Model similarity and classification agreement between Bayesian network models

Sung-Ho Kim[1] **and Geon Youp Noh**

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

Abstract

Consider a decision support system based on a Bayesian network (BN) 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. We introduce a similarity measure between BN models and describe how a BN model can be constructed which is similar to a given BN model. Then under the condition that all the variables are positively associated with each other, a method of obtaining an agreement level of predictions between two BN models is proposed. The agreement levels are obtained by a simulation experiment for some BN models.

Keywords: Agreement level; Bayesian network; Model similarity; Conditional probability; Positive association.

1 Introduction

Model-based decision support systems (DSSs) are preferred due to, among others, consistency in decision-making and due to time-efficiency in model evaluation and modification. 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. 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 are uploaded. We may not have enough time to go through the full model-building procedure to serve the users.

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, [11, 17, 18]). Kim [10] proposed a robust

[1]Tel: +82-42-869-2737; fax: +82-42-869-5710. E-mail address: shkim@amath.kaist.ac.kr. URL: http://amath.kaist.ac.kr/∼slki.

classification method as an alternative when exact or satisfactory values are not available. He considered a model-based DSS where all the variables involved are binary, the probability that a certain variable is equal to 1 (1 for success and 0 for failure) is categorized, and predictions are made for the variable in terms of category levels. He showed that the category levels are robust to the probability values provided that the variables are positively associated among themselves.

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) are 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 (BN) [15, 7] and call them BN models.

We will call the graphical model of knowledge states and task performance a task performance model. 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, one of the best ways is to use 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 to calculate it.

In this paper, we will introduce the notion of similarity between BN models and propose a method of constructing a BN whose prediction is robust when the prediction is made for a variable in terms of category levels of probability.

This paper is organized in 6 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. Section 3 then introduces the notion of similarity between BN models and proposes a BN model which is most similar to a given BN model under some condition. In section 4, we propose a method of computing agreement levels of predictions between a pair of BN models and then results of a simulation experiment are presented. In section 5, we discuss how the agreement levels are affected in a BN model. Section 6 concludes the paper with some remarks

on applications.

2 A brief review on positive association

All the variables considered in this paper are binary, taking on the values 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 indicator of possessing a certain knowledge and X an item-score. Vectors are bold-faced. For a pair of n-vectors $\mathbf{u}$ and $\mathbf{v}$ of the same length, we write $\mathbf{u} \preceq \mathbf{v}$ when $u_i \leq v_i$ for $i = 1, \cdots, n$, and write $\mathbf{u} \prec \mathbf{v}$ if $\mathbf{u} \preceq \mathbf{v}$ and $u_i < v_i$ for at least one $i = 1, \cdots, n$.

In a BN such as the graph in Fig. 1, if a pair of nodes a and b are connected by an arrow with the arrow heading towards b from a, we call node a a parent node of node b and call node b a child node of node a. If a node does not have any parent node, it is called a root node. U_1 is the only root node in the figure. If two nodes are connected by an arrow we say that the two nodes are neighbors or that they are connected directly to each other.

Theorem 1. *Let U and X be binary variables, taking on the values 0 or 1. If*

$$0 < P(U = 1) < 1 \quad \textit{and} \quad 0 < P(X = 1) < 1 \tag{1}$$

then the following two inequalities are equivalent:

$$P(X = 1|U = 0) < P(X = 1|U = 1) \tag{2}$$

$$P(U = 1|X = 0) < P(U = 1|X = 1). \tag{3}$$

Proof. See Theorem 1 in [10] □

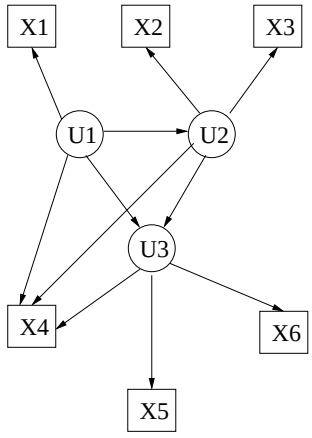

Figure 1: A BN where U variables are latent and X variables observable.

Under condition (2), we have

$$\frac{P(X=1|U=1)P(X=0|U=0)}{P(X=0|U=1)P(X=1|U=0)} > 1,$$

that is, U and X are positively associated.

We can extend this result to a situation where $X_i, i = 1, 2, \cdots, I$, are influenced by multiple U's, i.e., the conditional probability of X_i is subject to the states of some of $U_k, k = 1, 2, \cdots, K$.

Theorem 2. *Let $\boldsymbol{X} = (X_1, \cdots, X_I)$ and $\boldsymbol{U} = (U_1, \cdots, U_K)$ where all the X_i's and U_k's are binary, taking on 0 or 1. Then the following two statements are equivalent.*

(i) *For $i = 1, 2, \cdots, I$,*

$$P(X_i = 1|\mathbf{u}) < P(X_i = 1|\mathbf{v}), \quad \text{when} \quad \mathbf{u} \prec \mathbf{v}. \tag{4}$$

(ii) *For $k = 1, 2, \cdots, K$,*

$$P(U_k = 1|\mathbf{x}) < P(U_k = 1|\mathbf{y}), \quad \text{when} \quad \mathbf{x} \prec \mathbf{y}. \tag{5}$$

The strict inequality($<$) in both Eqs.(4) and (5) may be replaced by the plain inequality ($\leq$).

Proof. See Theorem 2 in [10] □

Inequality (4) is equivalent to that

$$\frac{P(\mathbf{v}|X_i = 1)}{P(\mathbf{u}|X_i = 1)} > \frac{P(\mathbf{v})}{P(\mathbf{u})} > \frac{P(\mathbf{v}|X_i = 0)}{P(\mathbf{u}|X_i = 0)}.$$

This inequality says that $\mathbf{U} = \mathbf{v}$ is more likely than $\mathbf{U} = \mathbf{u}$ when $X_i = 1$ than when $X_i = 0$. Furthermore, we can compare the likeliness of $U_k = 1$ for individual U_k's as in Theorem 2.

