

Using the K-means Method for Diagnosing Cancer Stage Using the Pandas Library

Ievgen Meniailov^[0000-0002-9440-8378], Kseniia Bazilevych^[0000-0001-5332-9545],
Kirill Fedulov^[0000-0001-9619-0299], Sergey Goranina^[0000-0001-8988-3935],
Dmytro Chumachenko^[0000-0003-2623-3294]

National Aerospace University "Kharkiv Aviation Institute", Chkalova str., 17, Kharkiv, 61070,
Ukraine

j.menyailov@khai.edu, k.bazilevych@khai.edu,
fedulov.kirill172@gmail.com, sgoranin@gmail.com,
dichumachenko@gmail.com

Abstract. The characteristics of the patients have a great influence on the determination of the probability of the stage of cancer. To determine the significant factors for assessing the degree of influence, statistical methods of data analysis are most often used. Recently, however, Data Mining methods have become widely used in medicine, which, with large amounts of information and complex relationships, can provide more accurate estimates, especially with a large number of similar characteristics. This paper discusses the problem of clustering data to determine the stages of cancer of patients with similar characteristics. To solve the problem, the k-means method with normalization was used, and the Python language and the Pandas library were chosen to implement the algorithm. The developed software module allows the visualization of the algorithm. This system also supports downloading and uploading data service according to safety requirements of data.

Keywords. K-means; data mining; cluster analysis; differential diagnosis.

1 Introduction

In various areas of human activity (economics, finance, medicine, business, geology, chemistry, etc.), every day there is a need to solve the problems of analysis, prediction and diagnosis, identify hidden dependencies and support optimal decision-making. Due to the rapid growth in the volume of information, the development of technologies for its collection, storage and organization in databases and data warehouses, accurate methods for analyzing information and modeling the objects under study often lag behind the needs of real life. It requires universal and reliable approaches suitable for processing information from various fields, including solving problems that may arise in the near future. The technologies and approaches of the mathematical theory of recognition and classification can be used as a similar basis.

Indeed, these approaches use as a source of information only sets of descriptions of objects, objects, situations or processes (sample of precedents), with each individual observation-precedent being recorded as a vector of values of its individual attribute properties. Samples of feature descriptions are the simplest standardized representations of primary source data that arise in various subject areas in the process of collecting information of the same type and which can be used to solve the following tasks:

- recognition (classification, diagnostics) of situations, phenomena, objects or processes with justification of decisions [1-4];
- prediction of situations, phenomena, processes or states by sampling dynamic data [5-7];
- cluster analysis [8,9] and data structure research [10,11];
- identification of essential features and finding the simplest descriptions;
- finding empirical patterns of various types;
- construction of analytical descriptions of sets (classes) of objects;
- finding non-standard or critical cases;
- formation of reference descriptions of images.

Diagnostics plays an important role in medicine, and diagnosis requires a great deal of skill, knowledge and intuition from a doctor.

The accuracy of the diagnosis and the speed with which it can be made depend on very many factors: on the patient's condition, on the available data on the symptoms of the disease and the results of laboratory tests, as well as on the total amount of medical information about the observation of such symptoms in various diseases, qualifications of the doctor himself [12, 13].

A timely and accurate diagnosis at an early stage often facilitates the choice of treatment method and significantly increases the probability of recovery of the patient [14,15].

The development of software systems for analyzing data and forecasting precedents is actively carried out in leading foreign countries.

First of all, these are statistical data processing and visualization packages (SPSS, STADIA, STATGRAPHICS, STATISTICA, SYSTAT, Forecast Expert, and others), which are based on the methods of various sections of mathematical statistics - testing statistical hypotheses, regression analysis, variance analysis, time series analysis, etc.

The use of statistical software products has become a standard and effective tool for data analysis, and, above all, the initial stage of research, when the values of various averaged indicators are found, the statistical reliability of various hypotheses is checked, and regression dependencies are found.

However, statistical approaches have significant drawbacks. They make it possible to estimate (under certain conditions) the statistical reliability of the value of the predicted parameter, hypothesis or dependence, however the methods for calculating predicted values, hypothesizing or finding dependencies themselves have obvious limitations [16].

