USE OF CLUSTERING FOR CREATING ECONOMIC-MATHEMATICAL MODEL OF AWEB PORTAL

R. Kratochvíl[*], *M. Jánešová*[†]

Abstract: This article describes the mathematical economic model of a communication web portal. To create the model, we use Cluster analysis as one of the areas of artificial intelligence. Based on real data obtained from the operation of the communication web portal and the subsequent identification of the individual data clusters, a model is created that mathematically describes the dependence of economic variables (income from sales of services and selling price of services) on other variables (time of sales of services and field of offered services). Using this analysis, the model clusters together all data to parameterized sets of given properties. The main purpose of creating a model is the suitable classification of data. Consequently, it is possible to streamline the sale of the services and maximize the profits of web portals offering this type of service.

Key words: *web portal, economic-mathematical model, cluster analysis*

Received: 2016-12-15 **DOI:** 10.14311/NNW.2019.29.005
Revised and accepted: 2019-04-15

1. Introduction

The main purpose of creating this model is the suitable classification of data. Consequently, it is possible to streamline the sale of the services and maximize the profits of web portals offering this type of service. Using the model, we should determine when (day / hour) and at what prices it is best to offer the counseling services, so that the income is as high as possible while optimizing the work of consultants (quantity reduction of services offered – saving time for consultants).

By using the economic-mathematical model it will be possible with some degree of probability (according to the Normal distribution using the Gaussian function) to assign fundamental economic quantities of the new service to already created parameterized clusters. These clusters determined by the center and defined by given properties are based on knowledge of the real historical data using the Cluster analysis method [1].

[*]Radek Kratochvil – Corresponding author; Czech Technical University in Prague, Faculty of Transportation Sciences, Konviktská 20, CZ-110 00 Prague 1, Czech Republic, E-mail: kratochvil.radek@seznam.cz

[†]Mária Jánešová; Czech Technical University in Prague, Faculty of Transportation Sciences, Konviktská 20, CZ-110 00 Prague 1, Czech Republic, E-mail: janesova@fd.cvut.cz

Individual clusters in the model consist of data records with similar properties and in a certain way define the given set. The basic parameter is a required income, which is here dependent on the time of sale of the service and the price of the provided service. With the knowledge of this model it is possible to indicate closer or even optimally adjust relevant quantities (especially the time and price of the provided service) depending on the desired income or according to pre-defined prices and times to include the given service in a certain parameterized set.

The real data are obtained from the testing operation of the web communication portal [2]. The web portal offers paid online audio-video consultations that are provided by experts in individual branches of science.

2. Methodology of model creation

To create a mathematical-economic model we have available historical data from a testing version of a web communication portal offering online web consultations [3]. All consultations in individual branches have been in the testing period of one year set at 2 different sales prices, with which the applied analysis subsequently works. These real historical data form the basis for the model creation.

In the calculation and further description of the model we work with incomes / revenues [4], because the notion of profit is rather a part of accounting and in addition its calculation would be relatively difficult (we would have to take into account, for example charges for transferred data depending on the total time segment of the sales and other quantities) [5].

The aim of the model is to create parameterized sets of data records created from real data, which will then form the basis for the inclusion of future data records (here a data record means input values of the new service). Furthermore, the model is able to find a particular optimum (i.e. the best possible solution) to achieve maximum incomes of the web portal. To create the model is used one of the methods of artificial intelligence – namely the Cluster analysis. This method looks for certain relationships within a given set of data records. For input quantities of the model see the Tab. I below.

Index	description	explanation	value
x	price of consultation	CZK per hour	$100, 150 \ldots 800$
y	time (hour)	start of consultation	$9, 10, 11 \ldots 20$
z	commission (income)	commission (20 %)	$1, 2, 3 \ldots$

Tab. I *Basic quantities.*

The aim of the article is to optimize the prices of provided e-consultations [6] with regard to the times and the given profession with the resulting maximizing revenues [7] for the operators of this kind of web portals [8]. The Scilab application [9] was used for the mathematic calculation, which is a freeware for numerical computations with the option of graphical representation.

3. Cluster analysis

Cluster analysis works in two phases. In the first phase, so called preparatory phase – learning, clusters in the data space is generated on the collected data, that characterize a certain way similar data records and thus a certain working mode of the monitored system. In the second phase, the newly arrived data records are compared to the individual clusters and these records are then assigned to clusters that have the highest similarity (in terms of the likelihood function) [1].

The algorithm of model estimation of mixture of distributions was used for classification. Each component here describes a particular cluster. Components are considered as normal. The pointer that is part of the model, indicates the active component relative to the measured new data record. Based on the collected data classification models for each profession offered were created and their classification ability was tested on an independent date.

