Identifying At-Risk Students for Early Interventions—A Time-Series Clustering Approach

M.Sivashankar1, Dr.M.Sreedevi2
1PG Student, Department of Computer Science, S.V.University, Tirupati—INDIA
2Assistant Professor, M.C.A, M.Phil., PhD Department of Computer Science, S.V.University, Tirupati—INDIA.

Abstract—The purpose of this study is to identify at-risk online students earlier, more often, and with greater accuracy using time-series clustering. The case study showed that the proposed approach could generate models with higher accuracy and feasibility than traditional frequency aggregation approaches. The best performing model can start to capture at-risk students from week 10. In addition, the four phases in student’s learning process detected holiday effect and illustrates at-risk students’ behaviors before and after a long holiday break. The findings also enable online instructors to develop corresponding instructional interventions via course design or student-teacher communications.

Index Terms—Clustering, classification, and association rules, Feature extraction or construction, Mining methods and algorithms, Time-Series analysis, LMS, predictive modeling.

I. INTRODUCTION

According to Online Learning Consortium’s Survey of Online Learning Report in 2013 [1], over 7.1 million students were taking at least one online course, and the number of students taking at least one online course has continued to grow at a rate far in excess of overall enrollments. Furthermore, 65% of higher education institutions now say that online learning is a critical part of their long-term strategy [1] [2]. With the exponential growth of online courses in higher education, retention is an area of great concern. Thus, it is imperative for institutions to develop practices and interventions that can contribute to student retention in fully online programs [3] [4]. One of the approaches is to harness the predictive power of Learning Management System (LMS) data to develop an early warning system or tools that identify at-risk students and allow for more timely pedagogical interventions in improving student retention [5]. An effective early warning system could provide formative grade feedback to online students and could help online programs take proactive steps to intervene before a student drops out or falls behind in the course. A key problem is how to identify at-risk

• J. Hung is with the Dept. of Educational Technology, Boise State University, Boise, ID 83725. E-mail: an@boisestate.edu • M. Wang is with the Dept. of Statistics, University of Central Florida, Orlando, FL 80523. E-mail: chung-ching.wang@ucf.edu • S. Wang is with the Dept. of Curriculum, Instruction and Special Ed, The University of Southern Mississippi, Hattiesburg, MS 80309. Email:shuyan.wang@usm.edu • M. Ab is a PhD student in the Dept. of Computer Science, Old Dominion University, Norfolk, VA 23529, USA. E-mail Email:yaohang@cs.odu.edu • W. He is with the Department of Information Technology & Decision Sciences, Old Dominion University, Norfolk, VA 23529, USA. E-mail: whe@odu.edu.

Online students in time to help them, even when they do not seek assistance. Based on these needs, a few predictive algorithms and early-warning systems aimed at identifying at-e (e.g.[5][8]) identify at-risk students, prediction accuracy is the most important attribute during the model construction. However previous studies were primarily based on aggregated data. Students with similar aggregated behavior frequencies or time spent will be predicted as having similar learning performance. Therefore, the aggregation approaches cannot generate the most accurate predictions because of the following reasons. First, most studies utilized aggregated data from the whole semester to construct predictive models, such
as total frequency of logins, total frequency of content access, and accumulated time spent in the course (e.g. [5] [9]). However, these studies failed to consider variances in learning patterns. For example, two students had identical aggregated time spent per week. One student might evenly allocate his/her learning time across weekdays but the other student might spend longer time on weekends only.

Second, the aggregation approaches failed to consider differences in student learning preferences. For example, at-risk students might engage well on topics of personal interest but might fail to maintain minimum participation levels on topics they do not like.

Third, the aggregation approaches failed to consider variances in course activity requirements across different course modules [10]. While a course is in progress, students may be asked to participate in different course activities during any one week or module. At-risk situations may occur when a student fails to meet the changing labeled nodes to unknown nodes. Based on passing time during random walks with bounded lengths, Callut et al. [29] and Newman [30] introduce a novel technique, called D-walks, to handle semi-supervised classification problems in large graphs. Zhou and Schlkopf [31] define calculus on graphs by using spectral graph theory, and propose a regularization framework for classification problems on graphs. However, many semi-supervised learning methods rely heavily on the assumption that the net-work exhibits homophily, i.e., nodes belonging to the same class tend to be linked with each other [12]. Meanwhile, the implementation of semi-supervised learning algorithm often requires a large amount of matrix computation, and thus is infeasible for processing large datasets [25]. Many methods have been developed to overcome these limitations. For example, Tong et al. propose a fast random walk with restart algorithm [32] to improve the performance on large-scale dataset. Lin et al. propose a highly scalable method, called Multi-Rank-Walk (MRW), which requires only linear computation time in accordance to the number of edges in the network [12]. Mantrach et al. [33] design two iterative algorithms which can be applied in networks with millions of nodes to avoid the computation of the pairwise similarities between nodes. Gallagher et al. [13] design an even-step random walk with restart (Even-step RWR) algorithm, which mitigates the dependence on network effect.

