

Predicting Learners' Performance in an E-Learning Platform Based on Decision Tree Analysis

Badr HSSINA¹, Abdelkrim MERBOUHA², and Belaid BOUIKHALENE¹

¹TIAD laboratory, Computer Sciences Department Sultan Moulay Slimane University, FST
Beni-Mellal, Morocco

²LMACS laboratory, Mathematics Department Sultan Moulay Slimane University, FST
Beni-Mellal, Morocco

Abstract: The ability to predict learners' performance on an e-learning platform is a decisive factor in the current educational systems. Indeed, learning through decision trees uses more sophisticated and efficient algorithms based on the use of predictive models. A decision tree is a decision support tool for assessing the value of a characteristic of a population based on the observation of other characteristics of the same population. As our research is focused on how to help a tutor to monitor the learners' activities on e-learning systems (B. HSSINA and al.2015 [1], B. HSSINA and al.2014 [2] [3]), we propose a predictive model based on the algorithm ID3, C4.5 and CART to predict the level of learners in their learning path in a training. The choice of the most efficient algorithm is done based on a comparative study between different algorithms of decision trees, which leads us to confirm that the most powerful and most convenient is C4.5. The data used by the latter is harvested from the e-learning platform on which learners are enrolled.

Keywords: E-learning, Data mining, Decision trees.

1. Introduction

Exploration of educational data is an emerging and promising field which allows the development of methods that explore the data that comes from e-learning platforms. Data mining can provide answers to more abstract questions like "find learners who can pass the exams". Traditional database queries can only answer questions such as "find the students who failed the exams". One of the main objectives of the data mining application in e-learning is the development of students' models that would predict students' performance in their schools [4]. The ability to predict the performance of students with high accuracy is very important for a teacher to identify students with low achievements who are likely to have weak results [5]. Indeed, the idea of developing an accurate prediction model based on decision tree to automatically identify weak students is a very interesting and auspicious work. Recognizing the level of students, tutors are able to provide additional support to their learners during their training to improve the quality of education.

2. Related Works

Data Mining is used in the education sector in order to improve our understanding of the learning process and focus on the identification, extraction and evaluation of variables related to the learning process of learners [6].

Bray 2007, in his study on private tutoring and its implications [7], compared the percentage of students receiving private tutoring in India, Malaysia, Singapore, Japan, China and Sri Lanka. He observed that there was an enhancement of performance and the variation of intensity of private tutoring which depends on the socio-economic conditions.

Kabra discussed the use of decision trees in predicting the performance of students by using the data of 346 students and the results showed that the generated tree is only 209 instances out of 346 were correctly classified [8].

Galit and all. [9] used, in a case study, data from learners to analyze their learning behavior so as to predict and notify the students' learning level.

Decision Tree was applied to e-learning in [18], [19], [20], [21]. An e-Learning platform for the personalization of courses, based both on the learner's needs and capabilities, was described in [22]. Personalized learning paths in the training were

modeled using graph theory [23]. Likewise in [24], an automatic tool, based on the level of communication and learning preferences of learners, for the generation and discovery of simple study models has been described, with the ultimate goal of creating a personalized educational environment.

Association Rules for machine learning, applied to e-learning, have been explored in the areas of learning recommendation systems [25], [26], learning material organization [27], student learning assessments [28], [29], [30], course adaptation to the students' behavior [31], [32], and evaluation of e-learning Web sites [33].

3. Decision trees

Decision trees are very effective methods of supervised learning [13]. They partition the dataset into more homogenous groups as possible from the variable to be predicted. We take in the input set of ordered data, and output shaft is provided wherein each end node (leaf) is a decision (a class) and each non-end node (middle) shows a test.

To build a tree, several algorithms exist: ID3 [14], CART [15], C4.5 [16], RF [17] among others. It usually starts with choosing the best attribute and this attribute is assigned to the root. For each value of this attribute, we create a new son node of the root and thereafter we classify the examples in the son nodes. If all examples of a son node are homogeneous, it affects their class nodes. But if not, we start from that node.

3.1. Choosing the segmentation variable

To choose the best segmentation variable from the attributes of the data, we calculate the "GAIN" of each attribute depending on the different values taken by this attribute. This measure is based on research in theory of information carried by C. Shannon [12].

