Final Report

Zhuqi Wang

August 1, 2014

1 Describing the problem

This problem is about finding error data and missing data in a given sample with a given template. The given template is an integer stream with length N, and each of its elements is in [0,4095]. The given sample is should be a similar stream with length M. The whole process could be described as: Imagine that a transmitter sends the template recurrently and only sends an element for each time. Notice that we may start sending data from each possible position of template stream, which leads to the problem more complicated. Then there is also a receiver receiving these sent data continually. And this is not a stable process which means we have chance to get some wrong data or miss some of them. And the p denotes missing rate and q denotes error rate.

2 Experiment Environment

Table 1: Experiment Env.

Env. Details	
Item	Detail
OS	Win7 64-bit
CPU	Inter(R) Core(TM) i5-3210M 2.50GHz
RAM	4.00GB(3.70GB is available)
Dev.Env.	C++(vs2012)

3 Algorithm

Our algorithm is based on the solution for edit distance problem, which takes dynamic programming strategy. The details about our solution is showed following,

Algorithm 1 Solution for detecting wrong and missing bit

Input:
Template Data **mfs**;
Sample Data **sample**;
Output:
The predicted operation;
1: Generate possible received stream;
2: Calculate edit distance between **sample** and all possible received stream;
3: Find the case where the edit distance is minimum;
4: Recover the edit operation;
5: **return** predicted operation;

The complexity of our algorithm is $O(nm^2)$ where n represents the length of template stream and m represents the length of sample data.

4 Experiment

I have conducted two experiments. The first one is focus on how length of stream impact on result. The other one is about how do miss rate and error rate impact on the result. And we mainly measured the precision and recall rate which are denoted by p, q. This part is based on **matplotlib**, a python 2D plotting library.

4.1 Corresponding to length

When we learn from these two graph. It is easy to find that the short template tends to lead higher precision, though a small part of data does not

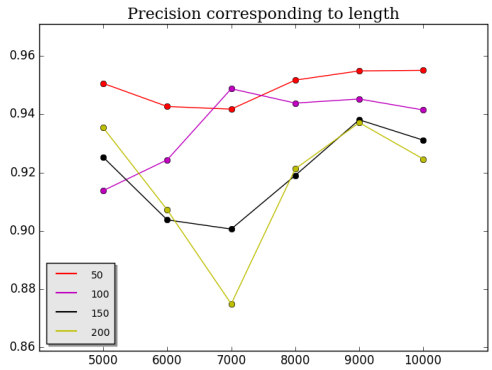

Figure 1: precision influenced by length

support this conclusion. I think it is because each data is the average result of 10 inference and ten times may not be large enough to get the stable result. However, does length of template have an impact on our algorithm is not definite, and we will take a discuss in some following sections.

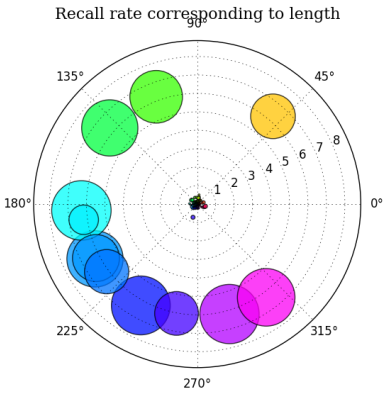

Figure 2: recall rate influenced by length

In addition, we can find a interesting phenomenon that recall rate data distributes in an extreme form. All recall rate tends to be zero or one, and it is hard to explain why this case occurs. And we will try to find the cause in the future work.

4.2 Corresponding to p and q

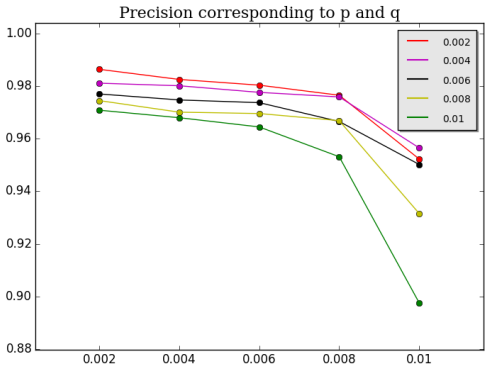

Figure 3: precision influenced by p&q

From **Figure.3**, we could easily find that precision increases apparently with the lower miss rate and error rate. It is reasonable while the lower miss rate and error rate will make the sample stream similar to the template, which makes it more likely to get correct result under edit-distance algorithm.

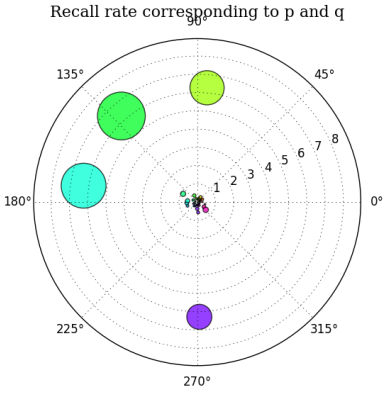

Figure 4: recall rate influenced by p&q

The same case occurs while there is almost no medium recall rate. Another

noticeable phenomenon is less high recall rate compared to **Figure.2**.

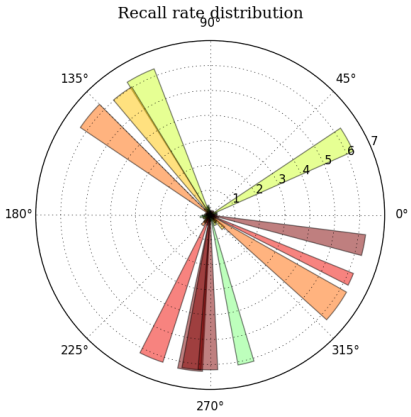

Figure 5: recall rate distribution

The **Figure.5** is the distribution of recall rate. Color is related to length of samples and quadrant is related to length of template. It is obvious that high recall rate rarely locates in the first quadrant which represents a 50-digit template. And in the second experiment, we fix the template to be a 50-digit template. When the template is short, it provides less features, which makes it difficult for us to recognize the incorrect bit. Otherwise, we choose the least edit distance as the most optimal solution, cause we don't know the miss rate and error rate. The most reasonable choice is the solution which makes a most proximate missing rate and error rate. However, this aim will never be achieved because of having no chance to know them in any case.

5 Future work

I think edit distance should not be the best strategy. In fact, this problem is a really intractable problem and our future work is to find a new method to increase the recall rate. In addition, the current algorithm is time-consuming and we should find a more efficient algorithm. If we want to apply it into practice, we should also design a time series algorithm for time-sequential data.