Defence Science Journal, Vol. 74, No. 6, November 2024, pp. 885-892, DOI : 10.14429/dsj.74.19846

Hybrid-Neuro Bandit: A Bandit Model for Online Recommendation

Nipun Bansal[#,*], Manju Bala[$] and Kapil Sharma[#]

[#]*Delhi Technological University, Delhi - 110 042, India*
[$]*I.P. College for Women, Delhi University, Delhi - 110 054, India*
[*]*E-mail: nipunbansal@dtu.ac.in*

ABSTRACT

Contextual Multi-Arm Bandit (CMAB) is a popular framework for sequential decision-making problems where an agent must repeatedly choose among multiple actions, each with an unknown reward distribution. The CMAB agent aims to maximize its cumulative reward over a finite or infinite horizon. At the same time, a high level of accountability is required, and there is a need to understand the underlying mechanism so that the user can trust the model's decisions. In this direction, a novel Hybrid Neuro Bandit (HNB) model is proposed, which infuses the expert advice from the existing contextual multi-arm bandits into one combined unit, thereby exploiting the different CMAB algorithm merits and providing personalized recommendations to the user's liking. The proposed HNB model decisions can be easily understandable by the user as the HNB agent ignores the non-performing bandit experts and considers the opinion of the majority of the bandit experts. The HNB model has been empirically compared with the existing state-of-the-art contextual bandit models over nine performance metrics, namely recall, specificity, precision, prevalence, F1 score, Matthews Correlation Coefficient (MCC), Fowlkes–Mallows index (FM), Critical Success Index (CSI) and accuracy.

Keywords: Explainable AI(XAI); Adaptive neuro-fuzzy inference system; Contextual multi-arm bandit; Personalization; Reinforcement learning

1. INTRODUCTION

Consider an online advertisement(ad) model, like YouTube, where a user visits the website, queries a request, and gets rendered with an advertisement along with the requisite content on the web page. The system aims to show the most relevant ad on the web page to maximize the likelihood of the user clicking on the displayed ad. If the user clicks on the ad, the system receives the reward. Feedback for the system is the action taken by the user. The system, however, cannot observe the user's actions if other advertisements are shown. Therefore, the system needs to trade-off between the exploration and exploitation dilemma, i.e. either keep showing the ad earlier clicked by the user (exploitation) every time the user requests a webpage or take the risk to explore the new ad from the set of available ads. The above-mentioned scenario can be modelled as a contextual multi-arm bandit problem[25] where each trial is a user visit to the website, and each arm is the ad displayed. For each trial, the context is announced, and the agent chooses an arm from the set of predefined arms [k]= {1, 2, ...k} and observes the reward associated with the selected arm. Over several iterations, the agent finds a relation between the context and rewards obtained, i.e. arm reward distribution. The ultimate goal is to maximize the cumulative rewards over n trials or improve the accuracy of predicting the correct arm in each trial by updating the model with the feedback received from the user after each trial. This involves the exploration of a new ad to cater to stochastic user behaviour over time. For instance, the system showed the relevant ad to the user, and the user clicked on it and bought the concerned item. After that, the user doesn't want to see the same ad for that item, and the system needs to adapt to this change in user behaviour. In literature, both unguided exploration (e.g., greedy[26] and epoch-greedy[11]) and guided exploration (e.g., LinUCB[27], EXP4[28], and Thompson sampling[29]) algorithms have been proposed and are used widely in various real-world applications such as clinical trials to reduce patient losses by figuring out the best medicine for a given context of symptoms, marketing optimization to improve click-through rates, website layout optimization as an alternative to A/B testing, adaptive routing to minimize network delays and dynamic allocation of resources in different funding research projects with the given risk and the probability of the positive payoffs. However, in practice, several challenges are faced where the most predominately are exploration-exploitation trade-off (i.e. to explore new options for better payoff or choose the known option), non-stationarity (i.e. reward distributions of the arms changes over time), contextual inference, scalability (i.e. the ability to handle high-dimensional and large datasets) and, robustness.

Another significant hurdle encountered is the cold start predicament, where numerous recently introduced items exhibit minimal historical interactions with users, rendering personalized recommendations nearly impossible. Additionally, there is a need to elucidate the rationale behind the decisions made by state-of-the-art contextual bandit

Received : 23 January 2024, Revised : 06 May 2024
Accepted : 07 May 2024, Online published : 25 November 2024

models. This constitutes a major bottleneck in the autonomous deployment of the model, as users find it challenging to trust the model's decisions.

To meet the above challenges, a novel Hybrid Neural Bandit (HNB) model is proposed, where each arm mimics an existing contextual multi-arm bandit model and acts as a bandit expert. The decisions formulated by these bandit experts are integrated to capitalize on the capabilities of contemporary contextual bandit algorithms. This involves assigning weights to individual bandit experts and subsequently fine-tuning these weights through a neural network, incorporating pertinent user feedback within the system. Furthermore, the suggested Hybrid Neural Bandit (HNB) model delivers tailored recommendations based on user preferences, ultimately elevating the overall user experience and interaction with the applications. Furthermore, the HNB model is interpretable, i.e. a user can easily understand the decision made. So, a user can trust the proposed HNB model's decisions and deploy the model at scale.

