The Thirty-First International Florida
Artificial Intelligence Research Society Conference (FLAIRS-31)

Predictive Models of User Performance for Marksmanship Training

Mary Jean Blink,[1] Ted Carmichael,[1,2] Jennifer Murphy,[3] Michael Eagle[1,4]

[1]TutorGen, Inc., Fort Thomas, KY, USA
[2]University of North Carolina at Charlotte, Charlotte, NC, USA
[3]Quantum Improvements Consulting, Orlando, FL, USA
[4]Carnegie Mellon University, Pittsburgh, PA, USA
mjblink@tutorgen.com, tedsaid@gmail.com, jmurphy@quantumimprovements.net, maikuusa@gmail.com

Abstract

How the Army conducts rifle marksmanship training is undergoing a number of positive changes. Despite this, challenges to conducting and coordinating this critical training remain. One challenge to assessing training effectiveness is a lack of persistent records of soldier performance; too often soldier data are purged shortly after training events for convenience and in order to ensure privacy. This paper reports on our efforts to research the feasibility of collecting, analyzing, and storing data from multiple training systems, in order to accelerate and improve marksmanship training. We do this through the use of cognitive, psychomotor, and affective constructs; and the use of predictive modeling techniques in order to forecast marksmanship qualification scores. These models successfully predicted scores on a 40-point scale with a root mean square error (RMSE) of less than three, using models that are robust to changing input variables. Future improvements and directions for this research are also discussed.

Introduction

The Army's success relies on the ability to effectively apply innovative training solutions that result in adaptable and ready soldiers. To meet this goal, the Army has invested in training technology to provide effective training while maximizing budget and schedule efficiencies. However, the effectiveness of these training platforms is rarely assessed, and too often soldier data are purged shortly after training events, for convenience and in order to ensure privacy. Providing improved access to soldier data has potential benefits for multiple audiences involved in the training process. Instructors can utilize a deeper understanding of their learners' performance to improve personalization of training and feedback. Developers and researchers could use these data to evaluate the effectiveness of training programs and investigate best practices in training delivery through experimentation. Resource managers could more effectively coordinate the manpower and matériel required to conduct training events and track schedules and budgets.

Thus, the goal of this project was to identify the human performance marksmanship measures and metrics relevant to training technologies, and to develop a predictive model of marksmanship performance to serve as the basis for adapting training. As rifle marksmanship is a complex skill comprised of cognitive, psychomotor, and affective components (Chung et al. 2006), these collected data can be categorized as one of these three types; although demographic data are also included. Cognitive attribute data can include measures of general cognitive ability, such as with the Armed Forces Qualification Test (AFQT); or domain knowledge measures that are specific to marksmanship and useful as a metric of learning (Maier 1993). Psychomotor data include results from eye dominance tests (James and Dyer 2011), and handedness. When considered separately neither eye dominance nor handedness show trends in marksmanship performance; however, a meta-analysis from the literature suggests approximately one-third of the population is cross-dominant, meaning the dominant hand (left vs. right) is different than the dominant eye (McManus et al. 1999), and this has implications for marksmanship training. For affective data a survey was developed during the early stages of this project, including questions about attitudes and feelings, in order to create constructs of general affective states such as resiliency, self-efficacy, initiative, and perceived stress, based on previous work in the literature (Aude et al. 2014; Bandura 2006; Cohen, Kamarck, and Mermelstein 1983; Duckworth, Peterson, and Matthews 2007; Frese et al. 1996).

As a demonstration of the usefulness of this data, and in preparation for future work in creating adaptive and personalized marksmanship training systems, we created a predictive model of soldier performance on the standard marksmanship qualification exam and compared the model outputs to actual exam performance.

Background

By conducting surveys and interviewing subject matter experts (SMEs) during our User Needs Analysis we determined a wide range of either available or attainable data on soldier abilities, experiences, attitudes, and other individual measures. The analysis was conducted from September 28, 2015 to October 2, 2015 at Fort Benning, GA. Interviews and focus groups were conducted with SMEs from groups involved in Army marksmanship training who might benefit from improved access to soldier performance data. These included: Research psychologists from the U.S. Army Research Institute (ARI), Engagement Skills Trainer (EST) proponent instructors, 194th Armor Brigade (AR BDE) instructors, Marksmanship Master Trainer Course (MMTC) trainers, Training and task developers from the Maneuver Center of Excellence (MCoE) Department of Training Development (DOTD), Simulation center managers, Range control operations personnel, Ammunition resource managers, and Drill Sergeants.

