Journal of Theoretical and Applied Information Technology
15th July 2022. Vol.100. No 13

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

CLUSTERING AND DATA MINING ON THE EXAMPLE OF HIV-INFECTED PEOPLE DATA

AIGUL DAULETOVNA KUBEGENOVA[1], AIZAT GARIFULLAYEVNA ZHAKHIENA[2], SAYA KUBAIDULIEVNA BAIGUBENOVA[3], GULNARA SHAVKATOVNA UTYASHEVA[4], AKYLBEK NURLYBEKOVICH OMAROV[5]

[1,2,3]Higher School "Information Technologies", Kazakhstan Agrarian and Technical University named after Zhangir Khan, 51 Zhangir Khan str., Uralsk, 090009, Republic of Kazakhstan

[4,5]Department of "Engineering and Technology" of the Faculty of Engineering and Humanities, West Kazakhstan Innovation and Technology University, 44/1 Ihsanov str., Uralsk, 090009, Republic of Kazakhstan

E-mail: aigul.d.kubegenova@gmail.com

ABSTRACT

The paper discusses aspects of data research, in-depth data analysis, knowledge acquisition, methods of data processing in the knowledge base, methods of intellectual analysis, and application of data mining in the field of medicine. A group of HIV-infected patients was identified, an analysis with a medical history was carried out, models and an algorithm of actions (input data) were developed, and analysis and experiments with data search methods were carried out. All diseases were presented as a set of numerical vectors and were grouped into clusters, according to the described methods, and with the help of this distribution, the Hopkins statistics value was calculated. Clustering itself was carried out using the usual tools of the sklearn library. Various methods of representation of multidimensional data in a two-dimensional plane are proposed, such as the method of basic components, the Kohonen line, etc. Two different clustering methods were considered, namely, the k-Medium method (using the Kmeans function from the Python sklearn library) and density-based clustering methods with autoconfiguration (from the HDBSCAN function from the Python Hdbscan library). In the case of comparison, the cluster structure is evaluated by changing various parameters of one algorithm (for example, the number of k groups); a model (or several) is built on the received and prepared objects, and its parameters are adjusted. After that testing and analysis of the results were carried out.
Keywords: *Clustering, Vectorization, Correlation, Sklearn, Manipulation*.

1. INTRODUCTION

The development of modern methods of data storage and processing leads to a rapid growth of accumulated information requiring analysis. Such a large amount of accumulated data does not allow them to be processed by human forces, and it is obvious that among these raw data there is information necessary for making important decisions.

Therefore, it will be necessary to use data mining for automatic data analysis.

The intellectual analysis method called data mining means the detection of implicit patterns in data sets, and is usually translated into Russian as "intellectual data analysis".

As a scientific field, it began to develop actively in the 1990s, which was caused by the widespread use of automated information processing technologies and the accumulation of large amounts of data in computer systems [1,2].

We know that data mining is a multidisciplinary field that has emerged and is developing based on such sciences as applied statistics, knowledge recognition, artificial intelligence, database theory, etc.

Data mining is a decision support process based on finding hidden patterns (patterns of information) from data.

Data mining can be described as a technology designed to search for a large volume of fuzzy, objective, and useful patterns in practice:

- fuzzy because the detected patterns cannot be determined by standard methods of information processing or by expert means;
- objective, since the revealed patterns fully correspond to reality, unlike expert knowledge, which is always subjective;

- it is useful because the conclusions have a real meaning that can be applied in practice.

Data analysis is also used very successfully in medicine. Examples of this can be found in the analysis of examination results, diagnosis, comparison of the effectiveness of treatment methods and medications, analysis of diseases and their prevalence, identification of side effects. Such data mining technologies as associative rules and chain patterns are successfully used to determine the relationship between medications intake and their side effects.

When visualizing and identifying hidden complex relationships between diagnostic features of different groups of patients, various types of algorithms related to intelligent data analysis are used [3].

Data mining allows detecting patterns in medical data that form the basis of these rules. Based on this, one can find out the diagnosis of the patient and the method of their treatment from the description and combination of various disease symptoms.

Data mining is based on classification and clustering methods, modeling and forecasting, genetic and evolutionary algorithms, and soft computing methods.

It is also possible to note the beginning of the development of applied statistics methods, regression and correlation analyses, discriminant and factor analysis in data mining [4]. For processing medical information the following algorithms are used: the C4.5 algorithm; the k-means method; the Apriori algorithm; the PageRank algorithm; the Ada Boost algorithm; the k-nearest neighbors algorithm (kNN); the naive Bayesian classifier, etc. These algorithms are included in data mining tools and are used for processing big data.

