JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Data mining approaches for e e-learning

Author 1:
P.Vijayakumar,
Assistant Professor & Research Scholar,
Department of Computer Science,
Karuppannan Mariappan College,
Muthur-638105,TamilNadu,India.
E-mail:vijaykmcollege@gmail.com

Author 2:
Dr P.Parameswari
Associate Professor,
Department of Computer Science,
Karuppannan Mariappan College,
Muthur-638105,TamilNadu,India.
E-mail:paramtech20@gmail.com

Abstract:

Data Mining (DM), sometimes called Knowledge Discovery in Databases (KDD), is a powerful new technology with great potential to help companies focus on the most important information in the data they have collected via transactions. In the education field, the prediction of students learning performance, detection of inappropriate learning behaviours, and development of student profile may be considered e-learning problems where data mining can successfully solve them.

In this paper, the authoress analyses the possibilities to apply data mining approaches in e-learning context, to predict the students' status referring to their activities and the interest in using advanced tutoring tools. The experiments were performed on the basis of data provided by an e-learning platform (Moodle) regarding the logging parameters of students enrolled on Interactive Tutoring Systems discipline during the second semester of current year.

Keywords: e-learning, data mining, decision, classification, regression.

I. INTRODUCTION

Many of the typical pedagogies provide little immediate feedback to students regarding educational content and require teachers (tutors) to spend hours grading routine assignments. This type of learning isn't very proactive about showing students how to improve comprehension and increase the performance, and fail to take advantage of digital resources that can significantly improve the learning process. Data collected from learning process inside in a Leaning Management System (LMS), Course Management System (CMS) or Virtual Learning Environment (VLE) so-called "big data" make it possible to mine learning information for insights regarding student performance and tutoring approaches [13][2].

A current trend in education is to replace (or combine) teacher-centred learning with student-centred learning. It is more important for instructors (teachers) to be able to analyse what students know and what approaches are most effective for each student, rather than classify students relying on periodic test performance.

Online tools (e.g. e-learning platforms) enable the evaluation of a much wider range of student activities, such as how long they spent to readings, if they need supplementary documentation, where they get electronic resources, and how quickly they learn and understand the key concepts.

To accomplish these tasks, the data mining tools are adequate and can offer the answer to many questions. Discovering new patterns and future behaviours of students, predicting the exam passing rate, identifying the student profile for a certain discipline are just a few problems for data mining in the educational field. In literature, this data mining approach in education is known as Educational Data Mining (EDM).

In this article, the authoress investigates the potential for improving e-learning in context of data mining.

The main goal is to predict the future status of students regarding the interest about Interactive Tutoring System discipline, using an e-learning tool – a Moodle platform. The logging data were collected, processed and analysed with classification and regression data mining algorithms. The results were compared and the best classification model was identified.

The paper is organized as follows: section 2 gives information regarding application of data mining in the educational field, section 3 describes the predictive data mining algorithms used to solve the e-learning problem, section 4 contains the experiments and interpretation of results and the final section presents the conclusions about the advantages of data mining approaches for e-learning improvement.

II.RELATED WORK

Data mining is considered the process of discovering new patterns in large data sets involving methods at the intersection of many disciplines: artificial intelligence, statistics, mathematical modelling, pattern recognition, machine learning and database systems [16][22].

Data Mining can be used to extract knowledge from e-learning systems through the analysis of the information available in the form of the enormous volume of data generated by their users (teachers, tutors, students, guests). In this case, the main objective becomes finding the student profile for a certain discipline, predicting the students abandon, predicting the student's performance and, perhaps most importantly, discovering the students' learning behaviour patterns.

To answer to the question "Where does Data Mining fit in e-learning processes?", the authoress presents examples of various data mining applications retrieved in the literature.

In [14], [19], studies on how DM approaches could successfully be incorporated into e-learning environments and how they could improve the learning tasks were described. A combination of multiple classification algorithms, for the classification of students and the prediction of their final grades, based on features discovered from logging data in an e-leaning platform, was described in [13]. The correct classification rate and prediction accuracy are improved through the weighting of the data feature vectors using a Genetic Algorithm technology. In [11][12], decision trees as classification models were applied. Also in [12], an automatic tutoring tool, based on the students' learning performance and communication preferences, for the designing of simple student profiles was described, with the supplementary goal of creating a personalized education environment.

