doi:10.20944/preprints202007.0078.v1

Peer-reviewed version available at *Mathematics* **2020**, *8*, 1211; doi:10.3390/math8081211

Article

Data Driven Analytics for Personalized Medical Decision Making

Nataliia Melnykova[1], Natalya Shakhovska[1*], Michal Gregus[2], Volodymyr Melnykov[3], Mariana Zakharchuk[4], Olena Vovk[1]

[1] Artificial intelligence department, Institute of computer sciences and information technologies, Lviv Polytechnic National University, Lviv, Ukraine, melnykovanatalia@gmail.com
[2] Department of Information Systems, Comenius University in Bratislava, Bratislava, 81499, Slovakia; michal.gregusml@fm.uniba.sk
[3] General Surgery Department, Danylo Halytsky Lviv National Medical University, Lviv, Ukraine; v.melnikov2013@gmail.com
[4] Linguistically Education Center, Lviv Polytechnic National University, Lviv, Ukraine, mariana.y.zakharchuk@lpnu.ua
* Correspondence: nataliya.b.shakhovska@lpnu.ua

Abstract: The study was conducted on applying machine learning and data mining methods to personalizing the treatment. This allows investigating individual patient characteristics. Personalization is built on the clustering method and associative rules. It was suggested to determine the average distance between instances for optimal performance metrics finding. The formalization of the medical data pre-processing stage for finding personalized solutions based on current standards and pharmaceutical protocols is proposed. The model of patient data is built. The paper presents the novel approach to clustering built on ensemble of cluster algorithm with better than k-means algorithm Hopkins metrics. The personalized treatment usually is based on decision tree. Such approach requires a lot of computation time and cannot be paralyzed. Therefore, it is proposed to classify persons by conditions, to determine deviations of parameters from the normative parameters of the group, as well as the average parameters. This made it possible to create a personalized approach to treatment for each patient based on long-term monitoring. According to the results of the analysis, it becomes possible to predict the optimal conditions for a particular patient and to find the medicaments treatment according to personal characteristics.

Keywords: personalization, decision making, medical data, artificial intelligence, Data-driving, Big Data, Data Mining, Machine Learning

---

1. Introduction

Recently everyone is seeing that modern medicine is on the verge of change in many countries. It is very important to support medical decisions regarding analysis, prediction of the patient's condition at different stages of his treatment. Thus, in recent years, the health policy has been changing the form of intervention aimed at preventing and self-responsibility of people by taking measures to increase the level of medical literacy. The various programs are developed to facilitate the widespread adoption of integrated digital solutions that will improve the quality of life of citizens while demonstrating a significant improvement in the efficiency of healthcare delivery. Particularly important for the implementation of these programs are the literacy and knowledge of medical staff, the completeness and timeliness of receiving information about patients, the ability to monitor the patient not only in the hospital. This is why information technology, smart devices and systems are being integrated into medical and social areas. It is important that the technologies used for such solutions are tailored to meet the requirements of international standards within personalized

Peer-reviewed version available at Mathematics 2020, 8, 1211; doi:10.3390/math8081211

primary care for patients or comprehensive solutions for the diagnosis or treatment of various diseases [1, 2, 3].

As an example, the system of primary healthcare serves almost 70% of the population, treating about 90% of common and local problems. Therefore, for qualified medical care, it is necessary to develop intelligent systems for personalized treatment decisions. Such intelligent systems will be responsible for collection, processing and transmitting a large amount of data using set of microcontrollers, communication equipment, sensors based on Internet of Things (IoT) solution. Although artificial intelligence (AI) will help to solve the preprocessing tasks, such as gaps imputation, outliers analysis, complexity reduction, hidden dependencies finding etc. [4, 5, 6, 7]. Artificial intelligence and Big Data will be used together for improving the prediction accuracy of spatial and temporal development of the disease in patients [8, 9, 10].

The aim of the paper is to develop the novel approach for supporting of medical decision based on ensemble of clustering methods. The optimal condition of the patient will be estimated as distance between centroid in cluster and particular instance.

The structure of the paper is following. The first part, related sources, presents approach for medical data analysis. After that, the model of patient data and description of existing dataset is given. The estimation of quality metrics for the existing method is provided. The novel approach based on ensemble of the clustering methods is developed. The results of proposed approach is discussed. The conclusion section underline the novelty of the proposed approach.

2. Materials and Methods

2.1. Related sources

Let us take into account the following aspects of medical data processing:

- Missing data imputation,
- Diagnostics (classification and prediction),
- Clustering and personalizing the treatment.

P. Bidiuk [11] uses decision tree-based computation procedures and regression approach to predict missing data, analyze the nature of data gaps and fill out these gaps. Dangare [12] obtained similar results for associative rules mining in medical data.

