# Labelfix

Presented by Gentry Atkinson

# Source:

- "Identifying Mislabeled Instances in Classification Datasets"
- NM Müller, K Markert
- Cognitive Security Technologies Fraunhofer AISEC  Garching, Germany
- IEEE International Joint Conference on Neural Networks
- Code available at: github.com/mueller91/

# Summary:

- The performance of any classier can be greatly improved by removing mislabeled instances from the data set.
- An estimated 5% of data in real-world data sets is mislabeled.
- One method of identifying mislabeled data would be to classify all points with an appropriate classifier, sort every point by how close the predicted label is to the true label, and remove the most "bad" instances.
- A user could input an appropriate percentage of the points to remove.

# Notation:

- D is some data set composed of X vector and Y vector.
- X vector is all data point
- Y vector is all labels expressed as one-hot vectors
- g a classifying model
- $\langle y_n, \bar{y}_n \rangle$the cosine distance between label y and the label predicted by g
- $\alpha$ the percentage of X to mark as mislabeled
- $I_\alpha$ the subset of X identified as mislabeled

# Preprocessing:

- User calls preprocess_x_y_and_shuffle python function. Function takes a data set and a label set. No other parameters are needed.
- System automatically identifies data type as numerical, image, or language.
- Numerical data is normalized.
- Image data is standardized (set mean to 0 and divide by std deviation)
- NL is mapped to a 300-d embedding and summed.

# Classification:

- User calls check_dataset on the preprocessed x and y. Hyper-parameters are optional.
- g is chosen based on the on the data type identified by pre-processing:
  - Numerical and textual data is classified using a dense NN.
  - Image data is classified using  a CNN with 48 2x2 kernels.
- $\bar{y}_n$ is generated for every $y_n$ by classifying $x_n$

# Sorting:

- $\langle y_n, \bar{y}_n \rangle$ is calculated as the inner product of the real label and predicted label.
  - Represents the probability that $x_n$ is assigned the label $y_n$
  - arcos($y_n$ dot $\bar{y}_n$ / ($|y_n|$ * $|\bar{y}_n|$))
- $I_\alpha$ is generated such that:
  - $\Sigma_{I\alpha} \langle y_n, \bar{y}_n \rangle$ is minimized
  - $|I_\alpha| = \alpha N$ where N is $|X|$

# Experimental Method:

- 29 data sets where chosen
    - 22 real world
    - 7 synthetic produced by sklearn
- Researchers altered labels for $\mu$=3% of the set, and trained labelfix to detect altered labels.
- Precision = $|I_a \cap I| / |I_a|$ = TP / TP + FP
- Recall = $|I_a \cap I| / |I|$ = TP / TP + FN

TABLE II
OVERVIEW OF THE DATASETS.

| Dataset | Size | Type | Classes |
|---|---|---|---|
| adult | (32561, 14) | numerical | 2 |
| breast_cancer | (569, 30) | numerical | 2 |
| cifar10 | (50000, 32, 32, 3) | image | 10 |
| cifar100 | (50000, 32, 32, 3) | image | 100 |
| cifar100, at random | (50000, 32, 32, 3) | image | 100 |
| cifar100, subset aqua | (2500, 32, 32, 3) | image | 5 |
| cifar100, subset flowers | (2500, 32, 32, 3) | image | 5 |
| cifar100, subset household | (2500, 32, 32, 3) | image | 5 |
| credit card default | (30000, 23) | numerical | 2 |
| digits | (1797, 64) | numerical | 10 |
| fashion-mnist | (60000, 28, 28, 3) | image | 10 |
| forest covertype (10%) | (58101, 54) | numerical | 7 |
| imdb | (25000, 100) | textual | 2 |
| iris | (150, 4) | numerical | 3 |
| mnist | (60000, 28, 28, 3) | image | 10 |
| pulsar_stars | (17898, 8) | numerical | 2 |
| sloan-digital-sky-survey | (10000, 17) | numerical | 3 |
| sms spam | (5572, 300) | textual | 2 |
| svhn | (73257, 32, 32, 3) | image | 10 |
| synthetic 1 | (10000, 9) | numerical | 3 |
| synthetic 2 | (10000, 9) | numerical | 5 |
| synthetic 3 | (10000, 45) | numerical | 7 |
| synthetic 4 | (10000, 45) | numerical | 15 |
| synthetic 5 | (10000, 85) | numerical | 15 |
| synthetic 5 | (10000, 85) | numerical | 7 |
| synthetic blobs | (4000, 12) | numerical | 12 |
| twenty newsgroup | (18846, 300) | textual | 20 |
| twitter airline | (14640, 300) | textual | 3 |
| wine | (178, 13) | numerical | 3 |

All data sets.

