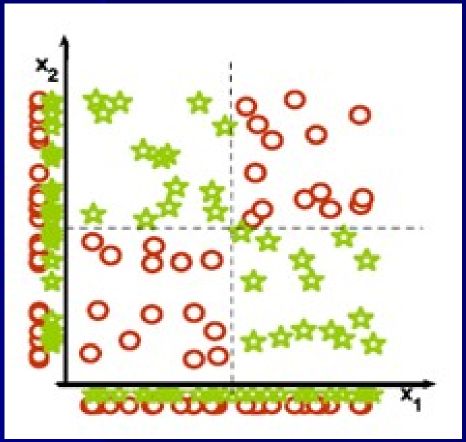
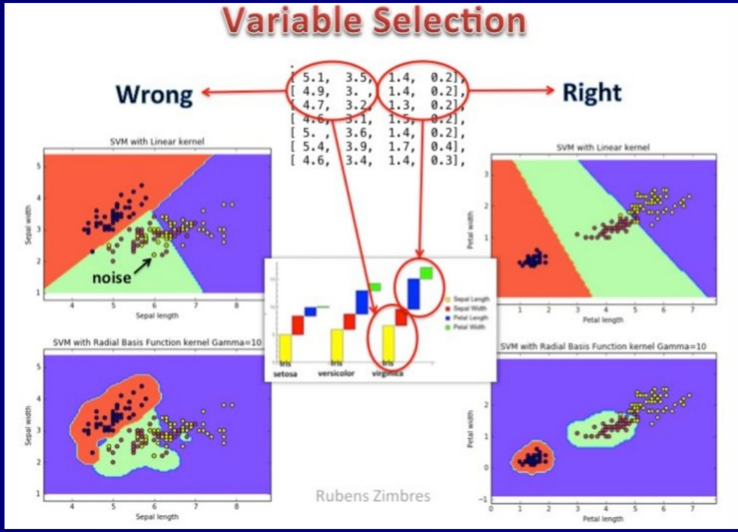
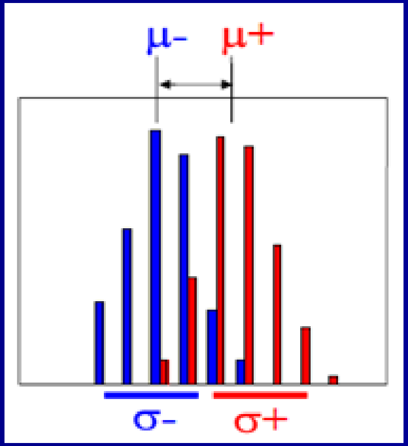
DIMENSIONALITY REDUCTION BY FEATURE SELECTION



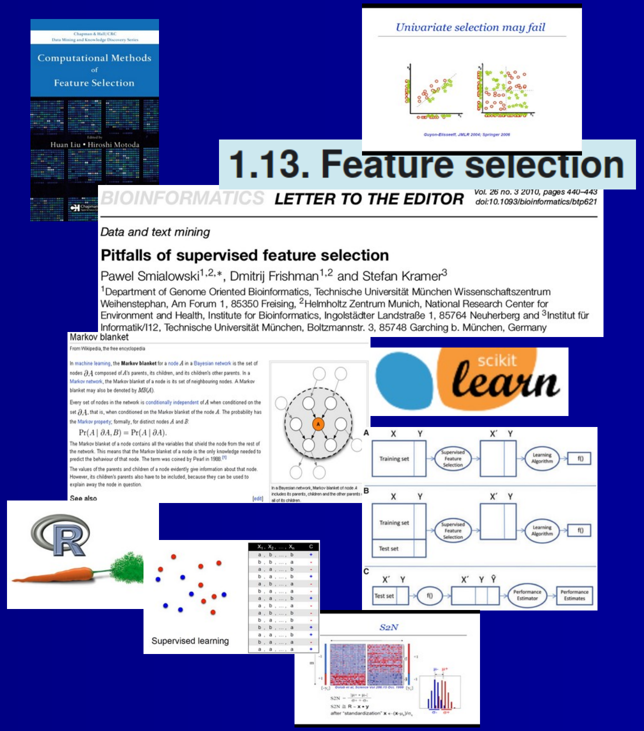




Iñaki Inza
Intelligent Systems Group, www.sc.ehu.es/isg
Computer Science Faculty
University of the Basque Country, Donostia - San Sebastian

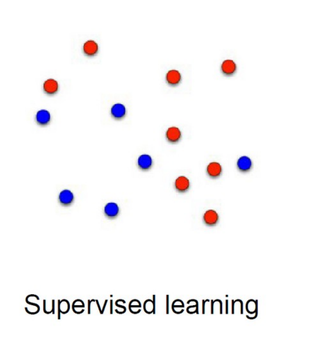
OUTLINE

- The context: FS in supervised classification
- FS versus feature extraction
- Types of techniques
- Final remarks and ideas
- References and software

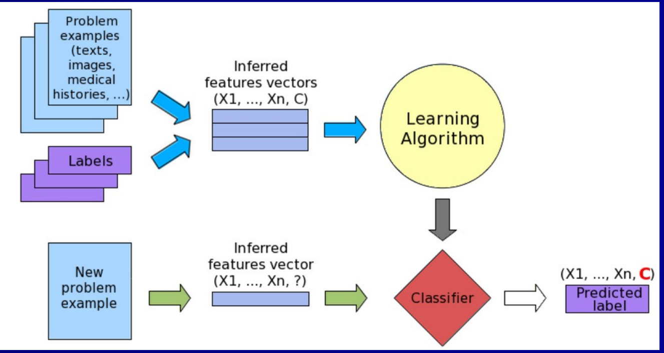


FEATURE SELECTION (FS) FOR SUPERVISED CLASSIFICATION

- Fix the learning scenario: supervised classification



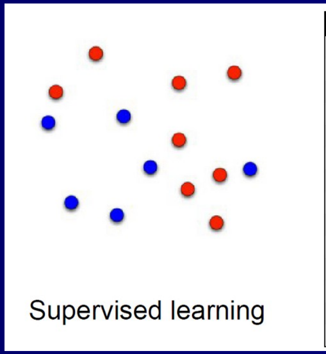
$X_1, X_2, \dots, X_n$	C
a , b , ... , b	+
b , b , ... , a	-
a , a , ... , b	-
b , a , ... , b	+
a , b , ... , a	-
b , a , ... , a	-
a , a , ... , b	+
a , b , ... , a	-
a , b , ... , b	-
b , a , ... , b	-
b , b , ... , a	+
a , a , ... , b	+
b , a , ... , a	-
a , a , ... , a	+



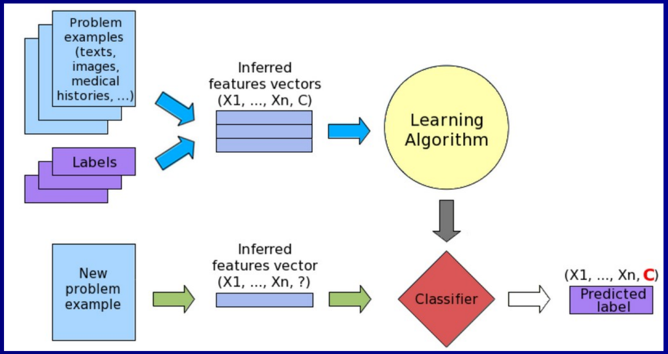
- Reduce the number of original features → $(X_1, X_2, \dots, X_d)$
- Irrelevancy?
- Redundancy?

