Using Prior Information in Bayesian Inference – with Application to Fault Diagnosis

Anna Pernestål and Mattias Nyberg

Dept. of Electrical Engineering, Linköping University, Linköping, Sweden
{annap, matny}@isy.liu.se

Abstract. In this paper we consider Bayesian inference using training data combined with prior information. The prior information considered is response and causality information which gives constraints on the posterior distribution. It is shown how these constraints can be expressed in terms of the prior probability distribution, and how to perform the computations. Further, it is discussed how this prior information improves the inference.

Keywords: Bayesian Classification, Prior Information, Bayesian Inference, Fault Classification

INTRODUCTION

In this paper we study the problem of making inference about a state, given an observed featrue vector. Traditionally, inference methods rely either on prior information only or on training data consisting of simultaneous observations of the class and the feature vector [1], [2], [3]. However, in many inference problems there are both training data and prior information available. Inspired by the problem of fault diagnosis, where the feature vector typically is a set of diagnostic tests, and the states are the possible faults, we recognize two types of prior information. First, there may be information that some values of the features are impossible under certain states. In the present paper this information is referred to as *response information*, which for example can be that it is known that a test never alarms when there is no fault present. Second, it may be known that certain elements of the feature vector are equally distributed under several states, here referred to as *causality information*. In the fault diagnosis context this means that a diagnostic test is not affected by a certain fault.

The type of prior information studied in the present work typically appears in previous works on fault diagnosis. The response information is used for example in [4], [5], and [6]. The causality information is an interpretation of the Fault Signature Matrix (FSM) used for example in [7] and [8]. The main difference between these previous works and the present is that here we combine the prior information with training data instead of relying on prior information only.

To compute this posterior probability for the states in the case of training data only is, although previously well studied, a nontrivial problem, see e.g. [9], [10], and [11]. In these previous works the computations are based on training data only. In the present work we go one step further, and discuss how the prior information in terms of response and causality information can be integrated into the Bayesian framework.

INFERENCE USING TRAINING DATA

We begin by introducing the notation used, and summarizing previous results on inference using training data alone. Let $\mathbf{Z} = (\mathbf{X}, C)$ be a discrete variable, where the feature vector $\mathbf{X} = (X_1, \ldots X_R)$ is R-dimensional and the state variable C is scalar. The variables $\mathbf{X}$ and C can take K and L different values respectively, and hence $\mathbf{Z}$ can take $M = KL$ values. Use $\mathbf{z} = (\mathbf{x}, c) = ((x_1, \ldots, x_R), c)$ to denote a sample of $\mathbf{Z}$. Let $\mathbb{X}$, $\mathbb{X}_i$, $\mathbb{C}$, and $\mathbb{Z} = \mathbb{C} \times \mathbb{X}$ be the domains of $\mathbf{X}$, X_i, C and $\mathbf{Z}$ respectively. Enumerate the elements in $\mathbb{Z}$, and use $\zeta_i, i = 1, \ldots, M$, to denote the ith element. We use $p(\mathbf{X} = \mathbf{x}|\mathbf{I})$, or simply $p(\mathbf{x}|\mathbf{I})$, to denote the discrete probability distribution for $\mathbf{X}$ given the current state of knowledge $\mathbf{I}$. For continuous probability density functions we use $f(\mathbf{x}|\mathbf{I})$.

Let $\mathscr{D}$ be the training data, i.e. a set of simultaneous samples of the feature vector and the state variable. In the inference problem, the probability distribution $p(c|\mathbf{X} = \mathbf{x}, \mathscr{D}, \mathbf{I})$ is to be determined. Note that for a given feature vector $\mathbf{x}$, the posterior probability for a state is proportional to the joint distribution of c and $\mathbf{x}$, $p(c|\mathbf{x}, \mathscr{D}, \mathbf{I}) = p(c, \mathbf{x}|\mathscr{D}, \mathbf{I})/p(\mathbf{x}|\mathscr{D}, \mathbf{I}) \propto p(c, \mathbf{x}|\mathscr{D}, \mathbf{I}) = p(\mathbf{z}|\mathscr{D}, \mathbf{I})$. Therefore we can study the probability distribution $p(\mathbf{z}|\mathscr{D}, \mathbf{I})$. The computations of $p(\mathbf{z}|\mathscr{D}, \mathbf{I})$ are, under certain assumptions, given in detail for example in [9], [10], and [11]. In these references the arguments for the underlying assumptions are also discussed. Here we summarize them in the following theorem.

