




Automated Machine Learning

André C P L F de Carvalho
Universidade de São Paulo, Brazil







Acknowledgements

- Alex Freitas, Kent
- Ana Carolina Lorena, ITA
- André Rossi, UNESP
- Bruno A Pimentel, USP
- Bruno F de Sousa, UFMA
- Carlos Soares, U. Porto
- Davi P Santos, U. Porto
- Edésio Alcobaça, USP
- Joaquin Vanchoren, TU/E
- João Moreira, U. Porto
- Jorge Kanda, UFAM
- Luis P Garcia, USP
- Márcio Basgalupp, UNIFESP
- Osvaldo Anacleto Jr, USP
- Péricles Miranda, URFPE
- Rafael Mantovani, USP
- Ricardo Cerri, UFSCar
- Ricardo Prudêncio, UFPE
- Rodrigo Barros, PUCRS
- Taciana Gomes, UFPE
- Victor Hugo Barella, USP

Andre Ponce de Leon de Carvalho



Summary

- Introduction
- Machine learning
- AutoML
- AutoML at ICMC-USP
- Responsible Data Science
- Conclusion

Andre Ponce de Leon de Carvalho

3



Introduction

- Data Scientists use knowledge from different areas, like
 - Databases
 - Linear algebra
 - **Machine Learning**
 - Statistics
 - High-performance computing
 - Visualization

André P L F de Carvalho

4



Introduction

- Machine Learning (ML) is strongly related to Artificial Intelligence
- Pressreader, November 2018
 - Canadian news company that produce and distribute digital newspapers
 - Each US\$ 1 invested in good quality Artificial Intelligence (AI)
 - Can bring a return of 44

Andre Ponce de Leon de Carvalho

5



Introduction

- In many Data Science applications, we need to create data models from data
 - Machine Learning algorithms have been often used
 - High performance in several applications
 - There is a high demand for Machine Learning experts
 - But few number of experts available

Andre Ponce de Leon de Carvalho

6

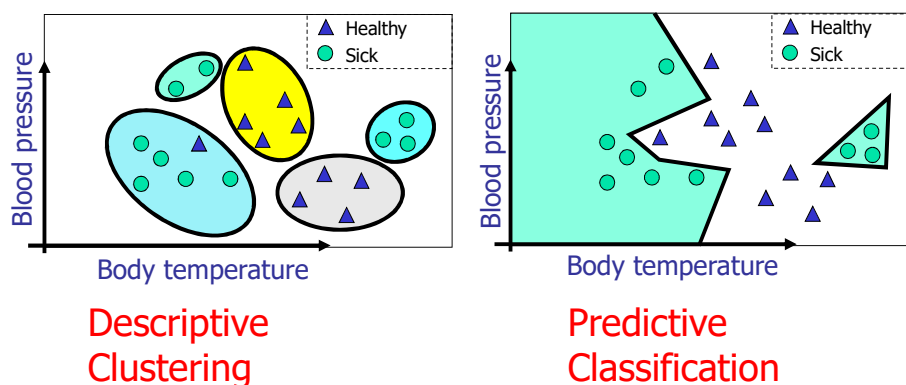
Machine Learning

- Investigate algorithms able to learn models from data
 - Automatically, reducing (removing) human interference
- Has been successfully used in many data modelling tasks
 - Descriptive
 - Predictive

© André de Carvalho - ICMC/USP

7

Machine Learning



André P L F de Carvalho

8



Machine Learning

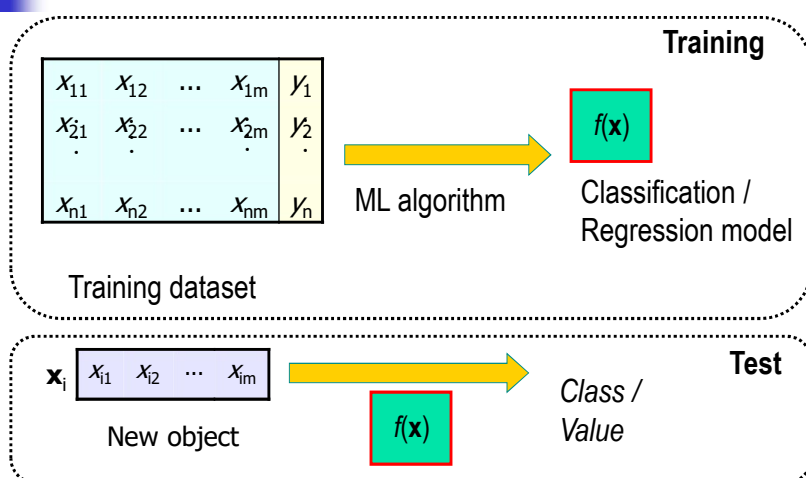
- For this talk, assume the predictive task context
 - Concepts and methodologies can be easily adapted to the descriptive task context
- Conventional predictive learning process
 - Application of a ML algorithm to a dataset induces a model (hypothesis, function)
 - E.g. ANN, SVM, NB (base algorithm)
 - Induced model \leftrightarrow Predictive function

Andre Ponce de Leon de Carvalho

9



Predictive task



Andre Ponce de Leon de Carvalho

10



Dataset tabular format

Predictive attributes				Diagnosis
Body Temp.	Age	Weight		
Examples (objects, instances)	37.0	70	94	Healthy
	39.2	30	40	Unhealthy
	38.5	70	85	Unhealthy
	37.4	15	60	Healthy
	40.1	90	78	Unhealthy
				Target attribute

Examples
(objects,
instances)

Andre Ponce de Leon de Carvalho

11



End-to-End Machine Learning

- ML successful application includes more than just model induction
 - Every step from creation of a dataset to design of a ML-based system, including:
 - Data integration, curation and preprocessing
 - Model selection
 - Model induction
 - Hyperparameter tuning
 - Code parallelization
 - ...
 - Pipeline definition