The applied algorithm operates with the parametric estimation of the probability density. Based on ~~a~~ some prior information, a certain type of probability distribution is considered whereas in the own calculation it only works with the necessary parameters such as normal distribution median value (μ) and standard deviation (σ) [10].

One of the most important clustering tasks is the initial initialization, respectively, the estimation of mixture of components. It is necessary to properly define the initial centers of components, which are actually the centers of the individual Gaussian curves [1]. Another important step is to determine the width of the covariance, respectively, a description of the covariance matrix. Incorrect setting of these values may lead to a total erroneous result and calculated values of weights w_t approaching zero (values xe^{-yy}). We used several calculation methods to a set of initial centers of clusters – homogeneous distribution, distribution with random selection, distribution with random selection in a block. During the testing of each variant of the initial initialization of the determination of the centers of clusters, it has been verified that the distribution with own choice is the most appropriate. The following calculations will be working with this initialization.

The model performs the entire clustering process on the basis of the algorithm repeatedly on the basis of so-called iteration cycles, when the final centers of all clusters move and refine their optimal / best location. After a certain number of cycles the center of clusters has not changed much and we can interrupt an iterative process and determine these values for the final and the best possibility.

4. Use of the clustering algorithm

During my own testing of various options of the initial initialization of the determination of the centers of clusters we verified that the best option is with the distribution by own choice. Therefore, in consequent calculations we work with this initialization. For the own calculation, we use obtained input data specified by quarters (aggregate data in 4 quarters). At the same time, it is always necessary for each variant of the calculation (for individual fields of consultation and the resulting price differences of consultations) to have the covariance matrix set correctly.

We consequently use the method of cluster analysis separately for individual professions. In relation to the number and distribution of measured data, we choose a reasonable number of clusters (range 3–6) for the calculation. Their number can be freely changed according to the current needs and requirements. However, it is necessary to adjust the number of clusters appropriately, as for example in case of a small number of data and a large number of clusters some of these clusters could be occupied by one data record, which is in terms of statistics an entirely unacceptable value. Individual clusters are in charts always marked by different signs. Furthermore the chart illustrates the movement of cluster centers, which vary during the given calculation.

Clustering is created separately for individual fields of consultation, because each area has a totally different time indication, customers want to buy every area of consulting at different times and even end users themselves are another target group of people.

We perform the analysis for two offered fields – legal consulting and English language teaching. These fields differ from each other both in terms of focusing on the target group of people and the price of offered service. Therefore, the results of the performed analysis in these two fields are different and also more vivid.

The analysis carried out five fields of consultation, this data was collected over 1.5 years and the data was used for classification over a period of 1 year. Data are classified by individual days over a period of 12 months. Each day includes times from 9 am to 8 pm, ie. 12 lines of data records (time, cost and number of consultations purchased). The data file has about 4300 data records in each field. The figures show reduced data with total income / profit for the portal for the whole year.

In the calculations it is assumed that the consultant has 80 % of the cost of the consultation and the web portal has 20 %. Always considered in the calculations is income for the web portal, the sum for the whole year.

5. Clustering – legal consulting

In this field, very different testing selling prices of consultations have been set (CZK 400/hour and CZK 800/hour). We therefore chose the total number of clusters to be 6 with the assumption of subsequent distribution of clusters being even between the two price ranges (3 clusters in one and three clusters in the other one).

During each iteration cycle, the final centers of clusters move and refine their optimal / best location. We have completed the iteration process after the third cycle. This number of cycles was sufficient.

The final state after the third iteration is shown below. Of course, using testing you can check that at the suitably chosen initialization even after the first iteration the data are appropriately classified and with a repeated iterative process, centers of clusters do not really move much.

