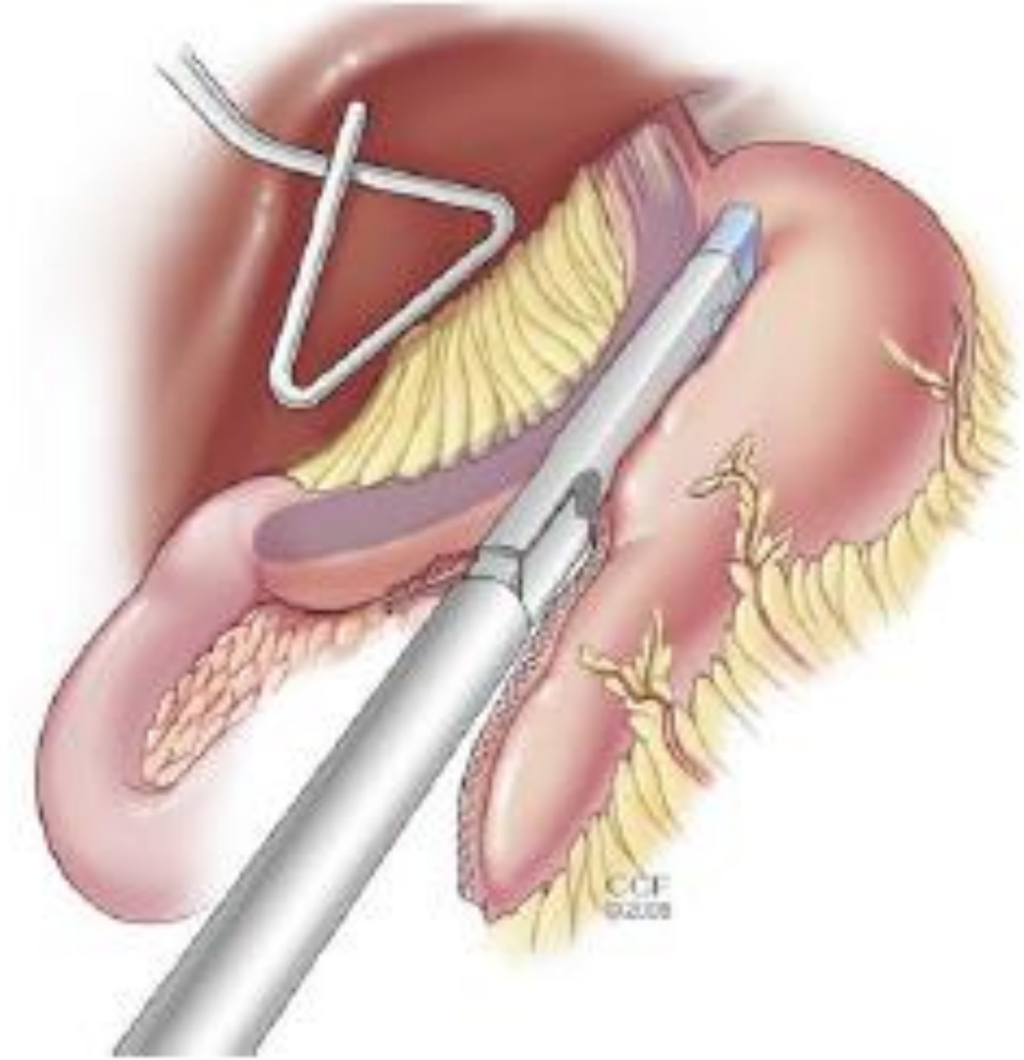
Technische Universiteit Eindhoven University of Technology