A DM approaches for classification is association rules [8] and may cover various directions in e-learning context, such us: investigating learning recommendation systems [2], improving and optimization of learning material [9][10], student learning assessments [7][10], optimization of courses adapted to the students' behaviour [5], finding the best implementation of e-learning strategies and evaluation of educational web sites [18].

An interesting topic in e-learning is the prediction problem that may be associated with classification and regression problems in data mining. The forecasting of students' behaviour and performance when using an LMS, CMS or VLE imply the potential of facilitating the improvement of virtual courses in e-learning frameworks. Course log-files stored in databases could be mined by teachers using predictive data mining algorithms (e.g. decision trees, regression, neural networks etc.) to discover important relationships and patterns, with the main scope of discovering relationships between students' knowledge levels, e-learning portal usage times and students' grades. In [6], a study that aimed to detect and identify the sources of error in the prediction of students' knowledge behaviour was developed.

Linear regression was applied in [1] to predict whether the student's next response would be correct or incorrect, and how long the student would need to formulate that answer. In [9], a set of experiments was applied in order to predict the students' performance in e-learning courses, as well as to assess the significance of the attributes involved in the data mining model. In this approach, several DM algorithms were tested, including: Naïve Bayes, kNN, Neural Network, C4.5, Logistic Regression and Support Vector Machines.

The examples could go on, but one thing is certain: data mining can be successfully applied in the educational field.

In the next section, a short description of data mining algorithms regarding prediction problem is presented.

III.PREDICTIVE DATA MINING

The two approaches of prediction problems are classification and regression. Samples of past experience (historical data) with known answers are examined and based on them, the labels are generalized to future cases. For classification, the answer is true or false. For example, finding the answer to the question of whether or not students have passed an exam is a classification problem. For regression, the answer is a number. In the prediction problem, the available cases are necessary to train the prediction model. Then, additional test data are needed for evaluating the performance of the data mining model.

"Predictive data mining is goal directed" [21]. Representative samples of data with known responses, synthesizing past experiences in order to accomplish the established goals, are used in predictive mining. A classical problem has the following formulation: given a situation and a data warehouse, the prediction of a class variable must be done. Predictive data mining can be divided into four important tasks: data preparation, data reduction, data modelling and prediction, and case and solution analyses [19].

In the current article, six predictive data mining models were used in e-learning context: logistic regression model, multilayer perceptron model and decision tree model (J48, Simple CART, JRIP and REPTree).
A short description for each data mining model is given below.

Logistic regression is a data mining model that measures the relationship between the categorical dependent variable and one or more independent variables, by estimating probabilities. Logistic regression is used to predict the odds of being a case based on the values (measurements) of the independent variables (known as predictors). The odds are defined as the probability that a particular outcome (answer) is a case divided by the probability that it is a non-case.

A multilayer perceptron represents a feedforward artificial neural network model that maps sets of input data into a set of appropriate output. The network of perceptron is defined by: input layer, hidden layer(s), and output layer; each connection has a weight (a number); each node performs a weighted sum of its inputs and thresholds the result.

J48 is an implementation of the Quinlan algorithm (C4.5) [17], known as an improvement of the basic ID3 algorithm. The main goal of this classification data mining algorithm is to build a decision tree for the given dataset, using the concept of information entropy. The decision tree nodes represent discrimination rules acting on selective patterns by recursive partitioning of data, using depth-first strategy. While building a tree, J48 ignores the missing values, meaning that the value of that item can be predicted based on what is known about the attribute values for the other records (from historical data). The basic idea of this classifier is to divide the data into ranges based on the attribute values for that item that are found in the training sample.

Simple CART (Classification and Regression Trees) [20] is a classification method which uses historical data to construct decision trees, when the number of classes are a priori known. The inputs of this predictive data mining model can be both numerical and categorical variables. The decision tree is built in accordance with splitting rule (the rule that performs the splitting of learning sample into smaller parts). In practice, each time data (all the observations) have to be divided into two parts with maximum homogeneity [20]. The resulted decision tree is used to classify new data. An important practical characteristic of Simple CART is that the structure of its classification or regression trees is invariant with respect to monotone transformations of independent variables.

JRIP is an optimized version of IREP and represents a data mining model that implements in Java a propositional rule learner (Repeated Incremental Pruning to Produce Error Reduction (RIPPER)), proposed in [4]. This model is based on the construction of a rule set in which all positive examples are covered. In this algorithm, the discovered knowledge is represented in the form of IF THEN ELSE prediction rules.

