# Bayesian Treatment of Incomplete Discrete Data applied to Mutual Information and Feature Selection


Marcus Hutter & Marco Zaffalon

IDSIA

Galleria 2, 6928 Manno (Lugano), Switzerland

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

{marcus,zaffalon}@idsia.ch

## Keywords

Incomplete data, Bayesian statistics, expectation maximization, global optimization, Mutual Information, Cross Entropy, Dirichlet distribution, Second order distribution, Credible intervals, expectation and variance of mutual information, missing data, Robust feature selection, Filter approach, naive Bayes classifier.

## Abstract

Given the joint chances of a pair of random variables one can compute quantities of interest, like the mutual information. The Bayesian treatment of unknown chances involves computing, from a second order prior distribution and the data likelihood, a posterior distribution of the chances. A common treatment of incomplete data is to assume ignorability and determine the chances by the expectation maximization (EM) algorithm. The two different methods above are well established but typically separated. This paper joins the two approaches in the case of Dirichlet priors, and derives efficient approximations for the mean, mode and the (co)variance of the chances and the mutual information. Furthermore, we prove the unimodality of the posterior distribution, whence the important property of convergence of EM to the global maximum in the chosen framework. These results are applied to the problem of selecting features for incremental learning and naive Bayes classification. A fast filter based on the distribution of mutual information is shown to outperform the traditional filter based on empirical mutual information on a number of incomplete real data sets.

# 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

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

  - 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 $\iota$ if class $\gamma$ does not depend on it: $I(\mathbf{p}) = 0$
  - Or: remove $\iota$ 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

$$n_{ij} = \#\text{ of times } (i,j) \text{ occurred}$$
$$n_{i+} = \sum_j n_{ij} = \#\text{ of times } i \text{ occurred}$$
$$n_{+j} = \sum_i n_{ij} = \#\text{ of times } j \text{ occurred}$$
$$n = \sum_{ij} n_{ij} = \text{dataset size}$$

| $j \backslash i$ | 1 | 2 | ... | $r$ |
|---|---|---|---|---|
| 1 | $n_{11}$ | $n_{12}$ | ... | $n_{1r}$ |
| 2 | $n_{21}$ | $n_{22}$ | ... | $n_{2r}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| $s$ | $n_{s1}$ | $n_{s2}$ | ... | $n_{sr}$ |

  - 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 $P(I > e|\mathbf{n})$?

# Incomplete Samples

- Missing features/classes
  - Missing class: (i,?) $\rightarrow$ $n_{i?}$ = # features i with missing class label
  - Missing feature: (?,j) $\rightarrow$ $n_{?j}$ = # classes j with missing feature
  - Total sample size $N_{ij}=n_{ij}+n_{i?}+n_{?j}$
- MAR assumption: $\pi_{i?}=\pi_{i+}$ , $\pi_{?j}=\pi_{+j}$
  - General case: missing features and class
    - EM + closed-form leading order in $N^{-1}$ expressions
  - Missing features only
    - Closed-form leading order expressions for Mean and Variance
    - Complexity $O(rs)$

# We Need the Distribution of MI

- Bayesian approach
  - Prior distribution $p(\mathbf{p})$ for the unknown chances (e.g., Dirichlet)
  - Posterior: $p(\mathbf{p}|\mathbf{n}) \propto p(\mathbf{p}) \prod_{ij} p_{ij}^{n_{ij}} \prod_{i} p_{i+}^{n_{i?}} \prod_{j} p_{+j}^{n_{?j}}$
- 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 using mode and approximate variance

