Methods to Evaluate Temporal Cognitive Biases in Machine Learning Prediction Models

Christopher G. Harris
School of Mathematical Sciences
University of Northern Colorado
Greeley, Colorado USA
christopher.harris@unco.edu

ABSTRACT

When asked to rank or rate a list of items, human assessors suffer from cognitive biases, which can make their decisions inconsistent over time. These inconsistencies often become part of machine learning prediction algorithms trained on human judgments, leading to misalignment and consequently affecting the metrics used to evaluate their correctness. In this paper, we build upon existing metrics to propose new statistics-based and decision support-based accuracy metrics, each of which addresses the varying nature of human judgment and is designed to evaluate the importance of decisions that change over time.

CCS CONCEPTS

• Computing methodologies → Machine learning

KEYWORDS

Data Science; Fairness; Machine Learning; Decision Making; Temporal Evaluation; Cognitive Bias;

ACM Reference format:

FirstName Surname, FirstName Surname, and FirstName Surname. 2018. Insert Your Title Here: Insert Subtitle Here. In *Proceedings of ACM Woodstock conference (WOODSTOCK'18). ACM, New York, NY, USA, 2 pages.* https://doi.org/10.1145/1234567890

1 Introduction

When humans make judgments or decisions, they frequently use *heuristic strategies* or decisional short cuts. These heuristics often lead to *cognitive biases*, which are systematic and predictable errors in judgment that result from a reliance on these heuristics. These cognitive biases often detract from optimal decision making. One example of a cognitive bias affecting human decision making is the *decoy effect* in which the choice between two options, A and B, is influenced when a third, relatively unattractive option, C, is introduced. [1]. When humans are presented with feedback or with new information on previously-made choices, this new information has a disproportionately large effect on their future choices which can lead to inconsistencies in decision making. Surprisingly, humans are not alone in this regard; the decoy effect has also been observed in animals, including ants, honeybees, and several bird species [2].

The decoy effect is not the only form of cognitive bias that affects human decision making [3]. Many other cognitive biases such as the *confirmation bias* (the tendency to focus on information that confirms already held preconceptions) [4] and the *anchoring effect* (where decision-makers rely too heavily on the first piece of information they receive) [5], can also be observed.

This human decision data, replete with biases, is used as inputs to train algorithms to model human-like decisions, from user recommendations, tasks involving reinforcement learning, and judgments that have a material impact on people's lives and well-being, such as in hiring, advertising, criminal justice, granting credit, personalized medicine, and targeted policymaking. In each case, the data derived from these decisions can change over time with new information: recommending movies can be affected by newly-released critics reviews, Academy Award nominations, or advice from friends (i.e., the *misinformation effect*) [6]; collaborative filtering on websites such as Amazon.com can reweigh product choices when new unrelated product options are introduced (i.e., the *decoy effect*); reinforcement learning is affected when choices between two options are inconsistent when made at different times; and pairwise preferences can be changed based on the current mood of a decision-maker (i.e., the *anchoring effect* or *confirmation bias*).

Each of these cognitive biases impacts decision-making models in different ways depending on the algorithm's purpose:

- Models that prioritize fairness use human decision data as inputs to train algorithms; however, this data typically does not account for the transitory nature of human decision-making.
- Models try to mimic human decisions to better understand human choices; however, they often fail to account for the inconsistent temporal nature of human preferences.

[*] Article Title Footnote needs to be captured as Title Note
[†] Author Footnote to be captured as Author Note

WOODSTOCK'18, June, 2018, El Paso, Texas USA

In this paper, we suggest changes to metrics that will allow us to measure the temporal aspects of human judgments. These metrics either measure statistical accuracy of a model's ability to choose a correct class through statistical accuracy (MAE, RMSE and correlation) or make correct decisions between classes (Precision, Recall, F- score, and nDCG). We also make suggestions about which metrics would be most appropriate to address which types of temporal cognitive biases.

2 Metrics for Statistical Accuracy

Statistical accuracy metrics are most suitable for recommender system predictions as well as other types of prediction models that involve classification (i.e., predicting the risk level of a potential borrower: high-risk, medium-risk, or low-risk). Consider the following 4-tuple, *TS*, which corresponds to a user, an item, the current time, and an ordinal rating for a specific product, respectively:

$$T \quad = \{(u, i, t, r_{u,i,t}) : u \in U, i \in I, t \in T, r_{u,i,t} \in R\} \quad (1)$$

In the case of a prediction, this 4-tuple would refer to the user, the category, the time, and the list of possible categories.

Two statistical measures evaluate the average deviation between an item's predicted rating or category ($\hat{r}_{u,i}$) and the user's true rating for the item or category ($r_{u,i}$), provided $U_T \quad = \{u : u \in T \quad \}$.

