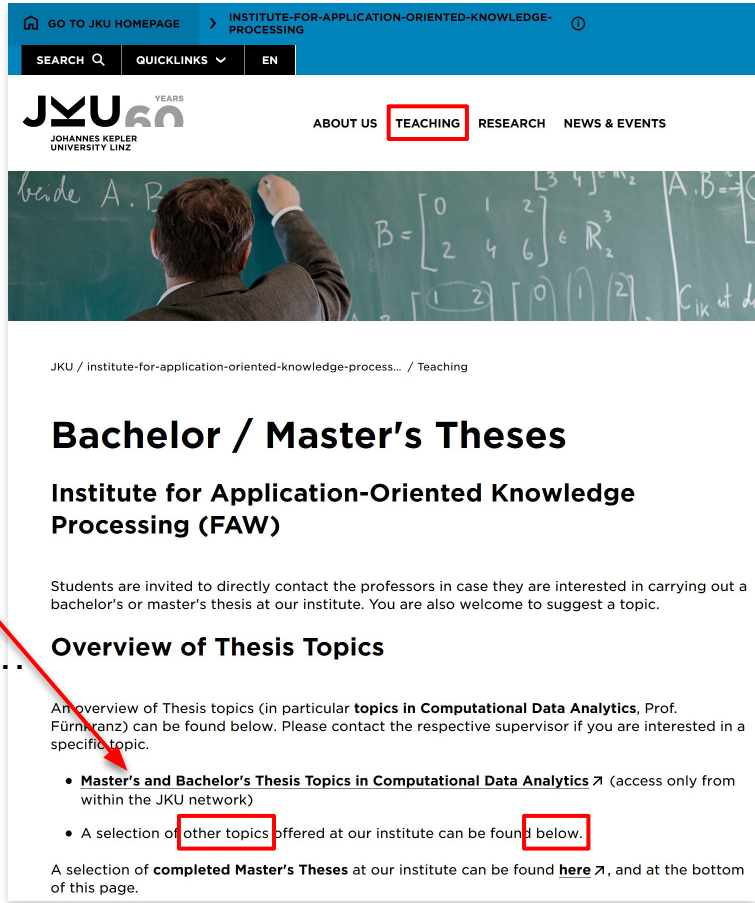
JKU

General Information

- Here we only present a selection of topics
  - with the goal of illustrating our research directions
  - a current list (these slides) can be found at https://teaching.faw.jku.at/Theses/Student Topics - FAW.pdf
  - concrete topics will be fixed in a personal meeting

- You can also suggest your own topic!
  - no guarantee, depends on our capacity, interest, …
  - needs to fit into our general research directions
  - we can then discuss whether it is suitable
  - also, there may be fees for industry projects

http://www.jku.at/faw/



Aktuelle Themen – Josef Küng

see also https://www.jku.at/en/institute-for-application-oriented-knowledge-processing/teaching/bachelor-masters-theses/

- Zugriffskontrolle in Graph-Datenbanken
  - Zugriffsschutzmechanismen in Graph-Datenbanken decken nicht alle Anforderungen ab, vor allem nicht solche, die über einfachen Standard hinausgehen. In der Bachelorarbeit sollen Möglichkeiten für einen verbesserten Zugriffsschutz recherchiert und mindestens eine davon prototypisch implementiert werden.
- Zugriffskontrolle in Workflow-Systemen
  - Recherche und prototypische Implementierung für verbesserten Zugriffsschutz.
- Assoziativspeicher
  - Eine alte, am Institut entwickelte Methode soll mit neuer Technologie implementiert werden.
- Datenintegration im Medizin-Bereich
  - Analyse, Design und prototypische Implementierung eines Client in C# (Webservice mit REST-API ist bereits vorhanden).

Aktuelle Themen – Birgit Pröll

- **iVolunteer** – Digitale Plattform informeller Kompetenzen im Freiwilligenbereich (VMS)
  - Community/Social”-Aspekte in VMS
  - Gamification in VMS
  - Chatbot für Kompetenzermittlung
  - Kompetenzextraktion aus Texten
  - Open Innovation Website
  - Communication Hub
  - Matching von Volunteers und Aufgaben
  - etc.

