International Journal of Approximate Reasoning 51 (2010) 695–717

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

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

Feature selection for Bayesian network classifiers using the MDL-FS score

Mădălina M. Drugan [a,*], Marco A. Wiering [b]

[a] Department of Information and Computing Sciences, Utrecht University, P.O. Box 80.089, 3508 TB Utrecht, The Netherlands
[b] Department of Artificial Intelligence, University of Groningen, 9700 AK Groningen, The Netherlands

ARTICLE INFO

Article history:
Received 12 October 2007
Received in revised form 3 February 2010
Accepted 3 February 2010
Available online 18 February 2010

Keywords:
Feature subset selection
Minimum Description Length
Selective Bayesian classifiers
Tree augumented networks

ABSTRACT

When constructing a Bayesian network classifier from data, the more or less redundant features included in a dataset may bias the classifier and as a consequence may result in a relatively poor classification accuracy. In this paper, we study the problem of selecting appropriate subsets of features for such classifiers. To this end, we propose a new definition of the concept of redundancy in noisy data. For comparing alternative classifiers, we use the Minimum Description Length for Feature Selection (MDL-FS) function that we introduced before. Our function differs from the well-known MDL function in that it captures a classifier's conditional log-likelihood. We show that the MDL-FS function serves to identify redundancy at different levels and is able to eliminate redundant features from different types of classifier. We support our theoretical findings by comparing the feature-selection behaviours of the various functions in a practical setting. Our results indicate that the MDL-FS function is more suited to the task of feature selection than MDL as it often yields classifiers of equal or better performance with significantly fewer attributes.

1. Introduction

Many real-life problems, such as medical diagnosis and troubleshooting of technical equipment, can be viewed as a classification problem, where an instance described by a number of features has to be classified in one of several distinct predefined classes. For many of these classification problems, instances of every-day problem solving are recorded in a dataset. Such a dataset often includes more features, or attributes, of the problem's instances than are strictly necessary for the classification task at hand. When constructing a classifier from the dataset, these more or less redundant features may bias the classifier and as a consequence may result in a relatively poor classification accuracy. By constructing the classifier over just a subset of the features, a less complex classifier is yielded that tends to have a better generalisation performance [1]. Finding a minimum subset of features such that the selective classifier constructed over this subset is optimal for a given performance measure, is known as the *feature subset selection* problem [2–4]. The feature subset selection problem unfortunately is NP-hard in general [5–8].

We begin by providing a new definition of the concept of redundancy of attributes, where the redundancy is viewed within some allowed amount of noise in the data under study. It allows us to study feature selection for different types of Bayesian network classifier more specifically. With our definition we distinguish between different *levels of redundancy* for an attribute. The levels depend on the cardinality of the (sub)sets of other attributes with which the attribute is combined so that the attribute is not useful for the classification task. We will argue that these levels of redundancy provide for relating the problem of feature subset selection to the types of dependence that can be expressed by a Bayesian network classifier. By allowing noise for the various levels, our concept of redundancy provides for studying feature selection in a practical setting.

* Corresponding author.
E-mail address: M.M.Drugan@uu.nl (M.M. Drugan).

0888-613X/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ijar.2010.02.001

For constructing a selective classifier, generally a heuristic algorithm [9] is used that searches the space of possible models for classifiers of high quality. Because of its simplicity, its intuitive theoretical foundation and its associated ease of computation, the MDL function and its variants [10] have become quite popular as quality measures for constructing Bayesian networks from data, and in fact for constructing Bayesian network classifiers [11]. The function in essence weighs the complexity of a model against its ability to capture the observed probability distribution. While the MDL function and its variants are accepted as suitable functions for comparing the qualities of alternative Bayesian networks, they are not without criticism when constructing Bayesian network classifiers. The criticism focuses on the observation that the functions capture a joint probability distribution over the variables of a classifier, while it is the conditional distribution over the class variable given the attributes that is of interest for the classification task [11–17].