When the $<$ and $\prec$ in expression (4) are replaced by $\leq$ and $\preceq$, respectively, we have

$$P(X_i = 1|\mathbf{u}) \leq P(X_i = 1|\mathbf{v}), \quad \text{when} \quad \mathbf{u} \preceq \mathbf{v}. \tag{6}$$

This expression actually means that the conditional distribution of X_i given $\mathbf{U} = \mathbf{u}$ is stochastically larger than that of X_i given $\mathbf{U} = \mathbf{v}$. Holland and Rosenbaum[6] discuss properties concerning positive association among the X variables when the X variables are conditionally stochastically ordered given $\mathbf{U}$. Junker and Ellis [8] characterize such X variables in more generic terms such as conditional association and vanishing conditional dependence. We will call expression (6) the condition of positive association (or PA condition for short) among the variables $X_1, \cdots, X_I, U_1, \cdots, U_K$.

3 Similarity between BN models

Consider two BN models which are the same in graph but not necessarily in distribution. Let the graph be denoted by $\mathcal{G} = (V,\ E)$ where V and E are the set of vertices and the set of arrows between

vertices. If two vertices u and v are connected by an arrow from u to v, the arrow is represented by (u, v). So, (v, u) cannot be in E if $(u, v) \in E$. Such a graph is called a directed acyclic graph (DAG). As for a DAG $\mathcal{G}$, we define a set $pa(v)$ for a vertex v, as $pa(v) = \{u;\ (u, v) \in E\}$.

For two BN models, BN_1 and BN_2, of $\mathbf{X}_V$, assume that they have the same graph, say $\mathcal{G}$, and denote their probability distributions by $\mathcal{D}_1$ and $\mathcal{D}_2$ respectively. In other words, the BN model, $BN_i,\ i = 1, 2$, is defined as $BN_i = (\mathcal{G}, D_i)$. The probability function P^m of $\mathcal{D}_m,\ m = 1, 2$, is expressed as

$$P^m(x_V) = \prod_{v \in V} P^m_{v|pa(v)}(x_v|x_{pa(v)}).$$

In other words, $\mathcal{D}_m$ is defined in terms of the conditional distributions of X_v conditional on its parent variables. In this respect, we may say that the two distributions are similar to each other if the conditional distributions are for each variable between the two models. The similarity measure between BN_1 and BN_2 can be defined by

$$\sum_{v \in V} \sum_{x_v} \sum_{x_{pa(v)}} \left(P^1_{v|pa(v)}(x_v|x_{pa(v)}) - P^2_{v|pa(v)}(x_v|x_{pa(v)}) \right)^2.$$

We denote the support of X_v by $\mathcal{X}_v$ and let, for $A \subseteq V$, $\mathcal{X}_A = \prod_{v \in A} \mathcal{X}_v$. If all the variables are binary, the similarity measure can be re-expressed as

$$\sum_{v \in V} \sum_{\mathbf{x} \in \mathcal{X}_{pa(v)}} \left(P^1_{v|pa(v)}(1|\mathbf{x}) - P^2_{v|pa(v)}(1|\mathbf{x}) \right)^2. \tag{7}$$

We will denote this measure by $\psi(BN_1,\ BN_2)$.

We can think of a probability distribution for $P^1_{v|pa(v)}(1|y),\ y \in \mathcal{X}_{pa(v)}$, regarding $P^1_{v|pa(v)}(1|y)$ itself as a random variable, in such a way that the PA condition (6) is satisfied for $X_v,\ v \in V$. We will denote the distribution of $P^1_{v|pa(v)}(1|y),\ y \in \mathcal{X}_{pa(v)}$, by $\mathcal{D}^v_P$ and the set $\{\mathcal{D}^v_P,\ v \in V\}$ by $\mathcal{D}_P$.

Now suppose that $\mathcal{D}_2$ is not known and that we want to find a distribution, say $\mathcal{D}_*$, for which $E(\psi(BN_1,\ BN_*))$ is minimized where the expectation is made with respect to the distribution $\mathcal{D}_P$. By the definition of the similarity measure in (7), it follows that $E(\psi(BN_1,\ BN_*))$ is minimized when $\mathcal{D}_*$ consists of the conditional probabilities of $X_v,\ v \in V$ conditional on $\mathbf{X}_{pa(v)}$ which are given by

$$E(P_{v|pa(v)}(1|\mathbf{x}_{pa(v)}))$$

where the expectation is made with respect to $\mathcal{D}_P$.

For X_v, we denote by $\delta_v(\mathbf{x})$ the number of 1's in the vector of $\mathbf{x}_{pa(v)}$ and by $\kappa_v(\mathbf{x})$ the number of vectors $\mathbf{y} \in \mathcal{X}_{pa(v)}$ such that $\delta_v(\mathbf{y}) = \delta_v(\mathbf{x})$. If we let $P^*_{v|pa(v)}$ depend on $\delta_v(\mathbf{x})$ rather than on $\mathbf{x} \in \mathcal{X}_{pa(v)}$, we have the following result.

Theorem 3. *Let the expression in (7) be modified as*

$$s_0 = \sum_{v \in V} \sum_{\mathbf{x} \in \mathcal{X}_{pa(v)}} \left(P^1_{v|pa(v)}(1|\mathbf{x}) - g_v(\delta_v(\mathbf{x})) \right)^2$$

for some real valued functions g_v. If we regard $P^1_{v|pa(v)}(1|\mathbf{x})$ as random variables with distribution $\mathcal{D}_P$, then s_0 is minimized when

$$g_v(\delta_v(\mathbf{x})) = \frac{1}{\kappa_v(\mathbf{x})} \sum_{\substack{\mathbf{y} \in \mathcal{X}_{pa(v)} : \\ \delta_v(\mathbf{y}) = \delta_v(\mathbf{x})}} E(P^1_{v|pa(v)}(1|\mathbf{y})). \tag{8}$$

Proof: Since s_0 is quadratic in $g_v(\delta(\mathbf{x}))$, we have

$$\begin{aligned} \frac{ds_0}{dg_v(\delta_v(\mathbf{x}))} &= \frac{d}{dg_v(\delta_v(\mathbf{x}))} \sum_v \sum_{\mathbf{y} \in \mathcal{X}_{pa(v)}} E\left(P^1_{v|pa(v)}(1|\mathbf{x}) - g_v(\delta_v(\mathbf{x})) \right)^2 \\ &= 2 \sum_{\substack{\mathbf{y} \in \mathcal{X}_{pa(v)} : \\ \delta_v(\mathbf{y}) = \delta_v(\mathbf{x})}} E\left(P^1_{v|pa(v)}(1|\mathbf{y}) - g_v(\delta_v(\mathbf{x})) \right) = 0. \end{aligned}$$

