The Problem of Accuracy as an Evaluation Criterion

Edith Law EDITH@CMU.EDU
Machine Learning Department, School of Computer Science, 5000 Forbes Ave, Pittsburgh, PA 15213 USA

Abstract

In this position paper, I argue that for a particular class of problems, the widely adopted evaluation criterion, accuracy, is an *incorrect* measure. I describe the general characteristics of this class of problems and why accuracy is not a suitable evaluation method, using examples from computer vision, machine translation and music information processing.

1. Introduction

The typical way to evaluate the quality of a learning algorithm is to compare its output against some previously collected ground truth via a loss function. More formally, suppose there exists a true function $f(x)$ that a set of algorithms $\mathcal{A} = \{A_1, A_2, \ldots, A_k\}$ all try to learn. Let $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$ be a set of test examples. An algorithm A_i which outputs a hypothesis $\hat{f}_i(x)$ is considered superior to algorithm A_j if $l_{0/1}(\hat{f}_i(x), f(x)) < l_{0/1}(\hat{f}_j(x), f(x))$, where $l_{0/1}(\hat{f}(x), f(x)) = \sum_{n=1}^{N} \mathbf{1}(\hat{f}(x_n) \neq f(x_n))$. The 0/1 loss function $l_{0/1}$, also known as *accuracy*, has become one of the most popular de-facto methods in evaluating the quality of learning algorithms.

As machine learning techniques are applied to increasingly more complex domains, there emerges a class of problems for which measuring accuracy is not a correct evaluation approach. Using examples from computer vision, natural language processing and music information processing, I will illustrate the general characteristics of this class of problems and explain why accuracy is not a suitable method of evaluation.

2. Examples from Three Domains

2.1. Salient Region Detection for Images

Visual attention has been confirmed by numerous psychological and neuroscientific studies to be an integral part of human vision. The idea is that there exists a pre-recognition phase where humans usually attend to certain regions of interest (ROI) that are the most salient. Exploiting this fact, many algorithms have since been developed for automatically detecting salient regions in images, which are then applied to domains such as object recognition, adaptive image compression (Ouerhani et al., 2001), automatic cropping and information retrieval.

In a recent study (Liu et al., 2007), the authors compare their approach of salient object detection against two other algorithms (Itti & Baldi, 1998; Ma & Zhang, 2003). The ground truth set consists of a set of images and their associated set of rectangular regions, indicating the presence of salient objects. To evaluate the relative merits of the algorithms, three familiar measures are used: precision (proportion of pixels in the learned salient regions that are found in ground truth salient regions), recall (the proportion of pixels in the ground truth salient regions that are found in the learned salient regions) and F-measure (harmonic mean of precision and recall).

The idea of using *overlap*, or rectangular intersection, as a measure of distance between the learned and ground truth salient regions is used extensively. Some variations of this measure include ratio of spatially unmatched points (Lin & Yang, 2007), percentage score of under and over extraction (Ko et al., 2004) and proportion of point-wise matches (Kadir et al., 2004).

The main issue with this evaluation approach is correctness. There is little guarantee, without human inspection, that a region that has a larger overlap with the ground truth salient region is actually perceptually more meaningful than a region that has a smaller overlap with the ground truth salient region.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

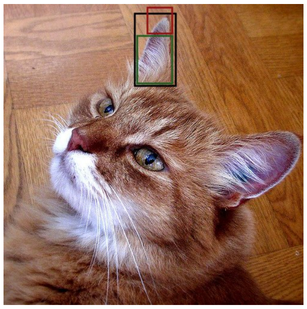

Figure 1. The black rectangle is the ground truth salient region, the green and red rectangles are two salient regions to be evaluated.

Consider the example in Figure 1. Although the green rectangle has greater overlap with the black ground truth salient region, it is arguable whether it is more salient than the red rectangle, which contains the tip of the cat's ear. Interpreting small differences in the accuracy of multiple algorithms is especially problematic. It is possible for the detected regions output by different algorithms to have varying degree of overlap with the ground truth region, all of which are equally salient when evaluated by human judges.

