Expert with Clustering: Hierarchical Online Preference Learning Framework

Tianyue Zhou ZHOUTY1@SHANGHAITECH.EDU.CN
ShanghaiTech University

Jung-Hoon Cho JHOONCHO@MIT.EDU
Massachusetts Institute of Technology

Babak Rahimi Ardabili BRAHIMIA@CHARLOTTE.EDU
Hamed Tabkhi HTABKHIV@CHARLOTTE.EDU
University of North Carolina at Charlotte

Cathy Wu CATHYWU@MIT.EDU
Massachusetts Institute of Technology

Editors: A. Abate, M. Cannon, K. Margellos, A. Papachristodoulou

Abstract

Emerging mobility systems are increasingly capable of recommending options to users, to guide them towards personalized yet sustainable system outcomes. Even more so than the typical recommendation system, it is crucial to minimize regret, because 1) the mobility options directly affect the lives of the users, and 2) the system sustainability relies on sufficient user participation. In this study, we thus consider accelerating user preference learning by exploiting a low-dimensional space that captures the mobility preferences of users within a population. We therefore introduce a hierarchical contextual bandit framework named Expert with Clustering (EWC), which integrates clustering techniques and prediction with expert advice. EWC efficiently utilizes hierarchical user information and incorporates a Loss-guided Distance metric. This metric is instrumental in generating more representative cluster centroids, thereby enhancing the performance of recommendation systems. In a recommendation scenario with N users, T rounds per user, and K options, our algorithm achieves a regret bound of $O(N\sqrt{T \log K} + NT)$. This bound consists of two parts: the regret from the Hedge algorithm, and the average loss from clustering. To the best of the authors knowledge, this is the first work to analyze the regret of an integrated expert algorithm with k-Means clustering. This regret bound underscores the theoretical and experimental efficacy of EWC. Experimental results highlight that EWC can substantially reduce regret by 27.57% compared to the LinUCB baseline. Our work offers a data-efficient approach to capturing both individual and collective behaviors, making it highly applicable to contexts with latent hierarchical structures. We expect the algorithm to be applicable to other settings with layered nuances of user preferences and information.

Keywords: Online preference learning, Contextual bandit, Clustering, Eco-driving recommendation, Expert advice

1. Introduction

Emerging mobility systems are increasingly pivotal in designing efficient and sustainable transit networks. These systems with advanced technologies have the potential to revolutionize how we navigate urban environments, thereby enhancing the overall efficiency of transportation systems. Recent work aims to minimize emissions through eco-driving recommendation systems, thus contributing to environmental sustainability (Tu et al. (2022); Chada et al. (2023)).

The challenge lies in the complex nature of drivers' preferences, which are shaped by various factors such as personal schedules, environmental concerns, and the unpredictability of human

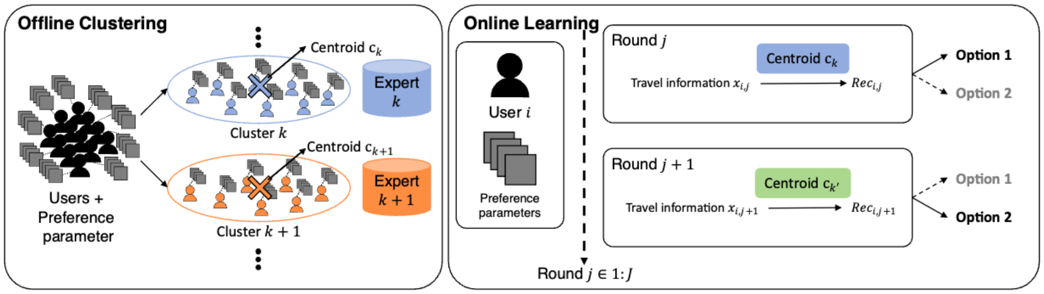

Figure 1: Illustrative figure for Expert with Clustering algorithm.

behavior. To tackle this, we advocate for the contextual bandit algorithm, a framework adept at learning from and adapting to the driver's unique context. This algorithm is poised to accurately capture the subtle preferences of drivers, offering nuanced insights into their decision-making processes in ever-changing environments.

Our problem deviates from a classical contextual bandit scenario. It resembles supervised online learning, as the driver's choice is known after the system provides recommendations. However, this choice may be influenced by the recommendation itself, leading to varying observed rewards. This variation closely mirrors the structure of a bandit problem, where actions influence observed outcomes. Thus, our problem can be considered a specialized variant of the bandit problem, incorporating elements of both supervised learning and adaptive decision-making typical of bandit scenarios. Our research focuses on refining online learning algorithms tailored to the contextual bandit framework, enhancing their ability to discern traveler preferences and predict responses to routing and transit mode recommendations.

In this paper, we propose the Expert with Clustering (EWC) framework, a novel approach that synergizes clustering and prediction with expert advice. The fundamental concept involves using clustering to discern hierarchical information among users, with each cluster acting as an 'expert' representing common user preferences. This approach transforms the online preference learning problem into one of prediction with expert advice. For each user, an expert is selected to approximate their preferences, enabling accurate recommendations.

1.1. Related Works

Preference learning for drivers. The field of preference learning for drivers has focused on adaptive models that cater to individual driving behaviors and decision-making processes. Jalota et al. (2022) contribute to this domain with the online learning approach, adjusting traffic tolls based on aggregate flows to influence drivers towards more efficient routes, thus minimizing the need for detailed personal travel information. This method resonates with the utilitarian perspective of Chorus et al. (2009), delving into how drivers' preference with advice is shaped by their personal preferences and perceptions of travel time uncertainty. Sadigh et al. (2017) explore the learning of human preferences from comparative choices between trajectories, eschewing the need for direct reward signals. Collectively, these works highlight a trend toward machine learning that is not just reactive but anticipatory and adaptable to the nuanced spectrum of human driving preferences.

