Optimizing probabilistic rule lists with evolutionary algorithms

Objective: Use of an evolutionary algorithm (EA) to optimize probabilistic rule lists over discrete data using the minimum description length (MDL) principle as optimization criterion.

The overall goal of this project is to apply successful methods for optimizing fuzzy systems to the optimization of rule lists using an MDL criterion, to obtain better solutions than simple greedy search and faster than exhaustive search. The context of this idea will be explained in the next paragraphs.

Supervisors: Hugo Proença, Hao Wang, Matthijs van Leeuwen

A fuzzy rule system is a set of fuzzy rules that can be applied to a continuous variable dataset for classification or regression. Genetic and evolutionary algorithms have, for more than two decades, been successfully used for optimizing fuzzy rule systems [1]. Their success arises from two facts. First, the problem of combining rules is NP-complete, which makes it impossible, except for simple cases, to solve the problem optimally. Second, heuristic approaches tend to converge to local optima and miss crucial parts of the space. Usually authors argue that fuzzy rule systems are interpretable, but in practice they tend to be hard to understand for a lay person. The ease of understanding is what makes our desired model class, i.e., rule lists, attractive.

A rule list is a set of ordered rules of the form: if x then y, else if x' then y', that finishes with a default else clause, and that, contrary to fuzzy rules, is usually applied to categorical variables (or to discretized continuous variable). This allows them to be more interpretable, as a statement of the form "if *feathers = yes* then *animal = bird*", is more precise than an fuzzy interval. Recent advances in computational power have made it possible to mine all the possible rules from a dataset, and to try and combine them optimally in a rule list. Most recent approaches use an optimality criterion drawn from Bayesian probability theory [2]. The problem of the optimal combination is that it only works when the number of rules mined from a dataset is small [3].

The MDL principle, originating from information theory (the theory of communication of information), can be seen as a probabilistic framework similar to Bayes theory that can be used to solve the problem of model selection, i.e., it naturally takes into account the goodness of fit (or performance) of the model and the complexity of that model in one single value. This attractive property renders it unnecessary to use multiple objectives to obtain the model, and uniquely reduces the computational burden of optimizing the problem, even though the total number of possible models is, in general, too large to solve optimally. MDL was successfully applied to solve the problem of compactness and performance of rule lists through Classy, a greedy algorithm that searches for the best probabilistic rule lists according to its MDL criterion [4].

Using the existing MDL formulation of probabilistic rule lists [4] and given a set of pre-mined rules, the goal is to possible to apply an evolutionary algorithm to find better models than the current greedy approach does.

Possible steps for the project

1. Literature overview of potential methods to be applied.
2. Adapt existing GA frameworks of Genetic Fuzzy Systems (GFS) to rule lists.
3. Use an already existing MDL formulation of rule lists as the single objective.
4. (extra) Use an interesting rule mining algorithm (from frequent pattern mining literature) to reduce the number of rules mined – example: emerging patterns/rules algorithms.
5. (extra) Construct potential bounds to reduce the search space.
6. Empirically compare the results against the state of the art.
7. Write down the results found.

Requirements/good to have: Python, knowledge of genetic algorithms and data mining/machine learning.

Existing Genetic Fuzzy Systems (GFS) approaches:

Best known methods:
1. Pittsburg approach (a chromosome as a whole rule list)
2. Iteratively rule mining (a chromosome as a rule)

"A historical review of evolutionary learning methods for Mandani-type fuzzy rule-based systems: Designing interpretable genetic fuzzy systems" (2011) - https://dl.acm.org/citation.cfm?id=1975375

"Revisiting Evolutionary Fuzzy Systems: Taxonomy, applications, new trends and challenges" (2015) - https://www.sciencedirect.com/science/article/abs/pii/S0950705115000209

Extra notes on Genetic Fuzzy Systems: it is important to note that fuzzy systems act over continuous data and probabilistic rule lists are usually over discrete data. The ideas from one topic can also be reused in the other but the formulation as to be slightly adapted.

References

[1] Genetic Fuzzy Systems: Overview of the field (2015): https://www.sciencedirect.com/science/article/abs/pii/S0950705115000209

[2] Scalable Bayesian Rule lists (2017) – optimization through sampling from a Markov Chain Monte Carlo (MCMC) https://dl.acm.org/citation.cfm?id=3306086

[3] Mixed integer programming for rule lists (2018): "Learning customized and optimized lists of rules with mathematical programming" - https://link.springer.com/article/10.1007/s12532-018-0143-8

[4] Minimum Description Length (MDL) principle formulation of rule lists (2019): "Interpretable multiclass classification by MDL-based rule lists" https://arxiv.org/abs/1905.00328