Survey on Applications of Multi-Armed and Contextual Bandits

1st Djallel Bouneffouf
IBM Research
Djallel.bouneffouf@ibm.com
New York, USA

2nd Irina Rish
IBM Research
irish@ibm.com
New York, USA

3rd Charu Aggarwal
IBM Research
charu@us.ibm.com
New York, USA

***Abstract*—In recent years, the multi-armed bandit (MAB) framework has attracted a lot of attention in various applications, from recommender systems and information retrieval to healthcare and finance. This success is due to its stellar performance combined with attractive properties, such as learning from less feedback. The multi-armed bandit field is currently experiencing a renaissance, as novel problem settings and algorithms motivated by various practical applications are being introduced, building on top of the classical bandit problem. This article aims to provide a comprehensive review of top recent developments in multiple real-life applications of the multi-armed bandit. Specifically, we introduce a taxonomy of common MAB-based applications and summarize the state-of-the-art for each of those domains. Furthermore, we identify important current trends and provide new perspectives pertaining to the future of this burgeoning field.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

Many practical applications require sequential decision-making problems, where an agent must choose the best action out of several alternatives. Examples of such applications include clinical trials [1], recommender systems in temporal settings [2] and anomaly detection [3]. In some cases, side information, or *context*, is associated with each action (e.g., a user's profile), and the feedback, or *reward*, is limited to the chosen option. For example, in clinical trials [**?**], [1] the context is the patient's medical record (e.g., health condition, family history, etc.), the actions correspond to the treatment options being compared, and the reward represents the outcome of the proposed treatment (e.g., success or failure). An important aspect affecting the long-term success in such settings is finding a good trade-off between *exploration* (e.g., trying a new drug) and *exploitation* (choosing the best known drug so far).

This inherent trade-off between exploration and exploitation exists in many sequential decision-making problems, and is traditionally formulated as the *multi-armed bandit (MAB)* problem, which is stated as follows: Given K possible actions, or "arms", each associated with a fixed but unknown reward probability distribution [4], [5], at each iteration (time point) an agent selects an arm to play and receives a reward, sampled from the respective arm's probability distribution independently from the previous actions. The task of an agent is to learn how to choose its actions so that the cumulative rewards over time are maximized. Note that the agent needs to try different arms in order to learn their rewards (i.e., explore the payoff), and also use this learned information in order to receive the best payoff (exploit the learned payoffs). There is a natural trade-off between exploration and exploitation. For example, trying each arm exactly once and then playing the best one among them forever is often likely to lead to highly suboptimal solutions when the rewards from the arms are uncertain. Different solutions have been proposed for this problem, based on a stochastic formulation [4]–[6] and a Bayesian formulation [7]; however, these approaches did not account for the context or side information available to the agent.

It is noteworthy that the multiarmed bandit problem can be seen as the simplest form of reinforcement learning, in which the agent is *stateless*. When the system is not stateless, the actions causes changes in states and the rewards also depend on the states. Therefore, in general reinforcement learning, the rewards in different steps are not independent of one another. In fact, the classical algorithms for reinforcement learning (with states) often use solutions to the multiarmed bandit problem as subroutines for defining policies in (general) reinforcement learning. For example, the well-known ϵ-greedy algorithm in multi-armed bandits is often com-

Identify applicable funding agency here. If none, delete this.

978-1-7281-6929-3/20/$31.00 ©2020 IEEE

bined with Bellman's dynamic programming algorithm for reinforcement learning in order to define choices of actions. Furthermore, many reinforcement learning algorithms, when applied to stateless systems, reduce to multi-armed bandit algorithms.

A particularly useful version of the MAB is the *contextual multi-arm bandit (CMAB)*, or simply the *contextual bandit* problem, where at each iteration, before choosing an arm, the agent observes an N-dimensional *context*, or *feature vector*. The agent uses this context, along with the rewards of the arms played in the past, to choose which arm to play in the current iteration. Over time, the agent's aim is to collect enough information about the relationship between the context vectors and rewards, so that it can predict the next best arm to play by looking at the current context [8], [9]. Different algorithms were proposed for the general case, including LINUCB [10], Neural Bandit [11] and Contextual Thompson Sampling (CTS) [9], where a linear dependency is typically assumed between the expected reward of an action and its context.