FEATURE SELECTION (FS) FOR SUPERVISED CLASSIFICATION

- Fix the learning scenario: supervised classification



X_1 , X_2 , … , X_n	C
a , b , … , b	+
b , b , … , a	-
a , a , … , b	-
b , a , … , b	+
a , b , … , a	-
b , a , … , a	-
a , a , … , b	+
a , b , … , a	-
a , b , … , b	-
b , a , … , b	-
b , b , … , a	+
a , a , … , b	+
b , a , … , a	-
a , a , … , a	+



- Reduce the number of original features → $(X_1, X_2, \dots, X_d)$

- Improve accuracy
- Reduce costs
- Computational cost

FEATURE EXTRACTION

- Feature selection ≠≠≠≠ Feature construction-extraction
- Feature extraction → PCA, SVD, PLS...
- Mathematical properties
- Intuition - interpretation

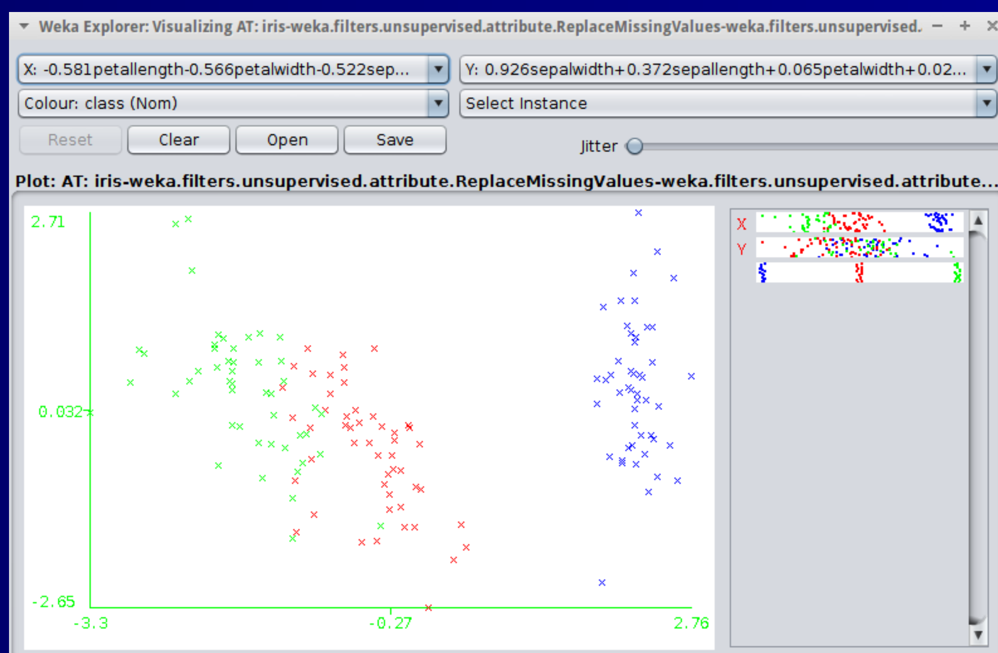
eigenvalue	proportion	cumulative	
2.91082	0.7277	0.7277	-0.581petallength-0.566petalwidth-0.522sepallength+0.263sepalwidth
0.92122	0.23031	0.95801	0.926sepalwidth+0.372sepallength+0.065petalwidth+0.021petallength

```
Ranked attributes:
 0.2723  1 -0.581petallength-0.566petalwidth-0.522sepallength+0.263sepalwidth
 0.042   2 0.926sepalwidth+0.372sepallength+0.065petalwidth+0.021petallength
```

FEATURE EXTRACTION
PRINCIPAL COMPONENT ANALYSIS

eigenvalue	proportion	cumulative	
2.91082	0.7277	0.7277	-0.581petallength-0.566petalwidth-0.522sepallength+0.263sepalwidth
0.92122	0.23031	0.95801	0.926sepalwidth+0.372sepallength+0.065petalwidth+0.021petallength

```
Ranked attributes:
 0.2723  1 -0.581petallength-0.566petalwidth-0.522sepallength+0.263sepalwidth
 0.042   2 0.926sepalwidth+0.372sepallength+0.065petalwidth+0.021petallength
```



A MATURE FIELD
APPLICATION DOMAINS

Artificial Intelligence Review (2021) 54:6149–6200
https://doi.org/10.1007/s10462-021-09970-6

Feature selection methods for text classification: a systematic literature review

Julliano Trindade Pintas[1] · Leandro A. F. Fernandes[1] ·
Ana Cristina Bicharra Garcia[2]

(2003) 1289-1305 Submitted 5/02; Published 3/03

**An Extensive Empirical Study of
Feature Selection Metrics for Text Classification**

George Forman GFORMAN@HPL.HP.COM
Hewlett-Packard Labs
Palo Alto, CA, USA 94304

**Analysis of Local Appearance-based Face Recognition:
Effects of Feature Selection and Feature Normalization**

Hazım Kemal Ekenel, Rainer Stiefelhagen
Interactive Systems Labs, Computer Science Department, Universität Karlsruhe (TH)
Am Fasanengarten 5, 76131, Karlsruhe, Germany
{ekenel,stiefel}@ira.uka.de

Journal of Bioinformatics and Computational Biology
Vol. 3, No. 2 (2005) 185–205
© Imperial College Press
Imperial College Press
www.icpress.co.uk

MINIMUM REDUNDANCY FEATURE SELECTION FROM
MICROARRAY GENE EXPRESSION DATA

CHRIS DING† and HANCHUAN PENG‡
†*Computational Research Division and* ‡*Life Sciences/Genomics Division*
Lawrence Berkeley National Laboratory, University of California
Berkeley, CA, 94720, USA

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Pattern Recognition Letters
ELSEVIER
Pattern Recognition Letters 25 (2004) 1377–1388
www.elsevier.com/locate/patrec

Feature selection in the independent component subspace
for face recognition

H.K. Ekenel *, B. Sankur
Department of Electrical and Electronic Engineering, Bogazici University, Bebek 34342, Istanbul, Turkey

BIOINFORMATICS REVIEW Vol. 23 no. 19 2007, pages 2507–2517
doi:10.1093/bioinformatics/btm344

Gene expression

A review of feature selection techniques in bioinformatics

Yvan Saeys[1,*], Iñaki Inza[2] and Pedro Larrañaga[2]
[1]Department of Plant Systems Biology, VIB, B-9052 Ghent, Belgium and Bioinformatics and Evolutionary Genomics group, Department of Molecular Genetics, Ghent University, B-9052 Ghent, Belgium and [2]Department of Computer Science and Artificial Intelligence, Computer Science Faculty, University of the Basque Country, Paseo Manuel de Lardizabal 1, 20018 Donostia - San Sebastián, Spain

TYPES OF FS TECHNIQUES

- Filter → selection independent of classifier
- Wrapper → ad-hoc features selected for a classifier

- Univariate filter

- Multivariate filter
- Multivariate wrapper

- Univariate → score individual features
- Multivariate → score feature subsets

UNIVARIATE FILTER

- ORDER - RANK features
- By their correlation with the class-label

- Choose top-k features → learn classifier

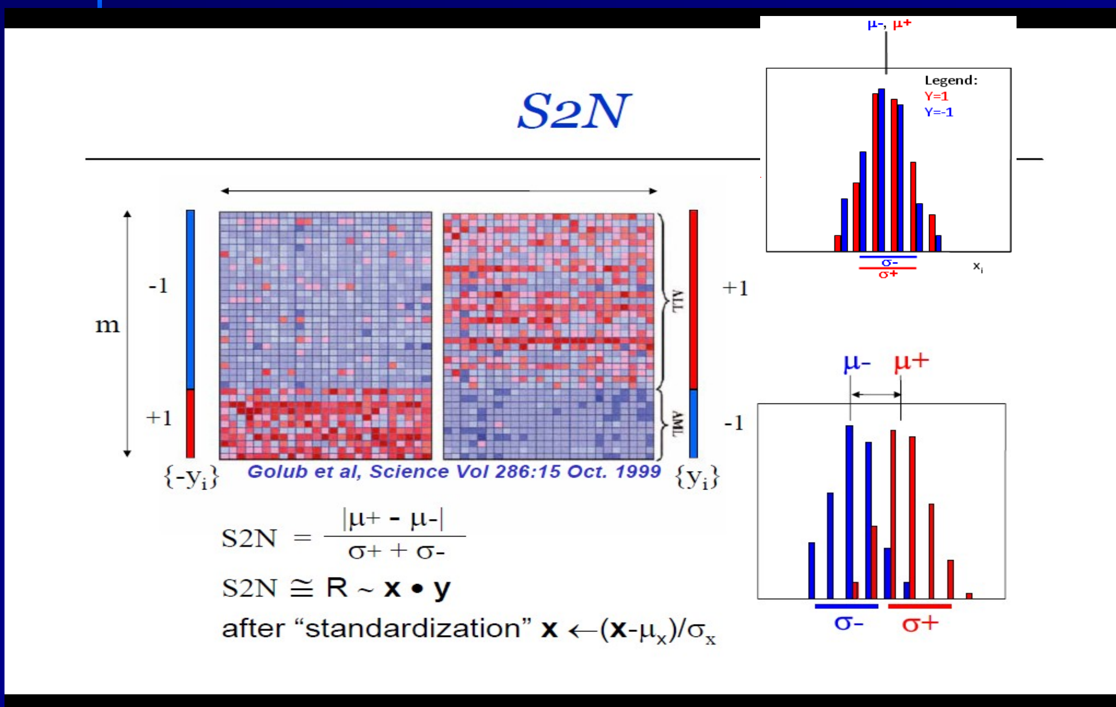
- Correlation measures
- Chi-square, G^2, t-test, Fisher Criterion Scoring, Entropy-based metrics (mutual info)...

