Applied aspects of data mining for decision support at the regional health system

G G Rapakov¹, A A Sukonshchikov¹, A N Shvetsov¹, V A Gorbunov¹ and O Ja Kravets²

¹ Vologda State University, 15, Lenin str., Vologda, 160000, Russian Federation
² Voronezh State Technical University, 14 Moscow ave., Voronezh, 394026, Russian Federation

E-mail: grapakov@yandex.ru

Abstract. The article presents the results of research of Data Mining methods with Microsoft SQL Server. Microsoft Clustering algorithm was used for improving the effectiveness of medical prevention and treatment in a cohort of patients with arterial hypertension. There are rationales for monitoring of cardiovascular risk and desire to correct the risk with Data Mining at medical decision support systems. Authors used medical and sociological monitoring data from regional clinical hospital. The segmentation of arterial hypertension patients was performed using Microsoft Clustering algorithm. As a result, a quantitative assessment of the population profile for patients with arterial hypertension was obtained. The authors presented diagrams and profiles of clusters. They were compared. The developed approach is applied for decision support at regional health information management system for reduce of cardiovascular risk.

1. Introduction
Since 1998 in the Vologda region acts Regional Target Program (RTP) «Prevention and treatment of the arterial hypertension among the population of the Vologda region». RTP allowed to begin operations on early detection and well-timed treatment of patients with an arterial hypertension (AH), and also to shaping of register AH. AH is the risk factor of development the circulatory system diseases (CSD): an coronary heart disease, cerebrovascular illness, a lesion of peripheral arterial system. Abundance AH among adult population makes about 35–40 %. At an estimation of a level of economic damage from CSD it is accepted to part a factor cost and losses in the economy, associated with CSD. Into composition of a factor cost enter: emergency call expenditures, outpatient clinic and hospitalization, rendering of high-technology medical aid and medicamental therapy. Losses in economy include: losses of gross national product owing to death in able-bodied age, a temporary invalidity and payments of disability pensions. The summarized economic damage from the diseases associated only with an acute coronary syndrome in the Russian Federation exceeded 70 billion rubles a year [1].

The purpose of the present research is to study the possibilities of using the Data Mining approach for implementation in real medical practices and to advance outcomes of regional health decision support systems. The practical significance of the work is determined by the possibility of using the intelligence systems for efficiency improvements of prevention and treatment of the arterial hypertension among the population of the Vologda region and creating a health-preserving space of active longevity. The use of computer modeling makes it possible to increase the effectiveness of the
regional health information management system (IMS) for the purposes of medical prevention of arterial hypertension in the system of medical and social support of the population and when creating a health-preserving space of active longevity, which causes the novelty of the work.

2. Methods

The Department of Public Health Services of the region was developed an expert estimation cards of treatment efficiency for the patient with AH. It gives the estimate of therapeutic effectiveness the measures for decrease of cardiovascular risk by the medical–sociological research. The efficiency of estimation is fulfilled with usage of an ordinal scale. Necessary accuracy of an estimation within 0.05 with a confidence probability $\alpha = 0.95$ is ensured.

In the end of the XX–th century for traditional methods of statistic analysis there was an alternative in the form of approaches of Data Mining or Knowledge Discovering in Databases. The causal factor was the growth of volumes of the stored inhomogeneous information. The main tasks of intellectual data analysis are: detection of the latent rules and regularities (descriptive) and creation of the prognosis (predictive). Extracted knowledge is presented in the form of models (clusters, decision trees, etc). For creation of models use the methods of artificial intelligence. Cluster analysis is the descriptive task of Data Mining [2, 3, 4].

In the present research was exercise the segmentation of patients AH with usage basis of medical–sociological indices by the algorithm of clustering in Microsoft implementation — Microsoft Clustering (figure 1). The cleaning and preprocessing of the data was executed for support of its demanded quality. Also the counterparts test, contradictions and anomalous value check was carried out.

![Figure 1. Charts and profiles of clusters.](image-url)
Intellectual data analysis has been fulfilled with Microsoft SQL Server 2012 Business Intelligence. The algorithm parameters were varied in research. By means of the heuristic approach the data set has been parted into three homogeneous groups that come to agreement with an expert estimation. The method of a scalable maximization of waiting in MS Clustering can be applied to small samplings. The expectation–maximization algorithm uses a likelihood function logarithm for match making of probabilistic model to data. Taking into account a sample size the value used for determination of the model truncation has been reduced to two. Cross validation showed model fastness.

The diagram of clusters shows degree of relationship between clusters. Color filling with attribute Efficiency Estimation displays similarity measure (figure 1). Comparing of clusters allows to match the probable values of attributes mutually. The classifying is executed in the order of importance level for cluster attribute (figure 2) [5, 6, 7].

Figure 2. Comparing of clusters.

3. Results and discussion

On basis of clusters profiles that contains the distributions of attributes discrete values the interpretation of the grouping was performed (figure 1). For clusters #1 (Cure rate is sufficient) and #3 (Cure rate is not sufficient) the likelyhood values of attributes are tabulated (table 1).

Thus, on the basis of attributes importance (the probability of their appearance in a cluster) should draw following conclusions:

- sufficient cure rate is appreciably caused by information about arterial hypertension corrected by non-drug means to patients (0.949 – for the Cluster #1 and 0.253 – for the Cluster #2). At each visit of the patient to an outpatient department, admission to hospital or discharge from the hospital, it is necessary for doctor to pay attention to practical usage of non-drug means by AH patients;
- self–preservation activity of patient demands availability a tonometer and permanent self–checking of arterial pressure;
- efficiency of antihypertensive therapy depends from quality of attainment of arterial pressure target level by patient;
efficiency of antihypertensive therapy increases in case of detection of AH patients in an outpatient department at early stages of disease;

- the patients learning at schools of health and AH register maintenance favour the decrease of cardiovascular risk.

