ITMO UNIVERSITY

Datasets Meta-Feature Description for Recommending Feature Selection Algorithm

Andrey Filchenkov and Arseniy Pendryak

Computer Technology Chair

Computer Technologies Lab

ITMO University

afilchenkov@niuitmo.ru

AINL-ISMW FRUCT'15. Nov 14, 2015, St. Petersburg, Russia

Algorithm selection problem

- Many domain fields: (number grows!)
- Many datasets: (number grows!)
- Extremely many algorithms: (number dramatically grows)
- Very high demand in data scientist: (will also grow due to the universities are not capable to create as much programs as it is required to fulfill the growing demand).

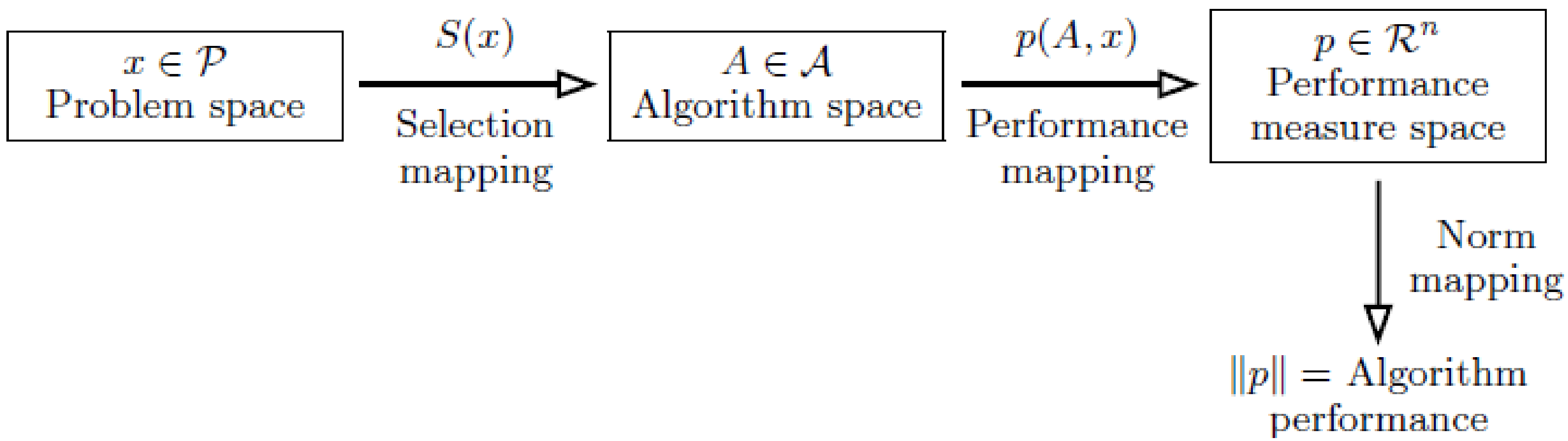
Approaches to choose algorithms

- Use a single hopefully good algorithm
- Try all the algorithms from a predefined set (implemented in the library you use)
- Apply your experience and intuition to choose a single / rank some algorithms
- Use some algorithms and combine them into an ensemble or a mixture

Framework for algorithm selection problem

Rice's algorithm selection problem:

The objective is to determine $S(x)$ [the mapping of problems to algorithms] maximizing high algorithm performance."