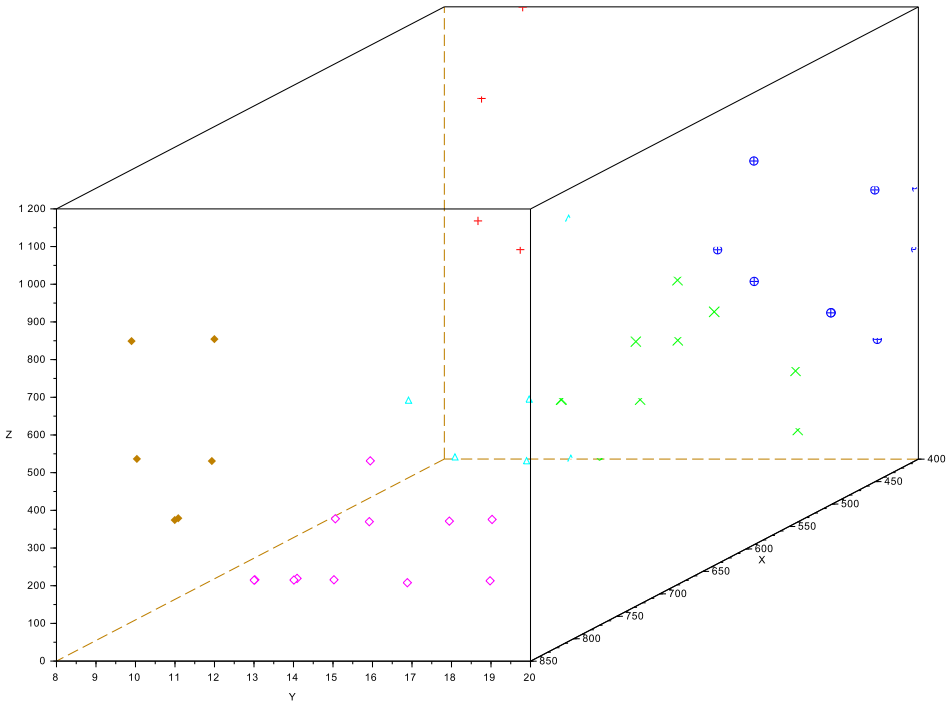

Fig. 1 *3D representation, legal consulting, creating 6 clusters after the 3rd iteration.*

6. Clustering – english teaching

Since in this field a different distribution of sales prices of consultations has been set up (CZK 180/hour and CZK 230/hour), with only very little difference in prices, we decided to implement two variants of analysis implementation. The first option is to set a smaller number of clusters (selected $6/2 = 3$) and higher width of covariance. Only in the second variant is there a higher number of clusters (selected 6) and a smaller width of the covariance. The initial initialization data determining centers of clusters must of course be different.

A) For the first option we consider all the data to determine the best price of the provided consultation. Individual clusters include data from various price levels (only 3 clusters common to both levels (CZK 180 and CZK 230).

   Initializing setting of the covariance matrix:

$$\text{Est} \cdot Cy\,(j) \cdot sd = [50 \times 50, 0, 0;\ 0, 2 \times 2, 0;\ 0, 0, 150 \times 150]$$

   The final state after the fourth iteration is shown in Fig. 2.

B) For the second option we also consider all the data, but we estimate the inclusion of data in clusters in various price ranges (6 clusters – 3 times in CZK 180 and 3 times in CZK 250).

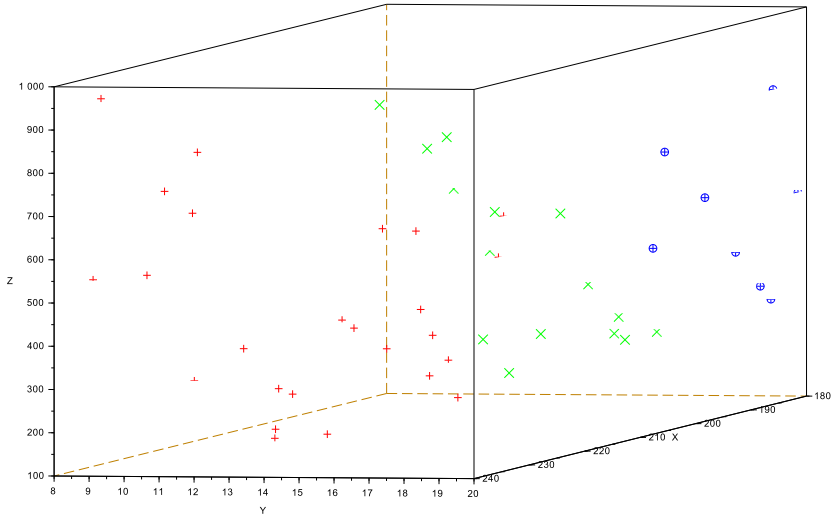

Fig. 2 *3D representation, English teaching, creating 3 clusters after the 4th iteration.*

Initializing setting of the covariance matrix:

$$\text{Est} \cdot Cy\,(j) \cdot sd = [5 \times 5, 0, 0;\ 0, 2 \times 2, 0;\ 0, 0, 150 \times 150]$$

The final state after the third iteration is shown in Fig. 3.