Reduced Error Pruning Tree (REPTree) is a simple and fast procedure for learning and pruning decision trees. The main task of this classifier is to produce a decision or regression tree using information gain as the splitting criterion and prune trees using REP. This data mining model only sorts values for numeric attributes once Error! Reference source not found..

IV.EXPERIMENTS AND RESULTS

The data used to build predictive data mining models were collected from an open-source e-learning platform (Moodle). This portal is used by authoress as a tutoring e-learning support for students from Petroleum-Gas University of Ploiesti. The data stored refers to logging information regarding students who have studied Interactive Tutoring Systems discipline during the second semester of the current year: access time, IP address, user name, action, information, number of forum posting, number of files upload, number of discussion etc.

Only six attributes with a significant role were used by classification and regression algorithms, as are presented in the table I.

Access_time refers to the period when a participant uses the e-learning platform, by accessing the available resources. Participant represents the "actor" who plays a role on the e-learning platform and owns more or less access rights corresponding to the associated role. Action refers to the available activities on the e-learning platform and provides information about the most interesting topics (exams, attendances, courses etc.). The number of visits on the e-learning platform is stored in the attribute acces_counting.

TABLE I. DESCRIPTION OF VARIABLES USED BY DATA MINING MODEL

Attribute
Type
Value
access_time
nominal
exam_day
less_than_one_week
one_week_befor_exam
two_weeks_befor_exam between_one_and_two_weeks,
more_than_two_weeks
participant
nominal
administrator
student
teacher
guest
action
nominal
resource_view
course_view
attendance_student_view assignment_view
forum_view
upload
posting_messages
manage_account
manage_resources
information

nominal

exam_topics

attendances_number Interactive_Tutoring_Systems_syllabus

student_mark

project

forum_news

course_ITS1

course_ITS2

course_ITS3

course_ITS4

course_ITS5

course_ITS6

course_ITS7

UML

access_counting

numeric

participant_status

nominal

active

passive

The target attribute is participant_status with two possible values: active, passive. A participant (a student, for example) is considered active if he has the following profile: he recently accessed the portal (less than a week), the number of portal visits is more than 20, and his action on the portal is not attendance_student_view or student_mark. In other cases, the student status is noticed as passive.

Applying data mining algorithm was possible through WEKA (Waikato Environment for Knowledge Analysis) software, an open-source machine learning [24]. The data captured from e-learning platform was processed and edited in a standard format accepted by WEKA.

After source file loading, the classifier is chosen (J48) and the model is run according to testing options. As a result, a decision tree is built and the evaluation measures are calculated. The procedure is the same in the case of the rest of data mining models (Logistic regression, Simple CART, Multilayer Perceptron, JRIP and REPTree).

The interpretation of evaluation measures (correctly classified instances, incorrectly classified instances, Kappa Statistic, mean absolute error, root mean squared error) offers information about the performance associated to the considered data mining models (table II).

TABLE II. COMPARISON OF STATISTICAL PARAMETERS

Logistic regression

J48

Simple CART

Multilayer Perceptron

JRIP

REPTree

Correctly Classified Instances

86,22%

(144 instances)

91,61%

(153 instances)
89,22%
(149 instances)
88,02%
(147 instances)
88,02%
(147 instances)
89,22%
(149 instances)
Incorrectly Classified Instances
13,77%
(23 instances)
8,38%
(14 instances)
10,77%
(18 instances)
11,97%
(20 instances)
12,77%
(20 instances)
10,77%
(18 instances)
Kappa
Statistic
0,66
0,79
0,74
0,70
0,72
0,73
Mean absolute error
0,17
0,13
0,13
0,13
0,16
0,15
Root mean squared error
0,34
0,26
0,30
0,32
0,32
0,30

The comparison of the results indicates a better classification in case of J48, Simple CART and REPTree models.

In contrast with JRIP, which is another classifier based on discovering the rules, J48 presented the best performance for unseen data (test set). JRIP produced only four simple rules for classifying the flow patterns (fig. 1)

Fig. 1 JRIP Rules

The method based on neural networks (Multilayer Perceptron) has the same performance as JRIP. The logistic regression algorithm has the lowest number of correct classified instances (144 instances) and a mean absolute error of classification about 0.17.

After building predictive data mining models for the current problem, the next step is to test the performance of the models.

The procedure is the following: after a new test file ARFF (e-learning_test.arff) is edited, with the same structure as the training file, but only with one instance (fig.3), the training file (e-learning.arff) is loaded and then a classifier is selected (e.g. Multilayer Perceptron).

