Contextual Bandit for Active Learning: Active Thompson Sampling

Djallel Bouneffouf, Romain Laroche, Tanguy Urvoy, Raphael Feraud, Robin Allesiardo

Orange Labs, 2, avenue Pierre Marzin, 22307 Lannion, France
{djallel.bouneffouf,
romain.laroche,
tanguy.urvoy,
raphael.feraud,
robin.allesiardo}@Orange.com

Abstract. The labelling of training examples is a costly task in a supervised classification. Active learning strategies answer this problem by selecting the most useful unlabelled examples to train a predictive model. The choice of examples to label can be seen as a dilemma between the exploration and the exploitation over the data space representation. In this paper, a novel active learning strategy manages this compromise by modelling the active learning problem as a contextual bandit problem. We propose a sequential algorithm named Active Thompson Sampling (ATS), which, in each round, assigns a sampling distribution on the pool, samples one point from this distribution, and queries the oracle for this sample point label. Experimental comparison to previously proposed active learning algorithms show superior performance on a real application dataset.

Keywords: Contextual Bandits, Active learning, Thompson sampling

1 Introduction

Active Learning (AL) has emerged as a popular approach for solving machine learning problems with limited labelled data [8]. In this approach the learning algorithm is "active", and is allowed to query, an oracle O, for the label of points that are maximally informative for the learning process. The result is that, by using few but well chosen labels, the active learning algorithm is able to learn as well as a passive learning algorithm that has access to more labelled data.

In selective sampling, the choice of examples to be labelled can be seen as the dilemma between exploration and exploitation (exr/exp) on the training data. On one hand, an active learning strategy that just exploits the data will be specialized in certain areas of the input space X but will be very poor in generalization. On the other hand, a strategy which uses that exploring data does not focus on regions where X is known to improve the predictive model. These two situations illustrate the need for an active learning to find a compromise between exr/exp of labelling data strategy.

In [5], a similar analysis of the problem led the authors to model the active learning problem as a multi-armed bandit problem. They suppose that the different hypotheses of distribution $h \in H$ of the data are the arms and use an adapted UCB (upper confident bound) to select the most promising hypothesis of distribution to select the points to be labelled. The drawback of this approach is in the non consideration of the context or the different features that characterize the points. For example the number of points in a area space, the proportion of labelled points, the ratio of the classes in the area or the density of points in the area can be useful to determine the most interesting to label.

To tackle this problem, we propose to model the active learning as a contextual bandit problem, where we have at first clustered the input space: each cluster is considered as an arm and the different features of the cluster are the context of the arm. Then, we implement a novel algorithm named Active Thompson Sampling (ATS) adapting the Thompson Sampling to the active learning problem. Finally, we evaluate ATS on actual data and find out that ATS outperforms all other algorithms in our panel.

The remaining of the paper is organized as follows. Section 2 reviews related works. Section 3 describes our multi armed contextual bandit model and the ATS algorithm. The experimental evaluation is illustrated in Section 4. The last section concludes the paper and points out possible directions for future works.

2 Related Work

We refer, in the following, recent works that address Active Learning problem and the exr/exp trade-off (bandit algorithm).

Active Learning. A variety of AL algorithms have been proposed in the literature employing various query strategies. One of the most popular strategy is called uncertainty sampling (US), where the active learner queries the point whose label is the most uncertain [6]. Usually the uncertainty in the label is calculated using variance of the label distribution [8]. The authors in [9] introduced the query-by-committee (QBC) strategy where a committee of potential heterogeneous models, is learnt from the labelled data, and used to select for querying, the point where most committee members disagree. Other strategies include the maximum expected reduction in error [11] or variance reducing query strategies [10] to querying the optimal point. All above proposed approaches only exploit the data.

Contextual Multi-armed Bandit. Multi-armed bandit (MAB) problems model the exr/exp trade-off inherent in many sequential decision problems. A particularly useful version is the contextual multi-armed bandit problem. In this problem, in each iteration, an agent has to choose between arms. Before making the choice, the agent sees a d-dimensional feature vector (context vector), associated with each arm. The learner uses these context vectors along with the rewards of the arms played in the past to make the choice of the arm to play in the current iteration. Overtime, the learner's aim is to collect enough information

about how the context vectors and rewards relate to each other, so that it can predict the next best arm to play by looking at the feature vectors.

