Decision functions for chain classifiers based on Bayesian networks for multi-label classification

Gherardo Varando *, Concha Bielza, Pedro Larrañaga

Dept. of Artificial Intelligence, Universidad Politécnica de Madrid, Campus de Montegancedo, Madrid, Spain

A B S T R A C T

Multi-label classification problems require each instance to be assigned a subset of a defined set of labels. This problem is equivalent to finding a multi-valued decision function that predicts a vector of binary classes. In this paper we study the decision boundaries of two widely used approaches for building multi-label classifiers, when Bayesian network-augmented naive Bayes classifiers are used as base models: *Binary relevance method* and *chain classifiers*. In particular extending previous single-label results to multi-label chain classifiers, we find polynomial expressions for the multi-valued decision functions associated with these methods. We prove upper boundings on the expressive power of both methods and we prove that chain classifiers provide a more expressive model than the binary relevance method.

1. Introduction

We consider a multi-label classification problem [24,20] over categorical predictors, that is, mapping every instance $\mathbf{x} = (x_1, \ldots, x_n)$ to a subset of h labels:

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

where $\Omega_i \subset \mathbb{R}$, $|\Omega_i| = m_i < \infty$. As usual the problem 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 $\mathbf{c} = (c_1, \ldots, c_h) \in \{-1, +1\}^h$:

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

where $c_i = +1$ (-1) means that the ith label is present (absent) in the predicted label subset Y. 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 [8], diagnosis of multiple diseases from common symptoms and identification of multiple biological gene functions [3,23].

The easiest way to approach a multi-label classification problem is to divide it into a set of single-label classification problems (equivalent to binary classification problems). Each binary problem is then solved independently and thus h binary

* Corresponding author.
E-mail addresses: gherardo.varando@upm.es (G. Varando), mcbielza@fi.upm.es (C. Bielza), pedro.larranaga@fi.upm.es (P. Larrañaga).

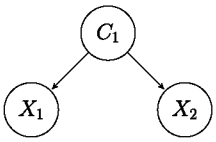

Fig. 1. Naive Bayes classifier structure in Example 1.

classifiers, one for each class variable C_i, are built. Each binary classifier is learned from predictor variables and C_i data only. At the end the results are combined to form multi-label prediction. Known as *binary relevance*, this method is easily implementable, has low computational complexity and is fully parallelizable. Therefore it is scalable to a large number of classes. However, it completely ignores dependencies among labels and generally does not represent the most likely set of labels.

Chain classifiers [18,6] relax the independence assumption by iteratively adding class dependencies in the binary relevance scheme. The kth classifier in the chain predicts class C_k from $X_1, \ldots, X_n, C_1, \ldots, C_{k-1}$. Sucar et al. [19] employed naive Bayes within chain classifiers.

In this paper, we study differences in the *expressive power* of these two methods when Bayesian network (BN) classifiers [1] are used. Expressive power of a classifier over categorical variables could be seen simply as the number of distinct decision functions that a given type of classifier induces.

In Varando et al. [22] the expressive power of one-dimensional binary, or one-label classifiers has been studied. In particular, the results of Minsky [11] and Peot [14] about the decision boundary of naive Bayes have been extended to a broader class of Bayesian network classifiers. A polynomial representation of the decision functions induced by Bayesian network-augmented naive Bayes classifier is described, and in absence of V-structures a stronger characterization is shown to hold. In this paper, we extend these results to multi-label classifiers. Moreover, we suggest some theoretical reasons why the simple binary relevance method can perform poorly when relationships among labels exist, and we prove that chain classifiers provide more expressive models. A broader chain classifiers class than in Varando et al. [21] is considered and studied extensively and a bounding on the expressive power of those models is proved. Moreover we present novel illustrative examples both about the one-dimensional results and about multi-label ones.

In Section 2 we review previous work on one-dimensional binary classifiers. We describe the binary relevance method and compute its expressive power in Section 3. We analyse chain classifiers in Section 4. In Section 5 we compare the two methods, proving that actually chain classifiers are more expressive than binary relevance and in Section 6 we present our conclusions and some ideas for future research.

2. Expressive power of one-dimensional BN classifiers

We report here previous results on the decision boundary and expressive power of one-label, or equivalently one-dimensional binary, BN classifiers [22]. We restrict to binary classifier and we can assume that the class variables takes its values on $\{-1, +1\}$. Classifiers where the class variable takes more than two values are more complex to study, the associated decision functions could be seen as combinations of binary decision functions and thus some of the results of this section could probably be extended. In the present work we prefer to remain in the binary case. Moreover binary classes are the variables needed to define multi-label classification problems.

In particular, we look at Bayesian network-augmented naive Bayes (BAN) classifiers [7].

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 $\mathcal{G}$ can be a general BN. We observe that every BAN classifier is determined by the predictor sub-graph $\mathcal{G}$, because the class variable C is superposed as parent of every variable of $\mathcal{G}$. As we focus only on Bayesian network, we will use the word graph to refer only to a directed acyclic graph, the structure of a Bayesian network (For general notations see Table 2).

For every BAN classifier, the induced decision function is

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

and $P(C = c, X_1 = x_1, \ldots, X_n = x_n)$ is factorized according to BN theory [13] as

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

where $\mathbf{X}_{\mathbf{pa}(i)}$ are the 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 $\mathbf{X}_{\mathbf{pa}(i)}$.

Example 1. Consider a naive Bayes classifier (structure in Fig. 1), that is, the simplest BAN, over predictor variables $X_1 \in \{0, 1, 2\}$, $X_2 \in \{0, 1\}$. In this case the joint probability over (C, X_1, X_2) is factorized as

Table 1
Conditional probability tables for X_1 and X_2 in Example 1.

$P(X_1\|C)$		X_1		
		0	1	2
C	-1	0.3	0.3	0.4
	$+1$	0.1	0.7	0.2

$P(X_2\|C)$		X_2	
		0	1
C	-1	0.5	0.5
	$+1$	0.1	0.9

$$P(C=c, X_1=x_1, X_2=x_2) = P(C=c)P(X_1=x_1|C=c)P(X_2=x_2|C=c).$$

We consider a uniform prior probability over the class $P(C=+1)=0.5$, $P(C=-1)=0.5$, and conditional probabilities tables as in Table 1.

The induced decision function $f^{NB}(x_1,x_2)$, defined in Equation (1), could be computed easily and it is exactly:

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

We describe decision functions through polynomial representations, in particular we use the following concept [12]:

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 $\mathbf{\Omega}$.

Where the sign function $sgn(t)$ is defined as,

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

Example 2. We consider $\mathbf{\Omega} = \{-1,2\} \times \{0,4\}$ and the decision function over $\mathbf{\Omega}$

$$f(x_1,x_2) = \begin{cases} +1 & \text{if } (x_1,x_2) = (-1,0),(2,0),(2,4) \\ -1 & \text{if } (x_1,x_2) = (-1,4). \end{cases}$$

We have that the polynomial $r(x_1,x_2) = 2x_1^2 - x_2 + 1$ sign-represents f over $\mathbf{\Omega}$, that is,

$$r(-1,0) = 2 > 0, \quad r(2,0) = 9 > 0, \quad r(2,4) = 5 > 0 \text{ and } r(-1,4) = -1 < 0.$$

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

$$\ell_j^{\Omega_i}(x) = \prod_{k \neq j} \frac{(x - \xi_i^k)}{(\xi_i^j - \xi_i^k)} \text{ for every } j = 1, \ldots, m_i \text{ and } x \in \mathbb{R}. \tag{2}$$

