Expressive Power of Binary Relevance and Chain Classifiers Based on Bayesian Networks for Multi-label Classification

Gherardo Varando, Concha Bielza, and Pedro Larrañaga

Departamento de Inteligencia Artificial,
Universidad Politécnica de Madrid,
Campus de Montegancedo, 28660 Boadilla del Monte, Madrid, Spain
gherardo.varando@upm.es, {mcbielza,pedro.larranaga}@fi.upm.es
http://cig.fi.upm.es

Abstract. Bayesian network classifiers are widely used in machine learning because they intuitively represent causal relations. Multi-label classification problems require each instance to be assigned a subset of a defined set of h labels. This problem is equivalent to finding a multi-valued decision function that predicts a vector of h binary classes. In this paper we obtain the decision boundaries of two widely used Bayesian network approaches for building multi-label classifiers: Multi-label Bayesian network classifiers built using the *binary relevance method* and Bayesian network *chain classifiers*. We extend our previous single-label results to multi-label chain classifiers, and we prove that, as expected, chain classifiers provide a more expressive model than the binary relevance method.

1 Introduction

We consider a multi-label classification problem [19] over categorical predictors, that is, mapping every instance $\mathbf{x} = (x_1, \ldots, x_n)$ to a subset of h labels:

$$\Omega_1 \times \cdots \times \Omega_n \to Y \subset \mathcal{Y} = \{y_1, \ldots, y_h\},$$

where $\Omega_i \subset \mathbb{R}$, $|\Omega_i| = m_i < \infty$. This could be transformed into a multi-dimensional binary classification problem, that is, finding an h-valued decision function $\mathbf{f}$ that maps every instance of n predictor variables $\mathbf{x}$ to a vector of h binary values $\mathbf{c} = (c_1, \ldots, c_h) \in \{-1, +1\}^h$:

$$\begin{aligned} \mathbf{f}: \quad \mathbf{\Omega} = \Omega_1 \times \cdots \times \Omega_n &\to \{-1, +1\}^h \\ (x_1, \ldots, x_n) &\mapsto (c_1, \ldots, c_h), \end{aligned}$$

where $c_i = +1$ (-1) means that the ith label is present (absent) in the predicted label subset. Moreover, we consider the predictor variables $X_1, \ldots, X_n$ and the

binary classes $C_i \in \{-1, +1\}$ as categorical random variables. Real examples include classification of texts into different categories by counting selected words, diagnosis of multiple diseases from common symptoms and identification of multiple biological gene functions.

The simplest method to build a multi-label classifier is to consider h single-label binary classifiers, one for each class variable C_i. Each classifier f_i is learned from predictor variables and C_i data, and the results are combined to form multi-label prediction. This method, called *binary relevance* [6], is easily implementable, has low computational complexity and is fully parallelizable. Hence it is scalable to a large number of classes. However, it completely ignores dependencies among labels and generally it does not represent the most likely set of labels.

Chain classifiers [14] relax the independence assumption by iteratively adding class dependencies in the binary relevance scheme, that is, the kth classifier in the chain predicts class C_k from $X_1, \ldots, X_n, C_1, \ldots, C_{k-1}$.

We study differences in the expressive power of these two methods when Bayesian network (BN) classifiers [1] are used. Sucar *et al.* [15] employed naive Bayes within chain classifiers. We use the results on the decision boundaries and expressive power of one-dimensional BN classifiers. (a) For naive Bayes classifiers, Minsky [9] proved that the decision boundaries are hyperplanes if binary predictors are used. (b) Peot [11] observed that Minsky's results could be extended to categorical predictors. (c) Recently, we have developed a method [18] to compute decision boundaries for a broad class of BN classifiers. In this paper we extend these results to multi-label classifiers. Moreover, we suggest some theoretical reasons why the binary relevance method performs poorly and prove that chain classifiers provide more expressive models.

The paper is organized as follows. In Sect. 2 we give some definitions and report our results on one-label classifiers. We describe the binary relevance method in Sect. 3 and chain classifiers in Sect. 4. In Sect. 5 we compare the decision boundaries, and expressive power of the two methods. In Sect. 6 we present our conclusions and some ideas for future research.

2 Expressive Power of One-Dimensional BN Classifiers

We first report some results on the decision boundary and expressive power of one-label, or equivalently one-dimensional binary, BN classifiers [18]. In particular we look at Bayesian network-augmented naive Bayes (BAN) classifiers [5].

BAN classifiers are Bayesian network classifiers where the class variable C is assumed to be a parent of every predictor and the predictor sub-graph can be a general BN. The decision function induced by the BAN classifier is

$$f_{\mathcal{G}}^{BAN}(x_1, \ldots, x_n) = \arg \max_{c \in \{-1,+1\}} P(C = c, X_1 = x_1, \ldots, X_n = x_n),$$

where $P(C = c, X_1 = x_1, \ldots, X_n = x_n)$ could be factorized according to BN theory [10] as

$$P\left(C = c\right) \prod_{i=1}^{n} P\left(X_i = x_i | C = c, \mathbf{X}_{\mathbf{pa}(i)} = \mathbf{x}_{\mathbf{pa}(i)}\right),$$

where $\mathbf{X}_{\mathbf{pa}(i)}$ stands for the vector of parents of X_i in the predictor sub-graph $\mathcal{G}$. Moreover, $\mathbf{pa}(i)$ denotes the set of indexes defining the parents of X_i that are not C and $\mathbb{M}_i = \times_{s \in \mathbf{pa}(i)} \{1, \ldots, m_s\}$, the set of possible configurations of the parents of X_i.