Contextual bandits. The contextual bandits framework has emerged as an efficient approach for recommendation systems. Originally introduced by Auer (2002), this framework delved into the utilization of confidence bounds within the exploration-exploitation trade-off, specifically focusing on contextual bandits with linear value functions. Building upon this foundation, Li et al. (2010) expanded the application of this concept to personalized news recommendations and proposed LinUCB algorithm which has since become a benchmark in the field.

Expert algorithm. The prediction with expert advice is a fundamental problem in online learning. The Hedge algorithm, also recognized as the Weighted Majority Algorithm and initially introduced by Littlestone and Warmuth (1994), presents an efficient approach to addressing this challenge. The insights offered by the Hedge algorithm have significantly informed our development of the Expert with Clustering (EWC) framework. In terms of theoretical performance, Freund and Schapire (1997) has established that the Hedge algorithm's regret bound is $O(\sqrt{T \log K})$. This regret bound provides a foundation for theoretical analysing the EWC framework.

Contextual bandits with clustering. K-Means clustering, a classic unsupervised learning algorithm, was first introduced by Lloyd (1982). With the evolution of both contextual bandits and clustering techniques, the concept of clustering users within contextual bandits was proposed by Gentile et al. (2014), utilizing a graph to represent user similarities. To enhance the utilization of similarity information, Li et al. (2016) combined collaborative filtering with contextual bandits to cluster both users and items. Furthermore, Gentile et al. (2017) introduced context-aware clustering, where each item could cluster users. Existing works in contextual bandits typically overlook user choice due to their classical framework. In contrast, our unique problem structure incorporates user choice, offering insights into the comparative utility of different options. This distinct approach allows for accelerating the preference learning within limited data.

Objective function in clustering. The classic K-Means algorithm focuses on minimizing the within-cluster sum of squares, which may not fit all clustering needs. For instance, K-Means assumes that clusters are spherical and roughly equal in terms of size, which may not always be the case in real-world data. Addressing these limitations, researchers in fields such as federated learning and system identification have devised bespoke objective functions to enhance clustering methodologies. For example, Ghosh et al. (2020) proposed a new framework which alternates between identifying clusters and minimizing a customized loss function for federated learning. Similarly, Sattler et al. (2021) uses the geometric properties of loss surfaces to group clients in federated multitask learning. Additionally, Toso et al. (2023) applies a clustering algorithm to derive linear system models, aiming to minimize the aggregate cost function within each cluster. Building on these developments, we propose a loss-guided distance metric tailored for online preference learning.

1.2. Contributions

The primary contributions of this work are outlined as follows:

1. We introduce the novel hierarchical contextual bandit framework, Expert with Clustering (EWC), which integrates clustering and prediction with expert advice to address the online preference learning problem. This framework effectively utilizes hierarchical user information, enabling rapid and accurate learning of user preferences.
2. We propose a distance metric, Loss-guided Distance, tailored for the online preference learning problem, which enhances the representativeness of centroids. This advancement improves the performance of the EWC framework, demonstrating its practical effectiveness.
3. We establish the regret bound of EWC as a sum of two components: the regret from the Hedge algorithm and the bias introduced by representing users with centroids, indicating superior theoretical performance in the short term and enhanced overall experimental performance compared to the LinUCB algorithm (Li et al. (2010)).

2. Problem Formulation

Consider the scenario where a social planner is tasked with recommending mobility options $\mathcal{R}$, where $A := |\mathcal{R}|$, to a population of drivers. Each mobility option is parameterized by a travel information vector $x_{i,t} \in \mathbb{R}^d$ specifying relevant travel metrics, where i and t indicates the index of driver and decision round. which specifies relevant travel metrics. For simplicity, we consider two mobility options ($A = 2$), each with two relevant travel metrics ($d = 2$), although the framework extends gracefully to more options and metrics. Thus, at each decision point for a user, the social planner faces a choice between two route options: route 1, the standard route with regular travel time and emissions, and route 2, an eco-friendly alternative that, while offering reduced emissions, comes with an increased travel time. Intuitively, in this simplified example, the social planner seeks to quickly identify users who prefer travel time or environmental impact while ensuring user participation, in order to best achieve the system sustainability. In the future, considerations of multiple system objectives and incentives to shape user choices can be included.

For each decision round t, the user i compare two routes in terms of their relative travel time and emissions. Let's denote the travel time and emissions for route 2 relative to route 1 as $\tau_{i,t}$ and $e_{i,t}$, respectively. For example, $[\tau_{i,t}, e_{i,t}] = [1.2, 0.9]$ means 120% of travel time and 90% of emission. Travel information vector for this decision round, $x_{i,t}$, is defined with two components for two routes: $x_{i,t}(1) = [1, 1]$ and $x_{i,t}(2) = [\tau_{i,t}, e_{i,t}]$. Based on this information, we issue a recommendation $Rec_{i,t} \in \{1, 2\}$, whereupon we receive feedback in the form of the user's choice $y_{i,t} \in \{1, 2\}$. The objective of our system is to minimize the total regret: $\sum_{i=1}^{N} \sum_{t=1}^{T} |Rec_{i,t} - y_{i,t}|^2$, where N represent the number of users and T denote the total number of decision rounds.

3. Expert with Clustering (EWC)

3.1. General Framework

We introduce the Expert with Clustering (EWC) algorithm, a novel hierarchical contextual bandit approach. EWC transforms an online preference learning problem into an expert problem and utilizes the Hedge algorithm to identify the most effective expert.