2. LITERATURE REVIEW

In the works of S. Kabanikhin and O.Krivorotko, monitoring and forecasting of the spread of the epidemic in the region were carried out and a mathematical model was built, taking into account the population and visualization and the processing of big data. They analyzed a numerical method for solving the inverse epidemiology problem based on a genetic algorithm and traditional optimization ideas.

After all these factors had been taken into account, a model and a forecast were made regarding the number of people who were expected to be infected during the epidemic, the duration of the epidemic, and the maximum incidence rate [5].

In the study of A. Bushnits and G. Bocharov, the authors formulated a multiscale model of acute HIV infection, which combines the processes of infection spread and immune responses in the lymph nodes (LN) and links with the dynamics of HIV observed in the blood.

A multiscale model of HIV infection formulated in the study was based on several simplifying assumptions, among which the following can be distinguished:

1. the spatial dynamics of cells and cytokines in LNs is considered in a two-dimensional regular domain;
2. the model is limited to primary acute HIV infection and concomitant reactions of cytotoxic T cells;
3. intracellular regulation of cell fate using multiple cytokine signaling is described through a hierarchy of activation thresholds:

Elementary high-resolution mechanistic modules and their integration into the developed multiscale model framework allowed us to study the effectiveness of multimodal approaches to the treatment of HIV infection combining antiretroviral therapy (ART), antifibrosis, and immunological treatment methods, modulating medications using sensitivity analysis. This study has helped clinicians move towards the ambitious goal of ideal long-term control for the treatment of this infection with minimal side effects [6].

The paper presents a numerical algorithm for constructing an individual mathematical model of HIV dynamics at the cellular level, examines the problem of determining the parameters of HIV infection and immune response using additional measurements of concentrations of T-lymphocytes, free virus, and immune effectors at fixed points in time for a mathematical model of HIV dynamics.

The goal of setting the parameters of a mathematical model (the inverse problem) is thus reduced to the goal of minimizing the objective function describing the deviation of the simulation results from experimental data. A genetic algorithm for solving the problem of minimizing the least-squares function has been implemented and investigated. The results of the numerical solution of the inverse problem have been analyzed [7].

E.O.Omondi in his paper analyzed and formulated mathematical compartment models of HIV transmission within and between two age groups in Kenya. Using the Monte Carlo method, he adjusted the model to the data and derived the parameters by estimating the reproduction numbers in the transmission aisles in age groups and between age groups of transmission of the main reproduction

numbers. The analysis of the data showed that there was a significant difference in the average number of new cases of HIV infection between men and women in two age groups.

To a greater extent, the result showed that in the majority of cases women were infected with HIV, and the rate of HIV transmission per capita was the highest due to the interaction between young people and adults. Sensitivity analysis showed that reproduction rates depended mainly on the probability of infection [8].

The significance of my study is to identify groups of HIV-infected patients using specialized data analysis software. The importance of investigating this problem lies in the timely establishment of the correct diagnosis and the necessary treatment corresponding to one of their clinical forms. The lack of treatment leads to deterioration of health, and complications and the further development of the disease may occur. Thus, I needed to study the methods of processing medical data, create a relationship and a pattern between symptoms, and evaluate and build various prognostic models. The results obtained can be used by specialists to make decisions when making a diagnosis.

3. MATERIALS AND METHODS

The archive of medical data contains a lot of information about various cases of specific diseases, methods of their diagnosis. The search for samples is one of the tasks of many medical studies. Methods of automatic data analysis are often used to solve such problems.

The concept of data mining is used to denote a set of methods for determining practically useful knowledge in data necessary for decision-making in various professional fields, including medicine. Data search is a vast field that has emerged and is developing in the generation of sciences such as statistics, machine learning, artificial intelligence.

Let us look at a brief description. Statistics is a science that collects information and carefully studies objects for their further analysis and processing. Machine learning can be described as a process that studies methods for constructing algorithms capable of learning. Artificial intelligence is a scientific field that develops methods that allow solving intellectual problems on an electronic computer if these problems are solved by people.

Data mining combines such methods and algorithms as artificial neural networks, decision trees, correlation, cluster analysis, linear regression, Bayesian networks, and much more. It can be used for such tasks as classification, clustering, and forecasting [9].

Statistical methods are often reduced to solving linear regression equations. However, with this approach, it is not always possible to find a contact. In such cases, machine learning methods are used. In medicine, there are many expert systems for diagnosis based on patterns and rules describing the combination of symptoms of various diseases.

