Educational Data Mining for Future Educational Employees*

Xenia Piotrowska[1] krp62@mail.ru

Ekaterina Terbusheva[1] terbushevae@gmail.com

[1] Herzen State Pedagogical University of Russia
Saint Petersburg, Russian Federation

Abstract

The didactical engineering of educational data mining teaching for students of pedagogical universities is described in details. We propose the authors' methodical system which based on analysis of requirements and expectations to research competence level, data analysis skills and modern education, comparison and analysis of the content of educational programs, books and courses on data mining and related disciplines, as well as generalization of pedagogical experience. Didactical content, learning activities, forms and learning tools are under discussion. Using of proposed didactical model allows to increase the level of research competence and to develop data analysis competence of students with an average knowledge in Applied Math and IT disciplines. The effectiveness of suggested strategy of data mining teaching is demonstrated.

Keywords: *data mining, didactical engineering, educational data mining, research competence, flipped learning, concentric model of content, iterative model of learning.*

1 Introduction

New digital age requires an adequate revision of pedagogical approaches. Well-known traditional learning theories - behaviorism, constructionism and cognitivism are complemented by new approaches, such as connectivity. Its includes learning in the process of communication and communication in a distributed network. Future strategies for the digital age teachers should be developed in accordance with the following principles: personalized learning, broadening experience, deepening knowledge and learning in a global context. One of the innovations in development of digital teacher's competence is the processing of data in the professional sphere, which requires new effective methodological approaches enriched with engineering didactics and e-learning tools.

The application of engineering approaches to didactics was called didactic engineering. It was proposed (since 1991) as a tool for research and development in order to study

the educational process and to build effective learning [Artigue et al., 1991]. According to this approach, it is necessary to acquire a new type of didactics, which binds in itself the content, the didactic itself and the technics of didactic modeling (Didactics Engeneering). But earlier in Soviet Union in 1975 math teacher A. Pyshkalo suggested simple, but very effective model for constructing didactic (methodical) systems. Since this time we can talk about didactical engineering in Russia (see. fig. 1, [Pyshkalo, 1975]). Lending our Russian didactic tradition and transforming it with the help of new techniques of digital educational age, we'll aim to construct modern didactics of educational data mining (EDM) teaching for students of pedagogical universities[1] with the purpose to develop research competences of future teachers. According to A. Pyshkalo learning objectives, educational content, methods, forms and learning tools are must be constructed in closely connection. We are realizing this didactic scheme with the help of new technologies of the digital education era, which are conceptually different from the classical didactics. The most important distinction is that ICT and engineering tools play critical role in digital didactics. Traditional and new digital didactics can be characterized by united goals and theoretical basis, but ideologically different teaching characteristics: delivery format, teacher's and student's roles, learning and teaching spaces, dominating mode of learning, representation and format of instructional materials, visualization usage, dominating mode of communication and assessment format, primary communication means and information access. Digital didactics could be defined as ICT-integrated didactics with a focus on engineering of learning (see. Table 1, [Tchoshanov, 2013]).

Further in Section 2, we'll propose the implementation of a methodological system for educational data mining teaching, based on the Pyshkalo's scheme. In Section 3, we'll discuss the first results of the influence of the methodology on improving the skills of students in the field of research and data analysis.

2 Didactics Tools and Implementation

EDM: overview of students' contingent. First of all, we would like to discuss the variants of students' contingent, which can be suitable to the learning process in EDM-field. We would like to highlight three different approaches in students' contingent selection for EDM-learning.

1. **IT-students:** IT-students often have in their curricula such courses as Data mining, Machine learning, Neural networks. It is assumed that further IT specialist will be able to work with any data from any field of knowledge.

Risks: Usually, the specific features of the educational data analysis are not considered in the learning process. Personal pedagogical experience, knowledge of the educational public demand, trends and standards, the specifics of educational systems and data are completely absent from these students. An open question is the readiness and desire of IT professionals to work in education area.