3.1.1. Shannon Entropy

In general, if we are given a probability distribution $P = (P_1, P_2, \dots, P_n)$ and a sample S , then the Information carried by this distribution, also called the entropy of P , is given by:

$$Entropy(S) = - \sum_{i=1}^n p_i \times \log_2(p_i)$$

The advantage of using entropy is that the algorithm operates on symbolic data that are categorical variables (such as color) or discrete digital.

3.1.2. The gain information $G(S, Q)$

We have functions that allow us to measure the degree of heterogeneity of classes in all samples and any position of the tree in construction. It remains to define

a function to select the test that must label the current node.

It defines the gain for a set S and an attribute Q :

$$G(S, Q) = E(S) - \sum_{i=1}^k p_i E(S, Q_i)$$

The values p_i are the set of all possible values for attribute Q . We can use this measure to rank attributes and build the decision tree where at each node is located the attribute with the highest information gain among the attributes not yet considered in the path from the root.

The disadvantage of this method is that to preserve the effectiveness of learning and relevance of the product model, continuous variables must be discretised before the implementation of the algorithm.

This is a recursive structure by cutting successively the set of observations. If all samples are in one class, one places a sheet of this class. If one chooses a most discriminating issue possible, we divide all examples according to this issue. For each set we build a decision sub-tree

3.2.C4.5 algorithm

This algorithm was proposed in 1993, again by Ross Quinlan [10], from a training sample composed of a target variable or predicted variable Y and at least a learning variable or predictor variables $X = \{x_1, x_2, x_3, \dots\}$. C4.5 produces a type of decision tree model. This model can predict for an individual to the estimated value of the variable $\hat{y}_i$ objective based on values taken by the "predictor" variables x_i . The C4.5 algorithm is based on a measure of entropy in the sample learning to produce the graphical model.

```

Algorithm: BUILD-TREE C4.5( $D$ )
input : dataset  $D$ 
output : Tree
If  $D$  is "pure" OR other stopping criteria met Then
    Terminate
End if
For all attribute  $a \in D$  do
    Compute information-theoretic criteria if we split on  $a$ 
End for
 $a_{best}$  = Best attribute according to above computed criteria
Tree = Create a decision node that tests  $a_{best}$  in the root
 $D_v$  = Induced sub-datasets from  $D$  based on  $a_{best}$ 
for all  $D_v$  do
    Tree $_v$  = C4.5( $D_v$ )
    Attach Tree $_v$  to the corresponding branch of Tree
End for
Return Tree
    
```

Figure 1.C4.5 algorithm to construct a decision tree

Theory of Shannon is at the base of the ID3 algorithm and thus C4.5. Entropy Shannon is the best known and most applied. It first defines the amount of information provided by an event: the higher the probability of an event is low (it is rare), the more information it provides is great [11].

4. Description of our system

Learners log into the e-learning platform (Moodle in our case) for training. Our system integrates the algorithms of decision trees to evaluate the performance of learners and their collaborative work based on data taken from the Moodle database, as shown in Figure2. The results of prediction greatly facilitate the task of tutors.

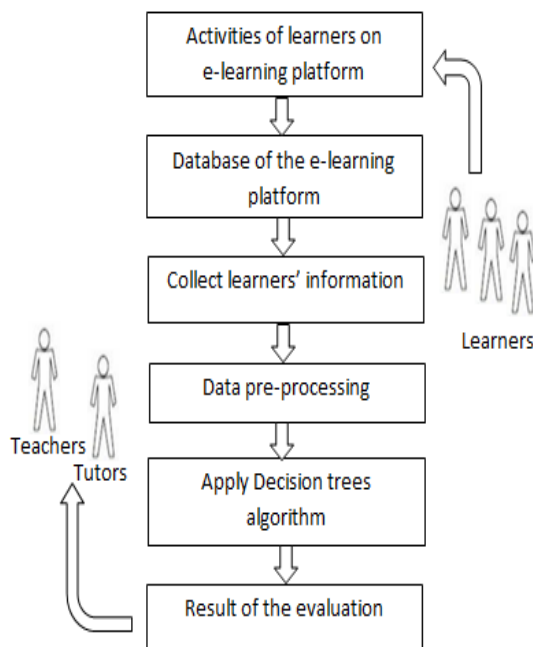


Figure 2. Architecture of our system

In fact, our system helps to have an idea on the level of learners and evaluate their performance during training on the e-learning platform.

