Managing Innovation and Diversity in Knowledge Society Through Turbulent Time
25–27 May 2016
Timisoara • Romania

make learn
TIIM

Management, Knowledge and Learning
Joint International Conference 2016
Technology, Innovation and Industrial Management

THE USE OF CLUSTERING METHODS AND MACHINE LEARNING ALGORITHMS IN THE TRADING ENTERPRISE FOR CUSTOMER SEGMENTATION

Mieczysław Pawłowski
Onninen sp. z o.o., Poland
mieczyslaw.pawlowski@onninen.pl

Jarosław Banaś
Maria Curie-Sklodowska University in Lublin, Poland
jaroslaw.banas@umcs.pl

Abstract:
In the activity of any enterprise, it is essential to prepare a specific offer tailored to the needs of commercial customers. Large operation scale of businesses often makes it impossible to prepare individual offers for all customers, mainly for economic and logistical reasons. Therefore, it is important to appropriate customer grouping for the preparation of a proper offer to each group. Nowadays, it is difficult to separate the relevant groups characterized by a specific purchasing profile due to the dynamism of events in the modern economy and frequent changes in customer preferences. In order to maintain the current divisions, this classification must be done relatively often. The search for appropriate models and benchmarks is a continuous process. Enterprises use various methods to classify their customers. These methods are characterized by various degrees of complexity and varying effectiveness. This paper presents the results of analyses of customer segmentation in a trading enterprise, using clustering methods.

Keywords: clusters, clustering methods, customer segmentation, customized offer, enterprise.

1. INTRODUCTION

Division of a customer database into groups and their assignment to predetermined segments is a complex process. Determination of the optimum number of customers is in itself an interesting challenge for researchers and entrepreneurs, due to implications to management organization and engagement of financial, human and organizational resources into customer service, with consideration of specificity of customer groups. On the other hand, accepting the aforesaid challenge provides an opportunity to learn about customer preferences, offer a wider choice of products, expand the business scale, as well as better customize the services to meet the requirements of particular segments.

The paper provides an analysis of purchases made by a group of customers over 12 months, and it focuses on identification of customer groups displaying a specific purchase profile. Here, the analysis pertained to the number of conducted transactions, in order to ignore the financial volume and focus on the very demand for specific products in transactional terms. In the future, other qualities are planned to be analyzed as well, such as e.g. the volume of purchases, shopping in physical/traditional stores and online shops, speed of express deliveries, or deliveries to the head office.

For research purposes, the sample of over 12,000 transactions was collected. The sample of such great volume enabled authors to perform cluster analysis. Cluster analysis was performed using the k-means algorithm. It allows to determine relatively homogeneous groups from the set of available observations.

Grouping has been highly popular among researchers and entrepreneurs. Practical implementations of segmentation research which serve as the basis for entrepreneurs to minimize customer service costs, with simultaneous pursue for their maximum satisfaction are an example here. Determination of the number of customer groups and assignment of customers to each determined group gives an opportunity to optimize enterprise resources necessary for customer service, in particular: human, financial and material resources.
Internet trends analysis points to a large number of grouping-related inquiries, e.g. using the k-means algorithm (Pic. 1).

Picture 1: Interest in the term ‘k-means clustering’ in the period Jan 2006-Jan 2016

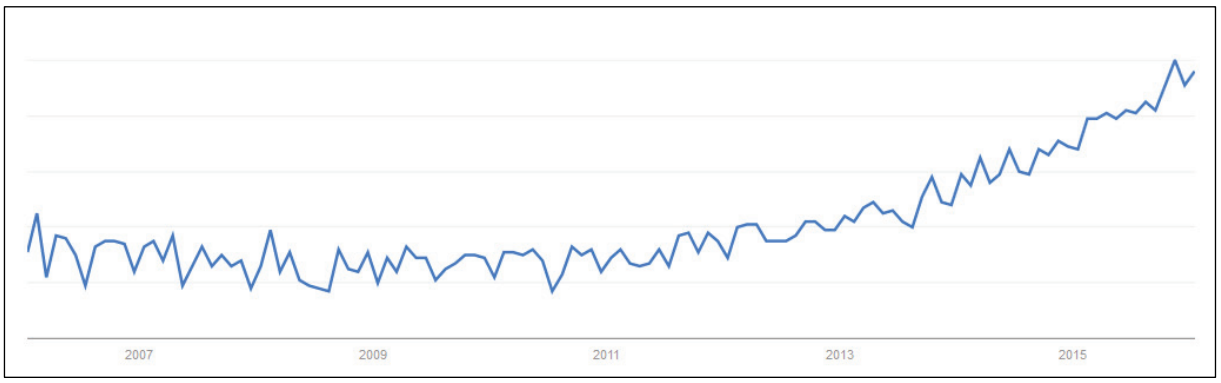
Source: Google Trends, https://www.google.pl/trends.

