Evaluating the Efficacy of Influence Functions in Identifying Mislabeled Census & Medical Data

Jia'Ao (Jason) Sun
University of Toronto
jsun@cs.toronto.edu

Shalmali Joshi
Vector Institute
shalmali@vectorinstitute.ai

Marzyeh Ghassemi
University of Toronto, Vector Institute
marzyeh@cs.toronto.edu

Abstract

Influence functions are a classic technique from robust statistics that can be used to trace backwards a model's predictions to its training data to identify training points most impactful towards a model's predictions. Influence functions have been previously investigated to be a means of identifying mislabeled data. In this paper, we further test the ability of the influence functions previously demonstrated by Koh et al. as a credible method of mislabel identification. We experiment further on multiple census and medical datasets, and show that while influence functions can accurately approximate leave-one-out retraining on various types of datasets, they are not a viable tool for identifying mislabeled minority groups or mislabeled data in general.

1 Introduction

Machine learning (ML) advancements have transformative potential to improving the efficiency and efficacy of medical treatment. Intelligent medical systems have applications on all three pillars of healthcare: prevention, diagnosis, and treatment [1]. Implementing powerful machine learning methods into medical systems can help integrate smart technology into the healthcare environment. ML algorithms have previously shown capabilities to solve diagnostic and prognostic problems at a relatively high accuracy in a variety of medical domains. However, achieving a level of artificial intelligence capable of facilitating and effectively enhancing the work of medical experts requires both state-of-the-art machine learning algorithms, as well as large quantities of high-quality data. Medical data is already hard to obtain due to patient confidentiality. Datasets, especially those of the medical nature, are also often flawed due to incompletion (missing parameter values), inaccuracy (systematic or random noise in the data), and/or sparsity (few and/or non-representable patient records available) [2]. Furthermore, datasets with minority groups can cause machine learning models to develop discriminatory biases and exhibit reduced performance on historically disadvantaged groups [3]. These reasons, among others, are factors that impede current progression in machine learning in healthcare.

Our goal in this paper is three-fold: 1) assess whether influence functions can accurately approximate iterative influence calculations in leave-one-out retraining in various real-life datasets (census and medical), 2) evaluate the extent of influence in which protected groups such as females and black people exhibit on a trained model, and 3) investigate the ability of influence functions to identify purposely mislabeled protected groups in real sets of data.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

The results of this investigation indicate that while influence functions proposed by Koh et al. can accurately approximate leave-one-out retraining and is useful for identifying mislabeled samples for some datasets, it is not a viable tool to be used for all datasets [4]. For clinical data, mislabeling errors can be correlated with marginalized identity. We demonstrate that in such cases, label correction does not perform well on minority samples as they are not as highly influential as expected.

2 Related Work

Koh et al. presented influence functions as a method to identify training points that are most responsible for a model's given prediction, which can be successfully applied for multiple purposes [4]. Koh et al. showed that in the case of email spam classification, influence functions effectively identified an abundance of mislabeled training data as highest priority in order to be checked for mislabeling. In that situation, fixing those mislabeled data helped repair the dataset and improve the accuracy of the model. In this work, the use of influence functions is investigated as a method of detecting and correcting mislabeled data in real-life census and medical datasets. Three different datasets are tested using influence functions and we find that contrary to Koh et al.'s results, influence functions do not work as well on all datasets and are not viable tools of mislabeling correction. Furthermore, we discover that minority groups in the data do not influence the model any more than majority groups.

3 Data

Adult Census Dataset

The Adult Census Dataset [5] contains 48,842 instances of 14 attributes such as age, education, race, sex, etc., of adult census information. The prediction task is to determine whether a person's income exceeds $50,000 per year. In data preprocessing, we drop all data points that contain missing values.

