Enhancing the Clustering of Student Performance using the Variation in Confidence

Ani Aghababyan[1] and Nicholas Lewkow[1] and Ryan S. Baker[2]

[1] McGraw-Hill Education
[2] University of Pennsylvania

Abstract. While prior research has typically treated student self-confidence as a static measure, confidence is not identical in all situations. We study the degree to which confidence varies over time using entropy, investigating whether high variation in confidence is more characteristic of highly confident or highly uncertain students, using data from 118,000 students working within 8 courses within the LearnSmart adaptive platform. We find that more confident students are also more consistent in their confidence. Confident students were more likely to answer correctly but also more likely to be overconfident, making unexpected mistakes. Finally, we develop interpretable clusters of students based on their confidence entropy, degree of over/underconfidence, and related variables.

Keywords: Confidence, Variance, Entropy, Adaptive Learning, Clustering

1 Introduction

As academic work becomes more and more reliant on remote or out of classroom participation, being able to account for learner characteristics that can affect their motivation and performance becomes vital. One of such motivational variables is one's self-belief expressed as self-confidence. As defined in research, confidence refers to one's beliefs in oneself and one's perceived abilities to succeed in a specific activity. Confidence refers to the strength of one's belief or the degree of confidence in a judgement.

Considerable research has shown connections between confidence and knowledge and has shown that confidence influences academic performance and outcomes [6]. Prior research has typically treated confidence as static, looking at overall levels of confidence, or confidence measured at a single time point. However, confidence is not identical in all situations, even for a given topic. Instead, it may be warranted to study the degree to which confidence varies over time and understand how variation in confidence relates to its overall levels. One possible way to represent how values vary is standard deviation, but this metric is poor at handling high variation and non-normal data. Other ITS researchers have used dynamic analyses to capture the variance of different student characteristics across contexts [9], but have not yet applied this method to study variation in student confidence over time. In this study, we investigate whether studying confidence entropy can enhance understanding of student performance. Confidence entropy could be beneficial to analysis of ITS in several ways, including the analysis of how it relates to learning and performance, and also through incorporating it into clusters of students that can be used to differentiate learning experiences for different groups of students. As such, this paper will investigate whether confidence entropy can be a meaningful contributor to a successful and predictive set of clusters.

Thus, the goal of this paper is to better understand student confidence entropy and how it can contribute to enhancing clustering of students into meaningful groups. More specifically we plan to investigate the following:

- Is there variance in student confidence (confidence entropy) reports or do students generally experience and report consistent confidence levels over time? How does this variance correlate to students' average confidence level and to performance more broadly?
- Does confidence entropy meaningfully contribute to student clustering based on performance?

We hypothesize that student confidence entropy will contribute to a better-quality set of clusters that has better goodness metrics and can better predict student accuracy.

2 Data Set & Content

Our data comes from the LearnSmart adaptive platform that offers personalized learning and self-assessment adaptive paths. The platform provides immediate feedback on the accuracy of each answer along with an explanation of the correct answer. If the learners understand the content and are able to demonstrate knowledge, they progress quickly. If the learners are lacking knowledge, they will need to spend more time working through the questions. Since the courses we studied did not have a final grade within the platform, we used students' overall accuracy score instead, which is the ratio of student's correctly answered questions to their total number of questions answered.

LearnSmart measures student confidence by asking the learner to self-report their confidence after each question. Immediate ratings of confidence are used to reduce the frequency of inaccurate responses as a result of recall bias due to retrospection [5]. With each question, the platform prompts the student to select one of the confidence buttons from a four-level confidence scale: "I know it", "Think so", "Unsure", "No Idea". The system records these reports as "3", "2", "1", "0" respectively.

For this study, we harvested data from eight courses from the Spring 2015 academic semester. We selected four humanities/social science courses and four physical/life science courses both with the largest usage. Additionally, we verified that the selected courses were comparable in terms of the number of total questions answered throughout the semester. Hence, the participants in the current study included 118,291 college students who took one of the eight courses taught via LearnSmart. Combined, these students completed 93,800,984 million questions.

3 Analysis 1: Confidence Entropy

Our first analysis attempts to better understand the variation of confidence, operationalized as Shannon entropy to find the distribution of confidence across its possible values [3]. The Shannon entropy equation provides a way to estimate the average minimum number of bits needed to encode a string of symbols, based on the frequency of the symbols. The entropy index is calculated by the following formula:

$$h(p_1, \ldots, p_a) = -\sum_{i=1}^{a} p_i \log(p_i)$$

When entropy is zero, the learner's confidence never varies. If the entropy is the maximum value (2 in our case – the base-2 logarithm of the four possible outcomes), the learner used the four confidence levels in the same proportion; there is maximal uncertainty as to the student's confidence. In other words, higher entropy means higher variability in confidence reported, and lower entropy indicates consistency in the learner's confidence. Note that entropy calculation does not consider order.

