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

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!
  - if it fits within the general research directions of the institute
  - we can then discuss whether it is suitable for a thesis

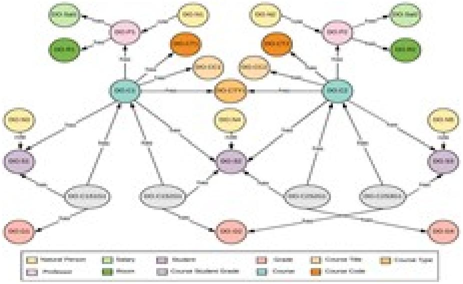
Seminar / Project / Thesis

There are two possible paths

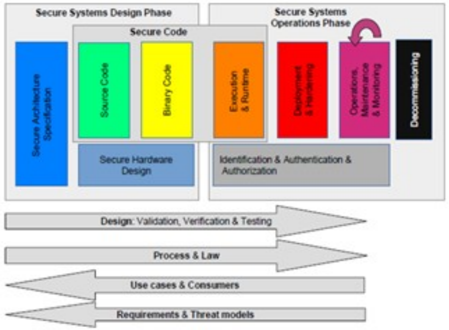
- You compile seminar / project / thesis (or two of the three) into a single package, typically
    - start with giving a presentation (seminar)
    - implement or work with state-of-the-art techniques (project)
    - investigate a new interesting question (thesis)
- You do all of them separately
    - 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

Current Topics – Josef Küng



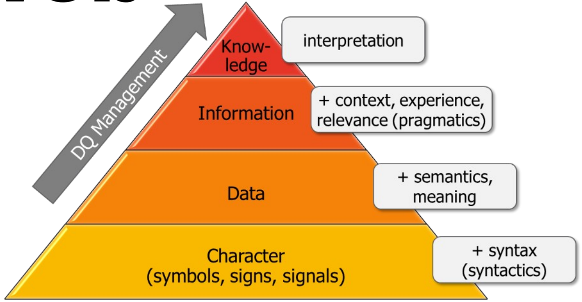
- Access Control in Graph Databases
    - Terms and Concepts around Access Control
    - Extending ABAC with graph specific access control expressions
- Mental Model Graphs
    - Managing & Mining Knowledge Work via a Gradually Externalised Mental Model Graph
- LIT – Secure and Correct Systems Lab
    - 9 Institutes work together to foster research in security
    - Supply Chain Security is the focus point for this funding period
    - FAW is working on access control to information systems


Secure Systems Design Phase
Secure Systems Operations Phase
Secure Code
Secure Architecture Specification
Source Code
Binary Code
Execution & Runtime
Deployment & Hardening
Operations, Maintenance & Monitoring
Decommissioning
Secure Hardware Design
Identification & Authentication & Authorization
Design: Validation, Verification & Testing
Process & Law
Use cases & Consumers
Requirements & Threat models

Aktuelle Themen – Josef Küng

- 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).

Current Topics – Wolfram Wöß



- Data Quality (DQ) Measurement
  - Using explicit constraints to capture and define domain knowledge
  - Using machine learning or AI-based methods to scale DQ measurement to real-world applications
- Knowledge Graphs (KGs) and Ontologies
  - Measuring the quality of KGs (e.g., with SHACL constraints)
  - Graph database extensions
- Data Profiling and Data Summarization (in Big Data)
- Data Catalogs - Technical Metadata Annotated with Business Context

DQD vocabulary: https://zenodo.org/records/8418173 (ontology to model DQ measurements)
DSD vocabulary: https://zenodo.org/records/7773862 (data source description vocabulary)

Current Topics – Birgit Pröll

iVolunteer – Digital Platform for informal Competences in Volunteering (VMS)

- Community/Social"-Aspects in VMS
- Gamification in VMS
- Chatbot for competence investigations
- Competence extraction from text
- Open Innovation Website
- Communication Hub
- Matching of volunteers and tasks
- etc.