Thus, s_0 is minimized when

$$g_v(\delta_v(\mathbf{x})) = \frac{1}{\kappa_v(\mathbf{x})} \sum_{\substack{\mathbf{y} \in \mathcal{X}_{pa(v)} : \\ \delta_v(\mathbf{y}) = \delta_v(\mathbf{x})}} E\left(P^1_{v|pa(v)}(1|\mathbf{y}) \right).$$

□

Suppose that $|pa(v)| = 3$. Then

$$\mathcal{X}_{pa(v)} = \{\underbrace{(0,0,0)}_{0}, \underbrace{(0,0,1),(0,1,0),(1,0,0)}_{1}, \underbrace{(0,1,1),(1,0,1),(1,1,0)}_{2}, \underbrace{(1,1,1)}_{3}\}, \tag{9}$$

where the elements ($\mathbf{x}$'s) are grouped according to $\delta(\mathbf{x})$. For four vectors, $\mathbf{x}^0, \mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3$, in (9), with $\delta(\mathbf{x}^i) = i$, we have that

$$\mathbf{x}^0 \prec \mathbf{x}^i \prec \mathbf{x}^3, \ i = 1, 2. \tag{10}$$

But not every pair in $\mathcal{X}_{pa(v)}$ is ordered. For instance, $(0,1,0)$ and $(1,0,1)$ are not ordered.

Recall that, under the PA condition, $P_{v|pa(v)}(1|\mathbf{x}) \leq P_{v|pa(v)}(1|\mathbf{y})$ for $\mathbf{x}, \mathbf{y} \in \mathcal{X}_{pa(v)}$, when $\mathbf{x} \preceq \mathbf{y}$. In constructing a BN model for $\mathbf{X}_V$ under the PA condition for $P_{v|pa(v)}$, we can think of assigning probability values as follows:

> For simplicity of argument, we consider the case that $|pa(v)| = 3$. If we are given 8 values between 0 and 1 in a random manner from some distribution, we assign these values to $P_{v|pa(v)}(1|\mathbf{x})$, $\mathbf{x} \in \mathcal{X}_{pa(v)}$, in the order of the elements in (9) from small to large allowing order distortions between some of $P_{v|pa(v)}(1|\mathbf{x})$ values for which the $\mathbf{x}$'s are not comparable in the sense of $\preceq$.

Since the values are assigned in a random manner to $P_{v|pa(v)}(1|\mathbf{x})$'s where the $\mathbf{x}$'s are not comparable with each other, Theorem 3 says that, if the conditional probability of X_v conditional on $\mathbf{X}_{pa(v)}$ depends upon $\delta_v(\mathbf{X})$ only in a BN model and minimizes $E(s_0)$, then the desired conditional probability is given by the right-hand side of (8). A BN model for which g_v values are used will be called an s-BN model. We will compare the performance of the s-BN model with the actual BN model in next section.

4 s-BN models for some distributions of $P^1_{v|pa(v)}(1|\mathbf{x})$

In this section, we will compare two BN models which share the same DAG but the conditional probabilities, $P_{v|pa(v)}(1|\mathbf{x})$, $v \in V$, may be different between the models. The comparison will be made in the context of predictions by the models, where the predictions are made in the form of a class level of probability values which is obtained by classifying the probability values into a finite number of subgroups in accordance with their magnitudes. A detailed description of this is given in section 3 of Kim [10].

We will call one of the two BN models for which a $\mathcal{D}_P$ is used an actual BN model and we use the g_v values corresponding to $\mathcal{D}_p$ for the s-BN model of the actual model. The agreement level of the predictions between the two models is defined as follows:

Suppose we compare predictions for subject j with regard to X_k and denote the predicted levels from the actual BN model and the s-BN model, respectively, by Y^a_{jk} and Y^s_{jk}. We define $D_{jk} = Y^s_{jk} - Y^a_{jk}$. So, if the class levels range from 1 through L, then $-L+1 \le D_{jk} \le L-1$. Suppose that there are N subjects for whom predictions are to be made. Then we can define the agreement level up to a difference d ($d > 0$) by

$$\alpha^k_d = \sum_{h:\ |h| \le d} r_{kh} \tag{11}$$

where

$$r_{kh} = \frac{\text{the number of cases that } D_{jk} = h}{N}.$$

We will use α^k_0 and α^k_1 as the two main measures of agreement. As for $\mathcal{D}_P$, we will consider the uniform distribution between 0 and 1 (denoted by $\mathcal{U}(0,\ 1)$), the beta distribution with parameters a and b (denoted by $\mathcal{B}(a, b)$), and some variations of the latter for $P^1_{v|pa(v)}(1|\mathbf{x}),\ \mathbf{x} \in \mathcal{X}_{pa(v)}$. Since $P^1_{v|pa(v)}(1|\mathbf{x})$ are ordered according to $\mathbf{x}$ under the PA condition, we may use a set of order statistics for $P^1_{v|pa(v)}(1|\mathbf{x})$ which are obtained from such a distribution as mentioned above.

