Available online at www.sciencedirect.com
SCIENCE DIRECT®
Decision Support Systems 39 (2005) 253–266
Decision Support Systems
www.elsevier.com/locate/dsw

Stochastic ordering and robustness in classification from a Bayesian network

Sung-Ho Kim*

Division of Applied Mathematics, Korea Advanced Institute of Science and Technology, Daejeon 305-701, South Korea

Received 7 August 2002; received in revised form 16 October 2003; accepted 16 October 2003
Available online 27 November 2003

Abstract

Consider a model-based decision support system (DSS) where all the variables involved are binary, each taking on 0 or 1. The system categorizes the probability that a certain variable is equal to 1 conditional on a set of variables in an ascending order of the probability values and predicts for the variable in terms of category levels. Under the condition that all the variables are positively associated with each other, it is shown in this paper that the category levels are robust to the probability values. This robustness is illustrated by a simulated experiment using a variety of model structures where a set of probability values is proposed for a robust classification. A robust classification method is proposed as an alternative when exact or satisfactory probability values are not available.

Keywords: Agreement level; Basic structures of model; Conditional probability; Graphical model; Positive association

1. Introduction and motivation

One of the most significant trends over the past 20 years has been the evolution from individual stand-alone computers to the highly interconnected telecommunication network environment of today. This network environment has enhanced availability of decision support systems (DSSs) to a very large number of users, allowing more rapid exchange of information among the users [14]. Shim et al. [16] point to the importance of model-based DSSs as a powerful tool for decision aid in the web-based information sharing environment. Classification is a form of decision making under a certain loss structure [4], and DSSs for these purposes are in general model-based (see, for example, Refs. [11,17,18]).

Although model-based DSSs are preferred due to, among others, consistency in decision-making and due to time-efficiency in model evaluation and modification, they are of no use unless they are ready when needed. Constructing a model may take time if a number of random variables are involved in the model and the model structure is not simple. Suppose that a group of users want model-based classifications from a web-based DSS soon after the information about the model structure and a corresponding data set is uploaded. We may not have enough time to go through the full model-building procedure to serve the users. However, we may be able to make reasonable model-based classifications not from a model which is totally based on data but from a model which is based on data

* Tel.: +82-42-869-2737; fax: +82-42-869-5710.
E-mail address: shkim@amath.kaist.ac.kr (S.-H. Kim).
URL: http://amath.kaist.ac.kr/~slki.

0167-9236/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2003.10.010

in part and satisfies some condition that will be described in Section 2. We will show in this paper that the classification from the latter model is robust under that condition. We will consider student diagnosis in education as a running example of the classification problem and will elaborate below on the problem in the context of educational testing. Of course, the problem domain can be extended to other classification problems.

In educational testing, test results are used for guessing students' knowledge states. The need for better understanding of knowledge states calls for statistical technologies for linking performance outcomes to knowledge states [13]. Some of the technologies are used in the form of graphical models [19] whose model structures are represented in graphs, each of which consists of vertices and edges. The vertices represent random variables and the edges associative or causal relationships among the variables. The edges are directed if the relationships between the variables can be interpreted as causal and not directed otherwise. Since the relationship between abilities or knowledge units (KUs) is causal or hierarchical and the relationship between task performance and knowledge is causal, we will consider graphical models whose model structures are represented in the form of a Bayesian network [7,15].

We will call the graphical model of knowledge states and task performance a task performance model. All the variables considered in this paper are binary. The outcome of the task performance is classified as success (1) or failure (0) and the knowledge state good enough (1) or poor (0) for a given set of test items. If a student possesses a good enough knowledge for a test item, he or she has a high probability of a successful answer; otherwise, the probability will be low. When we diagnose a student's knowledge state based on his or her test result, a best way is using the conditional probability that a certain KU is in a good enough state given his or her test result. A statistical technique for computing the conditional probability is what is called evidence propagation [12] and computer programs such as HUGIN [1] and ERGO [5] are available for the computing.

In reality, building a task performance model is, in most cases, time-consuming and the quality of the probability estimates for the model may often be unsatisfactory. However, if we are interested in diagnosing a student's knowledge state in terms of class levels rather than the conditional probability, this concern may be safely resolved. As an example in this line of work, Kim [9] developed a task performance model [13] based on a test data set from a Mathematics test for a group of the 7th grade students and diagnosed the students for nine cognitive attributes or KUs. The diagnosis was carried out by classifying the students into one of five levels of the knowledge state for each KU. About 76% of the students said that the model-based diagnosis was helpful in their catch-up efforts. Of course, the test quality must be good enough for a successful diagnosis, and if the diagnosis is served sooner after the test, we can expect the better effect of the diagnosis.

In this paper, we are interested in a classification problem where the class levels are in accordance with the rank order of the probability values of a random variable. Thus, we have only to deal with relative magnitudes of the probability. This leads us to the notion of stochastic ordering which, incorporated with the rank-based classification, will play an important role in addressing the issue of robustness in classification. It is anticipated that the level of robustness varies according to the model structure. In order to see a possible range of the robustness in classification, a simulation experiment is carried out over a variety of models.

This paper is organized in four sections. Section 2 presents theorems showing that positive association among a set of binary variables preserves a stochastic ordering among the conditional probabilities of the binary variables. This result is carried over to Section 3 in the form of a simulated experiment in an effort to fathom the robustness of classification which is made based on the ordering of the conditional probabilities of an interested variable for a given data set. The simulation result shows a very high level of robustness when the variables are positively associated. Section 4 concludes the paper with a brief guideline of the proposed robust classification method.

2. Positive association and order preservation

All the variables considered in this paper are binary, taking on 0 or 1. We will use U for unobservable variables and X for observable variables. In educational testing, U may be regarded as a random