Prediction with expert advice is a classic online learning problem introduced by Littlestone and Warmuth (1994). Consider a scenario where a decision-maker has access to the advice of K experts. At each decision round t, advice from these K experts is available, and the decision maker selects an expert based on a probability distribution $\mathbf{p}_t$ and follows his advice. Subsequently, the decision maker observes the loss of each expert, denoted as $\mathbf{l}_t \in [0, 1]^K$. The primary goal is to identify the best expert in hindsight, which essentially translates to minimizing the regret: $\sum_{t=1}^{T} (< \mathbf{p}_t, \mathbf{l}_t > -\mathbf{l}_t(k^*))$, where k^* is the best expert throughout the time.

We cast the online preference learning problem into the framework of prediction with expert advice in the following way. Assume that each user has a fixed but unknown preference parameter $\boldsymbol{\theta}_i \in \mathbb{R}^d$. Given $\boldsymbol{\theta}_i$, we can make predictions using a known function $\hat{y}(\boldsymbol{\theta}_i, x_{i,t})$. The EWC algorithm operates under the assumption of a cluster structure within the users' preference parameters $\{\boldsymbol{\theta}_i\}_{i\in[N]}$. Utilizing a set of offline training data $\mathcal{D} = \{\{x_{i,t}\}_{i\in[N'],t\in[T']}, \{y_{i,t}\}_{i\in[N'],t\in[T']}\}$ where N' and T' are number of users and decision rounds in training data, we initially employ a learning framework (such as SVM or nonlinear regression) to determine each user's $\boldsymbol{\theta}_i$. Despite differences between training and testing data, both are sampled from the same distribution. This allows for an approximate determination of $\boldsymbol{\theta}_i$, providing insights into the hierarchical structure among users, albeit with some degree of approximation. Subsequently, a clustering method is applied to identify centroids $\{\mathbf{c}_k\}_{k\in[K]}$.

Each centroid is considered as an expert. Using the Hedge algorithm, we initialize their weights and, at every online decision round, select an expert $E_{i,t} \in [K]$. An expert $E_{i,t}$ provides advice suggesting that a user's preference parameters closely resemble the centroid $\mathbf{c}_{E_{i,t}}$. Consequently, we use this centroid to estimate the user's preferences. The recommendation $Rec_{i,t} = \hat{y}(\mathbf{c}_{E_{i,t}}, x_{i,t})$ is then formulated. Upon receiving the user's chosen option $y_{i,t}$, we calculate the loss for each expert and update the weights in Hedge based on this loss. The loss for each expert k is determined by a known loss function $\mathbf{l}_{i,t}(k) = l(\hat{y}(\mathbf{c}_k, x_{i,t}), y_{i,t}) \in \mathbb{R}$, e.g., $l(\hat{y}(\mathbf{c}_k, x_{i,t}), y_{i,t}) = \mathbb{1}_{\hat{y}(\mathbf{c}_k, x_{i,t}) \neq y_{i,t}}$. The details of this process are encapsulated in Algorithm 1.

Algorithm 1 Expert With Cluster

```
Require: Number of clusters $K$, offline training data $\mathcal{D}$, learning rate $\eta$
  Train with data $\mathcal{D}$, receive $\{\boldsymbol{\theta}_i\}_{i \in [N']}$
  Apply clustering on $\{\boldsymbol{\theta}_i\}_{i \in [N']}$, receive centroids $\{\mathbf{c}_k\}_{k \in [K]}$
  Initialize weight $\mathbf{p}_{i,1}(k) \leftarrow \frac{1}{K}$ for all $i \in [N], k \in [K]$
  for $t = 1, \ldots, T$ do
    for $i = 1, \ldots, N$ do
      Receive $x_{i,t}$
      Sample $E_{i,t} \sim \mathbf{p}_{i,t}$, submit $Rec_{i,t} = \hat{y}(\mathbf{c}_{E_{i,t}}, x_{i,t})$,
      Receive $y_{i,t}$, compute loss $\mathbf{l}_{i,t}(k) = l(\hat{y}(\mathbf{c}_k, x_{i,t}), y_{i,t})$ for all $k \in [K]$
      $\mathbf{p}_{i,t+1}(k) \leftarrow \frac{\mathbf{p}_{i,t}(k) e^{-\eta \mathbf{l}_{i,t}(k)}}{\sum_{k' \in [K]} \mathbf{p}_{i,t}(k') e^{-\eta \mathbf{l}_{i,t}(k')}}$ for all $k \in [K]$
    end for
  end for
```

3.2. EWC for Online Preference Learning

We implement the EWC algorithm for online preference learning in a driving context. For each user i, a linear decision boundary is posited, characterized by parameters $\boldsymbol{\theta}_i = [b_i, s_i, o_i]$. Here, b_i and s_i denote the bias and slope of the line, and o_i denotes the orientation of the decision boundary that differentiates the affiliation of user's choice. $\boldsymbol{\theta}_i$ classifies the data points $[\tau_{i,t}, e_{i,t}]$ into two categories: opting for the regular route ($y_{i,t} = 1$) or the eco-friendly route ($y_{i,t} = 2$). This example illustrates the meaning of $\boldsymbol{\theta}_i$. Assume that the eco-friendly route is described by $x_{i,t}(2) = [\tau_{i,t}, e_{i,t}] = [1.1, 0.85]$. Consider a user with preference parameters $\boldsymbol{\theta}_i = [b_i, s_i, o_i] = [2, -1, 1]$, which sets the decision boundary at $\tau = -e + 2$, signifying a preference for lower travel times and emissions. This user will opt for the eco-friendly route since it is within the decision boundary. However, if $\boldsymbol{\theta}_i = [3, -2, 1]$ which means user pays more attention to travel time, this eco-friendly route will not be preferred.

