Combining Empirical and Machine Learning Techniques to Predict Math Expertise using Pen Signal Features

Jianlong Zhou
National ICT Australia
(NICTA), Australia
jianlong.zhou@nicta.com.au

Kevin Hang
University of New South
Wales, Australia
kevin.hang07@gmail.com

Sharon Oviatt
Incaa Designs
Bainbridge Island,
WA. 98110,
oviatt@incaadesigns.org

Kun Yu, Fang Chen
National ICT Australia
(NICTA), Australia
{kun.yu, fang.chen}
@nicta.com.au

ABSTRACT

Multimodal learning analytics aims to automatically analyze students' natural communication patterns based on speech, writing, and other modalities during learning activities. This research used the Math Data Corpus, which contains time-synchronized multimodal data from collaborating students as they jointly solved problems varying in difficulty. The aim was to investigate how reliably pen signal features, which were extracted as students wrote with digital pens and paper, could identify which student in a group was the dominant domain expert. An additional aim was to improve prediction of expertise based on joint bootstrapping of empirical science and machine learning techniques. To accomplish this, empirical analyses first identified which data partitioning and pen signal features were most reliably associated with expertise. Then alternative machine learning techniques compared classification accuracies based on all pen features, versus empirically selected ones. The best unguided classification accuracy was 70.8%, which improved to 83.3% with empirical guidance. These results demonstrate that handwriting signal features can predict domain expertise in math with high reliability. Hybrid methods also can outperform black-box machine learning in both accuracy and transparency.

Categories and Subject Descriptors

H.5.2 [**User Interfaces**]: Information interfaces and presentation - *user-centered design, interaction styles, evaluation/ methodology*

General Terms

Experimentation, Human Factors, Performance

Keywords

Multimodal Learning Analytics; Math Data Corpus; Pen Signal Analysis; Domain expertise; Prediction methodology; Machine learning

1. INTRODUCTION

Multimodal learning analytics (MMLA) is a newly emerging field that aims to assess and improve the learning process through automatic analysis of students' communication and activity patterns using natural modalities, such as video, audio, and digital pen [13]. Progress in MMLA can improve our understanding of how learning occurs, and how to stimulate it better through the development of new curricula, teaching strategies, and educational technologies. It can lead to new expertise prediction techniques that are relevant to both interpersonal and computer-mediated learning exchanges. It also will support more rapid feedback and responsive interventions, and facilitate learning in a more diverse range of students and contexts [8].

Of all modalities, writing and the use of pen tools is common work practice during classroom learning activities. Analysis of students' writing behavior potentially can play a valuable role in identifying differences among students in domain expertise [7,11], especially in areas like mathematics that include a great deal of spatial content like symbols and diagrams. The present research aims to develop automatic prediction of domain expertise in math by analyzing pen stroke signal features, such as number of strokes, stroke distance, duration, speed, and pressure. It accomplishes this in two steps: (1) by empirically analyzing how pen signal features relate to student expertise, and (2) by pursuing machine learning classification of expertise using the data when unguided, and also when guided by empirical results on the most valuable pen signal features and data partitioning.

To support the research outlined above, the Math Data Corpus was used, which contains multimodal data (digital pen, speech, vision) from collaborating students as they jointly solved math problems [13]. The dataset is structured to include problems varying in difficulty level (i.e., easy to very hard). Students' handwriting during problem solving also is coded by type of written representation (i.e., numeric, symbolic, diagrammatic, linguistic). As a result, the present work is able to explore expertise prediction as a function of problem difficulty and type of representation. It also examines the impact of data partitioning using these parameters on machine learning classification accuracies.

The paper is organized as follows: Section 2 presents related literature on multimodal learning analytics, handwriting and expertise, and pen interfaces as a means of stimulating human cognition. It also describes related theory, and outlines specific objectives and hypotheses of the present research. Section 3 presents the Math Data Corpus and process of collecting these data. Section 4 presents results of empirical analyses performed on pen stroke signal data. Section 5 summarizes different machine learning classification results when unguided, versus guided by empirical findings. Section 6 discusses and interprets these findings, and highlights future directions for research. Section 7 summarizes research conclusions.

MLA'14, November 12-16, 2014, Istanbul, Turkey.
Copyright © 2014 ACM 978-1-4503-0488-7/14/11…$15.00.
http://dx.doi.org/10.1145/2666633.2666638.

2. BACKGROUND & RESEARCH AIMS

Recent research has shown that pen interfaces can substantially facilitate students' ability to produce domain-appropriate ideas, solve problems correctly, and make accurate inferences about information [12]. The magnitude of these effects across different studies ranges from 9% to 38%. The cognitive advantages of pen input are largely due to their expressive power and flexibility in conveying all types of representation, especially spatial ones (e.g., diagrams, symbols). As a result, pen interfaces and multimodal ones that incorporate them are a promising tool for developing new educational technologies [9]. In the future, when students use digital pens during educational activities, corresponding learning analytics techniques will be required based on written signal or representational metrics [11].