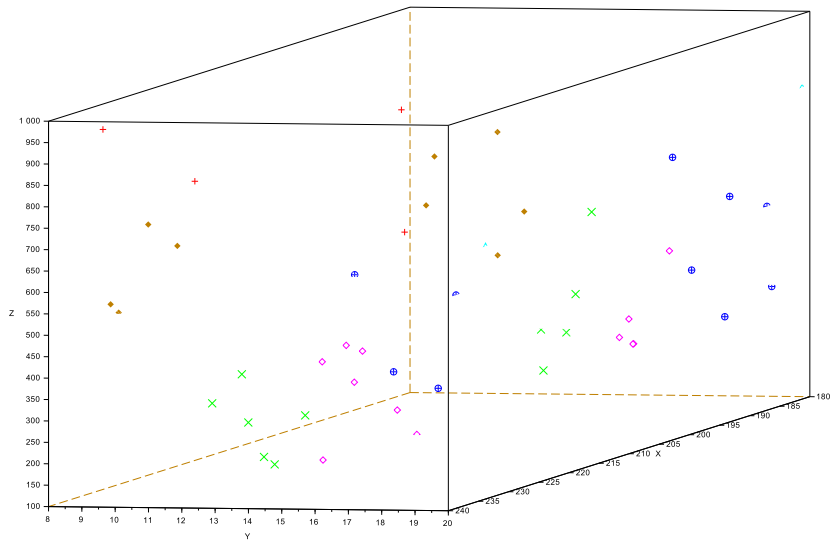

Fig. 3 *3D representation, English teaching, creating 6 clusters after the 3rd iteration.*

7. Evaluation of results

Based on the cluster analysis we created for two selected branches of offered services (legal consulting and English language teaching) sets of data records, can consequently form the basis for the correct setting or including the required parameters of the newly offered services. Data records in each of these clusters always have very similar properties and at the same time quite visibly differ from other clusters.

7.1 Legal consulting – determining of 6 clusters

The analysis obviously indicates that always three clusters are created in each level of value $x1$ (i.e. 3 times for $x1$ = CZK 400 and 3 times for $x1$ = CZK 800), and it is clearly seen that the clusters are also distributed according to the amount of revenues with a clear spacing (e.g. for $x1 = 400$ values it is the distribution of centers at 257.1, 579.5 and 880.8).

The performed cluster analysis indicates that the highest incomes are achieved for the price of CZK 400/hour around the center of the cluster with the coordinates [400.1, 10, 880.8] and for the price of CZK 800/hour around the center of the cluster with the coordinates [799.9, 18.5, 625.2], i.e. around 10 hours / 18–19 hours.

7.2 English teaching – determining of 3 clusters

The analysis shows that each cluster includes the data records from various price levels $x1$ (CZK 180 and CZK 230), and it is clearly seen that the clusters are divided according to the amount of income with a distinct spacing.

The performed cluster analysis indicates that the highest incomes are achieved at the average price of CZK 208/hour around the center of the cluster with the coordinates [208.1, 10.1, 696.1].

7.3 English teaching – determining of 6 clusters

The analysis obviously indicates that always three clusters are created in each level of value $x1$ (i.e. 3 times for $x1$ = CZK 180 and 3 times for $x1$ = CZK 230), and it is clearly seen that the clusters are also distributed according to the amount of revenues with a clear spacing (e.g. for $x1 = 180$ values it is the distribution of centers at 356.8, 546.0 and 707.6).

Clusters	centers	probability
1	[400.1, 10.0, 880.8]	0.0700112
2	[400.3, 14.1, 257.1]	0.2728525
3	[400.2, 18.1, 579.5]	0.1472782
4	[799.8, 10.6, 476.9]	0.1472782
5	[799.9, 15.0, 270.0]	0.2310426
6	[799.9, 18.5, 625.2]	0.1197723

Tab. II *Legal consulting – determining of 6 clusters.*

Clusters	centers	probability
1	[208.1, 10.1, 696.1]	0.2296971
2	[205.0, 14.5, 331.3]	0.5114558
3	[204.3, 18.5, 528.1]	0.2588471

Tab. III *English consulting – determining of 3 clusters.*

The performed cluster analysis indicates that the highest incomes are achieved for the price of CZK 180/hour around the center of the cluster with the coordinates [180, 10.1, 707.6] and for the price of CZK 230/hour around the center of the cluster with the coordinates [230, 10.3, 670.0], i.e. around 10 hours for both.

