Comparing Expert and Metric-Based Assessments of Association Rule Interestingness

Diego A. Luna Bazaldua
+1 (917) 543 7674
dal2159@tc.columbia.edu

Ryan S. Baker
+1 (412) 983-3619
baker2@exchange.tc.columbia.edu

Maria Ofelia Z. San Pedro
+1 (508) 330-0410
mzs2106@tc.columbia.edu

Department of Human Development, Teachers College, Columbia University
525 W. 120th Street, New York, NY 10027

ABSTRACT

In association rule mining, interestingness refers to metrics that are applied to select association rules, beyond support and confidence. For example, Merceron & Yacef (2008) recommend that researchers use a combination of lift and cosine to select association rules, after first filtering out rules with low support and confidence. However, the empirical basis for considering these specific metrics to be evidence of interestingness is rather weak. In this study, we examine these metrics by distilling association rules from real educational data relevant to established research questions in the areas of affect and disengagement. We then ask three domain experts to rate the interestingness of the resultant rules. We finally analyze the data to determine which metric(s) best agree with expert judgments of interestingness. We find that Merceron & Yacef (2008) were right. Lift and cosine are good indicators of interestingness. In addition, the Phi Coefficient, Convinction, and Jaccard also turn out to be good indicators of interestingness.

Keywords

Association Rules, Interestingness, Cosine, Phi Coefficient, Human Rating

1. INTRODUCTION

In recent years, Association Rule Mining has become a central method in the field of Educational Data Mining. It plays a prominent role in reviews of the field, including reviews by Romero & Ventura (2007, 2010), Baker & Yacef (2009), Scheuer & McLaren (2012), and Baker & Siemens (in press), referred to this method as a core type of relationship mining. In Association Rule Mining, algorithms search for patterns where a set of values of variables (the "if-clause") predict another variable's value (the "then-clause"). (It is also possible for a then-clause to have multiple variables, but less common).

In these reviews, it was noted that Association Rule Mining has several potential applications. It is excellent for generating

hypotheses to study further, and for finding unexpected connections within data.

Association Rule Mining has been applied to several applied research problems within the educational data mining community and related research communities. Some notable examples include: Freyberger and colleagues have used association rules to analyze interactions between students and intelligent tutoring systems, in order to find models that predict student's success (Freyberger, Heffernan & Ruiz, 2004); Lu (2004) used association rules to match suitable learning materials based on each student learning needs; Garcia, Romero, Ventura & De Castro (2009) have used association rules to make recommendations to instructors for how to improve the effectiveness of a web adaptive course; in a similar example, association rules have been implemented to provide information to teachers about students' behavior in intelligent tutoring systems (Ben-Naim, Bain & Marcus, 2009).

A subset of Association Rule Mining, Sequential Pattern Mining, has also seen extensive use in the educational data mining community, as well as being highlighted in reviews of the field (e.g. Romero & Ventura, 2007; Baker & Yacef, 2009; Scheuer & McLaren, 2012; Baker & Siemens, in press). Sequential Pattern Mining consists of finding association rules where the contents of the then-clause occur temporally after the contents of the if-clause (Agrawal & Srikant, 1995). In the case of educational data mining, Kinnebrew, Loretz, & Biswas (2012) have used Sequential Pattern Mining to analyze how students engage in the different activities within an intelligent tutoring system over time, in particular studying the different sequences seen in high-performing and low-performing students. In another example, Perera et al. (2009) used Sequential Pattern Mining to analyze how groups of students use online tools, studying the work patterns of successful and unsuccessful groups, in order to provide feedback to the groups about their work strategies. One more example in education comes from the research done by Romero, Ventura, Delgado & De Bra (2007), who integrated Sequential Pattern Mining techniques in an algorithm within an educational system in order to provide personalized recommendations to students about possible links they should explore.

Association rules are typically initially selected on the basis of rules' confidence and support (Agrawal & Srikant, 1995). The support of a rule corresponds to the percentage of data points that contain both the if-clause and then-clause. The confidence of the rule is expressed as the percentage of data points that contain both

the if-clause and also includes the then-clause, divided by the number of data points that contain the if-clause (Garcia, Romero, Ventura & Calders, 2007).

However, the combination of support and confidence is insufficient to select good association rules. By definition, support and confidence find variable values that are frequently seen together. As such, these metrics often end up selecting combinations of variable values that are trivially associated, such as finding that students who take advanced biology probably took introductory biology, or finding that students who fail a course's exams fail the course as well.

What is desirable is to instead find association rules that are novel, that are surprising, that are unexpected. Frequently, after rules are filtered by looking for all rules with a minimum support and confidence, the next step is to use an alternate metric that can give some indicator of novelty; that can determine if an association rule is *interesting*.