Let us recall that the sign function $sgn(t)$ is defined as

$$sgn(t) = \begin{cases} +1 & \text{if } t > 0 \\ 0 & \text{if } t = 0 \\ -1 & \text{if } t < 0. \end{cases}$$

We define [18]:

Definition 1. *Given a decision function $f : \mathbf{\Omega} \to \{-1, +1\}$, where $\mathbf{\Omega} \subset \mathbb{R}^n$, $|\mathbf{\Omega}| < \infty$ and $r : \mathbb{R}^n \mapsto \mathbb{R}$ is a polynomial, we say that r sign-represents f if*

$$f(\mathbf{x}) = sgn(r(\mathbf{x})) \text{ for every } \mathbf{x} \in \mathbf{\Omega}.$$

Moreover, given a set of polynomials $\mathcal{P}$, we denote by $sgn(\mathcal{P})$ the set of decision functions that are sign-representable by polynomials in $\mathcal{P}$ and by $\{-1, +1\}^{|\mathbf{\Omega}|}$, the set of all $2^{|\mathbf{\Omega}|}$ decision functions over $\mathbf{\Omega}$.

Example 1. We consider $\mathbf{\Omega} = \{0, 2\} \times \{-3, 1\}$ and the decision function over $\mathbf{\Omega}$

$$f(x_1, x_2) = \begin{cases} +1 & \text{if } (x_1, x_2) = (0, -3), (2, -3), (0, 1) \\ -1 & \text{if } (x_1, x_2) = (2, 1). \end{cases}$$

We have that the polynomial $r(x_1, x_2) = -x_1^2 - x_2 + 3$ sign-represents f over $\mathbf{\Omega}$, precisely:

$$r(0, -3) = 6 > 0, \quad r(2, -3) = 2 > 0, \quad r(0, 1) = 2 > 0 \text{ and } r(2, 1) = -2 < 0.$$

Next let us recall the definition of the Vapnik-Chervonenkis (VC) dimension [17].

Definition 2. *Given a subset of decision functions $\mathcal{F} \subset \{-1, +1\}^{|\mathbf{\Omega}|}$, we say that $\mathcal{F}$ shatters $\mathbf{\Omega}_0 \subset \mathbf{\Omega}$ if for every $g \in \{-1, +1\}^{|\mathbf{\Omega}_0|}$ there exists a decision function $f \in \mathcal{F}$ such that $f_{|\mathbf{\Omega}_0} = g$, where $f_{|\mathbf{\Omega}_0}$ indicates the restriction of f over $\mathbf{\Omega}_0$.*

That is, $\mathcal{F}$ shatters $\mathbf{\Omega}_0$ if every decision over $\mathbf{\Omega}_0$ is representable by some elements of $\mathcal{F}$. The cardinality of the largest subset shattered by $\mathcal{F}$ is called the VC dimension of $\mathcal{F}$. It indicates the maximum number of points that can be discriminated by $\mathcal{F}$.

Definition 3. *The VC dimension of* $\mathcal{F} \subset \{-1,+1\}^{|\mathbf{\Omega}|}$*, denoted by* $d_{VC}(\mathcal{F})$*, is defined by*

$$d_{VC}(\mathcal{F}) = \max\{|\mathbf{\Omega}_0| \text{ s.t. } \mathbf{\Omega}_0 \text{ is shattered by } \mathcal{F}\}.$$

For every predictor variable $X_i \in \Omega_i = \{\xi_i^1, \ldots, \xi_i^{m_i}\}$, we define the Lagrange basis polynomials over Ω_i

$$\ell_j^{\Omega_i}(x) = \prod_{k \neq j} \frac{(x - \xi_i^k)}{(\xi_i^j - \xi_i^k)} \text{ for every } j = 1, \ldots, m_i \text{ and } x \in \mathbb{R}. \tag{1}$$

Then we have [18]:

Lemma 1. *If* f *is the decision function induced by a BAN classifier for a classification problem with* n *categorical predictor variables* $\{X_i \in \Omega_i \subset \mathbb{R}, |\Omega_i| = m_i\}_{i=1}^n$*, then there exists a polynomial of the form*

$$\sum_{i=1}^{n} \sum_{j=1}^{m_i} \ell_j^{\Omega_i}(x_i) \sum_{\mathbf{k} \in \mathbb{M}_i} \beta_i(j|\mathbf{k}) \prod_{s \in \mathbf{pa}(i)} \ell_{k_s}^{\Omega_s}(x_s)$$

that sign-represents f*, where we write* $\sum_{\mathbf{k} \in \mathbb{M}_i} \beta_i(j|\mathbf{k}) \prod_{s \in \mathbf{pa}(i)} \ell_{k_s}^{\Omega_s}(x_s) = \beta_i(j)$ *when a variable does not have parents different from* C*, that is,* $\mathbf{pa}(i) = \emptyset$*.*

The proof of Lemma 1 [18] is constructive and the coefficients $\beta_i(j|\mathbf{k})$ of the built polynomial are related to the conditional probability tables of the BAN. Precisely we have that

$$\beta_i(j|\mathbf{k}) = \ln \frac{P(X_i = \xi_i^j | X_s(i) = \xi_s^{k_s}, \forall s \in \mathbf{pa}(i), C = +1)}{P(X_i = \xi_i^j | X_s(i) = \xi_s^{k_s}, \forall s \in \mathbf{pa}(i), C = -1)}, \tag{2}$$

where $\mathbf{k} = (k_s)_{s \in \mathbf{pa}(i)}$, $k_s \in \{1, \ldots, m_s\}$.

When the predictor sub-graph $\mathcal{G}$ does not contain V-structures, the inverse implication of Lemma 1 is provable and thus the following theorem [18] holds.