Example 3. The Lagrange basis polynomials over $\Omega = \{0,1,2,3\}$ are

$$\ell_1^{\Omega}(x) = \frac{(x-1)(x-2)(x-3)}{(0-1)(0-2)(0-3)} = \frac{(x-1)(x-2)(x-3)}{-6}$$
$$\ell_2^{\Omega}(x) = \frac{x(x-2)(x-3)}{1(1-2)(1-3)} = \frac{x(x-2)(x-3)}{2}$$
$$\ell_3^{\Omega}(x) = \frac{x(x-1)(x-3)}{2(2-1)(2-3)} = \frac{x(x-1)(x-3)}{-2}$$
$$\ell_4^{\Omega}(x) = \frac{x(x-1)(x-2)}{3(3-1)(3-2)} = \frac{x(x-1)(x-2)}{6}$$

We have the following result, that describes in polynomial form the decision function induced by a BAN classifier [22]:

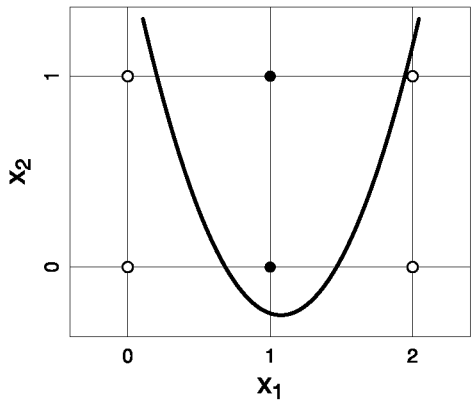

Fig. 2. Decision boundary of NB classifier in Example 1.

Lemma 1. *If f is the decision function induced by a BAN classifier for a binary 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 X_i does not have parents different from C, that is, $\mathbf{pa}(i) = \emptyset$.

In particular when the prior probability over the class is uniform, the coefficients $\beta_i(j|\mathbf{k})$ could be chosen as

$$\beta_i(j|\mathbf{k}) = \ln\left(\frac{P(X_i = \xi_i^j | X_s = \xi_s^{k_s},\ \forall s \in \mathbf{pa}(i), C = +1)}{P(X_i = \xi_i^j | X_s = \xi_s^{k_s},\ \forall s \in \mathbf{pa}(i), C = -1)}\right), \tag{3}$$

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

Example 4. We show now how to compute the polynomial that sign-represents the decision function of NB in Example 1, more examples could be found in Varando et al. [22]. For NB classifiers Equation (3) reduces to the more simpler form

$$\beta_i(j) = \ln\left(\frac{P(X_i = \xi_i^j | C = +1)}{P(X_i = \xi_i^j | C = -1)}\right) \tag{4}$$

since the NB predictors sub-graph has no connections among the nodes. Thus

$$\begin{aligned} p(x_1, x_2) = {} & \beta_1(1)\ell_1^{\Omega_1}(x_1) + \beta_1(2)\ell_2^{\Omega_1}(x_1) + \beta_1(3)\ell_3^{\Omega_1}(x_1) \\ & + \beta_2(1)\ell_1^{\Omega_2}(x_2) + \beta_2(2)\ell_2^{\Omega_2}(x_2), \end{aligned}$$

where $\ell_1^{\Omega_1}, \ell_2^{\Omega_1}, \ell_3^{\Omega_1}$ are the Lagrange basis polynomials over $\Omega_1 = \{0, 1, 2\}$ and $\ell_1^{\Omega_2}, \ell_2^{\Omega_2}$ are those over $\Omega_2 = \{0, 1\}$. Using the definition of $\beta_i(j)$ given in (4) and the values of Table 1 we obtain,

$$\begin{aligned} p(x_1, x_2) = {} & \ln\left(\frac{0.1}{0.3}\right)\ell_1^{\Omega_1}(x_1) + \ln\left(\frac{0.7}{0.3}\right)\ell_2^{\Omega_1}(x_1) + \ln\left(\frac{0.2}{0.4}\right)\ell_3^{\Omega_1}(x_1) \\ & + \ln\left(\frac{0.1}{0.5}\right)\ell_1^{\Omega_2}(x_2) + \ln\left(\frac{0.9}{0.5}\right)\ell_2^{\Omega_2}(x_2) \\ = {} & \ln\left(\frac{0.1}{0.3}\right)\frac{(x_1-1)(x_1-2)}{2} + \ln\left(\frac{0.7}{0.3}\right)\frac{x_1(x_1-2)}{-1} + \ln\left(\frac{0.2}{0.4}\right)\frac{x_1(x_1-1)}{2} \\ & + \ln\left(\frac{0.1}{0.5}\right)\frac{x_2-1}{-1} + \ln\left(\frac{0.9}{0.5}\right)\frac{x_2}{1} \end{aligned}$$

In Fig. 2 the decision boundary correspondent to $p(x_1, x_2)$ is shown.

Definition 2. Given a directed acyclic graph $\mathcal{G}$, a V-structure [5] in $\mathcal{G}$ is a triplet of nodes X_1, X_2, X_3 in $\mathcal{G}$ such that both X_1 and X_2 are parents of X_3 and X_1, X_2 are not directly connected in $\mathcal{G}$.

When the predictor sub-graph $\mathcal{G}$ does not contain V-structures, the inverse implication of Lemma 1 is shown to be true and the following theorem [22] holds.

Theorem 2. *Let $\mathcal{G}$ be a directed acyclic graph with nodes X_i, $i \in \{1, 2, \ldots, n\}$ and f a decision function over categorical predictor variables $X_i \in \Omega_i = \{\xi_i^1, \ldots, \xi_i^{m_i}\}$. Assume 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 $\mathcal{G}$.

Theorem 2 applies in a lot of practical cases as naive Bayes (NB) classifier [11], tree augmented naive Bayes (TAN) classifier [7] and super-parent one-dependence-estimator (SPODE) classifier [9], because the corresponding predictor sub-graphs do not contain V-structures.

Theorem 2 is useful because it completely characterizes the set of decision functions induced by BAN with a given structures $\mathcal{G}$ with no V-structures. In particular, the theorem implies that when $\mathcal{G}$ does not contain V-structures the family of polynomials $\mathcal{P}_{\mathcal{G}}$, defined as

$$\mathcal{P}_{\mathcal{G}} = \left\{ r(\mathbf{x}) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \ell_j^{\Omega_i}(x_i) \sum_{\mathbf{k} \in \mathbb{M}_i} \beta_i(j|\mathbf{k}) \prod_{s \in \mathbf{pa}(i)} \ell_{k_s}^{\Omega_s}(x_s) \text{ s.t. } \beta_i(j|\mathbf{k}) \in \mathbb{R} \right\} \tag{5}$$

sign-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 a graph without any arc, we have that

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

is exactly the set of polynomials used to sign-represent decision functions induced by NB classifiers as in Theorem 2.

The set $\mathcal{P}_{\mathcal{G}}$, when $\mathcal{G}$ does not contain V-structure, is a vectorial space of dimension

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

As in Varando et al. [22], it is useful to consider spaces $\mathcal{P}_{\mathcal{G}}$ as subspaces of the vector space of polynomials that can interpolate every function over $\boldsymbol{\Omega}$. That is, the space

$$\mathcal{P}_{FBN} = \left\{ \sum_{k \in \mathbb{M}} \gamma_{\mathbf{k}} \delta_{\mathbf{k}}(\mathbf{x}) \text{ s.t. } \gamma_{\mathbf{k}} \in \mathbb{R} \right\},$$

where *FBN* stands for full Bayesian classifier, $\mathbb{M} = \times_{i=1}^{n} \{1, \ldots, m_i\}$ and $\delta_{\mathbf{k}}(\mathbf{x}) = \prod_{i=1}^{n} \ell_{k_i}^{\Omega_i}(x_i)$ are the polynomials that interpolate the Dirac's delta over $\boldsymbol{\Omega}$, that is,

$$\forall \mathbf{k} \in \mathbb{M}, \quad \delta_{\mathbf{k}}(\mathbf{x}) = \prod_{i=1}^{n} \ell_{k_i}^{\Omega_i}(x_i) = \begin{cases} 1 \text{ if } \mathbf{x} = (\xi_1^{k_1}, \ldots, \xi_n^{k_n}) \\ 0 \text{ if } \mathbf{x} \in \boldsymbol{\Omega} \setminus \{(\xi_1^{k_1}, \ldots, \xi_n^{k_n})\}. \end{cases}$$

Observe that in this case $\mathcal{P}_{FBN} = \mathcal{P}_{\mathcal{G}}$, where $\mathcal{G}$ is a full Bayesian network (FBN) over the predictors, it is a Bayesian network with the maximum possible number of arcs. The polynomials $\{\delta_{\mathbf{k}}(\mathbf{x})\}_{\mathbf{k} \in \mathbb{M}}$ form a basis of $\mathcal{P}_{FBN}$ (we show an example of the basis construction in Example 5). Therefore the dimension of $\mathcal{P}_{FBN}$ is equal to $\prod_{i=1}^{n} m_i$, where m_i is the number of values the ith predictor assumes.

