Improving of Personal Educational Content Using Big Data Approach for Mooc in Higher Education

Yunus Santur, Mehmet Karaköse, Erhan Akın
Fırat University, Computer Engineering Department, 23119
Elazig, Turkey
{ysantur, mkarakose, eakin}@firat.edu.tr

***Abstract*—Today, web-based education technologies such as e-learning, distance learning, online course, virtual classrooms and interactive learning are commonly used outside the traditional education systems. Massive Open Online Courses, which were adopted with the evolution of these systems in 2008, have become the most popular education systems of today and access to a very large audiences with many modules such as video courses, documents, interactive learning activities and quizzes in themselves. In this study, a machine learning and big data based approach has been presented for the mentioned online education systems. With the proposed approach, it is aimed to develop course contents in online education, offer student-specific learning activities, perform an analysis according to criteria such as age, gender, occupation, education level and location, and to obtain decisions with strategic importance such as determining the course prerequisites to be developed by big data-based analysis.**

Keywords—big data; massive open online course; mooc; higher education;machine learning; deep learning

I. INTRODUCTION

In today's world, the speed of access to data and the amount of data produced along with the developing and expanding the use of technology are rapidly increasing. The proper interpretation of data such as social network interactions, blogs, photos, search engine entries and access logs created by user with mobile devices and computers along with big data analysis allows for taking strategic decisions in the industry [1]. Sectors such as health, e-government, food and finance can take user-specific, location-based, product or market-based and effective decisions with strategic importance by interpreting these data. The use of big data in education is relatively a new field compared to other sectors.

In short, Massive Open Online Course (Mooc) that can be defined as web-based online courses are different from the classic e-learning systems by the following characteristics [2-5].

1) Massive: Unlike traditional online education systems, it aims large audiences worldwide.

2) Open: There is no prerequisite or formal procedure for participation in courses. They are open to the world-wide because they are mostly free of charge or only certain modules are paid such as certification.

3) Online: All educational contents and educational activities such as assignments and exams are online and accessible via the Internet as web-based.

4) Course: The course content includes the following criteria compared to e-learning systems.

- Instructional: With the purpose of achieving the highest instructional level, presentations, documents and additional resources are given in addition to 8-10-minute video contents appropriate to the pedagogical format which are generally divided into modules.
- Certification: They provide certified job-ready courses along with technology companies.
- Learning Activities: Especially the courses in the field of informatics include offline or online interactive encoding activities to increase the instructional level. In addition to this, activities such as assignment, quizzes, final exam and final project are also conducted depending on a particular calendar.
- Learning network: They provide the formation of e-learning networks in which activities such as problem solving, mentor support and discussion are conducted for each course for the development of courses and for trainees to conduct learning activities together.

In 2012, coursera that emerged as one of the Moocs reached 1.7 million trainees in a semester and surpassed facebook which is the world's biggest social network when the rate of growth is taken into account in this range. Today, popular web-based systems such as coursera, udemy, edX give service to millions of trainees with thousands of courses within themselves [6]. The video contents and documents created by the instructor for each course, periodical learning activities such as course tracking, quizzes, surveys, assignments, final exam and project of millions of trainees following the courses, explanation of data such as discussions, social network analysis and personal information by interpreting with big data are extremely important for these applications [7]. Performing the big data analysis of these data is of great importance in terms of course productivity, changing the course contents, the development of the learning activities of the trainees and the development of new course contents.

978-1-5090-0778-3/16/$31.00 ©2016 IEEE

For this purpose, a big-data based structure that can be adapted to the said online education systems has been presented in this study. Along with this structure presented, a software framework has been defined to perform the intended analyses using machine learning.

A. Mooc and Big Data

The concept of Big Data is expressed by five main components which are briefly known as 5V in the literature, these are Volume, Variety, Velocity, Value and Verification. The subject of big data can be associated as in Table I in accordance with the Mooc and 5V components [8, 9].

TABLE I. BIG DATA AND MOOC RELATION

Component	Big Data Relation	MOOC Relation
Volume	Represents the data size.	The daily Terabyte amounts are accessed along with the video contents of each course, the number of trainees and daily accesses.
Variety	Represents the diversity of data.	It has a diversity of data such as training documents like online course video, text and presentation, the discussion and collaborative learning activities of trainees on the learning network, exam/assignment evaluations, performing interactive coding activities, in information courses.
Velocity	Represents the speed of access to data.	Hundreds of thousands of trainees have a periodical real-time access to some individual course contents.
Verification	Represents the data validity	This information refers to big data's development of personalized training as in banking and health sectors.
Value	Represents the value to be created by this information as a result of data analysis.	The value to be created for online courses represents the trainees' quick and effective learning, finding a job or improving the existing career via certification program, the development of course contents using student data.

