INFORINO 2020
Moscow, Russia
14-17 April, 2020

Educational Data Mining: Current Problems and Solutions

Speacker's
Evgenia
Muntyan
Author

Authors:

(Sergey Kovalev, Rostov branch JSC NIIAS, Rostov-on-Don, Russia)

(Anna Kolodenkova, Samara State Technical University, Samara, Russia)

(Evgenia Muntyan, Southern Federal University, Taganrog, Russia)

Educational Data Mining: Current Problems and Solutions

PROBLEM FORMULATION:

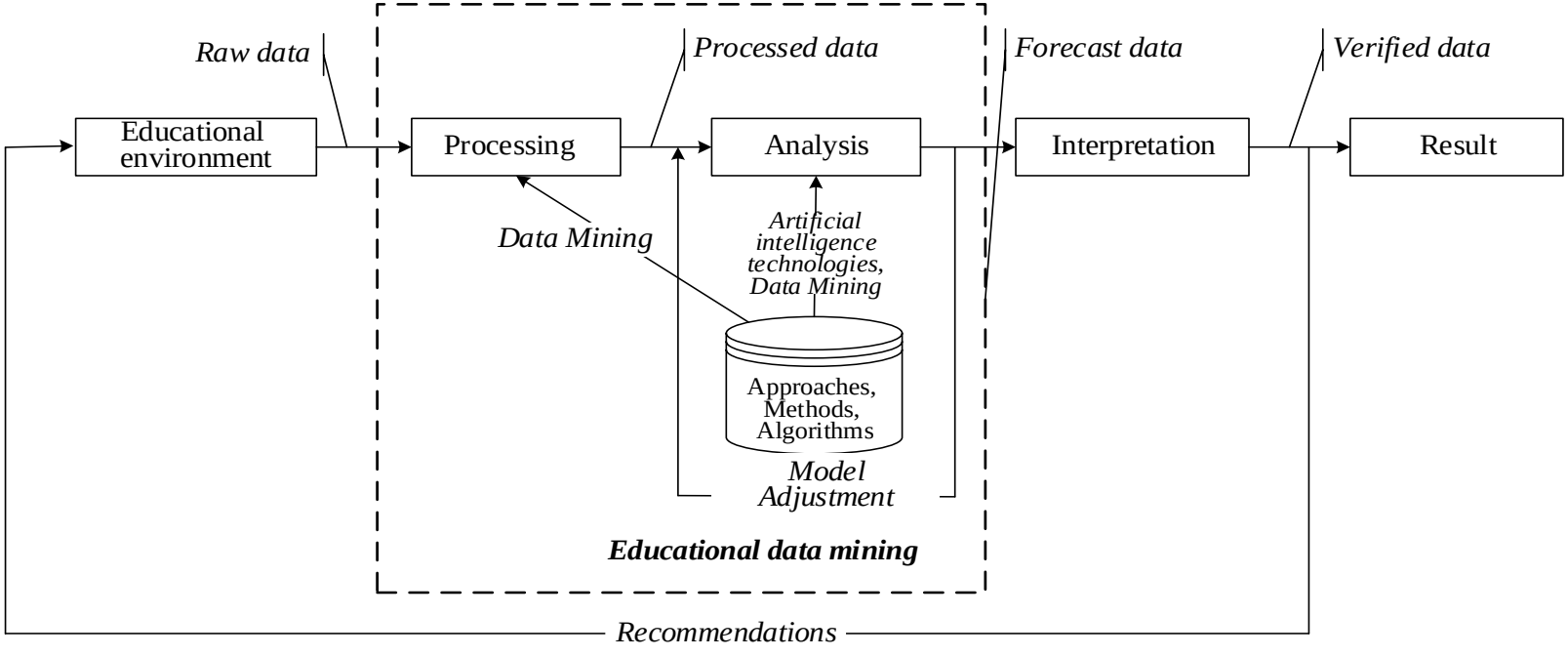
Many Russian and foreign authors focused on the study of the educational data mining. However, despite a significant number of works, this problem still remains open due to a number of tasks not solved yet. Firstly, most of the developed methods are focused on working with statistical data without taking into account the human factor associated with the inaccuracy and unclear presentation of linguistic information and data.

Secondly, the existing models are of recommendatory nature from a pedagogical point of view and do not provide automating the individual learning process. Thirdly, the software systems developed are focused on specific educational institutions of the Russian Federation, not taking into account the features of each university individually.

The authors of the paper propose a number of methods aimed at solving the above problems in the development of EDM area.

Educational Data Mining: Current Problems and Solutions

Generalized scheme of the educational data mining process:



The electronic information and educational environment AIS (Automated Information System) of Samara State Technical University (SamSTU) was used as a source of educational data mining.