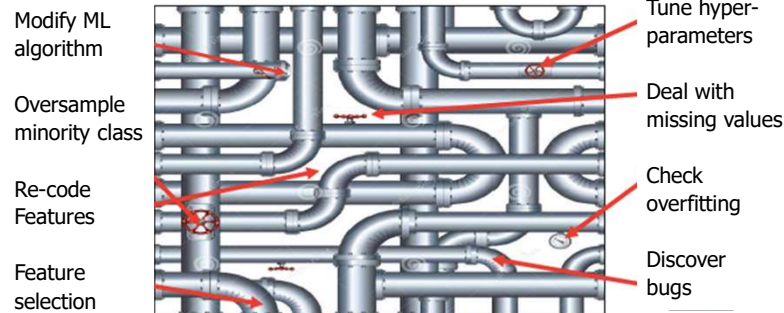
Andre Ponce de Leon de Carvalho

12



End-to-End Machine Learning

Machine Learning (Research) Pipeline



Adapted from Rick Caruana, Research opportunities in AutoML Microsoft Research

Andre Ponce de Leon de Carvalho

13



Machine Learning algorithms

- Tens of thousands
 - Hundreds more every year
 - Taking new aspects into account
 - Using new clever heuristics
 - Each new algorithm comes with a set of hyperparameters to be tuned
 - Not mentioning other techniques used
 - E.g. feature selection, noise detection, ...

Andre Ponce de Leon de Carvalho

14



Key question

- How to have the best performance?
 - Which ML algorithm (and hyperparameter values) can induce the best model for a given dataset?

Andre Ponce de Leon de Carvalho

15



Key question

- How to have the best performance?
 - Which ML algorithm (and hyperparameter values) can induce the best model for a given dataset?
 - Two hypothesis:
 - There is a master algorithm
 - The most suitable algorithm is domain (data) dependent

Andre Ponce de Leon de Carvalho

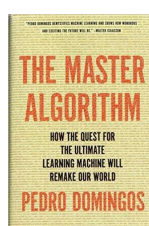
16



The Master algorithm

An unique ML algorithm can beat all others in any data analysis task

Pedro Domingos, University of Washington



"Yoda algorithm"

Andre Ponce de Leon de Carvalho

17



The Master algorithm

- What is known:
 - A small subset of ML algorithms can efficiently cover most of the tasks
 - Different from programming languages, the same ML algorithm can be efficiently used in several different tasks
- Can one ML algorithm induce the best possible model for any task (dataset)?

Andre Ponce de Leon de Carvalho

18



The Master algorithm

- Can one ML algorithm learn everything that is possible to learn from any dataset?
 - Universal ML algorithm
 - Assumption: if a ML algorithm has enough adequate data, it can learn anything
 - Even if it has to be infinite data
- Do we (our brain) use a single algorithm to learn?

Andre Ponce de Leon de Carvalho

19



The Master algorithm

- Does not need to be as computationally efficient as the most suitable algorithm
 - Considers generalization to be more important than computing cost
- To learn, can use more data than a "specialized" ML algorithm

André de Carvalho - ICMC/USP

20



Machine Learning

- Selective superiority (Brodley 1995)
 - No free lunch
 - Each algorithm can perform better than others on a subset of tasks
 - Each ML algorithm has an inductive bias
 - Necessary for learning to occur
 - Preference for selecting one particular hypothesis over other hypotheses
 - Affects predictive performance in different datasets

Andre Ponce de Leon de Carvalho

21




The most suitable algorithm

- It is possible to select the most suitable algorithm for a given task (dataset)
 - And its hyperparameter values
 - And any other pre-processing (or post-processing) technique
- Automated Machine Learning (AutoML) approach


Andre Ponce de Leon de Carvalho

22


AutoML tools




Auto-WEKA




MLJAR




DataRobot




Amazon Rekognition




Auto-Sklearn




H2O.ai




Cloud AutoML Vision




CreateML Apple



TPOT



AutoML.org Freiburg



clarifai

Andre Ponce de Leon de Carvalho

23

Challenges (Competitions)

- Workshop@ICML2014
- Workshop@NIPS2015
- Bootcamp@Stanford2015
- MLIschool@Petersburg2015
- Hackathon@ICML2015
- Workshop@ICML2015
- Hackathon@ESPCI2015
- Competitions@WCCI2016
- Workshop@ICML2016
- GPUtrack@ICML2016
- Workshop@ICML2017
- Workshop@PRICAI2018
- Workshop@ICML2018
- Workshop@PAKDD2018

PAKDD 2018 Data Competition

Automatic Machine Learning Challenge 2018: Towards AI for Everyone
(Provided and Sponsored by the Fourth Paradigm Inc. and ChaLearn)

NIPS 2018 Challenge

The 3rd AutoML Challenge: AutoML for Lifelong Machine Learning
(Provided and Sponsored by 4Paradigm, ChaLearn, Microsoft and Acadia University)

Andre Ponce de Leon de Carvalho

24



AutoML

- Fully Automatic Machine Learning without ANY human intervention
- Related with issues like:
 - Meta-learning
 - Optimization
 - Transfer learning
 - Pipeline definition
 - ...

Andre Ponce de Leon de Carvalho

25



AutoML

- Main approaches
 - Optimization
 - Algorithms and/or hyperparameters
 - Design of new algorithms
 - Meta-learning
 - Algorithms and/or hyperparameters
 - Propose the use of one or more existing algorithms
 - Hybrid

Allow correct and efficient use of End-to-End ML by non-experts

Support ML experts with insights and relieving them from repetitive tasks

Andre Ponce de Leon de Carvalho

26



Optimization

- There are several optimization algorithms (methods)
 - Exact methods
 - Heuristics
 - Meta-heuristics
 - Single-based
 - Population-based
 - Evolutionary algorithms
 - ...