Diabetes Dataset

The Diabetes Dataset [6] comprises of 101,766 hospital admissions from patients with diabetes from 130 U.S. hospitals between the years 1999-2008. Data points contain 50 features representing patient and hospital outcomes. The prediction task is whether a patient with diabetes will be readmitted to the hospital within 30 days. Data points with missing values in attributes race, diagnosis 1, diagnosis 2, and diagnosis 3 are dropped because these are key attributes to model training. We drop encounter ID and patient number as they are unique identifier values that have no effect on the model. We drop weight, payer code, and medical specialty because they are missing for more than half of the data. Patients readmitted in less than 30 days ("<30") and patients readmitted in more than 30 days (">30") are collectively relabeled as "readmitted" to differentiate them from patients who are not readmitted to binarize classification.

Heart Dataset

The Heart Disease Dataset [7] contains data from 303 patients. There are 76 attributes, but a subset of 14 of them are used, which is typical for machine learning research using this dataset. The prediction task is to determine the presence of heart disease in a patient. There are no missing values.

Categorical attributes from the three datasets are one-hot-encoded and combined with their numerical attributes to be normalized to a 0-1 scale.

4 Methods

Influence functions are a classic technique from robust statistics that measure how much a model's parameters change if one of its training points is upweighted by an infinitesimal amount [4, 8]. This method can be used to trace a model's predictions through the learning algorithm and back to its training data to identify training points that are the most responsible for a given prediction. Koh et al. demonstrated that this can further be used to help identify mislabeled data by ranking training points based on their influence to the model, such that points with higher mislabel potential have

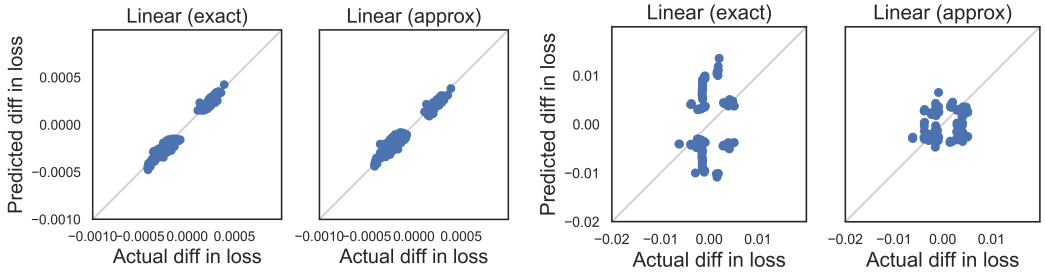
higher influence. The hindrance to influence functions however, is that they require expensive second derivative calculations, so they would not scale well with large sets of data. To address this issue, Koh et al. used second-order optimization techniques, which showed to be an efficient and accurate method to approximate influence calculations.

We use the same influence approximations as Koh et al. in this paper. We define $\sigma(t) = 1/(1 + \exp(-t))$, θ to be the model's parameters, $L(z, \theta) = \log(1 + \exp(-y\theta^\top x))$ to be the loss, and $H_\theta = \frac{1}{n}\sum_{i=1}^{n} \sigma(\theta^\top x_i)\sigma(-\theta^\top x_i)x_i x_i^\top$ to be the Hessian. The influence of upweighting training point $z = (x, y)$, where x is the input image and y is the output label, on the loss at a test point z_{test}, $\mathcal{I}_{\text{up, loss}}(z, z_{\text{test}})$, is:

$$-y_{\text{test}} y \cdot \sigma(-y_{\text{test}}\theta^\top x_{\text{test}}) \cdot \sigma(-y\theta^\top x_i) \cdot {x_{\text{test}}}^\top H_\theta^{-1} x \tag{1}$$

We train a logistic regression model on each of the datasets for our experiments.