B. Mooc and Machine Learning

Machine learning can be defined as performing analysis and producing result on large and/or complex data which cannot be analyzed by human intervention using intelligent and learning algorithms. Today, machine learning is commonly used in many areas such as finding fault in face detection, predicting economic data, spam detection, speech recognition and industrial applications. Machine learning basically investigates the problems in three classes including regression, classification and clustering [10, 11].

1) Regression: It represents making prediction on new data based on the historical data learned. Estimating the future data in economy and making prediction based on the person's data in health can be given as examples to the practical application areas.

2) Classification: It is the process of finding the class to which data belongs or the nearest class. Face and voice recognition can be given as examples to the practical application area.

3) Clustering: It is the process of clustering data by taking care of the relationships between them. Its main difference from classification is that class label and features are clear by performing classification process. In clustering process, classification of data can be performed without knowing these features. Grouping news content in semantic processing can be given as an example to the practical application area.

In addition to these, it is necessary to know the concepts of deep learning and dimension reduction contributing these three methods.

4) Deep Learning (DP): Classical machine learning algorithms may not be sufficient for the inference process by making process on some data. Image classification, speech recognition and making inference from labeled video images are the examples for this area. The main differences are the use of multi-layer structures for the accuracy performance and the use of graphic processors instead of processors for speed although the general approach does not change.

5) Dimension Reduction (DR): Especially in DP applications, making processing over high-dimensional data such as image, audio and video constitutes disadvantages in terms of both run-time and processing load and memory usage. Therefore, DR algorithms such as Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) enable to obtain lower dimensional eigenvalues by performing pre-processing on large data.

Which method can be used for the relationship of machine learning algorithms with mooc is given in Table II by examples [12-14].

TABLE II. MACHINE LEARNING AND MOOC

Machine Learning	Mooc Relation
Regression	Making the success estimation from the details of the trainees and survey, exam, project activities and access data.
Classification	Classification of personalized learning activities and trainees trends for new courses
Clustering	Analysis of access to the course content, grouping of results
Deep learning	Making sense of the unlabeled videos by analyzing, generating label
Dimension Reduction	Performing dimension reduction on materials such as video and sound for deep learning

C. Hadoop

The interpretation of the data of big data has led to the emergence of new technologies because it is not possible by traditional methods.

Open source which was developed for this purpose and supported by the companies such as Hadoop Microsoft, Facebook, Twitter, Linkedin, Ibm is a framework and includes several technologies in itself [7]. The descriptions related to Hadoop main modules and Hadoop-based popular big data projects are given in Table III [15].

TABLE III. HADOOP TECHNOLOGIES

Hadoop Technology	Description
Hadoop Common	Common modules supporting Hadoop modules
Hadoop Distributed File System	Hadoop-based distributed file system like Google big table. It allows distributed data on more than one cluster to appear as a whole single data.
Hadoop Mapreduce	High-performance parallel programming for the Distributed structure
Data Base	High-performance noSql databases to use big data that can store non-relational data as well as relational data: such as Hbase, Cassandra, Mongodb
Spark	Hadoop-based high-performance machine learning framework: The framework that enables to perform classification, clustering, association rule learning procedures very quickly on big data. It is 100 times faster than the classic Hadoop.

II. PROPOSED METHOD

A web-based service that provides the development of course contents with the use of machine learning and big data in high education, the development of personal educational systems, the analysis of non-relational data such as access logs to students' all materials such as age, education, gender, location, course to determine new course contents, social media analysis with Apache Hadoop Spark has been proposed. The proposed system is shown in Fig 1.

Firstly, no-sql based Mongodb and Spark installation was performed for application. A virtual course content, course participants and content were generated using random data. The methods shown in Table IV on raw data were coded in Matlab environment, and accuracy performances were tested using Roc analysis. (1) shows the application accuracy percentage according to confusion matrix method.

Performances were compared by running the algorithm, which was run on DP, both on entral Process Unit (CPU) and Nvidia Graphical Process Unit (GPU). Comparison of CPU and GPU is presented in Fig.2. Although traditional CPUs can have 2, 4, 8 core, GPUs may have hundreds of cores, thus performance increase up to 30 times can be achieved. This situation is the most important reason of the fact that GPU programming predominates the parallel programming over CPU programming.