To this end, researchers have tried to decide which metrics best capture an association rule's interestingness, both in general (Tan, Kumar & Srivastava, 2004), and in the specific case of educational data mining (Meceron & Yacef, 2008). Merceron and Yacef (2008) recommend Lift/Added Value (Lift and Added Value are mathematically equivalent) and Cosine as excellent interestingness measures for educational data because their meaning is easily understood even to people not expert in data mining (e.g., teachers, school administrators, and so on); in addition, Cosine does not depend on the data set size. In particular, they recommend that researchers consider an association rule to be interesting if it has a high value for either of these measures.

Moreover, there are additional metrics identified that have the potential to measure interestingness. Tan et al. (2004) review the potential candidates for an interestingness measure, finding over twenty in the published literature. Their list includes lift and cosine, but also includes the Phi coefficient, Goodman-Kruskal's, the Odds ratio, Yule's Q, Yule's Y, Cohen's Kappa, Mutual information, the J-Measure, the Gini Index, Laplace, Conviction, Piatetsky-Shapiro, Certainty Factor, Added Value, Collective strength, Jaccard, and Klosgen. Such variety of possible interestingness measures has made it complicated to identify which one is the most appropriate.

Further complicating the matter of choosing an appropriate interestingness measure (or measures) is the fact that the research on interestingness measures has thus far been mathematical or intuitive: interestingness measures have been selected based on their mathematical properties, and in some cases based on the intuitive perceptions of expert data miners.

In this paper, we consider an alternate strategy for selecting interestingness measures: using data mining to determine which interestingness measure is best, based on expert judgments of interestingness. In other words, instead of selecting a metric formally or intuitively, we can actually collect data on which association rules are seen as being the most interesting by domain experts, the population that could best take advantage of new hypotheses and unexpected findings in a domain. We then analyze this data to determine which metrics, or combination of metrics, best matches the domain experts' perception of specific rules' interestingness.

In the following sections, we take real data from online learning. We then distill association rules for that data relevant to established research questions in the field. We then ask three domain experts to rate the interestingness of the resultant rules. We finally analyze the data to determine which metric(s) best agree with expert judgments of interestingness. In doing so, we will explicitly compare our findings to claims in Merceron & Yacef (2008) as to which metrics best represent interestingness.

2. Method
2.1 Data
In order to study domain experts' assessments of which association rules are interesting, we generated association rules from real student data, relevant to established research questions in the field. We use domain experts, under the hypothesis that what experts consider interesting may be different than what novices consider interesting (and we believe that finding rules that are interesting for an expert is a more valuable use of association rule mining, though opinions could differ). We use genuine data to create these rules rather than simulated data, due to the concern that the metrics that predict the interestingness of genuine data may not be the same as the metrics that predict interestingness in simulated data. This would be a particular concern if the simulated data were to produce association rules that were actually false; and using generic operators would eliminate the potential to leverage domain expertise.

To this end, we used models that assess student affect and disengaged behaviors within a widely-used online learning environment, to examine association rules about the relationships between student´s affect and disengaged behaviors. The study of student disengagement and affect has been a research topic of considerable interest to researchers in EDM and related fields. Sabourin, Rowe, Mott, & Lester (2011) have analyzed the relation between engaged and disengaged behaviors with positive and negative affective states in students while interacting with a learning system, finding that different patterns of affect correlate to engaged and disengaged behaviors. Hershkovitz, Baker, Gobert, & Nakama (2012) have found evidence that boredom mediates between the student´s tendency to avoid novelty and off-task behavior. Baker, D'Mello, Rodrigo & Graesser (2010) find that gaming the system is often preceded and followed by boredom. Chauncey & Azevedo (2010) show a relationship between induced affect and cognitive engagement/meta-cognition, leading to differences in performance.

These rules were generated from data from the ASSISTments system (Razzaq, Heffernan, Feng & Pardos, 2007). ASSISTments is an educational web-based system that provides students with intelligent tutor-based online problem solving activities, while providing teachers with dynamic formative assessment of the students' mathematical abilities. The system has been found to be effective at enhancing student learning. (Razzaq et al., 2007), and is used by over 50,000 students a year. Figure 1 shows a screen shot of the ASSISTment system.