The remainder of this paper is organised as follows. Section 2 comprises the literature review of various contextual multi-arm bandit models, their limitations and applications. Section 3 presents a detailed study of the proposed Hybrid Neuro Bandit model to predict arm reward distribution. In Section 4, a comparison of the performance of our proposed model with the traditional state-of-the-art contextual bandit algorithms is made. Finally, Section 5 concludes and discusses extensions of this work.

2. LITERATURE REVIEW

A neural network's ability to process information is constrained by its number of parameters. Introducing a Sparsely-Gated Mixture-of-Experts layer[1] (MoE), housing tens of thousands of feed-forward sub-networks, addresses this constraint. An iterative approach commonly used for estimating parameters with maximum likelihood is the Expectation-Maximization[2] (EM) algorithm. Empirically, they demonstrated that the EM algorithm produces considerably faster convergence for these structures than gradient ascent. The Dempster-Shafer theory of evidence supports a combination method presented in[3] that uses statistics about the relative classification strengths of various classifiers. A tree-structured architecture[4] for supervised learning is the underlying statistical model of the design, which is a hierarchical mixture model with generalised linear models for both the mixture coefficients and the mixture components (GLIMs). Learning is viewed as a maximum likelihood problem. The process[5] of merging the classification abilities of multiple classifiers is recognised as a generic problem in various pattern recognition application domains, and a thorough examination has been conducted. Four algorithms were independently created[6], each of these techniques as an expert because of their excellent recognition rates. Combining them enables us to see how well they work together. The resulting multi-expert system demonstrates that the consensus of various approaches tends to preserve individual strengths while making up for individual weaknesses. Recommendation[7] of news articles and advertisements in web portals, in dynamic circumstances, using implicit input and clicks from users, provide the system with feedback in these dynamic settings. This feedback must be swiftly tapped into to enhance the following recommendations. It is understood that the multi-armed bandit problem[8] is an exploration/exploitation trade-off problem where the user's happiness is maximised by choosing the best arm or exploitation while investigating

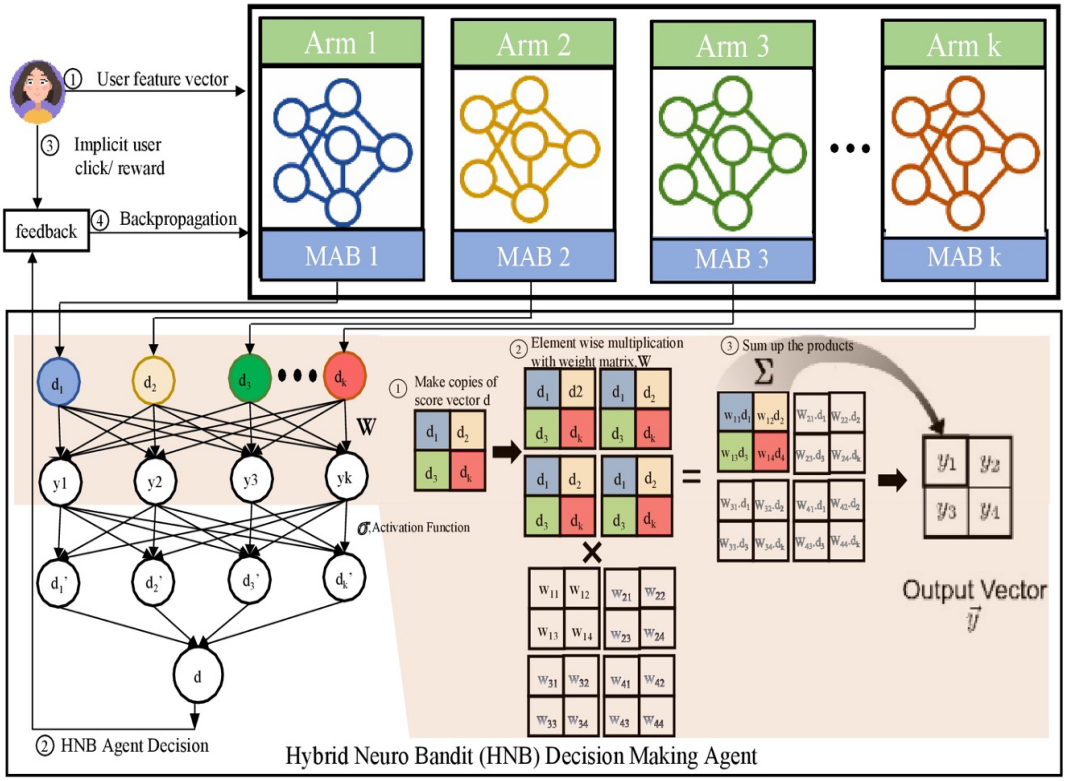

Figure 1. Hybrid neuro bandit model architecture.

the new options/arms for uncertainty in the user's interests. While Thompson Sampling[9,10] is the heuristic algorithm that has handled the exploration-exploitation trade-off by keeping probability distributions for each arm and then sampling from them on each trial to choose the one that predicts greater rewards, Epsilon-greedy[12] and Epoch Greedy[11] are famous algorithms for random exploration. A linear link between the predicted reward and the situation was presupposed by LinUCB[13-14]. Nevertheless, Chapelle and Li[9] demonstrated that Thompson Sampling (TS) defeats UCB. The value of innovation and diversity is increasingly being incorporated into evaluation practices[15-16]. Many contexts, including movies[17], tags[18], and adverts[19], have seen the successful application of multi-objective recommender systems. Multi-objective Multi-armed Bandit (MO-MAB) algorithms-based techniques merit our attention when handling online settings. In MO-MAB, depending on the set of objectives, several arms (items) are candidates for the best solution[20-22]. In this work, we adopt the Ranked Bandits approach[23-24], as we are interested in offering various things to users while considering multiple objectives. A rating of objects is learned via the Ranked Bandits online learning algorithm based on user behaviour[24]. The authors[23] presented several Ranked Bandits iterations that consider side information regarding the similarity of items provided by search engines.