We will now provide an extensive overview including various applications of the bandit framework, both in real-life problem setting arising in multiple practical domains (healthcare, computer network routing, finance, and beyond), as well as in computer science and machine-learning in particular, where bandit approaches can help improve hyperparameter tuning and other important algorithmic choices in supervised learning, active learning and reinforcement learning.

II. Real-Life Applications of Bandit

As a general mathematical framework, the stochastic multi-armed bandit setting addresses the challenges associated with the presence of uncertainty in sequential decision-making. This type of uncertainty has a complex interplay with the exploration versus exploitation dilemma, and therefore provides a natural formalism for most real-life online decision-making problems.

A. Healthcare

Clinical trials. Collecting data for assessing treatment effectiveness on animal models during the full range of disease stages can be difficult when using conventional random treatment allocation procedures, since poor treatments can cause deterioration of subject's health. The authors in [1] aim to design an adaptive allocation strategy to improve the efficiency of data collection by allocating more samples for exploring promising treatments. They cast this application as a contextual bandit problem and introduce a practical algorithm for exploration vs. exploitation in this framework. The work relies on sub-sampling to compare treatment options using an equivalent amount of information. They extend the sub-sampling strategy to the contextual bandit setting by applying sub-sampling within Gaussian Process regression.

Warfarin is the most widely used oral anticoagulant agent in the world; however, administering an accurate dosage remains a significant challenge, as the appropriate dosage can be highly variable among individuals due to various clinical, demographic and genetic factors. Physicians currently follow a fixed-dose strategy: they start patients on 5mg/day (which is the appropriate dosage for the majority of patients) and slowly adjust the dose over the course of a few weeks by tracking the patient's anticoagulant levels. However, an incorrect initial dosage can result in highly adverse consequences such as stroke (if the initial dose is too low) or internal bleeding (if the initial dose is too high). Thus, the authors in [12] tackle the problem of learning and assigning an appropriate initial dosage to patients by modeling the problem as a multi-armed bandit with high-dimensional covariates, and propose a novel and efficient bandit algorithm based on the LASSO estimator.

Brain and behavioral modeling. Drawing inspirations from the behavioral studies of human decision making in both healthy controls and patients with different mental disorders, the authors in [13] propose a general parametric framework for multi-armed bandit problem which extends the standard Thompson Sampling approach to incorporate reward processing biases associated with several neurological and psychiatric conditions, including Parkinson's and Alzheimer's diseases, attention-deficit/hyperactivity disorder (ADHD), addiction, and chronic pain. They demonstrate empirically, from the behavioral modeling perspective, that their parametric framework can be viewed as a first step towards a unifying computational model capturing reward processing abnormalities across multiple mental conditions.

B. Finance

In recent years, sequential portfolio selection has been a focus of increasing interest at the intersection of the machine learning and quantitative finance. The trade-off between exploration and exploitation, with the goal of maximizing cumulative reward, is a natural formulation of the portfolio choice problems. In [14], the authors proposed a bandit algorithm for making online portfolio choices via exploiting correlations among multiple arms.

By constructing orthogonal portfolios from multiple assets and integrating their approach with the upper-confidence-bound bandit framework, the authors derive the optimal portfolio strategy representing a combination of passive and active investments according to a risk-adjusted reward function. In [15], the authors incorporate risk-awareness into the classic multi-armed bandit setting and introduce a novel algorithm for portfolio construction. Through filtering assets based on the topological structure of financial market and combining the optimal multi-armed bandit policy with the minimization of a coherent risk measure, they achieve a balance between risk and return.