Data was obtained from the logs of 724 middle school students from the Northeastern United States, who answered different problems that measure 70 different mathematics skills. Within this data set, there were a total of 107,382 problems solved by students within the ASSISTment software. Student actions in this data set were classified in terms of affective states and disengaged behaviors from machine-learned affect and behavior detectors. The detectors inferred if the student:

- was detected as being bored or not,
- was detected as being concentrated or not,
- was detected as being frustrated or not,
- was detected as being confused or not,
- was detected as being on task or off task,
- was detected as being gaming the system or not,

The following additional features were also included in the data set:
- the student providing a correct answer
- the student providing an incorrect answer
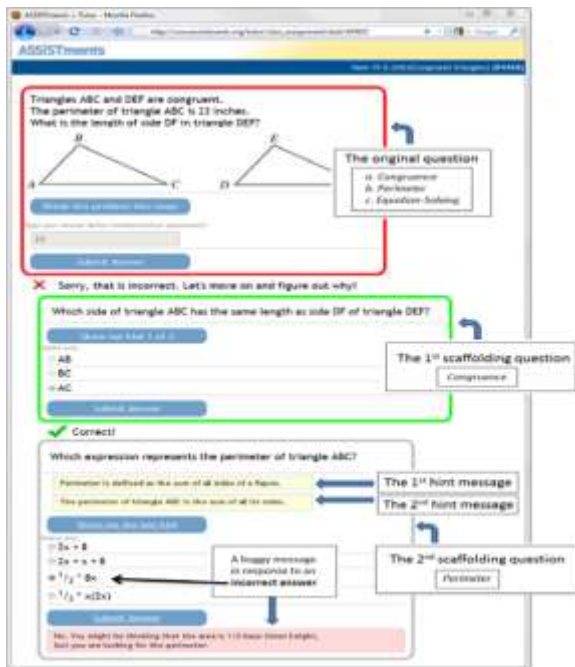- the student asking for a hint.

Figure 1. Example of an ASSISTment item

The detection of these binary categories of affective states/behaviors was done using the detectors presented in Pardos et al (2013). These detectors were developed by distilling features of the students´ interactions with the software, and synchronizing those features with field observations collected by two trained coders during the students' interactions with ASSISTments. The log data entry and the field observations were synchronized and segmented in 20 second windows to develop the detectors.

Detector performance was evaluated using student-level cross-validation (5-fold). All detectors performed substantially better than chance, being able to distinguish each affective state/behavior between 63%-82% of the time (the A' statistic), performance that was 23%-51% better than chance (the Kappa statistic). The detectors provide confidence values of the probability that an affective state or behavior occurred. To support the association rule mining analyses discussed below, we convert these probabilities into binary predictions, using a 50% probability threshold (the Kappa values listed above represent the model goodness when this transformation is used). Pardos et al. (2013) and San Pedro et al. (2013) provide a detailed description of the detectors and their use in multiple discovery with models analyses. Table 1 summarizes the frequency and proportion of

each of these behaviors/affective states. Regarding table 1, it shows some of the average confidences are higher than what should be expected. Here we point out that, as it is indicated in San Pedro et al. (2013), some detectors used in the current research presented some systematic error in prediction, which impacted in a higher or lower average confidence of the resultant models compared to the proportion of the affective states in the original data set. This type of bias does not affect correlation to other variables since relative order of predictions is unaffected, neither affects A' or Kappa, but it can reduce model interpretability. We did not rescaled the detectors, as it is proposed in Pardos et al. (2013) since we are considering final binary predictions from the detectors, where Kappa is the relevant goodness statistic, we use non-rescaled confidences in this paper.

The association rules were created in way that each rule described how a set of the affective states/ behaviors seen in the first attempt at a problem was associated with a single affective state or behavior in the student's first action on the next problem. In this analysis, simple association rules were created that predicted affect or behavior from a combination of the elements at the previous action.

2.2 Generation of Association Rules
Association rules were created using the arules package (Hahsler, Gruen, & Hornik, 2005; Hahsler et al., 2009) in R version 2.15.2 (R Development Core Team, 2012). In specific, the apriori algorithm implemented within the arules package was used to discover the association rules (Agrawal et al., 1994). This process in R resulted in a list of 431,768 rules, for which support, confidence, and lift were automatically computed. A total of 120 different association rules were selected from the 431,768 measures obtained; these 120 rules were selected to be the rules with the highest support and confidence that were representative of different numbers of elements in the if-clauses and were representative of all variables in the then-clauses of the rules. All rules selected had a support over 0.05 and confidence over 0.1; most were considerably higher.

Table 1. Frequency and average confidence for each affective /behavioral state in the data

	Frequency	Percentage	Rescaled Average Confidence
Bored	52080	48.49	0.2469
Engaged concentration	47854	44.56	0.5160
Frustrated	10929	10.17	0.0988
Confused	20308	18.91	0.1372
Off-Task	18135	16.88	0.0406
Gaming the system	9805	9.13	0.0182
Used Hints	16216	0.15	
Answer was Correct	45116	0.42	

2.3 Expert Rating of Association Rules
Once the rules had been created, they were rated for their interestingness by domain experts. In specific, four scientific researchers with scientific expertise in the areas of affect and disengagement in online learning. They rated the extent to which each of the 120 association rules was "scientifically interesting".

