A methodology for Internet Customer segmentation using Decision Trees

Y. B. Cho and S. H. Kim

Graduate School of Management, Korea Advanced Institute of Science and Technology
207-43 Cheongryangri-Dong, Dongdaemun, Seoul, 130-012, South Korea
Tel : 82-2-958-3684 Fax : 82-2-958-3604
e-mail : ybcho@kgsm.kaist.ac.kr, seekim@kgsm.kaist.ac.kr

Abstract

Application of existing decision tree algorithms for Internet retail customer classification is apt to construct a bushy tree due to imprecise source data. Even excessive analysis may not guarantee the effectiveness of the business although the results are derived from fully detailed segments. Thus, it is necessary to determine the appropriate number of segments with a certain level of abstraction. In this study, we developed a stopping rule that considers the total amount of information gained while generating a rule tree. In addition to forwarding from root to intermediate nodes with a certain level of abstraction, the decision tree is investigated by the backtracking pruning method with misclassification loss information.

Key words: Electronic market; Customer classification; Decision tree

1. Introduction

The rapid proliferation of the Internet has created a fast-growing electronic channel for marketing, and the number of on-line stores has increased at an unprecedented rate [1]. The immediate communication afforded by the Internet enables a company to respond quickly to market and customer changes. The Internet allows a company to provide convenient access to a broader customer base that traditional "bricks-and-mortar" outlets, as Internet access is not limited by any physical location and is available 24 hours a day. In addition, the Internet is highly suitable for gathering information on consumer usage preferences. The click-streams that contain a sequence of URLs visited, and for how long and at what time customers viewed product information enable a company to analyze customers' behavior on an electronic basis. Moreover, the ever-increasing computing capabilities for databases and data processing, coupled with the latest data mining and warehousing techniques, offer an opportunity to obtain a great deal of knowledge about the on-line shopping process [2].

Identification of a group of customers with an increased likelihood of the target behavior will allow a company to conduct marketing campaigns adapted for those customers' behavior. If targeting criteria for an e-mail campaign are chosen well, the company can mail a much smaller group of people while obtaining an almost equivalent number of responses. The increased concentration of the desired targets in such targeted campaigns can be expected to produce more satisfactory results. Thus, customer classification is an important avenue of research in the field of electronic commerce [3]. However, there are two opposing views concerning Internet customer classification. First, advocates of personalization argue that it will become possible to eventually reach and target a segment of one on the Internet [4]. This should be justified by the expected benefits of reaching individual consumers to satisfy their unique needs and wants in the best way possible [5]. However, several shortcomings of this approach have been highlighted: for example the economic inviability of the creation of massive amounts of customized content, and the costs of maintaining the service for personalized interaction over time [6]. Further, source data considered for analysis may contain some degree of uncertainty due to noise in the measurements or to the presence of factors that are difficult to measure but are necessary to include, as is often the case in Internet customer classification. Overloaded and inefficiently captured results, possibly based on imprecise source data, may not guarantee the effectiveness of the business although they are derived from fully detailed segments. Thus, between the two extreme viewpoints, it is necessary to find an appropriate number of segments with a certain level of abstraction without trying only to find fully discovered segments.

The decision tree method with an appropriate pruning algorithm is one of the simplest, yet most powerful methods for Internet customer classification. The pruning decision tree with misclassification loss has business implications in that it actively reflects operational knowledge for maintaining the Internet market, although the information is somewhat dependent on the subjective judgments of operational managers and the situations encountered.

This paper is organized as follows: issues regarding segmentation in the Internet marketplace and use of a decision tree as a classification tool are presented in Section 2. Section 3 describes the proposed classification approach, which is followed by a discussion of the application of our methodology to a real-world Internet customer classification problem in Section 4. The final section draws some conclusions from the results and summarizes the findings of the study.

