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

Marcus Hutter and **Marco Zaffalon**

IDSIA, Galleria 2, CH-6928 Manno-Lugano, Switzerland
{marcus,zaffalon}@idsia.ch

24 June 2003

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.

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.

*This work was supported in parts by the NSF grants 2000-61847.00 and 2100-067961.02.

1 Introduction

Let $\pi_{\imath\jmath}$ be the joint chances of a pair of random variables $(\imath,\jmath)$. Many statistical quantities can be computes if $\boldsymbol{\pi}$ is known; for instance the *mutual information* $I(\boldsymbol{\pi})$ used for measuring the stochastic dependency of $\imath$ and $\jmath$. The usual procedure in the common case of *unknown chances* π_{ij} is to use the *empirical probabilities* $\hat{\pi}_{ij}=n_{ij}/n$ as if they were precisely known chances. This is not always suitable: (a) The point estimate $\hat{\pi}_{ij}$ does not carry information about the reliability of the estimate. (b) Samples (i,j) may be incomplete in the sense that in some samples the variable i or j may not be observed.

The *Bayesian* solution to (a) is to use a (second order) prior distribution $p(\boldsymbol{\pi})$ over the chances $\boldsymbol{\pi}$ themselves, which takes account of uncertainty about $\boldsymbol{\pi}$. From the prior $p(\boldsymbol{\pi})$ and the likelihood $p(\boldsymbol{D}|\boldsymbol{\pi})$ of data $\boldsymbol{D}$ one can compute the posterior $p(\boldsymbol{\pi}|\boldsymbol{D})$. The traditional solution to (b) is to assume that the data are *missing at random* [LR87]. A (local) maximum likelihood estimate for $\hat{\boldsymbol{\pi}}$ can then be obtained by the *expectation-maximization* (EM) algorithm [CF74].

In this work we present a full Bayesian treatment of incomplete discrete data with Dirichlet prior $p(\boldsymbol{\pi})$ and apply the results to *feature selection*. This work is a natural continuation of [ZH02], which focused on the case of complete data and, by working out a special case, provided encouraging evidence for the extension of the proposed approach to incomplete data. Here we develop that framework by creating a very general method for incomplete discrete data, providing the complete mathematical derivations, as well as experiments on incomplete real data sets. In particular, Section 2 derives expressions (in leading order in $1/n$) for $p(\boldsymbol{\pi}|\boldsymbol{D})$. In the important case (for feature selection) of missingness in one component of $(\imath,\jmath)$ only, we give closed form expressions for the mode, mean and covariance of $\boldsymbol{\pi}$. In the general missingness case we get a self-consistency equation which coincides with the EM algorithm, that is known to converge to a local maximum. We show that $p(\boldsymbol{\pi}|\boldsymbol{D})$ is actually unimodal, which implies that in fact *EM always converges to the global maximum*. We use the results to derive in Section 3 closed-form leading order expressions of the distribution of mutual information $p(I|\boldsymbol{D})$. In case of complete data, the mean and variance of I have been approximated numerically in [Kle99] and analytically in [Hut02]. The results are then applied to feature selection in Section 4. A popular *filter approach* discards features of low empirical mutual information $I(\hat{\boldsymbol{\pi}})$ [Lew92, BL97, CHH$^+$02]. We compare this filter to the two filters (introduced in [ZH02] for complete data and tested empirically in this case) that use *credible intervals* based on $p(I|\boldsymbol{D})$ to robustly estimate mutual information. The filters are empirically tested in Section 5 by coupling them with the *naive Bayes classifier* [DHS01] to incrementally learn from and classify incomplete data. On five real data sets that we used, one of the two proposed filters consistently outperforms the traditional filter.

2 Posterior Distribution for Incomplete Data

Missing data. Consider two discrete random variables, class $\imath$ and feature[1] $\jmath$ taking values in $\{1,...,r\}$ and $\{1,...,s\}$, respectively, and an i.i.d. random process with samples $(i,j)\in\{1,...,r\}\times\{1,...,s\}$ drawn with joint probability π_{ij}. In practice one often has to deal with incomplete information. For instance, observed instances often consist of several features plus class label, but some features may not be observed, i.e. if i is a class label and j is a feature, from the pair (i,j) only i is observed. We extend the contingency table n_{ij} to include $n_{i?}$, which counts the number of instances in which only the class i is observed (= number of $(i,?)$ instances). Similarly, $n_{?j}$ counts the number of $(?,j)$ instances, where the class label is missing. We make the common assumption that the missing-data mechanism is ignorable (missing at random and distinct) [LR87], i.e. the probability distribution of class labels i of instances with missing feature j is assumed to coincide with the marginal $\pi_{i+}:=\sum_j\pi_{ij}$. Similarly, given an instance with missing class label, the probability of the feature being j is assumed to be $\pi_{+j}:=\sum_i\pi_{ij}$.