2.1 Related Theory

During problem solving, major sources of working memory load include problem difficulty level, a person's expertise status, and the demands associated with planning and executing related communications. Working memory theory states that human cognition, including problem solving and learning, are highly constrained by attention and memory [1]. Expertise effectively expands working memory capacity by integrating more and higher-order information into a finite number of "chunks" [6]. For example, a chess master perceives individual moves as one integrated strategy play, which occurs following repeated activity when he or she begins to apprehend the isolated moves as a meaningful whole. As a result, becoming a domain expert circumvents the fixed-capacity limitation of working memory, and makes it possible to retain and retrieve more information relevant to solving a problem.

With respect to communication demands, limited-resource linguistic theories have characterized people as adaptively conserving energy at both the signal and lexical levels to minimize load. In chapter 1, Oviatt [9] summarizes:

> "From an evolutionary viewpoint, virtually all human communication systems have progressed toward greater simplicity and reduced length and human effort, including spoken, signed, and written languages."

For example, people adapt their speech production phonologically along a continuum to support intelligibility by their listener. They reserve effort associated with hyper-clear articulation for "high-risk" listeners [5]. At the lexical level, they also reduce noun phrase descriptions as they establish topical common ground with a listener, which enables them to reserve effort associated with lengthy descriptions for new communication partners or topics [3]. These economies in linguistic expression function to conserve working memory capacity, so communicators can maintain higher task performance. Evidence also indicates that domain experts use briefer noun phrase descriptions and technical terms than do non-experts [4, 15]. Domain experts in mathematics also write more compact domain-specific symbols than do lower-performing students [10].

2.2 Related Research Findings

With respect to past work on writing and dynamic pen signal features, Cheng and Rojas-Anaya [2] discovered that writing fluency is an indicator of math expertise. They showed that math experts have fewer and briefer pauses when writing math formulas, compared with non-experts. They interpreted this difference as a reflection that working memory consolidated into larger integrated chunks as math expertise developed.

In previous research involving the Math Data Corpus, Oviatt and Cohen [11] discovered that the dominant math expert in student groups were four-fold more active initiating math problem solutions, compared with non-experts. As a result, reliable identification of math expertise could be based simply on activity analysis, without any content analysis required. Furthermore, experts could be distinguished even more reliably as the difficulty level of math problems increased [11].

In other directly related work based on the Math Data Corpus, Ochoa et al. [7] reported that pen stroke writing speed is a predictor of math expertise:

> "How fast the student writes is an indicator of how certain the student is about how to solve the problem." And "The experts quickly established correspondence between external events and internal models of these events, which in turn correlates to the quick establishment/writing of a solution." (p. 589)

However, they did not conduct empirical analyses of writing behaviors that are significantly associated with expertise. In addition, their analyses involved automated techniques that were limited to assessing low-level geometric primitives of strokes (e.g., lines, circles), which provided only a rough approximation method. These automated techniques were not able to analyze higher-level shapes, or to examine whether differences in writing speed between experts and non-experts occurred across different types of written representation or problem difficulty levels.

2.3 Research Objectives

The pen signal features that students exhibit while solving math problems provide a window on the extent to which they have achieved communication and working memory economies that usually reflect domain expertise. The present research specifically investigates whether students who are domain experts in math, compared with non-experts, conserve communicative energy at the signal level when handwriting by making pen strokes with a shorter average distance, briefer average duration, lighter average pressure, and faster average speed. It also assesses whether experts make fewer pen strokes and write for a shorter time during a problem. Using the Math Data Corpus, handwriting is compared as students varying in expertise use a digital pen and paper to solve math problems. Analyses of pen stroke data are controlled for type of written representation and problem difficulty level, two parameters known to be important in evaluating expertise [11]. A second methodological objective of this work is to combine empirical with machine learning techniques to create a hybrid approach that can classify expertise more accurately and also function more transparently.

3. DATA COLLECTION

The data analyzed in this paper are from the Math Data Corpus (MDC) [13], which includes video, audio, and digital pen information from high-school students collaborating on solving math problems together. In addition, the lexical content of their speech and coded written representations are available for analysis. The multi-stream data in this corpus involve time-synchronised multimodal recordings from high school student groups as they collaborated to solve 16 math problems representing four difficulty levels (e.g., easy, moderate, hard, very hard problems).

Participants included 18 high school students, 9 female and 9 male, who ranged in age from 15 to 17 years old. All had recently completed Introductory Geometry at a local high school. They represented a range of geometry skills from low to high performers. Participants were divided into six groups, each containing three students. Each group worked on the math problems over two sessions, resulting in a total of twelve sessions of data. Within each group of three students, there was an expert and an appointed leader who coordinated the group's activity, with problems displayed on a desktop computer. All data was collected with ethics approval.

During data collection, video cameras were positioned to record various views of the students' work. Three video cameras provided a close-up view of each of the three students, one video camera provided a wide-angle view of the students working together, and the final video camera displayed a top-down view of the students' writing. An example of each of the three different views can be seen in Figure 1. To record students' speech patterns, digital audio recordings were procured using close-talking microphones.

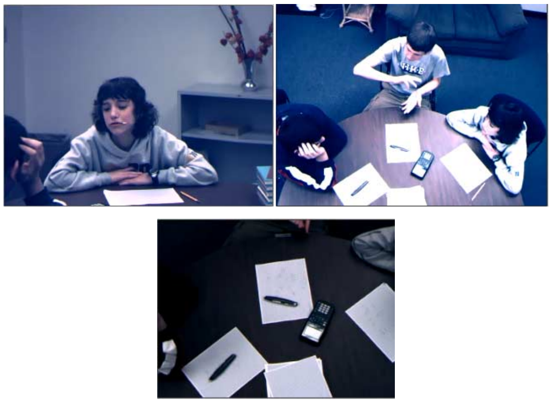

Figure 1. Example of close-up, wide angle and tabletop views.

