

“© 2013 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.”

A hybrid model for migrating customer segmentation with missing attributes

Jun Ma

SMART Infrastructure Facility
University of Wollongong
Wollongong, NSW 2522, Australia
j.ma@uow.edu.au

Hua Lin, Jie Lu, Guangquan Zhang

Centre for Quantum Computation and Information Systems
School of Software, Faculty of Engineering and IT
University of Technology, Sydney (UTS)
Sydney, NSW 2007, Australia
{Jie.lu, guangquan.zhang}@uts.edu.au

Abstract - Due to missing attributes in an enterprise's database, migrating customer segmentation results from external dataset to enterprise database is difficult. In this paper, a hybrid model, called HMCS model, is presented. This model artificially generates values of missing attributes based on external dataset and populates them to enterprise database. Based on this model, an application in a telecom application is reported. Application indicates the presented model can produce acceptable segmentation results on the enterprise dataset which is with missing attributes.

I. INTRODUCTION

Migrating customer segmentation results obtained on an enterprise's external data sources to enterprise data is a challenging topic in enterprise customer relationship management. Customer segmentation can be conducted on an enterprise's internal and external data sources such as samples from enterprise historical customer transactions in database or data warehouse [13], demographic and socioeconomic data from government agencies, and customer survey [1][10][16]. Internal data sources are golden mines of customer profiles and behaviour patterns; however, they are often missing some important features for effective customer segmentation. For example, a telecom can hold information about a customer's contract, billing history, as well as handset models in its enterprise database; but it can hardly store a customer's professional background and industrial sectors which contribute more information for a better customer segmentation. Hence, an enterprise often needs to collect customer-related information from external data sources, conduct customer segmentation on it, and then migrate the segmentation result to its internal data sources. However, it is quite common that customer-related attributes in external data sources seldom matching those in internal data. This results in difficulties when applying segmentation results obtained on external data sources to enterprise data. Therefore, migrating segmentation results on external data sources to enterprise data is a necessary solution of practical significance.

Migrating customer segmentation results is of particular importance in enterprise application. Existing segmentation results are generally built on enterprise internal data and external data separately. Both of them are gold mines for an enterprise. Since customer segmentation is a costly business process, which cannot be replicated often, reuse existing segmentation results is a realistic choice; however, it is by no

means an easy task. Enterprise internal data is often objective information such as a customer's contract terms, billing history, and spending amount; but, it cannot indicate a customer's subjective expectation or preference changes. To obtain a customer's subjective expectation and preference, customer survey is widely used and customer segmentation is often conducted on the survey data. If participants in a customer survey can be identified in an enterprise's internal data, it becomes easier to migrate customer segmentation result on survey data to enterprise data. However, this is prohibited by privacy law or privacy policy in most situations. Hence, the directed link between external and internal data sources cannot be built easily. Furthermore, even if the link exists, the migration still needs to solve the inconsistent attributes between two data sources. It is very common that a customer segmentation on survey data uses attributes which are not in an enterprise's internal data. Due to these reasons as well as other unlisted reasons, the migration of customer segmentation often fallen flat in real applications.

In this paper, we focus on the problem of migrating customer segmentation result from external data to an enterprise's internal data which we called the MCS problem, and develop a five-step hybrid model, we call it HMCS model, to resolve it. The model combines classification techniques and fuzzy set techniques to populate values of missing attributes to enterprise data; and then migrates customer segmentation result from an enterprise's external data source to its internal data source. The model is implemented and applied to a real application in a customer segmentation problem. The reminder of the paper is organized as below. Section 2 briefly reviews related methods and techniques in customer segmentation. Section 3 gives a formal definition of the MCS problem through an example. Section 4 outlines the main steps of the five-step HMCS model and Section 5 illustrates an application of the HMCS model in a service provider's customer segmentation. Section 6 summaries the presented work and discusses future works.