C. Dynamic Pricing

Online retailer companies are often faced with the dynamic pricing problem: the company must decide on real-time prices for each of its multiple products. The company can run price experiments (make frequent price changes) to learn about demand and maximize long-run profits. The authors in [16] propose a dynamic price experimentation policy, where the company has only incomplete demand information. For this general setting, authors derive a pricing algorithm that balances earning an immediate profit vs. learning for future profits. The approach combines multi-armed bandit with partial identification of consumer demand from economic theory. Similar to [16], authors in [17] consider high-dimensional dynamic multi-product pricing with an evolving low-dimensional linear demand model. They show that the revenue maximization problem reduces to an online bandit convex optimization with side information given by the observed demands. The approach applies a bandit convex optimization algorithm in a projected low-dimensional space spanned by the latent product features, while simultaneously learning this span via online singular value decomposition of a carefully-crafted matrix containing the observed demands.

D. Recommender Systems

Recommender systems are frequently used in various applications to predict user preferences. However, they also face the exploration-exploitation dilemma when making a recommendation, since they need to exploit their knowledge about the previously chosen items the user is interested in, while also exploring new items the user may like. The authors in [18] approach this challenge by using the multi-armed bandit setting, especially for large-scale recommender systems that have a really large or an infinite number of items. They propose two large-scale bandit approaches in situations when no prior information is available. Continuous exploration in their approaches can address the cold start problem in recommender systems. In context-aware recommender systems, most existing approaches focus on recommending relevant items to users, taking into account contextual information, such as time, location, or social aspects. However, none of those approaches has considered the problem of user's content evolution. In [19], the authors introduce an algorithm that takes this dynamics into account. It is based on dynamic exploration/exploitation and can adaptively balance the two aspects, deciding which situation is most relevant for exploration or exploitation. In this sense, [20] propose to study the ”freshness” of the user's content through the bandit problem. They introduce the *Freshness-Aware Thompson Sampling* algorithm for recommendation of fresh documents.

E. Influence Maximization

The authors in [21] consider influence maximization (IM) in social networks, which is the problem of maximizing the number of users that become aware of a product by selecting a set of “seed” users to expose the product to. They propose a novel parametrization that not only makes the framework agnostic to the underlying diffusion model, but also statistically efficient to learn from data. They give a corresponding monotone, submodular surrogate function, and show that it is a good approximation to the original IM objective. They also consider the case of a new marketer looking to exploit an existing social network, while simultaneously learning the factors governing information propagation. For this, they develop a LinUCB-based bandit algorithm. The authors in [22] also study the online influence maximization problem in social networks but under the independent cascade model. Specifically, they try to learn the set of “best seeds or influencers” in a social network online while repeatedly interacting with it. They address the challenges of combinatorial action space, since the number of feasible influencer sets grows exponentially with the maximum number of influencers, and limited feedback, since only the influenced portion of the network is observed. They propose and analyze IMLinUCB, a computationally efficient UCB-based algorithm.

F. Information Retrieval

The authors in [23] argue that Information Retrieval iterative selection process can be naturally modeled as a contextual bandit problem. The multi-armed bandit model leads to highly effective methods for document adjudication. Under this bandit allocation framework, they propose seven new document adjudication methods, of which five are stationary methods and two are non-stationary methods. This comparative study includes existing methods designed for pooling-based evaluation and existing methods designed for metasearch. In mobile information retrieval, the authors in [24] introduce an algorithm that tackles this dilemma in Context-Based Information Retrieval (CBIR) area. It is based on dynamic exploration/exploitation and it can adaptively balance the two aspects by deciding which user's situation is most relevant for exploration or exploitation. Within a deliberately designed online framework, they conduct evaluations with mobile users.

G. Dialogue Systems

Dialogue response selection. Dialogue response selection is an important step towards natural response generation in conversational agents. The existing work on conversational models mainly focuses on offline supervised learning using a large set of context-response pairs. In [25], the authors focus on online learning of response selection in dialog systems. They propose a contextual multi-armed bandit model with a nonlinear reward function that uses distributed representation of text for online response selection. A bidirectional LSTM is used to produce the distributed representations of dialog context and responses, which serve as the input to a contextual bandit. They propose a customized Thompson sampling method that is applied to a polynomial feature space in approximating the reward.

