Decision Support Systems 48 (2009) 180–190

Contents lists available at ScienceDirect

Decision Support Systems

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

Histogram distance-based Bayesian Network structure learning: A supervised classification specific approach

B. Sierra *, E. Lazkano, E. Jauregi, I. Irigoien

Dept. of Computer Science and Artificial Intelligence, University of the Basque Country, Spain

ARTICLE INFO

Article history:
Received 16 May 2008
Received in revised form 12 June 2009
Accepted 22 July 2009
Available online 6 August 2009

Keywords:
Bayesian Network
Histogram distance
Supervised classification
Machine learning
Structure learning

ABSTRACT

In this work we introduce a methodology based on histogram distances for the automatic induction of Bayesian Networks (BN) from a file containing cases and variables related to a supervised classification problem. The main idea consists of learning the Bayesian Network structure for classification purposes taking into account the classification itself, by comparing the class distribution histogram distances obtained by the Bayesian Network after classifying each case. The structure is learned by applying eight different measures or metrics: the Cooper and Herskovits metric for a general Bayesian Network and seven different statistical distances between pairs of histograms.

The results obtained confirm the hypothesis of the authors about the convenience of having a BN structure learning method which takes into account the existence of the special variable (the one corresponding to the class) in supervised classification problems.

1. Introduction

Almost any practical intelligent application requires dealing with uncertainty. This uncertainty may be motivated by the inherent complexity of the problem, by the technical limitation of the data collection machines (Interval Error in single data, low resolution in images, etc.), by safety considerations (radioactive trace of a patient could give good information, but it is not applicable), or it could be due to the impossibility to collect or to manage all the data needed to perform the reasoning.

Until recently, the application of strict probabilistic approaches to reasoning was considered impractical due to the problem of computing the joint probability distribution of a large number of random variables involved in reasoning. However, the emergence of the concept of conditional (in)dependency allowed to simplify the calculus involved and made the evolution of automatic methods for reasoning under uncertainty based on probability theory possible.

The last decade has seen significant theoretical advances and an increasing interest in probabilistic graphical models (PGMs), the most widely used of the probability based methods. These models represent dependency relationships within a set of random variables, where the random variables are represented as nodes in a graph. The absence of arcs in the graph corresponds to independence and its presence means possible dependence between two variables. One of the most popular types of graphical models is Bayesian Networks (BN). In these type of models arcs are directed, and there should not be a directed cycle in the whole graph [11,37,46]. Much research has been devoted to the BN structure learning task [13,14,25]. However, the problem of acquiring good BN structures in general, and in particular for structures which would serve as classification models, remains open.

Most of the structure learning algorithms need two components (score + search): the learning algorithm itself that guides the search, and the metric for evaluating the structure at each time step of the learning process. The objective of this work is to look for new metrics for Bayesian Network structure learning algorithms. We concentrate on supervised classification problems and, more specifically, on the use of histogram distances for computing the differences between the class a posteriori distribution a given net structure should give to a case, and the real category that case belongs to.

2. Motivation

Supervised classification tasks are related to classification problems in which the perfect classifier does not exist or is not known; these models belong to the machine learning (ML) area [43]. In the supervised classification process the goal is to distinguish the kind (class, category) of examples or cases. Given a database, experiments are usually carried out using a known learning algorithm to induce the corresponding classifier from the data; then the obtained model is used to classify new cases of the same problem. For example, given a database of patients, where each patient (case) is labelled as having or not having a specific disease, a classifier for the cases in this database is constructed, and the obtained model is used to classify new patients (cases) with the aim of helping the physician in the process of

* Corresponding author.
E-mail address: b.sierra@ehu.es (B. Sierra).
URL: http://www.sc.ehu.es/ccwrobot (B. Sierra).

0167-9236/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2009.07.010

diagnosis. Hence, it is only after the classifier is constructed that the obtained model can be used to classify new cases.

In the machine learning area, there are three main approaches to learn a classifier model:

1. Obtain a model by using a given measure for constructing the classifier; classification trees and rule inducers belong to this kind of models. With respect to the Bayesian Network induction, Naive Bayes models could also be considered as belonging to this category. Many machine learning/data mining applications use entropy or (conditional) mutual information as metrics for selecting features and/or structure in these models. And many common algorithms for learning decision trees use mutual information to select attributes for internal nodes of the trees [48] as well.
2. Obtain the classifier by maximizing some probability measure given the data. A learning procedure typically attempts to take out the parameters of the distribution P(Given case|Class) to maximize the likelihood of the training data. Most of the Bayesian Network structure learning algorithms work in this manner, for instance the *K2* algorithm and the BIC approach, both described later on.
3. Obtain a classification model by maximizing the classification power itself. Neural Networks and Support Vector Machines are members of this third group, as well as most of the so-called multiclassifier systems [41].

