Available online at www.sciencedirect.com
SCIENCE DIRECT®
European Journal of Operational Research 156 (2004) 508–523
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH
www.elsevier.com/locate/dsw

Computing, Artificial Intelligence and Information Technology

Bayesian network classifiers for identifying the slope of the customer lifecycle of long-life customers

Bart Baesens [a], Geert Verstraeten [b], Dirk Van den Poel [b,*], Michael Egmont-Petersen [c], Patrick Van Kenhove [b], Jan Vanthienen [a]

[a] *Department of Applied Economic Sciences, K.U. Leuven, Naamsestraat 69, B-3000 Leuven, Belgium*
[b] *Department of Marketing, Faculty of Economics and Business Administration, Ghent University, Hoveniersberg 24, B-9000 Ghent, Belgium*
[c] *Institute of Information and Computing Sciences, Utrecht University, Padualaan 14, De Uithof, The Netherlands*

Received 18 December 2001; accepted 5 December 2002

Abstract

Undoubtedly, customer relationship management has gained its importance through the statement that acquiring a new customer is several times more costly than retaining and selling additional products to existing customers. Consequently, marketing practitioners are currently often focusing on retaining customers for as long as possible. However, recent findings in relationship marketing literature have shown that large differences exist within the group of long-life customers in terms of spending and spending evolution. Therefore, this paper focuses on introducing a measure of a customer's future spending evolution that might improve relationship marketing decision making. In this study, from a marketing point of view, we focus on predicting whether a newly acquired customer will increase or decrease his/her future spending from initial purchase information. This is essentially a classification task. The main contribution of this study lies in comparing and evaluating several Bayesian network classifiers with statistical and other artificial intelligence techniques for the purpose of classifying customers in the binary classification problem at hand. Certain Bayesian network classifiers have been recently proposed in the artificial intelligence literature as probabilistic white-box classifiers which allow to give a clear insight into the relationships between the variables of the domain under study. We discuss and evaluate several types of Bayesian network classifiers and their corresponding structure learning algorithms. We contribute to the literature by providing experimental evidence that: (1) Bayesian network classifiers offer an interesting and viable alternative for our customer lifecycle slope estimation problem; (2) the Markov Blanket concept allows for a natural form of attribute selection that was very effective for the application at hand; (3) the sign of the slope can be predicted with a powerful and parsimonious general, unrestricted Bayesian network classifier; (4) a set of three variables measuring the volume of initial purchases and the degree to which customers originally buy in different categories, are powerful predictors for estimating the sign of the slope, and might therefore provide desirable additional information for relationship marketing decision making.

* Corresponding author. Tel.: +32-9-264-8980; fax: +32-9-264-4279.
E-mail addresses: bart.baesens@econ.kuleuven.ac.be (B. Baesens), geert.verstraeten@rug.ac.be (G. Verstraeten), dirk.vandenpoel@rug.ac.be (D. Van den Poel), michael@cs.uu.nl (M. Egmont-Petersen), patrick.vankenhove@rug.ac.be (P. Van Kenhove), jan.vanthienen@econ.kuleuven.ac.be (J. Vanthienen).

0377-2217/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0377-2217(03)00043-2

Keywords: Artificial intelligence; Bayesian network classifiers; Marketing; CRM; Customer loyalty

1. Introduction

Undoubtedly, customer relationship management (CRM) has gained its importance through the statement that acquiring a new customer is several times more costly than retaining and selling additional products to existing customers [2,20,46]. This simple rule-of-thumb has led to what many authors refer to as 'the paradigm shift in marketing' [4,25], implying that brand strategies are being replaced by customer strategies [3], and more and more voices rise to replace the traditional brand managers by customer (segment) managers [37,47]. Hence, it has become increasingly important to make informed marketing decisions on a customer level, and the customer loyalty of individual consumers has rapidly grown to become the focal point of relationship marketing (see, e.g., [22,31, 40,41]).

In order to ensure the success of a CRM strategy, it is crucial that customers remain, at least to a certain extent, loyal to the company in case. However, recent research suggests large heterogeneity in terms of spending and spending evolution within the group of long-life customers [44]. Responding to this finding, in the following section of the paper, we elaborate upon the relevance of an accurate indication of a customer's future spending evolution for improving relationship marketing decision making for long-life customers. Consequently, we try to account for the heterogeneity within the group of long-life customers by adding information about estimated future spending evolutions.

In this study, we limit the focus to estimating whether newly acquired customers will increase or decrease their future spending. Whereas, to the best of our knowledge, no published study has attempted to forecast this variable, we argue in the following section that the recently evolving literature around the loyalty issue has motivated us to do so. To this end, we will use and compare different recently developed classification techniques for optimally classifying the customers into the two relevant groups (i.e. customers with decreasing versus increasing spending). We hereby focus on techniques that besides yielding good classification accuracy also represent the marginal and conditional independence relations between the variables and how they jointly affect the classification decision.

In recent artificial intelligence literature, Bayesian networks have been suggested as probabilistic white-box models that are able to capture even higher-order dependencies between sets of variables. These networks can then also be efficiently adopted for classification purposes. In this paper, we will evaluate and compare several Bayesian network classifiers for the purpose of classifying customers in the binary classification problem at hand. Using the Naive Bayes classifier as a point of origin, we will gradually remove the restrictions put on the network structure and investigate Tree Augmented Naive Bayes classifiers (TANs) followed by completely unrestricted Bayesian network classifiers. Comparisons will be made with statistical and other artificial intelligence techniques. All classifiers will be evaluated by looking at their classification accuracy and the area under the receiver operating characteristic curve (AUROC). The latter basically illustrates the behavior of a classifier without regard to class distribution or misclassification cost, so it effectively decouples classification performance from these factors. Furthermore, we will also look at the complexity of the trained classifiers because from a marketing viewpoint, parsimonious, yet accurate and self-explanatory models are to be preferred.

This paper is organized as follows. In Section 2, we elaborate on the recent literature on relationship marketing that has provided motivation for investigating the predictability of the customer's spending evolution. To this end, we use Bayesian network classifiers which are discussed in Section 3. The design of the study, including both the data set description and the used performance criteria, are presented in Section 4. Section 5 presents the

results of the experiments. Finally, Section 6 concludes the paper.

2. Relevance of the estimation of a customer's spending evolution

Advocates of traditional relationship marketing attribute several advantages to loyal customers. Most importantly, these are expected to raise their spending (and contribution to the company) over their relationship with the company in case [43]. In the most optimistic settings, they are said to generate new customers by their positive word-of-mouth [22], ensure diminished costs to serve [31], exhibit reduced consumer price sensitivities [42] and have a salutary impact on the company's employees [43]. Since, from a database-driven approach, customer tenure (i.e. the length of a customer's relationship with a company) has often been used to approximate the loyalty construct [22,44,45], relationship marketing thrives on the idea that raising the length of the customer–company relationship is the main lever for a company's financial success [43].

