Yuanming Luo

IDEA AND THEORY

An important feature of the audio data is that the slicing of audio is also an audio. The slicing indeed increase the size of the data. But there is no free lunch. The sliced audio may not having the same genre label as the original song. A lot of songs have the segments of others mixed in. For example, rap are embedded in a lot of songs. If we just do the naive way to increase the size of the data by labeling the slices as the same as the original song, then we are feeding the model with incorrect data. There might be a huge percentage of data been misclassified.
How do minimize the loss while enlarge the data set?
This suggests that we cannot just slice the songs into m slices and hope the dataset that is m times larger can automaticly give you a better result. We need to filter the wrong one out. Here we introduce several methods that I thought about. Some of them are built upon each others.

(1) Majority Rule: We call this the mode strategy. Since we know that a lot of the slices have being labeled incorrectly, how do we still make the most correct decision on the non-sliced song? We know that a model using the incorrect data will perform bad. But most of, which we believed in, the data is still good. Let's say we have a portion of data p_m is correctly labeled where the footnote m is the number of slices that we slice one song into. Then what is the expected accuracy of using the mode to label one song? Now we have m random variables $X_1, X_2, ..., X_m$. Each of them have a label. Denote the true lable to be t then

$$E[I_t(X_i)] = p_m = \sum_{l \in L} I_t(X_i) P(X_i = l) = P(X_i = t)$$

where $I_t(x)$ is the indicator function which equals to 1 if $x = t$ is true and 0 otherwise. Let L be the set of labels. then let $N_l = \sum_{i=1}^{m} I_l(X_i)$ be the number of slices having label l. Then the mode is l that $N_l = \max_{l \in L} N_l$. Then for the prediction to be true, we need $N_t = \max_{l \in L} N_l$. The the probability of having the song to be predicted correctly is

$$P(N_t = \max_{l \in L} N_l) = \sum_{i=1}^{m} P(N_t = \max_{l \in L} N_l = i)$$

Notice that for $i \geq \frac{m}{2}$ we know that $P(N_t = \max_{l \in L} N_l = i) = P(N_t = i)$
What we also know is $P(N_t = i) = \binom{m}{i} p_m^i (1 - p_m)^{m-i}$

Then

$$P(N_t = \max_{i\in L} N_l) = \sum_{i=1}^{m} P(\max_{l\in L} N_l = i|N_t = i)P(N_t = i) = \sum_{i=1}^{m} P(\max_{l\in L} N_l = i|N_t = i)P(N_t = i)$$

$$= \sum_{i=1}^{\frac{m}{2}-1} P(\max_{l\in L} N_l = i|N_t = i)P(N_t = i) + \sum_{i=\frac{m}{2}}^{m} P(\max_{l\in L} N_l = i|N_t = i)P(N_t = i)$$

$$= \sum_{i=1}^{\frac{m}{2}-1} P(\max_{l\in L} N_l = i|N_t = i)P(N_t = i)+(1-F(\frac{m}{2}-1, m, p_m)) \geq 1-F(\frac{m}{2}-1, m, p_m)$$

But

$$\sum_{i=1}^{\frac{m}{2}-1} P(\max_{l\in L} N_l = i|N_t = i)P(N_t = i) \leq \sqrt{\sum_{i=1}^{\frac{m}{2}-1} P(\max_{l\in L} N_l = i|N_t = i)^2 \cdot \sum_{i=1}^{\frac{m}{2}-1} P(N_t = i)^2}$$

$$\leq \sqrt{(\frac{m}{2}-1)\sum_{i=1}^{\frac{m}{2}-1} P(N_t = i)^2}$$

$$P(N_t = \max_{i\in L} N_t) \leq \sqrt{(\frac{m}{2}-1)\sum_{i=1}^{\frac{m}{2}-1} P(N_t = i)^2} + 1 - F(\frac{m}{2}-1, m, p_m)$$