Information Sciences 260 (2014) 120–148

Contents lists available at ScienceDirect

Information Sciences

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

Domains of competence of the semi-naive Bayesian network classifiers

M. Julia Flores, José A. Gámez, Ana M. Martínez *

Computing Systems Department, Instituto de Investigación en Informática de Albacete (I3A), University of Castilla-La Mancha, 02071 Albacete, Spain

ARTICLE INFO

Article history:
Received 13 June 2012
Received in revised form 1 September 2013
Accepted 1 October 2013
Available online 19 October 2013

Keywords:
Domains of competence
Semi-naive Bayesian network classifiers
Naive Bayes
AODE
Complexity measures
Discretization

ABSTRACT

The motivation for this paper comes from observing the recent tendency to assert that rather than a unique and globally superior classifier, there exist local winners. Hence, the proposal of new classifiers can be seen as an attempt to cover new areas of the complexity space of datasets, or even to compete with those previously assigned to others. Several complexity measures for supervised classification have been designed to define these areas. In this paper, we want to discover which type of datasets, defined by certain range values of the complexity measures for supervised classification, fits for some of the most well-known semi-naive Bayesian network classifiers. This study is carried out on continuous and discrete domains for naive Bayes and Averaged One-Dependence Estimators (AODE), which are two widely used incremental classifiers that provide some of the best trade-offs between error performance and efficiency. Furthermore, an automatic procedure to advise on the best semi-naive BNC to use for classification, based on the values of certain complexity measures, is proposed.

1. Introduction

The use of complexity measures (CMs) for supervised classification has received increasing attention since their formal definition in Ho and Basu [28]. These measures have been defined to determine the intrinsic characteristics of real-world classification problems. They aim at characterising the complexity of a classification problem by a set of geometrical descriptors, such as: the degree of linear separability amongst classes, Fisher's discriminant ratio or the width of the class boundary. Hence, many subsequent studies have applied these CMs to find the domains of competence of different classifiers [3,4,56,38,39,41]. Other works have also attempted to generalise these measures to problems with multiple classes [47,49], as the original definition only covers binary-class problems, although there is still no agreement on the set of measures to use in the multi-class setting.

There is no work, however, relating to semi-naive Bayesian network classifiers (BNCs), or to discrete domains, since these CMs have traditionally been applied to numeric domains. The simplest of the BNCs is naive Bayes (NB), which assumes all the attributes are independent given the class. In spite of its naive assumption, it performs surprisingly well in certain domains [13]. Hence, numerous techniques have been proposed that aim to improve the accuracy of NB by alleviating the attribute interdependence problem. We refer to them as semi-naive Bayesian network classifiers, a term first introduced by Kononenko [34]. It is now used to group together the BNCs that extend the naive Bayes graphical structure by adding extra arcs with the goal of alleviating the naive Bayes' independence assumption in a computationally efficient manner. Hence, they

* Corresponding author. Tel.: +34 967 599200x2677; fax: +34 967 599224.
E-mail addresses: julia.flores@uclm.es (M.J. Flores), jose.gamez@uclm.es (J.A. Gámez), anam.martinezf@gmail.com (A.M. Martínez).

0020-0255/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2013.10.007

either do not perform a structural search, or if they do it is very simple (usually quadratic in the number of attributes in the worst case).[1]

The family of semi-naive Bayesian network classifiers is being extensively used for machine learning and data mining in a variety of scientific applications, especially the Averaged One-Dependence Estimator[2] (AODE) [69]. These two classifiers, namely NB and AODE, have already been supported by the research community, and are considered good representatives of the family of semi-naive BNCs. Surprisingly, their domains of competence have remained unexplored until this paper. We also focus on this family of classifiers also because they are able to naturally deal with uncertainty with similar complexity demands. They provide relatively good error rates with low computational requirements, and they are able to learn incrementally, an important property for learning from big data as well.

