The Use of Neural Networks and Rule Induction for Customer Segmentation and Target Market Profiling

J Z Bloom

Department of Business Management, University of Stellenbosch

ABSTRACT

Inadequate market segmentation and clustering problems could cause an enterprise to either miss a strategic marketing opportunity or not cash in on a tactical campaign. The need for in-depth knowledge of customer segments and to overcome the limitations of non-linear problems require a different approach. The objectives of the research are (1) to consider the use of self-organising feature (SOM) neural networks for segmenting tourist markets and (2) to assess the use of inducing decision trees to obtain rules for profiling existing and classifying new respondents. The findings of the SOM neural network modelling indicate three definitive natural clusters. The induction of rules from decision trees were used to obtain a broad indication of a segment profile on the basis of a rule set and also enables the segment classification of customers from follow-up surveys.

JEL M30

1 INTRODUCTION AND BACKGROUND[1]

Marketing strategists often encounter the problem of how to segment and compile profiles of an enterprise's existing customers. Market segmentation is a process of dividing a market into distinct groups of buyers who might require separate products or marketing mixes (Venugopal & Baets, 1994: 36). Segmentation is based on various consumer characteristics such as demographics, socio-economic factors, geographic location, and product related behavioural characteristics like purchasing and consumption behaviour and attitudes towards and preference for products and services (Dibb & Simkin, 1991: 5). Target marketing is a strategy that aims at grouping a major market into segments in order to target one or more of these segments or to develop products and marketing programmes tailored to each segment (Kotler, 2000: 256-58).

Customer clustering and segmentation are two of the most important data mining methods used in marketing and customer relationship management (Saarenvirta, 1998: 1). Behavioural clustering and segmentation help drive strategic marketing initiatives, while sub-segments based on demographic and lifestyle characteristics could also be determined and used for tactical marketing efforts.

However, inadequate market segmentation and clustering together with a limited understanding of the characteristics of a segment profile, could cause an enterprise to either miss a strategic marketing opportunity or not cash in on a tactical campaign. Market segmentation has not only developed as a tool to segment markets and identify target markets, but could also be used at a higher level to obtain more in-depth knowledge of the segment characteristics and further assist an enterprise to understand the relationship with its customers.

The need for in-depth knowledge of customer segments and to overcome the limitations of non-linear problems requires a different approach. For instance, neural network models based on artificial intelligence technologies, can be developed to create clusters based on combinations of natural characteristics present in a set of customer data (e.g. purchase history, demographic attributes, phsychographic characteristics, etc.). However, many of the neural network modeling applications are used to develop individual models for a specific research problem. The sequencing of modeling applications such as using the output of a neural network modeling application as an input or output for another artificial intelligence technique e.g. the induction of decision trees to obtain rules) also enhances the knowledge base of customers' behavioural characteristics.

To illustrate the applications of a self-organising feature map (SOM) neural network for segmentation and the induction of decision trees to obtain rules, the data of a tourist survey of domestic tourists to the Western Cape isused. The use of SOM neural networks and induction of decision trees in travel and tourism is not widespread. Applications of neural networks in the tourism industry refer to, among others, customer analysis and holiday package targeting (Ryman-Tubb, 1993) and forecasting tourist behaviour (Pattie & Snyder, 1996). There is an increased need, however, for tools and techniques which could provide further knowledge and understanding of dynamic tourist behaviour. It has become essential for national and provincial tourism organisations to extract knowledge from the data obtained from tourist surveys. The findings of tourist surveys to be mostly descriptive and provide little or no indication of behavioural segments, or the underlying profile of the tourist characteristics that form part of a segment, or future segment classification.

In the light of the above, the primary objectives of the research are (1) to consider the use of SOM neural networks for segmenting tourist markets and (2) to assess the use of inducing decision trees to obtain rules for profiling existing and classifying new respondents by using the output provided by SOM neural networks.

The article firstly considers the nature and scope of a SOM neural network for learning and grouping data and the induction of decision trees to obtain rules. A conceptual comparison is provided of Cluster Analysis and SOM neural networks, and the advantages and disadvantages of decision trees are also discussed. A tourism industry application of using SOM neural networks and the induction of decision trees to obtain rules, is provided by means of a discussion of the methodology related to the modeling process. The outcomes of the modeling process in terms of the ability of a SOM neural network to naturally group data and the use of induction to obtain rules from decision trees, are also discussed. A conclusion is provided in the final section.

2 NATURE AND SCOPE OF SOM NEURAL NETWORKS AND INDUCING RULES FROM DECISION TREES

Artificial neural networks (ANN) have evolved as a technique to solve a variety of problems in the business field; from classification, grouping and forecasting, to portfolio optimisation, credit scoring and stock picking. Neural networks are described as information processing technology, which is inspired by the human brain and mimics its problem solving processes (Klimasauskas, 1996: 45). ANNs exhibit certain features such as the ability to learn complex patterns in a set of data and generalise the learned pattern (Venugopal & Baets, 1994: 30).

There are a variety of neural network algorithms for solving complex business and marketing problems. The appropriate use of a learning algorithm depends primarily on the type of problem which needs to be modelled. A taxonomy of learning algorithms has been proposed in the literature (Lippman, 1989: 10). The distinction is primarily differences in the input format, i.e. binary-valued input or continuous-valued input. Each of these categories could be further subdivided into supervised learning and unsupervised learning techniques.

Unsupervised learning algorithms use patterns that are typically redundant raw data, having no labels regarding their class membership or association. During this mode of learning, the network must discover for itself any possible existing patterns, regularities and separating properties. The parameters of the network undergo changes when discovering the properties mentioned previously and self-organising occurs (Turban & Trippi, 1996: 16). During unsupervised learning only input stimuli are presented to the network. Examples of this type of learning are adaptive resonance theory and Kohonen self-organising feature maps (Kohonen, 1988).