5 Experiments & Results



(a) Adult Dataset (b) Diabetes Dataset

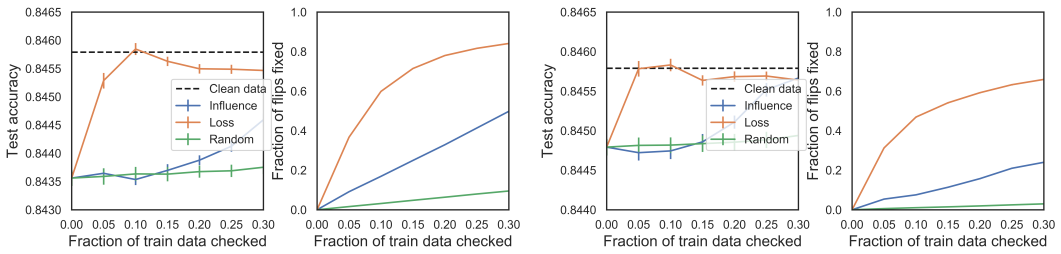
Figure 1: Influence compared to leave-one-out retraining

We compare influence functions' approximation of leave-one-out retraining (effect of removing a training point and retraining the model) to assess influence functions' accuracy. We compare $-\frac{1}{n}\mathcal{I}_{\text{up, loss}}(z, z_{\text{test}})$ with $L(z_{\text{test}}, \hat{\theta}_{\text{-z}}) - L(z_{\text{test}}, \hat{\theta})$ on each of the three datasets (left plots of Fig 1. (a) and (b)), and find that for the adult and heart (Appendix Fig 3) datasets, the influence-predicted loss difference is very similar to the actual loss difference from leave-one-out retraining depicted by a clear positive correlation. But for the diabetes dataset, there is greater discrepancy and no clear correlation between influence-predicted loss difference and leave-one-out retraining's loss difference. The stochastic approximation method [9] (right plots of Fig 1. (a) and (b), and Fig 3.) shows similarly accurate predicted difference results for the three datasets. From these results, we can say that influence functions can be used for efficient approximations of parameter change on some datasets, but not all.

Table 1: Fractions of females vs males and blacks vs whites for Diabetes dataset

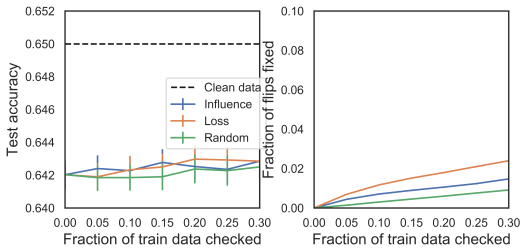
Method	**Fraction of train data checked**	**Sex**		**Race**	
		Female (%)	Male (%)	Black (%)	White (%)
Influence	0.05	2353 (56.0)	1847 (44.0)	766 (18.2)	3218 (76.6)
	0.10	4674 (55.6)	3726 (44.4)	1545 (18.4)	6462 (76.9)
	0.20	9225 (55.0)	7575 (45.1)	3162 (18.8)	12830 (76.4)
Loss	0.05	2234 (53.2)	1966 (46.8)	904 (21.5)	3051 (72.6)
	0.10	4460 (53.1)	3940 (46.9)	1641 (19.5)	6312 (75.1)
	0.20	8976 (53.4)	7824 (46.6)	3251 (19.4)	12785 (76.1)
Random	0.05	2287 (54.5)	1913 (45.5)	802 (19.1)	3227 (76.8)
	0.10	4590 (54.6)	3810 (45.4)	1586 (18.9)	6469 (77.0)
	0.20	9106 (54.2)	7694 (45.8)	3185 (19.0)	12918 (76.9)
	Total	45208 (53.8)	38792 (46.2)	16150 (19.2)	64326 (76.6)

In datasets, minority groups have higher error rates than majority groups based on historical records. We test influence functions to see if the assumed inaccurate data points of minority groups can be prioritized for checking. We show in Table 1 for the Diabetes dataset that minority groups "Female" and "Black" for attributes "Sex" and "Race" are not being picked up more than majority groups "Male" and "White". The quantity ratios of the groups in each particular protected attribute seem to be very similar to the original ratio ("Total"). Influence functions are further shown to not be better at prioritizing minority groups compared to using highest train loss or sorting points at random since there are only slight differences in quantity ratios for each prioritizing method. We show for the Adult and Heart datasets (Appendix Table 2 and 3) that influence is a poor method for prioritizing minority groups for those cases as well. Thus, influence functions seem to have no or negative effect on identifying minority groups in general, regardless of how inaccurate the minority group data may actually be.

