2010 IEEE International Conference on Data Mining Workshops

Using SOM-Ward clustering and predictive analytics for conducting customer segmentation

Zhiyuan Yao[1,2], Tomas Eklund[1], Barbro Back[1]

1. Department of Information Technologies, Åbo Akademi University, Turku, Finland
E-mail: {zyao, toeklund, bback}@abo.fi
2. Turku Centre for Computer Science (TUCS), Turku, Finland

***Abstract*— Continuously increasing amounts of data in data warehouses are providing companies with ample opportunity to conduct analytical customer relationship management (CRM). However, how to utilize the information retrieved from the analysis of these data to retain the most valuable customers, identify customers with additional revenue potential, and achieve cost-effective customer relationship management, continue to pose challenges for companies. This study proposes a two-level approach combining SOM-Ward clustering and predictive analytics to segment the customer base of a case company with 1.5 million customers. First, according to the spending amount, demographic and behavioral characteristics of the customers, we adopt SOM-Ward clustering to segment the customer base into seven segments: exclusive customers, high-spending customers, and five segments of mass customers. Then, three classification models - the support vector machine (SVM), the neural network, and the decision tree, are employed to classify high-spending and low-spending customers. The performance of the three classification models is evaluated and compared. The three models are then combined to predict potential high-spending customers from the mass customers. It is found that this hybrid approach could provide more thorough and detailed information about the customer base, especially the untapped mass market with potential high revenue contribution, for tailoring actionable marketing strategies.**

Keywords- customer segmentation; predictive analytics; self-organizing map (SOM); Ward's clustering; support vector machine (SVM) ; neural network (NN) ; decision tree

I. INTRODUCTION

Companies have long been diverting their attention from products to customers [1]. The Pareto principle that 20% of the customers create 80% of the profit or revenue [2-4] is commonly agreed upon in the industry. Reichheld and Teal [5] also claim that an increase in customer retention would result in a significant rise in company profit. Hence, companies should try to develop their analytical customer relationship management (CRM) to identify valuable customers, and customers with growth potential, to increase the aggregate value of their customer base. The application of enterprise resource planning (ERP) systems and customer data warehouses present companies with an opportunity to use data mining techniques to convert data and information into knowledge with the aim of improving customer relationships and facilitating decision making. Customer segmentation is an important part of analytical CRM, in which the customer base is divided into different groups based on similarity [6]. Effective customer segmentation enables companies to interact with customers in each segment collectively, and allocate limited recourses to various customer segments according to corporate strategies.

Data mining techniques, cluster analysis in particular, could assist companies in conducting customer segmentation [4, 7]. Cluster analysis is a collection of statistical and machine learning methods capable of dividing heterogeneous data into multiple more homogeneous clusters in a data-driven way, and it is often the first step in customer segmentation [7]. In the present study, we propose a two-step approach to carrying out customer segmentation. We first adopt the SOM-Ward method [8] (i.e., the Self-Organizing Map (SOM) [9] combined with Ward's clustering [10]) to conduct cluster analysis of a customer base. The customer base is divided into distinct groups of customers with similar characteristics and behavior, which allows us to specify the characteristics that distinguish high-spending customers from mass customers. Then, three classification models - the support vector machine (SVM), the artificial neural network (ANN) and the decision tree - will be used respectively to classify high-spending and low-spending customers. Although the three models have been applied in marketing individually, their suitability and performance in marketing have not been fully understood, nor has a comparison been examined in detail. In the present study, we will evaluate and compare the performance of the three methods. Furthermore, the three models combined will be used to reveal potential mass customers who share similar characteristics with high-spending customers, thus pinpointing the segments with high-value potential. By extension, this type of analysis allows companies to take into consideration the current value and potential migration pattern of the customers concurrently while creating marketing strategies. By doing so, companies could not only be more focused in maintaining the relationship with those high-spending customers but could also be more effective in launching cross-selling, up-selling, and deep-selling campaigns among mass customers.

The remainder of this paper is organized as follows. Section II introduces the methodology used in the study, i.e., SOM-Ward clustering and the three classification methods for conducting predictive analytics. Section III presents the data used in this study. Section IV describes the training and analysis of the SOM-Ward model. Section V documents the training process of the three classification models and the classification results; the performance of the three models is

978-0-7695-4257-7/10 $26.00 © 2010 IEEE
DOI 10.1109/ICDMW.2010.121

evaluated and compared. In Section VI, the trained classification models are adopted to analyze mass customers to identify their market potential. Finally, in Section VII, a conclusion is drawn and the empirical comparison of the three classification models is discussed.