3. METHODOLOGY

This paper proposes a novel Hybrid Neuro Bandit (HNB) model that optimizes decisions for each user based on the previous user's likings and preferences in an online setting. This enables the proposed model to run on the fly with minimal training, thereby making the model fast, optimal (fewer resources are required), and usable in various real-time applications where much training data is unavailable. At the same time, the user can also understand the decisions made by the HNB model simply and intuitively, thereby allowing the user to trust the model's decisions.

As shown in Fig.1, the proposed model consists of k arms, where each arm corresponds to an existing contextual multi-arm bandit algorithm and acts as the bandit expert. Each bandit expert has merit and can be chosen based on their proven efficacy in the literature. The HNB model amalgamates these bandit algorithms into a cohesive unit, leveraging their respective merits and properties to overcome the challenges encountered by individual bandit algorithms. The amalgamation is done by fusing the decision made by each bandit expert by assigning the weight to each bandit expert's opinion; in turn, this allows the HNB agent to ignore the opinion of any bandit expert in the final decision. This ensures that if any particular bandit expert consistently provides inaccurate predictions or underperforms in multiple trials, the HNB model disregards the input from that non-performing expert. Instead, it gives more weight to the opinions of other bandit experts that demonstrate better performance, thereby enhancing the overall accuracy and reliability of the model. A crucial element for effectively achieving this is optimally and verifiably training the weights assigned to bandit experts. This is significant not only for explaining the decisions of the HNB model to the user but also for making the model interpretable. This transparency allows the HNB model to be accountable for its decisions, fostering user trust—an essential quality often absent in many existing state-of-the-art bandit models.

To understand the HNB model agent's decision, let us consider a training sample snippet as shown in Fig. 2. decision vectors d1, d2, and d3 represent the decisions outputted by the existing bandit algorithm experts in the proposed HNB model (also called bandit experts). The $\overrightarrow{ud}$ vector represents the user's liking of the choice made and is represented as one if liked; otherwise, 0. Similarly, if the individual bandit expert decision is correct, i.e., the choice predicted by the bandit expert is the same as the user liking the corresponding decision vector, say d1 is 1, otherwise 0. At the beginning of the HNB model network training, let the initial weights and bias be zero, and the value of learning rate i.e., α = 0.5. 'z' will be the output

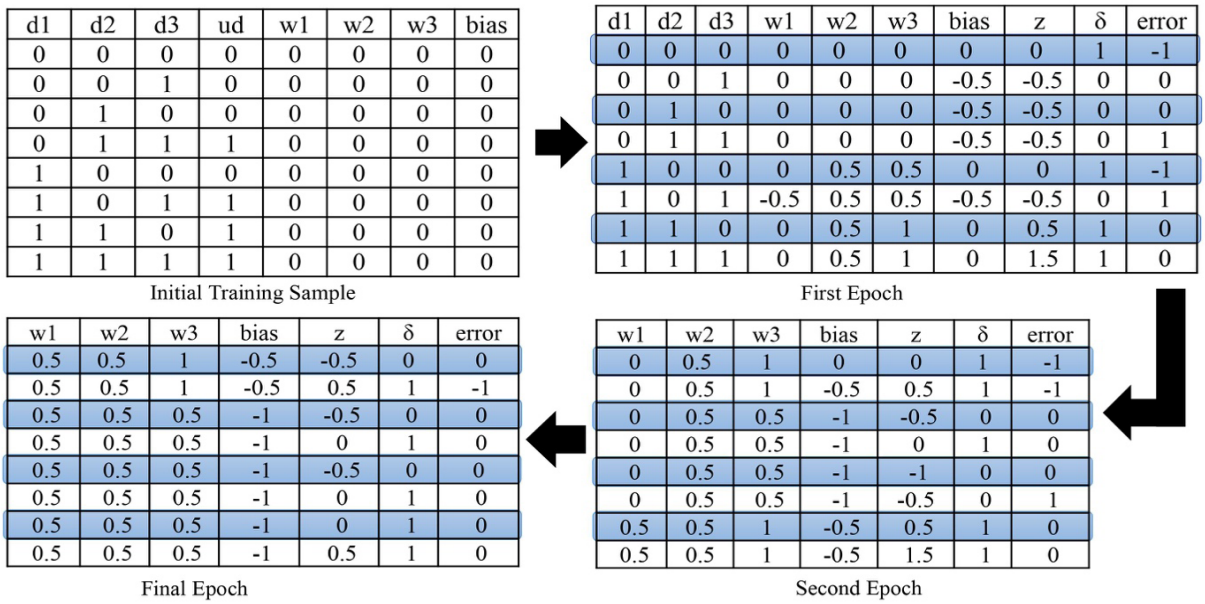

Figure 2. The weight matrix updation in the proposed HNB model.