Bayesian networks, ANN, k-means algorithms are used in [13] for heart disease prediction. However, Bayesian networks are too slow for processing the huge amount of data and for on-line diagnostic. That is why Y. Tang [14] developed a method of paralleling Bayesian networks. However, even in the case of parallelism, it is advisable to use Bayesian networks in combination with other methods of machine learning for multi-parameter, large-scale and dynamic medical data flows. The apparatus of artificial neural networks, with using of fuzzy logic, is also actively used for the analysis of various medical data. Thus, in the works of E. Bodiansky, I. Perova [15 – 18] a system of rapid medical diagnostics based on auto-associative neuro-fuzzy memory was proposed. However, one of the urgent tasks remains to improve the accuracy of the results of the classification problem. In addition, the problem of imbalance of input data, as well as small samples of data collected manually by medical staff, impose a number of restrictions on the use of existing methods and tools of computational intelligence to solve such problems [19].

The cluster analysis is widely used for outliers finding. The outliers in medical domain mean the differences between optimal patient’s conditions based on the local protocol and individual features. One of the simples clustering algorithms are partitioning methods. K-means algorithm builds k clusters located at great distances from each other. The main type of problems solved by the k-means algorithm is the assumption (hypotheses) about the number of clusters, and the diversity of the instances in different clusters. The choice of k number may be based on the results of previous studies and theoretical considerations [20].

For BIRCH algorithm (Balanced Iterative Reducing and Clustering using Hierarchies), the clustering rate increases due to the generalized appearance of clusters. This algorithm implements a two-stage clustering process. During the first stage, a preliminary set of clusters is formed. In the second step, other clustering algorithms, suitable for RAM, are applied to the detected clusters [21].

The DBSCAN algorithm is used to find clusters of different sizes and shapes. If the object boundary contains more points than the minimum number of objects, a new cluster with a root object is created. DBSCAN iteratively collects objects directly close to the root objects that can merge several tightly accessible clusters. DBSCAN does not require predefining the number of clusters received, unlike partitioning methods. Although there is a need for guidance on the values of the radius parameters of any object and the minimum number of objects that directly affect the clustering result. The optimal values for these parameters are difficult to determine, especially for multidimensional data spaces [22].

Therefore, the search for personalized solutions involves a number of interrelated processes, namely:

- data collecting (physical parameters of the patients) from smart sensors, for example, electronic glucometers, automatic tonometers, holters, etc.;
- data consolidation is focused on defining the data models, as well as the formation of data flows in accordance with the defined tasks;
- data processing consists in validating data, storing data like a database, knowledge base or data warehouses, while ensuring data protection;
- data analysis process provides sorting of data by defined parameters, taking into account Data Mining methods such as cluster analysis methods, including k-means, k-median, algorithms; methods of data visualization, etc.,
- the forecasting process involves the construction of a test object that allows the use of artificial intelligence methods, namely: artificial neural networks; decision trees; Bayesian networks; linear regression, correlation-regression analysis; methods for finding associative rules, including the Apriori algorithm; it is final to carry out an assessment of the decisions made on their basis.

The schema of the medical data processing procedure is given on Figure 1.

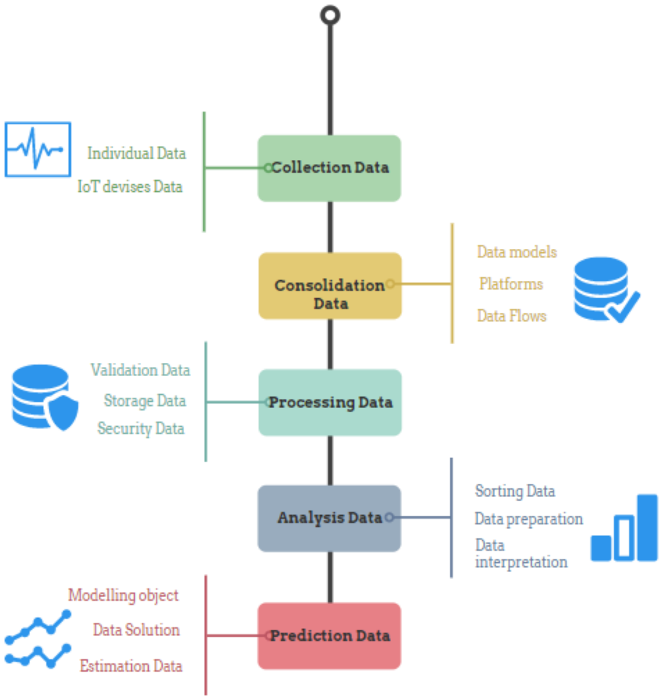

Figure 1. The dependency of patient medical data research stages for their intellectualization on personalized decision-making.

The predicting the next state of an object consists of the following tasks:

- the monitoring of the patient's condition during long time,

Peer-reviewed version available at Mathematics 2020, 8, 1211; doi:10.3390/math8081211

- finding the optimal condition, based on an assessment of patient's quality of life,
- searching for personalized treatment and predicting the results of its use, taking into account the patient's time-dependent and time-independent characteristics, international treatment standards and current pharmaceutical protocols.

2.2. Model of patient data

The patient data is characterized by heterogeneity, which complicates the analysis stage. That is why it is necessary to formalize the physical state of the patient (FS) by the set of parameters A, consists of time-dependent A_t and time-independent A_{in} parameters.

$$FS = f(A), \quad (1)$$

$$A_t \cup A_{in} = A,$$

Dataset consists of information about 36 patients of Lviv hospital. Data is collected during their hospitalization period. The list of time-independent characteristics is following: ID, sex, age, weight, diagnose (list of values), sub-diagnose (list of values), flora (list of values), active substance (list of values), medicament (list of values). The task is to find the most appropriative medicaments. The medicaments list is based on medical protocols and diagnose. For each diagnose the amount of medicaments is more than 50 items. We must eliminate the medicaments from this list based on analysis of other characteristics. The list of time-dependent characteristics are the following: temperature, blood pressure, heart rate.