- How to choose *k*?
- Correlations among selected features?
- Complementarity?

Ranked attributes:

0.39065	5	odor
0.25795	8	gill-size
0.23312	12	stalk-surface-above-ring
0.21818	20	spore-print-color
0.20716	19	ring-type
0.19644	4	bruises?
0.19433	13	stalk-surface-below-ring
0.15815	7	gill-spacing
0.1376	9	gill-color
0.13106	14	stalk-color-above-ring
0.12204	15	stalk-color-below-ring
0.12137	17	veil-color
0.10081	21	population
0.09141	18	ring-number
0.08182	6	gill-attachment
0.06895	22	habitat
0.02952	1	cap-shape
0.02848	11	stalk-root
0.01815	2	cap-surface
0.01436	3	cap-color
0.00762	10	stalk-shape
0	16	veil-type

UNIVARIATE FILTER CORRELATION by t-test



X_1	X_2	C
8	3	c_1
9	4	c_1
8	5	c_1
9	6	c_1
14	3	c_2
15	4	c_2
14	6	c_2
15	6	c_2

$$\text{t-test}(X_1, C) = \text{S2N}(X_1, C) = \frac{|8.5 - 14.5|}{0.57 + 0.57} = 5.26$$

$$\text{t-test}(X_2, C) = \text{S2N}(X_2, C) = \frac{|4.5 - 4.75|}{1.29 + 1.5} = 0.089$$

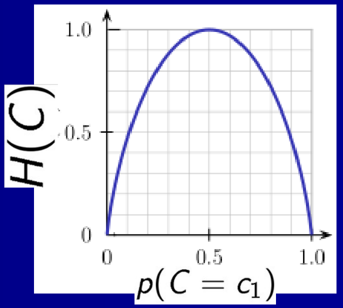
e.g.: $\bar{X}_2^{c=c_1}$

$$t_{test}(X_i, C) = \frac{|\mu_{X_i}^{c_1} - \mu_{X_i}^{c_2}|}{\sigma_{X_i}^{c_1} + \sigma_{X_i}^{c_2}}$$

UNIVARIATE FILTER
CORRELATION by MUTUAL-INFORMATION

- Based on the Entropy concept of a random variable

$$H(C) = -\sum_{c_i} p(c_i)\log_2 p(c_i)$$



- For example: information gain - mutual information, symmetrical uncertainty

$$MI(X_i, C) = IG(X_i, C) = H(C) - H(C|X_i)$$

$$SU(X_i, C) = \frac{2 \times MI(X_i, C)}{H(X_i) + H(C)}$$

UNIVARIATE FILTER
CORRELATION by MUTUAL-INFORMATION

X_1	X_2	C
a	a	c_1
b	b	c_1
a	b	c_1
a	a	c_1
b	a	c_2
b	a	c_2
a	b	c_2
b	b	c_2

$$H(C) = -\sum_c p(c) \cdot \log_2 (p(c))$$

$$H(C) = -(0.5 \times \log_2 0.5 + 0.5 \times \log_2 0.5) = 1$$

$$I(X, C) = H(C) - H(C|X) = \begin{matrix}\text{mutual-inf.} \\ \text{inf.-gain}\end{matrix}$$

Entropy reduction in C

$$H(C|X) = \sum_x p(x) \cdot H(C|X=x)$$

$$I(X_1, C) = H(C) - H(C|X_1) = 1 - \left[\sum_{x_1} p(x_1) \cdot \overbrace{-\sum_c p(c|x_1) \cdot \log_2 p(c|x_1)}^{H} \right] =$$

$$= 1 - \left[p(X_1=a) \cdot -\sum_c p(c|X_1=a) \cdot \log_2 p(c|X_1=a) + p(X_1=b) \cdot -\sum_c p(c|X_1=b) \cdot \log_2 p(c|X_1=b) \right] =$$

$$= 1 - \Big[p(X_1=a) \cdot -\big(p(c_1|X_1=a) \cdot \log_2 p(c_1|X_1=a) + p(c_2|X_1=a) \cdot \log_2 p(c_2|X_1=a) \big) +$$

$$+ p(X_1=b) \cdot -\big(p(c_1|X_1=b) \cdot \log_2 p(c_1|X_1=b) + p(c_2|X_1=b) \cdot \log_2 p(c_2|X_1=b) \big) \Big] =$$

$$= 1 - \left[4/8 \times -(3/4 \cdot \log_2 3/4 + 1/4 \cdot \log_2 1/4) + 4/8 \times -(1/4 \cdot \log_2 1/4 + 3/4 \cdot \log_2 3/4) \right] =$$

$$= 1 - [4/8 \times -0.8075 + 4/8 \times -0.8075] = 0.1925$$

$$I(X_2, C) = H(C) - H(C|X_2) = \emptyset$$

UNIVARIATE FILTER CORRELATION by CHI-SQUARE

The value of the test-statistic is

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i},$$

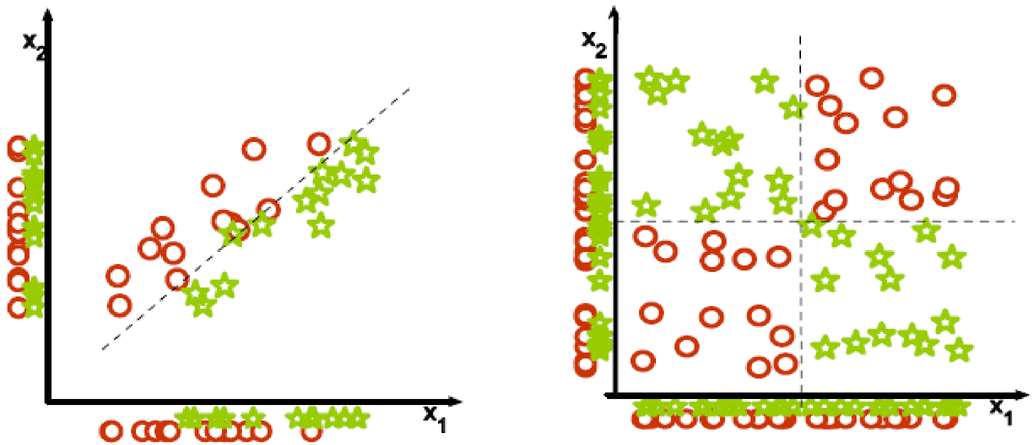
where

X^2 = Pearson's cumulative test statistic, which asymptotically approaches a χ^2 distribution.
O_i = an observed frequency;
E_i = an expected (theoretical) frequency, asserted by the null hypothesis;
n = the number of cells in the table.