II. RELATED WORKS

Customer segmentation is a crucial issue in enterprise's customer relationship management (CRM) for gaining better performance and corporate reputations from providing customers with expected products and services in competitive

markets [4]. Customer segmentation has been extensively studied in all kinds of industrial sectors, such as telecom [15], finance [8], insurance [9], information and communication technology (ICT) facilities [7], airline industry [19], healthcare [11], as well as tourists [12]. These methods are built on both enterprise internal and external datasets and are with assumption that the dataset used can provide all segmentation-required customer-related variables. However, in real applications, this assumption is hard to be hold. Hence, many of such methods have "fallen flat when used in marketing and advertising campaigns" [5].

Customer segmentation is a procedure to "recognize groups of customers who share the same or similar needs" [13]. Essentially, customer segmentation is a clustering problem; therefore, majority existing methods are based on clustering techniques and algorithms, for example, the K-means clustering algorithm. Moreover, experience description and statistical analysis are also the main basis of a segmentation method [3][4]. On the other hand, because combining multiple clustering or learning algorithms often outperforms single algorithm, many works adopt more than one algorithms in one method.

Customer segmentation result is hard to be migrated from one dataset to another. A customer segmentation method is often designed based on customer demography, geographic locations, behaviors, benefit-cost relations, as well as lifestyles. These variables may change frequently and then affect the consistency of a segmentation [13]. Moreover, these attributes contain huge amount of a customer's subjective expectation or preference; for which it is hard to find a counterpart in a real dataset. Currently, customer segmentation is conducted separately on different datasets and no report has been given on how to migrate a segmentation result from one dataset to another.

III. PROBLEM DESCRIPTION AND FORMALISATION

In this section, the MCS problem is described through an example and then formalised.

A service provider plans to develop several new service products to retain existing customers and attract potential customers which are small- or medium-sized companies. The service provider notes roughly that the industrial sector background, number of employees, and company owner's preference impact a customer's selection of a certain telecom product or service. Hence, the service provider wants to segment its focal customers into several groups and develops corresponding products and services for each group. Due to various historical, legal or technical reasons, the service provider lacks some important demographic, behaviour and preference data of the targeted customers in its enterprise database. Hence, the telecom appoints a third-party consultant company to survey a number of randomly selected customers and conduct a customer segmentation on the collected survey data. Because the service provider's enterprise data is ill-match with the survey data, particularly it lacks some customer-related indicators which are used in the segmentation model, the telecom cannot apply the segmentation result to its

enterprise data directly and needs to find a way to use that result. Therefore, a practical problem arises that how to migrate a customer segmentation result from one dataset to another. We call this kind of problem as Migrating Customer Segmentation (MCS) problem.

The MCS problem is existing in many industrial sectors such as finances and insurances. Generally, an MCS problem is briefly described as below.

Suppose C is an enterprise's internal dataset, which contains customer-related records. Each customer-related record is depicted through m attributes a_1, \dots, a_m . Let S be another dataset, the customer survey date in above example, obtained externally from the enterprise. On S , K customer segments are defined through n attributes x_1, \dots, x_n and labelled as G_1, \dots, G_K , i.e., $G_k = g_k(x_1, \dots, x_n)$ for any $k = 1, \dots, K$. The MCS problem needs to answer the question that how to migrate G_1, \dots, G_k from S to C under the constraint that $\{a_1, \dots, a_m\} \cap \{x_1, \dots, x_n\} \neq \emptyset$ and $\{a_1, \dots, a_m\} \neq \{x_1, \dots, x_n\}$.

Furthermore, an MCS problem can be formalised in a more generalised form.

Definition 2: Let S and T be the source and target datasets, respectively. Elements of S and T are represented by attribute sets X and A , respectively; and $X \neq A$ and $X \cap A \neq \emptyset$. Let G be a label set with K labels G_1, \dots, G_K , which represents some knowledge learnt from S . A mapping g is defined on S such that for any $s \in S$, $g(s) \in G$. An MCS problem is how to define a mapping f on T such that for any $t \in T$, $f(t) \in G$.

IV. A FIVE-STEP HYBRID MODEL (HMCS)

In this section, a hybrid model, called HMCS model, is presented to solve the MCS problem. This model contains five steps as described below. Fig. 1 gives the main steps of it.