Algorithm 1 Hybrid Neuro Bandit (HNB) Algorithm

Input: Given data-set of the training sample $\{x(t),t\}_{i=1}^{N} \in \{R^d \times R\}$

Initialize $MAB_1, MAB_2 ... MAB_k$ for each arm ▷ MAB_i is any existing multi arm bandit algorithm
for epoch $i = 1,2,\ldots$ **do**
  for $t = 1,2,\ldots$ N **do**
    // Calculate the decision vector,$\vec{d}$ from the initialized *MAB* experts
    if MAB_i decision is correct **then**
      $d_i \leftarrow 1$
    else
      $d_i \leftarrow 0$
    // Input the decisions predicted by the MAB experts into Neural Network
    Initialize random weight matrix W for the *MAB* expert's decision
    $HNB\_decision[t] \leftarrow feedforward(\vec{d}, \mathrm{W})$ ▷ $z = sigmoid(d.W + bias)$
    get HNB agent action, $\vec{a}(t) \leftarrow HNB\_decision[t]$
    // display action chosen by the HNB agent to the user chosen action, a(t)
    if $\vec{a(t)} = a(t)$ **then** error = 0
    else
      error =1
    update(W, error)

of neuron i.e., $z = \sum d_i \cdot w_i + bias$ and the output of the proposed model will be decided by a step activation function δ (if the value is less than zero then the output will be 0; otherwise, it will be 1).

Let us examine the weight updates during the first epoch of training. The first instance that the perceptron processes is $\vec{d} = (\vec{d1}, \vec{d2}, \vec{d3}) = (0,0,0)^{\mathrm{t}}$. In this instance, the perceptron's net input is: $z = w^{\mathrm{t}}\vec{d} + b = 0 \times 0 + 0 \times 0 + 0 \times 0 + 0 = 0$. Thus, its output is $\vec{o} = 1$, as the step function produces 1 when the input is ≥ 0. However, the target label in this instance is y=0, resulting in an error of $\vec{ud} - \vec{o} = -1$ for the perceptron. Following the perceptron learning rule, each weight w_i is updated by adding $\alpha(\vec{d} - \vec{o})x_i = -0.5x_i$ to it. Since all the inputs in this instance are 0, except for the bias neuron (x_0=1), only the bias is modified to -0.5 instead of 0. The same procedure is repeated for the other seven training examples, and weight updates are shown in the first epoch table in Fig. 2. It can be seen in the table that during the first epoch, the perceptron makes four errors, and the value of the weight vector after the first epoch is w=(0,0.5,1)^t and of the bias is 0. During the second epoch of the training sample, similar operations are performed by the perceptron. This time, the perceptron only makes three errors. After the second epoch, the weight vector is w=(0.5,0.5,1)^t and the bias is -0.5. After updating the second example in this epoch, the perceptron converges to the weight vector that solves this classification problem: w=(0.5,0.5,0.5)^t and b = -1. Since all the weights are equal, the perceptron only fires when at least two of the inputs are 1, resulting in their weighted sum being greater or equal to 1, which is greater or equal to the absolute value of the bias (-1), thereby ensuring that the net input of the perceptron is non-negative

3.1 Hybrid Neuro Bandit Ranking Algorithm

The proposed HNB model simulates contextual multi-arm bandit (CMAB) model settings where it inputs a finite n-dimension user context feature vector, $x(t)$={$x_1(t)$, $x_2(t)$,… $x_n(t)$} and chooses an action, $a(t)$ from an alternate number of choices/actions for each trial t. Each action is associated with a reward unknown to the model and is revealed after the action is chosen. The proposed model observes a binary reward, $r_{a(t)}(t)$ i.e. +1 if the user accepts the rendered action; otherwise, 0. In order to maximize the cumulative rewards over total trials, the HNB model uses a Hybrid Neuro Bandit Ranking Algorithm. The HNB model consists of k HNB arms where each HNB arm mimics a contextual multi-arm bandit model acting as a bandit expert and inputs an n-dimension user context feature $x(t)$. For each trial, each bandit expert outputs a decision vector d_i which represents the final action chosen by the bandit expert to be rendered to the end user. The decision vector $\vec{d} = \{\mathrm{d_1(t)}, \mathrm{d_1(t)}, \ldots \mathrm{d_k(t)}\}$ is then assigned a random weight matrix and is inputted into the HNB neural network. The HNB neural agent then chooses the best decision from all the inputted decisions suggested by the bandit experts and displays the final action to the user. If the user accepts the rendered HNB decision, no error is computed. Otherwise, feedback is sent back, which allows the proposed model to explore new actions in the next iteration.

4. DATASET ANALYSIS & PREPROCESSING

This paper uses two publicly available datasets: Mushroom dataset and Adult Income from the UCI machine learning repository. These datasets have been widely used in the literature as benchmark datasets to measure the performance of various contextual multi-arm bandit algorithms. Table 1 shows the number of contexts, the number of context features and the number of choices/ actions in each dataset.

4.1 Mushroom Dataset

The Mushroom Dataset consists of 8124 instances of mushrooms as recorded data with 22 feature columns. The first column of the dataset is the classified mushroom type of that instance as a binary number (0/1), stating whether the mushroom belongs to the edible or poisonous category. The other features include cap-shape, cap-surface, cap-color, bruises, odor, gill-attachment, gill-spacing, gill-size, gill-color and stalk-shape. The aim is to correctly predict the category of mushroom, i.e., edible or poisonous.

