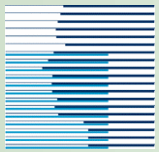
Robust Feature Selection by Mutual Information Distributions

Marco Zaffalon & Marcus Hutter



IDSIA

Galleria 2, 6928 Manno (Lugano), Switzerland

www.idsia.ch/~{zaffalon,marcus}

{zaffalon,marcus}@idsia.ch

Mutual Information (MI)

- Consider two discrete random variables (ι, γ)

$$p_{ij} = \text{joint chance of } (i, j), \quad i \in \{1, \ldots, r\} \text{ and } j \in \{1, \ldots, s\}$$

$$p_{i+} = \sum_j p_{ij} = \text{marginal chance of } i$$

$$p_{+j} = \sum_i p_{ij} = \text{marginal chance of } j$$

- (In)Dependence often measured by MI

$$0 \leq I(\mathbf{p}) = \sum_{ij} p_{ij} \log \frac{p_{ij}}{p_{i+} p_{+j}}$$

  - Also known as *cross-entropy* or *information gain*
  - Examples
    - Inference of Bayesian nets, classification trees
    - Selection of relevant variables for the task at hand

MI-Based Feature-Selection Filter (F)

Lewis, 1992

- Classification
  - Predicting the *class* value given values of *features*
  - Features (or attributes) and class = random variables
  - Learning the rule 'features ➔ class' from data
- Filters goal: removing irrelevant features
  - More accurate predictions, easier models
- MI-based approach
  - Remove feature ι if class γ does not depend on it: $I(\mathbf{p}) = 0$
  - Or: remove ι if $I(\mathbf{p}) < e$
    - $e \in \Re^+$ is an arbitrary threshold of relevance

Empirical Mutual Information

a common way to use MI in practice

- Data ($\mathbf{n}$) ➔ contingency table

n_{ij} = # of times (i,j) occurred

$n_{i+} = \sum_j n_{ij}$ = # of times i occurred

$n_{+j} = \sum_i n_{ij}$ = # of times j occurred

$n = \sum_{ij} n_{ij}$ = dataset size

$j \backslash i$	1	2	…	r
1	n_{11}	n_{12}	…	n_{1r}
2	n_{21}	n_{22}	…	n_{2r}
⋮	⋮	⋮	⋱	⋮
s	n_{s1}	n_{s2}	…	n_{sr}

  - Empirical (sample) probability: $\hat{p}_{ij} = n_{ij} / n$
  - Empirical mutual information: $I(\hat{\mathbf{p}})$

- Problems of the empirical approach
  - $I(\hat{\mathbf{p}}) = 0$ due to random fluctuations? (finite sample)
  - How to know if it is reliable, e.g. by $P(I > e | \mathbf{n})$?

We Need the Distribution of MI

- Bayesian approach
  - Prior distribution $p(\mathbf{p})$ for the unknown chances (e.g., Dirichlet)
  - Posterior: $p(\mathbf{p}|\mathbf{n}) \propto p(\mathbf{p}) \prod_{ij} p_{ij}^{n_{ij}}$
- Posterior probability density of MI:

$$p(I|\mathbf{n}) = \int d(I(\mathbf{p}) - I) p(\mathbf{p}|\mathbf{n}) d\mathbf{p}$$