Once an acceptable mapping solution is obtained, it is possible to use the output obtained from the SOM neural network to develop an induction model for obtaining rules from decision trees. Decision trees are a way to represent mathematical regularities or relationships underlying a set of observations obtained for a particular problem. In addition, decision trees are hierarchical structures that partition the set of observations to explicitly relate a number of independent variables, known as the attribute, to one or more discrete dependant variables, or classes (Gray, 1990: 41-42).

A decision tree consists of nodes and branches. Each node is either a decision node that consists of a test on an attribute that partitions the current subset of observations into two or more smaller subsets, or a terminal node that classifies the remaining subset of observations in that node with a particular class label. Branches indicate the path that must be followed as decisions are made at each decision node until a terminal node is reached (Ben-David & Mandel, 1995: 110).

Decision trees are compact and are therefore readily understood. However, complex classification problems may cause the decision tree to become unwieldy with a large number of nodes and branches. Although such a tree may be complete and accurate it is often difficult to understand. It is possible to simplify the tree and make it more intelligible by expressing the tree model in terms of so-called If...Then rules, also referred to as production rules (Crusader Systems, 1998b: 115). Production rules have been widely used to represent knowledge in expert systems and they have the advantage of being easily interpreted by human experts because of their modularity, that is, a single rule can be understood in isolation and does not require a reference to other rules (Kamber, Winstone, Gong, Cheng & Han, 1997: 10).

3 NEURAL NETWORKS, INDUCTION OF DECISION TREES AND MULTIVARIATE STATISTICAL TECHNIQUES

SOM neural networks provide an opportunity to extract knowledge from data and offer improved performance by overcoming various limitations associated with multivariate statistical techniques such as Cluster Analysis. Although SOM neural networks also have limitations in respect of explanation, they offer advantages in terms of learning ability, flexibility, adaptation and knowledge discovery (Goonatilake, 1995: 21). The nearest neighbour algorithm, which is the premise for a SOM neural network, is a refinement of existing cluster techniques, in the sense that both use distance in some feature space to create structure in the data. The nearest neighbour algorithm offers more refinement, as part of the algorithm provides a way of automatically determining the weighting of the importance of predictors and how distance will be measured within the feature space (Berson, Smith and Thearling, 2000: 144-45). In the context of this research a comparison is provided of SOM neural networks and Cluster Analysis, while the advantages and disadvantages of inducing rules from decision trees are also highlighted.

The primary advantages of SOM neural networks over Cluster Analysis include the following:

- SOM neural networks are more robust than cluster techniques. The use of Cluster Analysis will provide a cluster solution even if no natural clusters exist in the data (Mitchell, 1994: 8).
- Various assumptions about the underlying distribution of the data are required to use Cluster Analysis, while SOM neural networks do not require any assumptions.
- The number of clusters requires specification when using Cluster Analysis, while SOM neural networks cluster data naturally based on assigning an incoming signal to the segment having the nearest weight vector (Venugopal & Baets, 1994: 36-37).

The relevance of SOM neural networks for modeling tourism data stems from the need to provide a segmentation solution that will enable decision-makers to allocate the scare financial resources for more focused target marketing. Besides overcoming the limitations of Cluster Analysis, SOM neural networks enable more refined analysis of tourist behaviour and also provides a level of predictive ability to track changes in tourist profiles and behaviour.

The goal of classification trees is to predict or explain responses on a categorical dependent or class variable and on this basis has much in common with multivariate techniques like Discriminant Analysis. For instance, the hierarchical nature of classification trees is illustrated by a comparison of the decision-making procedure employed in Discriminant Analysis. Superficially, the Discriminant Analysis and classification tree decision processes might appear similar, because both involve coefficients and decision equations. However, the difference of the simultaneous decisions of Discriminant Analysis from the hierarchical decisions of classification trees should be noted (Statsoft, 2001).

For the purposes of this article, the advantages and disadvantages of tree-based methods are discussed as a means to enhance the classification and profiling of customer segments. The advantages of decision trees and rule induction in mining data for classification, profiling and decision-making, include:

- Decision trees and rules explicitly represent the relationships discovered by the decision tree or rule induction algorithm and can be effectively understood and analysed by decision makers.
- The decision path followed by the tree or rule set can be easily followed when determining the class of a new observation, leaving nothing unknown or implied.
- If...Then rules can be imported into the rule base of decision support systems and can be integrated with a knowledge database.
- The influence and relevance of attributes with regard to classification of the set of training data is shown.
- The induction of decision trees is very flexible as it easily copes with discrete and continuous attributes. Regression and neural network classification require the encoding of discrete attributes as either a number of a continuous scale or as orthogonal unit vectors (Crusader Systems, 1998b: 119-20).

The disadvantages of decision trees are the following:

- The algorithms to induce trees are dependent on the quality and appropriateness of the attributes of the training data. If the attributes chosen do not adequately represent the underlying relationships in the data then the induction algorithm will group most of the data into one class with few nodes.
- The divide-and-conquer strategy partitions training data into smaller and smaller sub-sets. The algorithm uses less and less information about the entire training data as tree growth continues. The foundation for decisions based on the attributes lower down in the tree is diluted due to a smaller basis of information.
- The induced rules may require further processing and analysis after the initial induction cycle to improve comprehension (Crusader Systems, 1998b: 120-121).

The limitations of Cluster Analysis provide the rationale for the use of artificial SOM neural networks. Decision tree models are inexpensive to construct, easy to interpret and are easy to integrate with database systems and therefore their use for classification and profiling customer segments outweigh the limitations of inducing decision trees.

3 METHOD

On the basis of the research problem specification and the objectives of the research, the use of an unsupervised learning algorithm to group the data into segments and a decision tree algorithm for the profiling of tourist segments was required. Decision trees and rules are also generally used for classification purposes. A SOM neural network is used for the initial grouping of tourists, while the univariate-split decision tree algorithm is used for profiling the segments.