2.2. Machine Translation

Given a sentence in a source language, the task of a machine translation algorithm is to generate a sentence with equivalent meaning in the target language. Consider the following example (Figure 2):

1. At least 12 people were killed in the battle last week.
2. At least 12 people lost their lives in last week's fighting.
3. Last week's fight took at least 12 lives.
4. The fighting last week killed at least 12.
5. The battle of last week killed at least 12 persons.
6. At least 12 persons died in the fighting last week.
7. At least 12 died in the battle last week.
8. At least 12 people were killed in the fighting last week.
9. During last week's fighting, at least 12 people died.
10. Last week at least twelve people died in the fighting.

Figure 2. Human translations of the same Chinese sentence from the Multiple Translation Corpus (MTC).

It is apparent in the above example that evaluation of machine translation algorithms is difficult because for any given source sentence, there are many possible *good* candidate translations of that sentence. In addition, the relative merit of the candidate translations is not easy to judge manually, let alone automatically.

In order to foster a tighter loop between research and evaluation, several metrics, including the F-measure (J. Turian & Melamed, 2003), BLEU (Papineni et al., 2002), NIST (Doddington, 2002), WER (word error rate) and METEOR (Lavie et al., 2004), have been developed, with BLEU being the most prevalent. Many of these metrics are based on some form of matching between the words in the translated sentence and the source sentence, with other parameters allowing for variation in word choice, phrase order and length of the sentences.

Although the BLEU score is shown to be correlated with human judgments, this correlation is not *guaranteed* (Callison-Burch et al., 2006). The most alarming evidence is the fact that it is possible for millions of variations of a candidate translation to receive the same BLEU score, even though not all these variations are "equally grammatical or semantically plausible." In addition, BLEU is found to underestimate the quality of some translation systems that do not use N-gram techniques (Lee & Przybocki, 2005; Callison-Burch et al., 2006).

2.3. Automatic Annotation of Music

With the popularity of the web as a medium for finding and discovering music, accurate and semantically relevant descriptions of music have become increasingly important for use in search and recommendation. In the music information retrieval (MIR) community, there have been some recent efforts in automatically generating tags for music (Turnbull et al., 2007; Eck et al., 2007; Bergstra et al., 2006). For any given piece of music, there can be many possible descriptions, including artist, genre, instrumentation, mood and purpose. As such, the automatic generation of tags for music can be thought of as a multi-class supervised learning problem where the goal of the algorithm is to map each song to a set of tag classes.

Last.fm	classic rock, rock, 70s, singer-songwriter, british, pop, oldies, john lennon, piano, the beatles, relaxing, sad, protest, political, melancholic, beautiful, classic, peace, ballad, acoustic, alternative, anti-war, awesome, best songs ever, blues, britpop, calm, chill

Figure 3. Tags for *Imagine* by *John Lennon.*

The typical approach for evaluating tags is to measure the proportion of correctly (as predefined by some ground truth set) classified songs, using precision and recall (Turnbull et al., 2007) or comparisons of ranked lists (Eck et al., 2007).

The following is an example of why this approach can be, in many cases, incorrect. Suppose that we were to compare two algorithms A_1 and A_2 in their ability to produce a set of five tags that accurately describes

John Lennon's *Imagine* and suppose that the tags from Last.fm in Figure 3 is the ground truth set.

A_1	A_2
the beatles	the beatles
piano	piano
calm	**aggressive**
country	**heavy metal**
60s	70s

Figure 4. Comparison of two hypothetical automatic tag generation algorithms. Tags in bold fonts are mistakes made by the algorithms.

Under the accuracy criterion, algorithms A_1 and A_2 (Figure 4) would be deemed equal because they made the same number of mistakes. However, to a user of a music retrieval system, seeing John Lennon's *Imagine* returned as a search result of the *aggressive heavy metal* query would be much more perplexing than as a search result of the *country songs in the 60s* query.

2.4. Common Properties, Common Problems

The three examples – salient region detection, machine translation, and automatic music tag generation - share some common characteristics. First, the function being learned is a one-to-many function: one image to many salient regions, one sentence to many translated sentences, one song to many tags. The problem in using accuracy to compare learned and ground truth data is that we are comparing sets of things without explicitly stating which subset is more desirable than another. In order for accuracy to be a *correct* measure of the quality of algorithms, the ground truth set must contain a ranked list of all possible subsets of outputs as a reference point for learning algorithms, which is clearly infeasible to collect. Accuracy is not a correct measure of quality unless we have a ground truth set that is of exponential size, which makes evaluation a highly inefficient process.

