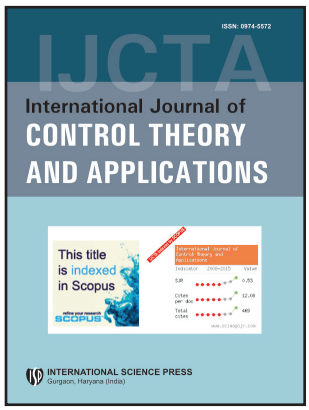


International Journal of Control Theory and Applications

ISSN : 0974-5572



Volume 9 • Number 44 • 2016

# Identifying Disengaged Learners using Educational Data Mining Techniques

**A.V. Senthil Kumar**[a]

[a]*Director of MCA, Hindusthan College of Arts and Science, Coimbatore*

***Abstract:*** On-line learning is a new emerging technology which is dynamic and potentially enriching forms of learning but attrition remains a serious problem. Motivation towards learning is affected by the learner's self-efficacy, locus of control, goal orientation and perceived task difficulty. In a traditional classroom environment, tutors infer learners' levels of motivation from several cues, including speech, behaviour, attendance, body language or feedback, and offer interventional strategies aimed at increasing motivation. Similarly, Online learning system also needs an ability to recognize when the learner is becoming de-motivated and to intervene with effective motivational strategies. Being able to automatically detect disengaged learners would offer the opportunity to make online learning more efficient, enabling tutors and systems to target disengaged learners, to reengage them, and thus, to reduce attrition. Analysing data from log-file is an efficient method for automatic analysis, whereas it has certain level of fuzziness in order to retrieve desired information in robust fashion. Generally, the log files have more information about the learner's attitude and it does not include the results of the assessments they attend. The main purpose of extracting the disengaged students is to prevent from disengagement. To do so, the log file analysis alone could not have enough data to support our aim. Thus integration of log file information with database gives meaningful insights. Thus our proposed methodology predicts the disengagement based on the learner's attitude and Assessment performance in connection with time spent on learning based on the regional index.

***Index Terms:*** Educational Data Mining, Online learning, Log File Analysis, Identifying Unmotivated learners, Region based Classification, Quasi Framework.

## 1. INTRODUCTION

Educational Data Mining (EDM) is a field that exploits Statistical, Machine-learning, and Data-mining algorithms over the different types of educational data. Its main objective is to analyse these types of data in order to resolve educational research issues. EDM has been considered as fast grown independent research area in recent years. Educational Data mining can be implemented in many techniques such as decision trees, neural networks, *k*-nearest Neighbour, Naive Bayes, support vector machines and many others. using these methods many kind of knowledge can be discovered such as association rules, classification, clustering, pruning the data.

EDM can handles both offline and online learning research issues. Here traditional classroom learning is considered as offline learning. In an offline learning, EDM is used to analyse the student's behaviour and predicts the unmotivated students. This helps the instructor to identifying the drop outs and students who need special attention and allow the teacher to provide appropriate counselling / Advising. In Addition to this, accurately predicting student performance is useful in many different contexts like identifying exceptional students for scholarships is an essential part of the admissions process and identifying weak students who are likely to fail is also important for allocating limited tutoring resources. In an online learning, analysing the learning behaviour is consider to be an important and difficult task, because the learner and instructor are in different end. Hence there is need for automatic mechanism to record and analyse the learning behaviour of the learners.

The learner's actions preserved in log files have been recently discovered as a valuable source of information and several approaches to motivation detection and intervention have used log-file analysis. An important advantage of log-file analysis over self-assessment approaches is the unobtrusiveness of the assessment process, similar to the classroom situation where a teacher observes that a learner is not motivated without interrupting his/her activities.