4.2 Adult Income Dataset

The adult income dataset consists of 14 features such as age, education, fnlwgt, workclass, education-num, marital-status, occupation, relationship, race, sex, capital- gain, capital-loss, hours-per-week and native-country of an adult that determines the income of an adult. There is a total of 32561 instances of recorded data where the last column of the dataset is the classified adult income of that instance as a binary number (0/1).

Table 1. Description of datasets

Datasets	# Contexts	# Features	# Actions
Mushroom	8124	22	2
Adult Income	32561	13	2

The original datasets have missing values and consist of categorical and numerical features that need pre-processing for better prediction and outcome. Firstly, all the rows with missing values are dropped from the dataset, and standardization is performed to convert categorical features into numerical ones. One hot encoding and Ordinal Encoding are the two most popular standardization techniques used. Ordinal Encoding maps each unique feature to an integer value and is used as a known ordinal relationship between categorical features in the datasets. It transformed the dataset with a mean of zero and a standard deviation of one. Then, the anomalies and outliers are removed from all the datasets. For that, a correlation matrix of all the dataset's features is made, and the least relevant features are removed. For e.g., the feature fnlwgt is least relevant in the adult dataset as the correlation value is negative and, thus, is safely dropped from the dataset. Also, all the entries in the adult dataset having various anomalies, such as context feature horizontal and vertical distance, can't be negative and are removed. There are approximately 200 outliers found in the dataset. Lastly, the features are scaled to a given range (default at (0,1)) with the help of a MinMax/ Robust scaler to prevent the optimisation from getting stuck in local optima and make training faster.

5. EXPERIMENT AND RESULTS

In this section, the proposed HNB model is compared with the existing CMAB algorithms; first is the uniform sampling model, which randomly makes the choice and ignores the context vector. Next are the banditron and dropout models, which use an epsilon greedy policy for exploration and Multi-Layer Perceptron (MLP) variants to mimic each arm as a neural model to exploit the best choice. Lastly, the RMS model is chosen based on the highest score predicted, i.e., it acts greedily for the current context. The HNB model, alongside state-of-the-art contextual bandit models, is evaluated on ten performance metrics—accuracy, specificity, recall, precision, prevalence, F-score, Matthews Correlation Coefficient, Fowlkes–Mallows Index, and Threat score. These metrics are crucial in informed decision-making, comprehensively assessing and optimizing targeted outcomes in diverse contexts.

Table 2. Confusion matrix

	Predicted class	
Actual class	True Positive (TP)	False Negative (FN)
	False Positive (FP)	True Negative (TN)

5.1 Accuracy and Regret

Let, $a^*(t)$ be the optimal arm which would yield the highest reward, $r_{a^*(t)}(t)$ and $r_{a(t)}(t)$ be the reward observed corresponding to the arm, $a(t)$ chosen by the contextual bandit model for the user context feature vector $x(t)$ Then, regret, as calculated in Eqn. (1), will be the cumulative sum of the difference between the maximum reward possible and the actual reward observed by the contextual bandit model over n trials, i.e.

$$regret(t) = r_{a^*(t)}(t) - r_{a(t)}(t) \mid x(t) \quad (1)$$

The overall goal is to minimize the regret. In applications such as an online advertisement model, the contextual bandit model renders a relevant advertisement, i.e. $a(t)$ to the user from the set of advertisements and if the user likes the chosen

