The educational content and methods for big data courses including big data cluster analysis

Meruert Serik†, Gulmira Nurbekova†, Meiramgul Mukhambetova‡ & Zhandos Zulpykhar†

L.N. Gumilyev Eurazian National University, Nur-Sultan, Kazakhstan†
Kh. Dosmukhamedov Atyrau University, Atyrau, Kazakhstan‡

ABSTRACT: Big data technology is currently one of the most dynamically developing areas in information technology. In recent years, big data have been particularly visible in global economic and technological development. Keeping pace with the global trends in big data, Kazakhstan has come up with its strategic agenda until 2030. Along with the processing of unstructured data, the study of big data technology for the management of industrial and social processes is a priority. In order to train competitive specialists in big data, educational and methodological foundations for teaching large amounts of data in universities have to be developed. Therefore, the study presented here has been focused on developing the required educational and methodological support for the processing and analysis of big data with the goal of offering optimal solutions for the implementation of analysis through clustering of big data. In the scope of the article is described the algorithm of selection of number of clusters for the algorithm k-means and its course. The method of implementing the algorithm is universal, and the code can be changed according to one's needs.

INTRODUCTION

The main methods and technologies that are needed when processing large volumes of unstructured data are mastering algorithms for parallel processing and the clustering of data. Large systems, such as Hadoop, are based on parallel data processing and clustering algorithms. In this article, is presented an overview of training courses on large volumes of data in Kazakh and world universities, as well as the content of educational and methodological support for conducting a training course based on the acquired experience. Among the rich content of the educational and methodological support, is highlighted the implementation of the algorithm for selecting the number of clusters for the k-means algorithm in the R software environment.

The study presented here has been particularly focused on the development of effective educational and methodological support for determining the foundations of teaching large-scale data to university students and the practical implementation of that support in the educational process.

LITERATURE REVIEW

There are some educational practices regarding the implementation of big data in in Kazakhstan universities that need to be highlighted. The content of the educational programme *Big Data Analysis* offered by Astana IT University in Kazakhstan focuses on the formation of skills in the following areas: analysis of large amounts of information; data management of the organisation, industry; introduction to new technologies for data processing and analysis; and development of new models of the information infrastructure of the organisation taking into account the capabilities of big data technologies [1].

The purpose of the educational programme *7M06106-Big Data Systems (Big Data)* of *D. Serikbayev* East Kazakhstan State Technical University also in Kazakhstan is to train specialists able to assess the impact of big data technology on the activities of large enterprises and offer options for the effective use of these technologies in enterprise management [2].

The rapid development of distributed data technologies marked the beginning of introducing new content, technologies and methods to the training of specialists in the field of education [3]. These changes were necessary to form the adequate competencies of future specialists in information and communication technologies, and therefore a whole module containing new disciplines has been developed and introduced into the educational programme at the university [3].

When analysing the practice of introducing large amounts of data into the content of university education across different countries, the authors of this article have identified several problems and achievements. Researchers from the

Department of Didactics and School Organisation of the Valencian International University of Spain and the University of Granada, noted in their bibliometric review on big data in education the increasing popularity of processing large amounts of data for the analysis of a particular activity over the decade 2010-2020 [4]. This phenomenon, that is the ever-increasing amount of data is commonly referred to as big data. The authors of that review also observed that the analysis of large amounts of data had been mainly introduced and performed by students. In the review, the authors proposed to analyse scientific products from big platforms that provide access to multiple databases in education, such as the Web of Science (WOS), Scopus, ERIC and PsycINFO. The bibliometric review covered 1,491 scientific documents. Among the results, the increase in the number of publications since 2017, the increase in the journals and countries covered, and the creation of links to topics by authors were particularly highlighted [4].