Current Topics – Wolfram Wöß

see also https://www.jku.at/en/institute-for-application-oriented-knowledge-processing/teaching/bachelor-masters-theses/

- Data Quality Tool DQ-MeeRKat (Java-based)
  - Extensions for "multivariate outlier detection models", "ML models for duplicate detection", …
  - https://github.com/lisehr/dq-meerkat
- Knowledge Graphs
  - Measuring the Quality of Knowledge Graphs
  - Graph Database Extensions
- Data Profiling und Data Summarization (in Big Data)
- Data Catalogs, Metadata Management

Computational Data Analytics

coda

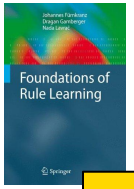
Inductive Rule Learning

```
IF   MaritalStatus = single
 AND Sex = male
THEN Approved = no

IF   MaritalStatus = married
THEN Approved = yes

IF   MaritalStatus = divorced
 AND HasChildren = yes
THEN Approved = no
```

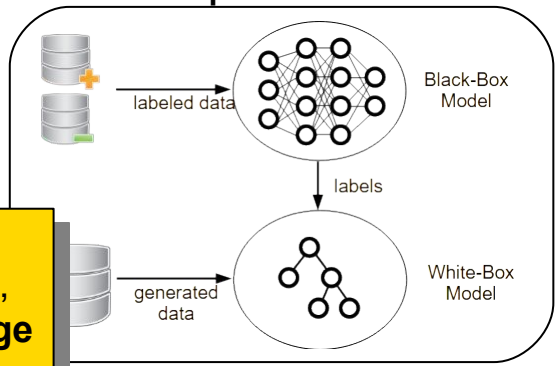
yes (0/9) | no (3/5)

yes (4/9) | no (0/5)

yes (0/9) | no (2/5)



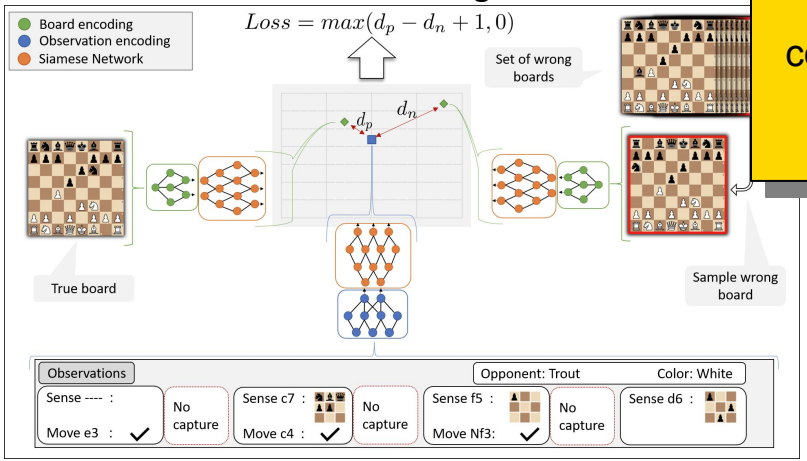
Explainable AI



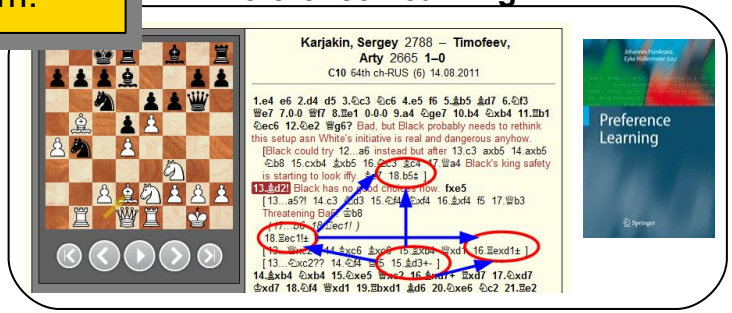
Goal

Acquisition of explicit, formalizable **knowledge** from sources which contain **information** in implicit, not directly accessible form.