Theorem 1. *Let* $\mathcal{G}$ *be a directed acyclic graph with nodes* X_i *for* $i \in \{1, 2, \ldots, n\}$ *and* f*, a decision function over predictor variables* $X_i \in \Omega_i = \{\xi_i^1, \ldots, \xi_i^{m_i}\}$*. Suppose that* $\mathcal{G}$ *does not contain* V*-structures, then we have that* f *is sign-represented by a polynomial of the form*

$$r(\mathbf{x}) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \ell_j^{\Omega_i}(x_i) \sum_{\mathbf{k} \in \mathbb{M}_i} \beta_i(j|\mathbf{k}) \prod_{s \in \mathbf{pa}(i)} \ell_{k_s}^{\Omega_s}(x_s),$$

if and only if f *is induced by a BAN classifier whose predictor sub-graph is* $\mathcal{G}$*.*

The above result applies in a lot of practical cases as naive Bayes (NB) classifier [9], tree augmented naive Bayes (TAN) classifier [5] and super-parent

one-dependence-estimator (SPODE) classifier [8], because the corresponding predictor sub-graphs do not contain V-structures. Moreover, Theorem 1 implies that when $\mathcal{G}$ does not contain V-structures the family of polynomials

$$\mathcal{P}_{\mathcal{G}} = \left\{ r(\mathbf{x}) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \ell_j^{\Omega_i}(x_i) \sum_{\mathbf{k} \in \mathbb{M}_i} \beta_i(j|\mathbf{k}) \prod_{s \in \mathbf{pa}(i)} \ell_{k_s}^{\Omega_s}(x_s) \text{ s.t. } \beta_i(j|\mathbf{k}) \in \mathbb{R} \right\} \tag{3}$$

completely represents the set of decision functions induced by BAN classifiers, that is, $sgn(\mathcal{P}_{\mathcal{G}})$ is exactly the set of decision functions induced by BAN classifiers whose predictor sub-graph is $\mathcal{G}$.

Remark 1. In the simplest NB classifier case, that is, when the predictor sub-graph $\mathcal{G}$ is an empty graph, we have that

$$\mathcal{P}_{\mathcal{G}} \equiv \mathcal{P}_{NB} = \left\{ r(\mathbf{x}) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \alpha_i(j) \ell_j^{\Omega_i}(x_i) \text{ s.t. } \alpha_i(j) \in \mathbb{R} \right\}$$

is exactly the set of polynomials that sign-represent the decision function induced by NB classifiers.

We can prove that the set $\mathcal{P}_{\mathcal{G}}$ is a vector space of dimension

$$d = \sum_{i=1}^{n} \left((m_i - 1) \prod_{s \in \mathbf{pa}(i)} m_s \right) + 1$$

and that the VC dimension of $sgn(\mathcal{P}_{\mathcal{G}})$ is precisely d. Theorem 1 also places an upper bound on the number of decision functions representable by BAN classifiers without V-structures [18].

Corollary 1. *Consider a BAN classifier over predictor variables $X_i \in \Omega_i$, $|\Omega_i| = m_i$ for every $i = 1, \ldots, n$. Moreover suppose that the predictor sub-graph $\mathcal{G}$ does not contain V-structures. Then we have*

$$|sgn(\mathcal{P}_{\mathcal{G}}^{BAN})| \leq C(M, d) = 2 \sum_{k=0}^{d-1} \binom{M-1}{k},$$

where $d = \sum_{i=1}^{n} \left((m_i - 1) \prod_{s \in \mathbf{pa}(i)} m_s \right) + 1$ and $M = \prod_{i=1}^{n} m_i$.

Remark 2. If $\mathbf{\Omega} = \Omega_1 \times \cdots \times \Omega_n$, we observe that $|\{-1, +1\}^{|\mathbf{\Omega}|}| = 2^{|\mathbf{\Omega}|} = 2^M$. Thus Corollary 1 implies that in the case of the NB classifier the quotient of decision functions representable by NB classifiers over 2^M becomes vanishingly small as the number of predictors increase. Figure 1 shows the number of total decision functions ($2^{|\mathbf{\Omega}|}$) and the bounding of Corollary 1 for NB classifiers with n binary predictors, $C(M, d)$. Observing that the scale of the graph is logarithmic, the graph shows that the number of decision functions induced by NB classifiers is *small* compared with all possible decision functions over $\mathbf{\Omega}$.

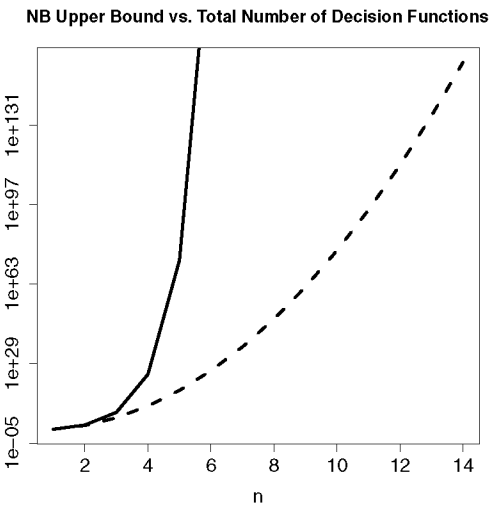

Fig. 1. Total number of decision functions over n binary predictors (gray) and the bounding $C(M, d)$ of Corollary 1 (dashed black) for NB classifiers

Remark 2 could be extended to every type of BAN classifier, such that for every variable the number of parents is bounded (Corollary 17 in Varando *et al.* [18]), that is, $|\mathbf{pa}(i)| < K$.

Remark 3. When the predictor sub-graph $\mathcal{G}$ of a BAN classifiers contains V-structures, Lemma 1 is still valid and exists a polynomial that sign-represents the induced decision function. The problem is that the associated family of polynomials is not a linear space as in (3), thus is not possible to employ the same techniques as in Varando *et al.* [18] to prove the bounding in Corollary 1.

3 Binary Relevance Method

We consider the binary relevance method with BAN classifiers, that is, for every class C_i we build a BAN classifier with predictor sub-graph $\mathcal{G}$. Thus every one-dimensional classifier has the same predictor structure and differs with respect to the values of the conditional probability tables that define the BAN models. From a practical point of view, the advantages of this method are that the structure of the predictor sub-graph has only to be learned once and the parameters of the BN are then fitted to the different data sets related to each class.