5. Description of learners

Our goal is to predict learners' performance (decision is the variable to predict) based on the learners' data (Mark_exam, quiz_p, wiki ...).

The following table shows the different variables used in our study.

Table 1. Description of the meaning of attributes characterizing learners.

variables	Description	Possible values
T_wiki	Type of wiki done	{A,B,C,D}
Nb_assignment	Number of assignment done	an integer
Quiz_p	Number of quiz passed	an integer
Quiz_f	Number of quizzes failed	an integer

msg_forum	Number of messages sent to the forum	an integer
msg_read	Number of messages read to the forum	{A,B,C}
Indication_tutor	The grade given by the tutor to learners	an integer
T_time	Total time used on quizzes	an integer
Mark_exam	Examination mark is higher than 10	Yes, No
Sex	Learners sex	{M,F}
Age	Learners age	an integer
Specialty_L	Specialty of learners.	{ chemistry, mathematics, physics, computer science}
decision	Boolean	{accept or reject}

The decision tree is created from a learning base. We have a sample of measures with a series of parameters chosen for their connection (and discriminative power) with the parameter that it is sought to find. A decision tree thus makes it possible to test a parameter and the results are generally expressed as a probability of satisfying or not the parameter tested.

6. Experience and evaluation

Our learning base consists of 270 learners. Each learner is described by 13 variables; the objective is to automatically classify a learner's performance on the basis of their activities and in particular on their collaborative work on the e-learning platform. In this study we have two classes that are defined by the 13th variable value (decision). The other 13 variables can take different values.

The predictive model derived from our knowledge base by the C4.5 algorithm is illustrated in Figure 3 :

```

Mark_exam = non
|   Specialty = D
|   |   Quiz_P <= 4: Refuser (2.0)
|   |   Quiz_P > 4: Acceptor (10.0/1.0)
|   Specialty = A: Refuser (120.0/18.0)
|   Specialty = B
|   |   sexe = masculin: Acceptor (21.0/6.0)
|   |   sexe = feminin: Refuser (10.0/1.0)
|   Specialty = C
|   |   msg_read = C: Acceptor (9.0/1.0)
|   |   msg_read = A
|   |   |   sexe = masculin: Acceptor (4.0/1.0)
|   |   |   sexe = feminin: Refuser (5.0)
|   |   msg_read = B: Acceptor (0.0)
Mark_exam = oui
|   T_wiki = D
|   |   Quiz_P <= 7
|   |   |   Specialty = D: Acceptor (2.0)
|   |   |   Specialty = A: Refuser (9.0/2.0)
|   |   |   Specialty = B: Acceptor (6.0/1.0)
|   |   |   Specialty = C: Acceptor (1.0)
|   |   Quiz_P > 7: Acceptor (52.0/2.0)
|   T_wiki = C
|   |   age <= 35: Refuser (8.0/1.0)
|   |   age > 35: Acceptor (3.0)
|   T_wiki = B: Refuser (4.0/1.0)
|   T_wiki = A: Refuser (4.0/1.0)

```

Number of Leaves : 18

Size of the tree : 28

Figure 3. C4.5 rules

Comparison of the accuracy and execution time of algorithms of construction of decision trees ID3, C4.5 and CART:

Table 2. Comparison of the accuracy and execution time of ID3, C4.5 and CART.

	ID3	C4.5	CART
Execution time (Sec)	0,01	0,05	0,27
Recognition rate (%)	0,7519	0,8730	0,8641
Error rate (%)	0,2481	0,1370	0,1259

In this work, we analyzed the data of 28 enrolled students in the first year of the master system and Computer Engineering at the University of Sultan Moulay Slimane Beni Mellal in Morocco.

Figure 4 below shows a comparison of the three algorithms of construction of decision trees that produce predictive models with the best class judicious accuracy.