Obviously, the test options can be specified: Use training set – the loaded file is used for testing the data mining model;

Supplied test set - means that the user can choose a file with the test data (e-learning-test.arff), Cross-validation - means that the classification results will be evaluated by cross-validation. In this mode the number of folds can be modified, Percentage split - means that classification results will be evaluated on a test set that is a part of the original data. The default split is 66%, which means that 66% of the data go for training and 34% for testing.

The test file contains 194 instances. The results indicate a performance for Multilayer Perceptron classifier as 97.42% (189 instances correct classified) and 0.03 mean absolute error.

The results of testing the rest of the data mining models considered in this article (Logistic regression, J48, Simple CART, JRIP and REPTree) are compared in the table III.

TABLE III. COMPARISON OF STATISTICAL PARAMETERS AFTER TESTING DATA MINING MODELS

Logistic regression
J48
Simple CART
Multilayer Perceptron
JRIP
REPTree
Correctly Classified Instances
91,75%
(178 instances)
91,23%
(177 instances)
90,72%
(176 instances)
97,42%
(189 instances)
91,75%
(178 instances)
89,22%
(149 instances)
Incorrectly Classified Instances
8,24%
(16 instances)
8,76%
(17 instances)

9,27%
(18 instances)
2,57%
(5 instances)
8,24%
(16 instances)
10,77%
(18 instances)
Kappa
Statistic
0,79
0,78
0,78
0,93
0,81
0,73
Mean absolute error
0,12
0,14
0,10
0,03
0,13
0,15
Root mean squared error
0,24
0,28
0,27
0,13
0,26
0,30

The process of classifying new instances follows the same procedure: first a test file is edited with the new instances, then a classifier is selected (Multilayer Perceptron) and the test file is loaded. Assume that the new instance is the following:

{less_than_one_week, student, course_view,
Interactive_Tutoring_Systems_syllabus, 79, ?} (1)

The class attribute is “?” (unknown) because the classification data mining must find the value for the target attribute (active or passive).

In the test file, the guess value for class attribute participat_status is active (fig. 2).

Fig. 1 The test file with a single instance

After selecting a classifier and setting its parameters, the output from the classifier is presented in the Classifier output window. Correctly Classified Instances tells that the assumed value of the class attribute (participant_status=active) was correct (according to Multilayer Perceptron).

The predictive data mining models can help teachers to identify the factors that influence the student participation on the e-learning platform and to increase the interest in using modern tutoring tools. Furthermore, the behaviour of a student with a certain profile may be predicted and various classes of students may be identified, such as: regular student, visitor student, bad student, good student etc.

V. CONCLUSION

This research paper discusses about data mining approaches applied in the educational field. Predictive models (classification and regression models) are proposed to solve various problems in e-learning environment. Data used in the experiments were provided by an e-learning platform (Moodle) and represents logging data of students' activities on the portal. Various decision tree models were trained and tested to predict the students' status referring to their intervention on the portal corresponding to the Interactive Tutoring Systems discipline. A comparison of six predictive data mining model (logistic regression, multilayer perceptron, J48, Simple CART, JRIP, REPTree) was conducted to find the classifier with the best performance for the formulated problem. In the training phase, J48 algorithm obtained the best correct classification rate (91.61%), but in testing phase Multilayer Perceptron model was ranked first (97.42%), proving once again self-learning nature of neural networks. Future work will focus on identifying the attributes which most influence the students' interest in using an e-learning environment, how to improve the students' capabilities in acquiring knowledge and to increase the students' performance, in the context of data mining.

REFERENCES

[1] J.E. Beck, B.P. Woolf, High-Level Student Modeling with Machine Learning, Gauthier, G., et al. (eds.): Intelligent Tutoring Systems, ITS 2000. Lecture Notes in Computer Science, Vol. 1839. Springer-Verlag, Berlin Heidelberg New York, 2000, pp. 584-593.

[2] F. Castro, A. Vellido, A. Nebot, F. Mugica, Applying Data Mining Approaches to e-Learning Problems, Studies in Computational Intelligence, Volume 62, 2007, pp. 183-221.

[3] K. Chu, M. Chang, Y. Hsia, Designing a Course Recommendation System on Web based on the Students' Course Selection Records, World Conference on Educational Multimedia, Hypermedia and Telecommunications, 2003, pp.14-21.