The processing and analysis of data to find personalized medical solutions APD is represented as a pair:

$$APD = \{GS, A \cup D\}, \quad (2)$$

where $GS = \bigcup_{i=1}^{n} FS_i$, GS - is the set of the patient's physical states, A is the set of the patient's personal data, n - is the total number of patient characteristics, D is the set of personalized decisions.

The physical condition is represented as

$$a_i: GS_i \rightarrow V_d, \quad (3)$$

where V_d is the set of values of the attribute d_i according to the treatment protocol.

Attributes of the set A are called conditions, and D is the solution, d_i - is the value of the i-th solution obtained from the set D.

The equation of the state looks like given below:

$$\vec{FS}(t) = A * \vec{FS}(t) + B * \vec{S}(t), \quad (4)$$

where $\vec{FS}(t)$ is the state vector of the dimension of space n, which includes object variables that uniquely determine its state:

$$\vec{FS}(t) = \begin{bmatrix} FS_1(t) \\ FS_2(t) \\ \dots \\ FS_n(t) \end{bmatrix}, \quad (5)$$

$\vec{S}(t)$ is the control vector that displays signals affecting the system from the outside, through proposed solutions to determine the therapeutic scheme of treatment:

$$\vec{S}(t) = \begin{bmatrix} S_1(t) \\ S_2(t) \\ \dots \\ S_n(t) \end{bmatrix}, \quad (6)$$

A, B are the parameter matrices consist of object parameters [23, 24].

So, the medical data is represented of multidimensional dataset. Therefore, the next step is optimal state evaluation based in time-independent characteristics.

2.3. The analysis based on existing clustering method

In this section, the existing methods was compared with proposed one. The outliers were analyzed for personalized treatment.

First, the existing methods of clustering are used. The k-mean algorithm allow us to evaluate the shape of the clusters and to find the outliers. To do this, the gap statistics is used for estimating of the most appropriative number of clusters. The best number is equal 3 (Fig 2). The total sum of square value is 43.5 %.

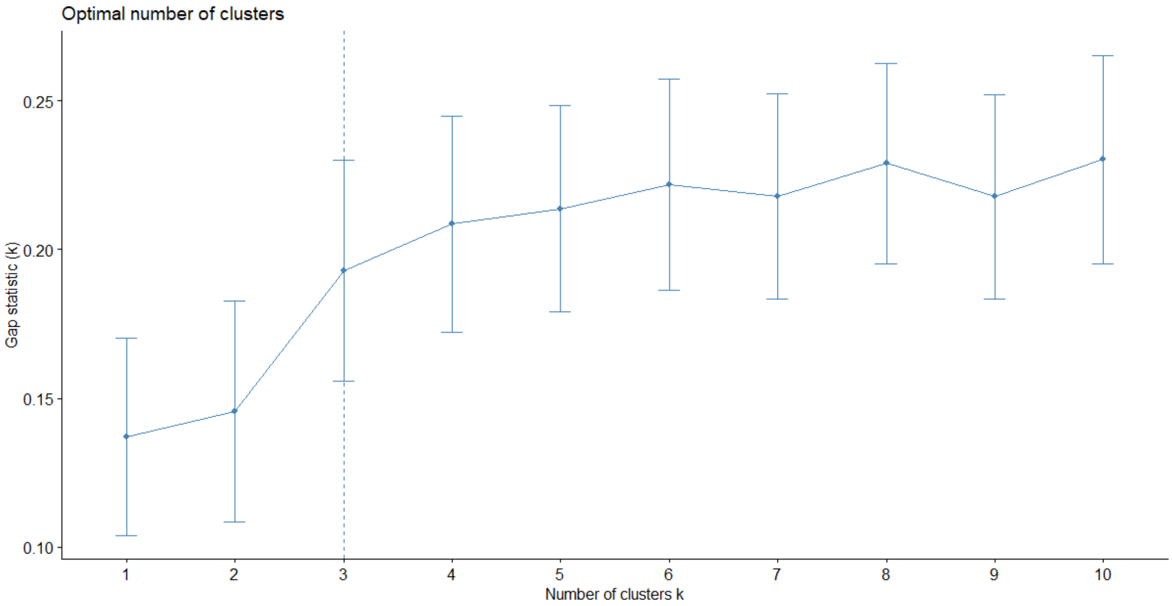

Figure 2. The applying the gap-statistics method to find the optimal number of clusters.

The realization was performed using RStudio. The shape of three clusters is given on Figure 3.