A Likert scale was used in rating, ranging from 1 to 5, where 1 was "*Not at all interesting*" and 5 was "*Extremely interesting*". Based on these expert ratings, the average inter-rater interestingness value was calculated for each rule, giving an indicator of how interesting the experts found each rule. In addition, measures of the degree of agreement between the experts were calculated, and are discussed in Section 3.1.

2.4 Computing Association Rule Metrics

After the expert coders rated the 120 selected association rules, additional interestingness measures from Tan et al. (2004) were computed in Microsoft Excel. The following metrics were computed for each rule:

- Phi Coefficient
- Cosine
- Piatetsky-Shapiro
- Jaccard, Laplace
- Certainty Factor
- Added Value
- Klosgen

- Odds Ratio
- Cohen's Kappa
- Gini Index
- Conviction
- J Measure
- Collective Strength

In addition, non-standard metrics were created, under the hypothesis that these metrics might also capture some key aspects of expert perception of interestingness in this domain, where an expert might be looking for evidence of successful students or unsuccessful students:

- The number of elements in a rule with values equal to Yes, Correct, and/or On task behavior.

- The number of elements in a rule with values equal to No, Incorrect, and/or Off task behavior.

3. Results

The findings of the research are presented in this section. First, examples of some association rules rated as very interesting, not interesting, and with mixed rating, are presented. Then, results about the inter-rater agreement are included. Finally, correlations between the experts´ ratings and the association metrics are described, and regression models are presented that make combined predictions of expert ratings from a combination of association metrics.

3.1 The Most and Least Interesting Rules

As discussed in the previous section, each rule was rated for perceived interestingness by each of the four expert coders. Below, we present some of the most interesting and least interesting rules, in their perception. Note that each rule represents a transition from time t_1 (left side of rule) and time t_2 (right side of rule). Note also that rules are presented with the exact same operators as generated by the algorithm, which means that some redundancy is present.

The most interesting rules according to the experts (e.g. the rules with the highest average interestingness) were:

{Got incorrect answer, not frustrated} → {Gaming the system}

{Gaming the system, bored, not in engaged concentration, got the incorrect answer and did not request a hint}} → {Confused}

{Off-task, confused, not bored, got the correct answer, and did not request a hint} → { Off-task}

The following rules were rated as least interesting by the experts (in terms of average rating).

{In engaged concentration, did not request a hint, not bored or frustrated or confused or off-task or gaming the system} → {Off-task}

{In engaged concentration, got correct answer, did not request a hint, not frustrated or confused or off-task or gaming } → {Not gaming the system}

{In engaged concentration, got correct answer, did not request a hint, not bored or frustrated or confused or off-task or gaming the system} → {Not frustrated}

However, some rules obtained a high rating from two experts but low rating from the other two:

{Got incorrect answer, did not request a hint, not in engaged concentration or frustrated or off-task or gaming} → {Confused}

{Got incorrect answer, did not request a hint, bored, not concentrated or frustrated or confused or gaming} → {Not being frustrated}

The first of these rules was rated as not interesting by two members of the same research group (experts 2 and 3 below) but rated as very interesting by two members of other research groups. The second rule, however, was rated highly by experts 1 and 2, who belong to different research groups, and it was rated as less interesting by experts 3 and 4.

3.2 Agreement among raters

Though there was generally good agreement between experts, some rules led to disagreement between the coders in terms of interestingness, as shown above. To see the degree of agreement (and to evaluate whether it was feasible to use these expert codes as a basis for studying which metrics best evaluate interestingness), we checked to make sure there was consistency among the four domain experts, using multiple metrics. The estimated Cronbach´s Alpha coefficient for the consistency in rating among the four experts was 0.845, which indicates there is a high covariation among experts in their ratings of interestingness of different rules. The general Intraclass Correlation for the agreement among the four raters was 0.487, which indicates a moderate agreement among the experts (Bartko, 1966). It is worth noting that while Cronbach's Alpha expresses a measure of covariation in the ratings among experts, Intraclass Correlation estimates reliability as the magnitude of disagreement/agreement among the experts (Hallgren, 2012). Hence, the difference among both measures reflects a discrepancy of what each statistic estimates. In the context of our results, these statistics mean that while the experts showed consistency in the way they rated each rule, only a moderate agreement among experts was achieved.

Additionally, Spearman correlation coefficients were calculated to determine the degree of agreement between each pair of experts based on their rating of interestingness to the 120 association rules. Results of the Spearman correlation coefficients are included in the table 2, which indicate there was a significant degree of consistency among the four experts. As this table shows, all four experts had a reasonable degree of consistency, but experts 1 and 2 showed higher agreement with each other, while experts 3 and 4 had higher agreement with each other. Overall, there was moderate to high agreement among the experts in their rating of interestingness of different association rules.