Time spent attribute is considered to be a key value on identifying the disengaged learners in an online learning, whereas marks secured attribute based on number of correct answers and no of wrong answers will be considered as traditional method to identifying the disengaged learners in an offline learning. Time spent and marks secured attributes alone is not enough to identify the disengaged learners but finding the disengagement based on the both attributes will give better prediction results. In addition to this, still there is need to classify the student log data for better prediction of disengaged learners. Generally, the log files have more information about the learner's attitude and in rare cases it includes results, demographic information and student's personal information etc., The main purpose of extracting the disengaged students is to prevent from disengagement. To do so, the log file analysis alone could not have enough data to support our aim. Thus integration of log file information with database gives meaningful insights. Demographic profile of students and their academic performances are key information about the students, which is not preferred to store it in log file. Thus our proposed methodology quasi framework predicts the disengagement based on the learner's attitude and Assessment performance in connection with time spent on learning based on the regional index.

The rest of the work is structured as follows: Section-2 provides related works on predicting the student behaviour and factors which are used to classify the students log data. Section-3 discuss about the methodology of proposed work. Section-4 describes the experimental results and finally in Section 5 conclusions are outlined.

## 2. RELATED WORKS

Although Educational data mining is a recent research field, there are many works are already done in this area. That is because of its potential to educational institutes. [1][2][3] used educational data mining to analyse students learning behaviour of disengaged students and to warn students at risk before they entering into final exam. Al-Radaideh [4] study helps instructor on identifying the dropouts and students who need special attention and allow the teacher to provide appropriate advising. Shaeela [5] concluded that students grade in senior secondary exam, living location, medium of teaching, mother's qualification, student's other habits, family annual income and student's family status were highly correlated with the student academic performance. Bharadwaj [6] used simple linear regression analysis and it was found that the factors like mother's education and student's family income were highly correlated with the student academic performance. Hijazi [7] conducted study on the student performance using association rule technique and find the interestingness of student in opting class teaching language. Praveen [8] used naive Bayes algorithm to predict students' academic performance and the Bayes model help lecturers to indicate weak students for special care.

David [9] clearly indicates that rural and urban students are different in their learning styles and it suggest that students in rural schools appear to be more concerned and engaged in the educational process than urban students. Barcinas [10] indicates that students from the two areas are quite different in ethnicity. The rural students appear to be quite homogeneous, however the urban students seemed to have a greater mix of race and cultures. The lack of opportunity for rural students to interact with persons of varying backgrounds may be a limiting factor in their educational and sociological development and the educational level of the parents was higher in urban areas than in rural areas. Urban parents were more likely to expect their children to advance their education beyond high school. All these factors shows that the difference in their social context between rural and urban areas. These differences help to explain the aspirations of students and finally the work suggests that students from rural areas should learn to live and work in an urban area. Rekha et. al., [11] study suggests that these facilities of school and area of residence influence students' academic self-concept and academic achievement. Praveen et. al., [12] has conducted a survey on Regional Bench mark of rural and urban students and concluded that the students who have completed their past studies in Uniform area (Either Rural/Urban) has no significant difference but the students who completed in Mixed Area has found a significant difference. The Results of [13] show that education performance of rural children and migrants' children is significantly lower than that of their urban counterparts. Praveen et. al., [14] states that region wise classification of online learners will identify the disengaged learners more accurately. The review of literature concludes that, there are so many factors which leads to predict the student academic performance and classification of students based on their living area also helps better prediction on disengaged learners.

## 3. DISENGAGEMENT DETECTION METHODOLOGY

The proposed work identifies the disengaged learners using log file analysis. In a log file analysis, Time spent is considered to be an important factor of detecting the disengaged learners. According to claypool[15] examines various attributes in their analysis and concludes that time spent on reading a page is an important indicator for finding disengagement behaviour of a learner. Gowda [16] states that the fast moving on pages, as well as students who spend long pauses are more likely to learn just shallowly. They did not gain the knowledge of the learning material and most probably they seem to be disengaged. The learner's engagement can be identified based on the average session duration and time on task percentage [17]. According to Monterrat[18] engagement is determined based on the two metrics, i.e. too short time to read texts and to answer questions or taking too long time read or answering questions. According to Oskouei[19] analysed the behaviour of the student in terms of average time spent in online and category of visited websites by them along with their academic performance.