- How to compute it?
  - Fitting a curve by the exact mean, approximate variance

Mean and Variance of MI

Hutter, 2001; Wolpert & Wolf, 1995

- Exact mean

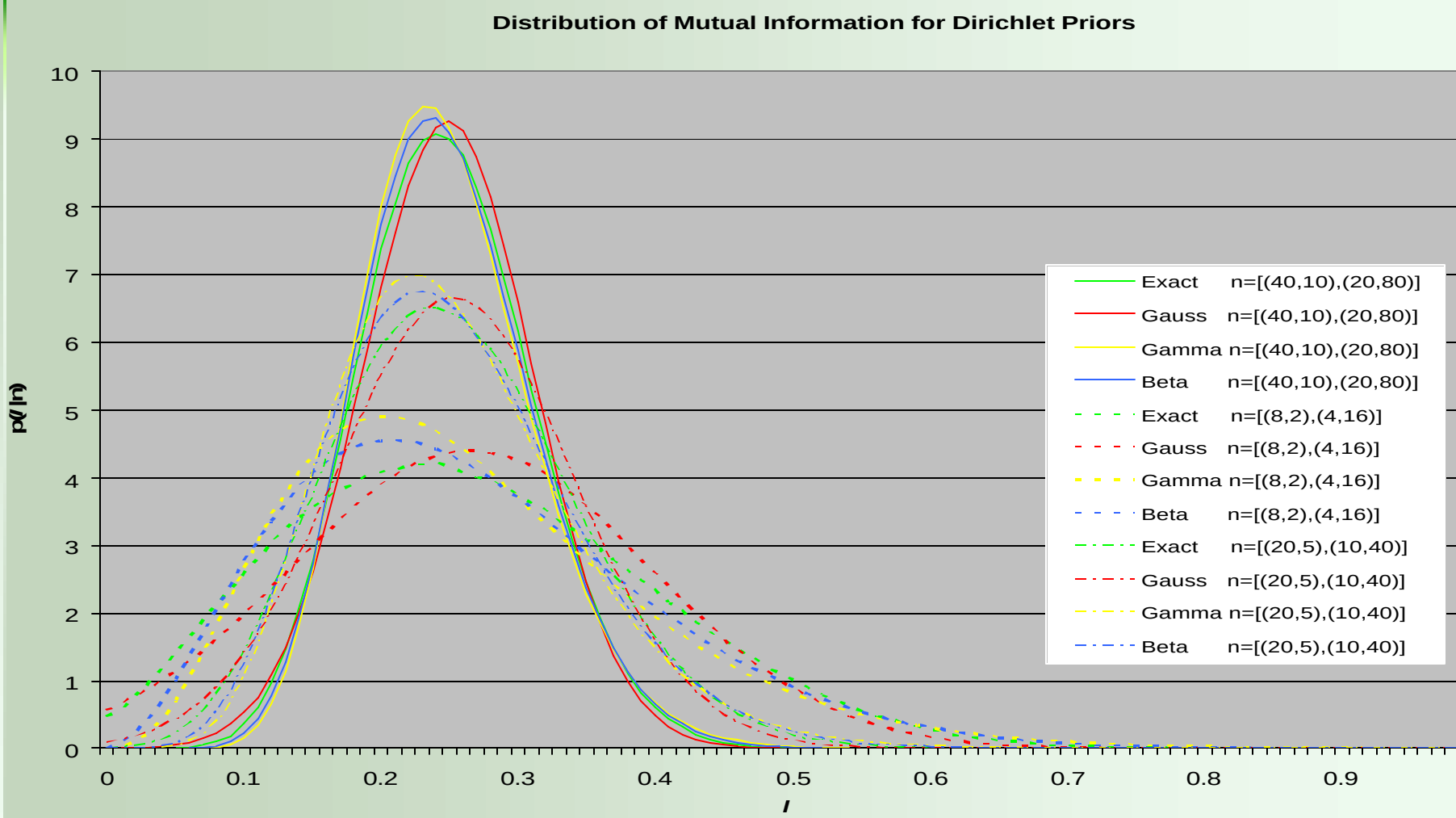
$$E[I] = \frac{1}{n}\sum_{ij} n_{ij}\left[y(n_{ij}+1) - y(n_{i+}+1) - y(n_{+j}+1) + y(n+1)\right], \quad y(n) = \sum_{k=1}^{n-1}\frac{1}{k}$$

- Leading and next to leading order term (NLO) for the variance

$$VAR[I] = \frac{1}{n}\sum_{ij}\frac{n_{ij}}{n}\left(\log\frac{n_{ij}n}{n_{i+}n_{+j}}\right)^2 - \frac{1}{n}\left(\sum_{ij}\frac{n_{ij}}{n}\log\frac{n_{ij}n}{n_{i+}n_{+j}}\right)^2 + \text{NLO} + O\left(n^{-3}\right)$$