Maximum likelihood estimate of $\boldsymbol{\pi}$. The likelihood of a specific data set $\boldsymbol{D}$ of size $N=n+n_{+?}+n_{?+}$ with contingency table $\boldsymbol{N}=\{n_{ij},n_{i?},n_{?j}\}$ given $\boldsymbol{\pi}$, hence, is $p(\boldsymbol{D}|\boldsymbol{\pi},n,n_{+?},n_{?+})=\prod_{ij}\pi_{ij}^{n_{ij}}\prod_i\pi_{i+}^{n_{i?}}\prod_j\pi_{+j}^{n_{?j}}$. Assuming a uniform $p(\boldsymbol{\pi})\sim 1\cdot\delta(\pi_{++}-1)$, Bayes' rule leads to the posterior[2]

$$p(\boldsymbol{\pi}|\boldsymbol{D})=p(\boldsymbol{\pi}|\boldsymbol{N})=\frac{1}{\mathcal{N}(\boldsymbol{N})}\prod_{ij}\pi_{ij}^{n_{ij}}\prod_i\pi_{i+}^{n_{i?}}\prod_j\pi_{+j}^{n_{?j}}\;\delta(\pi_{++}-1), \tag{1}$$

where the normalization $\mathcal{N}$ is chosen such that $\int p(\boldsymbol{\pi}|\boldsymbol{N})d\boldsymbol{\pi}=1$. With missing features and classes there is no exact closed form expression for $\mathcal{N}$.

In the following, we restrict ourselves to a discussion of leading-order (in N^{-1}) expressions, which are as accurate as one can specify one's prior knowledge [Hut02]. In leading order, the mean $E[\boldsymbol{\pi}]$ coincides with the mode of $p(\boldsymbol{\pi}|\boldsymbol{N})$ (=the maximum likelihood estimate) of $\boldsymbol{\pi}$. The log-likelihood function $\log p(\boldsymbol{\pi}|\boldsymbol{N})$ is

$$L(\pi|\boldsymbol{N})=\sum_{ij}n_{ij}\log\pi_{ij}+\sum_i n_{i?}\log\pi_{i+}+\sum_j n_{?j}\log\pi_{+j}-\log\mathcal{N}(\boldsymbol{N})-\lambda(\pi_{++}-1),$$

where we have replaced the δ function by a Lagrange multiplier λ to take into account the restriction $\pi_{++}=1$. The maximum is at $\frac{\partial L}{\partial\pi_{ij}}=\frac{n_{ij}}{\pi_{ij}}+\frac{n_{i?}}{\pi_{i+}}+\frac{n_{?j}}{\pi_{+j}}-\lambda=0$.

[1] The mathematical development is independent of the interpretation as class and feature, but it is convenient to use this terminology already here.

[2] Most (but not all) non-informative priors for $p(\boldsymbol{\pi})$ also lead to a Dirichlet posterior distribution (1) with interpretation $n_{ij}=n'_{ij}+n''_{ij}-1$, where n'_{ij} are the number of samples (i,j), and n''_{ij} comprises prior information (1 for the uniform prior, $\frac{1}{2}$ for Jeffreys' prior, 0 for Haldane's prior, $\frac{1}{rs}$ for Perks' prior, and other numbers in case of specific prior knowledge [GCSR95]). Furthermore, in leading order in $1/N$, any Dirichlet prior with $n''_{ij}=O(1)$ leads to the same results, hence we can simply assume a uniform prior. The reason for the $\delta(\pi_{++}-1)$ is that $\boldsymbol{\pi}$ must be constrained to the probability simplex $\pi_{++}:=\sum_{ij}\pi_{ij}=1$.

Multiplying this by π_{ij} and summing over i and j we obtain $\lambda = N$. The maximum likelihood estimate $\hat{\boldsymbol{\pi}}$ is, hence, given by

$$\hat{\pi}_{ij} = \frac{1}{N}\left(n_{ij} + n_{i?}\frac{\hat{\pi}_{ij}}{\hat{\pi}_{i+}} + n_{?j}\frac{\hat{\pi}_{ij}}{\hat{\pi}_{+j}}\right). \tag{2}$$

This is a non-linear equation in $\hat{\pi}_{ij}$, which, in general, has no closed form solution. Nevertheless (2) can be used to approximate $\hat{\pi}_{ij}$. Eq. (2) coincides with the popular expectation-maximization (EM) algorithm [CF74] if one inserts a first estimate $\hat{\pi}^0_{ij} = \frac{n_{ij}}{N}$ into the r.h.s. of (2) and then uses the resulting l.h.s. $\hat{\pi}^1_{ij}$ as a new estimate, etc.