Obviously we have that for every Bayesian network structure $\mathcal{G}$ over predictor variables $X_1, \ldots, X_n$,

$$\mathcal{P}_{NB} \subseteq \mathcal{P}_{\mathcal{G}} \subseteq \mathcal{P}_{FBN},$$

and

$$sgn(\mathcal{P}_{NB}) \subseteq sgn(\mathcal{P}_{\mathcal{G}}) \subseteq sgn(\mathcal{P}_{FBN}) = \{-1,+1\}^{\Omega}.$$

We can now define the interpolating polynomial of every function over $\boldsymbol{\Omega}$ as follows,

Definition 3. Given a function f over $\boldsymbol{\Omega} = \Omega_1 \times \cdots \times \Omega_n$, with $\Omega_i = \{\xi_i^1, \ldots, \xi_i^{m_i}\}$, we define the interpolating polynomial

$$\pi_f(\mathbf{x}) = \sum_{\mathbf{k} \in \mathbb{M}} f\left(\xi_1^{k_1}, \ldots, \xi_n^{k_n}\right) \delta_{\mathbf{k}}(\mathbf{x}),$$

where $\mathbb{M} = \times_{i=1}^{n} \{1, \ldots, m_i\}$. And we have $\pi_f(\mathbf{x}) = f(\mathbf{x})$ for every $\mathbf{x} \in \boldsymbol{\Omega}$.

Polynomial π_f is just a way to *see* function f as an element of the vectorial space $\mathcal{P}_{FBN}$.

Example 5. We show here an example of interpolating polynomial, we consider f the following decision function over $\boldsymbol{\Omega} = \{0,1\} \times \{4,6\}$,

$$f(x_1, x_2) = \begin{cases} -1 & \text{if } (x_1, x_2) \in \{(0,4),(1,6)\} \\ +1 & \text{if } (x_1, x_2) \in \{(0,6),(1,4)\}. \end{cases}$$

The Lagrange basis over $\{0,1\}$ is composed by,

$$\ell_1^{\{0,1\}}(x_1) = 1 - x_1 \quad \ell_2^{\{0,1\}}(x_1) = x_1,$$

and the Lagrange basis over $\{4,6\}$ is composed by,

$$\ell_1^{\{4,6\}}(x_2) = \frac{6 - x_2}{2} \quad \ell_2^{\{4,6\}}(x_2) = \frac{x_2 - 4}{2}.$$

Thus the basis of $\mathcal{P}_{FBN}$, $\{\delta_{\mathbf{k}}(\mathbf{x})\}_{\mathbf{k} \in \mathbb{M}}$ in this case is formed by the following four polynomials in (x_1, x_2).

$$\delta_{(1,1)} = \ell_1^{\{0,1\}}(x_1)\ell_1^{\{4,6\}}(x_2) = \frac{(1 - x_1)(6 - x_2)}{2} \quad \delta_{(1,2)} = \frac{(1 - x_1)(x_2 - 4)}{2},$$
$$\delta_{(2,1)} = \frac{x_1(6 - x_2)}{2} \quad \delta_{(2,2)} = \frac{x_1(x_2 - 4)}{2}.$$

We now compute the interpolating polynomial directly with Definition 5,

$$\begin{aligned} \pi_f(x_1, x_2) &= f(0,4)\delta_{1,1} + f(0,6)\delta_{1,2} + f(1,4)\delta_{2,1} + f(1,6)\delta_{2,2} \\ &= \frac{(2x_1 - 1)(10 - 2x_2)}{2}. \end{aligned}$$

As we can see from substitution, $\pi_f = f$ over $\boldsymbol{\Omega}$.

Remark 2. We observe that if $f : \boldsymbol{\Omega} \mapsto \{-1,+1\}$ is a decision function, obviously the interpolating polynomial π_f sign-represents f. But there exist a lot of polynomials that sign-represent f without interpolating it over $\boldsymbol{\Omega}$. A polynomial sign-represents a decision function if it agrees on the sign of f over $\boldsymbol{\Omega}$, while interpolating refers to actually having the same values over the points of $\boldsymbol{\Omega}$. For example consider $p \in \mathcal{P}_{NB}$ and $f = sgn(p)$, thus f is induced by a naive Bayes classifier. Consider now π_f. Could it be interesting to know if $\pi_f \in \mathcal{P}_{NB}$? This question is important when studying the expressive power of chain classifiers, and in Lemma 6 we will answer it completely.

Thanks to Theorem 2 it is possible to place an upper bound (Corollary 3) on the number of decision functions representable by BAN classifiers without V-structures [22].

Corollary 3. *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*

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

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

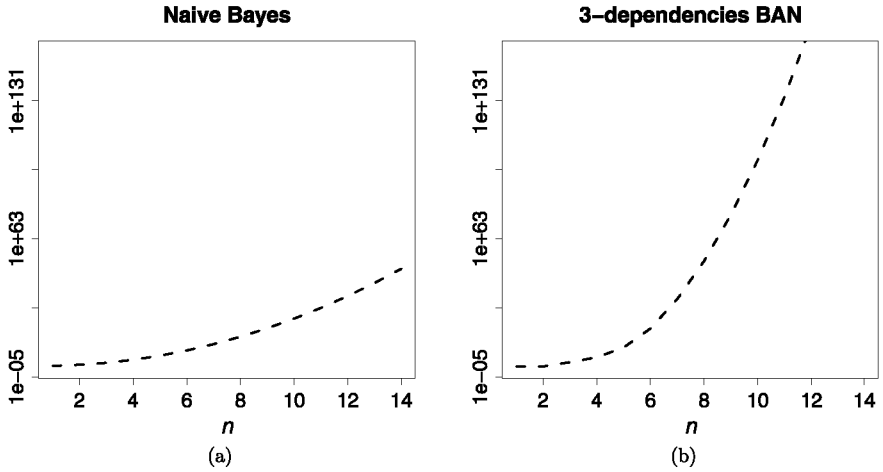

Fig. 3. Total number of decision functions over n binary predictors (solid grey) and the bounding $C(M,d)$ of Corollary 3 (dashed black) for NB classifiers (a) and for 3-dependence BAN classifiers (b).

Table 2
Table of notations.