II. Methodology

A. The SOM and SOM-Ward clustering

As a widely used unsupervised neural network for clustering tasks [11, 12], the SOM has been applied as an analytical tool in many industry applications [13-15], including market segmentation [12-14]. It is capable of projecting the relationships between high-dimensional data onto a two-dimensional display, where similar input records are located close to each other [11]. By adopting an unsupervised learning paradigm, the SOM conducts clustering tasks in a completely data-driven way [11, 16], i.e., it works with little *a priori* information or assumptions concerning the input data. In addition, the SOM's capability of preserving the topological relationships of the input data and its excellent visualization features motivated the authors to apply it in the present study.

A SOM is typically composed of two layers: an input and an output layer. Each input field is connected to the input layer by exactly one node, which is fully connected with all the nodes in the output layer [7, 17]. When the number of nodes in the output layer is large, the adjacent nodes need to be grouped to conduct clustering tasks. Accordingly, Vesanto and Alhoniemi proposed a two-level approach [18], e.g., the SOM-Ward clustering, to perform clustering tasks. The dataset is first projected onto a two-dimensional display using the SOM, and the resulting SOM is then clustered. Several studies [19-21] have shown the effectiveness of the two-level SOM, especially the superiority of the SOM-Ward over some classical clustering algorithms.

As mentioned previously, the SOM-Ward clustering is a two-level clustering approach that combines local ordering of the SOM and Ward's clustering algorithm to determine the clustering result. Ward's clustering is an agglomerative (bottom-up) hierarchical clustering method [10, 22]. The SOM-Ward starts with a clustering where each node is treated as a separate cluster. The two clusters with the minimum Euclidean distance are merged in each step, until there is only one cluster left on the map. The distance follows the SOM-Ward distance measure, which takes into account not only the Ward distance but also the topological characteristics of the SOM. In other words, the distance between two non-adjacent clusters is considered infinite, which means only adjacent clusters can be merged. A low SOM-Ward distance value represents a more natural clustering for the map, whereas a high value represents a more artificial clustering [8]. In this way, the users can flexibly choose the most appropriate number of clusters for their data mining tasks.

B. Predictive Analytics

Predictive analytics refers to a variety of techniques that deal with the prediction of future events by analyzing historical and current data [23, 24]. A prediction model starts by applying data mining techniques to historical data, trying to search for relationships between explanatory variables and a response variable. Once created and validated, the model could be used for a new dataset that shares the explanatory variables, to predict the possibility of a response variable. If the response variable is categorical, one should choose classification algorithms to construct a prediction model [23]. Three classification algorithms, i.e., the SVM, the ANN and the decision tree, selected for the present study, are elaborated upon in the following subsections. After the performance of the three models has been evaluated, they will be combined into an ensemble model for more accurate predictive analytics. We used three ensemble methods: voting, confidence-weighted voting and highest confidence win. In voting, the number of times each possible target value appears is summed and the one with the highest total is chosen as the prediction. Confidence-weighted voting works in similar way as voting except the confidence of prediction is taken into account and the votes are weighted by the confidence. In "confidence-weighted voting", the prediction with the highest confidence is chosen as the prediction of the ensemble model.

1) Support Vector Machine

The SVM, introduced by Vapnik [25], is a kernel-based method capable of conducting classification and regression tasks [26]. The use of kernel transformation effectively overcomes the problem of the "curse of dimensionality" [27], which enables the application of the SVM to a wide range of datasets. Furthermore, the SVM is based on structural risk minimization (SRM), i.e., a principle for model selection based on a trade-off between model complexity and training error [28]. Therefore, it guarantees a global unique optimal solution and can reduce the risk of overfitting [26]. All the advantages mentioned above have made the SVM an extensively applied technique in many industries [29].

In classification tasks, the SVM works by transforming input data using the kernel function into a high-dimensional feature space in which the classes of the data can be separated by a hyperplane. This hyperplane can be used to predict which category the new data belong to. In addition, on each side of the hyperplane, the SVM locates the maximum-margin hyperplanes, i.e., two parallel hyperplanes that maximize the distance between the data classes. The larger the distance between the maximum-margin hyperplanes is, the less likely the model will be prone to overfitting [26]. Readers can refer to [27, 30] for technical details and the algorithm of the SVM.

2) Artificial Neural Network

As its name implies, an ANN is designed to mimic the architecture of the human brain in a simplified way, to process information and learn from examples. ANNs have been widely applied in many business areas. In most of those studies, the ANN exhibits a performance as good as, if not better than, that of other methods [31]. The Back-propagation (BP) neural network, one of the currently most widely used neural network algorithms [32], is chosen for our classification tasks in the present study. As a type of a supervised ANN, the BP neural network functions by adjusting the weight among network nodes by constantly reducing the difference between predicted and actual values during the network training process. This process of feeding training examples and updating weights can iterate many times until the overall error is minimized. Thus, the network gradually becomes capable of understanding how inputs affect outputs. Once the network has been trained, its prediction capability should be evaluated using a test set, after which the network can be applied to predict outcomes for unknown examples. The algorithm of the BP neural network is well-known, so there will be no further explanation of it in this paper. For more details on the algorithm, readers could refer to [32].

