Market Segmentation of Inbound Business Tourists to Thailand by Binding of Unsupervised and Supervised Learning Techniques

Anongnart Srivihok
Department of Computer Science, Faculty of Science
Kasetsart University, Bangkok, Thailand
Email: fsciang@ku.ac.th

Wirot Yotsawat
Department of Computer Science, Faculty of Science
Kasetsart University, Bangkok, Thailand
Email: g5314401258@ku.ac.th

***Abstract*— Market segmentation is an important tool, for driving an organization to achieve its goals. This study proposes a market segmentation technique with the binding of unsupervised and supervised learning techniques. The method aims to cluster international tourists who arrived in Thailand for business proposes, and to classify business tourists by using the products of an unsupervised learning technique as class labels. A Self-Organizing Map (SOM), K-Means and Hierarchical clustering were applied to find the best quality of segmentation guided by the computation of the Silhouette index. Segment labels were used to supervise the learning part as class labels. Multilayer Perceptron (MLP), J48 decision tree, Decision Table, OneR and Naïve Bayes classifiers were used to classify the business tourist data set, and the best performance technique was preferred. The experimental results designated that K-Means outperformed the other clustering techniques and provided five different segments. Moreover, the Naïve Bayes classifier gave the best performance among the other classifiers based on the business tourist variables. Thus, this model can be used to predict the segment of new arrival business tourists.**

***Index Terms*—market segmentation, tourism, K-Means, unsupervised learning, supervise learning, Naïve Bayes**

I. Introduction

Each year many inbound tourists travel to Thailand for business purposes such as exhibitions, conferences and meetings. The business tourists are important because they may bring investment from their country to Thailand. Even though we don't know their type of business, the quantity of investment, the profit or loss rate or when the businesses will operate, they produce an effect on demand and supply in Thailand. Moreover, in the future, these business people are likely to come back again for other business reasons such as following up on their businesses and expansion growth. Consequently, associated organizations must have better policies and planning in order to attract and maintain this tourist market. They have to understand the characteristics of the tourists. Market segmentation techniques can be used to produce this knowledge for the organization.

Past research focused on inbound tourist market segmentation [1] and proposed data mining of tourists by using two step clustering and classification. The research found that the K-Means technique gave higher quality information than SOM and Fuzzy C-Means (FCM) for tourist partitioning based on international tourists to Thailand. Moreover, MLP can predict the characteristics of new tourists as part of the production from clustering technique. The results of this research indicated that each tourist segment was different in terms of economics. However, the mean value of daily expenditure of business tourist was higher than nonbusiness tourists. These expenses were computed from the data on overall tourists provided by the Department of Tourism in the Ministry of Tourism and Sports, Thailand [2]. In the current study, we focus on inbound tourists who come to Thailand for business proposes.

The primary objectives of this study are (1) to compare the performances of K-Means, SOM neural network and Hierarchical clustering techniques in order to segment business tourists and (2) to compare the performance of classifiers namely, Decision Tree, Decision Table, OneR, MLP and Naïve Bayes, in order to predict the segment of new business tourists as part of the production from the clustering technique.

The remainder of this paper is organized as follows. Section 2 provides a literature review on market segmentation. The related algorithms of unsupervised and supervised learning and the measurements are presented in Section 3. Section 4 details the experimental design and its results. Finally, the conclusion and future work are in Section 5.

II. Market Segmentation

Market segmentation is a methodological process of dividing markets which are comprised of individuals into

doi:10.4304/jsw.9.5.1334-1341

smaller groups with homogenous characteristics within each segment and heterogeneity between segments, based on an identified set of attributes [3]. There are many techniques for market segmentation. However, the most important technique for identify segments is clustering. Boone et al. [4] suggested that there are over 50 clustering techniques which can be applied for market segmentation. K-Means and Hierarchical clustering (Ward method) are the most popular in market segmentation as suggested by S. Dolnicar [5]. Moreover, a SOM neural network can be applied for market segmentation and several different fields [6]. Some researchers proposed market segmentation techniques with combined algorithms. R.J. Kuo et al. [7] proposed the integration of SOM and K-Means for market segmentation. They found that their proposed technique provided slightly better partitioning than the conventional methods such as the integration of hierarchical and K-Means or only SOM. However, no clustering techniques were suitable for all data.