X_i	categorical predictor variable
C, C_i	binary class variable
Ω_i	i-th predictor values space
m_i	cardinality of Ω_i
ξ_i^j	j-th element of Ω_i
$\boldsymbol{\Omega}$	$\times_{i=1}^n \Omega_i$ predictors space
$\mathbb{M}$	$\times_{i=1}^n \{1,\ldots,m_i\}$ indexes of predictors space $\boldsymbol{\Omega}$
$\mathbf{X}_{\mathbf{pa}(i)}$	vector of parents of X_i
$\mathbf{pa}(i)$	subset of $\{1,\ldots,n\}$ relative to the parents of X_i
$\mathbb{M}_i$	$\times_{s\in\mathbf{pa}(i)}\{1,\ldots,m_s\}$ configuration of $\mathbf{X}_{\mathbf{pa}(i)}$
$\mathcal{G}, \mathcal{G}_i, \mathcal{H}_i$	predictor sub-graph
$\mathcal{P}_{NB}$	space of polynomials sign-representing NB classifiers
$\mathcal{P}_{\mathcal{G}}$	space of polynomials sign-representing BAN with sub-graph $\mathcal{G}$
$\mathcal{P}_{FBN}$	space of polynomials sign-representing all classifiers over $\boldsymbol{\Omega}$
$\ell_j^{\Omega_i}$	j-th Lagrange polynomial over Ω_i
$\delta_{\mathbf{k}}$	$\mathbf{k}$-th polynomial interpolating the Dirac's delta over $\boldsymbol{\Omega}$
π_f	polynomial interpolating f over $\boldsymbol{\Omega}$

Remark 3. If $\boldsymbol{\Omega} = \Omega_1 \times \cdots \times \Omega_n$, we observe that $|\{-1,+1\}^{\boldsymbol{\Omega}}| = 2^{|\boldsymbol{\Omega}|} = 2^M$. Thus Corollary 3 implies that in the case of the NB classifier the quotient of the number of decision functions representable by NB classifiers over 2^M becomes vanishingly small as the number n of predictors increase. Fig. 3a shows the total number of decision functions, $2^{|\boldsymbol{\Omega}|}$ (solid grey) and the bounding of Corollary 3 for NB classifiers with n binary predictors, $C(M,d)$ (dashed black). Note that in this case $d = \sum_{i=1}^n (m_i - 1) + 1 = n + 1$. 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 $\boldsymbol{\Omega}$.

Remark 3 could be extended to every type of BAN classifier, such that for every variable X_i, the number of parents is bounded (Corollary 19 in Varando et al. [22]), that is, $|\mathbf{pa}(i)| < K$. Fig. 3b shows the total number of decision functions (solid grey) and the bounding of Corollary 3 (dashed black) for a BAN structure such that $|\mathbf{pa}(i)| \leq 3$.

Remark 4. When the predictor sub-graph $\mathcal{G}$ of a BAN classifier contains V-structures, Lemma 1 is still valid and there exists a polynomial sign-representing the induced decision function. The problem is that the associated family of polynomials is not a linear space as in (5), thus is not possible to employ the same techniques as in Varando et al. [22] and thus prove the bounding in Corollary 3.

3. BAN binary relevance classifiers

We consider the binary relevance method built upon BAN classifiers as base models, that is, for every class variable C_i we learn a BAN classifier with predictor sub-graph $\mathcal{G}_i$. Thus we actually transform our multi-label problem into a number of single binary-class problems. The results of last section are then straightforwardly applied.

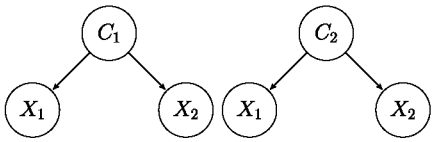

Fig. 4. Two NB classifiers in Example 6.

Table 3
Conditional probability tables in Example 6 for the NB of C_1.

$P(X_1\|C_1)$		X_1	
		0	1
C_1	-1	0.25	0.75
	$+1$	0.5	0.5

$P(X_2\|C_1)$		X_2		
		2	3	4
C_1	-1	0.1	0.7	0.2
	$+1$	0.3	0.5	0.2

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

$$p_1(\mathbf{x}) \in \mathcal{P}_{\mathcal{G}_1}, \ldots, p_h(\mathbf{x}) \in \mathcal{P}_{\mathcal{G}_h},$$

such that $f_k(\mathbf{x}) = sgn(p_k(\mathbf{x}))$ for every $k \in \{1, \ldots, h\}$. We have then that the multi-valued decision function has a polynomial representation as,

$$\mathbf{f}(\mathbf{x}) = \big(sgn(p_1(\mathbf{x})), \ldots, sgn(p_h(\mathbf{x}))\big).$$

When we also assume that the predictor sub-graphs $\mathcal{G}_1, \ldots, \mathcal{G}_h$ contain no V-structures, we have that, for every single binary-class problem, Theorem 2 apply. Thus, in Lemma 4, we bound the number of multi-valued decision functions representable by the BAN binary relevance method, when the predictor sub-graphs $\{\mathcal{G}_k\}_{k=1}^h$ do not contain V-structures.

Lemma 4. *Consider h BAN classifiers to predict h binary classes. Suppose that the predictor sub-graphs are $\mathcal{G}_1, \ldots, \mathcal{G}_h$ respectively and they contain no V-structures. We have that $N(\mathcal{G}_1, \ldots, \mathcal{G}_h)$, the number of h-valued decision functions representable by the BAN binary relevance method, satisfies*

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

where $C(M,d) = 2\sum_{k=0}^{d-1}\binom{M-1}{k}$, $d_k = \sum_{i=1}^{n}\Big((m_i - 1)\prod_{s\in\mathbf{pa}_k(i)} m_s\Big) + 1$, $\mathbf{pa}_k(i)$ is the set of X_i parents in $\mathcal{G}_k$ and $M = \prod_{i=1}^{n} m_i$.

Proof. The proof is a straightforward application of Corollary 3. □

Remark 5. We consider now, for visualization purposes, a simpler version of the above models. In particular when the predictors sub-graphs are all the same, that is, $\mathcal{G}_j = \mathcal{G}$. 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

$$\frac{N(\mathcal{G}_1, \ldots, \mathcal{G}_h)}{2^{hM}} \leq \left(\frac{C(M,d)}{2^M}\right)^h.$$

Thus, as in Remark 3, 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.

Example 6. 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. 4. Next, we list the conditional probability tables for both classifiers (Tables 3 and 4). Moreover, we consider uniform prior probabilities for both classes C_1 and C_2.

From the representation of Theorem 2 we have that there exist two polynomials p_1, p_2 that sign-represent the decision functions induced by the two NB classifiers

Table 4
Conditional probability tables in Example 6 for the NB of C_2.

$P(X_1\|C_2)$		X_1	
		0	1
C_2	-1	0.4	0.6
	$+1$	0.7	0.3

$P(X_2\|C_2)$		X_2		
		2	3	4
C_2	-1	0.6	0.2	0.2
	$+1$	0.1	0.1	0.8

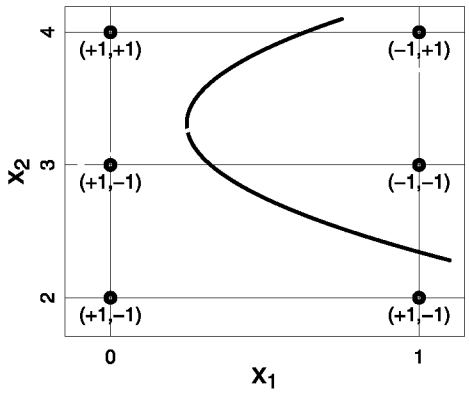

Fig. 5. Decision boundaries for the two NB classifiers in Example 6, black for C_1 and grey for C_2. The value of the predicted classes is reported.

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

and

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

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. Fig. 5 shows the decision boundaries of the two classifiers (black for C_1 and grey 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. The value of the respective predicted class changes when one of the decision boundaries is crossed.