Let V be partitioned into $\{V_1, V_2\}$ and suppose that we are interested in predicting for $\mathbf{X}_{V_2}$ given an outcome of $\mathbf{X}_{V_1}$. In the following subsections, we obtain the α values for $\mathbf{X}_v,\ v \in V_2$, by a

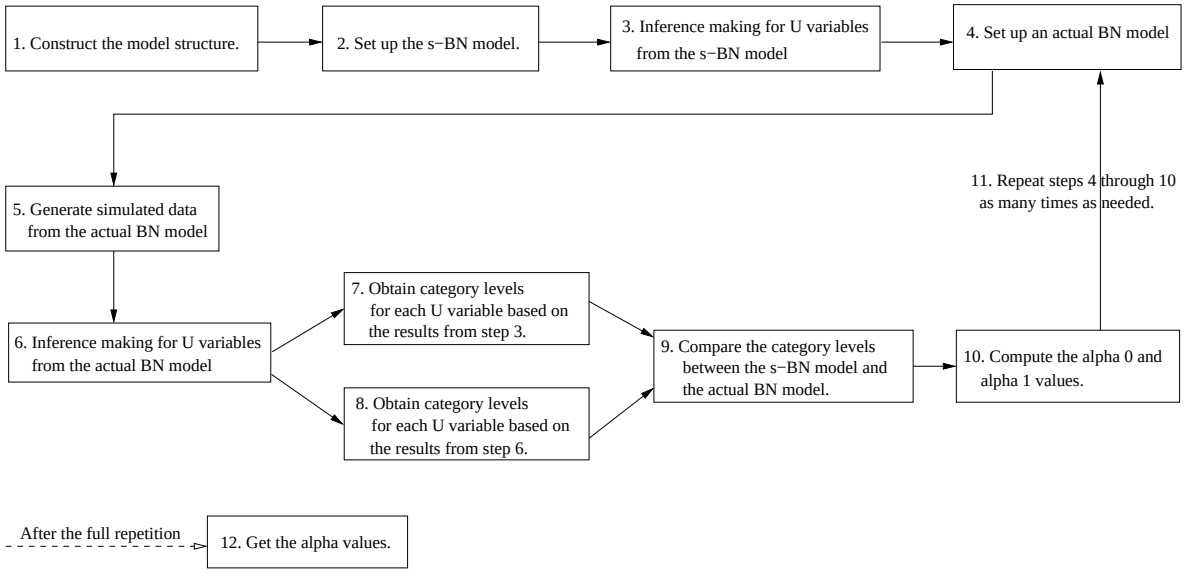

Figure 2: The process for the α values as in (12). Predictions are made for U variables.

simulation experiment for a given distribution $\mathcal{D}_P$.

In the simulation, we obtained the α values as follows:

(i) For a given $\mathcal{D}_P$, construct an s-BN model.

(ii) Generate the conditional probabilities, $\{P^1_{v|pa(v)}(1|\mathbf{x}),\ v \in V\}$ from $\mathcal{D}_P$ and set up an actual BN model using the conditional probabilities,

(iii) Generate N vector values of $\mathbf{X}_V$ from the actual model,

(iv) For each $k \in V_2$, obtain r_{kh}, $-L+1 \le h \le L-1$.

(v) Repeat steps (ii)-(iv) B times and compute the average of the B r_{kh} values for each k.

We denote by $\bar{r}_{kh}$ the average of the r_{kh} values and replace the r_{kh} in (11) by $\bar{r}_{kh}$ as in

$$\alpha^k_d = \sum_{h:\ |h| \le d} \bar{r}_{kh}. \tag{12}$$

Since B different sets of conditional probabilities, $\{P^1_{v|pa(v)}(1|\mathbf{x}),\ v \in V\}$, are used from the distribution, $\mathcal{D}_P$, the α values as in (12) can be regarded as a measure of the robustness of the predictions by the s-BN model when $\mathcal{D}_P$ is the true distribution for $\{P^1_{v|pa(v)}(1|\mathbf{x}),\ v \in V\}$. The process for the α values is depicted as a flowchart in Figure 2. In the simulation experiment below, we set $N = 10,000$ and $B = 100$.

4.1 When the $\mathcal{D}_P$ is a uniform distribution

The jth order statistic, R_j, of a random sample of size n from $\mathcal{U}(0,1)$ has a $\mathcal{B}(j, n-j+1)$ distribution[14]. From this it follows that

$$E(R_j) = \frac{j}{n+1}. \tag{13}$$

Suppose for X_v that $|pa(v)| = 3$ and consider the $\mathcal{X}_{pa(v)}$ in expression (9). If we have eight values randomly selected from the distribution $\mathcal{U}(0,\ 1)$ and order them as $R_1 < R_2 < \cdots < R_8$, we can assign them to $P^1_{v|pa(v)}(1|\mathbf{x}),\ \mathbf{x} \in \mathcal{X}_{pa(v)}$ in such a way that the condition of positive association is satisfied. By (10), we assign R_1 to $P^1_{v|pa(v)}(1|(0,0,0))$ and R_8 to $P^1_{v|pa(v)}(1|(1,1,1))$. The remaining six values, $R_2, \cdots, R_7$, are assigned to $P^1_{v|pa(v)}(1|\mathbf{x}^1)$ and $P^1_{v|pa(v)}(1|\mathbf{x}^2)$ under the PA condition.

Note that there is no ordering between the vectors, $(1,1,0)$ and $(0,0,1)$. So it is possible that $P^1_{v|pa(v)}(1|(1,1,0)) < P^1_{v|pa(v)}(1|(0,0,1))$ or that $P^1_{v|pa(v)}(1|(1,1,0)) > P^1_{v|pa(v)}(1|(0,0,1))$. Since there are two $\mathbf{x}$'s with $\delta(\mathbf{x}) = 1$ which are ordered with each $\mathbf{x}$ for which $\delta(\mathbf{x}) = 2$, it follows that R_2 and R_3 are assigned to $P^1_{v|pa(v)}(1|\mathbf{x})$, $\delta(\mathbf{x}) = 1$. And by the same argument, R_6 and R_7 must be assigned to $P^1_{v|pa(v)}(1|\mathbf{x})$, $\delta(\mathbf{x}) = 2$. One of R_4 and R_5 can be assigned to $P^1_{v|pa(v)}(1|(1,0,0))$ or $P^1_{v|pa(v)}(1|(0,1,1))$, to $P^1_{v|pa(v)}(1|(0,1,0))$ or $P^1_{v|pa(v)}(1|(1,0,1))$, or to $P^1_{v|pa(v)}(1|(0,0,1))$ or $P^1_{v|pa(v)}(1|(1,1,0))$. We can easily check that R_4 can be assigned to one of $P^1_{v|pa(v)}(1|\mathbf{x})$, $\delta(\mathbf{x}) = 2$ one third of the times of its assignment to one of $P^1_{v|pa(v)}(1|\mathbf{x})$, $\delta(\mathbf{x}) = 1$. From this and (13) it