3) *Decision Tree*

A decision tree is another classification model that can divide data into smaller and more homogeneous sets by recursively applying two-way and/or multi-way splits [7]. The greatest advantage the decision tree possesses is that it can produce a transparent predictive model, i.e., its output is a set of straightforward and explainable rules describing the relationships between explanatory variables and the response variable. Moreover, the decision tree can identify the significant variables for predicting the response variables; it is relatively insensitive to outliers; and it can deal with missing values [7, 33, 34]. In this study, we use the C5.0 decision tree [35] (modified and improved after the well-known C4.5 [36]) with the boosting algorithm [37] to classify high- and low-spending customers. The C5.0 boosting model works by constructing multiple decision tree models. The initial model is trained as usual, and in each of the following boosting rounds, all models are built in such a way that they focus on the records that were incorrectly classified by the previous models. Finally, records are classified by combining the weighted classification results of individual models, according to each model's performance, into one overall classification.

III. THE DATA

The data used in this study are from a service provider that belongs to a large, multiservice Finnish corporation. Through a loyalty card system, the corporation provides customers with various discounts and rewards based on the bonus points accumulated. Personal information of the cardholders is collected when they apply for the card, and their transactions are recorded in the system. The dataset with a total number of 1,522,701 customers was obtained through the loyalty card system. It contains aggregated sales information from several branches of the service provider in Finland, for the period 2007-08. The dataset consists of ten variables that fall into two bases: demographic and behavioral variables.

The demographic variables are shown as follows.

- Age
- Gender: 0 for male, 1 for female.
- Mosaic group: Mosaic is a household-based socio-economic ranking system developed by Experian PLC that classifies all Finnish households and neighborhoods (250-by-250 meter map grid cell) into one of nine unique segments. Each segment is described by demographic information, cultural/ethnic composition, lifestyle, purchase habits and so on.
- Mosaic class: Based upon the Mosaic group, the nine Mosaic groups are further divided into 33 subclasses.
- Estimated probability of children: Based upon the probability of their having children at home, this variable divides households into 10 groups of equal size. The higher the value of this variable is, the more likely there are children living in the households. The value ranges from 1 to 10.
- Estimated income level: This variable predicts customers' income level. The higher the value, the wealthier the household is considered to be. Possible values are 1, 2 and 3.

The behavioral variables are shown as follows.

- Loyalty point level: Based on the average purchase amount per customer in the corporate chain (the case company is one service provider in the corporation), this variable divides customers into five classes: 0, 1, 2, 3, and 4. A higher value in loyalty point level indicates a larger spending amount in the entire corporate chain.
- Customer tenure: This indicates how many years the customer has been a cardholder.
- Service level: The variable measures the number of service providers in the corporate chain that the customer has used in the last 12 months.
- The spending amount: The variable shows the total spending amount of each customer during the period 2007-08.

IV. THE SOM-WARD MODEL

A. *Training of the SOM-Ward model*

Viscovery SOMine 5.0, which builds upon the batch SOM algorithm [8], was used to train the SOM-Ward Model. As a user-friendly SOM implementation, SOMine includes three alternative clustering algorithms: SOM-Ward, Ward and SOM Single Linkage [15].

First, we preprocessed the data to ensure the quality and validity of the clustering result. Since the SOM requires a numeric input, we converted the Mosaic group into nine binary dummy variables (either 0 or 1). In addition, the Mosaic class variable was excluded from the SOM-Ward model as each of the 33 sub-classes of the Mosaic class would have required a dummy variable. This would have made visualization extremely difficult.

All the variables included in the training process were scaled to comparable ranges in order to prevent variables with large values from dominating the result. Viscovery SOMine offers two forms of scaling, linear and variance scaling. Linear scaling is simply a linear scaling based upon the range of the variable, and is suggested as default when the range of the variable is greater than eight times its standard deviation. Otherwise, variance scaling, i.e., the well-known normalization, is used. In this study, range scaling was applied to the variables spending amount, customer tenure and Mosaic groups, while variance scaling (normalization) was applied to the others.