A. Active Learning
In active learning [34], the number of known labels required for accurate learning is reduced by intelligently selecting to-be-labeled nodes to achieve improved classification performance in sparsely labeled networks. Lewis and Catlett [35] propose a method based on uncertainty reduction, which selects the data with lowest certainty for querying. However, the method will fail when there are a certain number of outliers. The outliers have high uncertainty in the network, but getting their labels doesn’t help to inference the rest data. To handle this limitation, Roy and McCallum [36] design a method to determine the impact on the expected error of each potential labeling request by using Monte Carlo approach. In the active learning process, the feature of linked data in the network can also be taken into account. Bilgic and Getoor [18] propose several ways of adapting existing active learning methods to network data. Maeskassy [37] designs a novel hybrid approach by using community detection and social network analytic centrality measures to identify the candidates for labeling. When network structure and node attribute information are available, Bilgic et al. [19] apply several classic active learning strategies such as disagreement and clustering to select samples for labeling, which has shown significant improvements over baseline methods. Active learning is able to overcome the sparse labeling problem to some extent, but it still requires the participation of experts and lacks an automatic learning process.

II. LITERATURE REVIEW

2.1 Predictive modeling for at-risk student detection
Almost all educational data are recorded in database systems today, and thereby, lead to a substantial and dramatic growth in stored data. This is especially true in online education. During the online instruction process, students can choose to interact with course materials and with instructors or other students via multiple communication channels. All related activities are tracked and stored in back-end database systems and server logs. The stored data offers a greater opportunity to extract hidden knowledge that can then be used to inform instruction and pedagogy. This also drives the developments of Learning
Analytics (LA) [18] and Educational Data Mining (EDM) [19]. Among research topics in LA and EDM studies, at-risk student prediction is absolutely one of the hottest topics [20]. Generally speaking, at-risk signals identified by previous literatures can be classified into two categories: (1) student profiles, such as gender [21] [22], age [21] [22], ethnicity [23], cumulated GPA [23] [24] [25], standardized scores [23] [24] [25], current credit load [26], and course satisfaction [27] [28] [29]; and (2) student’s engagement level in the online courses, such as frequency of logins [30] [31] [33], frequency of course material accessed [30] [31] [33], number of discussions posted [30] [31] [33], percentages of grade earned [26] [32], and total time spent [31] [34]. However, results of previous studies are difficult to generalize as common profile signals because findings really depend on what profile variables can be collected and analyzed. On the other hand, studies on student’s engagement level in online courses have led to pretty consistent findings. The literature shows that, in general, student’s performance is highly related to their engagement level in any given course. Almost all related studies found higher engagement level usually leads to higher performance (e.g. [5] [6] [22] [26] [31] [33]). The most similar node with u, and thus, leading to lower performance.

Requirements of a course at a specific point in time
Since behavior feature can provide a different kind of information that may be useful in sparsely labeled networks, we propose a novel Behavior-based Collective Classification method (BCC) in this paper to handle the sparse labeling problem. The process of BCC in network data consists of four steps: behavior feature extraction, screening valuable nodes, classification by voting and collective inference.

III. DATA ANALYSIS

3.1 Data source
In this case study, data was collected from an online graduate program in the United States. The program offers approximately 20 graduate-level courses, hosted in Moodle. Dynamic data was collected from Moodle logs that contain 12 courses with 25 course sections and 509 enrollments in the spring semester of Spring 2014. Some courses might incorporate unique learning activities, such as Blog, Glossary, and Wiki. These unique activities were filtered out and obtained the following four common course behaviors for analysis (1) frequency of course material accessed, (2) frequency of forum read, (3) number of discussions posted, and (4) number of replies posted. After initial cleaning, the dynamic dataset contains 427,382 logs in the time period of 16 weeks. Student static data contains student demographics retrieved from the institution’s data warehouse. Student’s final grade is a nominal variable which contains multiple levels A+, A, A-, B+, B, B-, C+, C, C-, and F (failed). To avoid the curse of dimensionality, the final grade was consolidated into three levels A (A+, A, and A-), B (B+, B, and B-), and F (C+, C, C-, and F). Table 1 lists all dynamic and static variables for the analysis.