With the help of intelligent systems, in addition to using big data, digital transformation has led to profound changes in higher education at the Prince Sultan University, College of Business Administration in Riyadh, Saudi Arabia [5]. The author of this paper pointed out that the use of big data and digital technologies had impacted on the basic institutional value of education that is to better meet the needs of students. As part of the technological impact, a machine learning algorithm was used in the analysis of learning results. The author had achieved positive results in the use of large amounts of data in teaching and learning, which was recognised by the university [5].

Along with the development of information and computer technology, cloud computing technologies are also rapidly expanding. Currently, the modern education reform is deepening as the original teaching materials have not been able to satisfy modern educational requirements. Especially in recent years, the importance of content and methods for teachers and students has been increasing, which is aligned with the increase in the number and types of educational resources. The era of *we*-media in China, often referred to as *self*-media, is another factor discussed in relation to big data and ideological and political education in colleges and universities in China and abroad [6].

Methods of using big data in remote locations is considered by Wen et al who stressed the importance of using big data in this type of work [7]. Also, researchers from Chile indicated - which concurs with other studies - that big data is only getting better, and conditions are being created for training highly competent specialists for the future [8].

COURSE DESIGN

In course design, the following points have to be considered:

- Increasing the motivation to study among all students, can be influenced by the inclusion of bid data in the study content. In educational programmes, the overwhelming majority of students, when deciding on study subjects for next academic year, choose a subject due to the big data inclusion.
- Increasing the knowledge and skills of all students, including undergraduates can be influenced by bid data. Since working with large amounts of data is associated with working in the university's local network and in a large-scale network, creating a database in a remote disk space - a cloud platform - allows students to increase knowledge and skills in organising big data in this database.

Teaching large amounts of data is related to mathematics and statistics, computer science, information and communication technologies, physics and process modelling. This, in turn, indicates that science, technology, engineering, mathematics (STEM) training is currently widely supported in the field of education [9-11].

In this regard, it is necessary to take into account the current state of economy in a given country; and in the specific context of this article - the Republic of Kazakhstan. In particular, in order to increase students' motivation to learn, teachers should actively use various approaches to organising the educational process, including those used in human resource (HR) management, noting that big data analytics tools, similar to HR management, help to increase students' interest and improve the quality of education [12].

STUDY RESULTS

All students, including undergraduates involved in this study participated in it voluntarily. Special courses introduced into the educational process were chosen by students themselves. The study was carried out in the educational programmes: *6B01511-Informatics*, *7M01511-Informatics* and *7M01525-STEM-Education* offered by the Faculty of Information Technologies at *L.N. Gumilyov* Eurasian National University in Nur-Sultan, Kazakhstan, in the amount of five credits (one lecture, two practical classes, two independent works).

The purpose of the educational and methodological support is to form students' professional competence in the development and use of systems for processing and analysing large amounts of data. This goal is related to the purpose of the educational programme, in particular, to the technology of developing specialised software systems responsible for processing big data. In the course of mastering the content of this educational and methodological support, students are prepared to perform the following professional tasks:

- Formulate the problem for data analysis.

- Pre-processing of data.
- Data visualisation.
- Development, implementation and application of methods for intelligent data analysis for a large range of data.
- Presentation of work results.

During the development of the educational and methodological package *Collection and processing of big data*, lectures on the following topics were considered: basics of big data, problems with big data, the content and tasks of big data management, technology for working with big data, big data analytics, the basics of data mining, the concept of Hadoop, MapReduce, big data processing on the Google Cloud Platform (GCP), and large data processing on a network controller. Also, practical work as the basics of the R programming language that is, making statistical forecasts, starting with its installation and configuration and work on its visualisation were considered [13-15].

As a result, students should be able to:

- understand the phenomenon of big data, the scientific and technical problems and opportunities associated with their emergence, and be familiar with trends in storage technologies;
- provide reasons for the emergence of the big data trend, and be familiar with the processes of big data analysis, the main methods of processing large data arrays and the basics of the R language;
- formulate algorithms in the MapReduce paradigm, choose the appropriate big data analytics tool and choose the appropriate big data storage technology.