2. Segmentation in the Internet market- place and decision trees

2.1 Segmentation in the Internet marketplace

Companies communicate with their customers through

various media. Traditionally, these media follow a passive one-to-many communication model, whereby a company reaches many current and potential customers, in either a segmented or a non-segmented manner. Developments in Internet communications over the past several years have dramatically altered this traditional view of advertising and communication media. The Internet is a new marketing medium that has the potential to radically change the way companies do business with their customers [7]. The Internet presents an opportunity to reach new groups of consumers and provides an alternative way to offer low-price and value-added products or services. The Internet can aid in value-added marketing by using technology to gain market share with low cost and timely performance [8]. Marketing on the Internet is achieved through electronic mail and the World Wide Web. The WWW provides two of the most important aspects of modern marketing philosophies: the ability to target selected groups of buyers and to open interactive dialog with customers[9]. The Web is not a one-way communication channel between a vendor and a buyer: unlike traditional marketing channels, the Web provides a means of communication that facilitates a more intimate relationship between buyer and seller. In addition, the Web supports hypermedia technology through which a marketer can trace the customer's footsteps on his web site. This tracing of customers' behavioral characteristics on the Web is a valuable source of information to facilitate segmentation. For example, we can infer the customer's preferences through information about which web sites the customer visits, how long the customer stays in certain web sites, and finally on what web sites the customer purchases products.

Studies of segmentation of Internet customers have used various methods, including fuzzy clustering, neural networks, and conjoint analysis. Ozer [10] presented an application of fuzzy clustering in the area of on-line music services. Fuzzy clustering was used to categorize potential customers of music services and to explain the characteristics of the categorized segments. The analyses indicated differences among the segments with respect to their attitude, interests, and opinions about music services. Vellido *et al.* [6] carried out exploratory segmentation of customers in an on-line shopping mall. They extracted various aspects of shopping structure and characterized the on-line shopping adoption process. A self-organizing map (SOM) was used to perform segmentation in combination with factor analysis of the observable variables. This method has been applied to discover clusters or segments and to visualize the segments in which the data are naturally organized. However, these studies used only survey data concerning WWW users. Lee *et al.* [11] proposed that electronic market segmentation requires new data preparation procedures, for example, click stream, different to those used in traditional market segmentation. The click stream is automatically gathered from the web server and reveals the consumer's preference toward the products and/or services represented by the Internet site. Lee *et al.* [11] also suggested an adaptation of the conjoint model for Internet customer segmentation. There have been few previous studies on Internet customer segmentation using click stream data. It is not easy to use click stream data for such purposes because the volume of data is very large and must be preprocessed using text-processing techniques to translate qualitative information into a quantitative form that is useful from a marketing standpoint [2].

2.2 Decision trees

The decision tree is a widely used symbolic modeling technique for classification tasks in machine learning. Since ID3 [12], a number of algorithms for decision tree induction have been proposed. ID3 uses an information theoretic measure, and continues the partitioning process until all the examples in a subclass are partitioned. Other algorithms using various measures for tree-structure classification have been proposed, such as chi-square statistics [13],[14] and the pseudo chi-square statistic G^2 [15],[16].

Pruning algorithms, removing branches with little statistical validity, have been introduced in many studies. Pre- and post-pruning are two standard methods for pruning a decision tree. The pre-pruning method stops further splitting of a given node based on stopping criteria, rather than building a decision tree until it is complete with given training data. Examples of these stopping criteria include information gain [14] and chi-square test [17]. However, these pre-pruning approaches have critical drawbacks in that stopping criteria are based on local information and intermediate estimates may be far from optimal. It is difficult to say what would happen a few levels below the node being investigated for pruning, and thus pre-pruning is locally optimal and globally inefficient [18]. On the other hand, post-pruning is used to build an overall tree, which it then prunes backwards. Several post-pruning methods have been reported, including cost-complexity pruning [14], the minimum error method [19], pessimistic error pruning [17], and error-based pruning [20]. Post-pruning is based on estimating the sensitivity of sub-trees to some measure (*e.g.*, error rate or misclassification cost) and removing those sub-trees that have minimal impact on the measure. When the predicted error of the leaf replacing the sub-tree is lower than that of the sub-tree, a single leaf node replaces the sub-tree. Of course, the post-pruning method is not without its drawbacks. There is considerable computational difficulty in evaluating combinations of removals, and most post-pruning algorithms create trees that are larger than necessary if error minimization is an evaluation criterion [21]. In many practical problems, the costs of misclassification errors are not equal, and Pazzani *et al.* [22] and Bradford *et al.* [23] suggested a pruning algorithm that involves non-uniform cost rather than uniform cost of misclassification,which is discussed in detail in Section 3.

3. Decision tree approach for segmenting Internet customers

3.1 Constructing decision trees

Construction of a decision tree involves a series of processes of selecting attributes that classify objects with class values.