The multi-layer artificial neural network used for DP is presented in Fig.3. y output value of each class for DP was formulated according to (2). The output activation function according to (3) was selected as sigmoid.

Which machine learning algorithms are used with the proposed method in accordance with Table II is presented in Table IV [10].

TABLE IV. MACHINE LEARNING ALOGRITHMS

Model	Algorithm
Regression	Lineer regression
Classification	Random Forest
Clustering	K-Means
Deep learning	Multi-layer artificial neural network
Dimension Reduction	SVD

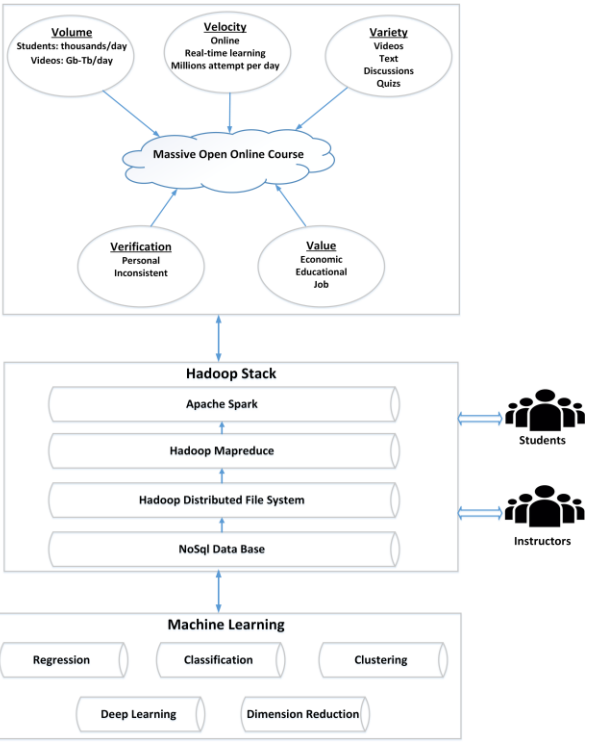

Fig. 1. Overall block diagram of the proposed method

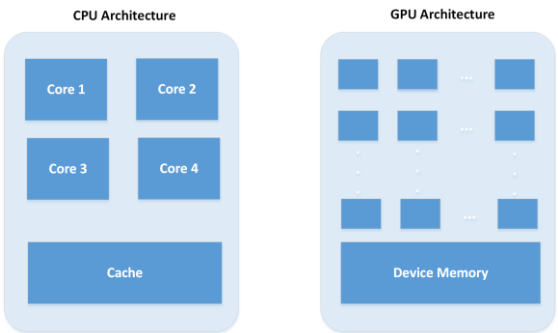

Fig. 2. Cpu vs Gpu

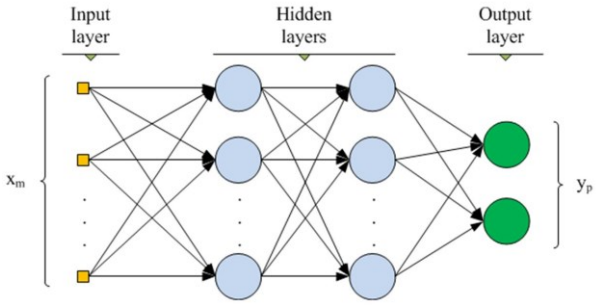

Fig. 3. Multi layer artificial neural network for DP

$$Accuary = \frac{TP+TN}{N} \quad (1)$$

$$y_i = \mathrm{g} * \left(\sum_j x_j * w_k + b_i\right) * \mathrm{d} \quad (2)$$

$$y_i(\mathrm{sig}) = \frac{1}{1+\mathrm{e}^{-\mathrm{Bx}}} \quad (3)$$

III. EXPERIMENTAL RESULTS

In the study, all data types that make up the application entry were converted from raw data format into numerical values more suitably for the neural network input. In the scenario created for the application, classification was performed to determine the users' satisfaction in order to prepare new course contents. The obtained results of this scenario are presented on confusion matrix in Table V. After 75% of the generated data sets were used in the education algorithm, the remaining 25% was used for testing purposes.

In the generated scenario, 80% accuracy performance was achieved as it can be seen from confusion matrix.