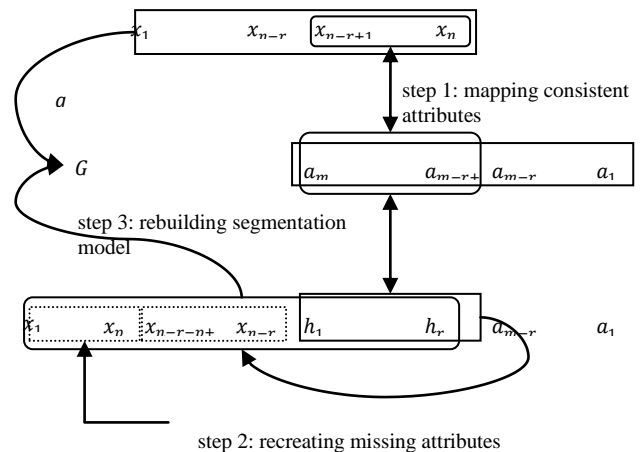


FIG 1 Main steps of the HMCS model

TABLE 1
OUTLINE OF THE HMCS MODEL

Outline of the HMCS model

-
- Step 1: Mapping consistent attributes between source and target datasets
 - Step 2: Recreating missing attributes and populating their values
 - Step 3: Rebuilding segmentation model on source dataset
 - Step 4: Applying segmentation model to target dataset
 - Step 5: Evaluating segmentation model
-

Step 1: Mapping consistent attributes between source and target datasets.

We will say that an attribute x in the source dataset is consistent with an attribute a in the target dataset if both x and a refer to the same feature of a customer and may have different value forms though. For instance, "month spending" is an attribute used in most telecom customer survey and often given in the form of a number of spending ranges (intervals of spending amounts). In a telecom company's enterprise database, a customer's "monthly billing amount" records the real spending of the customer and is often recorded as a real number. Although they are expressed in different forms, these two attributes describe the same thing, i.e., a customer's spending on telecom service approximately in a month. Hence, "month spending" (from a source dataset) is consistent with "monthly billing amount" (in a target dataset). In the following sections, two consistent attributes are called matching to each other.

Below, we use the same symbol to replace the consistent attributes between X and A ; and rewrite X and A as: $X = \{x_1, \dots, x_{n-r}, h_1, \dots, h_r\}$, $A = \{h_1, \dots, h_r, a_{r+1}, \dots, a_m\}$, where h_1, \dots, h_r are matched attributes between X and A . Let $H = \{h_1, \dots, h_r\}$, where h_r is a matching attribute and H is called the matching attribute set.

Each matching attribute indicates a common customer feature in both source and target datasets. For each h_r , we build a mapping m_r such that:

(1) If h_r has categorical values in both source and target datasets.

$$m_r(V_S(h_r)) \subseteq V_T(h_r)$$

where $V_S(h_r)$, $V_T(h_r)$ are the values of h_r occurs in S and T , respectively.

(2) If h_r has categorical values in source dataset but continuous values in target dataset.

$$m_r(V_T(h_r)) \subseteq V_S(h_r).$$

By this step, the matching attributes are aligned.

Step 2: Recreating missing attributes on source dataset and populating values of missing attributes to target dataset.

A missing attribute in the target dataset is an attribute which only exists in the source dataset but without consistent (matching) attribute in the target dataset. A typical example is a customer's "gender". A customer's gender is a common attribute used in many customer-oriented survey datasets; but it is seldom an attribute stored in an enterprise's database.

Because a missing attribute does not exist in the target dataset but it is used in the segmentation mapping g , this step tries to build a mock one for the target dataset. Consider the matching attribute set H is shared between the source and target datasets, we use H to generate the missing attribute.

Without loss of generality, suppose $n - r - p$ missing attributes x_{p+1}, \dots, x_{n-r} can be generated from H . For each x_j , $j = p + 1, \dots, n - r$, a subset S_{x_j} of the source dataset with attributes $H \cup x_j$ is obtained where H can be seen as condition attributes and x_j can be seen as decision attribute (class/category attribute). Therefore, a classification algorithm L_j , such as the decision tree or support vector machine [20], can be implemented to learn x_j from H , which can be then used on the target dataset to populate values of missing attributes x_{p+1}, \dots, x_{n-r} .