Educational Data Mining: Current Problems and Solutions

Main stages of the educational data mining process:

- the Processing stage (preprocessing) is needed to obtain more useful, complete and accurate data and knowledge;
- the Analysis stage follows the careful data processing using DM-KD and artificial intelligence technologies in order to develop scientifically sound solutions under heterogeneous data to improve the quality of the educational process;
- The Interpretation block provided to transform the knowledge extracted from data into the following possible representations by: the production model; semantic networks; the frame model; the ontological model; the propositional calculus model.

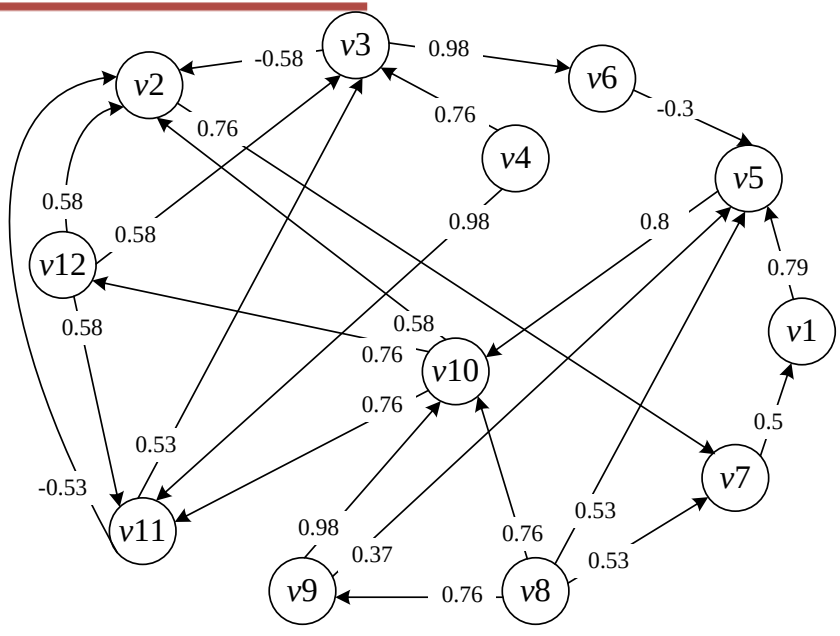
Educational Data Mining: Current Problems and Solutions

1. DATA DM-KD IN THE EDUCATIONAL PROCESS:

There is a wide range of DM-KD tools for the analysis and processing of educational data. But the authors highlight the methods based on graph models and fuzzy-logical modeling. In particular, ,they are used to evaluate the qualifications of university graduates. There are 48 parameters that affect the formation of graduate qualifications, then, they are reduced to 10-15 without loss of information in order to increase the visibility of the model.

- There is a hypergraph, that takes into account 48 parameters, that influence the formation of the graduate's qualification. A fuzzy hypergraph is a model formally defined in the form H' = (G'v, G'e, H'e), where G'v – set of vertices of the hypergraph with weights, G'e – set of links of the graph with weights, H'e – set of hyperedges.
- Made convolution of hypergraph. The use of the convolution algorithm made it possible not only to reduce the number of vertices, but also to make the transition from hypergraph to graph.
- The assessment of factors of professional development of the graduate was performed using the fuzzy cognitive model. The fuzzy cognitive model is understood as the fuzzy cognitive map in which vertexes present factors, and transitions between vertexes – fuzzy relationships of cause and effect between *Gfuzzy*= <*V*, *W*>, where *V* = {*vi*} – a set of vertexes, $vi \in V$, $j = \overline{1,h}$, *h* – quantity of vertexes; *W* – fuzzy relationships of cause and effect between vertexes (*wij* elements characterize the direction and force of influence between vertexes of *vi* and *vj* ($wij \in W$).

Educational Data Mining: Current Problems and Solutions



The information obtained allowed the expert (management of the educational institution) to improve the quality of students' training process, with the possibility of their further guaranteed employment.

Here the vertices v1-v9 correspond to factors influencing the graduate qualification: v1 – organizational; v2 – personal; v3 – social; v4 – economic; v5 – qualitative; v6 – communicative; v7 – informational; v8 – material and technical; v9 – professional and pedagogical.

The vertices v10-v12 correspond to the state level of the graduate: v10 – graduate qualification; v11 – graduate employment; v12 – the graduate success in the profession.

The following results of the analysis of the FCM structure were obtained:

1) A number of factors have a very strong influence on the graduate qualification including professional and pedagogical, qualitative, material, technical and organizational factors.

2) Vertices v11 and v6 are most affected by the FCM. This suggests that the influence of the system on these vertices can extinguish any negative impact from the outside.