Pro-activity dialogue systems. An objective of pro-activity in dialogue systems is to enhance the usability of conversational agents by enabling them to initiate conversations on their own. While dialogue systems have become increasingly popular recently, current task-oriented dialogue systems are mainly reactive, as human users tend to initiate conversations. The authors of [26] propose to introduce the paradigm of contextual bandits as framework for proactive dialog systems. Contextual bandits have been the model of choice for the problem of reward maximization with partial feedback since they fit well to the task description, they also explore the notion of memory into this paradigm, where they propose two differentiable memory models that act as parts of the parametric reward estimation function. The first one, Convolutional Selective Memory Networks, uses a selection of past interactions as part of the decision support. The second model, called Contextual Attentive Memory Network, implements a differentiable attention mechanism over the past interactions of the agent. The goal is to generalize the classic model of contextual bandits to settings where temporal information needs to be incorporated and leveraged in a learnable manner.

Multi-domain dialogue systems. Building multi-domain dialogue agents is a challenging task and an open problem in modern AI. Within the domain of dialogue, the ability to orchestrate multiple independently trained dialog agents, or skills, to create a unified system is of particular significance. In [27], the authors study the task of online posterior dialogue orchestration, where they define posterior orchestration as the task of selecting a subset of skills which most appropriately answers a user input using features extracted from both the user input and the individual skills. To account for the varied costs associated with extracting skill features, they consider online posterior orchestration under a skill execution budget. This setting is formalized as Context-Attentive Bandit with Observations, a variant of context-attentive bandits, and then evaluate it on simulated non-conversational and proprietary conversational datasets.

H. Anomaly Detection

The problem of anomaly detection on attributed networks finds nodes whose behaviors deviate significantly from the majority of nodes. The authors in [3] investigate the problem of anomaly detection in an interactive setting by allowing the system to proactively communicate with the human expert in making a limited number of queries about ground truth anomalies. Their objective is to maximize the true anomalies presented to the human expert after a given budget is used up. Along with this line, they formulate the problem through the principled multi-armed bandit framework and develop a novel collaborative contextual bandit algorithm, that explicitly models the nodal attributes and node dependencies seamlessly in a joint framework, and handles the exploration-exploitation dilemma when querying anomalies of different types. Credit card transactions predicted to be fraudulent by automated detection systems are typically handed over to human experts for verification. To limit costs, it is standard practice to select only the most suspicious transactions for investigation. The authors in [28] claim that a trade-off between exploration and exploitation is imperative in enabling adaptation to changes in behavior. Exploration consists of the selection and investigation of transactions with the purpose of improving predictive models, and exploitation consists of investigating transactions detected to be suspicious. Modeling the detection of fraudulent transactions as rewarding, they use an incremental regression tree learner to create clusters of transactions with similar expected rewards. This enables the use of a *contextual* multi-armed bandit (CMAB) algorithm to provide the exploration/exploitation trade-off.

I. Telecommunication

In [29], a multi-armed bandit model was used to describe the problem of best wireless network selection by a multi-Radio Access Technology (multi-RAT) device, with the goal of maximizing the quality perceived by the final user. The proposed model extends the classical MAB model in a twofold manner. First, it foresees two different actions: to measure and to use; second, it allows actions to span multiple time steps. Two new algorithms designed to take advantage of the higher flexibility provided by the muMAB model were also introduced. The first one, referred to as measure-use-UCB1 is derived from the UCB1 algorithm, while the second one, referred to as Measure with Logarithmic Interval, is appositely designed for the new model so to take advantage of the new measure action, while aggressively using the best arm. The authors in [30] demonstrate the possibility to optimize the performance of the Long Range Wide Area Network technology. The authors suggest that nodes use multi-armed bandit algorithms, to select the communication parameters (spreading factor and emission power). Evaluations show that such learning methods allow to manage the trade-off between energy consumption and packet loss much better than an Adaptive Data Rate algorithm adapting spreading factors and transmission powers on the basis of Signal to Interference and Noise Ratio values.

