Paper

Multicriteria analysis for behavioral segmentation

Janusz Granat

Abstract—Behavioral segmentation is a process of finding the groups of clients with similar behavioral patterns. The basic tool for segmentation is a clustering algorithm. However, the clusters generated by the algorithm depend on the preprocessing steps as well as parameters of the algorithm. Therefore, there are many possibilities of dividing the clients into segments and it is a subjective process. In this paper we will focus on application on multicriteria analysis for selecting the best partition of clients into segments.

Keywords—segmentation, clustering, multicritera analysis, telecommunications.

1. Introduction

The knowledge about clients becomes one of the most important assets in making various business decisions. In this paper we will focus on telecommunication industry. However, the methods that we will present are also valid for other industries. The only requirement is that they have operational databases that are sources for building behavioral features of the clients. Behavioral segmentation is a process of finding the groups of clients with similar behavioral patterns. It is one of the key data mining tasks for marketing departments of telecommunication operators [1, 6, 9]. In this paper we will focus on multicriteria application in segmentation process.

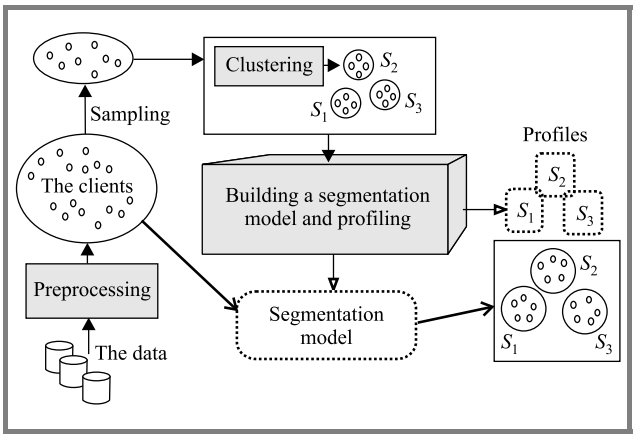

Fig. 1. The segmentation process–analyst view.

We can distinguish two views on segmentation process. The first, analyst view (see Fig. 1), focuses on preprocessing, selection of samples, clustering, building the segmentation model and profiling. Preprocessing consist of loading the data from various sources and transforming it into a table where each row corresponds to a client and columns correspond to the selected features of clients. In classical segmentation the features might be divided into the following groups: geographic (e.g., area, region, city, size),

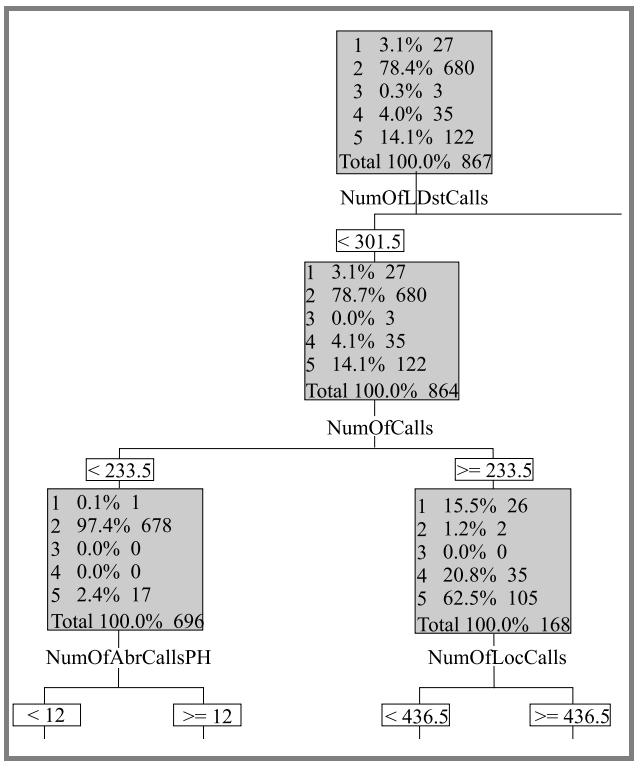

Fig. 2. Model as the cluster profile tree (SAS system).

```
IF      31.5 <= NumOfLDstCalls
THEN
 NODE:    3
 N    :   3
 1    :   0.0%
 2    :   0.0%
 3    : 100.0%
 4    :   0.0%
 5    :   0.0%

IF      12 <= NumOfIntCalls
AND NumOfCalls < 23.5
AND NumOfLDstCalls < 31.5
THEN
 NODE:    7
 N    :   1
 1    : 100.0%
 2    :   0.0%
 3    :   0.0%
 4    :   0.0%
 5    :   0.0%
```

Fig. 3. Model as the set of rules (SAS system).