Andre Ponce de Leon de Carvalho

27



Evolutionary Algorithms

- Computational techniques for problem solving based on:
 - Genetics
 - Natural selection
- Research started in the 1950s
 - Independently from around 10 groups in a 15 years time span

28



Evolutionary Algorithms

- Works can be grouped in 4 main areas
 - Genetic Algorithms
 - Genetic Programming
 - Evolution Strategies
 - Evolutionary Programming
- Features from each area are being assimilated by the others
 - It is difficult to draw borders

29



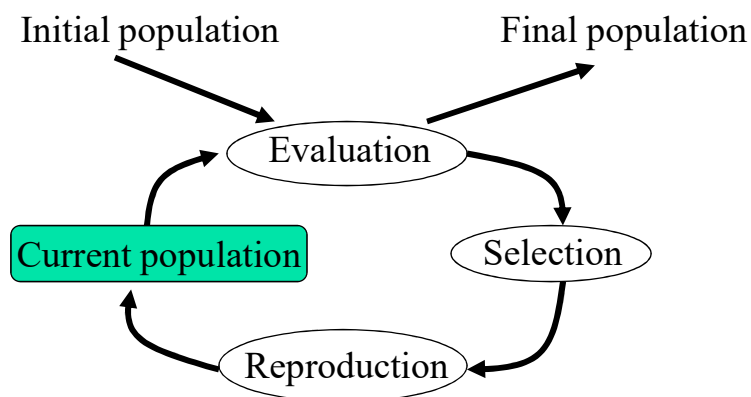
Genetic Algorithms

- Search and optimisation method
 - Based in genetics and natural selection
 - Use a population of candidates (individuals)
 - Optimisation occurs through several generations
 - At each generation
 - Selection mechanism chooses the fittest individuals (chromosomes)
 - Genetic operators produce new individuals from those selected

30



Genetic Algorithms



31



Reproduction

- Produces new generations of individuals
 - Individuals should improve their fitness at each new generation
- Genetic operators transform the population
 - Crossover
 - Mutation
 - Elitism

32

Crossover

- Recombines parents features
 - Allows heritage of desirable features by the next generations

↖ Crossover point

Parent 1	0.7	0.3	22	16	12
Parent 2	0.2	0.5	10	8	5
Offspring 3	0.7	0.3	22	8	5
Offspring 4	0.2	0.5	10	16	12

33

Mutation

- Responsible for the introduction and maintenance of genetic diversity in the population
 - Modifies genes at random
 - Reduces incidence of local minima

↖ Mutation point

Before mutation:	0.2	0.5	18	8	5
After mutation:	0.2	0.5	10	8	5

34



Genetic Programming

- Evolve computer programs in an evolutionary process
 - Allow automatic design of new programs
- Individuals are programs
 - Represented by trees instead of codes
 - Can have different size and formats
 - Evaluated by how well, when run, they solve a given task

05/04/2019

André de Carvalho - ICMC/USP

35



Genetic Programming

- The position of a node in the tree defines its role in the program
 - Internal nodes
 - Arithmetic and logic functions, programming commands, function calls
 - External nodes (leaves)
 - Variables, constants, input/output
- Alternative: grammar representation

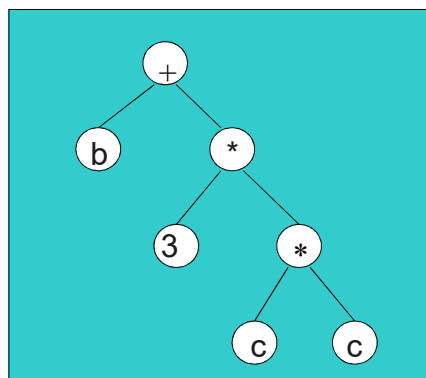
05/04/2019

André de Carvalho - ICMC/USP

36

Exemple

- Tree to calculate: $b + 3 * c^2$

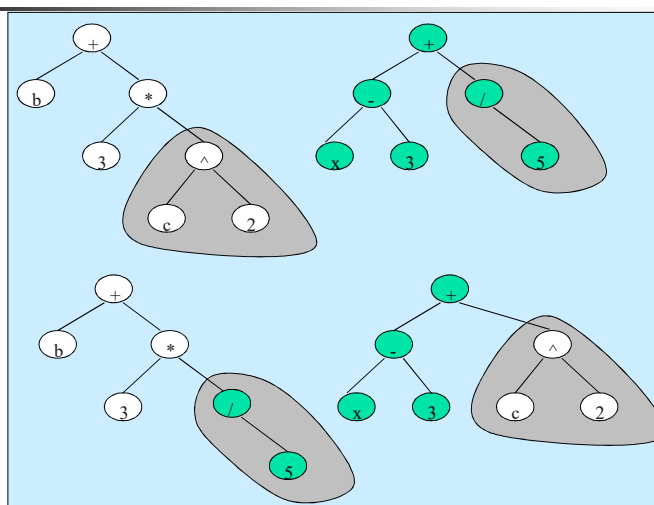


05/04/2019

André de Carvalho - ICMC/USP

37

Crossover

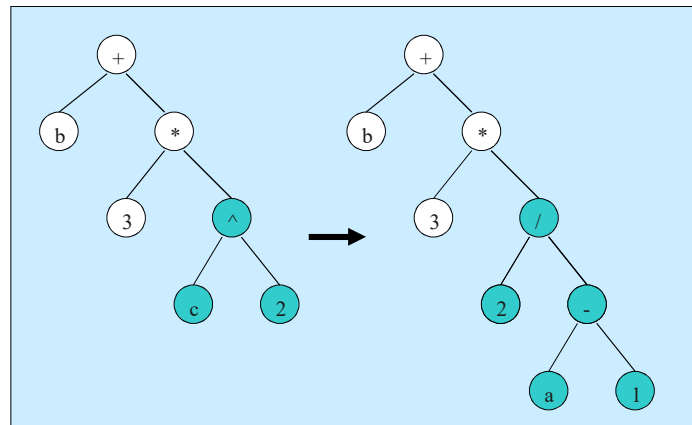


05/04/2019

André de Carvalho - ICMC/USP

38

Mutation



05/04/2019

André de Carvalho - ICMC/USP

39

Optimization

- Hyperparameter tuning
 - Neural networks
 - SVMs
 - ...
- Design of new algorithms
 - Rule learning algorithms (RLAs)
 - Decision tree induction algorithms (DTIAs)
 - Bayesian classification algorithms (BCAs)

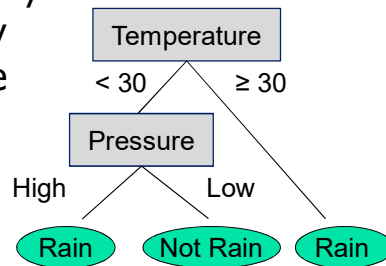
Andre Ponce de Leon de Carvalho

40



Case study: HEAD-DT algorithm

- Hyper-heuristic Evolutionary Algorithm to automatically Designing – Decision Tree induction algorithms
- Automates the design of full top-down DTIAs
 - Different from evolutionary design of DTs