Of prime importance is the measure used to evaluate the classification capabilities of particular attributes, and rough set theory and Quinlan's C4.5 [20] are used for this purpose. In rough set theory, a rough set can be replaced by a pair of crisp sets called the lower and the upper approximations, and the ratio of their cardinality is the critical measure of further inclusion of attributes for classifying objects [24]. On the other hand, Quinlan's C4.5 is based on Shannon's [25] information theory, and thus the entropy measure is used to explain the uncertainty of the decision system. The remainder of this paper is in line with the concept of the latter approach. In C4.5, an attribute that contributes to maximum information gain is selected as the root node and then subsequent attributes are considered with the criterion of information gain under a partially constructed tree. This process is applied to each sub-tree recursively until it has completed the tree. Information gain is the amount of information that can be obtained when an attribute is selected as a splitting attribute. The amount of information gained by addition of an attribute is equal to the amount of removed uncertainty. Hence, the "attribute that contributes to the greatest information gain" can be expressed in other ways, such as the "attribute yielding the least uncertainty," and "information gain can attain the maximum value when the attribute is associated." In this study, we used information gain as a selection criterion for attributes for further development of our methodology.

3.2 Total amount of Information Gained

In the customer classification problem, gathered information can be considered the amount of reduction in market uncertainty, as the overall market is divided into several sub-markets. Recursive selection of splitting attributes with maximum information gain implies removal of market uncertainty through segmentation. However, segmentation using existing decision tree algorithms, such as C4.5, seeks a local optimum but ignores information gained in constructing decision trees from the root node to intermediate nodes. Thus, to apply the decision tree algorithms to segmentation problems, a new approach that can reduce local optimality and use all of the information gained, and moreover that can control the number of segments according to the marketer's preferences, is required.

For this purpose, it is necessary to first define some terminology. Let IG^t be information gain at t^{th} splitting stage, where splitting has a breadth-first sequence that descends a level after selecting the splitting attribute in the same level. Note that IG^t equals the maximum information gain in the t^{th} stage. We define TG^t as the *total amount of information gained* until t^{th} splitting, $TG^t = TG^{t-1} + (n^t / N) IG^t$, where cardinality of C^t, which is the set of objects at t^{th} splitting, is n^t and N is the total number of objects. TG^t is a weighted summation of information gained until the t^{th} splitting stage. The difference between TG^t and TG^{t-1} represents the reduction in uncertainty due to splitting of the subclass C^t. Also, TG^t increases monotonically as N, n^t and IG^t are non-negative values. The minimum value of TG^t is 0, because it occurs in cases where classification is undone: *i.e.*, the initial class C is not split at all. The maximum value of TG^t is $H(C)$, where $H(C)$ is the entropy of the initial set of objects C. The upper limit of TG^t occurs when all the terminal subclasses have only one kind of object, because if we assume that an initial attribute totally determines the initial class, then $TG^t = IG^1 = H(C)$.

According to the above properties of TG, we define *confident percentage* $TG^t / H(C)$, which also increases monotonically and ranges between 0 and 1. If a certain level of this measure is determined subjectively, *e.g.*, 65%, 85%, or 95%, the process of selecting attributes will end when $TG^t / H(C)$ first exceeds a given value. This implies that the marketer willingly accepts a level of uncertainty of 35%, 15%, or 5%, respectively.

This measure can be used to determine the appropriate stopping level at which further splitting does not proceed. This method may be useful when a marketer has sufficient knowledge about the market such as return rate of direct mail, retention rate of customers, monthly arriving rate of new customers, etc.

Let us consider other properties of TG^t to determine a heuristic stopping level of confident percentage. The graph of TG^t has various patterns according to classifier performance. With superior (inferior) performance of a classifier, the graph of TG^t skews more to the northwest (southeast) if using the same data set. For example, the TG_1 graph shown at the top of Figure 1 has better classification performance than TG_2 shown at the bottom, as the classifier of the former has more compact decision trees with smaller nodes than the latter.