First of all, the values averaged over the sample are found, which can be a fairly rough idea of the parameters being analyzed or predicted. Any statistical model uses

the concepts of "random events", "distribution functions of random variables", etc., while the relationship between the various parameters of the objects, situations or phenomena under investigation are deterministic.

The very use of statistical methods implies the presence of a certain number of observations for the validity of the final result, especially for accurate diagnosis. At the same time, the problem of processing and analyzing information obtained in the course of the medical activity of a medical institution is currently one of the most pressing.

2 Rationale and Purpose of the Research

It should also be noted that the methodology of using mathematical classification methods in medical diagnostics tasks is not yet sufficiently developed, there is no methodical justification for using classification algorithms, especially regarding the study of cluster structure and its identification, a number of issues of evaluation and interpretation of the diagnoses remain unresolved.

All this greatly hinders the widespread implementation of the results of solving classification problems in the practice of medical institutions and at the same time makes the study of this problem relevant [17].

To date, the creation and maintenance of modern computerized database of patient characteristics, the course of diseases, laboratory tests and treatment is not a task that is difficult for specialists in the information sphere. The problem is the lack of an effective information technology for processing and analyzing data that would enable the medical analyst to identify hidden patterns and interrelationships of various factors in medical data, which ultimately would increase the effectiveness of treatment by choosing the intensity of therapy adequate to the state of the patient's body on the basis of the identified risk factors [18-19].

To solve these problems, it is proposed to apply mathematical modeling as well as appropriate software supporting libraries for data analysis that are suitable for processing large volumes of data at high speed. To solve this problem, Python was chosen, which has a number of convenient libraries of machine learning and scientific calculations: Pandas, NumPy, SciPy, Scikit-Learn, which allow you to quickly build working models in the field of Data Science. Such an approach to diagnosing tasks can satisfy the need for flexibility, scalability, speed, speed up response to changes, and optimize the data processing processes that need to be used in medical practice.

The purpose of the study is to analyze the statistical dependence between the variables that determine the condition of patients; determination of the patient's belonging to a certain class (oncological disease stage) on the basis of data of registered state variables.

Object of study: the process of diagnosing the state of elements of dynamic systems. Subject of research: mathematical models and methods for solving problems of statistical data analysis and classification of states of elements of dynamic systems.

The main goal of cluster analysis is to find groups of similar objects in a data sample. These groups are conveniently called clusters. There is no generally accepted or

simply useful definition of the term "cluster", and many researchers believe that it is too late or there is no need to try to find such a definition. Despite the lack of definition, it is clear that clusters have some properties, the most important of which are density, dispersion, size, shape, and separability.

Cluster methods form seven main families:

- hierarchical agglomerative methods;
- iterative grouping methods;
- methods for finding modal density values;
- factor methods;
- condensation methods;
- methods using graph theory.

These families correspond to different approaches to creating groups, and applying different methods to the same data can lead to very different results. In specific branches of science, certain families of methods may be particularly useful [20].

3 Experiments and Results of the Modeling

To solve the problem of analyzing the statistical dependence between the variables that determine the state of the patients, the k-means method was chosen. Consider the stages of solving the problem:

First, a preliminary division of the sample of objects into groups is carried out. The k most distant points are selected and the objects are distributed into groups as sets of objects for which one of the selected points is the nearest. The proximity function is calculated by the user-specified metric.

Then iterative optimization of the penalty functional is carried out - the sum of the intraclass variations by the formula (1):

$$J = \sum_{p=1}^k J_p, \quad (1)$$

$$J_p = \frac{1}{|T_p|} \sum_{\mathbf{x}_i \in T_p} \rho(\mathbf{x}_i, \mathbf{y}_p)^2,$$

where, \mathbf{y}_p is center of pth group T_p , to which the object is referred \mathbf{x}_i . J_p equal to the mean square of the distance from the objects assigned to the p-th group to its center (intraclass spread). At each iteration grouping is selected T_p , object \mathbf{x}_i and grouping T_q such that when transferring the object \mathbf{x}_i from T_p into T_q functional J decreases by the maximum value.

The process is completed when no subsequent transfer reduces the functionality (a local minimum is obtained) or the maximum number of iterations specified by the user is reached.