For certain analysis purposes, some variables possess a higher priority than others do. Assigning a higher priority factor to variables gives them additional weight and importance in the training and segmentation processes. By default, the priority factor value is set to 1. The importance of a variable is reduced if its priority factor is set to less than 1, and accordingly, a variable with a priority factor value of 0 has no influence on the training process. The model is intended to explore the relationship between customers' spending amounts and the other characteristics; therefore, we decided to give the variable spending amount more weight in the training process. We assigned its priority value to 1.3 through experimenting with different values. In the pilot tests, it was discovered that the Mosaic group binary variables dominated the cluster formation, leading to clusters exclusively defined by some particular Mosaic group, and therefore, a biased segmentation result. To avoid this, we set the priority factor of the Mosaic group variable to 0.1 to ensure that the Mosaic group data had little impact on the segmentation result but that their distributions in the segments could still be investigated while training the map. Because the variables of the estimated probability of children and estimated income level both involved some estimates, their priority factor values were set to 0.5. In view of the results of the pilot tests, the priority factor values of other variables were somewhat adjusted to obtain a more interpretable segmentation result. The priority factors of age, gender, and customer tenure were set to 1.1, 0.9 and 0.8, respectively.

B. *Analysis of the SOM-Ward model*

The resulting SOM-Ward model consists of seven clusters. The component planes (Fig. 1) show the distributions of each variable across the map, on which the color scale visualizes the distribution of each variable over different segments. Cold colors indicate low values, while warm ones indicate high values, e.g., high-spending customers were mainly found in Segments Six and Seven, while long-standing customers were mainly found in Segment Five. Apart from the component planes, a series of error bar charts (Fig. 2) demonstrate the characteristics of each variable of the seven clusters. In each error chart there is a horizontal reference line, indicating the mean value of the variable in the whole customer base. The mean value of the variable in each segment is represented by a circular marker, and the 95% confidence interval of a variable's mean for each segment is represented by another two parallel red lines. For Mosaic group, we used a bar chart (Fig. 3) to illustrate its distributions in the seven clusters. The nine Mosaic groups are represented through the bars that are displayed in the same sequence in each segment. The height of a bar measures the extent to which the mean value of a variable in a cluster deviates from that of the entire data set. The unit of the y-axis is the standard deviation of the entire data set. Specifically, the green and yellow bars (the first and second bar) in Cluster One which are below zero level, show that the proportion of customers belonging to Mosaic groups B and E in Cluster One is less than that of the customer base in general. The key figures and important characteristics of each segment are summarized in Table IV.

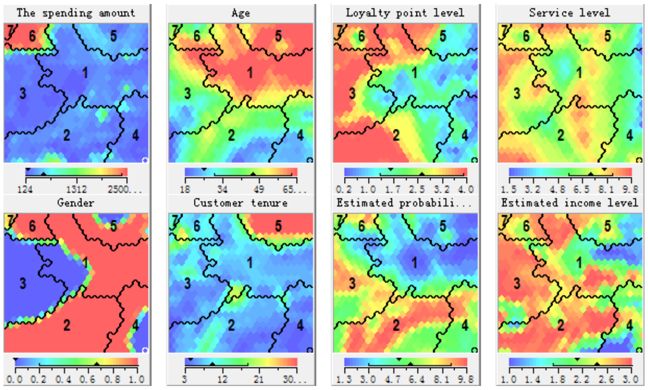

Figure 1. The component planes of the map.

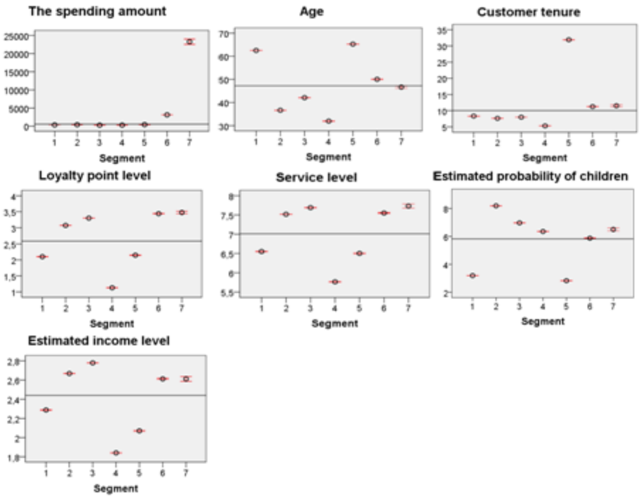

Figure 2. The error bar charts for each variable.

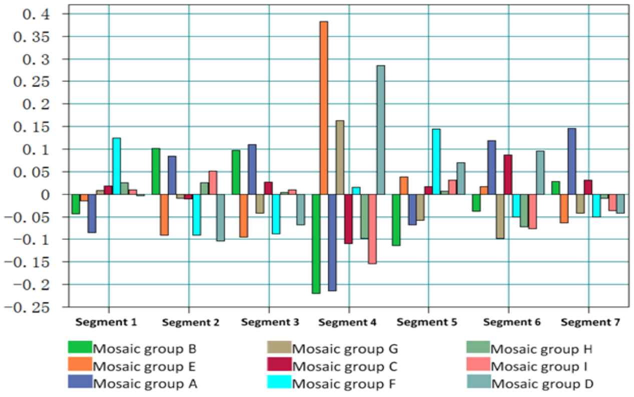

Figure 3. The bar chart illustrating the distributions of Mosaic groups in the seven segments.