The research design is divided into two stages. The first stage involves the development of a SOM neural network and the second stage involves using the segment classification obtained from the SOM neural network model as output to create a decision tree from which rules are induced for segment profiling. The modeling process uses two different learning algorithms to group the respondents and allow the induction of rules from a decision tree. The rules obtained from the decision tree induction could also be useful for the classification of new tourists from data obtained in follow-up surveys.

3.1 Sampling procedure, data capture and scope of data

The sample was geographically stratified in line with the distribution of the urban population nationally, and starting points were selected using a geo-demographic sampling grid. Persons that reside in urban areas within South Africa and had visited the Western Cape at least once during 1997 and 1998 qualified for inclusion in the survey. A record was kept of unsuccessful contacts in order to be able to calculate the proportion of the urban population that visit the province. A total of 5 642 contacts were required to achieve the final sample of 1 630 respondents. Personal in-home interviews were conducted. The questions represented a broad mix of trip, demographic, socio-economic and geographic characteristics. The nature of the data was predominantly categorical and nominal.

The data was captured using a software package, Survey System (Creative Research Systems, 2001). The Survey System is tailored for survey research conducted using questionnaires. The software handles all phases of survey projects, from creating questionnaires through data entry, interviewing to producing tables, graphics and text reports.

The data used for the research presented in this paper is applicable to the domestic tourist market in South Africa. It is assumed that the behaviour of domestic tourists will not differ significantly since the period of the survey to the time the analysis was conducted. However, it may be relevant for Western

Cape Tourism to conduct a second survey in the near future to ascertain whether or not the profile and behaviour of tourists are changing.

3.2 Nature and design of a self-organising neural network model and the creation and induction of decision trees

The design of a SOM neural network is divided into several distinct steps. Several authors including Deboeck (1995), Masters (1993), Blum (1992) and McCord-Nelson & Illingworth (1993) have outlined a series of steps for building a neural network model. The eight-step procedure proposed by Kaastra and Boyd (1996: 219), which encompasses many of the steps proposed by the abovementioned authors, is adapted for SOM neural networks and presented in Table 1. The steps for the creation of decision trees and the induction of rules are also presented in Table 1, as adapted from Brodley & Utgoff (1995: 49).

Table 1 A sequence of steps used to design a SOM neural network model and create a decision tree for the induction of rules

Steps for SOM neural network specification	Steps for specification of decision trees and rule induction
Step 1: Data collection	Step 1: Data collection
Step 2: Variable selection (number of inputs)	Step 2: Attribute selection (number of inputs, and output)
Step 3: Data pre-processing (e.g. normalising, log transformation, standardisation, scaling)	Step 3: Data pre-processing (e.g. removal of outliers and transformations)
Step 4: Selection of training, validation and test sets	Step 4: Selection of training, validation and test sets
Step 5: Specification of SOM neural network training parameters and configuration values	Step 5: Specification of decision tree training parameters and configuration values
-Number of input neurons	-Number of input neurons
-Percentage iterations for constant initial learning rate	-Stopping rules (Uniform class, uniform vector, minimum observations)
-Learning rate increment	-Splitting criteria (Gain, gain ratio or gain ratio/log (Depth)
-Neighbourhood radius	-Pruning algorithm (error based pruning, reduced error pruning)
-Radius increment	

Table 1 continued

Steps for SOM neural network specification	Steps for specification of decision trees and rule induction
Step 6: SOM neural network training specifications	Step 6: Rule induction specifications
-Initial weights (upper and lower bound)	-Rule generalisation algorithm (upper confidence limit, correctness generalising and best difference)
-Presentation of records	-Rule set optimising algorithm (Minimum Description Length Principle)
-Number of iterations (epochs)	
Step 7: Evaluation criteria	Step 7: Evaluation criteria for both the decision tree and rule induction (before and after pruning on the training data and validation set)
Step 8: Model deployment	Step 8: Model deployment

Source: Adapted from Kaastra and Boyd (1996: 219) and Brodley and Utgoff (1995: 49)

The development of any neural network or rule induction model, is based on a thorough knowledge of the research problem. In addition, the procedure described in Table 1 is not one-off, but may involve revisiting steps between the training and validation of the model to reassess the input variables and network parameters.

All the variables included in the survey that had no missing data, were considered for the modeling process. An exploratory data analysis (EDA) was conducted to provide an indication of the distribution of the variables. Both Box and Whisker plots and histograms were used to visualise the distribution of the data variables. These descriptive statistical techniques were used together with several descriptive statistics (i.e. mean, median, standard deviation, skewness and kurtosis) to further describe the data. The EDA is important to ensure non-inclusion of data variables with substantial positive or negative skewness, as the natural clustering algorithm of the SOM neural network would tend to consider the data as one group and distinguishing between groups would be more difficult due to the nature of the algorithm.

Based on the outcomes of the exploratory data analysis and the importance of selecting the right mix of variables for creating a SOM neural network model, 10 variables, which include trip, demographic, socio-economic and geographical characteristics, are included for the SOM neural network modeling process. The data captured in Survey System was exported in a delimited format into Excel, which also facilitated the importation of the data into the modeling software. In order to enhance the description of the segment profile, six additional variables were included for the creation of the decision tree and the induction of the rules. Table 2 lists the 10 variables used for the SOM neural network modeling and the 16 variables for the induction of rules from a decision tree.

Table 2 Variables used for the construction of the decision tree and rule induction

Variables used for SOM modeling procedure	Variables used for creation and induction of decision trees
Main purpose of visit	Main purpose of visit
Region visited in the Western Cape province	Region visited in the Western Cape province
Duration of stay	Duration of stay
Total spent on first trip	Total spent on first trip
Likelihood to visit the Western Cape again	Likelihood to visit the Western Cape again
Ethnic group	Ethnic group
Occupation	Occupation
Origin of tourist	Origin of tourist
Income group	Income group
Age	Age
Education	Education
	Number visits to the Western Cape in past 2-years
	Time of the year visit to the Western Cape
	Information sources of the Western Cape
	Arrangements for visit
	Likelihood to visit the Western Cape province again
	Gender

The software package used for part of the exploratory data analysis and the modeling procedure is Basic Modelgen (Crusader Systems, 1998a). In addition, the software package, Statistica, is used for most of the exploratory data analysis (Statsoft, 1998).

