ISSN 2518-170X (Online),
ISSN 2224-5278 (Print)

ҚАЗАҚСТАН РЕСПУБЛИКАСЫ
ҰЛТТЫҚ ҒЫЛЫМ АКАДЕМИЯСЫНЫҢ
Satbayev University

Х А Б А Р Л А Р Ы

ИЗВЕСТИЯ

НАЦИОНАЛЬНОЙ АКАДЕМИИ НАУК
РЕСПУБЛИКИ КАЗАХСТАН
Satbayev University

N E W S

OF THE ACADEMY OF SCIENCES
OF THE REPUBLIC OF KAZAKHSTAN
Satbayev University

SERIES
OF GEOLOGY AND TECHNICAL SCIENCES

2 (446)

MARCH – APRIL 2021

THE JOURNAL WAS FOUNDED IN 1940

PUBLISHED 6 TIMES A YEAR

ALMATY, NAS RK

NAS RK is pleased to announce that News of NAS RK. Series of geology and technical sciences scientific journal has been accepted for indexing in the Emerging Sources Citation Index, a new edition of Web of Science. Content in this index is under consideration by Clarivate Analytics to be accepted in the Science Citation Index Expanded, the Social Sciences Citation Index, and the Arts & Humanities Citation Index. The quality and depth of content Web of Science offers to researchers, authors, publishers, and institutions sets it apart from other research databases. The inclusion of News of NAS RK. Series of geology and technical sciences in the Emerging Sources Citation Index demonstrates our dedication to providing the most relevant and influential content of geology and engineering sciences to our community.

Қазақстан Республикасы Ұлттық ғылым академиясы "ҚР ҰҒА Хабарлары. Геология және техникалық ғылымдар сериясы" ғылыми журналының Web of Science-тің жаңаланған нұсқасы Emerging Sources Citation Index-те индекстелуге қабылданғанын хабарлайды. Бұл индекстелу барысында Clarivate Analytics компаниясы журналды одан әрі the Science Citation Index Expanded, the Social Sciences Citation Index және the Arts & Humanities Citation Index-ке қабылдау мәселесін қарастыруда. Webof Science зерттеушілер, авторлар, баспашылар мен мекемелерге контент тереңдігі мен сапасын ұсынады. ҚР ҰҒА Хабарлары. Геология және техникалық ғылымдар сериясы Emerging Sources Citation Index-ке енуі біздің қоғамдастық үшін ең өзекті және беделді геология және техникалық ғылымдар бойынша контентке адалдығымызды білдіреді.

НАН РК сообщает, что научный журнал «Известия НАН РК. Серия геологии и технических наук» был принят для индексирования в Emerging Sources Citation Index, обновленной версии Web of Science. Содержание в этом индексировании находится в стадии рассмотрения компанией Clarivate Analytics для дальнейшего принятия журнала в the Science Citation Index Expanded, the Social Sciences Citation Index и the Arts & Humanities Citation Index. Web of Science предлагает качество и глубину контента для исследователей, авторов, издателей и учреждений. Включение Известия НАН РК. Серия геологии и технических наук в Emerging Sources Citation Index демонстрирует нашу приверженность к наиболее актуальному и влиятельному контенту по геологии и техническим наукам для нашего сообщества.

«ҚР ҰҒА Хабарлары. Геология және техникалық ғылымдар сериясы».

ISSN 2518-170X (Online),
ISSN 2224-5278 (Print)

Меншіктенуші: «Қазақстан Республикасының Ұлттық ғылым академиясы» РҚБ (Алматы қ.).

Қазақстан Республикасының Ақпарат және қоғамдық даму министрлігінің Ақпарат комитетінде 29.07.2020 ж. берілген **№ KZ39VPY00025420** мерзімдік басылым тіркеуіне қойылу туралы куәлік.

Тақырыптық бағыты: ***геология және техникалық ғылымдар бойынша мақалалар жариялау.***

Мерзімділігі: жылына 6 рет.
Тиражы: 300 дана.

Редакцияның мекен-жайы: 050010, Алматы қ., Шевченко көш., 28, 219 бөл., тел.: 272-13-19, 272-13-18

http://www.geolog-technical.kz/index.php/en/

Типографияның мекен-жайы: «Аруна» ЖК, Алматы қ., Муратбаева көш., 75.

«Известия НАН РК. Серия геологии и технических наук».

ISSN 2518-170X (Online),
ISSN 2224-5278 (Print)

Собственник: Республиканское общественное объединение «Национальная академия наук Республики Казахстан» (г. Алматы).

Свидетельство о постановке на учет периодического печатного издания в Комитете информации Министерства информации и общественного развития Республики Казахстан **№ KZ39VPY00025420**, выданное 29.07.2020 г.

Тематическая направленность: ***публикация статей по геологии и техническим наукам.***

Периодичность: 6 раз в год.
Тираж: 300 экземпляров.

Адрес редакции: 050010, г. Алматы, ул. Шевченко, 28, оф. 219, тел.: 272-13-19, 272-13-18

http://www.geolog-technical.kz/index.php/en/

Адрес типографии: ИП «Аруна», г. Алматы, ул. Муратбаева, 75.

News of the National Academy of Sciences of the Republic of Kazakhstan. Series of geology and technology sciences.

ISSN 2518-170X (Online),
ISSN 2224-5278 (Print)

Owner: RPA "National Academy of Sciences of the Republic of Kazakhstan" (Almaty).

The certificate of registration of a periodical printed publication in the Committee of information of the Ministry of Information and Social Development of the Republic of Kazakhstan **No. KZ39VPY00025420**, issued 29.07.2020.

Thematic scope: ***publication of papers on geology and technical sciences.***

Periodicity: 6 times a year.
Circulation: 300 copies.

