

## Prediction Method based DRSA to Improve the Individual Knowledge Appropriation in a Collaborative Learning Environment: Case of MOOCs

Sarra Bouzayane

University of Picardie Jules Verne, Amiens, France  
Higher Institute of Computer Science and  
Multimedia, Sfax, Tunisia  
Sarra.bouzayane@u-picardie.fr

Inès Saad

Amiens Business School, France  
University of Picardie Jules Verne, Amiens, France  
ines.saad@esc-amiens.com

### Abstract

*This paper proposes a prediction method that relies on the Dominance-based Rough Set Approach (DRSA) to improve the individual knowledge appropriation when the learning process occurs in a collaborative environment such as the Massive Open Online Courses (MOOCs). This method is based on two phases: the first has to be applied at the end of each week of the MOOC and aims at inferring a preference model resulting in a set of decision rules; the second is applied at the beginning of each week of the same MOOC and consists of classifying each learner in one of the three defined decision classes, which are C11 of the “At-risk Learners”, C12 of the “Struggling Learners” and C13 of the “Leader Learners”, based on the previously inferred preference model. This method runs weekly. It has been validated on real data of a French MOOC proposed by a Business School in France.*

### 1. Introduction

A MOOC is a model of educational delivery that is, to varying degrees, massive, with theoretically no limit to enrollment; open, allowing anyone to participate, usually at no cost; online, with learning activities typically taking place over the web; and a course, structured around a set of learning goals in a defined area of study [9]. As a product of the digital age revolution and a form of distance learning, the MOOC has become an alternative to the traditional higher education courses [26]. Since 2008, when the first MOOC has been coined by Downes and Siemens [8], the number of MOOCs and of their platform providers has rapidly increased around the world, especially in 2015 where the total number of MOOCs reached 4,200 [19].

However, despite their increasing popularity and proliferation, MOOCs are faced with big limitations

such as the high dropout rate that usually reaches 90% [25]. Researchers in this domain link the learners’ dropout behavior to several factors such as the lack of interaction with the instructor and the course content difficulty [12]; the lack of time, the lack of digital skills and the late starting [18] and the voluntary mode of participation [17]. This excessive dropout rate has encouraged researchers to think of the methods for early predicting the learners who are at risk to dropout MOOCs in order to help them carry on following the courses.

Several prediction models were proposed in literature based on many learning machine techniques. In this context, we think that when focusing on the dropout rate, it is important to address the problem of the tutor’s absence that obviously degrades the quality of the learning process and so encourages learners to dropout the course. Hence, the quality of the learning process in a context of MOOCs depends on the quality of the knowledge transfer process between the learners and the pedagogical team. According to Davenport and Prusak [7], the knowledge transfer process is based not only on the knowledge transmission by the transmitter but also on the knowledge appropriation by the receiver. In this work, we focus on the knowledge appropriation process that occurs when the transmitted information is absorbed, so interpreted by an individual cognitive process into a tacit knowledge which will be used thereafter.

Thus to deal with this issue, we propose a method based on the Dominance-based Rough Set Approach (DRSA) [10] in order to weekly predict first, the learners who are likely to dropout the course during the current week of the MOOC, called the “At-risk Learners”; second, the learners who do not intend to leave the MOOC but who have some difficulties with the learning process, called the “Struggling Learners”, and finally the learners characterized by a profile and a behavior permitting them to support the

two other groups of learners by providing them with the accurate and the effective knowledge, called “Leader Learners”.

This method helps not only weekly predict the dropout rate, but also improve the learning process. It consists of two phases. The first aims to construct a preference model and comprises three steps which are: First the identification of a training sample of learners, then the construction of a coherent family of criteria to characterize the learner’s profile, and finally the inference of a preference model resulting in a set of decision rules. The second consists of the classification of the new learners called “Potential Learners” based on the previously inferred preference model. The method has been validated on real data coming from a French MOOC.

The remainder of this paper is organized as follows: section 2 shows the related work. Section 3 sets the background. Section 4 describes the method. Section 5 is dedicated to the application of this method. Section 6 concludes the work and advances some prospects.

## 2. Related Work

This section consists of two parts. The first is about the methods proposed to enhance the collaborative learning process. The second presents the models proposed to predict the dropout rate.

### 2.1. Methods proposed to enhance the collaborative learning process

Many works in the literature are concerned with the improvement of the learning process especially when it occurs via the technological tools.

Authors in [14] aim to improve the learning process taking place on the online discussion forums where learners share the same articles but do not have similar preferences. Hence, they consider the learners’ preferences, expressed by a rating tool, in order to recommend an appropriate set of articles to each of them. This method applies the k-means clustering approach to group similar learners. Then, it infers a set of association rules, related to each learner, and classifies them according to a descending order based on the confidence of each rule. Finally, it recommends to each learner the N -Top articles associated with his preferences. This method helps learners choose the most appropriated articles to construct a deep knowledge.

Authors in [13] seek to generate valuable insights about the appropriation process of the collaborative learning. Their study is based on the adaptive

structuration theory and the linguistic approach of the macro coding level scheme allowing to analyze larger samples to gain insights regarding the appropriation and the structuring activities. Applied to the context of flipped classrooms, this coding scheme has identified nine types of junctures (e.g. dissatisfaction, faithfulness and irony) to categorize the electronic conversations. The junctures categorization identified four models that determine the cases where the appropriation can affect the collaborative learning. These are the conflicts with technology, the domineering group members, the inanimate appropriation and the determined discussions. These models must be considered by designers when designing a collaborative learning environment to ensure a successful learning process.

In the context of MOOCs, authors in [16] propose a model to identify what they called the “leader learners”. The corresponding method is based on the Support Vector Machine as well as the language accommodation measure. It relies on the lexical analysis of the forums posts in order to identify the students by whom the language of the struggling students is influenced. The students whose language influences positively the other students are called “leaders of the struggling students” and will be mobilized to answer their questions on the forums to support their learning process.

Finally, we cite the work of Chaturvedi et al. [5] that proposes a model to identify the situation when the MOOC instructor has to intervene in the forum threads. The purpose is to help students get an answer from the instructor in order to provide them with the effective knowledge they need. This work uses the Chain Markov Model that takes as inputs both the features about the thread and those about each post in the thread. The thread structure and the lexical analysis of the posts are also considered.