3.2.1 Self-organising neural networks

The following discussion of the SOM neural network modeling process refers to the steps listed in Table 1. Data variables identified for modeling were scaled or recoded to assume values between 0 and 1 or 0 and 1 respectively. In this manner each data record is considered by the network as continuous valued input or binary-valued input. In order to create the training, validation and test sets, the data set (1 630 records) was randomly sub-divided so that 70 per cent of the data points were allocated to the training set, 10 per cent to the validation set and 20 per cent to the test set. This classification is based on heuristics and on the principle that the size of the validation set must strike a balance between obtaining a sufficient sample size to evaluate both the training and test sets (Kaastra & Boyd, 1996: 223).

The training parameters used for final SOM neural network modeling process entails the following measures together with the relevant values.

Parameter and Values

Learning decrement (per cent)	0.1
Initial weights (per cent)	-0.5 (Lower bound); 0,5 (Upper bound)
Scaling	0.1 (Lower bound); 0,9 (Upper bound)
Neighbourhood radius	5
Radius increment	Linear
Side N of one map side:	40 which refers to 1600 data point images on a two-dimensional grid

Configured values

Presentation of data	Random
Training cases	70 per cent
Test cases	20 per cent
Validation cases	10 per cent
Number of inputs	10 neurons.

The SOM neural network was trained for 1000 iterations and the records were presented to the network in a random manner. The findings of the SOM neural network modeling procedure are presented in a following section.

3.2.2 Induction of rules from decision trees

The second stage of the analysis entailed the induction of rules from a decision tree. The decision tree was created using 16 variables including the 10 variables that were also used for the SOM modeling process. In order to specify an output variable each respondent belonging to a specific segment was mapped back to the original data set. The model specification for the decision tree included 16 input neurons and a single output neuron representing the segment classification of each respondent. The primary aim of this stage of the research is to develop a trained rule induction model which could be used to provide an indication of the profile of each segment and also assist with the classification of new respondents based on the exiting rule set for each segment.

The parameters and configured values used to create the decision tree and induce rules are as follows:

Decision tree and rule induction specifications

Stopping rules	Uniform class Minimum observations (2)
Splitting criteria	Gain ratio
Pruning algorithm	None specified
Rule generalisation algorithm	Correctness generalising

Model specification

Training cases	90 per cent
Test cases	10 per cent
Number of inputs	16 neurons
Number of outputs	1 neuron.

Many algorithms are proposed in literature, to induce a decision tree and thereafter use the tree structure as a basis for deriving a set of production rules. It is beyond the scope of this text to describe each of these algorithms, especially those that induce rules from scratch. For the purposes of this article, the well-known C4.5 algorithm is used for the induction of the decision tree (Quinlan, 1993). The C4.5 algorithm generates a classification-decision tree for the given data set by recursive partitioning of data. The algorithm considers all the possible tests that can split the data set and selects a test that gives the best information gain. For each discrete attribute, one test with outcomes as many as the number of distinct values of the attribute is considered. In addition, for each continuous attribute, binary tests involving every distinct value of the attribute are considered. In order to gather the entropy gain of all these binary tests efficiently, the training data set belonging to the node under consideration is sorted for the values of the continuous attribute and the entropy gains of the binary cut based on each distinct value are calculated in one scan of the sorted data. This process is repeated for each continuous attribute (Kamber, *et al.*, 1997: 4).

For the purposes of this research, the uniform class together with the minimum observation stopping rules are used. The uniform class stopping rule is generally used in classification problems and will fire if all the cases at a node are of the same class. The minimum observations stopping rule will fire when the number of cases in a node are less than or equal to the specified number, which is two in this case. The use of two stopping rules is to prevent a large overtrained tree.

The Gain ratio is used as the splitting criteria in the C4.5 algorithm and is considered as a measure of impurity (Quinlan, 1993 & Breiman, 1996). It is also a commonly used criterion and generally provides good results. This measure determines the ratio of the extent of information contained within the new smaller subsets of data after a given split in the data has been performed over that contained in the previous larger sub-set of data. The attribute and its associated value on which to split are chosen as the pair that minimises the gain ratio (Quinlan, 1993).

The rule generalisation algorithm used in this study simplifies the predicate of a rule. The correctness generalising algorithm keeps the optimal predicates to maximise the number of correct cases identified by a rule (Crusader Systems, 1998b: 71). The C4.5 algorithm uses a criterion that pessimistically estimates the expected classification error rate of a generalised rule to decide which part of the antecedent to remove. The change in the rule that produces the lowest error rate is selected (Quinlan, 1993). The outcome of the induction of the decision tree is discussed in a following section.

4 FINDINGS OF THE MODELING PROCEDURE

4.1 Findings of the SOM neural network model

It was possible to distinguish three clear segments among domestic tourists that visit the Western Cape. Figure 2 is an illustration of the self-organising feature map, which is obtained through the unsupervised SOM neural network modeling process. Each record in the training set corresponds to a single unit, namely the best matching one. Thus, the unit represents the record's image on the feature array (consisting of 40x40 units), which implies 1600 output units in this case. The self-organising feature map is two-dimensional and depicts the natural segments obtained from the interrelationships between the 10 variables used for the modeling of the data.