Theorem 1 *Let $p(\mathbf{z}|\mathscr{D}, \mathbf{I})$ be discrete, and assume that there are parameters $\Theta = (\theta_1, \ldots, \theta_M)^T$ such that*

$$p(\mathbf{Z} = \zeta_i|\Theta, \mathbf{I}) = \theta_i, \quad i = 1, \ldots, M, \tag{1a}$$

$$\theta_i > 0, \quad \sum_{\zeta_i \in \mathbb{Z}} \theta_i = 1. \tag{1b}$$

Assume that $f(\Theta|\mathbf{I})$ is Dirichlet distributed,

$$f(\Theta|\mathbf{I}) = \frac{\Gamma(\sum_{i=1}^{M} \alpha_i)}{\prod_{i=1}^{M} \Gamma(\alpha_i)} \prod_{i=1}^{M} \theta_i^{\alpha_i - 1}, \quad \alpha_i > 0, \tag{2}$$

where $\Gamma(\cdot)$ is the gamma function, i.e. fulfills $\Gamma(n+1) = n\Gamma(n)$ and $\Gamma(1) = 1$ and the parameters $\alpha = (\alpha_1, \ldots, \alpha_M)$ are given. Assume that the samples in the training data are independent, and let n_i be the count of samples in $\mathscr{D}$ where $\mathbf{Z} = \zeta_i$, and let $N = \sum_{i=1}^{M} n_i$ and $A = \sum_{i=1}^{M} \alpha_i$. Then it holds that

$$p(\mathbf{Z} = \zeta_i|\mathscr{D}, \mathbf{I}) = \frac{n_i + \alpha_i}{N + A}. \tag{3}$$

In the following sections we will now discuss how the results from Theorem 1 can be extended to take the response and causality information into account.

INFERENCE USING RESPONSE INFORMATION

Consider the case where some values of the feature vector are known to be impossible in certain states of the system. We refer to this kind of information as *response information*.

TABLE 1. Example of response information, where "•" means that the value of the feature is possible.

	$C=c_1$	$C=c_2$	$C=c_3$
$x_1=0$	•	•	•
$x_1=1$		•	•
$x_1=2$		•	

Formally, it means that there are sets $\gamma_{i,c} \subset \mathbb{X}_i$ representing "forbidden values" under state c, i.e.

$$p(x_i|c,\mathscr{D},\mathbf{I}_{\mathscr{R}}) = 0, \text{ for } x_i \in \gamma_{i,c},$$

where we have used $\mathbf{I}_{\mathscr{R}}$ to denote that $\mathbf{I}$ includes response information.

To exemplify how the sets $\gamma_{i,c}$ can be determined, consider the following example with a three-valued feature X_1 with domain $\mathbb{X}_1 = \{0,1,2\}$. Assume that the information is given that in state c_1, the feature X_1 can only take the value 0. In state c_2 all values are possible, while in state c_3 all values except 2 are possible. This information is summarized in Table 1, where "•" means that the value of the feature is possible. This information gives the sets $\gamma_{1,c_1} = \{1,2\}$, $\gamma_{1,c_2} = \{\emptyset\}$, $\gamma_{1,c_3} = \{2\}$.