Unimodality of $p(\boldsymbol{\pi}|\boldsymbol{N})$. The $rs \times rs$ Hessian matrix $\boldsymbol{H} \in \mathbb{R}^{rs\cdot rs}$ of $-L$ and the second derivative in direction of the rs dimensional column vector $\boldsymbol{v} \in \mathbb{R}^{rs}$ are

$$\boldsymbol{H}_{(ij)(kl)}[\boldsymbol{\pi}] \;:=\; -\frac{\partial L}{\partial \pi_{ij}\partial \pi_{kl}} = \frac{n_{ij}}{\pi_{ij}^2}\delta_{ik}\delta_{jl} + \frac{n_{i?}}{\pi_{i+}^2}\delta_{ik} + \frac{n_{?j}}{\pi_{+j}^2}\delta_{jl},$$

$$\boldsymbol{v}^T\boldsymbol{H}\boldsymbol{v} \;=\; \sum_{ijkl} v_{ij}\boldsymbol{H}_{(ij)(kl)}v_{kl} \;=\; \sum_{ij}\frac{n_{ij}}{\pi_{ij}^2}v_{ij}^2 + \sum_i \frac{n_{i?}}{\pi_{i+}^2}v_{i+}^2 + \sum_j \frac{n_{?j}}{\pi_{+j}^2}v_{+j}^2 \;\geq\; 0.$$

This shows that $-L$ is a convex function of $\boldsymbol{\pi}$, hence $p(\boldsymbol{\pi}|\boldsymbol{N})$ has a single (possibly degenerate) global maximum. L is strictly convex if $n_{ij} > 0$ for all ij, since $\boldsymbol{v}^T\boldsymbol{H}\boldsymbol{v} > 0$ $\forall\, \boldsymbol{v} \neq 0$ in this case[3]. This implies a unique global maximum, which is attained in the interior of the probability simplex. Since EM is known to converge to a local maximum, this shows, that in fact *EM always converges to the global maximum.*

Covariance of $\boldsymbol{\pi}$. With

$$\boldsymbol{A}_{(ij)(kl)} := \boldsymbol{H}_{(ij)(kl)}[\hat{\boldsymbol{\pi}}] = N\left[\frac{\delta_{ik}\delta_{jl}}{\rho_{ij}} + \frac{\delta_{ik}}{\rho_{i?}} + \frac{\delta_{jl}}{\rho_{?j}}\right],$$

$$\rho_{ij} := N\frac{\hat{\pi}_{ij}^2}{n_{ij}}, \quad \rho_{i?} := N\frac{\hat{\pi}_{i+}^2}{n_{i?}}, \quad \rho_{?j} := N\frac{\hat{\pi}_{+j}^2}{n_{?j}}. \tag{3}$$

and $\boldsymbol{\Delta} := \boldsymbol{\pi} - \hat{\boldsymbol{\pi}}$ we can represent the posterior to leading order as an $rs-1$ dimensional Gaussian:

$$p(\boldsymbol{\pi}|\boldsymbol{N}) \sim e^{-\frac{1}{2}\boldsymbol{\Delta}^T\boldsymbol{A}\boldsymbol{\Delta}}\delta(\Delta_{++}). \tag{4}$$

The easiest way to compute the covariance (and other quantities) is to also represent the δ-function as a narrow Gaussian of width $\varepsilon \approx 0$. Inserting $\delta(\Delta_{++}) \approx \frac{1}{\varepsilon\sqrt{2\pi}}\exp(-\frac{1}{2\varepsilon^2}\boldsymbol{\Delta}^T\boldsymbol{e}\boldsymbol{e}^T\boldsymbol{\Delta})$ into (4), where $\boldsymbol{e}_{ij} = 1$ for all ij (hence $\boldsymbol{e}^T\boldsymbol{\Delta} = \Delta_{++}$), leads to a full rs-dimensional Gaussian with kernel $\tilde{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{u}\boldsymbol{v}^T$, $\boldsymbol{u} = \boldsymbol{v} = \frac{1}{\varepsilon}\boldsymbol{e}$. The covariance of a Gaussian with kernel $\tilde{\boldsymbol{A}}$ is $\tilde{\boldsymbol{A}}^{-1}$. Using the Sherman-Morrison formula $\tilde{\boldsymbol{A}}^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\frac{\boldsymbol{u}\boldsymbol{v}^T}{1+\boldsymbol{v}^T\boldsymbol{A}^{-1}\boldsymbol{u}}\boldsymbol{A}^{-1}$ [PFTV92, p73] and $\varepsilon \to 0$ we get

$$\mathrm{Cov}_{(ij)(kl)}[\boldsymbol{\pi}] := E[\Delta_{ij}\Delta_{kl}] \simeq [\tilde{\boldsymbol{A}}^{-1}]_{(ij)(kl)} = \left[\boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1}\boldsymbol{e}\boldsymbol{e}^T\boldsymbol{A}^{-1}}{\boldsymbol{e}^T\boldsymbol{A}^{-1}\boldsymbol{e}}\right]_{(ij)(kl)}, \tag{5}$$

