Texture Analysis in Multi-class Image Classification

Chao Wang

7/26/2020

Introduction

One tricky thing about colorectal cancer (CRC), one of the most prevalent cancer types, is that tumour architecture varies within the period of tumour progression and is also related to patient prognosis.

Although it is still indispensable to manually identify and classify different tissues in clinical routine, recent years has witnessed an eruption of automatic image processing and classification regarding texture analysis, which can greatly boost efficiency of clinical trial and save doctors' time for something more valuable.

Research development on this topic has evolved quickly from considering only two categories of tissues (tumour and stroma) to classifying up to eight types of texture images almost perfectly. As for the paper (https://www.nature.com/articles/srep27988) about eight-class texture analysis, there is still room for potential improvement, even though it has already achieved a reasonably good and impressive progress.

In this report, I proposed two main proposals with respect to a possible re-finement of the paper's work. On one hand, the paper adopted serveral distinct sets of texture descriptors mainly in a gesture to reduce the dimensionality and meanwhile maintain as much information such as variability and directions as possible. To name a few, histogram features such as mean, standard deviation and central moments represent the overall characteristics of images, whereas local binary patterns (LBP), a concept within computer vision, is referred to as the summary of images' local patterns including the consideration of different directions. As a result, I incorporated one more set of features, Local Phase Quantization (LPQ), which should render the performance even better.

On the other hand, the paper only applied four machine learning algorithms, signifying that I could try different methods and utilize emsemble in the end to yield more robust outcomes.

Material and Methods

1. Dataset

There were two types of dataset available, original images (250,000px) and compressed ones (4,096px). Given the time and hardware limit, I decided to choose the latter one first to get a basic idea of whether the improvement is feasible. Time permitting, the original data can be involved to produce a more professional research project. Before analyzing, the dataset was shuffled to ensure randomness.

The eight categories are: (a) tumour epithelium, (b) simple stroma, (c) complex stroma, (d) immune cell, (e) debris, (f) mucosal gland, (g) adipose tissue, (h) background. Each has 625 images.

2. Method: Histogram-Feature Analysis

2.1 Motivation Texture analysis has complicated images that usually require taking all pixel information into account, thus rendering direct gray-scale analysis inefficient. Hence, histogram-feature analysis is called

for, which can be viewed as a special case of principal component analysis but is based more on the image context itself and thus more efficient.