TPOT's performance for Biomedical Data

Tim Beishuizen

Supervisor: Joaquin Vanschoren



Introduction

- Computational Biology
    - Data analysis framework

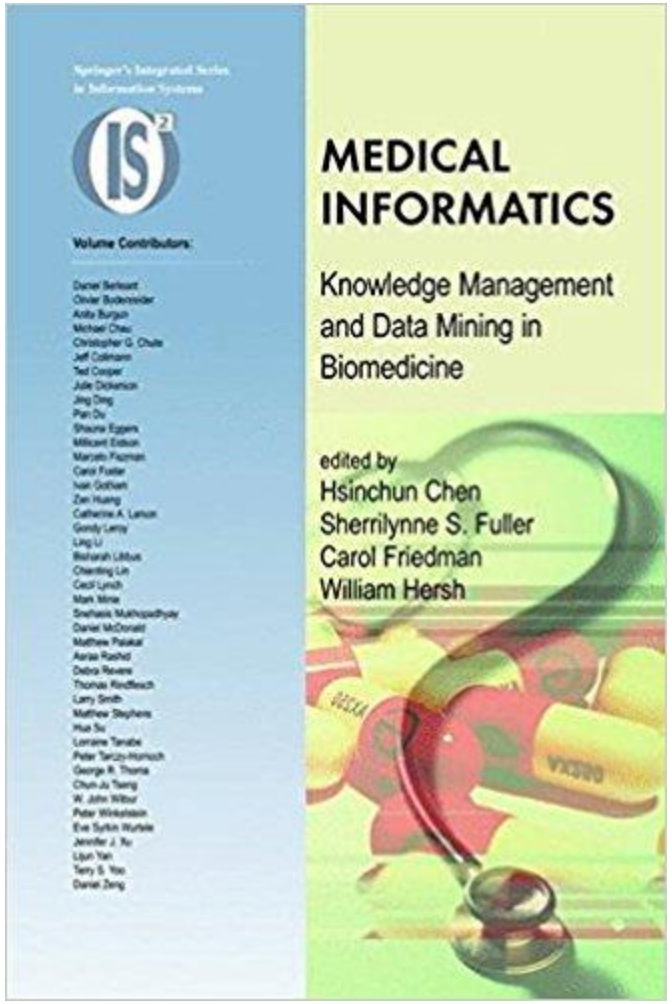
Content

- Biomedical Data
    - Bariatric dataset

- Automated Machine learning
    - TPOT

Biomedical Data

- Challenges
  - Volume
  - Dimensionality
  - Complexity
  - Heterogeneity
  - Quality



Chen, Hsinchun, et al., eds. *Medical informatics: knowledge management and data mining in biomedicine*. Vol. 8. Springer Science & Business Media, 2006.

Bariatric Data set

- Bariatric surgery
  - Catharina Hospital

- Two data sets
  - DATO set
    - Co-morbidity presence
    - Basic health variables
  - Lab set
    - Biomarkers

- Challenges
  - Combining two data sets
  - Missing values
  - High error rate (5%)