Some researchers have worked on tourism market segmentation such as J. Z. Bloom [8] who proposed tourist market segmentation using a SOM neural network for partitioning the tourist data set. Also, a BP neural network was used for predicting the segmentation of new tourists as part of already existing segments. M. Najmi et al. [9] proposed the segmentation of inbound tourism market in Iran with a concentration on culture. In the first stage, they used agglomerative hierarchical clustering to extract the core segments. Subsequently, common sense segmentation was performed to obtain the final segments. Moreover, J. Chang [10] proposed the segmentation of tourists who travelled in the Rukai tribal area, Taiwan. The research used Hierarchical clustering (Ward method) to partition the data and used K-Means to obtain the final segments.

Some tourism market segmentation research has focused on a specific market such as J. Kim et al. [11] who considered senior Australian tourists who were older than 49 years. They used SOM to partition the data set into four segments. E. M. Garcia et al. [12] focused on tourists in Spain who flew on low-cost airlines to Girona Airport, Spain. They used two step clustering which was proposed by M. Wedel et al. [13]. In the first step, Hierarchical clustering (Ward method) was performed to partition the data, then K-Means was applied to obtain the final segments. H. L. T. Trang [14] proposed an inbound tourism market segmentation in Thailand. The researcher focused on the tourists of the Andaman cluster (Phuket, Phang-Nga and Krabi), Thailand. Hierarchical clustering (ward method) was used to partition the data set. Then, K-Means was applied to obtain the final segments. Moreover, J. G. Brida et al. [15] used the integration of SOM and K-Means, proposed by J. Vesanto et al. [16], to partition tourism data. They focused on three different Christmas markets in Northern Italy.

In the current study, we focused on the segmentation of inbound business tourists to Thailand. We applied SOM, K-Means and Hierarchical clustering in order to partition the data set. The best quality technique was selected. Then, classification techniques were applied to predict the segments of new tourists as part of the production from clustering technique. The best performance classifier was selected.

III. Related Algorithms

A. Unsupervised Learning

In this study, we obtain the segments of the business tourist market by using the clustering method. K-Means clustering is a simple partition method. It is the most popular method in cluster analysis [17] and was proposed by James MacQueen in 1967. The basic concept of K-Means is the partitioning of n data points into k clusters by the nearest mean. Thus, the important input requirement of K-Means is the number of clusters. Some important problems of K-Means are the optimum number of clusters and the outliers of input data. To avoid the problems of K-Means, we find the optimum number of k by computation of the Silhouette coefficient and preprocess the data set. Subsequently, outliers and missing values are handled before input to the K-Means method. The basic K-Means algorithm [17] can be described as follows.

1. Select k points as initial centroids.
2. Repeat
3. from k clusters by assigning each point to its closest centroid.
4. Recompute the centroid of each cluster.
5. until centroids do not change.

Hierarchical clustering techniques are a second important type of clustering method [17]. In the current study, Agglomerative Hierarchical clustering is applied to segment the business tourist data set. Agglomerative Hierarchical clustering starts with each point as a cluster. The closest pairs of clusters are merged until only one cluster remains. The proximity measurement between clusters is computed by different methods. In our case, four most of the commonly used methods, namely the Single, Complete, Average and Ward methods, are used to measure the proximity candidate. The basic Agglomerative Hierarchical clustering technique [17] can be described as follows.

1. Compute the proximity matrix, if necessary.
2. Merge the closest two clusters.
3. Update the proximity matrix to reflect the proximity between the new cluster and the original clusters.
4. Repeat step 2 and 3 until only one cluster remains.

O. Kaski et al. [6] suggested that an SOM could be used in different fields including market segmentation. The SOM algorithm was proposed by Kohonen in 1982 [18]. It uses a feed-forward neural network algorithm which consists of input and output layers. The goal of SOM is to find a set of centroids and to assign each point in the data set to the centroid that provides the best approximation of that point [17]. The SOM procedure consists of the following stages.

1. Randomly initialize all weights.
2. Select the input vector.
3. Calculate the distance between input vectors and all weights to find the closest output node.
4. Define a neighborhood function that allows the identification of the output node to be updated in the next step.
5. Update the winner's weight.
6. Repeat steps 2 to 5 until the algorithm converges.