4. BAN 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

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

In the predicting phase we will then use the predictor values and the previous predicted classes values $\hat{c}_1,\ldots,\hat{c}_{k-1}$ to predict class C_k. From Lemma 1 we have that there exist h polynomials $p_1,\ldots,p_h$

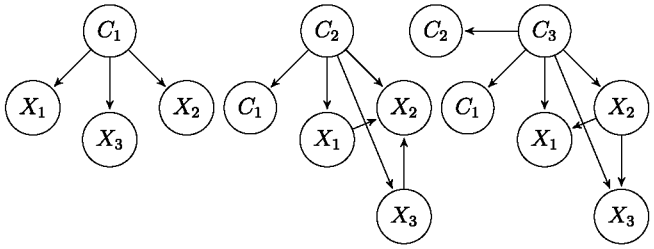

Fig. 6. Example of naive BAN chain classifier with three classes and three predictor variables.

$$p_k(\mathbf{x}, \hat{c}_1, \ldots, \hat{c}_{k-1}) : \mathbb{R}^{n+k-1} \to \mathbb{R}$$
$$p_k \in \mathcal{P}_{\mathcal{G}_k},$$

such that, if $\mathbf{f} = (f_1, \ldots, f_h)$ is the multi-valued decision function associated with a chain classifier we have that,

$$f_k(\mathbf{x}) = sgn(p_k(\mathbf{x}, f_1(\mathbf{x}), \ldots, f_{k-1}(\mathbf{x}))) = sgn(p_k(\mathbf{x}, \pi_{f_1}(\mathbf{x}), \ldots, \pi_{f_{k-1}}(\mathbf{x})) \tag{6}$$

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

From now on we will focus on a particular and simpler form of BAN chain classifier, where the previous predicted classes are present in a naive way in the predictor sub-graph. That is, $C_1, \ldots, C_{k-1}$ are not connected among them neither with other predictors in the sub-graph $\mathcal{G}_k$. We refer to this kind of chain classifier as *naive BAN chain classifier*, we show an example in Fig. 6. As we will see those naive models have a more simpler representation of multi-valued decision functions and permit a deeper analysis. We observe that more complex chain models could be addressed in a similar way, using the interpolating polynomials to represent the decision functions of the already predicted classes. In more complex model however the analysis of the decision function is more difficult and not all the following results can be extended directly.

For a naive BAN chain classifier for $C_1, \ldots, C_h$, over $X_1, \ldots, X_n$ we denote by $\mathcal{H}_k$ the sub-graph of the k-th BAN restricted to the original predictors $X_1, \ldots, X_n$.

Since classes C_j are binary, expanding Equation (6) we obtain the following sign-representation of the k-th decision function in a naive BAN chain classifier:

$$\begin{aligned}
f_k(\mathbf{x}) &= sgn(p_k(\mathbf{x}, \pi_{f_1}(\mathbf{x}), \ldots, \pi_{f_{k-1}}(\mathbf{x})) \\
&= sgn\left(\sum_{i=1}^{n}\sum_{j=1}^{m_i} \ell_j^{\Omega_i}(x_i) \sum_{\mathbf{k}\in\mathbb{M}_i} \beta_i(j|\mathbf{k}) \prod_{s\in\mathbf{pa}(i)} \ell_{k_s}^{\Omega_s}(x_s)\right. \\
&\quad \left. + \sum_{j=1}^{k-1}\left[\beta_j(-1)\ell_{-1}^{\{-1,+1\}}(\hat{c}_j) + \beta_j(+1)\ell_{+1}^{\{-1,+1\}}(\hat{c}_j)\right]\right) \\
&= sgn\left(\hat{q}_k(\mathbf{x}) + \sum_{j=1}^{k-1}\left[\beta_j(-1)\ell_{-1}^{\{-1,+1\}}(\hat{c}_j) + \beta_j(+1)\ell_{+1}^{\{-1,+1\}}(\hat{c}_j)\right]\right),
\end{aligned}$$

where $\hat{q}_k \in \mathcal{P}_{\mathcal{H}_k}$, $\hat{c}_j = f_j(\mathbf{x}) = \pi_{f_j}(\mathbf{x})$ is the predicted value of the previous classifier expressed by the interpolating polynomial as a function of $\mathbf{x}$, $\ell_{-1}^{\{-1,+1\}}(c) = \frac{c-1}{-2}$ and $\ell_{+1}^{\{-1,+1\}}(c) = \frac{c+1}{2}$ are the Lagrange basis polynomials over $\{-1, +1\}$ and $\beta_j(c) = \ln\left(\frac{P(C_j=c|C_k=+1)}{P(C_j=c|C_k=-1)}\right)$. Rearranging the terms in the sum we obtain that the following polynomial sign-represents f_k,

$$q_k(\mathbf{x}) = \hat{q}_k(\mathbf{x}) + \sum_{j=1}^{k-1}\left(a_j\pi_{f_j}(\mathbf{x}) + b_j\right), \tag{7}$$

where f_j are the decision functions of the previous predicted class in the chain, $\hat{q}_k$ is the polynomial related to the sub-graph $\mathcal{H}_k$ as in Theorem 2 and

$$a_j = \frac{1}{2}\ln\left(\frac{P(C_j=+1|C_k=+1)P(C_j=-1|C_k=-1)}{P(C_j=+1|C_k=-1)P(C_j=-1|C_k=+1)}\right) \tag{8}$$

$$b_j = \frac{1}{2}\ln\left(\frac{P(C_j=+1|C_k=+1)P(C_j=-1|C_k=+1)}{P(C_j=+1|C_k=-1)P(C_j=-1|C_k=-1)}\right) \tag{9}$$

Observe that we can omit constants b_j in Equation (7) if analysing the expressive power. In fact constants could be included in the polynomial $\hat{q}_k$ using elementary properties of Lagrange basis polynomials, see Varando et al. [22]. The following lemma describes the set of decision functions induced by the kth step of the naive BAN chain classifier.

Lemma 5. *Consider a multi-label classification problem over predictors $X_1, \ldots, X_n$ and a naive BAN chain classifier with predictor sub-graphs $\mathcal{H}_1, \ldots, \mathcal{H}_h$ for classes ordered as $C_1, \ldots, C_h$. Assume that the predictor sub graphs do not contains V-structures. For every $k \in \{2, \ldots, h\}$ we have that, if $f_1, \ldots, f_{k-1}$ are the decision functions for $C_1, \ldots, C_{k-1}$ respectively, then the following set of polynomials sign-represent every decision function for class C_k,*

$$\mathcal{P}_{\mathcal{H}_k} + < \pi_{f_1}, \ldots, \pi_{f_{k-1}} >,$$

where $\pi_{f_1}, \ldots, \pi_{f_{k-1}}$ are the interpolating polynomials, $< \ldots >$ denotes the span of the included vectors and the sum is intended as the sum of two vectorial space, that is, the vectorial space which includes all the possible sum of elements of the two spaces, $\mathcal{P}_{\mathcal{H}_k}$ and $< \pi_{f_1}, \ldots, \pi_{f_{k-1}} >$.

Proof. The proof of the result is just an application of Theorem 2 and Equation (7). □

We have furthermore, that the set $sgn\left(\mathcal{P}_{\mathcal{H}_k} + < \pi_{f_1}, \ldots, \pi_{f_{k-1}} >\right)$ is equal to the set of decision functions representable by the k-th BAN classifier of the naive BAN chain classifier if the graphs $\mathcal{H}_k$ do not contain V-structures. Intuitively, from an expressive-power point of view, we have the addition of the previous predicted classes in the kth step of a naive BAN chain classifier being the equivalent to the *enrichment* of the space of polynomials $\mathcal{P}_{\mathcal{H}_k}$, related to the original predictors, by a subspace generated by the interpolating polynomial of the previous induced decision functions. To analyse if and how the enlarged space is indeed a bigger space, in other words, that it has a grater dimension, we have to understand when an interpolating polynomial π_f does not belong to a polynomial space of the type $\mathcal{P}_{\mathcal{G}}$ for some graph $\mathcal{G}$. Thus, in this case, adding $< \pi_f >$ to $\mathcal{P}_{\mathcal{G}}$ will actually increase the dimension.

First of all we define the set of relevant variables for a given decision function.