Recently, the contextual bandit has been used in different domains such as recommender system (RS) and information retrieval. For example, in [3, 2], authors model RS as a contextual bandit problem. The authors propose an algorithm called Contextual-ϵ-greedy which sequentially recommends documents based on contextual information about the users. In [4], authors analyse the Thompson Sampling (TS) in contextual bandit problem. The study demonstrates that it has a better empirical performance compared to the state-of-art methods. The TS is one of the oldest heuristics for multi-armed bandit problems and it is a randomized algorithm based on Bayesian ideas. Authors in [4, 3] describe a smart way to balance exr/exp, but do not study the contextual bandit in the active learning problem.

Multi-armed Bandit for Active Learning. To our knowledge there has been only two papers bridging the world of active learning and MAB. [7] adapted the EXP4 algorithm which is a MAB algorithm with expert advice, where the different active learning algorithms are the various experts and the different points in the pool are the arms of the MAB. An each iteration, every expert provides a sampling distribution on the pool. EXP4 maintains an estimation of the error rate for each expert, and uses exponential weight to select the optimal sampling distribution on the pool. Authors in [5] propose an adaptation of UCB called LCB algorithm, the authors suggested minimizing an unbiased estimator of risk of h, and a sampling distribution that was in proportion to the entropy of the prediction on the pool. The authors consider the arms of the bandit as the different hypothesis, and querying a data point, as the process of improving their estimation of the risk of the different hypothesis.

Our Contributions. As it is observed above, none of the described works has dressed the active learning problem from a contextual bandits view, although the consideration of the pool context might be a very informative feature for an active learning algorithm. This is precisely what we intend to do by exploiting the following new features: (1) We model the active learning as a contextual bandit problem, where each cluster of points in the space is an arm and the different features of the cluster are the context of the arms. (2) We propose a new algorithm named Active Thompson Sampling (ATS), that adapts the TS to the active learning problem. (3) We evaluate it against other methods form the state-of-the-art.

3 Key Notions and Proposed Model

This section focuses on the proposed model, beginning by introducing the key notions used in this paper.

In pool based AL we are provided with a pool $U_0 = \{x_1, ...x_n\}$ of unlabelled points, an empty set of labelled points $L_0 = \{\}$ and a labelling oracle O, which when queried for the label of x, returns y. Algorithms in the pool based setting have to decide which points to query by looking at the entire pool.

Definition (Contextual Bandit Problem with Linear Payoffs). In a contextual bandits problem with linear payoffs, there are N arms. At time $t = 1, 2, ...,$ a context vector $b_i(t) \in R^d$, is revealed for every arm i. History $Q_{t-1} = \{a_\tau, r_\tau, b_i(\tau), i = 1, ..., N, \tau = 1, ..., t-1\}$ where a_τ denotes the arm played at time τ and that triggered reward r_τ. Given $b_i(t)$, the reward for arm i at time t is generated from an (unknown) distribution with mean $b_i(t)^\top \mu$, where $\mu \in R^d$ is a fixed but unknown parameter. An algorithm for the contextual bandit problem needs to choose, at every time step t, an arm a_t to play, using history Q_{t-1} and current contexts $b_i(t), i = 1, ..., N$.

Let a_t^* denote the optimal arm at time t, i.e. $a_t^* = \underset{i}{argmax} b_i(t)^\top \mu$. And let $\Delta_i(t)$ be the difference between the mean rewards of the optimal arm and the arm i played at time t, i.e., $\Delta_i(t) = b_{a_t^*}(t)^\top \mu - b_i(t)^\top \mu$. Then, the regret at time t is defined as $regret(t) = \Delta_{a_t}(t)$. The objective is to minimize the total regret $R(T) = \sum_{t=1}^{T} regret(t)$. The time horizon T is finite and known in our case.

To model the active learning problem as a contextual bandit with linear payoffs we need to define the arms of the bandit, the rewards of the environment and the context of each arms.