The two latter works propose to help only the learners who participate in the forum. However, when following the MOOC, the majority of learners does not participate in the forum. Thus, the method that we propose must consider the learners in their integrality.

### 2.2. Dropout prediction models

The commonly adapted principle when addressing the dropout prediction issue within the MOOCs is to apply one or more machine learning techniques on a set of static and/or dynamic attributes.

Balakrishnan in [1] proposes a model to predict the students’ retention in MOOCs using two kinds of Hidden Markov Model (HMM) techniques; HMM

with a single feature and HMM with multiple features. Prediction is based on the cumulative percentage of the lecture videos watched, the number of threads viewed on the forum, the number of posts made on the forum and the number of times the course progress page was checked. This model aims at predicting the dropout of a student for the following week of the MOOC on the basis of his data for the current week. Experiments showed that the multiple features HMM gives more reasonable results than those provided by the single feature one. In addition, the percentage of the videos watched is the most efficient when using the single feature HMM.

Chaplot et al. [4] proposed a model based on the Neural Network to predict a student's attrition to MOOCs. Other than the classical attributes, such as the number of clicks made by the learner and the weekly number of the forum pages viewed, authors integrated a sentiment score attribute. This is calculated using a lexicon-based sentiment analysis of the forum posts. Authors proved that the analysis of the students' sentiment is an important indicator of their dropout intention. This model permits to estimate whether or not the student will dropout the course in the next week of the MOOC.

Xing et al. [24] propose a temporal modeling approach to predict students who are at risk to dropout the MOOC, using the General Bayesian Network and the Decision Tree. The used features are the number of discussion posts, the number of forum views, the number of quiz views and the degree of social network. Authors showed the importance of using the appended features input and applied the Principle Component Analysis to predict the dropout behavior of students in a chronological order throughout the MOOC weeks.

The authors in [21] developed a multiple linear regression model for predicting the performance of learners in a future homework proposed by the MOOC. The prediction is based on the click number made by the learner. Clicks are categorized according to their types. Thus, they can be subdivided into six categories that are those linked to a session, a video, a quiz, the activities between the last quiz and the last homework and finally, the activities between two consecutive homeworks. Two categories of students were considered: those who have finished all the activities and those who have partially finished it. The experiments showed that considering data from all the previous weeks as inputs is more efficient than considering those from only the last week. In addition, the number of sessions and that of videos and quizzes views are the most important.

The authors in [23] proposed a model to predict the future interactions between pairs based on their

history of interaction on a MOOC forum. The prediction method is based on a directed graph where the nodes are the learners and arches are the interactions between these learners. An arc is labeled with a strength that represents the number of comments added by the learner source to answer the learner recipient. The predicted strength between the pairs is based on the sum of these existing strengths. Four possible categories of friendship were identified: a nonexistent friendship, a persistent friendship, a friendship through another learner and an isolated friendship where the learner has never received answers. The value of this strength should allow to predict whether or not the learner will abandon the MOOC.

In this context of MOOCs, the models based on machine learning techniques are usually faced with the problem of imbalanced data which can degrade the prediction efficiency. In effect, because of the weekly dropout rate that is relatively steep, the data used when training the prediction model are highly imbalanced towards the negative class.

### 3. Dominance-based Rough Set Approach

The approach DRSA is a method of supervised learning. It was proposed by Greco et al. [10] and inspired from the Rough Sets Theory [20]. It allows to compare objects through a dominance relation and takes into account the preferences of a decision maker to extract a preference model resulting in a set of decision rules. According to the DRSA, a data table is a 4-tuple  $S = \langle K, F, V, f \rangle$ , where:

- $K$  is a non-empty finite set of reference objects,
- $F$  is a non-empty finite set of criteria,
- $V_g$  is the domain of the attribute  $g$ .  $V = \bigcap_{g \in F} V_g$ ,
- $f: K \times F \rightarrow V$  is the information function defined such that  $f(x, g) \in V_g$  for each object  $x \in K$  and criterion  $g \in F$ .

$F$  is often divided into a subset  $C \neq \emptyset$  of condition attributes and a subset  $D \neq \emptyset$  of decision attributes such that  $C \cup D = F$  and  $C \cap D = \emptyset$ . In this case,  $S$  is called a decision table (cf. Table 1).

**Table 1. Example of a Decision Table**

	C						D
	$g_1$	$g_2$	..	$g_k$	..	$g_{ F -1}$	$d$
$K_1$	$V_{1,1}$	$V_{1,2}$	..	$V_{1,k}$	..	$V_{1, F -1}$	Clt
$K_2$	$V_{2,1}$	$V_{2,2}$	..	$V_{2,k}$	..	$V_{2, F -1}$	Clt
..	..	..	..	..	..	..	..
$K_{ K }$	$V_{ K ,1}$	$V_{ K ,2}$	..	$V_{ K ,k}$	..	$V_{ K , F -1}$	Clt

In multicriteria decision-making, the scale of condition attributes should be ordered according to a decreasing or an increasing preference of a decision maker. Such attributes are called criteria. We also assume that the decision attribute set  $D = \{d\}$  is a singleton. It partitions  $K$  into a finite number of decision classes  $Cl = \{Cl_t; t \in \{1..n\}\}$  such that each  $x \in K$  belongs to one and only one class. Furthermore, we suppose that the decision classes are preference-ordered, i.e., if  $r > s$ , then objects from  $Cl_r$  are preferred to those from  $Cl_s$ .

In this work, the learners enrolled in the MOOC are the objects and the pedagogical team represents the decision makers in this MOOC. Learners have to be evaluated according to a set of criteria. Then, based on this evaluation, each learner will belong to a decision class. Here, we consider only three preference-ordered decision classes that are:  $Cl_1$  of the “At-risk Learners”,  $Cl_2$  of the “Struggling Learners” and  $Cl_3$  of the “Leader Learners”.

Once the decision table is complete, we have to calculate the dominance relation, the P-dominating set, the P-dominated set and the upward and downward approximations (cf. Table 2). These sets will be used to infer a set of decision rules that permits to automatically assign each new learner to one of the three predefined decision classes.