Benefits: Students have adequate technical competences for the qualitative solutions in the field of educational data analysis problems after studying the relevant subject area.

2. **Students receiving teacher education and specializing in Math, Applied Math or Computer Science:** this group of students is closely connected with the educational system. They are familiar with trends, existing problems and demands of

[1]Pedagogical Universities in Russia graduate teachers as well as other specialists.

Table 1: Characteristics of traditional and digital didactics[Tchoshanov, 2013]

Characteristics	Traditional Didactics	Digital Didactics
Dominating focus	Science and art of teaching	Engineering of learning
Primary goal	Quality of teaching and learning, students' competency and proficiency	Quality of teaching and learning, students' competency and proficiency
Theoretical basis	Research-based guiding principles of learning	Research-based guiding principles of learning
Delivery format	Face-to-face, hybrid	Hybrid, online, e-Learning
Primary teacher's role	Transmitter of knowledge	Engineer of learning
Primary student's role	Information receiver	Connected leaner
Dominating mode of learning	Passive,	Active Interactive
Primary learning and teaching space	Physical classroom, auditorium	Learning management systems, virtual space
Instructional material representation	Text, graphics	Hypertext, media
Instructional material format	Hardcopy	Softcopy
Usage of graphics and visualization	Static and illustrative	Dynamic and interactive
Dominating mode of communication	Verbal	Written
Primary means of communication	Classroom discourse	Online discussion boards, chats, social networks
Information access	Limited by the textbook	Unlimited by ICT resources
Primary mode of scaffolding	Training, instructing	Screencasting, videostreaming
Dominating assessment format	Paper-and-pencil assessment	On-line assessment, e-folio

education, learning activities and educational data. On the other hand, specialization in the field of mathematics and computer science will allow completing the course on EDM, professionally penetrating into the existing approaches and algorithms.

Risks: It is necessary to develop an EDM course and adjust the curriculum taking into account the weak technical skills of students. There are not enough teachers in Pedagogical Universities ready to support this discipline.

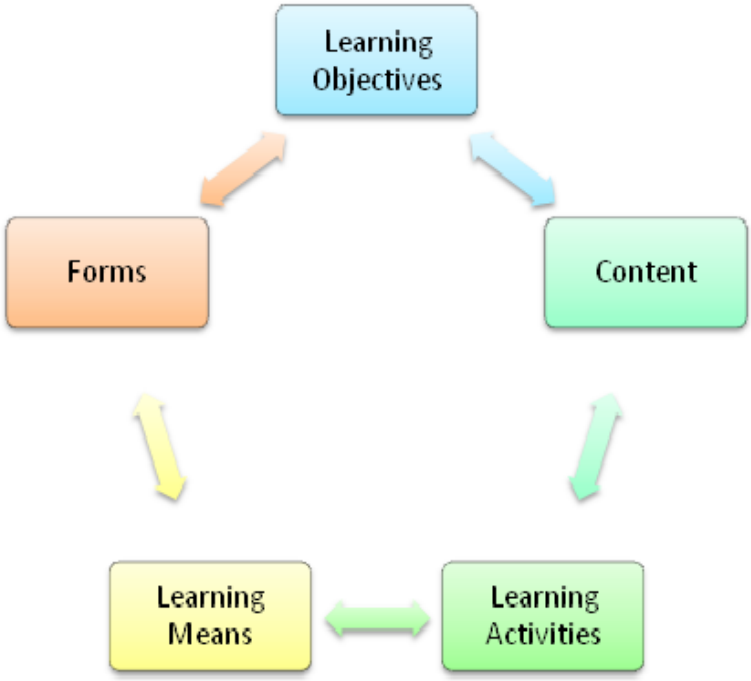

Figure 1: Pyshkalo's scheme of Methodical System

Benefits: Students are interested in working in the field of education and studying modern data mining technologies.

3. **Future and present specialists of education area:** Here we assume the training of data mining of all future (in the program of higher pedagogical education) and present (in the post-graduate programs) specialists in the field of education.