Each participant's writing was collected using Nokia digital pens and large sheets of Anoto digital paper. The pens provided the unique identification of each student, even when they wrote on another student's paper [13]. Through the use of digitized pens and paper, it was possible to obtain pen stroke data for each student. The timestamp, coordinates, and pressure of each student's pen strokes, herein defined as the period between the pen nib's point of contact with the paper and later removal of contact, was available for analysis. Typically, several pen strokes clustered together would compose a meaningful representation, such as a word, number, symbol, or diagram. Using the timestamps, each student's pen strokes could be mapped to specific written representations and also to specific problems varying in difficulty level (easy, moderate, hard, very hard).

Figure 2 shows an example of pen strokes during problem-solving from this dataset. Each of the student's pen strokes were coded for type of written representation and semantic content. Over 10,000 written representations were coded and analyzed [11]. There were four distinct types of written representations: Diagrammatic (D), Linguistic (L), Numeric (N) and Symbolic (S). Strokes that involved content corrections, such as misspelled words, misshapen symbols, or crossed out content, were coded as disfluent. Due to expected differences in signal characteristics, such disfluencies were filtered out before conducting analyses in the present research. Task irrelevant written content, such as doodles, also were filtered out for the present purposes.

4. PEN STROKE ANALYSIS

Pen signal analysis potentially can play a valuable role in predicting domain expertise during educational activities. This is especially true since pen input represents existing work practice in classrooms, so automatic data analysis based on using digital pen tools is expected to be minimally disruptive [11].

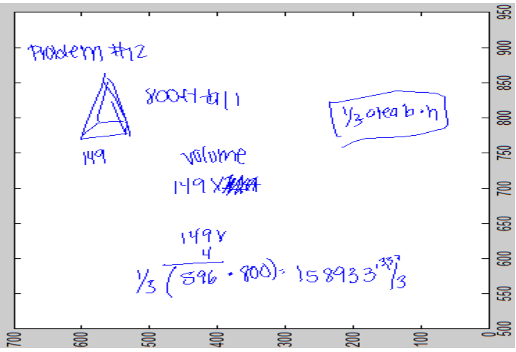

Figure 2. An example of written content of pen strokes during problem-solving.

The following pen signal features were analyzed across both problem difficulty levels and types of written representation for students identified as experts versus non-experts (see [8] for ground-truth coding of expertise and written representations):

- Average number of pen strokes during a problem;
- Average total writing time during a problem;
- Average pen stroke distance, or accumulated distance from start to end of a pen stroke;
- Average stroke duration, or time from start to end of a pen stroke;
- Average writing speed when forming a pen stroke, or stroke distance divided by duration;
- Average writing pressure when forming a pen stroke.

Preliminary analyses revealed that experts and non-experts did not differ significantly in their average number of strokes or total writing time when solving problems, so the sections that follow focus on the remaining stroke-level pen signal features.

4.1 Average Stroke Distance

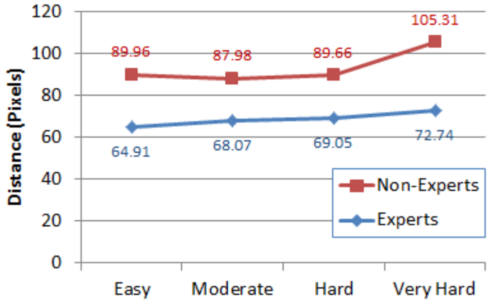

Figure 3. Average stroke distance for problem difficulty.

Average Stroke Distance (ASDis) measured the accumulated average distance of a pen stroke from pen down to pen up, which accurately reflected stroke distance involving curvatures. This feature was analyzed separately with the data partitioned by problem difficulty levels and also by type of written representations. Figure 3 shows average ASDis across the four problem difficulty levels for both experts and non-experts. A two-way ANOVA was conducted that examined the effect of problem difficulty levels and expertise on ASDis. There was no statistically significant interaction between the problem difficulty levels and expertise, or among problem difficulty levels, *F*s < 1. However, experts and non-experts differed in ASDis, *F* (1,88)=17.45, *p* < 0.001. Experts' pen stroke distance averaged 30.9% shorter than non-experts during math problem-solving.

Further analyses of ASDis across various types of written representations are shown in Figure 4. A two-way ANOVA was conducted that examined the effect of types of written representation and expertise on ASDis. There was a statistically significant interaction between type of written representation and expertise, *F*(3, 88)=2.71, *p*<0.05. Follow-up analyses revealed that experts' pen strokes averaged significantly shorter for L (*t*(22)=-3.28, *p*<0.003, two-tailed), N (*t*(22)=-2.51, *p*<0.02, two-tailed), and S (*t*(22)=3.52, *p*<0.002, two-tailed), but not for D (*t*(22)=0.44, *p*>0.665, two-tailed). Overall, experts averaged 22.3% shorter strokes than non-experts. Further analyses showed that ASDis for D averaged significantly longer than for L (*t*(46)=4.19, *p*<0.000, two-tailed), N (*t*(46)=5.18, *p*<0.001, two-tailed), and S (*t*(46)=4.93, *p*<0.001, two-tailed).