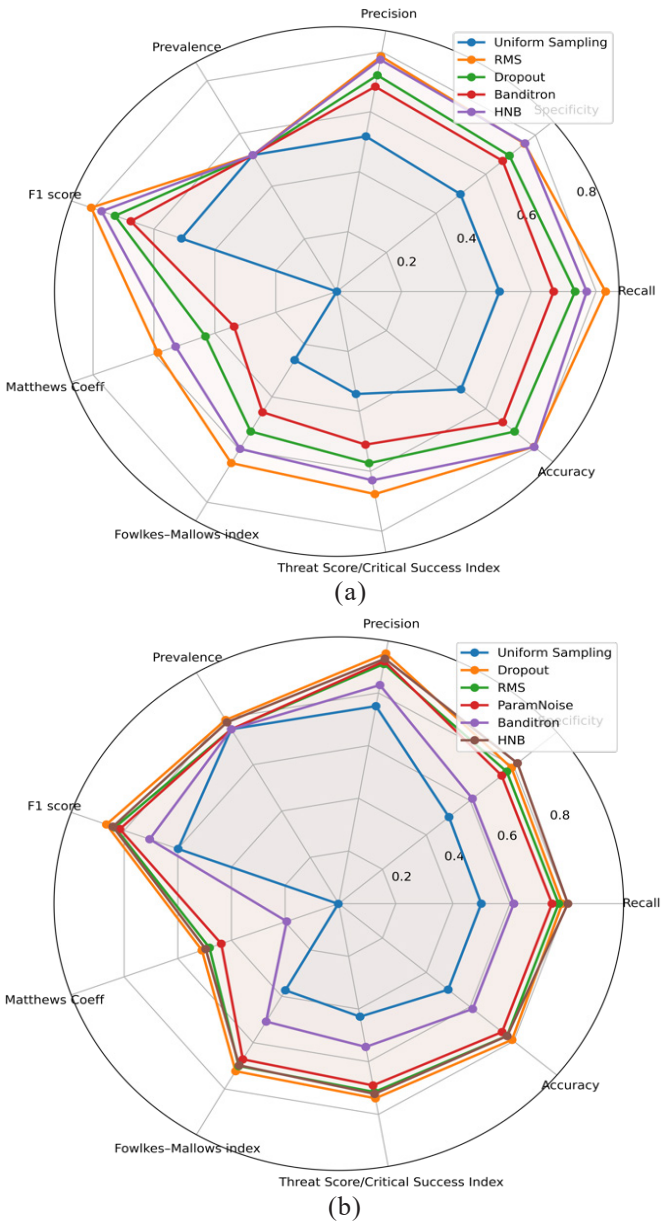

Figure 3. Performance metric analysis of various contextual multi-arm bandit algorithms on different datasets. (a) Mushroom dataset, and (b) Adult dataset.

advertisement, the model gets a reward, $r_{a(t)}(t)$ as +1 otherwise 0. The maximum reward possible for each trial will be 1. In such applications, where the reward is either 0 or 1 based on whether the user liked the choice given by the model or not, the accuracy of the contextual bandit model can be defined as the ratio of the total number of choices made by the contextual bandit model liked by the user, i.e. correct predictions to the total number of the choices made or trials. Accuracy provides a straightforward measure of predictive success; however, it has a downside in that it can be misleading in scenarios with imbalanced class distributions. The accuracy can be calculated using Eqn. (2). The proposed HNB model has performed significantly better than the uniform sampling, dropout and banditron model on all the datasets. However, the HNB model has comparable accuracy with the RMS model.

$$Accuracy = \frac{Correct\ predictions}{Total\ predictions} \quad (2)$$

5.2 Specificity

Specificity is a crucial performance metric that provides insights into a model's ability to identify true negatives correctly. It is calculated using Eqn. (3). Fig. 3 shows that the HNB model has the highest specificity among the compared CMAB models on all the datasets. This signifies that the HNB model predicts fewer false positives, demonstrating its precision in identifying instances of the negative class.

$$Specificity = \frac{Correctly\ Negatives\ Predicted}{Total\ Negative\ Data} = \frac{TN}{TN+FP} \quad (3)$$

5.3 Recall (Sensitivity)

Recall is the ratio of correct positives predicted over total positive data and can be successfully computed using Eqn. (4). The HNB model has the second highest recall value in the studied CMAB models, minimizing the risk of missing positive cases. Thus, the proposed HNB model captures and recognizes a high number of true positive instances among other models.

$$Recall = \frac{Correctly\ Positives\ Predicted}{Total\ Positive\ Data} = \frac{TP}{TP+FN} \quad (4)$$

5.4 Precision

Precision is defined as the ratio of correct positives predicted over total positives predicted, as shown in Eqn. (5), reflecting the precision of the model in identifying positive instances. The proposed HNB model has better precision value than various CMAB models like dropout, banditron and uniform sampling and is useful in multiple applications where false positives carry significant consequences.

$$Precision = \frac{Correctly\ Positives\ Predicted}{Total\ Positives\ Predicted} = \frac{TP}{TP+FP} \quad (5)$$

5.5 Prevalence

Prevalence, within the realm of performance metrics, denotes the proportion of positive instances within a dataset or population in binary classification, as depicted in Eqn. (6). It is a critical factor influencing the interpretation of model performance. In scenarios with significant class imbalances, where one class prevails, prevalence impacts the reliability of metrics like accuracy. The proposed HNB model scores high prevalence among its competing CMAB models, making it effective in real-world applications.

$$Prevalance = \frac{Total\ Positive\ Data}{Total\ Predictions} \quad (6)$$

5.6 F-Score

The F-score evaluates a model's performance by combining precision and recall and is valuable in scenarios where dataset balance is crucial, especially when one class is underrepresented. F-score is highly relevant in applications ranging from search engines to personalized recommendation systems, making it an essential performance metric. The F-score is calculated using Eqn. (7). From Fig 3, it can be observed that uniform sampling has the minimum, and the RMS model has the highest F-score among all the models on all the datasets.

$$F\ score = 2 * \left(\frac{Precision * Recall}{Precision + recall}\right) \quad (7)$$

5.7 Matthews Correlation Coefficient (MCC)

Unlike other metrics, such as accuracy or precision, which can be biased towards one class or another, MCC provides an overall measure of classification performance that is representative of both positive and negative classes. As shown in Fig. 4, the output range of the MCC index is from -1 to 1, and the more it goes near 1 indicates better performance; the more it is around 0, it shows the random performance of the model, and the more value is around -1, it shows the performance is even worse than 0. MCC is usually used for imbalanced datasets, where the number of values for one classification is much more than the other. Because the MCC considers both true positives and negatives, it is unaffected by class imbalance and provides a more accurate performance measure for such datasets. From Fig. 3, the Uniform Sampling model has an MCC value of zero. In contrast, the RMS, HNB, Dropout and Banditron models have an MCC value near 1 in descending order on all the datasets.

