Real Time Decision Making on the Kaggle Platform

Michael S. Kim
Virginia Tech
Falls Church, VA
mikeskim(AT)gmail(dot)com

ABSTRACT

Kaggle is a platform that allows the crowdsourcing of predictive analytics problems. Clients submit data sets with a well defined ground truth. Data scientists[1] compete to obtain the best possible score on a private test set given a predefined scoring function. Often there is prize money involved for the top 3 teams on the final leaderboard.

We propose to extend the Kaggle platform for real time decision making. This would typically involve forecasting problems in fields such as energy, finance, sports, sales, and weather[2]. Data scientists would make forecasts and a running score would rank them. Each contest would be for some fixed length of time[3]. Prizes would be awarded at the end of the contest based upon total score. An ensembler and optimal controller would be built into the Kaggle system to combine the top predictions[4] and make a decision. One possible system would involve predicting energy use across some fixed region of units. At every time period t, Kagglers would make a prediction file for period t+1 using all data possible[5]. Each Kaggler would then be scored on the following time period once the ground truth is realized. The Kaggle platform would update the running scores on a public leaderboard, provide a new dataset with additional ground truth data, and make a decision on energy allocation based upon an ensemble of top Kagglers as determined by the running leaderboard.

While leakage is not an issue since participants are always scored on a ground truth that has yet to be realized at the time of prediction, there is still a possibility of collusion between teams [6]. This system has various issues involving participation entry deadlines, selection of scoring functions, and implementation of decision making based upon Kaggler input data. It will likely take some experimentation to work out the bugs in the system. However, we believe a correct implementation of this system will allow for better forecasting and decision making.

[1] http://www.kaggle.com/users/64626/mike-kim

[2] Any field that requires forecasting would be a candidate for this system.

[3] All participants would have to enter before scoring would start.

[4] A simple median of the top 5 leaderboard predictions would likely work well in practice without much computational overhead.

[5] This process could be automated by Kagglers so that one can predict multiple time periods into the future

[6] It might be possible to bootstrap this system for Kaggle competition cheating detection

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: [Human Computation]

General Terms

NIPS 2014 workshop

Keywords

Challenges in Machine Learning

1. REFERENCES

[1] Kaggle (2014). *Challenges in Machine Learning workshop at NIPS*. Retrieved September 28, 2014 from http://www.kaggle.com/forums/t/10468/challenges-in-machine-learning-workshop-at-nips

[2] R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.