From Lemma 1 it follows that if $\mathbf{f} = (f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_h(\mathbf{x}))$ is the multi-valued decision function induced by the h BAN classifiers, then there exist

$$p_1(\mathbf{x}), \ldots, p_h(\mathbf{x}) \in \mathcal{P}_{\mathcal{G}},$$

such that $f_i(\mathbf{x}) = sgn(p_i(\mathbf{x}))$ for every $i \in \{1, \ldots, h\}$. Thus, in Lemma 2, we bound the number of multi-valued decision functions representable by the BAN

binary relevance method, when the predictor sub-graph does not contain V-structures.

Lemma 2. *Consider h BAN classifiers, whose predictor sub-graph $\mathcal{G}$ contains no V-structures, to predict h binary classes. We have that $N(\mathcal{G}, h)$, the number of h-valued decision functions representable by the BAN binary relevance method, satisfies*

$$N(\mathcal{G}, h) \leq C(M, d)^h,$$

where $C(M,d) = 2\sum_{k=0}^{d-1} \binom{M-1}{k}$, $d = \sum_{i=1}^{n} \left((m_i - 1)\prod_{s\in\mathbf{pa}(i)} m_s\right) + 1$ *and* $M = \prod_{i=1}^{n} m_i$.

Proof. The proof is a straightforward application of Corollary 1. □

Remark 4. The total number of h-valued decision functions over n categorical predictors is $2^{h\prod m_i} = 2^{hM}$. Then the fraction of h-valued decision functions representable by the BAN binary relevance method is bounded by

$$\frac{N(\mathcal{G}, h)}{2^{hM}} \leq \left(\frac{C(M,d)}{2^M}\right)^h.$$

Thus, as in Remark 2, we have that if we fix the structure of the predictor sub-graph, and it does not contain V-structures, the number of representable multi-valued decision functions becomes vanishingly small as the number of predictors increase. Moreover, using the binary relevance method, the *speed* at which the ratio between representable multi-valued decision functions and the total number of multi-valued decision functions drops to zero, is exponential in h, the number of classes.

The above bound could also be computed when each of the h BAN classifiers is built with different structures, that is, the kth classifier to predict class C_k is a BAN classifier whose predictor sub-graph $\mathcal{G}_k$ does not contain V-structures. Then if we denote $N(\mathcal{G}_1, \ldots, \mathcal{G}_h)$ the number of h-valued decision functions built with h BAN classifiers whose predictor sub-graph is $\mathcal{G}_1, \ldots, \mathcal{G}_h$ respectively, we have that

$$N(\mathcal{G}_1, \ldots, \mathcal{G}_h) \leq \prod_{k=1}^{h} C(M, d_k),$$

where $d_k = \sum_{i=1}^{n} \left((m_i - 1)\prod_{s\in\mathbf{pa}_k(i)} m_s\right) + 1$, $\mathbf{pa}_k(i)$ is the set of X_i parents in $\mathcal{G}_k$ and $M = \prod_{i=1}^{n} m_i$.

Example 2. We consider two binary classes C_1, C_2 and two predictor variables $X_1 \in \{0, 1\}$ and $X_2 \in \{2, 3, 4\}$. Using the binary relevance method we build two independent NB classifiers, see Fig. 2.

Next, we list the conditional probability tables for both classifiers (Tables 1a and 1b).

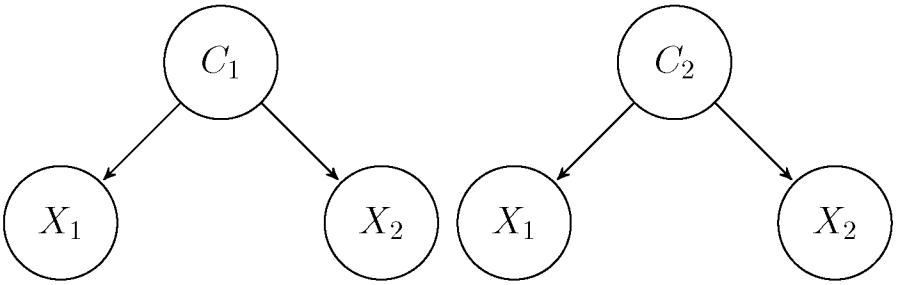

Fig. 2. Two NB classifiers in Example 2

Table 1. Conditional probability tables in Example 2 for the two NB classifiers

(a) NB for C_1

X_1	$C_1 = +1$	$C_1 = -1$
0	0.5	0.25
1	0.5	0.75

X_2	$C_1 = +1$	$C_1 = -1$
2	0.3	0.1
3	0.5	0.7
4	0.2	0.2

(b) NB for C_2

X_1	$C_2 = +1$	$C_2 = -1$
0	0.7	0.4
1	0.3	0.6

X_2	$C_2 = +1$	$C_2 = -1$
2	0.1	0.6
3	0.1	0.2
4	0.8	0.2

From the representation of Theorem 1 we have that there exist two polynomials p_1, p_2 that sign-represent the decision functions induced by the two NB classifiers

$$\begin{aligned} p_1(x_1, x_2) = &\ln\left(\frac{0.5}{0.25}\right)\frac{x_1 - 1}{-1} + \ln\left(\frac{0.5}{0.75}\right)\frac{x_1}{1} \\ &+ \ln\left(\frac{0.3}{0.1}\right)\frac{(x_2 - 3)(x_2 - 4)}{2} + \ln\left(\frac{0.5}{0.7}\right)\frac{(x_2 - 2)(x_2 - 4)}{-1} \\ &+ \ln\left(\frac{0.2}{0.2}\right)\frac{(x_2 - 2)(x_2 - 3)}{2} \end{aligned}$$

and

$$\begin{aligned} p_2(x_1, x_2) = &\ln\left(\frac{0.7}{0.4}\right)\frac{x_1 - 1}{-1} + \ln\left(\frac{0.3}{0.6}\right)\frac{x_1}{1} \\ &+ \ln\left(\frac{0.1}{0.6}\right)\frac{(x_2 - 3)(x_2 - 4)}{2} + \ln\left(\frac{0.1}{0.2}\right)\frac{(x_2 - 2)(x_2 - 4)}{-1} \\ &+ \ln\left(\frac{0.8}{0.2}\right)\frac{(x_2 - 2)(x_2 - 3)}{2}. \end{aligned}$$

We have that

$$\mathbf{f}(\mathbf{x}) = \Big(sgn\big(p_1(\mathbf{x})\big), sgn\big(p_2(\mathbf{x})\big) \Big)$$

is the bi-valued decision function that predicts C_1, C_2 from X_1, X_2. Figure 3 shows the decision boundaries of the two classifiers (black for C_1 and gray for C_2). We observe that the predictor space $\mathbf{\Omega} = \{0,1\} \times \{2,3,4\}$ is partitioned into four subsets corresponding to the four different predictions of the two binary classes. Moreover, the value of the respective predicted class changes when one of the decision boundaries is crossed.