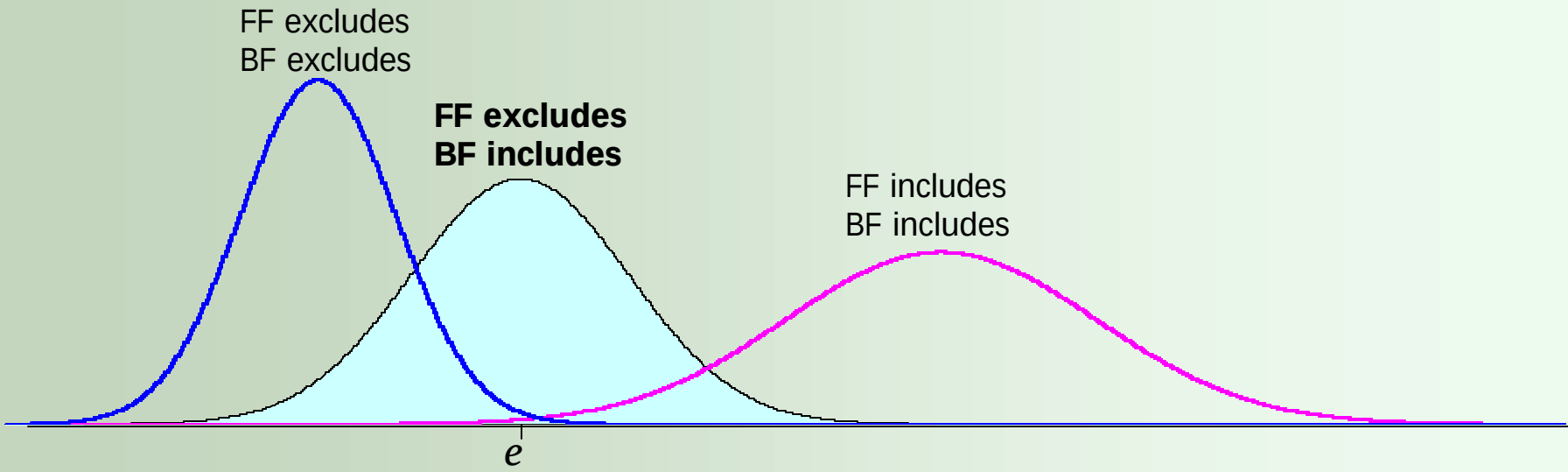
- Computational complexity $O(rs)$
  - As fast as empirical MI

MI Density Example Graphs



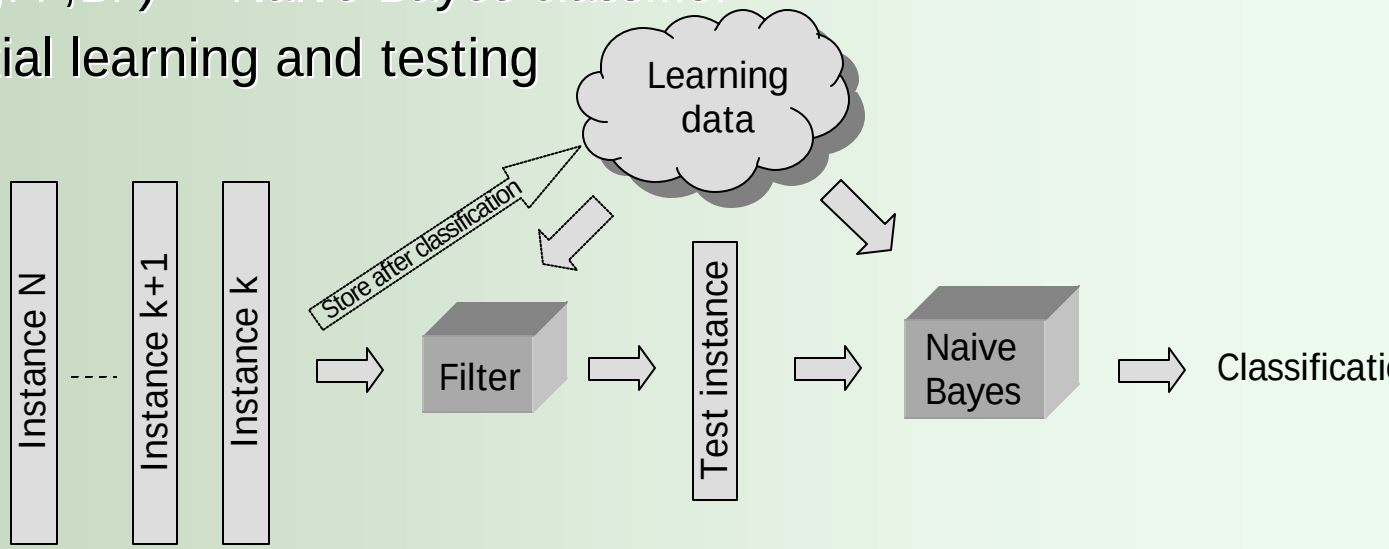
Robust Feature Selection

- Filters: two new proposals
  - FF: include feature ι iff $P\left(I > e \middle| \mathbf{n}\right) > 0.95$
    - (include iff "proven" relevant)
  - BF: exclude feature ι iff $P\left(I \leq e \middle| \mathbf{n}\right) > 0.95$
    - (exclude iff "proven" irrelevant)
- Examples



Comparing the Filters

- Experimental set-up
  - Filter (F,FF,BF) + Naive Bayes classifier
  - Sequential learning and testing



- Collected measures for each filter
  - Average # of correct predictions (prediction accuracy)
  - Average # of features used