These rules will help determine the course of the patient's illness, what treatment to prescribe, predict the result of the prescribed treatment, and study the causes of various pathologies. Data retrieval technologies allow finding medical data, such as rules and schemas. The development of diagnostic methods is a relevant task of medicine, which, in turn, refers to the tasks of classification.

A case study using survey data demonstrated approaches to identifying and implementing indirect interventions and analytical approaches to researching how gender norms can affect health. Research showed the importance of collecting statistics and the need to collect data regularly to identify trends in gender norms and gain insight into their impact on people's behavior and health outcomes. The main findings, the strengths and weaknesses of norms proxies, and the choice of methodology were summarized.

When clustering individual data for use as a norm proxy, reported relationships or derived behaviors can be considered. The most common solution for building norm proxies is data clustering. The decision to cluster indicators of attitudes or behaviors as proxies for gender norms should be made considering the contextual characteristics of the measure of interest.

Attitudinal clustering may be more appropriate when the behavior is difficult to detect and researchers can guess with confidence.

Methods for obtaining and using proxy indicators can help develop effective analysis strategies [10].

With the rapid development of computing power and machine learning algorithms, clustering methods have become a powerful tool for gaining information and discovering patterns in datasets. Clustering techniques are especially important for big data applications where the dimensionality is high and the amount of data is too large for human analysis. A comprehensive review of modern clustering methods and their latest advances is conducted, and their performance, when applied to a high-measure expression-level data set in multiple environments, is analyzed. PCA and MDS are often

used to cluster environments based on expression level, and traditional UMAP and t-SNE methods are used as well. By combining the ideas and advantages of several clustering methods, it is possible to develop new clustering methods that are more powerful and versatile [11].

Spatial scanning methods are extremely popular for identifying clusters of diseases using data on the number of diseases. The original circular scanning method proposed by Kulldorff is easy to implement and computationally inexpensive and has high power for detecting ring clusters. However, it may be difficult to identify non-circular clusters using it. Many extensions of the original method can be proposed for better detection of irregularly shaped clusters. Popular extensions of the spatial scanning method include Top Level Set, Flexible Form, Dynamic Minimum Spanning Tree, Fast Subset, etc. We compared the performance of different methods using power, sensitivity, positive predictive value, and overall accuracy by applying the methods to 126 publicly available reference datasets based on 46 different cluster shapes. Comparisons were considered more in-depth and included more methods. The comprehensiveness of the study allowed us to draw reliable conclusions and give specific recommendations regarding the most effective methods. R packages and scripts were provided to ensure reproducible results [12].

The importance of this analysis lies in the timely establishment of the correct diagnosis and the necessary treatment corresponding to one of their clinical forms. Untimely treatment leads to deterioration of health and can result in a complication of the disease.

4. RESULTS

The object of the analysis is the data collected from patients with various clinical forms of HIV infection, which is used to build a predictive model and conduct experiments using data retrieval methods.