The resulting groupings of objects $T_p, p=1,2,\dots,k$, are considered the desired clusters.

As it is known, cluster analysis works best on a set of normalized vectors, therefore, rationing is necessary before starting clustering. To do this, we use this formula (2):

$$X^* = \frac{X - X_{\min}}{X_{\max} - X_{\min}}, \quad (2)$$

where X^* is new value of cell, X_{\min} is minimum value of vector, X_{\max} is maximum value of vector. This formula lays out the values of all vectors in the range from 0 to 1 inclusive. For the work required to normalize the data in the interval $[-3; 3]$, for this it is necessary to change the formula (3):

$$X^* = (|a| + |b|) \frac{X - X_{\min}}{X_{\max} - X_{\min}} - |a|, \quad (3)$$

where a is left spacing limit, b is right spacing limit.

This algorithm was implemented using the Pandas library and its functions for working with data sets. As an input, a data set was used with information about patients with prostate cancer at different stages of the disease (data provided by the Kharkiv Regional Oncology Center). A complete list of parameters can be seen in Table.

Table – Full list of parameters

Parameter name	Description	Data type
ID	points	string
Age	years	count
KarnovskyScale	points	100-40
VASVASScale	points	0-10
UrinationCount	times	count
Number of urgency	times	count
Nighturination	times	count
Strangury	present or not	0/1

Parameter name	Description	Data type
OZM	present or not	0/1
HZM	present or not	0/1
Residualurine	times	count
Bilateral-Inflam	present or not	0/1
ProstateVolume	sm ³ /mm	1/2/3
PSA	ng/ml	float
Hemoglobin	gr/l	integer
ESR	mm/hour	integer
Leukocytes	10 ⁹ /l	float
Lymphocytes	%	count
SpecificGravity	gr/ml ³	integer
Eritrotsyty	instances in sight	count
LeukocytesUrine	instances in sight	count
Lymphadenopathy	present or not	0/1
Bones	present or not	0/1
Vertebrates	points	integer
G	points	1/2/3
Glisson	points	1-10

In Fig. 1, you can see a part of the data on patients with prostate cancer imported from the data set; a more detailed definition of the characteristics (data columns) is presented in Table.

	0	1	2	3	4	5	6	7	8	9	...	14	15
0	56.0	1.0	11.0	0.5	5.0	0.5	1.0	0.0	300.0	0.5	...	8.8	10.0
1	63.0	1.0	10.0	-0.5	5.0	-0.5	1.0	0.0	250.0	-0.5	...	8.4	25.0
2	60.0	1.0	9.0	0.5	4.0	0.5	1.0	0.0	400.0	0.5	...	6.2	25.0
3	58.0	1.0	9.0	-0.5	3.0	-0.5	1.0	0.0	250.0	-0.5	...	9.9	26.0
4	65.0	1.0	10.0	0.5	4.0	0.5	1.0	0.0	400.0	0.5	...	10.0	35.0

Fig. 1. Imported data

The result of the normalization of the imported data can be seen in Fig. 2. As mentioned above, this is required to obtain the most accurate results.

	0	1	2	3	4	5	6	7	8	9
0	-2.142857	-1.5	-1.090909	-2.636364	-0.5	1.0	3.0	-3.0	-0.750	1.0
1	-0.942857	-1.5	-1.363636	-3.000000	-0.5	-3.0	3.0	-3.0	-1.125	-3.0
2	-1.457143	-1.5	-1.636364	-2.636364	-1.0	1.0	3.0	-3.0	0.000	1.0
3	-1.800000	-1.5	-1.636364	-3.000000	-1.5	-3.0	3.0	-3.0	-1.125	-3.0
4	-0.600000	-1.5	-1.363636	-2.636364	-1.0	1.0	3.0	-3.0	0.000	1.0

Fig. 2. Normalized data

It can be seen that after normalization, the data is indeed in the interval $[-3; 3]$. In Fig. 3, you can see the scatter plot of normalized data.

