Artificial Prediction Markets for Online Prediction

Fatemeh Jahedpari*
University of Bath
Bath, UK
F.Jahedpari@bath.ac.uk

Abstract

In this dissertation, we propose an online learning technique to predict a value of a continuous variable by (i) integrating a set of data streams from heterogeneous sources with time varying compositions including (a) changing the quality of data streams, (b) addition or deletion of data streams (ii) integrating the results of several analysis algorithms for each data source when the most suitable algorithm for a given data source is not known *a priori* (iii) dynamically weighting the prediction of each analysis algorithm and data source on the system prediction based on their varying quality.

1 Introduction

In some situations, it is desirable to integrate heterogeneous data sources where the quality of data streams is variable over time. In addition to varying quality, the availability of data streams is an issue. They may be permanently or temporarily deleted/ added. In addition, it may be desirable to analyse data streams with more than one analysis algorithm and combine their results. Given that the quality of a data stream is subject to change, the suitability of an analysis algorithm for a particular data stream may vary time to time. Hence, a mechanism to integrate data streams, each analysed with a variety of analysis algorithms, by dynamically weighting them based on their current quality, in these situations, is necessary.

One example of these situations is Syndromic Surveillance Systems. The main objective and challenge of a syndromic surveillance system is the earliest possible detection of a disease outbreak within a population. Various data sources such as medical absentee rates at schools, over-the-counter pharmacy sales and Internet and open source information can be used to identify a disease outbreak. However, their data quality fluctuates over time. For example, Google Flu Trends (GFT) may show false alerts as a result of a sudden increase in Influenza Like Illness (ILI) related queries due to unusual events, such as a drug recall for a popular cold or flu remedy [Ginsberg *et al.*, 2008; Copeland *et al.*, 2013], or a specific social media platform might become less popular over time. In addition to losing or gaining quality, data source availability can also change over time, in that a particular data stream might become available, or unavailable, for whatever circumstances. Given that the quality of data changes over time, and the most suitable algorithm for a given data source is not known *a priori*, a reasonable mechanism is analysing each data source with a variety of algorithms and integrate their results.

Therefore, our goal in this dissertation is to develop a technique which (i) integrates various data sources by adapting to dynamic environments where the availability and quality of data sources may change over time. More specifically, it should shift focus in response to changes in quality of each individual (combination of a data stream and analysis algorithm) prediction, (ii) is independent to each individual. Therefore, temporary or permanent deletion of an individual, for whatever circumstances, does not affect the system performance hugely, (iii) is resilient to different proportions of low- and high-performing individuals and does not impose any limits on the number of individuals or their quality, (iv) acts like an adaptive ensemble algorithm, (v) its overall performance is (at least) as good as the best individual performance.

In our model, we adopt the concept of prediction market. Prediction markets aim to utilise the aggregated wisdom of the crowd in order to predict the outcome of a future event [Ray, 2006]. In these markets, participants behaviour transfers their private information and beliefs about the possible outcomes by purchasing and selling instruments, also called securities, whose payoffs are tied to the occurrence of future events. A prediction market is run by a market-maker who interacts with traders to buy and sell securities. Physical prediction markets are typically used to predict discrete events such as the winner of an election, for example. We adapt the principle of the prediction market to support the prediction of a real value in a continuous domain to maintain high accuracy.

2 Progress to Date

To achieve our goal, mentioned in Section 1:

Initial Model: we put forward a model for an artificial prediction market whose input is a record (in the terminology of the machine learning literature) and output is a prediction.

*Supervised by: Marina De Vos and Julian Padget, University of Bath, UK. Special thanks to Sattar Hashemi, University of Shiraz, Iran and Benjamin Hirsch, EBTIC, UAE .

The model comprises market participants (called agents) and a market maker. The latter runs the market, deals with agent transactions and determines the market prediction. Each market includes a number of rounds, where each agent submits its bids to the market maker. Each bid contains a prediction value and the amount the agent is wagering on its prediction.