Our research team identified, through literature review and user needs analysis, measures of the knowledge, skills, and abilities (KSAs) relevant to successful marksmanship performance. A prototype battery of measures of these KSAs was developed, with an eye for easy implementation due to the already overburdened schedule during training (which necessarily covers much more than marksmanship training). These measures can be used to not only tailor the individual soldier's training, but also to address the needs of instructors, researchers, and resource managers. For example, instructors and training developers in Army schoolhouses could use the data to track student performance, evaluate their instructors, and compare courses over time. Personnel involved in simulator maintenance could leverage the data to ensure equipment consistency. Acquisitions personnel could use it to calculate return on investment, throughput, and other metrics of program success. Finally, research personnel could benefit from having the data available to them for experimentation.

Another potential means of maximizing training effectiveness is the implementation of adaptive training technologies. To investigate the extent to which intelligent tutoring can be developed in a highly authorable way, Army Research Lab (ARL) is developing GIFT (Generalized Intelligent Framework for Tutoring), a modular framework that is designed to increase authoring efficiency (Sottliare et al. 2012). Successful adaptation of instructional material and formative feedback depends upon a robust student model of performance. Therefore, the first step towards this goal is the development of predictive models based on the cognitive, psychomotor, and affective data for each learner. Such a predictive model will allow us to identify the most important attributes for marksmanship skill, and will lead to improved methods of integrated training and more efficient instruction and feedback, as well as lay the groundwork for automated training systems that do more of the basic data collection, progress monitoring, and fundamental analyses, leaving more time for meaningful one-on-one interactions with the expert trainers.

Methodology

The input data are attributes in four general categories: demographic, cognitive, psychomotor, and affective. These include survey results, simulation training data, self-reported qualification and fitness test results: more than 60 data fields in all. Each trainee also has up to five qualification scores. The highest of these are used to calibrate and validate the models, as only one score is needed to pass qualification, and the highest is used also for soldier ranking. In summary, there are 84 subjects with qualification scores that can range from 0-40, with most scores falling between 20 and 40). Out of this cohort there were (based on the highest qualification score) 10 Experts, 47 Sharpshooters, 26 Marksmen, and 1 UQ (unqualified).

In order to test the measurements identified and developed, a team of researchers were onsite for one week with four platoons of recently commissioned officers. Each platoon cycled through various training exercises that included some data collection events. These observations were critical for better understanding the methods and training protocols being used in marksmanship drills and exercises. Data assessing hand-eye coordination through an application on tablets was conducted. Data related to the simulators was collected during the course of regular instruction. Soldiers completed a new exercise that included F/N Expert, a tool that allows both dry-fire and live-fire training and provide immediate feedback to both the shooter and coach, including details about rifle movement and location of hit and miss. We also collected data from the range during the live-fire practice rounds. Finally, soldiers completed their live-fire marksmanship qualification exam while observers were on site.

Results

We used a regression model known as LASSO: Least Absolute Shrinkage and Selection Operator. LASSO is a regression analysis method that uses both variable selection and regularization. In short, LASSO is designed to minimize the number of contributing attributes while avoiding the problem of overfitting to the extant data. Detailing the exact workings of LASSO is outside of the scope of this report; however, it should be noted that the regression analysis is done through an iterative process using randomly assigned subsets of data and coefficient values – about 100,000 simulations at each step – in order

to determine the best fit coefficients, as well as to reduce to zero coefficients for some variables, effectively removing them from the model. The effect of this methodology is that, for any particular run of the LASSO regression technique, the resultant model will have some slight variability from subsequent runs; and variables that are close to the same value, in terms of their contribution to the models, may change slightly in rank order from one regression output to the next. (If the regression technique instead produced the exact same output each time it was computed, this would actually be a sign of overfitting to the data.)

These results can be refined in various ways, depending on the requirements of how the model will be used. For example, it may be more important to identify all potential expert shooters, even though this would err on the side of also including some who will not score so high; conversely, it may be more important to *correctly* identify these expert shooters, even though doing so will inevitably overlook some who are borderline-rated but score very high.