Clusters	centers	probability
1	[180, 10.1, 707.6]	0.0821087
2	[180, 14.1, 356.8]	0.2661987
3	[180, 18.5, 546.0]	0.1414885
4	[230, 10.3, 670.0]	0.1464508
5	[230, 15.3, 330.7]	0.2753323
6	[230, 18.7, 556.6]	0.0884210

Tab. IV *English consulting – determining of 6 clusters.*

8. Conclusion

A model mathematically describing the web communication portal was created using the cluster analysis method. This analysis is applied for two selected services provided via online consultations, namely for legal consulting and teaching English [11]. Clusters created through the analysis contain data records describing the prices of consultations, the times when the service is provided, and the resulting income from this realized service.

This analysis creates for each field of provided services several parameterized sets (pre-defined 3 or 6 clusters) having some common characteristics, while differing from other sets as much as possible. These resulting sets of data records determined by the center and having a certain probability of inclusion define how often is the offered service used and what is the income arising from it. At the same time these individual clusters determine sets with the highest or lowest incomes.

The aim of the model is to create thus parameterized sets of data records created from real data, to which it is consequently possible with a certain degree of probability (according to the Normal distribution) to assign the input quantities of the new service.

Using the model, we can determine when and at what prices it is best to offer the given counseling services so that the income is as high as possible. Simultaneously, we can optimize the work of consultants, for example by reducing their services. With the proper setting of the number of clusters, we can divide in a desired way

the time period (in our case a day) to individual zones / sets of data records. These resulting sets are crucial in determining the generated income.

The article analyzes 2 offered services – English language teaching and legal counseling. For the inclusion of new services, a model illustrating similar services can be used. For example, a model for the English language may be used for German language teaching, the result will probably be very similar. On the contrary, the legal counseling model can be used for accounting and financial advice. This was verified by a service analysis. Similarly, we can create a model for mathematics for university or high school, this model will be very similar for physics or chemistry for university or high school. It is very important to identify a target group of buyers – students, retirees, people of working age, people / companies, etc.

By analyzing and subsequently creating the model it was verified that the number of variables was sufficient. Using the model, we can optimize the times and costs of consultations and save time for consultants. Creating a model has positive benefits for the provider of a communications web portal in terms of maximizing profits and saving time.

With knowledge of the model, selling prices and offered times of provided services can be preset in a sophisticated way with an emphasis on maximizing incomes or classify the given service according to predetermined prices and times into a parameterized group. This model is – in terms of its focus – designed for providers of online consultation Web portals or vendors of other similar services.

References

[1] NAGY I. *O bayesovském učení*. Ivan Nagy. Available from: http://eridanus.cz/id32402/ekonomika/pru%287mysl/vy%281poc%282etni%281_technika/software/ume%282la%281_inteligence/expertni%281_syste%281my/au070256.pdf

[2] KRATOCHVIL R. *Videokonzultace* Videokonzultace [viewed 2015-10-15]. Available from: http://www.videokonzultace.cz

[3] QUITTKAT C. The European Commission's Online Consultations: A Success Story? *JCMS: Journal of Common Market Studies*. 2011, 49(3), pp. 653–674, ISSN 1468-5965.

[4] MARTIN B., XUEMEI T. *Books, Bytes and Business: The Promise of Digital Publishing*. Farnham, GB: Routledge, 2016. ISBN 978-0-7546-9653-7.

[5] ALTMANNJ,VEIT D., ed. *Grid economics and business models: 4th international workshop, GECON 2007, Rennes, France, August 28, 2007: proceedings*. Berlin; New York, Springer, 2007. Lecture notes in computer science, 4685. ISBN 978-3-540-74428-3.

[6] HAYTHORNTHWAITE C. *E-learning theory and practice*. London: Sage Publications, 2011, ISBN 978-1-84920-471-2.

[7] DUCHOŇ B. *Inženýrská ekonomika*. Praha, 2007, ISBN 978-80-7179-763-0.

[8] BELALEM G., BOUAMAMA S., SEKHRI L. An Efficient Economic Model user-oriented for Cloud Computing. *International Journal of Computer Applications*. 2011, 15(2), [viewed 2016-06-21]. ISSN 09758887. Available from: https://www.researchgate.net/publication/50946216_An_Efficient_Economic_Model_user-oriented_for_Cloud_Computing

[9] PECHERKOVÁ P., NAGY I. *Vybrané funkce v programu Scilab z oblasti pravděpodobnost a statistiky*. Praha, 2015, learning material.

[10] PILMANN J. *Diplomová práce - Data a jejich klastrování* Pilmann. Available from: https://is.cuni.cz/webapps/zzp/detail/64298.

[11] CAFFERY L., SMITH C.A. SCUFFHAM P.A. *An economic analysis of email-based telemedicine: A cost minimisation study of two service models. BMC Health Services Research* [online]. Available from: http://bmchealthservres.biomedcentral.com/articles/10.1186/1472-6963-8-107.