TABLE III
PRECISION AND RECALL VALUES FOR ARTIFICIALLY ADDED 3% NOISE, AVERAGED OVER FIVE RUNS.

| **Dataset** | **Runtime** | **$\alpha$-precision** | | | **$\alpha$-recall** | | |
|---|---|---|---|---|---|---|---|
| | | $\alpha = 0.01$ | $\alpha = 0.02$ | $\alpha = 0.03$ | $\alpha = 0.01$ | $\alpha = 0.02$ | $\alpha = 0.03$ |
| adult | 2.1 min | 0.80 | 0.63 | 0.51 | 0.27 | 0.42 | 0.51 |
| breast_cancer | 34.0 sec | 0.76 | 0.80 | 0.74 | 0.22 | 0.52 | 0.74 |
| cifar10 | 9.47 min | 0.98 | 0.88 | 0.72 | 0.33 | 0.59 | 0.72 |
| cifar100 | 13.07 min | 0.94 | 0.82 | 0.67 | 0.31 | 0.54 | 0.67 |
| cifar100, at random | 11.48 min | 0.43 | 0.35 | 0.31 | 0.14 | 0.23 | 0.31 |
| cifar100, subset aqua | 20.2 sec | 0.61 | 0.38 | 0.32 | 0.20 | 0.25 | 0.32 |
| cifar100, subset flowers | 32.8 sec | 0.63 | 0.43 | 0.34 | 0.21 | 0.29 | 0.34 |
| cifar100, subset household | 48.6 sec | 0.62 | 0.46 | 0.37 | 0.21 | 0.30 | 0.37 |
| credit card default | 1.9 min | 0.18 | 0.17 | 0.18 | 0.06 | 0.12 | 0.18 |
| digits | 51.8 sec | 0.98 | 0.95 | 0.86 | 0.31 | 0.63 | 0.86 |
| fashion-mnist | 10.71 min | 0.99 | 0.98 | 0.90 | 0.33 | 0.66 | 0.90 |
| forest covertype (10%) | 4.6 min | 1.00 | 0.95 | 0.74 | 0.33 | 0.63 | 0.74 |
| imdb | 3.71 min | 0.70 | 0.61 | 0.51 | 0.23 | 0.41 | 0.51 |
| iris | 26.9 sec | 1.00 | 0.53 | 0.55 | 0.25 | 0.40 | 0.55 |
| mnist | 3.74 min | 1.00 | 1.00 | 0.97 | 0.33 | 0.67 | 0.97 |
| pulsar_stars | 51.8 sec | 0.91 | 0.86 | 0.78 | 0.30 | 0.57 | 0.78 |
| sloan-digital-sky-survey | 1.5 min | 0.80 | 0.71 | 0.63 | 0.27 | 0.47 | 0.63 |
| sms spam | 1.44 min | 0.85 | 0.86 | 0.79 | 0.28 | 0.57 | 0.79 |
| svhn | 13.6 min | 0.92 | 0.90 | 0.83 | 0.31 | 0.60 | 0.83 |
| synthetic 1 | 2.05 min | 1.00 | 0.98 | 0.89 | 0.33 | 0.66 | 0.89 |
| synthetic 2 | 2.74 min | 1.00 | 0.99 | 0.89 | 0.33 | 0.66 | 0.89 |
| synthetic 3 | 3.79 min | 1.00 | 0.99 | 0.91 | 0.33 | 0.66 | 0.91 |
| synthetic 4 | 4.9 min | 0.98 | 0.90 | 0.74 | 0.33 | 0.60 | 0.74 |
| synthetic 5 | 3.53 min | 0.95 | 0.84 | 0.70 | 0.32 | 0.56 | 0.70 |
| synthetic 6 | 3.58 min | 1.00 | 0.98 | 0.86 | 0.33 | 0.65 | 0.86 |
| synthetic blobs | 37.8 sec | 1.00 | 1.00 | 0.98 | 0.33 | 0.67 | 0.98 |
| twenty newsgroup | 3.2 min | 0.79 | 0.73 | 0.63 | 0.26 | 0.49 | 0.63 |
| twitter airline | 2.39 min | 0.66 | 0.52 | 0.43 | 0.22 | 0.34 | 0.43 |
| wine | 28.1 sec | 1.00 | 1.00 | 0.88 | 0.20 | 0.60 | 0.88 |
| **Averages** | | 0.84 | 0.77 | 0.68 | 0.27 | 0.51 | 0.68 |

Results

# Conclusions:

- Paper claims best overall precision=0.84 when α=0.01 and μ=0.03.
- When μ=α=0.03, total overall precision falls to 0.68 with recall of 0.68
- Intent is to mark suspicious instances for review.
- System was able to detect some **actual** mislabeled data in real world sets:

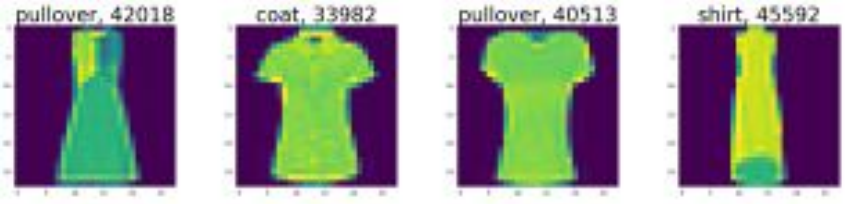


Fig. 2. Mislabeled instances in the Fashion-MNIST training set.

# Proposed future work:

- Expand preprocessor and g to accommodate time series and bio-signal data. Preprocessor could be same as numerical if features are extracted traditionally. g could include a 1d-CNN.
- g could be expanded to be an ensemble method. 3 (or whatever number) of models could classify an instance. $I_a$ would be generated using:
    - $\langle y_n, \bar{y}_n \rangle = \langle y_n, \bar{y}_{n1} \rangle + \langle y_n, \bar{y}_{n2} \rangle \langle y_n, \bar{y}_{n3} \rangle$