**Dominance relation.** Let  $P \subseteq C$  be a subset of attributes. The dominance relation  $D_p$  associated with  $P$  is defined for each pair of objects  $x$  and  $y$  thus:  $\forall (x, y) \in K, x D_p y \Leftrightarrow f(x, g_j) \succcurlyeq f(y, g_j) \forall g_j \in P$ . To each object  $x \in K$ , are associated:

- P-dominating set  $D_p^+(x) = \{y \in K: y D_p x\}$  containing objects that dominate  $x$  and
- P-dominated set  $D_p^-(x) = \{y \in K: x D_p y\}$  containing objects dominated by  $x$ .

**Approximating downward and upward class unions.** In DRSA, the represented knowledge is a collection of downward unions  $Cl_t^{\leq}$  and upward unions  $Cl_t^{\geq}$  of classes such that:

$$Cl_t^{\leq} = \bigcup_{s \leq t} Cl_s, Cl_t^{\geq} = \bigcup_{s \geq t} Cl_s; t \in \{1..n\}$$

The assertion “ $x \in Cl_t^{\geq}$ ” means that “ $x$  belongs to at least the class  $Cl_t$ ”, while “ $x \in Cl_t^{\leq}$ ” means that “ $x$  belongs to at most the class  $Cl_t$ ”.

**Example:** we consider three decision classes  $Cl = \{Cl_t; t \in \{1, 2, 3\}\}$ . The downward class unions that we can obtain are  $Cl_1^{\leq} = \{Cl_1\}$ ,  $Cl_2^{\leq} = \{Cl_1, Cl_2\}$  and,  $Cl_3^{\leq} = \{Cl_1, Cl_2, Cl_3\}$ . Moreover, the upward class unions are  $Cl_1^{\geq} = \{Cl_1, Cl_2, Cl_3\}$ ,  $Cl_2^{\geq} = \{Cl_2, Cl_3\}$  and  $Cl_3^{\geq} = \{Cl_3\}$ .

The P-lower and P-upper approximations of  $Cl_t^{\geq}$  with respect to  $P \subseteq C$ , respectively denoted  $\underline{P}(Cl_t^{\geq})$  and  $\overline{P}(Cl_t^{\geq})$  are defined thus:

- $\underline{P}(Cl_t^{\geq}) = \{x \in K: D_p^+(x) \subseteq Cl_t^{\geq}, \forall t \in \{1..n\}\}$

- $\overline{P}(Cl_t^{\geq}) = \{x \in K: D_p^-(x) \cap Cl_t^{\geq} \neq \emptyset, \forall t \in \{1..n\}\}$

The P-lower and P-upper approximations of  $Cl_t^{\leq}$  with respect to  $P \subseteq C$ , respectively denoted  $\underline{P}(Cl_t^{\leq})$  and  $\overline{P}(Cl_t^{\leq})$  are defined:

- $\underline{P}(Cl_t^{\leq}) = \{x \in K: D_p^-(x) \subseteq Cl_t^{\leq}, \forall t \in \{1..n\}\}$
- $\overline{P}(Cl_t^{\leq}) = \{x \in K: D_p^+(x) \cap Cl_t^{\leq} \neq \emptyset, \forall t \in \{1..n\}\}$

The P-lower approximation of  $Cl_t^{\geq}$  (resp.  $Cl_t^{\leq}$ ) contains all objects whose P-dominating (resp. P-dominated) set is assigned with certainty to classes that are at most as good as  $Cl_t$ . The P-upper approximation of  $Cl_t^{\geq}$  (resp.  $Cl_t^{\leq}$ ) contains objects whose P-dominating (resp. P-dominated) set is assigned to a class at least as good as  $Cl_t$ .

The P-boundaries of  $Cl_t^{\geq}$  and  $Cl_t^{\leq}$  are:

- $Bnp(Cl_t^{\geq}) = \overline{P}(Cl_t^{\geq}) - \underline{P}(Cl_t^{\geq})$
- $Bnp(Cl_t^{\leq}) = \overline{P}(Cl_t^{\leq}) - \underline{P}(Cl_t^{\leq})$

The boundaries group objects that can be ruled neither inside nor outside as members of  $Cl_t$ .

**Table 2. summary on concepts**

Concept: Symbol	Meaning
P-dominating set: $D_p^+(x)$	Objects that dominate $x$
P-dominated set: $D_p^-(x)$	Objects that are dominated by $x$
P-lower approximations of $Cl_t^{\geq}$ : $\underline{P}(Cl_t^{\geq})$	Objects whose P-dominating set is assigned with certainty to classes at most as good as $Cl_t$ .
P-upper approximations of $Cl_t^{\geq}$ : $\overline{P}(Cl_t^{\geq})$	Objects whose P-dominating set is assigned to a class at least as good as $Cl_t$ .
P-lower approximation of $Cl_t^{\leq}$ : $\underline{P}(Cl_t^{\leq})$	Objects whose P-dominated set is assigned with certainty to classes at most as good as $Cl_t$ .
P-lower approximation of $Cl_t^{\leq}$ : $\overline{P}(Cl_t^{\leq})$	Objects whose P-dominated set is assigned to a class at least as good as $Cl_t$ .
P-boundaries. $Bnp(Cl_t^{\geq})$ and $Bnp(Cl_t^{\leq})$	P-doubtful region: Objects that are uncertainly classified in $Cl_t$

**Decision rule.** A decision table may look as a set of “if...then...” decision rules, where the condition part specifies values assumed by one or more condition attributes and the decision part specifies an assignment to a decision class. An object  $x \in K$  supports a decision rule if its description matches both the condition and the decision parts of the rule. Decision rules are represented as follows:

If  $f(x, g_1) \leq r_1 \wedge \dots \wedge f(x, g_n) \leq r_n$  then  $x \in Cl_t^{\leq}$ , such that  $(r_1 \dots r_n) \in (V_{g_1} \dots V_{g_n})$

If  $f(x, g_1) \geq r_1 \wedge \dots \wedge f(x, g_n) \geq r_n$  then  $x \in Cl_t^{\geq}$ , such that  $(r_1 \dots r_n) \in (V_{g_1} \dots V_{g_n})$

The approach DRSA proposes an algorithm called DOMLEM that inputs the P-upper and P-lower approximations and outputs a set of decision rules. Each decision rule is characterized by its force, which is the number of objects supporting this rule.

