Combining Human and Machine Intelligence for Making Predictions

Yiftach Nagar and Thomas W. Malone
Center for Collective Intelligence
Massachusetts Institute of Technology
Cambridge, MA 02142
ynagar at mit.edu; malone at mit.edu

Abstract

Statistical models almost always yield predictions that are more accurate than those of human experts. However, humans are better at data acquisition and at recognizing atypical circumstances. We use prediction markets to combine predictions from groups of humans and artificial-intelligence agents and show that they are more robust than those from groups of humans or agents alone.

1 Introduction

How can we make predictions about actions or behaviors in complex social systems? Recent advances in artificial-intelligence enable artificial agents to relatively successfully identify patterns even in complex scenarios [e.g. 1, 2]. Substantial evidence from multiple domains suggests that models usually yield better (and almost never worse) predictions than do individual human experts [3, 4]. Whereas models (or machines) are better at information processing and are consistent [5], humans suffer cognitive and other biases that make them bad judges of probabilities [6, 7]. In addition, factors such as fatigue can produce random fluctuations in judgment [3]. Indeed models of judges often outperform the judges themselves [8-10]. When working in groups, humans often exhibit effects such as groupthink [11] and group polarization [12] that negatively affect their judgment. Nevertheless, humans are still valuable in real-life prediction situations, for at least two good reasons. First, humans are still better at tasks requiring the handling of various types of information – especially unstructured information – including retrieval and acquisition [5, 13], categorizing [14], and pattern recognition [15, 16]. Second, humans' common-sense is required to identify and respond to "broken-leg" situations [17] in which the rules normally characterizing the phenomenon of interest do not hold. Therefore, combining human and machine predictions may help in overcoming the respective flaws of each. The scarcity of both theoretical and empirical work to that end is conspicuous. Previous work [5, 18, 19] emphasized the complementary nature of humans and models, but did not stress the potential of improving predictions by combining predictions from *multiple* humans and models. We know, however, that combining forecasts from multiple independent, uncorrelated forecasters leads to increased forecast accuracy whether the forecasts are judgmental or statistical [20-22]. Further, because it may be difficult or impossible to identify a single forecasting method that is the best, "it is less risky in practice to combine forecasts than to select an individual forecasting method" [23]. We conjecture, therefore, that in situations where rules are fuzzy or difficult to discern, and where some data are hard to codify, combining predictions from groups of humans and artificial-intelligence agents together can be more accurate and robust than those from groups of either type alone.

2 Method

To test this hypothesis, we used prediction markets to combine the predictions made by humans and artificial-neural-network agents of the plays in an American football game. This enabled us to emulate a realistic situation where humans and agents would have access to different information (specifically, humans had access to video information that is difficult or costly to codify for the agents). We hypothesized that 'hybrid' markets of humans and computers would do better than either markets of only computer-agents or only humans.

We conducted 20 laboratory sessions in which groups of 15 – 19 human subjects participated in prediction markets, both with and without computer agents. Overall there were 351 subjects, recruited from the general public via web advertising. The prediction markets used the Zocalo open source software platform (available at http://zocalo.sourceforge.net/). Simple neural net agents were developed using the JOONE open-source package (available at http://sourceforge.net/projects/joone/).

For each play, the agents had three pieces of previously coded information: the down number, the number of yards to first down, and whether the previous play was a run or pass. The agents were trained on a similar dataset of plays from one previous game. In addition, the agents considered the market price and traded only if they were confident about their prediction. After initial explanation and training rounds, each experimental session included 20 plays from the same football game. For each play, a short video excerpt from the game was shown to participants. The video was automatically stopped just before the play was about to start. Then, an online prediction market was opened, and the participants (either humans only, or humans plus AI agents) started trading contracts of RUN and PASS (other plays were eliminated from the video). After 3.5 minutes of trading, the market was closed, and the video continued. The video revealed what play had actually occurred and then continued until just before the next play.