In this study, we assessed the quality of clusters by comparing their Silhouette indices [19]. The Silhouette width is a composite index reflecting the compactness and separation of the clusters. The value of the Silhouette index lies in the range of [-1, 1]. The best quality of clustering has a value near 1; on the other hand, poorly clustered data have a value near -1. The computation of the Silhouette index is the average dissimilarity within its own cluster (a_i) compared to dissimilarity in the nearest neighboring cluster (b_i) as follows.

$$a_i = \frac{1}{|C_j|}\sum_{r\in c_j} d_p(a_i, x) \qquad b_i = \frac{1}{|C_j|}\sum_{r\in c_h}^{h\neq j} d_p(a_i, x)$$

$$s_i = \frac{b_i - a_i}{\max(b_i - a_i)} \tag{1}$$

where d_p is the distance of vector i to another vector in the same cluster. Further, x is an interest vector. C_j is j^{th} cluster where $j \in \{1,..., k\}$ from the overall k cluster and C_h is the h^{th} cluster where $h \in \{1,..., k\}$ from the overall k cluster and $h \neq j$.

The best quality of segmentation and the best performance technique is indicated by the maximum value of the average Silhouette index calculated as follows.

$$\bar{s} = \frac{1}{K}\sum_{k=2}^{K}\frac{1}{n_k}\sum_{i=1}^{n_k} s(i,k) \tag{2}$$

B. Supervised Learning

The market segmentation results can be used for planning market strategies. This study uses the segment label, which each point belong, as class labels in the supervised learning phase. The classification techniques used and tested in this work were J48 Decision Tree, MLP, Naïve Bayes, Decision Table and OneR. J48 Decision Tree is widely used in data classification. It is modified version of C4.5 and implemented in WEKA data mining software [20]. We used the decision tree to predict the segments of business tourists based on the input variables. The decision tree is formed by splitting the source set into subsets based on an attribute value test and is completed when the subset at a node has all the same values of the target variable. The decision tree model consists of one root, a number of branches, a number of nodes and a number of leaves. Each node involves one attribute, each branch node represents a choice between a number of alternatives, and each leaf node represents a class. Further computation of this method can be accessed in P.N. Tan et al. [17].

Multilayer Perceptron (MLP) is a modification of the standard Perceptron. It is a popular form of the feed-forward artificial neural network model. In our case, it can be applied for data classification by assigning segment labels to represent each class. The MLP model consists of three or more layers (an input layer, one or more hidden layers, and an output layer). Each node in the MLP model is connected by weights and output signals which are a function of the sum of the inputs to the node modified by an activation function. The MLP classifier is implemented on WEKA with training by the back-propagation learning algorithm which is summarized below [21].

1. Randomly initialize all weights.
2. Present the first input training vector to the network.
3. Propagate the input vector through the network to obtain an output.
4. Calculate an error signal by comparing actual output to the target output.
5. Back-propagate the error signal through the network.
6. Adjust weights to minimize the overall error.
7. Repeat steps 2-7 with the next input vector until the overall error is satisfactorily small.

The Naïve Bayes algorithm is a classification algorithm base on Bayes theorem. The Naïve Bayes Classifier gives high accuracy and speed when applied to the large data bases. It assumes the conditional independence of attributes given a class. The conditional independence assumption can be illustrated as follows.

$$P(a_1, a_2, ..., a_n \mid class) = \prod_{i=1}^{n} P(a_i \mid class) \tag{3}$$

$$P(a_1, a_2, ..., a_n \mid class) = P(a_1 \mid class)P(a_2 \mid class)...P(a_n \mid class) \tag{4}$$

Further calculations of the Naïve Bayes algorithm are available in P.N. Tan et al.[17].

The decision table is one type of classifier that is implemented on WEKA. It consists of a set of features that is included in the table, and class labels which defined by the features. Creating a decision table might involve selecting some of the attributes. It searches the optimum feature subsets using the best-first search and uses cross-validation for evaluation. Further detail on the decision table can be accessed in Ron Kohavi's research [22].

OneR or the 1-Rule is a very simple classification rule from a set of data. It generates a set of rules through all attribute tests. Thus, each attribute will provide a different set of rules, with one rule for every value of the attribute. The error rate of each attribute rule is evaluated. OneR will choose the attribute that produces rules with the least error rate. The algorithm of OneR [21] can be described as follows.

1. Select an attribute input.
2. Build a rule for each value of the attribute by
   2.1 Counting the frequent of each class appeared.