	C=0	C=1	
X=low	16	2	18
X=medium	3	6	9
X=high	1	22	23
	20	30	50

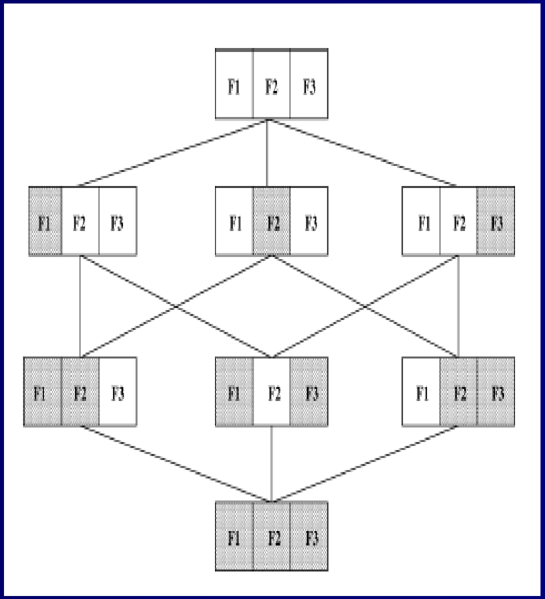
- If both variables were independent (predictor and class) → number of "expected" cases in each cell would be the product of the marginal, divided by the total number of cases:
  - e.g., "expected" number of cases in $cell_{11}$ is: (18x20)/50=7.2
  - while the "observed" number of cases in $cell_{11}$ is: 16
- Larger differences between expected and observed values in each cell → indicative of a "dependence-relationship" between predictor and class

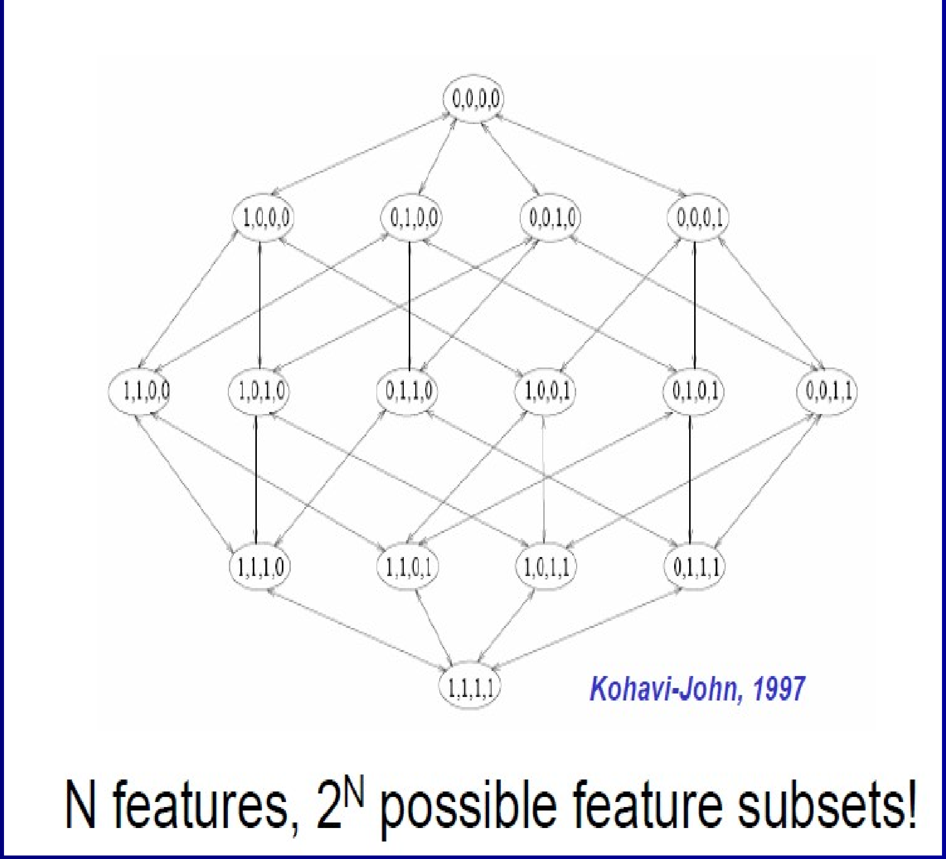
Univariate selection may fail



Guyon-Elisseeff, JMLR 2004; Springer 2006

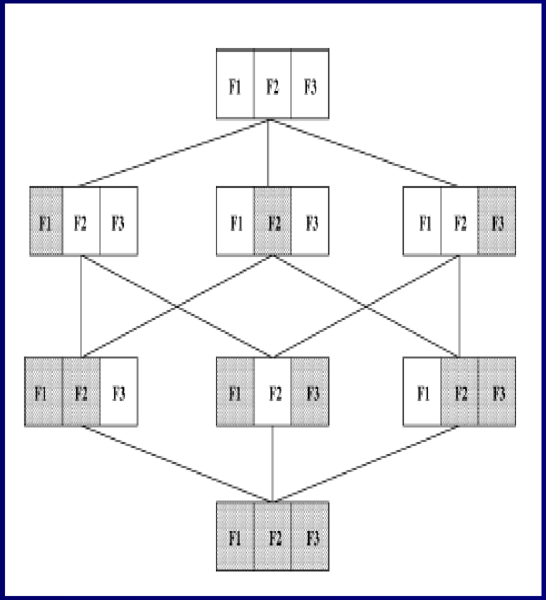
MULTIVARIATE FEATURE SELECTION NUMBER OF FEATURE SUBSETS

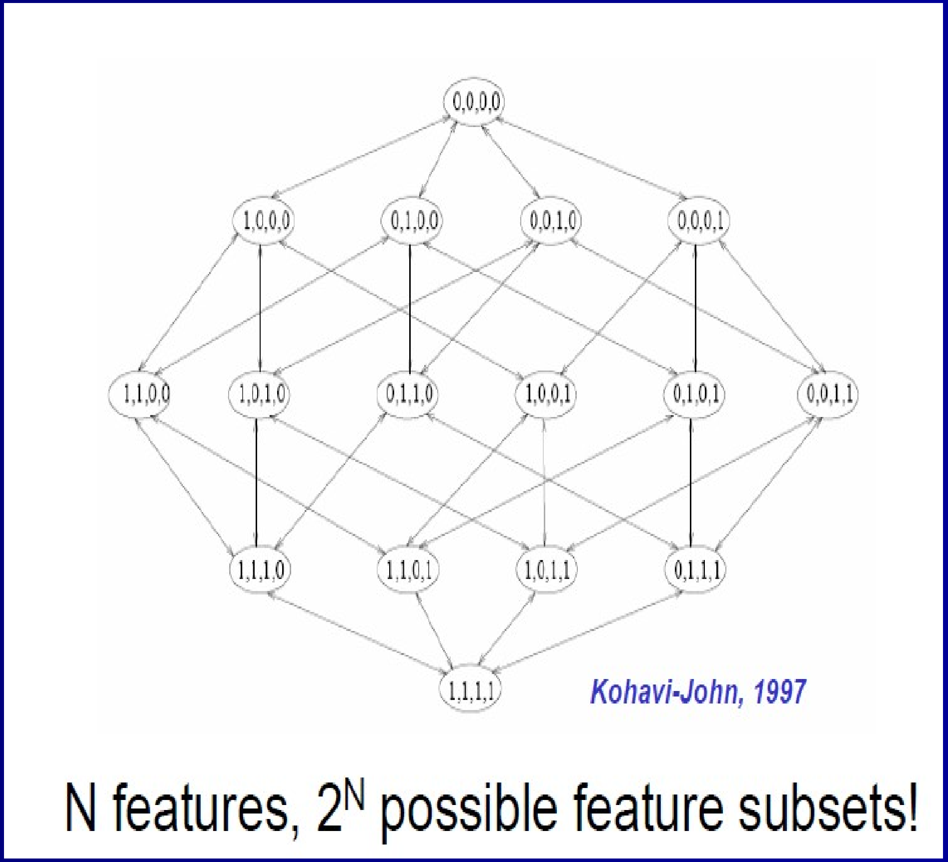




- Can the goodness of all feature subsets be evaluated? Hardly…
- 2^{10}=1024; 2^{20}≈1,048,576; 2^{50}=1,125,899,906,842,624

HOW TO EVALUATE THE "GOODNESS" OF A FEATURE SUBSET?