V. The Classification Models

A. *Creation of the binary target variable*

A binary target variable, i.e., the variable of high- and low-spending customers, was created for the classification models. It was revealed from the analysis of the SOM-Ward model that high-spending customers were mainly distributed in Segments Six and Seven, especially in the shaded area in Figure 4, representative of 72,337 customers. After eliminating the records containing missing values, there were altogether 66,535 customers left, whom we labeled as high-spending customers. In SPSS Modeler, SVM only works with non-missing records, records with missing values for any input or output field are excluded from the estimation of the model. In ANN, missing values are handled by substituting neutral values for the missing ones. For this reason, we decided to remove the records with missing values so as to be able to compare all models regardless of the constraints of one modeling technique. Then we conducted an RFM analysis of Segments One to Five composed of mass customers. RFM analysis is a well-known method for analyzing customer purchase behavior and their tendency for buying more products [4]. Here, R (recency) represents the period since the last transaction. The lower the value of R is, the more likely the customers are to purchase again. Therefore, we used reverse recency, i.e. 365*2-recency to calculate the RFM score. F (frequency) represents the number of times customers purchased within a certain period, and M (monetary value) represents the spending amount of customers within a certain time period. To calculate the RFM score, we needed to assign different weights to the three RFM attributes. We found that the purchase amount of most low-spending customers was no more than 15 Euros, and that they had made only few transactions in two years, but their reverse recency varied from 1 to 730. Therefore, to obtain an unbiased result, we set the weight of recency to 0.01, and frequency and monetary values to 1, respectively. By doing this, we prevented the customers with high reverse recency from dominating the RFM scores. After working out the RFM scores, we selected from each segment the records with the lowest RFM score and with no missing values. From the RFM analysis, we located those customers with low spending amount and low potentiality of purchasing in the near future. The number of customers selected from each segment was proportional to the segment size, and customers with a low RFM score, whom we labeled as low-spending customers, added up to 66,535. We have thus constructed a dataset that contained an equal number of high-spending and low-spending customers. This balanced dataset enabled the model to identify important relationships in the underlying data [24]. In this dataset, the average purchase amount of high-spending customers was 4,649 Euros, while that of low-spending customers was 6 Euros.

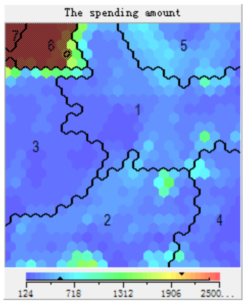

Figure 4. The high-spending customers (shaded area) chosen for the binary target variable.

B. *Training of the classification models*

The training of the three classification models was implemented using SPSS Modeler 13. In the classification models, we decided to include the Mosaic class in the training process because the three models can handle nominal data well. Since the Mosaic group gives the same, but less detailed, information as Mosaic class, we decided not to use it in this part of analysis. For SVM, SPSS Modeler provides four types of kernel functions: linear function, polynomial function, radial basis function (RBF), and sigmoid function. In this study, we used RBF, which is considered a natural choice when employing SVM [24]. The training of SVM using other kernel types takes an extremely long time. We experimented with different pairs of values for regularization parameter (C) and RBF gamma (γ). The higher the values of these two parameters are, the higher the classification accuracy of the training data is, but at the same time the risk of overfitting is higher. The prediction accuracies of the models were estimated by a ten-fold cross-validation and the model with the highest accuracy rate was chosen.

In employing the ANN, we used a multilayer perceptron with two hidden layers to explore the relationship between the predictors and target variables. The number of hidden layers and the units in each of them are automatically determined by the software. It starts with a large network and prunes the weakest units. By default, the network training process stops if there is no improvement between cycles. Users can also preset the desired accuracy rate and training time as stopping criteria. The default stopping criteria made the training last long because of the large dataset, so we preset training cycles as the stopping criteria. After experimenting with different training cycles, we chose a model with 1,000 training cycles being the stopping criteria according to the accuracy rate of cross-validation.

As mentioned in Section II, we used the boosting method to improve the prediction accuracy of the decision tree. Here we employed 10 boosting rounds, i.e., the method combining 10 base models of C5.0 that complement each other, with each model focusing on the records incorrectly handled by previously constructed ones. We also pruned the trees to prevent overfitting.