The researches [20], [21], [22], [23] and [24] detects the disengaged learners based on their learning attitude. According to their dataset, they have maintained some fixed minimum and maximum threshold values. Each and every log file sequences are monitored and if the timespent value of the log sequence is beyond those threshold values means then the log sequence is assigned as the engaged, or else it is considered as the disengaged sequence. The 2/3 value of the engaged status assigned in the log sequences are considered as the overall status of the learner. The researches [14], [25] states that the learning alone cannot be enough to detect the disengagement. They integrate the time spent attribute with marks secured attribute for better prediction results. As discussed in review of literature, there are so many key attributes are used to predict the student's log data. Through this work, we classify the student log data based on their area of residence and checks whether such classification make any impact in student's log data. The rest of the paper is organized by redefining the threshold values based on the region wise classification of our dataset.

### 3.1. Redefining Threshold Values

The time interval for reading and number of pages in each interval for our dataset is presented in Table 1.

Table 1 indicates that most of the pages require less than 240 seconds to read a page, similarly 6182 which means less than 1% of pages requires more than 720 seconds. Hence we assign the range for finding minimum threshold is less than 240 seconds and maximum threshold is greater than 720.

**Table 1**
**Time Interval for number of pages in read in each Interval**

| *Time Interval* | *Rural* | *Urban* | *Total No of page's read* |
|---|---|---|---|
| < 240 seconds | 192070 | 284794 | 4,76,864 |
| >240 and <480 seconds | 115178 | 93202 | 2,08,380 |
| >480 and <720 Seconds | 11826 | 9728 | 21,554 |
| >720 Seconds | 4966 | 1216 | 6182 |

The formula for finding Minimum threshold value is,

$$\sum_{i=m}^{n} a_i = a_m + a_{m+1} + a_{m+2}, + a_{m+3} + \ldots a_{n-1} + a_n. \tag{1}$$

$$\mu = \frac{1}{n} \sum_{i=1}^{n} a_i \tag{2}$$

For minimum Threshold, where $i = 1$, $n$ = Total number of pages read on the given threshold value, and $a$ = total time spent for the given minimum threshold value.

Similarly, for maximum threshold, where $i = 720$, $n$ = Total number of pages read on the given threshold value and a is the total timespent of a page on given maximum threshold values.

$$\mu_1(r) = \frac{1}{192070} \times 2796540 = 14.56$$

$$\mu_2(r) = \frac{1}{4966} \times 4409792 = 888$$

$$\mu_1(u) = \frac{1}{284794} \times 4871434 = 17.105$$

$$\mu_2(u) = \frac{1}{1216} \times 934241 = 768.29$$

Exact Pages read = Total no of pages read – (Total no of pages above threshold
+ Total no of pages below threshold) (3)