Risks: Poor knowledge in Applied Math and Computer Science.

Benefits: The quality of teacher education can be improved by developing of EDM competences.

Point 2, when EDM-learners are the students of Math and Computer science from pedagogical universities, looks optimally (Fig. 2). In this case, we can prepare specialists which can professionally deal with the modern analytics in the educational sphere, acting as intermediaries between the demands of the educational environment and IT specialists.

Forms and Learning Activities. As the main form of implementation of the educational process, we have chosen the flipped learning format. This format allows teacher to free up classroom time for active learning and practice that is especially important in computer science field. During class time students can be focused on higher forms of cognitive activity in accordance with Bloom's inverted taxonomy, i.e. on application, analysis, evaluation and creation, which contribute to the development of research competence. At the same time the lowest levels of cognitive activities, such as remembering and understanding, are practiced during the time of student's independent work. We have identified several successive stages in the students' study of new material: 1) homework, 2) testing in class, 3) discussion, 4) class activities, 5) individual work (fig. 3). Planning of the learning process is carried out in such a way that classroom activities (stages No. 2 - 5) must take place in 4 academic hours (180 minutes). In our methodical system we

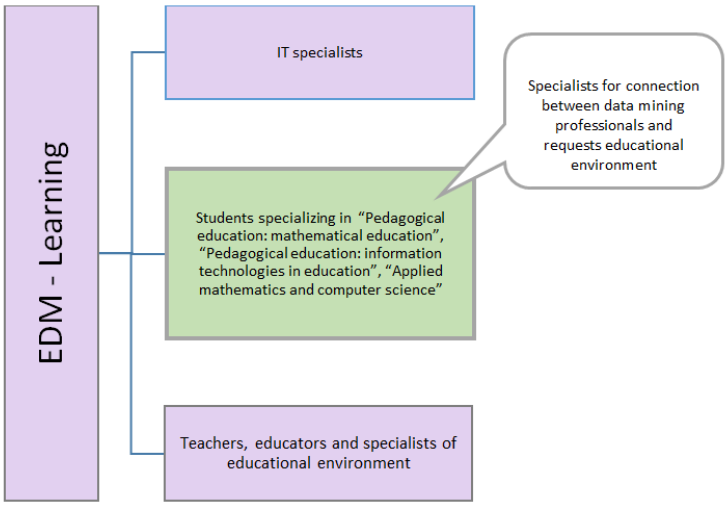

Figure 2: Student's contingent for EDM learning

use such educational activities that contribute to the formation of research competence. Some of selected and applied by us activities are shown in Table 2.

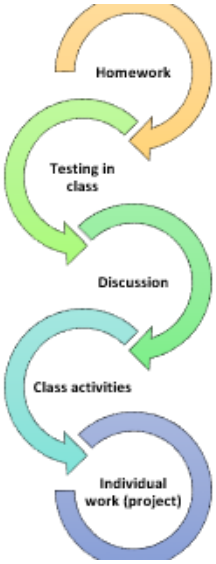

Figure 3: Successive stages in the students' study of new material

Learning Means. As learning means we use existing software for data mining, which we divide into three groups (see Table 3).

For the first acquaintance with the capabilities of data mining, it was decided to use programs designed specially for solving data mining problems (group No 2). This allows us to focus student's attention on familiarity with data mining methods:

- without overloading various functional of statistical packages,
- without diving into the details of data mining algorithms,
- without learning new programming language.

Table 2: Students' learning activities in the training cycle stages

Stages of the training cycle	Learning activities
Homework	Studying of digital materials given by the teacher
Testing	Students perform testing on the materials of homework at the beginning of each classroom. The following types of test questions are used: information reproduction, selection and formulation of basic ideas, understanding of what was studied, comparison, application. We consider testing as a tool for increasing learning motivation, but not as an assessment.
Discussion	After the testing phase, verbal teaching methods, such as discussion and explanation are applied. Test questions that caused difficulties are considered. Students offer their answers, explain to each other, discuss. The teacher guides the discussion, provides explanations if it is necessary.
Class activities	Teacher gives some explanations on the work in some data mining system. For example, if the concept and algorithms of classification were studied in the home material, the possibilities of the system for solving classification problems are considered in the classroom. Further, the results obtained by execution of a specific sequence of actions in data-mining system are discussed in the training group.
Individual work	Such research methods as comparison, analysis, synthesis, experiment are applied. Students solve research tasks with hidden problematic situations.