Definition 4. Given a decision function

$$f(x_1, \ldots, x_n) : \mathbf{\Omega} = \Omega_1 \times \cdots \times \Omega_n \to \{-1, +1\}$$

we say that a variable X_i is irrelevant for f if

$$f(x_1, \ldots, x_n) = g(\mathbf{x}_{-i}) = g(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \quad \forall (x_1, \ldots, x_n) \in \mathbf{\Omega},$$

where we denote with $\mathbf{x}_{-i}$ the $(n-1)$ dimensional vector obtained from $\mathbf{x}$ by eliminating the i-th component (in general $\mathbf{x}_{-I}$ will denote the vector obtained eliminating the components indexed by I). A variable is said to be relevant for f if it is not irrelevant, and we indicate with $\mathcal{X}(f)$ the set of relevant variables for f.

As we will see relevant variables are important in order to determine if the interpolating polynomial of a given decision function belongs or not to some polynomial space. In real applications the task of finding relevant variables of a decision function is computationally expensive and moreover in reality we usually know just an estimation of a decision function or its value on a set of random points. The presented analysis is thus intended as a theoretical analysis.

Example 7. We show some example of decision functions and their respective set of relevant variables.

1. If f_1 is a decision function over $\{0, 1, 2\} \times \{-3, -2\}$, such that
$$f_1(x_1, x_2) = \begin{cases} -1 & \text{if } (x_1, x_2) = (0, -3) \text{ or } (0, -2) \\ +1 & \text{otherwise.} \end{cases}$$
Then obviously $f_1(x_1, x_2) = g(x_1)$, where $g(x_1) = -1$ if $x_1 = 0$ and $+1$ otherwise. Thus X_2 is irrelevant for f and $\mathcal{X}(f_1) = \{X_1\}$.
2. If f_2 is the xor-function over $\{0, 1\} \times \{0, 1\}$, defined as follows
$$f_2(x_1, x_2) = \begin{cases} -1 & \text{if } (x_1, x_2) = (0, 0) \text{ or } (1, 1) \\ +1 & \text{if } (x_1, x_2) = (0, 1) \text{ or } (1, 0). \end{cases}$$
Then $\mathcal{X}(f_2) = \{X_1, X_2\}$ and f_2 do not have irrelevant variables.
3. If f_3 is the function over $\{0, 1\} \times \{0, 1\}$ such that,
$$f_3(x_1, x_2) = \begin{cases} -1 & \text{if } (x_1, x_2) = (0, 0) \\ +1 & \text{otherwise.} \end{cases}$$
Then also in this case $\mathcal{X}(f_3) = \{X_1, X_2\}$.

We can now state the following result about the interpolating polynomial of decision functions.

Lemma 6. *Consider, for a graph $\mathcal{G}$ without V-structures, and categorical predictors $X_1, \ldots, X_n$, the space of polynomials $\mathcal{P}_{\mathcal{G}}$ defined in (5). For every decision function f we have that,*

$$\pi_f \in \mathcal{P}_{\mathcal{G}} \Leftrightarrow \textit{ Variables } \mathcal{X}(f) \textit{ are completely connected in } \mathcal{G},$$

where a set of variables is said to be completely connected in graph $\mathcal{G}$ if for every couple of variables in the set, they are in a parent–child relationship in the graph $\mathcal{G}$. In other words, it is not possible to add any arcs among this set of variables respecting the acyclic property of the graph.

Proof. If the relevant variables for f are completely connected in the graph $\mathcal{G}$, then we have that the polynomials in $\mathcal{P}_{\mathcal{G}}$ could interpolate, over $\boldsymbol{\Omega}$ any function of variables in $\mathcal{X}(f)$ only. In particular, there exists a polynomial $p(\mathbf{x}) \in \mathcal{P}_{\mathcal{G}}$ such that $f(\mathbf{x}) = p(\mathbf{x})$, $\forall \mathbf{x} \in \boldsymbol{\Omega}$ and thus $\pi_f \in \mathcal{P}_{\mathcal{G}}$.

To prove the other implication we observe that if two variable X_i and X_j are not directly connected in the graph $\mathcal{G}$, each polynomial $p(\mathbf{x}) \in \mathcal{P}_{\mathcal{G}}$ could be split into

$$p(\mathbf{x}) = p_1(\mathbf{x}_{-\{i,j\}}, x_i) + p_2(\mathbf{x}_{-\{i,j\}}, x_j). \tag{10}$$

To prove the above equality we just observe that each polynomial p in $\mathcal{P}_{\mathcal{G}}$ has the following expression

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

Thus two variables appear in the same product of different Lagrange polynomial basis if and only if they are directly connected, that is, if and only if one variable belongs to the parents of the other. It is clear now that the sum in Equation (10) is therefore valid.

So we have only to prove that a decision function f with two relevant variables $X_1 \in \Omega_1, X_2 \in \Omega_2$ could not be equal, over $\Omega_1 \times \Omega_2$, to the sum of two functions $p_1(x_1)$ and $p_2(x_2)$. Since X_1 and X_2 are relevant variable, there exist $s, s' \in \Omega_1$ and $t, t' \in \Omega_2$ such that,

$$f(s,t) = -f(s,t') \quad \text{and} \quad f(s,t) = -f(s',t)$$

Suppose $f(x_1, x_2) = p_1(x_1) + p_2(x_2)$, then we have,

$$\begin{aligned} f(s',t') &= p_1(s') + p_2(t') \\ &= p_1(s') + p_2(t) + p_1(s) + p_2(t') - p_1(s) - p_2(t) \\ &= f(s',t) + f(s,t') - f(s,t) = -3f(s,t). \end{aligned}$$

And we get $|f(s',t')| \neq 1$ which is absurd given that f is a decision function. □

We return to points 2 and 3 of Example 7. In both cases the functions f_2 and f_3 do not have irrelevant variables, thus from Lemma 6 we have that $\pi_{f_2}, \pi_{f_3} \notin \mathcal{P}_{NB}$. But $f_2 \notin sgn(\mathcal{P}_{NB})$ (see the results of [10]) while $f_3 \in sgn(\mathcal{P}_{NB})$ (see proof of Theorem 8). As observed in Remark 2, there is a clear difference between sign-representing and interpolating.

Thanks to Lemma 6, we have the following result.

Lemma 7. *Consider a multi-label classification problem over categorical predictors $X_1, \ldots, X_n$, for binary classes ordered as $C_1, \ldots, C_h$. Given a sequence of predictor sub-graphs $\mathcal{H}_1, \ldots, \mathcal{H}_h$ without V-structures, let us consider $\mathbf{f} = (f_1, \ldots, f_h)$ the h-valued decision functions of the corresponding naive BAN chain classifier. Then, for every $1 \leq k \leq h$, we have that*

$$\left| sgn\left(\mathcal{P}_{\mathcal{H}_k} + < \pi_{f_1}, \ldots, \pi_{f_{k-1}} >\right) \right| \leq C(M, d_k + s) \leq C(M, d_k + k - 1),$$

where $M = |\boldsymbol{\Omega}| = \prod_{i=1}^{n} m_i$, $d_k = \dim(\mathcal{P}_{\mathcal{H}_k})$, and s is equal to the number of functions among $f_1, \ldots, f_{k-1}$ such that their relevant variables are not completely connected in $\mathcal{H}_k$.

Proof. Suppose, $f_{i_1}, \ldots, f_{i_s}$ are the decision functions among $f_1, \ldots, f_{k-1}$ such that their relevant variables are not completely connected in $\mathcal{H}_k$. From Lemma 6 we have that,

$$\pi_{f_{i_1}}, \ldots, \pi_{f_{i_s}} \notin \mathcal{P}_{\mathcal{H}_k},$$

and that

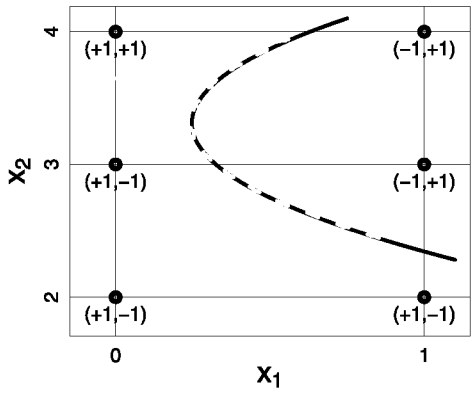

Fig. 7. Decision boundaries for the chain NB classifier in Example 8. The value of the predicted classes is reported.