The accuracies of the three models based on resubstitution and ten-fold cross validation, and area under the ROC curve (AUC) are listed in Table I. In resubstitution, all records are used in training and the accuracy is also based on these data. Therefore, the result could be overoptimistic. AUC is a criterion used to measure the accuracy of the model in discriminating between target variable values. If the value of AUC is 0.5, it shows that the model does not perform better than guessing; if the value of AUC is 1, it shows that it is a perfect model. We found that the estimated prediction accuracy rates of the three models were high and close to each other. Though the accuracy rate of the boosted decision tree is higher than that of the other models, it tends to overfit the data whereas the support vector machine has the lowest risk of overfitting.

TABLE I. THE COMPARION OF THE THREE MODELS.

Classification models	***Accuracy (%) (Resubstitution)***	***Accuracy (%) (Cross-validation)***	***Area under the ROC curve***
Support vector machine	81.25	80.79	0.89
Neural network	80.31	79.60	0.87
Boosted decision tree	82.49	80.58	0.88
Decision tree without boosting	81.73	80	0.86

VI. PREDICTING POTENTIAL HIGH-SPENDING CUSTOMERS

The three models established earlier were then combined into an ensemble model to predict the segments with development potential among those composed primarily of mass customers. SVM cannot generate output probability for records with missing values in SPSS Modeler, which uses the method combining sigmoid and SVM [38], so it cannot be used alone for prediction tasks. We thus decided to incorporate SVM, ANN, and the boosted decision tree to predict the purchasing potential of mass customers. As discussed in Section II, we adopted three ensemble methods: voting, confidence-weighted voting, and highest confidence win. The three methods were used to classify the training data, and the accuracies of the three ensemble methods were compared. The method with the highest accuracy was then used to predict mass customers' potentiality. The confidence here refers to the posterior probability estimate for the predicted target class. As mentioned, the standard SVM does not generate a confidence value, which was estimated using a SVM plus sigmoid method here [38]. The backpropagation ANN uses the sigmoid function and the output activation values range from 0 to 1, with 0.5 being the cutoff value. The confidence was calculated as output activation value, which is twice the absolute difference value of 0.5. In the decision tree, each terminal node was a mixture of high-spending customers and low-spending customers. The predicted value of an unknown record was the category with the highest percentage of cases in the terminal node it belongs to. This percentage value was then used as the confidence value. For the boosted decision tree, the confidence values from each base model were then combined, while each model was assigned a weight that was proportional to its performance. Table II lists the overall accuracy of the three ensemble methods, and accuracy rates of predicting high-spending and low-spending customers. It was discovered that the three models performed better in classifying the high-spending customers than low-spending customers. Furthermore, the three ensemble methods slightly raise the classification accuracy rate of SVM and ANN on the training data. Although their overall accuracy is slightly lower than that of the boosted decision tree, the latter is more prone to overfitting data. For this reason, we decided to use the ensemble method with the highest overall accuracy, i.e., the confidence-weighted voting, to predict mass customers and locate those customers with development potential.

TABLE II. THE COMPARISON OF THE THREE ENSEMBLE METHODS.

Ensemble method	**Training data**		
	Overall accuracy (%)	***Accuracy of predicting high-spending customers (%)***	***Accuracy of predicting low-spending customers (%)***
Voting	82.37	84.83	79.92
Confidence-weighted voting	82.39	84.22	80.55
Highest confidence win	82.29	83.46	81.11

We then ran all the cases in Segments One to Five through the ensemble model using confidence-weighted voting. As for each customer, every classification model generates a confidence value for the predicted target class. The ensemble model aggregates the confidence values of the predicted classes generated by the classification models. The predicted class of the ensemble model is the one with the highest aggregate confidence value and its output probability is the aggregated confidence value of the class predicted divided by three. Through this output probability, we could obtain the propensity score of a customer being a high-spending customer. If the predicted result of the ensemble model is a high-spending customer, then the propensity score is the same as the output probability of the ensemble model; if the predicted result of the ensemble model is a low-spending customer, then the propensity score is the output probability subtracted by one. Table III lists the present value rank and the average propensity score for being high-spending customers for each segment, which could be used as an index for measuring customer potential. Analyzed together with Table IV, we could find out that customers in Segment Two were most likely to be potential high-spending customers while customers in Segment Four were the least likely. Though customers in Segment Two have smaller purchasing amounts than those in Segment Five, they possess a higher propensity score. Therefore, Segment Two plays no less important role than Segment Five does in the process of devising market strategies. Moreover, this result is visually confirmed in Figure 2, where Segment Two is most similar to Segments Six and Seven based on the variables shown. The average values of customer tenure, loyalty point level, service level and estimated income level of Segment Two are close to those of Segments Six and Seven.

