Predicting Learner Performance Using Data-Mining Techniques and Ontology

Alla Abd El-Rady(✉), Mohamed Shehab, and Essam El Fakharany

Arab Academy for Science, Technology & Maritime Transport, Cairo, Egypt
alla.abdelrady@gmail.com, melemam9@gmail.com,
essam.fakharany@gmail.com

Abstract. The high rates of learners' dropout and failures in different courses that are offered by many universities and educational institutions through the use of e-learning or online learning systems have been a serious concern. Analyzing and studying learners' learning data in order to predict their future performance can support both tutors and e-learning systems to determine learners' progress or status and spot those with low performance. Thus they can offer learners with personalized learning resources and activities designed to each one in order to maximize their learning outcomes and overcome their learning gaps. This paper presents a methodology that uses semantic web technologies as well as data mining techniques to predict learners' future performance based on data produced by learners through their interaction with LMS (Learning Management System) and social networks.

Keywords: E-learning · Semantic web · Ontologies · SWRL · Education data mining · Social networks · LMS

1 Introduction

Over the last years there has been a quick change in the web technologies and internet. This quick change has driven changes in various fields such as economy and education. In education, web technologies play an essential role in changing the way of learning as well as the learning procedures. They change the way of how learning contents are delivered to learners and prompt a new way of learning called e-learning, online learning or distance learning.

E-learning is generally referred to as "web based learning" [1]. What is unique about e-learning is that it facilitates the learning process through the use of computers, networks and web technologies to ease delivery, sharing and management of different e-learning resources and activities to learners. It also aids learners to learn in ways that are much faster, easier and cheaper than the traditional way of learning in order to enhance their performance as well as to improve their final outcomes [1].

Today people use e-learning systems as a normal way of learning. However with the rapid growth of e-learning systems or online learning systems there has been a rising concern over number of problems related to learners such as the high dropout rates [2, 3] and failure to get a degree or pass a course. A lot of researches and studies

A.E. Hassanien et al. (eds.), *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2016*, Advances in Intelligent Systems and Computing 533, DOI 10.1007/978-3-319-48308-5_63

in the last years are concentrating on how to solve these problems through the analysis of learning process in order to improve it.

In general the massive amount of data produced by learners through their learning process is an important asset to different educational stakeholders such as tutors, parents and learners themselves. Highlighting different features and characteristics of learners' data is an important process as it can help tutors in getting better image of learners' progress so they can adapt their learning process according to learners' needs [4].

Different e-learning systems have begun to use educational data mining techniques to analyze learners' data. Therefore tutors can offer those learners personalized and adapted learning activities designed for each one to enhance their final performance.

Moreover, a lot of educational studies and researches have begun to develop personalized e-learning systems using semantic web technologies such as ontology and semantic web rule languages (SWRL) [5]. Utilizing both semantic web technologies and educational data mining techniques in developing e-learning systems is an essential part in the process of analyzing learners' data as they enrich each other by interacting over the time [6].

In this paper we propose a methodology that uses semantic web technologies such as ontology and SWRL as well as data mining techniques to predict learners' final performance according to their learning data. The data is derived from learners' engagement with both learning management system (LMS) and social networks (Facebook). This methodology can help and support both tutors and e-learning systems to analyze the learners' performance and determine those who are underachieving and have high possibility to fail, so they can offer them extra educational materials and activities to avoid their failure.

2 Background Knowledge

E-learning systems, data mining techniques and ontology are three areas of interest in building our system. Brief description of each one is given in this section.

2.1 E-learning

E-learning is a web based learning which introduces a new environment and a new way of learning through the use of internet, interactive multimedia and different web technologies. It's defined as "interactive learning which permits learning through the deployment of computers as an education medium" [7]. Using e-learning, learners can easily start learning independently from others.

E-learning technologies have been changed and evolved over the last years by utilizing internet and web technologies. They can be classified into e-learning 1.0, e-learning 2.0 and e-learning 3.0 [8].