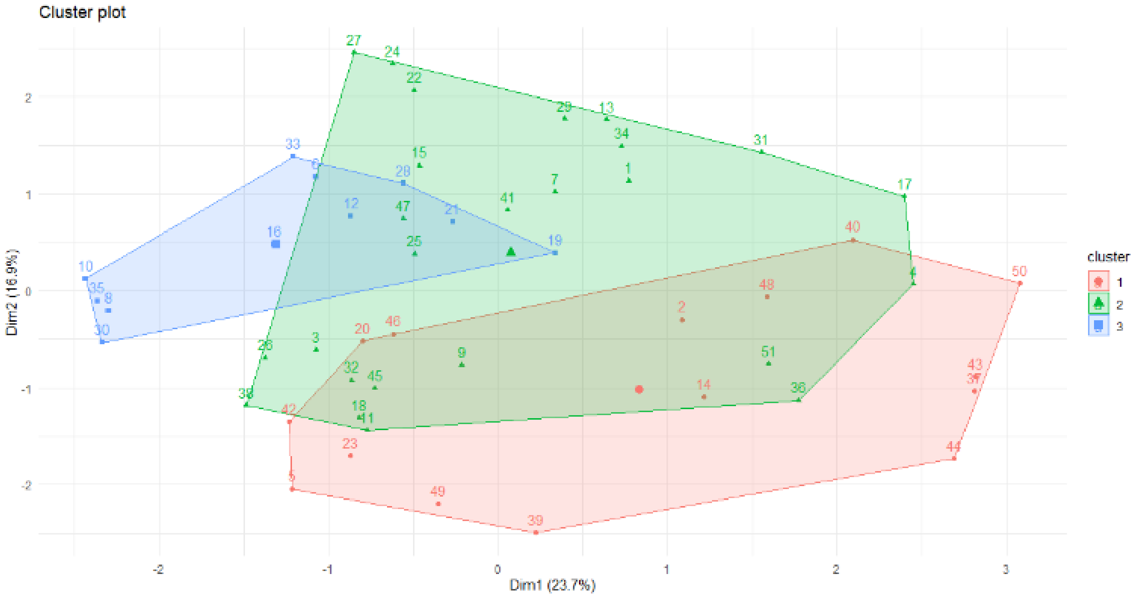

Figure 3. The applying the k-means method to determine the distribution of elements between clusters.

The overlapping of the object is presented. That is why fuzzy c-mean was used for data analysis. The membership function is given below. The minimum difference between values of the membership function for one instance is 0.01. We can see the overlapping of the clusters for the following objects:

```
Memberships:

          1           2           3
```

Peer-reviewed version available at Mathematics 2020, 8, 1211; doi:10.3390/math8081211

```
 [3,] 0.16711487 0.35000991 0.48287522
[15,] 0.23002867 0.32066047 0.44931086
[16,] 0.39445021 0.21628065 0.38926914
[22,] 0.38334426 0.32979465 0.28686109
[32,] 0.30323223 0.29632674 0.40044103
[40,] 0.27174345 0.43061645 0.29764011
```

Let us evaluate the quality of clustering. Visual Assessment of cluster Tendency metric (VAT) [24] shown low propensity to form groups (Fig. 3).

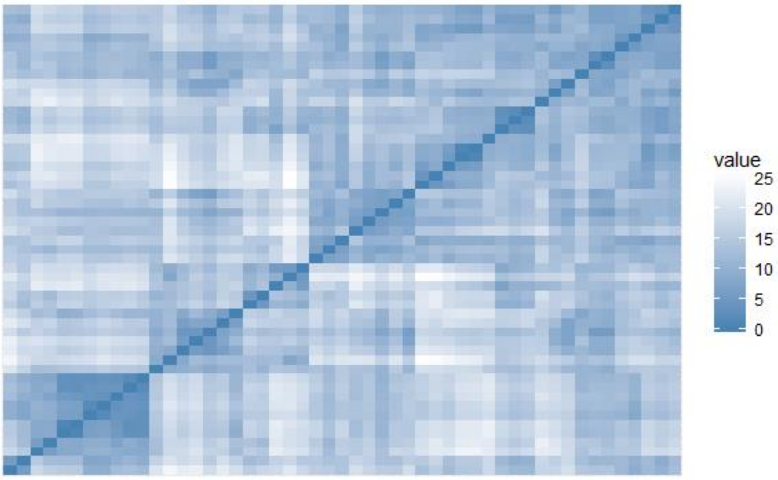

Figure 4. The quality of clustering based on VAT.

The cofenetic correlation can also be calculated between the original distance matrix and the cofenetic distance matrix, and then it can serve as a measure of the adequacy of the cluster solution to the initial data. Let us evaluate five hierarchical clusters by this indicator, which are compared among themselves in the previous section.

```
##             Complete Single Average Centroid Ward.D2
## W-Mantel    0.698   0.541    0.718    0.607    0.698
## P-value     0.001   0.001    0.001    0.001    0.001
```

Practical research on the effects of certain parameters of the object in his description of his state shown on Figure 5.

Thus, in order to find personalized solutions, it is suggested to take into account the average values of a particular cluster, which will determine the optimal state of the patient. Using the chart, it is possible to determine the impact of the main components for each patient.

Applying the naive Bayes method, we obtained a calculation of the value of a random variable, which achieves an a posteriori maximum only for some classes 1, 4, 5 (Table 1).