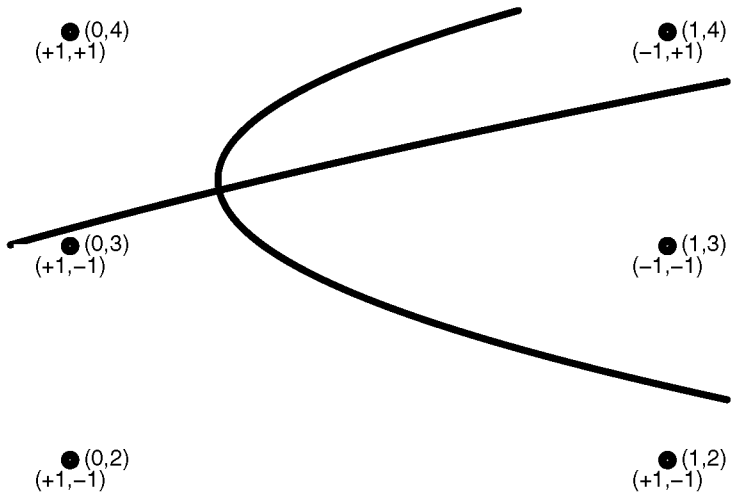

Fig. 3. Decision boundaries for the two NB classifiers in Example 2. The value of the predicted classes and the coordinates of the points are reported.

4 BN Chain Classifiers

The easiest way to relax the strong independence assumption of the binary relevance method is to gradually add the predicted classes to the predictors. Specifically, suppose that we have to predict h binary classes $C_1, \ldots, C_h$ from n predictor variables $X_1, \ldots, X_n$. We consider h BAN classifiers such that the kth BAN classifier predicts C_k from the variables

$$X_1, \ldots, X_n, C_1, \ldots, C_{k-1}.$$

From Lemma 1 we have that there exist h polynomials $p_1, \ldots, p_h$ such that

$$p_k(\mathbf{x}, c_1, \ldots, c_{k-1}) : \mathbb{R}^{n+k-1} \rightarrow \mathbb{R}$$

$$p_k \in \mathcal{P}_{\mathcal{G}_k},$$

where $\mathcal{G}_k$ is the predictor sub-graph related to the kth BAN classifier over $X_1, \ldots, X_n, C_1, \ldots, C_{k-1}$.

If we consider only naive Bayes classifiers, we state

$$\mathcal{P}_k = \left\{ \begin{array}{r} r(\mathbf{x}) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \alpha_i(j) \ell_j^{\Omega_i}(x_i) + \sum_{i=1}^{k-1} \beta_i(+1) \ell_{+1}^{\{-1,+1\}}(c_i) + \beta_i(-1) \ell_{-1}^{\{-1,+1\}}(c_i) \\ \text{s.t. } \alpha_i(j), \beta_i(+1), \beta_i(-1) \in \mathbb{R} \end{array} \right\}, \tag{4}$$

for the set of polynomials sign-representing the decision function of the kth classifier in the chain, that is, the NB classifier that predicts C_k from $X_1, \ldots, X_n$ and $C_1, \ldots, C_{k-1}$. Moreover, observe that

$$\ell_{+1}^{\{-1,+1\}}(c_i) = \frac{c_i + 1}{2} = \begin{cases} 1 & \text{if } c_i = +1 \\ 0 & \text{if } c_i = -1 \end{cases}$$

$$\ell_{-1}^{\{-1,+1\}}(c_i) = \frac{1 - c_i}{2} = \begin{cases} 0 & \text{if } c_i = +1 \\ 1 & \text{if } c_i = -1 \end{cases}$$

For the first class C_1, we have that the first classifier is a NB over $X_1, \ldots, X_n$ and so the decision function for C_1 is

$$f_1(\mathbf{x}) = sgn\big(p_1(\mathbf{x})\big), \tag{5}$$

where $p_1(\mathbf{x}) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \alpha_i(j) \ell_j^{\Omega_i}(x_i) \in \mathcal{P}_1$. For the second class C_2, we have a NB classifier over $X_1, \ldots, X_n, C_1$. Thus $f_2(\mathbf{x})$, the decision function for C_2, is

$$f_2(\mathbf{x}) = sgn\Big(p_2\big(\mathbf{x}, c_1\big)\Big), \tag{6}$$

where $p_2 \in \mathcal{P}_2$ and $c_1 = f_1(\mathbf{x})$. Substituting (5) in (6), we obtain

$$f_2(\mathbf{x}) = sgn\Big(p_2\big(\mathbf{x}, sgn(p_1(\mathbf{x}))\big)\Big).$$

