Evolution of Bias in Human and Machine Learning Algorithm Interaction

Wenlong Sun[1], Advisor: Olfa Nasraoui[1], Patrick Shafto[2]
[1] Knowledge Discovery & Web Mining Lab, University of Louisville
[2] Depts. Mathematics and Computer Science, Rutgers University - Newark

Introduction

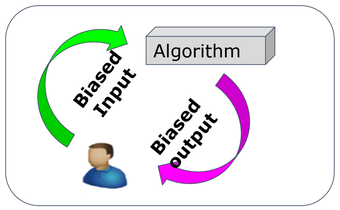
- **Human algorithm interaction:**

people are now affected by the output of all types of machine learning algorithms

  - social media, blogs, social networks, and other services and applications.

- **Motivation**
  - ML algorithm relied on reliable labels from experts to build prediction.
    - However, ML algorithm started to receive data from the more general population.

•The interaction leads to biased result which is caused by ingesting unchecked information from general population, such as biased samples and biased labels.



Objectives

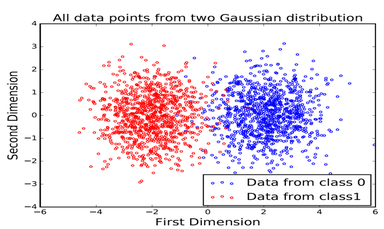
•Develop an iterated-learning framework to study the interaction between machine learning algorithms and users.

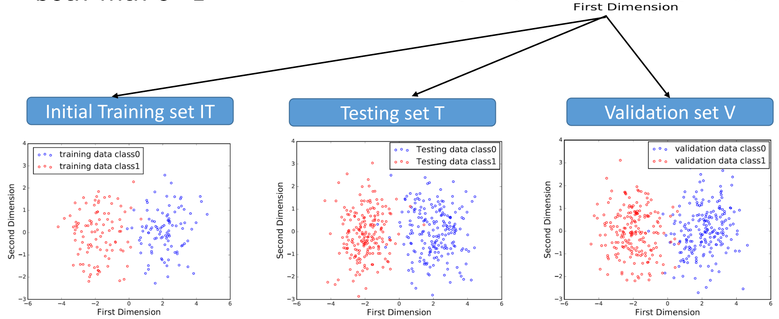
- The process by which people select information to label.
- The process by which an algorithm selects the subset information to present to people.

• Develop different types of metrics to measure the impact of interaction between machine learning algorithms and humans

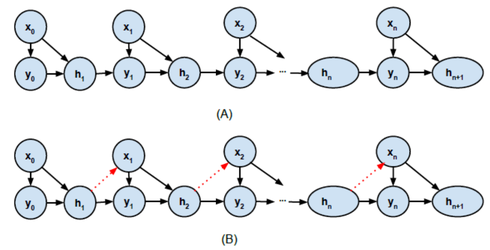
Data

- Synthetic data ({*x1,x2*}, *y*):
  - 2 Gaussian distributions
- class y = 0 and 1
  - non-relevant and relevant
- Each class contained 1000 points
  - centered at (-2,0) and (2,0)
  - both with σ=1





Methodology



- **Human-Algorithm Interaction Mechanism:**

In simple iterated learning the next data input **x is independent of the previous inferred hypothesis (model) h** [1]. In our proposed learning framework, the next input Xn+1 depends on the pervious hypothesis, hn.

$$p(\mathbf{x}_n|h_n) = (1-\epsilon)p_{seen}(\mathbf{x}_n|h_n) + \epsilon q(\mathbf{x}_n)$$

$q(\mathbf{x}_n)$: The prior distribution of *x*
ϵ : The weight between prior and algorithm controlled probability
h_n : Current hypothesis learned
$p_{seen}(\mathbf{x}_n|h_n)$: Probability of being seen given current learned model

- **Iterated Filter Bias**

$$p_{seen}(\mathbf{x}_n|h_n) = \frac{p(\mathbf{y}_n = 1|\mathbf{x}_n, h_n)}{\sum_{\mathbf{x}_i} p(\mathbf{y}_i = 1|\mathbf{x}_i, h_n)}$$

- **Iterated Active-learning Bias**

$$p_{active}(\mathbf{x}|h) \propto 1 - p(\hat{\mathbf{y}}|\mathbf{x}, h)$$

$$\hat{\mathbf{y}} = \arg\max_y (p(y|\mathbf{x}, h))$$

$\hat{y}$: the predicted y value. That is, x values that are selected are the least certain about $\hat{y}$.

- **Metric Used**
  - **Blind spot:** Data at risk to be hidden or not visible to the user

$$\mathbf{D}_\delta^{\mathbf{F}} = \{\mathbf{x} \mid p_{seen}(\mathbf{x}|h) < \delta\}$$

  - **Gini Coefficient:** Used to study distribution of a data set, most commonly used measure of inequality.

$$G = \left(\frac{\sum_{i=1}^{n}(2i - n - 1)p_i}{n\sum_{i=1}^{n} p_i}\right)$$

  - **Boundary shifts:** indicates how different biases affect algorithm performance

$$b = \sum_{i=1}^{N} L_i$$

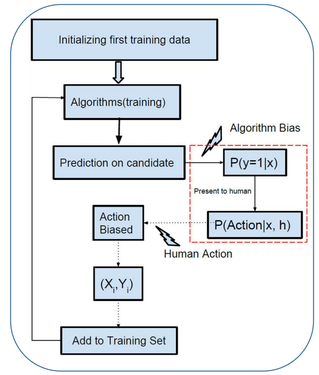
p_i : value of the sorted target array

- N = number of testing instances. L_i = predicted label of specific class

Acknowledgements

This research is supported by the National Science Foundation through grant NSF INSPIRE (IIS)- Grant #1549981.

Results



Experiment Steps:

1. Select initial training data (*Training*$_0$)
   a. using randomly selected instances
   b. using class-imbalanced selection (e.g. more instances from relevant class : *y=1*)
2. Train a ML model (*Model*$_{init}$)
   a. Naive Bayes classifier
   b. Others: Logistic regression, etc
3. Apply the learned model to candidate set $\mathbb{C}$ to get prediction probability
4. Use iterated bias to select unlabeled instances x_n to present to user
5. Update training set with human action (label y_n) (*Training*$_{n+1}$ = *Training*$_n$ ∪ {(x_n,y_n)}
6. Use new data to update model (*Model*$_{n+1}$)
7. Repeat 3-6 until maximum iteration

Research Questions	Answer	Filter Bias P-value	Active Learning P-value	Random Selection P-vale
RQ 1: Do Different iterated bias modes have different effects on the boundary shifting at significance level 0.05 ?	Yes	0.00	0.84	0.99
RQ 2: Do different biases lead to different trends of Gini Coefficient during iterations given the same initialization?	Yes	1.94e-7	7.29e-7	4.0e-25
RQ 3: Does the iterated bias affect the size of the class 1-blind spot and the all-classes-blind spot, i.e. is the initial size of the blind spot significantly different compared to its size in the final iteration?	Yes	5.4e-8	0.05	0.42
RQ 4: Does Initialization bias affect the boundary learned during iterative learning given a fixed iterated bias model?	Yes	6.6e-9	7.3e-16	3.6e-14
RQ 5: Does human action (labeling data when requested to by the machine learning algorithm) affect the boundary shift?	Yes	1.0e-5	0.08	0.75

Conclusion

- Develop a theoretical and simulation framework for studying bias evolution in interactive learning.
- Extreme filtering affects number of items which can be seen by users.
- More heterogeneity of predicted relevance ⇒ more inequality between predicted relevance of different items.
- Significant impact of extreme filtering on the number of items that can be seen by the user within iterated human machine-learning interaction.
- More frequent human action ⇒ more significant effect on the boundary shift.

Reference

[1] Kirby, Simon, Tom Griffiths, and Kenny Smith. "Iterated learning and the evolution of language." *Current opinion in neurobiology* 28 (2014): 108-114.