#### 4. Prediction method based-DRSA

The method we propose relies on the supervised learning approach DRSA. It is based on the learners' demographic and dynamic data of the previous week that would help predict their decision class of the following week. It consists of two phases: (i) constructing a preference model, and (ii) using the decision rules to classify new learners.

First, we introduce some new notations. Let  $W = \{W_1, \dots, W_i, \dots, W_t\}$  be the set of weeks making a MOOC such that  $t \geq 2$  is the number of weeks a MOOC holds and  $W_i$  is the  $i^{\text{th}}$  week of the MOOC. We note  $S_i = \langle K_i, F_i, V_i, f_i \rangle$  the information table build at the end of the week  $W_i$  such that  $K_i$  and  $F_i$  are respectively the non-empty finite set of reference objects and the non-empty finite set of criteria selected at the end of the  $i^{\text{th}}$  week of the MOOC to build the information table  $S_i$  and  $f_i: K_i \times F_i \rightarrow V_i$  is the information function.

##### 4.1 Phase 1: Preference model construction

This phase inputs a set of assignment examples and outputs a set of decision rules generalizing the decision makers' preferences. It is made of three steps: the first is to identify assignment examples of learners, the second is to construct a family of criteria for the learners' profiles characterization and the third is to infer a preference model.

**Step 1.1: Definition of a set of "Learners of Reference".** Given the massive number of learners involved in a MOOC, it is difficult to analyze and to evaluate all of them. Hence, it is necessary to define a training sample including an adequate number of representative examples for each decision class; the decision class C11 of the "At-risk Learners", the decision class C12 of the "Struggling Learners" and the decision class C13 of the "Leader learners". In order to comply with the terminology used in the DRSA approach, we call the training examples, "Learners of Reference".

As noted above, the approach DRSA strongly involves the human dimension in the decision making process. However, from a psychological point of view [15], a human decision maker is characterized by a channel capacity that represents the upper limit

on the extent to which he can match his responses to the stimuli we give him. So, to meet the channel capacity of the pedagogical team of the MOOC, we do not focus on the number of learners in the training sample but rather on their quality. Otherwise, a large training sample can degrade the quality of the decisions made by the pedagogical team, a thing which eventually affects the efficiency of the preference model. Hence, our purpose is to build a set of "Learners of Reference" both of a high quality and of a reasonable quantity, in harmony with the pedagogical team's channel capacity, in order to ensure an efficient set of decision rules.

Nonetheless, since during a MOOC the learners can enter or dropout it at any time, the training sample  $K_i$  can not be stable over many weeks. Thus, at the end of each week  $W_i$  of the MOOC, we organize a direct meeting with the pedagogical team to define a new set  $K_i$  of "Learners of Reference".

**Step 1.2: Construction of a family of criteria.** In this step, we use a constructive approach based on a deepened literary review to construct a criteria family that permits to characterize the learners' profiles and behavior within a MOOC. The criteria can be either static or dynamic. Static criteria are provided by the learner when filling the registration form proposed by the platform broadcasting the MOOC. Dynamic criteria are supplied by the tracking tool that manages the MOOC. Both types of criteria provide insights about the richness of the cultural background of the learner, his sharability [2], his absorptive capacity [6], his autonomy [11], etc.

It is important to note that compared to an attribute, a criterion must allow the measuring of the decision maker's preferences according to a personal viewpoint [22]. In other words: criterion = attribute + decision maker's preferences. To this end, direct meetings have to be conducted with the pedagogical team of the concerned MOOC in order to elicit its ordered preferential information for each attribute. For example, the "Study level" is an attribute identified from a literary review (cf. Table 3). After a meeting with the pedagogical team, four increasing ordered scales are defined upon this attribute: 1: Scholar student; 2: High school student; 3: PhD Student; 4: Doctor. At this level, "Study level" is a criterion. In this work, we have retained eight static and four dynamic attributes that are presented in column 1 of Table 3. These attributes should serve as a preliminary list on which the pedagogical team can rely to build a family of criteria that meets its preferences.

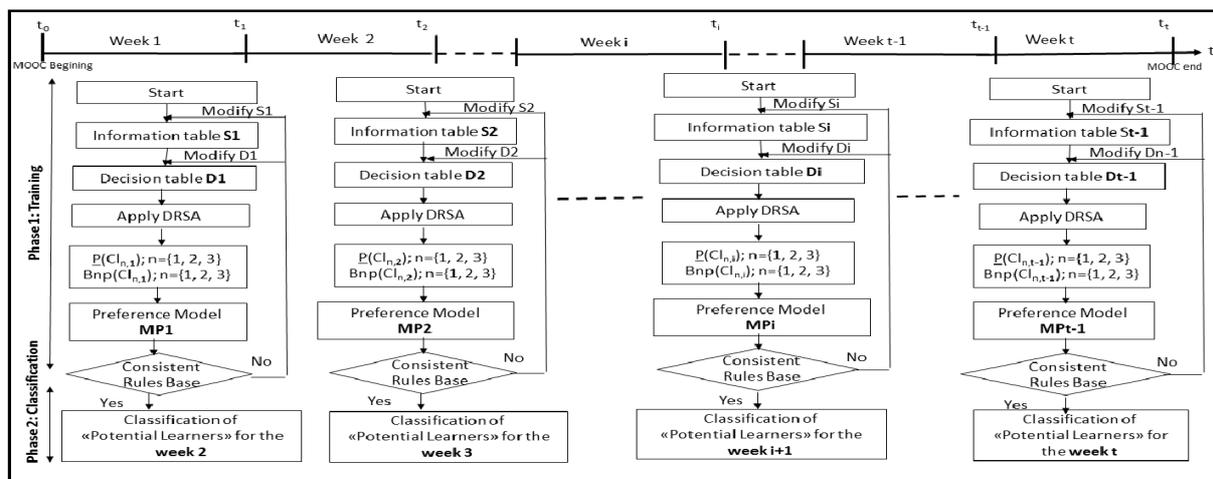


Figure 1: Weekly prediction method based on DRSA

This list must be either validated or updated by the pedagogic team of the MOOC in question each time we apply this method. The construction phase of the criteria family is detailed in [3].

