From: **Irena Koprinska** irena@jedm.educationaldatamining.org
Subject: [JEDM] Editor Decision
Date: May 6, 2018 at 11:53 PM
To: Seoyeon Park pseoyeon5@gmail.com, Noboru Matsuda noboru.matsuda@tamu.edu

Dear Seoyeon and Naboru,

Thank you for submitting your work to the JEDM track of EDM.

Your contribution is very relevant to EDM and JEDM. However, in its present form, it requires **major revisions**. Your paper will be subject to another round of reviews.

The reviewer concerns include clarifying your methodology, addressing the time and skill dependency, including a comparison with the previously reported method for every opportunity, including a baseline, improving the discussion and writing of the paper.

Given the timeline and the current state of the paper, it appears unlikely that the paper can be accepted in time to JEDM so that it is presented at EDM 2018. In order for a JEDM paper to be presented at EDM 2018, it has to be re-submitted by May 14th and fully accepted by May 28th. However, papers that miss these deadlines will still be considered as regular submissions to JEDM. So please submit a carefully revised version of your contribution, together with a response to the reviewer comments.

We hope that you find the reviews helpful and constructive, and we wish you success with your work!

Kind regards,

Irena Koprinska and Min Chi, JEDM Track Co-Chairs (EDM 2018)

Andrew Olney, Editor, Journal of Educational Data Mining

Reviewer 1:
Recommendation: Revisions Required

How relevant is this submission to the scope of JEDM? (if applicable: to the targeted special issue?)

> relevant

How novel is the described research? Are the authors aware of related work?

> Novelty is somewhat limited because logistic regression and the gradient boosted decision tree model are standard machine learning methods.

What is the scientific contribution of this submission? Is it clearly explained, in terms of how the paper advances the EDM field or contributes to related fields?

> The contribution is mainly on feature engineering and using standard machine learning algorithms to address the specific problem of wheel-spinning.

Is the work technically sound? Are there enough methodological details? Are claims convincingly substantiated, either through theoretical argument or empirical data?

> Yes but can be improved. In particular, the gradient boosted decision tree model needs

more explanations.

Do the authors describe the limitations of their approach in a satisfactory manner?

Yes.

How significant is the research? Will the paper be likely to have an impact on the community?

The wheel-spinning detection problem itself is interesting and significant. This paper provides a possible way to solve it.

Does the title of this paper clearly and sufficiently reflect its contents?

Yes

Are the presentation, organization and length satisfactory?

Mostly satisfactory.

Can you suggest additions or amendments or an introductory statement that will increase the value of this paper?

Can you suggest any reductions in the paper, or deletions of parts?

Are the illustrations and tables necessary and acceptable?

Are the key words and abstracts/summary informative?

Please list any other general comments or specific suggestions below.

This paper develops methods to detect wheel-spinning in student performance on cognitive tutors. Six features are created to build the detector, including "M/W on each opportunity", and average probability of correct first attempt per student, skill, student-skill pair, problem type, and student-problem type pair. Two models are developed that apply these features. The first uses logistic regression, which shows somewhat similar but improved results in its accuracy, precision and recall rate. To learn more rapidly, a second model trained using gradient boosted decision trees is also developed.

1. It is unclear to me what "practices opportunities", "skill", "attempt" exactly mean when stating the problem. The authors are suggested to give more concrete examples to clarify these concepts. Moreover, it is also important to make clear if there are any temporal dependencies among the sequence of opportunities.

2. The authors point out that their detector can have negative examples of the M/W features. That is, if a student already achieves M before the 10th opportunity (which is already observed), the detector can still predict W on him/her on the 10th opportunity. How could that happen? Isn't there a naïve way to overcome this deficiency? How about training the detector based on only the remaining five features and then combining the output with the "M/W on each opportunity", as the latter is actually a definite indicator of M.

3. The gradient boosted decision tree model needs more explanations. Maybe a more formal problem formulation and some figures can help understand the adopted method.