In e-learning 1.0 the learning contents are kept and viewed online so learners can easily access different learning entities and resources. Moving forward to learning 2.0 it became more advanced by allowing learners to access diverse learning materials passively. It also allows learners to share their own beliefs through writing comments

or notes. Finally in e-learning 3.0 learning process and environment can be adjusted or personalized through the use of sematic web technologies [9].

2.1.1 E-learning and Social Networks

Lately, terms like "Social Media" and "Social networks" became so widespread. It gained a strong position in different educational studies and researches as they are being accessed and used by most tutors and learners frequently. Researchers have begun to explore the opportunities and challenges of using social media in educational systems [10].

Social networks are used by different learners for communicating each other as well as for discovering, sharing and exchanging knowledge. Facebook, Twitter, LinkedIn, Instagram, MySpace and Google+ are the most common social networks used around the world. There are over than 1.65 billion active Facebook users [11]. Because of Facebook popularity between learners and tutors we have nominated it for our study as a source of learners' social data. Learners' activities at course Facebook groups such as number of comments, shares, posts and likes can be used in determining those with active engagement and interaction with course which consequently can be used in predicting leaner's future performance.

2.2 Educational Data Mining Techniques

Data Mining (DM) is referred to as the method of analyzing massive amount of raw data from different perceptions and perspectives for the purpose of discovering novel, non-trivial, understandable patterns and valuable information [11].

Educational Data Mining (EDM) is defined as the process of analyzing and studying raw data derived from educational systems in order to discover and extract valuable information and useful knowledge. This knowledge can be presented to diverse participants or stakeholders such as tutors, learners, system developers, parents and educational researchers [12].

Data mining techniques have been categorized and classified from different views and perspectives. Ryan Baker [13] has classified it into: "Prediction, Clustering, Relationship mining, Distillation of data for human judgment and Discovery with models".

For the time being there is an increasing interest in employing data mining techniques in e-learning systems [14]. Digging the vast amount of data made by learners through their learning process and extracting knowledge from it can support tutors in identifying learners' progress [15].

2.3 Ontology

Basically, Semantic Web is defined as an extension of the existing web. Information in semantic web is given a well-defined meaning which can permit both people and computers to work together. It supports people in getting accurate answers to their inquiries by allowing different web agents to reason the multiple web resources and contents [16].

Ontology is an important and essential part of semantic web layer cake. It is defined as "Explicit specification of conceptualization" [17]. It is broadly used in artificial intelligence, knowledge engineering and computer science related applications. These applications are related to different fields such as knowledge discovery, e-commerce and education.

Ontology has been well utilized in e-learning in diverse fields for the purpose of representing learning domain and learner profile, personalizing and recommending learning materials, evaluating learning process and planning course syllabus and contents [17]. It offers a way of discovering and extracting new knowledge through the use of its inference mechanism such (reasoner). Semantic web rule language (SWRL) is used to extend ontology reasoning capabilities. As it is generally used to express different types of relationships and conditions that cannot be expressed or defined using ontological reasoning only.

3 Proposed Methodology

The main purpose of this study is to build a methodology that can be used to predict learners' performance by analyzing data generated from their interaction with both learning management system and Facebook groups. In this study both data mining techniques and semantic web technologies are being used.

3.1 Dataset Description

In this study an educational data of 140 learners with different variables that highlight learners' different aspects was analyzed. The data set was obtained from "UCI Machine Learning Repository" which is an online data sets repository. Modifications were made on it in order to meet our requirements.

The dataset contains variables about learner age, address, sex, family members, average time spent on learning, number of previous failures, learner activities (curriculum related or unrelated), attended sessions, exercises grades, midterm grade and final grade. It also contains variables about average number of comments, posts and likes submitted by leaner on course Facebook groups.

Our objective is to analyze the data and then classify learner into one of two classes "Pass" or "Fail" based on his/her final grade.