In LearnSmart, the average confidence entropy was 0.78, suggesting that students' self-reported confidence does not vary much. It was also more common for a student to have very low entropy (0.1 or lower), 8.5% of students, than very high entropy (1.8 or higher). Only 2% of students have exactly 0 entropy. Of those 2% of students, 93% reported the highest confidence for every question, just under 7% reported the middle two confidence levels, and only 3 reported the lowest confidence. Students with 0 entropy also had a higher average accuracy than those with entropy above 0.

Across the distribution of students, a student's confidence entropy correlated to several other metrics. More confident students varied less in their self-reports: Confidence entropy and average confidence were correlated at r= -0.66. The majority of the learners who report only one level of confidence are also the learners who report high confidence. There is, however, a second group of low-entropy students who have an average confidence in the middle. Relatedly, students who varied less in their self-report were more likely to answer correctly; there was s a negative correlation between confidence entropy and student accuracy (r= -0.35). Previous research [1] found that learners with higher accuracy are also likely to have higher average confidence, as well as a higher proportion of overconfidence.

4 Analysis 2: Confidence Entropy

4.1 K-Means Clustering Method

In our second analysis, we investigate whether students separate into relatively distinct groups based on their confidence entropy and other relevant performance characteristics captured by our set of variables. Thus, we use clustering analysis to build groups from the set of variables described below without including student accuracy, as we will correlate the clusters to this metric afterwards. We engineered performance features for each learner as input for this analysis. We chose the following features that are descriptive of a learner's performance but not dependent on their accuracy score:

1. Confidence entropy - variation of confidence described in analysis 1.
2. Overconfidence ratio - the proportion of incorrectly answered questions where the student reported the highest confidence.

3. Underconfidence ratio - the proportion of correctly answered questions where the student reported the lowest confidence.
4. Average confidence – mean confidence values for each student
5. Average number of questions answered - depending on the accuracy of their answer, each student may see 1 or more questions per learning objective
6. Average time taken to respond to the question (in 1/100ths of a second)
7. Average time taken to report confidence (in 1/100ths of a second) after prompt is displayed (after the student answers the question)

When using a clustering approach, several issues must be considered: the selection of clustering algorithm, the number of clusters, the statistical difference between clusters, cluster stability, and the interpretation of the clusters. Prior to including #5 as a feature, we verified that it was not a proxy for student accuracy ($r = -0.16$). To group the students into clusters we used the K-means clustering algorithm, which partitions the input into k distinct groups based on cluster centroid locations. We compared cluster consistency of the k-means to hierarchical clustering using silhouette validation [2] and k-means outperformed hierarchical clustering. Additionally, we used one-way ANOVA to compare the cluster mean for each feature in each cluster to make sure the average values of each cluster's features are significantly different from each other to render meaningfully different groups. Finally, for interpretation we came up with descriptive labels for each cluster and computed the average accuracy scores for each cluster to see whether the scores matched with our interpretation of the cluster performance based on the cluster characteristics.

4.2 K-means cluster Results

Our cluster features had different scales, so prior to using the k-means algorithm, we converted them to z-scores. We used within-set sum of squared error between points in clusters to choose our cluster number. As a result, 4 was the highest number of clusters where within-set sum of squared errors was decreasing substantially. It also has reasonably-sized clusters. We then conducted one-way ANOVA. All 4 of our clusters are significantly different from each other on all seven of the features (all feature means had $p < 0.01$). Given that the probability of this pattern being obtained by chance is 0.01^{42}, further post-hoc correction is not needed. Finally, in order to enhance interpretability, we assigned low, medium, high, very high to the average values for each feature within each cluster based on their value (see Table 1).

Table 1. Mean and (standard deviation) of features for each cluster.