Table 3: Data mining software

No	Learning Means	Examples
1	Statistical programs that include data mining capabilities as a part of its functionality	Statistica, Matlab, Deductor, SPSS Statistics
2	Specialized programs for data mining	RapidMiner, Orange, Weka, Knime
3	Programming software environment for solving data mining problems	R, Python

Based on the above the Weka program `https://www.cs.waikato.ac.nz/ml/weka/` was chosen. It's a free well-known software program developed at the University of Waikato, New Zealand [Frank et al, 2016]. The program is characterized by a non-overloaded intuitive interface and at the same time strong functionality. Weka implements a number of methods for preprocessing data (discretization, normalization, attribute type conversion, adding noise, missing values replacement, manipulating attributes and values, etc.), classification (lazy methods — k nearest neighbors, etc., based on trees - C4.5,

random forest, etc., based on the rules - decision table, etc., Bayesian methods, methods based on functions and rules, etc.), regression (linear regression, logistic regression, SMO, etc.), clustering (k-means, EM, hierarchical clustering and others.), association rule detection (Apriori, FPG). Weka also provides additional opportunity for results visualization, workflows building and working through the command line. Unlike similar software, there is an "Experimenter" panel that allows you to compare different approaches when conducting serial experiments. Weka is often used for small researches in the field of education [Hajizadeh, 2014, Aher et al., 2012], as well as in other areas [Shaukat et al., 2015, Elghazaly et al., 2017].

Students can get acquainted with other, different from Weka, data mining tools. They prepare presentations about the selected software: Knime, RapidMiner, Orange, Deductor, Matlab, Statistica, Excel, Minitab, R, Python, etc. Presentation requirements are as follows: program name, official website, software access conditions, a brief description of the functionality, own opinion on the work with the system should be reflected. The using of the discussed programs is supposed in the further course tasks.

Content. The basic principles of educational content creation process are as follows: compliance with the objectives of learning, professional orientation, fundamentality, development mental learning, accessibility of content, interdisciplinary and practical orientation. In the course there are theoretical and practical parts. The concentric structure of the content was chosen. This approach suggests:

- returning to previously learned knowledge,
- studying the same problem several times with a gradual expansion of its content,
- in each new iteration enriching students with new information about relations and dependencies of course content [Podlasuy, 2007].

Our aim is to facilitate students' perception of a fundamentally new and difficult course. The concentric model of learning is ideologically similar to the successful iterative model of software development. We have identified three iterations. Each iteration deals at different level with key course topics, such as preprocessing, associative rules and sequential patterns discovery, clustering, regression and classification. Iterations are as follows:

Iteration No 1. *Theory:* Introduction to data mining: basic definitions, tasks and methods. *Practice:* introduction to Weka system, development of searching skills of new information, that is necessary to work in the program.

Iteration No 2. *Theory:* In-depth study of the tasks identified at the first iteration. *Practice:* confident using of Weka system, deep overview of the capabilities of the system; application of the studied methods of data mining to solve problems; skills development of comparing different approaches for problem solving, comparing various forms of results presentation, analyzing and critical understanding of results.

Iteration No 3. *Theory:* Educational data mining. *Practice:* Educational data processing. Implementation of projects / research tasks, that require independent goal setting, choice of approach to solution and result analysis.

The content of theory part in detail for each iteration is listed in Table 4.