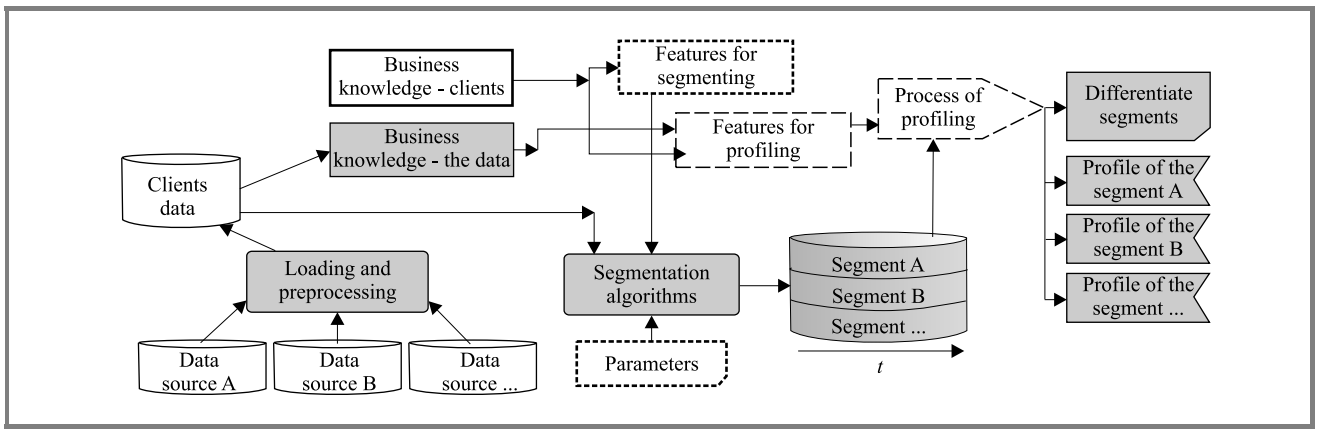

Fig. 4. The segmentation process–business view.

demographic (age, life stage, marital status), socioeconomic (e.g., income, education), psychographic (e.g., personality, lifestyle) [7]. The organizations must buy this type of data or gather it by questionnaires [10].

Telecommunication operators have the data stored in billing databases and others dedicated to specific services. The data about each call is stored in call detail records (CDR). These records contain the following data: caller number, called number, timestamp, the length of the call, tariff units per call, etc. It should be stressed that this data has incomparable quality in comparison to demographic, socioeconomic data, etc. This leads to much better modeling results. CDR records are transformed into client's behavioral features. The examples of behavioral features are the following: number of calls, number of calls within peak hours, number of calls within off-peak hours, number of different called parties, number of specific type connections (local, long distance, international, mobile, premium rate numbers with prefix 0700, toll-free numbers with prefix 0800, internet provider), total length of calls, total length of calls within peak hours, total length of calls within off-peak hours, number of not answered calls, etc. In practical modeling there are more then 1000 such features. If we have, e.g., 5 million of clients the resulting behavioral segmentation table is huge. Therefore, we choose a sample of clients for building the segmentation model. Next, we apply segmentation algorithms several times and generate various instances of the segmentation model. The model might have various views, e.g., in SAS system the model is represented as:

- the cluster profile tree (Fig. 2),
- the set of rules (Fig. 3),
- the SAS 4GL code,
- the C language code.

A cluster profile tree has the percentages and numbers of cases assigned to each cluster and the threshold values of each input variable is displayed as a hierarchical tree. The set of rules is a text file that lists all the rules used to create the profile tree.

In fact, segmentation model belongs to the class of the classification models. We generate several models. Then we usually choose subjectively one of them. This paper shows how this phase might be supported by multicriteria analysis. The model is then applied to whole population of clients. After that, profiles for all segments are built.

The segmentation process in the company has to take into account knowledge of business people about the data, the features that might be used for clustering and the features of clients that might be used for profiling (see Fig. 4).

The business people set up goals of segmentation and then actively participate in the segmentation process. We have quite different results of segmentation, even if it is based on the same data, by setting different goals. As a result of segmentation in a business environment there are profiles of each segment (which describe the features that are common for the segments), as well as so called differentiate profiles (which describe the features that differentiate the segments).