For comparing the qualities of alternative classifiers, we propose the *Minimum Description Length for Feature Selection* (MDL-FS) function [18]. The MDL-FS function is closely related to the well-known Minimum Description Length (MDL) function. It differs from the MDL function only in that it encodes the conditional probability distribution over the class variable given the various attributes. Upon using the function as a measure for comparing the qualities of Bayesian network classifiers therefore, this conditional distribution has to be learned from the available data. Unfortunately, learning a conditional distribution is generally acknowledged to be hard [19–21], since it does not decompose over the graphical structure of a Bayesian network classifier as does the joint distribution. Our MDL-FS function approximates the conditional distribution by means of an auxiliary Bayesian network which captures the strongest relationships between the attributes. With the function, both the structure of the Bayesian network classifier over all variables involved and the structure of the auxiliary network over the attributes are learned using a less demanding generative method. The conditional log-likelihood of the classifier then is approximated by the difference between the unconditional log-likelihood of the classifier and the log-likelihood of the auxiliary network.

This paper is organised as follows: In Section 2 we provide some background on Bayesian networks and on Bayesian network classifiers more specifically; we further review the MDL function and present our notational conventions. In Section 3 we introduce the problem of feature subset selection and provide a formal definition of the concept of redundancy. We introduce our new MDL-FS function and study its relationship with the MDL function in Section 4. In Section 5, we investigate the feature-selection behaviour of the MDL-FS function in general and we compare it with the behaviour of the MDL function. In Section 6 we study the use of the MDL-FS function in constructing selective Naïve Bayes and TAN classifiers from data. In Section 7 the feature-selection behaviour of the MDL-FS and MDL functions and other state of the art feature selection algorithms are compared in a practical setting. Our results indicate that the MDL-FS function indeed is more suited to the task of feature subset selection than the MDL function or other feature selection algorithms as it yields classifiers of comparably good or even significantly better performance with fewer attributes. The paper ends with our concluding observations and remarks in Section 8.

2. Background

In this section, we provide some preliminaries on Bayesian networks and on Bayesian network classifiers more specifically. We conclude this section with a discussion of the MDL function.

2.1. Bayesian networks and Bayesian network classifiers

We consider a set V of stochastic variables $V_i,\ i = 1,\ldots,n,\ n \geqslant 1$. We use $\Omega(V_i)$ to denote the set of all possible (discrete) values of the variable V_i; for ease of exposition, we assume a total ordering on the set $\Omega(V_i)$ and use v_i^k to denote the kth value of V_i. For any subset of variables $S \subseteq V$, we use $\Omega(S) = \times_{V_i \in S}\Omega(V_i)$ to denote the set of all joint value assignments to S. A *Bayesian network* over V now is a tuple $\mathcal{B} = (G, P)$ where G is a directed acyclic graph and P is a set of conditional probability distributions. In the digraph G, each vertex models a stochastic variable from V. The set of arcs captures probabilistic independence: for a topological sort of the digraph G, that is, for an ordering $V_1,\ldots,V_n$, $n \geqslant 1$, of its variables with $i < j$ for every arc $V_i \to V_j$ in G, we have that any variable V_i is independent of the preceding variables $V_1,\ldots,V_{i-1}$ given its parents in the graphical structure. Associated with the digraph is a set P of probability distributions: for each variable V_i are specified the conditional distributions $P(V_i|p(V_i))$ that describe the influence of the various assignments to the variable's parents $p(V_i)$ on the probabilities of the values of V_i itself. The network defines a unique joint probability distribution $P(V)$ over its variables with

$$P(V) = \prod_{V_i \in V} P(V_i|p(V_i))$$

Note that the thus defined probability distribution factorises over the network's digraph into separate conditional distributions. Bayesian network classifiers are Bayesian networks of restricted topology that are tailored to solving classification problems. In a classification problem, instances described by a number of features have to be classified in one of several distinct predefined classes. We consider to this end a set A of stochastic variables A_i, called *attributes*, that are used to describe the features of the instances. We further have a designated variable C, called the *class variable*, that captures the various possible classes. *Bayesian network classifiers* now are defined over the set of variables $A \cup \{C\}$. Like a Bayesian network in general,