The possibility of object grouping is a valuable tool for all stakeholders analyzing large data sets. There are many adjustment possibilities, examples of which include: global political economy (Koutsoukis, 2015), knowledge economy and country classification (De la Paz-Marín et al., 2015), web browsing patterns (Song & Shepperd, 2006), clustering e-commerce search engines (Lu et. al., 2006) etc.

In addition, a tendency to use machine learning for grouping purposes has been observed. Here, it is the system that qualifies objects to specific groups. Such solution may enable automation of the assignment (classification) process of a relevant instance to a given group, and allow for better performance due to constant learning – gaining the experience. The use of machine learning allows to perform the analyses faster and use the full range of algorithms determining the assignments.

The Internet trend analysis points to a growing number of queries concerning the term “machine learning” (Pic. 2).

Picture 2: Interest in the term ‘machine learning algorithms’ in the period Jan 2006-Jan 2016



Source: Google Trends, https://www.google.pl/trends.

2. CONCEPTUAL BACKGROUND AND METHOD

For purposes of customer segmentation, the sample of 12 870 transactions of company customers was obtained, who made purchases in the following product groups:

- power industry
- gas and water
- electrical wiring
- hydraulic systems
- heating
- technical stores
- telecommunications
- industrial

Input data for further analyses was the number of conducted transactions in each of the listed groups over the analyzed period. For easier comparison of the volume of transactions in each of the eight groups, their per cent share was determined, according to the principle:

$$\frac{}{}$$

where:

– percent value of transactions in group ,

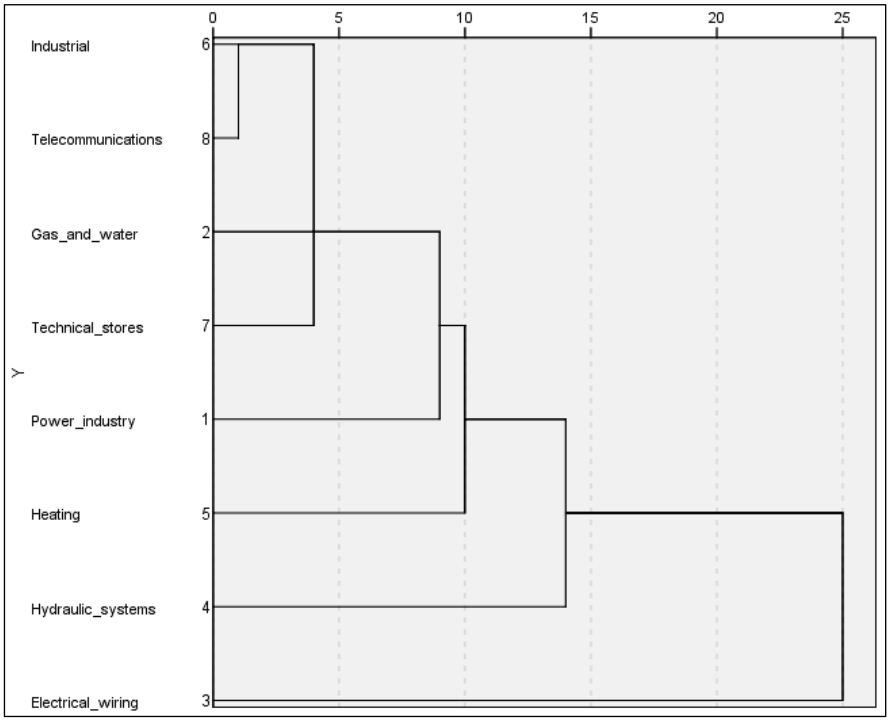
– number of transactions in group .

In further analyses, percent value of a transaction ( ) was applied in each group, to each purchase. Hierarchical clustering was employed for the variables. In subsequent stages of the research, observations were partitioned into clusters using the k-means method, with the adopted number of clusters. Ultimately, visualization of selected clusters was presented.

3. THE HIERARCHICAL CLUSTER ANALYSIS

In order to find similarities between the variables, hierarchical cluster analysis was employed, using the method based on average distance between the clusters. Clusters referred to the variables (Picture 3). The analysis indicates that four product groups are similar to each other: industrial, telecommunications, gas and water, and technical stores. Other product categories differ from each other and from the previously determined group.