Editorial address: 28, Shevchenko str., of. 219, Almaty, 050010, tel. 272-13-19, 272-13-18,

http://www.geolog-technical.kz/index.php/en/

Address of printing house: ST "Aruna", 75, Muratbayev str, Almaty.

N E W S
OF THE NATIONAL ACADEMY OF SCIENCES OF THE REPUBLIC OF KAZAKHSTAN
SERIES OF GEOLOGY AND TECHNICAL SCIENCES
ISSN 2224-5278
Volume 2, Number 446 (2021), 114 – 121 **https://doi.org/10.32014/2021.2518-170X.42**

UDC 637.523

M. A. Nikitina[1], I. M. Chernukha[1], Ya. M. Uzakov[2], D. E. Nurmukhanbetova[3]

[1] V. M. Gorbatov Federal Research Center for Food Systems of RAS, Moscow, Russia;
[2] Almaty Technological University, Almaty, Kazakhstan;
[3] Narxoz University, Almaty, Kazakhstan.
E-mail: dinara.nurmukhanbetova@narxoz.kz

CLUSTER ANALYSIS FOR DATABASES TYPOLOGIZATION CHARACTERISTICS

Abstract. The article deals with basic concepts of cluster analysis and data clustering. The authors give brief information on the history of cluster analysis and its first applications. The article gives the classification of methods by the way of data processing and analysis in cluster analysis. The detailed description of the popular, non-hierarchical K-means algorithm is given. When developing databases, their structure should provide for the division of products into clusters based on various characteristics. It is necessary to consider the division into clusters based on other characteristics, such as allergenicity (whether the product contains an allergic component or not) or carbohydrate content (important for diabetics). The content of protein, potassium and phosphates should be taken into account when developing diets for those suffering from kidney diseases. The presence of specific amino acids - for metabolic diseases, etc. In this way, food composition data and product clustering across different categories allow nutritionists to create interchangeable lists of meals with portion sizes, or lists of permitted and prohibited food products in terms of various diseases. The authors give the clustering of the database fragment of chemical composition of food products on the example of cottage cheese products and confectionary by one of the signs – the content of carbohydrates – in the R software environment by k-means. Food clusters based on carbohydrate content are very important in shaping the diet for diabetics. A visual gradation of products into clusters is demonstrated in the form of a dendrogram showing the degree of proximity of individual clusters. The resulting dendrogram contains 5 clusters. Cluster 4 includes the largest number of products (170 items) with an average carbohydrate content of 1.8 g with a variation range from 0 to 7.1 g. Food products and dishes that fall into this cluster are the least dangerous for people with diabetes. Cluster 5 includes only 8 products with a distribution of carbohydrates within the cluster from 62.60 to 80.40 g. This category of food should be excluded when preparing a diet for people with diabetes.

Keywords: cluster, proximity measure, clustering methods, k-means, dendrogram, characteristic.

Introduction. Clustering means combining objects into groups (clusters) based on the similarity of features for objects in the same group and differences between the groups. Most clustering algorithms do not rely on traditional statistical assumptions; they can be used in conditions where there is almost no information about data distribution laws. Thus, the task of cluster analysis is to divide the initial set of objects into groups that are similar and close to each other. These groups are called clusters or taxons. In addition to the term "clustering", there are a number of terms with similar meanings, such as automatic classification, numerical taxonomy, botryology, and community detection. It is believed that the term "cluster analysis" was first used in the work of the American psychologist Robert C. Tryon from the University of Berkeley [1].

The first works on clustering in the 30-40s of the last century can be attributed to the field of anthropology - Driver and Kroeber [2], psychology - Joseph Zubin [3], Robert Tryon [1] and to the classification of traits in personality psychology [4]. However, the publication of the book Sokal R. R., and Sneath P. H. A. "Principles of Numerical Taxonomy" [5] in 1963 served as an impetus for the development of various methods of cluster analysis. To date, a large number of different clustering algorithms and

their modifications have been developed. For the first time, the monograph of Hartigan J.A. [6] provides an overview of classical (first) methods and algorithms in clustering.

The results obtained by cluster analysis methods are applied in various fields. For example, in the field of medicine, clustering of diseases and symptoms of diseases leads to classifications used to select treatment methods. To create an adequate diet, it is necessary to process large amounts of data related to the chemical composition of food products and dishes. Information such as rules should be stored in databases. The database structure should be divided into clusters, such as "Porridge", "Soups", "Vegetables", etc. Clusters are necessary for the subsequent distribution of food products and dishes included in the diet to separate meals according to time. Along with this, it is necessary to take into account the division into clusters based on various characteristics, such as allergenicity (whether the product contains an allergic component or not), etc. Food clusters can be based on, for example, carbohydrate content for diabetics, or protein, potassium and phosphates for kidney disease sufferers, or specific amino acids for metabolic diseases, and so on. Thus, product composition data and product clustering across different categories allow nutritionists to create interchangeable lists of meals with portion sizes, or lists of permitted and prohibited food products in terms of various diseases.

The purpose of this research is to analyze existing groups of algorithms for classifying food products, ingredients, dishes and diets, and to conduct cluster analysis on the example of cottage cheese products and confectionery.

Organization and research methods. The initial information for clustering is the observation matrix:

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{12} & x_{22} & \dots & x_{2n} \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix}$$

each of its line represents the values of n attributes of one of the Mclusterization objects. The clustering task is to break objects from X into several subsets (clusters), where the objects are more similar to each other than to objects from other clusters. In metric space, the "similarity" is usually defined in terms of a distance. The distance can be calculated either between the source objects (lines of the X matrix), or between these objects to the cluster prototype. Usually, prototype coordinates are not known in advance – they are found simultaneously with data breaking into clusters.

