Paper Review

Shaily Roy

December 2021

1. A Contextual-Bandit Approach to Personalized News Article Recommendation

- **What?** It proposes an efficient computational bandit approach to recommend articles in yahoo based on user interests and can work both online and offline using available records. The paper combines the feature-based exploration process and contextual side information of the articles in order to recommend new articles to the user.

  Dataset was collected from random bucket considering three components: (i) the random article chosen to serve the user, (ii) user/article information, and (iii) whether the user clicks on the article at the story position. Each article was represented by a raw feature vector of about 100 categorical features constructed in the same way. These features include: (i) URL categories: tens of classes inferred from the URL of the article resource; and (ii) editor categories: tens of topics tagged by human editors to summarize the article content.

- **Why?** Recommending correct articles is important to maximize users. This paper wants to show result based on who is using it and so stores previous article records and new article ratings. They also encountered the number of clicks an article got to measure its popularity and then recommend it based on the user's previous activity. It increases the accuracy of recommending new articles in a search engine. The process starts with 0 step and start making new states based on the previous states and gives better flexibility in grouping users with the same taste and work more efficiently.

- **How?** Contextual bandit algorithm chooses an user to observe his/her activities and use a feature vector to store the summarized information, then based on the previous data it interprets the next activity with a regret value which depends on the current user and his/her current activity. It continues observing the user activities, feature vectors and previous pay off values to maximize the regret score. Feature vectors are parameterized based on states and activities. When the information is low, the regret score is random and not up to the mark. However, with training, it starts exploring more, produces expected recommendations and confidence bound starts getting tight.

  They proposed contextual bandit algorithm Lin UCB which is UCB with a linear parameterization assumed for the value functions.

$$E[r_{t,a}|x_{t,a}] = {x_{t,a}}^T {\theta^*}_a \quad (1)$$

The equation explains Expectation value 'E' of a particular user activity 'a' based on the feature vector 'r' equals multiplication of a random co-efficient vector 'θ' powered by total number of trials 'T' and all the feature vectors '$x_{t,a}$' of each activity. As the parameters are not shared among the different arms, the set is disjoint. Then the coefficient is calculated applying ridge regression to the training data using a matrix. They have shown a complete pseudo code of the algorithm which is quite self-explanatory. The input parameter of this algorithm can be large based on problem set and can create complexity in payoffs. Its computational complexity is linear in the number of arms and at most cubic in the number of features, and works better in dynamic arm set with small activity set.

They improve the algorithm using hybrid lin-UCB approach.

$$E[r_{t,a}|x_{t,a}] = {x_{t,a}}^T {\theta^*}_a + {x_{t,a}}^T \beta^* \quad (2)$$

where a linear term ${x_{t,a}}^T \beta^*$ is added. β is an unknown n coefficient vector common to all arms.This model is hybrid in the sense that some of the coefficients β are shared by all arms, while others ${\theta^*}_a$ are not.The algorithm 2 includes some more calculations to decrease complexity and find out an exact block dimension. Algorithm 3 shows an evaluation approach which takes policy and events as input , stream the logged events of user, based on the data they choose an arm and update the expected value. then events which are not included in data stream are completely ignored. Because the logging policy chooses each arm uniformly at random, each event is retained by this algorithm with probability exactly 1/K, independent of everything else. This means that the events which are retained have the same distribution as if they were selected by D. As a result, they proved that two processes are equivalent: the first is evaluating the policy against T real-world events from D, and the second is evaluating the policy using the policy evaluator on a stream of logged events.

After collecting dataset, they have implemented encoding method to process categorical data and clustered users with similar interests. Then parameter tuning was done based on the value of CTR in learning bucket and deployment bucket using greedy, UCB and omniscient algorithms where UCB performed better in both bucket. An algorithm's CTR is defined as the ratio of the number of clicks it receives and the number of steps it is run. They tried different size of buckets to find out the best performance. It is clear from Table 1 that, by considering user features, both epsilon-greedy (seg/disjoint/hybrid) and UCB methods (UCB (seg) and linUCB (disjoint/hybrid)) were able to achieve a CTR lift of around 10%, compared to the baseline epsilon-greedy. Then some graphs were visualized to see the performance of these algorithms in extreme cases where both UCB and linUCB outperformed others. According to the authors it is not a co-incidence. it was happened because features in the disjoint model are actually normalized membership measures of a user in the five clusters.linUCB shows advantages when data are sparse.

- **Observation** In order to maximize regret value and produce an optimal result, traditional bandit overlooks sub-optimal cases which can be the optimal approach for that particular problem set. To solve this issue, they have used multi-armed contextual bandit approach but it can generate exploitation (what has already been learned) vs exploration(to learn which behavior gives best result) dilemma.On the one hand the learner wants to exploit what has already been learned to behave in a way that will maximize rewards but on the other hand, if you always act in the way that you think is best based on your experience you might be missing out on other great ways of behaving new articles to be discovered whatever and so it needs to do some exploration so that it can learn new behaviors which might give better results.so that's a dilemma how to balance between those two. To develop that exploration process for syntactic data, meta-learning can be a good strategy to be included. They have jumped from equation 2 to 3 and the explanation was not really clear.

---

2. Improved Algorithms for Linear Stochastic Bandits

- **What?** The paper proposes an improved algorithm of linear UCB bandit approach. The smaller confidence sets one is able to construct, the better regret bounds one obtains for the resulting algorithm, and, more importantly, the better the algorithm performs empirically. So they vastly reduced the size of the confidence sets. Their target is to maximize the regret value.
- **Why?** while working with multi-arm bandit, The arms which are unexplored have low confidence bound and new articles that enters the pool of previous news articles, have never been explored previously. So probability of not getting expected new news becomes higher. It means everybody who comes to the page will get shown the same. UCB can solve this problem but the regret function is random. This paper proposed an approach which will produce efficient and constant regret function. it will improve the regret bound by logarithmic factor.
- **How?** the algorithms that we have read in the previous paper are based on the same underlying idea—the optimism-in-theface-of-uncertainty (OFU) principle. The OFU principle elegantly solves the exploration-exploitation dilemma inherent in the problem. The basic idea of the principle is to maintain a confidence set for the vector of coefficients of the linear function. In every round, the algorithm chooses an estimate from the confidence set and an action so that the predicted reward is maximized, i.e., estimate-action pair is chosen optimistically. In the regret function calculation , they remove the additive noise using a simple calculation which ultimately looks like the previous equations but not the same , because the new equation has no noise. then they apply OFU approach. Because of random decision matrix, activities can be arbitrary as well which may cause stochastic dependencies. To solve this activity vector can be normalized.After that confidence sets are constructed using theorem 2 where they used linear relations. They have analyzed the regret score of OFUL by giving bound to regret

with the confidence set. They save all the computations with log(n) time complexity. By using multi armed band it problem they prove high probability constant regret factor. Visualizing by a graph, they have shown UCB with the traditional confidence factor performs poorly than their proposed confidence bound.

- **Observation** the new algorithm balances exploration and exploitation. They have theoretically proved the improvement, no real-life application or result analysis are shown. but as the paper is theoretically well explained, it should be enough.In Lemma 8, the equation is not that much well explained,

---