Figure 1 Two-dimensional self-organising feature-map of the trained data

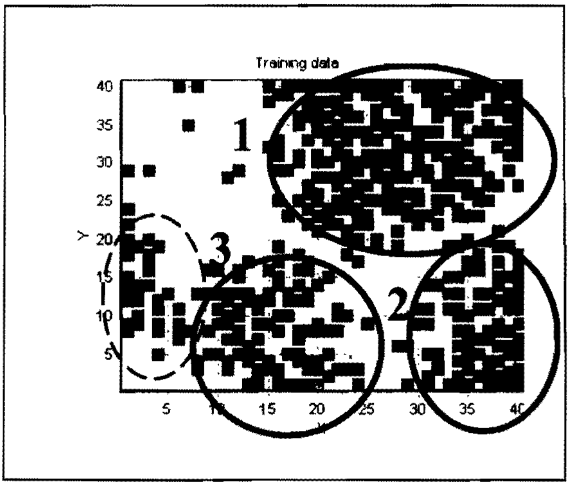

Figure 1 indicates four possible segments, three of which are definitive, while the fourth is smaller and appears to be developing. However, for the purposes of this analysis, the respondents that form part of the "emerging" market segment are included together with respondents classified as part of segment 3. Table 3 indicates the size and number of respondents classified per segment.

Table 3 Number of respondents per segment for the traveller group

Segment number	Total respondents per segment	Percentage contribution
1	693	42.52
2	421	25.82
3	516	31.66
Total	**1630**	**100.00%**

Table 3 shows that Segment 1 is the larger of the three segments and that 42.52 per cent of the respondents have a similar profile of attributes, views and characteristics. Among the remaining respondents, 31.66 per cent have a similar profile and are grouped in Segment 3, while 25.82 per cent of the respondents in Segment 2 could be considered as a homogeneous group. The three segment profiles obtained from the classification of each tourist based on the SOM neural network is presented in Table 4 with acronyms for each of the segments based on the profile.

Table 4 Profiles of the three segments identified from the self-organising feature map

Profile of segment 1: "Adventurists"

The largest majority of tourists who form part of segment 1 either visit the Western Cape for holiday purposes or visit friends and relatives. This group also spends most of their vacation period in Cape Town and the Cape Peninsula. Other areas of interest to this group are the West Coast and Garden Route. An average duration of stay for this segment is 16 days (median 14 days), while most of the tourists spend between R3 000 and R4 000 on average (median R3 000 to R4 000), irrespective of the purpose of visit. The majority of this segment is from the White population group, of which almost 40 per cent is self employed or work as clerical or sales staff. A small portion (11 per cent) of this group occupies professional positions in the business fraternity. Over 63 per cent of the tourists from this segment come from Gauteng. These tourists earn an average of between R10 500 and R12 999 per month (median R8 500 to R10 499) and are between the ages of 35 and 39. Over 50 per cent of these individuals have a matric qualification, while 37 per cent have an additional diploma or university degree.

Profile of segment 2: "Yuppies"

Tourist classified under segment 2, visit the Western Cape for reasons other than VFR, leisure or business. However, almost a fifth do also visit the Western Cape for business. They visit primarily Cape Town or the Cape Peninsula in the Western Cape and spend approximately 12 days (median 10 days) in the province. They spend a low average total of between R2 001 and R3 000 (median R1 001 to R2 000) during this time. Interestingly, the ethnic spread among this segment is similar to that of segment 3, for White, Black and Coloured/Asian groups. Almost 40% of this group is professionals, while about 37% is middle management, self employed, clerical or sales staff. Approximately a third of these tourists come from the Eastern Cape, while close to 30% come from Kwazulu-Natal. These tourists earn an average of between R13 000 and R15 999 per month (median R10 500 to R12 999) and are between the ages of 35 and 39 (median 35 – 39 years of age). Sixty per cent of this segment has a matric qualification or diploma and almost 40% a university degree.

Profile of segment 3: "Content"

Tourists classified as part of segment 3, visit the Western Cape either to see friends or relatives or to have a holiday. Almost 15 per cent of this segment also visits the province for business purposes. More than 80 per cent of these tourists visit Cape Town or the Cape Peninsula during their visit to the Western Cape and spend approximately 12 days (median 10 days). This segment spends a low average total of between R1001 and R2000 (median R1001 to R2000). As mentioned in segment 2, the ethnic spread among this segment is similar regarding White, Black and Coloured/Asian groups. Almost 40 per cent of this group are housewives, students, retired persons or semi-skilled workers. Approximately 40 per cent of these tourists come from Gauteng, while almost a third are from Kwazulu-Natal and the Free State. These tourists earn an average of between R4 000 and R4 999 per month (median R4 000 to R4 999). The average age of this group is between 40 and 44 years (median 40 – 44 years of age), while more than three-quarters has a matric, Std 8, Std 9 or Std 9 with a diploma.

Note: The basis for the creation of the segment profile, is frequency tables compiled for each variable included in the SOM neural network modeling procedure and several descriptive statistics. These segment profiles should be considered as a means to distinguish between the different segments on an overall basis.

The "Adventurists" seem to be scattered throughout the province and visit other areas in addition to Cape Town and the Peninsula. They are less trendy than the "Yuppies", who prefer to spend their holiday period in areas like Cape Town and the Peninsula that seem to have a certain vacation atmosphere. Interestingly, the "Yuppies" earn more than the "Adventurists" but spend less during their time in the Western Cape. The tourists in the "Content" segment are, by virtue of its profile, older than those of the other segments; retired, or students and less qualified. Their expenditure over the vacation period is also lower than the other two segments. The "Content" appear to be more content with life and are less energetic and adventurous than tourists whose profile fits in the other two segments.

4.1.1 Implications of the segment classification for management decision making

The three profiles above provide three different behavioural segments, portrayed by the acronym assigned to each. Consequently, decision makers, could develop more focused marketing strategies instead of the less generic ones often derived from the supply-side tourism industry (e.g. attractions). These segment profiles could be incorporated into existing tourism positioning strategies, by more specifically indicating to decision makers the kind of tourists the Western Cape would target in the context of the existing and envisaged tourism framework of the province.

In addition, the research conducted in this paper will also allow decision makers to track the behaviour of tourists by conducting similar surveys in the future and also determine the changing profile of key market segments for the Western Cape. Decision makers could also, from a similar analysis performed on tourists that do not visit the Western Cape, determine corresponding profiles and identify potential tourists that fit the profile of those tourists the industry decision makers in the Western Cape would want to attract.