Peer-reviewed version available at *Mathematics* **2020**, *8*, 1211; doi:10.3390/math8081211

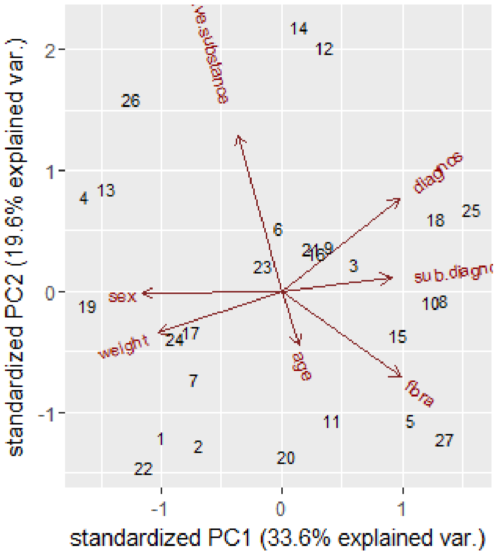

Figure 5. The impact of main components for each patient

Table 1. Results of the naive Bayes method

Class Statistics	1	2	3	4	5
Sensitivity	0.00000	0.8000	0.6000	0.00000	1.0000
Specificity	1.00000	0.7647	0.7647	1.00000	0.9167
Pos Pred Value	0.00000	0.6667	0.6000	0.00000	0.6000
Neg Pred Value	0.96296	0.8667	0.7647	0.92593	1.0000
Prevalence	0.03704	0.3704	0.3704	0.07407	0.1111
Detection Rate	0.00000	0.2963	0.2222	0.00000	0.1111
Detection Prevalence	0.00000	0.4444	0.3704	0.00000	0.1852
Balanced Accuracy	0.50000	0.7824	0.6824	0.50000	0.9583

The Kohonen map shows the proximity of the patient's weight and age attributes, as well as the diagnosis and concomitant diagnosis, and medication and hospital stay, which testifies to the affinity of the data and their impact on the object under study (Figure 6).

2.3. The proposed clustering method

To sum up, the quality metrics for different clustering algorithms are no too high for the analyzed dataset.

We propose the ensemble based on k-means clustering algorithm. The main purpose of our modification is to find the anomalous clusters. The following steps present the algorithm:

1. *Preprocessing stage*. The "reference" point $FS = (FS_1, \ldots, FS_n)$ is fixed (for example, at the center of gravity of the "cloud" of objects), which is then converted to 0 by subtracting it from all points representing the objects.
2. *Initialization stage*. The point farthest from 0 is taken as the initial centroid.
3. *Cluster update*. An object is assigned to the centroid c if and only if $d\ (y_i,\ c) < d\ (y_i,\ 0)$; cluster S is formed from all such objects.
4. *Update the centroid*. The center of gravity $c`$ for cluster S is calculated, which is then compared with the previous centroid c. If $c\ ' \neq c$, stage 3 is performed again after c is replaced by c'. Otherwise, a decision is issued - cluster S, its centroid c, as well as the values of the

Peer-reviewed version available at *Mathematics* **2020**, *8*, 1211; doi:10.3390/math8081211

contributions to the data spread - they are necessary for making a decision on the continuation of the cluster analysis process.

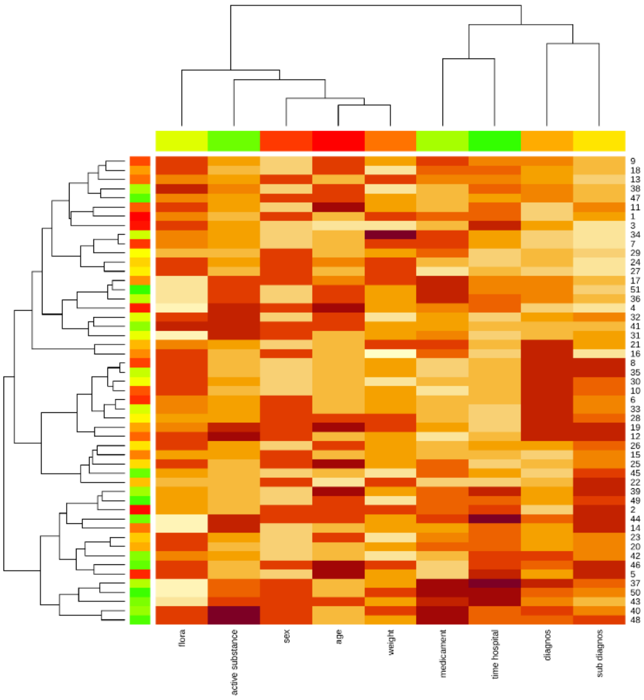

Figure 6. Patient placement in component space

The state space of the object in this study is given as a Euclidean space. It allows to present state space as a multidimensional system in the time domain. In order to describe the relationships between them, it is advisable to analyze the criterion of divergence between patients. To do this, we use the measure of distance L_i., calculated at stage number 3