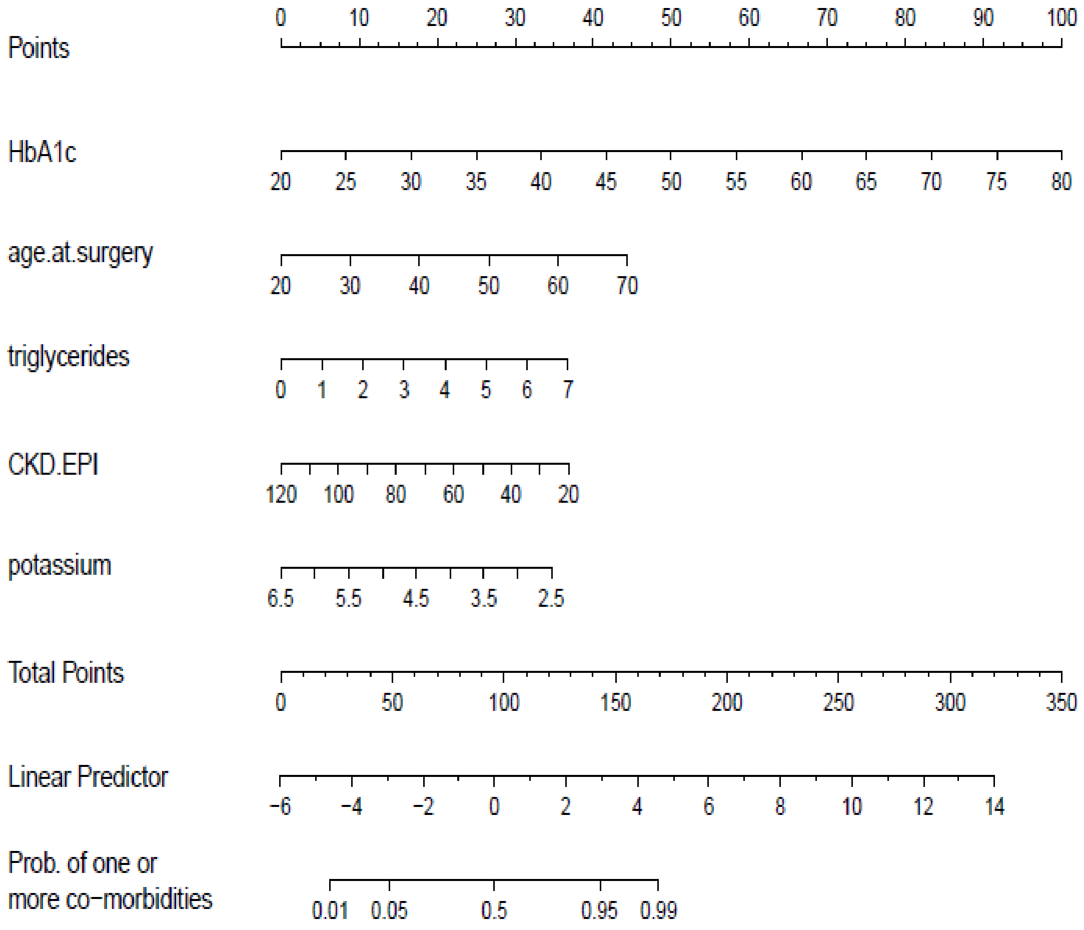
Bariatric Data set

- Previous project

- Challenges
  - Dropped non matching data
  - Dropped missing values
  - Ignored missing data

- Logistic regression

- Model nomogram

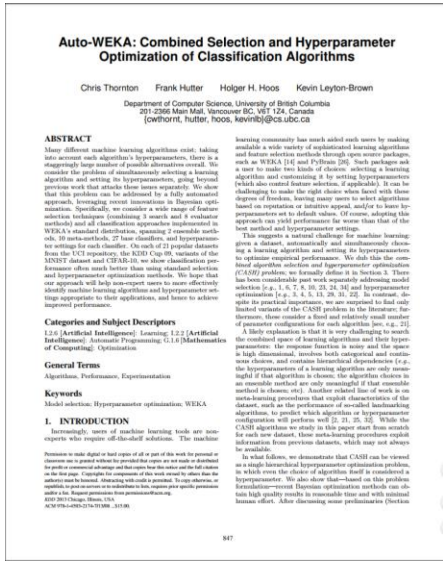


Automated Machine Learning

- Machine learning pipelines
  - Algorithm selection
  - Hyperparameter Optimization
  - Preprocessing

- Combined Algorithm Selection and Hyperparameter Optimization (CASH)
  - Find the best automatically
  - Meta-learning approach
  - Hyperparameter optimization
  - Preprocessing

- Auto-WEKA



Auto-WEKA: Combined Selection and Hyperparameter Optimization of Classification Algorithms