SUBSET EVALUATION (i) MULTIVARIATE FILTER SELECTION "CORRELATED FEATURE SELECTION (CFS)"

Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning

Mark A. Hall MHALL@CS.WAIKATO.AC.NZ
Department of Computer Science, University of Waikato, Hamilton, New Zealand

$$Merit_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \quad (1)$$

where $Merit_S$ is the heuristic "merit" of a feature subset S containing k features, $\overline{r_{cf}}$ the average feature-class correlation, and $\overline{r_{ff}}$ the average feature-feature intercorrelation. Equation 1 is, in fact, Pearson's correlation, where all variables have been standardized. The numerator can be thought of as giving an indication of how predictive a group of features are; the denominator of how much redundancy there is among them. The heuristic handles irrelevant features as they will be poor predictors of the class. Redundant attributes are discriminated against as they will be

SUBSET EVALUATION (i) MULTIVARIATE FILTER SELECTION "CORRELATED FEATURE SELECTION (CFS)"

- The *Merit*$_S$ function is calculated for each found feature subset *S*, for example:

  $$Merit_S\,(X_3, X_6, X_8)?$$

- Relevance concept, r_{cf}: to be augmented
- Correlation of each {predictor ~ class} → enhances the merit-metric, e.g., corr(X_3,*Class*), corr(X_6,*Class*), corr(X_8,*Class*)
- Irrelevant features → hurt e.g. K-NN

- Redundancy concept, r_{ff}: to be diminished
- Correlation among pairs of predictors → reduces the merit-metric, e.g., corr(X_3, X_6), corr(X_6, X_8), corr(X_8, X_3)
- Redundant features → hurt e.g. naïve Bayes

- Any type of correlation measure can be used (t-test, MI...)!

$$Merit_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}$$

SUBSET EVALUATION (i)
MULTIVARIATE FILTER SELECTION
"MAX-RELEVANCE + MIN-REDUNDANCY"

Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy

Hanchuan Peng, *Member*, *IEEE*, Fuhui Long, and Chris Ding

$$\max D(S,c), \quad D = \frac{1}{|S|} \sum_{x_i \in S} I(x_i; c)$$

$$\min R(S), \quad R = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i, x_j)$$

$$\max \Phi(D,R), \Phi = D - R$$

MULTIVARIATE FILTER SELECTION …MORE TECHNIQUES…

Criterion	Full name	Authors
MIM	Mutual Information Maximisation	Lewis (1992)
MIFS	Mutual Information Feature Selection	Battiti (1994)
KS	Koller-Sahami metric	Koller and Sahami (1996)
JMI	Joint Mutual Information	Yang and Moody (1999)
MIFS-U	MIFS-'Uniform'	Kwak and Choi (2002)
IF	Informative Fragments	Vidal-Naquet and Ullman (2003)
FCBF	Fast Correlation Based Filter	Yu and Liu (2004)
AMIFS	Adaptive MIFS	Tesmer and Estevez (2004)
CMIM	Conditional Mutual Info Maximisation	Fleuret (2004)
MRMR	Max-Relevance Min-Redundancy	Peng et al. (2005)
ICAP	Interaction Capping	Jakulin (2005)
CIFE	Conditional Infomax Feature Extraction	Lin and Tang (2006)
DISR	Double Input Symmetrical Relevance	Meyer and Bontempi (2006)
MINRED	Minimum Redundancy	Duch (2006)
IGFS	Interaction Gain Feature Selection	El Akadi et al. (2008)
SOA	Second Order Approximation	Guo and Nixon (2009)
CMIFS	Conditional MIFS	Cheng et al. (2011)

Journal of Machine Learning Research 13 (2012) 27-66

Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection

Gavin Brown GAVIN.BROWN@CS.MANCHESTER.AC.UK
Adam Pocock ADAM.POCOCK@CS.MANCHESTER.AC.UK
Ming-Jie Zhao MING-JIE.ZHAO@CS.MANCHESTER.AC.UK
Mikel Luján MIKEL.LUJAN@CS.MANCHESTER.AC.UK
School of Computer Science
University of Manchester
Manchester M13 9PL, UK

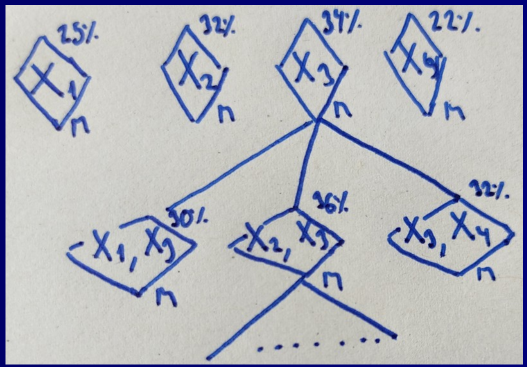
$$J_{mrmr}(X_k) = I(X_k;Y) - \frac{1}{|S|}\sum_{j\in S} I(X_k;X_j)$$

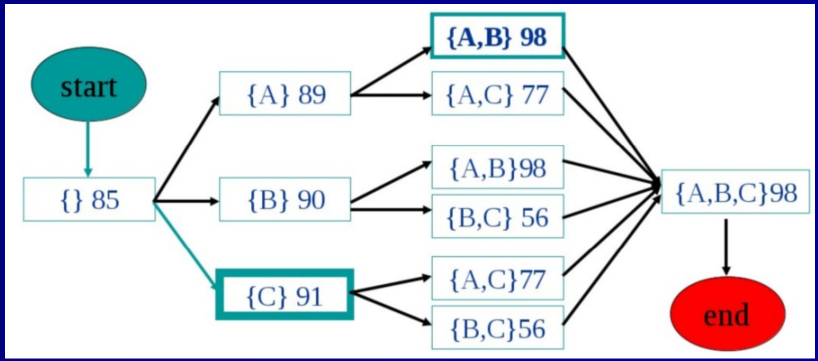
$$J'_{cmi}(X_k) = I(X_k;Y) - \sum_{j\in S} I(X_j;X_k) + \sum_{j\in S} I(X_j;X_k|Y)$$

SUBSET EVALUATION (ii) WRAPPER FEATURE SELECTION

- Fixed a <u>classification algorithm</u>
- Fixed a way to validate classifiers
- {Learn + Validate} classifier with the feature subset

FEATURE SET	CLASSIFIER	PERFORMANCE
{A,B,C}	*M*	98 %
{A,B}	*M*	98 %
{A,C}	*M*	77 %
{B,C}	*M*	56 %
{A}	*M*	89 %
{B}	*M*	90 %
{C}	*M*	91 %
{}	*M*	85 %



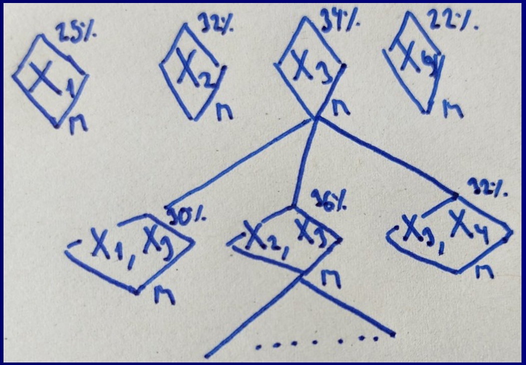


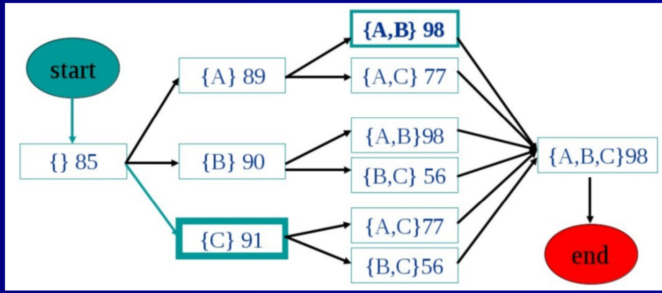
SUBSET EVALUATION (ii) WRAPPER FEATURE SELECTION

- Fixed a classification algorithm
- Fixed a way to validate classifiers
- {Learn + Validate} classifier with the feature subset

- Be careful! CPU cost!
- Computational costs → naïve Bayes ‹‹‹ K-nearest neighbour ‹‹‹ neural networks…

- Accuracy estimation → k-fold cross-validation? → CPU cost!