Suppose, a group of instances belonging to class G_i contains M examples. The diagnostic measure of distance is selected and then the distance from the point a_{ij} to the point a_{im} included in the group of verified examples (quadratic measure) is defined as,

$$L_i = \sum_{i=1}^{n} {k_i}^2 \Delta {A_i}^2, \tag{7}$$

where

$$A = \{A_1, A_2, \dots, A_n\} \ ,$$
$$A_1 = \{a_{11}, a_{12}, \dots, a_{1m}\},$$
$$A_2 = \{a_{21}, a_{22}, \dots, a_{2m}\},$$
$$\dots \dots \dots \dots \dots \dots \dots \dots \dots \dots ..,$$
$$A_n = \{a_{n1}, a_{n2}, \dots, a_{nm}\}.$$

So

$$\Delta A_i = a_{ij} - a_{im},$$

where

$$\Delta A = \begin{bmatrix} a_{11}\, a_{12} \cdots a_{1n} \\ a_{21}\, a_{22} \cdots a_{2n} \\ \cdots \quad \cdots \; \cdots \; \cdots \\ a_{n1}\, a_{n2} \cdots a_{nn} \end{bmatrix} \tag{8}$$

ΔA_i– the difference between parameter values, n - is the number of parameters, m - is the number of patients that are being clustered, k_i – normalization factor.

Peer-reviewed version available at Mathematics 2020, 8, 1211; doi:10.3390/math8081211

The unequal attributes can have different ranges of represented entities in the selection. That is why the distance can be dependent on attributes with large ranges. Therefore, data must be normalized. The optimal distance to the optimal parameter value can be used for each defined cluster. It is possible to determine the average distance from point a_{ij} to points of a training sequence belonging to $G_i$class:

$$L_{ij} = \frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{n} {k_{ij}}^2 \Delta {A_{ij}}^2. \tag{9}$$

3. Results

We create three clusters based on proposed clustering method. The minimum difference between values of the membership function for one instance is 0.23. This value demonstrates the better result of clustering than for fuzzy c-mean.

In addition, the Hopkins statistics [25] was used for quality of clustering algorithm evaluation (Table 2):

$$H_{ind} = \frac{\sum_n w_i}{\sum_n q_i + \sum_n w_i}.$$

Table 2. Results of Hopkins statistics for clustering methods evaluation.

Method Numb. of clusters	k-means	Fuzzy c-means	Kohonen map	Proposed method
3	0.35	0.35	0.32	0.22
5	0.4	0.4	0.38	0.25
7	0.5	0.45	0.5	0.38

The Hopkins statistics value bigger than 0.5 for objects w_i and q_i is corresponded to the null hypothesis that w_i and q_i are similar, and the grouped objects are distributed randomly and uniformly. A value of H_{ind} <0.25 at a 90% confidence level indicates a tendency to data grouping.

So, the Hopkins statistics of the proposed methods is not dominated by the k-means, fuzzy c-means and Kohonen maps for different numbers of clusters.

The next stage is outliers finding. The outliers can affect the result of clustering by shifting the cluster centers. However, it is most important for us to find outliers for personalized treatment. The outliers can be evaluated based on the distance (9).

The algorithm for outliers mining and personalizing the treatment is following:

1) To create the clusters.
2) For each cluster to find the objects with distance L_{ij} higher than 75% of the average value.
3) For each object from step 2:
    - Find the relations between time-independent and time-dependent parameters using Apriori algorithm [28],
    - Create the list of most appropriative medicaments from medicaments with support level higher than min_support.

The results of algorithm for outliers mining and personalizing the treatment have been evaluated by the experts with more than 20 years' experience in purulent surgical infection of soft tissues, antibiotic therapy in the treatment of diseases caused by purulent surgical infection. The personalized treatment was proposed for seven form 36 patients (step 2 of the algorithm). The min_support was set 40%. Number of frequent rules (relations between time-independent and time-dependent parameters) was nine (step 3). The reduction of the medicaments rate was 52%. The experts confirmed the correctness of medicaments treatment for six of the seven patients.

So, the confirmed the correctness level of the proposed medical decisions is 86%.

Discussion

Peer-reviewed version available at *Mathematics* **2020**, *8*, 1211; doi:10.3390/math8081211

According to the results of research, it is possible to take into account such factors as heterogeneity of data, subjective assessments of psychological indicators, as well as temporal changes in the characteristics of the general state of the investigated object when working with medical data. All this requires the improvement of algorithms for solving analytical problems of estimation and forecasting. The processing of medical data also requires the development of appropriate solutions that allow for the collection, transmission, storage and protection of data. [19, 20, 26, 27]

To date, most existing algorithms that allow to you effectively solve current problems do not cover the entire list of problems associated with the personalization of decisions during the maintenance of the treatment process. Against this background, there is a need for new solutions that allow valid processing and storage of data without losing the quality of the results.

In this article, the methods of patient clustering were applied, which provided for determining the conditional measure of similarity of the studied objects, combining them into clusters so that similar objects were in the same cluster. The patient's state is assumed the average cluster score for the facility. The Euclidean distance is usually used as a measure of proximity. Using this method the number of clusters is unknown and selected by the researcher, and the quality of the clustering depends on the primary separating by elbow method [22, 23, 24].

So let's add the following disadvantages:

- the specificity of each user is not taken into account - all users are divided into classes (templates);
- if there is no characteristic value in the cluster, it will not be possible to obtain a recommendation.

