Accuracy measures and voting methods

Le Bich Phuong

Hanoi University of Mining and Geology

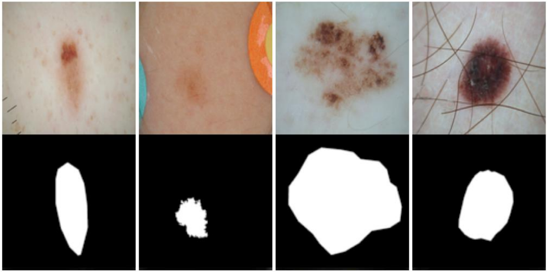
Joint work with "Torus Actions" team,
under the supervision of Prof. Nguyen Tien Zung

Introduction

Most problems in artificial (and human) intelligence may be formulated as a combination of **binary decision** (or classification) problems.

(*A complicated information = many bits of data, each bit is binary*).

For example, the question "Is the skin lesion in these images a **skin cancer or not**?" is a binary problem. Also, the **lesion segmentation** problem for these dermatoscopic images may be viewed as a combination of binary problems, one for each pixel: *does the pixel belong to the lesion?*



Introduction

In practice, in order to increase the accuracy in binary decision problems, one often uses **voting methods**.

The basic idea is to have a group of (*human or AI*) experts and make them vote. Intuitively *and hopefully*, the accuracy of the collective decision via voting will be better than the accuracy of any particular expert (*voter*) in the group.

This idea works very well in practice. Nevertheless, we are faced with the following questions:

1. What is the correct measure of accuracy?
2. What about theoretical gains and limits of voting methods? *(The accuracy can't go to 100% even if the number of voters goes to* ∞*)*.
3. What are the best voting methods for each problem?

Introduction

In this talk, I would like to present three results (*that our team "Torus Actions" obtained under the guidance of Prof. Nguyen Tien Zung*), which address the above three problems, namely:

1. First, the notion of **cost-adjusted**, or **cost-wise accuracy**
2. Second, an **asymptotic formula for accuracy improvement** by voting when the number of voters tends to infinity
3. Last, our **topological voting method**, which *significantly* outperforms the usual average voting method in many image segmentation problems.

Cost-wise accuracy

Recall that, in a binary classification problem, there are not one but two kinds of errors: **false positives** and **false negatives**.

As an illustration, assume that we have a herd of 10000 cows of which 10 are mad, and we have to diagnose them.
- Mad cow diagnosed as healthy: false negative
- Healthy cow diagnosed as mad: false positive

The often-used **naive binary accuracy** score S_{naive}, defined by

$$S_{naive} = P(\Omega^0 \cap \Omega_E^0) + P(\Omega^1 \cap \Omega_E^1) = 1 - P(\Omega^0 \cap \Omega_E^1) - P(\Omega^1 \cap \Omega_E^0)$$

is often very misleading.
Here Ω is the total probability space (the set of all cows), P is a natural probability measure (frequency), Ω^0 is the true negative set (the set of cows which are not-mad), Ω_E^0 is the set of elements classified as negative by some expert E (the set of cows diagnosed as not-mad), and so on. *(Ω^1 is the true positive set (the set of cows which are mad), , and Ω_E^1 is the set of elements classified as positive by E (the set of cows diagnosed as mad)).*

Cost-wise accuracy

For example, if "expert" E says that all cows are healthy, he's compeletely useless, even though is naive binary accuracy is 99,9% *(because 99,9% of cows are healthy, but the other 0,1% are deadly)*.

A more reasonable accuracy measure, used by some people, is the **balanced binary accuracy** (to compensate for imbalances in the data):

$$S_{balanced} = \frac{1}{2}\left(\frac{P(\Omega^0 \cap \Omega_E^0)}{P(\Omega^0)} + \frac{P(\Omega^1 \cap \Omega_E^1)}{P(\Omega^1)}\right)$$

In our opinion, the most relevant accuracy measure is the cost-adjusted, or **cost-wise accuracy** score:

$$S_{cost\text{-}wise} = \mathbf{P}(\Omega^0 \cap \Omega_E^0) + \mathbf{P}(\Omega^1 \cap \Omega_E^1) = 1 - \mathbf{P}(\Omega^0 \cap \Omega_E^1) - \mathbf{P}(\Omega^1 \cap \Omega_E^0)$$

where **P** is the **cost distribution** (instead of case distribution): the weight of each case is equal to the cost that it will incur if wrongly classified.