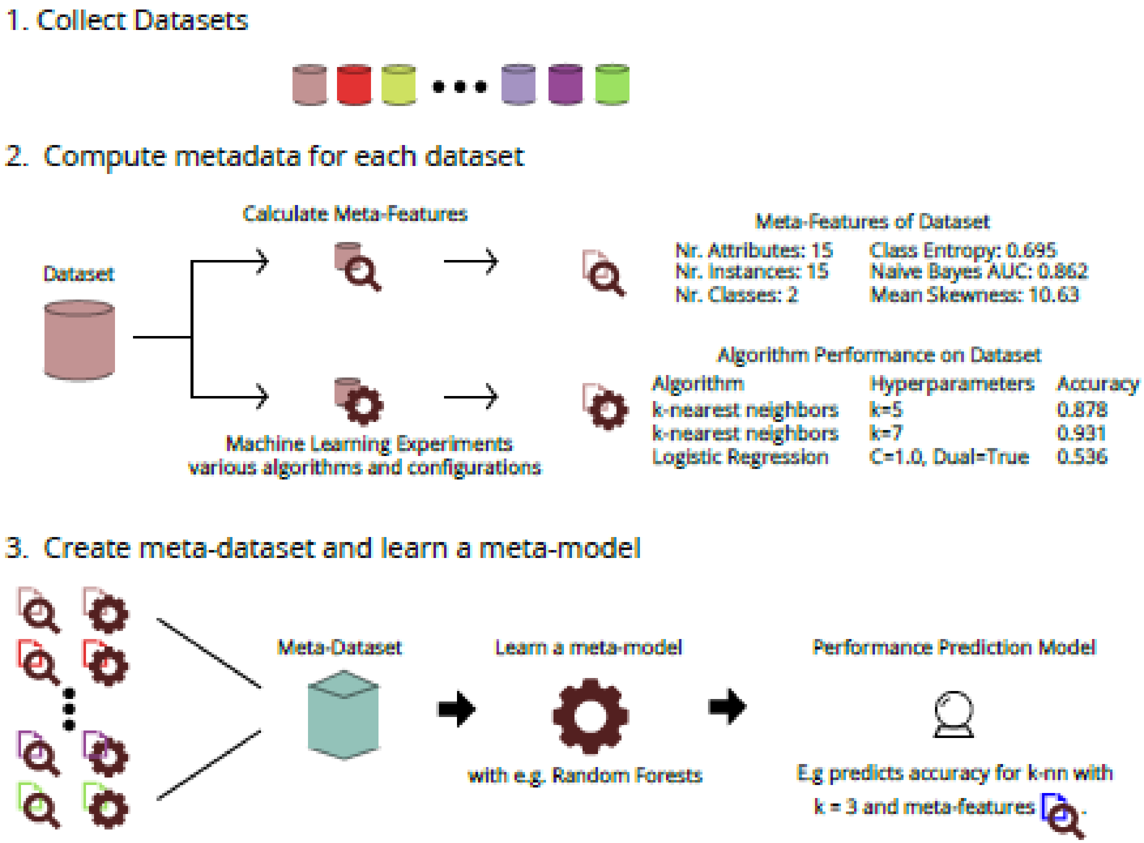
Chris Thornton Frank Hutter Holger H. Hoos Kevin Leyton-Brown

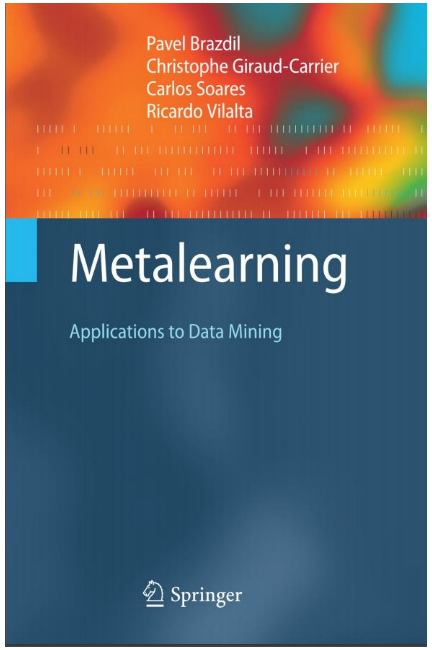
Department of Computer Science, University of British Columbia
201-2366 Main Mall, Vancouver BC, V6T 1Z4, Canada
{cwthornt, hutter, hoos, kevinlb}@cs.ubc.ca

Thornton, Chris, et al. "Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms." *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013.

Automated Machine Learning

- Meta-learning
  - Find performance changes
  - Time and space constraints

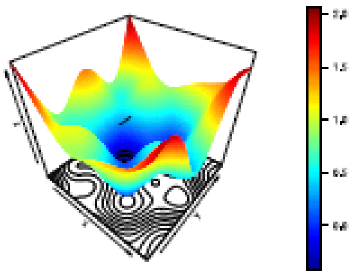




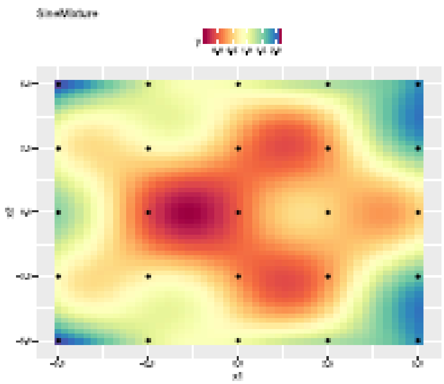
Brazdil, Pavel, et al. "Development of metalearning systems for algorithm recommendation." *Metalearning: Applications to Data Mining* (2009): 31-59.