Machine Learning in Games



Preference Learning



Current Topics – J. Fürnkranz

- We offer seminar/project/thesis topics in Computational Data Analysis
- In the following, you can find **sample topics** in a few areas
  - Machine Learning in Games
  - Interpretability and Inductive Rule Learning
  - Multi-label Classification

A full list of topics can be found at https://teaching.faw.jku.at/Theses/Student Topics - FAW.pdf

- If you are interested in similar problems, you can also propose your own topic
- Prerequisites
  - Some basic knowledge (and ideally practical experience) in machine learning and data mining is assumed

Seminar / Project / Thesis

With Prof. Fürnkranz, there are two possible paths

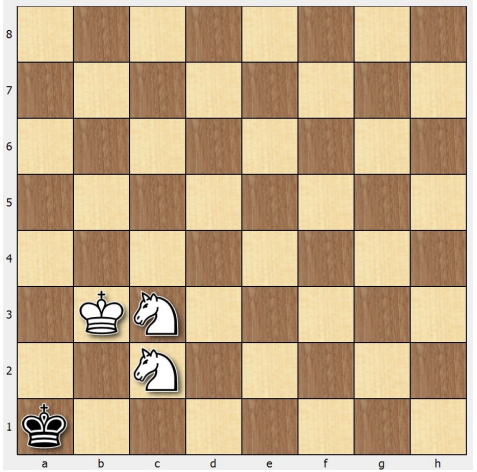
- You compile seminar / project / thesis (or two of the three) into a single package, typically
  - start with giving a presentation (seminar)
  - implement or work with state-of-the-art techniques (project)
  - investigate a new interesting question (thesis)
- You do all of them separately
  - availability depends on the amount of interest
  - seminar: several talks by different students on an over-arching topic
  - project: group work on some problem (often a competition)

You can switch between the two models inbetween

1. Game Playing

- We are generally interested in using AI technology for game playing
  - typically conventional board or card games, but dynamic video games are also possible
- Topics could involve questions such as
  - design and implement a strong player for a new game
  - learn a player from human game playing databases or from self play
  - analyze human decisions in game databases
  - gain knowledge about the game by analyzing game databases
  - ….
- Some example projects are on the following slides
  - If you have an interesting game project, feel free to talk to us

Retrograde Analysis for Endgames with Two Knights



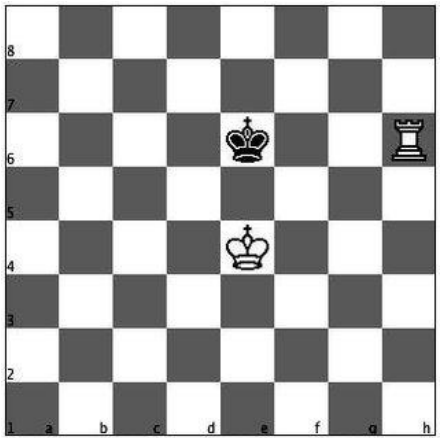
- It is impossible to force a mate with two knights (the shown position is mate, but the previous must have been a stale-mate, i.e., a draw)
- However, it is possible to sometimes force a mate if the board has one square more or less

Task:

- implement retrograde analysis to solve KNNK endgames on various boards with one square added or removed
- tabulate the results

Counterfactual Explanations for Chess Endgames

- Counterfactual explanations are “near-misses”, i.e., positions that are similar to the current one but have a different evaluation
- explore whether they can be used for a tutoring system for simple chess endgames like KRK
  - either directly or by modifying rule-based strategies



Literature:

- Johannes Rabold, Michael Siebers, Ute Schmid: Generating contrastive explanations for inductive logic programming based on a near miss approach. *Machine Learning* 111(5): 1799-1820 (2022)