# Mean and Variance of p and I

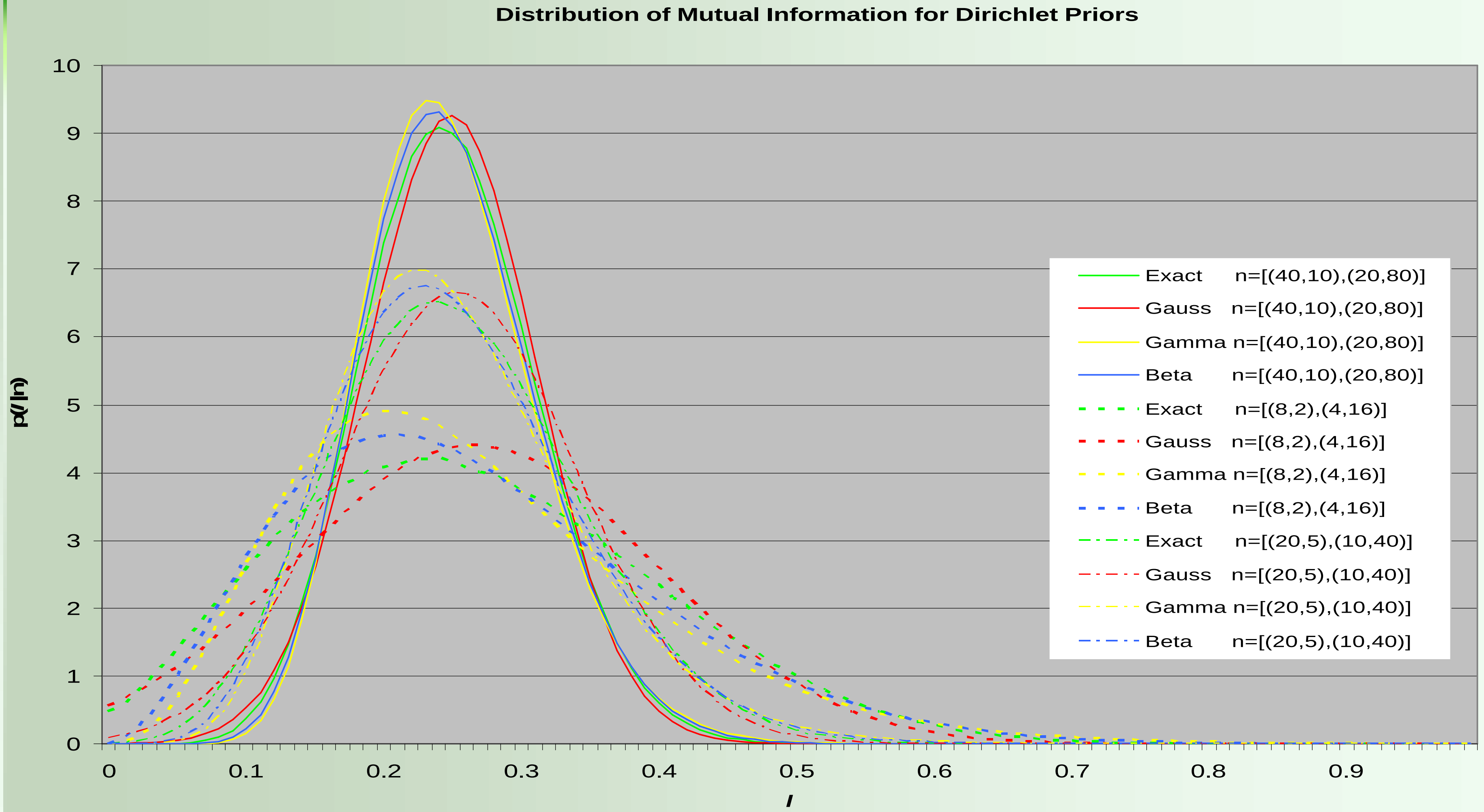
## (missing features only)

- Exact mode $\hat{p}_{ij} = \frac{N_{ij}}{N}\frac{n_{ij}}{n_{i+}} = E[p] + O(N^{-1})$ = leading mean
- Leading covariance: $Cov_{(ij)(kl)}[p] \cong \frac{1}{N}[r_{ij}d_{ik}d_{jl} - \frac{r_{ij}r_{kl}}{r_{i+}+r_{i?}}d_{ik} - \frac{r_{ij}Q_{i?}r_{kl}Q_{k?}}{Q}]$

  with $Q_{i?} := \frac{r_{i?}}{r_{i?}+r_{i+}}, \quad Q := \sum_i r_{i+}Q_{i?}, \quad r_{ij} = N\frac{\hat{p}_{ij}^2}{n_{ij}}, \quad r_{i?} = N\frac{\hat{p}_{i+}^2}{n_{i?}}$
- Exact mode = $I(\hat{p}) = E[I] + O(N^{-1})$ = leading order mean
- Leading variance: $Var[I] \cong \frac{1}{N}[K - J^2/Q - P], \quad K := \sum_{ij} r_{ij}(\log\frac{\hat{p}_{ij}}{\hat{p}_{i+}\hat{p}_{+j}})^2$