On the other hand, probability based classifiers can be of two different types:

1. Generative models where all variables and responses are obtained from the joint probability distribution functions. Naive Bayes, Hidden Markov Models and Bayesian Networks are of this type of models.
2. Discriminative models, which only optimize a mapping from inputs to desired outputs [44], identifying outgoing parameters to maximize the ability of the model to discriminate between the classes. Examples of this kind of paradigms are logistic regression, Support Vector Machines, Neural Networks and *K*-nearest neighbor algorithm.

The histogram distance-based BN learning approach presented in this paper makes use of distances among the a posteriori distributions of the class variable to guide the search of the structure of BNs with classification purposes. This new method does not belong fully to any of above mentioned approaches. In fact, it can be considered a mixture. On the one hand, it looks for the structure that maximizes a measure and it performs a computation over the a posteriori probabilities of the class variable. On the other hand, it looks for a generalization model by means of a metric, with a clearly discriminative approach.

The main motivation of this approach is the hypothesis, based on previous experiences, that the relation between some metric values is not adequate when the future use of the BN is to classify new cases in supervised classification problems. We are concerned about the correctness of the metric calculation process, but we do realize that this does not guarantee the discriminative capability of the obtained model. The underlying idea is to have a measure of the classification capabilities of the BN, which at the same time should give good generalization capabilities.

The main characteristic of this new method is that it is intended to model the behavior of the class variable not only in the majority class value, but in all its amplitude. The obtained model should take into account the characteristic of the BNs — and of the probabilistic methods in general — of giving a certainty measure in relation to the classification assigned. In other words, the result a BN gives when classifying a case is a vector containing the a posteriori probabilities for all the values the class variable can take. For instance, if the class variable takes 4 different values, the classification response is a 4 element vector (x_1, x_2, x_3, x_4) where $\sum x_1 = 1$. Therefore, the vector contains the value distribution of the class variable for the given case. Typically, the case would be classified as belonging to the class x_j with the highest a posteriori probability. However, taking into account the a posteriori distribution of the class variable during the BN structure leaning process should in principle allow to obtain models with higher classification capabilities.

The rest of the paper is organized as follows: Section 3 reviews how BNs are used as supervised classifiers. Section 4 briefly reviews the concept of BNs and it is fully devoted to the description of how the structure or graph of BNs can be automatically acquired from data. Section 5 presents the new proposed approach as a metric to measure how adequate a given structure is for a classification task; Section 6 shows the three phase experimental setup we have designed to evaluate the performance of the new approach, and the obtained results are presented in Section 7. Finally, in Section 8 conclusions are given and further work lines are pointed out.

3. Related work: Bayesian Networks as classifiers

There is a lot of work devoted to the Bayesian Network structure learning for classification purposes. The related work shows that some structural learning approaches do take into account the existence of the class variable, and probably the most extended approach is to use the classification accuracy of the net as the metric value [1,58]. But none of the approaches treat the class variable distribution as the method proposed here treats it. Several approaches acquire the structure by representing the joint probability of all the variables involved in the model. Thus, [51] present a parameter learning for BNs devoted to classification tasks, maximizing the conditional (supervised) likelihood instead of the joint (unsupervised) one; [21] present a structural learning method that needs to take into account the existence of the class variable and obtains a tree-shaped structure, known as a Tree Augmented Network (TAN), in which the class variable is the root node. Keogh and Pazzani [29,30] present an approach to learn TAN structures not by means of probability, but guided by the accuracy; it is a greedy approach in which the concept of SuperParent method is presented. Greiner and Zhou [23] present the ELR algorithm. This algorithm maximizes the conditional likelihood of the class node to augment the discriminative capabilities of the acquired Bayesian Network. Grossman and Domingos [24] present the BNC algorithm to learn the structure of a BN maximizing the conditional likelihood; it is a greedy algorithm similar to that presented by Heckerman et al. [25] which combines user knowledge and statistical data.

Other authors take into account the conditional independence among the variables. For instance, [1] presents a local search in a space consisting of Partially Directed Acyclic Graphs (PDAGs), combining the two types of DAG equivalences: classification equivalence and independence equivalence; and [33] presents an approach that uses independence assumptions to learn BN finding subsets of predictor variables and augmenting the NB model using the dependencies found.

The interest in BNs as classifiers spreads to real applications such as microarray data analysis [34] or real world data treatment on information technology [27]. [57] uses BNs for the survival prediction of patients suffering from malignant skin melanoma. They extend the TAN approach, eliminating the restriction of the class variable to be the root node, but keeping all the predictor variables within the Markov Blanquet of the class variable. [39] uses PGMs to perform the diagnosis and control of autonomous vehicles in which the existence of the class variable is taken into account by the group of experts responsible for constructing the models. Similarly, [36] presents a real application of BNs for guiding a robot in door crossing behaviors using sonar sensor readings (which are a source of high uncertainty for the model); they show an attempt of integrating the expert knowledge of