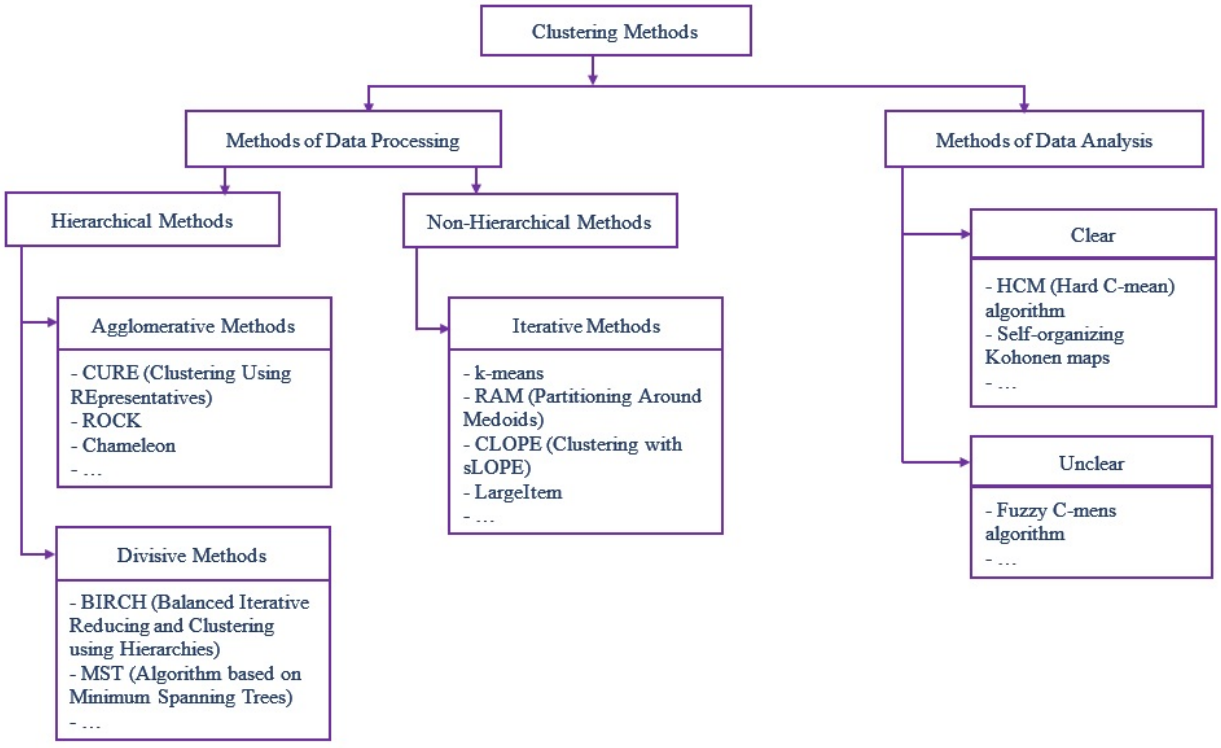


Figure 1 – Classification of methods by the method of data processing and analysis in cluster analysis

Data of the observation matrix must be normalized, i.e., reduced to dimensionless values. If the variables are not normalized, they will not affect the distance equally (i.e. if we measure vitamins in milligrams and micrograms, and protein, fat, and carbohydrates - in grams, then protein, fat, and carbohydrates will not be able to affect anything at all).

There are many clustering methods that can be classified as follows: 1) by the method of data processing [7-14]; 2) by the method of data analysis [15-16]; 3) by the number of applications of clustering algorithms [17-22]; 4) by the possibility of expanding the scope of processed data [23-24]; 5) by the time of clustering [25-26]. A more detailed classification of methods by the method of data processing and analysis in cluster analysis is shown in (figure 1).

All clustering methods work with data as vectors in a multidimensional space. Each vector is determined by the values of several directions, while the directions themselves are the characteristics we know (the content of protein, fat, amino acids, vitamins, etc. in the product). Characteristics can be both quantitative and qualitative, and the art of a data mining specialist is to correctly select and normalize these characteristics, and then choose the appropriate measure of distances. Only then clustering algorithms shall be applied.

Results and discussion. The criterion for determining the similarity and difference of clusters is the distance between points on the scattering diagram. There are several ways to determine the measure of distance between clusters, called the proximity measure: 1) *Mahalanobis distance* (general view); 2) *ordinary Euclidean distance*; 3) *"weighted" Euclidean distance*; 4) *Hemming distance*; 5) *Chebyshev distance*; 6) *power distance*; 7) *percent disagreement*. Depending on the research purpose, one or another formula is chosen to determine the measure of proximity. The authors have analyzed the most used algorithms. One of the most popular non-hierarchical algorithms is the K-means algorithm. It was invented in the 1950s by the mathematician Hugo Steingauz [27] almost simultaneously with Stuart Lloyd [28]. It became particularly popular after McQueen's work [29].

The k-means algorithm is popular due to its ease of implementation and speed [9-10]. Its main drawback is its convergence to the local minimum and dependence of the result on the initial distribution. You also need to know the estimated number of k clusters in advance. The main idea of the k-means algorithm is that the center of mass for each cluster obtained in the previous step is recalculated at each iteration, then vectors are divided into clusters again taking into account the closest of the new centers by the chosen metric. The algorithm ends when there is no change in the intra-cluster distance at some iteration. This happens in a finite number of iterations, since the number of possible breakings of a finite set is finite, and at each step the total square deviation decreases, thus, looping is impossible.