Cluster Features	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Cluster size	8,744	31,491	31,915	46,141
Avg. Confidence	M/H 2.5 (0.3)	L 1.9 (0.3)	M/H 2.52 (0.2)	VH 2.9 (0.1)
Entropy	M 0.86 (0.4)	VH 1.12 (0.5)	H 1.11 (0.3)	L 0.29 (0.2)
Avg. Number of Questions per Chapter	VH 168 (61)	H 83 (40)	L 68 (29)	M 75 (35)
Avg. Answer Time	L 584 (797)	M 755 (1427)	VH 1124 (21K)	H 760 (1618)
Avg. Confidence Report Time	L 561 (790)	M 723 (1404)	VH 1098 (21K)	H 728 (1597)
Overconfidence Ratio	H 0.53 (0.3)	L 0.08 (0.1)	M 0.42 (0.2)	VH 0.87 (0.1)
Underconfidence Ratio	H 0.0023 (0.01)	VH 0.008 (0.05)	M 0.002 (0.01)	L 0.0005 (<0.001)
Cluster Label	Rapid and Thought-less	Realistically Inconsistent and Entropic	Realistically Knowledgeable and Thoughtful	Consistently Confident
Accuracy (not clustered on)	60.9 (16.1)	63.1 (14.3)	69.6 (12.1)	72.6 (11.9)

Based on Table 1, Cluster 1 students ("Rapid & Thoughtless") show signs of low effort, both on answering questions and reflecting on their confidence, spending the lowest amount of time thinking about the questions. These students do not appear to be realistic in their expectations as they have one of the highest average confidence scores, despite being required to complete double as many items as other students due to making many errors, and had the second-highest overconfidence ratio. Cluster 2 students ("Realistically Inconsistent/Entropic") have the lowest average confidence among all the Clusters, are least likely to be overconfident, and most likely to be underconfident (although underconfidence was still rare). In addition, these students have the highest entropy, varying considerably in their answers about their confidence, and using the middle confidence buttons more often than the extremes. Cluster 2 students spend an adequate amount of time answering questions but make many errors and have to answer more questions than average. Students in Cluster 3 ("Realistically Knowledgeable and Thoughtful") completed chapters with the fewest number of questions of any cluster but spent the longest responding to questions, suggesting that these students put extra effort into their work. These students are confident, but unlike Cluster 1, have high entropy in their reports of their confidence. Finally, Cluster 4 students ("Consistently Confident") have the highest confidence and the lowest entropy, mostly choosing the highest confidence button. On average these students answer a relatively small number of questions per chapter, due to successful performance. These students spend a moderate amount of time answering questions. However, these students had the highest overconfidence ratio by a substantial amount.

After labeling the groups, we correlated cluster membership to students' actual accuracy. As Table 1 shows, Cluster 4, Consistently Confident, had the highest mean accuracy, 72.6 (SD=11.88), a finding in line with their very high average confidence, low entropy, and a very high overconfidence ratio (overconfidence has been found to be associated with good academic performance [1]). Cluster 1, Rapid & Thoughtless, was the lowest performing group, unsurprising given their low average time spent per question and very high average number of questions answered per chapter. Clusters 2 and 3 were in the middle. These findings suggest that our clustering approach including entropy led to a meaningful and interpretable set of clusters that corresponded closely to student accuracy, despite not having actual accuracy information to cluster on.

5 Discussion & Future Work

In this paper, we explored students' confidence variability, operationalizing this as confidence entropy. We then examined the variability of students' self-confidence, and analyzed its relationship with performance and confidence strength. Our results show that average confidence and confidence entropy are highly negatively correlated, suggesting that more confident students are also more consistent in their confidence. More consistent confidence is also associated with higher accuracy. We then developed meaningful, interpretable clusters using entropy in combination with other behavior variables. Confidence and confidence entropy could be used in several ways in future ITS research and practice, including using time-based confidence entropy to predict if a student is losing interest or changing their outcome expectations. The clusters developed here could also be used to provide differential learning experiences. Future work should take context into account as well, investigating if some students' confidence varies more in specific situations or for specific material. As such, the work here is only a step towards better understanding how confidence shifts over time, and how this understanding can be used to improve learning.

References

1. Aghababyan, A., Lewkow, N., & Baker, R.: Exploring the asymmetry of metacognition. In: Proceedings of the Seventh International Learning Analytics & Knowledge Conference, pp. 115-119. ACM, New York, NY, USA (2017).
2. Rousseeuw P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics 20, 53-65 (1987).
3. Shannon, C.E.: Prediction and entropy of printed English. Bell Labs Technical Journal 30(1), 50-64 (1951).
4. Snow, E., Jacovina, M., Varner, L., Dai, J., McNamara, D.: 2014. Entropy: A stealth measure of agency in learning environments. In: Proceedings of the 7th International Conference on Educational Data Mining, pp. 241-244, Springer, Heidelberg, Berlin, Germany (2014)
5. Stone, A., Shiffman, S., Atienza, A., Nebeling, L.: The science of real-time data capture: Self-reports in health research. Oxford University Press (2007).
6. Zusho, A. Pintrich, P.R., Coppola, B.: Skill and will: The role of motivation and cognition in the learning of college chemistry. International Journal of Science Education 25(9), 1081–1094 (2003).