In the offline training phase, using the dataset $\mathcal{D}$, a linear Support Vector Machine (SVM) is initially employed to differentiate the two classes of data points for each user i. This process yields the parameters $\{\boldsymbol{\theta}_i\}_{i \in [N']}$. Subsequently, K-Means clustering is applied to ascertain the centroids $\{\mathbf{c}_k\}_{k \in [K]}$ of the set $\{\boldsymbol{\theta}_i\}_{i \in [N']}$, where each centroid is represented as $\mathbf{c}_k = [\bar{b}_k, \bar{s}_k, \bar{o}_k]$. K serves as a hyperparameter. We select the value of K that yields the minimum regret on the offline training set.

In the online learning stage, first, the weight $p(k)$ is initialized for each expert. For every decision instance t pertaining to user i, we collect action data $x_{i,t}$. Utilizing the Hedge Algorithm, an expert $E_{i,t}$ is selected. The recommendation is then formulated as $Rec_{i,t} = \hat{y}(\mathbf{c}_{E_{i,t}}, x_{i,t})$, with $\hat{y}(\mathbf{c}_k, x_{i,t}) = 1 + \mathbb{1}_{\bar{o}_k(\tau_{i,t} - \bar{s}_k e_{i,t} - \bar{b}_k) > 0}$. Upon obtaining the user's choice $y_{i,t}$, the loss $\mathbf{l}_{i,t}(k) = |\hat{y}(\mathbf{c}_k, x_{i,t}) - y_{i,t}|^2$ is computed, leading to an adjustment of each expert's weight accordingly.

3.3. Clustering with Loss-guided Distance

The core parameter influencing the regret in our model is the set of centroids $\{\mathbf{c}_k\}_{k\in[K]}$. An accurately representative set of centroids can significantly reflect users' behaviors, whereas poorly chosen centroids may lead to suboptimal performance. In our simulations, we observed limitations with centroids generated by the standard K-Means algorithm. For instance, a centroid $\mathbf{c}_k$ that differs slightly from $\boldsymbol{\theta}_i$ in the bias term but exceeds the decision boundary can misclassify many points, resulting in higher regret. This implies that centroids with similar $\boldsymbol{\theta}_i$ values do not necessarily yield comparable performances. To address this issue, we introduce a distance metric guided by the loss function which is tailored for the online preference learning problem. Our objective is to ensure that θ_i values within the same cluster exhibit similar performance. Thus, we replace the traditional L$_2$ norm distance with the prediction loss incurred when assigning $\mathbf{c}_k$ to user i. Here, we define: $\mathbf{x}_i = [x_{i,1}, x_{i,2}, ..., x_{i,T'}] \in \mathbb{R}^{T'\times A\times d}$ and $\mathbf{y}_i = [y_{i,1}, y_{i,2}, ..., y_{i,T'}] \in \mathbb{R}^{T'}$, while $\hat{\mathbf{y}}(\mathbf{c}_k, \mathbf{x}_i) = [\hat{y}(\mathbf{c}_k, x_{i,1}), \hat{y}(\mathbf{c}_k, x_{i,2}), ..., \hat{y}(\mathbf{c}_k, x_{i,T'})] \in \mathbb{R}^{T'}$. The Loss-guided Distance is defined as $dist(i, \mathbf{c}_k) = ||\hat{\mathbf{y}}(\mathbf{c}_k, \mathbf{x}_i) - \mathbf{y}_i||^2$. The detailed clustering is presented in Algorithm 2.

Algorithm 2 K-Means with Loss-guided Distance

```
Require: $\{\boldsymbol{\theta}_i\}_{i\in[N']}$
  Randomly initialize centroids $\{\mathbf{c}_k\}_{k\in[K]}$
  while $\{\mathbf{c}_k\}_{k\in[K]}$ not converged do
    $dist(i, \mathbf{c}_k) \leftarrow ||\hat{\mathbf{y}}(\mathbf{c}_k, \mathbf{x}_i) - \mathbf{y}_i||^2$ for all $i \in [N'], k \in [K]$
    $r_{i,k} \leftarrow \mathbb{1}_{k=\arg\min_{k'} dist(i, c_{k'})}$ for all $i \in [N'], k \in [K]$
    $\mathbf{c}_k \leftarrow \frac{\sum_{i=1}^{N} r_{i,k}\boldsymbol{\theta}_i}{\sum_{i=1}^{N} r_{i,k}}$ for all $k \in [K]$
  end while
  return $\{\mathbf{c}_k\}_{k\in[K]}$
```

4. Regret analysis

4.1. Regret Bound of EWC

Before describing our theoretical findings, we first introduce some background and definitions. In the expert problem, spanning T total rounds with K experts, we denote the best expert throughout the duration as k^*. The regret bound, as established by Freund and Schapire (1997), is expressed as:

$$R_{Hedge} = \sum_{t=1}^{T} (\langle \mathbf{p}_t, \mathbf{l}_t \rangle - \mathbf{l}_t(k^*)) \leq 2\sqrt{T \log K} \tag{1}$$