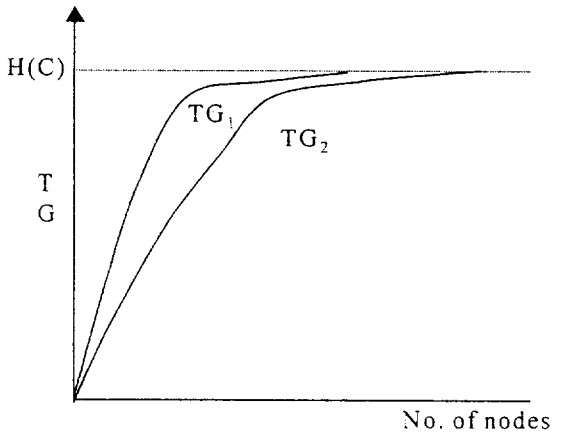

Figure 1. Two examples of TG graphs

As TG^t indicates total information gained from the root node to the t^{th} node and IG^t is the maximum information gain at the t^{th} splitting, the remaining information decreases rapidly in advance of splitting. If the amount of additive information that can be obtained through further splitting is quite small, it would rather not go on further split.

After depicting the TG^t graph of a classifier, we can determine a heuristic stopping level at which the slope of the TG^t graph rises steeply and then becomes approximately horizontal. The number of segments becomes one plus the number of nodes corresponding to this stopping level, assuming binary splitting of the data set.

3.3 Misclassification loss

The amount of loss incurred due to misclassification of Internet customers is high because Internet customers are much more proactive than off-line customers. Customers are empowered by the information richness of the medium and it is possible to perform comparisons between products, services, and their prices with only a few keystrokes. Thus, Internet customers show their immediate responses and further leave to competitors' sites when unsympathetic to the on-line vendors' marketing campaigns. On-line vendors have to make extra efforts to differentiate their offers and attract on-line browsers towards their on-line stores.

For simplicity, assuming two classification classes, *i.e.*, "positive" (prone to buy) and "negative" (reluctant to buy), there are two types of misclassification errors: false positive and false negative. False positive corresponds to the case in which a customer is predicted to be positive, but is actually negative, and false negative corresponds to the opposite case. Note that losses incurred from misclassification errors may have different measurement units depending on the situation considered. In some problems, loss represents dollars expended or opportunity cost in real terms. In this paper, loss is defined as opportunity cost due to inappropriate marketing. Uniform loss implies that classification errors have symmetric unit cost, and thus false positives and false negatives have the same value, denoted as 1 in the upper right and lower left of Table 1, respectively. In real-world problems, this rarely happens, because classifications lead to actions that might have different consequences. Indeed, it is difficult to find a domain in which a learning system may be indifferent to whether a false positive or false negative error is made [23] and thus a loss matrix associated with classification errors consists of asymmetric elements [26].

A more general loss matrix is shown in Table 2, where three types of misclassification losses are involved. In the first column, when a non-purchasing customer is wrongly predicted to be a purchaser, let misclassification loss be 1 as a basis for comparative reference. Similarly, if a non-purchaser is wrongly predicted to be a purchaser that is prone to buy product A (or B), loss in this case will be greater than 1 as the company offered many kinds of marketing events on the Internet (*e.g.*, commercial e-mail and pop-up windows), which are of little or no interest to the potential customer. If a purchaser who would tend to buy product A (or B) in reality is taken to be a purchaser buying product B (or A), the loss will be less than 1 because customers who have bought product A (or B) have some preference for that on-line store and may buy another product (*e.g.*, cross-selling or up-selling). It is relatively likely that these customers will respond to a given promotion.

Then, we must compute loss incurred by misclassification. This can be done by two methods: *i.e.*, direct and indirect. The direct method estimates loss from the actual cost of a marketing campaign. In e-mail marketing, for example, the direct method will be as follows. After calculating the total cost of sending e-mails (*i.e.*, labor, depreciation of the machine(s) used to send the e-mail, *etc.*), total cost may be divided into each type of mis-classified object.

Table 1. Uniform loss matrix

		Predicted class	
		N	P
Actual class	N	0	1
	P	1	0

N: Non-purchaser
P: Purchaser

Table 2. Loss matrix suitable for customer classification in an on-line retail store

		Predicted class		
		N	A	B
Actual class	N	0	> 1	> 1
	A	1	0	<1
	B	1	< 1	0

N: Non-purchaser
A: Purchaser of product type "A"
B: Purchaser of product type "B"

However, it is difficult to measure loss precisely using the direct method, as loss due to spam mail or mis-recommendation mail is much greater than direct marketing costs. Thus, the indirect method in which mis-classification loss is estimated based on the experience of domain experts is more desirable. Domain experts, *i.e.*, web marketers or web masters, have some knowledge about misclassification loss and they can determine the relative importance of each misclassification case using pair-wise comparisons with an ordinal scale.