Embedding chess playing styles

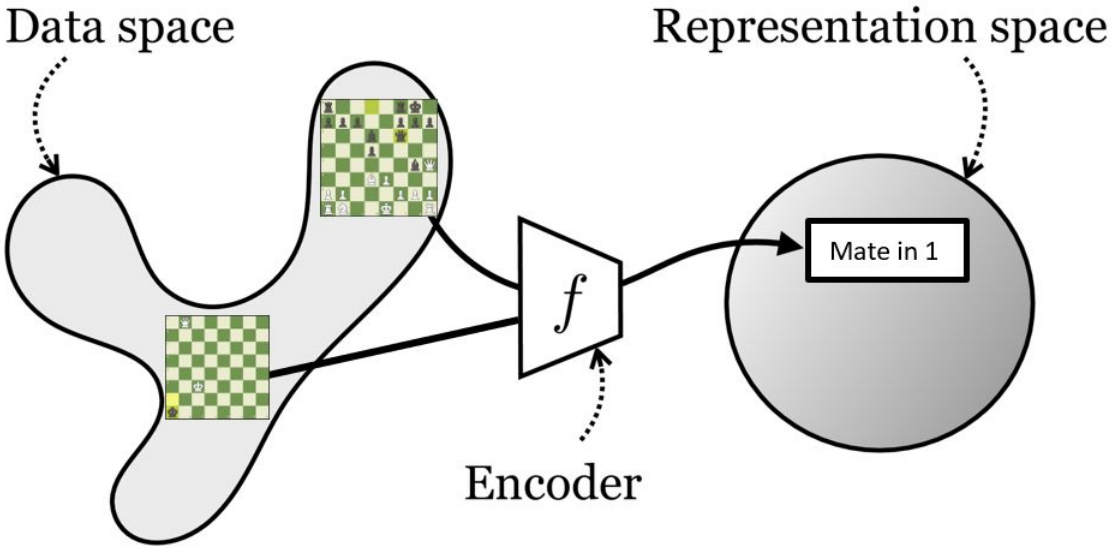
- Train a neural network on chess games that predicts the playing style of various grandmasters
- Alternatively: Learn a general embedding and try to discover playing styles within the encodings

Q: https://www.youtube.com/watch?v=GJr5UtM4TyU

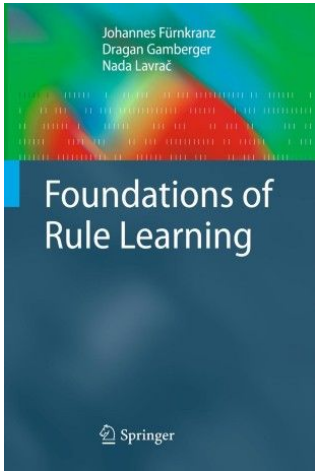
Further topics of interest

Particularly in games

- Representation learning
- World models
- Explainable AI



2. Inductive Rule Learning



- Logical IF-Then rules are commonly used in many applications
  - Are very interpretable but often not very accurate
- Many topics that can be worked on, such as
  - **Combining Rule Predictions**
    - Many approaches have been proposed on how to combine the predictions of multiple rules (e.g., in an ensemble) -> compare them!
  - **Interpreting Multi-Label Predictions**
    - Current works on interpretable machine learning focus on single-label predictions tasks -> methods for explaining joint predictions are needed
  - **Comparison of Extracted Explanations vs. Learned Models**
    - Do local rules extracted from a deep network outperform rules that are directly learned from the data?

Meta-Learning Structured Models using Deep Neural Networks

Implement and evaluate the following idea:

- given a fixed structure of the target model (e.g., a template for a NNF formula)
- for a given dataset and structure generate a large number of possible models randomly
- re-label the examples of the dataset according to the model
- train a deep neural networks that tries to predict the model structure from a given labeling
- test the predicted model on the “real” labeling

Learning m-of-N Networks

m-of-N concepts are true if more than m of the N input nodes are 1

- $m = 1$ corresponds to a logical OR, $m = N$ to a logical AND

Task:

- implement a network architecture where each node is an m-of-N concept
  - initialized randomly where each node has exactly N inputs
- design and implement a learning algorithm for optimizing the thresholds m
- evaluate and compare the result
  - e.g., to difflogic, which is an alternative way of trading off between conjunction and disjunction

Literature:

- Florian Beck, Johannes Fürnkranz, Van Quoc Phuong Huynh: Generalizing Conjunctive and Disjunctive Rule Learning to Learning m-of-n Concepts. ITAT 2023: 8-13