MULTIVARIATE FEATURE SELECTION SEARCH PROBLEM

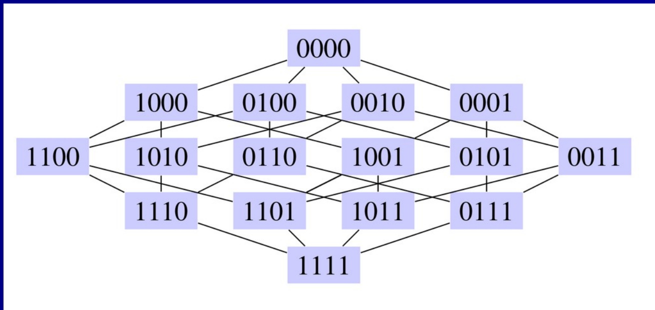
- Multivariate feature selection → search for the optimal feature subset

- How many different feature subsets? → 2^d
- 2^{50}=1,125,899,906,842,624 !!!!
- NP-hard problem

- Search heuristics → allowed
- Return → "suboptimal" solution

- Guarantee "optimal" solution → exhaustive search
- Computationally unfeasible!

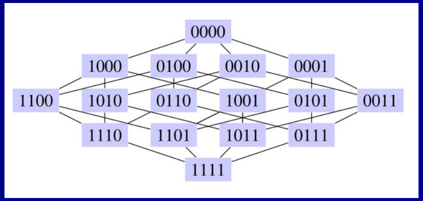
- Common approach:
  - univariate filter + multivariate

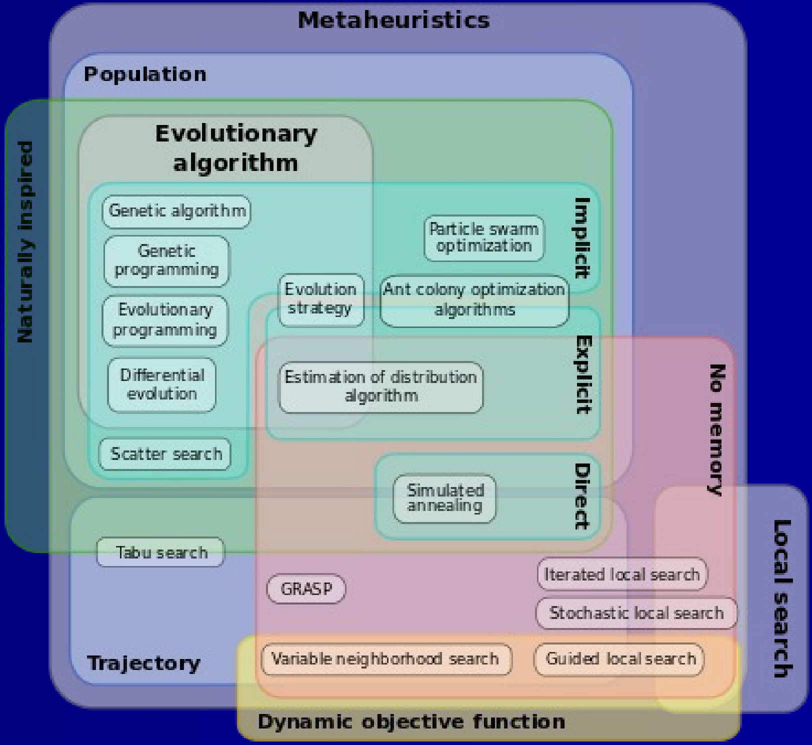


MULTIVARIATE FEATURE SELECTION SEARCH HEURISTICS

- Incremental - “local” heuristics
- Forward feature selection
- Backward feature elimination
- GRASP
- Simulated Annealing
- …

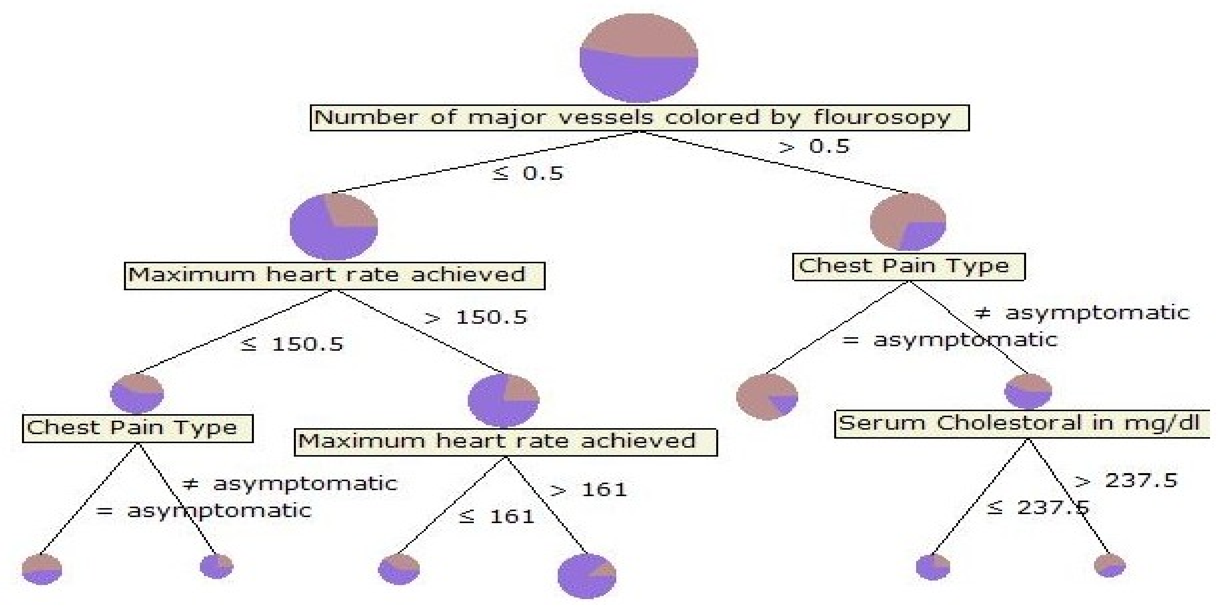
- “Population-based” - “Global”
- Genetic algorithms
- EDAs
- Ant-colony optimization
- …





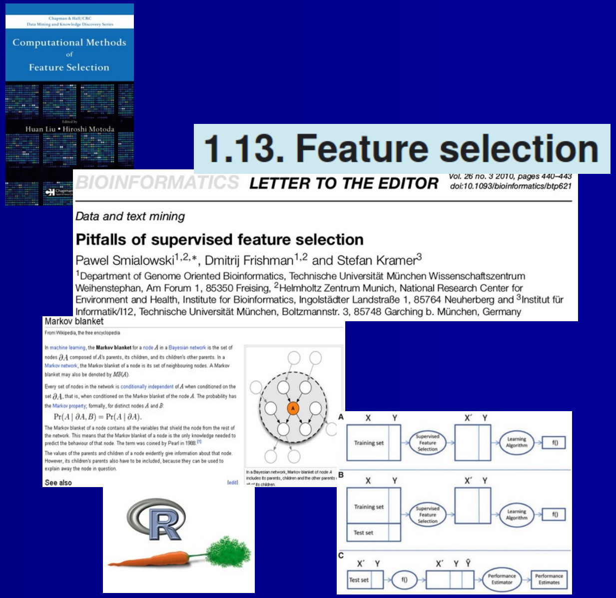
EMBEDDED FEATURE SELECTION

- Several classification algorithms
- “Embed” the capacity to discard initial features
- e.g. decision trees, random forest...