Each agent, using its assigned data source, analysis algorithm and accumulated knowledge, analyses its data and predicts the true value of the record and participates in the market. Subsequently, the market maker calculates the market prediction by combining all the individual predictions. Once the true value of the record is known and the market is over, agents are informed of the correct answer (the true value of the record) and are given an amount of revenue. The market maker uses a reward function to reward participants. Each agent learns from each market, based on the revenue they receive and the losses they make, in addition to finding out the correct answer. Consequently, they can, if desired, update their strategy, analysis algorithm and beliefs for future markets.

Integration and reward: in order to predict a continuous variable, we designed an integration function and a reward function. Our designed integration function combines each individual prediction by applying more weight to predictions backed by higher investments. Therefore, participants who accrue more capital, due to their success in earlier markets, have the opportunity to invest more and so get greater influence in the market.

The market maker rewards participants based on their prediction accuracy and their invested amount. Consequently, agents are incentivised to submit accurate prediction, and invest according to their confidence on their prediction. In addition, the agents with low capital (indicating low past performance) cannot invest and influence the market prediction as much as high performing agents, who acquire more capital over time.

Individual trading strategy: we also designed two trading strategies to be utilised by agents when participating in a market. The first one is *Constant Strategy* where agents simply dedicate a fixed ratio of their capital to bid in each round. The second one is *Q-Learning based strategy* [Watkins, 1989]. In this strategy, each agent estimates the expected reward for following different betting functions. It also advises the agent to what extent rely on the market prediction, the aggregated prediction of all market participants, as another source of information in order to improve its prediction in the subsequent rounds of a market. Consequently, high performing participants learn to ignore market predictions and low performing participants learn to minimise the amount of noise (low accurate prediction) they send to the market maker.

Therefore, using Q-Learning based strategy, agents learn in two levels (i) at the end of each record by updating their classifier/analysis algorithms with the correct answer of the record and updating their trading strategy based on how much they could earn if behaving differently, (ii) at the end of each round by updating their bids based on market prediction of the previous round. The market maker learns as it indirectly updates the weighting of each agent prediction on the final market prediction.

Evaluation: we evaluated the performance of our model by applying it to syndromic surveillance in the USA. In this context, the event to predict is the disease activity level of influenza-like illnesses on a specific date in the whole USA using publicly available data sources. The data used contains more than 100 real data streams from different sources including Google Flu Trends and Centers for Disease Control and Prevention (CDC), Google Trend, etc. The prediction of the system is then compared with CDC ILI rate.

The attained results are very promising, demonstrating: (i) The overall performance of the system is higher than that of the best performing market participant (combination of a data source and an analysis algorithm). (ii) The model is resilient to different proportions of low- and high-performing participants. (iii) The model outperforms well-known classifiers and ensembles. (iv) Adopting the Q-learning trading strategy, compared to the constant strategy, improves the system performance. (v) Adopting Q-learning based trading strategy improves each participating agent's performance.

We also compared our system prediction with Google Flu Trend prediction. We found out our system outperformed GFT in many situations, for example, during the 2012-2013 flu season where GFT overestimated influenza-like illness [Lazer *et al.*, 2014].

3 Future Plans

Our future plan is firstly to develop an intelligent market that can self-select the appropriate parameters for the market based on the characteristics of market participants and their data sources. Secondly, we want to apply our proposed model on different domains, such as, for example, stock market and cancer predictions to fully demonstrate its machine learning capabilities.

References

[Copeland *et al.*, 2013] Patrick Copeland, Raquel Romano, Tom Zhang, Greg Hecht, Dan Zigmond, and Christian Stefansen. Google disease trends: an update. *Nature*, 457:1012–1014, 2013.

[Ginsberg *et al.*, 2008] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2008.

[Lazer *et al.*, 2014] David M Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. The parable of google flu: traps in big data analysis. 2014.

[Ray, 2006] Russ Ray. Prediction markets and the financial "wisdom of crowds". *Journal of Behavioral Finance*, 7(1):2–4, 2006.

[Watkins, 1989] Christopher John Cornish Hellaby Watkins. *Learning from delayed rewards*. PhD thesis, University of Cambridge England, 1989.