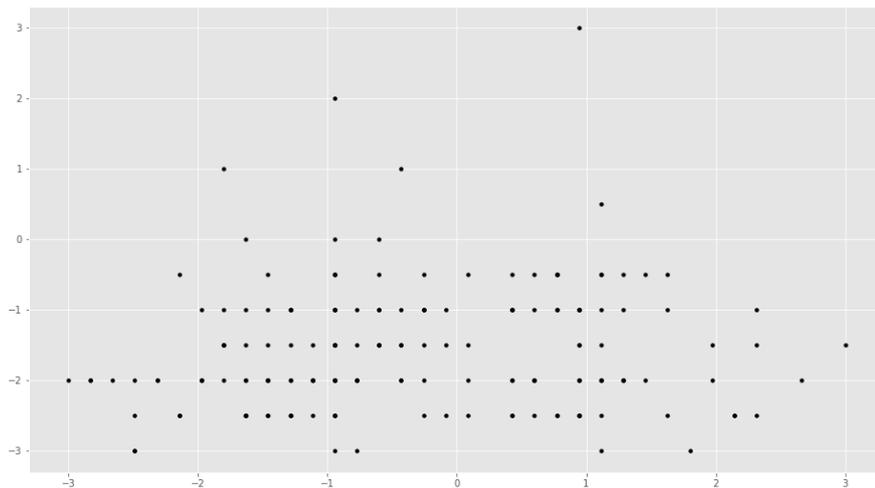


Fig. 3. Scatter plot of normalized data

The first iteration of the k-means method randomly selects k points from the data column. In Fig. 4, you can see the result of selecting centers.

[[1.8000002	-1.5]
	[-0.942857	-2.]
	[-0.25714278	-1.5]
	[-0.25714278	-2.5]

Fig. 4. Mass centers at the first iteration

As you can see, the number of points is 4 because initially 4 clusters were indicated ($k = 6$). In Fig. 5, these centers of mass are represented as green stars (*).

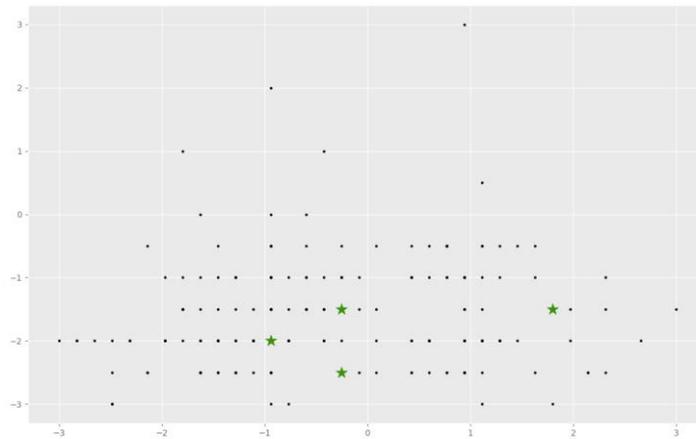


Fig. 5. Scatter plot of normalized data with mass centers

After picking K random points as cluster centers called centroids, algorithm assigns each x_i to nearest cluster by calculating its distance to each centroid, then it finds new cluster center by taking the average of the assigned points and repeats this steps until none of the cluster assignments change.

The final result of the algorithm can be seen in Fig. 6.

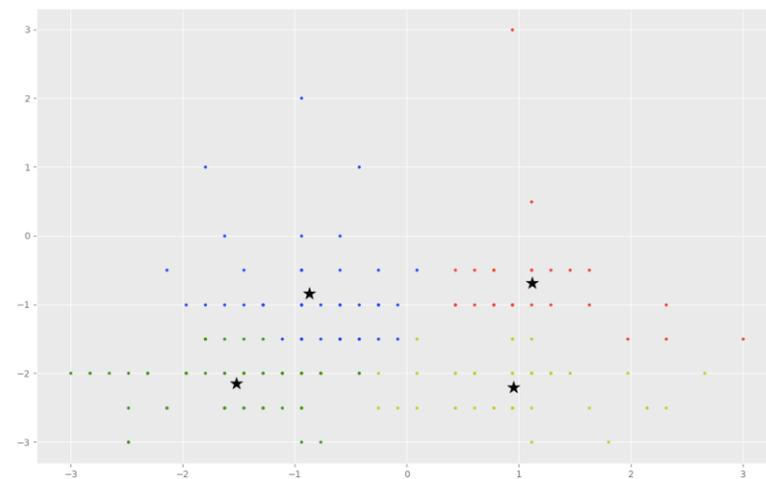


Fig. 6. The end result of the algorithm

4 Conclusion

The study analyzed the statistical dependence between the variables that determine the condition of patients. It is shown how it is possible to determine the patient's belonging to a certain class on the basis of statistical data - registered state variables.