Automated Machine Learning

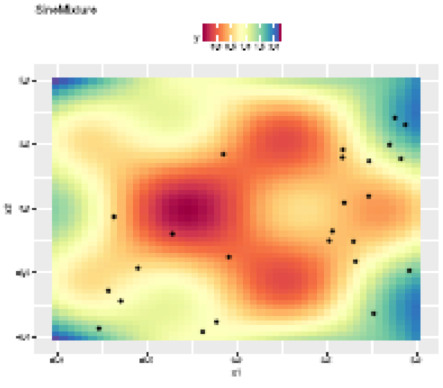
- Hyperparameter optimization
    - Grid search
    - Random search
    - Bayesian optimization



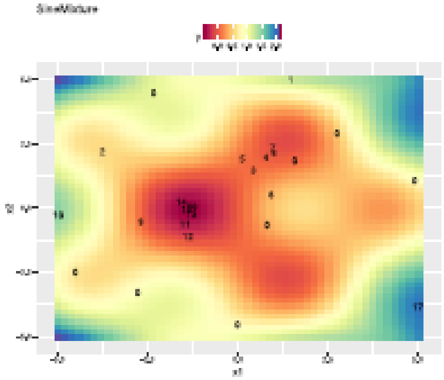
(a) Function to optimize: $f(x, y) = sin(x) * 0.5 * cos(y) + 0.04 * (x^2 + y^2)$.



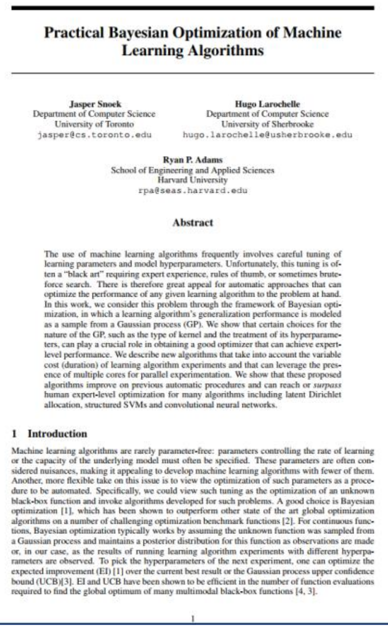
(b) Example sample points of grid search.



(c) Example sample points of random search.



(d) Example sample points of search with Bayesian optimization.



Practical Bayesian Optimization of Machine Learning Algorithms

Jasper Snoek
Department of Computer Science
University of Toronto
jasper@cs.toronto.edu

Hugo Larochelle
Department of Computer Science
University of Sherbrooke
hugo.larochelle@usherbrooke.edu

Ryan P. Adams
School of Engineering and Applied Sciences
Harvard University
rpa@seas.harvard.edu

Abstract

The use of machine learning algorithms frequently involves careful tuning of learning parameters and model hyperparameters. Unfortunately, this tuning is often a "black art" requiring expert experience, rules of thumb, or sometimes brute-force search. There is therefore great appeal for automatic approaches that can optimize the performance of any given learning algorithm to the problem at hand. In this work, we consider this problem through the framework of Bayesian optimization, in which a learning algorithm's generalization performance is modeled as a sample from a Gaussian process (GP). We show that certain choices for the nature of the GP, such as the type of kernel and the treatment of its hyperparameters, can play a crucial role in obtaining a good optimizer that can achieve expert-level performance. We describe new algorithms that take into account the variable cost (duration) of learning algorithm experiments and that can leverage the presence of multiple cores for parallel experimentation. We show that these proposed algorithms improve on previous automatic procedures and can reach or *surpass* human expert-level optimization for many algorithms including latent Dirichlet allocation, structured SVMs and convolutional neural networks.