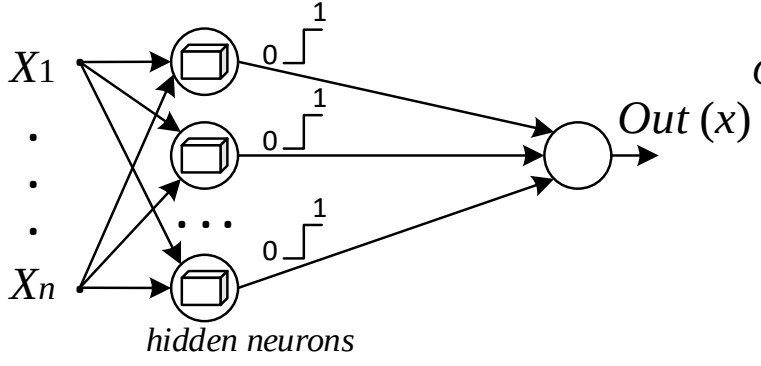
3) The consonance of most influences of the vertices is high, since it is close to 1 and allows one to conclude that the final effects of the vertices on each other are reliable.

Educational Data Mining: Current Problems and Solutions

2. AUTOMATIC KNOWLEDGE ACQUISITION FROM EDUCATIONAL DATA:

In this paper the approach to the automatic knowledge acquisition from educational data is based on the use of a special type of neural network models conditionally called hyperneural networks (HNN).

The HNN structure:



$$Out(x) = f\left(\sum_{j=1}^{J} Out_j(x) - \eta\right), \quad Out_j(x) = f(net_j(x)),$$

$$net_j(x) = \sum_{i=1}^{n} f((M_{ji} - x_i)(x_i - m_{ji})) - n$$

$$f(x) = \begin{cases} 1, & if \quad x \geq 0, \\ 0, & if \quad x < 0, \end{cases}$$

where M_{ji} and $m_{ji} \in R$ – configurable synaptic connections of the j-th hidden neuron, $x = (x_1, \ldots, x_n)^T$ – input image, η – positive constant ($\eta < 1$), $Out_j(x)$ – output function of the j-th hidden neuron, and $Out_j(x)$: $R^n \rightarrow \{0, 1\}$ – output signal of a two-level HNN with J hidden neurons.

HNN learning is carried out by adding hidden neurons as needed. The initial image of the target class is selected in the form of a minimal area in the attribute space, covering an arbitrary example of this class. Next, the selected area is expanded to include a new example from this class, and then it is narrowed to exclude examples of other classes.

Educational Data Mining: Current Problems and Solutions

3. KNOWLEDGE INTERPRETABILITY ASSESSMENT:

At the structural level reflecting the readability of the knowledge model based on production rules, the following main criteria of interpretability can be distinguished: the number of terms and variables in the description of the rules, the total number of rules in the model, the distinguishability of the rules, the completeness of the coverage of the characteristic scales by the rules, the uniform distribution of the membership functions of the terms on the characteristic scale. Consider the formalization of these criteria in relation to fuzzy systems, most often acting as knowledge models:

- The completeness criterion reflects the property of each element of the characteristic scale to be taken into account by at least one rule, that is, to belong to at least one fuzzy set from this scale: $\forall x \in U$: $\mu_A(x) > 0$. More strictly, this criterion is formalized as a definition of α - completeness:

$$CD_\alpha = \frac{\int_{\{x \in U : \max \mu_A(x) > \alpha\}} dx}{\int_U dx}.$$

- Uniform granulation is expressed in the similarity of powers of all fuzzy sets of the scale under consideration, that is, $\forall A, B$: $|A| \approx |B|$. When this criterion is implemented within the framework of HNN learning, it is convenient to present it in the form of minimizing the parametric penalty function:

$$Gr(p) = \frac{1}{1 + \exp(-(p - p_{\min})/\sigma_p)} - \frac{1}{1 + \exp((p - p_{\max})/\sigma_p)}.$$

Educational Data Mining: Current Problems and Solutions

The developed HNN meets the following principles:

- the rationality principle of thinking;
- the hierarchy principle;
- the non-linearity principle.

Additional terms thus formed and included in the descriptions of generalized classification rules as exclusive signs allow the compact presentation of rather complex data areas belonging to different partition classes in the attribute space using the minimum number of generalized and detailed rules.

4. BY AUTHORS THE DEVELOPED SOFTWARE PACKAGES FOR DATA MINING:

- to build a graph, a software package in C++ using the Qt library has been developed;
- the software for building FCM and analysis of its structure using the system indicators calculation in Java.

Thus, the software developed allows reducing the time spent on the educational data mining by 45%, as well as reducing the number of participants involved in the analysis by 2 times.

Educational Data Mining: Current Problems and Solutions

CONCLUSION:

The proposed methods and models allow increasing the effectiveness of decision-making process at all stages of the educational process management. The novelty of the work results is a new approach to the extraction of interpretable knowledge based on the use of a special class of neural network models, as well as the hybridization of several approaches developed as part of data mining into a single learning environment in order to implement an end-to-end learning strategy, including the steps of data collecting and analyzing, knowledge extracting and management.

INFORINO 2020

Moscow, Russia
14-17 April, 2020

Thank you for attention!

Speacker's contacts:

Name Surname: Evgenia Muntyan

Affiliation: author

e-mail: ermuntyan@sfedu.ru

web-site

IEEE

МЭИ NATIONAL RESEARCH UNIVERSITY

CROC