Nevertheless, in their recent article, Reinartz and Kumar [45] report a series of studies across industries that challenges most claims of the loyalty advocates. In these studies, they have found no evidence to suggest that long-life customers with steady purchase behavior are necessarily cheaper to serve, less price sensitive, or more effective in bringing new business to the company, such as through word-of-mouth referrals. Additionally, in a previous article, Reinartz and Kumar [44] showed that the contributions of long-life customers were generally declining, although the analysis of this issue was not the focus of their discussion. Finally, the authors pointed out that, at least for a non-contractual setting, short-life but high-revenue customers accounted for a sizeable amount of profits for the mail-order company in case [44].

In the article mentioned above, Reinartz and Kumar clearly illustrate the pitfalls involved with spending a large slice of the marketing budget on customers that have been good customers in the past over a short-period of time, yet tend to show a decreasing spending pattern (i.e. customers that have been labelled 'butterflies') [45]. In the example of a mail-order setting, it is generally known that repurchase behavior can—and has—effectively been modeled by using an (often linear) combination of RFM variables, representing the recency of a customer's last purchase, the average frequency of the customer's purchases and the average monetary value spent on the customer's purchase occasions [12,50]. Hence, the group of customers called 'butterflies', being customers with a high historical monetary value, will tend to be over selected for mailing campaigns [45]. An estimation of the future slope of the customer lifecycle (i.e. a customer's spending evolution) would then likely be able to deliver the required insights to the decision-making process and the understanding of the relationship between the slope and other variables, such as customer spending, might generate rich qualitative information for marketers. For instance, for this group of customers, the company might decide to attempt to improve its return on (direct) marketing investments by shifting its focus from long-term investments to investments or promotions on which a short-term return is possible. Alternatively, the company might even consider abandoning investments in these customers altogether. Thus, in this customer-based view, the a priori knowledge of the slope of the customer lifecycle would be useful information.

In this research study we limit our attention in terms of marketing contribution to proving that it is possible to predict the slope of the customer lifecycle of long-life customers. Accordingly, due to the limitations that are extensively documented in Section 7 of this paper, it is not within the scope of this paper to devise, implement and test an optimal marketing strategy for a specific company in case, nor for an array of companies in industries with different characteristics. In this attempt, we will compare different techniques for the estimation problem, which can in its essential form be transformed into a binary classification problem: 'Will newly acquired customers increase or decrease their spending after their first purchase experiences?'

In the marketing literature, binary classification problems have typically been tackled by using traditional statistical methods (e.g. discri-

minant analysis and logistic regression [2,50]), non-parametric statistical models (e.g. k-nearest neighbour [50] and decision trees [49,50]) and neural networks [2,50]. In this paper, we will adopt Bayesian network classifiers which have been recently introduced in the artificial intelligence literature. This is motivated by the fact that Bayesian network classifiers are probabilistic white-box models which facilitate a clear insight into the underlying dependencies pertaining to the domain under study. They are based on solid probabilistic reasoning and offer a great potential for knowledge discovery in data in a marketing context. Unfortunately, despite their attractive properties, their application for business decision making and marketing purposes is still limited. In the following section, we will elaborate on the basic concepts of Bayesian network classifiers and discuss some recently suggested structure learning algorithms.

3. Bayesian networks for classification

A Bayesian network (BN) represents a joint probability distribution over a set of discrete, stochastic variables. It is to be considered as a probabilistic white-box model consisting of a qualitative part specifying the conditional (in)dependencies between the variables and a quantitative part specifying the conditional probabilities of the data set variables [36]. Formally, a Bayesian network consists of two parts $B = \langle G, \Theta \rangle$. The first part G is a directed acyclic graph consisting of nodes and arcs. The nodes are the variables $X_1, \ldots, X_n$ in the data set whereas the arcs indicate direct dependencies between the variables. The graph G then encodes the independence relationships in the domain under investigation. The second part of the network, Θ, represents the conditional probability distributions. It contains a parameter $\theta_{x_i|\Pi_{x_i}} = P_B(x_i \mid \Pi_{x_i})$ for each possible value x_i of X_i, given each combination of the direct parent variables of X_i, Π_{x_i} of Π_{X_i}, where Π_{X_i} denotes the set of direct parents of X_i in G. The network B then represents the following joint probability distribution:

$$P_B(X_1, \ldots, X_n) = \prod_{i=1}^{n} P_B(X_i \mid \Pi_{X_i}) = \prod_{i=1}^{n} \theta_{X_i|\Pi_{X_i}}. \quad (1)$$

The first task when learning a Bayesian network is to find the structure G of the network. Once we know the network structure G, the parameters Θ need to be estimated. In general, these two estimation tasks are performed separately. In this paper, we will use the empirical frequencies from the data D to estimate these parameters: [1]

$$\theta_{x_i|\Pi_{x_i}} = \hat{P}_D(x_i \mid \Pi_{x_i}). \quad (2)$$

It can be shown that these estimates maximise the log likelihood of the network B given the data D [21]. Note that these estimates might be further improved by a smoothing operation [21].

A Bayesian network is essentially a statistical model that makes it feasible to compute the (joint) posterior probability distribution of any subset of unobserved stochastic variables, given that the variables in the complementary subset are observed. This functionality makes it possible to use a Bayesian network as a statistical classifier by applying the winner-takes-all rule to the posterior probability distribution for the (unobserved) class node [15]. The underlying assumption behind the winner-takes-all rule is that all gains and losses are equal (for a discussion of this aspect see, e.g., [15]). In what follows, we will discuss several structure learning algorithms for developing Bayesian network classifiers.

3.1. The Naive Bayes classifier

A simple classifier, which in practice often performs surprisingly well, is the Naive Bayes classifier [15,30,33]. This classifier basically learns the class-conditional probabilities $P(X_i = x_i \mid C = c_l)$ of each variable X_i given the class label c_l. A new test case $(X_1 = x_1, \ldots, X_n = x_n)$ is then classified by using Bayes' rule to compute the posterior probability of each class c_l given the vector of observed variable values:

$$P(C = c_l \mid X_1 = x_1, \ldots, X_n = x_n) = \frac{P(C = c_l) P(X_1 = x_1, \ldots, X_n = x_n \mid C = c_l)}{P(X_1 = x_1, \ldots, X_n = x_n)}. \quad (3)$$

[1] Note that we hereby assume that the data set is complete, i.e. no missing values.

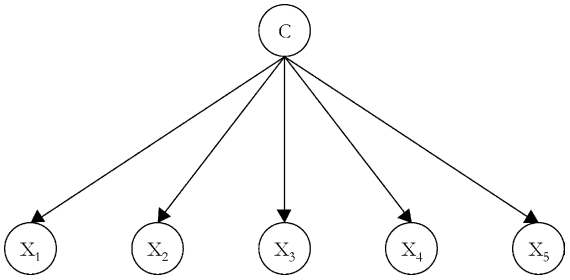

Fig. 1. The Naive Bayes classifier.