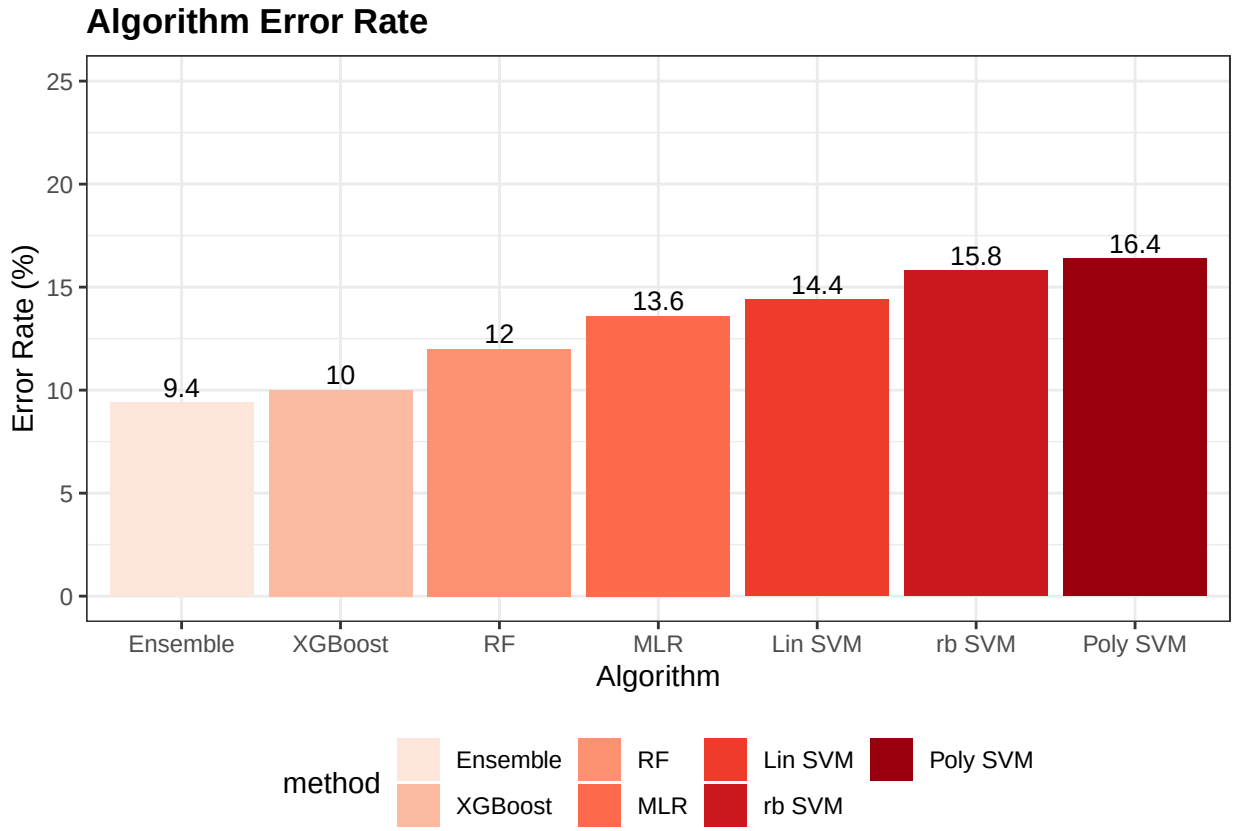
2.2 Overview

- Feature Set
  Instead of training on and analyzing pixel values of images, the histogram-feature analysis is interested in extracting specific image features via distinct approaches.
  - Lower-Order Histogram Features
    Lower-order statistics can be used to describe texture. In particular, this set contains the mean, variance, skewness, kurtosis, and the 5th central moment of the histogram.
  - Higher-Order Histogram Features
    The higher-order feature set consists of the central moments from 2nd to 11th. It is invariant to changes in the average gray-scale intensity of the input image because it does not contain the mean.
  - Local Binary Pattern Histogram Fourier Features (LBP-HF)
    The LBP-HF operator considers the probability of occurence of all possible binary patterns that arise from a neighbourhood of predefined shape and size, and applies Fourier transformation to prevent the negative effect of image rotations.
  - Local Phase Quantization (LPQ)
    LPQ, a texture descriptor that is robust to image blurring, utilizes decorrelated phase information computed locally in a window for every image position.
- Algorithm
  I adopted seven classification strategies: 1) Multinomial Logistic Regression, 2) Linear SVM, 3) Radial Basis SVM, 4) Polynomial SVM, 5) Random Forest, 6) XGBoost, 7) Ensemble.
  - Multinomial Logistic Regression (MLR)
    I used multinomial logistic regression as a baseline method so as to get a basic idea of how the features extracted performs.
  - Support Vector Machine (SVM)
    I employed support vector machine with one-vs-one class decisions in terms of linear SVM, radial basis SVM and polynomial SVM, with radial basis SVM being the winning method in the paper. The features were automatically standardized before training to maintain equal mean and variance.
  - Random Forest
    Random forest is also considered a fast and robust technique in which it is particularly effective on skewed data. Although it is not the case here, it is still worth trying given its quickness and robustness.
  - XGBoost
    Tree boosting is a highly effective and widely used machine learning method which can achieve state-of-the-art results on many machine learning challenges.
  - Ensemble
    The core idea of ensemble is to carry out a majority vote system for prediction, incorporating all results of the algorithms mentioned above. Theoretically, it is guaranteed to yield a better outcome.

Results

For this data set, I split out 90% of training set and 10% of test set, as it is relatively not a very large one in terms of each category. And then, by training all 7 machine learning algorithms over the combination of all 4 feature sets, here is the plot of error rates.

```
##
## Ensemble of All 4
## ================================================
##          Sensitivity Specificity Precision  F1
## ------------------------------------------------
## Class: 1    0.912       0.980       0.852   0.881
## Class: 2    0.853       0.970       0.817   0.835
## Class: 3    0.800       0.961       0.738   0.768
## Class: 4    0.862       0.995       0.966   0.911
## Class: 5    0.914       0.989       0.914   0.914
## Class: 6    0.910       0.993       0.953   0.931
## Class: 7    0.953         1           1     0.976
## Class: 8      1         0.998       0.984   0.992
## ------------------------------------------------
```



As is shown in the histogram, the accuracy is quite high with the best scoring over 90%. Other performance metrics, such as specificity and sensitivity, are all satisfactory according to the summary table.

Conclusion

This project demonstrates that histogram-feature analysis is fairly powerful when it comes to classifying complex images. The four feature sets, including lower-order, higher-order, LBP-HF and LPQ can work together for great achievement.

What's more, this idea is actually a better way to reduce dimensionality. No matter how many pixels an image have, they can always be converted into this frame required, thus boosting the development of not only health industry, but all other fields where complicated image classification is needed.

As for the limitation, the sample size is relatively small in this context, which limits the further study of the model.

Future work could be focused on looking for more efficient histogram feature sets and more powerful algorithms.