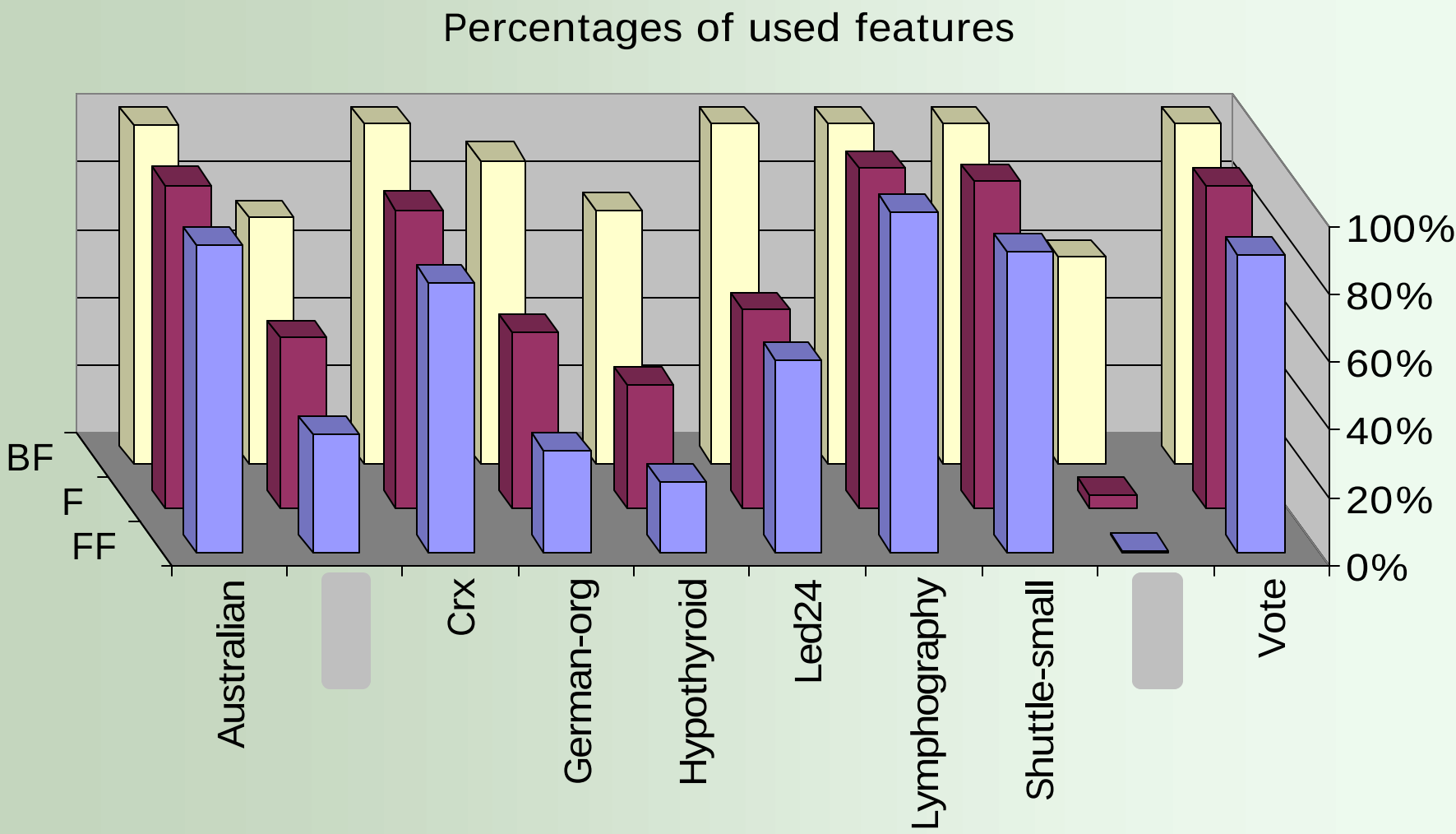
Results on 10 Complete Datasets

- # of used features

# Instances	# Features	Dataset	FF	F	BF
690	36	Australian	**32.6**	34.3	35.9
3196	36	**Chess**	**12.6**	18.1	26.1
653	15	Crx	**11.9**	13.2	15.0
1000	17	German-org	**5.1**	8.8	15.2
2238	23	Hypothyroid	**4.8**	8.4	17.1
3200	24	Led24	**13.6**	14.0	24.0
148	18	Lymphography	**18.0**	18.0	18.0
5800	8	Shuttle-small	**7.1**	7.7	8.0
1101	21611	**Spam**	**123.1**	822.0	13127.4
435	16	Vote	**14.0**	15.2	16.0

