Towards understanding learning behavior

Joaquin Vanschoren

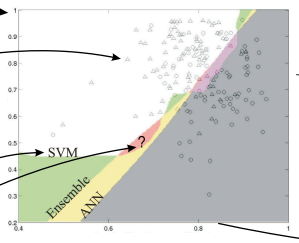
Meta-learning

How to select the right learning algorithm for a given dataset?

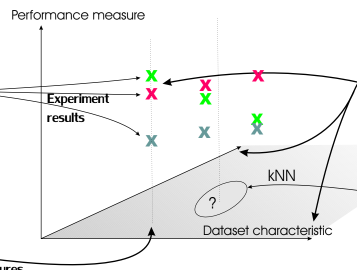
Base-Learning

Problem domain →

1) Define instance space
2) Measure instances (build dataset)
3) Preprocess data (sampling, discretizing, feature selection,...)
4) Choose learning algorithm (+ parameter settings)
5) Predict new instances/ derive rules



Dataset meta-features



Meta-Learning

Experience experiments →

1) Define meta-features for datasets (reflecting restrictions of various learning algorithms)
2) Measure algorithm performance in experiments (build meta-dataset)
3) Preprocess experiment data
4) Choose learning algorithm (+ parameter settings)
5) Predict algorithm performance/ derive rules, rankings

"Traditional" approach

Dataset characteristics			Performance measures			
Size	nb. Attr.	...	Algorithm	Acc	RunT	...
2300	34		C4.5	.92	43	

Publicly available datasets

Often only default parameter settings

Limitations:

1) The Curse of Dimensionality → We need many more datasets
2) Generalisation over algorithms? → Algorithms should be characterized
3) Only when, not why... → More thorough investigation needed
4) What about preprocessing? → Effects should be included

Towards Descriptive Meta-learning

Experiment database

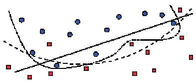
			Performance measures		
Algo impl.	Param. sett.	Dataset	TP	FP	...
C4.5 v.1	C451 - 1	Dataset1	945	84	

Dataset characteristics

Id	nb. samples	...
Dataset1	2300	

C4.5 v.1 parameter sett.

Id	min. leaf size	...
C451 - 1	2	

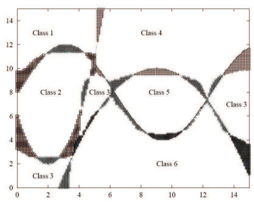


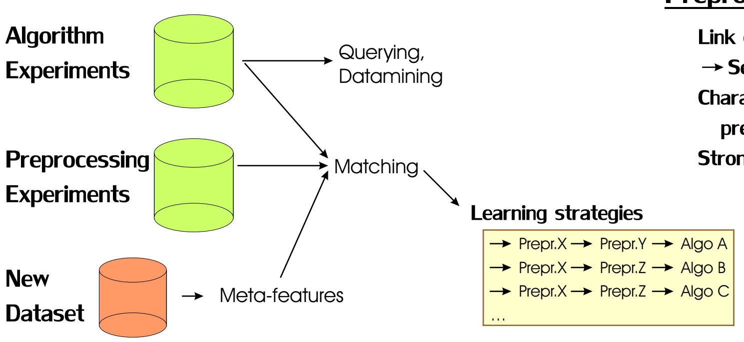
general algorithm properties

Id	model	linear?	...
C4.5 v.1	DT	no	

Performance measures

TP	FP	...	bias err	variance err
945	84		43	62





Experiment databases

Make experiments reusable, reproducible: log all experiment details and results
Querying or datamining on stored results
→ Thorough investigation of interactions between algorithm parameters, dataset characteristics and performance measures

Synthetic datasets

Unbiased: hide a large range of different kinds of concepts in the data, and characterise: model characteristics, concept variation, example cohesiveness,...
"Natural": approximate characteristics of natural datasets: complex relations between attributes, noise, irrelevant attributes, missing values,...
Coverage: control characteristics to cover meta-feature space: experiment design

Algorithm characterization

Parameter settings: parameters and techniques used (1 table/algorithm)
General algorithm properties: representation model, dependency on linear separability or conditional independence, use of data fragmentation or attribute summations, ability to handle fine-grained concepts or local relevance,...

Understanding inductive performance

Averaged perf. measures do not explain *why* an algorithm failed/succeeded
→ Decompose the misclassification error: *bias error* vs. *variance error*

repr. bias	comp. bias	bias error	variance error
appropriate	too strong	high	low
appropriate	ok	low	low
appropriate	too weak	low	high
inappropriate	too strong	high	low
inappropriate	ok	high	average
inappropriate	too weak	high	high

link to dataset/ algorithm characteristics
→ advice possible improvements

Preprocessing steps

Link dataset characteristics to effect of preprocessing techniques
→ Separate experiment database for preprocessing experiments
Characterise dataset and predict changes after preprocessing, or advice useful preprocessing for optimizing performance of specific algorithm
Strong link with bias/variance error

Declarative Languages and Artificial Intelligence research group, Department of Computer Science, K.U.Leuven, Belgium