2.2 Finding the most frequent class.
2.3 Making the rule assign that class to this attribute value.
3. Calculate the error rate of the rules.
4. Repeat steps 1-3 until coverage of all attribute-values.
5. Choose the rules with the least error rate.

To evaluate the performance of classifiers, in our case, we divide the business tourist data set into ten independent subsets (K=10) of size n/10. Nine subsets were trained by the classification technique and one subset was used for the test set. Training was repeated and tested ten times until every subset was a test set. Accuracy of prediction was defined as the mean value over all the ten process times. The process is called K folds cross validation [17]. The performance of classifiers is measured by the values of measurement as follows.

$$\text{Recall} = (TP) / (TP+FN) \quad (5)$$

$$\text{Precision} = (TP) / (TP+FP) \quad (6)$$

$$\text{FP Rate} = (FP) / (FP+TN) \quad (7)$$

$$\text{F Measure} = \frac{2 \times (\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (8)$$

where TP is the number of True Positive instances, TN is the number of True Negative instances, FP is the number of False Positives instances and FN is the number of False Negative instances.

IV. Experimental Design and Results

A. Data Set

The data set in this paper consists of 5,926 tourists who visited Thailand for business proposes and was selected from 72,413 tourist records for people who travelled to Thailand for many purposes during the period 2008 to 2010. The overall tourist data set was obtained from the Department of Tourism in the Ministry of Tourism and Sports, Thailand [2]. The attributes of the data are the length of stay (days), age, annual incomes (US Dollars), average expenditure (Baht per day), type of accommodation, occupations, tourist origin and place of residence. The characteristics of the business tourists are presented in Appendix A.

B. Study Framework

The specification of this study required the use of unsupervised and supervised learning algorithms to partition the data into segments and to predict the segments of unseen data, respectively. Thus, the research design is divided into two phases namely an unsupervised learning phase and a supervised learning phase. The unsupervised learning phase involves partitioning by the best clustering technique, while the supervised learning phase involves using segments to predict the cluster of new tourists by the best classifier. The study framework can be explained as follows.

1. A data preprocessing step is crucial. In this step, we removed the outliers, missing values and unreliable data. Some attribute values were transformed in order to qualify for the requirements of the algorithms. Thus, the values of each element were normalized in the range [0, 1] and then de-normalized after the processing was completed.

2. After the data had been preprocessed, the unsupervised learning phase began. We used SOM, K-Means and four of the most common hierarchical algorithms (Single, Complete, Average and Ward) to partition the business tourist data set. The Silhouette index was calculated to assess all cluster partitions from 2 to 10. The best technique was selected.

3. After the unsupervised learning phase was finished, we analyzed the characteristics of each cluster according to the results from the unsupervised learning phase, and predefined the cluster labels of tourists for the supervised learning phase.