2. The formal model

Modeling is based on the dynamic information system [2, 8] defined as follows:

$$IT = (X, F, V, \rho, T, R), \tag{1}$$

where: T is a nonempty set whose elements are called moments of time, R is a order on the set T (here we assume a linear order), X is the finite and nonempty set of n objects or observations, F is finite and nonempty set of features of the objects, $V = \bigcup_{f \in F} V_f$, V_f is a set values of feature $f \in F$, called the domain of f, ρ is an information function: $\rho : F \times X \times T \to V$.

The dynamic information system considers explicitly time. Segmentation is a dynamic process because clients change their behavior. In this paper we are focusing on multicriteria selection of partition of segments at a specific moment of time t so the time might not be considered.

There is a set of objects X of a dynamic information system IT and the similarity measure between objects $x_i, x_j \in X^{id}$, $i \neq j$:

$$\varphi(x_i, x_j).$$

If there are linguistic, nominal, boolean, and interval-type of features, along with quantitative attributes, the symbolic similarity between the objects is applied [5].
Moreover, clustering depends on parameters of the preprocessing steps Γ and parameters of the algorithms Ω. Let us Δ ($\Delta = \Gamma \cup \Omega$).
Clustering algorithms divide the set of objects into m subsets of similar objects (based on the similarity measure):

$$X \mapsto_{\varphi(x_i,x_j),\Delta} \{X_{S_{1,\Delta}}, X_{S_{2,\Delta}}, \ldots \cup X_{S_{m,\Delta}}\},$$

$$X_{S_{i,\Delta}} \subset X, X_{S_{i,\Delta}} \cap X_{S_{j,\Delta}} = \emptyset \text{ for } i \neq j, \bigcup_i X_{S_{i,\Delta}} = X.$$

For the set of parameters Δ, we have the corresponding identifiers of the clusters:

$$\mathrm{Seg}_\Delta = \{S_{1,\Delta}, S_{2,\Delta}, \ldots, S_{m,\Delta}\}.$$

For a huge data set we have to find the clusters of objects for a training set $X^{Train} \subset X$ and then we build a classification model that can be applied to X set. Let us consider the selected f_S feature of the object (where $S \in FS$, FS is the index of cluster feature), called cluster feature, and the subsets of input features f_I ($I \in FI$, $FI = F \setminus FC$, FI is the index set of input features, F is the index set of all object features).
The model is defined as follows:

$$\rho(f_S, x_i) = M_S\big(\rho(f_{k1}, x_i), \rho(f_{k2}, x_i), \ldots, \rho(f_{kk}, x_i)\big),$$

where: x_i is an object identifier, $k1, k2, \ldots, kk \in FI$.

3. The multicriteria analysis

The analyst divides the set of clients into various segments by setting various sets of parameters Δ. The multicriteria analysis might be applied to final selection of the segments. For each of the set of parameters Δ_i we have a partition $\mathrm{Seg}_{\Delta i}$ and a vector of criteria that evaluate this segmentation $\mathbf{q}_i = (q_{i,1}, q_{i,2}, \ldots q_{i,r})$ (see Table 1). Let the set of segmentation results $\mathrm{Seg} = \{\mathrm{Seg}_{\Delta_1}, \mathrm{Seg}_{\Delta_2}, \ldots, \mathrm{Seg}_{\Delta_n}\}$. The multicriteria problem is defined as follows:

$$\min_{\mathbf{Seg}}, \max, \mathrm{stab}\,\mathbf{q}.$$

Table 1
Table of evaluations of segmentations

Partition	q_1	q_2	$\ldots$	q_r
Seg_{Δ_1}	$q_{1,1}$	$q_{1,2}$	$\ldots$	$q_{1,r}$
Seg_{Δ_2}	$q_{2,1}$	$q_{2,2}$	$\ldots$	$q_{2,r}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Seg_{Δ_n}	$q_{n,1}$	$q_{n,2}$	$\ldots$	$q_{n,r}$

The following examples of criteria might be considered:

- *Number of segments*:

$$q_{i,1} = N\mathrm{Seg}_{\Delta_i}.$$

- *Outliers frequency*:

$$q_{i,2} = OF_{\Delta_i} = \frac{n_{\mathrm{Seg}_{\Delta_i}} - no_{\mathrm{Seg}_{\Delta_i}}}{n_{\mathrm{Seg}_{\Delta_i}}},$$

  where: $n_{\mathrm{Seg}_{\Delta_i}}$ is the number of objects in all segments for Δ_i, $no_{\mathrm{Seg}_{\Delta_i}}$ is the number of deleted outliers.

