UDC 004.852

EDUCATIONAL DATA MINING

OLEG RACHITSKY, ARKADY OSKIN
Polotsk State University, Belarus

This article describes the effectiveness of using machine learning tools for educational data mining.

Making decisions in the field of education is a complex, multifaceted process that usually involves a lot of people. A big and effective part in this process plays the analysis of information coming from participants of the educational process at its various stages. Huge quantities of educational data about various aspects of the educational process are accumulated: about students and their academic performance, teachers and their scientific and educational work, distance learning courses, educational forums, testing and questioning systems for students, and much more. So, a lot of data has been acquired in the recent years and is continuing to accumulate.

The interest in new methods and approaches in the automated detection of new, sometimes hidden, relationships in data and their interpretation has arisen in connection with the growing use of information technologies in education. There are many fields where the methods of statistics, machine learning and knowledge extraction are useful for all participants of the educational process: students, teachers, developers of training courses, methodologists, administrative personnel. However, the theoretical basis for the application of these methods in practical activities is not sufficiently developed at the moment.

Educational data mining (EDM) is a set of methods for detecting previously unknown but useful information about the educational process and its participants in order to support decision making. The aim of EDM is to process and analyze data obtained within the framework of the educational process for finding hidden patterns. Methods, algorithms, tools for intellectualizing the solution of applied problems in education are being developed within the framework of EDM. Sources of educational data are information systems such as computer educational programs, university information systems, social networks, logs, test results, work programs of disciplines, etc. For a large audience of courses, the application of algorithms of EDM becomes particularly important. In connection with the active growth of information flows and the intensive accumulation of information, the task of processing and analyzing information and making rational decisions becomes important. Applying modern machine learning technologies is the solution for this task.

Machine learning is a field of computer science, the main point of which is to grant computers the ability to learn without being explicitly programmed for any particular task. This was coined by Arthur Samuel, an American pioneer in the field of computer gaming and artificial intelligence, back in 1959. And since then it has evolved from the study of pattern recognition and computational learning theory into artificial intelligence research. Machine learning is closely related to computational statistics, which also focuses on prediction-making through the use of computers and explores algorithms that can learn from and make predictions on data, of which we have a lot in the case of educational data. Machine learning algorithms overcome strictly static program instructions by making data-driven predictions or decisions, through building a model from sample inputs. All this makes machine learning an absolute leader for solving tasks of EDM.

Machine learning perfectly conflate with data mining, where the latter subfield focuses more on exploratory data analysis and is known as unsupervised learning. Within the field of educational data analytics, machine learning can be used as a method to devise complex models and algorithms that lend themselves to predictions of very complex things like a student's performance. These analytical models could allow us to produce reliable, repeatable decisions and results and uncover hidden insights in all entropy of educational data through learning from historical relationships and trends in the data from the past years.

Machine learning tasks are typically classified into two broad categories, depending on whether there is a learning signal or feedback available to a learning system:

3.Supervised learning: The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs. In special cases, the input signal can be only partially available, or restricted to special feedback.

4.Semi-supervised learning: the computer is given only an incomplete training signal: a training set with some (often many) of the target outputs missing.

5.Active learning: the computer can only obtain training labels for a limited set of instances (based on a budget), and also has to optimize its choice of objects to acquire labels for. When used interactively, these can be presented to the user for labeling.

6.Reinforcement learning: training data (in the form of rewards and punishments) is given only as feedback to the program's actions in a dynamic environment, such as driving a vehicle or playing a game against an opponent.

7.Unsupervised learning: No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in the data) or a means towards an end (feature learning).

If we look at this from the perspective of EDM, we can say that it's possible to create a model and train it using supervised learning where we feed data from the past years and the results we got from this data (student papers and some final mark of their research). Also, if we deal with flexible input data and we cannot even find a direct result for a given set, we can use unsupervised learning, allowing the model to determine some statement for the data given that can be used further.

As a scientific endeavor, machine learning grew out of the quest for artificial intelligence. Already in the early days of AI as an academic discipline, some researchers were interested in having machines learn from data. They attempted to approach the problem with various symbolic methods, as well as what was then termed "neural networks"; these were mostly perceptrons and other models that were later found to be reinventions of the generalized linear models of statistics.

Artificial neural networks are computing systems inspired by biological brains and their natural neural networks. Such systems progressively improve performances on tasks by considering examples without task-specific programming as it is described in machine learning definition. An artificial neural network is based on a collection of connected units or nodes called artificial neurons. Each connection between artificial neurons can transmit a signal from one to another. The artificial neuron that receives the signal can process it and then signal artificial neurons connected to it. In the real implementation of artificial neural network, the signal at a connection between artificial neurons is a real number, and the output of each artificial neuron is calculated by a non-linear function of the sum of its inputs. Artificial neurons and connections typically adjust weights as learning proceeds. The weight increases or decreases the strength of the signal in the connection. Typically, artificial neurons are organized in layers. Different layers may perform different kinds of transformations on their inputs. Signals travel from the first layer, that is considered to be an input layer, to the last layer, which represents the desirable format of output, possibly after traversing the layers multiple times. The original goal of the artificial neural network approach is to solve problems in the same way that a human brain would, like finding connections in data.

Machine learning and data mining often employ the same methods and overlap significantly, but machine learning focuses on prediction, based on the known properties learned from the training data.

Data mining focuses on the discovery of unknown properties in the data. Educational data mining can use many machine learning methods, but with different goals.

REFERENCES

1. Samuel, A. Some Studies in Machine Learning Using the Game of Checkers / Samuel, Arthur // IBM Journal of Research and Development. – 1959.
2. Mitchell, T. Machine Learning / T. Mitchell // McGraw Hill. – 1997. – P. 2.
3. McCulloch, W. A Logical Calculus of Ideas Immanent in Nervous Activity : bulletin of Mathematical Biophysics / McCulloch, Warren; Walter Pitts. – 1943. – 5 (4). – P. 115–133.
4. Kleene, S.C. Representation of Events in Nerve Nets and Finite Automata / S.C. Kleene // Annals of Mathematics Studies. – Princeton University Press, 1956. – P. 3–41.
5. Kohavi, R. Glossary of terms / Kohavi Ron; Provost Foster // Machine Learning. – 1998. – 30. – P. 271–274.
6. Mannila, Heikki. Data mining: machine learning, statistics, and databases : Int'l Conf. Scientific and Statistical Database Management / Mannila, Heikki / IEEE Computer Society. – 1996.