The following paragraphs focus on the ways to solve some of the most complex tasks considered in the content of the educational and methodological support.

In the process of processing large amounts of data, it is difficult to analyse data using the *for-loop*. So, in this case, a clustering method is used to process large amounts of data. Clustering is the division of a given object model into indivisible subsets called clusters. There are two groups of clustering algorithms: k-means method and hierarchical clustering.

In this case, the k-means method was considered. The algorithm of the k-means method was included in a publication by Lloyd [16]. The basic structure of the algorithm is as follows:

- data and k (the number of clusters in which this data should be divided) are entered into the algorithm;
- random k-points (centroids) are selected and the closest distance from these points to the centroids is calculated - the points closest to some centroids form a cluster;
- in order for the distance from all points to the new centroid to be small, a new centroid is created based on the points included in the cluster;
- part of the points approaches the new centroid and enters its cluster, and part of the centroid leaves and begins to enter other clusters;
- all these steps are repeated until the position of the centroids changes.

The practical implementation of the work was carried out as an analysis of the k-means cluster using the *Iris* package of the R programming environment. The *Iris* flower data set has four descriptors (length and width of the leaves) consisting of 50 specimen of three Iris species (Iris setosa, Iris virginica and Iris versicolor). These indicators were used to create a linear discriminant model for classifying species.

Before considering the k-means cluster analysis, it is necessary to install the necessary packages.

```
// stats the package contains functions for statistical calculations and generating random numbers.
install.packages("stats")
// package for data processing
install.packages("dplyr")
// package for working with graphics
install.packages("ggplot2")
// the program has all the necessary equipment
install.packages("devtools")
// Data visualization tools for statistical analysis results
install.packages("ggfortify")
library(stats)
library(dplyr)
library(ggplot2)
library(devtools)
library(ggfortify)
```

After installing the necessary packages, the Iris view command was used to display the Iris data on the screen (Figure 1).

Filter

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3.0	1.4	0.1	setosa
14	4.3	3.0	1.1	0.1	setosa
15	5.8	4.0	1.2	0.2	setosa
16	5.7	4.4	1.5	0.4	setosa
17	5.4	3.9	1.3	0.4	setosa
18	5.1	3.5	1.4	0.3	setosa

Figure 1: Iris database.

As can be seen in Figure 1, the Iris database consists of five columns containing Iris data: sepal.length, sepal.width, petal.length, petal.width, species.

To create a work from the first four columns in the Iris database, one creates a mydata object:

```
mydata=select(iris,c(1,2,3,4))
```

To select the optimal number of clusters, one generates the wide-sense stationary (WSS) plot function. The WSS function is the sum of the distances between points and the corresponding centroids for each cluster:

```
wssplot<-function(data,nc=15, seed=1234)
{
  wss<-(nrow(data)-1)*sum(apply(data,2,var))
  for (i in 2:nc) {
    set.seed(seed)
    wss[i]<-sum(kmeans(data,centers = i)$withinss)}
  plot(1:nc,wss,type="b",xlab="numbers of clusters",
       ylab="within grops sum of squares")}
```

In this study, the authors implement the function in the R scripts block, depending on the desired object:

```
wssplot(mydata)
```

As a result, they obtained the optimal number of clusters (Figure 2).