The loss of K-Means algorithm is define as $\mathcal{L} = \sum_{i=1}^{N} ||\mathbf{c}_{k(i)} - \boldsymbol{\theta}_i||^2$, where $k(i)$ is the cluster centroid assigned to $\boldsymbol{\theta}_i$. Consider $\{\mathbf{c}_k\}_{k\in[K]}$ be any set of centroids, P as any distribution on $\mathbb{R}^d$ with mean $\boldsymbol{\mu} = \mathbb{E}_P[\boldsymbol{\theta}_i]$ and variance $\sigma^2 = \mathbb{E}_P[||\boldsymbol{\theta}_i - \boldsymbol{\mu}||^2]$. Assuming finite Kurtosis (4th moment) $\hat{M}_4 < \infty$ and given $\epsilon \in (0, 1)$, $\delta \in (0, 1)$ and a sample size m from P, we establish that for $m \geq \frac{12800(8+\hat{M}_4)}{\epsilon^2\delta}\left(3 + 30K(d+4)\log 6K + \log \frac{1}{\delta}\right)$, the Uniform deviation bound of K-Means, as proven by Bachem et al. (2017), holds with at least $1 - \delta$ probability:

$$|\mathcal{L} - \mathbb{E}_P[\mathcal{L}]| \leq \frac{\epsilon}{2}\sigma^2 + \frac{\epsilon}{2}\mathbb{E}_P[\mathcal{L}] \tag{2}$$

We define the regret of EWC as the performance difference between EWC and Oracle $\boldsymbol{\theta}_i$:

$$R_{EWC} = \sum_{i=1}^{N}\sum_{t=1}^{T}\left(\langle \mathbf{p}_{i,t}, \mathbf{l}_{i,t}\rangle - |\hat{y}(\boldsymbol{\theta}_i, x_{i,t}) - y_{i,t}|^2\right) \tag{3}$$

Since the study in Bachem et al. (2017) shows the performance of K-Means clustering using the L_2 norm distance, we similarly adopt the L_2 norm distance to analyze regret in our framework. What follows is our main theoretical result. Here we slightly abuse the notation $\hat{\mathbf{y}}(\boldsymbol{\theta}_i, \mathbf{x}_i) \in \mathbb{R}^T$ and $\mathbf{y}_i \in \mathbb{R}^T$ to be the prediction and user's choice vector in testing data.

Theorem 4.1 (Regret Bound of EWC) *Let P be any distribution of $\boldsymbol{\theta}_i \in \mathbb{R}^d$ with $\boldsymbol{\mu} = \mathbb{E}_P[\boldsymbol{\theta}_i]$, $\sigma^2 = \mathbb{E}_P[||\boldsymbol{\theta}_i - \boldsymbol{\mu}||^2]$, and finite Kurtosis. Let $\{\mathbf{c}_k\}_{k\in[K]}$ be any set of centroids, $k^*(i)$ be the best expert for user i, $\mathcal{L} = \sum_{i=1}^{N} ||\mathbf{c}_{k^*(i)} - \boldsymbol{\theta}_i||^2$ be the total squared distance of clustering, and $\hat{\mathbf{y}}(\boldsymbol{\theta}_i, \mathbf{x}_i) \in \mathbb{R}^T$ be the prediction function. If $\hat{\mathbf{y}}(\cdot, \mathbf{x}_i)$ is Lipschitz continuous for all $\mathbf{x}_i$ with Lipschitz constant L, L_2 norm distance, and dimension normalization, then with probability at least $1-\delta$, the regret of EWC is bounded by:*

$$R_{EWC} \le \overline{R}_{EWC} = 2N\sqrt{T\log K} + TL\left(\frac{\epsilon}{2}\sigma^2 + (\frac{\epsilon}{2}+1)\mathbb{E}_P[\mathcal{L}]\right) \tag{4}$$

Proof

$$\begin{aligned}
R_{EWC} =& \sum_{i=1}^{N}\sum_{t=1}^{T}\left(\langle \mathbf{p}_{i,t}, \mathbf{l}_{i,t}\rangle - |\hat{y}(\boldsymbol{\theta}_i, x_{i,t}) - y_{i,t}|^2\right)\\
=& \sum_{i=1}^{N}\sum_{t=1}^{T}\left(\langle \mathbf{p}_{i,t}, \mathbf{l}_{i,t}\rangle - |\hat{y}(\mathbf{c}_{k^*(i)}, x_{i,t}) - y_{i,t}|^2\right)\\
&+ \sum_{i=1}^{N}\sum_{t=1}^{T}\left(|\hat{y}(\mathbf{c}_{\mathbf{k^*(i)}}, x_{i,t}) - y_{i,t}|^2 - |\hat{y}(\boldsymbol{\theta}_i, x_{i,t}) - y_{i,t}|^2\right)\\
=& \sum_{i=1}^{N}\sum_{t=1}^{T}\left(\langle \mathbf{p}_{i,t}, \mathbf{l}_{i,t}\rangle - \mathbf{l}_t(k^*(i))\right) + \sum_{i=1}^{N}\left(||\hat{\mathbf{y}}(\mathbf{c}_{k^*(i)}, \mathbf{x}_i) - \mathbf{y}_i||^2 - ||\hat{\mathbf{y}}(\boldsymbol{\theta}_i, \mathbf{x}_i) - \mathbf{y}_i||^2\right)
\end{aligned} \tag{5}$$

By the regret bound of Hedge and triangle inequality,

$$R_{EWC} \le 2N\sqrt{T\log K} + \sum_{i=1}^{N} ||\hat{\mathbf{y}}(\mathbf{c}_{k^*(i)}, \mathbf{x}_i) - \hat{\mathbf{y}}(\boldsymbol{\theta}_i, \mathbf{x}_i)||^2 \tag{6}$$

By the Lipschitz condition, $\exists L$ s.t. $\forall i, \forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2$, $\frac{1}{T}||\hat{\mathbf{y}}(\boldsymbol{\theta}_1, \mathbf{x}_i) - \hat{\mathbf{y}}(\boldsymbol{\theta}_2, \mathbf{x}_i)||^2 \le L||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2||^2$

$$\begin{aligned}
\sum_{i=1}^{N} ||\hat{\mathbf{y}}(\mathbf{c}_{k^*(i)}, \mathbf{x}_i) - \hat{\mathbf{y}}(\boldsymbol{\theta}_i, \mathbf{x}_i)||^2 &\le TL\sum_{i=1}^{N} ||\mathbf{c}_{k^*(i)} - \boldsymbol{\theta}_i||^2\\
&= TL\mathcal{L} \le TL(|\mathcal{L} - \mathbb{E}[\mathcal{L}]| + \mathbb{E}[\mathcal{L}])
\end{aligned} \tag{7}$$