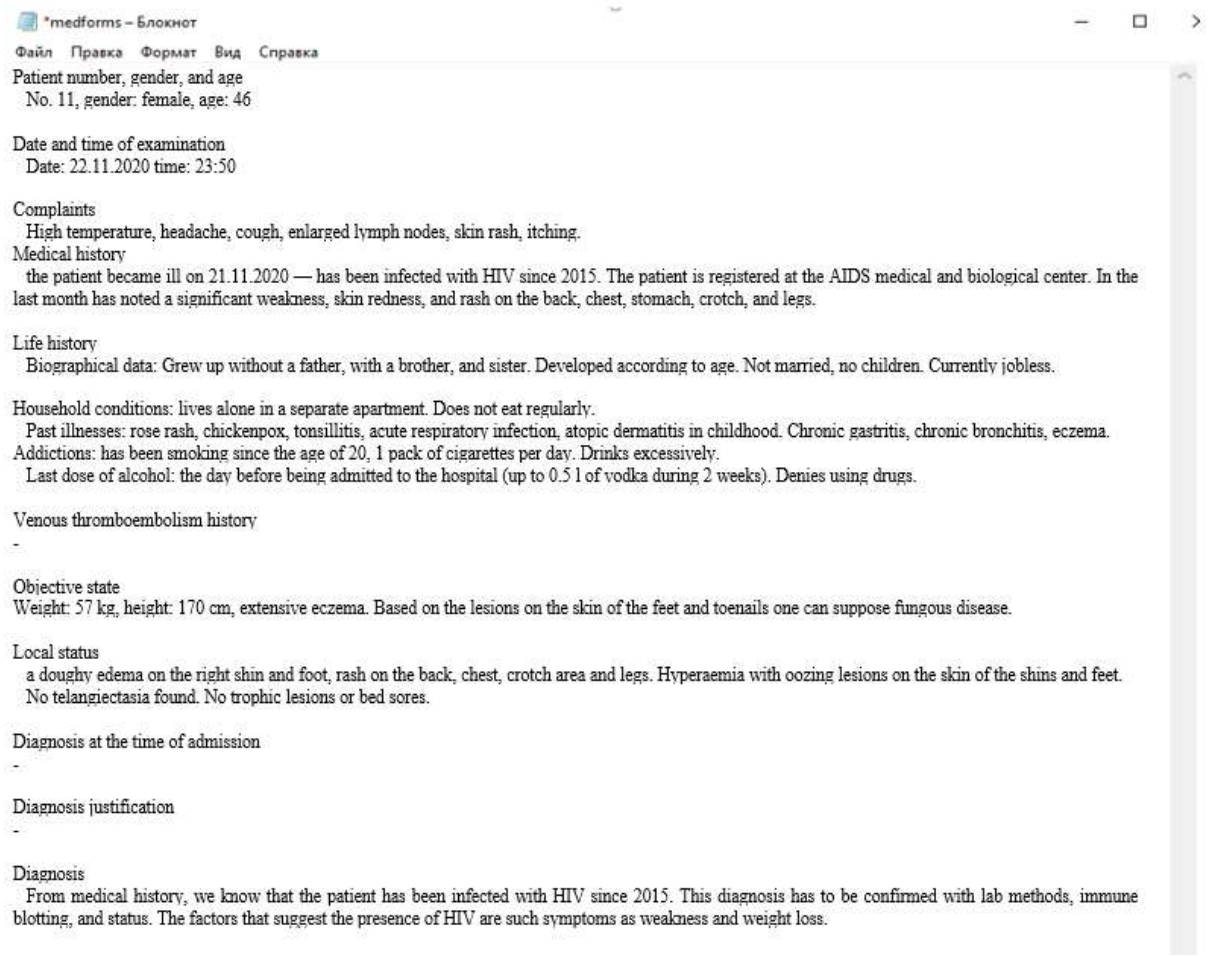
The purpose of the analysis is to identify a group of patients with HIV infection using specialized data analysis software.

The results obtained can be used by specialists to make a decision when making a diagnosis. In the process of developing models, an algorithm of actions is built and input data is entered.

First of all, the medical history of HIV-infected patients receiving treatment was analyzed as initial data.

Mandatory analysis was performed for each patient. Figure 1 shows an extract from one of the documents.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195


*medforms – Блокнот

Файл Правка Формат Вид Справка

Patient number, gender, and age
No. 11, gender: female, age: 46

Date and time of examination
Date: 22.11.2020 time: 23:50

Complaints
High temperature, headache, cough, enlarged lymph nodes, skin rash, itching.
Medical history
the patient became ill on 21.11.2020 — has been infected with HIV since 2015. The patient is registered at the AIDS medical and biological center. In the last month has noted a significant weakness, skin redness, and rash on the back, chest, stomach, crotch, and legs.

Life history
Biographical data: Grew up without a father, with a brother, and sister. Developed according to age. Not married, no children. Currently jobless.

Household conditions: lives alone in a separate apartment. Does not eat regularly.
Past illnesses: rose rash, chickenpox, tonsillitis, acute respiratory infection, atopic dermatitis in childhood. Chronic gastritis, chronic bronchitis, eczema.
Addictions: has been smoking since the age of 20, 1 pack of cigarettes per day. Drinks excessively.
Last dose of alcohol: the day before being admitted to the hospital (up to 0.5 l of vodka during 2 weeks). Denies using drugs.

Venous thromboembolism history
-

Objective state
Weight: 57 kg, height: 170 cm, extensive eczema. Based on the lesions on the skin of the feet and toenails one can suppose fungous disease.

Local status
a doughy edema on the right shin and foot, rash on the back, chest, crotch area and legs. Hyperaemia with oozing lesions on the skin of the shins and feet.
No telangiectasia found. No trophic lesions or bed sores.

Diagnosis at the time of admission
-

Diagnosis justification
-

Diagnosis
From medical history, we know that the patient has been infected with HIV since 2015. This diagnosis has to be confirmed with lab methods, immune blotting, and status. The factors that suggest the presence of HIV are such symptoms as weakness and weight loss.