Table 1: The g_v values of (8) for $|pa(v)| \le 3$ when $\mathcal{D}_P$ is $\mathcal{U}(0,1)$

$$|pa(v)| = 0: \quad g_v(0) = \frac{1}{2}$$

$$|pa(v)| = 1: \quad g_v(i) = \begin{cases} \frac{1}{3} & \text{if } i = 0 \\ \frac{2}{3} & \text{if } i = 1 \end{cases}$$

$$|pa(v)| = 2: \quad g_v(i) = \begin{cases} \frac{1}{5} & \text{if } i = 0 \\ \frac{1}{2} & \text{if } i = 1 \\ \frac{4}{5} & \text{if } i = 2 \end{cases}$$

$$|pa(v)| = 3: \quad g_v(i) = \begin{cases} \frac{1}{9} & \text{if } i = 0 \\ \frac{37}{108} & \text{if } i = 1 \\ \frac{71}{108} & \text{if } i = 2 \\ \frac{8}{9} & \text{if } i = 3 \end{cases}$$

Table 2: α and $\bar{r}_{kh}$ values for the model in Figure 1 when the $\mathcal{D}_P$ is $\mathcal{U}(0,1)$.

U_i	$\bar{r}_{k,-4}$	$\bar{r}_{k,-3}$	$\bar{r}_{k,-2}$	$\bar{r}_{k,-1}$	α_0	$\bar{r}_{k1}$	$\bar{r}_{k2}$	$\bar{r}_{k3}$	$\bar{r}_{k4}$	α_1
U_1	0.000	0.004	0.038	0.152	0.567	0.195	0.036	0.006	0.001	0.914
U_2	0.000	0.001	0.023	0.136	0.635	0.171	0.030	0.004	0.000	0.941
U_3	0.000	0.001	0.020	0.138	0.647	0.157	0.033	0.003	0.000	0.942
Average	0.000	0.002	0.027	0.142	**0.616**	0.174	0.033	0.004	0.001	**0.932**

follows that the g_v of (8) is given by

$$\begin{aligned} g_v(0) &= \tfrac{1}{9}, \quad g_v(1) = \tfrac{1}{3}\left\{\tfrac{2}{9} + \tfrac{3}{9} + \left(\tfrac{4}{9}\cdot\tfrac{3}{4} + \tfrac{5}{9}\cdot\tfrac{1}{4}\right)\right\} = \tfrac{37}{108} \approx \tfrac{1}{3}. \\ g_v(2) &= \tfrac{1}{3}\left\{\left(\tfrac{4}{9}\cdot\tfrac{1}{4} + \tfrac{5}{9}\cdot\tfrac{3}{4}\right) + \tfrac{6}{9} + \tfrac{7}{9}\right\} = \tfrac{71}{108} \approx \tfrac{2}{3}, \quad g_v(3) = \tfrac{8}{9}. \end{aligned} \tag{14}$$

This is summarized in Table 1.

When $|pa(v)| > 3$, the computational load for g_v increases exponentially and it is found that, when $|pa(v)| = k$, $g_v(i)$, $0 \le i \le k$, is close to

$$g_v^0(i) = \frac{\sum_{j=s_{i-1}+1}^{s_i} E(R_j)}{s_i - s_{i-1}}$$

where $s_j = \sum_{l=0}^{j} \binom{k}{l}$. This is illustrated in (14), where $|g_v(1) - g_v^0(1)| = |g_v(2) - g_v^0(2)| = 1/108$.

It is indicated in Table 2 that, on average, the predictions from the two models exactly agree for about 61.6% of the cases (see the column of α_0 in the table) and the average of the three values in the last column is 0.932, suggesting that about 93% of the predictions are on average different from the true values by at most 1.

4.2 When the $\mathcal{D}_P$ is a beta distribution

$\mathcal{U}(0,1)$ is a special form of a beta distribution, given by $\mathcal{B}(1,1)$. In this subsection, we will take as an example $\mathcal{B}(0.2, 0.4)$ for $\mathcal{D}_P$. When the conditional probabilities, $\{P^1_{v|pa(v)}(1|\mathbf{x}),\ v \in V\}$, are given as order statistics from $\mathcal{U}(0,1)$, the $P^1_{v|pa(v)}$ values tend to be evenly dispersed over the interval between 0 and 1. But the $P^1_{v|pa(v)}(1|\mathbf{x})$ values are more likely to appear near the end points of the

Table 3: The g_v values of (8) for $|pa(v)| \le 3$ when $\mathcal{D}_P$ is $\mathcal{B}(0.2, 0.4)$

$|pa(v)| = 0:\quad g_v(0) = 0.3333$

$$|pa(v)| = 1:\quad g_v(i) = \begin{cases} 0.1328 & \text{if } i = 0 \\ 0.5338 & \text{if } i = 1 \end{cases}$$

$$|pa(v)| = 2:\quad g_v(i) = \begin{cases} 0.0299 & \text{if } i = 0 \\ 0.2783 & \text{if } i = 1 \\ 0.7567 & \text{if } i = 2 \end{cases} \qquad |pa(v)| = 3:\quad g_v(i) = \begin{cases} 0.0035 & \text{if } i = 0 \\ 0.0801 & \text{if } i = 1 \\ 0.5072 & \text{if } i = 2 \\ 0.9012 & \text{if } i = 3 \end{cases}$$

Table 4: α and $\bar{r}_{kh}$ values for the model in Figure 1 when the $\mathcal{D}_P$ is $\mathcal{B}(0.2, 0.4)$.

U_i	$\bar{r}_{k,-4}$	$\bar{r}_{k,-3}$	$\bar{r}_{k,-2}$	$\bar{r}_{k,-1}$	α_0	$\bar{r}_{k1}$	$\bar{r}_{k2}$	$\bar{r}_{k3}$	$\bar{r}_{k4}$	α_1
U_1	0.004	0.025	0.052	0.166	0.540	0.183	0.025	0.004	0.000	0.889
U_2	0.002	0.006	0.035	0.153	0.617	0.172	0.014	0.001	0.000	0.942
U_3	0.000	0.006	0.034	0.161	0.613	0.154	0.029	0.002	0.000	0.928
Average	0.002	0.013	0.040	0.160	**0.590**	0.170	0.023	0.002	0.000	**0.920**

interval when they are from $\mathcal{B}(0.2, 0.4)$. In the case of educational testing, if X_v is an item score (1 for correct answer, 0 otherwise) and $\mathcal{X}_{pa(v)}$ is a vector of the knowledge states (1 for good enough, 0 otherwise) of the knowledge units that are required for the test item, then the $P^1_{v|pa(v)}$ values tend to be small when $\delta(\mathbf{x}_{pa(v)}) < |pa(v)|$ and near 1 otherwise (See, for example, [13]). In this respect, the beta distribution is a reasonable choice for $\mathcal{D}_P$.