This chain classifier over two classes is equivalent to the bi-valued decision function

$$\mathbf{f} = \big(f_1(\mathbf{x}), f_2(\mathbf{x})\big).$$

Iterating the above computations, we have that the kth decision function that predicts class C_k is given by

$$f_k(\mathbf{x}) = sgn\Big(p_k\big(\mathbf{x}, f_1(\mathbf{x}), \ldots, f_{k-1}(\mathbf{x})\big)\Big),$$

where $p_k \in \mathcal{P}_k$. More explicitly, we have that

$$f_k(\mathbf{x}) = \begin{cases} sgn\Big(q_k(\mathbf{x}) + \gamma(+1,+1,\ldots,+1)\Big) & \text{if } f_1(\mathbf{x}) = +1, \ldots, f_{k-1}(\mathbf{x}) = +1 \\ \vdots & \vdots \\ sgn\Big(q_k(\mathbf{x}) + \gamma(\sigma_1,\sigma_2\ldots,\sigma_{k-1})\Big) & \text{if } f_1(\mathbf{x}) = \sigma_1, \ldots, f_{l-1}(\mathbf{x}) = \sigma_{k-1} \\ \vdots & \vdots \\ sgn\Big(q_k(\mathbf{x}) + \gamma(-1,-1,\ldots,-1)\Big) & \text{if } f_1(\mathbf{x}) = -1, \ldots, f_{k-1}(\mathbf{x}) = -1 \end{cases} \quad (7)$$

where $q_k(\mathbf{x}) \in \mathcal{P}_1$ and $\gamma(\sigma_1, \ldots, \sigma_{k-1}) \in \mathbb{R}$ for every $(\sigma_1, \ldots, \sigma_{k-1}) \in \{-1, +1\}^{k-1}$. In other words, the kth decision function, in every subset of $\mathbf{\Omega}$ defined by the previous $k-1$ decision functions, is sign-represented by a polynomial in $\mathcal{P}_1$ or equivalently by a NB classifier over the original predictors. The only difference between these polynomials is the additive coefficients. Precisely the additive coefficients $\gamma(\sigma_1, \ldots, \sigma_{k-1})$ are obtained from the representation in (4) as follows:

$$\gamma(\sigma_1, \ldots, \sigma_{k-1}) = \sum_{i=1}^{k-1} \beta_i(\sigma_i),$$

where

$$\beta_i(\sigma_i) = \ln \frac{P(C_i = \sigma_i | C_k = +1)}{P(C_i = \sigma_i | C_k = -1)}.$$

Figure 4 shows two examples of decision boundaries of a NB chain classifier for two classes. The predictor domain in both examples is $\{0,1,2,3\} \times \{0,1,2,3\}$. We observe that the decision boundaries related to the second class in the chain C_2 (dashed black line) are dependent on the decision boundaries of the first class C_1 (gray line).

Remark 5. For simplicity's sake, we have presented the computation of the decision boundaries in the NB case. The same arguments as used above could be applied to a broader class of chain classifiers, specifically to every model where a BAN classifier with predictor sub-graph $\mathcal{G}_k$ is built in the kth step of the chain. If the previously predicted classes $C_1, \ldots, C_{k-1}$ are added in a *naive* way, that is, they have only one parent, C_k and they have no children, we have that the form of the kth decision function is similar to (7), where the previously predicted classes contribute in the form of additive constants.

Example 3. We use a chain NB classifier over the prediction problems of Example 2. The NB classifier for predicting class C_1 is the same as in Example 2 (see Fig. 2 left and Table 1a). The predictors of the NB classifier for predicting C_2 now include C_1. We consider the same conditional probability tables as in Example 2 (Tables 1a and 1b). Moreover we have to specify the conditional probabilities of C_1 given C_2 in the NB that predicts C_2. We set

$$P(C_1 = +1 | C_2 = +1) = 0.3 \text{ and } P(C_1 = -1 | C_2 = +1) = 0.7$$

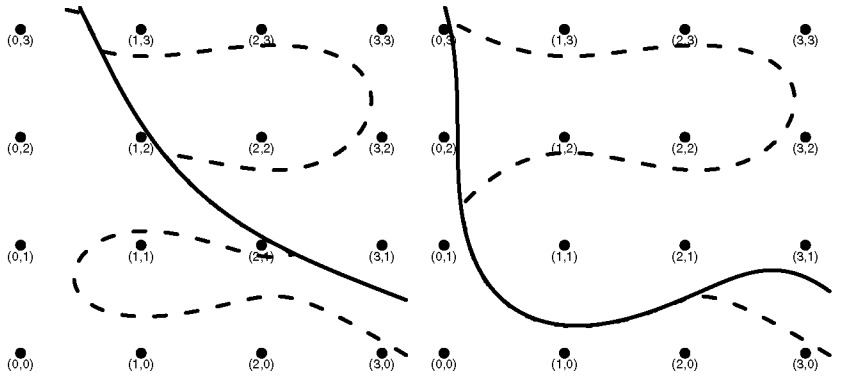

Fig. 4. Decision boundaries for NB chain classifiers with two predictor variables

$$P(C_1 = +1|C_2 = -1) = 0.9 \text{ and } P(C_1 = -1|C_2 = -1) = 0.1$$

And, thus, coefficients $\beta_1(+1)$ and $\beta_1(-1)$ as defined in (4) are given by

$$\beta_1(+1) = \ln\left(\frac{0.3}{0.9}\right) \quad \text{and} \quad \beta_1(-1) = \ln\left(\frac{0.7}{0.1}\right).$$

We have that the decision function to predict C_2 is given by

$$f_2(x_1, x_2) = \begin{cases} sgn\Big(p_2(x_1, x_2) + \beta_1(+1)\Big) & \text{if } p_1(x_1, x_2) > 0 \\ sgn\Big(p_2(x_1, x_2) + \beta_1(-1)\Big) & \text{if } p_1(x_1, x_2) < 0 \end{cases}$$

where p_1 and p_2 are the polynomials defined in Example 2. The decision boundaries of the two classes are shown in Fig. 5. We observe that the two boundaries are no longer independent; the decision boundary for the second class C_2 (dashed black line) depends on the predicted value of the first class C_1.

5 Binary Relevance vs. Chain Classifier

We denote the set of multi-valued decision functions representable by a NB chain classifier over $X_1, \ldots, X_n$ and by a multiple independent NB classifiers built as in the binary relevance method by $\mathcal{F}$ and $\mathcal{D}$, respectively. We can prove the following lemma.

