A new structure and criterion for dataset of image classification

Xinbin Zhang
Sydney Machine Learning Study Group
abner.zhangxinbin.anz@gmail.com

Abstract

What is the correct label for an image in Machine Learning? I investigate the process of label assignment of MNIST, CIFAR-10 and Fashion-MNIST and the results show that even in these simple datasets classes are not always as easy to distinguish as lived and died, and sometimes even impossible. The reasons include: lack of mutually exclusive definition for each class, and the definition lack of features to distinguish the classes, and the instances are incoherent to the definition and overlap, and the label can't present uncertainty. In current research, these images and labels are treated as noise label. To give a clearer definition for these cases, I propose a new structure of dataset including: mutually exclusive definition of classes and the label of labels, which classify labels into coherent, wrong and uncertain (including multi-objects, mid-object, unknown and unclear), and instances of different labels, and criterion to assess algorithms performance. Such structure could also apply to more complex dataset such as ChestXray14 in medical field to make learning and prediction more accurate and meaningful.

1 Introduction

In the machine learning community, there are two standardized image classification datasets (MNIST and CIFAR-10) that are widely used. The list of the state-of-the-art on them is important in allowing a more holistic approach to assessing and characterizing the performance of an algorithm or model. Convolutional nets have already achieved very high accuracy on MNIST and CIFAR-10, which leave less place to improve. At same time some techniques such as Bayesian Neural Network and Adversarial Network raise the question of dealing with uncertainty and attack. And some new datasets appear, such as Fashion-MNIST, which is the same structure as MNIST and wants to replace the position of MNIST. But current structure of dataset of MNIST, CIFAR-10 and Fashion-MNIST has problems such as lack of mutually exclusive definition for each class, and lack of features to distinguish the class, and the instances are incoherent to the definition and overlap, and the label can't present uncertainty, and criterions could not assess the performance on uncertainty and anti-attack.

In this paper, I discuss related work in Section 2. Section 3 introduces the new structure of the dataset and new criterions. Section 4 shows the exploration in MNIST, CIFAR-10 and Fashion-MNIST and explain why new structure and new criterions are needed. Section 5 provides some advice on how to build dataset in such structure. Section 6 is conclusion.

2 Related Work

Current studies are focused on the noise label. There exists a large literature about class noise, such as different definitions and consequences of class noise.

In the survey, [Frenay *et al.*, 2014] defines that label noise is considered to be a stochastic process, i.e., the case where the labeling errors may be intentionally and maliciously induced by an adversary agent is not considered. And he also discusses several definitions and sources of label noise, and the potential consequences of label noise and three types of approaches to deal with label noise: label noise-robust, label noise cleansing, and label noise-tolerant methods.

In [Brodley *et al.*, 1999], labeling error can occur for several reasons including subjectivity, data-entry error, or inadequacy of the information used to label each object.

In [R. J. Hickey, 1996], noise is anything that obscures the relationship between the features of an instance and its class.

In [J. R. Quinlan, 1986], noise is also described as consisting of nonsystematic errors. In the literature, two types of noise are distinguished: feature (or attribute) and class noises.

3. The new structure

3.1 Mutually exclusive explicit definition

In some datasets, definition of classes is missed, which will cause confusion. For example, no definition in MNIST, in CIFAR-10 only definition between automobile and truck, and Fashion-MNIST only provides the name of classes, such as Label 0, description T-shirt/Top, label 6, description shirt. On Wikipedia, T-shirt is a subclass of shirt. If I take the T-shirt as a class and the others include: dress shirt, camp shirt, Polo shirt, Poet shirt, Baseball shirt, Top shirt as the shirt class. It still causes confusion, because of "Top shirt – a

long-sleeved collarless polo shirt" mixed features of T-shirt and shirt. And there is no definition about female's shirt.
Since I am not expert in fashion field, I just give a simple example of definition for T-Shirt and Shirt for male.
T-Shirt: a garment for the upper part of the body, normally short sleeves and a round neckline, lacks a collar.
Shirt: a garment for the upper part of the body, usually long--sleeved, and having a collar and a front opening.