The simplifying assumption behind the Naive Bayes classifier then assumes that the variables are conditionally independent given the class label. Hence,

$$P(X_1 = x_1, \ldots, X_n = x_n \mid C = c_l) = \prod_{i=1}^{n} P(X_i = x_i \mid C = c_l). \quad (4)$$

This assumption simplifies the estimation of the class-conditional probabilities from the training data. Notice that one does not estimate the denominator in expression 3 since it is independent of the class. Instead, one normalises the nominator term $P(C = c_l)P(X_1 = x_1, \ldots, X_n = x_n \mid C = c_l)$ to 1 over all classes. Naive Bayes classifiers are easy to construct since the structure is given a priori and no structure learning phase is required. The probabilities $P(X_i = x_i \mid C = c_l)$ are estimated by using the frequency counts for the discrete variables and a normal or kernel density based method for continuous variables [30]. Fig. 1 provides a graphical representation of a Naive Bayes classifier.

3.2. Tree Augmented Naive Bayes classifiers

In [21] Tree Augmented Naive Bayes classifiers (TANs) were presented as an extension of the Naive Bayes classifier. TANs relax the independence assumption by allowing arcs between the variables. An arc from variable X_i to X_j then implies that the impact of X_i on the class variable also depends on the value of X_j. An example of a TAN is presented in Fig. 2. In a TAN network the class variable has no parents and each variable has as parents the class variable and at most one other variable. The variables are thus only allowed to

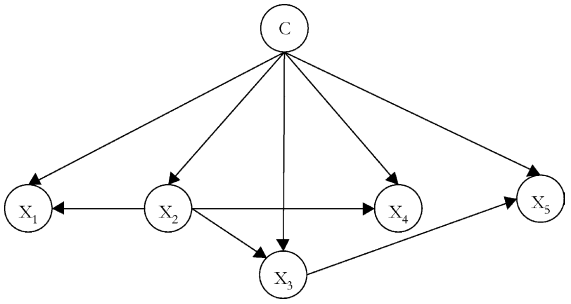

Fig. 2. The Tree Augmented Naive Bayes classifier.

form a tree structure. In [21], a procedure was presented to learn the optional arrows in the structure that forms a TAN network. This procedure is based on an earlier algorithm suggested by Chow and Liu (CL) [11]. The procedure consists of the following five steps.

1. Compute the conditional mutual information given the class variable C, $I(X_i; X_j \mid C)$, between each pair of variables, $i \neq j$. $I(X_i; X_j \mid C)$ is defined as follows:
$$I(X_i; X_j \mid C) = \sum_{x_i, x_j, c_l} P(X_i = x_i, X_j = x_j, C = c_l) \times \log \frac{P(X_i = x_i, X_j = x_j \mid C = c_l)}{P(X_i = x_i \mid C = c_l)P(X_j = x_j \mid C = c_l)}. \quad (5)$$
This function is an approximation of the information that X_j provides about X_i (and vice versa) when the value of C is known.
2. Build a complete undirected graph in which the nodes are the variables. Assign to each arc connecting X_i to X_j the weight $I(X_i; X_j \mid C)$.
3. Build a maximum weighted spanning tree.
4. Transform the resulting undirected tree to a directed one by choosing a root variable and setting the direction of all arcs to be outward from it.
5. Add the classification node C and draw an arc from C to each X_i.

We used Kruskal's algorithm in step 3 to construct the maximum weighted spanning tree [32]. In [21], it was proven that the above procedure builds TANs that maximise the log likelihood of

the network given the training data and has time complexity $O(n^2 \cdot N)$ with n the number of variables and N the number of data points. Experimental results indicated that TANs outperform Naive Bayes with the same computational complexity and robustness [21].

3.3. General Bayesian Network classifiers

Many algorithms have been proposed that can learn the structure of a General Bayesian Network (GBN) from a set of (complete) data [5,29]. Some algorithms impose restrictions onto the direction of the arcs that connect the nodes whereas other algorithms omit such restrictions. In this paper, we use the learning algorithm of Cheng et al. [7,9], which assumes an a priori ordering of the variables. Before we discuss the different steps of this algorithm, we first elaborate on the concept of d-separation because this plays a pivotal role in the structure learning algorithm.

Let X, Y and Z be mutually disjoint sets of nodes in a directed acyclic graph G. The set Y is said to d-separate the sets X and Z in G if for every node $X_i \in X$ and every node $X_j \in Z$, every chain (of any directionality) from X_i to X_j in G is blocked by Y [51]. We say that a chain s is blocked by a set of nodes Y if s contains three consecutive nodes X_1, X_2, X_3, for which one of the following conditions holds [51]:

1. arcs $X_1 \leftarrow X_2$ and $X_2 \rightarrow X_3$ are on the chain s, and $X_2 \in Y$;
2. arcs $X_1 \rightarrow X_2$ and $X_2 \rightarrow X_3$ or $X_1 \leftarrow X_2$ and $X_2 \leftarrow X_3$ are on the chain s, and $X_2 \in Y$;
3. arcs $X_1 \rightarrow X_2$ and $X_2 \leftarrow X_3$ are on the chain s and X_2 and the descendants of X_2 are not in Y.

It can be shown that if sets of variables X and Z are d-separated by Y in a directed acyclic graph G, then X is independent of Z conditional on Y in every distribution compatible with G [24,52]. It is precisely this property that will be exploited in the algorithm of Cheng to learn the Bayesian network structure.

The algorithm consists of four phases. In a first phase, a draft of the network structure is made based on the mutual information between each pair of nodes. The second and third phase then add and remove arcs based on the concept of d-separation and conditional independence tests. Finally, in the fourth phase, the Bayesian network is pruned and its parameters are estimated.

The algorithm proceeds as follows [7,9].

Phase 1: Drafting

1. Initiate a graph $G(X, A)$ where $X = \{X_1, X_2, \ldots, X_n, C\}$ and $A = \{\}$. Initiate two empty ordered sets S and R.
2. Compute the (non-parametric) mutual information $I(X_i; X_j)$ between each pair of variables where $X_i, X_j \in X$, $i \neq j$. $I(X_i; X_j)$ is defined as follows:
$$I(X_i; X_j) = \sum_{x_i, x_j} P(X_i = x_i, X_j = x_j) \times \log \frac{P(X_i = x_i, X_j = x_j)}{P(X_i = x_i) P(X_j = x_j)}. \quad (6)$$
The mutual information $I(X_i; X_j)$ is the amount of information gained about X_i when X_j is known, and vice versa ($I(X_i; X_j) = I(X_j; X_i)$). Hence, $I(X_i; X_j) = 0$ if and only if X_i and X_j are independent.
3. Sort all pairs of nodes where $I(X_i; X_j)$ is greater than ϵ from large to small and put them into an ordered set S. In our experiments, we set $\epsilon = 0.008$ which is an appropriate value for large data sets [9].
4. Add arcs to A according to the first two pairs of nodes in S and remove them from S. The direction of the arcs is decided by the a priori node ordering.
5. Get the first pair of nodes remained in S and remove it from S. If there is no open path between the two nodes, add the corresponding arc to A. Otherwise, add the pair of nodes to the end of an ordered set R. Note that an open path is a chain with no collider nodes whereby a collider node is a node having two incoming arcs.
6. Repeat step 5 until S is empty.

Phase 2: Thickening

7. Get the first pair of nodes in R and remove it from R.

8. Find a cut-set that can d-separate these two nodes in the current network. Use a conditional independence test (see Eq. (5)) to see if these two nodes are conditionally independent given the cut-set and using a threshold value of 0.008. If so, go to the next step, otherwise, connect the pair of nodes by an arc.
9. Repeat step 7 until R is empty.