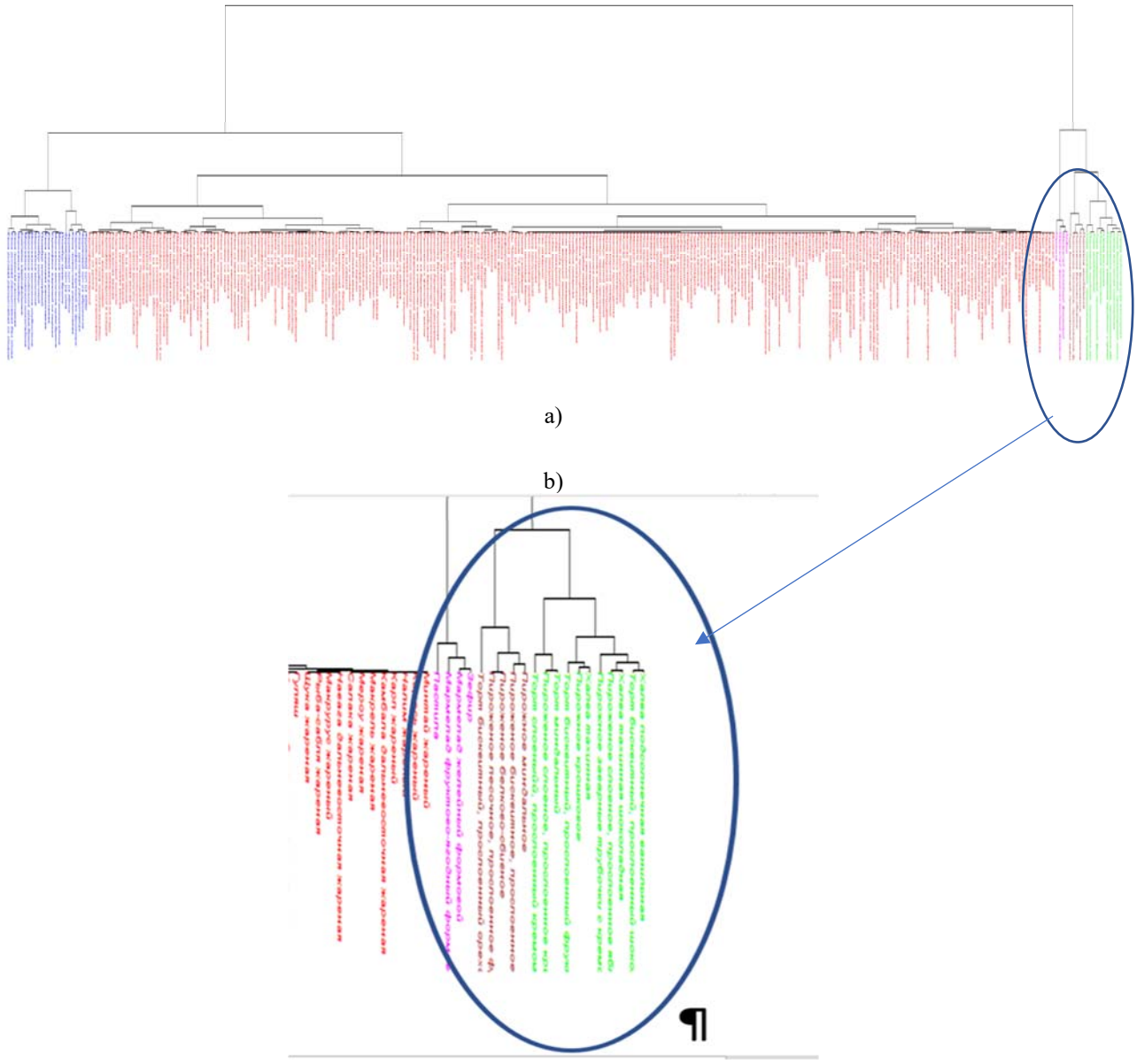
```
# To upload and prepare data
# To output data elements
head(prod)
                                                      Carbohydrates
  Borshch with fresh cabbage and tomato                    9.8
  Borshchwithsauerkraut                                    8.8
  Navy-style borshch with meat                            11.5
  Borshch with fresh cabbage, potatoes and meat            5.7
  Krasnodar borshch with meat                              8.3
# To normalize data
prod <- scale(prod)
# To build a dendrogram
hc<- hclust(dist(prod))
ph<- as.phylo(hc)
groups5 <- cutree(hc, k = 5)
colors = c("red", "blue", "green", "brown","magenta")
plot(ph, tip.color = colors[groups5], cex = 0.6)
# To conduct clusterization using the k-means method
kmeans(prod, 5, 1000000)
```

Figure 2 – Program code listing in the R software environment

It is a well-known fact that to support a healthy lifestyle and maintain health, it is necessary to prepare a diet that meets the needs and capabilities of the human body and is balanced in all indicators of nutritional and biological value. That is, taking into account the human metabolism. For this purpose, decision support systems are being developed. The information basis of such systems is a database of products, ingredients, and dishes that are most common and sold in large cities and megacities. The database structure should provide for the distribution of products into clusters based on various criteria. Dividing products into clusters will allow excluding "undesirable" products from the diet when creating a menu. So, for example, for patients with diabetes – this is the quantitative content of carbohydrates in products. The largest amount of carbohydrates is mainly found in dairy and confectionery products [30-31].

The authors have conducted a clustering of the database fragment by one of the signs – carbohydrate content – on the example of cottage cheese products and confectionary in the R software environment. The software code fragment is shown in (figure 2).

At the beginning of the study, a dendogram was formed. A dendrogram is a visualization of results of hierarchical clustering. It allows visually assessing the degree of proximity of individual objects and clusters, as well as graphically demonstrating the sequence of their association or separation. The number of dendrogram levels corresponds to the number of steps for merging or dividing clusters.



a)

b)

Figure 3 –Dendrogram: a) general view, b) cluster 1 (green font color)

The general view of the dendrogram for clustering products based on carbohydrate content is shown in (figure 3). Color zones (red, blue, green, brown, pink) in the dendogram visually display the division of products into clusters. Note that in the R environment, numbering goes from the right to the left. As can be seen from (figure 3), after classification, food products and dishes are grouped into 5 clusters with different carbohydrate content. The largest number of products included in cluster 4 is of red color.