[3]Note that $n_{i?} > 0$ $\forall i$ is not sufficient, since $v_{i+} \equiv 0$ for $\boldsymbol{v} \neq 0$ is possible. Actually $v_{++} = 0$.

where $\simeq$ denotes $=$ up to terms of order N^{-2}. Singular $\boldsymbol{A}$ are easily avoided by choosing a prior such that $n_{ij}>0$ for all ij. $\boldsymbol{A}$ may be inverted exactly or iteratively, the latter by a trivial inversion of the diagonal part $\delta_{ik}\delta_{jl}/\rho_{ij}$ and by treating $\delta_{ik}/\rho_{i?}+\delta_{jl}/\rho_{?j}$ as a perturbation.

Missing features only, no missing classes. In the case of missing features only (no missing classes), i.e. for $n_{?j}=0$, closed form expressions for $\mathrm{Cov}[\boldsymbol{\pi}]$ can be obtained. If we sum (2) over j we get $\hat{\pi}_{i+}=\frac{N_{i+}}{N}$ with $N_{i+}:=n_{i+}+n_{i?}$. Inserting $\hat{\pi}_{i+}=\frac{N_{i+}}{N}$ into the r.h.s. of (2) and solving w.r.t. $\hat{\pi}_{ij}$ we get the explicit expression

$$\hat{\pi}_{ij} = \frac{N_{i+}}{N}\frac{n_{ij}}{n_{i+}}. \tag{6}$$

Furthermore, it can easily be verified (by multiplication) that $\boldsymbol{A}_{(ij)(kl)}=N[\delta_{ik}\delta_{jl}/\rho_{ij}+\delta_{ik}/\rho_{i?}]$ has inverse $[\boldsymbol{A}^{-1}]_{(ij)(kl)}=\frac{1}{N}[\rho_{ij}\delta_{ik}\delta_{jl}-\frac{\rho_{ij}\rho_{kl}}{\rho_{i+}+\rho_{i?}}\delta_{ik}]$. With the abbreviations

$$\tilde{Q}_{i?} := \frac{\rho_{i?}}{\rho_{i?}+\rho_{i+}} \quad \text{and} \quad \tilde{Q} := \sum_i \rho_{i+}\tilde{Q}_{i?} \tag{7}$$

we get $[\boldsymbol{A}^{-1}\boldsymbol{e}]_{ij}=\sum_{kl}[\boldsymbol{A}^{-1}]_{(ij)(kl)}=\frac{1}{N}\rho_{ij}\tilde{Q}_{i?}$ and $\boldsymbol{e}^T\boldsymbol{A}^{-1}\boldsymbol{e}=\tilde{Q}/N$. Inserting everything into (5) we get

$$\mathrm{Cov}_{(ij)(kl)}[\boldsymbol{\pi}] \;\simeq\; \frac{1}{N}\left[\rho_{ij}\delta_{ik}\delta_{jl} - \frac{\rho_{ij}\rho_{kl}}{\rho_{i+}+\rho_{i?}}\delta_{ik} - \frac{\rho_{ij}\tilde{Q}_{i?}\rho_{kl}\tilde{Q}_{k?}}{\tilde{Q}}\right]. \tag{8}$$

Expressions for the general case. The contribution from unlabeled classes can be interpreted as a rank s modification of $\boldsymbol{A}$ in the case of no missing classes. One can use Woodbury's formula $[\boldsymbol{B}+\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T]^{-1}=\boldsymbol{B}^{-1}-\boldsymbol{B}^{-1}\boldsymbol{U}[\boldsymbol{D}^{-1}+\boldsymbol{V}^T\boldsymbol{B}^{-1}\boldsymbol{U}]^{-1}\boldsymbol{V}^T\boldsymbol{B}^{-1}$ [PFTV92, p75] with $\boldsymbol{B}_{(ij)(kl)}=\delta_{ik}\delta_{jl}/\rho_{ij}+\delta_{ik}/\rho_{i?}$, $\boldsymbol{D}_{jl}=\delta_{jl}/\rho_{?j}$, and $\boldsymbol{U}_{(ij)l}=\boldsymbol{V}_{(ij)l}=\delta_{jl}$ to reduce the inversion of the $rs\times rs$ matrix $\boldsymbol{A}$ to the inversion of only a *single* s-dimensional matrix. The result (which may be inserted into (5)) can be written in the form

$$[\boldsymbol{A}^{-1}]_{(ij)(kl)} = \frac{1}{N}\left[F_{ijl}\delta_{ik} - \sum_{mn} F_{ijm}[\boldsymbol{G}^{-1}]_{mn}F_{kln}\right], \tag{9}$$

$$F_{ijl} := \rho_{ij}\delta_{jl} - \frac{\rho_{ij}\rho_{kl}}{\rho_{i?}+\rho_{i+}}, \qquad G_{mn} := \rho_{?n}\delta_{mn} + F_{+mn}.$$