REMARKS ON FS

- FS by Markov Blanket
- Honest model evaluation when applying FS
- Stability on FS
- MultiObjective FS
- FS in other learning scenarios?
- References and software



MULTIVARIATE FILTER: MARKOV BLANKET

- Idea: selecting the features involved in the Markov blanket of the class variable in the learned Bayesian network structure

- Markov Blanket is a concept of Bayesian network theory
- Constructed using an algorithm for Bayesian network learning

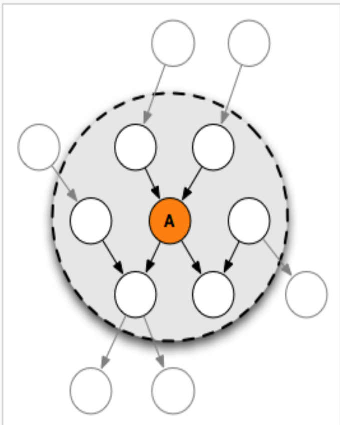
Markov blanket

From Wikipedia, the free encyclopedia

In machine learning, the **Markov blanket** for a node A in a Bayesian network is the set of nodes ∂A composed of A's parents, its children, and its children's other parents. In a Markov network, the Markov blanket of a node is its set of neighbouring nodes. A Markov blanket may also be denoted by $MB(A)$.

Every set of nodes in the network is conditionally independent of A when conditioned on the set ∂A, that is, when conditioned on the Markov blanket of the node A. The probability has the Markov property; formally, for distinct nodes A and B:

$$\Pr(A \mid \partial A, B) = \Pr(A \mid \partial A).$$



In a Bayesian network, Markov blanket of node A includes its parents, children and the other parents of all of its children.

Toward Optimal Feature Selection

Daphne Koller
Gates Building 1A
Computer Science Department
Stanford University
Stanford, CA 94305-9010
koller@cs.stanford.edu

Mehran Sahami
Gates Building 1A
Computer Science Department
Stanford University
Stanford, CA 94305-9010
sahami@cs.stanford.edu

HONEST ACCURACY ESTIMATION

Journal of Machine Learning Research 3 (2003) 1371-1382 Submitted 5/02; Published 3/03

Overfitting in Making Comparisons Between Variable Selection Methods

Juha Reunanen JUHA.REUNANEN@FI.ABB.COM
ABB, Web Imaging Systems
P.O. Box 94, 00381 Helsinki, Finland

BIOINFORMATICS ***LETTER TO THE EDITOR*** *Vol. 26 no. 3 2010, pages 440–443*
doi:10.1093/bioinformatics/btp621

Data and text mining

Pitfalls of supervised feature selection

Pawel Smialowski[1,2,*], Dmitrij Frishman[1,2] and Stefan Kramer[3]
[1]Department of Genome Oriented Bioinformatics, Technische Universität München Wissenschaftszentrum Weihenstephan, Am Forum 1, 85350 Freising, [2]Helmholtz Zentrum Munich, National Research Center for Environment and Health, Institute for Bioinformatics, Ingolstädter Landstraße 1, 85764 Neuherberg and [3]Institut für Informatik/I12, Technische Universität München, Boltzmannstr. 3, 85748 Garching b. München, Germany

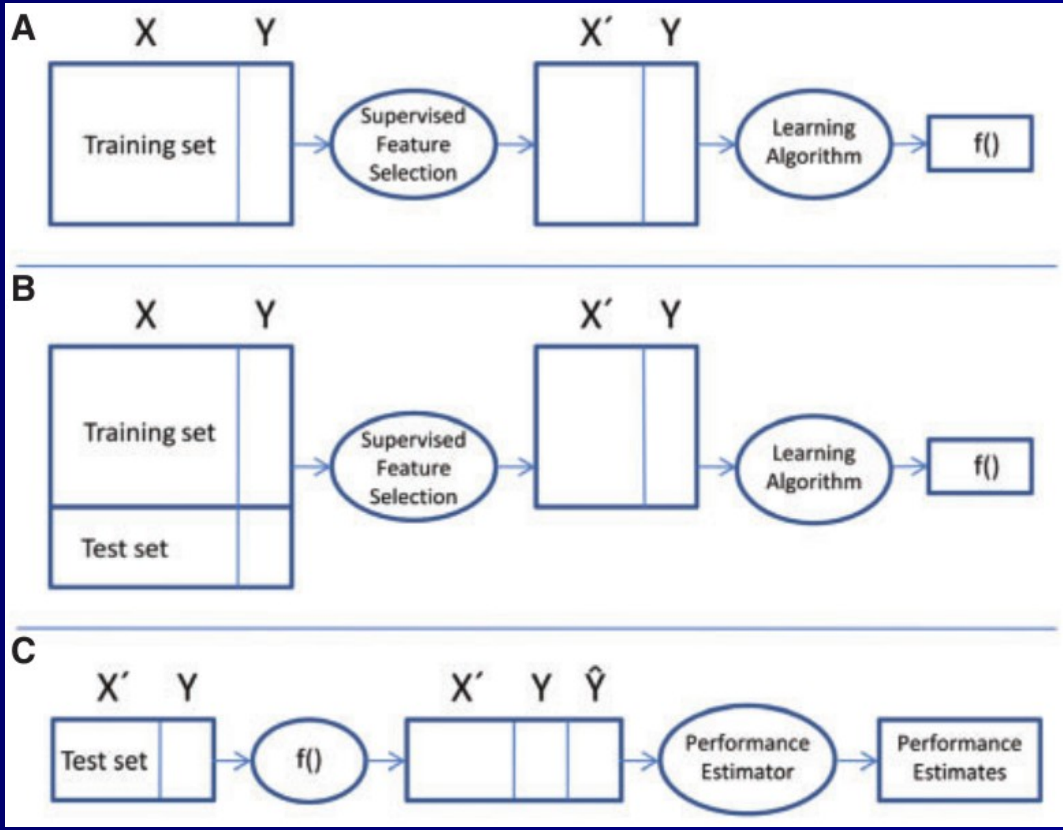
Received and revised on October 7, 2009; accepted on October 26, 2009
Advance Access publication October 29, 2009
Associate Editor: Martin Bishop

Pattern Recognition 81 (2018) 660–673
ELSEVIER
Contents lists available at ScienceDirect
Pattern Recognition
journal homepage: www.elsevier.com/locate/patcog

On feature selection protocols for very low-sample-size data

Ludmila I. Kuncheva[a], Juan J. Rodríguez[b,*]
[a] Bangor University, Dean Street, Bangor Gwynedd LL57 1UT, United Kingdom
[b] Universidad de Burgos, Escuela Politécnica Superior, Avda. de Cantabria s/n, Burgos 09006, Spain

HONEST ACCURACY ESTIMATION



Correct: A+C **Incorrect: B+C**

STABILITY ON FEATURE SELECTION

- When applying different feature selection techniques
- Or using different seeds to randomly partition data

Joint European Conference on Machine Learning and Knowledge Discovery in Databases
ECML PKDD 2016: Machine Learning and Knowledge Discovery in Databases pp 442-457

Measuring the Stability of Feature Selection

Authors | Authors and affiliations

Sarah Nogueira, Gavin Brown

- Variablity on the subsets of selected features

Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM'05)

Stability of Feature Selection Algorithms

Alexandros Kalousis, Julien Prados, Melanie Hilario
University of Geneva, Computer Science Department
Rue General Dufour 24, 1211 Geneva 4, Switzerland
{kalousis, prados, hilario}@cui.unige.ch

- Proposing metrics to measure the *stability* on the subsets of selected features

$$S_S(A,B) = 1 - \frac{|A| + |B| - 2|A \cap B|}{|A| + |B| - |A \cap B|} = \frac{|A \cap B|}{|A \cup B|}$$

- Learning a consensus subset?

$$I_C(A,B) = \frac{r - \frac{k^2}{n}}{k - \frac{k^2}{n}} = \frac{rn - k^2}{k(n-k)}$$

$$S_H(A,B) = 1 - \frac{|A \setminus B| + |B \setminus A|}{n}$$

MULTIOBJECTIVE FEATURE SELECTION

- Optimization of 2 conflicting objectives:
  - accuracy
  - subset dimension

- Not a single solution
- Pareto Front: non-dominated solutions

- Why not other objective apart from accuracy?