J. Bandit in Real-Life Applications: Summary and Future Directions

TABLE I
BANDIT FOR REAL LIFE APPLICATION

	MAB	Non-stationary MAB	CMAB	Non-stationary CMAB
Healthcare	√		√	
Finance	√			
Dynamic pricing		√		
Recommendr system	√	√	√	√
Maximization	√			
Dialogue system			√	
Telecomunication	√			
Anomaly detection	√			

Table I provides a summary of bandit problem formulations used in various domain-specific applications. The choice of bandit model is often domain-specific. For example, it is evident that non-stationary bandit was not used in healthcare applications, as significant changes are not expected to the process of making the treatment decisions, i.e. no transition in the state of the the patient; such transitions, if they occurred, would be better modeled using reinforcement learning rather than non-stationary bandit. There are clearly other domains where the non-stationary bandit is a more appropriate setting, but it looks like this setting was not yet been significantly investigated in healthcare domains. For example, anomaly detection, is a domain where non-stationary contextual bandit could be used, since in this setting the anomaly could be adversarial, which means that any bandit applied to this setting should have some kind of drift condition, in-order to adapt to new types of attacks. Another observation is that none of the existing work tried to develop an algorithm that could solve these different tasks at the same time, or apply the knowledge obtained in one domain to another domain, thus opening a direction of research on *multitask* and *transfer learning* in bandit setting. Furthermore, given an online nature of bandit problem, *continuous*, or *lifelong learning* would be a natural next step, adapting the model learned in the previous tasks to the new one, while still remembering how to perform earlier task, thus avoiding the problem of ”catastrophic forgetting”.

III. BANDIT FOR BETTER MACHINE LEARNING

In this section we are describing how bandit algorithms could be used to improve other algorithms, e.g. various machine-learning techniques.

A. Algorithm Selection

Algorithm selection is typically based on models of algorithm performance, learned during a separate offline training sequence, which can be prohibitively expensive. In recent work, they adopted an online approach, in which a performance model is iteratively updated and used to guide selection on a sequence of problem instances. The resulting exploration-exploitation trade-off was represented as a bandit problem with expert advice, using an existing solver for this game, but this required using an arbitrary bound on algorithm runtimes, thus invalidating the optimal regret of the solver. In [31], a simpler framework was proposed for representing algorithm selection as a bandit problem, using partial information and an unknown bound on losses.

B. Hyperparameter Optimization

Performance of machine learning algorithms depends critically on identifying a good set of hyperparameters. While recent approaches use Bayesian optimization to adaptively select optimal hyperparameter configurations, they rather focus on speeding up random search through adaptive resource allocation and early-stopping. [32] formulated hyperparameter optimization as a pure-exploration non-stochastic infinite-armed bandit problem where a predefined resources, such as iterations, data samples, or features are allocated to randomly sampled configurations. This work introduced a novel algorithm, Hyperband, for this framework and analyze its theoretical properties, providing several desirable guarantees. Furthermore, Hyperband wascmpared with popular Bayesian optimization methods on a suite of hyperparameter optimization problems; it was observed that Hyperband can provide more than an order-of-magnitude speedup over its competitors on a variety of deep-learning and kernel-based learning problems.

C. Feature Selection

In a classical online *supervised learning* the true label of a sample is always revealed to the classifier, unlike in a bandit setting were any wrong classification resuls into zero reward, and only the single correct classification yields reward 1. The authors of [33] investigate the problem of Online Feature Selection, where the aim is to make accurate predictions using only a small number of active features using epsilon greedy algorithm. The authors of [34] tackle the online feature selection problem by addressing the combinatorial optimization problem in the stochastic bandit setting with bandit feedback, utilizing the Thompson Sampling algorithm.