Phase 3: Thinning

10. For each arc in A, if there are other paths besides this arc between the two nodes, remove this arc from A temporarily and find a cut-set that can d-separate the two nodes in the current network. Use a conditional independence test to see if the two nodes are conditionally independent given the cut-set and again using a threshold value of 0.008. If so, remove the arc permanently, otherwise add the arc back to the network.

Phase 4: Prune and learn the parameters of the Bayesian network classifier

11. Find the Markov Blanket of the classification node. The Markov Blanket of a node X_i consists of the union of X_i's parents, X_i's children and the parents of X_i's children [36].
12. Delete all the nodes that are outside the Markov Blanket.
13. Learn the parameters of the conditional probability tables and output the Bayesian network classifier.

Note that in steps 8 and 10, it is important to find cut-sets that are as small as possible in order to avoid conditional independence tests with large condition sets. In [1], a correct algorithm is presented to find minimum cut-sets between two nodes. In this paper, we will use the heuristic algorithm suggested by Cheng et al. [7].

It can be shown that when the values of the variables in the Markov Blanket of the classification node are observed, the posterior probability distribution of the classification node is independent of all other variables (nodes) not in the Markov Blanket [34]. Hence, in step 12, all variables outside the Markov Blanket can be safely deleted because they will have no impact on the classification node and thus will not affect the classification accuracy. In this way, the Markov Blanket results in a natural form of variable selection.

Note that this algorithm requires $O(N^2)$ mutual information tests and is linear in the number of cases N. An extension has been presented in [8] in case no node ordering is given. In this paper, we will simply treat the classification node as the first node and order the other nodes based on their correlation with the classification node from large to small.

3.4. Multinet Bayesian network classifiers

Both TANs and GBNs assume that the relations between the variables are the same for all classes. A multinet Bayesian network allows for more flexibility and is composed of a separate, local network for each class and a prior probability distribution of the class node [10,21,23,28]. Thus, for each value c_i of the classification node C a Bayesian network structure B_i is learned. The multinet M then defines the following joint probability distribution:

$$P_M(C, X_1, \ldots, X_n) = P_C(C) \cdot P_{B_i}(X_1, \ldots, X_n). \quad (7)$$

A new instance is then assigned to the class that maximises the posterior probability $P_M(C \mid X_1, \ldots, X_n)$ conform the winner-takes-all rule. Since we have

$$P_M(C \mid X_1, \ldots, X_n) = \frac{P_M(C, X_1, \ldots, X_n)}{P_M(X_1, \ldots, X_n)}, \quad (8)$$

and the denominator is the same for all classes, we can assign the instance to the class that maximises the value of Eq. (7). The term $P_C(C)$ may then be estimated by the empirical frequency of the class variable in the training set $\hat{P}_D(C)$. Note that for multinet classifiers the number of parameters that need to be estimated per training instance inevitably increases. As the parameters are estimated from a limited number of instances, learning a separate multinet structure per class instead of one overall structure results in more unreliable parameter estimates and, hence, a higher probability of

overgeneralization. This effect is closely related to the so-called peaking phenomenon, for a discussion see, e.g., [53].

In this paper, we consider both CL multinets and GBN multinets. CL multinets are multinets which are built using the procedure of Chow and Liu [11]. This is essentially the same procedure as the one outlined in Section 3.2 with the exception that step 5 is now omitted and in step 1 the conditional mutual information is replaced by the mutual information (see Eq. (6)). This procedure is then executed separately for each value c_i of the class node C using only the training data D_i whereby D_i contains all instances of D for which $C = c_i$. The resulting multinet then consists of an ensemble of tree structured Bayesian networks. The GBN multinets are trained using the approach of Cheng discussed in Section 3.3 with the exception that the classification node is now omitted in the structure learning phase. Again, the algorithm is executed for each class on the corresponding training data.

4. Design of the study

4.1. Data set

We conducted our research on UPC scanner data of a large Belgian DIY (Do-It-Yourself) retail chain. The data we used for our models were all gathered by the customer loyalty cards, which have been in use since January 1995. Due to some restrictions (cf. infra), we were able to use four complete years of information.

Since we are interested in examining the behavior of long-life customers, we imposed three conditions on the data: firstly, we only used customers who started purchasing before February 1997. Secondly, to ensure the data was not left-censored (i.e. to ensure the customers in our database really started their relationship with the company at the time of our first observation), we only used information of customers who had not purchased before. We thus used the first two years of information in the customer database only to check that the customers in our sample were new customers. Thirdly, using a database containing eight six-month periods of information for all customers of the company, we have selected all customers who purchased in five or more periods. Hence, we arrived at a database containing an approximate sample of the company's long-life customers. In order to assess the quality of our models, we have randomly divided the database into two parts. While 2/3 of the observations were used for learning the classifiers, the remaining 1/3 was used as a test set for estimating the generalization behavior of the classifiers. Table 1 displays the characteristics of our data set.

By performing a linear regression model on the historical contributions of each customer, we were able to capture the slope of the lifecycle of each individual customer. This slope, after being discretised into positive or negative to represent increasing or decreasing spending, was henceforth used as the dependent variable in the study (SlopeSign). It is interesting to note that the finding of Reinartz and Kumar that the slope of long-life customers was generally decreasing [44] was validated in our study by the fact that only 28% of those customers in the database exhibited a positive slope. In this case, we have used a set of 15 continuous variables computed on the first six months of information, in order to predict the sign of the evolution of the customer's contribution (i.e. the customer lifecycle) for the remaining 42 months of the relationship. While the variables computed are presented in Table 2, the time schedule is given in Fig. 3.

The independent variables can be divided into four major logical groups. A first group of variables is constructed to measure the volume of the purchases the subject made during his or her first six months as a customer. These contain TotCont, TotRev, NumbArt and NumbTick. Note that the variable TotCont represents the intercept of the customer lifecycle. It is merely the first of the eight

Table 1
Data set characteristics

Data set size	3827 observations
Training set size	2551 observations
Test set size	1276 observations
Number of attributes	15

Table 2
Variables used in the study

1	Total contribution	TotCont
	Total revenues	TotRev
	Total number of articles bought	NumbArt
	Total number of visits to the store (tickets)	NumbTick
2	Amount of different categories purchased	DiffCat
	Amount of different products purchased	DiffProd
	Maximum percentage of products bought in one product family	MaxPerc
3	Mean margin of articles purchased	MeanMarg
	Mean price of articles purchased	MeanPrice
	Maximum price paid for an article	MaxPrice
	Total value of received discounts/ total revenues	PercDisc
	Articles bought in discount/total amount of articles bought	ArtDisc
4	Slope of the 'customer lifecycle' during the first six months	Lifec6m
	Contribution in the sixth month	LastCont
	Date the maximum price was paid	DateMaxPrice

data points forming the customer lifecycle. While this first set of attributes can be regarded as the "depth" of the customer purchases, the second group of variables contains the variables that measure the "broadness" of the purchases. These are DiffCat, DiffProd and MaxPerc. The latter variable contains the percentage of products bought in the product category in which the customer has bought most of his or her products. In this way, it can be seen as a skewness indicator, a large indicator meaning that the customer only buys a certain category of products from the company. A third group of variables captures the 'bargaining tendency' and 'price sensitivity' of the customer. The relevant variables here are PercDisc, ArtDisc, MeanMarg, MeanPrice and MaxPrice. Finally, three measures are introduced to value evolutions within the first six months. These are Lifec6m, LastCont and DateMaxPrice.