Let $\gamma \subset \mathbb{Z}$ be the set of values such that if $x_i \in \gamma_{i,c}$, then $\mathbf{z} \in \gamma$. In our example we have $\gamma = \{(1,c_1),(2,c_1),(2,c_3)\}$. Assume that $p(\mathbf{z}|\Theta,\mathbf{I}_{\mathscr{R}})$ is parameterized by Θ as in (1a). By $\mathbf{I}_{\mathscr{R}}$ we have the following requirements on the parameters

$$\theta_i = 0 \quad \forall \zeta_i \in \gamma, \qquad \theta_i > 0 \quad \forall \zeta_i \in \mathbb{Z} \setminus \gamma, \qquad \sum_{\zeta_i \in \mathbb{Z}\setminus\gamma} \theta_i = 1. \tag{4}$$

We can now state the following theorem for the joint probability distribution when response information is available.

Theorem 2 *Assume that* $p(\mathbf{Z}|\Theta,\mathbf{I}_{\mathscr{R}})$ *is discrete and given by* (1a) *and* (4)*. Further, assume that* $f(\Theta|\mathbf{I}_{\mathscr{R}})$ *is Dirichlet distributed over the set* $\mathbb{Z}\setminus\gamma$*,*

$$f(\Theta|\mathbf{I}_{\mathscr{R}}) = \begin{cases} \frac{\Gamma(\sum_{\zeta_i\in\mathbb{Z}\setminus\gamma}\alpha_i)}{\prod_{\zeta_i\in\mathbb{Z}\setminus\gamma}\Gamma(\alpha_i)} \prod_{\zeta_i\in\mathbb{Z}\setminus\gamma} \theta_i^{\alpha_i-1}, \ \alpha_i > 0 & \text{if } \Theta \in \Omega_{\mathscr{R}} \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

Assume that the samples in the training data $\mathscr{D}$ *are idependent. Let* n_i *be the count of samples in* $\mathscr{D}$ *where* $\mathbf{Z} = \zeta_i$*, and let* $N = \sum_{i=1}^M n_i$ *and* $A = \sum_{i=1}^M \alpha_i$*. Then it holds that*

$$p(\mathbf{Z} = \zeta_i|\mathscr{D},\mathbf{I}_{\mathscr{R}}) = \begin{cases} 0, & \text{if } \mathbf{z} \in \gamma \\ \frac{n_i+\alpha_i}{N+A} & \text{otherwise.} \end{cases} \tag{6}$$

Proof: Apply Theorem 1 when $\mathbf{z} \in \mathbb{Z}\setminus\gamma$, and use that (5) gives probability 0 for all $\mathbf{z} \in \gamma_c$. A complete proof is given in [12]. □

INFERENCE USING CAUSALITY INFORMATION

Let us now turn to the case when there is information available that a certain feature is equally distributed in two states. We call this kind of information *causality information*.

In this section we show how this information can be integrated in the problem formulation, and we also discuss a method for solving the problem.

Computing the Posterior Using Causality Information

The causality information is formally represented by

$$p(x_i|c_j,\Theta,\mathbf{I}_{\mathscr{C}}) = p(x_i|c_k,\Theta,\mathbf{I}_{\mathscr{C}}), \tag{7}$$

where $\mathbf{I}_{\mathscr{C}}$ is used to denote that causality information is given by in the state of knowledge. Applying the product rule of probabilities on (7) we have

$$\frac{p(x_i,c_j|\Theta,\mathbf{I}_{\mathscr{C}})}{p(c_j|\mathbf{I}_{\mathscr{C}})} = p(x_i|c_j,\Theta,\mathbf{I}_{\mathscr{C}}) = p(x_i|c_k,\Theta,\mathbf{I}_{\mathscr{C}}) = \frac{p(x_i,c_k|\Theta,\mathbf{I}_{\mathscr{C}})}{p(c_k|\mathbf{I}_{\mathscr{C}})},$$