- *Segments frequency*:

$$q_{i,3} = SF_{\Delta_i} = \frac{\sum_{l=1}^{m} \frac{n_{S_{l,\Delta_i}}}{\max_{\mathrm{Seg}_{\Delta_i}}}}{N\mathrm{Seg}_{\Delta_i}},$$

  where: $n_{S_{l,\Delta_i}}$ is the number of objects in the segment, $\max_{\mathrm{Seg}_{\Delta_i}}$ is the maximal numer of objects.

- *Segments compactness* [4], that evaluates how well the segments are redistributed by the clustering algorithm, compared to the whole set X:

$$q_{i,4} = \frac{1}{m} \sum_{l=1}^{m} \frac{v(X_{S_{l,\Delta_i}})}{v(X)},$$

  where:

$$v(X) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} d^2(x_i, \bar{x})},$$

  $d(x_i, x_j)$ is a distance metric between two objects x_i and x_j, n is the number of objects in X, m is a number of segments, $\bar{x}$ is the mean of X, $v(X_{S_{l,\Delta_i}})$ is calculated for objects in the segment.
  A smaller $q_{i,4}$ indicates a higher homogeneity of the objects in the segments.

- *Separation of segments* [4]:

$$q_{i,5} = \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j=1, j\neq k}^{m} \exp\left(-\frac{d^2(x^c_{S_{i,\Delta}}, x^c_{S_{j,\Delta}})}{2\sigma}\right),$$

  where: σ is a Gaussian constant, m is the number of clusters, $x^c_{S_{i,\Delta}}$ is the centroid of the cluster $S_{i,\Delta}$, $d()$ is the distance metric used by the clustering algorithm, and $d(x^c_{S_{i,\Delta}}, x^c_{S_{j,\Delta}})$ is the distance between the centroid of segments.
  A smaller segment separation criterion indicates a larger overall similarity among the segments.

- $\ldots$ $q_{i,r}$ (several other criteria might be considered).

Table 2
Example of table of evaluations of segmentations

Partition	*NC*	*Outliers*	*FC*
$\text{Seg}_1 = \text{Seg}_{\Delta_1}$	10	1.0	0.115
$\text{Seg}_2 = \text{Seg}_{\Delta_2}$	6	1.0	0.171
$\text{Seg}_3 = \text{Seg}_{\Delta_3}$	5	1.0	0.208
$\text{Seg}_4 = \text{Seg}_{\Delta_4}$	5	0.62	0.208
$\text{Seg}_5 = \text{Seg}_{\Delta_5}$	15	0.98	0.082
$\text{Seg}_6 = \text{Seg}_{\Delta_6}$	15	0.87	0.119

Illustrative example (Table 2) has been prepared on sample of telecommunications data. We have generated six partitions of clients by the SAS system. Then, we calculated the values of three criteria $N\text{Seg}_{\Delta_i}$, OF_{Δ_i} SF_{Δ_i}. Then the multicriteria analysis has been applied.
For specification of the preferences, we have applied the ISAAP software, developed by Granat and Makowski [3]. The user specifies preferences, including values of criteria that he/she wants to achieve and to avoid. Those values are called *aspiration* and *reservation* levels, respectively. The graphs of the so called component achievement functions and the related solution are presented in Fig. 5. The solution, marked by rectangle, is projected into two-dimensional space, in which, for each criterion, its values (the x axis) and the degree of satisfaction (the y axis) of meeting preferences, expressed by aspiration and reservation levels, are reported. In the next step, an interactive procedure is used to assist the user in selecting a segment that best corresponds to his/her preferences. In our example, the user changes reservation point for *SF* criterion to value equal to 0.15. The new solution is presented in Fig. 6. The user can continue the interactive process in order to find another segment.

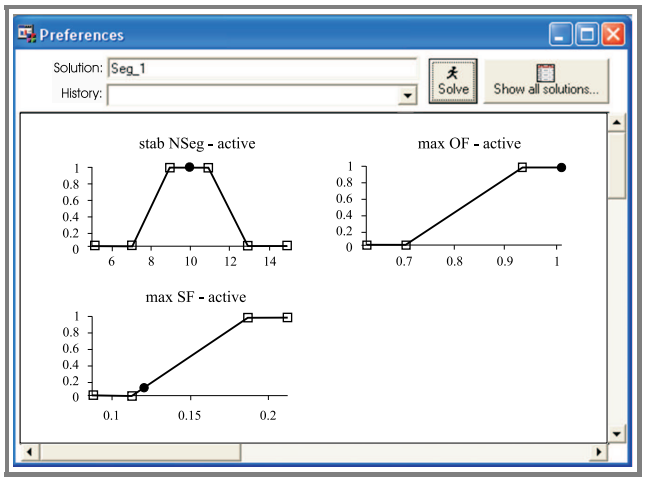

Fig. 5. The ISAAP screen.