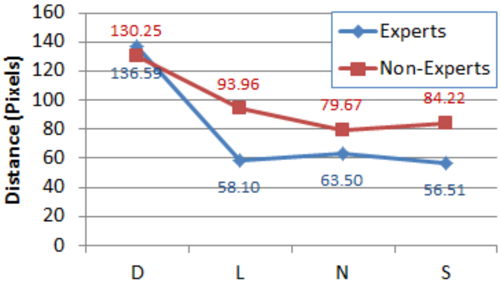

Figure 4. Average stroke distance for written representations.

4.2 Average Stroke Duration

Average Stroke Duration (ASDur) refers to time required to form a pen stroke written during math problem solving. The ASDur across the four problem difficulty levels for both experts and non-experts is shown in Figure 5. A two-way ANOVA was conducted that examined the effect of problem difficulty levels and expertise on ASDur. There was no statistically significant interaction between difficulty levels and expertise, *F* < 1. There also was no significant difference among problem difficulty levels, *F* (3,18)=1.94, *p*<0.129. However, experts and non-experts differed in stroke duration, *F*(1, 88)=14.67, *p*<0.001. Experts averaged 21.6% briefer strokes than non-experts.

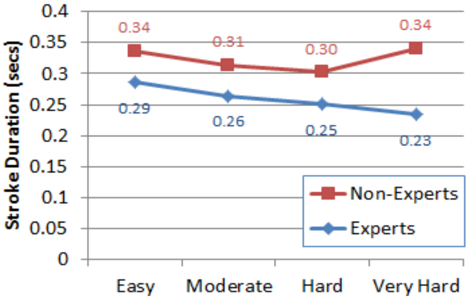

Figure 5. Average stroke duration for problem difficulty.

Figure 6 shows ASDur across four types of written representation for both experts and non-experts. A two-way ANOVA was conducted that examined the effect of types of written representations and expertise on ASDur. There was no statistically significant interaction between type of written representation and expertise, *F* < 1. There also was no significant difference between experts and non-experts, *F*(1,88)=1.88, *p*<0.173. However, different types of written representation significantly influenced stroke duration, *F*(3, 88)=30.80, *p*<0.001. Follow-up analyses revealed that ASDur for D averaged significantly longer than for L (*t*(46)=5.28, *p*<0.001, two-tailed), N (*t*(46)=5.14, *p*<0.001, two-tailed), and S (*t*(46)=7.28, *p*<0.001, two-tailed).

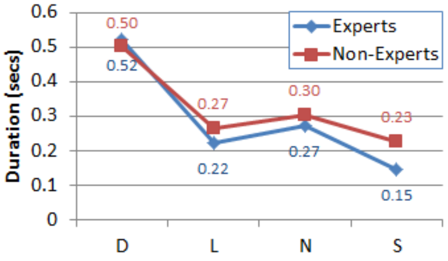

Figure 6. Average stroke duration for written representations.

4.3 Average Writing Speed

Further analyses evaluated whether student's writing speed, a derivative metric, provided predictive information on math domain expertise. Writing Speed (AWSpe) refers to how fast a student wrote a pen stroke on average, which was simply calculated by dividing ASDis by ASDur. Intuitively one might expect writing speed to be a distinguishing factor between experts and non-experts if its two ingredient factors (i.e., distance, duration) are significantly predictive. However, since experts tended to have both shorter distance and briefer duration of strokes, these proportional changes would not necessarily result in writing speed per se being any different between the two groups. For this reason it was evaluated separately.

Figure 7 shows AWSpe across the four problem difficulty levels. A two-way ANOVA was conducted that examined the effect of problem difficulty levels and expertise on AWSpe. There was no statistically significant interaction between problem difficulty level and expertise, *F*<1. There also was no significant difference between difficulty levels, *F*(3, 88)=2.11, *p*>0.104. However, experts and non-experts differed significantly in writing speed, *F*(1, 88)=7.39, *p*<0.008. Experts were actually 19.2% slower in average writing speed than non-experts.

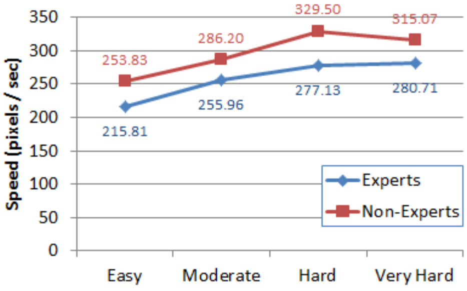

Figure 7. Average writing speed for problem difficulty.

Further analyses evaluated writing speed by expertise across various types of written representation, as shown in Figure 8. A two-way ANOVA was conducted that examined the effect of types of written representation and expertise on AWSpe. There was no statistically significant interaction between type of written representation and expertise *F*(3, 88)=1.12, *p*>0.345. However, experts and non-experts differed significantly in

AWSpe, $F(1, 88)=3.91$, $p<0.05$. Overall, experts averaged 16.9% slower writing speed than non-experts. In addition, significant differences were present between different types of written representation, $F(3, 88)=6.09$, $p<0.001$.