Construction of the Arms. We cluster corpus points $\{x_1, x_2, ..., x_n\}$. The resulted clusters $c \subset U$ are considered as the arms of the bandit.

Context of the arms. We consider a context vector $b_i(t)$ that describes the the arms (the clusters), and contains the features that characterise the clusters.

Reward. A metric is used to measure the variation of the hypothesis learned by the model between two iterations. More the hypothesis learned by the model varies more is the received reward. We now define the function $d(h_{t-1}, h_t)$ that we use to get the variation of the model. Let $U_0 = \{x_1, ..., x_n\} = L_t \cup U_t$ be the set of labelled and unlabelled training examples that we have. Then for each of the two real-valued hypotheses $h_{t-1}(.), h_t(.)$, we define the vectors $H_{t-1} = (h_{t-1}(x_1), h_{t-1}(x_2), ..., h_{t-1}(x_n))$ and $H_t = (h_t(x_1), h_t(x_2), ..., h_t(x_n))$, i.e. vectors of the real-valued predictions of h_{t-1} and h_t.

$$d(h_{t-1}, h_t) = \frac{H_{t-1}.H_t}{||H_{t-1}||.||H_t||} \tag{1}$$

In Eq. 1, we compute the cosine similarity between the two vectors H_{t-1} and H_t. Thus $d(h_{t-1}, h_t) \in [-1, +1]$ is the cosine of the angle between H_{t-1} and H_t, and we normalise the result in the interval [0, 1] using $y(t) = \frac{2.cos^{-1}(d(h_{t-1}|h_t))}{\pi}$.

Stationarity of the reward. We have observed that, more an area is sampled by the model less is the received reward (nonstationarity of the rewards). We have confirmed this common sense idea from an off-line evaluation (see Fig. 2). In order to circumvent this nonstationarity we assume that the reward function $y(t) = r_t \cdot D(t)$, where $D(t)$ is a decreasing function that follows the decreasing reward given by the environment and r_t is the stationary reward. The process for obtaining the function $D(t)$ is described in Section 4.

Contextual Bandit Algorithm. A Contextual bandits algorithm determines a cluster $c \subset U_t$ to be sampled at each time step t, based on the previous

observation sequence $Q_{t-1} = \{c_\tau, r_\tau, b_c(\tau), c = 1, ..., N, \tau = 1, ..., t-1\}$, and its current context $b_c(t)$.

3.1 Active Thompson Sampling

Thompson sampling is understood in a Bayesian setting as follows. The set of past observations Q is made of triplets $(c_t, r_t, b_c(t))$ and are modelled using a parametric likelihood function $Pr(r_t|\tilde{\mu})$ depending on some parameters $\tilde{\mu}$. Given some prior distribution $Pr(\tilde{u})$ on these parameters, the posterior distribution of these parameters is given by the Bayes rule, $Pr(\tilde{\mu}|r_t) \propto Pr(r_t|\tilde{\mu})Pr(\tilde{\mu})$.

From [1], we can say that the posterior distribution at time $t+1$, $Pr(\tilde{\mu}|r_t) \propto Pr(r_t|\tilde{\mu})Pr(\tilde{\mu})$ were given by a multivariate Gaussian distribution $\mathcal{N}(\hat{\mu}(t+1), v^2B(t+1)^{-1})$, where $B(t) = I_d + \sum_{\tau=1}^{t-1} b_{c_\tau}(\tau)b_{c_\tau}(\tau)^\top$ with d the size of the context vectors, $v^2 \in]0, 1]$ is a constant fixed to 0.25 according to [4] and $\hat{\mu} = B(t)^{-1}(\sum_{\tau=1}^{t-1} b_{c_\tau}(\tau)b_{c_\tau}(\tau))$. Every step t consists of generating a d-dimensional sample $\tilde{\mu}$ from $\mathcal{N}(\hat{\mu}(t), v^2B(t)^{-1})$, and solving the problem $\underset{c \subset U_t \wedge |c|>0}{argmax}\ b_c(t)^\top \tilde{\mu}$ (select the cluster c that maximizes $b_c(t)^\top \tilde{\mu}$. After that the algorithm selects randomly an individual $x \in c$, requests a labelling from the oracle O and observes reward $y(t)$.