In the literature we can find studies based on data complexity measures for classifiers belonging to different paradigms, e.g. decision trees, fuzzy classifiers, probabilistic classifiers, support vector machines, neural networks and nearest neighbour classification [3,4,56,38,39,41,54]. In this type of studies, classifiers of (very) different complexity are analyzed, NB being usually considered as the representative for probabilistic reasoning. The training/classification time and space required by these classifiers with respect to the number of attributes, instances, classes and values per attribute vary substantially. In this paper we focus on the probabilistic setting, and so we assume that the user wants to use a probabilistic classifier, which is very usual when having to deal with uncertain relations between the variables and/or a probabilistic outcome is desired besides the predicted class label. Furthermore, we want to compare the domains of competence of classifiers of similar complexity, and this is the reason why we set the focus on the semi-naive BNC family, where little or no effort is devoted to structural learning, leaving out more complex BNCs such as those based on directly learning a Bayesian network. Hence the two major contributions of this paper are (1) the description of the common characteristics of the ideal datasets for this family of classifiers and (2) a recommendation mechanism to select the *best* of these classifiers for a given dataset.

Thus, our main objective in this paper is to explore the behaviour of some of the semi-naive BNCs according to the values of the CMs in the literature for a particular group of datasets, where behaviour refers to the predictive power in terms of the accuracy obtained. Since the natural domain of BNCs comprises exclusively discrete attributes, we want to analyse as well the descriptive power of the CMs on discrete domains.

Our study begins with the definition of the domains of competence for NB and AODE, according to the values of a selected group of CMs. We select these two BNCs because NB represents the baseline and AODE is (perhaps) the most outstanding semi-naive BNC, they do not require structural learning and, in both cases, the learning algorithm has no tunable parameters, which allows for a cleaner and unbiased study. To do so, we obtain rules which describe both good and bad behaviours for these two classifiers. These rules take the values of data complexity metrics in their antecedents, so the behaviour of the classifiers is predicted from the values of several CMs of a particular dataset prior to their application. This study is carried out both on discrete and numeric domains. We analyse the global tendency followed by the values of these CMs when they are calculated on the discrete version of numeric datasets. Furthermore, we propose an automatic procedure to advise on the best semi-naive BNC to be used for prediction on a particular dataset. For this purpose we consider NB, AODE, the Hidden One-Dependence Estimator (HODE), Tree Augmented NB (TAN) and the k-dependence Bayesian classifier (KDB).

The rest of the paper is divided as follows: Section 2 contains four subsections devoted to preliminaries and previous work, that is, Section 2.1 describes the semi-naive BNCs considered; Sections 2.2 and 2.3 include a short background to data complexity and describe the different complexity measures respectively; and Section 2.4 briefly describes an algorithm to extract ranges of good and bad behaviour in a set of datasets given a particular complexity measure. Section 3 provides the first case study, we characterise NB and AODE (in discrete and continuous domains) based on several CMs. In Section 4 (case study II), a meta-classifier to predict the best semi-naive BNC, based on some of these CMs, is proposed and empirically tested. Section 5 presents the main conclusions and outlines future work. Finally, Appendix A includes several graphs showing bivariate relationships between some of the CMs on NB.

2. Preliminaries and previous work

2.1. Semi-naive Bayesian network classifiers: NB, AODE, TAN and KDB

The classification task consists of assigning one category c_i or value of the class variable C, with $\Omega_C = \{c_1, \ldots, c_c\}$ being the set of class labels, to a new object $\vec{e}$, which is defined by the assignment of a set of values, $\vec{e} = (a_1, a_2, \ldots, a_n)$, to the attributes $A_1, \ldots, A_n$. In the probabilistic case, this task can be accomplished in an exact way by the application of Bayes' theorem (Eq. (1)).

$$p(c|\vec{e}) = \frac{p(c)p(\vec{e}|c)}{p(\vec{e})}. \tag{1}$$

[1] The reader can refer to Flores et al. [15] for a review/survey on the family of semi-naive BNCs and Zheng and Webb [72] for a comparative study of these classifiers.

[2] A list of some publications that report research that uses AODE can be found at http://www.csse.monash.edu.au/webb/AODEapps.html