Evaluate Closed Classification Rules

- Classification rule learners typically strive for learning simple discriminative rules
  - “If you see an animal with a trunk, it is an elephant.”
- From an interpretability point of view (and maybe also for robustness) it may be preferable to learn more complex characteristic rules
  - “An Elephant is a very large and heavy animal that has a trunk, tusks, grey skin, big ears and four thick legs.”
- Task:
  - survey the literature on characteristic rule learning
  - implement an algorithm that takes discriminative rules as input and outputs the closure
  - devise a method for “approximate” closures, possibly also with a bias toward interesting conditions
  - evaluate it on various benchmark data w.r.t. accuracy, robustness, and interpretability

Evaluation of Rule Ensembles

- The goal of this Bachelor's thesis is to explore the rule learner LORD as a base classifier for ensemble methods such as bagging and boosting
- Tasks:
  - realize the integration of LORD in ensembles in a standard data mining environment such as WEKA or SciKit-Learn
  - evaluate its performance on a large number of benchmark datasets
  - compare its performance to random forests, bagged decision trees etc.
  - explore different parametrizations of the algorithm

Prerequisites:

- basic knowledge in ML
- Python or Java programming

Random Rule Forests

- design and implement a variant of random forests that focuses on classification rules instead of decision trees
- Possible approaches:
  - extend Ripper so that it not only uses random sampling on the examples but also on the features
  - extend LORD so that it includes random sampling on the features of an example
- evaluate the algorithm on a variety of benchmark datasets
  - compare it to other algorithms on these data

Prerequisites:

- good java programming skills

Deep Random Rule Forests

- Conventional rule learning algorithms learn only flat rule sets
- implement and refine an algorithm that can learn a deep structure using a random forest
  - based on existing rule learning algorithms such as BOOMER or LORD
  - the key idea for the algorithm is provided but needs some refinement
- evaluate the algorithm on a variety of benchmark datasets
  - compare it to other algorithms on these data

Prerequisites:

- good java programming skills
- basic knowledge in machine learning
- knowledge in inductive rule learning a plus

Transductive Rule Learning

- conventional rule learning algorithms learn a rule set on the basis of the training examples and use this for classifying test examples
- transductive learning would learn a new rule for each test example

Tasks:

- We have a rule learner that is able to do that
- However, the problem also requires the development of new heuristics that allow to assess the compactness of the covered examples
- Develop, implement, and evaluate an algorithm on that basis

Literature:

- Veloso, Meira Jr., Goncalves, de Almeida, Zaki: Calibrated Lazy Associative Classification, Information Sciences 181(13):2656-2670 (2011)
- Huynh, V.Q.P., Beck, F., Fürnkranz, J.: Efficient learning of large sets of locally optimal classification rules. Machine Learning (2023), in press.

Improve JRip

- Ripper is a classic rule learning algorithm that is still very hard to beat in terms of efficiency as well as compactness of the learned rules
- the currently best implementation available is **JRIP**
  - available in the Weka data mining environment (implemented in Java, https://www.cs.waikato.ac.nz/ml/weka/)
  - code of the original C++ implementation is also available
- However JRip could be made for flexible in various ways, such as
  - allow the use of different search heuristics
  - allow for different class orders, both static as well as dynamically selected
  - adjust for the use of mini-batches instead of train/test splits
  - analyze the rule optimization phase and possible find a cheaper alternative
  - interface with Python/scikit-learn
- excellent Java programming skills required

Re-Implement and Evaluate Classic Rule Learning Algorithms

- LORD is a state-of-the-art rule learning algorithm developed within our group
- it features an efficient framework for data structures that allow to summarize all information for a rule learning algorithm
- Goal of this Bachelor's Thesis is to efficiently re-implement and evaluate classic (e.g., AQ, CN2, Foil) and/or modern (e.g., IDS) rule learning algorithms within this framework
- excellent Java programming skills required

https://github.com/vqphuynh/LORD

Literature:

- Huynh, V.Q.P., Beck, F., Fürnkranz, J.: Efficient learning of large sets of locally optimal classification rules. Machine Learning (2023).

JKU coda

Local Optimal Rule Weighting

- In an effort to improve the predictive performance of Local Optimal Rule Learning (LORD), weighting every single rule in a rule set by LORD with a neural network is a potential approach.