**Step 1.3: inference of the decision rules.** This step is made of three sub-steps: (i) the construction of the information table, (ii) the construction of the decision table and (iii) the inference of a preference model. The information table is a matrix whose rows form the set of the “n” “Learners of Reference” identified in step 1.1, and whose columns represent the “m” evaluation criteria constructed in step 1.2. This matrix is about the evaluation function  $f_i(L_{j,i}, g_{k,i})$  of each learner  $L_{j,i} \in K_i$  on each criterion  $g_{k,i} \in F_i$  such that  $i \in \{1..t\}$ ,  $j \in \{1..n\}$  and  $k \in \{1..m\}$ . Variables t, n and m are respectively the number of weeks a MOOC holds, the size of the “Learners of Reference” set defined in the  $i^{\text{th}}$  week of the MOOC and the size of the criteria family built in the  $i^{\text{th}}$  week of the MOOC. Analogously, the variables  $L_{j,i}$  and  $g_{k,i}$  are respectively the  $j^{\text{th}}$  “Learner of Reference” in the set  $K_i$  and the  $k^{\text{th}}$  criterion in the set  $F_i$ .  $K_i$  and  $F_i$  are respectively the set of “Learners of Reference” and the family of criteria identified in the  $i^{\text{th}}$  week.

Once the information table  $S_i$  is achieved at the end of the  $i^{\text{th}}$  week, we construct the decision table with the pedagogical team during some meetings. It consists in adding a column to the information table, dedicated to the affectation of each “Learner of Reference” in one of the three decision classes:

- C11. The decision class of the “At-risk Learners” corresponding to the learners who are likely to dropout the course in the next week of the MOOC.

- C12. The decision class of the “Struggling Learners” reserved to the learners who have some difficulties but who are still active on the MOOC environment and who do not have the intention to leave it at least in the next week of the MOOC.
- C13. The decision class of the “Leader Learners” who are able to lead a team of learners by providing them with accurate and immediate answer to their questions.

The decision table is thus made of “n” rows and “m+1” columns. The decisions made by the pedagogical team about the classification of each “Learner of Reference” should be based on his/her assessment values on the set of all criteria. We call  $D_i = \{d_{1,i}, d_{2,i}, \dots, d_{j,i}, \dots, d_{n,i}\}$ , the vector of decisions of the affectation of each “Learner of Reference” in one of the three decision classes such that  $d_{j,i} \in \{C11, C12, C13\}$  is the classification of the  $j^{\text{th}}$  “Learner of Reference”  $L_{j,i} \in K_i$  in the one the three decision classes C11, C12 or C13.

Once the decision table of the  $i^{\text{th}}$  week  $W_i$  of the MOOC is complete, it will be provided as an input to the algorithm DOMLEM proposed by the DRSA approach. This algorithm outputs a preference model resulting in a set of decision rules. The preference model aims to classify learners at the beginning of the week  $W_{i+1}$  of the same MOOC.

This method runs weekly: the first phase runs at the end of each week  $W_i$  of the MOOC such that  $i \in \{1..t-1\}$  while the second phase runs at the beginning of each week  $W_{i+1}$  of the same MOOC, such that  $i \in \{2..t\}$  and t is the number of weeks that a MOOC holds. The second phase inputs the output of the first one (cf. Figure 1).

**Table 3. The coherent family of criteria**

<i>Criterion</i>	<i>Description</i>	<i>Scale</i>	<i>P</i>
$g_1$ : Study level	Indicates the actual study level of the learner or the last diploma he obtained	1: Scholar student; 2: High school student; 3: PhD Student; 4: Doctor	↑
$g_2$ : Level of technical skills	Indicates the extent to which the learner masters the use of the computer tools	1: Basic; 2: Average; 3: Expert	↑
$g_3$ : Level of proficiency in MOOC language	Indicates the extent to which the learner masters the language of the MOOC	1: Basic; 2: Average; 3: Good	↑
$g_4$ : Motivation for MOOC registration	Indicates the motivation behind the participation of the learner in the MOOC	1: To discover the MOOCs; 2: To exchange ideas or to have a certificate ; 3: To exchange ideas and to have a certificate	↑
$g_5$ : Previous experience with MOOCs	Indicates whether the learner has a previous experience on learning via MOOCs or not	0: No experience at all; 1: At least one experience	↑
$g_6$ : Mastery level of the subject of a MOOC	Indicates to which extent the learner masters both the topic and the theme of the MOOC	0 : No knowledge at all; 1: Average knowledge; 2: Deepened knowledge	↑
$g_7$ : Probability to finish the MOOC	Indicates the probability for a learner to carry-on the MOOC activities until the end	1: Very weak; 2: Weak; 3: Average; 4: Strong; 5: Very strong	↑
$g_8$ : Weekly availability	Indicates the estimative weekly availability of the learner to follow the MOOC	1: Less than 1 hour; 2: From 1 to 2 hours; 3: From 2 to 3 hours; 4: Four hours or more	↑
$g_9$ : Weekly number of forum posts	Indicates the number of responses or information added on the forum per week	1: $n=0$ ; 2: $n \in \{1, 2\}$ ; 3: $n \in \{3, 4\}$ ; 4: $n \geq 5$	↑
$g_{10}$ : Weekly number of forum questions	Indicates the number of questions asked on the forum per week	1: $n=0$ ; 2: $n \in \{1, 2\}$ ; 3: $n \in \{3, 4\}$ ; 4: $n \geq 5$	↓
$g_{11}$ : Weekly number of viewed resources	Indicates the weekly number of the viewed and / or downloaded resources and material courses	1: $n < 10$ ; 2: $10 \leq n < 20$ ; 3: $20 \leq n < 30$ ; 4: $30 \leq n < 40$ ; 5: $n \geq 40$	↑
$g_{12}$ : Weekly score	Indicates the weekly score the learner got on the set of activities he made	1: $0 \leq \text{Score} < 6$ ; 2: $6 \leq \text{Score} < 7$ ; 3: $7 \leq \text{Score} < 8$ ; 4: $8 \leq \text{Score} \leq 10$	↑

The purpose of this method is to identify the “Leader Learners” who will be mobilized to support the “At-risk Learners” and the “Struggling Learners” throughout their learning process. This support will decrease the dropout rate and help learners more appropriate the knowledge transmitted to them. Our aim is to improve the knowledge appropriation by each learner who interacts in a learning environment.