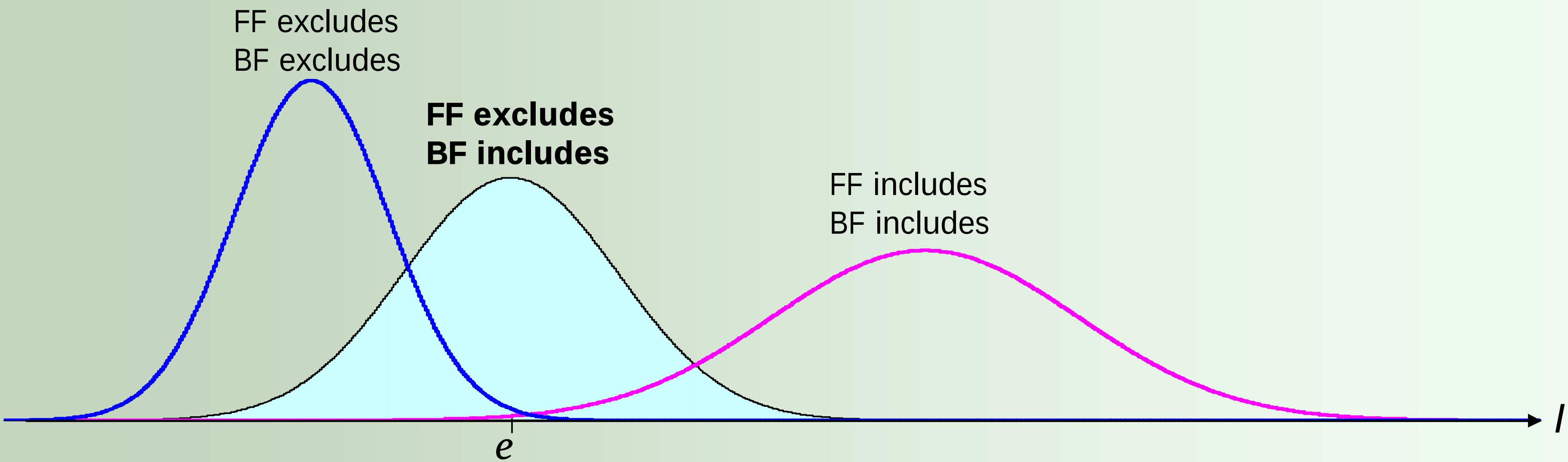
  $$P := \sum_i \frac{J_{i+}^2 Q_{i?}}{r_{i?}}, \quad J := \sum_i J_{i+}Q_{i?}, \quad J_{i+} := \sum_{ij} r_{ij}\log\frac{\hat{p}_{ij}}{\hat{p}_{i+}\hat{p}_{+j}}$$
- Missing features & classes: EM converges globally, since p(π|n) is unimodal

