

Efficient and effective automatic feature discovery

2022

Andreas Benedict Bien

How to make automatic feature discovery efficient and effective?



Efficiency

This is a common approach of feature selection: ranking features based on their utility. This approach is inefficient because it needs to evaluate all other features to compare each one.

This causes the algorithm to spend a lot of time on the previous step.

My heuristic makes things better: it only needs to evaluate a few features at a time.

Read more about my heuristic on my GitHub page.

Approach	Time	Efficiency	Complexity
Ranking	Fast	Slow	Very slow
Heuristic	Slow	Fast	Very fast

Approach	Time	Efficiency	Complexity
Ranking	Fast	Slow	Very slow
Heuristic	Slow	Fast	Very fast

Methodology

Approach

I developed a heuristic that uses statistical characteristics of the data to rank features. It then iteratively removes the least useful feature by calculating the likelihood of a feature being useful given the current state of the feature.

Process

1. Calculate feature importance
2. Remove feature with lowest importance
3. Calculate new importance for all remaining features
4. Repeat steps 2 and 3 until no more features are available
5. Calculate the change in importance
6. If change is less than a threshold, stop
7. Calculate new importance for all remaining features

Dataset

One dataset is selected at random to evaluate the performance of the heuristic.

Energy Data from per patient for different medical characteristics



On the right, I show the results of the heuristic against the ranking approach. The ranking approach is slower and less efficient.

This was done on a random dataset.

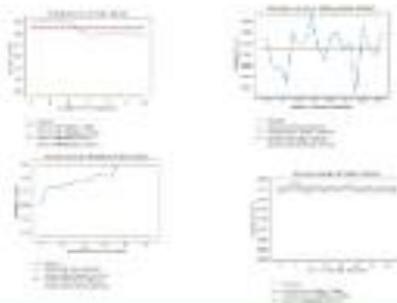


Experiments

Four experiments performed and efficient feature selection has been tested and proved to work.

Experiment	Description
Ranking	Ranking of features based on their utility.
Heuristic	Heuristic based on statistical characteristics of the data.
Random	Random selection of features.
Baseline	Baseline selection of features.

Main Results



Conclusion

The heuristic performs well. In 3/4 datasets it performs equally or better than all other experiments, with very K values less than 5.

Future work

Collect more data on the datasets for higher precision on further analysis.

A collection of data thousands of datasets instead of hundreds could allow for better analysis of the covariance between statistical characteristic scores, and the independence of the likelihoods.

Pair with machine learning:

The heuristic itself could be fine-tuned using a reinforcement learning algorithm, or possibly used as an input to train a machine learning model that predicts the utility of the feature.

Pair with other heuristics:

Pairing the model with heuristics that take into account the statistical values rather than just the rankings might synergize well with a heuristic that only considers ranking of features, as it's the heuristic I developed is less precise when all features similar statistical characteristics.