3.2 The label of labels

I classify labels into three types:
Coherent label, image is coherent to the definition of the class.
Wrong label, image is coherent to the definition of the other class.
Uncertain label, image is not coherent to the definition of the class, which causes confusion between readers.
The uncertain label could be classified into 4 subtypes based on the reason of incoherence.
Multi-objects label, there are more than 2 objects, which belong to more than 2 classes.
Mid-object label, the image is close to more than 2 classes.
Unknown label, the object is clear but doesn't belong to the classes in the list
Unclear label, the image itself is not clear to judge.

3.3. The instances

Table1 is an example of the number of each type label.

Table 1 the number of each type of instances

	Train	Test
Coherent	50000	10000
Wrong	100	0
Uncertain	1000	1000
Adversarial	0	100

3.4. New criterions

The criterions include the accuracy with coherent dataset (both in training and test), the accuracy with noise in training dataset (including uncertain and wrong), the prediction on uncertain instances, the prediction on adversarial instances, the numbers of training dataset used, and the complex of algorithm.
For example,
the score =
X1*w1+X2*w2+X3*w3+X4*w4+X5*w5+X6*w6+X7*w7
X1 = the number of correct prediction with coherent training dataset
X2 = the number of correct prediction by adding wrong labels
X3 = the number of correct prediction by adding uncertain labels
X4 = the number of detection of uncertain test dataset
X5 = the number of detection of adversarial test dataset
X6 = the number of training instances
X7 = the number of parameters
w1-w7 are the weights of each criterions.

4. Exploration

4.1 Exploration on MNIST

The MNIST database of handwritten digits, has a training set of 60,000 examples, and a test set of 10,000 examples. It is a subset of a larger set available from NIST Special Database including handwritten of digits and upper and lower case, which are hand checked. The handwritten digit is a symbol which is difficult to make a mutually exclusive definition, and even difficult to decide what is the correct label. Since there is no definition, I propose a possible definition used in NIST: in the context of handwritten digits recognition, based on the shape of handwritten, choose the closest to the print version between 0-9.
There are 3 parties involved the decision of the label in NIST Special Database, the writers, the image processors, and the label checkers.
The writers are required to write the handwritten below the print symbols as Figure 1

Figure1 the writer d3462's original form

The image processors do the image segmentation and assign a label based on the print symbols. They are involved, because some writers handwriting was not segmented correctly, such as writer d3462, many handwritings' the below part are missing.

Figure2 the writer d3462's handwritten digits' segmentation is not correct.

Figure1 the part of original form of writer d3462 shows the handwritings are close to the button line, which causes the wrong segmentation, and it also shows the handwritten of 2, which is different with d3462_15_00012.

The part of a symbol could be another symbol, as below shows part of 3 becomes to 2, part of 2 becomes 7.

Figure3 shows that different part of a handwritten could be another symbol.

The label checker notice that the image d3462_15_00012 is closer to 2, and change the label from 3 to 2. Because in the context of the original form, it is 3, and in the context of digital recognition in 0-9, it is closer to 2.

In MNIST, 3443 and 17229 are two examples, which are not segmented completely, it's label is 2, but it is part of 3.

 3443

 17229

It is also true when the image is changed in other ways. For instance, when adversarial noise is added into the image, the label should not be the same.

At same time, label checkers could make mistake. For example, below shows 4 wrong labels:

 2080 labeled as 3

 54915 labeled as 4

 30310 labeled as 5

 44960 labeled as 6

And sometimes it is impossible to give a correct label. For example, the given labels in below figure 4 are 1,4,4,7, but the labels could change based on different bias of different readers, and it could be difficult to reach an agreement on what's the correct label.