The developed software solution, based on the Python Pandas library, makes it possible to classify on the basis of training samples, which ensures a high percentage of recognition, as well as to identify the precedent belonging to one of several clusters.

It should be noted that the described approach is universal and can be used not only for biomedical systems, but also technical, economic, etc.

References

1. Baldi, P., Brunak, S.: *Bioinformatics: The Machine Learning Approach* (2nd ed.). MIT Press, 400 p. (2001).
2. Berry, M.W.: *Survey of Text Mining: Clustering, Classification, and Retrieval*. Springer, 244 p. (2003).
3. MacQueen, J.B.: Some Methods for classification and Analysis of Multivariate Observations, In: *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, pp. 281-297 (1967).
4. Bazilevych, K. et al.: Stochastic modelling of cash flow for personal insurance fund using the cloud data storage. In: *International Journal of Computing*, Vol. 17, Iss. 3, pp. 153-162 (2018).
5. Brockwell, P.J., Davis, R.A.: *Introduction to Time Series and Forecasting* (2nd ed.). Springer, 428 p. (2002).
6. Berthold, M., Hand, D.J.: *Intelligent Data Analysis: An Introduction* (2nd ed.). Springer-Verlag, 515 p. (2003).
7. Chatfield, C.: *The Analysis of Time Series: An Introduction* (6th ed.). Chapman and Hall, 352 p. (2003).
8. Arabie, P., Hubert, L.J., Soete, G.De.: *Clustering and Classification*. World Scientific, 490 p. (1996).
9. Chaturvedi, A., Green, P., Carroll, J.: K-modes clustering. *Journal of Classification* archive Vol. 18 (1), pp. 35-55 (2001).
10. Deshpande, M., Kuramochi, M., Karypis, G.: Automated approaches for classifying structures. In: *Proceedings 2002 Workshop on DataMining in Bioinformatics (BIOKDD'02)*, Edmonton, Canada, pp. 11-18 (2002).
11. Frakes, W., Baeza-Yates, R.: *Information Retrieval: Data Structures and Algorithms*. Prentice Hall, 512 p. (1992).
12. Ferlay, J., et al.: Cancer incidence and mortality worldwide: sources, methods and major patterns. In: *GLOBOCAN 2012*. Lyon, France, International Agency for Research on Cancer, Vol. 136 (5) (2014).
13. *Cancer control: early detection*. WHO Guide for effective programmes. Geneva: World Health Organization 2007.
14. Rubin, G., et al.: The expanding role of primary care in cancer control. In: *Lancet Oncol*. Elsevier Ltd., Vol. 16(12) (2015).

15. Bowers, N.L., et. al.: Actuarial mathematics. Schaumburg, Illinois, USA by Society Of Actuaries, 621 p. (1997).
16. Norman, T.J.: Bailey The mathematical approach to biology and medicine norman. John Wiley & Sons Ltd., 326 p. (1967).
17. Tom, I.E., et. al.: Technology analysis of medical data by statistical and neural network methods. In: Artificial Intelligence. Vol. 2. p.372-376 (2004).
18. Chumachenko D. et al.: On agent-based approach to influenza and acute respiratory virus infection simulation, In: 14th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering, TCSET 2018. Proceedings, pp. 192-195 (2018).
19. Chumachenko, D.: On Intelligent Multiagent Approach to Viral Hepatitis B Epidemic Processes Simulation. In: Proceedings of the 2018 IEEE 2nd International Conference on Data Stream Mining and Processing, DSMP 2018, pp. 415-419 (2018).
20. Kim, J.O., et al.: Factor, discriminant and cluster analysis. Moscow. Finance and Statistics, 215 p. (1989).