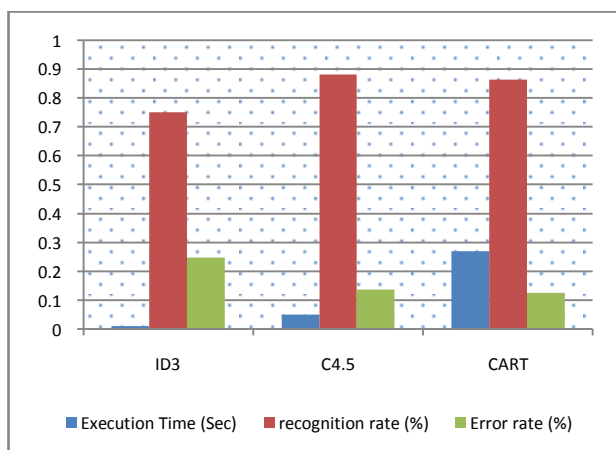


Figure 4. Comparison of the accuracy and execution time of ID3, C4.5 and CART

7. Conclusion

Decision trees are tools to help decision making and data mining. They can model graphically and quickly complex phenomenon. Decision trees provide effective methods that perform well in practice.

This study will help the tutor and the teacher to improve the performance of the Learners. This study introduces Machine Learning by decision trees that can be used in practical settings to predict learners' academic performance. The result of this study shows that the model based on C4.5 is the best predictor in terms of execution time (0.05 second) and recognition rate (87.30 %).

This study contributes to a better division of learners to identify those that require special attention to ensure an effective behavior at the right time by tutors.

References

- [1] Hssina, B., Bouikhalene, B., & Merbouha, A. (2015). Towards an E-Learning Platform Multi-Agent Based On the E-Tutoring for Collaborative Work. World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering, 9(4), 961-965.
- [2] Hssina, B., Merbouha, A. E., & Bouikhalene, B. (2014). Web services at the service of e-learning platforms. International Journal of Innovation and Applied Studies, 7(4), 1574.
- [3] HSSINA, B., MERBOUHA, A., EZZIKOURI, H., & ERRITALI, M. (2014). A comparative study of decision tree ID3 and C4. 5. International Journal of Advanced Computer Science and Applications, 4(2).
- [4] Adhatrao, K., Gaykar, A., Dhawan, A., Jha, R., & Honrao, V. (2013). Predicting Students' Performance using ID3 and C4. 5 Classification Algorithms. arXiv preprint arXiv:1310.2071.
- [5] Akhtar, S., Warburton, S., & Xu, W. The use of an online learning and teaching system for monitoring computer aided design student participation and predicting student success. International Journal of Technology and Design Education, 1-20.
- [6] Klinedinst, R. E. (1991). Predicting performance achievement and retention of fifth-grade instrumental students. Journal of Research in Music Education, 39(3), 225-238.
- [7] Mallick, P. K. (2016). Research Advances in the Integration of Big Data and Smart Computing.
- [8] Kabra, R. R., & Bichkar, R. S. (2011). Performance Prediction of Engineering Students using Decision Trees International Journal of Computer Applications (0975 8887) Volume 36-No. 11.
- [9] Ben-Zadok, G., Hershkovitz, A., Mintz, E., & Nachmias, R. (2009). Examining online learning processes based on log files analysis: A case study. In 5th International Conference on Multimedia and ICT in Education (m-ICTE'09).
- [10] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... & Zhou, Z. H. (2008). Top 10 algorithms in data mining. Knowledge and information systems, 14(1), 1-37.
- [11] Benjamin Devéze & Matthieu Fouquin, DATAMINING C4.5 – DBSCAN, PROMOTION 2005, SCIA Ecole pour l'informatique et techniques avancées.
- [12] C. F. L. Lima, F. M. d. Assis and C. P. d. Souza, "Decision Tree Based on Shannon, Rényi and