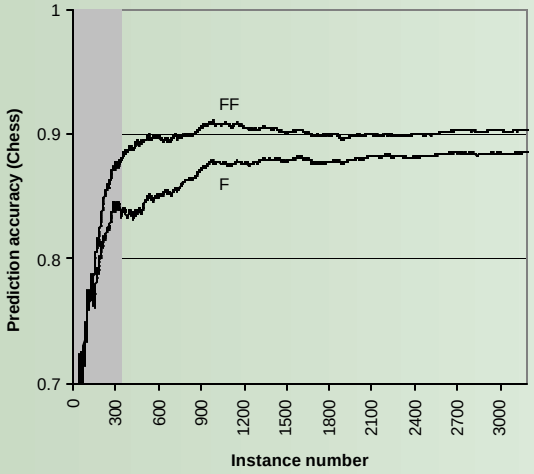
- Accuracies NOT significantly different
  - Except Chess & Spam with FF

Results on 10 Complete Datasets - ctd

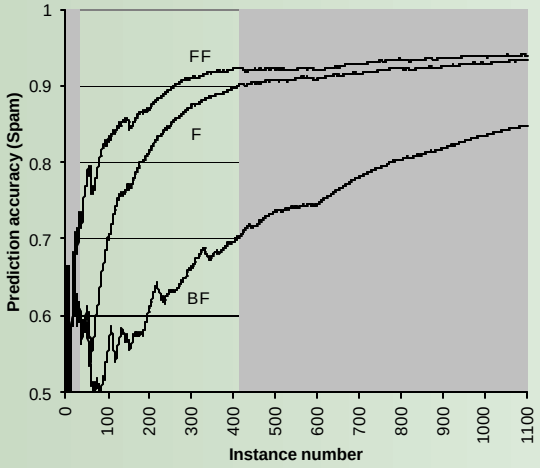


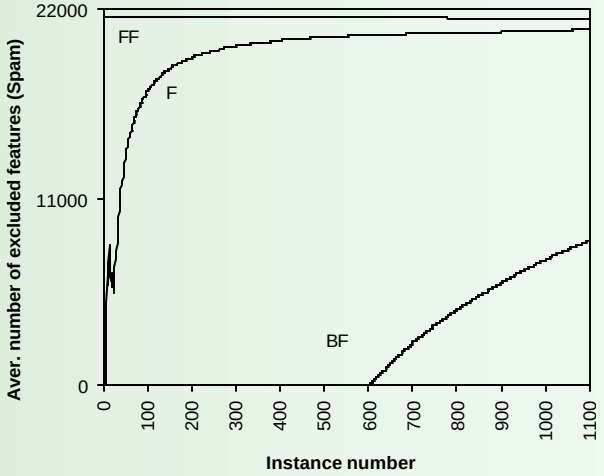
FF: Significantly Better Accuracies

- Chess



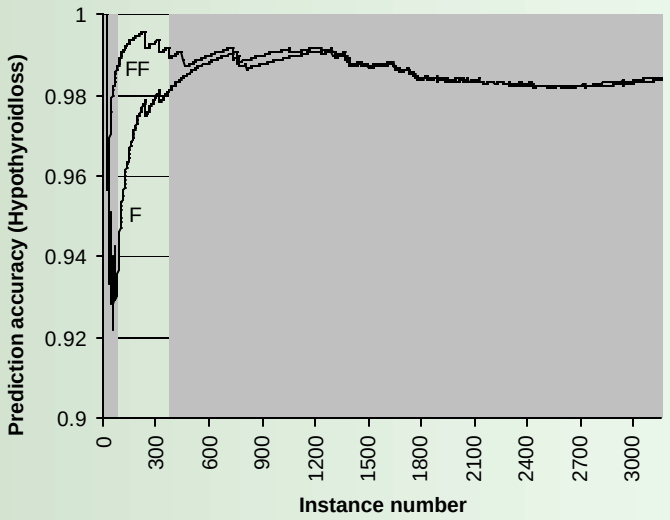
- Spam





Extension to Incomplete Samples

- MAR assumption
  - General case: missing features and class
    - EM + closed-form expressions
  - Missing features only
    - Closed-form approximate expressions for Mean and Variance
    - Complexity still $O(rs)$
- New experiments
  - 5 data sets
  - Similar behavior



Conclusions

- Expressions for several moments of MI distribution are available
  - The distribution can be approximated well
  - Safer inferences, same computational complexity of empirical MI
  - Why not to use it?
- Robust feature selection shows power of MI distribution
  - FF outperforms traditional filter F
- Many useful applications possible
  - Inference of Bayesian nets
  - Inference of classification trees
  - ...