Andre Ponce de Leon de Carvalho

41



HEAD-DT algorithm

- DTIAs algorithms takes months or years to be designed by ML researchers
- HEAD-DT can design new DTIAs in seconds
 - Even faster using a parallel architectures
- Combine components from existing DTIAs
 - Using Evolutionary Computation
 - Genetic Algorithms (GA) / Genetic Programming (GP)

Andre Ponce de Leon de Carvalho

42



HEAD-DT algorithm

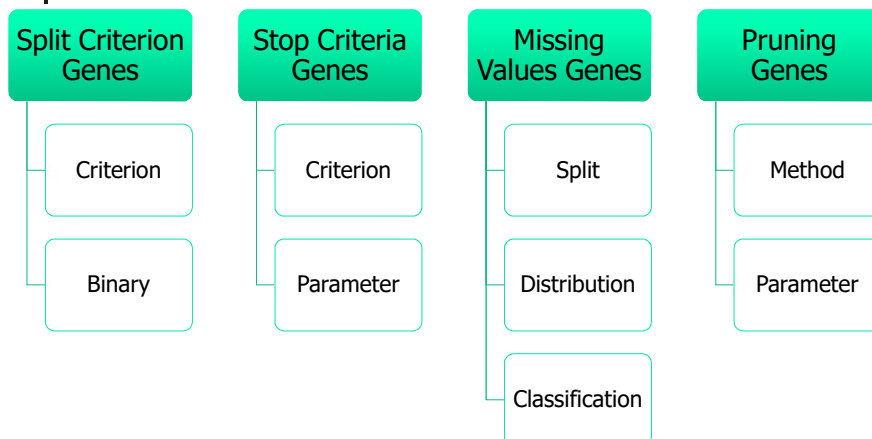
- Can evolve DTIAs for a:
 - Specific dataset
 - Group of datasets
- Represents DTIAs in 3 different ways:
 - Linear genome GA-like approach
 - Grammar-based approaches
 - Standard grammar-based GP
 - Grammatical Evolution

Andre Ponce de Leon de Carvalho

43




Linear representation



Andre Ponce de Leon de Carvalho

44



Algorithm representation

Split Criterion Genes


Criterion

Binary

- Information gain (Quinlan, 1986)
- Gini index (Breiman *et al.*, 1984)
- Global mutual information (Gleser & Colleen, 1972)
- G statistics (Mingers, 1987)
- Mántaras criterion (Mántaras, 1981)
- Hypergeometric distribution (Martin, 1997)
- Chanda-Varghese criterion (Chandra & Varghese, 2009)
- DCSM (Chandra *et al.*, 2010)
- χ^2 (Mingers, 1989)
- Mean posterior improvement (Taylor & Silverman, 1993)
- Normalized gain (Jun *et al.*, 1997)
- Orthogonal criterion (Fayyad & Irani, 1992)
- Twoing (Breiman *et al.*, 1984)
- CAIR (Ching *et al.*, 1995)
- Gain Ratio (Quinlan, 1993)

45

Andre Ponce de Leon de Carvalho



DTIA designed by HEAD-DT

Algorithm

1. Recursively split nodes using the **Chandra-Varghese criterion**
2. Aggregate nominal splits in **binary subsets**
3. Perform step 1 until **class-homogeneity** or **the minimum number of 5 instances** is reached
4. Perform **MEP pruning** with **m = 10**

If dealing with missing values:

1. Calculate the split of missing values by performing **unsupervised imputation**
2. Distribute missing values by **assigning the instance to all partitions**
3. To classify an instance with missing values, **explore all branches and combine the results**

46

Andre Ponce de Leon de Carvalho

Meta-learning

- Learn from previous learning experiences

- Learns a function (meta-model) associating:

Input Characteristics extracted from a dataset

Output Performance of ML algorithms applied to this dataset



- Meta-model can

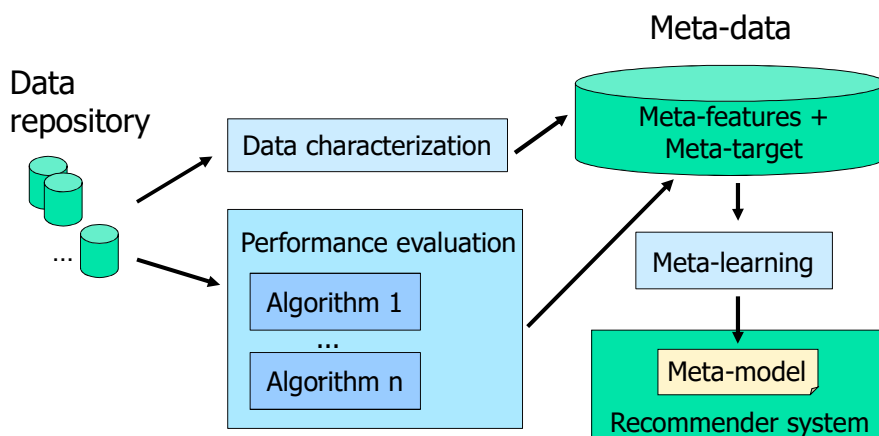
- Predict the best algorithm(s) for new datasets
 - Be part of a recommender system

- Combines base-level and meta-level learning

Andre Ponce de Leon de Carvalho

47

Meta-learning



Andre Ponce de Leon de Carvalho

48



Main steps

- Before
 - Selection of base ML algorithms
 - Acquisition of datasets
- During
 - Generation of meta-data
 - Induction of meta-model
 - Evaluation of meta-model (both levels)
 - Use meta-model in a recommender system

Andre Ponce de Leon de Carvalho

49



Generation of metadata

- Identify properties of datasets that can affect the performance of ML algorithms
- Meta-examples
 - Predictive attributes (predictive meta-features)
 - Dataset characteristics
 - Direct characterization
 - Model-based characterization
 - Landmarking
 - Target attribute (target meta-features, meta-target)
 - Performance obtained by a set of algorithms

Andre Ponce de Leon de Carvalho

50



Direct characterization

- Select data descriptions directly from each dataset
 - Describe the main aspects of the dataset
- Meta-features
 - General simple measures
 - Statistical measures
 - Information-theoretic measures