3.2 Methodology Architecture

Figure 1 shows the architecture of our proposed methodology. It consists of two main phases. In the following a description of each phase is given.

Phase 1: The main objective of this phase is to select and implement different data mining techniques in order to build our predictive model which predicts learner's future performance based on learner's data. Initially the data was collected and prepared for the analysis process through the using of different preprocessing and filtering

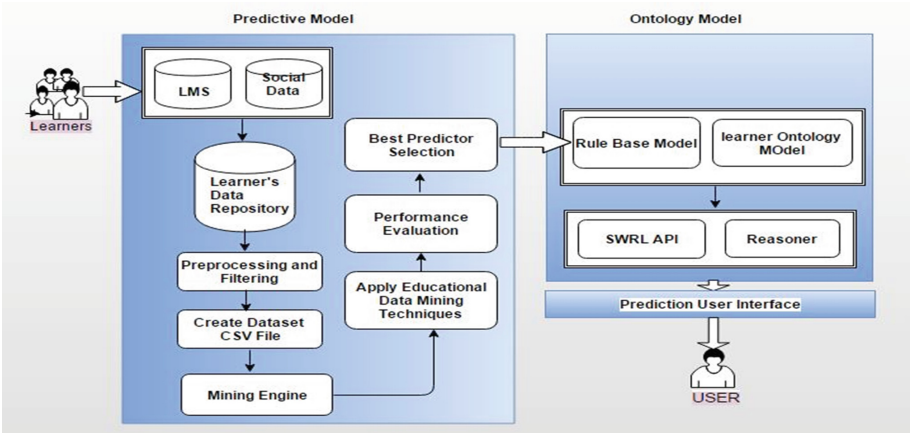

Fig. 1. Overall methodology architecture

techniques in order to remove any duplication in data and handle any missing values which could cause errors later.

The dataset was created as CSV file format (comma separated values file format which is supported by WEKA data mining tool) and then different classification techniques were chosen to be applied by utilizing WEKA data mining tool.

3.2.1 Applying Classification Techniques

Mainly the type of data and the problem domain affect the choice of the classification techniques. Twelve different data mining classification techniques form different groups were chosen to be implemented and evaluated through the use of WEKA data mining tool:

- Functions Family: SMO(Support Vector classifier) and Simple Logistics
- Bayesian Network Family: Naïve Bayes and Bayes Net
- Decision Tree Family: J48, Simple Cart, Random forest, and ADTree
- Rule Based Family: JRIP and OneR
- Lazy Classifiers Family: IB1 and KStar.

These classification techniques are frequently used by researchers in multiple educational researches and studies. They have high potential to yield to good results and high accuracy. Furthermore, these classification techniques used different methodologies in building their prediction models which can increase the chance of finding a prediction model with high accuracy and fewer errors. The results of classification techniques are given in Sect. 4. The output of this phase is the prediction rules which will be mapped and stored in rule base model.

Phase 2: The objective of phase2 is to build learner ontology model and develop our inference engine through the use of ontology reasoning and SWRL rules. The classification results which are the output of phase 1 will be mapped as the input of phase2.

Learner data and prediction rules extracted from the data mining output in phase1 is used to build learner ontology and build our rule base model which is the base of ontology inference engine. This phase consists of the following components:

Learner ontology model which is a key component of phase 2 is used to represent and model learner's learning data such as learner's personal information (gender, age, address) through the identification of different sets or classes, object properties or variables and relationships between these sets. OWL (web ontology language) is a language which is used for ontology encoding. It is used to describe and represent knowledge of each set. Figure 2 shows a part of learner ontology model implemented using Protégé software tool which is the software used for the implementation of phase2.