» iVolunteer Marketplace «
Vision von iVolunteer
» iVolunteer Encouragement «
» iVolunteer FootPrint »
iVolunteer
Solarys
VMS
CrAc

Wissen & Innovation

Sag mir, wo du helfen willst

Informatik. Freiwillige sind bei Naturkatastrophen wie Überflutungen oder auch in der Nachbarschaftshilfe eine wichtige gesellschaftliche Stütze. Neue IT-Werkzeuge sollen helfen, das Miteinander und die Vernetzung weiter zu verbessern.

VON ALICE SENARCLENS DE GRANCY

Das Engagement ist enorm. Knapp die Hälfte der österreichischen Bevölkerung leistet irgendeine Freiwilligentätigkeit in einer Institution oder einem Verein. Das belegte eine Erhebung des Sozialministeriums vor zwei Jahren. Damit leisten Österreichs Freiwillige einen entscheidenden Beitrag, um sogenannte kritische Infrastrukturen wie Katastrophenschutz, Rettungsdienst, Gesundheits- und Sozialwesen sowie die Lebensmittelversorgung aufrechtzuerhalten.

Einzig: Deren Organisation könnte von mehr Digitalisierung profitieren. „Ein Freiwilligenkoordinator hat uns kürzlich berichtet, dass sie beim Ausfüllen von Einstiegsformularen für neue Freiwillige noch mit Papier und Bleistift arbeiten und das selbst gern ändern würden", erzählt Birgit Pröll vom Institut für anwendungsorientierte Wissensverarbeitung der Johannes Kepler Uni Linz. Die Informatikerin befasst sich seit mehreren Jahren mit digitalen Lösungen für Freiwilligenorganisationen – von der Waldorfschule bis zum Europäischen Forum Alpbach. „Auch da helfen

Wetterextreme nehmen zu. Hier schützen Feuerwehrleute und Freiwillige mit Sandsäcken ein Restaurant (in Nordfrankreich) vor Überschwemmungen. [Picturedesk / Anthony Brzeski]

flexibel helfen. Das will man nutzen. Zudem würden immer wieder Menschen ihre Hilfe spontan an-

Überdies sei der Einsatz einer solchen Plattform für Nachbarschaftshilfe denkbar.

Anwerben und Kennenlernen Interessierter, aber nur zwei von zehn blieben, schilderte etwa ein Mit-

Ziel setzen zu können, für das man dann belohnt wird. Für die meisten Freiwilligen sei die persönliche