By inequation 2, with probability at least $1-\delta$,

$$\sum_{i=1}^{N} ||\hat{\mathbf{y}}(\mathbf{c}_{k^*(i)}, \mathbf{x}_i) - \hat{\mathbf{y}}(\boldsymbol{\theta}_i, \mathbf{x}_i)||^2 \le TL\left(\frac{\epsilon}{2}\sigma^2 + (\frac{\epsilon}{2}+1)\mathbb{E}[\mathcal{L}]\right) \tag{8}$$

$$R_{EWC} \le 2N\sqrt{T\log K} + TL\left(\frac{\epsilon}{2}\sigma^2 + (\frac{\epsilon}{2}+1)\mathbb{E}[\mathcal{L}]\right) \tag{9}$$

■

The Gaussian Mixture Model (GMM) aligns closely with our hypothesis of a hierarchical structure among users, which is a typical assumption in the analysis of clustering algorithms. By assuming that the distribution of users' preferences follows a GMM, we derive Corollary 4.1.1.

Corollary 4.1.1 *If P is a Gaussian Mixture Model (GMM) with K Gaussian distributions, each of which has weight π_k, mean $\boldsymbol{\mu}_k$, and covariance Σ_k, and the clustering outputs the optimal centroids where $\mathbf{c}_k = \boldsymbol{\mu}_k$. Define $l_{centroids} = L\frac{\epsilon}{2N}\sigma^2 + L(\frac{\epsilon}{2}+1)\sum_{k=1}^{K}\pi_k trace(\Sigma_k)$ be the average loss caused by centroids. With probability at least $1-\delta$, the regret of EWC is bounded by*

$$R_{EWC} \le \overline{R}_{EWC} = 2N\sqrt{T\log K} + TNl_{centroids} \tag{10}$$

Proof Since $\mathbf{c}_k = \boldsymbol{\mu}_k$, and $P = \sum_{k=1}^{K}\pi_k\mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$, the expected squared distance is $\mathbb{E}[||\boldsymbol{\theta}_i - \mathbf{c}_{k(i)}||^2] = \sum_{k=1}^{K}\pi_k trace(\Sigma_k)$. So, $\mathbb{E}[\mathcal{L}] = N\mathbb{E}[||\boldsymbol{\theta}_i - \mathbf{c}_{k(i)}||^2] = N\sum_{k=1}^{K}\pi_k trace(\Sigma_k)$. ■

4.2. Comparison

We compare the regret bound of the EWC algorithm with LinUCB and oracle Follow-the-Leader (oracle FTL). Follow-the-Leader (FTL) is a straightforward method that selects the option with the best historical performance up to the current time step $k = \text{argmin}_{k'}\sum_{t'=1}^{t}\mathbf{l}_{t'}(k')$. The oracle FTL is an oracle method that lets us know the best option up to time T in hindsight and always chooses it at decision rounds. Lemma 4.2.1 is the regret bound of SupLinUCB (a varient of LinUCB) which has been proved by Li et al. (2010), and lemma 4.2.2 is the regret bound of oracle FTL. Corollary 4.2.1 compares EWC with both LinUCB and Oracle FTL.

Lemma 4.2.1 (Regret Bound of SupLinUCB) *Assume $\forall i, t, \exists\theta_i^* \in \mathbb{R}^d$, s.t. $E[\mathbf{l}_{i,t}(a)|x_{i,t}(a)] = x_{i,t}(a)^\intercal\theta_i^*$. Define $R_{LinUCB} = \sum_{i=1}^{N}\sum_{t=1}^{T}\left(\mathbf{l}_{i,t}(a_{i,t}) - \mathbf{l}_{i,t}(a_{i,t}^*)\right)$ where $a_{i,t}^* = \arg\max_a x_{i,t}(a)^\intercal\theta^*$. If SupLinUCB runs with $\alpha = \sqrt{\frac{1}{2}\ln\frac{2TK}{\delta}}$, with probability at least $1-\delta$, $R_{LinUCB} < \overline{R}_{LinUCB} = O\left(N\sqrt{Td\ln^3(KT\ln T/\delta)}\right)$.*

Lemma 4.2.2 (Regret Bound of Oracle FTL) *Define the regret of oracle FTL be $R_{OracleFTL} = \sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{l}_{i,t}$. Let p_i be the proportion of choosing option 1 for each user i. The regret of oracle FTL is $R_{OracleFTL} = \sum_{i=1}^{N}T\min\{p_i, 1-p_i\} = O(TN)$.*

Proof Since we always choose the best one of options $\{1,2\}$ for each user, the number of wrong prediction should be $T\min\{p_i, 1-p_i\}$. So the total regret is the summation of all users. ■

Corollary 4.2.1 (Advantage of EWC) *1) Assume $\overline{R}_{LinUCB} = CN\sqrt{Td\ln^3(KT\ln T/\delta)}$, then when $T < (\frac{C-2}{l_{centroids}})^2$, $\overline{R}_{EWC} < \overline{R}_{LinUCB}$. 2) When $l_{centroids} < \frac{1}{N}\sum_{i=1}^{N}\min\{p_i, 1-p_i\} - 2\sqrt{\log(K)/T}$, $\overline{R}_{EWC} < R_{OracleFTL}$*