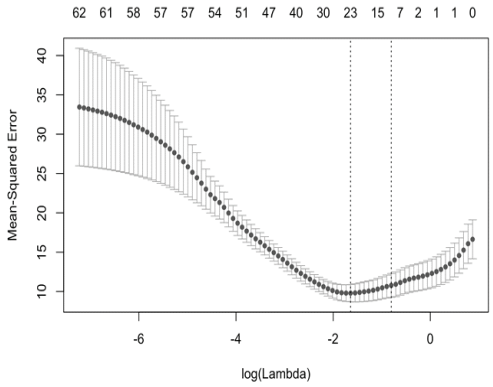

Figure 1. Iterations of the LASSO regression algorithm; from left to right the number of attributes is reduced (x-axis), with the predicted score shown on the y-axis.

Another way that the score could be calibrated differently, based on the end-user needs, would be to consider edge cases (very high or very low scores) as being more important than currently. The LASSO algorithm is conservative, meaning it tries to minimize the predicted error in such a way that disproportionately affects outliers in the data. This is because outliers, by definition, lie on the edges of the potential range of data and have many fewer data points. Thus, any algorithm that attempts to minimize error for the majority of data points will tend to overlook outliers, a process which – mathematically – will penalize the model the least. Thus, if edge case accuracy is determined to be more important, then the LASSO regression can be re-run using an adjusted error calculation, and thus giving more import to edge case situations. This will have the effect of improving edge case predictions (though at the cost of some accuracy in differentiating between other, more common classifications). It should be noted here that improving predictions for UQ scores is difficult, as there is only one UQ score in this dataset (out of n=84 cases). More data of the relevant scores is the best way to improve predictions for those scores.

MAIN Model (n=84)		**ALT Model (n=84)**	
(Intercept)	30.7976	(Intercept)	30.7976
APFT.Score	0.2458	APFT.Score	0.3681
Rifle	-0.0087	Law.Enforcement	0.0435
Law.Enforcement	0.0049	Individual.Sports	0.2879
Individual.Sports	0.2546	Large.Animal.Hunting	0.5786
Large.Animal.Hunting	0.4819	Mechanical.Work	0.1784
Mechanical.Work	0.2917	Initiative	-0.9614
Initiative	-0.8188	SelfEfficacyBehaviorTotal	0.6166
SelfEfficacyBehaviorTotal	0.2402	PerceivedStress	0.4108
PerceivedStress	0.1635	Conscientiousness	-0.1285
Conscientiousness	-0.1539	LocusOfControl	-0.5392
LocusOfControl	-0.6453	Knowledge	0.5101
Knowledge	0.0736	ESTPS	0.0679
Practice	1.6760	MOT.Mean.Latency	-0.0222
RTI.Five.choice.accuracy.score	0.2940	RTI.Five.choice.accuracy.score	0.1946
RTI.Mean.five.choice.movement.time	-0.0839	RTI.Mean.five.choice.movement.time	-0.2213
RTI.Simple.error.score..inaccurate.	0.0338	RTI.Simple.error.score..inaccurate.	0.0540
RTI.Mean.simple.reaction.time	-0.3945	RTI.Mean.simple.reaction.time	-0.0654
EducationMastePhD	-0.2530	EducationMastePhD	-0.3625

Table 1. List of attributes and their relative contributions for the MAIN model and the ALT model (for n=84).

Models

For the n=84 case we created two distinct models based on the LASSO regression analysis: one that included the practice scores and one that did not. (It was determined that "practice" might be too good at predicting final scores, as this data was recorded in the same way as the live-fire qualification test, and immediately preceding. Therefore an ALT model with practice data removed was created.)

The RMSE (Root Mean Squared Error) for the main model is 2.45, and for the alt model it is 2.88, meaning the average error for the predicted qualification score for each solider is 2.45 (or 2.88).

The above tables show the 20 attributes used in the main model, and the alt model, as well as the coefficient values for each. Each of the coefficients has been standardized using each attribute's mean and standard deviation, for easy comparison and ranking. Attributes that are negatively

correlated with the final score have a negative coefficient value.

Discussion and Future Work

The two models created here are general and do not take into account end-user refinements, as discussed in previous sections; nor do they take into account the cost – in time, money, difficulty, level of completeness, or privacy concerns – for collecting each of the attributes from the trainees. However, it is possible to repeat this analysis so that both the attributes relative contribution to the model as well as the cost of collecting that attribute are taken into account, instead of just the relative contribution alone. Thus, while the performance of the model may be reduced by a modest amount over the "pure," value-only analysis, the usefulness of the end result could be substantially enhanced simply by using a value-over-cost ratio (instead of value-only) on all applicable attribute data.