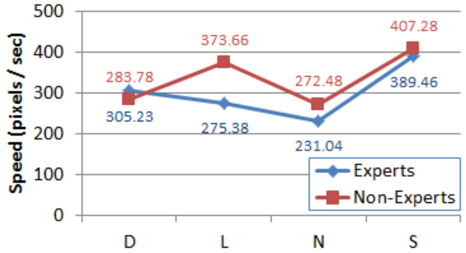

Figure 8. Average writing speed for written representations.

4.4 Average Writing Pressure

Since the digital pen also recorded writing pressure during problem solving, analyses investigated whether Average Writing Pressure (AWPre) was predictive of math domain expertise. Figure 9 shows the AWPre across four problem difficulty levels. It reveals that expert students exerted lower pressure on average than non-experts. A two-way ANOVA was conducted that examined the effect of problem difficulty levels and expertise on AWPre. There was no statistically significant interaction between problem difficulty and expertise, or between difficulty levels in AWPre, *F*s<1. However, experts and non-experts differed significantly in AWPre, $F(1, 88)=9.64$, $p<0.003$. Experts averaged 7.9% lower writing pressure than non-experts.

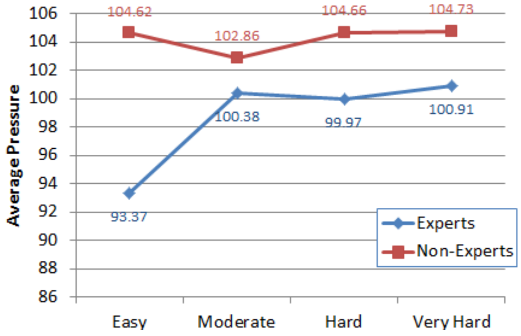

Figure 9. Average writing pressure for problem difficulty.

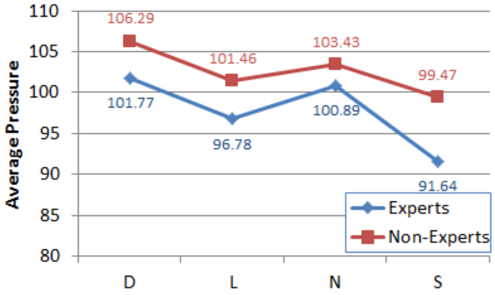

Figure 10. Average writing pressure for written representations.

Figure 10 illustrates the corresponding data broken down by written representations. A two-way ANOVA was conducted that examined the effect of type of written representation and expertise on AWPre. Once again, there was no statistically significant interaction between type of written representation and expertise, *F*<1. However, written representations differed significantly in AWPre, $F(3, 88)=3.71$, $p<0.014$. In addition, experts and non-experts differed significantly in AWPre, $F(1, 88)=15.96$, $p<0.001$. Experts averaged 7.1% lower pressure than non-experts.

5. EXPERTISE CLASSIFICATION

Following the statistical analyses conducted on the pen signal data in the previous section, Machine Learning (ML) algorithms were used to conduct two-class classification – experts against non-experts – to examine how accurately domain expertise could be predicted based on various sets of features. For all classifications performed in this section, Support Vector Machine (SVM), Random Forest (RF), and Naïve Bayes (NB) classifiers were compared for classification of expertise in order to assess the relative reliability of these alternative approaches.

Support Vector Machine (SVM), Random Forest (RF), and Naïve Bayes (NB) classifiers were selected because these three classifiers are the most widely used for feature classification in machine learning. SVM is potentially advantageous for capturing complex relations in the data without manual intervention. The Naïve Bayes classifier applies Bayes theorem, and it considers each feature to have contributed independently. While it can be trained very efficiently, it nonetheless contains oversimplified assumptions such as independence of the features. Random Forest works by combining multiple decision trees during the training of the data, and it predicts the class by taking the mode of the individual trees.

The classification performance (Acc: Accuracy, Sens: Sensitivity, Spec: Specificity) was compared for different feature combinations. Classification accuracy provides an estimate of the overall degree of accuracy in correctly distinguishing expertise from non-expertise. Sensitivity, also known as the true positive rate or recall, estimates the likelihood of correctly classifying experts. Specificity, also called the true negative rate or false positives, estimates the likelihood of correctly classifying non-experts as distinct from experts. The best possible classification accuracy, 100%, would yield no false positives or false negatives.

The leave-one-out method [14] was used for cross validation, which involves a single observation from the data serving as validation data while the remaining observations serve as training data. The model then conducts repeated estimates so each observation in the data serves once as validation data. Classification performance was compared for different feature combinations, as described in the sections that follow.

5.1 Unguided Classification Using All Pen Features

Table 1. Classification accuracies for expertise with all 48 possible pen signal features

	Expertise Classification		
	Acc.	**Sens.**	**Spec.**
SVM	0.667	0.583	0.750
RF	0.708	0.667	0.750
NB	0.667	0.583	0.750

Table 1 shows the expertise classification accuracies based on all possible 48 pen features and all data, including average number of pen strokes, average total writing time during a problem, average pen stroke distance, average stroke duration, average writing speed, and average writing pressure for both problem difficulty level and types of written representation. These performance levels were achieved without taking into account empirical results reported in the previous sections. Therefore, the performance levels in Table 1 provide a baseline for unguided machine learning results. The results show that Random Forest outperforms the other two classifiers, with the best performing classification accuracy of 70.83%.