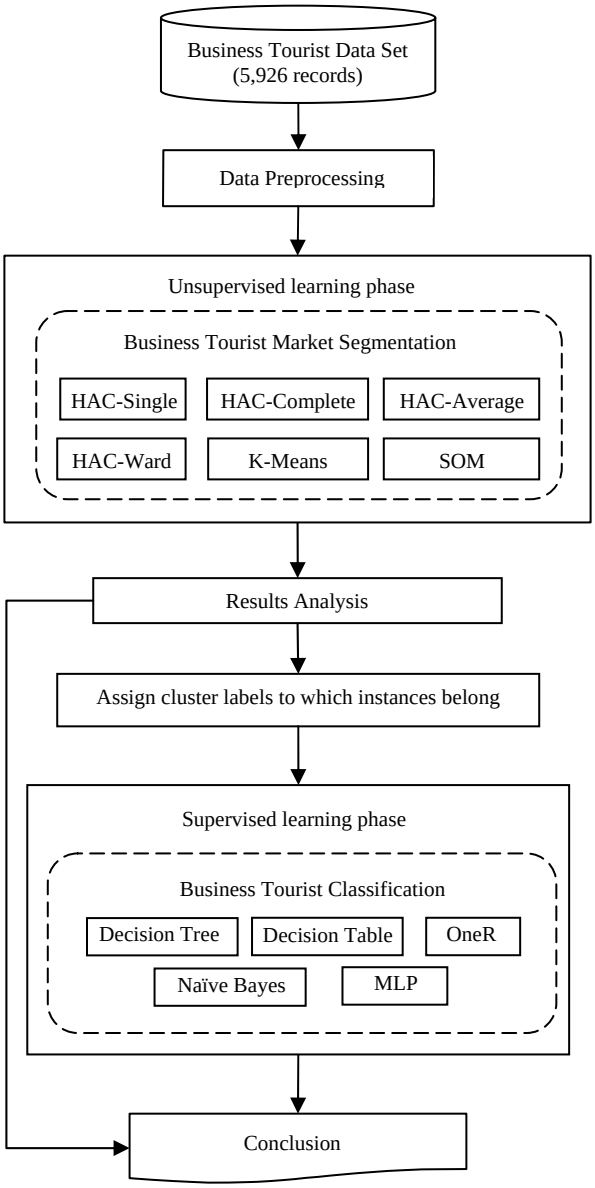

Fig. 1 Study framework

4. The tourists were assigned to segment labels according to the unsupervised learning results. The supervised learning phase was begun, using the four candidates (J48 Decision Tree, MLP, Decision tree, One

R and Naïve Bayes) to classify the business tourist data set. The performance of classifiers was compared to find the best technique. The study framework is illustrated in Fig. 1.

C. Unsupervised Learning Results

In this study, K-Means, a popular technique, was performed to partition the business data compared with the most commonly used hierarchical algorithms (Single, Complete, Average and Ward) and a SOM neural network. The optimum number of clusters was compared and is shown in Fig. 2

In general, the value of the Silhouette index, used to validate the quality of segmentation, must be greater than 0.65 [23]. In this study, the Silhouette index provided a lower value than this threshold (0.65) because it was calculated on the normalized value of the data. As we can see from Fig. 2, K-Means outperformed both SOM and the four most common methods of Hierarchical algorithm. Consequently, we chose the best partition obtained from the K-Mean at five clusters. The percentage of business tourist per cluster is illustrated in Table 1.

TABLE I
PERCENTAGE OF BUSINESS TOURISTS PER CLUSTER

Cluster	Number of tourists per cluster	Percentage contribution
1	1,156	19.51
2	957	16.15
3	1,288	21.73
4	426	7.19
5	2,099	35.42
Total	5,926	100.00

Table 1 indicates that segment 5 is relatively dominant among the 5 clusters. Cluster 5 is the biggest cluster. It comprises over 35% (n=2,099) of overall tourists (N=5,926) and can be considered as a homogeneous cluster. Cluster 4 is the smallest cluster and consists of 7.19% (n=426) which has a similar profile of features. On the whole, business tourists in all segments select hotels as their accommodations. Their business is mainly in Bangkok. There was no significant difference in the business tourist age data. The characteristics of the business tourist market segment are summarized as follows.

Cluster 1 consists of 19.51% of overall business tourists. This segment belonged to business tourists who have an occupation as clerical, sales or commercial. The business tourists in this cluster have the smallest mean age value. They mainly come from Southeast Asia and East Asia with 34.43% and 25.43%, respectively. Moreover, they stayed in Thailand for a short period (less than 10 days). The annual income of tourists in this cluster is less than 40,000 US Dollars per year (26.21% less than 20,000 and 39.88% less than 40,000 US Dollars per year).

Cluster 2 consists of 16.15% of overall business tourists. The tourists in this cluster stay in Thailand for a short period (less than 10 days). They have variety of occupations and come from a variety of countries. The annual income of this cluster is the least (42% less than 20,000 US Dollars per year) when comparing to the other segments.

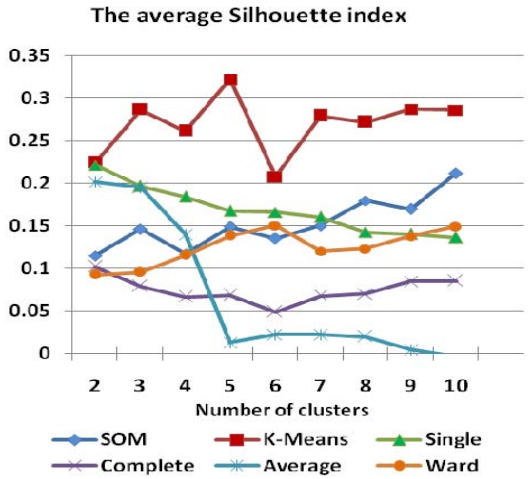

Fig. 2 Average Silhouette index provided by unsupervised learning partition from 2 to 10 clusters.