Since the fact that not all missing attributes can be generated by H , another method is also used to generate missing attributes. For missing attributes x_1, \dots, x_p which are not generated from H , we will populate their values to the target dataset based on their values' nature. If an attribute x focuses on a customer's objective feature, such as geographical location, then we populate its values following the probability distribution of those values. If an attribute y focuses on a customer's subjective feature, such as "how likely a customer will select a competitor's service", we will use fuzzy set and fuzzy logic technique [21] to summary its values, define a fuzzy set on them then, and populate the fuzzy memberships of those values. To explain this method, we give an example below.

Example. Suppose a missing attribute x in a telecom's customer survey is about "previous service provider" with values "company A", "company B", "company C", and "company D". Then we population all four values to the target dataset following their frequency distribution (probability). Suppose another missing attribute y in the same survey is about "how likely will you select another service provider?" and with five values "Definitely", "Very likely", "More likely", "Unlikely" and "Definitely not". Then we define a fuzzy set F on these five values as "degree of likely to leave", calculate the fuzzy membership degree of each value; and then populate the fuzzy membership degrees to the target dataset.

Considering that a fuzzy set is not uniquely determined, we can add a small disturbance ϵ to a fuzzy membership degree when populating it to the target dataset.

In step 2, we artificially generate attributes which are missing in the target dataset. After this step, the target dataset T contains all attributes in the source dataset S except that previously missed attributes take artificial values. Before using the original and generated attributes to implement customer segmentation, a model rebuilding (retraining) procedure on the source dataset is needed; this is the main task in step 3.

Step 3: Rebuilding segmentation model on source dataset.

In this step, a model retraining is implemented by using artificial data for some attributes and by using classification algorithm, which is conducted on the source dataset.

Noticed in FIG. 1, attributes x_{p+1}, \dots, x_{n-r} can be learned by attributes h_1, \dots, h_r , we replace them by

$$x_j = L_j(h_1, \dots, h_r), \quad j = p + 1, \dots, n - r.$$

For attributes x_1, \dots, x_p , we reassign artificially generated values to them based on either probability distribution or fuzzy membership degree following the method given in step 2. For these attributes, we use y_1, \dots, y_p to replace them. As the customer segmentation has been conducted and the segmentation result is known, an attribute z is, therefore, used to record the segmentation result.

Based on above preparation, a classification model g^* is built where attributes $h_1, \dots, h_r, y_1, \dots, y_p$ are condition attributes and the attribute z is the decision attribute, i.e.,

$$g^*(h_1, \dots, h_r, y_1, \dots, y_p) = z.$$

Note that it is completely built on attributes which are now existing in the target dataset, model g^* is therefore applied to the target dataset.

Step 4: Applying segmentation model on target dataset.

In this step, the model g^* is applied to the target dataset after populating values to those artificially generated attributes. Because the target dataset does not have those artificially generated attributes and it cannot provide any information about them, we firstly generate a value pool for each of those attributes based on the attribute's probability distribution in the source dataset, and then randomly pick a value from the generated value pool for each record in the target dataset to build an applicable record as the input of model g^* . Formally, the data population procedure is shown in Fig. 2.

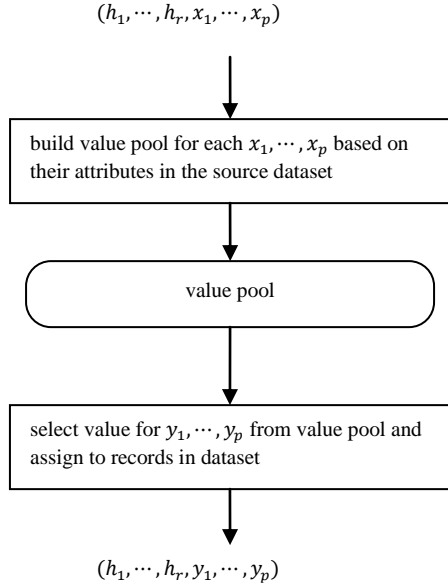


FIG.2 Data population from source dataset to target dataset

Step 5: Evaluating segmentation results.