In order to train the Bayesian network classifiers, we discretised all variables by using the discretisation algorithm of Fayyad and Irani with the default options [19]. This algorithm uses an information entropy minimisation heuristic to discretise the range of a continuous-valued attribute into multiple intervals. This discretisation procedure was performed using the Java Weka workbench. [2] Table 3 depicts how the attributes in our data set were discretised into intervals.

4.2. Performance criteria for classification

The performance of all trained classifiers will be quantified using both the classification accuracy and the AUROC. The classification accuracy is undoubtedly the most commonly used measure of performance of a classifier. It simply measures the percentage of correctly classified (PCC) observations. However, it tacitly assumes equal misclassification costs and balanced class distributions [38]. The receiver operating characteristic curve (ROC) is a two-dimensional graphical illustration of the sensitivity ('true alarms') on the Y-axis versus 1-specificity on the X-axis ('false alarms') for various values of the classification threshold [16,48]. It basically illustrates the behavior of a classifier without regard to class distribution or misclassification cost. The AUROC then provides a simple figure-of-merit for the performance of the constructed classifier. An intuitive interpretation of the AUROC is that it provides an estimate of the probability that a randomly chosen instance of class 1 is correctly rated (or ranked) higher than a randomly selected instance of class 0 [26].

We will use McNemar's test to compare the PCCs of different classifiers [17]. This chi-squared test is based upon contingency table analysis to detect statistically significant performance differences between classifiers. In [14], it was shown that this test has acceptable Type I error which is the probability of incorrectly detecting a difference when no difference exists. While Hanley and McNeil described a method for comparing ROC curves derived from the same sample [27], De Long et al. [13] developed a non-parametric chi-squared test by using the theory on generalized U-statistics and the method of structural components to estimate the covariance matrix of the AUROC. Hence, we will use the latter test to detect

[2] http://www.cs.waikato.ac.nz/ml/weka/.

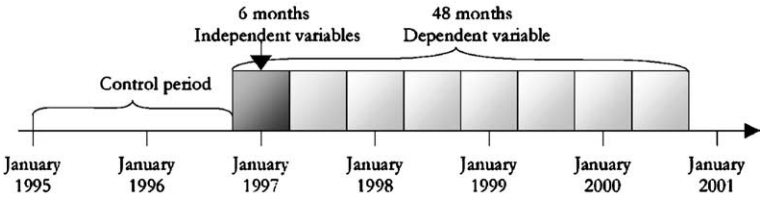

Fig. 3. Time schedule of our empirical study.

Table 3
Discretisation of the attributes

Attribute	Values	Encoding
TotCont	1,2,3,4	]−∞;241.74],]241.74;817.92],]817.92;3158.09],]3158.09;∞]
TotRev	1,2,3,4	]−∞;679.53],]679.53;2481.82],]2481.82;7410.12],]7410.12;∞]
NumbArt	1,2,3,4	]−∞;4],]4;13],]13;38],]38;∞]
NumbTick	1,2,3	]−∞;2],]2;5],]5;∞]
DiffCat	1,2,3,4	]−∞;2],]2;6],]6;13],]13;∞]
DiffProd	1,2,3	]−∞;4],]4;11];]11;∞]
MaxPerc	1,2,3,4	]−∞;0.49],]0.49;0.5],]0.5;0.98],]0.98;∞]
MeanMarg	1,2	]−∞;0.53],]0.53;∞]
MeanPrice	1,2	]−∞;118.16],]118.16;∞]
MaxPrice	1,2,3,4	]−∞;165],]165;549],]549;1095],]1095;∞]
PercDisc	1,2	]−∞;0.17],]0.17;∞]
ArtDisc	1,2,3	]−∞;0],]0;0.33],]0.33;∞]
Lifec6m	1,2,3	]−∞;-72.24],]-72.24;87.61],]87.61,∞]
LastCont	1,2,3	]−∞;621.85],]621.85;2010.85],]2010.85;∞]
DateMaxPrice	1,2	]−∞;13544],]13544;∞]

statistically significant AUROC differences between classifiers.

5. Results

We compared and contrasted the performance of the Naive Bayes, TAN, CL multinet, GBN, GBN multinet, C4.5, C4.5rules, linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) classifiers on our marketing data set. We included the decision tree induction algorithm C4.5 and its rules variant, C4.5rules, because they are also white-box classifiers giving besides a classification decision also a clear explanation why the particular classification is being made [39]. LDA and QDA were included because they are well-known benchmark statistical classifiers. To train the Naive Bayes, TAN, and CL multinet classifiers, we used the Matlab toolbox of Kevin Murphy [35]. For the GBN and GBN multinet classifiers, we used the PowerPredictor software of Cheng [6]. Table 4 depicts the classification accuracy of all classifiers on both the training and test set. The best test set performance is in bold face and underlined and those not statistically different from it according to McNemar's test (using a significance level of 5%) are in bold face. The GBN classifier achieved the highest classification accuracy on the test set. The classification accuracy of the TAN, C4.5 and LDA classifier was not statistically different from it. Table 5 depicts the AUROC of all classifiers and has the same setup as Table 4. Note that for the Bayesian network classifiers, the LDA and QDA classifier, the calculation of the AUROC values poses no problems since each of these classifiers yields class probabilities. For C4.5, we use the confidence at the

Table 4
Classification accuracy of the Bayesian network classifiers versus C4.5 and discriminant analysis

	Training set	Test set
Naive Bayes	71.0	72.5
TAN	74.9	**74.0**
CL multinet	74.2	72.3
GBN	75.3	**75.0**
GBN multinet	70.6	72.3
C4.5	76.7	**74.1**
C4.5rules	77.8	73.3
LDA	75.5	**74.1**
QDA	72.9	72.7

Table 5
Area under the receiver operating curve of the Bayesian network classificiers versus C4.5 and discriminant analysis

	Training set	Test set
Naive Bayes	75.9	**74.3**
TAN	77.8	73.6
CL multinet	77.0	72.6
GBN	77.5	**74.7**
GBN multinet	76.6	74.0
C4.5	76.5	73.8
C4.5rules	77.0	70.9
LDA	77.7	**75.9**
QDA	77.0	72.7

leaves as the class probability. For C4.5rules, we used the confidence of the first rule of the ordered C4.5rules rules set (ordered by class and then by confidence) that matches the instance as its class probability. In [18], it was shown that this is a feasible strategy for computing the AUROC of C4.5rules. Table 5 clearly indicates that the LDA classifier gave the best AUROC performance. However, there is no significant difference with the AUROC performance of the GBN and Naive Bayes classifier according to the test of De Long et al. and again using a significance level of 5%. Observe from Tables 4 and 5 that both multinet classifiers, QDA and C4.5rules never achieved good performance in terms of PCC and AUROC. Note that for all Bayesian network classifiers, we also investigated the impact of smoothing the parameter estimates. However, no significant performance increase in terms of either the PCC or AUROC values were found with parameter smoothing.