where $p(c_j|\mathbf{I}_{\mathscr{C}})$ and $p(c_k|\mathbf{I}_{\mathscr{C}})$ are the prior probabilities for the states c_j and c_k, and are assumed to be given by the background information $\mathbf{I}_{\mathscr{C}}$. The prior probabilities are known proportionality constants, and we can write $p(c_j|\mathbf{I}_{\mathscr{C}}) = \rho_{jk}p(c_k|\mathbf{I}_{\mathscr{C}})$ for a known constant ρ_{jk}. Thus, (7) means that $p(c_j,x_i|\Theta,\mathbf{I}_{\mathscr{C}}) = \rho_{jk}p(c_k,x_i|\Theta,\mathbf{I}_{\mathscr{C}})$. We have that

$$p(c_j,\xi_i|\Theta,\mathbf{I}_{\mathscr{C}}) = \sum_{\zeta_l\in\mathbb{Z}_{\xi_i,c_j}} p(\zeta_l|\Theta,\mathbf{I}_{\mathscr{C}}) = \sum_{\zeta_l\in\mathbb{Z}_{\xi_i,c_j}} \theta_l, \tag{8}$$

where $\mathbb{Z}_{\xi_i,c_j} = \{\zeta_l \in \mathbb{Z} : \zeta_l = ((x_1,\ldots,\xi_i,\ldots,x_R),c_j)\}$, i.e. the set of all possible values ζ_l of $\mathbf{Z}$ in which $x_i = \xi_i$ and $c = c_j$. Equations (7) and (8) give requirements in the form

$$\sum_{\zeta_l\in\mathbb{Z}_{\xi_i,c_j}} \theta_l = \rho_{jk} \sum_{\zeta_l\in\mathbb{Z}_{\xi_i,c_k}} \theta_l. \tag{9}$$

To exemplify, consider the following case with two states, $C \in \{c_1,c_2\}$, and one feature $\mathbf{X} \in \{0,1\}$. Define $\Theta = (\theta_1,\theta_2,\theta_3,\theta_4)$ by

$$p(\mathbf{X}=0,C=c_1|\Theta,\mathbf{I}) = \theta_1,\ p(\mathbf{X}=0,C=c_2|\Theta,\mathbf{I}) = \theta_2, \tag{10a}$$

$$p(\mathbf{X}=1,C=c_1|\Theta,\mathbf{I}) = \theta_3,\ p(\mathbf{X}=1,C=c_2|\Theta,\mathbf{I}) = \theta_4. \tag{10b}$$

Assume that the causality information $p(\mathbf{X},C=c_1|\mathbf{I}_{\mathscr{C}}) = p(\mathbf{X},C=c_2|\mathbf{I}_{\mathscr{C}})$ is given. Expressed in terms of the parameters this means that $\theta_1 = \rho_{12}\theta_2$ and $\theta_3 = \rho_{12}\theta_4$.

Let $L \geq 0$ be the number of constraints in the form (7) given by the causality information. Each constraint gives one equation in Θ for each possible value of the feature considered in the constraint. Let K_i be the number of possible values of the feature considered in the i:th constraint. Furthermore, Θ should fulfill the requirement (1b). All in all, there are $1+\sum_{i=1}^{L} K_i = l$ equations that Θ should fulfill. In matrix form we write

$$E\Theta = F, \tag{11}$$

where $E \in \mathbb{R}^{l\times M}$ and $F \in \mathbb{R}^l$. Note that (1b) requires that one row in E consists of ones only, and that the corresponding row in F is also a one. In the example with parameters as in (10), and with $\rho_{12} = 1$, the matrices becomes

$$E = \begin{bmatrix} 0 & 0 & -1 & 1 \\ 1 & -1 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \quad F = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}. \tag{12}$$

To compute $p(\mathbf{Z}|\mathscr{D},\mathbf{I}_{\mathscr{C}})$ marginalize over the set of parameters Ω that fulfill (1)

$$p(\mathbf{Z}|\mathscr{D},\mathbf{I}_{\mathscr{C}}) = \int_{\Omega} p(\mathbf{Z}|\Theta,\mathscr{D},\mathbf{I}_{\mathscr{C}}) f(\Theta|\mathscr{D},\mathbf{I}_{\mathscr{C}}) d\Theta. \tag{13}$$