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

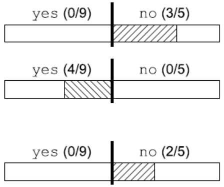
Computational Data Analytics

Inductive Rule Learning

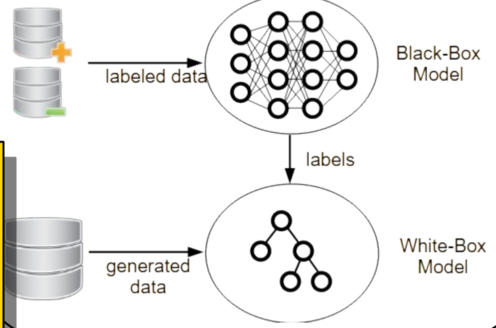
```
IF   MaritalStatus = single
 AND Sex = male
THEN Approved = no

IF   MaritalStatus = married
THEN Approved = yes

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



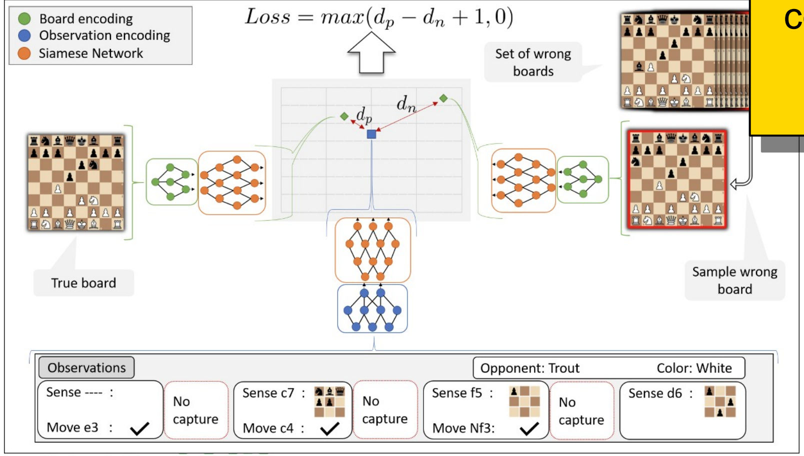
Explainable AI



Overall Objective

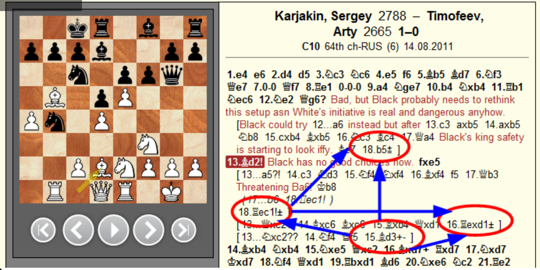
Acquisition of explicit, formalizable **knowledge** from sources which contain **information** in implicit, not directly accessible form.

Machine Learning in Games



$$Loss = max(d_p - d_n + 1, 0)$$

Preference Learning



Karjakin, Sergey 2788 – **Timofeev, Arty** 2665 **1–0**
C10 64th ch-RUS (6) 14.08.2011

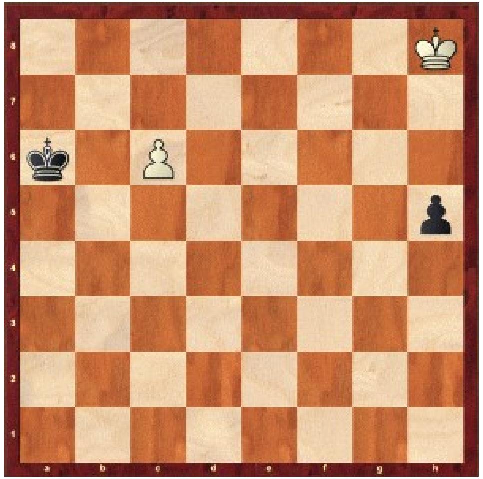
1.e4 e6 2.d4 d5 3.♘c3 ♘c6 4.e5 f6 5.♗b5 ♗d7 6.♘f3 ♕e7 7.0-0 ♕f7 8.♖e1 0-0-0 9.a4 ♘ge7 10.b4 ♘xb4 11.♖b1 ♘ec6 12.♘e2 ♕g6? Bad, but Black probably needs to rethink this setup asn White's initiative is real and dangerous anyhow.
[Black could try 12...a6 instead but after 13.c3 axb5 14.axb5 ♘b8 15.cxb4 ♗xb5 16.♘c3 ♗c4 17.♕a4 Black's king safety is starting to look iffy. ♗e7 18.b5±]
13.♗d2! Black has no good choices now. **fxe5**
[13...a5?! 14.c3 ♘d3 15.♘f4 ♘xf4 16.♗xf4 f5 17.♕b3 Threatening Ba6! ♔b8
(17...b6 18.♖ec1!)
18.♖ec1!±]
[13...♕xc2 14.♗xc6 ♗xc6 15.♖xb4 ♕xd1 16.♖exd1±]
[13...♘xc2?? 14.♘f4 ♕f5 15.♗d3+-]
14.♗xb4 ♘xb4 15.♘xe5 ♕xc2 16.♗xd7+ ♖xd7 17.♘xd7 ♔xd7 18.♘f4 ♕xd1 19.♖bxd1 ♗d6 20.♘xe6 ♘c2 21.♖e2

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

Discovering Interpretable Strategies for Chess Endgames

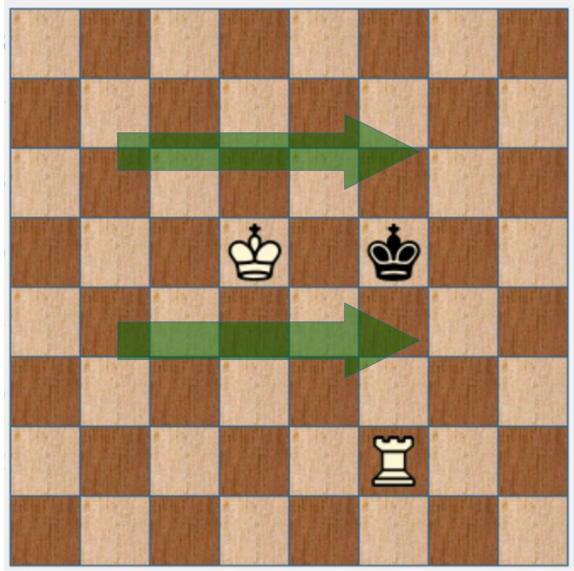
- Chess endgame databases provide tables that Contain optimal moves for certain endgames
- Human players, on the other hand, follow simple strategies



Tasks:

- Take a simple database like KRK (or later the more interesting KQKR)
- Try to work out ways for learning a human-like (i.e., simple, but possibly suboptimal) policy from the databases

Learning Progress in a Simple Chess Endgame



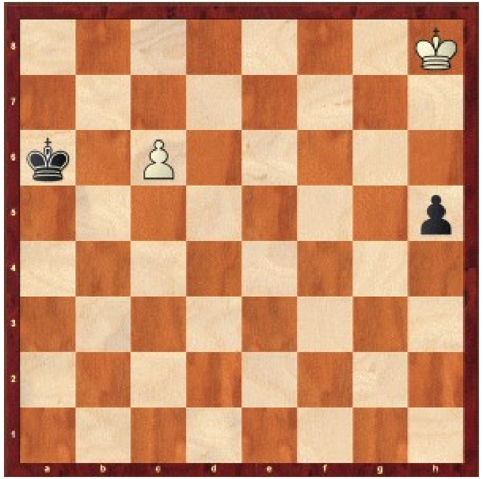
- A key difficulty in learning how to play a KRK endgame are symmetries