Andre Ponce de Leon de Carvalho

51



Direct characterization

- Examples of meta-features:
 - Number of classes
 - Number of attributes
 - #examples / #attributes
 - Correlation between predictive attributes
 - Corr. between pred. attr. and target attr.
 - Average class entropy

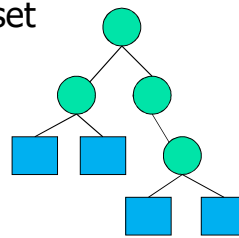
Andre Ponce de Leon de Carvalho

52



Model-based characterization

- Characterize a dataset by exploiting properties of an induced model
- Example of meta-features:
 - Properties of a decision tree induced by a ML algorithm for a given dataset
 - Number of leaf nodes
 - Tree shape
 - Maximum tree depth
 - Tree imbalance degree



Andre Ponce de Leon de Carvalho

53



Landmarking

- Exploit information obtained by running a set of fast and simple algorithms (landmarkers)
 - Run landmarkers for a short period of time
 - Landmarks must have different learning bias
 - Performance of these algorithm characterize a dataset
 - Datasets are similar when landmarkers applied to them present similar performances

Andre Ponce de Leon de Carvalho

54



Landmarking

- Examples of meta-features:
 - Recall for landmark algorithm 1
 - Precision for landmark algorithm 1
 - AUC for landmark algorithm 1
 - Recall for landmark algorithm 2
 - Precision for landmark algorithm 2
 - AUC for landmark algorithm 2
 - ...

Andre Ponce de Leon de Carvalho

55



Complexity measures

- Measures the complexity (difficulty) of a dataset for classification tasks
- Originally proposed in
 - Ho, T. and Basu, M.: Complexity measures of supervised classification problems. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(3):289-300, 2002

André de Carvalho - ICMC/USP

56



Complexity measures

- Can measure
 - Overlap of feature values for different classes
 - Data linear separability
 - Data geometry, topology, and density of manifolds
 - Binary classification performance
 - Adapted for multiclass classification

André de Carvalho - ICMC/USP

57



Meta-target

- Performance measures:
 - Predictive performance
 - Accuracy, AUC, F-measure, MSE, ...
 - Processing cost
 - Time (learning / application)
 - Storage cost
 - Model complexity
 - Knowledge extracted (interpretability)
 - Multi-objective

Andre Ponce de Leon de Carvalho

58



Dataset tabular format

Predictive attributes				Diagnosis
Body Temp.	Age	Weight		
37.0	70	94		Healthy
39.2	30	40		Unhealthy
38.5	70	85		Unhealthy
37.4	15	60		Healthy
40.1	90	78		Unhealthy

Examples (objects, instances)

Target attribute

Andre Ponce de Leon de Carvalho

59



Meta-dataset tabular format

Meta-features				Algorithm
MF01	MF02	MF03	MF04	
0.4	6	0.2	0.8	A
0.1	2	0.2	0.5	A
0.7	0	0.9	0.8	B
0.2	4	0.7	0.1	A
0.6	2	0.3	0.4	B

Meta-examples (meta-objects, meta-instances)

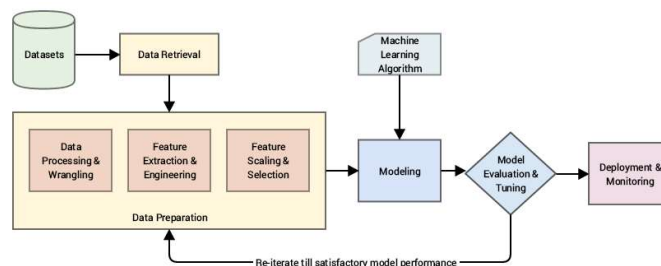
Meta-target

Andre Ponce de Leon de Carvalho

60

ML Pipeline

- Ordered sequence of operations or algorithms for end-to-end ML
 - Workflows are part of many AutoML tools



André de Carvalho - ICMC/USP

61

Pipeline for AutoML

- Distinct workflows can result in solutions with different performance
- Look for the best pipeline
 - Alternatives:
 - Evaluate all possible pipelines
 - Optimize
 - Predict
 - Lack of large number of previously assessed workflows
 - Expand only the most promising nodes/branches
 - Collaborative filtering

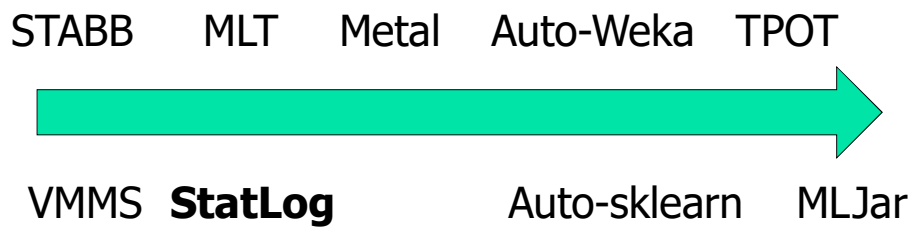
} Computational cost

André de Carvalho - ICMC/USP

62



AutoML along the time



Andre Ponce de Leon de Carvalho

63



StatLog

- Supported by European Community
 - Daimler-Benz AG, Brainware GmbH, Isoft and European Universities (1991-1994)
- First systematic and large-scale effort to relate:
 - Direct characterization measures
 - Predictive performance, running time, interpretability ...

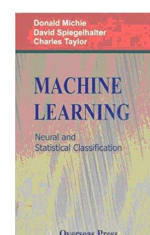
André de Carvalho - ICMC/USP

64



StatLog

- Classification tasks
 - 23 algorithms and 21 datasets
 - Most from UCI
- Results published in articles and a book
 - Machine Learning, Neural and Statistical Classification, D. Michie, D.J. Spiegelhalter, C.C. Taylor (eds.)