1 Introduction

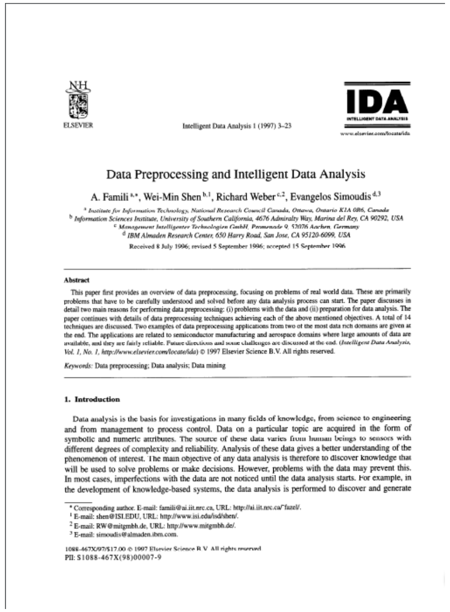
Machine learning algorithms are rarely parameter-free: parameters controlling the rate of learning or the capacity of the underlying model must often be specified. These parameters are often considered nuisances, making it appealing to develop machine learning algorithms with fewer of them. Another, more flexible take on this issue is to view the optimization of such parameters as a procedure to be automated. Specifically, we could view such tuning as the optimization of an unknown black-box function and invoke algorithms developed for such problems. A good choice is Bayesian optimization [1], which has been shown to outperform other state of the art global optimization algorithms on a number of challenging optimization benchmark functions [2]. For continuous functions, Bayesian optimization typically works by assuming the unknown function was sampled from a Gaussian process and maintains a posterior distribution for this function as observations are made or, in our case, as the results of running learning algorithm experiments with different hyperparameters are observed. To pick the hyperparameters of the next experiment, one can optimize the expected improvement (EI) [1] over the current best result or the Gaussian process upper confidence bound (UCB)[3]. EI and UCB have been shown to be efficient in the number of function evaluations required to find the global optimum of many multimodal black-box functions [4, 3].

1

Snoek, Jasper, Hugo Larochelle, and Ryan P. Adams. "Practical bayesian optimization of machine learning algorithms." *Advances in neural information processing systems*. 2012.

Automated Machine Learning

- Preprocessing
  - Problems
  - Techniques



ELSEVIER

Intelligent Data Analysis 1 (1997) 3–23

IDA
INTELLIGENT DATA ANALYSIS
www.elsevier.com/locate/ida

Data Preprocessing and Intelligent Data Analysis

A. Famili [a,*], Wei-Min Shen [b,1], Richard Weber [c,2], Evangelos Simoudis [d,3]

[a] Institute for Information Technology, National Research Council Canada, Ottawa, Ontario K1A 0R6, Canada
[b] Information Sciences Institute, University of Southern California, 4676 Admiralty Way, Marina del Rey, CA 90292, USA
[c] Management Intelligenter Technologien GmbH, Promenade 9, 52076 Aachen, Germany
[d] IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95120-6099, USA

Received 8 July 1996; revised 5 September 1996; accepted 15 September 1996

Abstract

This paper first provides an overview of data preprocessing, focusing on problems of real world data. These are primarily problems that have to be carefully understood and solved before any data analysis process can start. The paper discusses in detail two main reasons for performing data preprocessing: (i) problems with the data and (ii) preparation for data analysis. The paper continues with details of data preprocessing techniques achieving each of the above mentioned objectives. A total of 14 techniques are discussed. Two examples of data preprocessing applications from two of the most data rich domains are given at the end. The applications are related to semiconductor manufacturing and aerospace domains where large amounts of data are available, and they are fairly reliable. Future directions and some challenges are discussed at the end. (*Intelligent Data Analysis, Vol. 1, No. 1, http://www.elsevier.com/locate/ida*) © 1997 Elsevier Science B.V. All rights reserved.