D. Bandit for Active Learning

Labelling all training examples in supervised classification setting can be costly. Active learning strategies solve this problem by selecting the most useful unlabelled examples to obtain the label for, and 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 input space. In [35], a novel active learning strategy manages this compromise by modelling the active learning problem as a contextual bandit problem. they 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. The authors of [36] also propose a multi-armed bandit inspired, pool-based active learning algorithm for the problem of binary classification. They utilize ideas such as lower confidence bounds, and self-concordant regularization from the multi-armed bandit literature to design their proposed algorithm. In each round, the proposed algorithm assigns a sampling distribution on the pool, samples one point from this distribution, and queries the oracle for the label of this sampled point.

E. Clustering

[37] considers collaborative clustering, which is machine-learning paradigm concerned with the unsupervised analysis of complex multi-view data using several algorithms working together. Well-known applications of collaborative clustering include multiview clustering and distributed data clustering, where several algorithms exchange information in order to mutually improve each others. One of the key issue with multi-view and collaborative clustering is to assess which collaborations are going to be beneficial or detrimental. Many solutions have been proposed for this problem, and all of them conclude that, unless two models are very close, it is difficult to predict in advance the result of a collaboration. To address this problem, the authors of [37] propose a collaborative peer to peer clustering algorithm based on the principle of non stochastic multi-arm bandits to assess in real time which algorithms or views can bring useful information.

F. Reinforcement learning

Autonomous cyber-physical systems play a large role in our lives. To ensure that agents behave in ways aligned with the values of the societies in which they operate, we must develop techniques that allow these agents to not only maximize their reward in an environment, but also to learn and follow the implicit constraints assumed by the society. In [38], the authors study a setting where an agent can observe traces of behavior of members of the society but has no access to the explicit set of constraints that give rise to the observed behavior. Instead, inverse reinforcement learning is used to learn such constraints, that are then combined with a possibly orthogonal value function through the use of a contextual bandit-based orchestrator that picks a contextually-appropriate choice between the two policies (constraint-based and environment reward-based) when taking actions. The contextual bandit orchestrator allows the agent to mix policies in novel ways, taking the best actions from either a reward maximizing or constrained policy. The [39] tackles the problem of online RL algorithm selection. A meta-algorithm is given for input a portfolio constituted of several off-policy RL algorithms. It then determines at the beginning of each new trajectory, which algorithm in the portfolio is in control of the behaviour during the next trajectory, in order to maximise the return. A novel meta-algorithm, called Epochal Stochastic Bandit Algorithm Selection. Its principle is to freeze the policy

TABLE II
BANDIT IN MACHINE LEARNING

	MAB	Non-stationary MAB	CMAB	Non-stationary CMAB
Algorithm Slection		√		
Parameter Optimization	√			
Features Selection	√	√		
Active Learning	√		√	
Clustering	√			
Reinforcement learning	√	√	√	

updates at each epoch, and to leave a rebooted stochastic bandit in charge of the algorithm selection.

G. Bandit for Machine Learning: Summary and Future Directions

Table II summarizes the types of bandit problems used to solve the machine-learning problems mentioned above. We see, for example, that contextual bandit was not used in feature selection or hyperparameter optimization. This observation could point into a direction for future work, where side information could be employed in feature selection. Also, non-stationary bandit was rarely considered in these problem settings, which is also suggesting possible extensions of current work. For instance, the non-stationary contextual bandit could be useful in the non-stationary feature selection setting, where finding the right features is time-dependent and context-dependent when the environment keeps changing. Our main observation is also that each technique is solving just one machine learning problem at a time; thus, the question is whether a bandit setting and algoritms can be developed to solve multiple machine learning problems simultaneously, and whether transfer and continual learning can be achieved in this setting. One solution could be to model all these problems in a combinatorial bandit framework, where the bandit algorithm would find the optimal solution for each problem at each iteration; thus, combinatorial bandit could be further used as a tool for advancing automated machine learning.

IV. CONCLUSIONS

In this article, we reviewed some of the most notable recent work on applications of multi-armed bandit and contextual bandit, both in real-life domains and in automated machine learning. We summarized, in an organized way (Tables 1 and 2), various existing applications, by types of bandit settings used, and discussed the advantages of using bandit techniques in each domain. We briefly outlines of several important open problems and promising future extensions.