Tasks: The first two tasks can be suitable for Bachelor, all tasks for Master. Java & Python is required.

- Convert (on-the-fly) the training dataset to a new dataset in the covering-rule space.
- Build up and train a weighting model via a neural network.
- The LORD rule set is often large, reduce the rule set before the on-the-fly converting and the weighting.

Literature:

- Huynh, V.Q.P., Beck, F., Fürnkranz, J.: Efficient learning of large sets of locally optimal classification rules. Machine Learning (2023).

https://github.com/vqphuynh/LORD

LORD based X-AI

Local optimal rule learning (LORD) can explain the prediction of a classifier based on the ability to generate a rule(s) for a single example.

Tasks: Develop two algorithms to.

- Find a global optimal rule that explain a black-box's prediction for an example, in cases of not large datasets and/or not large numbers of features.
- Find near-reaching global optimal rules to explain a black-box's prediction for an example by local search on many random subspaces, in the cases of large datasets with large number of features.

Literature:

- Huynh, V.Q.P., Beck, F., Fürnkranz, J.: Efficient learning of large sets of locally optimal classification rules. Machine Learning (2023).

https://github.com/vqphuynh/LORD

Better Optimal Rule Learning

Local optimal rule learning (LORD) behaves well on multi-class classification and execution performance. Develop variants which improve the LORD's predictive performance.

Tasks: The first two tasks can be suitable for Bachelor, all tasks for Master.

- Develop a version to find global optimal rules that suitable for not large datasets and/or not large numbers of features.
- Develop a version to find near-reaching global optimal rules by local search on many random subspaces. This version can tackle large datasets with large number of features.
- Using these two variants to develop a lazy-classifier (transductive classifier) which does not learn a model but can generate a rule to classify an unseen example.

Literature:

- Huynh, V.Q.P., Beck, F., Fürnkranz, J.: Efficient learning of large sets of locally optimal classification rules. Machine Learning (2023).

https://github.com/vqphuynh/LORD

Rule Set Ensemble Learning

- Random Forests, ensembles of decision trees, have been well known for their high classification accuracy.
  - There are also available rule learners which basically generate rule set ensemble from decision trees.
- Tasks: Develop a method for learning an ensemble of rule sets directly from data while keeping in mind three following criteria.
  - As simple rule sets as possible
  - Comparative classification performance
  - Efficiency so that it is applicable for big data

Distributed Rule Learning

- Dealing with big and very big data sets is a crucial problem (running time and/or consumed memory) to the state-of-the-art rule learners.
- Tasks: Develop a parallel rule learning method for handling well very large-scale data sets on distributed environments e.g. clusters, local networks while keeping in mind the following criteria.
  - Remain or keep as same as possible with the rule set by the corresponding serial version.
  - Maintain high speed up when the number of computational nodes becomes larger.

Distributed Frequent Itemset Mining

- Frequent Itemset Mining (FIM) is a fundamental mining technique in Data Mining. It can be employed as a key calculation phase in other mining models such as Association Rules, Inductive rules, Classifications, Text Mining, etc.
- In the current era of Big Data, the distributed FIM algorithms have been proposed to dealing with very-large scale data sets. But the effort to improve the execution performance is always necessary because a distributed FIM algorithm likely suffers a kind(s) of adverse data sets.
- Tasks:
  - Quest for a distributed FIM algorithm that can perform stablly on diversity kinds of data sets and higher efficiency.

Binarization of Numeric Attributes

- rule learning algorithms typically deal with numeric attributes via simple threshold features of the form A > t
  - which check whether the value of attribute A is larger than the threshold t
- in many cases, these features are obtained via discretization
  - the result is a fixed set of non-overlapping interval features
- the goal of this master's thesis is to develop, implement, and evaluate alternative methods
  - e.g., features that check whether a numeric value is inside a certain quantile, is at the tail of the distribution, percentage of the value range, etc.

3. Other Topics

- We are also open to suggestions for other topics
  - typically they have to do with machine learning and/or knowledge discovery in databases
- Sometimes we also have other topics