For the detailed consideration of each formed cluster, one can build one's own dendrogram separately (to display it on the screen). For example, for cluster 1 (it is highlighted in green on the general dendrogram) containing 12 products, the dendrogram looks as follows (figure 4).

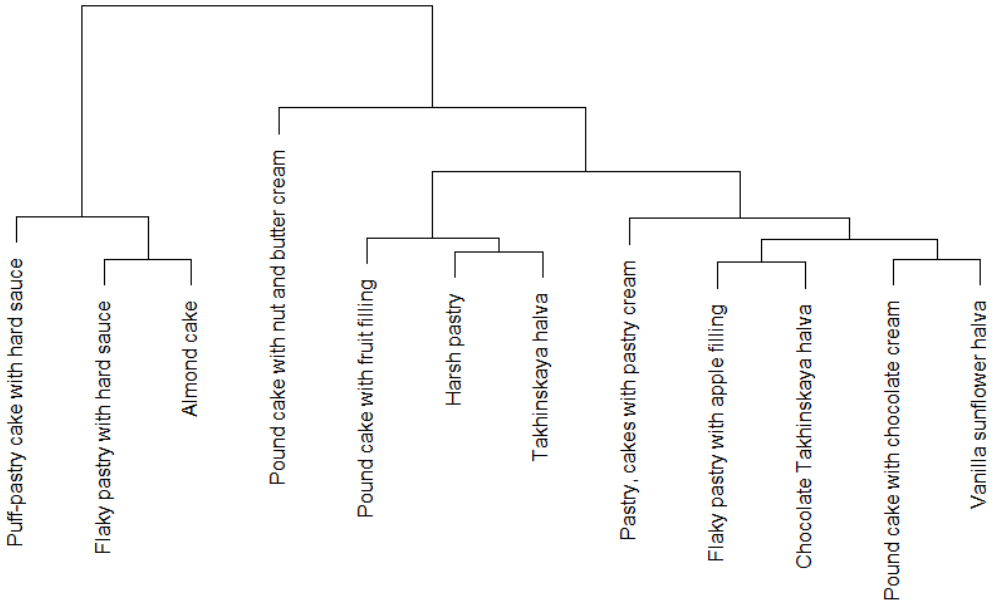


Figure 4 – Dendrogram of cluster 1

Along with the dendrogram, a table is formed with details for each cluster: 1) the number of products included in this cluster; 2) the average value of carbohydrates for this cluster; 3) the maximum content of carbohydrates in the product of this cluster; 4) the minimum content of carbohydrates in the product in this cluster. Thus, as a result of clustering using the k-means algorithm, we have obtained the following division of products provided in table 1, which can be interpreted by the carbohydrate content as products with low carbohydrate content (class 4), medium (class 2) and high carbohydrate content (class 5). The largest number of products (170 items) fell into cluster 4 with an average carbohydrate content of 1.8 g. with a variation range from 0 to 7.1 g. The smallest number of products (8 items) fell into cluster 5 with an average carbohydrate content of 70.91 g. with a variation range from 62.60 to 80.40 g.

Table1 – Clustering results

Cluster	Quantity by the Product field	Average by the Carbohydrates field	Maximum by the Carbohydrates field	Minimum by the Carbohydrates field
5	8	70.91	80.40	62.60
1	12	51.32	58.80	44.00
2	38	22.86	35.40	18.60
3	102	13.20	17.60	7.70
4	170	1.80	7.10	0.00
Grand total	330	11.22	80.40	0.00

Table 2 Products included in cluster 5

Cluster	Quantity by the Product field	Average by the Carbohydrates field	Maximum by the Carbohydrates field	Minimum by the Carbohydrates field
5	8	70.91	80.40	62.60
Marshmallows	1	78.30	78.30	78.30
Jelly shaped marmalade	1	77.70	77.70	77.70
Fruit and berry shaped marmalade	1	76.00	76.00	76.00
Paste	1	80.40	80.40	80.40
Protein-churned cake	1	62.60	62.60	62.60
Pound cake with fruit filling	1	64.20	64.20	64.20
Short pastry with fruit filling	1	62.60	62.60	62.60
Almond cake	1	65.50	65.50	65.50

A user can see more detailed information when each cluster is expanded. Which product subgroups or products are included in this cluster. Table 2 shows the products included in cluster 5.

Conclusion. The article shows the possibility of using cluster analysis to classify food products, ingredients, dishes and diets using the example of cottage cheese products and confectionery. There is a brief information on the history of cluster analysis and its first applications. The article gives the main terms and definitions, classification of methods and algorithms of cluster analysis. Based on the analysis, the k-means method was chosen as the implementation method. The advantage of this method is simplicity and speed of use as well as visibility when dividing elements (products) into clusters. Clustering of the database fragment of the chemical composition of products and dishes based on "carbohydrate content" was performed in the R software environment. The visual gradation of products into clusters in the form of a dendrogram is demonstrated. The resulting dendrogram contains 5 clusters. Cluster 4 in-cludes the largest number of products (170 items), with an average carbohydrate content of 1.8 g. with a variation range from 0 to 7.1 g. Food products and dishes that fall into this cluster are the least dangerous for people with diabetes. Cluster 5 includes only 8 products with a distribution of carbohydrates within the cluster from 62.60 to 80.40 g. This category of food should be excluded when preparing a diet for people with diabetes.

Acknowleadgment. This article is published as part of scientific research theme No. 0585-2019-0008 under the state assignment of the federal state budgetary scientific institution ‘V.M. Gorbatov Federal Research Centre for Food Systems’ of RAS.