Table 4: Learning objectives for Iterations

Iteration No	Theory content
1	Base concepts and tasks of data mining. Introduction to the Weka system. Data preprocessing (examples, data format conversion, discretization). Association rules (the concept of association rule, the definition of support and confidence of the rule, examples, the Apriori algorithm). Classification (difference from the regression problem, stages of classification, the concept of lazy and active classifiers, the simplest algorithms: ZeroR and OneR). Linear regression (problem statement, hypothesis and cost function, gradient descent method). Clustering (types of clustering algorithms: flat and hierarhical, fuzzy and non-fuzzy. K-means algorithm, metrics for determining the similarity of objects, hierarchical clustering and distance metrics for clusters).
2	Familiarity with a variety of data mining programs. Data preprocessing (examples of problems in the data (missing values, data inconsistency, anomalies, noise) and approaches to their elimination, data optimization). Sequential patterns (concept, algorithm for sequential patterns detection by [Agrawal, Srikant, 1995]). Classification (generalization of Bayes formula, naive Bayes classifier, example of the step-by-step operation of the algorithm). Clustering (types of clustering algorithms: partitioning methods, hierarchical methods, density-based, grid-based, model-based methods. EM (expectation-maximization) algorithm).
3	Educational data mining (EDM). Data preprocessing (attribute selection: methods for finding the optimal subset of attributes, schema-dependent and schema-independent methods for evaluating a subset of attributes). Text classification. Data mining for education (educational tasks and data mining methods for their solution, the study of scientific articles on EDM). Optional material (classification algorithm C4.5 based on decision trees, knowledge flow, different data mining system, etc.).

3 Results and Discussion

The participants of the study were students of 4th year bachelor program in Applied Mathematics and Computer Science of the Herzen State University of Russia. Data Mining Course (72 hours) was conducted during 2016-2018. The course was completed by 37 students. Training was held according to the mentioned methodology was in small groups of 7-11 persons. Unfortunately, our observation during educational process showed that at the beginning, students generally had demonstrated a low level of research competence. We have identify following typically difficulties in students' activities.

1. Facing a system error messages, students immediately ask teacher to help; they don't

try to understand the messages on their own, to find information in the professional Internet forums in order to solve the problem.

2. Discovering some unknown parameters during the implementation of new algorithms, students haven' t search for descriptions of these parameters, and leave the values set by default.

3. While homework doing, students have missed obscure theoretical materials or unfamiliar terms without trying to find additional information about it.

Students showed more interest, autonomy and perseverance in problem resolving, coping better with creative and research tasks at the last third iteration of our course. At the end of the educational data mining course, students were asked to assess the quality of their training and the degree of development of various research qualities during the course. According to students' opinion, the course significantly contributed to the improvement of knowledge in the field of data analysis, the development of such qualities as autonomy, result orientation, scientific erudition. The course also had some positive effect, but not so significant, on increasing motivation for researching, hardworking, critical thinking, reflexivity, purposefulness and the ability for information providing and analyzing. The results of the questionnaire regarding the evaluation of the training methodology and students' attitudes to our discipline are shown in the table 5.

Table 5: Evaluation of the training methodology and attitude to the discipline

No	Statement of the Questionnaire	Results (%)
1	Course is useful (yes / no)	100 / 0
2	Course was interesting for me	95 / 5
3	Material presentation is available (yes / partially / no)	95 / 5
4	Acquired knowledge could be useful for me (yes / no)	86 / 14
5	I am planning to study this subject in the future (yes / only if it would be necessary in professional activity / no)	16 / 52 / 32

Initially, the students had a negative attitude toward the need for an independent home study of theoretical material and further testing in a class that corresponds to the flipped learning format. However, in the learning process, they changed their opinion, realizing that the tests stimulate the preliminary mastery of the material and are needed to identify unclear points and their detailed consideration in the classroom. After the course about 20% of students expressed a desire to make a graduation project in the field of data analysis. Some students processed experimental data using data-mining tools in their graduation projects.