$\pi_{f_i} \in \mathcal{P}_{\mathcal{H}_k}$ for every $i \in \{1, \ldots, k-1\} \setminus \{i_1, \ldots, i_s\}$.

Thus we have

$$\mathcal{P}_{\mathcal{H}_k} + < \pi_{f_1}, \ldots, \pi_{f_{k-1}} > = \mathcal{P}_{\mathcal{H}_k} + < \pi_{f_{i_1}}, \ldots, \pi_{f_{i_s}} >,$$

and so

$$\dim\left(\mathcal{P}_{\mathcal{H}_k} + < \pi_{f_1}, \ldots, \pi_{f_{k-1}} >\right) \leq d_k + s \leq d_k + k - 1.$$

Analogously to Corollary 3 we have the corresponding bounding. □

Remark 6. We observe that changing the order of classes in which the chain classifier is built implies a change in the expressive power of the resulting multi-label classifier. If the chain classifier is built with the class ordering $C_1, \ldots, C_h$, we have that the kth classifier for C_k is *more expressive* than all the previous classifiers in the chain. In fact, from Equation (7), we have that if f is a decision function representable by the jth step of the chain classifier, then f is representable by every successive steps of the chain classifier.

Example 8. We use a NB chain classifier over the prediction problems of Example 6. The NB classifier for predicting class C_1 is the same as in Example 6 (see Fig. 4 left and Table 3). The predictors of the NB classifier for predicting C_2 now include C_1. We consider the same conditional probability tables as in Example 6 (Tables 3 and 4). Moreover we have to specify the conditional probabilities of C_1 given C_2 in the NB that predicts C_2. We set

$$P(C_1 = +1 | C_2 = +1) = 0.3 \text{ and } P(C_1 = -1 | C_2 = +1) = 0.7$$
$$P(C_1 = +1 | C_2 = -1) = 0.9 \text{ and } P(C_1 = -1 | C_2 = -1) = 0.1$$

And, thus, coefficients a_1 and b_1 as defined in (8) and (9) are given by

$$a_1 = \frac{1}{2} \ln\left(\frac{0.3 \times 0.1}{0.9 \times 0.7}\right) \quad \text{and} \quad b_1 = \frac{1}{2} \ln\left(\frac{0.3 \times 0.7}{0.9 \times 0.1}\right).$$

We have that the decision function to predict C_2 is sign-represented by

$$q_2(x_1, x_2) = p_2(x_1, x_2) + a_1 \pi_{f_1}(x_1, x_2) + b_1$$

where $f_1(x_1, x_2) = sgn(p_1(x_1, x_2))$ and p_2 are defined in Example 6. The decision boundaries of the two classes are shown in Fig. 7. We observe that the two boundaries are no longer independent; the decision boundary for the second class C_2 (dashed grey line) depends on the decision boundary of the first class C_1.

4.1. Extensions to classifier trellises

Classifier trellises (CT) are a novel paradigm to multi-label classification problems, recently introduced by Read et al. [17]. Basically CT work as chain classifiers, but instead of adding as predictors all the previous predicted classes, just some of them are considered in the new step of the classifier, thus reducing the complexity of the algorithm. We just observe here that our results about naive BAN chain classifier could easily be extended to CT (when BAN classifiers are used as base models), especially when, as in naive BAN chain classifier, the classes already predicted are added in a naive way.

5. Binary relevance vs. chain classifier

In this section, we compare the expressive power of BR and chain classifiers when BAN classifiers are used as based models. We recall that a full Bayesian network is a Bayesian network where all pairs of nodes are linked.

Thanks to Lemma 6, we can prove the following result that generalizes Lemma 3 in Varando et al. [21].

Theorem 8. *Consider a multi-label classification problem over categorical predictors $X_1 \in \Omega_1, \ldots, X_n \in \Omega_n$, for binary classes ordered as $C_1, \ldots, C_h$. Given a sequence of predictor sub-graphs $\mathcal{H}_1, \ldots, \mathcal{H}_h$ without V-structures and such that they are not full Bayesian networks, consider $\mathcal{F}$ to be the set of h-valued decision functions induced by the naive BAN chain classifier and $\mathcal{D}$ the set of h-valued decision functions induced by the corresponding binary relevance method. We have that,*

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

In other words, naive BAN chain classifiers are more expressive than the corresponding BAN binary relevance method.

Proof. From the results of the previous sections we have that,

$$\mathcal{D} = \left\{(f_1, \ldots, f_h) \text{ s.t. } f_k = sgn(p_k),\ p_k \in \mathcal{P}_{\mathcal{H}_k}\right\}$$

$$\mathcal{F} = \left\{(f_1, \ldots, f_h) \text{ s.t. } f_k = sgn\left(p_k + \sum_{j=1}^{k-1} a_j \pi_{f_j}\right),\ p_k \in \mathcal{P}_{\mathcal{H}_k}, a_1, \ldots, a_{k-1} \in \mathbb{R}\right\}$$

Among the decision functions for the first class C_1 we can always choose for every $\mathbf{k} = (k_1, \ldots, k_n) \in \mathbb{M} = \{1, \ldots, m_1\} \times \cdots \times \{1, \ldots, m_n\}$, $f_{\mathbf{k}}(\mathbf{x})$ such that

$$f_{\mathbf{k}}(\mathbf{x}) = \begin{cases} +1 & \text{if } \mathbf{x} = (\xi_1^{k_1}, \ldots, \xi_n^{k_n}) \\ -1 & \text{if } \mathbf{x} \in \boldsymbol{\Omega} \setminus \{(\xi_1^{k_1}, \ldots, \xi_n^{k_n})\} \end{cases}$$

To prove the above fact is sufficient to observe that for every $\mathbf{k} \in \mathbb{M}$, $f_{\mathbf{k}}$ belongs to $sgn(\mathcal{P}_{NB}) \subseteq sgn(\mathcal{P}_{\mathcal{H}_1})$. In fact we have that $f_{\mathbf{k}} = sgn(p(\mathbf{x}))$ where

$$\mathcal{P}_{NB} \ni p(\mathbf{x}) = \sum_{i=1}^{n} \ell_{k_i}^{\Omega_i}(x_i) - \sum_{i=1}^{n} \sum_{j \neq k_i} n \ell_j^{\Omega_i}(x_i),$$

as it is possible to check by substitution.

Since $\mathcal{X}(f_{\mathbf{k}}) = \{X_1, \ldots, X_n\}$ and $\mathcal{H}_k$ is not complete, we have, from Lemma 6, $\pi_{f_{\mathbf{k}}} \notin \mathcal{P}_{\mathcal{H}_k}$. Thus the space $\mathcal{P}_{\mathcal{H}_k} + < \pi_{f_{\mathbf{k}}} >$ has one dimension more than $\mathcal{P}_{\mathcal{H}_k}$, and so $sgn(\mathcal{P}_{\mathcal{H}_k} + < \pi_{f_{\mathbf{k}}} >)$ contains at least two more decision functions than $sgn(\mathcal{P}_{\mathcal{H}_k})$. So we have that there exist some h-valued decision functions that belong to $\mathcal{F}$ but not to $\mathcal{D}$. □

We can also have a roughly estimation of the gain in expressibility from BAN binary relevance to naive BAN chain classifier.