[4] W. Cohen, Fast effective rule induction, in Proceedings of the 12th International Conference on Machine Learning, Lake Tahoe, Calif, USA, 1995, pp. 115–123

[5] M.F. Costabile, A. De Angeli, T. Roselli, R Lanzilotti, P. Plantamura, Evaluating the Educational Impact of a Tutoring Hypermedia for Children, Information Technology in Childhood Education Annual, 2003, pp. 289-308.

[6] M. Feng, N. Heffernan, K. Koedinger, Looking for Sources of Error in Predicting Student's Knowledge, The Twentieth National Conference on Artificial Intelligence by the American Association for Artificial Intelligence, AAAI'05, Workshop on Educational Data Mining. July 9-13, Pittsburgh, Pennsylvania, 2005, pp.54-61.

[7] G.J. Hwang, C.L. Hsiao, C.R. Tseng, A Computer-Assisted Approach to Diagnosing Student Learning Problems in Science Courses, Journal of Information Science and Engineering 19, 2003, pp. 229-248.

[8] L. C. Jain, R. A. Tedman, D. K. Tedman, Evolution of Teaching and Learning Paradigms in Intelligent Environment, Studies in Computational Intelligence, 62, 2007, ISBN: 978-3-540-71973-1 (Print) 978-3-540-71974-8 (Online).

[9] S.B. Kotsiantis, C.J. Pierrakeas, P.E. Pintelas, Predicting Students' Performance in Distance Learning Using Machine Learning Approaches. Applied Artificial Intelligence 18(5), 2004, pp. 411-426.

[10] A. Kumar, Rule-Based Adaptive Problem Generation in Programming Tutors and its Evaluation, 12th International Conference on Artificial Intelligence in Education. July 18-22, Amsterdam, 2005, pp. 36-44

[11] A. Liang, W. Ziarco, B. Maguire, The Application of a Distance Learning Algorithm in Web-Based Course Delivery, Ziarko, W., Yao, Y. (eds.): Second International Conference on Rough Sets and Current Trends in Computing. Lecture Notes in Computer Science. Springer-Verlag, Berlin Heidelberg New York, 2000, pp. 338-345.

[12] O. Licchelli, T.M. Basile, N. Di Mauro, F. Esposito, Machine Learning Approaches for Inducing Student Models, 17th International Conference on Innovations in Applied Artificial Intelligence, IEA/AIE 2004. LNAI Vol. 3029. Springer-Verlag, Berlin Heidelberg New York, 2004, pp. 935-944.

[13] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A. Byers, Big Data: The Next Frontier for Innovation, Competition, and Productivity, McKinsey Global Institute, May, 2011.

[14] H. Margo, Data Mining in the e-Learning Domain., Computers & Education 42(3), 2004, pp. 267-287.

[15] B. Minaei-Bidgoli, W.F. Punch, Using Genetic Algorithms for Data Mining Optimization in an Educational Web-based System, Cantu, P.E., et al. (eds.): Genetic and Evolutionary Computation Conference, GECCO 2003, 2003, pp. 2252-2263.

[16] G. Piatetsky-Shapiro, R.J. Brachman, T. Khabaza, W. Kloesgen, E. Simoudis, An Overview of Issues in Developing Industrial Data Mining and Knowledge Discovery Applications, In KDD, 96, pp. 89-95, 1996.

[17] R.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Francisco, USA, 1993.

[18] M.L. Dos Santos, K. Becker, Distance Education: a Web Usage Mining Case Study for the Evaluation of Learning Sites, The 3rd IEEE International Conference on Advanced Learning Technologies, ICALT'03. IEEE Computer Society. Athens Greece, 2003, pp. 360-361.

[19] T.Y. Tang, G. McCalla, Smart Recommendation for an Evolving e-Learning System: Architecture and Experiment, International Journal on e-Learning 4(1), 2005, pp. 105-12

[20] R. Timofeev, Classification and regression trees (CART) theory and applications, 2004.

[21] S.M. Weiss, N. Indurkhya, Predictive data mining: a practical guide, Morgan Kaufmann, 1998.

[22] I.H. Witten, M. Frank, M.A. Hall, Data Mining: Practical Machine Learning Tool and Approaches with Java Implementation, Morgan Kaufmann, San Francisco, USA, 3rd edition, 2011.

[23] http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/REPTree.html

[24] http://www.cs.waikato.ac.nz/ml/weka/