3 Distribution of Mutual Information

Mutual information I. An important measure of the stochastic dependence of $\imath$ and $\jmath$ is the mutual information

$$I(\boldsymbol{\pi}) \;=\; \sum_{i=1}^{r}\sum_{j=1}^{s}\pi_{ij}\log\frac{\pi_{ij}}{\pi_{i+}\pi_{+j}} \;=\; \sum_{ij}\pi_{ij}\log\pi_{ij} - \sum_i \pi_{i+}\log\pi_{i+} - \sum_j \pi_{+j}\log\pi_{+j}.$$

The point estimate for I is $I(\hat{\boldsymbol{\pi}})$. In the Bayesian approach one takes the posterior (1) from which the posterior probability density of the mutual information can, in principle, be computed:[4]

$$p(I|\boldsymbol{N}) = \int \delta(I(\boldsymbol{\pi}) - I) p(\boldsymbol{\pi}|\boldsymbol{N}) d^{rs}\boldsymbol{\pi}. \tag{10}$$

[5]The $\delta(\cdot)$ distribution restricts the integral to $\boldsymbol{\pi}$ for which $I(\boldsymbol{\pi}) = I$. For large sample size, $N \to \infty$, $p(\boldsymbol{\pi}|\boldsymbol{N})$ is strongly peaked around the mode $\boldsymbol{\pi} = \hat{\boldsymbol{\pi}}$ and $p(I|\boldsymbol{N})$ gets strongly peaked around the frequency estimate $I = I(\hat{\boldsymbol{\pi}})$. The (central) moments of I are of special interest. The mean

$$E[I] = \int_0^\infty I \cdot p(I|\boldsymbol{N})\, dI = \int I(\boldsymbol{\pi}) p(\boldsymbol{\pi}|\boldsymbol{N}) d^{rs}\boldsymbol{\pi} \;=\; I(\hat{\boldsymbol{\pi}}) + O(N^{-1}) \tag{11}$$

coincides in leading order with the point estimate, where $\hat{\boldsymbol{\pi}}$ has been computed in Section 2. Together with the variance $\mathrm{Var}[I] = E[(I - E[I])^2] = E[I^2] - E[I]^2$ (computed below) we can approximate (10) by a Gaussian[6]

$$p(I|\boldsymbol{N}) \sim \exp\left(-\tfrac{(I - I(\hat{\boldsymbol{\pi}})^2}{2\mathrm{Var}[I]}\right) \sim \exp\left(-\tfrac{(I - E[I])^2}{2\mathrm{Var}[I]}\right) \tag{12}$$

In a previous work we derived higher order central moments (skewness and kurtosis) and higher order (in N^{-1}) approximations in the case of complete data [Hut02].

Variance of I. The leading order variance of the mutual information $I(\boldsymbol{\pi})$ has been related[7] in [Hut02] to the covariance of $\boldsymbol{\pi}$:

$$\mathrm{Var}[I] \;\simeq\; \sum_{ijkl} \log \frac{\hat{\pi}_{ij}}{\hat{\pi}_{i+}\hat{\pi}_{+j}} \log \frac{\hat{\pi}_{kl}}{\hat{\pi}_{k+}\hat{\pi}_{+l}} \mathrm{Cov}_{(ij)(kl)}[\boldsymbol{\pi}] \tag{13}$$

Inserting (8) for the covariance into (13) we get for the variance of the mutual information in leading order in $1/N$ in the case of missing features only, the following expression:

$$\mathrm{Var}[I] \;\simeq\; \frac{1}{N}[\tilde{K} - \tilde{J}^2/\tilde{Q} - \tilde{P}], \qquad \tilde{K} := \sum_{ij} \rho_{ij} \left(\log \frac{\hat{\pi}_{ij}}{\hat{\pi}_{i+}\hat{\pi}_{+j}}\right)^2, \tag{14}$$

---

[4]$I(\boldsymbol{\pi})$ denotes the mutual information for the specific chances $\boldsymbol{\pi}$, whereas I in the context above is just some non-negative real number. I will also denote the mutual information *random variable* in the expectation $E[I]$ and variance $\mathrm{Var}[I]$. Expectations are *always* w.r.t. to the posterior distribution $p(\boldsymbol{\pi}|\boldsymbol{N})$.