4 Conclusions

Comparison and analysis of the educational programs contents, books and courses on educational data mining and related disciplines, generalization of pedagogical experience, analysis of requirements and expectations to the research competence level, as well as data analysis skills and modern educational strategies allowed us to construct methodical system of educational data mining teaching (including goals, content, methods, forms and learning tools), which is adequate to the modern tendencies of digital didactics. Our methodical system is based on iterative content model, flipped learning, data mining tasks and exercises developing researcher features. Our practice demonstrates how to increase the level of research and data analysis competences for students with an average knowledge in Applied Math and Computer Science. The effectiveness of suggested didactical engineering of data mining teaching was tested by the educational process monitoring, students questioning and statistical processing of questionnaires data. Further prospects of this project we see in creation of interrelated courses system that ought to prepare a qualified specialist in the field of Applied Math and Computer Science for useful activities in the field of educational data mining.

5 Acknowledgements

The research was supported by the Russian Science Foundation (RSF), Project "Digitalisation of the high school professional training in the context of education foresight 2035" № 19-18-00108.

References

[Artigue et al., 1991] Artigue, M. Perrin-Glorian, M. (1991). Didactic engineering, research and development tool: Some theoretical problems linked to this duality. For the Learning of Mathematics, 11, Pp. 13-17.

[Agrawal, Srikant, 1995] Agrawal R., Srikant R.(1995) Mining sequential patterns// Proceedings of the Eleventh International Conference on Data Engineering, 6-10 March 1995, Taipei, Taiwan, 1995.Pp. 3–14 doi: 10.1109/ICDE.1995.380416

[Pyshkalo, 1975] Pyshkalo A.M.(1975) Metodicheskaya sistema obucheniya geometrii v nachal'noy shkole: Avtorskiy doklad po monografii «Metodika obucheniya elementam geometrii v nachal'nykh klassakh», predstavlennoy na soiskaniye ... d-ra ped. nauk. = Methodical system of teaching geometry in primary school: Authors report on the PhD dissertation «Methods of teaching geometry elements in primary school». M.: Academy ped. sciences USSR. - 60 p. (In Russ.)

[Tchoshanov, 2013] Tchoshanov M. Engineering of Learning: Conceptualizing e-Didactics. M.:UNESCO Institute for Information Technologies in Education, 2013. - 187 p.

[Podlasuy, 2007] Podlasuy I.P. Pedagogika. Kniga 2 = Pedagogy. Book 2. 2 edition. Moscow: Humanitarian publishing center VLADOS; 2007. - 575 p. (In Russ.)

[Frank et al, 2016] Frank E., Hall M.A., Witten I.H. Data Mining: Practical Machine Learning Tools and Techniques. 4 edition. Morgan Kaufmann; 2016. - 655 p. doi:10.1186/1475-925X-5-51

[Hajizadeh, 2014] Hajizadeh N., Ahmadzadeh M. Analysis of factors that affect students' academic performance - Data Mining Approach. //International Journal of advanced studies in Computer Science and Engineering. 2014; 3 (8). URL: http://www.ijascse.org/volume-3-issue-8/Data_mining_approach.pdf

[Aher et al., 2012] Aher S.B., Lobo L. Applicability of data mining algorithms for recommendation system in e-learning. //Proceedings of the International Conference on Advances in Computing, Communications and Informatics, ICACCI 2012, 2012 Aug. 3-5. Chennai, India. ACM, New York, NY, USA. P. 1034-1040. doi:10.1145/2345396.2345562

[Shaukat et al., 2015] Shaukat K., Masood N., Mehreen S., Azmeen U. Dengue Fever Prediction: A Data Mining Problem. //J Data Mining Genomics Proteomics. 2015; 6 (3). doi:10.4172/2153-0602.100018

[Elghazaly et al., 2017] Elghazaly T., Mahmoud A., Hefny H.A. Political Sentiment Analysis Using Twitter Data. //Proceedings of the International Conference on Internet of things and Cloud Computing, ICC '16, 2016 Mar. 22-23, Cambridge, United Kingdom. ACM, New York, NY, USA, Article 11, - 5 p. doi:10.1145/2896387.2896396