André de Carvalho - ICMC/USP

65



StatLog meta-features

- Log of number of examples
- Log of number of attributes
- Log of number of classes
- Mean absolute skewness
- Mean kurtosis
- Geometric mean ratio of the standard deviations of individual populations to the pooled standard deviations
- First canonical correlation
- Proportion of total variation explained by the first canonical correlation
- Normalized class entropy
- Average absolute correlation between continuous attributes per class

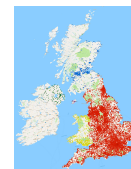
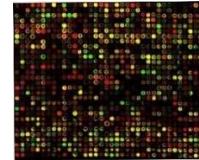
André de Carvalho - ICMC/USP

66



Case studies

- Classification
 - Gene expression analysis
- Regression
 - Travel duration prediction
- Optimization
 - Travelling Salesman Problem (TSP)



Andre Ponce de Leon de Carvalho

67



Classification task

- Gene expression analysis
 - Functional genomics
- Tissue diagnosis
 - Expression levels of thousands of genes collected from different tissues
 - Several ML algorithms were used
- First use of meta-learning for algorithm recommendation in a single domain

Andre Ponce de Leon de Carvalho

68



Experiments

- 60 cancer related datasets
 - Mainly disease diagnostics related
 - Large number of attributes
 - Normalized to mean 0, variance 1
- Different data characterization approaches
 - Statlog
 - Clustering based

Andre Ponce de Leon de Carvalho

69



Experiments

- Base algorithms
 - Traditional ML algorithms
 - Easy availability (e.g. in R)
 - Moderate computational burden
 - SVM Linear, SVM RBF, DLDA, DQDA, PAM, 3-NN, RF
 - Default parameters
 - Predictive performance estimation
 - Bootstrap with 50 runs

Andre Ponce de Leon de Carvalho

70



Experiments

- Default ranking as baseline
 - Average of rankings in the training meta-examples
- Evaluation of metamodel
 - Leave-one-out (**LOO**)
 - Correlation metrics for ranking accuracy:
 - Spearman's rank correlation (RS)
 - Log Ranking Accuracy (LRA)

Andre Ponce de Leon de Carvalho

71



StatLog data characterization

- Subgroup of 10 measures
 - Adequate to predictive attributes with numerical values
 - Log of number of examples
 - Log of number of attributes
 - Log of number of classes
 - Mean kurtosis
 - First canonical correlation
 - Normalized class entropy
 - ...

Andre Ponce de Leon de Carvalho

72



Validity data characterization

- 10 measures from clustering validation
 - Degree each object belongs to a cluster
 - Cluster compactness
 - Cluster separation
 - Dispersion in the clusters
 - Cluster of neighbour objects
 - Data distribution of objects in two clusters
 - ...

Andre Ponce de Leon de Carvalho

73



Data characterization

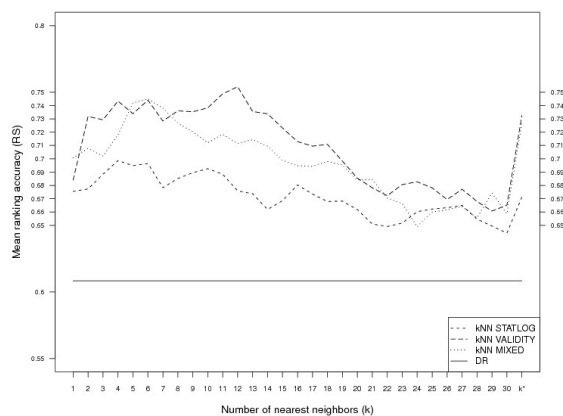
- Extraction of data characterization measures from microarray data
 - high dimensionality \Rightarrow high computational cost
 - Partial least squares (PLS) regression was applied to each dataset
 - Preserve main characteristics of the dataset

Andre Ponce de Leon de Carvalho

74



kNN performance



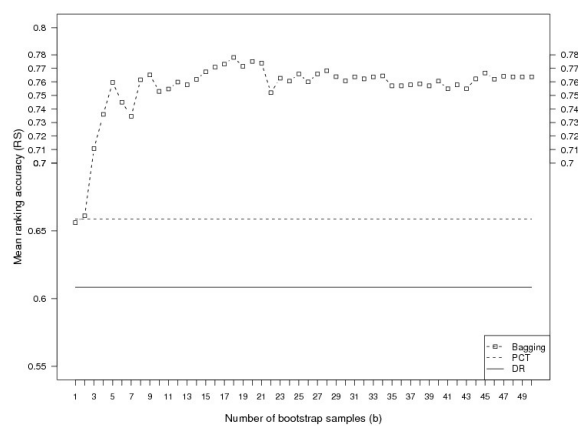
DR: default ranking

Andre Ponce de Leon de Carvalho

75



Ranking methods - mixed



Andre Ponce de Leon de Carvalho

76



Data streams main features

- Data arrive sequentially and, usually, at high speed
- There is no control on the arrival order
 - Or size of intervals between arrivals
- Stream usually have unlimited size
- Data distribution may change along the time
- Arriving objects are usually unlabelled



Andre Ponce de Leon de Carvalho

77



Data stream mining requirements

- Data must be accessed only once
 - Data cannot be stored in memory
 - After processed, must be discarded
- Decision model must be able to
 - Be continuously and fast updated
 - Deal with unlabelled data
 - Detect novelties

Andre Ponce de Leon de Carvalho

78



Regression task

- Data streams
 - Many real-world systems generate data continuously
 - Computer networks
 - Public transport traffic
 - Electricity monitoring
 - Underlying distribution that generate these data can change along the time
 - Good model now may not be a good model later
- Dynamically select regression algorithm

Andre Ponce de Leon de Carvalho

79



Travel time prediction (TPP)

- Used for travel planning of bus companies
 - Three days ahead is the limit to re-define buses and drivers
 - Better predictions can reduce operational costs and improve operational planning
- Benefits:
 - Definition of crew's duties
 - Real time travel adjustments
 - Reducion of travel delays