Proof 1) Since $\overline{R}_{EWC} = 2N\sqrt{T\log K} + TNl_{centroids}$, $\overline{R}_{EWC} < \overline{R}_{LinUCB}$ is equivalent to $\sqrt{T}l_{centroids} < c\sqrt{d\ln^3(KT\ln T/\delta)} - 2\sqrt{\log K}$. So when $\sqrt{T}l_{centroids} < c-2$, the condition above is satisfied. 2) Dividing $2N\sqrt{T\log K} + TNl_{centroids}$ and $\sum_{i=1}^{N}T\min\{p_i, 1-p_i\}$ by NT, we can get the second result. ■

As highlighted in Corollary 4.2.1, EWC demonstrates superior theoretical performance compared to LinUCB when T is relatively small. This advantage is contingent upon $l_{centroids}$ which is the average loss incurred when using the centroids $\mathbf{c}_k$ as representations of the users' preference parameters $\boldsymbol{\theta}_i$. EWC outperforms the oracle Follow-the-Leader (FTL) when the loss due to employing centroids is less than the loss from consistently selecting the fixed best arm. The term $2\sqrt{\log(K)/T}$ represents the average loss associated with the process of identifying the best expert. This loss is negligible since it decreases rapidly as T increases.

5. Experiments

In this section, we assess the Expert with Clustering (EWC) algorithm through experiments designed to evaluate its performance in learning driver preferences online, particularly its adaptability to new data and accuracy in making eco-friendly route recommendations.

5.1. Experimental Setup

Community survey. This study involved a community survey conducted in July 2023 on the University of North Carolina at Charlotte campus, and a total of 43 individuals participated. Participants provided the driving choice preferences as well as demographic data covering age, gender, ethnicity, and educational level. The survey's main component involved a series of questions assessing willingness to adhere to route recommendations under varying scenarios with distinct travel times and carbon dioxide emission levels. Participants rated their likelihood of following these recommendations in the Likert scale, offering insight into their decision-making criteria. For example, participants were asked on their likelihood to opt for an eco-friendly route offering a 10% reduction in CO_2 emissions in exchange for a 5–15% increase in travel time.

Mobility user simulation. To better represent a diverse driving population, we expanded our dataset. We use the Bayesian inference model that resembles the original distribution from the survey data (Andrieu et al. (2003)). We refined this approach by implementing distinct utility functions for demographic segments differentiated by gender, age, and household car ownership. Drawing samples from the posterior distribution of model parameters, we populated the dataset with individuals exhibiting a range of features and compliance behavior. This methodology allowed us to produce 2000 individual user choice records for the synthetic dataset, with parameters set at $N = 800$, $N' = 1200$, $T = T' = 40$, and $K = 6$. The optimal K was selected based on regret minimization. The synthetic dataset features a mix of route choices that reflect various driving preferences and behaviors, providing a rich foundation for evaluating our EWC algorithm.

Baselines. Our approach is benchmarked against a selection of well-established baseline algorithms. *Follow-the-Leader (FTL)* predicts the future actions based on historically rewarding choices. The *Linear Upper Confidence Bound (LinUCB)* algorithm adapts the upper confidence bound method for linear payoffs, optimizing the trade-off between exploring new actions and exploiting known ones. *Oracle Follow-the-Leader (Oracle FTL)* always chooses the historically optimal option. The *Oracle Cluster* algorithm makes predictions using precise cluster assignments to incorporate collective behaviors within a user's group for decision-making. Lastly, *Oracle* $\boldsymbol{\theta}_i$ leverages a perfect understanding of user preferences and behaviors to anticipate the most likely user action.

5.2. Results

Figure 2 compares the regret of various online learning algorithms on a synthesized dataset based on driving preferences and eco-driving recommendations. Regret measures how much worse an

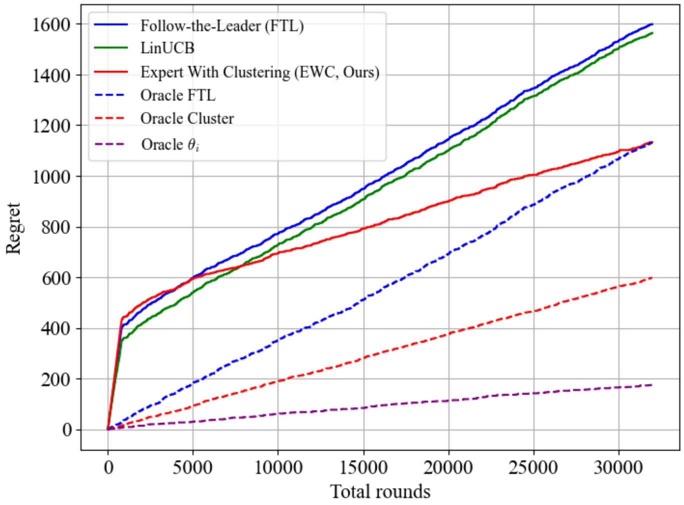

Figure 2: Comparative regret analysis of online learning algorithms: Expert with Clustering (EWC, Ours) shows lower regret than the baseline algorithms (Follow-the-Leader, LinUCB, Oracle FTL) and approaches the consistency of the Oracle methods.

algorithm performs compared to the best possible action over time. Oracle θ_i shows the lowest regret, indicating it almost perfectly predicts the best action due to its assumption of perfect user preference information. Oracle Cluster also performs well, benefiting from knowledge of user clusters. Oracle FTL exhibits a high slope value comparable to that of standard FTL. LinUCB and FTL algorithms experience higher regret. FTL, relying solely on historical action frequency, performs worst among baselines. LinUCB's expressiveness is limited, which leads to a similar performance with FTL. In the early rounds, LinUCB and the EWC algorithm start with similar levels of regret, suggesting that initially, both algorithms perform comparably in predicting the best action. This could be because, in the initial stages, there's less historical data to differentiate the predictive power of the algorithms, or the correct action is more obvious. EWC surpasses non-oracle methods, showing clustering's effectiveness in capturing user preferences. Its long-term regret slope mirrors that of Oracle Cluster, suggesting rapid identification of optimal user group affiliations. The EWC algorithm's performance gradually improves, indicating that it is increasingly predicting the optimal actions, reducing regret by 27.57% compared to the LinUCB at the final rounds. This could be due to the inherent advantages of its clustering-based predictive model, which, despite lacking perfect foresight, benefits from insights that approach the prescience of the Oracle methods.