The first factor in the integral (13) is independent of $\mathscr{D}$ since Θ is known. Thus, we have $p(\mathbf{Z}|\Theta,\mathscr{D},\mathbf{I}_{\mathscr{C}}) = p(\mathbf{Z}|\Theta,\mathbf{I}_{\mathscr{C}})$, which is given by (1). To determine the second factor in the integral (13), apply Bayes' theorem

$$f(\Theta|\mathscr{D},\mathbf{I}_{\mathscr{C}}) = \frac{p(\mathscr{D}|\Theta,\mathbf{I}_{\mathscr{C}}) f(\Theta|\mathbf{I}_{\mathscr{C}})}{\int_{\Omega} p(\mathscr{D}|\Theta,\mathbf{I}_{\mathscr{C}}) f(\Theta|\mathbf{I}_{\mathscr{C}}) d\Theta}.$$

Since the N samples in training data are assumed to be independent, and by using (1) we have that $p(\mathscr{D}|\Theta,\mathbf{I}_{\mathscr{C}}) = \prod_{i=1}^{N} p(d_i|\Theta,\mathbf{I}_{\mathscr{C}}) = \theta_1^{n_1} \dots \theta_M^{n_M}$, where $\sum_{i=1}^{M} n_i = N$.

To determine the probability $f(\Theta|\mathbf{I}_{\mathscr{C}})$, we investigate the prior information $\mathbf{I}_{\mathscr{C}}$. It consists of two parts, $\mathbf{I}_{\mathscr{C}} = \{\mathbf{I},\mathbf{I}_E\}$. The first part, $\mathbf{I}$, is the basic prior information, stating that the probability is parameterized by Θ, that Θ is Dirichlet distributed, and knowledge about the prior probabilities for the classes. The second part, $\mathbf{I}_E$, includes the information that Θ satisfies (11), as well as the values of E and F. By using Bayes' theorem we have that $f(\Theta|\mathbf{I}_{\mathscr{C}}) = f(\Theta|\mathbf{I},\mathbf{I}_E) \propto f(\Theta|\mathbf{I}) f(\mathbf{I}_E|\Theta,\mathbf{I})$, where $f(\Theta|\mathbf{I})$ is given by (2), and $f(\mathbf{I}_E|\Theta,\mathbf{I}) = f_{E\Theta=F}(\Theta)$ is the distribution where all probability mass is uniformly distributed over the set $\Omega_E = \{\Theta : \Theta \in \Omega, E\Theta = F\}$. Thus, we have

$$p(\mathbf{Z} = \mathbf{z}_i|\mathscr{D},\mathbf{I}_{\mathscr{C}}) = \frac{\int_{\Omega_E} \theta_1^{n_1+\alpha_1-1} \dots \theta_i^{n_i+\alpha_i} \dots \theta_M^{n_M+\alpha_M-1} f_{E\Theta=F}(\Theta) d\Theta}{\int_{\Omega_E} \theta_1^{n_1+\alpha_1-1} \dots \theta_i^{n_i+\alpha_i-1} \dots \theta_M^{n_M+\alpha_M-1} f_{E\Theta=F}(\Theta) d\Theta.} \tag{14}$$

We will now give one example of how this integral can be solved using variable substitution.

A Solution Method Based on Variable Substitution

To solve the integrals in (14) substitute variables $\Theta = B + Q\Phi$, where Φ are new variables parameterizing the set of Θ fulfilling $E\Theta - F = 0$. The matrix $E \in \mathbb{R}^{l \times M}$ has full row rank (otherwise there would be redundant information about the parameters Θ, and rows could be removed from E). Thus, we can find a permutation matrix P such that $EP = \tilde{E} = [\tilde{E}_l \quad \tilde{E}_{M-l}]$ where $\tilde{E}_l \in \mathbb{R}^{l \times l}$ has full rank. The requirement (11) is transformed to