Picture 3: Dendrogram



Source: Author's own work; Combined clusters – oversized distances.

4. CLUSTER ANALYSIS USING K-MEANS METHOD

Clusters were derived using WEKA software[1]. Partitioning into groups was conducted using the k-means clustering (SimplekMeans), where the distance function was Euclidean distance (distanceFunction → EuclideanDistance), with the fixed (assumed) number of clusters → 8 (numClusters → 8). The number of clusters was selected assuming that since 8 groups based on the knowledge about intended use of the range of specialist products were identified, then it is expected that there are significant customer groups interested in these products. Researchers were interested whether customer groups purchase different product groups, or whether they are homogeneous, that is they focus their purchases in single groups. Table 1 presents clustered instances.

Table 1: Clustered instances (no., %)

Cluster / Instances	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Instances (no.)	2983	1382	344	2259	787	3789	937	389
Instances (%)	23%	11%	3%	18%	6%	29%	7%	3%

Source: Author's own work, .

As a result of the cluster analysis, all instances (transactions) were assigned to the eight groups. The largest number of instances (3789; 29%) was classified in cluster 5, 2983 (23%) instances were assigned to cluster 0, whereas to cluster 3 – 2259 instances (18%). As presented in the analysis, three clusters (0, 3 and 5) had the total of 9031 instances assigned (70%). The remaining 3839 instances (30%) were assigned to clusters 1, 2, 4, 6 and 7.

Table 2: Final clusters (%)

Cluster / Attribute	Full data	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Power industry	14,0	1,6	0,6	3,9	63,3	0,6	6,7	0,9	9,0

[1] For more information about WEKA software, please visit: http://www.cs.waikato.ac.nz/ml/weka.

Gas and water	7,4	6,3	2,3	0,9	0,9	1,6	0,8	71,6	0,5
Electrical wiring	31,8	4,4	2,9	9,8	25,9	6,5	82,9	3,2	21,9
Hydraulic systems	19,8	71,4	6,9	1,9	3,0	4,7	2,6	11,5	2,2
Heating	13,3	11,8	83,8	3,0	1,0	4,3	1,4	8,8	0,9
Industrial	3,5	0,8	0,8	77,7	2,3	0,9	1,8	0,8	2,8
Technical stores	7,1	3,2	2,5	1,4	1,0	80,8	2,3	3,0	1,5
Telecommunications	2,9	0,4	0,2	1,4	2,5	0,7	1,4	0,2	61,1

Source: Author's own work, .

The above table (Table 2) presents assignment of relevant instances to assumed eight groups. Cluster 0 is dominated by universal HEPAC installers (71.4%) and heating installers (11.8%). Cluster 1 – by heating installers (83.3%). Industrial companies (77.7%) were assigned to cluster 2. Cluster 3 includes power systems installers (63.3%) and internal systems installers (25.9%). A separate group are technical stores. They prevail in cluster 4 (80.8%). Also, there is a large group of universal electricity installers (cluster 5; 82.9%). The largest group in cluster 6 are installers of gas and water utilities (71.6%). They are complemented by universal HEPAC installers. The last fixed group is related to telecommunication companies (61.1%) and dealing with installation of electrical wiring (21.9%).

Available software also allows for visualization of the classifications. Picture 5 presents cluster visualizations where:

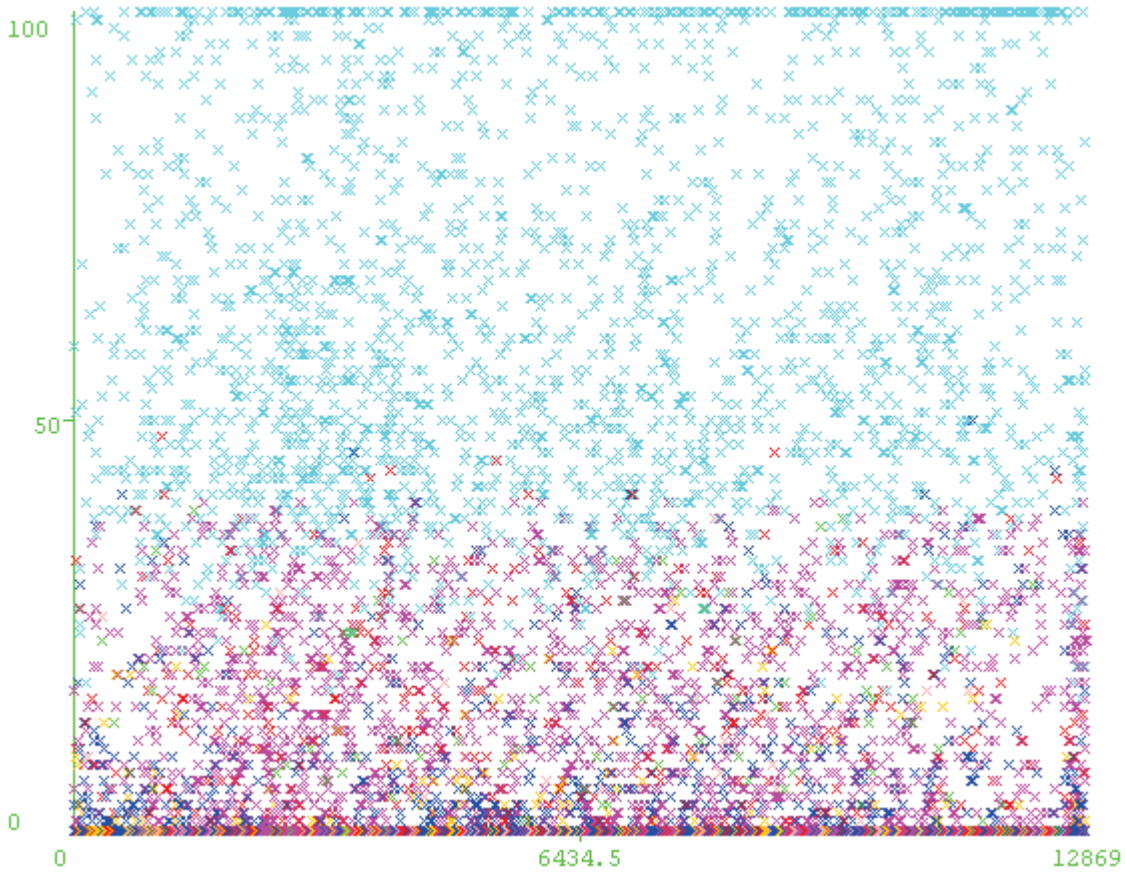
- X – Instance number (Num),
- Y – Power industry (Num),
- Color (see Picture 4): Cluster (Num).

Picture 4: Cluster Visualize – class colour

Source: WEKA software.

Picture 5 provides visualization of instances (X) assigned to relevant clusters (marked with colours in accordance with Picture 4) in the field of Power industry (Y). The picture clearly indicates assignment of a significant majority of instances to cluster 3.

Picture 5: Cluster Visualize - Power industry



Source: Author's own work.

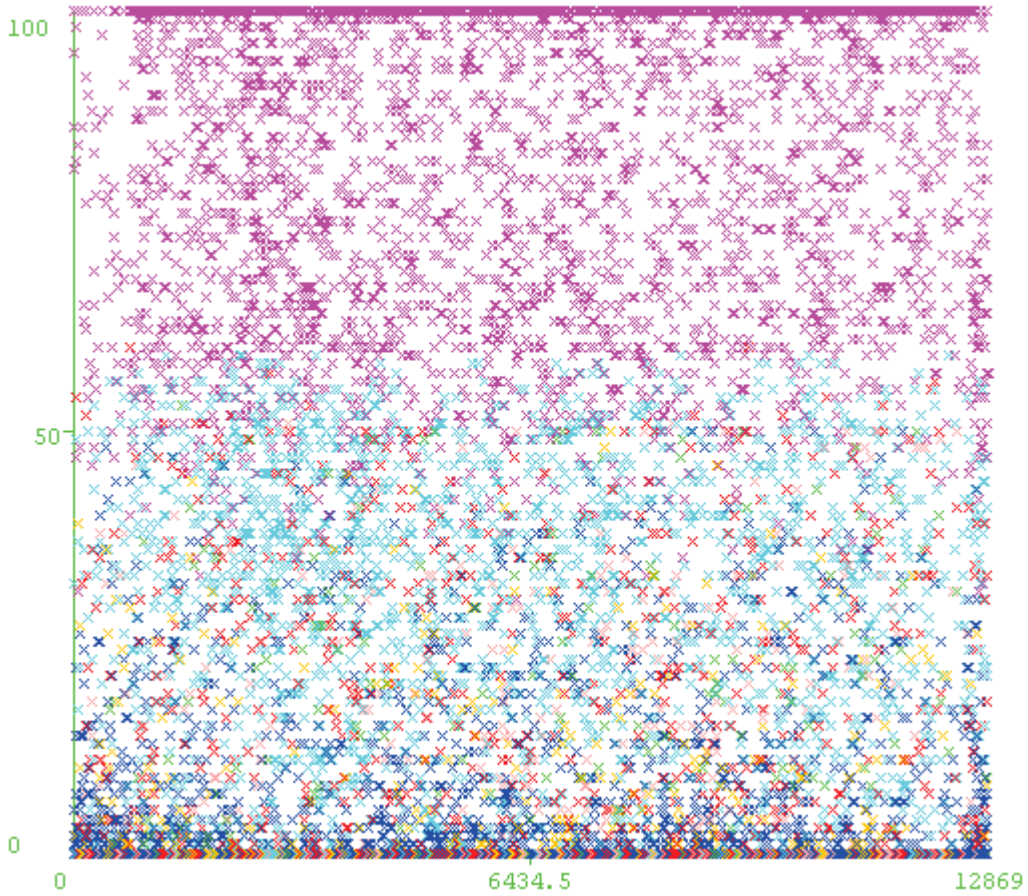
Picture below (Picture 6) presents cluster visualizations, where:

- X – Instance number (Num),
- Y – Electrical wiring (Num),
- Color (see Picture 4): Cluster (Num).

The picture below shows that a large number of instances was assigned to cluster 5. Significant amount of electrical installers was also assigned to cluster 3 and cluster 7 (see Picture 4 to find a corresponding colour).

Clicking on a diagram component allows to obtain numerical visualization of parameters of a given point. This allows for even more detailed examination of the point (a specific instance) in which the researcher/entrepreneur is interested and of the surrounding instances.

Picture 6: Cluster Visualize – Electrical wiring



Source: Author's own work.

5. LIMITATIONS

Performed analysis involved the following limitations:

- the sample pertained to selected groups of products offered by a specific enterprise,
- clusters were derived using a single algorithm – the k-means algorithm,
- the research assumed the existence of exactly 8 clusters.

6. CONCLUSIONS

Enterprise development and a growing number of transactions require proper customer management. Satisfaction of customer needs and expectations has involved the increasing employment of enterprise resources (in particular, human and financial resources). One possibility of reducing company expenses to the minimum is division of customers into a number of (predetermined) groups and adjusting the product offer and customer service to specific needs prevailing in these groups. Separation of a relevant number of groups and assigning customers to each of these groups is a difficult task due to existence of numerous parameters likely to determine the cooperation of the entrepreneur with his customers.

In the paper, the sample of 12 780 customers was used, who were assigned to the eight groups. Performed clustering using the k-means algorithm allowed to assign 70% of customers to three groups. The remaining 30% were assigned to five groups. Due to application of a proper software, numerical values as well as cluster visualizations were obtained (available from different angles), likely to serve as an excellent source of information for the entrepreneur to be used in the decision-making processes (e.g. decisions related to ongoing services, for purposes of marketing campaigns etc.).

7. FUTURE RESEARCH

The authors are planning to conduct further research using the obtained data sample in the following areas:

- clustering, enabling matching of the optimum number of clusters, e.g. using the Celiński-Harabasz Index; currently, research has been conducted as well in which from 5 to 12 clusters were derived,
- using other algorithms during assignment, e.g. EM etc.,
- wider employment of machine learning, especially after adding other qualities characteristic for a given cooperation, such as e.g. volume of purchases, purchases made in a traditional/physical store or online shop, shopping speed or delivery etc.,
- attempts to determine the 'standard sample', that is determination of N, e.g. 100 typical instances of customers, that is samples, including assignment of proper profile names, e.g. industrial companies, telecommunications companies, in terms of what and how they buy,
- launching of learning algorithms, in order to develop a customer classification method,
- performance of classification effectiveness assessments.

REFERENCE LIST

1. De la Paz-Marín, M., Gutiérrez, P.A., Hervás-Martínez, C. (2015). Classification of countries' progress toward a knowledge economy based on machine learning classification techniques, Expert Systems with Applications, 42, pp. 562–572.
2. K-means clustering, Google Trends, https://www.google.pl/trends.
3. Koutsoukis, N-S. (2015). Global political economy clusters: the world as perceived through black-box data analysis of proxy country rankings and indicators, *Procedia Economics and Finance*, 33, pp. 18–45.
4. Lu, Y., He, H., Peng, Q., Meng, W., Yu, C. (2006). Clustering e-commerce search engines based on their search interface pages using WISE-Cluster, *Data & Knowledge Engineering*, 59, 231–246.
5. Machine learning algorithms, Google Trends, https://www.google.pl/trends.
6. Song, Q., Shepperd, M. (2006). Mining web browsing patterns for E-commerce, *Computers in Industry*, 57, pp. 622–630.
7. WEKA, http://www.cs.waikato.ac.nz/ml/weka.