Table 2. Spearman correlation coefficients among experts.

	Expert 1	Expert 2	Expert 3	Expert 4
Expert 1	1			
Expert 2	.744	1		
Expert 3	.548	.590	1	
Expert 4	.580	.516	.674	1

3.3 Correlation between expert judgments and association metrics

Though there was some structure in terms of agreement (e.g. coders 1 and 2 agreed more, and coders 3 and 4 agreed more), the overall agreement between coders was sufficient to create a single metric representing the interestingness of each rule. This metric was created by taking the average of the four coders' ratings for each rule.

Next, Spearman correlation coefficients were calculated to analyze the degree of association between the expert ratings of interestingness and the metrics of interestingness computed in R (R Development Core Team, 2012; Gamer et al., 2012; Fletcher, 2010) and Excel. The resultant correlation coefficients are presented in Table 3. This table shows that the experts' ratings of interestingness were highly correlated with some association rule measures. 7 of the 24 metrics were more highly correlated with the expert ratings of interestingness than the experts' ratings of interestingness correlated with one another, on average. The most highly correlated metrics were Jaccard (r= -0.838), Cosine (r= -0.835), and Support (r = -0.82). As shown in Table 3, the metrics that agreed least well with expert ratings of interestingness were Added Value (r= -0.014) and Kappa (r=-0.029). Merceron & Yacef's (2008) recommendation to use Cosine agrees with our findings here; their recommendation to use Lift does not, at least initially. But they recommend using these metrics in concert, not individually. In the next section, we consider what mixture of metrics best predicts human judgments of interestingness.

3.4 Predicting Expert Perception of Interestingness from a Combination of Metrics

After looking at the predictive power of each metric, taken individually, we built a model that predicted expert judgments using a combination of metrics. Doing so may allow us to create a meta-metric that could be a better representation of interestingness than any single metric by itself.

A linear regression model was created to predict the average expert judgment of interestingness. For this full model, no variable selection was conducted – e.g. all metrics listed above were incorporated into this model. Although the model had statistically significant fit statistics (r= 0.938, r^2 = 0.879, Cross-validated r^2 =0.73, AIC = 123.2702, BIC = 181.8075; F(19, 100) = 38.24, p-value = 0.001), it also had a high degree of multicollinearity among the predictors, measured by the Variance Inflation Factor (VIF). Multicollinearity can lead to over-fitting, as well as making it very difficult to interpret the estimated values for the regression coefficients and their standard errors. This model is reported in table 4.

Table 3. Spearman correlation among inter-rater average and association rules metrics.

	Correlation to Inter-Judge Average	p-value
Jaccard	-0.838	<0.001
Cosine	-0.835	<0.001
Support	-0.82	<0.001
Certainty Factor	0.775	<0.001
Confidence	-0.747	<0.001
Laplace rule	-0.647	<0.001
Count var. of 1´s	-0.609	<0.001
Conviction	-0.432	<0.001
Count var. of 0´s	-0.368	<0.001
Klosgen	-0.327	<0.001
Gini Index	-0.32	<0.001
Odds Ratio	-0.31	0.001
Yule's Q	-0.31	0.001
Yule's Y	-0.31	0.001
Piatetsky-Shapiro	-0.303	0.001
J Measure	-0.303	0.001
Collective Strength	-0.298	0.001
Phi Coefficient	-0.29	0.001
Lift	0.202	0.027
Kappa	-0.029	0.754
Added Value	-0.014	0.876

Table 4. Regression model with all association rules metrics and counting variables as predictors

Predictor	Coeff	S.E.	T	P-val	VIF
Intercept	106.44	33.13	3.21	0.001	
Count var. of 1´s	-0.042	0.097	-0.43	0.664	4.2
Count var. of 0´s	-0.01	0.081	-0.13	0.896	13.3
Support	44.085	19.83	2.22	0.028	2375.2
Confidence	0.899	1.617	0.55	0.579	230.6
Lift	-28.56	13.46	-2.12	0.036	2117.1
Phi Coefficient	47.673	26.51	1.79	0.075	1422.1
Cosine	34.443	47.58	0.72	0.470	24213.7
Piatetsky Shapiro	-80.69	509.0	-0.15	0.874	18302.8
Jaccard	-108.6	57.81	-1.87	0.063	16274.3
Laplace	-10.62	9.39	-1.13	0.260	3832.8
Certainty Factor	-17.37	8.473	-2.05	0.042	347.8
Added Value	49.036	36.95	1.32	0.187	2257.6
Klosgen	-78.83	187.3	-0.42	0.674	6543.2
Odds Ratio	-0.235	4.097	-0.05	0.954	10700.1
Kappa	172.97	70.26	2.462	0.015	6245.3
Gini Index	-437.2	283.4	-1.54	0.126	712.7
Conviction	-2.369	5.516	-0.42	0.668	7758.5
J Measure	1038.7	502.4	2.068	0.041	1851.9
Collective Strength	-68.02	36.70	-1.85	0.066	8265.7