Besides looking at the classification performance, we also investigated the complexity of the generated classification models because from a marketing viewpoint, easy to understand, parsimonious models are to be preferred. Table 6 presents the complexity of the generated Bayesian network and C4.5(rules) classifiers. We did not include LDA and QDA because they are basically mathematical models which give a rather limited insight into the relationships and patterns present in the domain under study. The Naive Bayes and TAN network classifiers did not prune any attributes because all attributes remained in the Markov Blanket of the classification node. The TAN added 14 arcs to the Naive Bayes classifier which resulted in a performance increase in terms of PCC (from 72.5 to 74.0) but a performance decrease in terms of AUROC (from 74.3 to 73.6). Hence, the effect of the added complexity was rather marginal in our case. Although the GBN multinet classifier seems attractive because of its simple structure, its performance according to Tables 4 and 5 was rather bad. Also the CL multinet classifier gave bad performance and has on top a complex structure. The tree induced by C4.5 is not easy to handle and interpret because of its large number of internal and leave nodes. Moreover, the C4.5 tree was able to prune only 2 of the 15 attributes. The rule set inferred by C4.5rules contains 18 rules. This might seem interesting but when considering Tables 4 and 5 the performance of C4.5rules in terms of both PCC and AUROC was rather bad. Note that while the C4.5 tree pruned two attributes, the

Table 6
Complexity of the Bayesian network classifiers and C4.5

Naive Bayes	16 nodes and 15 arcs
TAN	16 nodes and 29 arcs
CL multinet	Net 1: 15 nodes and 14 arcs Net 2: 15 nodes and 14 arcs
GBN	4 nodes and 6 arcs
GBN multinet	Net 1: 3 nodes and 2 arcs Net 2: 3 nodes and 2 arcs
C4.5	13 internal nodes 32 leave nodes
C4.5rules	18 rules

C4.5rules rules set still contained all attributes. This can be explained by the fact that C4.5rules starts generating and pruning the rules from the unpruned C4.5 tree. The GBN classifier was able to prune 12 attributes, leaving only three attributes in the model. Only six arcs where necessary to efficiently model the dependencies between the attributes and the classification node. Furthermore, it gave also a very good performance in terms of PCC and AUROC on the test set. The structure of the GBN classifier is depicted in Fig. 4.

This figure clearly illustrates that it is a compact, parsimonious and yet powerful model for decision making. By using only three variables compiled from purchase records of the first six months of the customer lifecycle, we have provided evidence that, in our DIY case, the SlopeSign of a lifecycle of 48 months can be predicted with a classification accuracy of 75%. The total contribution of the client (TotCont), the total number of articles bought (NumbArt) and the maximum percentage of products bought in one product family (MaxPerc) proved to be very powerful predictors for the sign of the customer lifecycle slope when using GBN classifiers. While the first two variables present a measure of the volume of the purchases made (the purchase "depth"), the latter variable is an estimator of the variety of product families bought (the purchase "broadness").

The knowledge that these variables are intensely related to the slope's evolution can be useful for marketing decision makers. In this Belgian DIY retail setting, the initial monetary amount spent at the company (TotCont) and the initial number of articles purchased (NumbArt) were found to be negatively related to the SlopeSign, whereas the maximum percentage of products purchased in one category (MaxPerc) was found to be positively related to the SlopeSign. This implies that customers that tend to increase their spending over their lifetime with the company initially spend less money on a lower number of articles, purchasing from a smaller set of product categories. Alternatively, customers spending a lot of money initially on a lot of articles and who purchase products across a lot of different categories tend to decrease their spending in the future. This information may prove valuable for the company in this case as a starting point for investigating why high-spending customers generally decrease their spending over time.

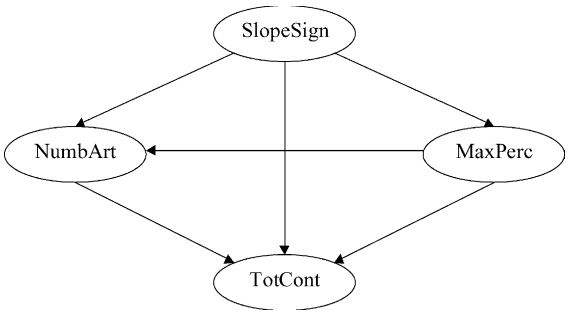

Fig. 4. Unrestricted Bayesian network constructed for marketing case.

To conclude, we can state that Bayesian network classifiers are performing well in predicting the future customer evolution and are able to contribute to an increased understanding of the relationship between the investigated variable and the most relevant explanatory variables. Hence, we have reached our goal to illustrate that Bayesian network classifiers can be considered to be a useful tool in the toolbox of marketing analysts in this application of identifying the slope of the customer lifecycle of long-life customers.

6. Conclusions

In the theoretical part of this paper, we have argued that long-life/loyal customers have been regularly regarded as a homogeneous group of the most profitable customers of a company. Building on more recent findings, in this study, we have tried to acknowledge the heterogeneity in the group of long-life customers by dividing the group into two subparts, essentially consisting of customers increasing versus decreasing their spending over their relationship with the company in case. Hence, it was the goal of this study to predict the sign of the slope—being the output of the estimation of a linear customer lifecycle—at the individual customer level using Bayesian network classifiers based on information from initial purchase occasions.

Bayesian network classifiers have been recently proposed in the artificial intelligence literature as probabilistic white-box models which allow to give a clear insight into the relationships between the variables of the domain under study. Starting from the Naive Bayes classifier, we gradually removed the restrictions put on the network structure and investigated TANs and GBN classifiers. The latter were learnt using the algorithm of Cheng et al. We compared the classification accuracy and the AUROC of all Bayesian network classifiers with discriminant analysis and the widely used C4.5 and C4.5rules algorithms. It was shown that general, unrestricted Bayesian network classifiers have a good performance in terms of both measures. Furthermore, using the Markov Blanket concept allowed us to prune a lot of attributes resulting in a compact, parsimonious, yet powerful Bayesian network classifier for marketing decision making.

In summary, we contribute to the literature by providing experimental evidence that: (1) Bayesian network classifiers offer an interesting and viable alternative for our customer lifecycle slope estimation problem; (2) the Markov Blanket concept allows for a natural form of attribute selection that was very effective for the case at hand; (3) the sign of the slope can be predicted with a powerful and parsimonious GBN classifier; (4) a set of three variables measuring the volume of initial purchases and the degree to which customers originally buy in different categories, are a powerful set of predictors for estimating the sign of the slope.