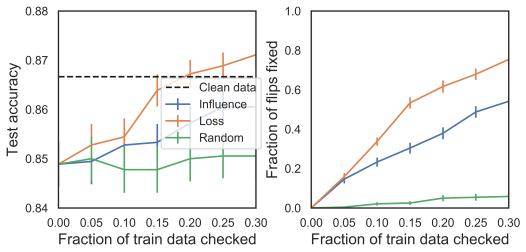


(a) Adult Females

(b) Adult Blacks



(c) Diabetes African Americans



(d) Heart Females

Figure 2: Fixing mislabeled examples

We subsequently flip the labels of a random 10% of individual minority groups of the training points to simulate guaranteed bias. We examine whether influence functions are now able to pick up those mislabeled minority group points. We show in Fig 2 and Fig 4 (Appendix) that across all three datasets, influence functions are still not effective when prioritizing mislabeled data to be checked. Our figures consistently show checking training points in order of highest train loss is more effective as it fixes more flipped points than influence, and increases the model's test accuracy more after fewer points are checked. So, we can say that if influence functions do not work on any of these datasets, it certainly cannot work in all scenarios.

6 Conclusion

In this work, we investigate influence functions in the case of mislabeled-data identification on purposely mislabeled minority group data. We train a logistic regression model on three real-life datasets and find that while influence functions were previously shown by Koh et al. to be effective at identifying mislabeled spam data, they cannot perform to the same degree for all datasets. It is therefore not a reliable tool to be used for repairing datasets.

Furthermore, our results show that although mislabeling errors can be correlated with marginalized identities, influence functions did not find minority samples to be highly influential.

In order to make influence functions more generalizable to more or all datasets, we propose that there be modifications done on the influence function algorithm itself, possibly also tailoring to individual datasets.

References

[1] Filippo A. Raso, Hannah Hilligoss, Vivek Krishnamurthy, Christopher Bavitz, Levin Kim. Artificial Intelligence & Human Rights: Opportunities & Risks. *Berkman Klein Center for Internet & Society at Harvard University*. page 32, 2018.

[2] George D. Magoulas, Andriana Prentza. Machine Learning in Medical Applications. In: Paliouras G., Karkaletsis V., Spyropoulos C.D. (eds) Machine Learning and Its Applications. ACAI 1999. Lecture Notes in Computer Science, vol 2049. *Springer, Berlin, Heidelberg*. page 301, 2001.

[3] Solon Barocas, Moritz Hardt. NIPS 2017 Tutorial on Fairness in Machine Learning. In *Advances in Neural Information Processing Systems*, 2017.

[4] Pang Wei Koh, Percy Liang. Understanding Black-box Predictions via Influence Functions. *the International Conference on Machine Learning (ICML)*, 2017.

[5] Ronny Kohavi, Barry Becker. Adult Census Income. Data Mining and Visualization. Silicon Graphics. 1996. From Dheeru Dua, Casey Graff. *UCI Machine Learning Learning Repository*. [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2019.

[6] Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore, "Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records," *BioMed Research International*, vol. 2014, Article ID 781670, 11 pages, 2014.

[7] Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D, University Hospital, Zurich, Switzerland: William Steinbrunn, M.D, University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D, V.A. Medical Center, Long Beach and Cleveland Clinic Foundation:Robert Detrano, M.D., Ph.D. Heart Disease Data Set. From Dheeru Dua, Casey Graff. *UCI Machine Learning Repository* [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2019.