Mean Absolute Error (MAE) measures the average magnitude of the errors in a set of predictions without considering their direction. It's the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight.

$$M \quad = \frac{\sum_{r_{u,i} \in T} \mid \hat{r}_{u,i} - r_{u,i} \mid}{|T \quad |} \quad (2)$$

Root Mean Squared Error (RMSE) emphasizes large errors between predicted and true ratings:

$$R \quad = \sqrt{\frac{\sum_{r_{u,i} \in T} (\hat{r}_{u,i} - r_{u,i})^2}{|T \quad |}} \quad (3)$$

As their names imply, both MAE and RMSE measure error size; thus, lower values indicate better accuracy. Note that neither standard definition of MAE or RMSE incorporates time, *t*.

Correlation measures the agreement between multiple vectors of data. Correlation can be used to measure how well a given variable (i.e., recommended item or classification category) can be predicted using a linear function of a set of other variables.

A correlation coefficient, such as the Pearson product-moment correlation coefficient can be used with this purpose. With more than two variables being related to each other, the value of the coefficient of multiple correlations depends on the choice of dependent variable. Contrary to MAE and RSME, a higher correlation value indicates high predictability of the dependent variable from the independent variables; thus, higher values imply better accuracy.

In 2009, the authors in [6] posed a temporal accuracy metric based on RMSE, which they call time-averaged RMSE, consisting simply in the RMSE computed on ratings made until a point of time $T \quad _t$.

$$R \quad _t = \sqrt{\frac{\sum_{r_{u,i} \in T \quad _t} (\hat{r}_{u,i} - r_{u,i})^2}{|T \quad _t|}} \quad (4)$$

Where $T \quad _{t'}$ is the set of user ratings, decisions made until time t'.

$$T \quad _t = \{(u, i, t', r_{u,i,t'}) : u \in U, i \in I, t \in T, r_{u,i,t} \in R \land t' \leq t\} \quad (5)$$

Although their definition places an examination of temporal accuracy, $T \quad _t$, where the time exists only up until t', provided $0 \leq t' \leq t$. The definition of marginal temporal accuracy, $\triangle T \quad _t$ for RMSE could exist between any two points, t' and t":

$$\triangle T \quad _t = (T \quad _{t'} - T \quad _{t'}) \quad (6)$$

provided $0 \leq t' \leq t" \leq t$.

Similarly, MAE could be adjusted to a temporal accuracy $M \quad _t$:

$$M \quad _t = \frac{\sum_{r_{u,i} \in T \quad _t} \mid \hat{r}_{u,i} - r_{u,i} \mid}{|T \quad _t|} \quad (7)$$

For correlation, at different points in time, the change in the correlation value can be calculated. Because they are timestamped, influential factors on the independent variables can be temporally evaluated; changes that influence correlations would imply changes due to factors such as bias that occur between the two timestamps being evaluated. For example, if modeling a prediction algorithm for a film review for a user, changes in correlation could be due to an external event (e.g., Academy Award nomination for the film), internal event (i.e., the user is tired of watching movies of that genre), or unexplained events (e.g., noise).

3 Metrics for Decision-Support Accuracy

Decision support accuracy metrics take into consideration how many relevant items are chosen by the prediction model out of the total number available. Usually, when items are ranked, these metrics are more pertinent.

Precision, a measure of exactness, is defined as, at a depth of *k* in the ranked list, the ratio of total number of items evaluated by the system, $I(u)$, to the total number items the model marks as relevant, $R \quad _{I(u)}@k$. A common interpretation of relevant items in recommender systems has been "items that would be rated highly by user" [7]. Thus, for prediction models, relevance would be "items correctly classified" by the algorithm.

$$P \quad @k = \frac{R \quad _{I(u)}@k}{I(u)} \quad (8)$$

Examining precision could be done at different time stamps, *t*.

$$P\quad {}_t@k = \frac{R\quad {}_{I(u),t}@k}{I(u)} \tag{9}$$

Recall, a measure of completeness is defined as the ratio of total number of k items in the dataset that are relevant to the number of items marked by the model as relevant.

$$R\quad @k = \frac{R\quad {}_{I(u)}@k}{T\quad R\quad {}_{I(u)}} \tag{10}$$

Like precision, recall can be taken at different timestamps and the difference between recall scores at different times could then be calculated.

$$R\quad {}_t@k = \frac{R\quad {}_{I(u),t}@k}{T\quad R\quad {}_{I(u)}} \tag{11}$$

F-score or *F-measure* is the weighted harmonic mean of precision and recall.

$$F = \frac{2}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} \tag{12}$$

Where α is the normalized weight placed on precision. This gives us a single number to compare different models. By using $P\quad {}_t$ and $R\quad {}_t$ instead of *Prec* and *Rec*, we can evaluate different F-scores for different timestamps and compare how they change over time.