5.2 Partially Guided Classification By Problem Difficulty or Type of Representation

Expertise classification was also examined separately when all of the data was partitioned by problem difficulty levels and by type of written representations. Table 2 shows these classification accuracies, based on 24 features apiece. Overall, Random Forest outperformed the other two classifiers. In comparison with classification based on all 48 pen features, accuracy decreased when pen features were based on problem difficulty, from 70.83% to 66.67%. However, it increased when pen signal attributes were classified by type of written representation, from 70.83% to 75%.

Table 2. Expertise classification accuracies with 24 features

	By Problem Difficulty			By Type of Representation		
	Acc.	Sens.	Spec.	Acc.	Sens.	Spec.
SVM	0.583	0.417	0.750	0.667	0.500	0.833
RF	0.667	0.667	0.667	0.750	0.667	0.833
NB	0.625	0.583	0.667	0.750	0.667	0.833

5.3 Classification by Problem Difficulty

In order to learn how expertise classification accuracy varies across problem difficulty and type representation, ML algorithms were applied on the attributes across these categories for all of the data. Table 3 shows results for expertise classification across problem difficulty levels. The highest classification accuracy of 79.2% was achieved with moderate difficulty problems using Naïve Bayes classification, compared with other problem difficulty levels during expertise classification.

Table 3. Expertise classification accuracies for problem difficulty levels

	Easy			Moderate		
	Acc.	Sens.	Spec.	Acc.	Sens.	Spec.
SVM	0.167	0.167	0.167	0.542	0.500	0.583
RF	0.542	0.417	0.667	0.708	0.667	0.750
NB	0.417	0.333	0.500	0.792	0.833	0.750
	Hard			**Very Hard**		
	Acc.	**Sens.**	**Spec.**	**Acc.**	**Sens.**	**Spec.**
SVM	0.583	0.583	0.583	0.542	0.500	0.583
RF	0.583	0.500	0.667	0.583	0.500	0.667
NB	0.542	0.500	0.583	0.625	0.583	0.667

5.4 Classification by Type of Representation

A similar analysis was performed to learn how classification accuracy varies across the four types of written representation, as shown in Table 4. Note that Table 4 highlights that the relatively small set of diagrams differed from all other categories in pen stroke formation during writing, which produced higher variance and lower prediction results.

Table 4. Classification accuracies across types of written representation

	Diagrammatic			Linguistic		
	Acc.	Sens.	Spec.	Acc.	Sens.	Spec.
SVM	0.208	0.333	0.083	0.667	0.750	0.583
RF	0.542	0.583	0.500	0.708	0.667	0.750
NB	0.417	0.333	0.500	0.750	0.750	0.750
	Numeric			**Symbolic**		
	Acc.	**Sens.**	**Spec.**	**Acc.**	**Sens.**	**Spec.**
SVM	0.583	0.500	0.667	0.833	0.750	0.917
RF	0.583	0.583	0.583	0.792	0.833	0.750
NB	0.583	0.500	0.667	0.792	0.833	0.750

By comparing Table 3 and Table 4, it is clear that expertise classification accuracy averaged higher when the data were broken down by type of written representation rather than problem difficulty level. In particular, symbolic content was the most sensitive for distinguishing domain expertise, with an accuracy of 83.3% using Support vector Machine classification. Furthermore, all techniques yielded 79.2% to 83.3% accuracies when symbolic content was the focus of analysis.

5.5 Empirically Guided Classification Using Significant Pen Features

Based on the eight statistically significant pen features identified in Section 4, targeted machine learning analyses were conducted using this advance guidance, with the aim of optimizing prediction accuracy for identifying expertise. Table 5 shows a summary of the prediction accuracy of each individual significant feature. Naïve Bayes outperforms the others for average stroke duration, writing speed, and writing pressure by problem difficulty levels, as well as average stroke distance, duration, and writing speed by type of written representation. However, Random Forest outperforms the other classifiers for average stroke distance by problem difficulty level and also average writing pressure by type of written representation. The highest classification accuracies obtained, 79.2%, were produced by Random Forest and Naïve Bayes techniques.

Table 5. Classification accuracies for expertise with individual significant features

	By Problem Difficulty				By Types of Representation[1]			
	ASDis	ASDur	AWSpe	AWPre	ASDis	ASDur	AWSpe	AWPre
SVM	0.375	0.000	0.375	0.333	0.667	0.500	0.167	0.125
RF	0.792	0.500	0.542	0.667	0.667	0.625	0.417	0.792
NB	0.667	0.583	0.625	0.750	0.792	0.750	0.583	0.708

[1] Diagrams removed from representations.

In order to determine how expertise prediction varies depending on whether the data is broken down by problem difficulty level or type of representation, expertise classification also was conducted across the combined eight significant pen features, as shown in Table 6. The results show that Naïve Bayes outperforms the other two classifiers for both cases, although accuracy was consistently highest at 83.3% when analyses were performed on data broken down by type of written representation.