## 5. Case study

This section provides at the beginning a brief description of the MOOC used to validate the proposed method then presents the weekly application of the two phases before discussing ultimately the obtained results.

### 5.1. Application field

The application field is a French MOOC offered by a Business School in France and broadcasted on a French platform. For reasons of anonymity, we were discreet about its name. The MOOC started with 2565 learners and lasted  $t = 5$  weeks. It required a weekly availability going to three hours and did not necessitate any prior knowledge. The first, the second and the fourth weeks ended with a quiz while the

third and the fifth were closed with a peer-to-peer (P2P) assessment. Data were saved in a CSV (Comma-Separated Values) file. However, only data about 1535 learners are used in these experiments. Learners who have been omitted from these results are those who have not completed the registration form. To obtain results, we have developed a decision support prototype using the JAVA language.

### 5.2. Method application

Phase 1: Construction of a preference model. At the end of each week  $W_i$  such that  $i \in \{1..4\}$ : First, the pedagogical team selected a sample  $K_i$  of  $n = 30$  representative examples of learners for each decision classes:  $C11$  of the “At-risk Learners”,  $C12$  of the “Struggling Learners” and  $C13$  of the “Leader Learners”. Second, with the pedagogical team we constructed the family of  $m = 12$  criteria that, in this case, remained stable over weeks (cf. Table 3). Third, we constructed the information table  $S_i$  and determined with the pedagogical team the decision vector  $D_i$  that classifies each learner in  $K_i$  in one of the three decision classes in question. An extract of the decision tables built at the end of each week  $W_i$  is shown in Table 4. Finally, we applied the algorithm DOMLEM and inferred a set of decision rules.

**Table 4. Extract from the decision tables**

Week	L_id	g <sub>1</sub>	g <sub>2</sub>	g <sub>3</sub>	...	g <sub>10</sub>	g <sub>11</sub>	g <sub>12</sub>	D
W <sub>1</sub>	14011	2	3	3	...	1	1	1	C11 <sup>≤</sup>
	43389	2	3	3	...	1	4	4	C12 <sup>≥</sup>
W <sub>2</sub>	36063	1	3	3	...	1	3	4	C12 <sup>≤</sup>
	36364	4	1	3	...	1	5	4	C13 <sup>≥</sup>
W <sub>3</sub>	18182	2	1	3	...	1	3	1	C11 <sup>≤</sup>
	36097	2	2	3	...	1	3	4	C13 <sup>≥</sup>
W <sub>4</sub>	35957	2	2	3	...	1	2	3	C11 <sup>≤</sup>
	36097	2	2	3	...	1	3	4	C13 <sup>≥</sup>

Table 5 shows an extract from the obtained decision rules over weeks. For example, the second rule of the first week, noted rule 1.2, can be translated as follows: If the learner’s motivation to participate in the MOOC is “at most” to discover the MOOC concept and the score that he obtained at the end of week W<sub>1</sub> is “at most” 6, then the learner is at-risk to dropout the course in week W<sub>2</sub> of the MOOC.

**Table 5. Extract from the decision rules**

	Rule_id	Rule	Force
W <sub>1</sub>	Rule <sub>1,1</sub>	<b>If</b> $f(L_{i,1}, g_4) \leq 1 \wedge f(L_{i,1}, g_{12}) \leq 1$ <b>then</b> $L_{i,1} \in C1_1^{\leq}$	28%
	Rule <sub>2,1</sub>	<b>If</b> $f(L_{i,1}, g_9) \geq 2 \wedge f(L_{i,1}, g_2) \geq 3$ $\wedge f(L_{i,1}, g_7) \geq 5$ <b>then</b> $L_{i,1} \in C1_3^{\geq}$	52%
W <sub>2</sub>	Rule <sub>1,2</sub>	<b>If</b> $f(L_{i,2}, g_9) \leq 1$ <b>then</b> $L_{i,2} \in C1_2^{\leq}$	79%
	Rule <sub>2,2</sub>	<b>If</b> $f(L_{i,2}, g_{11}) \geq 5 \wedge f(L_{i,2}, g_8) \geq 3$ $\wedge f(L_{i,2}, g_2) \geq 2$ <b>then</b> $L_{i,2} \in C1_3^{\geq}$	33%
W <sub>3</sub>	Rule <sub>1,3</sub>	<b>If</b> $f(L_{i,3}, g_{11}) \leq 1 \wedge f(L_{i,3}, g_6) \leq 0$ <b>Then</b> $L_{i,3} \in C1_1^{\leq}$	40%
	Rule <sub>2,3</sub>	<b>If</b> $f(L_{i,3}, g_6) \geq 2 \wedge f(L_{i,3}, g_7) \geq 4 \wedge$ $f(L_{i,3}, g_{12}) \geq 2$ <b>then</b> $L_{i,3} \in C1_3^{\geq}$	20%
W <sub>4</sub>	Rule <sub>1,4</sub>	<b>If</b> $f(L_{i,4}, g_9) \leq 1$ <b>then</b> $L_{i,4} \in C1_2^{\leq}$	85%
	Rule <sub>2,4</sub>	<b>If</b> $f(L_{i,4}, g_9) \geq 3 \wedge f(L_{i,4}, g_5) \geq 1 \wedge$ $f(L_{i,4}, g_{11}) \geq 3$ <b>then</b> $L_{i,4} \in C1_3^{\geq}$	30%

Phase 2: At the beginning of each week W<sub>i</sub> such that  $i \in \{2..5\}$ , we applied the previously inferred decision rules to classify each potential learner in one of the three decision classes.

### 5.3. Results and discussion

The DRSA requires that the rules have the form of “if condition (s), then decision” which allows the decision maker to understand the reason for his decision in a natural language. Compared to the machine learning techniques, the DRSA allows the intervention of human decision makers for decision making which gives a more sophisticated quality to the classification. In this work, experiments showed that the obtained decision rules were strong (the force

reaches 85%). Otherwise, to measure the performance of the preference model we calculated the precision that reflects the number of learners correctly predicted by the preference model; the recall that reflects the number of correctly predicted learners related to the positive examples and the F-measure that represents the harmonic average of precision and recall (cf. Figure 2). We note that the model has an F-measure that is generally satisfying.