Figure 1: Input Data

As we can see, there are unstructured records in the document (except for obvious signs, they can simply be broken into templates and analyzed). The data voluntarily submitted by the patient is entered into the database. In addition, some sections may be missing from the patient database. For example, not all documents contain detailed information about the diagnosis and complaints of the patient. The blocks of test results are presented in the form of several different templates from which parameters can be obtained, and such sections as complaints and medical history are filled in in a completely arbitrary form. In addition, before starting to build a cluster model, it is necessary to study the probabilistic solution of this problem.

Vectorization

The presentation of texts in a format that could be performed using such manipulations was carried out using the TfidfVectorizer function from the Python sklearn library. The term frequency-inverse document frequency (TF-IDF) statistical criterion underlying this method is used to assess the significance of a word. All diseases are presented as a set of numerical vectors with which further manipulations can be carried out [13].

Before starting clustering, it is necessary to determine the predisposition of the data grouped into the first clusters. To do this, Hopkins statistics have been selected. They are based on the null hypothesis that the data is not prone to grouping. To calculate the value based on a distribution with the same standard deviation as the original dataset, several randomly generated false datasets are created.

For each i observation, the average distance from n to k is calculated: ω_i between specific objects, q_i between artificial objects and their nearest actual neighbors (1).

$$H_{ind} = \frac{\Sigma_n \omega_i}{\Sigma_n q_i + \Sigma_n \omega_i} \quad (1)$$

Then the Hopkins statistics show that q is greater than 0.5 similarly, and the grouped objects

are divided into random and homogeneous ones and correspond to the null hypothesis.

The value H .25 ind < 0 reflects the trend towards data clustering at the 90% reliability level. If these statistics show that the null hypothesis is incorrect and our incomes tend to cluster, then we move on to clustering [14].

Clustering algorithms

Two different clustering methods were considered: the k-Medium method (using the Kmeans function from the Python sklearn library), and density-based clustering methods with Auto-configuration (using the HDBSCAN function from the Python Hdbscan library).

On the received and prepared text data objects prepared for vectorization, it is necessary to build a model (or several models) and adjust its parameters. Then testing and analysis of the results are carried out. Figure 2 shows the sequence of analysis work.

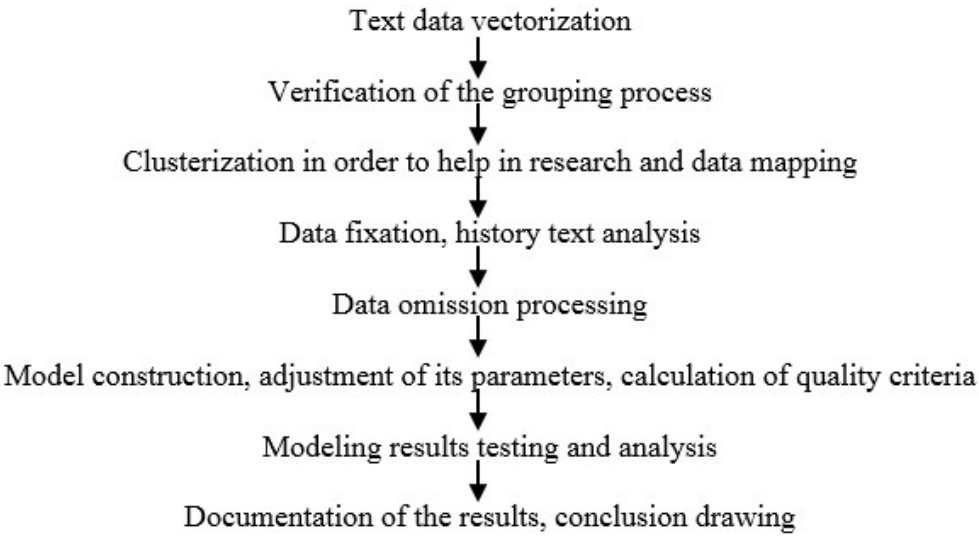


Figure 2: The Procedure for Conducting the Analysis

Visualization was required at different stages of work on the task, various functions based on the pyplot module in the Python matplotlib library were used [15].

In particular, the following:

1. NumPy, a fundamental library necessary for scientific computing in Python.
2. Matplotlib, a library for working with two-dimensional graphs.
3. Pandas, a tool for analyzing structural data and time series.
4. Scikit-learn, an integrator of classical machine learning algorithms.
5. SciPy, a library used in mathematics, science, and engineering.
6. Jupyter, an interactive computing environment.