A second linear regression model was tested including just statistically significant association metrics as predictors with small multicollinearity among them. The predictors excluded from this analysis were: Support, Confidence, Piatesky Shapiro, Jaccard, Laplace, Certainty Factor, Added Value, Klosgen, Odds Ratio, Kappa, Gini Index, J Measure, and Collective Strength. Those omitted predictors presented moderate to high correlations with one or more association metrics included in the model summarized in table 5. The criteria for exclusion were high correlations among the predictors that, consequently, resulted in VIF values higher than 10 for a given model.

Results of this second regression model showed that two association rule metrics –Lift and Conviction– had a positive prediction coefficient, while other two metrics – the Phi Coefficient and Cosine– had a negative coefficient. The model fit statistics were statistically significant and explained almost as much of the variance as the full model, which achieved a substantially higher cross-validated correlation ($r = 0.902$, $r^2 = 0.814$, Cross-validated $r^2 = 0.791$, AIC = 144.4186, BIC = 161.1436; F = 126.4, $df_1 = 4$, $df_2 = 115$, p-value = 0.001). Table 5 summarizes the second regression model. The lower values of BIC in the second model confirm it is a better and more simple model compared with the former one.

Table 5. Regression model with association rules metrics with restriction for multicollinearity

Predictor	Coeff	S.E.	T	P-val	VIF
Intercept	0.404	1.023	0.395	0.6937	
Lift	3.848	0.790	4.870	<0.001	5.477
Phi Coef.	-11.179	2.220	-5.034	<0.001	7.491
Cosine	-5.783	0.585	-9.880	<0.001	2.752
Conviction	0.469	0.116	4.013	<0.001	2.616

Although Jaccard presented the highest correlation with the inter-rater average score, it also presented a very high correlation with many other metrics, including Cosine ($r = 0.96$). Thus, many models that included Jaccard also presented a high degree of multicollinearity among the predictors; as a consequence, Jaccard was excluded in the combined model presented in table 5. Table 6 demonstrates a model similar to the model in table 5 but replacing Cosine with Jaccard. The model in this case was not better in terms of multicollinearity and was only slightly better in terms of goodness-of-fit ($r = 0.908$, $r^2 = 0.825$, Cross-validated $r^2 = 0.81$, AIC = 137.5014, BIC = 164.2263; F(4, 115) = 135.6, p-value = <0.001).

Table 6. Regression model including Jaccard instead of Cosine

Predictor	Coeff	S.E.	T	P-val	VIF
Intercept	0.547	0.986	0.556	0.579	
Lift	3.502	0.778	4.496	<0.001	5.638
Phi Coef.	-7.528	2.370	-3.177	0.002	9.038
Jaccard	-8.896	0.847	-10.49	<0.001	2.780
Conviction	0.207	0.121	1.721	0.088	2.945

Regression models were also computed for each individual metric used in the combined models. The results, which are summarized in table 7, show that single-feature models presented less

desirable fit statistics (i.e., r^2, AIC, and BIC) than the combined model. The model including just Jaccard as predictor has the best fit statistics among the single-variable models (with Cosine close behind), but the combined model is still superior.

.

Table 7. Regression models with single predictors

Predictor	Coeff – Intercepet (S.E.)	p	Fit Stats.
Lift	2.81*Lift – 0.80 (0.729)	<0.001	$R^2 = 0.112$ $CV\text{-}R^2 = 0.074$ AIC = 326.43 BIC = 334.79
Phi Coefficient	-6.744*Phi + 2.78 (1.754)	<0.001	$R^2 = 0.111$ $CV\text{-}R^2 = 0.098$ AIC = 326.56 BIC = 334.93
Cosine	-7.72*Cosine + 5.39 (0.387)	<0.001	$R^2 = 0.771$ $CV\text{-}R^2 = 0.754$ AIC = 163.71 BIC = 172.07
Conviction	-0.69*Conviction + 3.2 (0.152)	<0.001	$R^2 = 0.149$ $CV\text{-}R^2 = 0.119$ AIC = 321.24 BIC = 329.61
Jaccard	-11.56*Jaccard + 4.84 (0.552)	<0.001	$R^2 = 0.787$ $CV\text{-}R^2 = 0.779$ AIC = 154. 84 BIC = 163.21