Cluster 3 consists of 21.73% of overall business tourists. The tourists in this cluster are administrative or managerial. They stay in Thailand for the shortest period (less than 5 days) compared to other segments. They mainly come from Southeast Asia and East Asia (32.30% and 30.43%, respectively). The annual income of tourists in this cluster is in 20,001 – 40,000 US Dollars per year (35.64%) and 40,001 – 60,000 US Dollars, respectively.

Cluster 4 is the smallest cluster. It consists of only 7.19% of overall business tourists. The tourists in this cluster have the highest annual income by mean (40,001-60,000 US Dollars) which a variety of groups. They spend highest daily expenditure (4,000 – 6,000 Bath) by mean. All of the business tourists in this cluster are Europeans. Tourists in this cluster are mainly professionals (74.88%).

Cluster 5 is the biggest cluster. It consists of 35.42% of overall business tourists. The tourists in this cluster mainly come from Southeast Asia (41.26%) and 99.95% of tourists in this cluster are professionals. They stay in Thailand for less than 10 days.

D. Supervised Learning Results

In the supervised learning phase, the data classification is controlled by a supervised learning technique. It is used to predict the segment of new business tourists when the segments are produced. In the classification phase, we used WEKA [20] version 3.6 to learn from the training set. WEKA was employed on a Core 2 Duo processor with 4 GB RAM. We compared the performance of the J48 Decision Tree, Decision Table, OneR, MLP and Naïve Bayes classifiers. The class label of data was assigned according to which cluster the tourist belonged as provided by the results of the clustering phase. The correct classification or misclassification is determined by how well the training set can learn the pattern in the

data when compared to the test set which comprises unseen data or new arrival tourist data.

The performance of each classifier is suggested by the computation of Precision, Recall, F-Measure and FP-Rate as shown in Table. 2.

TABLE II
COMPARING THE PERFORMANCES OF CLASSIFIERS

Classifiers	Precision	Recall	F-Measure	FP-Rate
J48	0.985	0.985	0.984	0.003
Naïve Bayes	**0.990**	**0.990**	**0.990**	**0.002**
OneR	0.864	0.928	0.894	0.033
Decision Table	0.981	0.981	0.981	0.006
MLP	0.957	0.958	0.957	0.011

The experimental results indicated that the Naïve Bayes classifier gives the highest performance in terms of all four measurements. It gives slightly better performance than the J48 decision tree and Decision Table, respectively. It is clear that Naïve Bayes gives better results than MLP. Moreover, OneR produces a rule by choosing the occupation of tourists but it provides the least performance classification when compared to the other four techniques. Thus, Naïve Bayes can be used to predict the cluster of new business tourists as part of the five segments which are provided by K-Means. The confusion matrix of the Naïve Bayes classifier, based on business tourist variables, is presented in Table 3.

TABLE III
THE CONFUSION MATRIX OF NAÏVE BAYES CLASSIFIER

	Class1	Class2	Class3	Class4	Class5	Total
Cluster1	1,146	0	0	10	0	1,156
Cluster2	0	947	0	10	0	957
Cluster3	0	0	1,286	2	0	1,288
Cluster4	1	14	5	398	8	426
Cluster5	0	1	0	6	2,092	2,099
Total	1,147	962	1,291	426	2,100	5,926
Accuracy	0.998	0.996	0.999	0.991	0.997	
Weigh accuracy		0.997				

V. CONCLUSION

Market segmentation is an important tool, used for dividing markets, which are comprised of individuals, into smaller groups with homogenous characteristics within each segment, and heterogeneity between segments, based on an identified set of attributes [3]. This study proposes a market segmentation method for foreign tourists who come to Thailand for business purposes. The method consists of the evaluation of unsupervised learning techniques by computing the value of the average Silhouette index and comparing the performance of supervised learning techniques. The unsupervised learning techniques are compared using K-Means, SOM neural network and Hierarchical (single, complete, average and ward method) clustering. The experimental results indicated that K-Means outperforms all of SOM and the four most common Hierarchical clustering methods based on the business tourist's variables. Each market segment has different characteristics. However, some attributes are not significant such as age of the tourists, while some attributes influence each segment such as occupation. Thus, the associated organizations can use the segmentation method and the results to define their marketing plan or in other applications.