Second, the function is learned to approximate another quantity that is being sought after. In the salient region detection problem, what we are really after are regions of images that allow an individual to *recognize* an object. In the machine translation problem, what we are really after is translated text that conveys *comprehensible* ideas in the source text. In the automatic music tagging problem, what we are really after is a set of tags that can *sufficiently identify* a piece of music, so that it can be readily cataloged and retrieved by users. I would argue that in order to achieve these objectives, it is neither necessary nor sufficient to match the outputs of algorithms with the ground truth exactly.

2.5. An Alternative Approach

For this class of problems, an alternative evaluation approach is to measure *directly* the extent to which an algorithm captures the sought-after quantity. In the salient region detection problem, this quantity can be defined as the set of regions of a given image that allow humans to recognize the object in that image.

Figure 5 illustrates the difference between the two approaches. In the standard approach, an algorithm is evaluated based on the extent to which the output of its hypothesis $\hat{f}(x_i)$ matches that of the true function $f(x_i)$, where x_i is a particular image. In the alternative approach, the hypothesis $\hat{f}(x_i)$ is evaluated by the extent it allows humans to recognize the object o_i associated with the image, i.e. that it satisfies the equation $h(\hat{f}(x_i)) = o_i$ where h is a *human perception* function that maps a set of salient regions to an object.

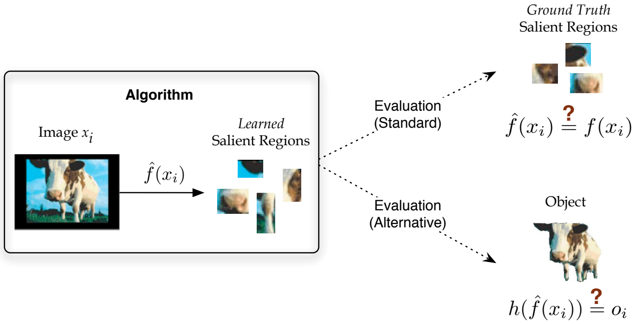

Figure 5. Evaluation by standard vs alternative approach

Evaluating algorithms using this alternative approach is more correct for two reasons. First, it is much easier for humans to recognize the object in an image (i.e. to execute the *human perception* function h) than for them to pinpoint exactly the salient regions of that image (i.e. to know the true function $f(x)$ explicitly), which is what is asked of them to produce the ground truth set. Second, learned salient regions that allow humans to recognize the object in an image are by *definition* salient. This alternative approach, therefore, can potentially reduce the amount of false positive (learned salient regions with large overlap with the ground truth salient regions that are in fact not salient) or false negative (learned salient regions with small overlap with the ground truth salient regions that are in fact salient) evaluations.

Furthermore, the alternative approach can be implemented in an efficient way. Peekaboom, a two-player online game, is one such implementation (Figure 6). In this game, the *boomer* is given an image and a label and asked to click on regions of the image that will make his partner, the *peeker*, guess the label. The regions that the boomer clicked on, which *enabled* the peeker to guess the object in the image, are by definition salient.

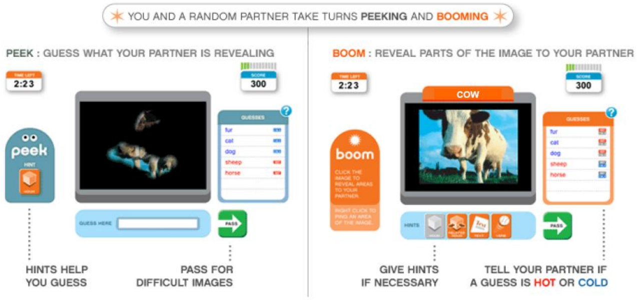
YOU AND A RANDOM PARTNER TAKE TURNS PEEKING AND BOOMING
PEEK : GUESS WHAT YOUR PARTNER IS REVEALING
2:23
300
peek
HINT
GUESSES
fur
cat
dog
sheep
horse
GUESS HERE
PASS
HINTS HELP YOU GUESS
PASS FOR DIFFICULT IMAGES
BOOM : REVEAL PARTS OF THE IMAGE TO YOUR PARTNER
2:23
300
COW
boom
CLICK THE IMAGE TO REVEAL AREAS TO YOUR PARTNER
RIGHT CLICK TO PING AN AREA OF THE IMAGE
GUESSES
fur
cat
dog
sheep
horse
HINTS
PASS
GIVE HINTS IF NECESSARY
TELL YOUR PARTNER IF A GUESS IS HOT OR COLD

Figure 6. Peekaboom

By replacing the boomer with a salient region detection algorithm, we can evaluate the algorithms based on the percentage of people who succeed in recognizing the object in the image using the salient regions that are generated by the algorithm. The argument for the efficiency of this evaluation method derives from the fact that a large number of people play this game on a repeated basis. For example, within a one-month period, over 14,153 different people played the game during this time, generating 1,122,998 pieces of data (von Ahn et al., 2006).