Note that Tables 5 and 6 show classification summary data with diagrams removed. Although diagrams only constituted 7% of all written representations [11], empirical analyses (see 4.1 and 4.2) and machine learning results (see Table 4) clarified that their written formation differed qualitatively from all other types of representation.

Table 6. Classification accuracies for expertise with combined significant features

	Significant Features by Problem Difficulty	Significant Features by Types of Representation[1]
SVM	0.625	0.792
RF	0.625	0.750
NB	0.708	0.833

[1] Diagrams removed from representations.

6. DISCUSSION

The present research demonstrates that expertise in mathematics is predictable based on dynamic signal features present in handwriting while students solve mathematics problems. Using unguided automatic machine learning, the dominant domain expert in a student group was identifiable with 70.83% accuracy using a Random Forest technique. In contrast, when machine learning was guided by empirical information on optimal data partitioning and feature set selection, classification accuracy improved to 83.3%.

Empirical analyses revealed a surprisingly large number of dynamic stroke signal-level features that significantly distinguished students who were expert versus non-expert in mathematics. These features included average stroke distance, duration, pressure, and speed. When feature sets were composed of all significant pen signal features, the highest classification accuracy was 83.3% using Naïve Bayes with the data partitioned by type of written representation. When individual signal features were used as the basis for classifying expertise, the highest accuracy of 79.2% was obtained using: (1) Naïve Bayes to analyze average stroke distance, with the data partitioned by type of representation; (2) Random Forest to analyze average pressure, with the data partitioned by type of representation; and (3) Random Forest to analyze average distance, with the data partitioned by problem difficulty level.

Overall, classification accuracy was higher when the data were partitioned by type of written representation, rather than problem difficulty level. Analysis of symbolic content supported the highest prediction accuracy, 83.3%, of the different types of written representation. This finding is consistent with previous research showing that higher-performing students express significantly more symbols when solving math problems than do lower-performing ones [10].

With respect to problem difficulty level, moderately difficult problems yielded the highest prediction accuracy, which was 79.2%. This may have occurred because more expert students often did not write at all during easy problems, which they could complete in their head. In contrast, the least expert students frequently did not write during hard or very hard problems, which they did not believe they could solve. As a result, participation was highest on problems of moderate difficulty, which distinguished between expert versus non-expert writing patterns most effectively.

From a theoretical viewpoint, the finding that markers of expertise include shorter stroke distance, briefer duration, and lighter pressure is consistent with experts conserving energy more effectively during communication. This reduction of effort expended on communication enabled them to reduce cognitive load and free up mental resources for math problem solving. The impact was an ability to solve harder problems and perform at an overall higher level than less expert students. Limited-resource linguistic theories have characterized people as adaptively conserving energy at both the signal and lexical levels to minimize their own cognitive load. This reduction in effort and increased efficiency has been documented previously in all communication modalities, including writing.

Interestingly, although it seems counter intuitive, experts' writing speed actually was significantly slower than that of non-experts. One explanation is that experts averaged both shorter stroke distance and briefer duration to convey the same information as non-experts. Since these adaptations were approximately proportional, the impact was to render writing speed an ineffective derivative predictor for distinguishing between the two groups.

This finding regarding slower average writing speed conflicts with Ochoa et al.'s previous report [7]. However, Ochoa and colleagues' work did not conduct any actual empirical analyses of expert behavior, and the automated analyses conducted only evaluated low-level geometric primitives of strokes (i.e., lines, circles). In contrast, the present work analyzed writing speed and other pen signal features empirically. It shows that experts wrote significantly slower than non-experts, and it also clarifies the relation of experts' writing speed with their pen stroke duration and distance. In addition, the present work provides analyses of these writing behaviors across different problem difficulty levels and types of written representation (symbols, letters, diagrams, numbers). These results confirm the generality and robustness of writing dynamics as a predictor of expertise. Finally, the present work conducted machine learning analyses that achieved higher expertise classification accuracy (83.3%) than prior work [7].

In the present corpus, diagrams constituted 7% of written representations [11]. As indicated by Table 4, diagrams supported lower classification accuracies than any of the other representations. They also differed systematically from all other representations in their stroke formation qualities. Figure 4 and 6 show that they averaged longer stroke distances and durations than either letters, numbers or words. As a result, classification accuracies were improved with minimal data loss by eliminating this type of representation.

This work presents a strong starting point for using machine learning techniques combined with empirical findings to leverage high classification accuracies for identifying domain expertise. Examining the machine learning analytic results, it is noteworthy that Random Forest and Naïve Bayes classifiers performed better than SVM in most cases. One implication is that we can conduct classification from the original feature, instead of mapping them into high dimensional space as does SVM. In addition, the similarly high classification rate by two different ML methods confirms that the feature sets used were consistently able to detect students' expertise levels. In the future, more complex classifiers than the basic ones used here are likely to improve upon this classification performance.

Since recent research has shown that pen interfaces can substantially facilitate students' cognition [12], corresponding learning analytics techniques will be required based on their written signal or representational metrics [11]. During classroom activities, writing and related pen tools are common work practice. As a result, their introduction into classrooms is less likely to disrupt educational activities than other alternatives. In summary, digital pen interfaces could be used in the future as a digital tool that stimulates students' cognition, to unobtrusively collect their writing samples, and also to evaluate and guide their learning progress.