In summary, the bandit framework, including both multi-arm and contextual bandit, is currently very active and promising research areas, and there are multiple novel techniques and applications emerging each year. We hope our survey can help the reader better understand some key aspects of this exciting field and get a better perspective on its notable advancements and future promises.

REFERENCES

[1] A. Durand, C. Achilleos, D. Iacovides, K. Strati, G. D. Mitsis, and J. Pineau, "Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis," in *Machine Learning for Healthcare Conference*, pp. 67–82, 2018.

[2] J. Mary, R. Gaudel, and P. Preux, "Bandits and recommender systems," in *Machine Learning, Optimization, and Big Data - First International Workshop, MOD 2015*, pp. 325–336, 2015.

[3] K. Ding, J. Li, and H. Liu, "Interactive anomaly detection on attributed networks," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, (New York, NY, USA), pp. 357–365, ACM, 2019.

[4] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[5] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, no. 2-3, pp. 235–256, 2002.

[6] D. Bouneffouf and R. Féraud, "Multi-armed bandit problem with known trend," *Neurocomputing*, vol. 205, pp. 16–21, 2016.

[7] S. Agrawal and N. Goyal, "Analysis of thompson sampling for the multi-armed bandit problem," in *COLT 2012 - The 25th Annual Conference on Learning Theory, June 25-27, 2012, Edinburgh, Scotland*, pp. 39.1–39.26, 2012.

[8] J. Langford and T. Zhang, "The epoch-greedy algorithm for multi-armed bandits with side information," in *Advances in neural information processing systems*, pp. 817–824, 2008.

[9] S. Agrawal and N. Goyal, "Thompson sampling for contextual bandits with linear payoffs," in *ICML (3)*, pp. 127–135, 2013.

[10] L. Li, W. Chu, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation," *CoRR*, 2010.

[11] R. Allesiardo, R. Féraud, and D. Bouneffouf, "A neural networks committee for the contextual bandit problem," in *Neural Information Processing - 21st International Conference, ICONIP 2014, Kuching, Malaysia, November 3-6, 2014. Proceedings, Part I*, pp. 374–381, 2014.

[12] H. Bastani and M. Bayati, "Online decision-making with high-dimensional covariates," *Available at SSRN 2661896*, 2015.

[13] D. Bouneffouf, I. Rish, and G. A. Cecchi, "Bandit models of human behavior: Reward processing in mental disorders," in *AGI*, pp. 237–248, Springer, 2017.

[14] W. Shen, J. Wang, Y.-G. Jiang, and H. Zha, "Portfolio choices with orthogonal bandit learning," in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[15] X. Huo and F. Fu, "Risk-aware multi-armed bandit problem with application to portfolio selection," *Royal Society open science*, vol. 4, no. 11, p. 171377, 2017.

[16] K. Misra, E. M. Schwartz, and J. Abernethy, "Dynamic online pricing with incomplete information using multi-armed bandit experiments," 2018.

[17] J. Mueller, V. Syrgkanis, and M. Taddy, "Low-rank bandit methods for high-dimensional dynamic pricing," *arXiv preprint arXiv:1801.10242*, 2018.

[18] Q. Zhou, X. Zhang, J. Xu, and B. Liang, “Large-scale bandit approaches for recommender systems,” in *International Conference on Neural Information Processing*, pp. 811–821, Springer, 2017.

[19] D. Bouneffouf, A. Bouzeghoub, and A. L. Gançarski, “A contextual-bandit algorithm for mobile context-aware recommender system,” in *International Conference on Neural Information Processing*, pp. 324–331, Springer, 2012.

[20] D. Bouneffouf, “Freshness-aware thompson sampling,” in *International Conference on Neural Information Processing*, pp. 373–380, Springer, 2014.

[21] S. Vaswani, B. Kveton, Z. Wen, M. Ghavamzadeh, L. V. Lakshmanan, and M. Schmidt, “Model-independent online learning for influence maximization,” in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3530–3539, JMLR. org, 2017.