- Human plan to which side on the board they will mate the king, but chess endgame tables have no notion of symmetry

Tasks:

- Train a simple classifier (e.g., a neural network) that predicts on which edge the king will be mated.
- This could be used
  - a) for rotating the position so that the mating edge is always at the bottom
  - b) as a predicate for establishing progress.

Counterfactual Explanations for Chess Endgames

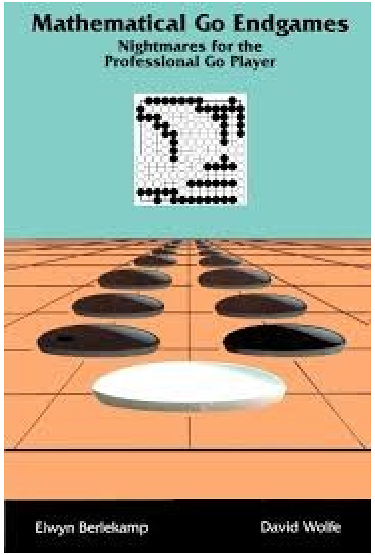
- Counterfactual explanations are “near-misses”, i.e., positions that are similar to the current one but have a different evaluation
- explore whether they can be used for a tutoring system for simple chess endgames like KRK
  - either directly or by modifying rule-based strategies



Literature:

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

Evaluation of Computer Endgame Play in NN-based Go Programs



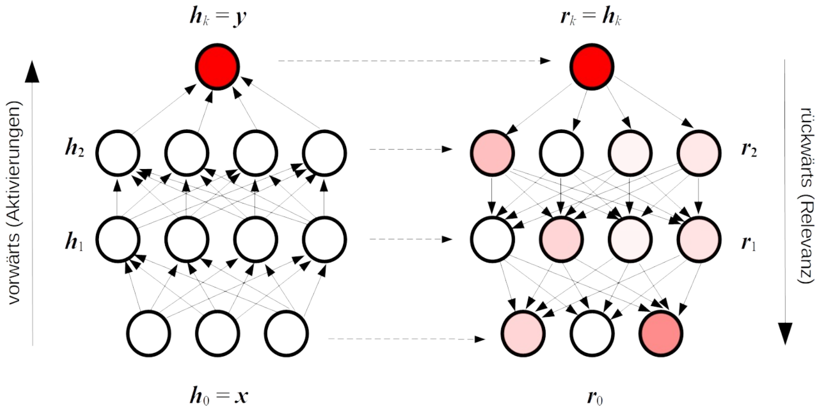
Mathematical Go Endgames
Nightmares for the Professional Go Player
Elwyn Berlekamp
David Wolfe

- Computer Go engines such as AlphaZero or its public variant Leela Zero make mistakes
- Endgame databases play perfectly
- Possible tasks:
  - Investigate and quantify the amount of mistakes these progams make
  - Try to characterize the type of mistakes they make
  - Compare mistakes in the evaluation vs. search
- Together with Martin Müller, an expert in Go & Computer Go

http://zero.sjeng.org/

Interpretation of NN-based Go or Chess Programs

- There are publicly available neural networks for Leela Go and others
- There are several recent techniques for making deep learning interpretable
  - LIME, SHAP, LRP



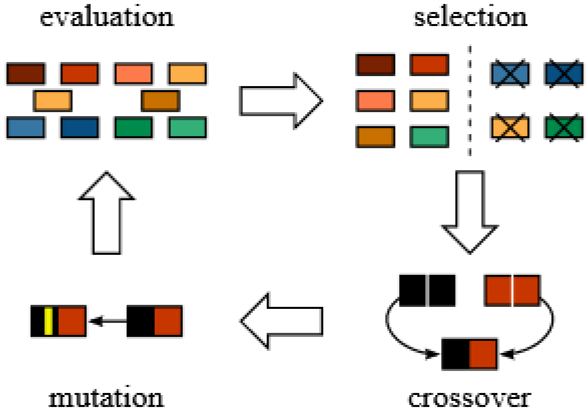
- Tasks:
  - Try these methods on trained Go or chess networks
  - See whether they help to interpret the learned knowledge

- Possibly together with Martin Müller, an expert in Go & Computer Go

http://zero.sjeng.org/

MCTS as a General Optimization Technique

- Monte-Carlo Tree Search (MCTS) is designed to find a sequence of actions to reach a goal
- In many problems, not the path is of interest, but the final result
  - E.g., in many puzzles or hard problems