# MI Density Example Graphs
## (complete sample)



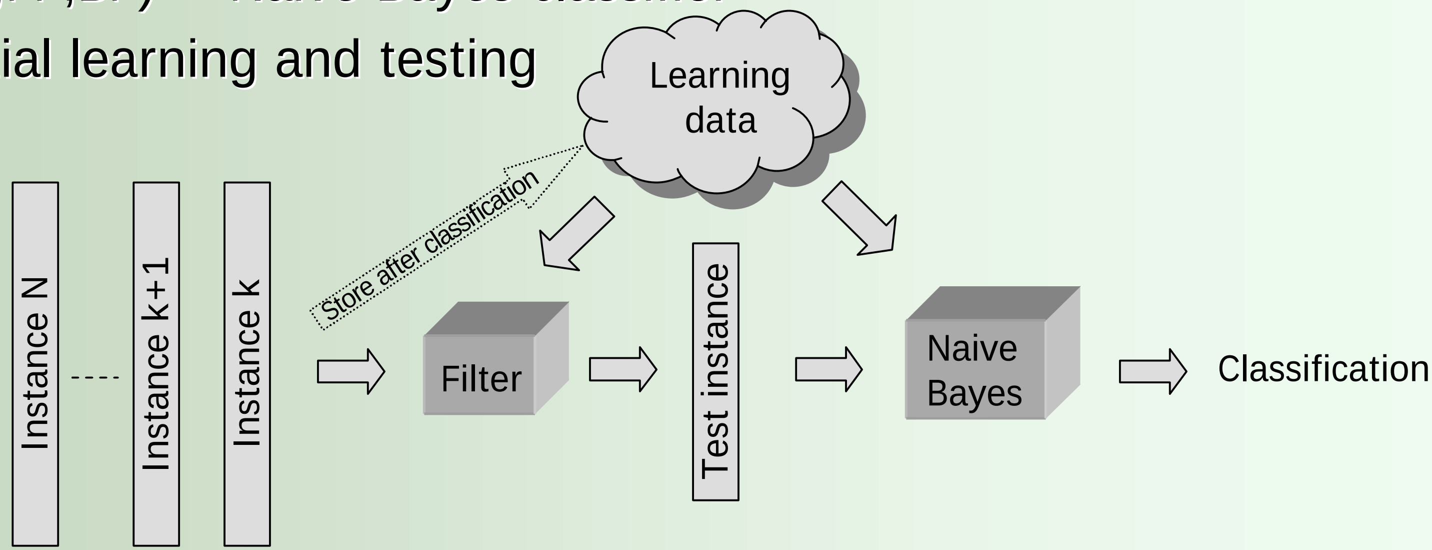
# Robust Feature Selection

- 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

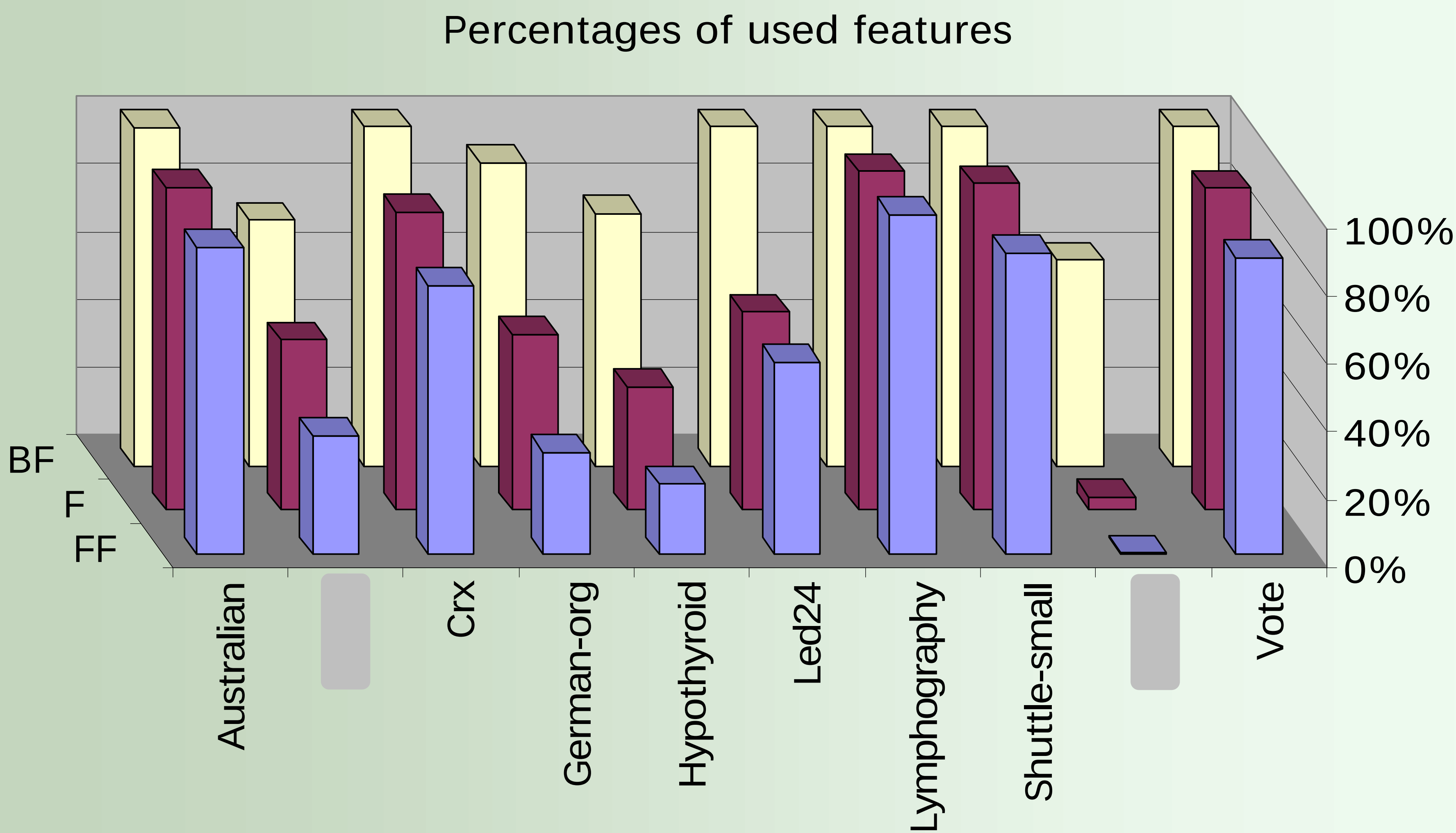
# Results on 10 Complete Datasets

- # 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

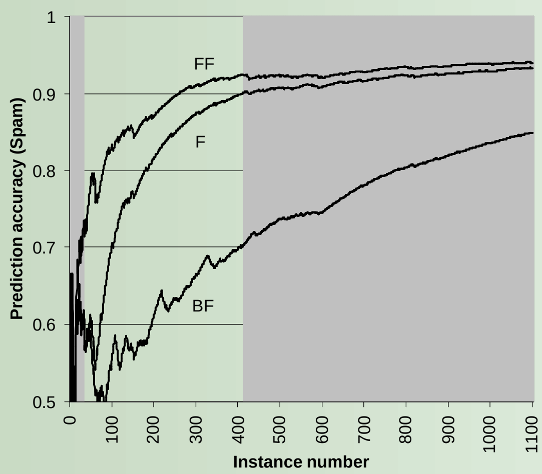
# Results on 10 Complete Datasets - ctd

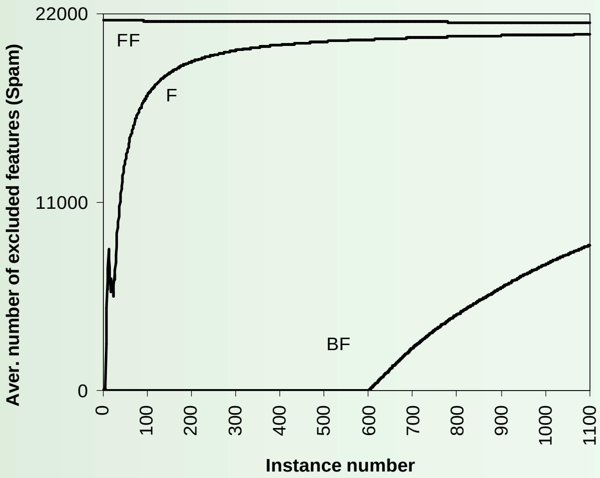