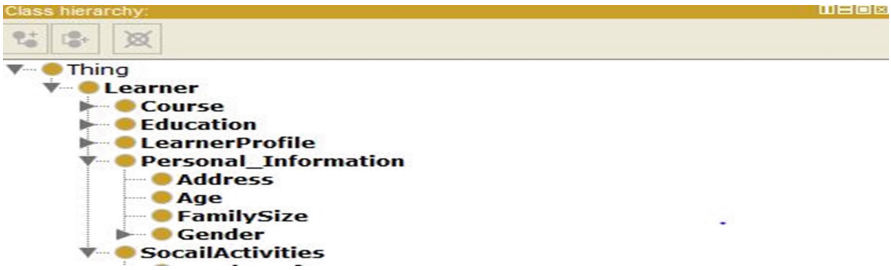

Fig. 2. Example of learner ontology implemented in Protégé

Rule base model stores the prediction rules which were produced from data mining results in phase 1. These rules can be driven for example from decision tree rules. As in the decision tree each branch could have a set of leaves. Each leaf represents a classification (prediction) rule. The following is an example of a developed prediction rule which illustrates some key performance indicators (KPIs) such as leaner midterm grade, session grade and average number of comments made by learner on course group at Facebook:

Example Rule 1: If (Midterm_Grade >= 10) && (SessionGrade_2 >= 2.25) && (AVG_No_Comments > 2) then Pass.

These rules then will be transformed and then represented as Semantic rules through the use of **SWRL API.** Using both SWRL prediction rules along with ontology reasoning will increase and extend the **ontology reasoning** capabilities by allowing creation of multiple conditional rules which cannot be expressed using ontology relationships only.

Both Semantic rules generated using SWRL API and ontology reasoning mechanism work as our inference engine as they enrich and complete each other. By implementing this methodology the predication of learner's final performance can be carried out directly from ontology online.

4 Experimental Results

The performance evaluation of different data mining techniques is given in this section. We used confusion matrix to evaluate the different data mining techniques [18]. Table 1 shows the confusion matrix used for this study.

Table 1. Confusion matrix

Data label	Correctly classified	Incorrectly classified
Correct	TP (True Positive)	FN (False Negative)
Incorrect	FP (False Positive)	TN (True Negative)

There are many measurement parameters which are used for the performance evaluation of different data mining techniques such as Accuracy, Precision, Recall, F-Measure, TP Rate (True Positives Rate) and FP Rate (False Positive Rate) [18, 19].

In our experiment we used those measurement parameters to evaluate 12 different classification techniques in order to determine the technique with the highest classification accuracy and fewest errors.

The training dataset was divided into 10-folds using cross validation to be used as testing data for the evaluation process. Figure 3 and Table 2 show the values of different measurement parameters. Random forest records the best values among other techniques except for FP Rate measurement. The results show that Simple Logistics, SMO, Bayes Net, Random Forest, JRip and ADtree record the best value for FP Rate respectively.

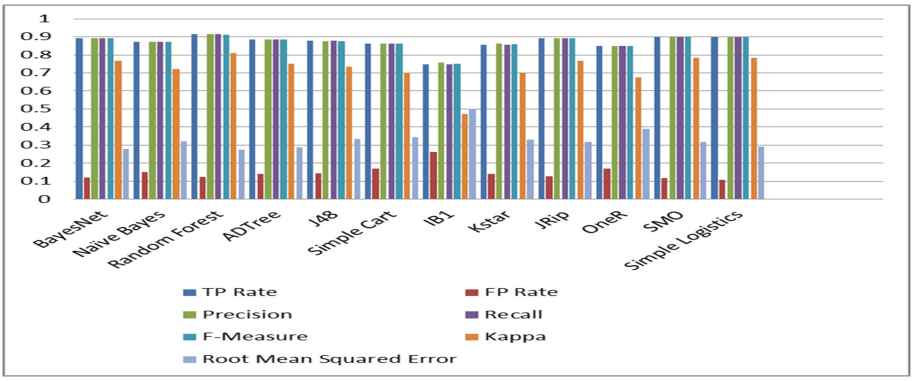

Fig. 3. Results of different classification techniques.