[5]Since $0 \le I(\boldsymbol{\pi}) \le I_{max}$ with sharp upper bound $I_{max} = \min\{\log r, \log s\}$, the domain of $p(I|\boldsymbol{n})$ is $[0, I_{max}]$, and integrals over I may be restricted to $\int_0^{I_{max}}$.

[6]For $I(\hat{\boldsymbol{\pi}}) \neq 0$ the central limit theorem ensures convergence of $p(I|\boldsymbol{N})$ to a Gaussian. Using a Beta distribution instead of (12), which also converges to a Gaussian, has slight advantages over (12) [ZH02].

[7]$\hat{\boldsymbol{\pi}}$ was defined in [Hut02] as the mean $E[\boldsymbol{\pi}]$ whereas $\hat{\boldsymbol{\pi}}$ has been defined in this work as the ML estimate. Furthermore the Dirichlet priors differ. Since to leading order both definitions of $\boldsymbol{\pi}$ coincide, the prior does not matter, and the expression is also valid for incomplete data case, the use of (13) in this work is permitted.

$$\tilde{P} := \sum_i \frac{\tilde{J}_{i+}^2 Q_{i?}}{\rho_{i?}}, \qquad \tilde{J} := \sum_i \tilde{J}_{i+} \tilde{Q}_{i?}, \qquad \tilde{J}_{i+} := \sum_j \rho_{ij} \log \frac{\hat{\pi}_{ij}}{\hat{\pi}_{i+}\hat{\pi}_{+j}}.$$

A closed form expression for $\mathcal{N}(\boldsymbol{N})$ also exists. Symmetric expressions for missing classes only (no missing features) can be obtained. Note that for the complete case $n_{?j} = n_{i?} \equiv 0$, we have $\hat{\pi}_{ij} = \rho_{ij} = \frac{n_{ij}}{n}$, $\rho_{i?} = \infty$, $\tilde{Q}_{i?} = 1$, $\tilde{J} = J$, $\tilde{K} = K$, and $\tilde{P} = 0$, consistent with [Hut02] (where J and K are defined and the accuracy is discussed).

There is at least one reason for minutely having inserted all expressions into each other and introducing quite a number definitions. In the so presented form all expressions involve at most a double sum. Hence, the overall computation time of the mean and variance is $O(rs)$ in the case of missing features only.

Expression for the general case. The result for the covariance (5) can be inserted into (13) to obtain the variance of the mutual information to leading order.

$$\mathrm{Var}[I] \;\simeq\; \boldsymbol{l}^T \boldsymbol{A}^{-1} \boldsymbol{l} - (\boldsymbol{l}^T \boldsymbol{A}^{-1} \boldsymbol{e})^2 / (\boldsymbol{e}^T \boldsymbol{A}^{-1} \boldsymbol{e}) \quad \text{where} \quad \boldsymbol{l}_{ij} = \log \frac{\hat{\pi}_{ij}}{\hat{\pi}_{i+}\hat{\pi}_{+j}}$$

Inserting (9) and rearranging terms appropriately we can compute $\mathrm{Var}[I]$ in time $O(rs)$ plus the time $O(s^2 r)$ to compute the $s \times s$ matrix $\boldsymbol{G}$ and time $O(s^3)$ to invert it, plus the time $O(\#rs)$ for determining $\hat{\pi}_{ij}$, where $\#$ is the number of iterations of EM. Of course, one can and should always choose $s \leq r$. Note that these expressions converge for $N \to \infty$ to the exact values. The fraction of data with missing feature or class needs not to be small.

In the following we apply the obtained results to feature selection for incomplete data. Since we only used labeled data we could use (11) with (6), and (14) with (7) and (3).

4 Feature Selection

Feature selection is a basic step in the process of building classifiers [BL97]. We consider the well-known filter (F) that computes the empirical mutual information $I(\hat{\boldsymbol{\pi}})$ between features and the class, and discards features with $I(\hat{\boldsymbol{\pi}}) < \varepsilon$ for some threshold ε [Lew92]. This is an easy and effective approach that has gained popularity with time.