6. Conclusion

In this paper, we introduce Expert with Clustering (EWC), a novel hierarchical contextual bandits algorithm designed to address the online learning challenges for drivers' preferences. EWC uniquely combines clustering techniques with prediction based on expert advice, effectively achieving low regret in online learning scenarios. Furthermore, EWC offers an efficient method for extracting insights into both population-wide and individual-specific behaviors, proving particularly effective in contextualized settings that exhibit hierarchical structures. In future work, we plan to refine EWC by incorporating more user-specific preference learning and investigating the preference for incentives, thereby enhancing the personalization and effectiveness of our recommendations.

Acknowledgments

This work was partially supported by the National Science Foundation (NSF) under grant number 2149511 and the Kwanjeong scholarship. The authors would like to thank Prof. Christos Cassandras, Prof. Andreas Malikopoulos, and Prof. Roy Dong for the insightful discussion.

References

Christophe Andrieu, Nando de Freitas, Arnaud Doucet, and Michael I. Jordan. An Introduction to MCMC for Machine Learning. *Machine Learning*, 50:5–43, January 2003.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Olivier Bachem, Mario Lucic, S. Hamed Hassani, and Andreas Krause. Uniform deviation bounds for unbounded loss functions like k-means, 2017.

Sai Krishna Chada, Daniel Görges, Achim Ebert, Roman Teutsch, and Shreevatsa Puttige Subramanya. Evaluation of the driving performance and user acceptance of a predictive eco-driving assistance system for electric vehicles. *Transportation Research Part C: Emerging Technologies*, 153: 104193, August 2023. ISSN 0968-090X. doi: 10.1016/j.trc.2023.104193. URL https://www.sciencedirect.com/science/article/pii/S0968090X23001821.

Caspar G. Chorus, Theo A. Arentze, and Harry J.P. Timmermans. Traveler compliance with advice: A Bayesian utilitarian perspective. *Transportation Research Part E: Logistics and Transportation Review*, 45(3):486–500, May 2009. ISSN 13665545. doi: 10.1016/j.tre.2008.10.004. URL https://linkinghub.elsevier.com/retrieve/pii/S1366554508001336.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997. ISSN 0022-0000. doi: https://doi.org/10.1006/jcss.1997.1504. URL https://www.sciencedirect.com/science/article/pii/S002200009791504X.

Claudio Gentile, Shuai Li, and Giovanni Zappella. Online clustering of bandits. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 757–765, Bejing, China, 22–24 Jun 2014. PMLR. URL https://proceedings.mlr.press/v32/gentile14.html.

Claudio Gentile, Shuai Li, Purushottam Kar, Alexandros Karatzoglou, Giovanni Zappella, and Evans Etrue. On context-dependent clustering of bandits. In *International Conference on machine learning*, pages 1253–1262. PMLR, 2017.

Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19586–19597. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/e32cc80bf07915058ce90722ee17bb71-Paper.pdf.

Devansh Jalota, Karthik Gopalakrishnan, Navid Azizan, Ramesh Johari, and Marco Pavone. Online Learning for Traffic Routing under Unknown Preferences, March 2022. URL http://arxiv.org/abs/2203.17150. arXiv:2203.17150 [cs, math].

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A Contextual-Bandit Approach to Personalized News Article Recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, April 2010. doi: 10.1145/1772690.1772758. URL http://arxiv.org/abs/1003.0146. arXiv:1003.0146 [cs].

Shuai Li, Alexandros Karatzoglou, and Claudio Gentile. Collaborative filtering bandits, 2016.

N. Littlestone and M.K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994. ISSN 0890-5401. doi: https://doi.org/10.1006/inco.1994.1009. URL https://www.sciencedirect.com/science/article/pii/S0890540184710091.

Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2): 129–137, 1982.

Dorsa Sadigh, Anca Dragan, Shankar Sastry, and Sanjit Seshia. Active Preference-Based Learning of Reward Functions. In *Robotics: Science and Systems XIII*. Robotics: Science and Systems Foundation, July 2017. ISBN 978-0-9923747-3-0. doi: 10.15607/RSS.2017.XIII.053. URL http://www.roboticsproceedings.org/rss13/p53.pdf.

Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE Transactions on Neural Networks and Learning Systems*, 32(8):3710–3722, 2021. doi: 10.1109/TNNLS.2020.3015958.

Leonardo F. Toso, Han Wang, and James Anderson. Learning personalized models with clustered system identification. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 7162–7169, 2023. doi: 10.1109/CDC49753.2023.10383950.

Ran Tu, Junshi Xu, Tiezhu Li, and Haibo Chen. Effective and Acceptable Eco-Driving Guidance for Human-Driving Vehicles: A Review. *International Journal of Environmental Research and Public Health*, 19(12):7310, June 2022. ISSN 1660-4601. doi: 10.3390/ijerph19127310. URL https://www.mdpi.com/1660-4601/19/12/7310.