First of all, we need to cluster the data for their detailed breakdown. To keep the data simple, all unnecessary punctuation marks have been removed, and the Python regular expression library is used to accomplish this task. The cleaned document is divided into empty symbols and separate words are used as a method of dividing documents into tokens.

For example:

```
def correct_known_words(story):
dict_ = {'conc. diagnosis': 'concomitant diagnosis',
'pls.':'pills',
'stages of sec.': 'stages of secondary',
'h/ache': 'headache'
```

After the manipulations performed in the previous step, it is enough to load a full-text modified and cleaned data frame and process all events, as well as the cell values of the data frame. The indexes are the names of the history files. The NearestNeighbors function from the Sklearn library is usually used to create uncontrolled and controlled models, which allows us to create a pseudo-object similar to our vectorized text [16].

5. DISCUSSION

The novelty of the study lies in the methods of data mining used as a tool for building a forecast

model and analyzing big data and digital technologies in the medical industry since in some areas they are difficult to implement to the right extent. Data mining technology allows one to find medical data, such as rules and charts. The development of such diagnostic methods as task classification is an urgent task in medicine.

Using this distribution, according to the methodology described above, we can calculate the value of Hopkins statistics. Clustering itself was carried out using the usual tools of the sklearn library. Various methods of representation of multidimensional data in a two-dimensional plane are proposed, such as the method of basic components, the Kohonen line, etc.

Principal component analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (tSNE) were used since these methods are based on different principles. Thus, if one does not show visual differences between clusters, the other may work better. Both methods are used in the sklearn library. In the Sklearn implementation, PCA and tSNE return a set of vectors corresponding to the number of axes that we specified when working.

In our case, working with a plane, it will be two vectors, each of which has a length equal to the number of rows within the data containing the event. These vectors should be passed to the visualization function, defining specific features of the class and cluster numbers.

As a result, a graph was created in which points in different clusters had different colors, and if we set them the necessary parameters, crosses of different colors would be above them.

After creating a cluster solution, the question arises about how stable and statistically important it is. Stable grouping should be preserved when clustering methods change: for example, if the percentage of matches in the results of hierarchical cluster analysis when grouping using the k-mean method is more than 70%, then tolerance to stability is accepted [17].

The comparative check evaluates the cluster structure by changing various parameters of one algorithm (for example, the number of k groups); this is done by the usual means of the sklearn library.

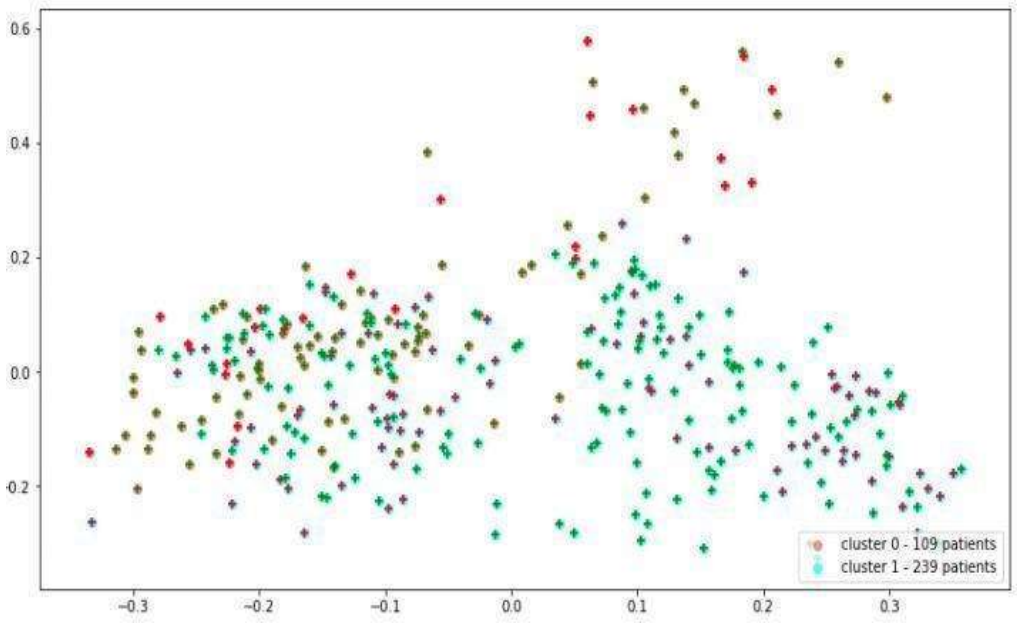


Figure 3: Example of Clustering

Calculations of Hopkins statistics on the original set of unstructured text data showed that in any experiment the value of H did not exceed 0.150, which indicates the correctness of our assumptions about the presence of a cluster structure in the data.