$$\tilde{E}\tilde{\Theta} = F, \tag{15}$$

where $P^T\Theta = \tilde{\Theta} = (\tilde{\theta}_1, \dots, \tilde{\theta}_M)^T$. Similarly for the counts of training data $n = (n_1, \dots, n_M)$ and the hypothetical samples we have $P^T n = \tilde{n} = (\tilde{n}_1, \dots, \tilde{n}_M)$ and $P^T\alpha = \tilde{\alpha} = (\tilde{\alpha}_1, \dots, \tilde{\alpha}_M)$. Multiply (15) by $\tilde{E}_l^{-1}$ to obtain

$$[I_l \quad \tilde{E}_l^{-1}\tilde{E}_{M-l}]\tilde{\Theta} = \tilde{E}_l^{-1}F \quad \Leftrightarrow \quad \tilde{\Theta}_{1:l} + \tilde{E}_l^{-1}\tilde{E}_{M-l}\tilde{\Theta}_{l+1:M} = \tilde{E}_l^{-1}F, \tag{16}$$

where $\tilde{\Theta}_{1:l}$ are the first l rows of $\tilde{\Theta}$ and $\tilde{\Theta}_{l+1:M}$ are the last $M-l$ rows. In in (16), augment $\tilde{\Theta}_{1:l}$ with $\tilde{\Theta}_{l+1:M}$ and let $\Phi = \tilde{\Theta}_{l+1:M}$. Then, rearranging the terms gives

$$\tilde{\Theta} = \underbrace{\begin{bmatrix} -\tilde{E}_l^{-1}\tilde{E}_{M-l} \\ I_{M-l} \end{bmatrix}}_{Q} \Phi + \underbrace{\begin{bmatrix} \tilde{E}_l^{-1} b \\ 0_{M-l\times 1} \end{bmatrix}}_{B}. \tag{17}$$

Let Q_i and B_i be the i:th rows in Q and B respectively. Then $\theta_i = Q_i\Phi + B_i$, and we can write the integrals in (14) as

$$\int_\Omega \tilde{\theta}_1^{\tilde{k}_1} \dots \tilde{\theta}_M^{\tilde{k}_M} \prod_{i=1}^{l} \delta(\tilde{\theta}_i - \tilde{\theta}_i^0(\Phi)) d\tilde{\Theta} = \int_{\Omega_\Phi} (Q_1\Phi + B_1)^{\tilde{k}_1} \dots (Q_M\Phi + B_M)^{\tilde{k}_M} d\Phi, \tag{18}$$

where $\delta(\cdot)$ is the dirac delta function, $\theta_i^0(\Phi)$ is the solution to the equation $Q_i\Phi + B_i = 0$, $\Omega_\Phi = \{\Phi : Q\Phi + B > 0\}$, and $\tilde{k}_j = \tilde{k}_j(\tilde{n}_j, \tilde{\alpha}_j)$.

The area of integration for the left hand side of (18) is determined by, for each ϕ_i in $\Phi = (\phi_i, \dots, \phi_{M-l})$, finding the lower boundary by solving the optimization problems

$$\begin{aligned} &\min_{\Sigma=(\sigma_1,\dots,\sigma_{M-l})} \sigma_i \\ &\text{subject to} \quad Q\Sigma > 0 \\ &\qquad\qquad \sigma_k = \phi_k, \quad k = 1, \dots, i-1. \end{aligned} \tag{19}$$

For the upper boundary, min is replaced by max in (19).

To investigate the computations in detail, return to the example with E and b given by (12). Here we use the identity matrix for P. Then the integral (18) becomes

$$\int_0^{0.5} (0.5-\phi_1)^{\tilde{k}_1} (0.5-\phi_1)^{\tilde{k}_2} \phi_1^{\tilde{k}_3} \phi_1^{\tilde{k}_4} d\phi_1 = \frac{1}{2^{1+\Sigma_{i=1}^4 \tilde{k}_i}} \frac{\Gamma(\tilde{k}_1 + \tilde{k}_2 + 1)\Gamma(\tilde{k}_3 + \tilde{k}_4 + 1)}{\Gamma(2 + \sum_{i=1}^4 \tilde{k}_i)}.$$

Although an analytical solution was easily found in the example considered here, this is generally not the case. To the authors knowledge, there is no closed formula for solving the integral on the right hand side in (18) in general. One possibility is to use Laplace approximation [13], where the integrand is approximated by an unnormalized Gaussian density function. See [12] for more details on the Laplace approximation applied to the current problem.