- Week 1-2: In this curve we note that the F-measure rate was low for both classes C11 and C13 because of the lurkers. These are the curious learners who participate just to discover the MOOC and who dropout it in the second assessment. At this level, their activities can not reflect their intention, a thing which degrades the quality of the precision measure.
- Week 2-3: Usually, the number of lurkers decreases noticeably just after the first assessment. This makes it easier to predict the “At-risk Learners”. This explains the high F-measure rate of the decision class C11. However, the F-measure rate of C13 remains poor. In fact, in this MOOC the second week was concluded by a quiz while at the third week a P2P activity was proposed. Obviously, a P2P activity needs more time and deeper skills than a simple quiz. Thus, the learner who is classified as leader, based on his assessment on the quiz, may not be as such if we consider the P2P assessment.
- Week 3-4: The results based on the data coming from the third week are generally satisfactory. In the third week a P2P activity was proposed. So, if the learners submitted their works during this week that means that they will remain engaged in the following one, a thing which explains the high rate of the F-measure of C11. Similarly, the decisions are made on the basis of their scores obtained on the peer-to-peer assessment. This type of assessment makes the learner evaluation more sophisticated, which explains the satisfactory rate of the C13 F-measure.
- Week 4-5: In this week, we share the same situation as the week 2-3. In fact, the learners are increasingly engaged and that explains the good quality of the C11 F-measure.

Finally, it is noteworthy to say that compared to the traditional learning models, this method based-DRSA has the advantage of enhancing the quality of the training set despite the mobility of its objects. It also overcomes the imbalanced data issue because of the human intervention in the decision making process as well as in the choice of the training set.

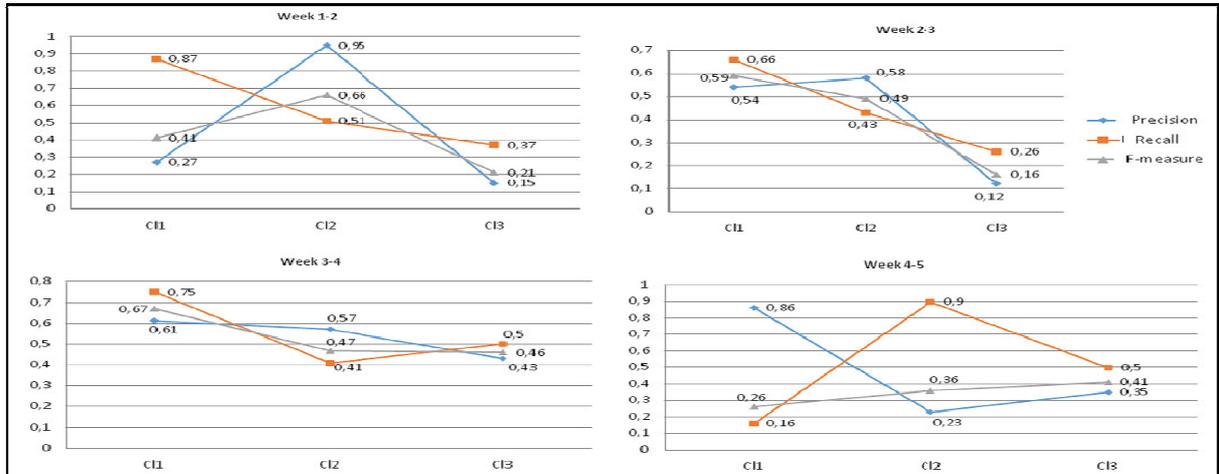


Figure 2. Measures of the preference model performance over the MOOC weeks

Figure 3 shows the dropout rates over weeks throughout the broadcasting the MOOC. We note that the overall dropout rate, as usual, exceeded 90%. The highest rate was registered in the second week because of the presence of lurkers. We notice as well that the number of participants increases in the fourth week of the MOOC concluded by only a quiz and decreases in the third and fifth ones ending with a peer-to-peer evaluations. This requires more time and more technological skills. As said in section 2, these two factors incite the learner to dropout the MOOC.

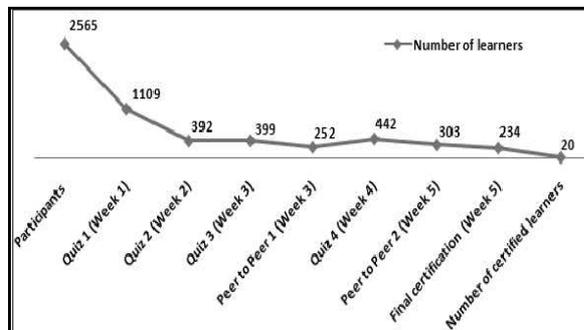


Figure3. Number of participants over weeks

## 6. Conclusion

In this paper, we have proposed a method based on the Dominance- based Rough Set Approach for the weekly prediction of the “At-risk Learners”, the “Struggling Learners” and the “Leader Learners” during a MOOC. It consists of two phases; first, the construction of a preference model resulting in a set of decision rules; second, the classification of the

MOOC learners based on this preference model. The second phase has to be applied periodically at the beginning of the current week of the MOOC on the basis of data provided during the first phase at the end of the previous week during this same MOOC. These data are made of both the static and the dynamic ones. The purpose is to identify the decision class to which each learner belongs: the decision class CI1 of the “At-risk Learners”, the decision class CI2 of the “Struggling Learners” or the decision class CI3 of the “Leader Learners”.

This method has two objectives: (i) minimizing the dropout rate through the early identification of the “At-risk Learners”, and so helping them carry on the MOOC; and (ii) improving the individual appropriation of the exchanged knowledge by the identification of the “Leader Learners” who will be mobilized to support the other learners throughout their training.

Thus the highlight of the proposed method is that it combines four issues at a time: predicting the At-risk learners; overcoming the issue of the absence of the tutor who will be replaced by the leaders; enhancing the individual appropriation process and finally minimizing the dropout rate by guiding the learners in difficulties to the appropriate knowledge.