Moreover, the supervised techniques can be used to predict a segment of new business tourist arrivals. The classes of data are the segments of tourist that are provided by the unsupervised learning method. The supervised learning techniques consist of J48Decision Tree, Decision Table, OneR, MLP and Naïve Bayes classifiers. The experimental results indicated that the Naïve Bayes performance better than the other techniques. Although MLP, Decision Tree and Decision Table each give a good classification, they are not as good Naïve Bayes. Moreover, OneR gives the poorest performance of all the techniques assessed. Thus, the associated organization can use Naïve Bayes to predict the segments of new business tourists as a part of the production from clustering technique.

Future studies on the marketing segmentation of tourism can focus on other interesting markets and employ other related features such as cultural and socio-economic variables. Some attributes can be summarized to narrow width values which occur and can add some related information to the data. Moreover, rules mining in the tourism data is challenging.

ACKNOWLEDGMENT

This work was supported by the Department of Computer Science, Faculty of Science, Kasetsart University.

REFERENCES

[1] W. Yotsawat and A. Srivihok, "Data mining of international tourist in Thailand by two step clustering and classification", *Advanced Science Letters*, vol.20, pp. 245-249, 2014, doi: 10.1166/asl.2014.5299.

[2] Department of Tourism, Ministry of Tourism and Sports, Thailand, http://www.tourism.go.th.

[3] A. Kara and E. Kaynak, "Markets of a single customer: exploiting conceptual developments in market segmentation", *European Journal of Marketing*, vol. 31, pp. 873-95, 1997, doi: 10.1108/03090569710190587.

[4] D. Boone and M. Roehm, "Retail segmentation using artificial neural networks", *International Journal of Research in Marketing*, vol. 19, pp. 287–310, September 2002, doi: 10.1016/S0167-8116(02)00080-0.

[5] S. Dolnicar, "Using cluster analysis for market segmentation typical misconceptions, established methodological weaknesses and some recommendations for improvement", *Australasian Journal of Market Research*, vol. 11, pp. 5-12, 2003.

[6] M. Oja, S. Kaski and T. Kohonen, "Bibliography of self-organizing map (SOM) papers: 1998–2001 addendum", *Neural Computing Surveys*, vol. 3, pp. 1–156, 2003.

[7] R.J. Kuo, L.M. Ho and C.M. Hu, "Integration of self-organizing feature map and K-means algorithm for market segmentation", *Computers and Operations Research*, vol. 29, pp. 1475-1493, September 2002, doi: 10.1016/S0305-0548(01)00043-0.

[8] J. Z. Bloom, "Tourist market segmentation with linear and non-linear techniques", *Tourism Management*, vol. 25, pp. 723–733, December 2004, doi: 10.1016/j.tourman.2003. 07.004.

[9] M. Najmi, A. Sharbatoghlie and A. Jafarieh, "Tourism Market Segmentation in Iran", *International Journal of Tourism Research*, vol. 12, pp. 497–509, January 2010, doi: 10.1002/jtr.768.

[10] J. Chang, "Segmenting tourists to aboriginal cultural festivals: An example in the Rukai tribal area, Taiwan", *Tourism Management*, vol. 27, pp. 1224–1234, December 2006, doi: 10.1016/j.tourman.2005.05.019.

[11] J. Kim, S. Wei and H. Ruys, "Segmenting the market of West Australian senior tourists using an artificial neural network", *Tourism Management*, vol. 24, pp. 25–34, 2003, doi: 10.1016/S0261-5177(02)00050-X.

[12] E. M. Garcia and M. R. Vela, "Segmentation of low-cost flights users at secondary airports", *Journal of Air Transport Management*, vol. 16, pp. 234–237, July 2010, doi: 10.1016/j.jairtraman.2010.01.003.

[13] M. Wedel and W.A. Kamakura, "Market Segmentation: Conceptual and Methodological Foundations", 2nd ed., Kluwer Academic Publishers, Norwell, 2000.

[14] H. L. T. Trang and P. Kullada, "Inbound tourism market segmentation of the Andaman cluster, Thailand", M.S. thesis, Prince of Songkla Univ., Songkla, Thailand, 2009.

[15] J. G. Brida, M. Disegna and L. Osti, "Segmenting visitors of cultural events by motivation: A sequential non-linear clustering analysis of Italian Christmas market visitors", *Expert Systems with Applications*, vol. 39, pp. 11349–11356, October 2012, doi: 10.1016/j.eswa.2012.03.041.