4.2 Findings of the creation and induction of a decision tree

The decision tree and rule induction modeling could also be used to obtain a broad profile of a segment based on a specific set of rules. In addition, the rules could also be used to classify tourists that take part in future or follow-up surveys. By using the variables indicated in Table 2 as inputs and considering an output variable representing the segment classification of each respondent, it is possible to compile a decision tree from which rules could be deduced. The creation of the decision tree and induction of rules are based on the set criteria listed in section 3.2.2[2].

The rules indicated in Table 5 and in the Appendix are interpreted as If...Then rules. This implies that should a tourist assume or adhere to the specifications for an individual rule, it would be possible to assign them to a certain segment classification. In this manner they could assume one of the profiles described in Table 4. For instance, consider Rule 13 in the Appendix and the first rule applicable to segment 3 in Table 5. A description of the rule indicates that 167 cases were correctly classified as segment 3, while one case was incorrectly classified. The accuracy of the rule is 99,4 per cent which is based on the test set (unseen data). Rule 13 could be described in the following manner:

IF the tourist spends less than R3000 *and* (s)he is a "non-professional" *and* predominantly African, Coloured or Asian, *and* has a matric or lower education qualification, *and* earns less than R6 500 per month *and* is from the Free State, Northern Cape, Kwazulu-Natal or the Eastern Cape, *THEN* the tourist would have a similar profile to the 421 respondents grouped within Segment 2. The individual rules for the other segments are used in the same manner to describe a particular segment or classification of a tourist.

Table 5 Examples of rules for the different segments

<table>
<tr><th>Rules for segment 3 [516]*</th><th>Rules for segment 1 [693]</th></tr>
<tr><td><i>IF</i> a tourist spends less than R3000 <i>and</i> is a non-professional <i>and</i> predominantly African, Coloured or Asian, <i>and</i> have a matric or lower education qualification, <i>and</i> earn less than R6 500 per month <i>and</i> are from the Free State, Northern Cape, Kwazulu-Natal or the Eastern Cape … <i>or</i>
<i>IF</i> the tourist visits Cape Town and the Peninsula <i>and</i> total spending is less than R2000 <i>and</i> is predominantly African, Coloured or Asian <i>and</i> has a matric or lower qualification <i>and</i> earns R6 500 or less <i>and</i> is from the Free State, Northern Cape, Kwazulu-Natal, or the Eastern Cape
<i>THEN</i> the tourist would have a similar profile to the 516 respondents grouped within Segment 3.</td><td><i>IF</i> the tourist is younger than 60 years of age <i>and</i> is likely to visit the Western Cape in the future <i>and</i> spent more than R3 000 during the visit, <i>and</i> is predominantly white <i>and</i> is from the Northern Provinces of the country <i>and</i> earns more than R8 500 per month … <i>or</i>
<i>IF</i> the tourist visits other areas than the Garden Route and Cape Town and the Peninsula, <i>and</i> is not from the Eastern Cape, but the other provinces, <i>and</i> earns more than R8 500 per month <i>and</i> is a white collar employee….
<i>THEN</i> the tourist would have a similar profile to the 693 respondents grouped within Segment 1.</td></tr>
<tr><th colspan="2">Rules for segment 2 [421]</th></tr>
<tr><td colspan="2"><i>IF</i> the tourist is a white collar worker, <i>and</i> has a matric or higher education, <i>and</i> is not from the northern provinces of South Africa <i>and</i> is predominantly African, Coloured or Asian, <i>and</i> spent more than R3 000 during the visit to the Western Cape…<i>or</i>
<i>IF</i> the tourist spends 15 days or less in the Western Cape, <i>and</i> is from the Free State, Northern Cape or the Eastern Cape, <i>and</i> has a university degree…
<i>THEN</i> the tourist has a similar profile to the other 421 tourists which forms part of segment 2.</td></tr>
</table>

* The value indicated in parenthesis represents the number of tourists for each of the segments.

Table 5 clearly indicates the class (segment), the "conditions" applicable to each of the attributes and the classification accuracy of the rule. These rules, which are determined in a quantitative manner, could also be combined with rule sets (qualitative information) obtained from experts. The use of SOM neural network models together with the induction of rules from decision trees requires clear research objectives, domain knowledge of the specific problem, representation of appropriate attributes and clarity on the definition and acquisition of data.

5 CONCLUSION

Inadequate market segmentation and clustering problems could cause enterprises to either miss a strategic marketing opportunity or not cash in on a tactical campaign. Market segmentation has not only developed as a tool to segment markets and identify target markets, but could also be used at a higher level to further assist an enterprise to understand the relationship with its customers. The need for in-depth knowledge of customer or tourist segments and the need to overcome the limitations of non-linear problems require a different approach. For instance, SOM neural network models based on artificial intelligence technology can be developed to create clusters based on combinations of natural characteristics within a set of customer data (e.g. purchase history, demographic attributes, phsychographic characteristics, etc.).

The research demonstrates the usefulness of artificial intelligence technology as a means of grouping respondents and for profiling existing respondents by using a rule set applicable to each segment. The SOM neural network application, which could be considered as an enabler, provided three segments (classes) that could be used to induce rules from decision trees. The rules provide decision makers with clear and concise indications of the segment profile of existing tourists based on an individual rule within a larger rule set. In addition, the rules sets for each segment could be used to classify tourists that partake in follow-up surveys.

The artificial intelligence application presented in this paper provides a mechanism for analysts to use if the objective of the research is to create customer segments and to describe the segments through a rule set applicable to each segment. The findings demonstrate that the analysis of data from surveys can be taken to a high level through the provision of knowledge represented in sets of rules also by overcoming the limitations of cluster techniques such as Cluster Analysis.

ENDNOTES

1 I would like to express my gratitude to Western Cape Tourism for providing the data to conduct the research presented in this article. The usual caveat applies.