We compare F to the two filters introduced in [ZH02] for the case of complete data, and extended here to the more general case. The *backward filter* (BF) *discards* a feature if its value of mutual information with the class is less than or equal to ε with high probability $\bar{p}$ (discard if $p(I \leq \varepsilon | \boldsymbol{N}) \geq \bar{p}$). The *forward filter* (FF) *includes* a feature if the mutual information is greater than ε with high probability $\bar{p}$ (include if $p(I > \varepsilon | \boldsymbol{N}) \geq \bar{p}$). BF is a conservative filter, because it will only discard features after observing substantial evidence supporting their irrelevance. FF instead will

Table 1: *Incomplete data sets used for the experiments, together with their number of features, instances, missing values, and the relative frequency of the majority class. The data sets are available from the UCI repository of machine learning data sets [MA95]. Average number of features selected by the filters on the entire data set are reported in the last three columns. FF always selected fewer features than F; F almost always selected fewer features than BF. Prediction accuracies where significantly different only for the Hypothyroidloss data set.*

Name	#feat.	#inst.	#m.v.	maj.class	FF	F	BF
Audiology	69	226	317	0.212	64.3	68.0	68.7
Crx	15	690	67	0.555	9.7	12.6	13.8
Horse-colic	18	368	1281	0.630	11.8	16.1	17.4
Hypothyroidloss	23	3163	1980	0.952	4.3	8.3	13.2
Soybean-large	35	683	2337	0.135	34.2	35	35

tend to use fewer features, i.e. only those for which there is substantial evidence about them being useful in predicting the class.

For the subsequent classification task we use the naive Bayes classifier [DH73], which is often a good classification model. Despite its simplifying assumptions (see [DP97]), it often competes successfully with much more complex classifiers, such as C4.5 [Qui93]. Our experiments focus on the incremental use of the naive Bayes classifier, a natural learning process when the data are available sequentially: the data set is read instance by instance; each time, the chosen filter selects a subset of features that the naive Bayes uses to classify the new instance; the naive Bayes then updates its knowledge by taking into consideration the new instance and its actual class. Note that for increasing sizes of the learning set the filters converge to the same behavior, since the variance of I tends to zero (see [ZH02] for details).

For each filter, we are interested in experimentally evaluating two quantities: for each instance of the data set, the average number of correct predictions (namely, the prediction accuracy) of the naive Bayes classifier up to such instance; and the average number of features used. By these quantities we can compare the filters and judge their effectiveness.

The implementation details for the following experiments include: using the Gaussian approximation (12) to the distribution of mutual information with the mean (11) using (6), and the variance (14) using (7) and (3); using natural logarithms everywhere; and setting the level $\bar{p}$ of the posterior probability to 0.95, and the threshold ε to 0.003 as discussed in [ZH02].

5 Experimental Analysis

Table 1 lists five data sets together with the experimental results. These are real data sets on a number of different domains. The data sets presenting non-nominal features have been pre-discretized by MLC++ [KJL$^+$94], default options (i.e., the

common entropy based discretization). This step may remove some features judging them as irrelevant, so the number of features in the table refers to the data sets after the possible discretization. The instances have been randomly sorted before starting the experiments.

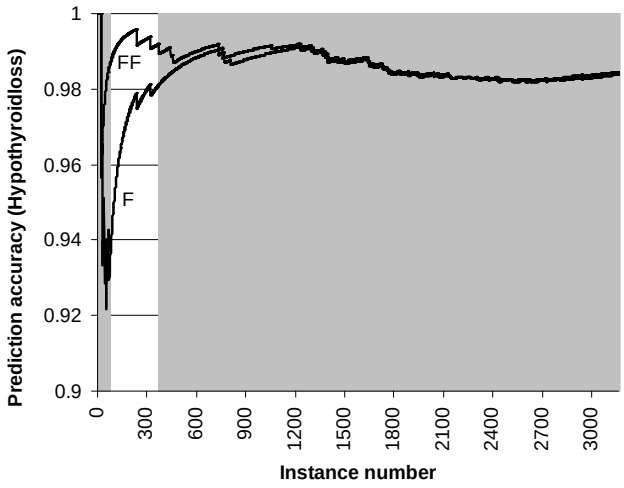

Figure 1: *Prediction accuracies of the naive Bayes with filters F and FF on the Hypothyroidloss data set. BF is not reported because there is no significant difference with the F curve. The differences between F and FF are significant in the range of observations 71–374 (white area). The maximum difference is achieved at observation 71, where the accuracies are 0.986 (FF) vs. 0.930 (F).*

The last three columns of Table 1 show that FF selects lower (i.e. better) number of features than the commonly used filter F, which in turn, selects lower number of features than the filter BF. We used the *two-tails paired t test* at level 0.05 to compare the prediction accuracies of the naive Bayes with different filters, in the first k instances of the data set, for each k. On four data sets out of five, both the differences between FF and F, and the differences between F and BF, were never statistically significant, despite the different number of used features, as indicated in Table 1. The reduction can be very pronounced, as for the Hypothyroidloss data set. This is also the only data set for which the prediction accuracies of F and FF are significantly different, in favor of the latter. This is displayed in Figure 1. Similar (even stronger) results have been found for 10 complete data sets analyzed in [ZH02].