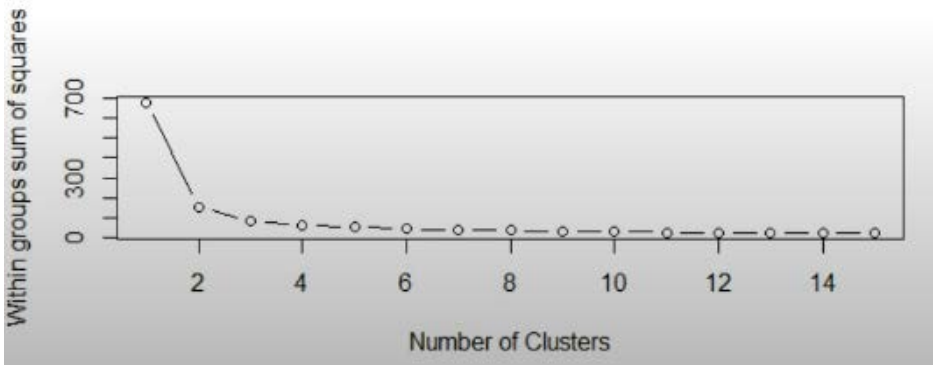

Figure 2: Graph of the optimal number of clusters.

To analyse the k-means cluster, the authors created a KM object.

KM=kmeans (mydata,2)

To evaluate the cluster analysis, one must construct a cluster graph (Figure 3). For this purpose:

Autoplot (KM,mydata,frame=TRUE)

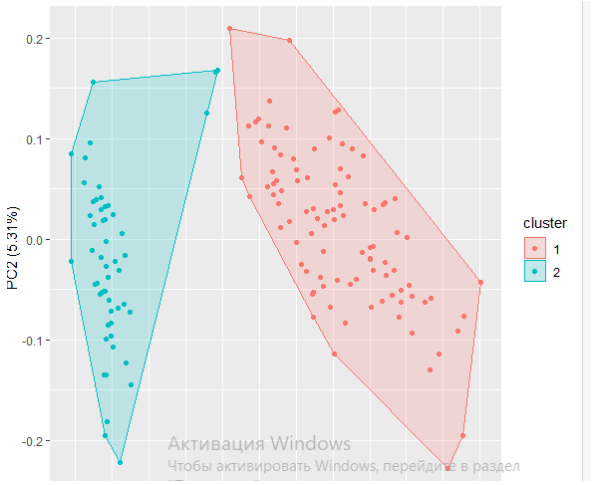

Figure 3: Cluster graph.

For a more accurate analysis of the k-means cluster, one must create cluster centers.

Cluster centers are the arithmetic mean of all points belonging to a cluster. To create the centers, the authors applied the cluster centers command to the KM object that they have created (Figure 4):

KM$centers

```
package 'ggfortify' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\User-NR\AppData\Local\Temp\RtmpcjTd05\downloaded_packages
> library(ggfortify)
> autoplot(KM,mydata,frame=TRUE)
> KM$centers
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1     6.301031    2.886598     4.958763    1.695876
2     5.005660    3.369811     1.560377    0.290566
>
```

Figure 4: Cluster centers values.

For example, the sepal for Class 1 - the average value for length is 6.301031. Considering the sepal for both classes according to the values in Figure 4 - sepal.length, sepal.width, petal.length, petal.width, one can see that the width has different values. One can also see that the cluster centers values of both classes do not overlap. Sometimes, with multiple groups, it is difficult to correctly mark the data using the *for-loop*. So, one has to learn the data from this cluster because this particular source of data is not marked. The selection can be done based on the cluster numbers.

In this study, the algorithm for selecting the number of clusters for the k-means algorithm was discussed. A quantitative study was conducted on real data reflecting the application of the proposed method. This solution allows to use the proposed approach in Web applications in the absence of access to similar data. This method can be applied to processes, such as assessing the future load on servers, creating a system for detecting fraud cases or predicting the company's profits. This method allows, first of all, to make changes and perform processes more efficiently and quickly. Therefore, taking into account the role of the k-means clustering method in predicting and analysing large amounts of data, there is a need to correctly communicate the meaning of the k-means method to university students through various teaching methods .

