International Journal of Approximate Reasoning 52 (2011) 705–727

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

journal homepage: www.elsevier.com/locate/ijar

Multi-dimensional classification with Bayesian networks

C. Bielza [a], G. Li [b], P. Larrañaga [a,*]

[a] Computational Intelligence Group, Departamento de Inteligencia Artificial, Universidad Politécnica de Madrid, Boadilla del Monte, 28660 Madrid, Spain
[b] Rega Institute and University Hospitals, Katholieke Universiteit Leuven, B-3000 Leuven, Belgium

ARTICLE INFO

Article history:
Received 30 July 2010
Revised 2 December 2010
Accepted 21 January 2011
Available online 16 February 2011

Keywords:
Multi-dimensional outputs
Bayesian network classifiers
Learning from data
MPE
Multi-label classification

ABSTRACT

Multi-dimensional classification aims at finding a function that assigns a vector of class values to a given vector of features. In this paper, this problem is tackled by a general family of models, called multi-dimensional Bayesian network classifiers (MBCs). This probabilistic graphical model organizes class and feature variables as three different subgraphs: class subgraph, feature subgraph, and bridge (from class to features) subgraph. Under the standard 0–1 loss function, the most probable explanation (MPE) must be computed, for which we provide theoretical results in both general MBCs and in MBCs decomposable into maximal connected components. Moreover, when computing the MPE, the vector of class values is covered by following a special ordering (gray code). Under other loss functions defined in accordance with a decomposable structure, we derive theoretical results on how to minimize the expected loss. Besides these inference issues, the paper presents flexible algorithms for learning MBC structures from data based on filter, wrapper and hybrid approaches. The cardinality of the search space is also given. New performance evaluation metrics adapted from the single-class setting are introduced. Experimental results with three benchmark data sets are encouraging, and they outperform state-of-the-art algorithms for multi-label classification.

1. Introduction

In this paper we are interested in classification problems where there are multiple class variables $C_1, \ldots, C_d$. Therefore the *multi-dimensional classification* problem consists of finding a function h that assigns to each instance given by a vector of m features $\mathbf{x} = (x_1, \ldots, x_m)$ a vector of d class values $\mathbf{c} = (c_1, \ldots, c_d)$:

$$h : \Omega_{X_1} \times \cdots \times \Omega_{X_m} \to \Omega_{C_1} \times \cdots \times \Omega_{C_d}$$

$$(x_1, \ldots, x_m) \mapsto (c_1, \ldots, c_d)$$

We assume that C_i is a discrete variable, for all $i = 1, \ldots, d$, with Ω_{C_i} denoting its sample space and $\mathcal{I} = \Omega_{C_1} \times \cdots \times \Omega_{C_d}$, the space of joint configurations of the class variables. Analogously, Ω_{X_j} is the sample space of the discrete feature variable X_j, for all $j = 1, \ldots, m$.

* Corresponding author. Tel.: +34 91 3367443; fax: +34 91 3524819.
E-mail addresses: mcbielza@fi.upm.es (C. Bielza), guangdi.li@rega.kuleuven.be (G. Li), pedro.larranaga@fi.upm.es (P. Larrañaga).

0888-613X/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ijar.2011.01.007

Many application domains include multi-dimensional classification problems: a text document or a semantic scene can be assigned to multiple topics, a gene can have multiple biological functions, a patient may suffer from multiple diseases, a patient may become resistant to multiple drugs for HIV treatment, a physical device can break down due to multiple components failing, etc.

Multi-dimensional classification is a more difficult problem than the single-class case. The main problem is that there is a large number of possible class label combinations, $|\mathcal{I}|$, and a corresponding sparseness of available data. In a typical scenario where an instance $\mathbf{x}$ is assigned to the most likely combination of classes (0–1 loss function), the aim is to compute $\arg\max_{c_1,\ldots,c_d} p(C_1 = c_1, \ldots, C_d = c_d|\mathbf{x})$. It holds that $p(C_1 = c_1, \ldots, C_d = c_d|\mathbf{x}) \propto p(C_1 = c_1, \ldots, C_d = c_d, \mathbf{x})$, which requires $|\mathcal{I}| \cdot |\Omega_{X_1} \times \cdots \times \Omega_{X_m}|$ parameters to be assigned. In the single-class case, $|\mathcal{I}|$ is just $|\Omega_C|$ rather than $|\Omega_{C_1} \times \cdots \times \Omega_{C_d}|$. Besides it having a high cardinality, it is also hard to estimate the required parameters from a (sparse) data set in this d-dimensional space $|\mathcal{I}|$. The factorization of this joint probability distribution when using a Bayesian network (BN) can somehow reduce the number of parameters required and will be our starting point.