FAULT DIAGNOSIS EXAMPLE

To illustrate the methods, consider the following fault classification example with two-dimensional feature vector $\mathbf{X} = (X_1, X_2)$, where $x_i \in \{0,1\}$, and the two faults (states) $C \in \{c_1, c_2\}$. To simplify notation, assume that the classes have equal prior probability. Enumerate the parameters as

C	1	2	1	2	1	2	1	2
X_1	0	0	1	1	0	0	1	1
X_2	0	0	0	0	1	1	1	1
$p(\mathbf{z}\|\Theta\mathbf{I}_{\mathscr{C}})$	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8

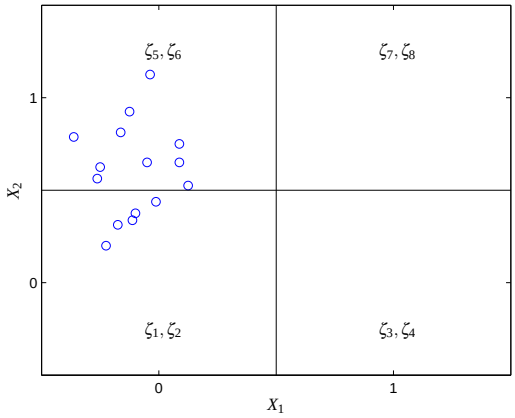

FIGURE 1. Example of training data from state c_2.

and assume that we are given the causality information

$$p(x_1|\Theta, c_1, \mathbf{I}_{\mathscr{C}}) = p(x_1|\Theta, c_2, \mathbf{I}_{\mathscr{C}}).$$

For this particular example, the integrals in (14) have the form

$$\int_{\Omega_E} (0.5 - \phi_1 - \phi_4 - \phi_5)^{\tilde{k}_1} (\phi_1 + \phi_4 - \phi_3)^{\tilde{k}_2} (0.5 - \phi_1 - \phi_4 - \phi_2)^{\tilde{k}_3} \phi_1^{\tilde{k}_4} \phi_2^{\tilde{k}_5} \phi_3^{\tilde{k}_6} \phi_4^{\tilde{k}_7} \phi_5^{\tilde{k}_8} d\Phi,$$

where we have used the permutation $\tilde{U} = [U_4\ U_1\ U_7\ U_2\ U_3\ U_5\ U_6\ U_8]$, where $U = n, \alpha, E, \Theta$. Let $\alpha_i = 1$, $i = 1, \ldots 8$ and consider for example the case when there is no data available from class c_1, i.e. $n_i = 0$, $i = 1, 3, 5, 7$, while there is training data $n_2 = 5$, $n_6 = 10$, $n_4 = n_8 = 0$ available. This example is plotted in Figure 1 and means that under class c_2 the observation $X_1 = 0$ is more likely than $X_1 = 1$. Since we have the causality information that X_1 is equally distributed under both classes we expect the observation $X_1 = 0$ to be more likely under class c_2 as well. This is verified by the computations