## 3.2. Enhanced Disengagement Detection with Region Wise Classification Algorithm (EDDA-R)

**Enhanced Disengagement Detection Algorithm with Regionwise Classification (EDDA-R)**

Initialize log file sequences $lf_1$, $lf_2$, $lf_3$, …, $lf_n$

**Output:** Preprocessed log file with engagement status

**Step 1:** Begin

**Step 2:** Group the student's log data based on Region wise.

**Step 3:** For each item in Region wise

**Step 4:** Calculate $\mu_1$, $\mu_2$ using Equ (1) and (2)

**Step 5:** For each sequences in grouped log file $lf_i$ do

**Step 6:** Assign Status = 'Disengaged'

**Step 7:** If timespent < $\mu_1$ then

**Step 8:** Goto Step 5

**Step 9:** Else if timespent > $\mu_2$ and NoS = 0 and NoMC = 0 then

**Step 10:** Goto Step 5

**Step 11:** Else assign Status = 'Engaged'

**Step 12:** End If

**Step 13:** End For

**Step 14:** For each item in Student Database do

**Step 15:** Calculate EPR using Equ (3)

**Step 16:** If ((EPR < (2/3 × NoP)) and (Noc >= NoQ/2))

**Step 17:** Assign Eng_status = 'Disengaged'

**Step 18:** Else Assign Eng_status = 'Engaged'

**Step 19:** End If

**Step 20:** End For

**Step 21:** End

Enhanced Disengagement Detection with Region Wise Classification Algorithm (EDDA-R) is used to construct and predict the disengagement based on new threshold values on learning and marks they scored in assessment. EDDA-R Algorithm first groups the student log data based on the region wise. Then calculate Minimum threshold ($\mu_1$), Maximum Threshold $\mu_2$ for both regions. Through log file analysis, each and every learning sequences is monitored. If the learning sequences is less than minimum threshold value, then the sequence is assigned as disengaged. Similarly, if the sequence is greater than the maximum threshold means then it has to check further condition that, whether there are any activities happened on those time spent (which includes mouse activities). If any activities happen on those sequences means then the system will have considered that sequence as slow learner and assigns that sequence as Engaged, or else the sequence is assigned as Disengaged. Once the learning sequences are monitored, the system has to find the Exact pages read by the learner and then the system checks if 2/3 of the total number of pages read is greater than the Exact Pages Read (EPR) and the learner has to get at least 50 % of Correct answers in their assessment means then the system will assign the Overall status of the learner as Engaged learner, otherwise the learner is considered as a disengaged learner.

## 4. EXPERIMENTAL RESULTS

In order to validate our approach, we have collected the log files of 247 users from an online learning system namely Quasi framework [26], where each learner has spent minimum of ten sessions for learning and ten sessions for exam activities. The proper login and logout is considered as session.

**Table 2**
**Attributes used for analysis**

| *Code* | *Attributes Description* |
|---|---|
| NoP | No of Pages Read |
| AvgTL | Average Time Spent for Learning |
| NoQ | Number of Questions Attended |
| AvgTQ | Average time spent on Assessment |
| NoC | Number of Correct Answers |
| NoW | Number of Wrong Answers |
| NoMC | Number of Mouse clicks used |
| NoS | Scrolls wheels used |

Totally 7,90,859 instances have been obtained. Out of those instances, 7,12,980 instances are identified as learning instances and 49,623 instances is identified as assessment instances and other activities like feedback, glossary, getting help has occurred 28,256 instances. Totally 33 Attributes are derived from logged events. The list of logged events is presented in [22]. The Hybrid PSO with Naïve Bayes classifier is used for feature selection [23]. After feature selection process, the selected attributes used for this analysis are listed in Table 2.

**Table 3**
**Experimental Results**

| *Performance Measures* | *F1* | *F2* |
|---|---|---|
| %correct | 93.33 | 94.74 |
| TP Rate | 0.926 | 0.953 |
| FP Rate | 0.045 | 0.058 |
| Precision | 0.962 | 0.945 |
| Error | 0.061 | 0.052 |
| Sensitivity | 0.962 | 0.945 |
| Specificity | 0.915 | 0.950 |
| F1 Score | 0.943 | 0.950 |
| Matthews Correlation Coefficient | 0.879 | 0.895 |

The Table 3 depicts the results with respect to the method of prediction as F1 and F2. The F1 dataset contains disengagement information of the learners based on Enhanced Disengagement Detection Algorithm, where minimum and maximum threshold values are calculated based on overall learners. The F2 dataset contains disengagement information based on Enhanced Disengagement Detection with respect to the region wise Classification, where separate minimum and maximum threshold values are calculated based on region wise classification.

**Table 4**
**Region wise Threshold values**

| | *Minimum Threshold* | *Maximum Threshold* |
|---|---|---|
| Rural | 14.56 | 888 |
| Urban | 17.105 | 768.29 |
| Overall | 16.08 | 864.45 |

The Table 4 shows the threshold values of learners based on EDDA and EDDA-R. The Table-4 clearly states that there is a huge variation between the region wise learners. While combining those values, it reduces the overall performance of the dataset.