М. А. Никитина[1], И. М. Чернуха[1], Я. М. Ұзақов[2], Д. Е. Нурмуханбетова[3]

[1]«В. М. Горбатов атындағы тағамдық жүйелердің федералдық ғылыми орталығы» РҒА, Мәскеу, Ресей;
[2]Алматы технологиялық университеті, Алматы, Қазақстан;
[3]Нархоз Университеті, Алматы, Қазақстан

ТАМАҚ ӨНІМДЕРІН ЖӘНЕ ТАҒАМДАРДЫ КЛАСТЕРЛІК ТАЛДАУ АРҚЫЛЫ ТИПТЕУ

Аннотация. Мақалада кластерлік талдаудың негізгі ұғымдар және кластерлеу деректері қарастырылған. Кластерлік талдаудың пайда болу тарихы, алғашқы салалардағы оның қолдануы бойынша қысқаша мәліметтер берілді. Кластерлік талдауда өңдеу тәсілі бойынша әдістерді жіктеу мен талдау деректері келтірілген. Қолданыстағы әдістер мен деректерді кластерлеу алгоритмдері: 1)өңдеу тәсілі бойынша деректер талданды; 2)деректерді талдау тәсілі бойынша талданды; 3)кластерлеу алгоритмдерін қолдану саны бойынша талданды; 4)өңделетін деректердің көлемін кеңейту мүмкіндігінше талданды; 5)кластерлеу уақыты орындалу бойынша талданды.

Белгілі k-орташа иерархиялықсыз алгоритмі тиянақты сипатталған. Барабар тамақтану рационын жасау кезде көлемі үлкен деректерді, кейде құрылымдық емес немесе әлсіз құрылымдыны пайдалану қажет. Деректер базасын жасау кезінде оның құрылымындағы өнімдерді әр түрлі сипаттамалар бойынша кластерлерге бөлуін қарастыру қажет. Басқа белгілер бойынша кластерлерге бөлінетінін ескеру қажет бөлу, мысалы, аллергендігі бойынша (өнімнің құрамында аллергиялық компонент бар немесе жоғын) немесе көмірсулардың болуы бойынша (диабетиктер үшін маңызды). Ақуыз, калий және фосфаттардың болуын бүйрек ауруларына

шалдыққан адамдардың рационын жобалау кезінде ескеру қажет. Метаболикалық аурулар және т.б. үшін нақты амин қышқылдардың қатысуы. Осылайша, өнімнің құрамы туралы мәліметтер мен әртүрлі санаттағы өнімдердің кластеризациясы диетологтарға порция өлшемдеріне қарай алмастырылатын тағам тізімдерін қалыптастыруға немесе әр түрлі аурулар тұрғысынан алғанда рұқсат етілген және рұқсат етілмеген өнімдерді қалыптастыруға мүмкіндік береді.

Өнімдер мен тағамдар химиялық құрамының деректер базасындағы фрагмент кластеризациясы келтірілген оны ірімшікті өнімдер мен кондитерлік бұйымдардың мысалынан бір қасиеті бойынша–көмірсулардың болуы – R бағдарламалық ортада k-орташа әдісімен көруге болады. Көмірсулар мәні бойынша өнімдердің кластерлері диабетиктер үшін тамақтану рационын қалыптастыру кезінде өте маңызды. Жеке кластерлердің жақын орналасқан дәрежесін көрсететін дендрограмма түрінде құрастыру арқылы өнімдердің кластерлерге визуалды градациясы көрсетілді. Кластерлеу кестесі пайдаланушыға әрбір кластерді ашу кезінде егжей-тегжейлі ақпаратты көруге мүмкіндік береді: нақты қай кластерге қандай топ өнімдері немесе өнімдер кіреді.

Түйін сөздер: кластер, жақындау шарасы, кластерлеу әдістері, k-means, дендрограмма.

М. А. Никитина[1], И. М. Чернуха[1], Я. М. Узаков[2], Д. Е. Нурмуханбетова[3]

[1]ФГБНУ «Федеральный научный центр пищевых систем им. В. М. Горбатова» РАН, Москва, Россия;
[2]Алматинский технологический университет, Алматы, Казахстан;
[3]Университет Нархоз, Алматы, Казахстан

КЛАСТЕРНЫЙ АНАЛИЗ ДЛЯ ТИПОЛОГИЗАЦИИ ПИЩЕВЫХ ПРОДУКТОВ И БЛЮД

Аннотация. В статье рассмотрены основные понятия кластерного анализа и кластеризации данных. Даны краткие сведения по истории возникновения кластерного анализа, первых областях его применения. Приведена классификация методов по способу обработки и анализу данных в кластерном анализе. Проанализированы существующие методы и алгоритмы кластеризации данных: 1) по способу обработки данных; 2) по способу анализа данных; 3) по количеству применений алгоритмов кластеризации; 4) по возможности расширения объема обрабатываемых данных; 5) по времени выполнения кластеризации.

Наиболее подробно описан популярный, неиерархический алгоритм k-средних. При составления адекватного рациона питания необходимо оперировать большим объемом данных, иногда не структурированным или слабоструктурированным. При разработке баз данных, в ее структуре следует предусмотреть деление продуктов на кластеры по различным характеристикам. Наряду с этим необходимо учесть деление на кластеры по другим признакам, например, аллергенности (содержит ли в своем составе продукт аллергический компонент) или содержании углеводов (важно для диабетиков). Содержание белка, калия и фосфатов следует учесть при проектировании рационов для страдающих заболеваниями почек. Присутствие конкретных аминокислот – для метаболических заболеваний и т.д. Таким образом, данные о составе продуктов и кластеризация продуктов по различным категориям позволяют диетологам формировать взаимозаменяемые списки блюд с размерами порций или списки разрешенных и неразрешенных продуктов с точки зрения различных заболеваний.