The mean of the jth largest among a random sample of size n from a beta distribution is not available in a closed form, and the g_v values are computed by a numeral approach. Table 3 lists the g_v values for $|pa(v)| \leq 3$.

It is indicated in Table 4 that on average, the predictions from the two models exactly agree for about 59% of the cases(see the column of α_0 in the table) and the average of the three values in the last column is 0.920.

It is worthwhile to note in Table 3 that the $g_v(0)$ values become very small as $|pa(v)|$ increases. In educational testing with multiple selection tests, these values are unreasonably small since multiple selection tests allow guessing for selecting response options. An example of the distribution $\mathcal{D}_P$ for this situation is

$$0.9 \cdot \mathcal{B}(a,\ b) + 0.1, \tag{15}$$

which means that guessing shrinks the range of the conditional probabilities from $\mathcal{B}(a, b)$ to $0.9 \cdot \mathcal{B}(a, b) + 0.1$. The g_v values under this transformation are listed in Table 5 for $|pa(v)| \leq 3$. In the table, we can see that the g_v values are larger than 0.1 but the differences in $g_v(i)$ get smaller

Table 5: The g_v values of (8) for $|pa(v)| \leq 3$ when $\mathcal{D}_P$ is $0.9 \cdot \mathcal{B}(0.2, 0.4)+0.1$

$|pa(v)| = 0: \quad g_v(0) = 0.4$

$$|pa(v)| = 1: \quad g_v(i) = \begin{cases} 0.2195 & \text{if } i = 0 \\ 0.5805 & \text{if } i = 1 \end{cases}$$

$$|pa(v)| = 2: \quad g_v(i) = \begin{cases} 0.1269 & \text{if } i = 0 \\ 0.3505 & \text{if } i = 1 \\ 0.7720 & \text{if } i = 2 \end{cases}$$

$$|pa(v)| = 3: \quad g_v(i) = \begin{cases} 0.1031 & \text{if } i = 0 \\ 0.1721 & \text{if } i = 1 \\ 0.5565 & \text{if } i = 2 \\ 0.9110 & \text{if } i = 3 \end{cases}$$

Table 6: α and $\bar{r}_{kh}$ values for the model in Figure 1 when the $P^1_{v|pa(v)}$ is from $0.9 \cdot \mathcal{B}(0.2, 0.4) + 0.1$

U_i	$\bar{r}_{k,-4}$	$\bar{r}_{k,-3}$	$\bar{r}_{k,-2}$	$\bar{r}_{k,-1}$	α_0	$\bar{r}_{k1}$	$\bar{r}_{k2}$	$\bar{r}_{k3}$	$\bar{r}_{k4}$	α_1
U_1	0.009	0.018	0.065	0.143	0.526	0.207	0.031	0.031	0.002	0.876
U_2	0.001	0.009	0.041	0.134	0.606	0.191	0.018	0.018	0.000	0.931
U_3	0.004	0.010	0.043	0.141	0.620	0.164	0.018	0.018	0.002	0.924
Average	0.005	0.012	0.050	0.139	**0.584**	0.187	0.022	0.001	0.000	**0.910**

between the two types of $\mathcal{D}_P$, $\mathcal{B}(0.2, 0.4)$ and $0.9 \cdot \mathcal{B}(0.2, 0.4) + 0.1$ as i increases. However, the α values (see Tables 4 and 6) are more or less the same between them.

5 Discussion

In this section we will investigate how α values at a node depend upon the location of the node in a given BN. For convenience, we denote by U the random variable which is to be predicted for, and by X the one to be conditioned on, when predictions are made. Our point of interest is how α values are affected with regard to the X variables in a BN. The effect of X variables on a U variable may depend on the lengths of the paths from the U variable to the X variables in the BN, on the number of the X variables in the BN, and the number of adjacent nodes of the U variable. It is very difficult to analyze theoretically the effect on α since the number of different BN's increases exponentially in $|V|$ as well as the model complexity. This is why we did the investigation by a simulation experiment.

In the simulation, we took $\mathcal{U}(0, 1)$ for $\mathcal{D}_P$, and $N = 10,000$ and $B = 1,000$ to obtain α values. The five BN's in Figure 3 were considered to see how the number of X variables affect α. In the graphs in panels (a), (b), and (d), the U variables can be grouped in such a way that those at the

Table 7: The averages of the α_0 and α_1 values for the U variables in the graphs in Figure 3. The sets, $g_1^a, \cdots, g_3^e$, are labelled such that for each panel the U variables in $g_i^{\cdot}$ are located closer to the X variables by one edge than the U variables in $g_{i+1}^{\cdot}$ are.

panel	grouping*	averages	
		α_0	α_1
a	$g_2^a = \{1\}$	0.638	0.923
	$g_1^a = \{2, 3\}$	0.713	0.958
b	$g_3^b = \{1\}$	0.531	0.903
	$g_2^b = \{2, 3\}$	0.582	0.913
	$g_1^b = \{4, 5, 6\}$	0.602	0.926
c	$g_2^c = \{1\}$	0.515	0.876
	$g_1^c = \{2, 3, 4, 5\}$	0.579	0.931
d	$g_4^d = \{1\}$	0.470	0.862
	$g_3^d = \{2, 3\}$	0.515	0.880
	$g_2^d = \{4, 5, 6\}$	0.560	0.910
	$g_1^d = \{7, 8, 9, 10\}$	0.547	0.919
e	$g_3^e = \{1\}$	0.497	0.875
	$g_2^e = \{2, 3, 4, 5\}$	0.581	0.932
	$g_1^e = \{6, 7, 8, 9\}$	0.528	0.921

* $\{i, j\}$ is a simplified expression of $\{U_i, U_j\}$.