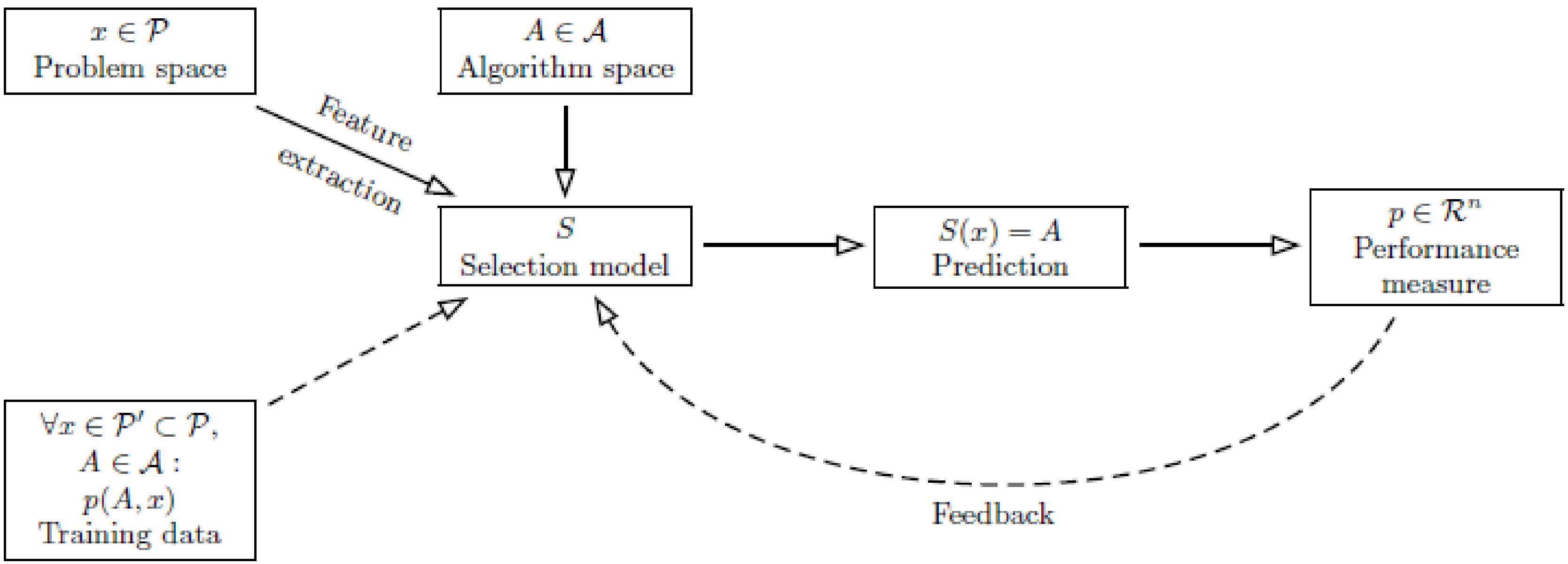


How to choose algorithms properly

- characterize problem space with **features**
- **learn** performance models for prediction
- use models to predict best algorithm from an **algorithm set** with respect to a **quality measure**

Meta-learning framework

- Rice's problem is rethought in meta-learning framework:



Feature selection algorithms

- Can be divided into **filters** and **wrappers**
- Feature selection algorithms are faster than feature extraction algorithms
- Filters are faster than wrappers
- Plethora of algorithms to select features
- Can be described as

  Subspace search algorithm + **Subspace estimator (measure)**

Meta-learning for feature selection

- Yes, but only one work:

G. Wang, Q. Song, H. Sun, X. Zhang, B. Xu and Y. Zhou (2013) "A Feature Subset Selection Algorithm Automatic Recommendation Method", Volume 47, pages 1-34.

- No explanation for meta-feature choice
- Only 13 meta-features.

Goal of this work

- Can choice of meta-features improves system performance?
- What are the best meta-features for dataset description for feature selection algorithms selection?

Goal of this work is to select optimal meta-feature space for meta-learning system recommending feature selection algorithms.

The paper could be titled "*Meta-feature selection for feature selection algorithm selection algorithm based on meta-learning*" (but we are merciful).

Experiment setup

- We use the same experiment setup
- 84 datasets
- 16 feature selection algorithms
- kNN algorithm on meta-level
- Algorithm ranking (not only the best algorithm) prediction

Feature selection algorithms

#	Algorithm	Search method	Evaluation measure
1	CFS-SFS	BestFirst + Seq. Forward Search	Dependency
2	CFS-SBS	BestFirst + Seq. Backward Search	Dependency
3	CFS-BiS	BestFirst + Bi-direction Search	Dependency
4	CFS-GS	Genetic Search	Dependency
5	CFS-LS	Linear Search	Dependency
6	CFS-RS	Rank Search	Dependency
7	CFS-SS	Scatter Search	Dependency
8	CFS-SWS	Greedy Stepwise Search	Dependency
9	CFS-TS	Tabu Search	Dependency
10	Cons-SFS	BestFirst + Seq. Forward Search	Consistency
11	Cons-BiS	BestFirst + Bi-direction Search	Consistency
12	Cons-GS	Genetic Search	Consistency
13	Cons-LS	Linear Search	Consistency
14	Cons-RS	Rank Search	Consistency
15	Cons-SWS	Greedy Stepwise Search	Consistency
16	Signific	Ranker	Prob. Significance

Meta-features

- Basic meta-features:
  - **general**: the number of samples in dataset, the number of features, the number of class, etc.
  - **statistical**: standard deviation, correlation coefficient, asymmetry coefficient, etc.
  - **information-theoretic**: mean feature entropy, mutual information of class and attribute, noise ratio, etc.
- **Landmark features**: performance value of a classifier trained on a small subsample of dataset.
- **Decision tree structure**

Novel meta-features:

- **Classifiers structure** (analogue of DT-based): numerical parameters with which the classifiers are described (like the sum of weights in perceptron).
- **Classifiers best parameters**: numerical parameters which helps to show the best results (like the best k for kNN).