The evaluation is conducted on the source dataset and also conducted manually on sample of the target dataset. On the source dataset, a cross validation is adopted. On the target dataset, a sample set is randomly picked up and evaluated by the domain experts mainly focusing on the approximate distribution of the segmentation result rather than the individual record.

The presented HMCS model has been implemented by using MySQL database and KNIME (the Konstanz Information Miner, www.knime.org) tool on a Dell Latitude D6500 laptop with 3GB RAM running Fedora 16 Linux system. In this section, we briefly introduce and analyse the experiment result.

A. Experiment data

The experiment data comes from an Australian service provider. The source data used in this experiment comes from the company's customer survey. The survey contains 42 customer related questions and covers total 2000 customers. Among the 2000 customers, 1542 customers are currently having contact with the company and 1519 customers have answered all relevant survey questions. Hence, we select all 1519 valid records to form the source dataset. The target datasets are three samples of the company's customer database with 102555 (target-1), 109743 (target-2), and 103013 (target-3) customer records, respectively .

B. Initial customer segmentation result

A customer segmentation has been conducted by a third-party consultant company on the 1519-record survey data. The total 1519 customers have been segmented into five groups which are labelled "segment-1", "segment-2", "segment-3", "segment-4", and "segment-5", respectively. TABLE 2 shows the record numbers of all the five groups.

TABLE 2
RECORD DISTRIBUTION OVER FIVE SEGMENTS IN SOURCE DATASET

Group Label	segment-1	segment-2	segment-3	segment-4	segment-5
Record Number	432	533	247	158	149
Percentage (%)	28.4	35.1	16.3	10.4	9.8

The segmentation result is built on five attributes (denoted by x_1, x_2, x_3, x_4, x_5) extracting from five questions in the total 42 survey. In the five attributes, three (x_3, x_4, x_5) have objective measurements and the other two (x_1, x_2) are subjective opinions. Among the three objective attributes, two (x_4, x_5) have counterparts in the target dataset. Furthermore, statistical analysis of correlation indicates that attributes x_1, x_2, x_3 cannot be estimated or learned from the attributes (x_4, x_5); therefore, we need to generate a value pool for each of them in order to populate their values to the target dataset as well as the source dataset as shown in step 2 and step 3.

C. Experiments results and analysis

To evaluate the presented model, three experiments are conducted. The first experiment, experiment 1, compares the model's segmentation result with the original segmentation result on the survey dataset through each segment's distribution. The second experiment (experiment 2) compares the model's segmentation result on the same target dataset (target-1) with different value pools in populating missing attributes' values. The third experiment, experiment 3,

compares the model's segmentation results on three sample target datasets.

The result of experiment 1 is shown in FIG. 3. The result indicates that the first three segments in both the original and the model' segmentation occupy the majority of the total 1519 records and with similar distributions, particularly the first two segments. Although as a whole the segment-4 and segment-5 in both segmentations occupy almost same percentages of the total dataset, the distributions of them in the two segmentations have significant difference. By checking individually the segmentation result, we noted that about 50-60% records are segmented to the same segment by both models.

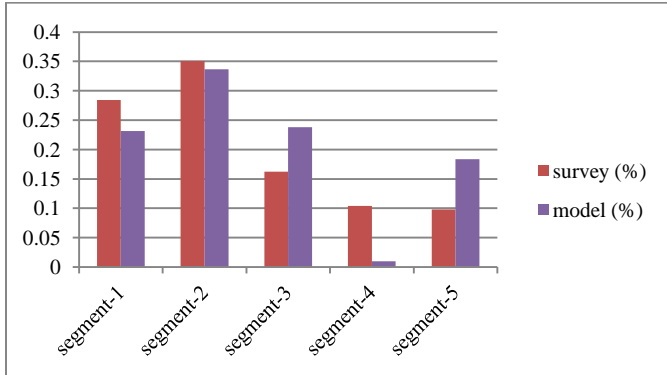


FIG. 3 Result of experiment 1.

The result of experiment 2 is shown in FIG. 4. The result indicates the model has produced similar segmentation results by using different value pools in populating missing values to the target dataset. However, we can still note the difference, in particular the segment-4.