2 The decision tree and the additional rules (i.e. with accuracy of less than 90 per cent) are available on request.

APPENDIX

Rules with accuracy of larger than 90%

(A legend is provided to assist with the interpretation of the rules)

1468 Training Cases and 162 Validation Cases

Segment 1 = A; Segment 2 = C; Segment 3 = B

Rule 13 ** B 167, 1 (99.4%) totspend < 4 occupation >= 6 race >= 2 education < 5 incgroup < 9 province < 6	Rule 9 ** B 88, 1 (98.9%) mainarea >= 8 totspend < 2 race >= 2 education < 5 incgroup < 9 province < 6	Rule 72 ** B 79, 3 (96.2%) education < 4 incgroup < 10 race >= 2 province >= 6 occupation < 8
Rule 153 ** A 189, 1(99.5%) age < 10 liklyvisit < 3 totspend >= 4 race < 2 province >= 6 incgroup >= 10 occupation >= 8	Rule 96 ** A 131, 2 (98.5%) incgroup >= 6 province >= 4 race < 2 mainarea < 5 incgroup < 10	Rule 39 ** A 208, 7 (96.6%) incgroup < 14 province >= 5 totspend >= 5 race < 2 incgroup >= 9
Rule 7 ** B 131, 2 (98.5%) education < 4 race >= 2 incgroup < 9 province < 6	Rule 114 ** A 179, 5 (97.2%) province >= 5 age < 6 incgroup >= 7 race < 2 totspend >= 4 incgroup < 10	Rule 48 ** B 109, 4 (96.3%) dayspent < 10 mainarea >= 7 incgroup < 8 education < 5 totspend < 4 province >= 6
Rule 23 ** A 117, 2 (98.3%) mainarea < 6 race < 2 province >= 3 incgroup >= 9 occupation < 8.	Rule 52 ** B 94, 4 (95.7%) age >= 5 totspend < 2 incgroup < 10 education < 5 province >= 6	Rule 120 ** C 50, 2 (96%) occupation < 13 education >= 5 province < 7 race >= 2 totspend >= 4 incgroup < 10
Rule 151 ** A 178, 3(98.3%) occupation < 12 totspend >= 3 incgroup >= 11 race < 2 province >= 6 occupation >= 8	Rule 50 ** B 73, 2 (97.3%) occupation >= 6 dayspent >= 10 mainarea >= 7 arrange >= 2 incgroup < 8 education < 5 totspend < 4 race < 2	Rule 127 ** B 95, 3 (96.8%) totspend < 4 age >= 4 incgroup < 11 occupation >= 10 province < 6 incgroup >= 10
Rule 117 ** A 18, 1 (94.4%) gender < 2 arrange >= 4 race < 2 totspend >= 4 incgroup < 10	Rule 69 ** A 297, 15(94.9%) purpose < 8 totspend >= 4 race < 2 province >= 6 occupation < 8	Rule 104 ** C 51, 3 (94.1%) dayspent < 15 province < 5 education >= 6 occupation >= 8
Rule 112 ** B 70, 4 (94.3%) occupation >= 12 incgroup < 7 mainarea >= 8	Rule 140 ** A 133, 7 (94.7%) mainarea < 7 race < 2 province >= 6 incgroup >= 10	Rule 95 ** B 124, 8 (93.5%) incgroup < 6 totspend < 4 occupation >= 8

Rule 2 ** B 17, 1 (94.1%) totspend < 3 province >= 3 race < 2 education < 5 incgroup < 9 province < 6 occupation < 8	Rule 25 ** B 139, 10(92.8%) incgroup < 10 purpose < 2 occupation >= 6 totspend < 3 province >= 3 province < 6	Rule 84 ** C 86, 7 (91.9%) purpose < 3 mainarea >= 8 incgroup >= 8 dayspent >= 5 education >= 5 race >= 2 occupation < 8
Rule 17 ** C 90, 8 (91.1%) province < 4 incgroup >= 8 education >= 5 incgroup < 9	Rule 40 ** A 12, 1 (91.7%) purpose < 2 incgroup >= 14 province >= 5 totspend >= 5 race < 2	Rule 57 ** A 237, 22 (90.7%) dayspent >= 5 totspend >= 3 incgroup >= 8 education < 5 race < 2 province >= 6 occupation < 8

Legend for the different variables included in the rule induction modeling:

Number of times visited the Western Cape: 1 Once 2 Twice 3 Three Times 4 Four Times 5 More than 4	**Purpose of visit to the Western Cape** 1 Visit friends or relatives 2 Holiday 3 Business purposes 4 Study purposes 5 Medical treatment 6 Conference 7 Other 8 Sports
Main area visited while in Western Cape 1 West Coast 2 Winelands 3 Breede River 4 Overberg 5 Central Karoo 6 Klein Karoo 7 Garden Route 8 Cape Town and Cape Peninsula	**Time of year of visiting the Western Cape** 1 December/January (Summer) 2 February/April (Autumn) 3 May/August (Winter) 4 September/November(Spring)
Information sources for the Western Cape 1 Travel Agency 2 Tourist Bureau 3 Friends/relations 4 AA 5 Magazines 6 Newspapers 7 Internet 8 Nowhere/myself 9 Previous visits/experience 10 Timeshare 11 Radio/TV 12 Pamphlets/Brochures 13 Church 14 Organised tours 15 Sports Organisation 16 Work/Business 17 School/Tech 18 · Other	**Total spent on visit irrespective of the purpose** 1 R0-1000 2 R1001-2000 3 R2001-3000 4 R3001-4000 5 R4001-5000 6 R5001-6000 7 R6001-7000 8 R7001-8000 9 8001-10,000 10 More than R10,000

Arrangements made for visit to the Western Cape

1	Travel Agent
2	Self/Other in party
3	Your company
4	Family/friends
5	Airline Office
6	Tour Operator
7	Other
8	Church
9	School teacher