Algorithm 1 The Active Thompson Sampling algorithm

1: **Require.** $B = I_d$ set $\hat{\mu} = 0_d, f = 0_d$.
2: **Foreach** $t = 1, 2, ..., T$ **do**
3: Sample $\tilde{\mu}$ from the $N(\hat{\mu}, v^2B^{-1})$ distribution.
4: Select cluster $c_t = \underset{c \subset U_t \wedge |c|>0}{argmax}\ b_c(t)^\top \tilde{\mu}$
5: $x_t = Random(c_t)$.
6: Query O for label y_t of x_t
7: Observe $y(t)$ and compute r_t
8: $B = B + b_{c_t}(t)b_{c_t}(t)^T$, $f = f + b_{c_t}(t).r_t$ **else** $\hat{\mu} = B^{-1}f$
9: **End**

4 Experimental Evaluation

To conduct our evaluation, we have got from our company a corpus containing French utterances collected from a commercial spoken dialogue system. There are 7 765 utterances annotated by human experts. The unannotated part consists of 3 911 695 utterances.

We use a corporate supervised algorithm (rule-based algorithm), being a part of a spoken language dialogue system. We simulate in the experiments an expert (oracle) on the unannotated corpus by using the rule-based algorithm which was designed using the 7 765 annotated utterances. In our experiments, the clustering algorithm (k-means in our case) uses the cosine similarity as a similarity metric

between utterances. We have considered different features in the context vector of the clusters $b_c(t) = (Mdis_c, Vdis_c, |c|, plb_{c,t}, MixRate_{c,t})$, where $Mdis_c$ and $Vdis_c$ are respectively the average distance between individuals in the cluster, and its variance. $|c|$ gives the number of points in the cluster. $plb_{c,t}$ gives the proportion of labelled individuals in the cluster at time t and $MixRate_{c,t}$ gives the ratio of the classes in the cluster at time t (the proportion of examples labelled in each class in the cluster). To obtain the decreasing function $D(t)$, we assume $D(t) = \alpha e^{-\beta t}$, and we compute the parameters α and β of $D(t)$ by sampling uniformly the different clusters and drawing the rewards in Fig.1, then we lit $D(t)$ on the reward curve. We obtain $\alpha = 0.61$ and $\beta = 0.12$

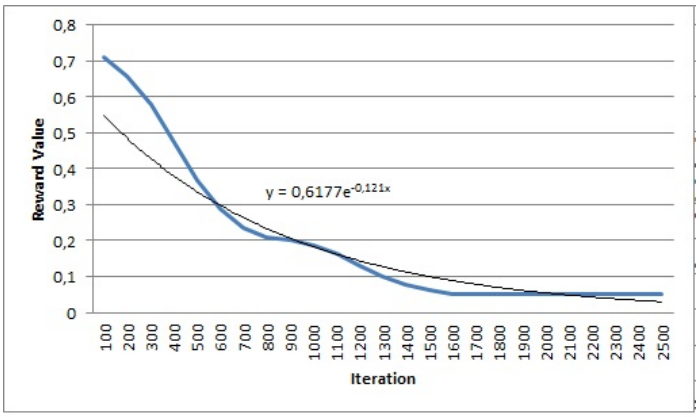

Fig. 1. Reward function

To evaluate the active learning algorithms, we have considered a version of the rule based algorithm without training. At each iteration the active learning selects from the unannotated corpus the relevant utterances to annotate and integrates it in the training set of the rule based algorithm. By relating the results to the newer versions, one can verify the usefulness of the proposed approach. We average the regret over 1000 times with a time horizon of 2000 sentences to label which correspond to our budget in term of labelling. To compute the regret, we have supposed that the optimal policy is given by the oracle. In addition to the random (baseline), we have compared our algorithm to the ones described in the related work (Sec. 2). QBC, US, and the different approaches that consider the bandit algorithms in the active learning as EXP4 used in [7] and LCB used in [5]. In Fig. 2, the horizontal axis represents the number of iterations and the vertical axis gives the cumulative regret (performance metric) which is the sum of the regrets from the first iteration to the current iteration.