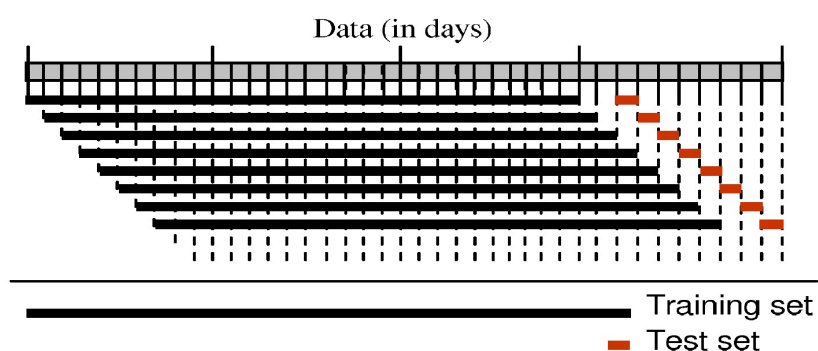
Andre Ponce de Leon de Carvalho

80



Travel time prediction

- Predicting travel time three days ahead
 - To re-define buses and drivers

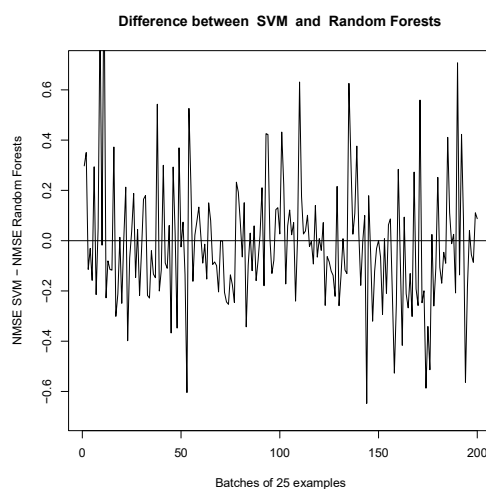


Andre Ponce de Leon de Carvalho

81



Why periodic selection



Andre Ponce de Leon de Carvalho

82



MetaStream

- Metalearning for data streams
 - Periodic algorithm selection in time-changing environments
 - Baselevel:
 - Different learning algorithms induce regression models using incoming data
 - Metalevel:
 - Metalearner relates characteristics extracted from the data to the predictive performance of the regression algorithms

Andre Ponce de Leon de Carvalho

83



Metafeatures

- 24 metafeatures extracted from a sliding window (100 examples from 12000)
 - Interquartile range
 - Skewness
 - Kurtosis
 - Coefficient of variation
 - Ratio of turning points
 - Correlation
 - Dispersion gain

Andre Ponce de Leon de Carvalho

84



Experiments

- Metalevel: algorithm selection
 - MetaStream
 - Default: majority class in the training data
 - Regression algorithm that won more often
- Baselevel: average error of the regressors
 - Selected regressor
 - MetaStream
 - Default
 - Ensemble: averages predictions made by different regressors

Andre Ponce de Leon de Carvalho

85



Experiments

- Baselevel
 - Classification and Regression Trees (CART)
 - Linear regression (LR)
 - Multivariate adaptive regression splines (MARS)
 - Project Pursuit Regression (PPR)
 - Random Forests (RF)
 - Linear regression (LR)
 - Support Vector Machines (SVM)
- Default parameter values from R packages

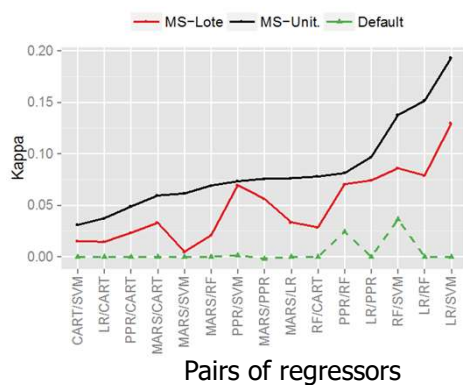
Metalevel

Andre Ponce de Leon de Carvalho

86



Results for RF metalevel



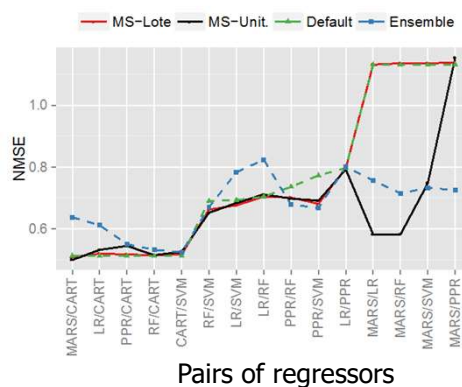
The higher
the better

Andre Ponce de Leon de Carvalho

87



Results for RF baselevel



The lower
the better

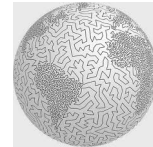
Andre Ponce de Leon de Carvalho

88



TSP

- Salesman needs to visit a group of locations (cities)
 - One location each time
 - Travelling the smallest overall distance
 - Returns to the first city
- Apparently simple
 - NP-complete
 - One of the most studied problems in optimization and artificial intelligence



Andre Ponce de Leon de Carvalho

89



TSP

- Theoretical and experimental studies
- Several variations
- Many techniques have been used/proposed
 - Mainly meta-heuristics (MHs)
- Which technique to choose for a given instance of a TSP?
 - Meta-learning

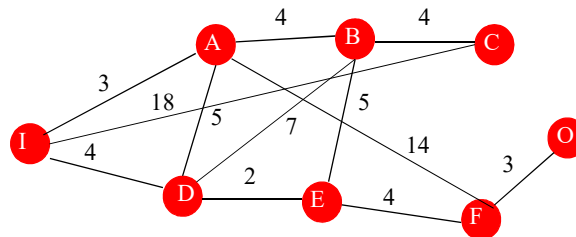
Andre Ponce de Leon de Carvalho

90



Computational solution

- TSP can be represented by:
 - Graphs
 - Nodes: cities
 - Weighted edges: distances between cities



- Maps with latitude and longitude

Andre Ponce de Leon de Carvalho

91



Experiments

- Four TSP scenarios:
 - Strongly / weakly connected and asymmetric / symmetric graph
- 600 TSP instances for each scenario
- Five MHs (TS, GRASP, SA, GA and ACO)
 - Each MH is run 30 times for each instance
 - MH performance = average/mode solution
- Four sets of metafeatures

Andre Ponce de Leon de Carvalho

92



Experiments

- Evaluation of different metafeatures
 - Edge and vertex measures (EVM)
 - 14 meta-features previously (Kanda et al,2011)
 - Meta-heuristic properties (MHP)
 - Subsampling landmarker properties (SLP)
 - Complex network measures (CNM)