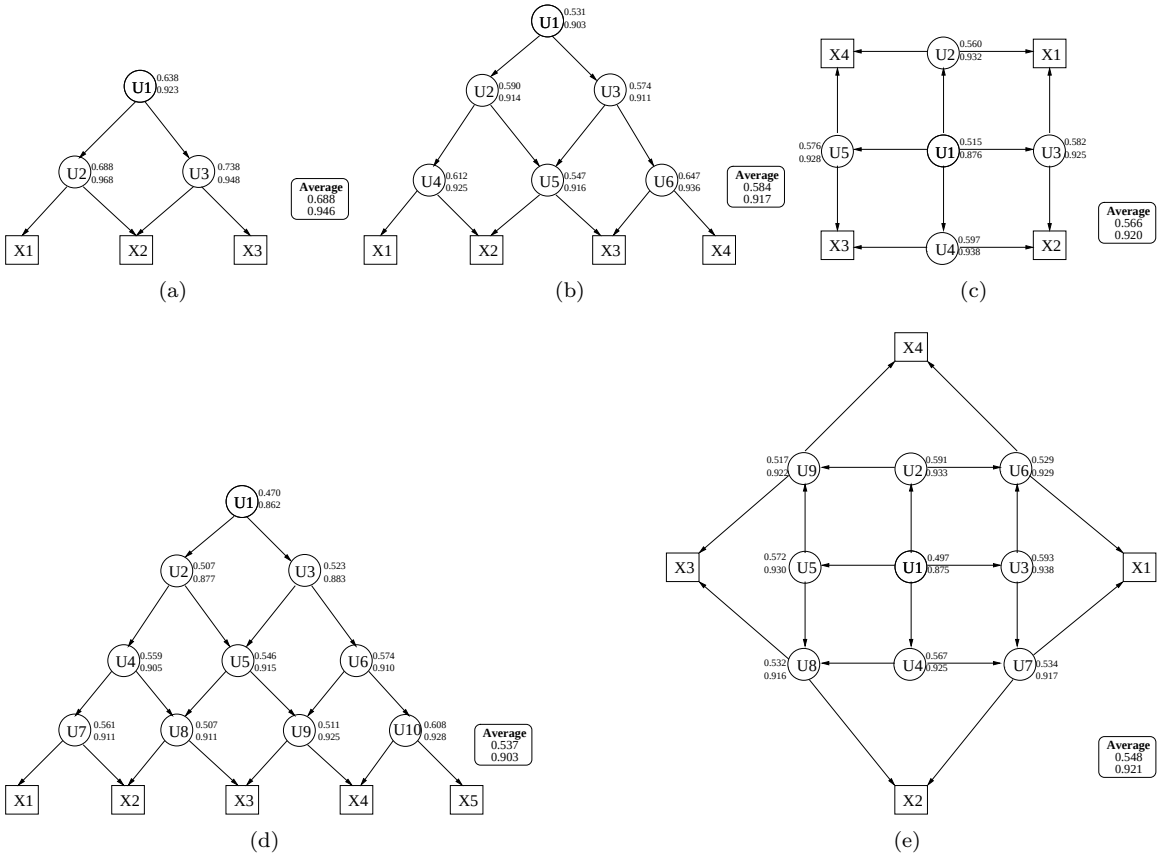

Figure 3: α_0 and α_1 values in five BN's. Predictions are made for U values conditional on the outcomes of all the X variables appearing in each of the BN's. The two values at each of the U nodes are α_0 and α_1 from top down.

same distance from the X variables are grouped together. For example, in panel (a), the grouping yields the sets, $\{U_1\}$ and $\{U_2, U_3\}$. The grouping for the other panels is summarized in the second column of Table 7. It is apparent in this table that α values decrease on average as the U variables appear farther away from the X variables. Note that, for panel (d), the U variables in set g_i^d are i edges away from the X variables if we count the number of edges of the shortest paths between one of the U variables and the X variables. It is interesting to see that the α values decrease in the order of g_1^a, g_1^b, g_1^d. This is due to the fact that the number of the X variables that affect the U variable increases in the order of the panels (a), (b), (d).

We consider panels (c) and (e) to compare the α values with those of panel (b), where each of these three graphs contains four X variables. It is noteworthy that the α values are almost the same between the sets, g_2^c and g_3^e, and between the sets, g_1^c and g_2^e or g_1^e. As for the two types of panels, panel (b) as one type and panels (c) and (e) as another, we can see that the α values depend on the

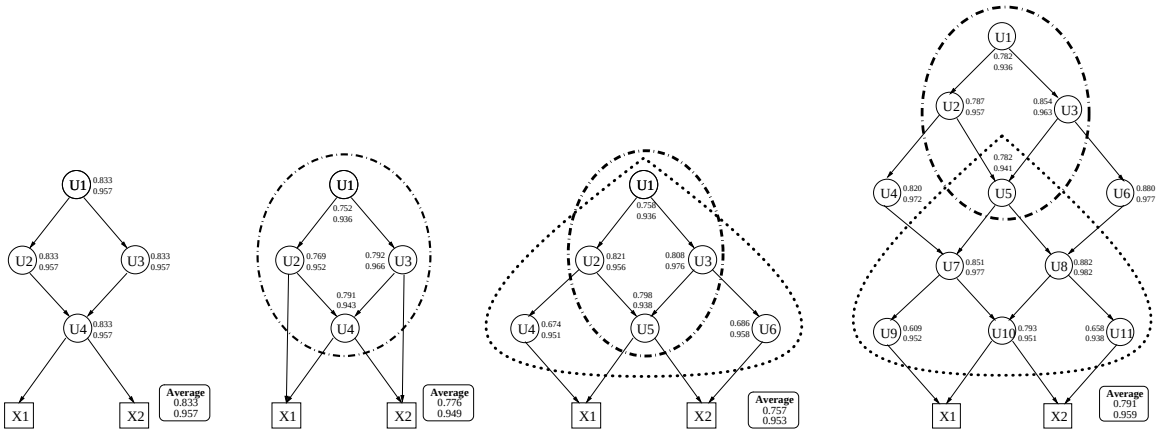

Figure 4: α_0 and α_1 values when predictions are made conditional on two variables only. Two patterns of subgraphs are enclosed in two types of lines, one in dotted chain lines and the other in dotted lines.

model structure also. While the α values of U_1 are more or less the same between panels (c) and (e), the α values in panel (b) ($\alpha_0 = 0.531,\ \alpha_1 = 0.903$) are different from those in either of panels (c) ($\alpha_0 = 0.515,\ \alpha_1 = 0.876$) and (e) ($\alpha_0 = 0.497,\ \alpha_1 = 0.875$).