- Tsallis Entropies for Intrusion Tolerant Systems," Internet Monitoring and Protection (ICIMP), 2010 Fifth International Conference on, Barcelona, 2010, pp. 117-122.
- [13] G. Holmes and L. Trigg, "A diagnostic tool for tree based supervised classification learning algorithms," Neural Information Processing, 1999. Proceedings. ICONIP '99. 6th International Conference on, Perth, WA, 1999, pp. 514-519 vol.2.
- [14] D. Thakur, N. Markandaiah and D. S. Raj, "Re optimization of ID3 and C4.5 decision tree," Computer and Communication Technology (ICCT), 2010 International Conference on, Allahabad, Uttar Pradesh, 2010, pp. 448-450.
- [15] Lausen, B., Sauerbrei, W., & Schumacher, M. (1994). Classification and regression trees (CART) used for the exploration of prognostic factors measured on different scales.
- [16] Polat, K., & Güneş, S. (2009). A novel hybrid intelligent method based on C4. 5 decision tree classifier and one-against-all approach for multi-class classification problems. Expert Systems with Applications, 36(2), 1587-1592.
- [17] Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS Journal of Photogrammetry and Remote Sensing, 67, 93-104.
- [18] Carchiolo, V., Longheu, A., Malgeri, M., & Mangioni, G. (2003, July). Courses personalization in an e-learning environment. In Advanced Learning Technologies, 2003. Proceedings. The 3rd IEEE International Conference on (pp. 252-253). IEEE.
- [19] Chang, C., & Wang, K. (2001). Discover sequential patterns of learning concepts for behavioral diagnosis by interpreting web page contents. In World Conference on Educational Multimedia, Hypermedia and Telecommunications. Chesapeake, VA (pp. 251-256).
- [20] CHIA-IChang, F. L. O. R. A., Hung, L. P., & Shih, T. K. (2003). A new courseware diagram for quantitative measurement of distance learning courses. Journal of Information Science and Engineering, 19, 989-1014.
- [21] Yoo, J., Yoo, S., Lance, C., & Hankins, J. (2006, March). Student progress monitoring tool using treeview. In ACM SIGCSE Bulletin (Vol. 38, No. 1, pp. 373-377). ACM.
- [22] Levin, T., & Wadmany, R. (2008). Teachers' views on factors affecting effective integration of information technology in the classroom: Developmental scenery. Journal of Technology and Teacher Education, 16(2), 233.
- [23] Castro, F., Vellido, A., Nebot, À., & Mugica, F. (2007). Applying data mining techniques to e-learning problems. In Evolution of teaching and learning paradigms in intelligent environment (pp. 183-221). Springer Berlin Heidelberg.
- [24] Saeed, N., Yang, Y., & Sinnappan, S. (2009). Emerging Web Technologies in Higher Education: A Case of Incorporating Blogs, Podcasts and Social Bookmarks in a Web Programming Course based on Students' Learning Styles and Technology Preferences. Educational Technology & Society, 12(4), 98-109.
- [25] Pazzani, M. J., & Billsus, D. (2007). Content-based recommendation systems. In The adaptive web (pp. 325-341). Springer Berlin Heidelberg.
- [26] Basu, C., Hirsh, H., & Cohen, W. (1998, July). Recommendation as classification: Using social and content-based information in recommendation. In AAAI/IAAI (pp. 714-720).
- [27] Ausubel, D. P. (1960). The use of advance organizers in the learning and retention of meaningful verbal material. Journal of educational psychology, 51(5), 267.
- [28] Stiggins, R. J., & Chappuis, J. (2012). An introduction to student-involved assessment for learning (pp. 29-30). Boston: Pearson.
- [29] Brown, G. A., Bull, J., & Pendlebury, M. (2013). Assessing student learning in higher education. Routledge.
- [30] Higgins, R., Hartley, P., & Skelton, A. (2002). The conscientious consumer: Reconsidering the role of assessment feedback in student learning. Studies in higher education, 27(1), 53-64.
- [31] Kandel, E. R., Schwartz, J. H., & Jessell, T. M. (Eds.). (2000). Principles of neural science (Vol. 4, pp. 1227-1246). New York: McGraw-hill.
- [32] Friend, M., & Bursuck, W. D. (2002). Including students with special needs: A practical guide for classroom teachers. Allyn & Bacon, A Pearson Education Company, 75 Arlington Street, Boston, MA 02116.
- [33] Khan, B. H. (Ed.). (2005). Managing e-learning: Design, delivery, implementation, and evaluation. IGI Global.