Lemma 9. *If $\mathcal{F}$ and $\mathcal{D}$ are defined as in Theorem 8 we have that*

$$|\mathcal{F} \setminus \mathcal{D}| > |\Omega| \left(3^{h-1} - 1\right).$$

Proof. As in the proof of Theorem 8 we can choose, among the decision functions for the first class C_1,

$$f_{\mathbf{k}}(\mathbf{x}) = \begin{cases} +1 & \text{if } \mathbf{x} = (\xi_1^{k_1}, \ldots, \xi_n^{k_n}) \\ -1 & \text{if } \mathbf{x} \in \boldsymbol{\Omega} \setminus \{(\xi_1^{k_1}, \ldots, \xi_n^{k_n})\} \end{cases}$$

Thus we have $|\boldsymbol{\Omega}| = |\mathbb{M}|$ possibility to choose the decision function for C_1. For every $f_{\mathbf{k}}$ we have two more decision functions representable for every other classes $C_2, \ldots, C_k$, thus counting all the combinations we get

$$|\mathcal{F} \setminus \mathcal{D}| > |\boldsymbol{\Omega}| \sum_{k=1}^{h-1} \binom{(h-1)}{k} 2^k = |\boldsymbol{\Omega}| \left(3^{h-1} - 1\right) \qquad \square$$

As we see from the proof, the estimation given by Lemma 9 is far from being sharp. However, it helps us to understand that chain classifiers are not just *more expressive* than binary relevance, the difference goes to $+\infty$ as the number of labels grows.

6. Conclusions and future work

In this paper we have extended 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 number of classes. In both models, we have expressed the multi-label decision boundaries in polynomial forms and we have also proved that chain classifiers provide more expressive models than the binary relevance method when the same type of BAN classifier is used as base classifier.

Extending our results to general multi-dimensional BN classifiers [4,15,2,16], that permit BN structures between classes and predictors, is however, a much more complicated task. In multi-dimensional BN classifiers, the multi-valued decision functions have to be found by a global maximum search over the possible classes values. This fact does not permit the employment of the same arguments used in this work. It would be interesting to extend the *geometric* study of BAN classifiers, such as the study of the space of polynomials associated with every particular BAN. A deeper comprehension of the structure of $\mathcal{P}_{\mathcal{G}}$ could help to precisely compute or estimate the effective gain in expressive power of chain classifier with respect to binary relevance when the same BAN classifiers are used as base model.

Acknowledgements

The authors thank the reviewers for comments, critics and corrections which significantly contributed to improve the paper. 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 and by the Regional Government of Madrid through the S2013/ICE-2845-CASI-CAM-CM project.

References

[1] Bielza Concha, Pedro Larrañaga, Discrete Bayesian network classifier: a survey, ACM Comput. Surv. 47 (1) (2014). Article 5.
[2] Concha Bielza, Guangdi Li, Pedro Larrañaga, Multi-dimensional classification with Bayesian networks, Int. J. Approx. Reason. 52 (2011) 705–727.
[3] Hendrik Blockeel, Leander Schietgat, Jan Struyf, Sašo Džeroski, Amanda Clare, Decision trees for hierarchical multilabel classification: a case study in functional genomics, in: Johannes Fürnkranz, Tobias Scheffer, Myra Spiliopoulou (Eds.), Knowledge Discovery in Databases: PKDD 2006, in: Lecture Notes in Computer Science, vol. 4213, Springer, Berlin, Heidelberg, 2006, pp. 18–29.
[4] Linda C. van der Gaag, Peter R. de Waal, Multi-dimensional Bayesian network classifiers, in: Milan Studený, Jirí Vomlel (Eds.), Third European Workshop on Probabilistic Graphical Models, 2006, pp. 107–114.
[5] David M. Chickering, Learning equivalence classes of Bayesian–Network structures, J. Mach. Learn. Res. 2 (2002) 445–498.
[6] Krzysztof Dembczynski, Weiwei Cheng, Eyke Hüllermeier, Bayes optimal multilabel classification via probabilistic classifier chains, in: Johannes Fürnkranz, Thorsten Joachims (Eds.), Proceedings of the 27th International Conference on Machine Learning (ICML-10), Haifa, Israel, Omnipress, June 2010, pp. 279–286.
[7] Nir Friedman, Dan Geiger, Moises Goldszmidt, Bayesian network classifiers, Mach. Learn. 29 (2–3) (1997) 131–163.
[8] Teresa Gonçalves, Paulo Quaresma, A preliminary approach to the multilabel classification problem of Portuguese juridical documents, in: Fernando-Moura Pires, Salvador Abreu (Eds.), Progress in Artificial Intelligence, in: Lecture Notes in Computer Science, vol. 2902, Springer, Berlin, Heidelberg, 2003, pp. 435–444.
[9] Eamonn J. Keogh, Michael J. Pazzani, Learning the structure of augmented Bayesian classifiers, Int. J. Artif. Intell. Tools 11 (04) (2002) 587–601.
[10] Charles X. Ling, Huajie Zhang, The representational power of discrete Bayesian networks, J. Mach. Learn. Res. 3 (2003) 709–721.
[11] Marvin Minsky, Steps toward artificial intelligence, in: Computers and Thought, McGraw-Hill, 1961, pp. 406–450.
[12] Ryan O'Donnell, Rocco A. Servedio, New degree bounds for polynomial threshold functions, Combinatorica 30 (3) (2010) 327–358.
[13] Judea Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann Publishers Inc., 1988.
[14] Mark A. Peot, Geometric implications of the naive Bayes assumption, in: Eric Horvitz, Jensen Finn (Eds.), Proceedings of the Twelfth International Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers Inc., 1996, pp. 414–419.
[15] Peter R. de Waal, Linda C. van der Gaag, Inference and learning in multi-dimensional Bayesian network classifiers, in: Khaled Mellouli (Ed.), ECSQARU, in: Lecture Notes in Computer Science, vol. 4724, Springer, 2007, pp. 501–511.
[16] Jesse Read, Concha Bielza, Pedro Larrañaga, Multi-dimensional classification with super-classes, IEEE Trans. Knowl. Data Eng. 26 (7) (2014) 1720–1733.
[17] Jesse Read, Luca Martino, Pablo M. Olmos, David Luengo, Scalable multi-output label prediction: from classifier chains to classifier trellises, Pattern Recognit. 48 (6) (2015) 2096–2109.
[18] Jesse Read, Bernhard Pfahringer, Geoff Holmes, Eibe Frank, Classifier chains for multi-label classification, in: Machine Learning and Knowledge Discovery in Databases, in: Lecture Notes in Computer Science, vol. 5782, Springer, Berlin, Heidelberg, 2009, pp. 254–269.
[19] Luis E. Sucar, Concha Bielza, Eduardo F. Morales, Pablo Hernandez-Leal, Julio H. Zaragoza, Pedro Larrañaga, Multi-label classification with Bayesian network-based chain classifiers, Pattern Recognit. Lett. 41 (2014) 14–22.
[20] Grigorios Tsoumakas, Ioannis Katakis, Multi-label classification: an overview, Int. J. Data Warehous. Min. 2007 (2007) 1–13.
[21] Gherardo Varando, Concha Bielza, Pedro Larrañaga, Expressive power of binary relevance and chain classifiers based on Bayesian networks for multi-label classification, in: Linda C. van der Gaag, Ad J. Feelders (Eds.), Probabilistic Graphical Models, in: Lecture Notes in Computer Science, vol. 8754, Springer, 2014, pp. 519–534.
[22] Gherardo Varando, Concha Bielza, Pedro Larrañaga, Decision boundary for discrete Bayesian network classifiers, J. Mach. Learn. Res. (2015), in press.
[23] Min-Ling Zhang, Zhi-Hua Zhou, Ml-knn: a lazy learning approach to multi-label learning, Pattern Recognit. 40 (7) (2007) 2038–2048.
[24] Min-Ling Zhang, Zhi-Hua Zhou, A review on multi-label learning algorithms, IEEE Trans. Knowl. Data Eng. 26 (8) (August 2014) 1819–1837.