3.4 Combining TG' with misclassification loss

Although TG' is useful in that it provides a heuristic measure to determine the number of segments even in cases where a marketer has sufficient knowledge about the market market, it may sometimes yield unexpected results that are difficult to apply to customer classification. That is, overfitting with regard to unseen objects, as an attribute with low classifiable performance can be used as the splitting attribute if it provides maximum information gain in the corresponding stage and TG' is not above the stopping threshold value. Thus, it is necessary to use additional procedures to prune attributes with low classifiable performance. Misclassification loss pruning can be a good alternative to remedy overfitting. Additional pruning with misclassification loss can also minimize predicted loss and improve predicted

accuracy of derived classifying rules.
Under loss information, which depends on the situation considered, the loss pruning process is similar to that of post pruning except that specific loss values are applied to wrongly classified objects. Thus, after subjectively determining an appropriate stopping level of the confident percentage of uncertainty, the decision criteria for loss pruning proceed; if the weighted sum of predicted losses at the leaf node is beyond the weighted predicted losses at its parent node, the leaf node is pruned, and the process proceeds until the contrary condition is met.

4. Application of the proposed approach

4.1 Overview of the company

We used real-world data to examine the performance of the proposed approach. To maintain confidentiality with regard to the databases used, we use the expressions "the company" and "this site" rather than their actual names.
The company opened this Korean web site in November 1997, directed toward female Internet users by providing information on beauty and supplying a hub for communication among them. The company manages an Internet retail shop to sell products for women, such as cosmetics and clothes. This site was ranked the 4th most frequently visited site among female-oriented web sites in a study carried out in March 2001 by Simf (http://simf.co.kr), a firm that evaluates Korean web sites. Around 3,000 Internet users visit the site daily, and on average 100 users order more than one product from this site per month.

4.2 Preparing the data set

Choice of appropriate data is crucial for customer classification because analysis of insufficient data may result in wrongly classified customers. In this study, we used Web log data obtained during the three-month period from February to April 2001, which signify customer's behavioral characteristics, such as number of clicks and session time. In contrast to previous studies, we did not take into account demographic attributes for market segmentation because most Internet retail stores, including the company discussed here, allow guests to log in and purchase products for extended sales performance and are thus easily accessed by users who do not register their demographic attributes. Further, non-purchasers who navigate through Internet retail stores sometimes leave no footprints except Web log data gathered automatically.
Interviews with operational managers indicated that they require information on four distinctive attributes, consisting of two continuous and two nominal attributes: the number of clicked pages (CP), session time in seconds (ST), BBS subscription (BBS), and marketing event participation (Event). The sample data ranges of attributes are shown in Table 3. The data showed that, on the average, purchasers show more activities than non-purchasers but the same was not always true for BBS participation. This is probably because customers visit and subscribe to the site to obtain a better understanding and to communicate with others concerning the products considered.

Table 3. Data ranges of attributes

		Purchaser	Non-purchaser
Session Time (ST:sec.)	Min.	6	163
	Max.	45,374	102,384
	Avg.	11,114	26,454
Clicked Pages(CP)	Min.	1	5
	Max.	1,972	4,245
	Avg.	319	906
BBS	No	109	91
	Yes	87	106
Event	No	131	84
	Yes	69	141

†: Number of persons

In general, as the relative percentage of purchasers to non-purchasers (*i.e.*, only visitors) is very small, the data set results in a very skewed class distribution and analysis with such data might produce very trivial outcomes. To circumvent this, we sampled almost equal numbers of purchasers and non-purchasers.
The contents of the on-line retail shop consist of largely three product categories: fundamental cosmetics, coloring cosmetics, and clothes.

- Purchasers: total = 193
  - Fundamental cosmetics: 103,
  - Coloring cosmetics: 70,
  - Clothes: 20
- Non-purchasers: total = 200

4.3. Results and discussions

While training the tree, the user's threshold value of *confident percentage of uncertainty* $TG'/H(C)$ is compared simultaneously and training is stopped when $TG'/H(C)$ is equal to or greater than the user's threshold value. Then, the nodes determined by $TG'/H(C)$ are pruned once again by misclassification loss.
We first grew the whole tree until all terminal nodes had only one class or attributes were no longer split. As a result, the decision tree consisted of 25 nodes. On the other hand, using the pessimistic error pruning method built into C4.5 with a significance level of 25%, we obtained a decision tree with 18 nodes.
The slope of total information gain increased sharply at the 8th node and became approximately horizontal after the 9th node, as shown in Figure 2. These observations implied that proceeding to the next split after the 8-node decision tree contributes little information gain relative to the effort expended in searching for other attributes. The *confident percentage of uncertainty*, $TG'/H(C)$, was 67.6% at the 8th node and increased by only an additional 7.4% at the 18th node, but the difference in predicted loss rate between the 8th and 18th nodes was not significant: predicted error rate at the 18th node was decreased by 2.5% as compared with that at the

8^{th} node.