CONCLUSIONS

The content of the educational and methodological package referred to in this article includes a large number of works based on the basics of parallelisation, processing of big data on the basis of distributed data technology, such as the algorithm for selecting the number of clusters for the k-means algorithm. The authors note that the method of

implementing the algorithm for selecting the number of clusters for the k-means algorithm used in this study is an important topic in the course of large-scale data. They offer this direction as a universal method of data processing and also note the ability to change the code in question according to one's needs. Clearly outlined program codes and visualisations of theoretical information are supported by practical tasks with visual descriptions, which allow students to gain a deeper understanding of the topics and contribute to the formation of the necessary knowledge and skills.

REFERENCES

1. Educational Program of Astana IT University Big Data Analysis (2021), 24 April 2022, www.astanait.edu.kz/big-data-analysis/
2. Educational Program 7M06106 Big Data Systems (Big data) at D. Serikbayev EKSTU (2021), 12 May 2022, www.univision.kz/edu-program/645.html
3. Semenyuk, O., Kuc, S., Sadykova, S., Arynov, K., Beloussova, E., Niyazbekova, S. and Suleimenova, B., New educational programmes as a factor in forming students' innovative competencies. *World Trans. on Engng. and Technol. Educ.*, 17, **3**, 367-372 (2019).
4. Marín-Marín, J-A., López-Belmonte, J., Fernández-Campoy, J-M. and Romero-Rodríguez, J-M., Big data in education. A bibliometric review. *Social Sciences*. 8, **8**, 223 (2019).
5. Park, Y.E., Uncovering trend-based research insights on teaching and learning in big data. *J. of Big Data*, 7, **1**, 1-17 (2020).
6. Wang-Sheng and Jie-Feng, W., Research on the innovation of ideological and political education of university students in the we-media and big data era. *9th Inter. Conf. on Measuring Tech. and Mech. Auto. (ICMTMA)*, IEEE, Changsha, China, 403-407 (2017).
7. Wen, J., Zhang, W. and Shu, W.A., Cognitive learning model in distance education of higher education institutions based on chaos optimization in big data environment. *The J. of Supercomputing*, 75, **2**, 719-731 (2019).
8. Vidal-Silva, C.L., Madariaga, E.A., Rubio, J.M. and Urzúa, L.A. Study of the reality and viability of the education in big data in the Chilean Academy. *Información tecnológica*. 30, **5**, 239-248 (2019).
9. Pusca, D. and Northwood, D.O., Creativity and its constraints in engineering education. *World Trans. on Engng. and Technol. Educ*, 17, **2**, 146-151 (2019).
10. Gerigk, M. Improvements to the STEAM-based teaching of architectural drawing. *World Trans. on Engng. and Technol. Educ.*, 19, **2**, 163-168 (2021).
11. Serik, M., Akhmetova, B., Shyndaliyev, N. and Mukhambetova, M., Supervising and managing STEM projects for school students by the school-university model. W*orld Trans. on Engng. and Technol. Educ.*, 20, **2**, 95-100 (2022).
12. Vichugova, A., Big Data Analytics and Machine Learning in Education: 5 Cases from Universities (2020), 24 May 2022, www.bigdataschool.ru/blog/big-data-analytics-education-cases.html
13. Sakhipov, A., Yermaganbetova, M., Latypov, R. and Ualiyev, N., Application of blockchain technology in higher education institutions. *J. of Theor. and Applied Inform. Tech.*, 100, **4**, 1138-1147 (2022).
14. Yerlanova, G., Serik, M. and Kopyltsov, A., High performance computers: from parallel computing to quantum computers and biocomputers. *J. of Physics: Conf. Series*, 1889, **3**, 032032 (2021).
15. Serik, M., Nurbekova, G. and Mukhambetova, M., Optimal organisation of a big data training course: big data processing with BigQuery and setting up a Dataproc Hadoop framework. *World Trans. on Engng. and Technol. Educ*., 19, **4**, 417-422 (2021).
16. Lloyd, S., Least squares quantization in PCM. *IEEE Trans. on Infor. Theory*, 28, **2**, 129-137 (1982).