4. The contribution of this paper is somewhat limited as it is mainly on feature engineering. Both logistic regression and the gradient boosted decision tree model are standard machine learning methods. Do the authors improve these methods in novel ways that are specific to the considered problem?

5. There are quite a few errors in the paper and the authors should carefully address. For example
Page 3, “provide us an insight” -> “provide us with an insight”
Page 6, “provide us additional information” -> “provide us with additional information”
Page 6, “how it affect students’ mastery” -> “how it affects students’ mastery”
Page 8, “the area under curve of the ROC” -> “the area under the curve of the ROC” or “the area under the ROC curve”
Page 8, “the trend of precision and recall rate” -> “the trend of precision and recall rates”
Page 11, “various types other datasets” -> “various types of other datasets”

Reviewer 2:
Recommendation: Revisions Required

How relevant is this submission to the scope of JEDM? (if applicable: to the targeted special issue?)

This paper studies the problem of early detecting so-called wheel-spinning of student performance recorded as a sequence of events. Because of that the paper is indeed relevant for JEDM.

How novel is the described research? Are the authors aware of related work?

The relevant work is well studied in the paper.

What is the scientific contribution of this submission? Is it clearly explained, in terms of how the paper advances the EDM field or contributes to related fields?

The main contribution is the described work on feature design in order to improve the overall performance of detecting student's wheel spinning.

Is the work technically sound? Are there enough methodological details? Are claims convincingly substantiated, either through theoretical argument or empirical data?

In my view the work is not technically sound due to several reasons.

First, the data includes records that might be dependent (for example when they are related to the same student in different times). This implies that iid modelling (Logistic regression, gradient boosted trees) proposed in this paper is not adequate for the data. Instead, longitudinal models have to be applied.

Second, the cross validation procedure proposed to estimate the performance of the models cannot be used due to the mentioned dependency present in the data. Since it was used in the experiments, the performance estimation is optimistic: for example when records for a student can be simultaneously in the training folds and test fold.

Third, the problem considered is imbalanced. This implies the need for balancing techniques such as cost sensitive learning, ROC analysis, etc.

Do the authors describe the limitations of their approach in a satisfactory manner?

In Section 6 two limitations are present: one-data set experiments and the measure for mastery

mastery?

How significant is the research? Will the paper be likely to have an impact on the community?

Yes, if adequate modelling is used.

Does the title of this paper clearly and sufficiently reflect its contents?

yes

Are the presentation, organization and length satisfactory?

yes

Can you suggest additions or amendments or an introductory statement that will increase the value of this paper?

see my comments above plus more experiments are needed.

Can you suggest any reductions in the paper, or deletions of parts?

no

Are the illustrations and tables necessary and acceptable?

yes

Are the key words and abstracts/summary informative?

yes

Please list any other general comments or specific suggestions below.

Minor remark: PCA is a unsupervised technique: it does not explain any dependent variable by definition.

Reviewer 3:
Recommendation: Revisions Required

How relevant is this submission to the scope of JEDM? (if applicable: to the targeted special issue?)

Very relevant.

How novel is the described research? Are the authors aware of related work?

Marginally novel. The main contributing is in the construction of six features, which are used together with logistic regression and boosted decision trees in a binary classification task: 2 classes – mastery (M) and wheel spinning (W).

In terms of related work, the focus is too narrow. The authors cite their previous work (Matsuda et al.,2016)and also the work of another research group - Beck and Gong (2013), Gong and Beck (2015) and Gong et al. (2016). The research should be situated in the broader area of predicting student performance.

the broader area of predicting student performance.

What is the scientific contribution of this submission? Is it clearly explained, in terms of how the paper advances the EDM field or contributes to related fields?

> The contribution is very limited. It shows how the constructed features (when used with the two chosen machine learning algorithms) perform, on a very specific problem, using one particular dataset.

Is the work technically sound? Are there enough methodological details? Are claims convincingly substantiated, either through theoretical argument or empirical data?