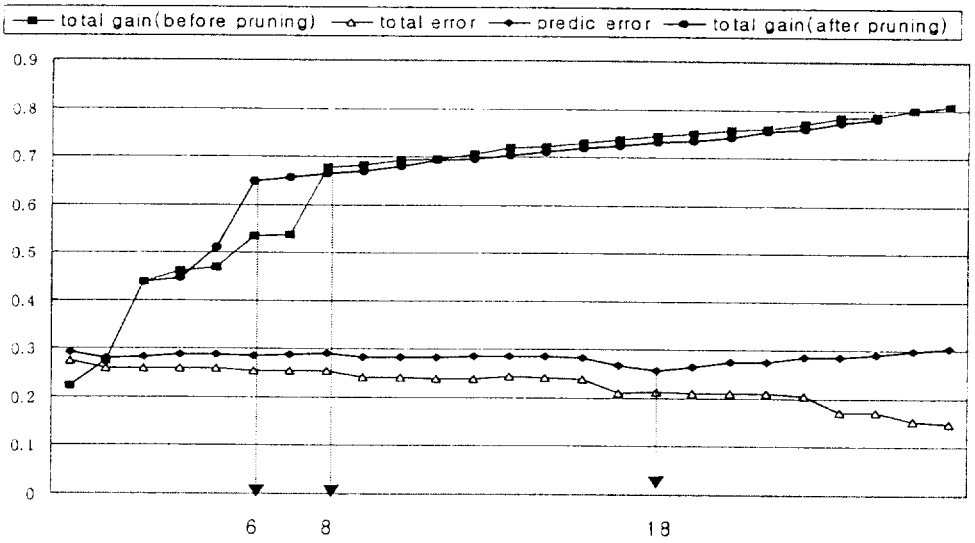

Figure 2. Comparison of TG, loss pruning, and pessimistic error pruning

Losses incurred due to false product category purchasers, false non-purchasers, and false purchasers were 0.5, 1, and 2 on a linear scale, respectively. In the 4^{th} and 8^{th} nodes, the predicted losses at leaf nodes were beyond those of the respective parent nodes. Thus, the number of segments was changed from 9 to 7 after pruning of the 4^{th} and 8^{th} leaf nodes. Through loss pruning, we can obtain a compact decision tree that is easy to apply to the classification problem but is not significantly different from the previous decision tree, as shown in Table 4. Existing pruning methods attempt to minimize prediction errors, and thus incremental reductions in errors cause expansion of the tree's branches, eventually leading to a bushy decision tree. In the customer classification problem, a bushy decision tree implies excessively subdivided segments and this can result in inefficient marketing strategies. Thus, derivation of the appropriate number of segments beyond the prescribed tolerance limits of information gained and administration of marketing strategies based on these conclusions can provide benefits to both marketers and customers.

Table 4. Indexes at three nodes

(Unit: %)

	6^{th} node 1	8^{th} node 2	18^{th} node 3	Diff. 3-2	Diff. 2-1
TG rate (before pruning)	53.3	67.6	75.0	7.4	14.6
TG rate (after pruning)	65.0	66.5	73.4	6.9	1.5
Predicted error rate	28.8	29.0	26.5	-2.5	-0.2