Lemma 3.

$$|\mathcal{F}| > |\mathcal{D}|.$$

In other words, NB chain classifiers are more expressive than the NB binary relevance method.

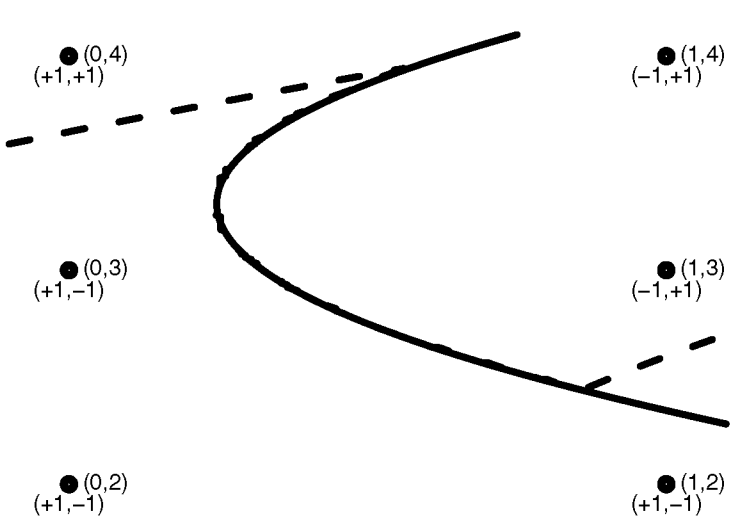

Fig. 5. Decision boundaries for the chain NB classifier in Example 3. The value of the predicted classes and the coordinates of the points are reported

Proof. We need only consider two class variables, since the result in the general case is proved analogously. If we define $\mathcal{P}_k$ for $k = 1, 2$ as in (4), we have that

$$\mathcal{D} = sgn(\mathcal{P}_1) \times sgn(\mathcal{P}_1) \subset sgn(\mathcal{P}_1) \times sgn(\mathcal{P}_2) = \mathcal{F}.$$

So, obviously, $|\mathcal{F}| \geq |\mathcal{D}|$. Thus to prove the lemma we just have to disprove the equality. Moreover, the VC dimension of $sgn(\mathcal{P}_1)$ (the cardinality of the maximum shattered subset) is equal to

$$d = \sum_{i=1}^{n} m_i - n + 1 < |\boldsymbol{\Omega}| = \prod_{i=1}^{n} m_i.$$

Then, by the definition of VC dimension, there exists $\boldsymbol{\Omega}_0 \subset \boldsymbol{\Omega}$ such that $|\boldsymbol{\Omega}_0| = d$ which is shattered by $sgn(\mathcal{P}_1)$. We now choose $\omega \in \boldsymbol{\Omega} \setminus \boldsymbol{\Omega}_0$ and find that there exists $p_0(\mathbf{x}) \in \mathcal{P}_1$ such that

$$p_0(\omega) < 0$$

and

$$p_0(\mathbf{x}) > 0 \text{ for every } \mathbf{x} \neq \omega.$$

Consider the bi-valued decision function $\mathbf{f} \in \mathcal{F}$ with the form

$$\mathbf{f} = \Big(sgn\big(p_0(\mathbf{x})\big), sgn\big(p_2(\mathbf{x}, sgn(p_0(\mathbf{x})))\big) \Big).$$

We observe from (4) that we have

$$p_2\big(\mathbf{x}, sgn(p_0(\mathbf{x}))\big) = \begin{cases} q(\mathbf{x}) + \beta_1(+1) & \text{if } p_0(\mathbf{x}) > 0 \\ q(\mathbf{x}) + \beta_1(-1) & \text{if } p_0(\mathbf{x}) < 0, \end{cases}$$

where $q(\mathbf{x}) \in sgn(\mathcal{P}_1)$. We now prove that the set of decision functions

$$\left\{ f_2 = sgn\left(p_2\left(\mathbf{x}, sgn(p_0(\mathbf{x}))\right)\right) \text{ s.t. } p_2 \in \mathcal{P}_2 \right\}$$

can shatter a subset of cardinality $d+1$ and thus cannot be represented by a NB classifier over predictors $X_1, \ldots, X_n$ alone. We have that $q(\mathbf{x}) + \beta_1(+1) \in \mathcal{P}_1$. Thus, by varying $q \in \mathcal{P}_1$, it can sign-represent every decision function over $\mathbf{\Omega}_0$ because of the choice of $\mathbf{\Omega}_0$. But the value of $f_2(\mathbf{x})$ over ω can be set independently by choosing $\beta_1(-1) \in \mathbb{R}$. So we have that choosing the polynomial $q \in \mathcal{P}_1$ and the real numbers $\beta_1(+1)$ and $\beta_1(-1)$, the defined decision functions $f_2(\mathbf{x})$ can shatter $\mathbf{\Omega}_0 \cup \{\omega\}$, a subset of cardinality $d+1$. □

Remark 6. As the number of classes grows, we see from (7) that the number of extra parameters, that is, the coefficients $\gamma(\ldots)$ that are added in a chain classifier model increase. Thus the chain NB classifier is considerably more expressive than a set of NB classifiers built with the binary relevance method.

From Remark 5, it follows that Lemma 3 could be extended to compare the expressive power of BAN chain classifiers versus the BAN binary relevance method, proving that BAN chain classifiers are in general more expressive than classifiers built using binary relevance.