One limitation of this study is the small size of the Math Data Corpus, which may have affected the machine learning classification accuracy and consistency. An interesting future direction would be to collect longitudinal data over time from the same students as they learn a new domain, so that within-subject tracking of expertise could be attempted in a more fine-grained way. More extended work of this type also would be likely to uncover other valuable metrics for more reliably, rapidly, and objectively identifying expertise.

This work in MMLA has the potential to improve future educational practice and technologies. It also has the potential to stimulate the development of more advanced hybrid engineering techniques for data analytics. New hybrid techniques are especially needed that can improve the generalizability and transparency of current approaches to machine learning. In addition, future work could examine other content domains, as well as students representing other cultural-linguistic groups. It also could combine the valuable pen signal features identified here with information derived from other modalities, including speech and visual analysis.

7. CONCLUSIONS

As investigated in this paper, empirical results combined with machine learning algorithms can enhance classification accuracies for predicting expertise. Whereas unguided machine learning identified the dominant domain expert in a student group with 70.83% accuracy using Random Forest, empirical guidance involving optimal data partitioning and feature set selection improved this performance to 83.3%. Empirical analyses revealed a large number of signal-level writing features that significantly distinguished students who were experts from non-experts in mathematics, including average stroke distance, duration, pressure, and speed. These results demonstrate that signal-level features of writing can predict domain expertise with high reliability. In addition, future hybrid methods that strategically incorporate empirical guidance could be developed that outperform unguided machine learning. Finally, the long-term learning analytics work outlined in this paper aims to eventually provide pragmatic guidance to teachers and students as they strive to achieve their learning goals, and to support the expansion of promising new mobile educational technologies.

8. ACKNOWLEDGMENTS

This research was supported by National ICT Australia (NICTA), and Incaa Designs (http://www.incaadesigns.org/). Additional support was provided for Dr. Oviatt by the Faculty of Engineering at University of New South Wales as part of their Emiment Visiting Scholars program.

9. REFERENCES

[1] Baddeley, A.D. and Hitch, G.J. 1974. Working memory. *The psychology of learning and motivation: Advances in research and theory*. G.A. Bower, ed. Academic. 47–89.

[2] Cheng, P.C.-H. and Rojas-Anaya, H. 2007. Measuring Mathematical Formula Writing Competence: An Application of Graphical Protocol Analysis. (Jan. 2007).

[3] Clark, H. and Brennan, S. 1991. Grounding in Communication. *Perspectives on Socially Shared Cognition*. L. Resnick, L. B, M. John, S. Teasley, and D., eds. American Psychological Association. 259–292.

[4] Glaser, R. and Chi, M.T.H. 1988. Overview. *The Nature of Expertise*. M.T.H. Chi, Glaser, and M.J. Farr, eds. Psychology Press. Xv–xxviii.

[5] Lindblom, B. 1990. Explaining Phonetic Variation: A Sketch of the H&H Theory. *Speech Production and Speech Modelling*. W.J. Hardcastle and A. Marchal, eds. Springer Netherlands. 403–439.

[6] Miller, G.A. 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*. 63, 2 (1956), 81–97.

[7] Ochoa, X., Chiluiza, K., Méndez, G., Luzardo, G., Guamán, B. and Castells, J. 2013. Expertise Estimation Based on Simple Multimodal Features. *Proceedings of the 15th ACM on International Conference on Multimodal Interaction* (2013), 583–590.

[8] Oviatt, S. 2013. Problem Solving, Domain Expertise and Learning: Ground-truth Performance Results for Math Data Corpus. *Proceedings of the 15th ACM on International Conference on Multimodal Interaction* (2013), 569–574.

[9] Oviatt, S. 2013. *The Design of Future Educational Interfaces*. Routledge.

[10] Oviatt, S., Arthur, A., Brock, Y. and Cohen, J. 2007. Expressive Pen-based Interfaces for Math Education. *Proceedings of the 8th International Conference on Computer Supported Collaborative Learning* (New Brunswick, New Jersey, USA, 2007), 573–582.

[11] Oviatt, S. and Cohen, A. 2013. Written and Multimodal Representations As Predictors of Expertise and Problem-solving Success in Mathematics. *Proceedings of the 15th ACM on International Conference on Multimodal Interaction* (2013), 599–606.

[12] Oviatt, S., Cohen, A., Miller, A., Hodge, K. and Mann, A. 2012. The Impact of Interface Affordances on Human Ideation, Problem Solving, and Inferential Reasoning. *ACM Transactions on Computer Human Interaction*. 19, 3 (Oct. 2012), 22:1–22:30.

[13] Oviatt, S., Cohen, A. and Weibel, N. 2013. Multimodal Learning Analytics: Description of Math Data Corpus for ICMI Grand Challenge Workshop. *Proceedings of the 15th ACM on International Conference on Multimodal Interaction* (2013), 563–568.

[14] Witten, I.H., Frank, E. and Hall, M.A. 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier.

[15] Worsley, M. and Blikstein, P. 2011. What's an Expert? Using learning analytics to identify emergent markers of expertise through automated speech, sentiment and sketch analysis. *Proceedings for the 4th Annual Conference on Educational Data Mining* (Netherlands, 2011), 235–240.