[22] Z. Wen, B. Kveton, M. Valko, and S. Vaswani, “Online influence maximization under independent cascade model with semi-bandit feedback,” in *Advances in neural information processing systems*, pp. 3022–3032, 2017.

[23] D. E. Losada, J. Parapar, and A. Barreiro, “Multi-armed bandits for adjudicating documents in pooling-based evaluation of information retrieval systems,” *Information Processing & Management*, vol. 53, no. 5, pp. 1005–1025, 2017.

[24] D. Bouneffouf, A. Bouzeghoub, and A. L. Gançarski, “Contextual bandits for context-based information retrieval,” in *International Conference on Neural Information Processing*, pp. 35–42, Springer, 2013.

[25] B. Liu, T. Yu, I. Lane, and O. J. Mengshoel, “Customized nonlinear bandits for online response selection in neural conversation models,” in *AAAI, 2018*, pp. 5245–5252, 2018.

[26] T. Silander *et al.*, “Contextual memory bandit for pro-active dialog engagement,” 2018.

[27] S. Upadhyay, M. Agarwal, D. Bounneffouf, and Y. Khazaeni, “A bandit approach to posterior dialog orchestration under a budget,” 2018.

[28] D. J. Soemers, T. Brys, K. Driessens, M. H. Winands, and A. Nowé, “Adapting to concept drift in credit card transaction data streams using contextual bandits and decision trees,” in *AAAI*, 2018.

[29] S. Boldrini, L. De Nardis, G. Caso, M. Le, J. Fiorina, and M.-G. Di Benedetto, “mumab: A multi-armed bandit model for wireless network selection,” *Algorithms*, vol. 11, no. 2, p. 13, 2018.

[30] R. Kerkouche, R. Alami, R. Féraud, N. Varsier, and P. Maillé, “Node-based optimization of lora transmissions with multi-armed bandit algorithms,” in *ICT 2018, Saint Malo, France, June 26-28, 2018*, pp. 521–526, 2018.

[31] M. Gagliolo and J. Schmidhuber, “Algorithm selection as a bandit problem with unbounded losses,” in *Learning and Intelligent Optimization, 4th International Conference, LION 4, Venice, Italy, January 18-22, 2010. Selected Papers*, pp. 82–96, 2010.

[32] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, “Hyperband: A novel bandit-based approach to hyperparameter optimization,” *arXiv preprint arXiv:1603.06560*, 2016.

[33] J. Wang, P. Zhao, S. C. Hoi, and R. Jin, “Online feature selection and its applications,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 3, pp. 698–710, 2014.

[34] D. Bouneffouf, I. Rish, G. A. Cecchi, and R. Féraud, “Context attentive bandits: Contextual bandit with restricted context,” in *IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pp. 1468–1475, 2017.

[35] D. Bouneffouf, R. Laroche, T. Urvoy, R. Féraud, and R. Allesiardo, “Contextual bandit for active learning: Active thompson sampling,” in *International Conference on Neural Information Processing*, pp. 405–412, Springer, 2014.

[36] R. Ganti and A. G. Gray, “Building bridges: Viewing active learning from the multi-armed bandit lens,” *arXiv preprint arXiv:1309.6830*, 2013.

[37] J. Sublime and S. Lefebvre, “Collaborative clustering through constrained networks using bandit optimization,” in *2018 International Joint Conference on Neural Networks, IJCNN 2018, Rio de Janeiro, Brazil, July 8-13, 2018*, pp. 1–8, 2018.

[38] R. Noothigattu, D. Bouneffouf, N. Mattei, R. Chandra, P. Madan, K. Varshney, M. Campbell, M. Singh, and F. Rossi, “Interpretable multi-objective reinforcement learning through policy orchestration,” *arXiv preprint arXiv:1809.08343*, 2018.

[39] R. Laroche and R. Féraud, “Algorithm selection of off-policy reinforcement learning algorithm,” *CoRR*, vol. abs/1701.08810, 2017.