Cost-wise accuracy

So we have: **Cost-wise accuracy = binary accuracy w.r.t. cost distribution** (instead of case distribution)

(*Cost-wise accuracy = balanced accuracy if the total cost of negatives is considered to be equal to the total cost of positives*)

For example, the cost of a non-mad cow is \$1 000, and the cost of a mad cow is \$10 000 000, which is 10000 times more than the cost of a non-mad cow. (*if a mad cow wrongly diagnosed as healthy, some people eat it and die, so the cost is extremely high*). In this case, the cost of 10 mad cows is about 10 times the cost of 9 990 non-mad cows. (*Mad cows are 10 times more cost-wise than non-mad cows*)

An expert who classifies every cow as mad still has a cost-wise accuracy of 91%. (*His recommendation to eliminate the whole herd of 10 000 cows is brutal but justified cost-wise*). If a test can detect all the mad cows as mad, plus also 5 000 non-mad cows as mad, then that test will have a cost-wise accuracy of about 96%, while its balanced accuracy is only 75%.

Let me now discuss the second topic. At first, assume that we have a group n completely **independent experts**, each with a (cost-wise) accuracy score $p > 1/2$ and **error rate** $q = 1 - p < 1/2$.

The distribution of the number of experts whose predictions are right is a binomial distribution $P(k) = C_n^k p^k q^{n-k}$ $(0 \leq k \leq n)$. By the central limit theorem, when n is large enough then this binomial distribution is approximately equal to the **normal distribution** with mean np and variance npq. It follows that the probability S_n of having at least $n/2$ correct predictions (*out of n experts*) is approximately

$$S_n \cong \Phi\left(\frac{\sqrt{n}(p - 1/2)}{\sqrt{pq}}\right) \quad \text{where} \quad \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-x^2/2} dx$$

is the cumulative distribution function of the Gaussian distribution. S_n is also the expected accuracy score of the collective decision by voting (*1 expert = 1 vote*). In particular, $\lim_{n\to\infty} S_n = 1$.

... with common blind spots

However, in practice, **we can't have truely independent experts**. They have **common blind spots**. *A fictive example: a Martian perfectly disguised as a human on Earth. No one can detect him.*

For simplicity, we consider a model with only two kinds of common blind spots: **half-blind** (random decisions), and **completely blind** (*everybody is brainwashed into believing in a lie*): the total set is divided into 3 parts

$$\Omega = \Omega_{blind} \cup \Omega_{hb} \cup \Omega_{l}$$

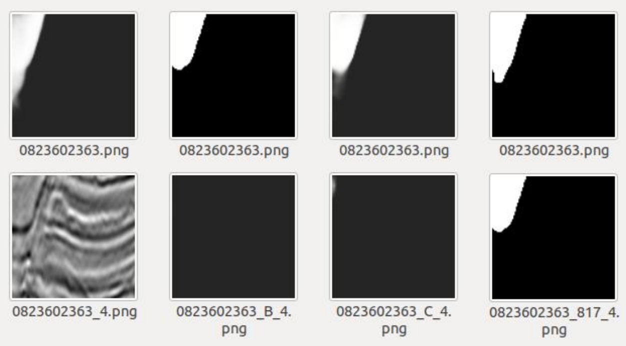
On Ω_{blind} everyone is wrong, on Ω_{hb} the decisions are like random, and on Ω_l (the learnable set) experts are independent and have error rate q (like in the previous slide). Then we have the following approximative formula for the accuracy score of the collective decision by voting:

$$S_n \cong \Phi\left(\frac{\sqrt{n}(p-1/2)}{\sqrt{pq}}\right)(1 - \mathbf{P}_{blind} - \mathbf{P}_{hb}) + \frac{1}{2}\mathbf{P}_{hb}$$

In particular, $\lim_{n\to\infty} S_n = 1 - \mathbf{P}_{blind} - \frac{1}{2}\mathbf{P}_{hb}$

Topological voting method

Let me mention now briefly the third topic. In a segmentation problem, different AI models may give different results, especially when the case is difficult.



Example: Segmentation of salt on a seismic image (shown on lower left corner), by some different models. (A Kaggle competition in 2018)

Topological voting method

Classical voting method: pixel-by-pixel, majority voting for each pixel.

Our **topological voting method** is as follows:
- Consider each proposed mask as a whole.
- Define distances $d(M_i, M_j)$ among the masks M_i.
- Take the mask which is closest to the others: M_k where

$$k = \mathrm{argmin}_i \sum_j d(M_i, M_j)$$

(Vote for the whole mask at once, not pixel-by-pixel) (*Vote for the whole team, not for individuals!*)

Claim: this works much better than the classical pixel-by-pixel voting method. (*Salt competition: our team participated in Kaggle for the first time, boosted the score from 0.84 to 0.87+ by topological voting, got a silver medal; top competitor 0.89*).

(*Explanation? Masks have logical topological structures. Pixel-by-pixel voting doesn't take into account such structures and may destroy them*).

THANK YOU FOR YOUR ATTENTION!