In addition we ran 10 "computer-only" experimental sessions with 10 neural net agents each. In these sessions, the agents traded with each other in separate markets for each of the 20 plays. We thus got a total of 600 observations: 10 observations for 20 plays in 3 conditions (humans only, computers only, and hybrid of humans and computers).

3 Results

As we predicted, the Hybrid markets were the most accurate, followed by Agents-only and then Humans-only (see **Table 1**). These differences were statistically significant ($p<0.05$) for the Log Scoring Rule (LSR). Statistical significance was not tested for the Mean Square Error (MSE) scoring rule because these scores were not normally distributed.

Table 1 - Accuracy and ex post Sharpe ratio results by type of prediction markets

	Accuracy		Sharpe Ratio	
	MSE	LSR	AMSE (Benchmark: 0.75)	ALSR (Benchmark: 1.70)
Humans-only Markets	0.19	0.25	0.41	0.41
Agents-only Markets	0.17	0.23	0.39	0.37
Hybrid Markets	0.15	0.21	0.74	0.72

However, measures of accuracy alone do not provide sufficient information to convey the complexity of the data; it is important to consider both the accuracy and the variability of the errors. To do this, we used the ex post Sharpe ratio [24], originally developed to compare reward-to-risk performance of financial investments. To keep with the familiar logic of the Sharpe ratio, where higher positive returns are better, we adjust our scoring rules such that the adjusted MSE score (AMSE) equals 1-MSE. The adjusted Log score is $\log_{10}(P)$ where P

is the prediction (market closing price) of the actual outcome. As a simple and straightforward benchmark, we use an "ignorant" predictor who bets 50% PASS all the time (and whose error variance is therefore zero). The corresponding AMSE and ALSR for the benchmark predictor are therefore 0.75 and 1.70, respectively. We summarize the results in **Table 1**. Clearly, the hybrid markets yield the highest Sharpe ratio, which means they offer a better tradeoff between prediction accuracy and error variability.

Our comparisons of accuracy, and of the Sharpe ratio, both rely on attaching values to prediction errors using scoring rules. While common, these rules may not represent the actual economic value of predictions (or corresponding errors), and in reality, it is not always possible to determine those values. The Receiver-Operating-Characteristic (ROC) is an established methodology for evaluating and comparing the performance of diagnostic and prediction systems [25], which does not rely on their unknown economic value, and hence, can provide additional support for our conclusions. ROC curves are obtained by plotting the hit rate (i.e., correctly identified events) versus the false alarm rate (incorrect event predictions) over a range of different thresholds that are used to convert probabilistic forecasts of binary events into deterministic binary forecasts. The area under the curve serves as a measure of the quality of the predictions, with a perfect predictor scoring 1. The ROC curves of our conditions are presented in **Figure 1** and the areas under the curves are depicted in **Table 2**. This result echoes our other findings, and yet again, suggests that the hybrid markets were more robust.

Figure 1: ROC Plots for Study 1

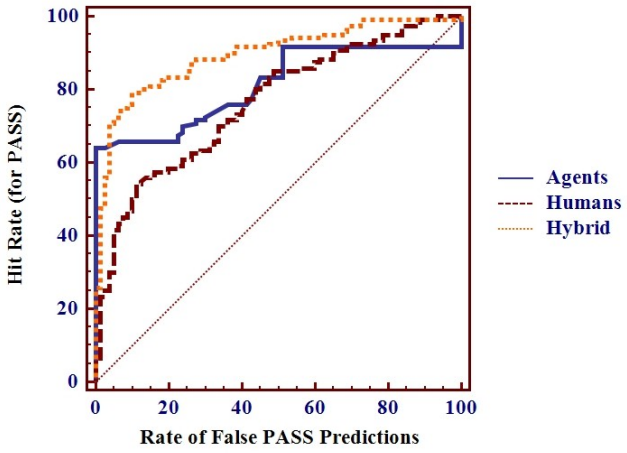

Table 2: Area under the ROC curves – all three conditions

	Area under ROC Curve	SE[1]
Humans-only Markets	0.76	0.03
Agents-only Markets	0.81	0.03
Hybrid Markets	0.90	0.02

[1] Standard errors were calculated using the method offered by DeLong, et al. [26]. However they may be inaccurate as we used repeated measurements.