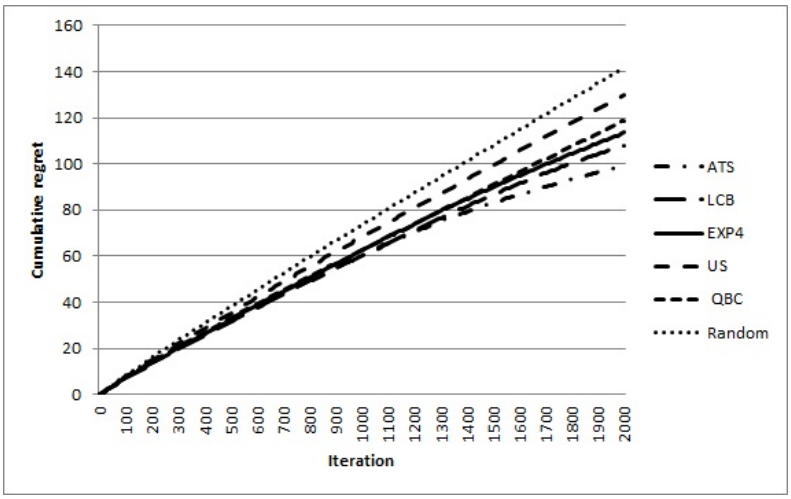

Fig. 2. Cumulative Regret for Active learning algorithms

From the Fig. 2 we observe that over all strategies gives better result than a random selection. Neither QBC nor US gives a good results. This confirms that a pure exploitation is not efficient, and it confirms the need of the exr/exp trade-off. While EXP4 algorithm gets a low cumulative regret, its overall performance is not as good as LCB and ATS. ATS and LCB indeed have the best cumulative regrets, ATS decreases the cumulative regret with 29% over the baseline and LCB, with 23%. The improvement comes from a dynamic exr/exp. These algorithms take full advantage of exploration from the beginning of the exploration rather than other strategies like uncertainty sampling or request by committee that need enough iteration to construct their models. Finally, as expected, ATS outperforms LCB, which is explained by the consideration of the context, and also that TS performs better exr/exp trade-off than UCB.

5 Conclusion

In this paper, we study the active learning problem from the side of contextual bandit and propose a new approach that adaptively balances exr/exp regarding the context of the cluster (arms). We have validated our work with data from real-world application and shown that the proposed algorithm offered promising results. This study yields to the conclusion that considering the contextual bandit model for the active learning significantly increases the results. Considering these results, we plan to study the theoretical regret of the proposed algorithm.

Bibliography

[1] S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. In *ICML (3)*, pages 127–135, 2013.

[2] D. Bouneffouf. *DRARS, A Dynamic Risk-Aware Recommender System.* PhD thesis, Institut National des Télécommunications, 2013.

[3] D. Bouneffouf, A. Bouzeghoub, and A. L. Gançarski. A contextual-bandit algorithm for mobile context-aware recommender system. In *ICONIP (3)*, pages 324–331, 2012.

[4] O. Chapelle and L. Li. An empirical evaluation of thompson sampling. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. C. N. Pereira, and K. Q. Weinberger, editors, *NIPS*, pages 2249–2257, 2011.

[5] R. Ganti and A. G. Gray. Building bridges: Viewing active learning from the multi-armed bandit lens. *CoRR*, abs/1309.6830, 2013.

[6] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, pages 3–12, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[7] T. Osugi, D. Kim, and S. Scott. Balancing exploration and exploitation: a new algorithm for active machine learning. In *Data Mining, Fifth IEEE International Conference on*, pages 8 pp.–, Nov 2005.

[8] B. Settles. Active Learning Literature Survey. Technical Report 1648, University of Wisconsin–Madison, 2009.

[9] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pages 287–294, New York, NY, USA, 1992. ACM.

[10] T. Zhang and F. J. Oles. A probability analysis on the value of unlabeled data for classification problems. In *17th International Conference on Machine Learning*, 2000.

[11] X. Zhu, J. Lafferty, and Z. Ghahramani. Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In *ICML 2003 workshop on The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining*, pages 58–65, 2003.