Andre Ponce de Leon de Carvalho

93



Experiments

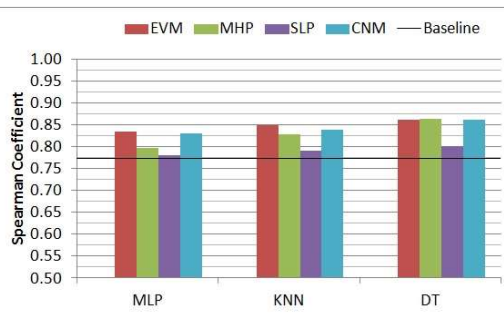
- Three ML algorithms for label ranking:
 - MLP, K-NN and DT
- Performance measured by accuracy ranking
 - Spearman Coefficient
- Baseline model = average ranking prediction for all examples

Andre Ponce de Leon de Carvalho

94



Weakly connected asymmetric



- Lack of some edges affected performance of the MHs
- Worst three performances are not statistically different from baseline
- SLP meta-features had poor results
 - MHs not able to find a feasible solution fast

Andre Ponce de Leon de Carvalho

95



Paje (ICMC-USP)

- End-to-end Machine learning
- Main focus
 - Data pre-processing
 - Non-structured data
 - Data quality
 - Dimensionality reduction
 - Explainable ML
 - Post-processing
 - Easily expandable

Andre Ponce de Leon de Carvalho

96



Research at Analytics USP

- Goal: End-to-End AutoML
 - Single-domain and multi-domain
 - Pre-processing
 - Modelling
 - Post-processing
 - Meta-learning + optimization
 - Pajé tool

Andre Ponce de Leon de Carvalho

97



Research at Analytics USP

- Design of new meta-features able to better describe datasets
- Pre-processing
 - Noise detection
 - Feature selection
 - Missing values imputation
 - Imbalanced data
 - Multilabel data

Andre Ponce de Leon de Carvalho

98



Research at Analytics USP

- Modelling
 - ML algorithm recommendation
 - Classification, regression and clustering
 - Active learning strategy
 - Ensemble
 - Data streams
 - Hyperparameter tuning
 - Recommender systems
 - Metalearning with and for optimization

Andre Ponce de Leon de Carvalho

99



Responsible Data Science

- Accountability
 - Who is in charge?
- Reproducibility
 - Data, code and experimental choices must be publicly available
- Privacy
 - With information of 300 likes, ML can predict someone personality better than her/his partner



André de Carvalho - ICMC/USP

100



Responsible Data Science

- Transparency
 - Right to explanation
 - General Data Protection Regulation (GDPR-EU)
 - Explainable AI (XAI)
- Fairness
 - Avoid decisions can be based on sensitive features
 - E.g. Citizenship, Gender, Race
 - Fair Information Practices

André de Carvalho - ICMC/USP

101



Reproducibility

- ML researchers compare new algorithms to existing alternatives
 - However, code of related alternatives are not often available
 - Reproducibility crisis
 - Medicine and Psychology went through a similar crises in the last decade

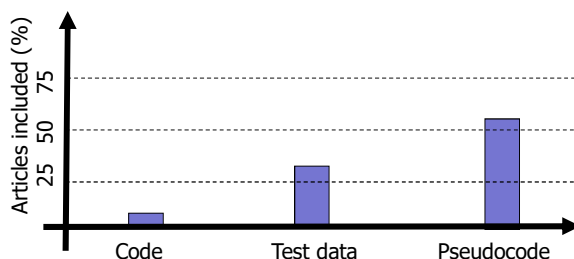
André de Carvalho - ICMC/USP

102



Reproducibility

- AI researchers do not share their code
 - Survey with 400 algorithms propose in the 2 main AI conferences



Fonte: <http://www.sciencemag.org/news/2018/02/missing-data-hinder-replication-artificial-intelligence-studiesF>

André de Carvalho - ICMC/USP

103



Reproducibility

- Main reasons for not sharing
 - Code is not finished
 - Code belongs to the company
 - Code depends on another code, not yet published
 - To keep ahead of competitors
 - Code lost because of computer is broken or was stolen
 - "My dog ate my program" reason
 - Nicolas Rougier, INRIA, France

André de Carvalho - ICMC/USP

104



Privacy protection

- Fair Information Practices (FIPs) for data
- Set of 10 principles for:
 - Collecting
 - Accessing
 - Sharing
 - Using

André de Carvalho - ICMC/USP

105



Transparency

- Often essential for the application and acceptance of a Data Science solution
- Most of the time brings benefits, but
 - People can inappropriately use the information obtained
 - To gain advantages
 - To cause harm
- GDPR Right to explanation

André de Carvalho - ICMC/USP

106



Brazilian Data Protection Law

- Based on EU GDPR
- Protect, but differentiate Personal and sensitive data
 - Personal data:
 - Can identify the person (name, photo, national identification, biometry,...)
 - Sensitive personal data:
 - Can lead to biased decisions (Race, gender, political preferences, religion, health, genetic, biometry, ...)
- National data protection agency

Andre Ponce de Leon de Carvalho

107



Fairness

- Decisions take by ML models can seriously affect individuals
 - Some decisions can be based on sensitive features
 - Citizenship
 - Gender
 - Race
 - Decisions based on sensitive features can lead to illegal or unfair discrimination of subgroups
 - How to deal with these features?

André de Carvalho - ICMC/USP

108



Ethical AutoML

- Fairness
 - Pre- and post-processing have a key role in ensure fairness in AutoML
- Transparency
 - AutoML must favour explainable algorithms
- Privacy
 - Importance of an explainable AutoML
- Accountability
- Reproducibility

Andre Ponce de Leon de Carvalho

109



Data Science for good

- University of Chicago summer program
- Non-profit movement
 - Bring social and economical benefits to people and communities
 - Some programs are funded by companies
- Adopted by other institutions
- Contribution for a fair society



André de Carvalho - ICMC/USP

110



Support

AMDA
Machine Learning
in Data Analysis

Algar
Telecom

CNPq
Conselho Nacional de Desenvolvimento
Científico e Tecnológico

CEPID
Centro de Pesquisa,
Inovação e Difusão

FAPESP

Intel

ICMC USP
SÃO CARLOS

Analytics
Machine Data Analysis Laboratory

CAPES



Questions?



Andre Ponce de Leon de Carvalho

113