In general, results of different data mining techniques depend on data to be classified and on its distribution. In our Experiment, Random Forest gives best accuracy (91.36 %) followed by Simple Logistics and SMO with accuracy (89.9281 %), then

Table 2. Classification techniques results

	Accuracy	TP Rate	FP Rate	Precision	Recall	F-Measure	Kappa	Root Mean Square Error
Bayes Net	89.2086	0.892	0.12	0.893	0.892	0.892	0.7687	0.2776
Naïve Bayes	87.0504	0.871	0.149	0.871	0.871	0.871	0.7213	0.3217
Random Forest	91.3669	0.914	0.124	0.915	0.914	0.912	0.8095	0.2734
AD Tree	88.4892	0.885	0.141	0.884	0.885	0.884	0.7502	0.2882
J48	87.7698	0.878	0.145	0.877	0.878	0.877	0.7357	0.3342
Simple Cart	86.3309	0.863	0.17	0.862	0.863	0.862	0.7021	0.3425
IB1	74.8201	0.748	0.261	0.758	0.748	0.751	0.4732	0.5018
Kstar	85.6115	0.856	0.141	0.862	0.856	0.858	0.6978	0.3321
JRip	89.2086	0.892	0.128	0.892	0.892	0.892	0.7668	0.3179
OneR	84.8921	0.849	0.17	0.85	0.849	0.849	0.6761	0.3887
SMO	89.9281	0.899	0.116	0.899	0.899	0.899	0.7832	0.3174
Simple Logistics	89.9281	0.899	0.108	0.9	0.899	0.9	0.785	0.2914

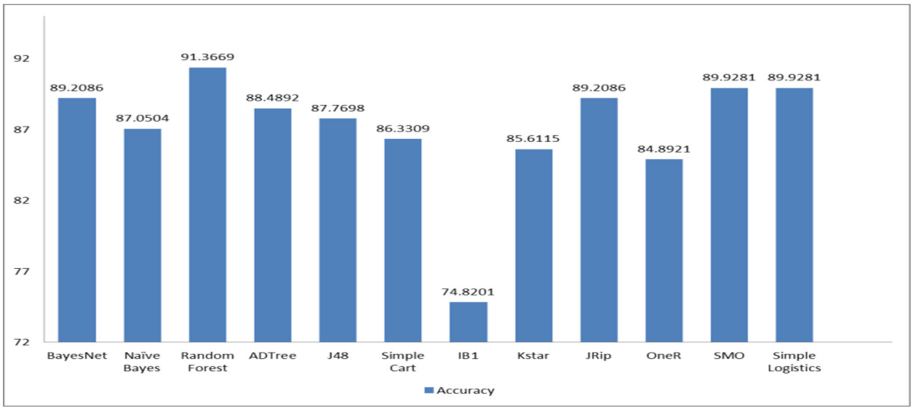

Fig. 4. Accuracy results of different classification techniques.

Bayes Net and JRip with same accuracy (89.2086 %), AD Tree with accuracy (88.4892 %), then J48 with accuracy (87.76 %) as shown in Fig. 4.

In our experiment, random forest technique shows the best result in predicting learner's class label. Random forest is built from various decision tree techniques. It uses the majority vote technique to predict the class label. Usually combining results of multiple techniques using an ensemble way give better results compared to using single technique for prediction.

5 Conclusion and Future Work

This paper proposed a methodology which can be used to predict learners' future performance and identify learners with high risk to drop out course or fail in the final exam. The prediction of learners' future performance can help learners to be aware of

their progress as well as tutors to improve their teaching procedures in order to engage underachieving learners in an appropriate learning process.

The data produced through learner engagement and interaction with both e-learning systems and social networks was analyzed through the utilization of different classification techniques in order to extract prediction rules and valuable knowledge. The extracted prediction rules along with learner's data were used to build learner ontology model and rule base model with the use of semantic web technologies.