4. Discussion and Conclusions

As seen in this paper, several standard association rule metrics can predict human expert ratings of interestingness of an association rule. Most commonly used interestingness metrics showed statistically significant correlations with the experts' ratings of interestingness, but not all of them were included in the final combined model given the high common variation among them. The best metrics – Jaccard, Cosine, and Support – achieved an absolute correlation higher than 0.80 with the average expert human judgment, which is higher than the average correlation of the ratings between experts. Hence, we see that these association metrics are a good substitute for human ratings of interestingness

In particular, our findings agree with Merceron and Yacef (2008) that Cosine and Lift are useful, as they were successful predictors in the final combined model in this data set. Taken individually, Cosine was good predictor, while Lift explained considerably less variance. The association metric Cosine consistently had a high negative correlation with the raters' scores of interestingness and significantly predicted expert ratings of interestingness, both in a single-predictor model and in combination with other association metrics. The association metric Lift had a positive correlation and significantly predicted the average score of interestingness among the experts in combination with other metrics and in a single-predictor model; however, Lift was relatively weak compared to other metrics when taken by itself.

However, one surprise is that Cosine, while important in both our findings and in Merceron & Yacef, was correlated to interestingness in the negative direction in our findings (i.e. low, while Merceron & Yacef recommend looking for high Cosine). This finding is surprising, and merits further study. One possibility is that once support and confidence are accounted for,

then interestingness links in some ways to rarity. Perhaps that is not surprising –facts that are already known are not particularly interesting– but it does show that the association rule mining conception of interestingness may not quite match intuitive notions of this construct. In our view, this finding is itself impressive. In general, this result suggests that Cosine is indeed important, but may reflect interestingness in a different way than previously understood.

In addition, other association rule metrics – the Phi Coefficient, Conviction and Jaccard – that have not been widely used in educational data mining also explained a significant proportion of the variance in the combined model and a in single-predictor models. Therefore, it might be useful to also consider these metrics in future research using association rule mining in educational data sets.

On the whole, results in this study show that the recommended metrics of interestingness proposed by Merceron and Yacef (2008) are useful, as well as other metrics not considered by those authors.

It is worth considering some limitations of this study. First, only linear correlations and linear regression models were considered. Although these approaches achieved good fit to the data, and explained much of the variance, it could be useful to consider models with non-linear relations among the association rules metrics and the expert ratings. Second, given the high correlation among different association rules metrics, other measures could be considered as alternative predictors of the inter-rater score of interestingness instead of the four measures chosen in the final regression model reported. Third, this paper represents a single analysis in a single educational research domain. Results might vary in a different educational research domain, or indeed outside of education. However, the fact that Cosine and Lift were prominent both in our models and in the recommendations in Merceron & Yacef (2008) is a positive sign, given that their work involved a very different area of educational research.

Overall, the use of association mining to understand complex and interesting relations among different variables is a method with a lot of potential in educational data mining research. Association rules can be understood at an intuitive level, and can provide useful information for a variety of stakeholders who are not experts in EDM, including students, teachers, administrators, and policy makers. However, given the huge numbers of association rules that can be generated, it is important to try to filter not just by support and confidence, but by interestingness as well. By using the metric or combination of metrics that matches an intuitive conception of interestingness, we can provide the most interesting information to users of association rules first, improving the efficiency of this method.

Acknowledgements

5. References

[1] Agrawal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference Very Large Data Bases* (pp. 487-499)

[2] Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., & Verkamo, A. I. (1996). Fast Discovery of Association Rules. *Advances in knowledge discovery and data mining*, *12*, 307-328.

[3] Agrawal, R., & Srikant, R. (1995, March). Mining sequential patterns. In *Proceedings of the Eleventh International Conference on Data Engineering* (pp. 3-14). IEEE.

[4] Baker, R. S., D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies, 68*(4), 223-241.

[5] Baker, R., & Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining, 1* (1), 3-17

[6] Baker, R., Siemens, G. (in press) Educational data mining and learning analytics. To appear in Sawyer, K. (Ed.) *Cambridge Handbook of the Learning Sciences: 2nd Edition*.Cambridge, UK: Cambridge University Press.

[7] Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological reports, 19*(1), 3-11.

[8] Ben-Naim, D., Bain, M., & Marcus, N. (2009, July). A User-Driven and Data-Driven Approach for Supporting Teachers in Reflection and Adaptation of Adaptive Tutorials. Paper presented at *Educational Data Mining*, (pp. 21-30)

[9] Chauncey, A., & Azevedo, R. (2010, January). Emotions and motivation on performance during multimedia learning: how do i feel and why do i care?. Paper presented at *Intelligent Tutoring Systems* (pp. 369-378). Springer Berlin Heidelberg.