The Confusion matrix of F1 and F2 dataset is listed in Table 5(a) and Table 5(b). Table 5(a) represents the confusion matrix of F1 dataset and Table 5(b) represents the confusion matrix of F2 dataset.

**Table 5**
**(a) Confusion matrix of F1 Dataset**

| *Engagement Status* | *Disengaged* | *Engaged* |
|---|---|---|
| Disengaged | 125 | 10 |
| Engaged | 5 | 107 |

**Table 5**
**(b) Confusion matrix of F2 Dataset**

| *Engagement Status* | *Disengaged* | *Engaged* |
|---|---|---|
| Disengaged | 121 | 6 |
| Engaged | 7 | 113 |

While in Overall prediction (F1 Dataset), 135 learners are identified as Disengaged and 112 learners are identified as Engaged. After classifying the learners based on region wise and new threshold values for each region, F2 dataset confirms that 127 learners are disengaged and 112 learners are confirmed as Engaged learners and 8 of the learners are getting advantage of region wise classification. They may fail on overall prediction values but still they got required threshold values on their own region. Thus, the algorithm identifies them as engaged learner. The Main goal of the work is to motivate the learners and make them as engaged as much as possible. Also, Table 5(b) shows that out of 127 disengaged learners, proposed method classified 121 learners correctly and 6 of the learners are wrongly identified as Engaged learner and out of 120 Engaged learners, 113 learners are correctly classified as Engaged and 7 of them are wrongly classified as Disengaged.

The evaluation measures of both dataset are displayed in Table-3 shows that overall classification has 93.33% and Region wise classification has 94.74% accuracy. Thus, the proposed work got better prediction results than the previous method.

True positive is considered as main point of reference to confirm the quality of prediction. Thus, it can be observed from the Table 3 that TP rate is found high at F2 dataset, which means that, adding region wise classification of students to EDD Algorithm makes high predictive quality.

## 5. CONCLUSION

Through this study, we discuss about the various factors which are related to identify the student's disengagement is discussed. The Disengaged behaviour is detected based on the learning behaviour and the marks secured in exams. The proposed work can detect the online and offline student's log data. The region wise classification of the student's log data will provide better accuracy than any other methods. In addition to the above factors for detection of disengaged learners, it may include the sentimental analysis for better prediction of learner's engagement level, getting feedback from learners will helps the tutors to increase the learner's engagement level and also online course content is designed using game theory or course content is explained through stories will increase the student's motivation.

## 6. REFERENCES


[1] Surjeet Kumar Yadav, Brijesh Bharadwaj, Saurabh Pal, "Data Mining Applications: A Comparative study for Predicting Students Performance," International Journal of Innovative Technology & Creative Engineering, 2011.

[2] Alaa el-Halees, "Mining student's data to analyze e-Learning behavior: A Case Study", 2009.

[3] Merceron, A. and Yacef, K., "Educational Data Mining: a Case Study" In Proceedings of the 12th International Conference on Artificial Intelligence in Education AIED 2005, Amsterdam, The Netherlands, IOS Press. 2005.

[4] Al-Radaideh, Q., Al-Shawakfa, E. and Al-Najjar, M. (2006), Mining Student Data Using Decision Trees, The 2006 International Arab Conference on Information Technology (ACIT'2006) – Conference Proceedings.

[5] Shaeela Ayesha, Tasleem Mustafa, Ahsan Raza Sattar, M. Inayat Khan, "Data mining model for higher education system", Europen Journal of Scientific Research, Vol. 43, No. 1, pp. 24-29, 2010.

[6] B.K. Bharadwaj and S. Pal. "Data Mining: A prediction for performance improvement using classification", International Journal of Computer Science and Information Security (IJCSIS), Vol. 9, No. 4, pp. 136-140, 2011.

[7] S. T. Hijazi, and R. S. M. M. Naqvi, "Factors affecting student's performance: A Case of Private Colleges", Bangladesh e-Journal of Sociology, Vol. 3, No. 1, 2006.