TABLE III. THE PROPENSITY SCORES FOR SEGMENTS WITH MASS CUSTOMERS

Segment ID	Value rank	Propensity score
7	Exclusive customers	
6	High-spending customers	
2	Mass customers	0.49
5	Mass customers	0.39
1	Mass customers	0.39
3	Mass customers	0.28
4	Mass customers	0.20

VII. CONCLUSION

A two-step approach combining the SOM-Ward clustering and predictive analytics has been proposed to conduct customer segmentation. SOM-Ward clustering is first used to divide customers into exclusive customers, high-spending customers and mass customers, with each segment possessing different demographic and behavioral characteristics. Because mass customers occupy a great percentage of the entire customer base, we continue by using classification models that are capable of distinguishing between high- and low-spending customers to identify those mass customers with development potential.

Based on an unsupervised neural network, the SOM-Ward clustering is a useful tool for exploratory data analysis, as in the case when no *a priori* classes have been identified. The SOM is a visual tool and possesses strong capabilities of dealing with non-linear relationships, missing data, and skewed distributions. However, while the clusters produced using unsupervised methods may warrant a good understanding of the current customer base, they cannot necessarily provide information about the potential value of a segment.

Predictive analytics, on the other hand, are tailored to a specific purpose through a supervised learning approach. The three classification models used in the study can effectively distinguish between high- and low-spending customers and identify mass customers who have similar characteristics as high-spending customers. Companies could thus efficiently manage customer segment migration with the provided actionable information. However, the starting point of predictive analytics inevitably requires more *a priori* knowledge than unsupervised learning does, making the knowledge gained in using the SOM-Ward model potentially significant.

The results of the analysis demonstrate that the combined method of SOM-Ward clustering and predictive analytics can potentially be effective in conducting customer segmentation. The results of market segmentation, incorporated with the results of predictive analytics, are more prospective and predictive. Thus more actionable information about the untapped mass customers for marketing purposes could be retrieved.

ACKNOWLEDGEMENTS.

The authors gratefully acknowledge the financial support of the National Agency of Technology (Titan, grant no. 40063/08) and the Academy of Finland (grant no. 127656). The case organization's cooperation is also gratefully acknowledged.

REFERENCES

[1] C. Rygielski, J.-C. Wang, and D. C. Yen, "Data mining techniques for customer relationship management," *Technology in Society,* vol. 24, pp. 483-502, 2002.

[2] S.-Y. Kim, T.-S. Jung, E.-H. Suh, and H.-S. Hwang, "Customer segmentation and strategy development based on customer lifetime value: A case study," *Expert Systems with Applications,* vol. 31, pp. 101-107, 2006.

[3] H.-S. Park and D.-K. Baik, "A study for control of client value using cluster analysis," *J. Netw. Comput. Appl.,* vol. 29, pp. 262-276, 2006.

[4] K. Tsiptsis and A. Chorianopoulos, *Data mining techniques in CRM: Inside customer segmentation*: John Wiley and Sons, 2010.

[5] F. F. Reichheld and T. Teal, *The loyalty effect: The hidden force behind growth, profits, and lasting value*. Harvard Business Press, 2001.

[6] J. Dyché, The CRM handbook: A business guide to customer relationship management. Addison-Wesley, 2002.

[7] M. J. A. Berry and G. Linoff, *Data mining techniques: For marketing, sales, and customer relationship management*. Wiley, 2004.

[8] Viscovery Software GmbH, Viscovery SOMine 5.0, http://www.eudaptics.com/.

[9] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE,* vol. 78, pp. 1464-1480, 1990.

[10] Jr, "Hierarchical Grouping to Optimize an Objective Function," *Journal of the American Statistical Association,* vol. 58, pp. 236-244, 1963.

[11] *Self-organizing maps*: Springer-Verlag New York, Inc., 1997.

[12] J. Han and M. Kamber, *Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems)*: Morgan Kaufmann, 2000.

[13] S. Kaski, J. Kangas, and T. Kohonen, "Bibliography of Self-Organizing Map (*SOM*) Papers 1981-1997," *Neural Computing Surveys* vol. 1, pp. 102-350 1998.

[14] M. Oja, S. Kaski, and T. Kohonen, "Bibliography of Self-Organizing Map (SOM) Papers: 1998-2001 Addendum," *Neural Computing Surveys,* vol. 1, pp. 1-176, 2002.

[15] *Visual Explorations in Finance*: Springer-Verlag New York, Inc., 1998.

[16] T. Kohonen, J. Hynninen, J. Kangas, and J. Laaksonen, "SOM PAK: The Self-Organizing Map program package," 1996.

[17] L. Wiskott and T. Sejnowski, "Constrained optimization for neural map formation: A unifying framework for weight growth and normalization (1998)," *Neural Computation,* vol. 10, pp. 671-716, 1998.