The use of the DBSCAN method shows the absence of densely filled sections of the state space. We found that the correlation between the attributes is negligible.

One of the simplest classification algorithms is the naive Bayes method. However, quite often it works better than more the sophisticated algorithms. The application of this method is based on the assumption that all the variables under consideration are independent of each other [14]. In the example given in Table 2, the a posteriori maximum of only a few classes is achieved.

The advantages can be considered:

- classification, including many classes, is made quickly;
- when the assumption of independence is fulfilled, the classifier outperforms other algorithms, such as logistic regression, while requiring less training data;
- the algorithm works better with categorical features than continuous ones.

The disadvantages of this method are:

- if there is some value of a categorical type in the test dataset that is not found in the training dataset, the model assigns a zero probability to this value and cannot make a prediction;
- although the algorithm is a good classifier, the values of predicted probabilities are not always sufficiently accurate.

Conclusions

The research was conducted on the application of methods of intellectual analysis of medical data to solving of the problem of clustering the objects. This made it possible to highlight the advantages and disadvantages of existing Date Mining methods.

It was suggested to determine the average distance between instances of classes in order to find priority signs of influence, which helps to find the best instance state of this cluster.

The formalization of the medical data pre-processing stage to find personalized solutions in accordance with current standards and pharmaceutical protocols is proposed.

According to the research highlights the main steps in the process of finding personalized solutions that provide approaches to data - driving during the diagnosis, treatment, or rehabilitation of patients.

Therefore using the methods of computational intelligence it is proposed to classify persons by state, to determine the deviation of the parameters from the normal parameters of the group, as well

Peer-reviewed version available at Mathematics 2020, 8, 1211; doi:10.3390/math8081211

as the average parameters. This made it possible to create a personalized approach to analysis the condition and making recommendations for each patient based on long-term follow-up and monitoring under the guidance of a physician.

According to the results of the analysis, it becomes possible to predict the optimal general state for a particular person, which will help to improve life satisfaction and ensure its continuation.

Author Contributions: Conceptualization, N.M. and N.S.; methodology, N.M.; software, N.S.; validation, N.M. and V.M.; formal analysis, N.S.; investigation, N.M.; resources, V.M.; data curation, V.M.; writing—original draft preparation, M.Z. and O.V; writing—review and editing, M.G. and O.V.; visualization, N.M. and M.Z.; supervision, N.S.; project administration, M.G; funding acquisition, M.G.

Funding: This research was partially supported by the Ukrainian Ministry of Science and Innovation, through the Information technology of psychophysical portrait formation in stressful situations, project № 0120U102194, and Information technology for processing personalized medical information, project № 0119U002257

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
IoT	Internet of Things
ANN	Artificial Neural Networks
BIRCH	Balanced Iterative Reducing and Clustering using Hierarchies
RAM	Random Access Memory
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
VAT	Visual Assessment of cluster Tendency metric

References

1. Waring J., et al. Automated Machine Learning: Review of the State-of-the-Art and Opportunities for Healthcare. *Artificial Intelligence in Medicine*, 104, Apr. **2020**, 101822. DOI.org (Crossref), doi:10.1016/j.artmed.2020.101822.
2. Kharkovyna O. Artificial Intelligence & Deep Learning for Medical Diagnosis, Nov 13, **2019**, https://towardsdatascience.com/artificial-intelligence-deep-learning-for-medical-diagnosis-9561f7a4e5f/
3. Ng, Man-Fai, et al. Predicting the State of Charge and Health of Batteries Using Data-Driven Machine Learning. *Nature Machine Intelligence*, 2(3), **2020**, 161–70. DOI.org (Crossref), doi:10.1038/s42256-020-0156-7.
4. Rytsar B. A New Method for Symmetry Recognition in Boolean Functions Based on the Set-Theoretical Logic Differentiation. I. *Control Systems and Computers*, 4 (282), Oct. **2019**, 3–13. DOI.org (Crossref), doi:10.15407/csc.2019.04.003.
5. Artificial Intelligence in Medicine | Machine Learning. https://www.ibm.com/watson-health/learn/artificial-intelligence-medicine. Accessed 20 Apr. 2020.
6. Democratizing Data-Driven Medicine / https://www.sophiage-netics.com/home.html Accessed 20 Apr. 2020.
7. Calvaneseand D., Calvanese, G,. Giacomo D., Lembo M., Lenzerini A., Poggi, R. Rosati Ontology-based database access. *In Proc. of SEBD–2007*. **2007**. 324–331
8. Tkachenko, R., Izonin, I.: Model and Principles for the Implementation of Neural-Like Structures based on Geometric Data Transformations. In: Hu, Z.B., Petoukhov, S., (eds) *Advances in Computer Science for Engineering and Education*. ICCSEEA2018. Advances in Intelligent Systems and Computing. Springer, Cham, 754, **2019**, 578-587,. https://doi.org/10.1007/978-3-319-91008-6_58
9. Tkachenko R., Izonin I., Kryvinska K., Chopyak V., Lotoshynska N., Danylyuk D. Piecewise-linear Approach for Medical Insurance Costs Prediction using SGTM Neural-Like Structure. In: Shakhovska N., Montenegro S., Estève Ya., Subbotin S., Kryvinska N., Izonin I.:. (Eds.): Informatics & Data-Driven