3.2 Model training
SAS Enterprise Miner 13.1 was employed to conduct analyses in this study. Model training was conducted using all static and dynamic variables listed in Table 1. Dynamic data was aggregated using two approaches: (1) traditional approach which aggregated frequencies by the whole semester, and (2) time-series approach that aggregated frequencies by day as data points of time-series. Therefore, each of the students had eight dynamic variables: four.
behaviors aggregated for the traditional approach, and the other four for the time-series approach. The traditional approach uses the aggregated dynamic and static variables in the predictive modeling. The time-series clustering approach adopted the most popular distance measure (Euclidean) and clustering methods (k-Means) to classify time-series data based on pattern similarity [52]. Then, the clustering results were combined with static variables in the predictive modeling. Stratified sampling was applied to a random selection of 60% of students for model training and the remaining 40% for validation. Six predictive algorithms were used in the analyses: Decision Tree, Boosting, Logistic Regression (forward, backward, & stepwise), and Rule 3.3 Results Figure 2 showed results of predictive modeling. The top five were all time-series models (TS xxx) with lowest misclassification rates on the validation dataset. Decision Tree was the best model with 0.101948 misclassification rate. Except for Decision Tree, the other two rule-based algorithms (Boosting and Rule Induction) in time-series also performed better than the rest of the models. However, the backward logistic regress model in time-series was not superior to traditional approaches. Based on results of the best model (decision tree), patterns of two discussion-related variables, number of replies posted (RP) and number of discussions posted (DP), were selected as important predictors. The next step is to test whether shorter time-series can still generate accurate predictions. Because the whole semester consists of 16 weeks, time-series data were chunked into various duration lengths. For example, week 1 contains time-series of the first week and week 2 contains timeseries of the first two weeks. Figure 3 compares prediction accuracy rates of time-series models from week 1 to week 16. Overall, Decision Tree is more stable and has higher Results Figure 2 showed results of predictive modeling. The top five were all time-series models (TS xxx) with lowest misclassification rates on the validation dataset. Decision Tree was the best model with 0.101948 misclassification rate. Except for Decision Tree, the other two rule-based algorithms (Boosting and Rule Induction) in time-series also performed better than the rest of the models. However, the backward logistic regress model in time-series was not superior to traditional approaches. Based on results of the best model (decision tree), patterns of two discussion-related variables, number of replies posted (RP) and number of discussions posted (DP), were selected as important predictors. The next step is to test whether shorter time-series can still generate accurate predictions. Because the whole semester consists of 16 weeks, time-series data were chunked into various duration lengths. For example, week 1 contains time-series of the first week and week 2 contains timeseries of the first two weeks. Figure 3 compares prediction accuracy rates of time-series models from week 1 to week 16. Overall, Decision Tree is more stable and has higher.

3.3 Results
Figure 2 showed results of predictive modeling. The top five were rates on the validation dataset. Decision Tree was the best model with 0.101948 misclassification rate. Except for (Boosting and Rule Induction) in time-series also performed better than the rest of the models. However, the backward logistic regress model in time-series was not superior to traditional approaches. Based on results of the best model (decision tree), patterns of two discussion-related variables, number of replies posted (RP) and number of discussions posted (DP), were selected as important predictors. The next step is to test whether shorter time-series can still generate accurate predictions. Because the whole semester consists of 16 weeks, time-series data were chunked into various duration lengths. For example, week 1 contains time-series of the first week and week 2 contains time-series of the first two weeks. Figure 3 compares prediction accuracy rates of time-series models from week 1 to week 16. Overall, Decision Tree is more stable and has higher.

That also explained why DP and RP were selected as predictors. In addition to visual comparisons, Linear Discriminant Analysis (LDA) was conducted by using CMA, FR, DP, and RP as feature vectors. As shown in Figure 9, the top eigenvector revealed that the key activities differentiating at-risk students from others were discussion related behaviors (DP and RP) in certain weeks. The peaks marked in Figure 9 represented weeks whose eigenvector values were higher or lower than 0.2 or -0.2 respectively. The results showed that at-risk students had unstable discussion participation levels (sometimes higher and sometimes lower than A and B students) in the period of weeks 1-5 (i.e., before week 11). Starting from peak 6 (i.e., week 11), at-risk students had significantly lower discussion participation levels at
IV. DISCUSSION

Performed better than the data aggregation models. Comparing with rule based models, logistic regression models did not appear to work study reveals that at-risk situation could occur suddenly during the semester. The time series clustering approach has potentials for identifying a risk situations in a real-time manner. The authors expected to identify more successful or at-risk learning patterns by incorporating learner’s demographics in the analysis. However, as none of student’s

V. LIMITATION AND FUTURE

Research

This study mainly compared time-series clustering with the traditional frequency aggregation approach. However, in addition to Euclidian distance and k-Means applied in this study, other similarity/clustering measures have been discussed by researchers on classifying time-series data [36] [49] [50]. Other similarity measures could possibly further improve outcomes of time-series clustering. In addition, this study discarded unique course learning activities that could influence models’ performances because performance variances resulted from other learning activities were not considered and analyzed. Finally, the dataset contains high percentage of A students as they are graduate courses. The time-series clustering might even perform better than a traditional approach in analyzing non-graduate courses with more normal distributed grade levels. Future research may focus on the following aspects: (a) identifying the most appropriate similarity measures for time-series clustering in online education; (b) validating holiday effect with more studies, especially in different disciplines; and (c) analyzing contents in online discussion. This study only considers behavioral frequencies. If discussion contents (text mining) can also be included in the analysis, model’s performance might be further improved in the forms of accuracy and/or timing.sparingly labeled networks,” in Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2008, pp. 256–264.


REFERENCES


[7] W. He, ”Examining students’ online interaction in a live video streaming


[26] R. Bukralia, A. V. Deokar and S. Sarnikar, ”Using Academic Analytics to Predict Dropout Risk in E-Learning Courses: Reshaping Society through Analytics,”


Dr. M. Sreedevi, M.C.A., M.Phil., Ph.D is Assistant Professor in the Department of Computer Science, Sri Venkateswara University, Tirupati.