Keywords: Data preprocessing; Data analysis; Data mining

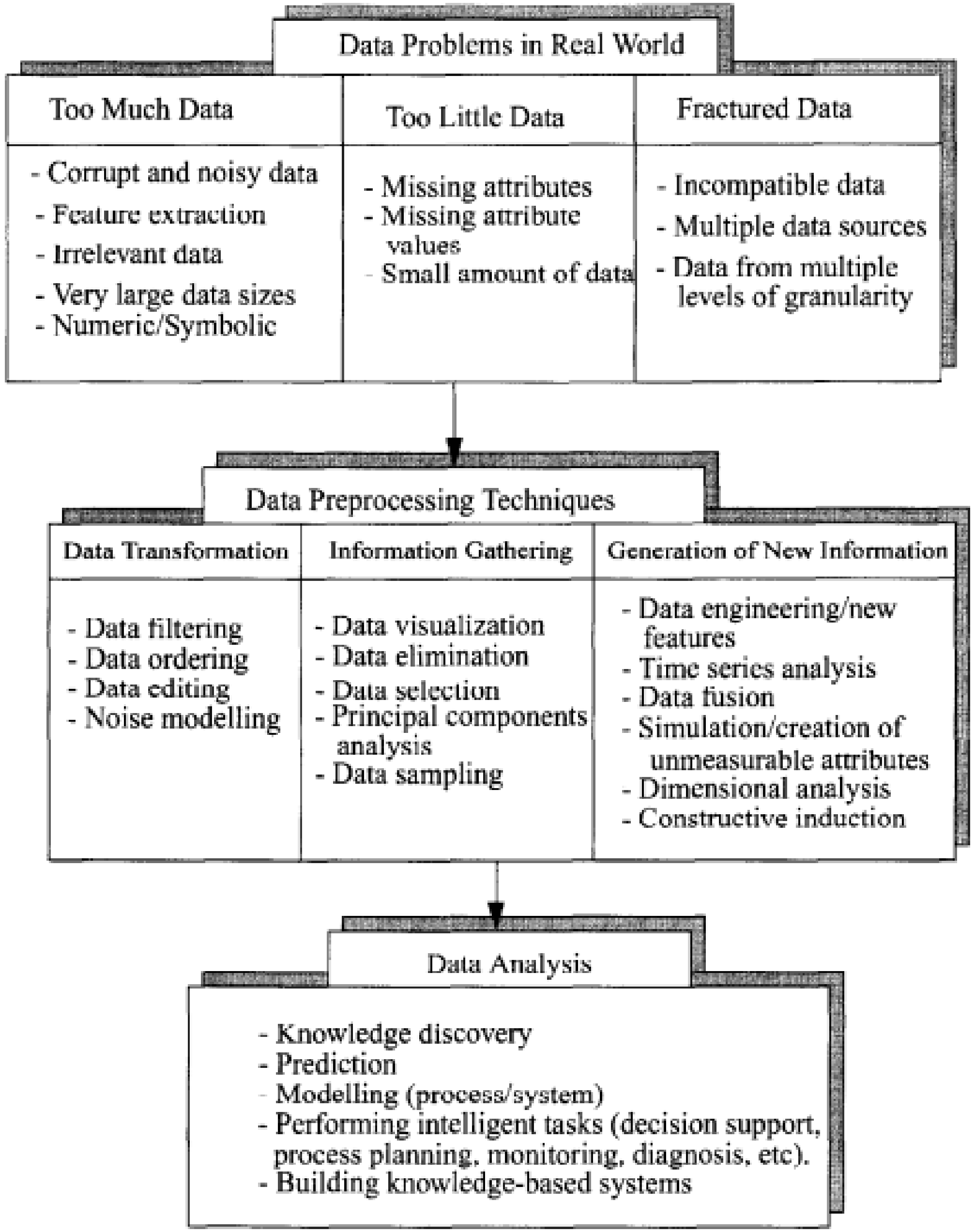
1. Introduction

Data analysis is the basis for investigations in many fields of knowledge, from science to engineering and from management to process control. Data on a particular topic are acquired in the form of symbolic and numeric attributes. The source of these data varies from human beings to sensors with different degrees of complexity and reliability. Analysis of these data gives a better understanding of the phenomenon of interest. The main objective of any data analysis is therefore to discover knowledge that will be used to solve problems or make decisions. However, problems with the data may prevent this. In most cases, imperfections with the data are not noticed until the data analysis starts. For example, in the development of knowledge-based systems, the data analysis is performed to discover and generate

* Corresponding author. E-mail: famili@ai.iit.nrc.ca, URL: http://ai.iit.nrc.ca/~fazel/.
[1] E-mail: shen@ISI.EDU, URL: http://www.isi.edu/isd/shen/.
[2] E-mail: RW@mitgmbh.de, URL: http://www.mitgmbh.de/.
[3] E-mail: simoudis@almaden.ibm.com.