This method has been validated on real data provided to us by a French MOOC proposed by a Business School in France and broadcasted on a French platform. It can be experienced either on similar MOOCs, that is to say MOOCs with the same pedagogical team and the same subject, or on different MOOCs. In the second case, we must mobilize the new pedagogical team in order to adapt the criteria family to it.

Our future work will focus on two points: (i) in this work we studied the case when the set of “Learners of Reference” and that of criteria are stable over the MOOC weeks. However, since the learners within a MOOC can enter or dropout it at any time while it is running, the set of “Learners of Reference” must also evolve over the MOOC period. Moreover, the criteria family can change since the preferences of the pedagogical team may vary because the MOOC characteristics keep evolving all the time (content complexity, technological skills needed, etc.). Thus, to take into account the dynamic aspect of the MOOC, we are proposing a prediction model that implements an incremental approach based on DRSA; and (ii) in this work, the prediction concerns only the following week during the MOOC. However, it is more interesting to capitalize in one MOOC in order to predict the exact week during which the learner will be at-risk of dropping out in similar MOOCs. To achieve this end, we plan to test this method on similar MOOCs in order to make a long-term prediction.

## 7. References

- [1] G. Balakrishnan, “Predicting student retention in massive open online courses using hidden markov models”, Tech. Rep. No. UCB/EECS-2013-109, University of California at Berkeley, 2014
- [2] I. Boughzala, and R.O. Briggs, “A value frequency model of knowledge sharing: an exploratory study on knowledge sharability in cross organizational collaboration”, *Electronic Markets*, 2012, Vol. 22, pp. 9-19
- [3] S. Bouzayane, and I. Saad, “Multicriteria Decision Aid Approach to Referent Learners Identification within a MOOC”, *Proceedings of the 4th Conference on European MOOCs Stakeholders Summit (EMOOCs 2016)*, Graz, Austria, 22- 24 February, 2016, pp. 237-247.
- [4] D. Chaplot, E. Rhim, and J. Kim, “Predicting student attrition in moocs using sentiment analysis and neural networks”, *Proceedings of AIED 2015 fourth workshop on intelligent support for learning in groups*, 2015
- [5] S. Chaturvedi, D. Goldwasser, and H. Daume II, “Predicting instructor’s intervention in mooc forums”, *Proceedings of the 52nd annual meeting of the association for computational linguistics (ACL)*, 2014, pp. 1501-1511.
- [6] W. Cohen, and D. Levinthal, “Absorptive capacity: A new perspective on learning and innovation”, *Administrative Science Quarterly*, 1990, Vol. 35, pp. 128-152
- [7] T.H. Davenport and I. Prusak, “Working knowledge: How organizations manage what they know”, Harvard Business School Press, Boston, 1998
- [8] S. Downes, “What makes a MOOC massive”, Blog post, 2013.
- [9] Educause. “7 things you should know about moocs II”, Tech. Rep, Educause Learning Initiative, 2013, (<https://library.educause.edu/media/files/library/2013/6/eli7097-pdf.pdf>)
- [10] S. Greco, B. Matarazzo, and R. Slowinski, “Rough sets theory for multicriteria decision analysis”, *European Journal of Operational Research*, 2001, pp.1-47.
- [11] J.R. Hackman, and G.R. Oldham, “Work Redesign”, United States of America: Addison-Wesly, 1980
- [12] K.S. Hone, and G.R. El Said, “Exploring the factors affecting MOOC retention”, A survey study. *Computers & Education*, 2016, Vol. 98, pp.157–168, (doi:10.1016/j.compedu.2016.03.016)
- [13] A. Janson, M. Söllner, and J.M. Leimeister, “The Appropriation of Collaborative Learning - Qualitative Insights from a Flipped Classroom”, *HICSS*, 2016, pp. 84-93
- [14] C.H. Liou, “Personalized Article Recommendation Based on Student’s Rating Mechanism in an Online Discussion Forum”, *HICSS*, 2016, pp. 60-65
- [15] G.A. Miller, “The magical number seven, plus or minus two: Some limits on our capacity for processing information”, *Psychological Review*, 1956, Vol.63, pp.81-97
- [16] S. Moon, S. Potdar, and L. Martin, “Identifying student leaders from MOOC discussion forums”, *EMNLP*, 2014, pp. 15–20
- [17] S.M. North, R. Ronny, and M. Max, “To adapt MOOCs, or not? that is no longer the question”, *Universal Journal of Educational Research*, 2014, Vol 2, pp. 69–72
- [18] D. F. O. Onah, J. Sinclair, and R. Boyatt, “Dropout rates of massive open online courses: behavioral patterns”, *proceeding of the 6th international conference on education and new learning technologies*, 2014, pp. 5825–5834
- [19] M. Patru and V. Balaji, “Making sense of MOOCs: a guide for policy makers in developing countries”, Report, UNESCO publication, p.100, 2016. Available at: <http://unesdoc.unesco.org/images/0024/002451/245122E.pdf>
- [20] Z. Pawlak, “Rough sets”, *International Journal & Computer Sciences*, 1992, pp. 341–356
- [21] Z. Ren, H. Rangwala and A. Johri, “Predicting Performance on MOOC Assessments using Multi-Regression Models”, *arXiv preprint arXiv:1605.02269*, 2016
- [22] B. Roy and V. Mousseau, “A theoretical framework for analysing the notion of relative importance of criteria”, *Journal of Multi-Criteria Decision Analysis*, 1996, Vol.5, pp.145-159
- [23] A. S. Sunar, N. Abdullah, S.White, and H.C. Davis, “Analysing and predicting recurrent interactions among learners during online discussions in a mooc”, *Proceedings of the 11th International Conference on Knowledge Management*, 2015
- [24] W. Xing, X. Chen, J. Stein, and M. Marcinkowski, “Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization”, *Computers in Human Behavior*, 2016, Vol.58, pp.119–129
- [25] D. Yang, T. Sinha, D. Adamson, and C.P. Rose, “Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses”, *Nips workshop on data driven education*, 2013
- [26] L. Yuan, and S. Powell, “MOOCs and disruptive innovation: Implications for higher education”, *e-learning papers, in-depth*, 2013, Vol. 33.