[10] Fletcher, T. D. (2010). Psychometric: Applied Psychometric Theory. R package version 2.2. Retrieved from URL http://CRAN.R-project.org/package=psychometric

[11] Freyberger, J. E., Heffernan, N., & Ruiz, C. (2004). *Using association rules to guide a search for best fitting transfer models of student learning* (Doctoral dissertation). Worcester Polytechnic Institute, Worcester, MA.

[12] Garcia, E., Romero, C., Ventura, S., & Calders, T. (2007, September). Drawbacks and solutions of applying association rule mining in learning management systems. In *Proceedings of the International Workshop on Applying Data Mining in e-Learning* (pp. 13-22).

[13] Garcia, E., Romero, C., Ventura, S., & De Castro, C. (2009). An architecture for making recommendations to courseware authors using association rule mining and collaborative filtering. *User Modeling and User-Adapted Interaction*,*19*(1-2), 99-132

[14] Hahsler, M., Grün, B., Hornik, K., & Buchta, C. (2009). Introduction to arules–A computational environment for

mining association rules and frequent item sets. *The Comprehensive R Archive Network*.

[15] Hahsler, M., Gruen, B., & Hornik, K. (2005). arules - A Computational Environment for Mining Association Rules and Frequent Item Sets. *Journal of Statistical Software. 14*(15). URL: http://www.jstatsoft.org/v14/i15/.

[16] Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. Tutorials in quantitative methods for psychology, 8(1), 23-34.

[17] Hershkovitz, A., Baker, R. S., Gobert, J., & Nakama, A. (2012). A Data-driven Path Model of Student Attributes, Affect, and Engagement in a Computer-based Science Inquiry Microworld. In *Proceedings of the International Conference on the Learning Sciences*.

[18] Kinnebrew, J. S., Loretz, K. M., & Biswas, G. (2012). A contextualized, differential sequence mining method to derive students' learning behavior patterns. *Journal of Educational Data Mining*, *5*(1), 190-219

[19] Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2012). irr: Various Coefficients of Interrater Reliability and Agreement. R package version 0.84. Retrieved from URL http://www.R-project.org.

[20] Merceron, A., & Yacef, K. (2008). Interestingness Measures for Associations Rules in Educational Data. *Educational Data Mining, 8*, 57-66.

[21] Lu, J. (2004). Personalized e-learning material recommender system. Paper presented at *International conference on information technology for application* (pp.374–379)

[22] Pardos, Z. A., Baker, R. S., San Pedro, M. O., Gowda, S. M., & Gowda, S. M. (2013, April). Affective states and state tests: Investigating how affect throughout the school year predicts end of year learning outcomes. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge* (pp. 117-124). ACM.

[23] Perera, D., Kay, J., Koprinska, I., Yacef, K., Zaiane, O. (2009) Clustering and Sequential Pattern Mining of Online Collaborative Learning Data. *IEEE Transactions on Knowledge and Data Engineering, 21*, 759-772.

[24] R Development Core Team. (2012). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from URL http://www.R-project.org.

[25] Razzaq, L., Heffernan, N., Feng, M., & Pardos, Z. (2007). Developing Fine-Grained Transfer Models in the ASSISTment System. *Journal of Technology, Instruction, Cognition, and Learning*, *5*(3), 289-304.

[26] Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications, 33*(1), 135-146.

[27] Romero, C., Ventura, S., Delgado, J. A., & De Bra, P. (2007). Personalized links recommendation based on data mining in adaptive educational hypermedia systems. In E. Duval, R. Klamma, and M. Wolpers (Eds.). *Creating New Learning Experiences on a Global Scale* (pp. 292-306). Berlin: Springer.

[28] Romero, C., & Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art. *IEEE transactions on systems, man and cybernetics. Part C, Applications and reviews*, *40*(6), 601-618.

[29] Sabourin, J., Rowe, J. P., Mott, B. W., & Lester, J. C. (2011, January). When off-task is on-task: the affective role of off-task behavior in narrative-centered learning environments. Paper presented at *the 15th International Conference In Artificial Intelligence in Education* (pp. 534-536).

[30] San Pedro, M. O. Z., Baker, R. S., Gowda, S. M., & Heffernan, N. T. (2013). Towards an Understanding of Affect and Knowledge from Student Interaction with an Intelligent Tutoring System. In *Proceedings of the 16th International Conference on Artificial Intelligence and Education*.

[31] Scheuer, O., & McLaren, B. M. (2012). Educational data mining. In N. M. Seel (Ed.). *Encyclopedia of the Sciences of Learning* (pp. 1075-1079). Springer US.

[32] Tan, P. N., Kumar, V., & Srivastava, J. (2004). Selecting the right objective measure for association analysis. *Information Systems, 29*(4), 293-313