[8] R. Dennis Cook, Sanford Weisberg. Characterizations of an Empirical Influence Function for Detecting Influential Cases in Regression. *Technometrics*, 22:495-508, 1980.

[9] Naman Agarwal, Brian Bullins, and Elad Hazan. Second-Order Stochastic Optimization for Machine Learning in Linear Time. In *Journal of Machine Learning Research*, 18(1), 2017.

7 Appendix

7.1 Additional results

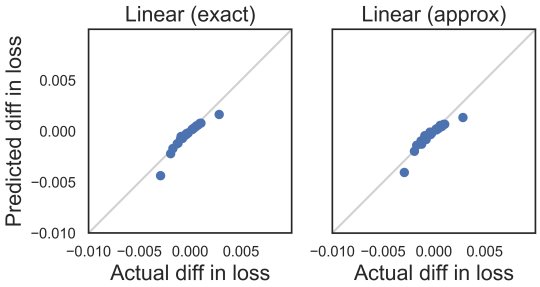


Figure 3: Influence matches leave-one-out retraining cont.

Table 2: Fractions of females vs males and blacks vs whites for Adult dataset

Method	**Fraction of train data checked**	**Sex**		**Race**	
		Female (%)	Male (%)	Black (%)	White (%)
Influence	0.05	425 (22.4)	1475 (77.6)	141 (7.4)	1254 (66.0)
	0.10	903 (23.8)	2897 (76.2)	333 (8.8)	2720 (71.6)
	0.20	1742 (22.9)	5858 (77.1)	676 (8.9)	5936 (78.1)
Loss	0.05	482 (25.4)	1418 (74.6)	127 (6.7)	1681 (88.5)
	0.10	725 (19.1)	3075 (80.9)	233 (6.1)	3402 (89.2)
	0.20	1206 (15.9)	6394 (84.1)	419 (5.5)	6885 (90.6)
Random	0.05	605 (31.8)	1295 (68.2)	183 (9.6)	1037 (85.6)
	0.10	1249 (32.9)	2551 (67.1)	365 (9.6)	2246 (85.7)
	0.20	2440 (32.1)	5160 (67.9)	713 (9.4)	4915 (85.8)
	Total	12378 (32.6)	25622 (67.4)	3554 (9.4)	32670 (86.0)

Table 3: Fractions of females vs males for Heart dataset

Method	**Fraction of train data checked**	**Sex**	
		Female (%)	Male (%)
Influence	0.05	2 (16.7)	10 (83.3)
	0.10	5 (20.8)	19 (79.2)
	0.20	9 (18.8)	39 (81.3)
Loss	0.05	3 (25.0)	9 (75.0)
	0.10	6 (25.0)	18 (75.0)
	0.20	9 (18.8)	39 (81.3)
Random	0.05	4 (33.3)	8 (66.7)
	0.10	9 (37.5)	15 (62.5)
	0.20	20 (41.7)	28 (58.3)
	Total	83 (34.6)	157 (64.4)

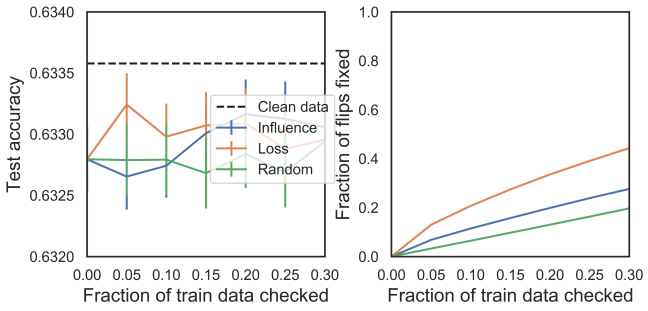


Figure 4: Fixing mislabeled Female examples in Diabetes Dataset