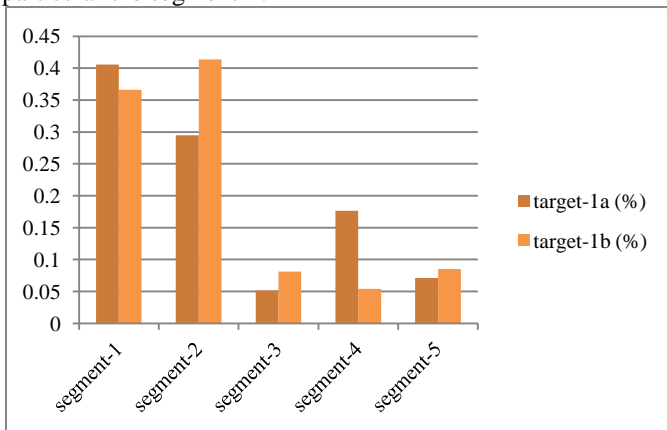


FIG. 4 Result of experiment 2.

The result of experiment 3 is shown in FIG. 5. It indicates that the model has produced similar segmentation results on different sample sets, although significant difference is still existing among these results.

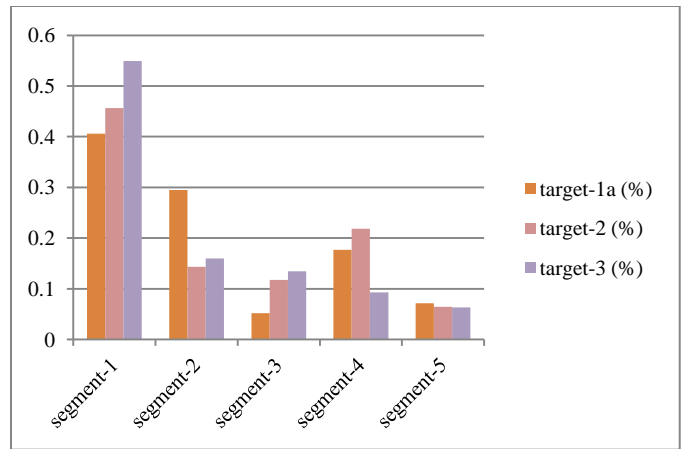


FIG. 5 Result of experiment 3.

In order to evaluate the model, we consultant some domain experts. They are satisfied with the segmentation result based on the service provider's enterprise data; but pointed the limitation on some segments such as segment-4 and segment-2.

Many reasons account for the model's limitation. The incompleteness of enterprise data may be the easiest one to blame for. Due to missing attributes in the target dataset, the value population procedure of those attributes is another one. Moreover, the inconsistent data between the source and target datasets also plays its role in the limitation. To overcome the limitation and improve the performance of the model, more works are needed.

VI. CONCLUSION AND FUTURE WORKS

Migrating customer segmentation from survey data to enterprise data is a challenging topic in enterprise customer relationship management. To solve this problem, this paper developed a five-step hybrid model, the HMCS model, which particularly focuses on missing attributes in the enterprise data. The model tries to generate missing attributes and populate their values to the target dataset. Experiments shown the capability of the model in solve this kind of problem. Due to complexity of the MCS problem, more works still need to be done in the presented HMCS model to improve its performance. Firstly, the model artificially generated values for missing attributes. A further theoretical analysis is required although the obtained segmentation results are acceptable. Secondly, the model's segmentation result has significant difference in some segments compared with the initial segmentation result. How to reduce the difference needs more studies. Finally, the essence of the MCS problem is a clustering algorithm. Hence how to build an appropriate clustering algorithm for this kind of problem is also an important issue to be studied.