[18] J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," *Neural Networks, IEEE Transactions on,* vol. 11, pp. 586-600, 2000.

[19] S.-C. Lee, J.-C. Gu, and Y.-H. Suh, "A Comparative Analysis of Clustering Methodology and Application for Market Segmentation: K-Means, SOM and a Two-Level SOM," in *Foundations of Intelligent Systems*, 2006, pp. 435-444.

[20] S. Samarasinghe, *Neural networks for applied sciences and engineering: from fundamentals to complex pattern recognition*: CRC Press, 2007.

[21] H. Li, B. Golden, E. Wasil, and P. Zantek, "A Comparison of Software Implementations of SOM Clustering Procedures," in *Intelligent Engineering Systems through Artificial Neural Networks*. vol. 12, A. B. Ca Dagli, Js Ghosh, M Embrechts,Os Ersoy, Ss Kercel, Ed.: ASME Press, 2002, pp. 447-452.

[22] C. Romesburg, *Cluster Analysis for Researchers*: Lulu.com, 2004.

[23] G. J. Myatt and W. P. Johnson, *Making Sense of Data II: A Practical Guide to Data Visualization, Advanced Data Mining Methods, and Applications*: John Wiley and Sons, 2009.

[24] J. F. H. Jr, "Knowledge creation in marketing: the role of predictive analytics," *European Business Review,* vol. 19, pp. 303 - 315, 2007.

[25] V. N. Vapnik, *The nature of statistical learning theory*: Springer-Verlag New York, Inc., 1995.

[26] D. L. Olson and D. Delen, *Advanced Data Mining Techniques*: Springer Publishing Company, Incorporated, 2008.

[27] N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines : and other kernel-based learning methods*: Cambridge University Press, 2000.

[28] V. Vapnik and A. Sterin, "On structural risk minimization or overall risk in a problem of pattern recognition," *Automation and Remote Control,* vol. 10, pp. 1495–1503, 1977.

[29] I. Guyon. SVM Application List. http://www.clopinet.com/isabelle/Projects/SVM/applist.html..

[30] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery,* vol. 2, pp. 121-167, 1998.

[31] A. Vellido, P. J. G. Lisboa, and J. Vaughan, "Neural networks in business: a survey of applications (1992-1998)," *Expert Systems with Applications,* vol. 17, pp. 51-70, 1999.

[32] R. Hecht-Nielsen, "Theory of the backpropagation neural network," in *Neural networks for perception (Vol. 2): computation, learning, architectures*: Harcourt Brace \&amp; Co., 1992, pp. 65-93.

[33] S. K. Murthy, "Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey," *Data Min. Knowl. Discov.,* vol. 2, pp. 345-389, 1998.

[34] L. Rokach and O. Z. Maimon, *Data mining with decision trees: theroy and applications*: World Scientific, 2008.

[35] R. R. P. Ltd, "Data Mining Tools See5 and C5.0," 2008.

[36] J. R. Quinlan, *C4.5: programs for machine learning*: Morgan Kaufmann Publishers Inc., 1993.

[37] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," in *Proceedings of the Second European Conference on Computational Learning Theory*: Springer-Verlag, 1995.

[38] J. Platt, "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods," in *Advances in Large Margin Classifiers*, 1999, pp. 61-74.

TABLE IV. SEGMENT PROFILE OF THE SOM-WARD MODEL.

ID	Value rank	Average spending amount per customer (€)	Size (%)	Contribution to overall purchase amount (%)	Segment Profile
7	Exclusive customers	23403	0.32	13.6	• Highest loyalty point level and service level • Relatively high estimated income level and estimated probability of having children • Mosaic groups A, B, and C
6	High-spending customers	3120	4.90	28.1	• Very high loyalty point level and service level • Relatively high estimated income level • Mosaic groups A, C, and D
5	Mass customers	421	8.57	6.6	• Highest age and customer tenure • Lowest estimated probability of having children • Relatively low estimated income level • Mosaic groups F, D, E, and I
2	Mass customers	381	28.56	20.0	• Relatively young customers • High loyalty point level, service level and estimated income level •Highest estimated probability of having children. •Mosaic groups B, A, I, and H
1	Mass customers	331	25.23	15.4	• Senior customers with low customer tenure • Relatively low loyalty point level. • Very low estimated probability of having children • Mosaic group F
4	Mass customers	285	11.68	6.1	• Youngest customers with lowest customer tenure • Lowest loyalty point level, service level and estimated income level • Mosaic groups E, D and G
3	Mass customers	270	20.73	10.3	• Very high loyalty point level, service level and estimated income level • Mosaic groups A and B