[16] J. Vesanto, and E. Alhoniemi, "Clustering of the self-organizing map". *IEEE Transactions on Neural Networks*, vol. 11, pp. 586–600, 2000, doi: 10.1109/72.846731.

[17] P.N. Tan, M. Steinbach and V. Kumar, *Introduction to Data Mining*, Pearson Education, Inc., USA., 2006.

[18] T. Kohonen, *Self-Organizing Maps*, Springer Inc., Berlin, 1995.

[19] V. Lucas, J. G. B. C. Ricardo and R. H.Eduardo, "On the Comparison of Relative Clustering Validity Criteria", *Statistical Analysis and Data Mining*, vol. 3, pp. 209–235, August 2010.

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

[21] Ian H. Witten and Eibe Frank, *Data mining: practical machine learning tool and techniques*, 2nd ed., Morgan Kaufmann Publishers, California, USA, 2005.

[22] R. Kohavi, "The power of decision tables", *Lecture Notes in Computer Science*, vol. 912, pp. 174-189, 1995, doi: 10.1007/3-540-59286-5_57.

[23] W. Liu, X. Di, G. Yang, H. Matsuzaki, J. Huang, R. Mei, et al., "Algorithms for large-scale genotyping microarrays", *Bioinformatics*, vol. 19, pp. 2397–2403, doi: 10.1093/bioinformatics/btg332.

Anongnart Srivihok is an associate professor at the Department of Computer Science, Faculty of Science, Kasetsart University, Bangkok, Thailand. She has a doctorate degree in Information Systems from Central Queensland University, Australia. Her research areas include data mining, knowledge management, decision support systems and intellectual capital.

Wirot Yotsawat received a B.S. in Computer Science in 2007 from Walailak University, Thailand. He is studying at the Department of Computer Science, Kasetsart University, Thailand. His research interests are data classification, data clustering and other related on data mining aspects.

APPENDIX A DESCRIPTIVE STATISTICS OF BUSINESS TOURIST ATTRIBUTES

Variables	Sample size	Percentage	Variables	Sample size	Percentage
Length of stay (day)			**Origin**		
1-5	3,723	62.82	Africa	182	3.07
6-10	1,477	24.92	America	316	5.33
11-15	376	6.34	East Asia	1,439	24.28
16-20	139	2.35	Europe	681	11.49
21 and above	211	3.56	Middle East	314	5.30
Age (years)			Oceania	223	3.76
Less than 15	9	0.15	South Asia	808	13.63
15-24	222	3.75	South East Asia	1,963	33.13
25-34	1,621	27.35	**Annual income (US dollars)**		
35-44	2,252	38.00	Less than 20,000	1,320	22.27
45-54	1,480	24.97	20,000-39,999	1,898	32.03
55-64	305	5.15	40,000-59,999	1,299	21.92
65 and above	37	0.62	60,000-79,999	636	10.73
			80,000 and above	773	13.04

APPENDIX A (CONT) DESCRIPTIVE STATISTICS OF BUSINESS TOURIST ATTRIBUTES

Variables	Sample size	Percentage	Variables	Sample size	Percentage
Accommodation			**Place of residence**		
Service apartment	121	2.04	Bangkok	4,543	76.66
Guest house	238	4.02	Chonburi	409	6.90
Hotel	5,169	87.23	Phuket	203	3.43
Sanitarium/Nursing home	196	3.31	Surat Thani	86	1.45
Resort	200	3.37	ChiangMai	195	3.29
Average expenditure (Baht per day)			Krabi	70	1.18
Less than 2,000	549	9.26	East (without Chonburi)	78	1.32
2,000 – 4,000	2,137	36.06	Central (without Bangkok)	87	1.47
4,000 – 6,000	1,704	28.75	West	47	0.79
6,000 – 8,000	819	13.82	South (without Phuket, Krabi and Surat Thani)	96	1.62
More than 8,000	717	12.10	North (without ChiangMai)	46	0.78
Occupation			Northeast	66	1.11
Professional	2,417	40.79			
Administrative and Managerial	1,301	21.95			
Clerical, Salesmen and Commercial Person	1,173	19.79			
Student and Children	40	0.67			
Housewife, Retired and Unemployed	91	1.54			
Laborer	335	5.65			
Other	569	9.60			