REFERENCES

- [1] A. Amiri, "Customer-oriented catalog segmentation: effective solution approaches," *Decision Support Systems*, 2006, vol. 42, pp.1860-1871
- [2] M. Böttcher, M. Spott, D. Nauck and R. Kruse, "Mining changing customer segments in dynamic markets," *Expert Systems with Applications*, 2009, vol. 36, pp. 155-164

- [3] Y. Chen, G. Zhang, D. Hu and C. Hu, "Customer segmentation based on survival character," *Journal of Intelligent Manufacturing*, 2007, vol. 18, pp. 513-517
- [4] B. Cooil, L. Aksoy and T. L. Keiningham, "Approaches to customer segmentation," *Journal of Relationship Marketing*, 2008, vol. 6, no. 3-4, pp. 9-39
- [5] P. Dalton, "Customer segmentation, makes marketing more effective," *ABA Banker News*, 2006, vol. 14, no. 9, pp. 1-2
- [6] M. S. Garver, Z. Williams, G. S. Taylor and W. R. Wynne, "Modelling choice in logistics: a managerial guide and application," *International Journal of Physical Distribution & Logistics Management*, 2012, vol. 42, no. 2, pp. 128-151
- [7] I. Gil-Saura and M. Ruiz-Molina, "Customer segmentation based on commitment and ICT use," *Industrial Management & Data Systems*, 2009, vol. 109, no. 2, pp. 206-223
- [8] C. Hillenbrand and K. Money, "Segmenting stakeholders in terms of corporate responsibility: implications for reputation management," *Australasian Marketing Journal*, 2009, vol. 17, pp. 99-105
- [9] M. B. Hosseini and M. J. Tarokh, "Customer segmentation using CLV elements," *Journal of Service Science and Management*, 2011, vol. 4, pp. 284-290
- [10] Abbas Keramati and Seyed M.S. Ardabili, "Churn analysis for an Iranian mobile operator," *Telecommunications Policy*, 2011, vol. 35, pp. 344-356
- [11] S. Khandelwal and A. Mathias, "Using a 360° view of customers for segmentation," *Journal of Medical Marketing: Device, Diagnostic and Pharmaceutical Marketing*, 2011, vol. 11, pp. 215-220
- [12] J. Kim, S. Wei and H. Ruys, "Segmenting the market of West Australian senior tourists using an artificial neural network," *Tourism Management*, 2003, vol. 24, pp. 25-34
- [13] G. Lefait and T. Kechadi, "Customer segmentation architecture based on clustering techniques," *2010 Fourth International Conference on Digital Society*, pp. 243-248, Feb. 10-16, 2010, St. Maarten, Netherlands Antilles
- [14] K.-N. Lemon and T. Mark, "Customer Lifetime value as the basis of customer segmentation," *Journal of Relationship Marketing*, 2006, vol. 5, no. 2-3, pp. 55-69
- [15] C. Mazzoni, L. Castaldi and F. Addeo, "Customer behavior in the Italian mobile telecommunication market," *Telecommunications Policy*, 2007, vol. 31, pp. 632-647
- [16] V.L. Miguéis, A.S. Camanho and J. Falcão Cunha, "Customer data mining for lifestyle segmentation," *Expert Systems with Applications*, 2012, vol. 39, pp. 9359-9366
- [17] J. A. Pesonen, "Segmentation of rural tourists: combining push and pull motivations," *Tourism and Hospitality Management*, 2012, vol. 18, no. 1, pp. 69-82
- [18] Y. Takano, A. Inoue, T. Kurosawa, M. Iwashita and K. Nishimatsu, "Customer segmentation in mobile carrier choice modeling," *9th IEEE/ACIS International Conference on Computer and Information Science*, 2010, pp. 111-116
- [19] T. Teichert, E. Shehu, I. von Wartburg, "Customer segmentation revisited: the case of the airline industry," *Transportation Research Part A*, 2008, vol. 42, pp. 227-242
- [20] X. Yang, G. Zhang, J. Lu and J. Ma, "A kernel fuzzy c-means clustering based fuzzy support vector machine algorithm for classification problems with outliers or noises," *IEEE Transactions on Fuzzy Systems*, 2011, vol. 19, no. 1, pp. 105-115.
- [21] G. Zhang and J. Lu, "A linguistic intelligent user guide for method selection in multi-objective decision support systems," *Information Sciences*, 2009, vol. 179, no. 14, pp. 2299-2308
- [22] X. Zhang, J. Zhu, S. Xu and Y. Wan, "Predicting customer churn through interpersonal influence," *Knowledge-Based Systems*, 2012, vol. 28, pp. 97-104