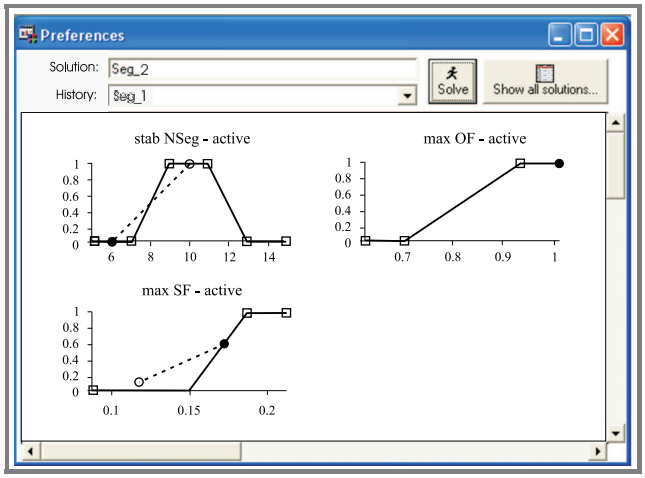

Fig. 6. The ISAAP screen–new preferences.

4. Conlusions

The quality of the behavioral segmentation process significantly influences the clients relationship managament. As we have shown, the process of behavior in business environment is very complex and requires various interactions of analysts and business representatives. Selection of the final partition of clients into segments that is well attuned to the business goals is usually a subjective task. There is a need for development of analytical tools that improve objective comparison of partition of clients into segments. This paper shows how to apply a multicriteria analysis in this process and it is a step into development of such tools.

References

[1] J.-L. Amat, "Using reporting and data mining techniques to improve knowledge of subscribers; applications to customer profiling and fraud management", *J. Telecommun. Inform. Technol.*, no. 3, pp. 11–16, 2002.

[2] J. Granat, "Data mining and complex telecommunications problems modeling", *J. Telecommun. Inform. Technol.*, no. 3, pp. 115–120, 2003.

[3] J. Granat and M. Makowski, "Interactive specification and analysis of aspiration-based preferences", *Eur. J. Oper. Res.*, vol. 122, no. 3, pp. 469–485, 2000.

[4] J. He, A.-H. Tan, C.-L. Tan, and S.-Y. Sung, "On quantitative evaluation of clustering algorithms", in *Clustering and Information Retrieval*, W. Wu, H. Xiong, and S. Shekhar, Eds. Kluwer, 2003.

[5] K. Mali, "Clustering and its validation in a symbolic framework", *Patt. Recogn. Lett.*, vol. 24, pp. 2367–2376, 2003.

[6] R. Mattison, *Data Warehousing and Data Mining for Telecommunications*. Boston, London: Artech House, 1997.

[7] M. McDonald and I. Dunbar, "Market segmentation. How to do it, how to profit from it". Palgrave Publ., 1998.

[8] E. Orłowska, "Dynamic information systems", in *Annales Societatis Mathematicae Polonae*, Series IV: *Fundamenta Informaticae*, vol. 5, no. 1, pp. 101–118, 1982.

[9] M. Shawa, C. Subramaniama, G. Tana, and M. Welgeb, "Knowledge management and data mining for marketing", *Decis. Supp. Syst.*, vol. 31, no. 1, pp. 127–137, 2001.

[10] P. Verhoef, P. Spring, J. Hoekstra, and P. Leeflang, "The commercial use of segmentation and predictive modeling techniques for database marketing in the Netherlands", *Decis. Supp. Syst.*, vol. 34, pp. 471–481, 2002.

Janusz Granat received his M.Sc. in control engineering (1996) and his Ph.D. (1997) in computer science from the Warsaw University of Technology. He holds a position as an Assistant Professor at the Warsaw University of Technology, and is the leader of a research group on applications of decision support systems at the National Institute of Telecommunications in Warsaw. He lectured decision support systems and various subjects in computer science. His scientific interests include data mining, modeling and decision support systems, information systems for IT management. Since 1988 he has been cooperating with IIASA. He contributed to the development of decision support systems of DIDAS family and the ISAAP module for specifying user preferences. He has been involved in various projects related to data warehousing and data mining for telecommunication operators. He was also involved in EU MiningMart project.

e-mail: J.Granat@itl.waw.pl
National Institute of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland
Institute of Control and Computation Engineering
Warsaw University Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland