Numerical and Computational Methods in Sciences & Engineering
An International Journal

http://dx.doi.org/10.18576/ncmse/020205

Adaptation of machine learning based fairness algorithm for real time decision in autonomous systems

I. A. Sulaimon[*], *A. Ghoneim and M. Alrashoud*

Department of Software Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia

Received: 23 Feb. 2020, Revised: 6 April 2020, Accepted: 17 April 2020
Published online: 1 July 2020

Abstract: Algorithmic bias has been a focus of research for many data scientist and machine learning researchers, but little research efforts have been dedicated to algorithmic bias in autonomous systems. The dynamic nature of autonomous systems makes it difficult to analyze for biases, hence we focus our research effort on the control loop of ML based autonomous systems. In this research, we adapted a machine learning based fairness algorithm designed for decision support systems in a real time and dynamic environment of an autonomous system. The final solution is in the form of a software module which provides access for auditing decision process of machine learning powered autonomous software systems. This, in turn, ensures fairness in the decision process of autonomous software systems.

Keywords: Software Engineering, Algorithmic-Bias (Algorithmic discrimination, Algorithmic Fairness), Artificial Intelligence (AI), Autonomous System, Machine Learning (ML), Reinforcement Learning.

1 Introduction

Biasness is to treat subjects differently, while fairness is to treat subjects similarly regardless of their defined protected attributes such as race, gender, social status, age etc. [1,2,3,4,5]. (Unfair) biases are based on classification, not individual merit. Human decisions include objective and subjective elements. Hence, they can discriminate. Algorithmic inputs include only objective elements, so it is ideal for algorithms not to discriminate. While human biases seem unavoidable, algorithmic bias should be controllable. Algorithmic biases in autonomous software systems originate either from data used in training a machine learning (ML) algorithm or the algorithm itself [6].

The main goal in autonomous system design is to construct algorithms that can predict certain target output. Hence, the learning algorithm is trained with some example dataset containing the intended relationships between the input and output values. The machine learning algorithm is expected to generate a correct output; despite been presented with similar inputs that have not been encountered during training.

For many obvious unfavorable reasons, AI systems have been making news headline for being bias. Recently in late October 2017, Facebook apologized after an error in its machine-translation service that saw Israeli police arrest a Palestinian man for posting an Arabic statement that translates to ”good morning” on his social media profile [4]. AI systems tend to associate some words with a specific group of subjects unfairly, FaceApp apologizes for building a racist AI when her hotness feature associate being hot to being white [5]. When a user taps the hotness button on FaceApp so as make a picture hot, the application makes the face in the picture white. Thereby, creating an impression that only a white face is hot regardless of other skin tones.

Bryson et.al [3] warned, ” AI has the potential to reinforce existing biases because, unlike humans, algorithms may be unequipped to consciously counteract learned biases”. In accordance with [6] Results for ”CEO” in Google Images: 11% female US CEOs, whereas the US has 27% female CEOs. Also, in Google Images, ”doctors” are mostly male, ”nurses” are mostly female. These data are not a true reflection of our reality. COMPAS [7] Prediction accuracy of recidivism errs by being too lenient with whites and too harsh with blacks: Blacks that did not re-offend were classified as high risk twice

* Corresponding author e-mail: engs.iunits@gmail.com & 437106948@student.ksu.edu.sa

as much as whites that did not re-offend. Whites who did re-offend were classified as low risk twice as much as blacks who did re-offend.

To a software engineer, biases in autonomous software systems are a kind of systematic errors, which are difficult to identify at development. Engineering an unbiased autonomous software system is a huge challenge to software engineers due to the dynamic nature of autonomic computing. Hence, we desire a software system which enable fairness audit of decision algorithms found in autonomous systems.

Here, we proposed a reinforcement learning based framework for algorithmic bias detection in ML powered autonomous software systems. We also proposed a modified MAPE-K [8] control loop which allows fairness check in decisions of autonomous software systems. The proposed model will enable software engineers to build quality autonomous software systems with consideration for fairness in the decisions made by such a system.

2 Related Works

Numerous researches have contributed to ensuring fairness in AI in general. Some contributions aimed at detecting or measuring biases in machine learning algorithms, while others focused on bias mitigation in machine learning algorithms. Few kinds of researches have also worked on both bias detection and mitigation.

2.1 Algorithmic Bias in AI

As a general belief, there is no one-size-fits-all to the concept of bias in machine learning (ML) algorithms.

Hence, each research aimed at ensuring fairness in ML algorithms is based on individual perspectives and definitions of fairness and bias. A section of AI fairness research worked on ensuring fairness through the preprocessing stage of ML pipeline by optimizing for fairness in data used to train the ML algorithm [6,9,10,11,12,13,14]. Another section prefers to ensure fairness through the in-process stage by optimizing the algorithm itself [15,16,17]. Other researches focused on post-processing by resolving bias in ML algorithms through optimization of the decisions made by such algorithm [18, 19].

There are varying notions of fairness in AI based on different perspectives of researchers. In [20], the authors introduced a comprehensive review of various notion of fairness based on researches so far. Fairness notions were classified into three groups with each having numerous subgroups.

Statistical Measures explains the statistical notion of fairness which forms the basis for other advanced notion of fairness in AI. Unlike Statistical measures of fairness which focused mainly on the sensitive attributes of the classified object, Similarity-Based Measures defined fairness considering other attributes of the classified object and how those attributes relate to the sensitive attributes. Causal reasoning defined fairness based on the relations between attributes and their influence on the outcome of a decision algorithm.

For the purpose of this research, the Similarity-Based Measures of fairness is adopted as it is well suited to the nature of decisions made in ML powered autonomous systems.

2.2 Algorithmic Bias detection and mitigation

Algorithm bias detection and mitigation have been challenging due to different notions of fairness in AI. Some AI fairness research community focused on bias detection and the mitigation process is left to the user to decide based on the insight provided by the detection tools [6,18,9,10,15,21,22,16,12]. Other researches [11,13,17,14,23] focused on both detection and mitigation by suggesting likely optimizations to the training data, algorithm, algorithmic prediction.

For the purpose of ML powered autonomous software systems, we adapt FairML [11] - one of the existing proven methods for detecting fairness in ML algorithms.

3 Proposed framework

Figure 1 below shows the Proposed Control Loop which is an adaptation of the traditional MAPE-K control loop for autonomous systems [8]. The MAPE-K framework comprises of a sensor, effector and four execution phases which Monitor, Analyze, Plan, Execute and a Knowledge module which houses the knowledge base. This framework forms the basis on which most self-adaptive system operates. The Sensor receives information/data from the environment, while the Monitor phase keeps a close watch on changes in the sensor data. The Monitor phase passes the data for analysis to the

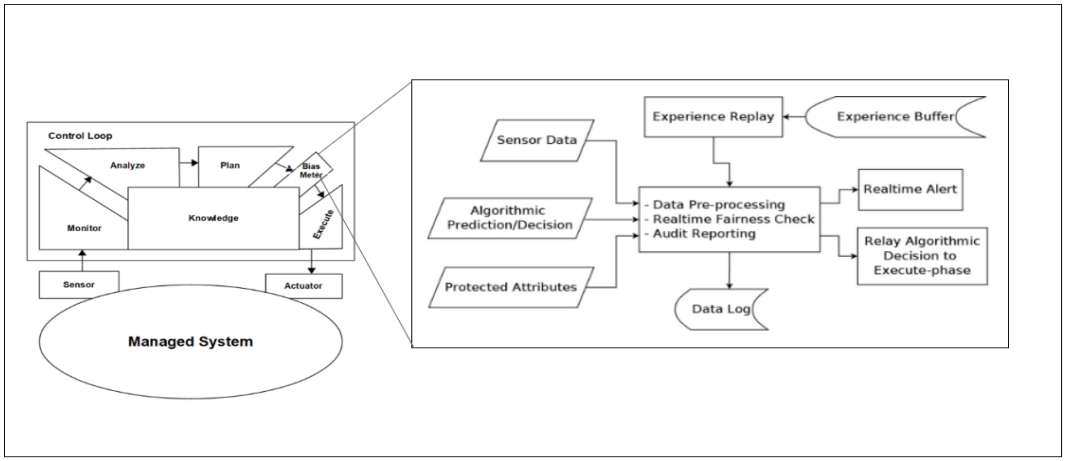

Figure 1: Machine Learning based Fairness algorithm for real time decision in autonomous systems

Analyze phase. This, in turn, relays the analyzed data to the Plan phase, while the Plan phase creates the plan for execution and relay execution commands to the Execute phase. All these process phases interact directly with the Knowledge-base. The proposed control loop has an extra phase (Bias-meter) between the Plan-phase and Execute-phase of the traditional control loop.

The enlarge bias-meter in Figure 1 gives an abstract overview of the proposed model. Bias-meter takes the following as input data;

1.Sensor data
2.Algorithmic predictions
3.Protected attribute
4.Result of an external process which performs experience replay based on the stored experience in the knowledge-base.

We used Q-learning with Combined Experience Replay (CER) to present dataset of past experience for fairness check. The experience replay technique we used specifically search for experiences with similar states variables. The purpose of experience replay is to ensure the fairness checking algorithm have access to appropriate data from past experience of the self-driving car algorithm.

After receiving the input data Bias-meter performs three main process steps below;

–Data Pre-processing
–Runtime fairness check
–Audit reporting

3.1 Data Pre-processing

All input data from the Plan-phase and experience buffer are passed through a data pre-processing stage to keep the data in a format compatible with the fairness detection algorithm. Sensor data, Algorithmic prediction and Experiences (documented through experience replay) were combined and discretized to suit input format of the fairness checking algorithm.

3.2 Runtime Fairness Check

For fairness check, we used an adapted form of FairML [11] in determining the relatedness of protected attributes to the decisions made by the self-driving car algorithm in a crash situation. Through this, our system derives insight in determining how bias the algorithm is based on the protected attributes. FairML does orthogonal feature transformation for all the features of the dataset. Our adaption of the FairML only does the orthogonal feature projection using the protected features of the dataset and consideration of the relationship that exist between the features of the dataset. We

Table 1: M-FairML performance evaluation

	Number of Attributes	Protected Attributes	Size	FairML Average Time (secs)	M-FairML Average Time (secs)	Average Time Difference (%)
Propublica Recidivism Dataset	13	3	6172	122.27	37.12	69.64
Adult Census Income Dataset	100	3	4999	3255.16	225.16	39.08
German Credit Score Dataset	60	2	1000	32.73	3.89	88.12

considered the relationship between the features so as to augment for accuracy in the estimation of predictive dependency of the black-box algorithm on each dataset feature.

Our choice of FairML as a fairness checking algorithm is due to flexibility of implementation and its consistency and robustness even in the presence of noise in the data. FairML used Iterative Orthogonal Feature Projection (IOFP) for measuring the relatedness of each attribute of a dataset to the decisions an algorithm made from it. This suits our notion of fairness for the research. The relatedness score generated by our modified FairML is used in our proposed framework to determine likely bias in the crash algorithm decision under consideration.

3.3 Audit Reporting

User-friendly alerts are raised at runtime if any algorithmic bias is detected. The output of the fairness check is also logged.

Finally, the algorithmic decision is relayed to the execution phase of the control loop.

4 Analysis and Discussion

In order to check the effectiveness of the modified fairness checking algorithm (M-FairML), we evaluate the algorithm with three different datasets of moderate sizes on varying level of noise. Basically, M-FairML is tested against similar dataset used in testing the original version of the fairness audit algorithm (FairML). No significant change in accuracy is noticed while execution time is greatly reduced when viewed under varying level of noise. Noise was introduced into each of the dataset by generating random noise data and appending it to the dataset. Due to the inaccessibility of the black-box algorithms used in making the predictions for the ML systems in consideration, we used common predictive algorithms such as Linear Logistic Regression, Non-Linear Logistic Regression, Random Forest, Neural Network, Linear Support Vector Machine (SVM) and Gaussian Process. These predictive algorithms have varying features and speed of execution as noticed with each dataset under consideration.

The result of the performance evaluation reported in Table 1 shows that M-FairML ensures more than sixty nine percent (69%) reduction in execution time across all the datasets used. In comparison with FairML, M-FairML report no significant difference in accuracy. This is noticed through the predictive dependency scores generated for each of the datasets. FairML and M-FairML reported the same predictive dependency scores for each of the protected attributes in the datasets considered.

5 Conclusion

We proposed a modified MAPE-K control loop which allows fairness check in decisions of autonomous software systems, while we adopt the experience replay mechanism of reinforcement learning to ensure fairness in decisions made by machine learning powered autonomous systems. Our proposed approach is novel. Hence, there is a need for further work on the proposed model. Firstly, the reliability of our model needs to be checked in larger autonomous systems. Secondly, there will be evaluation in comparison with similar frameworks which are yet to be discovered. Lastly, the model needs to be tested on a real autonomous system with near real accident scenarios.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

References

[1] K. Kirkpatrick, Battling Algorithmic Bias, *Communications of the ACM*, 59 (2016) 16-17.
[2] N. Byrnes, Are machine learning algorithms biased?" *MIT Technology Review, 2017*.
[3] A. Caliskan, J. Bryson and A. Narayanan, Semantics derived automatically from language corpora contain human-like biases, *Science*, 356 (2017) 183-186.
[4] A. Hern, Facebook translates 'good morning' into 'attack them', leading to arrest", *the Guardian*, 2018.
[5] FaceApp apologizes for building a racist AI, *TechCrunch*, 2018. https://techcrunch.com/2017/04/25/faceapp-apologises-for-building-a-racist-ai/.
[6] S. Hajian, F. Bonchi and C. Castillo, Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining, In 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 2125-2126.
[7] S. Jeff Larson, How We Analyzed the COMPAS Recidivism Algorithm - ProPublica, *ProPublica*, 2016. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.
[8] An architectural blueprint for autonomic computing, *A white paper* 31 (2006) 1-6. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.150.1011&rep=rep1&type=pdf
[9] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork., *Learning Fair Representations*, In 30th International Conference on International Conference on Machine Learning, 2013, pp. 325-333.
[10] D. Pedreschi, S. Ruggieri, and F. Turini, *A study of top-k measures for discrimination discovery*, In 27th Annual ACM Symposium on Applied Computing, 2012, pp. 126-131.
[11] J. A. Adebayo, *FairML: Toolbox for diagnosing bias in predictive modeling*, Master's thesis, Massachusetts Institute of Technology, 2016. https://dspace.mit.edu/handle/1721.1/108212.
[12] N. Bantilan, Themis-ml: A Fairness-Aware Machine Learning Interface for End-To-End Discrimination Discovery and Mitigation, *Journal of Technology in Human Services*, 36 (2018) 15-30.
[13] T. Calders and S. Verwer, Three naive Bayes approaches for discrimination-free classification, *Data Mining and Knowledge Discovery*, 21 (2010) 277-292.
[14] S. Friedler, C. Scheidegger and S. Venkatasubramanian, On the (im)possibility of fairness", *arXiv.org*, 2018. https://arxiv.org/abs/1609.07236.
[15] Y. Alufaisan, M. Kantarcioglu and Y. Zhou, *Detecting Discrimination in a Black-Box Classifier*, 2016 IEEE 2nd International Conference on Collaboration and Internet Computing (CIC), Pittsburgh, PA, 2016, pp. 329-338.
[16] M. Zehlike, et.al. *FAIR: A Fair Top-K Ranking Algorithm*. In ACM on Conference on Information and Knowledge Management - CIKM '17, 2017, pp. 1569-1578.
[17] F. Calmon, D. Wei, B. Vinzamuri, K. Ramamurthy and K. Varshney, Data Pre-Processing for Discrimination Prevention: Information-Theoretic Optimization and Analysis, *IEEE Journal of Selected Topics in Signal Processing*, 12 (2018) 1106-1119.
[18] S. Galhotra, Y. Brun, and A. Meliou, *Fairness Testing: Testing Software for Discrimination*, In 2017 11th Joint Meeting on Foundations of Software Engineering, 2017, pp. 498-510.
[19] M. Hardt, E. Price and N. Srebro, Equality of Opportunity in Supervised Learning", *arXiv.org*, 2018. https://arxiv.org/abs/1610.02413.
[20] S. Verma and J. Rubin, *Fairness Definitions Explained*, IEEE/ACM International Workshop on Software Fairness (FairWare), Gothenburg, 2018, pp. 1-7.
[21] F. Tramer et al., *FairTest: Discovering Unwarranted Associations in Data-Driven Applications*, 2017 IEEE European Symposium on Security and Privacy (EuroS&P), Paris, 2017, pp. 401-416.
[22] M. Raginsky, A. Rakhlin, Matthew Tsao, Y. Wu and Aolin Xu, *Information-theoretic analysis of stability and bias of learning algorithms*, 2016 IEEE Information Theory Workshop (ITW), Cambridge, 2016, pp. 26-30. doi:10.1109/ITW.2016.7606789
[23] S. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. Hamilton and D. Roth, A comparative study of fairness-enhancing interventions in machine learning, *arXiv.org*, 2018. https://arxiv.org/abs/1802.04422.