Standard (one-class) BN classifiers cannot be straightforwardly applied to this multi-dimensional setting. On the one hand, the problem could be transformed into a single-class problem if a compound class variable modeling all possible combinations of classes is constructed. However, this class variable would have too many values, and even worse, the model would not capture the structure of the classification problem (dependencies among class variables and also among class variables and features). On the other hand, we could approach the multi-dimensional problem by constructing one independent classifier for each class variable. However, this would not capture the interactions among class variables, and the most likely class label for each independent classifier – marginal classifications – after being assembled as a d-dimensional vector, might not coincide with the most likely vector of class labels of the observed data.

As we will show below, the few proposals found in the literature on multi-dimensional BN classifiers (MBCs) are limited. In this paper, we propose a comprehensive theory of MBCs, including their extended definition, learning from data algorithms that cover all the possibilities (wrapper, filter and hybrid score + search strategies), and results on how to perform total abduction for the exact inference of the most probable explanation (MPE). MPE computation is the main aim in 0–1 loss function classification problems but involves a high computational cost in the multi-dimensional setting. Several contributions are designed here to reduce this computational load: the introduction of special decomposed MBCs, their extension to non 0–1 loss function problems that respect this decomposition, and a particular and favorable way of enumerating all the $(c_1, \ldots, c_d)$ configurations instead of using a brute-force approach.

The paper is organized as follows: Section 2 defines MBCs. Section 3 covers different contributions for the MPE computation and introduces a restricted structure of decomposable MBCs where MPE is easier to compute. Section 4 extends these ideas to compute the Bayes decision rule with certain loss functions that we call additive CB-decomposable loss functions. Section 5 presents performance measures suitable for evaluating MBCs. Section 6 describes wrapper, filter and hybrid algorithms to learn MBCs from data. It also provides the cardinality of the MBC structure space where these algorithms search for. Section 7 shows experimental results on MPE with simulated MBCs. Section 8 contains experimental results with three benchmark data sets. Section 9 reviews the work related to multi-dimensional classification, with special emphasis on papers using (simpler) MBCs. Finally, Section 10 sums up the paper with some conclusions.

2. Multi-dimensional Bayesian network classifiers

A Bayesian network over a finite set $\mathcal{V} = \{Z_1, \ldots, Z_n\}$, $n \geq 1$, of discrete random variables is a pair $B = (\mathcal{G}, \Theta)$, where $\mathcal{G}$ is an acyclic directed graph whose vertices correspond to the random variables and Θ is a set of parameters $\theta_{z|\mathbf{pa}(z)} = p(z|\mathbf{pa}(z))$, where $\mathbf{pa}(z)$ is a value of the set of variables $\mathbf{Pa}(Z)$, parents of the Z variable in the graphical structure $\mathcal{G}$ [42,36]. B defines a joint probability distribution p_B over $\mathcal{V}$ given by

$$p_B(z_1, \ldots, z_n) = \prod_{i=1}^{n} p(z_i|\mathbf{pa}(z_i)). \tag{1}$$

A *multi-dimensional Bayesian network classifier* is a Bayesian network specially designed to solve classification problems including multiple class variables in which instances described by a number of features have to be assigned to a combination of classes.

Definition 1 (*Multi-dimensional Bayesian network classifier*)**.** In an *MBC* denoted by $B = (\mathcal{G}, \Theta)$, the graph $\mathcal{G} = (\mathcal{V}, \mathcal{A})$ has the set $\mathcal{V}$ of vertices partitioned into two sets $\mathcal{V}_\mathcal{C} = \{C_1, \ldots, C_d\}$, $d \geq 1$, of class variables and $\mathcal{V}_\mathcal{X} = \{X_1, \ldots, X_m\}$, $m \geq 1$, of feature variables ($d + m = n$). $\mathcal{G}$ also has the set $\mathcal{A}$ of arcs partitioned into three sets, $\mathcal{A}_\mathcal{C}$, $\mathcal{A}_\mathcal{X}$, $\mathcal{A}_{\mathcal{C}\mathcal{X}}$, such that:

- $\mathcal{A}_\mathcal{C} \subseteq \mathcal{V}_\mathcal{C} \times \mathcal{V}_\mathcal{C}$ is composed of the arcs between the class variables having a subgraph $\mathcal{G}_\mathcal{C} = (\mathcal{V}_\mathcal{C}, \mathcal{A}_\mathcal{C})$ –*class subgraph*– of $\mathcal{G}$ induced by $\mathcal{V}_\mathcal{C}$.
- $\mathcal{A}_\mathcal{X} \subseteq \mathcal{V}_\mathcal{X} \times \mathcal{V}_\mathcal{X}$ is composed of the arcs between the feature variables having a subgraph $\mathcal{G}_\mathcal{X} = (\mathcal{V}_\mathcal{X}, \mathcal{A}_\mathcal{X})$ –*feature subgraph*– of $\mathcal{G}$ induced by $\mathcal{V}_\mathcal{X}$.