> I am not sure if the paper is technically sound as there are important details that are missing. The following issues need to be addressed:
>
> 1. The paper is not self-content. It assumes that readers are familiar with previous work on wheel-spinning. It was not clear to me what “skill”, “attempt” and “practice opportunities” are. These concepts need to be clearly defined and illustrated with examples; they are very important for understanding the paper.
>
> I am also not sure what the definition of “cognitive tutor” is and what the relation between a cognitive tutor and the proposed method is. The title says “wheel-spinning detection in student performance on cognitive tutors”. Is the method applicable only to a cognitive tutors? I think that wheel-spinning is a problem for ITS in general.
>
> 2. The task is classification – predict M or W. What is the distribution of the two classes? The majority class will determine the baseline for the classification performance; it is very important to compare with a baseline.
>
> The skills are not independent – the more complex ones depends on the simpler ones; if a student hasn’t mastered the simpler ones, he/she will fail the more complex ones. This skill (and time) dependency is not taken into account when constructing the training data. How would it affect the training data and the task? Is it possible that the later data is mainly from class W (as the students haven’t mastered the previous required skills) and hence the task in the later time period becomes very imbalanced?
>
> 3. The authors need to properly compare with previous work, not simply mention previous results as 70-79% precision and 25-50% recall. This means that the results should be presented for each opportunity (from opp3 to opp9).
>
> 4. The discussion of the results is very shallow. It simply summarizes what the tables show without any insights.
>
> 5. Why is the logistic regression model called a “generic” model and the boosted trees called an “upgrated” model? Both are equally generic and a boosted tree classifier is not an upgrated versions of logistic regression.
>
> There are other many problems with the use of terminology and description of the methods, e.g. just to give 2 examples:
> --The summary of boosted trees on p. 9 is very poor - not clear, not coherent; it looks like sentences taken from various sources. Which are the “weak algorithms”?
> --On the same page – “We train gradient boosted decision trees with a”ten-fold cross validation by each opportunity…”. Cross validation is not a training algorithm but an evaluation procedure.

Do the authors describe the limitations of their approach in a satisfactory manner?

> Two limitation are listed.

How significant is the research? Will the paper be likely to have an impact on the community?

> Not likely unless the evaluation is extended to other datasets and the issues are addressed.

Does the title of this paper clearly and sufficiently reflect its contents?

> No.
>
> 1. “cognitive tutors”- is the work limited to cognitive tutors only? I think no.
>
> 2. “using an ensemble model” – 2 methods are used, not only an ensemble method – logistic regression and ensemble of decision trees. In addition, the contribution is not in the ensemble method which is simply taken “off-the-shelf” (DeepMiner), even its parameters are not optimised (it uses the default values). The contribution is in the feature construction.

Are the presentation, organization and length satisfactory?

> The presentation is not satisfactory – clarity, terminology, right level of detail, grammatical sentences – all these require work. See my comments above.
>
> Organisation – include a related work section, separate the conclusions from the discussion and include future work.
>
> Length – the paper is too short. It looks like a conference paper. It lacks the depth and insights required for a journal paper.

Can you suggest additions or amendments or an introductory statement that will increase the value of this paper?

> Define “skill”, “attempt”,“practice opportunities” and give an example to illustrate them.
> Define “cognitive tutors”.

Can you suggest any reductions in the paper, or deletions of parts?

> Please see my comment above - the paper is rather short and looks like a conference paper. It needs extensions.

Are the illustrations and tables necessary and acceptable?

> Table 2 requires more explanation – I did not find it useful in its current form.

Are the key words and abstracts/summary informative?

> There are no keywords.
>
> Abstract:
> -low recall is not "lack of detecting power"
> -generic and upgraded - incorrect terminology, see my comments above
> -more simplified and fast - a simpler (in what sense - smaller number of features), fast - I actually didn't see results supporting this claim

Please list any other general comments or specific suggestions below.

> Please see my comments above