Peer-reviewed version available at *Mathematics* **2020**, *8*, 1211; doi:10.3390/math8081211

Medicine (IDDM 2018). Proceedings of the 1st International Workshop IDDM 2018. Lviv, Ukraine, November 28-30, **2018,** .170-179, , CEUR-WS.org

10. Tkachenko, I. Izonin, P. Vitynskyi, N. Lotoshynska, and O. Pavlyuk Development of the Non-Iterative Supervised Learning Predictor Based on the Ito Decomposition and SGTM Neural-Like Structure for Managing Medical Insurance Costs," Data, 3 (4), **2018,** 1-14
11. Telenyk, S., Czajkowski, K., Bidiuk, P., & Zharikov, E. Method of Assessing the State of Monuments based on Fuzzy Logic. In *2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, 1, **2019**, 500-506. IEEE
12. Dangare, C. S., & Apte, S. S. Improved study of heart disease prediction system using data mining classification techniques. *International Journal of Computer Applications*, *47*(10), **2012**, 44-48.
13. Vijiyarani, S., & Sudha, S. Disease prediction in data mining technique–a survey. *International Journal of Computer Applications & Information Technology*, *2*(1), **2013**, 17-21.
14. Tang, Y., Wang, Y., Cooper, K. M., & Li, L. Towards big data Bayesian network learning-an ensemble learning based approach. In *2014 IEEE International Congress on Big Data*, **2014** ,. 355-357. IEEE.
15. Mulesa P., Perova I. Fuzzy Spacial Extrapolation Method Using Manhattan Metrics for Tasks of Medical Data Mining. Proc. of 12th International Conference on *Computer Science and Information Technologies* CSIT'2015. Lviv, Ukraine., **2015**, 104-106.
16. Bodyanskiy Ye., Perova I., Vynokurova O., Izonin I. Adaptive Wavelet Diagnostic Neuro-Fuzzy System for Biomedical Tasks. Proc. of 14th International Conference on *Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering* (TCSET), Lviv-Slavske, Ukraine, **2018**, 20 – 24,
17. Perova, O. Litovchenko, Ye. Bodyanskiy, Ye. Brazhnykova, I. Zavgorodnii, P. Mulesa. Medical Data-Stream Mining in the Area of Electromagnetic Radiation and Low Temperature Influence on Biological Objects. Proc. 2018 IEEE Second International Conference on *Data Stream Mining & Processing* (DSMP), August 21-25, **2018**, Lviv, Ukraine, 3-6
18. Perova I., Bodyanskiy Ye., Brazhnykova Ye., Mulesa P. Neural Network for Online Principal Component Analysis in Medical Data Mining Tasks IEEE First International Conference on System Analysis & Intelligent Computing (SAIC) 8-12 October **2018**, Kyiv, Ukraine,.150-154
19. Izonin, I. The combined use of the wiener polynomial and SVM for material classification task in medical implants production. *International Journal of Intelligent Systems and Applications*. 9,**2018**, 40-47
20. Melnykova, N., Shakhovska, N., & Sviridova, T. The personalized approach in a medical decentralized diagnostic and treatment. In *2017 14th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM)*, **2017**, 295-297. IEEE..
21. Bai, B. M., Nalini, B. M., & Majumdar, J. Analysis and detection of diabetes using data mining techniques—a big data application in health care. In *Emerging Research in Computing, Information, Communication and Applications*, **2019**. 443-455. Springer, Singapore.
22. Dudik, J. M., Kurosu, A., Coyle, J. L., & Sejdić, E. A comparative analysis of DBSCAN, K-means, and quadratic variation algorithms for automatic identification of swallows from swallowing accelerometry signals. *Computers in biology and medicine*, *59*, **2015**, 10-18.
23. Melnykova, N. Semantic search personalized data as special method of processing medical information. In: *Advances in Intelligent Systems and Computing*. Springer, Cham, **2017**. 315-325.
24. Havens, T. C., & Bezdek, J. C. An efficient formulation of the improved visual assessment of cluster tendency (iVAT) algorithm. *IEEE Transactions on Knowledge and Data Engineering*, *24*(5), **2011**, 813-822.
25. Hopkins, W., Marshall, S., Batterham, A., & Hanin, J. Progressive statistics for studies in sports medicine and exercise science. *Medicine+ Science in Sports+ Exercise*, *41*(1), **2009**, 3.
26. Arzubov, M., Shakhovska, N., & Lipinski, P. Analyzing ways of building user profile based on web surf history. In *2017 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT)* Vol. 1, **2017**, 377-380. IEEE.
27. Vidal, M. E., Endris, K. M., Jazashoori, S., Sakor, A., & Rivas, A. Transforming Heterogeneous Data into Knowledge for Personalized Treatments—A Use Case. *Datenbank-Spektrum*, *19*(2), **2019**, 95-106.
28. Cui, X., Yang, S., & Wang, D. An algorithm of apriori based on medical big data and cloud computing. In *2016 4th International Conference on Cloud Computing and Intelligence Systems (CCIS)*. **2016**, 361-365. IEEE