$$\begin{aligned} p(X_1 = 0, X_2 = 1, c = c_1|\mathscr{D}, \mathbf{I}_{\mathscr{C}}) &= p(\mathbf{Z} = \zeta_5|\mathscr{D}, \mathbf{I}_{\mathscr{C}}) = \\ &= \frac{\int_{\Omega_E} \phi_1^{n_2} \phi_3^{n_5} \phi_4^{n_6} d\Phi}{\int_{\Omega_E} \phi_1^{n_2} \phi_4^{n_6} d\Phi} \approx 0.41, \\ p(X_1 = 1, X_2 = 1, c = c_1|\mathscr{D}, \mathbf{I}_{\mathscr{C}}) &= p(\mathbf{Z} = \zeta_7|\mathscr{D}, \mathbf{I}_{\mathscr{C}}) = \\ &= \frac{\int_{\Omega_E} \phi_1^{n_2} (0.5 - \phi_1 - \phi_4 - \phi_2)^{n_7} \phi_4^{n_6} d\Phi}{\int_{\Omega_E} \phi_1^{n_2} \phi_4^{n_6} d\Phi} \approx 0.035, \end{aligned}$$

and similar for the case where $X_2 = 0$. If causality information is not used, the probabilities becomes $p(X_1 = 0, X_2 = 1, c = c_1|\mathscr{D}, \mathbf{I}) = p(X_1 = 1, X_2 = 1, c = c_1|\mathscr{D}, \mathbf{I}) = 1/23 \approx 0.043$ by Theorem 1.

CONCLUSION

In the present work, it has been shown how the probabilistic inference problem can be formulated using training data combined with prior information given in terms of response and causality information. This type of prior information appears for example in traditional fault diagnosis problems. It has been shown how this prior information can be expressed as requirements on the parameters in the distributions.

A theorem for using response information in the inference problem has been given. Furthermore, it has been shown how the causality information can be introduced in the computations, and it is discussed how to solve the computations conceptually.

In the present work response and causality information alone has been considered one a a time, but they can also be used together to improve the inference further.

Introducing the prior information to the fault inference problem can, as shown in an example, improve the results significantly. It has been shown that the causality information makes it possible to reuse training data from one state when considering other states. This is particularly helpful when there is only a limited amount of training data available as is often the case in fault diagnosis.

ACKNOWLEDGMENTS

We acknowledge Udo von Toussaint for interesting discussions, in particular on methods for solving the integrals.

REFERENCES

1. R. O. Duda, P. E. Hart, and D. G. Storch, *Pattern Classification, 2nd Edition*, Wiley and Sons, 2001.
2. L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer, 1996.
3. A. O'Hagan, and J. Forster, *Kendall's Advanced Theory of Statistics*, Arnold, 2004.
4. J. de Kleer, and B. C. Williams, "Diagnosis with Behavioral Modes," in *Readings in Model-based Diagnosis*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1992, pp. 124–130, ISBN 1-55860-249-6.
5. J. M. Koscielny, M. Bartys, and M. Syfert, "The Practical Problems of Fault Isolation in Large Scale Industrial Systems," in *proceedings IFAC SAFEPROCESS*, 2006.
6. S. N. G. Biswas, *IEEE Trans. on Systems, Man And Cybernetics. Part A* **37**, 348–361 (2007).
7. M. Blanke, M. Kinnaert, J. Lunze, M. Staroswiecki, and J. Schröder, *Diagnosis and Fault Tolerant Control*, Springer, 2003.
8. J. J. Gertler, *Fault Detection and Diagnosis in Engineering Systems*, Marcel Decker, 1998.
9. P. Kontkanen, P. Myllymaki, T. Silander, H. Tirri, and P. Grunwald, "Comparing predictive inference methods for discrete domains," in *Proceedings of the Sixth International Workshop on Artificial Intelligence and Statistics, Ft. Lauderdale, Florida.*, 1997, pp. 311–318.
10. D. Heckerman, D. Geiger, and D. M. Chickering, *Machine Learning* **20**, 197–243 (1995).
11. A. Pernestål, and M. Nyberg, "Probabilistic Fault Diagnosis Based on Incomplete Data with Application to an Automotive Engine," in *Proceedings of European Control Conference*, 2007.
12. A. Pernestål, Using Data and Prior Information in Bayesian Classification, Tech. Rep. LiTH-ISY-R-2811, ISY, Linköping University (2007).
13. D. J. MacKay, *Information Theory, Inference and Learning Algorithms*, Cambridge University Press, 2005.