Feature Selection Toolbox 3

version 3.1

FST3 web portal provides software, references, documents and links to all kinds of feature selection resources

Reference:
Introduction to FST 3 – The C++ Library for Subset Search, Data Modeling and Classification, UTIA Tech. Report No. 2287 (available at FST3 web)

Provided by:
Dept. of Pattern Recognition
UTIA, Academy of Sciences
Pod vodárenskou věží 4
Prague 8, 18208
Czech Republic

E-mail:
xaos@utia.cas.cz

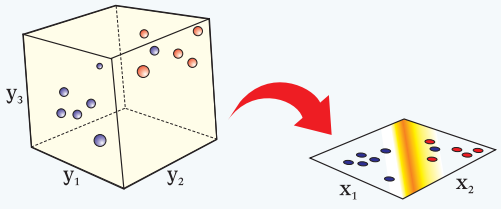
The FST project has been led by Petr Somol, Pavel Pudil and Jana Novovičová

Feature selection is applicable in a variety of fields, including:

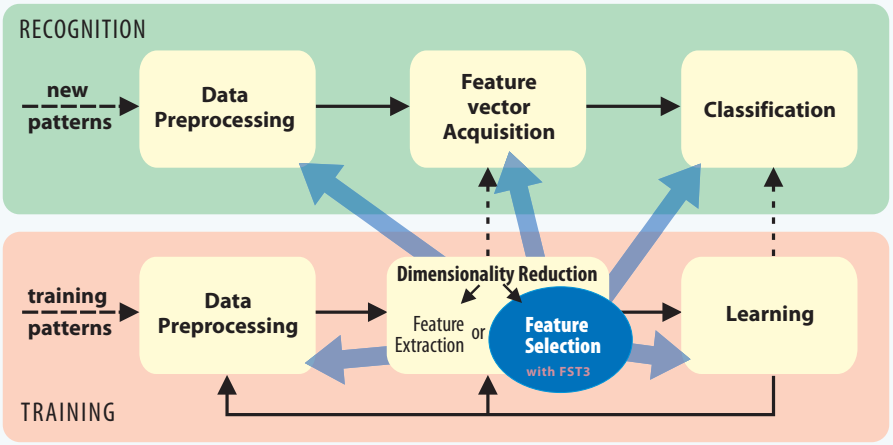
- ***medicine (diagnostic systems, gene search, etc.)***
- ***finance (evaluating trends, credit scoring, etc.)***
- ***governmental planning (analysis of remote sensing data, etc.)***
- ***text processing (keyword extraction, document categorization, etc.)***
- ***security (face or fingerprint recognition, etc.)***
- ***military (target spotting, etc.)***
- ***industry (defect detection, etc.)***

Feature Selection Toolbox 3 (FST3) is a standalone C++ library for feature subset search, freely available at

http://fst.utia.cz



Feature Selection (also known as attribute or variable selection) is capable of reducing problem dimensionality to maximize the accuracy of data models, performance of automatic decision rules as well as to reduce data acquisition cost.



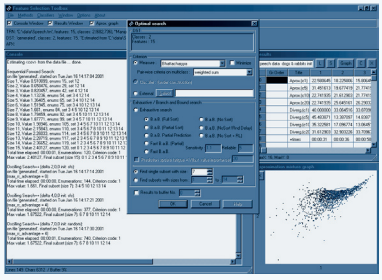
Feature Selection Toolbox 3 software library provides a selection of advanced tools focused primarily on solving the feature selection form of the dimensionality reduction problem, yet it addresses and interacts with all other stages of the machine learning and recognition process.

FST3 key functionality:

- (Threaded) highly effective subset search methods to tackle computational complexity
- Wrappers, filters & hybrid methods, deterministic and/or randomized
- Specialized methods for very-high-dimensional feature selection
- Anti-overfitting measures: criteria ensembles, result regularization, etc.
- Stability, similarity and bias evaluation

In more detail:

- **feature selection criteria**
  - **wrapper** – classifier accuracy estimation
    - normal Bayes classifier
    - k-Nearest Neighbor (various L distances)
    - Support Vector Machine (using external LibSVM library)
  - **filter** – normal model based
    - Bhattacharyya distance
    - Divergence
    - Generalized Mahalanobis distance
  - **filter** – multinomial model based
    - Bhattacharyya distance
    - Mutual Information
  - **criteria ensembles**
  - **hybrids**
- **feature selection methods**
  - **BIF, best individual features** (individual ranking)
  - **DAF, dependency-aware feature ranking**
  - **sequential search (sub-optimal)**
    - (G)SFS/SBS, sequential selection
    - (G)SFFS/SBFS, floating search
    - (G)OS, oscillating search
    - (G)DOS, dynamic oscillating search
    - (G)SFRS/SBRS, retreating search
  - **sequential search options:**
    - standard or generalized
    - threaded or sequential
    - *d*-parametrized or *d*-optimizing
    - deterministic or randomized
    - restricted or unrestricted, etc.
  - **Branch & Bound algorithms (optimal)**
    - BBB, Basic Branch & Bound
    - IBB, Improved Branch & Bound
    - BBPP, Branch & B. with Partial Prediction
    - FBB, Fast Branch & Bound
  - **exhaustive search (optimal)**
  - **Monte Carlo (pure random search)**
- **very-high-dimensional task specific methods**
  - BIF, best individual features
  - DAF, dependency-aware feature ranking
  - OS, oscillating search (w. low osc. depth)
- **supporting techniques**
  - subset size optimization
  - result regularization
  - feature acquisition cost minimization
  - feature sel. process stability evaluation
  - two-process similarity evaluation
  - classifier bias estimation
- **flexible data access and pre-processing**
  - **nested multi-level data splitting**
    - re-substitution · cross-validation
    - hold-out · leave-one-out
    - random sampling · etc.
  - **normalization/missing data substitution**
  - **ARFF (Weka) data format + TRN format**
- **customizable, templated, threaded C++ code**
- **free for non-commercial use**



FST1 as a tool less powerful but more suitable for quick experimenting and education is freely available as well.