Likelihood to visit the Western Cape again

1	Will definitely visit
2	Will probably visit
3	Might visit
4	Will probably not visit
5	Will definitely not visit

Income group

1	Up to 1499
2	1500-1799
3	1800-1999
4	2000-2499
5	2500-2999
6	3000-3999
7	4000-4999
8	5000-6499
9	6500-8499
10	8500-10499
11	10500-12999
12	13000-15999
13	16000-19999
14	20000-24999
15	25000-29999
16	30000-34999
17	35000-39999
18	40000 +
19	Confidential

Occupation

1	Professional
2	Senior Management
3	Middle Management
4	Junior Management
5	Self-employed
6	Clerical / Sales
7	Tradesman / Skilled
8	Semi-skilled
9	Unskilled
10	Housewife
11	Student
12	Pensioner/retired
13	Other
14	Unemployed/not working

Province of origin

1	W Cape
2	E Cape
3	N Cape
4	Free State
5	KZN
6	Northwest province
7	Gauteng
8	Mpumalanga
9	Northern Province

Age grouping

1	18-19
2	20-24
3	25-29
4	30-34
5	35-39
6	40-44
7	45-49
8	50-54
9	55-59
10	60-64
11	65 +
12	Confidential

Level of education

1	Less than Std6
2	Std 6-7
3	Std 8/9/9+Diploma
4	Matric
5	Matric/Diploma
6	University Degree

Gender

1	Male
2	Female

Ethnic group

1	White
2	Black
3	Coloured/Asian

REFERENCES

1 BEN-DAVID, A. & MANDEL, J. (1995) Classification Accuracy, Machine Learning vs. Explicit Knowledge Acquisition, *Machine Learning*, 18: 109-114.

2 BERSON, A., SMITH, S. & THEARLING, K. (2000) *Building Data Mining Applications for CRM*, New York: McGraw-Hill.

3 BLUM, A. (1992) *Neural Networks in C++: An Object-Orientated Framework for Building Connectionist Systems,* New York: Wiley.

4 BREIMAN, L. (1996) Technical Note: Some Properties of Splitting Criteria. *Machine Learning*, 24: 41-47

5 BRODLEY, C.E. & UTGOFF, P.E. (1995). Multivariate Decision Trees, *Machine Learning*, 19, 45-77.

6 CRUSADER SYSTEMS (1998a) *Basic Modelgen, Version 1.6.* Pretoria.

7 CRUSADER SYSTEMS (1998b) *Basic Modelgen Users Manual*: Pretoria.

8 DEBOECK, G.J. (1995) *Trading on the Edge: Neural, Genetic and Fuzzy Systems for Chaotic Financial Markets,* New York: Wiley.

9 DIBB, S. & SIMKIN, L. (1991) Targeting segments and positioning. *International Journal of Retail and Distribution Management.* 19(3): 4-10.

10 GOONATILAKE, S. (1995) Intelligent Systems for Finance and Business: An Overview, In S. Goonatilake and P. Treleaven, *Intelligent Systems for Finance and Business*: 1-28, New York: Wiley.

11 GRAY, N.A.B. (1990) Capturing Knowledge through Top-Down Induction of Decision Trees, *IEEE Expert*: 41-51.

12 KAASTRA, I., & BOYD, M. (1996) Designing a Neural Network for Forecasting Financial and Economic Time Series, *Neurocomputing*, 10: 215-36.

13 KAMBER, M., WINSTONE, L., GONG, W., CHENG, S, & HAN, J. (1997) "Generalization and Decision Tree Induction: Efficient Classification in Data Mining", Paper presented at the International Workshop on Research Issues on Data Engineering, Birmingham, England: 1-25.

14 KLIMASAUSKAS, C.C. (1996) "Applying Neural Networks", In R.R. Trippi and E. Turban, *Neural Networks in Finance and Investment*: 45-69, New York: McGraw-Hill.

15 KOHONEN, T. (1988) *Self-Organisation and Associative Memory,* New York: Springer.

16 KOTLER, P. (2000) *Marketing Management*, Englewood Cliffs: Prentice-Hall.

17 LIPPMAN, R.P. (1989) An Introduction to Computing with Neural Nets. *IEEE ASSP Magazine*: 4-22.

18 MASTERS, T. (1993) *Practical Neural Network Recipes in C++*. New York: Academic Press.

19 MCCORD-NELSON, M. & ILLINGWORTH, W.T. (1993) *A Practical Guide to Neural Nets,* New York: Academic Press.

20 MITCHELL, V. (1994) "How to Identify Psychographic Segments: Part 1", *Marketing Intelligence and Planning,* 12(7), 4-10.

21 PATTIE, D.C. & SNYDER, J. (1996) "Using a Neural Network to Forecast Visitor Behaviour", *Annals of Tourism Research*. 23(1): 151-64.

22 QUINLAN, J.R. (1993) *Programmes for Machine Learning,* San Mateo: Morgan Kaufmann.

23 RYMAN-TUBB. N. (1993) "The Use of Neural Networks to Identify the Characteristics of Holiday Markers", *The Journal of Database Marketing.* 1(2): 140-49.

24 SAARENVIRTA, G. (1998) "Mining Customer Data" *DB2 Magazine.* 3(3): 10-20.

25 STATSOFT (1998) *Statistica Version 5.5*. Tulsa, OK.

26 STATSOFT (2001) *Electronic Text Book - www.statsoft.com*, Tulsa. OK.

27 CREATIVE RESEARCH SYSTEMS (2001) *Survey Systems, Version 8.0*. Petaluma, CA.

28 TURBAN E. & TRIPPI, R.R. (1996) Neural Network Fundamentals for Financial Analysts, In R.R. Trippi and E. Turban, *Neural Networks in Finance and Investment*: 3-24, New York: McGraw-Hill.

29 VENUGOPAL, V. & BAETS, W. (1994) "Neural Networks and Statistical Techniques in Marketing Research: A Conceptual Comparison" *Marketing Intelligence and Planning,* 12(7), 30-38.