One way to include attribute cost would be to debrief data collection experts, and there are a number of ways to do this. Open-ended discussion might allow for both the classification of each attribute into rough classification "bins," as well as inviting these experts to suggest additional data streams that might prove useful, and can be tested. Another method would be to simply have the experts rank each collected attribute by difficulty level, in whatever way the experts would define "difficulty." A third way would be to assign a number to each attribute, if the eventual ranking represents a non-uniform distribution (i.e., interval, instead of ordinal, data). Regardless, further analysis that take attribute cost into account can not only be more useful from an operational standpoint, but also allows for easy recalibration as new data collection methods are developed, and provides a context for targeting innovation on attributes that are valuable to the model, but expensive (currently) to collect.

Another important consideration is that these models were built using a population of marksmanship trainees who have already been selected as officer candidates, and therefore these data might not have the range of useful analysis that a fuller dataset would have, one that includes many types of soldiers. For example, of these officer candidates only one did not produce a minimum qualifying score. And so the model for very-low-scoring trainees is not robust against the adverse effects of having too little data. In short, we can't predict classifications that have only a limited number of examples for our algorithms.

Further, more data – even in the cases for which we do have plenty of examples – can often enhance the analysis by allowing us to interrogate the data in a wider variety of ways. If we have the ability to ask more questions of the data, then we can provide more refined and granular analysis of attribute contributions towards the model outputs.

Finally, how the model and its outputs can be used, both during the training and the pre-training processes, is also important for shaping future model refinements and research directions. It may be that we can provide more value by suggesting insertion points for these model outputs, as well as developing an array of model types based on hypothetical uses. For example, a model that predicts performance based on training data collected right before the live-fire certification process will use a different set of attribute weights than a model designed to make predictions before training has even commenced. Also, future models might be built directly based on practice data that takes into account shot X-Y coordinates, in order to produce predictions of failure types and provide formative feedback, both to the trainers and the trainees. Certain patterns that manifest in a trainee's specific shot-by-shot performance might suggest incorrect posture or weapon handling, or might instead display the characteristics of eye-dominance misidentification. Regardless, it is likely that X-Y coordinate analysis of the shot pattern can reveal additional information that is not accessible by the calculated score alone.

References

Chung, G. K., Delacruz, G. C., de Vries, L. F., Bewley, W. L., Baker, E. L. (2006). New Directions in Rifle Marksmanship Research. Military Psychology, 18(2), 161.

Maier, M.H. (1993). Military aptitude testing: The past fifty years. Technical Report 93-007, Monterey, CA.

James, D. R., & Dyer, J. L. (2011). Rifle Marksmanship Diagnostic and Training Guide for the Behavioral and Social Sciences Dept of the Army. Research Product 2011-07, Arlington, VA.

McManus, I.C., Porac, C., Bryden, M.P., & Boucher, R. (1999). Eye-dominance, writing hand and throwing hand. *Laterality, 4(2),* 173-92.

Aude, S. N., Bryson, J., Keller-Glaze, H., Nicely, K., Vowels, C. L. (2014). Preparing Brigade Combat Team Soldiers for Mission Readiness Through Research on Intangible Psychological Constructs and Their Applications: Phase 1 (ARI Technical Report No. 1336). Fort Belvoir, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Bandura, A. (2006). Guide for constructing self-efficacy scales. Self-efficacy beliefs of adolescents, 5, 307-337.

Cohen, S., Kamarck, T., Mermelstein, R. (1983). A global measure of perceived stress. Journal of Health and Social Behavior, 24(4), 385–396.

Duckworth, A. L., Peterson, C., Matthews, M. D., Kelly, D. R. (2007). Grit: perseverance and passion for long-term goals. Journal of Personality and Social Psychology, 92(6), 1087–1101.

Frese, M., Kring, W., Soose, A., Zempel, J. (1996). Personal initiative at work: Differences between East and West Germany. Academy of Management Journal, 39(1), 37–63.

Sottilare, R. A., Brawner, K. W., Goldberg, B. S., & Holden, H. K. (2012). The Generalized Intelligent Framework for Tutoring (GIFT). http://www.gifttutoring.org/documents.