Moreover, the results obtained by the proposed approach can be used to develop a targeted marketing strategy. Attributes that classify purchasers and non-purchasers well are clicked pages (CP) and session time (ST), as shown in Table 5: CP $\leq$ 534 and ST $\leq$ 23,862 seconds for non-purchasers. This implies that non-purchasers are not willing to stay for a long time and do not visit the site frequently, as compared with purchasers. The most important point regarding non-purchasers is to increase the time they spend on the site. To do this, the company has to offer more attractive contents and services for which non-purchasers will be willing to visit the site. The next customer category, fundamental cosmetics purchasers, can be partitioned into two subgroups: the low-CP group and the mid-CP group. Note that customers in the low-CP subgroup have CP levels similar to those of non-purchasers but longer session times. We can predict that customers in the lower CP subgroup have no interest in other information provided by the company except that regarding fundamental cosmetics, as the unit session per clicked page was longer for this group as compared to the other groups. Thus, the company should provide compelling information about fundamental cosmetics. The other subgroup of customers interested in fundamental cosmetics, the mid-CP group, is comprised of people who search for information about fundamental cosmetics and perform comparisons between them as they click more pages than the low-CP group. This indicates that they are sensitive to marketing, *e.g.*, promotion events. Thus, we may recommend that the company provide appropriate information to convince these customers of product quality. Customers in the coloring cosmetics group had the highest number of clicked pages or quite long session times with a low number of clicked pages. This was due to the characteristics of the products, which involve ranges of products in many colors. The company sells more than 40 lipstick brands, each of which comes in from 3-11 colors. As customers interested in coloring cosmetics will perform comparisons among large arrays of colors, long session times are necessary. Therefore, the company should

provide high-resolution photographs of coloring cosmetics to allow customers to identify the differences between colors precisely. Finally, we found no significant features regarding clothes customers, suggesting that clothes customers do not have sufficient characteristics to discriminate clothes from other products.

5. Concluding remarks

The most important requirement in the customer classification problem is to determine the best splitting strategy. To minimize the errors incurred from wrong decisions, the number of segments should be neither too small nor too many [27]. If the number of segments is too high, then the decision tree is partitioned excessively resulting in increased marketing costs. Conversely, when the number of segments is too small, out-of-pocket costs due to incorrect marketing are incurred.

As described in this paper, it is necessary to modify the existing decision tree algorithm to prevent bushy rule trees, as the data used for segmentation may contain noisy or probabilistic components. Also, to manage the level of personalization on the Internet, it is necessary to control both the size and number of terminal subclasses of a resulting rule tree. For this purpose, we developed a stopping rule that considers the total amount of information gained while generating a rule tree. The proposed method is somewhat different from the existing pruning method that depends only on local information. The total amount of information gained prevents rule trees from being too sensitive to small frequencies, and builds rule trees that can reduce uncertainty. We also used Web log data to obtain a better understanding of customers' behavioral characteristics and further transactions in a real-world application. Our results suggest that pruning by loss information is suitable for customer classification in an on-line retail environment where opportunistic costs due to misclassification are a crucial problem. Error-based pruning results obtained by C4.5 were also discussed for the purpose of comparison. Our method may be useful in problem domains (*e.g.*, market segmentation) in which the whole picture is much more important than the precise details.

Although the method described here was developed to address a customer classification problem, it can also be used in other problem areas: for example, in pattern recognition or feature selection in data mining.

Table 5. Derived segments and examples of possible marketing opportunities

		Segment 7 (after loss pruning)	Segment 9 (before loss pruning)	Feature of segment	Marketing possibilities
Non-purchasers		- CP ≤534 and ST ≤23,862 (251, 82.9%)	- CP ≤534 and ST ≤23,862 and BBS is "Y" (103, 87.1%) - CP ≤534 and ST ≤23,862 and BBS is "N" and Event is "N" (79, 84.7%) - CP ≤534 and ST ≤23,862 and BBS is "N" and Event is "Y" (69, 84.6%)	Low CP, Low ST	Increase site visits
Purchasers	Funda-mental cosme-tics	- CP ≤534 and ST > 23,862 and BBS is "N" (11, 45.4%)		Low CP, High ST	Increase product information
		- CP 534 ~ 793 and Event is "N" (15, 73.3%) - CP > 534 and Event is "Y" and BBS is "N" (25, 44%) - CP > 534 and Event is "Y" and BBS is "Y" (51, 60.8%)		Mid CP	Precise product information
	Coloring cosme-tics	- CP ≤534 and ST > 23,862 and BBS is "Y" (9, 44.4%) - CP > 793 and Event is "N" (31, 48.4%)		High CP or Low CP, High ST	Product customization
	Clothes	-		No feature	-

References

[1] Liang T., Huang J. (1998). "The empirical study on consumer acceptance of products in electronic markets: transaction cost model," *Decision Support Systems*, Vol. 24, pp.29-43.

[2] Montgomery A.L. (2001). "Applying Quantitative Marketing Techniques to the Internet," *Interfaces*, Vol. 31, No. 2, pp. 90-108.