Although visualization by the main components method cannot visually construct a distribution line between clusters, we see that there is a certain tendency of visual separation [18,19].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

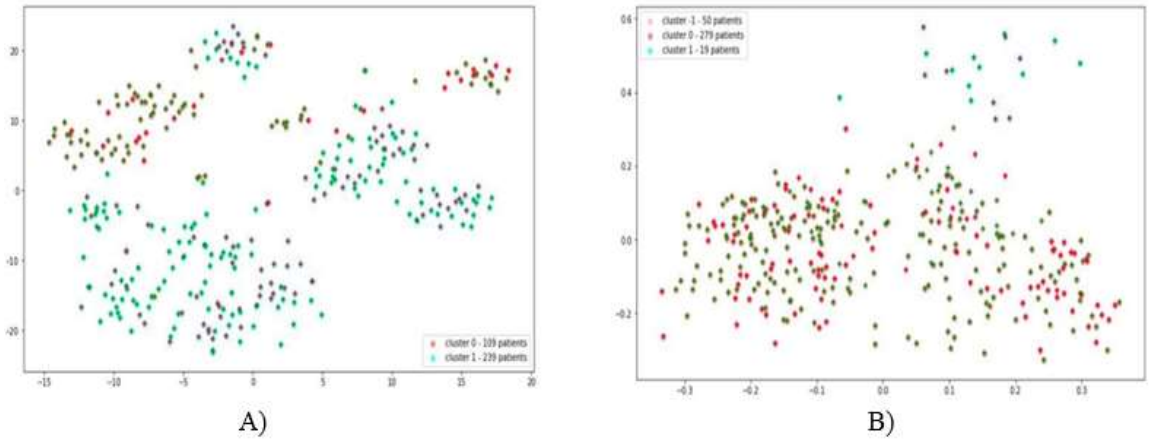


Figure 4: Cluster Experiment

No specific patterns have been revealed: A) and B) clusters at many points are similar and completely different, but they are located in the same and different clusters. In Figure 4. An attempt to perform a B) cluster experiment on visualization and distribution search for extracting information from data was carried out using the tSNE method.

This visualization method shows the pronounced severity of the cluster structure, but it is not possible to determine it by the method of neighbors. In this experiment, it is possible to determine the number of clusters equal to the number of areas highlighted for this visualization.

However, it is impossible to achieve compliance with visual clusters. At that time, a fundamentally different approach to clustering was used based on density, which automatically adjusts the distance (in particular, the hdbscan implementation) [20].

6. CONCLUSION

The use of data mining technology in the field of medicine as aspects of the application and methods of intellectual analysis is becoming more and more effective. The task of determining groups of HIV-infected patients, all the main points of implementation were recorded, and the Tfidf Vectorizer function from the Python sklearn library was used to evaluate the TF-IDF statistical criterion. Hopkins statistics were selected, the propensities of the grouped data were determined, and clustering of the patient according to the medical history was carried out.

Two types of clustering were considered. Models have been constructed from the prepared objects and a graph has been drawn up following the methods.

As a result, a certain tendency to visual division by the method of the main components has been revealed. All conclusions and criteria obtained have been recorded and analyzed.

All the solutions used and described in the work are used to find and implement groups of patients with other diseases that can be large-scale and also can be used in experiments.

REFRENCES:

[1] J. MacLennan, Zh. Tang, and B. Crivat, *Microsoft SQL Server 2008 Data mining. Intellektualnyi analiz dannykh* [Data Mining with Microsoft SQL Server 2008]. St. Petersburg, Russia: BKhV-Peterburg, 2009, 720 p.

[2] A.A. Barsegyan, M.S. Kupriyanov, I.I. Kholod, M.D. Tess, and S.I. Elizarov, *Analiz dannykh i protsessov: ucheb. posobie* [Data and process analysis: a manual], 3rd edition, revised and enlarged. St. Petersburg, Russia: BKhV-Peterburg, 2009, 512 p.

[3] U.R. Acharya, and W. Yu, "Data mining techniques in medical informatics", *The Open Medical Informatics Journal*, No. 4, 2010, pp. 21–22.

[4] A.A. Shumeiko, S.L. Sotnik, and E.A. Belaya, *Intellektualnyi analiz dannykh (Vvedenie v Data Mining): ucheb. posob*. [Data Mining (Introduction to Data Mining): a manual]. Dnepropetrovsk, Ukraine: Belaya Ye.A, 2012, 212 p.