The most prominent evidence from the experiments is the better performance of FF versus the traditional filter F. In the following we look at FF from another perspective to exemplify and explain its behavior. FF includes a feature if $p(I > \varepsilon|\boldsymbol{n}) \geq \bar{p}$, according to its definition. Let us assume that FF is realized by means of the Gaussian (as in the experiments above), and let us choose $\bar{p} \approx 0.977$. The condition $p(I > \varepsilon|\boldsymbol{n}) \geq \bar{p}$ becomes $\varepsilon \leq E[I] - 2 \cdot \sqrt{\mathrm{Var}[I]}$, or, in an approximate way,

$I(\hat{\boldsymbol{\pi}}) \geq \varepsilon + 2 \cdot \sqrt{\mathrm{Var}[I]}$, given that $I(\hat{\boldsymbol{\pi}})$ is the first-order approximation of $E[I]$ (cf. (11)). We can regard $\varepsilon + 2 \cdot \sqrt{\mathrm{Var}[I]}$ as a new threshold ε'. Under this interpretation, we see that FF is approximately equal to using the filter F with the bigger threshold ε'. This interpretation makes it also clearer why FF can be better suited than F for sequential learning tasks. In sequential learning, $\mathrm{Var}[I]$ decreases as new units are read; this makes ε' to be a self-adapting threshold that adjusts the level of caution (in including features) as more units are read. In the limit, ε' is equal to ε. This characteristic of self-adaptation, which is absent in F, seems to be decisive to the success of FF.

6 Conclusions

We addressed the problem of the reliability of empirical estimates for the chances $\boldsymbol{\pi}$ and the mutual information I in the case of incomplete discrete data. We used the Bayesian framework to derive reliable and quickly computable approximations for the mean, mode and the (co)variance of $\boldsymbol{\pi}$ and $I(\boldsymbol{\pi})$ under the posterior distribution $p(\boldsymbol{\pi}|D)$. We showed that $p(\boldsymbol{\pi}|\boldsymbol{D})$ is unimodal, which implies that EM always converges to the global maximum. The results allowed us to efficiently determine credible intervals for I with incomplete data. Applications are manifold, e.g. to robustly infer classification trees or Bayesian networks. As far as feature selection is concerned, we empirically showed that the forward filter, which includes a feature if the mutual information is greater than ε with high probability, outperforms the popular filter based on empirical mutual information in sequential learning tasks. This result for incomplete data is obtained jointly with the naive Bayes classifier. More broadly speaking, obtaining the distribution of mutual information when data are incomplete may form a basis on which reliable and effective uncertain models can be developed.

References

[BL97] A. L. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1–2):245–271, 1997. Special issue on relevance.

[CF74] T. T. Chen and S. E. Fienberg. Two-dimensional contingency tables with both completely and partially cross-classified data. *Biometrics*, 32:133–144, 1974.

[CHH+02] J. Cheng, C. Hatzis, H. Hayashi, M. Krogel, S. Morishita, D. Page, and J. Sese. KDD cup 2001 report. *ACM SIGKDD Explorations*, 3(2), 2002.

[DH73] R. O. Duda and P. E. Hart. *Pattern classification and scene analysis*. Wiley, New York, 1973.

[DHS01] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. Wiley, 2001. 2nd edition.

[DP97] P. Domingos and M. Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29(2/3):103–130, 1997.

[GCSR95] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis.* Chapman, 1995.

[Hut02] M. Hutter. Distribution of mutual information. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 399–406, Cambridge, MA, 2002. MIT Press.

[KJL+94] R. Kohavi, G. John, R. Long, D. Manley, and K. Pfleger. MLC++: a machine learning library in C++. In *Tools with Artificial Intelligence*, pages 740–743. IEEE Computer Society Press, 1994.

[Kle99] G. D. Kleiter. The posterior probability of Bayes nets with strong dependences. *Soft Computing*, 3:162–173, 1999.

[Lew92] D. D. Lewis. Feature selection and feature extraction for text categorization. In *Proc. of Speech and Natural Language Workshop*, pages 212–217, San Francisco, 1992. Morgan Kaufmann.

[LR87] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data.* John-Wiley, New York, 1987.

[MA95] P. M. Murphy and D. W. Aha. UCI repository of machine learning databases, 1995. http://www.sgi.com/Technology/mlc/db/.

[PFTV92] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge University Press, Cambridge, second edition, 1992.

[Qui93] J. R. Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann, San Mateo, 1993.

[ZH02] M. Zaffalon and M. Hutter. Robust feature selection by mutual information distributions. In A. Darwiche and N. Friedman, editors, *Proceedings of the 18th International Conference on Uncertainty in Artificial Intelligence (UAI-2002)*, pages 577–584, San Francisco, CA., 2002. Morgan Kaufmann.