5.8 Fowlkes–Mallows Index (FM)

It is a statistical measure designed to assess the quality of clustering or classification algorithms by quantifying the similarity between two sets of labelled data. FM index is computed with the help of Eqn. (8), and the output range

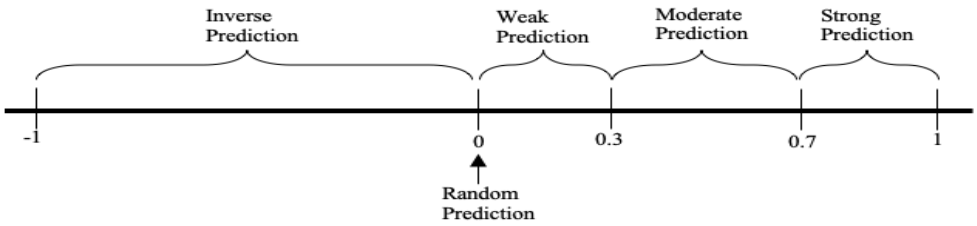

Figure 4. The working of the MCC index.

varies from 0 to 1, showing the similarity between the two sets. This makes FM particularly valuable in scenarios involving imbalanced datasets. Its use extends to applications where the identification of correctly clustered instances holds significance, such as image segmentation or biological data analysis. As an adaptable metric, the Fowlkes–Mallows index contributes to the ongoing refinement of clustering algorithms, aiding in selecting and optimizing methods based on their ability to group similar instances accurately.

$$FM = \frac{TP}{\sqrt{(TP+FP)*(TP+FP)}} \quad (8)$$

5.9 Critical Success Index (CSI) or Threat Score (TS)

CSI is an invaluable tool in decision-making processes and provides a holistic view of success within the context of predefined critical factors. CSI can be successfully calculated using Eqn. (9). A higher CSI value signifies a more successful outcome, while a lower value may indicate areas that require attention and improvement.

$$CSI = \frac{TP}{TP+FN+FP} \quad (9)$$

6. CONCLUSION

This paper proposes a novel Hybrid Neuro Bandit that consists of multiple multi-arm bandit experts; each individually produces a decision based on the user feature vector and is fused in the HNB decision agent model. The vector difference between the agent's and the user's decisions is backpropagated to the HNB model to minimize regret and achieve high accuracy. Also, the underlying mechanism of the proposed HNB model is easy for the user to understand. It can be successfully trained in offline and online settings for real-time applications where the rewards and data distribution can be dynamic. Further, a comparative study of the contextual bandit's models on nine performance metrics, namely accuracy, specificity, recall, precision, prevalence, F-score, FM index, MCC, and CSI, has been studied in detail on publicly available datasets. It is observed that none of the models has outperformed the other remaining models on different criteria in all the datasets. However, the HNB model has shown promising results on all the datasets.

REFERENCES

1. Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G. & Dean, J. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. 2017, arXiv preprint arXiv:1701.06538. doi: 10.48550/arXiv.1701.06538
2. Jordan, M.I. & Xu, L. Convergence results for the EM approach to mixtures of experts architectures. *Neural Netw.*, 1995, **8**(9), 1409-1431. doi:10.1016/0893-6080(95)00014-3
3. Rogova, G. Combining the results of several neural network classifiers. Classic works of the dempster-shafer theory of belief functions, 2008, **219**, 683-692. doi: 10.1007/978-3-540-44792-4_27
4. Jordan, M.I. & Jacobs, R.A. Hierarchical mixtures of experts and the EM algorithm. *Neural. Comput.*, 1994,**6**(2), 181-214. doi: 10.1162/neco.1994.6.2.181
5. Xu, L.; Krzyzak, A. & Suen, C.Y. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Trans. Syst. Man Cybern.*,1992, **22**(3), 418-435. doi: 10.1109/21.155943
6. Suen, C.Y. Nadal, C.; Legault, R.; Mai, T.A. & Lam, L. Computer recognition of unconstrained handwritten numerals. *In* Proceedings of the IEEE, 1992, 1162-1180,. doi: 10.1109/5.156477
7. Lacerda, A. Multi-objective ranked bandits for recommender systems. *Neurocomput*., 2017, **246**, 12-24. doi: 10.1016/j.neucom.2016.12.076
8. Robbins, H. Some aspects of the sequential design of experiments. *Bulletin of the Am. Mathematic. Soc.*, 1952, **58**(5), 527-535. doi: 10.1090/S0002-9904-1952-09620-8
9. Chapelle, O. & Li, L. An empirical evaluation of thompson sampling. *In* Proceedings of the 24th International Conference on Neural Information Processing Systems (NIPS'11), 2011, 2249–2257. doi: 10.5555/2986459.2986710
10. Agrawal, S. & Goyal, N. Analysis of thompson sampling for the multi-armed bandit problem. *In* Proceedings of the 25th Annual Conference on Learning Theory, Proceedings of Machine Learning Research, 39.1–39.26, June, 2012. doi: 10.48550/arXiv.1111.1797
11. Langford, J. & Zhang, T. Epoch-Greedy algorithm for multi-armed bandits with side information. *In* Proceedings of the 20th International Conference on Neural Information Processing Systems (NIPS'07), Curran Associates Inc., Red Hook, NY, USA, pp. 817–824,2007. doi: 10.5555/2981562.2981665
12. Tokic, M. Adaptive ε-greedy exploration in reinforcement learning based on value differences. *In* Annual Conference on Artificial Intelligence, Springer, Berlin, Heidelberg, 2010, 203-210. doi: 10.1007/978-3-642-16111-7_23
13. Li, L.; Chu, W.; Langford, J. & Schapire, R.E. A contextual-bandit approach to personalized news article recommendation. *In* Proceedings of the 19th International Conference on World wide web, 2010, 661-670. doi:10.1145/1772690.1772758
14. Chu, W.; Li, L.; Reyzin, L. & Schapire, R. Contextual bandits with linear payoff functions. *In* Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings, 2011, 208-214.
15. Ge, M.; Delgado-Battenfeld, C. & Jannach, D. Beyond accuracy: evaluating recommender systems by coverage and serendipity. *In* Proceedings of the fourth ACM conference on Recommender systems, Association for Computing Machinery, New York, NY, USA, 257–260. doi:10.1145/1864708.1864761
16. Vargas, S. & Castells, P. Rank and relevance in novelty and diversity metrics for recommender systems. *In* Proceedings of the fifth ACM conference on Recommender