64201 25792 32450 4433

Figure 4 confusion labels

Another example, the given labels in below figure 5 are 4,4,4,4, but it is difficult to say they are handwritten digits.

60033 9715 27776 24922

Figure 5 confusion labels

The uncertain might be the best label for these images.

The uncertain labels in MNIST are Mid-objects label, unknown label and unclear label.

Unknown labels:

 12958 in the context of form is 9, but in a large context of 0-9 and a-z, it should be 'g', which is unknown in 0-9.

 23652 is labeled as 4 and some readers may think it is closer to 9, but in a large context of 0-9 and a-z, it should be 'q', which is unknown in 0-9.

Mid-label:

 number 50598 labeled as 4, based on different reader's bias, it could be 7 as figure4 shows.

Unclear label:

 66571 labeled as 9

4.2 Exploration on CIFAR-10

CIFAR-10 is selected from Tiny Images dataset, which contains 80 million 32 × 32 color images collected from the webs, and it could explain that there are only 59863 unique value in 60000 images. And labels are given by university students based on Labeler instruction sheet "The only criteria for including an image were that the image contain one dominant instance of a CIFAR-10 class, and that the object in the image be easily identifiable as belonging to the class indicated by the image label."

CIFAR-10 only gives a definition between car and truck: "There is no overlap between automobiles and trucks. "Automobile" includes sedans, SUVs, things of that sort. "Truck" includes only big trucks. Neither includes pickup trucks.", which is different from the definition on Wikipedia: "A truck or lorry is a motor vehicle designed to transport cargo. Trucks vary greatly in size, power, and configuration; smaller varieties may be mechanically similar to some automobiles." And below instances are not coherent to definition and are overlapped.

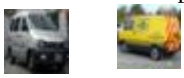

Image 16925 and 22490 are labeled as automobile.

Image 55416 and 5013 are labeled as truck.

4 Wrong labels:

 52405 labeled as cat

 52804 labeled as cat

 21347 labeled as cat

 17455 labeled as a cat

4 Multi-objects labels

 15696: a bird on a car

 56859 is a cat on a car

 8208: a deer in front of a car

 35829: a truck beside a car

4 Unknown labels:

 1569 labeled as a deer

 18310 labeled as truck

 5074 labeled as a deer

 52226 labeled as a bird

Unclear label:

 18839

And CIFAR-10 didn't give the definition between ship and car to prevent overlap or mid-object label.

Based on the definition on Wikipedia:

"A ship is a large watercraft that travels the world's oceans and other sufficiently deep waterways, carrying passengers or goods, or in support of specialized missions, such as defense, research and fishing."

"A car (or automobile) is a wheeled motor vehicle used for transportation. Most definitions of car say they run primarily on roads, seat one to eight people, have four tires, and mainly transport people rather than goods."

 4972 shows a new type of instance that could run on the road as an automobile and travel the deep waterways, so it is a mid-object label.

The uncertainty in test dataset

The accuracy is based on the number of predictions that are equal to the given label. If there are wrong labels and uncertain labels in the test dataset. The accuracy will not be accurate. For example: 52405's given label is cat, if an algorithm or a human predicts 52405 as a frog, the prediction will be judged as a wrong prediction but the judgment itself is wrong.

For uncertain labels, image 55416 could be an automobile or a truck based on training dataset. If it is predicted as a truck and the prediction will be judged as a wrong prediction but the judgment itself could be wrong. Image 56859 is a cat on an automobile, the predictions of cat and automobile are both right. Image 52226 is not in the list of classes, any prediction is wrong. That is why new criterions should include performance on uncertainty.

The importance of coherent between image and label

 8861 is labeled as a ship, but it is a Windsurfer. If it is in the training dataset, algorithms could only learn the high-level features of the environment. If it is in the test dataset and is predicted correctly, the prediction is not based on the high-level features of the ship, but the features of the environment such as water. And such learning and prediction is not the purpose of machine learning on object recognition.