The results of this study show that some variables have direct impact on learner performance such as average number of comments, midterm grade, study time, age and gender. In the future more variables could be examined in order to obtain prediction with fewer errors and higher accuracy. Also we could categorize learner's comments into positive and negative comments. Moreover we could consider the type of errors made by each learner in different tests in order to enrich our analysis and identify learner's weak points.

References

1. Surjono, H.D.: The design of adaptive e-learning system based on student's learning styles. Int. J. Comput. Sci. Inf. Technol. **5**, 2350–2353 (2011)
2. Yukselturk, E., Ozekes, S., Türel, Y.K.: Predicting dropout student: an application of data mining methods in an online education program. Eur. J. Open Dist. e-Learn. **17**(1), 118–133 (2014)
3. Dewan, M.A.A., Lin, F., Wen, D., Kinshuk.: Predicting dropout-prone students in e-learning education system. In: UIC-ATC-ScalCom-CBDCom-IoP, Beijing, China (2015)
4. Baradwaj, B.K., Pal, S.: Mining educational data to analyze students' performance. (IJACSA) Int. J. Adv. Comput. Sci. Appl. **2**(6), 63–69 (2011)
5. Qwaider, W.Q.: E-learning system based on semantic web technology. In: Second International Conference of E-learning and Distance Learning, Riyadh (2011)
6. Kazi, A., Kurian, D.T.: An ontology based approach to data mining. Int. J. Eng. Dev. Res. **2** (4), 3394–3397 (2014)
7. Titthasiri, W.: A comparison of e-learning and traditional learning: experimental approach. In: International Conference on Mobile Learning E-society and E-learning Technology (ICMLEET), Singapore, November 2013
8. Chung, H., Kim, J.: An ontological approach for semantic modeling of curriculum and syllabus in higher education. Int. J. Inf. Educ. Technol. **6**(5), 365–369 (2016)
9. Weber, P., Rothe, H.: Social networking services in e-learning. In: Proceedings of World Conference on E-learning in Corporate, Government, Healthcare, and Higher Education (2012). https://www.researchgate.net/publication/235975162. Accessed 11 May 2016
10. ZEPHORIA Digital Marketing: The Top 20 Valuable Facebook Statistics, April 2016. https://zephoria.com/top-15-valuable-facebook-statistics/. Accessed 24 May 2016
11. Srivastava, J., Srivastava, A.K.: Understanding linkage between data mining and statistics. Int. J. Eng. Technol. Manage. Appl. Sci. **3**(10), 4–12 (2015)
12. Lakshmi Prabha, S., Mohamed Shanavas, A.R.: Educational data mining applications. Oper. Res. Appl. Int. J. (ORAJ) **1**(1) (2014)
13. Elaal, S.A.E.A.: E-learning using data mining. Chin. Egypt. Res. J. (2011)

14. Prakash, B.R., Hanumanthappa, M., Kavitha, V.: Big data in educational data mining and learning analytics. Int. J. Innov. Res. Comput. Commun. Eng. **2**(12), 7515–7520 (2014)
15. Romero, C., Ventura, S., Espejo, P.G., Hervás, C.: Data mining algorithms to classify students. In: Proceedings of the 1 st International Conference on Educational Data Mining, Montreal, Quebec, Canada, pp. 20–21 (2008)
16. Kolovski, V., Galletly, J.: Towards e-learning via the semantic web. In: International Conference on Computer Systems and Technologies – CompSysTech 2003 (2003)
17. Al-Yahya, M., George, R.: A. Alfaries:"Ontologies in e-learning: review of the literature. Int. J. Softw. Eng. Appl. **9**(2), 67–84 (2015)
18. López, V., del Río, S., Benítez, J.: Cost-sensitive linguistic fuzzy rule based classification systems under the MapReduce framework for imbalanced big data. Sciencedirect Fuzzy Sets Syst. **258**, 5–38 (2014)
19. Gupta, D.L., Malviya, A.K., Singh, S.: Performance analysis of classification tree learning algorithms. Int. J. Comput. Appl. **55**(6), 0975–8887 (2012)