Normalized cumulative discounted gain (nDCG), measures the usefulness, or gain, of an item based on its position in the result list. The gain is accumulated from the top of the result list to the bottom, with the gain of each result discounted at lower ranks. nDCG evaluates against a perfectly-ranked list as a gold standard. As with precision and recall, this can be evaluated temporally to see how the rankings of items (products, candidates for a job position, etc.) compare to an ideally-ranked list over time. Like F-score, nDCG can also be calculated for different time stamps.

4 Evaluating Cognitive Biases

The primary purpose behind developing these temporal metrics is to evaluate cognitive biases in models (both human decisions and the corresponding trained algorithms) that have temporal components. Table 1 summarizes processes that may be used to evaluate measure different types of cognitive biases.

From Table 1, these biases can be measured using the temporal metrics described in Section 3 and 4. Determining the time window will require more trial and error and will depend on whether we are assessing human decisions or decisions made by algorithms as our model.

Table 1. Evaluation and Measurement of Cognitive Biases

Cognitive Bias	**Evaluation**	**Measurement**
Decoy effect	Examine the relative ratings/rankings made for a pair or series of items by a model at different points in time	Measure the effects of a decoy item on a decision between two items by subtracting the difference between a baseline model and one that offers a decoy.
Anchoring effect	Examine if a starting point of rating/ranking, either with the best choices or the worst choices, affect subsequent decisions within a time window	Introduce a low (high) rated item and see if this deflates (inflates) the ranking or rating of items within a specific time window.
Hindsight bias	Provide feedback to users after they make choices. Examine if this feedback increases or decreases the variance of future choices within a window of time	Evaluate the difference in accuracy within a time window immediately after the feedback. Compare this to models where feedback is not provided
Confirmation bias	Conduct a pre-test (or conduct a test on similar items) and determine the model's previously held beliefs. Determine how these beliefs shape future choices	Evaluate the difference in accuracy between items where the answer is clear from ones similar to those items within a window of time. Also, for humans, perform 'think aloud' methods of understanding the thought process

5 Remaining Questions

Cognitive biases are challenging to detect and measure. The following are questions that remain regarding

1. Are there better approaches to detect temporal biases other than the ones mentioned here?
2. How to determine the appropriate time window for measuring each type of bias?
3. Temporal cognitive biases may decrease over time. Is there a survival function that can model the decay of bias over time?
4. Is examining bias in an algorithm's decision any different from evaluating biases in a decision made by a human?
5. Cognitive biases are often complex; how can we best isolate any changes in these biases to simple, easy to understand model?

6 Summary

Detecting and evaluating cognitive biases in humans is challenging; in algorithms that involve complex interplay between numerous features is even more problematic. The metrics described in this paper provide a basis for measuring cognitive biases that may arise because a model is trained on data from human judgments. The recent focus on fairness, accountability and transparency in machine learning illustrates the growing desire for researchers in identifying these biases. There are numerous questions that need to be considered to design an appropriate study.

Cognitive biases are a result of humans applying heuristics, or decisional short cuts. In most cases, we want our algorithms to avoid these cognitive biases, but in some situations, we do not want to remove the cognitive biases, but instead identify and measure these biases in developing prediction models; for example, when algorithms are designed to mimic human behavior and decisions.

Formulating the right approach to measuring biases is essential to study design. In future work, we plan to design a series of experiments around the detection and measurement of biases in human judgment data and observe how these become propagated through the machine algorithms trained on this data.

REFERENCES

[1] Huber, J., Payne, J. W., & Puto, C. (1982). Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. Journal of consumer research, 9(1), 90-98.

[2] Kahneman, D., Slovic, S. P., Slovic, P., & Tversky, A. (Eds.). (1982). Judgment under uncertainty: Heuristics and biases. Cambridge university press.

[3] Kahneman, D. 2011. Thinking, fast and slow. New York, NY: Farrar, Straus, and Giroux.

[4] Pohl, R., & Pohl, R. F. (Eds.). (2004). Cognitive illusions: A handbook on fallacies and biases in thinking, judgment and memory. Psychology Press.

[5] Wilson, T. D., Houston, C. E., Etling, K. M., & Brekke, N. (1996). A new look at anchoring effects: basic anchoring and its antecedents. Journal of Experimental Psychology: General, 125(4), 387.

[6] Lathia, N., Hailes, S., Capra, L., & Amatriain, X. (2010, July). Temporal diversity in recommender systems. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval (pp. 210-217).

[7] Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE transactions on knowledge and data engineering, 17(6), 734-749.