4.3 Exploration on Fashion-MNIST

Unlike CIFAR-10, Fashion-MNIST doesn't provide the process of labeling and the mutually exclusive definition.

On Wikipedia, there is no definition of female t-shirt and female shirt, and the definition of male t-shirt and male shirt is not mutually exclusive.

The below figure is part of picture in Fashion-MNIST description, the above part is t-shirt/top, the below part is the shirt. There are overlap in the labeled female t-shirt/top and the labeled female shirt, and the labeled male t-shirt has collar, and the labeled male shirt is short sleeve and has no collar and no front opening.

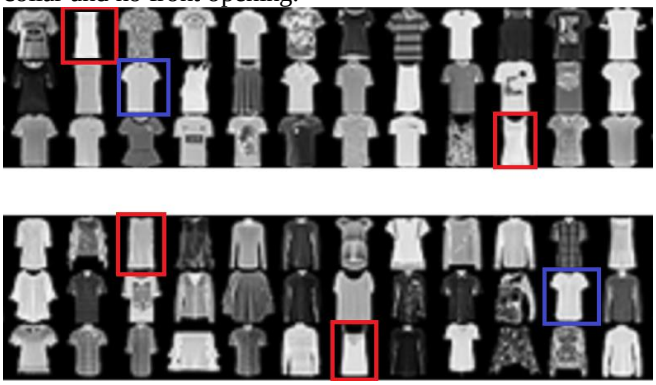

Fig 6 part of the picture in Fashion-MNIST description.

Lack of definition causes it difficult to decide wrong labels or mid-object labels. It needs a mutually exclusive definition as shown in Section 2.

 2693 a man wears t-shirt and trousers, labeled as trouser. And 19545,34248,27110 in class trousers are a model similar to 2693.

Image 37183,30191 in class pullover is a model.

Image 7388, 14274 in class dress is a model.

Image 21701 in class coat is a model.

Image 65249 in class shirt is a model.

unknown label:

 58983 is a carton man wear glass, labeled as bag

5 Building new structure dataset

The reliably labeled data are often expensive and time consuming to obtain. It is also true to labels' label.

It could be built like Wikipedia: The owners set up initial definition and use techniques like noise clean to set up initial instances, labels and labels' label. Then the datasets could be published and optimized by the Machine Learning communities. And new versions and new criterions could be updated with the development of techniques.

6 Conclusions

Uncertainty exists not only in real world but also in these widely used standardized datasets. The label can't present uncertainty, lack mutually exclusive definition, the instances are incoherent to the definition, it could be summarized as that if the question is not asked clearly, the accuracy could be

inaccurate and the result could be meaningless. The classes of MNIST and CIFAR-10 is simple, but the correct label structure could be complex. By building datasets in new structure, we can raise clear questions and assess the performance of algorithms. It is more important for real-world problems of image classification, such as medical images, which will affect patients welfare.

Reference

[1] [Brodley et al., 1999] C. E. Brodley and M. A. Friedl, “Identifying mislabeled training data,” J. Artif. Intell. Res., vol. 11, pp. 131–167, Jun. 1999.

[2] [Frenay et al., 2014] Classification in the Presence of Label Noise: a Survey, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 25, NO. 5, MAY 2014

[3] [R. J. Hickey, 1996], “Noise modelling and evaluating learning from examples,” Artif. Intell., vol. 82, nos. 1–2, pp. 157–179, 1996.

[4] [J. R. Quinlan, 1986], “Induction of decision trees,” Mach. Learn., vol. 1, no. 1, pp. 81–106, 1986.

[5] [Hinton, Geoffrey E., et al, 2012] Hinton, Geoffrey E., et al. "Improving neural networks by preventing co-adaptation of feature detectors." arXiv preprint arXiv:1207.0580 (2012).

[6] [Cohen, Gregory, et al, 2017]. "EMNIST: an extension of MNIST to handwritten letters." arXiv preprint arXiv:1702.05373 (2017).