7. Practical implications and issues for further research

While it has been the focus of this paper to demonstrate (i) the predictability of the sign of the slope and (ii) the performance of several Bayesian network classifiers versus statistical and other artificial intelligence techniques, here, we elaborate on possible applications of the knowledge of the sign of the slope for relationship marketing decision making. A number of future applications lie ahead. Firstly, the sign of the slope might prove to be a useful indicator in the decision upon the type or strength of the marketing investment that can be used vis-à-vis a certain consumer. For example, a company organizing a membership club, with special service offerings, special promotions, etc. might only want to deliver these benefits to consumers that are worthy of such a large investment. Thus, knowing that certain consumers will decrease their spending might be important for improving the return on the relationship marketing investment. Alternatively, a company might have two marketing incentives of unequal cost (e.g. a special promotion versus a small gift). Also in this case, it could be useful to assess the future spending of a customer in order to allocate the desired incentive to each customer. Secondly, the estimations may be used in an aggregated way, as a monitor of e.g. customer-acquisition policies. In this way, the percentage of customers that are expected to raise their spending in the future can be compared for different acquisition strategies and campaigns in order to select those target markets with higher potential for establishing enduring relationships. An additional benefit is derived from the fact that it was possible to predict the evolutions very early in the relationships, so acquisition campaigns can be evaluated in a time-effective way. Thirdly, the estimations might be used as a dimension for designing an a priori segmentation scheme for a company's customer base. Hence, it might be feasible to delineate a more customized customer strategy per segment. Two possible applications are summarized in Fig. 5. In the first segmentation scheme (a), the sign of the slope is used together with the tenure of customers in order to decide upon the relevant marketing message content and size. Whereas short-life customers only merit investments that can be regained during their limited relationship with the company, long-life customers might effectively be reached through more expensive marketing programs. For customers who are expected to increase the relationship with the company (who are likely to be more satisfied with the company in case) it might be beneficial to offer additional products according to their detected needs (detected e.g. through a cross-selling analysis) or extra value (e.g. through the membership to a club). Alternatively, customers who are expected to decrease their spending might be appropriately

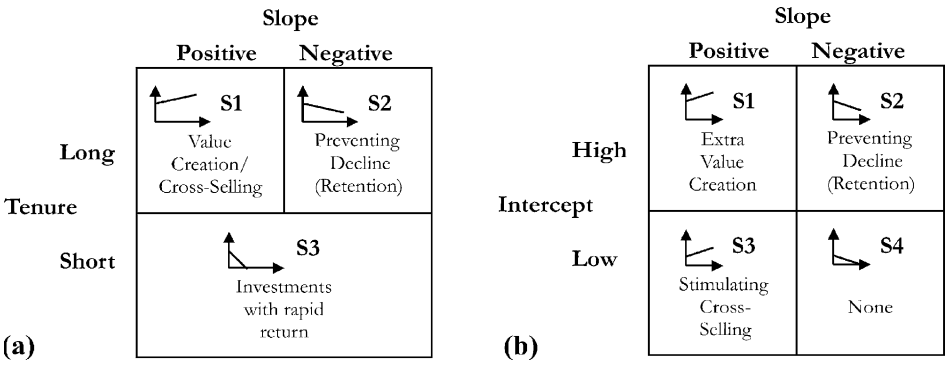

Fig. 5. Summary of possible a priori segmentation schemes.

managed with a retention program (e.g. focused on complaint detection and complaint handling). In the second segmentation scheme (b), the intercept and the slope of the customer lifecycle are used to delineate the segmentation. Also in this case, argumentation could be found to use specific marketing strategies to target the segments, where it could be argued that not all segments merit targeting (e.g. the segment of customers that starts as low-spending customers and are expected to decrease their spending even further).

The desired outcome of this line of research could consist in suggesting an optimal CRM strategy to different segments. However, in order to test the optimality of the proposed strategies, one would have to design and implement an experiment allocating strategies randomly to customers. It can be argued that several important practical problems would arise when attempting to implement such a study.

Firstly, a company that has not been performing a broad range of different CRM strategies would have to make large marketing investments in designing an appropriate tactic for each strategic goal (e.g. customer retention through satisfaction research, complaint handling, or other tactics). Secondly, and crucially, for optimally allocating a customer to a strategy, it would be necessary to assign a sizeable part of the customer base randomly to each of the strategies, implying that by definition, customers will be targeted with strategies that are inappropriate for them, implying large marketing expenses with low return on investment, confused and unsatisfied customer responses, especially within the group of high-spending customers that has been proven to expect preferential (or at least reasonable) treatment compared to other customers [45]. While this experimental setting would likely provide rich information to researchers, the costs involved are, especially while marketing management is aware of the long-term potential of customers, of a magnitude that is not acceptable to managers. Thirdly, even if a company would be interested in researching such an optimal segmentation scheme, the generalization capacity would probably be low, considering the specificity of the tactics used. Hence, the scientific outcome of the study might only be reached when validated with several tactics for each strategy, driving the required investments even further. Finally, in order to assess the effect of the approach, the results of the study can only be expected after several years, in order to measure the changes in the slope of the customer lifecycle. The four factors mentioned above all add to the difficulties of funding, designing and implementing an optimal experimental study.

Further research is needed in two major directions. In the domain of marketing, the creation of variables having still better predictive capabilities for predicting the sign of the slope of the linear lifecycle is an interesting research topic. Alternatively, a replication of this study over different customer bases in diverse industries and countries might deliver an insight into the stability of the findings. Eventually, if resources would be

available, testing and comparing different strategies (e.g. the frameworks presented in Fig. 5) 'in-the-field' can determine the full potential of the usage of customer spending evolutions for marketing decision making. Considering the Bayesian network classifiers, additional research is needed to investigate the power of other structure learning algorithms. Also the presence of hidden variables in the Bayesian network forms an interesting topic for further research.

References

[1] S. Acid, L.M. Campos, An algorithm for finding minimum *d*-separating sets in belief networks, in: Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence (UAI), Portland, Oregon, USA, 1996, pp. 3–10.

[2] B. Baesens, S. Viaene, D. Van den Poel, J. Vanthienen, G. Dedene, Using Bayesian neural networks for repeat purchase modelling in direct marketing, European Journal of Operational Research 138 (1) (2002) 191–211.

[3] R.C. Blattberg, J. Deighton, Manage marketing by the customer equity test, Harvard Business Review (July–Aug) (1996) 136–144.

[4] R.J. Brodie, N.E. Coviello, R.W. Brookes, V. Little, Towards a paradigm shift in marketing? an examination of current marketing practices, Journal of Marketing Management 13 (1997) 383–406.

[5] W. Buntine, A guide to the literature on learning probabilistic networks from data, IEEE Transactions on Knowledge and Data Engineering 8 (1996) 195–210.

[6] J. Cheng, Powerpredictor system, 2000. Available from <http://www.cs.ualberta.ca/jcheng/bnpp.htm>.

[7] J. Cheng, D.A. Bell, W. Liu, An algorithm for Bayesian belief network construction from data, in: Proceedings of the Sixth International Workshop on Artificial Intelligence and Statistics (AI and STAT), Fort Lauderdale, Florida, USA, 1997, pp. 83–90.

[8] J. Cheng, D.A. Bell, W. Liu, Learning belief networks from data: An information theory based approach, in: Proceedings of the Sixth ACM Conference on Information and Knowledge Management (CIKM), Las Vegas, Nevada, USA, 1997, pp. 325–331.

[9] J. Cheng, R. Greiner, Comparing Bayesian network classifiers, in: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI), Stockholm, Sweden, 1999, pp. 101–108.