# FF: Significantly Better Accuracies

- Chess


- Spam

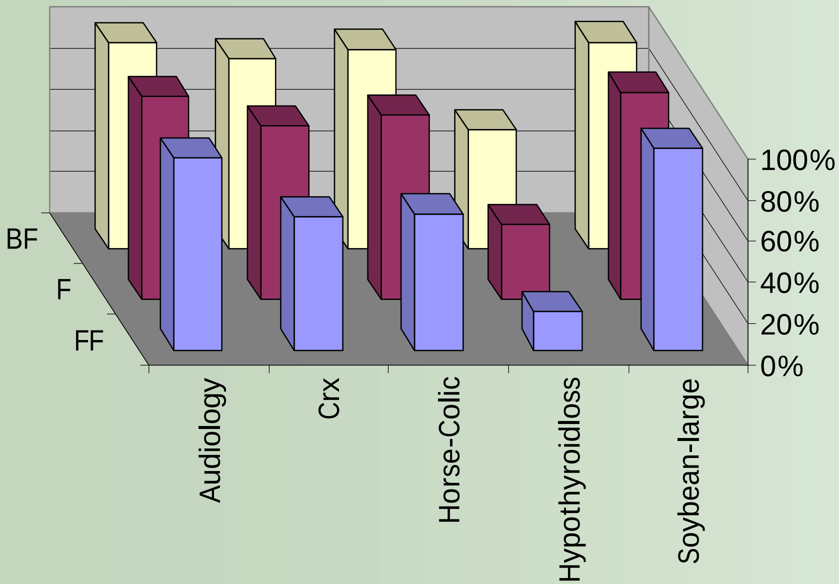


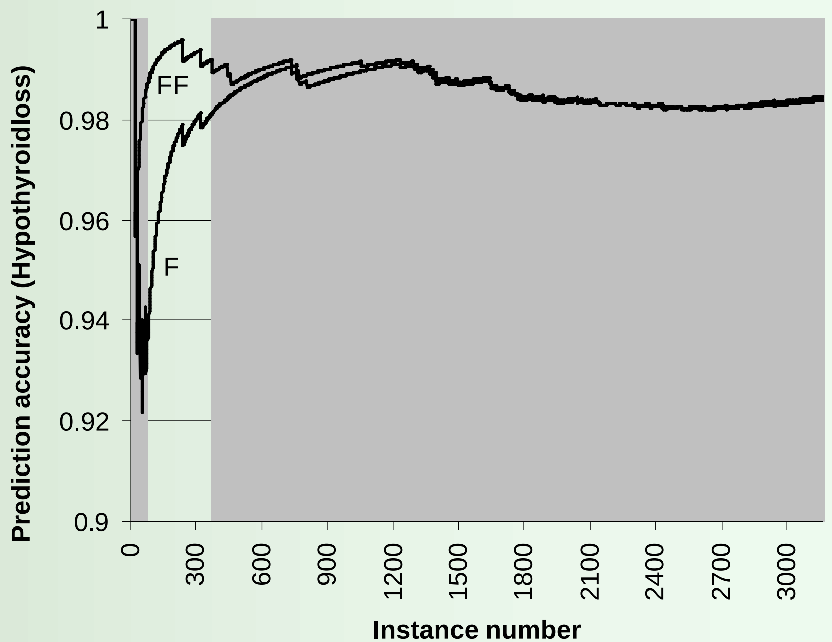



# Results on 5 Incomplete Data Sets

| # Instances | # Features | # miss.vals | Dataset | FF | F | BF |
|---|---|---|---|---|---|---|
| 226 | 69 | 317 | Audiology | **64.3** | 68.0 | 68.7 |
| 690 | 15 | 67 | Crx | **9.7** | 12.6 | 13.8 |
| 368 | 18 | 1281 | Horse-Colic | **11.8** | 16.1 | 17.4 |
| 3163 | 23 | 1980 | **Hypothyroidloss** | **4.3** | 8.3 | 13.2 |
| 683 | 35 | 2337 | Soybean-large | **34.2** | 35.0 | 35.0 |

Percentages of used features



# Conclusions

- Expressions for several moments of $\pi$ and MI distribution even for incomplete categorical data
  - 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
  - ...