In the second application scenario, classification was performed using DP on images, CPU and GPU performances realized in Fig.4. Feature extraction was performed by SVM before performing the classification process on images. In the application, five eigenvalues obtained by SVM were used for each frame to perform classification on images.

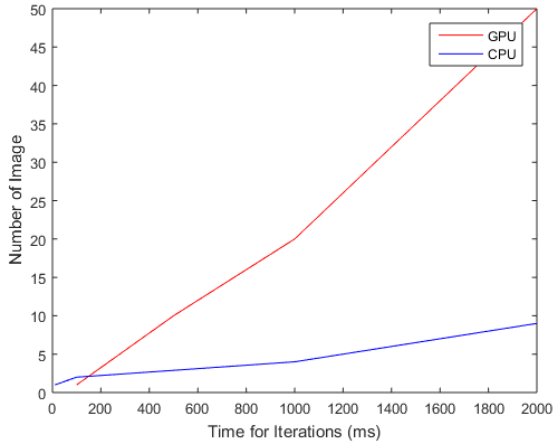

Fig. 4 Deep Learning with Cpu and Gpu

TABLE V. CONFUSION MATRIX

		Actual	
Predicted		P	N
	P	210	55
	N	45	190

IV. CONCLUSIONS

Moocs, one of today's most popular educational systems, comprise course materials and related learning activities such as video / text, quizzes and projects and similar data in proportion to millions of participants they reached since their introduction in 2012. The analysis of these data by traditional methods is almost impossible, sample scenarios were tested by presenting a big data and machine learning-based structure for the systems mentioned in this study.

REFERENCES

[1] E. Akın, M. Karaköse, "Bilgisayar Mühendisliği Eğitiminde Sanal Laboratuarların Kullanımı", Elektrik Elektronik Bilgisayar Mühendislikleri Eğitimi 1.Ulusal Sempozyumu, 166-169, Ankara, 2003.

[2] Y. Pang, T. Wang, N. Wang, "MOOC Data from Providers", In Enterprise Systems Conference (ES), pp.87-90, IEEE, 2014.

[3] Y. Pang, N. Wang, Y. Zhang, Y. Jin, "MOOC related data on navigators", In Computer Science & Education (ICCSE), 2015 10th International Conference on (pp. 558-561). IEEE, 2015.

[4] T. Daradoumis, R. Bassi, F. Xhafa, S. Caballé, "A review on massive e-learning (MOOC) design, delivery and assessment", In P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), 8.International Conference on (pp. 208-213), 2013.

[5] K. Park, M.C. Nguyen, H. Won, "Web-based collaborative big data analytics on big data as a service platform", In Advanced Communication Technology (ICACT), pp. 564-567, 2015.

[6] Y. Demchenko, E. Gruengard, S. Klous, "Instructional model for building effective Big Data curricula for online and campus education", In Cloud Computing Technology and Science (CloudCom), 2014 IEEE 6th International Conference on (pp. 935-941), 2014.

[7] T. Daradoumis, R. Bassi, F. Xhafa, S. Caballé, "A review on massive e-learning (MOOC) design, delivery and assessment", In P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), 2013 Eighth International Conference on (pp. 208-213), 2013.

[8] Y. Demchenko, E. Gruengard, S. Klous, "Instructional model for building effective Big Data curricula for online and campus education", In Cloud Computing Technology and Science (CloudCom), 2014 IEEE 6th International Conference on (pp. 935-941), 2014.

[9] G. Huang, G. Huang, S. Song, K. You, "Trends in extreme learning machines: a review", Neural Networks, 61, 32-48, 2015.

[10] Y. Santur, M. Karaköse, İ. Aydın, E. Akın, "IMU based adaptive blur removal approach using image processing for railway inspection", In 2016 International Conference on Systems, Signals and Image Processing (IWSSIP), pp.1-4, IEEE, 2016.

[11] P. Dillenbourg, "A Tutorial on Machine Learning in Educational Science", State-of-the-Art&Future Direc. of Smart Learning, 453, 2015.

[12] S. Cooper, M. Sahami, "Reflections on Stanford's MOOCs", Communications of the ACM, 56(2), 28-30, 2013.

[13] C. Romero, S. Ventura, "Data mining in education", Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3(1), 12-27, 2013.

[14] T. White, "Hadoop: The definitive guide", " O'Reilly Media

[15] X. Deng, Q. Liu, Y. Deng, S. Mahadevan, "An improved method to construct basic probability assignment based on the confusion matrix for classification problem", Information Sciences, 340, 250-261, 20116.