3. Conclusion

With the conclusion of this position paper, I argue that there exists an alternative approach using *human computation games* that is both more *correct* and *efficient* for evaluating algorithms that belong to the class of problems just described. Up until now, there has been much research focusing on getting human players to provide ground truth data as a by-product of having fun. In this paper, we propose the use of these games to involve humans in *directly* evaluating the quality of algorithms.

There are already implemented, or easily implementable, games that can evaluate algorithms for each of the problems described in this paper. For example, Tagatune is a game that presents two players with either the same piece of music or different pieces of music (Law et al., 2007). After seeing each other's descriptions, they must decide whether they are listening to the same thing or not. The extent to which the players can guess that they are listening to the same songs using tags generated by an algorithm is possibly indicative of how good a set of tags are for identifying a song.

Finally, it has been suggested by (Kulesza & Shieber, 2004; Reeder, 2004) to use a Turing Test to judge the quality of a translation, at least in terms of fluency. Their idea can easily be implemented as a game that can collect evaluations at an unparalleled rate.

References

Bergstra, J., Casagrande, N., Erhan, D., Eck, D., & Kégl, B. (2006). Aggregate features and AdaBoost for music classification. *Machine Learning*, *65*, 473–484.

Callison-Burch, C., Osborne, M., & Koehn, P. (2006). Re-evaluating the role of bleu in machine translation research. *EACL*.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. *HLT* (pp. 128–132).

Eck, D., Lamere, P., Bertin-Mahieux, T., & Green, S. (2007). Automatic generation of social tags for music recommendation. *NIPS*.

Itti, L., & Baldi, P. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on PAMI*, *20*, 1254–1259.

J. Turian, L. S., & Melamed, I. (2003). Evaluation of machine translation and its evaluation. *the MT Summit IX, New Orleans, USA* (pp. 386–393).

Kadir, T., Zisserman, A., & Brady, M. (2004). An affine invariant salient region detector. *ECCV*.

Ko, B., Kwak, S., & Byun, H. (2004). SVM-based salient region(s) extraction method for image retrieval. *ICPR* (pp. 977–980).

Kulesza, A., & Shieber, S. (2004). A learning approach to improving sentence-level MT evaluation. *Proceedings of ICTMIMT*.

Lavie, A., Sagae, K., & Jayaraman, S. (2004). The significance of recall in automatic metrics for mt evaluation. *AMTA* (pp. 128–132).

Law, E., von Ahn, L., Dannenberg, R., & Crawford, M. (2007). Tagatune: a game for music and sound annotation. *ISMIR*.

Lee, A., & Przybocki, M. (2005). Nist 2005 machine translation evaluation official results.

Lin, D., & Yang, S. (2007). Wavelet-based salient region extraction. *PCM* (pp. 389–392).

Liu, T., Sun, J., Zheng, N., Tang, X., & Shum, H. (2007). Learning to detect a salient object. *CVPR* (pp. 1–8).

Ma, Y., & Zhang, H. (2003). Contrast-based image attention by analysis by using fuzzy growing. *ICMM* (pp. 374–381).

Ouerhani, N., Bracamonte, J., Hugli, H., Ansorge, M., & Pellandini, F. (2001). Adaptive color image compression on visual attention. *Proceedings of ICIAP* (pp. 416–421).

Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2002). Bleu: a method for automatic evaluation of machine translation. *ACL* (pp. 311–318).

Reeder, F. (2004). Investigation of intelligibility judgments. *AMTA* (pp. 227–235).

Turnbull, D., Barrington, L., Torres, D., & Lanckriet, G. (2007). Towards musical query-by-semantic description using the cal500 data set.

von Ahn, L., Liu, R., & Blum, M. (2006). Peekaboom: a game for locating objects in images. *CHI* (pp. 55–64).