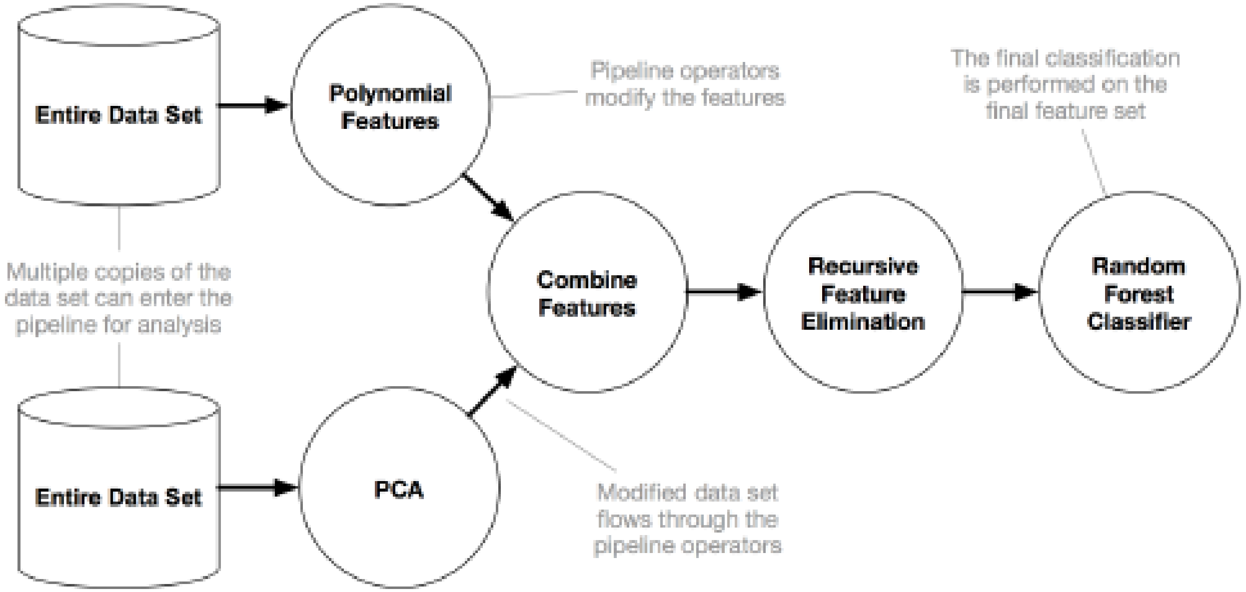
PII: S1088-467X(98)00007-9

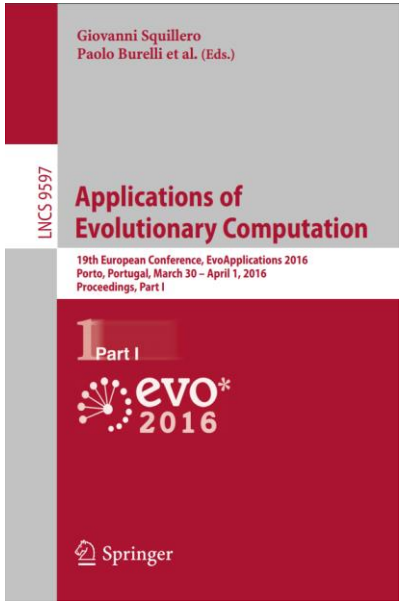
Famili, A., et al. "Data preprocessing and intelligent data analysis." *Intelligent data analysis* 1.1-4 (1997): 3-23.



TPOT

- Tree-based Pipeline Optimization Tool
  - Automated Machine Learning tool
  - Scikit-learn
  - Evaluation, selection and mutations





Olson, Randal S., et al. "Automating biomedical data science through tree-based pipeline optimization." *European Conference on the Applications of Evolutionary Computation*. Springer, Cham, 2016.

Research Question

- How does TPOT perform on specific biomedical data set problems and how can it be improved?
- Hypothesis
  - Combining data sets
  - Erroneous values
  - Missing values
- Improvement on missing values
  - Extrapolation
  - Nearest Neighbours