[5] S. Kabanikhin, O. Krivorotko, A. Takuadina, D. Andornaya, and Sh. Zhang, "Geo-information system of spread of tuberculosis based on inversion and prediction", *Journal of Inverse and III –*

posed Problems, Vol. 29, No. 1, 2021, pp. 65-79.

[6] A. Bouchnita, G. Bocharov, A. Meyerhans, and V. Volpert, Towards a multiscale model of acute HIV infection, *Computation*, Vol. 5, No. 1, 2017, Art. No. 6.

[7] H.Th. Banks, S.I. Kabanikhin, O.I. Krivorotko, and D.V. Yermolenko, "A numerical algorithm for constructing an individual mathematical model of HIV dynamics at the cellular level", *Journal of Inverse and ILL-Posed Problems*, Vol. 26, No. 6, 2018, pp. 859-873.

[8] E.O. Omondi, "A mathematical modeling study of HIV infection in two heterosexual age groups in Kenya", *Infectious Disease Modelling*, Vol. 4, 2019, pp. 83-98.

[9] A.V. Lisinin, and R.T. Faizulin, "Primenenie metaevristicheskikh algoritmov k resheniyu zadach klasterizatsii metodom k-srednikh" [Application of metaheuristic algorithms to solving clustering problems using the k-means method], *Kompyuternaya optika*, Vol. 39, No. 3, 2015, pp. 406–412.

[10] B. Cislaghi, A.M. Weber, H.B. Shakya, and S. Abdalla, "Innovative methods to analyse the impact of gender norms on adolescent health using global health survey data", *Social Science & Medicine*, Vol. 293, 2022, Art. No. 114652.

[11] K. Zhang, Z. Wang, and L. Liu, "Finding Clusters and Patterns in Big Data Applications: State-of-the-Art Methods in Clustering Environments", *Conference on Compute and Data Analysis*, 2021, pp. 8–12.

[12] J.P. French, M. Meysami, L.M. Hall, N.E. Weaver, M.C. Nguyen, and L. Panter, "A comparison of spatial scan methods for cluster detection", *Journal of Statistical Computation and Simulation*, 2022, Art. No. 2065676

[13] A. Myuller, Vvedenie v mashinnoe obuchenie s pomoshchyu Python. Rukovodstvo dlya spetsialistov po rabote s dannymi [An introduction to machine learning with Python. A guide for data scientists]. Moscow, Russia: Alfa-kniga, 2017, pp. 153-170.

[14] I.M. Neiskii, *Klassifikatsiya i sravnenie metodov klasterizatsii. Intellektualnye tekhnologii i sistemy. Sbornik uchebno-metodicheskikh rabot i statei aspirantov i studentov* [Classification and comparison of clustering methods. Intelligent technologies and systems. Collection of educational and methodical works and articles of graduate and undergraduate students]. Moscow, Russia: NOK CLAIM, 2006, Issue 8, pp. 130–142.

[15] A.A. Barsegyan, M.S. Kupriyanov, V.V. Stepanenko, and I.I. Kholod, *Metody i modeli analiza dannykh: OLAP i DataMining* [Methods and models of data analysis: OLAP and DataMining]. St. Petersburg, Russia: BKhV-Peterburg, 2008, 336 p.

[16] Yu.Yu. Petrunin, *Informatsionnye tekhnologii analiza dannykh* [Informational technologies for data analysis]. Moscow, Russia: KDU, 2010, pp. 180-200.

[17] J.V. Plas, *Python dlya slozhnykh zadach. Nauka o dannykh i mashinnoe obuchenie: Rukovodstvo* [Python data science handbook: Essential tools for working with data]. Moscow, Russia: Piter, 2018.

[18] M.E. Charikar, "Incremental clustering and dynamic information retrieval", *SIAM Journal on Computing*, Vol. 33, No. 6, 2004, pp. 1417–1440.

[19] E.E. Mokina, O.V. Marukhina, M.D. Shagarova, and I.A. Dubinina, "Ispolzovanie metodov Data Mining pri prinyatii meditsinskikh diagnosticheskikh reshenii" [Using data mining methods to make medical diagnostic decisions], *Fundamentalnye issledovaniya*, No. 5-2, 2016, pp. 269-274.

[20] A. Bühl, and P. Zöfel, *SPSS: Iskusstvo obrabotki informatsii. Analiz statisticheskikh dannykh i vosstanovleniye skrytykh zakonomernostey* [SPSS: The art of information processing. Analysis of statistical data and recovery of hidden patterns]. Moscow, Russia: DiaSoft, 2005, pp. 384-403.