4 Discussion

Predicting plays in a football game is a situation where the rules that determine a team's choice of plays are fuzzy or difficult to discern, and where some data about the game are hard to codify. In this context, we found that combining predictions from groups of humans and agents together led to overall predictions that were both more accurate and more robust (according to several measures of each) than predictions from groups of humans only or agents only.

While on average the agents were more accurate than humans, they had a higher number of big errors. This may have been due, in part, to the fact that the agents did not have access to as much information about the game as the humans did. For instance, informal interviews with the human subjects revealed that they indeed used information from the video that would have been difficult to code for agents (such as team formation, body language, and announcers' comments).

We thus provide a proof of concept of the existence of scenarios where combining predictions from groups of humans and artificial-intelligence agents can outperform groups of either type alone. We also show that prediction markets provide a useful way to combine predictions from humans and models, providing what we believe to be the first systematically studied attempt at using them for this purpose.

While prediction markets may offer only small improvements in accuracy over other methods of combining predictions (e.g. see [27]), they are appealing for additional reasons. Generally, they allow for dynamically updating predictions as new information becomes available, and they also incentivize gathering and sharing of knowledge. Further, people may be incentivized to have their expertise codified into agents, and then they still have an incentive to intervene in those cases where they think they can do a better job by trading manually (and not to do so when they do not).

Of course, additional work is required to identify and compare other ways of combining human and machine predictions, and to understand their respective advantages and disadvantages in different contexts. Future work should also examine our approach in more complex domains, and with more sophisticated, domain-specific agents. But we believe that the work reported here shows the potential value of combining predictions from humans and agents in situations with complex rules and difficult-to-codify information. We hope this initial work will encourage others to further investigate this promising direction.

Acknowledgments

We thank MIT Lincoln Laboratory for funding this project. We are grateful to John Willett for his patience, dedication and help with statistical analyses and to Chris Hibbert for software development, and education about prediction markets. We thank Sandy Pentland, Tomaso Poggio, Drazen Prelec, and Josh Tenenbaum for many discussions out of which this project originated, and benefited greatly. For help with software and experimental design we thank Jason Carver, Wendy Chang, Jeremy Lai and Rebecca Weiss. For their wise comments we thank John Carroll, Gary Condon, Robin Hanson, Haym Hirsh, Josh Introne, Ben Landon, Retsef Levi, David Pennock, Cynthia Rudin, Paulina Varshavskaya and two anonymous reviewers. Thanks also go to our research assistants Jonathan Chapman, Catherine Huang, Natasha Nath, Carry Ritter, Kenzan Tanabe and Roger Wong, and to Richard Hill and Robin Pringle for administrative support.

References

[1] A. Mannes, M. Michael, A. Pate *et al.*, "Stochastic Opponent Modeling Agents: A Case Study with Hezbollah," *Social Computing, Behavioral Modeling, and Prediction*, H. Liu, J. J. Salerno and M. J. Young, eds., pp. 37-45, 2008.

[2] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Transactions on pattern analysis and machine intelligence*, pp. 4-37, 2000.

[3] R. M. Dawes, D. Faust, and P. E. Meehl, "Clinical versus actuarial judgment," *Science,* vol. 243, no. 4899, pp. 1668-1674, 1989.

[4] W. M. Grove, D. H. Zald, B. S. Lebow *et al.*, "Clinical versus mechanical prediction: A meta-analysis," *Psychological Assessment,* vol. 12, no. 1, pp. 19-30, 2000.

[5] H. J. Einhorn, “Expert measurement and mechanical combination,” *Organizational Behavior and Human Performance,* vol. 7, no. 1, pp. 86-106, 1972.

[6] D. Kahneman, and A. Tversky, “On the psychology of prediction,” *Psychological review,* vol. 80, no. 4, pp. 237-51, 1973.