- Classical solutions methods for such problems include local search algorithms such as genetic algorithms
- Can (single-agent) MCTS be an alternative?
  - There are various variants that operate on sets instead of sequences
  - These need to be researched, evaluated, compared, improved.



Contextual Preference Ranking and Reinforcement Learning for Gin Rummy

- Gin Rummy is a card game where the player constantly has to evaluate whether a new card fits well into the hand she is holding
- contextual preference ranking can learn such information using Siamese neural networks
- reinforcement learning can be readily applied as the games are comparably short and there is a clear feedback signal

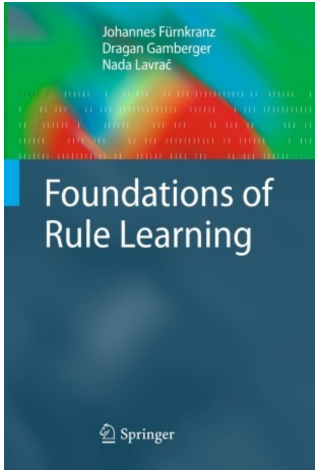
Tasks:

- Realize a Gin Rummy player based on a combination of contextual preference ranking and reinforcement learning

Literature:

- V. D. Nguyen, D. Doan, T. W. Neller: A Deterministic Neural Network Approach to Playing Gin Rummy. EAAI 2021.
- T. Bertram, J. Fürnkranz, M. Müller: Predicting Human Card Selection in Magic: The Gathering with Contextual Preference Ranking. IEEE CoG 2021: 1-8

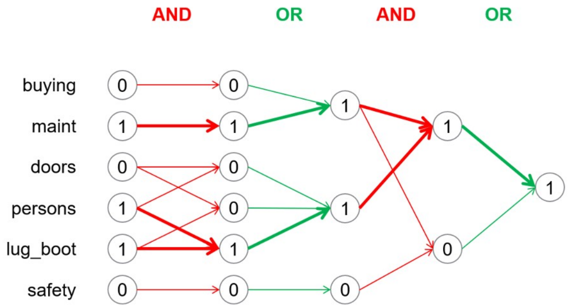
2. Inductive Rule Learning



- 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?

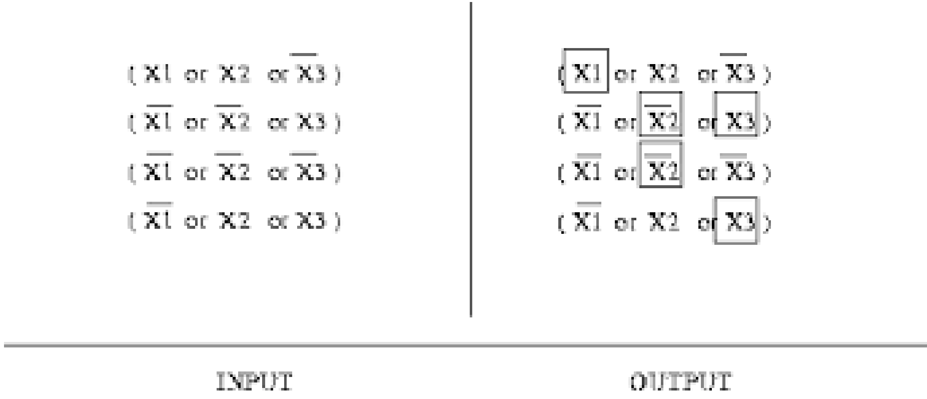
Improve Rule Networks

- Rule networks are neural networks with Boolean weights and activations
  - Each node can be interpreted as a rule
- Basic scikit-learn implementation developed in our group is provided
- Task:
  - Survey the literature on binary/ternary neural networks
    - How to adapt these approaches to the Boolean case?
  - Improve the existing implementation
    - Initialization of rule networks
    - Weight flipping algorithm
    - Multi-class classification
    - … (open for suggestions)



SAT for Inductive Rule Learning

- Several recent works have investigated the use of satisfiability solvers for decision tree learning and rule learning
- Task:
    - survey the field
    - implement and compare different approaches to each other and to SOA machine learning algorithms
    - investigate potential for noise handling
    - possibly develop a new improved approach