Moreover we observe that changing the order of classes in which the classifier is built implies a change in the expressive power of the resulting multi-label classifier. In fact we find that the first class in NB chain classifiers is predicted as in the binary relevance method, and from Lemma 3, we get that the chain classifier is more expressive than binary relevance over the second variable. In general it is possible to prove that if the chain classifier for classes $C_1, \ldots, C_h$, is built with the class ordering $j_1, \ldots, j_h$, we have that the kth classifier for C_{j_k} is more expressive than all the previous classifiers in the chain. So, by changing the order of the classes, we obtain a multi-label classifier with different expressive power. This last observation led us to formulate an easy expressiveness-based heuristic to select an ordering for the chain classifier. We built h classifiers, one for each class as in the binary relevance method. We sorted the classifiers according to some evaluation metric and we used the resulting order to build a chain classifier. Precisely we started with the classifier with the best prediction performance and we ended with the worst predicted classes. In other words, we tried to employ the more expressive classifiers in the chain for the classes that were predicted worst by the binary relevance model. Moreover, if the BAN chain classifier is built as suggested in Remark 5, that is, by adding the previously predicted classes in a naive way, we find that the above heuristic introduces a low computational complexity: once the binary relevance model is built we have only to compute the additive coefficient, corresponding to the previously predicted classes to build the chain classifier. In real problems, where the coefficient of the models have to be estimated, overfitting could be an issue, specially with a limited number of observations available. In those cases we have to check that the increased expressive power of the chain model does not increase the classification errors. This could be achieved estimating the errors with cross-validation techniques [7] or using structural risk minimization [16].

6 Conclusions and Future Work

In this paper we have extended our previous results on the decision boundaries and expressive power of one-label BN classifiers to two types of BN multi-label classifiers: BAN classifiers built with binary relevance method and BAN chain classifiers. We have given theoretical grounds for why the binary relevance method provides models with poor expressive power and why this gets worst for larger numbers of classes. In both models we have expressed the multi-label decision boundaries in polynomial forms, and we have proved that chain classifiers provide more expressive models than the binary relevance method when the same type of BAN classifier is used as the base classifier.

As possible future research, we would like to extend our results to general multi-dimensional BN classifiers [4,12,2,13]. Multi-dimensional BN classifiers permit BN structures between classes and predictors, and so the multi-valued decision functions have to be found by a global maximum search over the possible class values. This fact does not permit to employ the same arguments used in this work. *Class-Bridge* decomposable multi-dimensional BN classifiers [2,3] could be easier to study due to the factorization of the maximization problem into a number of maximization problems in lower dimensional spaces.

Acknowledgments. This work has been partially supported by the Spanish Ministry of Economy and Competitiveness through the Cajal Blue Brain (C080020-09) and TIN2013-41592-P projects.

References

1. Bielza, C., Larrañaga, P.: Discrete Bayesian network classifier: A survey. ACM Computing Surveys 47(1) (in press, 2015)
2. Bielza, C., Li, G., Larrañaga, P.: Multi-dimensional classification with Bayesian networks. International Journal of Approximate Reasoning 52, 705–727 (2011)
3. Borchani, H., Bielza, C., Larrañaga, P.: Learning CB-decomposable multi-dimensional Bayesian network classifiers. In: Petri, M., Teemu, R., Tommi, J. (eds.) Proceedings of the 5th European Workshop on Probabilistic Graphical Models (PGM 2010), pp. 25–32. HIIT Publications (2010)
4. van der Gaag, L.C., de Waal, P.R.: Multi-dimensional Bayesian network classifiers. In: Studený, M., Vomlel, J. (eds.) Third European Workshop on Probabilistic Graphical Models, pp. 107–114 (2006)
5. Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian network classifiers. Machine Learning 29(2-3), 131–163 (1997)
6. Godbole, S., Sarawagi Discriminative, S.: methods for multi-labeled classification. In: Advances in Knowledge Discovery and Data Mining, pp. 22–30. Springer (2004)
7. Kelner, R., Lerner, B.: Learning bayesian network classifiers by risk minimization. International Journal of Approximate Reasoning 53(2), 248–272 (2012)
8. Keogh, E.J., Pazzani Learning, M.J.: the structure of augmented Bayesian classifiers. International Journal on Artificial Intelligence Tools 11(04), 587–601 (2002)
9. Minsky, M.: Steps toward artificial intelligence. In: Computers and Thought, pp. 406–450. McGraw-Hill (1961)

10. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers Inc. (1988)
11. Peot, M.A.: Geometric implications of the naive Bayes assumption. In: Eric, H., Finn, J. (eds.) Proceedings of the Twelfth International Conference on Uncertainty in Artificial Intelligence, pp. 414–419. Morgan Kaufmann Publishers Inc (1996)
12. de Waal, P.R., van der Gaag, L.C.: Inference and Learning in Multi-dimensional Bayesian Network Classifiers. In: Mellouli, K. (ed.) ECSQARU 2007. LNCS (LNAI), vol. 4724, pp. 501–511. Springer, Heidelberg (2007)
13. Read, J., Bielza, C., Larrañaga, P.: Multi-dimensional classification with super-classes. IEEE Transactions on Knowledge and Data Engineering (2013)
14. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier Chains for Multi-label Classification. In: Buntine, W., Grobelnik, M., Mladenić, D., Shawe-Taylor, J. (eds.) ECML PKDD 2009, Part II. LNCS, vol. 5782, pp. 254–269. Springer, Heidelberg (2009)
15. Sucar, L.E., Bielza, C., Morales, E.F., Hernandez-Leal, P., Zaragoza, J.H., Larrañaga, P.: Multi-label classification with Bayesian network-based chain classifiers. Pattern Recognition Letter 41, 14–22 (2014)
16. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer (1995)
17. Vapnik, V.N., Chervonenkis, A.: On the uniform convergence of relative frequencies of events to their probabilities. Theory of Probability and Its Applications 16(2), 264–280 (1971)
18. Varando, G., Bielza, C., Larrañaga, P.: Decision boundary for discrete Bayesian network classifiers. Technical Report UPM-ETSIINF/DIA/2014-1, Universidad Politecnica de Madrid (2014), `http://oa.upm.es/26003/`
19. Zhang, M.-L., Zhou, Z.-H.: A review on multi-label learning algorithms. IEEE Transactions on Knowledge and Data Engineering (in press, 2014)