IEEE TRANSACTIONS ON CYBERNETICS 1

Particle Swarm Optimization for Feature Selection in Classification: A Multi-Objective Approach

Bing Xue, *Member, IEEE*, Mengjie Zhang, *Senior Member, IEEE*, and Will N. Browne

International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 17, No. 6 (2003) 903–929
© World Scientific Publishing Company

World Scientific
www.worldscientific.com

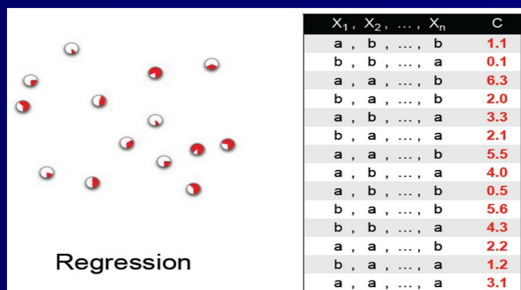
A METHODOLOGY FOR FEATURE SELECTION USING MULTIOBJECTIVE GENETIC ALGORITHMS FOR HANDWRITTEN DIGIT STRING RECOGNITION

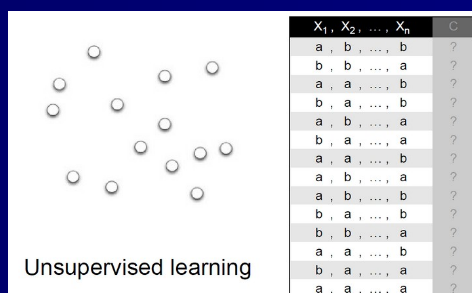
L. S. OLIVEIRA* and R. SABOURIN
École de Technologie Supérieure (ETS),
Laboratoire d'Imagerie, de Vision et d'Intelligence Artificielle (LIVIA),
Department de Génie de la Production Automatisée (GPA),
1100, rue Notre Dame Ouest, Montreal, Canada H3C 1K3
** soares@livia.etsmtl.ca*

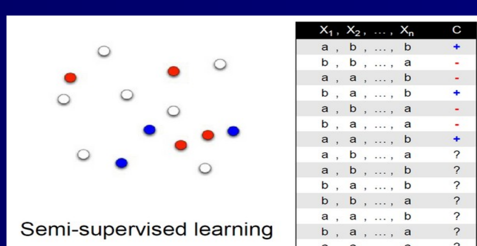
F. BORTOLOZZI
Pontifícia Universidade Católica do Paraná (PUCPR),
Rua Imaculada Conceição 1155, Prado Velho, 80215-901, Curitiba Pr, Brazil

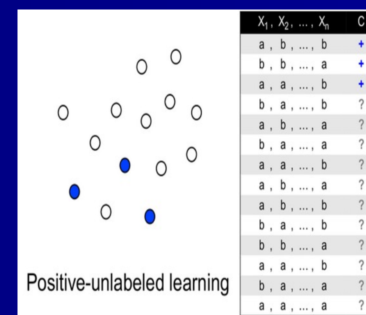
C. Y. SUEN
Centre for Pattern Recognition and Machine Intelligence (CENPARMI),
1455 de Maisonneuve Blvd. West, Suite GM 606 - Montreal, Canada H3G 1M8

FS IN OTHER LEARNING SCENARIOS



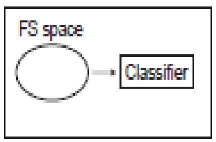
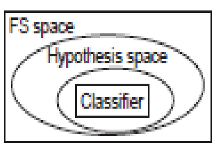
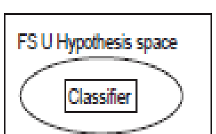




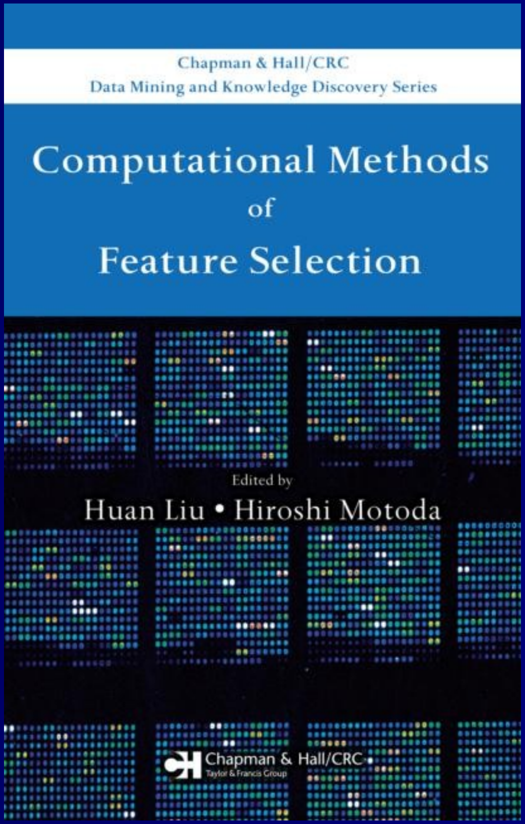


X_1	X_2	...	X_n	C_1	C_2	...	C_m
$x_1^{(1)}$	$x_2^{(1)}$	...	$x_n^{(1)}$	$c_1^{(1)}$	$c_2^{(1)}$	...	$c_m^{(1)}$
$x_1^{(2)}$	$x_2^{(2)}$	...	$x_n^{(2)}$	$c_1^{(2)}$	$c_2^{(2)}$	...	$c_m^{(2)}$
...	...	...	...	...	...	...	...
$x_1^{(N)}$	$x_2^{(N)}$	...	$x_n^{(N)}$	$c_1^{(N)}$	$c_2^{(N)}$	...	$c_m^{(N)}$

FILTER vs. WRAPPER vs. EMBEDDED

	Model search		Advantages	Disadvantages	Examples
Filter		Univariate	Fast Scalable Independent of the classifier	Ignores feature dependencies Ignores interaction with the classifier	Chi-square Euclidean distance t-test Information gain, Gain ratio [6]
		Multivariate	Models feature dependencies Independent of the classifier Better computational complexity than wrapper methods	Slower than univariate techniques Less scalable than univariate techniques Ignores interaction with the classifier	Correlation based feature selection (CFS) [45] Markov blanket filter (MBF) [62] Fast correlation based feature selection (FCBF) [136]
Wrapper		Deterministic	Simple Interacts with the classifier Models feature dependencies Less computationally intensive than randomized methods	Risk of over fitting More prone than randomized algorithms to getting stuck in a local optimum (greedy search) Classifier dependent selection	Sequential forward selection (SFS) [60] Sequential backward elimination (SBE) [60] Plus q take-away r [33] Beam search [106]
		Randomized	Less prone to local optima Interacts with the classifier Models feature dependencies	Computationally intensive Classifier dependent selection Higher risk of overfitting than deterministic algorithms	Simulated annealing Randomized hill climbing [110] Genetic algorithms [50] Estimation of distribution algorithms [52]
Embedded			Interacts with the classifier Better computational complexity than wrapper methods Models feature dependencies	Classifier dependent selection	Decision trees Weighted naive Bayes [28] Feature selection using the weight vector of SVM [44, 125]

REFERENCES



Journal of Machine Learning Research 3 (2003) 1157-1182 Submitted 11/02; Published 3/03

An Introduction to Variable and Feature Selection

Isabelle Guyon ISABELLE@CLOPINET.COM
Clopinet
955 Creston Road
Berkeley, CA 94708-1501, USA

André Elisseeff ANDRE@TUEBINGEN.MPG.DE
Empirical Inference for Machine Learning and Perception Department
Max Planck Institute for Biological Cybernetics
Spemannstrasse 38
72076 Tübingen, Germany

Journal of Machine Learning Research 13 (2012) 27-66 Submitted 12/10; Revised 6/11; Published 1/12

Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection

Gavin Brown GAVIN.BROWN@CS.MANCHESTER.AC.UK
Adam Pocock ADAM.POCOCK@CS.MANCHESTER.AC.UK
Ming-Jie Zhao MING-JIE.ZHAO@CS.MANCHESTER.AC.UK
Mikel Luján MIKEL.LUJAN@CS.MANCHESTER.AC.UK
School of Computer Science
University of Manchester
Manchester M13 9PL, UK

SOFTWARE FOR FS

1.13. Feature selection

the caret package