    - possibly also for alternative representations, such as decision graphs


$(X1 \text{ or } X2 \text{ or } \overline{X3})$
$(\overline{X1} \text{ or } \overline{X2} \text{ or } X3)$
$(\overline{X1} \text{ or } \overline{X2} \text{ or } \overline{X3})$
$(\overline{X1} \text{ or } X2 \text{ or } X3)$
$(\boxed{X1} \text{ or } X2 \text{ or } \overline{X3})$
$(\overline{X1} \text{ or } \boxed{\overline{X2}} \text{ or } \boxed{X3})$
$(\overline{X1} \text{ or } \boxed{\overline{X2}} \text{ or } \overline{X3})$
$(\overline{X1} \text{ or } X2 \text{ or } \boxed{X3})$
INPUT
OUTPUT

Learning Internal Disjunctions

- Internal disjunctions are sets of values for a condition
  - e.g., `weather = rainy || cloudy`
- Only a few learners can **explicitly** learn such rules
  - e.g., AQ or Bexa
- However, most rule learner can implicitly learn equivalent conditions as conjunctions of negations
  - e.g., `weather != sunny & weather != snowfall`
- Task:
  - Evaluate how rule learners work when *only* internal negated conditions are allowed
  - Possibly compare this to extension that explicitly handle internal disjunctions

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

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 developement 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 rule learning algorithms such as Ripper, AQ, CN2, Foil, etc. 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).

Local Optimal Rule Learning

- Local optimal rule learning (LORD) behaves well on single-label classification and execution performance
  - but the output rule set is often much larger than that by other approaches.

Tasks: at least one of the following two tasks should be done

- Quest for a technique to reduce the output rule set while keeping the classification performance not worse.
- Improve classification performance: from the set of covering rules for a new example, develop a method(s) to select a rule for classifying the new 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.

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

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 treshold 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.