[3] Chang S. (1998). "Internet Segmentation: state of art marketing application." *Journal of Segmentation in Marketing*, Vol. 2, No. 1, pp.19-34.

[4] Firat A.F., Shultz II C.J. (1997). "From segmentation to fragmentation: markets and marketing strategy in the postmodern era," *European Journal of Marketing*, Vol. 34, No. 3-4, pp.183-207.

[5] Kara A., Kaynak E. (1997). "Markets of a single customer: exploiting conceptual development in market

segmentation," *European Journal of Marketing*, Vol. 31, No.11-12, pp.873-895.
[6] Vellido A, Lisboa P.J.G., Meehan K. (1999). "Segmentation of the on-line shopping market using neural networks," *Expert Systems with Applications*, Vol.17, pp.303-314.
[7] Hoffman D.L., Novac T.P. (1996). "Marketing in hypermedia computer-mediated environments: conceptual foundations," *Journal of Marketing*, Vol.60, pp.50-68.
[8] Angelides M.C. (1997). Implementing the Internet for business: a global marketing opportunity. International Journal of Information Management 17 (6): 405-419.
[9] Herbig P., Hale B. (1997). "Internet: the marketing challenge of the twentieth century," *Internet Research: Electronic Networking Application and Policy*, Vol. 7, No. 2, pp.95-100.
[10] Ozer M. (2001). "User segmentation of online music services using fuzzy clustering", *Omega*, Vol. 29, pp.193-206.
[11] Lee D.H., Kim S.H., Ahn B.S. (2000). "A conjoint model for Internet shopping malls using customer's purchasing data," *Expert systems with Applications* Vol. 19, pp.59-66.
[12] Quinlan J.R. (1983). "Learning efficient classification procedures and their application to chess end games". In Machine Learning: An Artificial Intelligence Approach, Michalski RS. et al.,(eds). Palo Alto, California; pp. 463-482.
[13] Hart A. (1984). "Experience in the use of an induction system in knowledge engineering". In Research and Developments in Expert Systems, Bramer M(eds.), Cambridge University Press.
[14] Breiman L., Freiman J.H., Olshen R.A., Stone C.J. (1984). *Classification and Regression Trees*. Chapman & Hall/CRC.
[15] Mingers J. (1986). "Expert Systems-experiments with rule induction," *Journal of Operations Research Society*, Vol. 37, No. 11, pp.1031-1037.
[16] Mingers J. (1987). "Expert Systems-rule induction with statistical data," *Journal of Operations Research Society*, Vol. 37, No.1 pp.39-47.
[17] Quinlan J.R. (1986). "Induction of decision trees," Machine Learning vol.1, pp. 81-106.
[18] Geman D. (2001). "Model-based classification trees,". *IEEE Transactions on Information Theory* Vol. 47, No.3, pp.1075-1082.
[19] Niblett T. (1987). "Constructing decision tree in noisy domains," *In Proceedings of the 2nd European Working Session on Learning*, Sigma Press; 67-78.
[20] Quinlan J.R.(1993). *C4.5: programs for Machine Learning*, Morgan-Kaufman.
[21] Oates T., Jensen D. (1997). "The effects of training set size on decision tree complexity," *In Proceedings of the 14th International Conference of Machine Learning*, Morgan Kaufmann; 254-262.
[22] Pazzani M., Merz C., Murphy P., Ali K., Hume T., Brunk C. (1994). "Reducing misclassification costs," *In Proceedings of the 11th international conference of Machine Learning*, Morgan Kaufmann.
[23] Bradford J.P., Kunz C., Kohavi R., Brunk C., Brodley C.E. (1998). "Pruning decision trees with misclassification costs," *In proceeding on the 10th European Conference on Machine Learning*, Chemnitz, Germany.
[24] Pawlak Z. (1991). *Rough Sets: Theoretical Aspects of Reasoning about Data*, Dordrecht, Kluwer Academic Publishers.
[25] Shannon C.E.(1948). "A mathematical theory of communication," *Bell system Technology Journal* Vol.27, pp.379-423.
[26] Provost F., Fawcett T. (1997). "Analysis and visualization of classifier performance: comparison under class and cost distributions," *In Proceedings of the 3rd International Conference on KDD-97*. AAAI press.
[27] Levin N., Zahavi J. (2001). "Predictive modeling using segmentation," *Journal of Interactive Marketing* Vol. 15, No.2 pp.2-22.