- Since learnt classifiers are model of the data, they are good in descripting data properties

Meta-feature set

79 meta-features:

- General: 5
- Statistical: 5
- Information-theoretic: 6
- DT-based: 38 (19 for pruned and 19 for unpruned)
- Perceptron (neural): 13
- kNN based: 13

Results on subsets (group choice)

Meta-feature set	NaiveBayes	C4.5	PART	BayesNet	IB3
Recsys	**0.9738**	0.9653	0.9635	**0.9714**	0.9530
General	0.9636	0.9675	0.9589	0.9650	0.9530
Statistical	0.9391	0.9571	0.9532	0.9533	0.9448
Information theoretic	0.9691	0.9616	0.9529	0.9589	0.9520
Standard	0.9732	0.9643	0.9634	0.9711	0.9534
Pruned	0.9502	0.9569	0.9563	0.9655	0.9209
Unpruned	0.9639	0.9537	0.9528	0.9576	0.9403
Tree	0.9481	0.9578	0.9522	0.9645	0.9505
Neural	0.9549	0.9519	0.9477	0.9631	0.9513
kNN	0.9571	0.9540	0.9187	0.9515	0.9240
Classifier based	0.9575	0.9648	0.9574	0.9637	**0.9620**
All	0.9601	**0.9725**	**0.9653**	0.9642	0.9516

Results on feature selection

Manually selected feature selection algorithm

Meta-features	NaiveBayes	C4.5	PART	BayesNet	IB3
Original	0.9738	0.9653	0.9635	0.9714	0.9530
Group choice	0.9738	0.9725	0.9653	0.9714	0.9620
MFSO general	**0.9739**	**0.9841**	**0.9720**	**0.9755**	**0.9725**

Selected meta-features (18):

DataSetDimensionality	**UnprunedTreeLeavesNumber**
MeanSkewness	**UnprunedTreeMaxAtt**
MeanNormalizedFeatureEntropy	**UnprunedTreeMaxLevel**
MaxMutualInformation	**UnprunedTreeMinBranch**
PrunedTreeDevAttr	**UnprunedTreeDevClass**
PrunedTreeMinAttr	**MinSqrtBestK**
PrunedTreeMinBranch	**MinOneTenthBestK**
PrunedTreeDevClass	**MeanHalfBestK**
PrunedTreeMaxClass	**MinHalfPerceptronWeightSum**

Results on wrappers

	NaiveBayes	C4.5	PART	BayesNet	IB3
NaiveBayes	**0.9808**	0.9700	0.9665	0.9691	0.9600
C4.5	0.9639	**0.9923**	0.9542	0.9570	0.9422
PART	0.9631	0.9630	**0.9816**	0.9768	0.9697
BayesNet	0.9645	0.9601	0.9693	**0.9820**	0.9592
IB3	0.9581	0.9841	0.9587	0.9634	**0.9830**

Meta-features	NaiveBayes	C4.5	PART	BayesNet	IB3
Original	0.9738	0.9653	0.9635	0.9714	0.9530
Group choice	0.9738	0.9725	0.9653	0.9714	0.9620
MFSO general	0.9739	0.9841	0.9720	0.9755	0.9725
MFSO single	**0.9808**	**0.9923**	**0.9816**	**0.9820**	**0.9830**

Meta-feature set

79 meta-features:

- General: 5
- Statistical: 5
- Information-theoretic: 6
- DT-based: 38 (19 for pruned and 19 for unpruned)
- Perceptron (neural): 13
- kNN based: 13

What to do with time

Time-consuming evaluation of meta-features

- Use parameter tuning (SMBO) or surrogates
- Trade-off between significance and evaluation time
- Algorithm comparison when time is not so crucial

Summary

- Comparison of meta-features for feature selection algorithms
- Novel meta-feature selection types

Thank you!

Questions?

Andrey Filchenkov

aaafil@mail.ru
afilchenkov@mail.ru