[7] S. Lichtenstein, F. Baruch, and L. D. Phillips, "Calibration of probabilities: The state of the art to 1980," *Judgment under uncertainty: Heuristics and biases*, D. Kahneman, P. Slovic and A. Tversky, eds.: Cambridge University Press, 1982.

[8] L. R. Goldberg, “Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences,” *Psychological Bulletin,* vol. 73, no. 6, pp. 422-432, 1970.

[9] J. S. Armstrong, "JUDGMENTAL BOOTSTRAPPING: INFERRING EXPERTS' RULES FOR FORECASTING," *Principles of forecasting: A handbook for researchers and practitioners*, J. S. Armstrong, ed., Norwell, MA: Kluwer Academic Publishers, 2001.

[10] T. R. Stewart, "Improving reliability of judgmental forecasts," *Principles of forecasting: A handbook for researchers and practitioners*, J. S. Armstrong, ed., pp. 81-106: Kluwer Academic Publishers, 2001.

[11] I. L. Janis, *Victims of groupthink: A psychological study of foreign-policy decisions and fiascoes*: Houghton Mifflin Boston, 1972.

[12] R. Brown, *Social psychology*, 2nd ed., New York, NY: Free Press, 1986.

[13] B. Kleinmuntz, “Why we still use our heads instead of formulas: toward an integrative approach,” *Psychological Bulletin,* vol. 107, no. 3, pp. 296-310, 1990.

[14] L. von Ahn, and L. Dabbish, “Labeling images with a computer game,” in Proceedings of the ACM SIGCHI conference on Human factors in computing systems, Vienna, Austria, 2004.

[15] N. Mitra, J., H.-K. Chu, T.-Y. Lee *et al.*, “Emerging images,” in ACM SIGGRAPH Asia, Yokohama, Japan, 2009.

[16] L. von Ahn, M. Blum, N. Hopper *et al.*, "CAPTCHA: Using Hard AI Problems for Security," *Advances in Cryptology — EUROCRYPT 2003*, Lecture Notes in Computer Science: Springer Berlin / Heidelberg, 2003.

[17] P. E. Meehl, *Clinical vs. statistical prediction: A theoretical analysis and a review of the evidence*, Minneapolis: University of Minnesota Press, 1954.

[18] R. C. Blattberg, and S. J. Hoch, “Database Models and Managerial Intuition: 50% Model+ 50% Manager,” *Management Science,* vol. 36, no. 8, pp. 887-899, 1990.

[19] D. Bunn, and G. Wright, “Interaction of judgemental and statistical forecasting methods: Issues and analysis,” *Management Science,* vol. 37, no. 5, pp. 501-518, May, 1991.

[20] J. S. Armstrong, "Combining forecasts," *Principles of forecasting: a handbook for researchers and practitioners*, J. S. Armstrong, ed.: Kluwer Academic Publishers., 2001.

[21] R. L. Winkler, “Combining forecasts: A philosophical basis and some current issues,” *International Journal of Forecasting,* vol. 5, no. 4, pp. 605-609, 1989.

[22] S. Makridakis, “Why combining works?,” *International Journal of Forecasting,* vol. 5, no. 4, pp. 601-603, 1989.

[23] M. Hibon, and T. Evgeniou, “To combine or not to combine: selecting among forecasts and their combinations,” *International Journal of Forecasting,* vol. 21, no. 1, pp. 15-24, 2005.

[24] W. F. Sharpe, “The Sharpe ratio,” *Journal of portfolio management* no. Fall, pp. 49-58, 1994.

[25] J. A. Swets, “Measuring the accuracy of diagnostic systems,” *Science,* vol. 240, no. 4857, pp. 1285-1293, 1988.

[26] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, “Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach,” *Biometrics,* vol. 44, no. 3, pp. 837-845, 1988.

[27] S. Goel, D. M. Reeves, D. J. Watts *et al.* "Prediction Without Markets," Paper presented at the 11th ACM conference on Electronic commerce, Cambridge, MA, 2010.