[8] Sundar PVP, 'A Comparative Study for Predicting Student's Academic Performance Using Baysian Network Classifiers', IOSR Journal of Engineering, Vol. 03, No. 02, pp. 37-42, 2013.

[9] David E. Cox, Elizabeth Kendall Sprolesand George B. Sproles, "Learning Style Variations Between Rural and Urban Students", "Research in Rural Education", Volume 5, Number 1, 1988.

[10] Yang, S. W., Rural youths 'Decision to attend college: Aspirations and realizations. Paper presented at the annual meeting of the Rural Sociological Society, Guelph, Ontario, Canada (ERIC Document Reproduction Service No. ED207765), 1981.

[11] Rekha Srivastava, Shobhna Joshi, The Effect of School and Area on Academic Self-Concept and Academic Achievement of Adolescents, Delhi Psychiatry Journal, Vol. 14, No.2, PP: 331-336.

[12] Sundar PVP, Kumar AVS. "Evaluation of Regional Benchmark Impact in EDM", International Journal of Computer Science Issues (IJCSI), Vol 10, Issue 2, No 2, March 2013, PP: 531-535.

[13] Dandan Zhang, Xin Li and JinjunXue, "Education Inequality between Rural and Urban Areas of the People's Republic of China, Migrants' Children Education and Some Implications", Asian Development Review, Vol. 32, No. 1, pp. 196–224.

[14] Sundar PVP, Kumar AVS. "A Novel disengagement detection strategy for online learning using quasi framework," Advance Computing Conference (IACC), 2015 IEEE International, Bangalore, 2015, pp. 634-638. doi: 10.1109/IADCC.2015.7154784.

[15] Claypool M, Le P, Waseda M, Brown D. Implicit interest indicators. Proc International Conference on Intelligent User Interfaces; Santa Fe. 2001. p. 33–40.

[16] Gowda SM, Baker RS, Corbett AT, Rossi LM. Towards automatically detecting whether student learning is shallow. International Journal of Artificial Intelligence in Education. 2013 Nov; 23(1):50–70. Doi No.: 10.1007/s40593-013-0006-4.

[17] Hershkovitz A, Nachmias R. Developing a log-based motivation measuring tool. Baker RSJD, Barnes T, Beck JE (Eds.). Educational Data Mining 1st International Conference on Educational Data Mining; Montreal, Quebec, Canada. 2008 Jun 20-21. p. 226–33.

[18] Monterrat B, Desmarais M, Lavoue E, George S. A player model for adaptive gamification in learning environments. AIED; Madrid, Spain. 2015; 9112:297–306.

[19] Oskouei RJ, Askari M, Sajja PRP. Differential impact of web technology on various dimensions of students' academic and non-academic activities - A case study. Indian Journal of Science and Tehnology. 2013 Jul; 6(7):4912–22.

[20] Sundar PVP. Quasi Framework: A new student disengagement detection in online learning. International Journal of Engineering Research and Technology (IJERT). 2012 Dec; 1(10):1–6.

[21] Sundar PVP, Kumar AVS. Disengagement detection in online learning using Quasi framework. IOSR Journal of Engineering (IOSRJEN). 2015; 5(1):4–9.

[22] Sundar PVP, Kumar AVS. An enhanced disengagement detection in online learning using Quasi framework. Special Edition of International Journal of Applied Engineering Research (IJAER). 2015 Jun; 10(55):1298–302.

[23] Sundar PVP, Kumar AVS. A hybrid classification method for disengagement detection in online learning. International Journal of Education and Information Studies. 2015; 5(1):67–74.

[24] Farzan R, Brusilovsky P. Social navigation support in E-Learning: What are real footprints. Proceedings of IJCAI'05 Workshop on Intelligent Techniques for Web Personalization; Edinburgh, U.K. 2005. p. 49–56.

[25] Sundar PVP, Kumar AVS. A systematic approach to identify the unmotivated learners in online learning. Indian journal of science and technology. 2016. April; Vol 9(14): 1-6.

[26] Quasi Framework. Available from: www.quasiframework.com