Приведена кластеризация фрагмента базы данных химического состава продуктов и блюд на примере творожных продуктов и кондитерских изделий по одному из признаков – содержание углеводов – в программной среде R методом k-средних. Кластеры продуктов по содержанию углеводов очень важны при формировании рациона питания для диабетиков. Продемонстрирована визуальная градация продуктов на кластеры в виде построения дендрограммы, показывающая степень близости отдельных кластеров. Кластеризации позволяет пользователю увидеть более детальную информацию при раскрытии каждого кластера: какие подгруппы продуктов или продукты входят в данный кластер.

Ключевые слова: кластер, мера близости, методы кластеризации, k-means, дендрограмма.

Information about authors:

Nikitina M. A., candidate of technical sciences, docent, leading scientific worker, the Head of the Direction of Information Technologies, V. M. Gorbatov Federal Research Center for Food Systems of Russian Academy of Sciences, Moscow, Russia; m.nikitina@fncps.ru; https://orcid.org/0000-0002-8313-4105

Chernukha I. M., doctor of technical sciences, professor, Academician of the Russian Academy of Sciences, V.M. Gorbatov Federal Research Center for Food Systems of Russian Academy of Sciences, Moscow, Russia; imcher@inbox.ru; https://orcid.org/0000-0003-4298-0927

Uzakov Ya. M., doctor of technical sciences, professor, Almaty Technological University, Almaty, Kazakhstan; uzakm@mail.ru; https://orcid.org/0000-0003-4626-2471

Nurmukhanbetova D. E., candidate of technical sciences, associate professor, Narxoz University, Almaty, Kazakhstan; dinara.nurmukhanbetova@narxoz.kz; https://orcid.org/0000-0002-8939-6325

REFERENCES

[1] Tryon R.C. (1939) Cluster Analysis: Correlation Profile and Orthometric (factor) Analysis for the Isolation of Unities in Mind and Personality. Edwards Brothers.

[2] Driver H.E., Kroeber A.L. (1932) Quantitative Expression of Cultural Relationships // University of California Publications in American Archaeology and Ethnology, 31(4):211-256.

[3] Zubin J. (1938) A technique for measuring like-mindedness // Journal of Abnormal and Social Psychology, 33(4): 508-516. https://doi.org/10.1037/h0055441

[4] Cattell R.B. (1943) The description of personality: basic traits resolved into clusters // Journal of Abnormal and Social Psychology, 38(4): 476-506. https://doi.org/10.1037/h0054116

[5] Sokal R.R., Sneath P.H.A. (1963) Principles of Numerical Taxonomy. San Francisco: W.H. Freeman. ISBN: 0-7167-0697-0.

[6] Hartigan J.A. (1975) Clustering algorithms. N.Y.: John Wiley & Sons. ISBN 0-471-35645-X.

[7] Guha S., Rastogi R., Shim K. (1998) CURE: An Efficient Clustering Algorithm for Large Databases // ACM SIGMOD Record, 27(2): 73-84. https://doi.org/10.1145/276305.276312

[8] Zhang T., Ramakrishnan R., Livny M. (1996) BIRCH: An Efficient Data Clustering Method for Very Large Databases // ACM SIGMOD Record, 25(2):103-114. https://doi.org/10.1145/235968.233324

[9] Li Y., Wu H. (2012) A Clustering Method Based on K-Means Algorithm // Physics Procedia, 25: 1104-1109. https://doi.org/10.1016/j.phpro.2012.03.206

[10] Bai L., Liang J., Cao F. (2020) A multiple k-means clustering ensemble algorithm to find nonlinearly separable clusters // Information Fusion, 61:36-47. https://doi.org/10.1016/j.inffus.2020.03.009

[11] Paklin N.Category data clustering: CLOPE scalable algorithm.– Electronic resource – https://loginom.ru/blog/clope (last accessed 05.05.2020) (in Russ.).

[12] Yang Y., Guan H., You J. (2002) CLOPE: A fast and Effective Clustering Algorithm for Transactional Data // In Proceedings of SIGKDD'02, July 23-26, 2002, Edmonton, Alberta, Canada, 682-687.

[13] Wang K., Xu C., Liu B. (1999) Clustering transactions using large items // In Proceedings of CIKM'99, Kansas, Missouri, 483-490. https://doi.org/10.1145/319950.320054

[14] Paklin N. Clustering algorithms in the Data Mining service – Electronic resource – https://loginom.ru/blog/data-mining-clustering (last accessed 05.05.2020) (in Russ.).

[15] Barsegyan A.A., Kupriyanov M.S., Stepanenko V.V., Holod I.I. (2004) Data analysis methods and models: OLAP and Data Mining. SPb.: BHV-Petersburg. ISBN: 5-94157-991-8 (in Russ.).

[16] Bai L., Liang J., Du H., Guo Y. (2019) An information-theoretical framework for cluster ensemble // IEEE Transactions on Knowledge and Data Engineering, 31(8):1464-1477. https://doi.org/10.1109/TKDE.2018.2865954

[17] Huang D., Wang C., Wu J., Lai J., Kwoh C. (2020) Ultra-scalable spectral clustering and ensemble clustering // IEEE Transactions on Knowledge and Data Engineering, 32(6):1212-1226. https://doi.org/10.1109/TKDE.2019.2903410

[18] Yu Z., Zhu X., Wong H., You J., Zhang J., Han G. (2017) Distribution-based cluster structure selection // IEEE Transactions on Cybernetics, 47(11):3554-3567. https://doi.org/10.1109/TCYB.2016.2569529

[19] Yang Y., Jiang J. (2016) Hybrid sampling-based clustering ensemble with global and local constitutions // IEEE Transactions on Neural Networks and Learning Systems, 27(5):952-965. https://doi.org/10.1109/TNNLS.2015.2430821