systems, 2011, pp. 109–116.
doi:10.1145/2043932.2043955

17. Ribeiro, M.T.; Lacerda, A.; Veloso, A. & Ziviani, N. Pareto-efficient hybridization for multi-objective recommender systems. *In* Proceedings of the sixth ACM conference on Recommender systems, 2012, 19–26.
doi: 10.1145/2365952.2365962
18. Horsburgh, B.; Craw, S. & Massie, S. Learning pseudo-tags to augment sparse tagging in hybrid music recommender systems. *Artif. Intell.*, 2015, **219**, 25–39.
doi: 10.1016/j.artint.2014.11.004
19. Lacerda, A.; Veloso, A. & Ziviani, N. Exploratory and interactive daily deals recommendation. *In* Proceedings of the 7th ACM conference on Recommender systems. Association for Computing Machinery, NY, USA, 2013, 439–442.
doi: 10.1145/2507157.2507228
20. Drugan, M.M.; & Nowe, A. Designing multi-objective multi-armed bandits algorithms: a study. *In* International Joint Conference on Neural Networks (IJCNN), Dallas, TX, USA, 2013, 1-8.
doi: 10.1109/IJCNN.2013.6707036.
21. Yahyaa, S.Q.; Drugan, M.M. & Manderick, B. The scalarized multi-objective multi-armed bandit problem: an empirical study of its exploration vs. exploitation trade-off. *In* International Joint Conference on Neural Networks (IJCNN), Beijing, China, 2014, 2290-2297.
doi: 10.1109/IJCNN.2014.6889390
22. Yahyaa, S.; Drugan, M.M. & Manderick, B. Scalarized and pareto knowledge gradient for multi-objective multi-armed bandits. In Trans. Comput. Collect. Intell. XX, edited by N. T. Nguyen, R. Kowalczyk, B. Duval, J. van den Herik, S. Loiseau & J. Filipe, Lecture Notes in Computer Science, Springer, 2015. pp. 99-116.
doi: 10.1007/978-3-319-27543-7_5
23. Slivkins, A.; Radlinski, F. & Gollapudi, S. Ranked bandits in metric spaces: learning diverse rankings over large document collections. *J. Mach. Learn. Res.*, 2013,**14**(1), 399-436.
doi: 10.5555/2567709.2502595
24. Radlinski, F.; Kleinberg, R. & Joachims, T. Learning diverse rankings with multi-armed bandits. *In* Proceedings of the 25th International Conference on Machine Learning, pages, 2008, PP. 784–791.
doi: 10.1145/1390156.1390255
25. Vermorel, J. & Mohri, M. Multi-armed bandit algorithms and empirical evaluation. *In* European conference on machine learning. Springer, Berlin, Heidelberg, 2005, pp. 437-448.
doi: 10.1007/11564096_42
26. Tokic, M. Adaptive ε-greedy exploration in reinforcement learning based on value differences. *In* Annual Conference on Artificial Intelligence, Springer, Berlin, Heidelberg, 2010, pp. 203-210.
doi: 10.1007/978-3-642-16111-7_23
27. Li, L.; Chu, W.; Langford, J. & Schapire, R.E.A contextual-bandit approach to personalized news article recommendation. *In* Proceedings of the 19th International Conference on World wide web, 2010, pp. 661-670, .
doi:10.1145/1772690.1772758
28. Thompson, W.R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 1993,**25**(3/4), 285-294.
doi:10.2307/2332286
29. Chapelle, O. & Li, L. An empirical evaluation of thompson sampling. *In* Proceedings of the 24th International Conference on Neural Information Processing Systems, 2011, pp. 2249-2257.
doi:10.5555/2986459.2986710.

CONTRIBUTORS

Mr Nipun Bansal is an Assistant Professor in the Department of Computer Science at Delhi Technological University. He obtained his Master in Technology from IIIT Delhi.
In the current study, he designed and implemented the proposed Hybrid-Neuro Bandit model in the manuscript. He is responsible for drafting, editing and reviewing the complete manuscript.

Dr Manju Bala is working as an Associate Professor in the Department of Computer Science at I.P. University, Delhi University. She obtained her doctorate in Computer Science and Engineering from Jawaharlal Nehru University.
In the current study, she has provided guidance and mentorship to implement the Hybrid-Neuro Bandit model successfully.

Dr Kapil Sharma is a Professor in the Department of Information Technology at Delhi Technological University. He obtained his Doctoral Degree in the Faculty of Engineering and Technology at the M.D. University, Rohtak (Haryana), India.
In the current study, he has provided guidance and mentorship to design and implement a novel personalised contextual multi-arm bandit based model.