The BN models in Figure 4 are considered to check whether the α values are affected by model structures when a fixed set of X variables is involved in the models. In panel (a), the α values of the U variables are all the same since the information from X_1 and X_2 propagates to the U variables through U_4 only. As for the panels, (b), (c), and (d), we see two patterns of subgraphs, one enclosed in dotted lines and the other in dotted chain lines. For convenience, we will call the former pattern 1 and the latter pattern 2. Pattern 1 includes a root node and consists of four U variables and pattern 2 includes the nodes that are directly connected to the X variable nodes. The figure indicates that the α values in the same pattern are almost the same across the panels. The result suggests that the α values of a U variable are affected by the location of the U variables relative to the X variables.

In summary, the results from the simulation experiment suggest that the α values of U variables get smaller as more X variables are involved in a model and that, for a given BN model, the α values of a U variable depend on its location relative to the X variables in the network.

6 Concluding remarks

Suppose that the variables that are involved in a model are all binary and positively associated and that we are interested in predicting for the variables in terms of class levels of probability. Then we may apply the notion of similarity between models. We have considered BN models for model

structures and the similarity between models is measured by expression (7).

We have considered a uniform distribution, a beta distribution, and a variation of the beta distribution for the distributions of the conditional probabilities, $P_{v|pa(v)}$, $v \in V$. We may use other distributions in accordance with the characteristics of the random variables that are involved in the BN model.

When we use a similar BN model such as proposed in this paper, it is desirable to provide the agreement levels, α_0 and α_1, because the levels are dependent upon the location of the variables to be predicted, in a given BN model, relative to the variables based on whose outcomes predictions are made. We must interpret the prediction result for a certain predicted variable if its agreement level is relatively low. For example, in the case of medical diagnosis, a low agreement level for a certain variable may lead to a further consultation or medical examination with regard to the variable.

A similar BN model is an artificial BN model where the conditional distribution for each node is selected based on the assumption that the data set for the similar BN model is generated from a distribution, $\mathcal{D}_P$. When a BN model involves a lot of random variables, model building is time consuming and the estimates become more subject to their initial values as more latent variables are included in the model [9].

Although we have considered BN models only, the classification robustness would be valid for other forms of graphical models such as Markov networks [15] and a mixture of the two types. One of the reasons is that these two forms of graphs share the Markov properties that are essentially transformable with a few exceptions into the same graphical layout [2, 3].

The result of this paper is relevant to all the problems where classifications may be made based on the relative magnitudes of the conditional probabilities under the assumption that the variables are binary and are positively associated with each other. Similar BN models will save much of our time and effort provided that the positive association condition (6) is satisfied, and so enhance the utility of model-based DSSs.

References

[1] S.K. Andersen, F.V. Jensen, K.G. Olesen and F. Jensen, HUGIN: A shell for building Bayesian belief universes for expert systems [computer program] HUGIN Expert Ltd., Aalborg, Denmark, 1989.

[2] S.A. Andersson, D. Madigan and M.D. Perlman, On the Markov equivalence of chain graphs, undirected graphs, and acyclic digraphs, Scandinavian Journal of Statistics 24 (1997a) 81-102.

[3] S.A. Andersson, D. Madigan and M.D. Perlman, A characterization of Markov equivalence classes for acyclic digraphs. The Annals of Statistics 25(2) (1997b) 505-541.

[4] M.H. DeGroot, Optimal Statistical Decisions, McGraw-Hill Book Company, New York, 1970.

[5] ERGO [computer program] Noetic Systems Inc., Baltimore, MD (1991).

[6] P.W. Holland and P.R. Rosenbaum, Conditional association and unidimensionality in monotone latent variable models, The Annals of Statistics 14(4) (1986) 1523-1543.

[7] F.V. Jensen, An Introduction to Bayesian Networks, Springer-Verlag, NewYork, 1996.

[8] B.W. Junker and J.L. Ellis, A characterization of monotone unidimensional latent variable models, The Annals of Statistics 25(3) (1997) 1327-1343.

[9] S.-H. Kim, Calibrated initials for an EM applied to recursive models of categorical variables, Computational Statistics and Data Analysis, 40(1) (2002) 97-110.

[10] S.-H. Kim, Stochastic ordering and robustness in classification from a Bayesian network, Decision Support Systems 39 (2005) 253-266.

[11] O.I. Larichev, A.V. Kortnev, D. Y. Kochin, Decision support systems for classification of a finite set of multicriteria alternatives, Decision Support Systems 33 (2002) 13-21.

[12] S.L. Lauritzen and D.J. Spiegelhalter, Local Computations with Probabilities on Graphical Structures and their Application to Expert Systems, J. of Royal Statistical Soc. B 50(2) (1988) 157-224 .

[13] R.J. Mislevy, Evidence and inference in educational assessment, Psychometrika 59(4) (1994) 439-483.

[14] H.A.David and H.N.Nagaraja, Order Statistics, Wiley, 2003.

[15] J. Pearl, Probabilistic Reasoning In Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann, San Mateo, CA, 1988.

[16] J.P. Shim, M. Warkentin, J.F. Courtney, D. J. Power, R. Sharda, C. Carlsson, Past, present, and future of decision support technology, Decision Support Systems 33 (2002) 111-126.

[17] P.C. Verhoef, B. Conkers, Predicting customer potential value an application in the insurance industry, Decision support systems 32 (2001) 189-199.

[18] S. Walczak, W.E. Pofahl, R.J. Scorpio, A decision support tool for allocating hospital bed resources and determining required acuity of care, Decision Support Systems 34 (4) (2003) 445-456.r

[19] J. Whittaker, Graphical Models in Applied Multivariate Statistics, Wiley, New York, 1990.