[10] J. Cheng, R. Greiner, Learning Bayesian belief network classifiers: Algorithms and system, in: Proceedings of the Fourteenth Canadian Conference on Artificial Intelligence (AI), 2001.

[11] C.K. Chow, C.N. Liu, Approximating discrete probability distributions with dependence trees, IEEE Transactions on Information Theory 14 (3) (1968) 462–467.

[12] G.J. Cullinan, Picking them by their batting averages recency-frequency-monetary method of controlling circulation, Manual release 2103, Direct Mail/Marketing Association, NY, 1977.

[13] E.R. De Long, D.M. De Long, D.L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach, Biometrics 44 (1988) 837–845.

[14] T.G. Dietterich, Approximate statistical tests for comparing supervised classification learning algorithms, Neural Computation 10 (7) (1998) 1895–1924.

[15] R.O. Duda, P.E. Hart, Pattern Classification and Scene Analysis, John Wiley, New York, 1973.

[16] J.P. Egan, Signal Detection Theory and ROC Analysis. Series in Cognition and Perception, Academic Press, New York, 1975.

[17] B.S. Everitt, The Analysis of Contingency Tables, Chapman and Hall, London, 1977.

[18] T. Fawcett, Using rule sets to maximize roc performance, in: Proceedings of the IEEE International Conference on Data Mining, San Jose, California, USA, 2001.

[19] U.M. Fayyad, K.B. Irani, Multi-interval discretization of continuous-valued attributes for classification learning, in: Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence (IJCAI), San Francisco, CA, USA, Morgan Kaufmann, 1993, pp. 1022–1029.

[20] C. Fornell, B. Wernerfelt, Defensive marketing strategy by customer complaint management: A theoretical analysis, Journal of Marketing Research 24 (1987) 337–346.

[21] N. Friedman, D. Geiger, M. Goldszmidt, Bayesian network classifiers, Machine Learning 29 (1997) 131–163.

[22] J. Ganesh, M.J. Arnold, K.E. Reynolds, Understanding the customer base of service providers: An examination of the differences between switchers and stayers, Journal of Marketing 64 (2000) 65–87.

[23] D. Geiger, D. Heckerman, Knowledge representation and inference in similarity networks and Bayesian multinets, Artificial Intelligence 82 (1996) 45–74.

[24] D. Geiger, T.S. Verma, J. Pearl, Identifying independence in Bayesian networks, Networks 20 (5) (1990) 507–534.

[25] C. Grönroos, From marketing mix to relationship marketing—towards a paradigm shift in marketing, Management Decision 35 (4) (1997) 322–339.

[26] J.A. Hanley, B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (roc) curve, Radiology 148 (1983) 839–843.

[27] J.A. Hanley, B.J. McNeil, A method of comparing the areas under receiver operating characteristic curves derived from the same cases, Radiology 148 (1983) 839–843.

[28] D. Heckerman, Probabilistic Similarity Networks, MIT Press, Cambridge, MA, 1991.

[29] D. Heckerman, A tutorial on learning with Bayesian networks, Technical Report MSR-TR-95-06, Microsoft Research, 1995.

[30] G.H. John, P. Langley, Estimating continuous distributions in Bayesian classifiers, in: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI),

Montreal, Quebec, Canada, Morgan Kaufmann, San Francisco, CA, 1995, pp. 338–345.
[31] S. Knox, Loyalty-based segmentation and the customer development process, European Management Journal 16 (6) (1998) 729–737.
[32] J.B. Kruskal Jr., On the shortest spanning subtree of a graph and the travelling salesman problem, in: Proceedings of the American Mathematics Society, vol. 7, 1956, pp. 48–50.
[33] P. Langley, W. Iba, K. Thompson, An analysis of Bayesian classifiers, in: Proceedings of the Tenth National Conference on Artificial Intelligence (AAAI), San Jose, CA, USA, AAAI Press, 1992, pp. 223–228.
[34] S.L. Lauritzen, Graphical Models, Clarendon Press, Oxford, 1996.
[35] K. Murphy, Bayes net matlab toolbox, 2001. Available from <http://www.cs.berkeley.edu/~murphyk/Bayes/bnt.html>.
[36] J. Pearl, Probabilistic reasoning in Intelligent Systems: Networks for Plausible Inference, Morgan Kaufmann, San Francisco, CA, 1988.
[37] D. Peppers, M. Rogers, Enterprise One to One: Tools for Competing in the Interactive Age, Doubleday, New York, USA, 1997.
[38] F. Provost, T. Fawcett, R. Kohavi, The case against accuracy estimation for comparing classifiers, in: J. Shavlik (Ed.), Proceedings of the Fifteenth International Conference on Machine Learning (ICML), San Francisco, CA, USA, Morgan Kaufmann, San Francisco, CA, 1998, pp. 445–453.
[39] J.R. Quinlan, C4.5 Programs for Machine Learning, Morgan Kaufmann, San Francisco, CA, 1993.
[40] F.F. Reichheld, The Loyalty Effect, Harvard Business School Press, Cambridge, MA, 1996.
[41] F.F. Reichheld, Lead for loyalty, Harvard Business Review (July) (2001) 76–84.
[42] F.F. Reichheld, D.W. Kenny, The hidden advantages of customer retention, Journal of Retail Banking 4 (1990) 19–23.
[43] F.F. Reichheld, W.E. Sasser, Zero defections: Quality comes to services, Harvard Business Review (Sept–Okt) (1990) 105–111.
[44] W.J. Reinartz, V. Kumar, On the profitability of long-life customers in a noncontractual setting: An empirical investigation and implications for marketing, Journal of Marketing 64 (2000) 17–35.
[45] W.J. Reinartz, V. Kumar, The mismanagement of customer loyalty, Harvard Business Review (July) (2002) 4–12.
[46] L.J. Rosenberg, J.A. Czepiel, A marketing approach to customer retention, Journal of Consumer Marketing 1 (1984) 45–51.
[47] R.T. Rust, V.A. Zeithaml, K.N. Lemon, Driving Customer Equity: How Customer Lifetime Value is Reshaping Corporate Strategy, Free Press, New York, 2000.
[48] J.A. Swets, R.M. Pickett, Evaluation of Diagnostic Systems: Methods from Signal Detection Theory, Academic Press, New York, 1982.
[49] R.P. Thrasher, Cart: A recent advance in tree-structured list segmentation methodology, Journal of Direct Marketing 5 (1) (1991) 35–47.
[50] D. Van den Poel, Response Modeling for Database Marketing using Binary Classification, Ph.D. Thesis, K.U. Leuven, 1999.
[51] L.C. Van Der Gaag, Bayesian belief networks: Odds and ends, The Computer Journal 39 (2) (1996) 97–113.
[52] T. Verma, J. Pearl, Causal networks: Semantics and expressiveness, in: Proceedings of the Fourth Workshop on Uncertainty in Artificial Intelligence, Mountain View, CA, USA, 1988, pp. 352–359.
[53] W.G. Waller, A.K. Jain, On the monotonicity of the performance of a Bayesian classifier, IEEE Transactions on Information Theory 24 (3) (1978) 392–394.