[20] Iam-On N., Boongoen T. (2015) Comparative study of matrix refinement approaches for ensemble clustering // Machine Learning, 98:269-300. https://doi.org/10.1007/s10994-013-5342-y

[21] Gonzalez E., Turmo J. (2015) Unsupervised ensemble minority clustering // Machine Learning, 98:217-268. https://doi.org/10.1007/s10994-013-5394-z

[22] He L., Ray N., Guan Y., Zhang H. (2019) Fast large-scale spectral clustering via explicit feature mapping // IEEE Transactions on Cybernetics, 49(3):1058-1071. https://doi.org/10.1109/TCYB.2018.2794998

[23] Wu J.S., Zheng W.S., Lai J.H., Suen C.Y. (2018) Euler clustering on large-scale dataset // IEEE Transactions on Big Data, 4(4): 502-515. https://doi.org/10.1109/TBDATA.2017.2742530

[24] Liu H., Zhao R., Fang H., Cheng F., Fu Y., Liu Y.-Y. (2017) Entropy-based consensus clustering for patient stratification // Bioinformatics, 33(17): 2691-2698. https://doi.org/10.1093/bioinformatics/btx167

[25] Bryant A., Cios K. (2018) RNN-DBSCAN: A density-based clustering algorithm using reverse nearest neighbor density estimates // IEEE Transactions on Knowledge and Data Engineering, 30(6): 1109-1121. https://doi.org/10.1109/TKDE.2017.2787640

[26] Yu Z., Kuang Z., Liu J., Chen H., Zhang J., You J., Wong H.-S., Han G. (2017) Adaptive ensembling of semi-supervised clustering solutions // IEEE Transactions on Knowledge and Data Engineering, 29 (8): 1577-1590. https://doi.org/10.1109/TKDE.2017.2695615

[27] Steinhaus H. (1956) Sur la division des corps materielsen parties // Bulletin de l'academie ´polonaise des sciences, C1. III vol. IV(12): 801-804.

[28] Lloyd S.P. (1957) Least square quantization in PCM's // Bell Telephone Laboratories Paper.

[29] MacQueen J. (1967) Some methods for classification and analysis of multivariate observations // In Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability - University of California Press, Berkeley, California, 1:281-297.https://projecteuclid.org/euclid.bsmsp/1200512992

[30] Turovskaya S.N., Galstyan A.G., Petrov A.N., Radaeva I.A., Illarionova E.E., Ryabova A.E., Asembaeva E.K., Nurmukhanbetova D.E. (2018) Scientific and practical potential of dairy products for special purposes // News of National academy of sciences of the Republic of Kazakhstan. Series of geology and technical sciences. 432(6): 16-22. https://doi.org/10.32014/2018.2518-170X.3

[31] Lisitsyn A., Chernukha I., Nikitina M.(2020) Russian methodology for designing multicomponent foods in retrospect // Foods and raw materials, 8(1):2-11. https://doi.org/10.21603/2308-4057-2020-1-2-11

Publication Ethics and Publication Malpractice
in the journals of the National Academy of Sciences of the Republic of Kazakhstan

For information on Ethics in publishing and Ethical guidelines for journal publication see http://www.elsevier.com/publishingethics and http://www.elsevier.com/journal-authors/ethics.

Submission of an article to the National Academy of Sciences of the Republic of Kazakhstan implies that the described work has not been published previously (except in the form of an abstract or as part of a published lecture or academic thesis or as an electronic preprint, see http://www.elsevier.com/postingpolicy), that it is not under consideration for publication elsewhere, that its publication is approved by all authors and tacitly or explicitly by the responsible authorities where the work was carried out, and that, if accepted, it will not be published elsewhere in the same form, in English or in any other language, including electronically without the written consent of the copyright-holder. In particular, translations into English of papers already published in another language are not accepted.

No other forms of scientific misconduct are allowed, such as plagiarism, falsification, fraudulent data, incorrect interpretation of other works, incorrect citations, etc. The National Academy of Sciences of the Republic of Kazakhstan follows the Code of Conduct of the Committee on Publication Ethics (COPE), and follows the COPE Flowcharts for Resolving Cases of Suspected Misconduct (http://publicationethics.org/files/u2/New_Code.pdf). To verify originality, your article may be checked by the Cross Check originality detection service http://www.elsevier.com/editors/plagdetect.

The authors are obliged to participate in peer review process and be ready to provide corrections, clarifications, retractions and apologies when needed. All authors of a paper should have significantly contributed to the research.

The reviewers should provide objective judgments and should point out relevant published works which are not yet cited. Reviewed articles should be treated confidentially. The reviewers will be chosen in such a way that there is no conflict of interests with respect to the research, the authors and/or the research funders.

The editors have complete responsibility and authority to reject or accept a paper, and they will only accept a paper when reasonably certain. They will preserve anonymity of reviewers and promote publication of corrections, clarifications, retractions and apologies when needed. The acceptance of a paper automatically implies the copyright transfer to the National Academy of Sciences of the Republic of Kazakhstan.

The Editorial Board of the National Academy of Sciences of the Republic of Kazakhstan will monitor and safeguard publishing ethics.

Правила оформления статьи для публикации в журнале смотреть на сайте:

www:nauka-nanrk.kz

ISSN 2518-170X (Online), ISSN 2224-5278 (Print)

http://www.geolog-technical.kz/index.php/en/

Редакторы *Д. С. Аленов, М. С. Ахметова, Р. Ж. Мрзабаева*
Верстка *Д. А. Абдрахимовой*

Подписано в печать 15.04.2021.
Формат 70x881/8. Бумага офсетная. Печать – ризограф.
13,0 п.л. Тираж 300. Заказ 2.

Национальная академия наук РК
050010, Алматы, ул. Шевченко 28, т. 272-13-19, 272-13-18