- the cluster №3 differs the high value of attributes — Lack of Reduction for Influence of Principal Risk Factors – from 0.788 to 1.0;

- efficiency of treatment for AH patient especially depends upon the specialist of first-referral clinic. In case of unsufficient cure rate (Cluster #1 changes to Cluster #2) the probability of the patient observation in an outpatient department by therapist decreases with 0.716 to 0.461 and the likelihood of observation by family doctor increases with 0.166 to 0.395.

Table 1. The values of the likelihood for attributes of clusters.

| Value of attributes                                                                 | Cluster #1 | Cluster #3 |
|-------------------------------------------------------------------------------------|------------|------------|
| Antihypertensive therapy is assigned at outpatient department                        | 0.924      | 0.663      |
| Antihypertensive therapy is permanently                                             | 0.958      | 0.759      |
| The patient is included in AH register                                              | 0.952      | 0.449      |
| Attainment of arterial pressure target level (systolic pressure <140, diastolic pressure <90) | 0.794      | 0.255      |
| The patient is observed in outpatient department by the therapist                   | 0.716/0.166| 0.461/0.395|
| / the family doctor                                                                  |            |            |
| The patient uses a tonometer                                                        | 0.910      | 0.418      |
| Training course at school of health for AH patients is transited                     | 0.478      | 0.195      |
| Unregistered as a disabled person                                                    | 0.992      | 0.799      |
| Cure rate is sufficient                                                              | 1.000      | 0.098      |
| Arterial hypertension corrected by non-drug means is well-known                     | 0.949      | 0.253      |
| Regular medical check-up                                                             | 0.973      | 0.603      |
| The second stage of arterial hypertension                                           | 0.876      | 0.488      |

4. Conclusion

The application of Microsoft Clustering algorithm with Microsoft SQL Server 2012 Business Intelligence allowed to fulfill the quantitative estimation of a population profile for AH patient. Intellectual models of the analysis ensured cohorts typology of AH patients by the complex estimation of the data of medical–sociological monitoring. The results of computer modeling in the regional health information management system makes it possible to increase the efficiency of preparing, making and executing management decisions during the implementation of the concept of active longevity and the formation of a municipal health-saving environment. Further research prospects are related to the geospatial data mining of disease cases.

The results can be used for cloud technologies in functionality of policy control at the network edge [8], complex-structured objects optimization during modeling on the population algorithms adaptation basis [9], information security risk estimation for cloud infrastructure [10], multiagent technology for the application of a distributing function for load balancing in multiserver systems [11], SOA reference model [12], a way to accelerate the process of gathering information for decision-making [13], studying and characterization of the data flows in an IP-Based Network [14], a survey of informatization and privacy in the digital age and basic principles of the new regulation [15].

References

[1] Rapakov G G and Banshchikov G T 2014 Efficiency of implementation of the regional target program for the treatment of patients with arterial hypertension at the regional level (experience of the Vologda oblast) Economic and social changes: facts, trends, forecast 5(35)
206-21

[2] Wang J 2009 *Encyclopedia of Data Warehousing and Mining* (Hershey: Information Science Reference) 2122

[3] Zhang C and Zhang S 2002 *Association Rule Mining: Models and Algorithms* (Berlin: Springer-Verlag) 248

[4] Wu X 2009 *The Top Ten Algorithms in Data Mining* (Boca Raton: Chapman & Hall/CRC) 201

[5] Tang Z and MacLennan J 2005 *Data Mining with SQL Server 2005* (Indianapolis: Wiley Publishing) 460

[6] MacLennan J, Tang Z and Crivat B 2008 *Data Mining with Microsoft SQL Server* (Indianapolis: Wiley Publishing) 744

[7] Bassan A S and Sarkar D 2014 *Mastering SQL Server 2014 Data Mining* (Birmingham: Packt Publishing) 289

[8] Atanasov I, Pencheva E, Nametkov A and Trifonov V 2019 On Functionality of Policy Control at the Network Edge *International Journal on Information Technologies and Security* 3(11) 3-24

[9] Lvovich Ya E, Tishukov B N, Preobrazhenskiy A P, Pitolin A V and Kravets O Ja 2019 Complex-Structured Objects Optimization During Modeling on the Population Algorithms Adaptation Basis *International Journal on Information Technologies and Security* 3(11) 41-50

[10] Tsaregorodtsev A V, Kravets O Ja, Choporov O N and Zelenina A N 2018 Information Security Risk Estimation for Cloud Infrastructure *International Journal on Information Technologies and Security* 4(10) 67-76

[11] Kravets O Ja and Oleinikova S A 2014 Multiagent technology for the application of a distributing function for load balancing in multiserver systems *Automation and Remote Control* 75(5) 977-82

[12] Memeti A, Azizi A and Luma-Osmani S 2019 Human Resources Management System: SOA Reference Model. *International Journal on Information Technologies & Security* 4(11) 29-38

[13] Mihaylov D 2019 A Way to Accelerate the Process of Gathering Information for Decision-making *International Journal on Information Technologies and Security* 4(11) 39-50

[14] Nedyalkov I, Stefanov A and Georgiev G 2019 Studying and Characterization of the Data Flows in an IP-Based Network *International Journal on Information Technologies and Security* 1(11) 3-12

[15] Romansky R 2019 A Survey of Informatization and Privacy in the Digital Age and Basic Principles of the New Regulation *International Journal on Information Technologies and Security* 1(11) 95-106