Kernel methods - Data challenge report

Othmane ZARHALI
othmanezarhalimaster@gmail.com

Rayan DHOUIB
rayan.dhouib@gmail.com

Abstract

The challenge consists in performing a multiclass classification based on kernel methods over images from 10 distinct classes where a non linear preprocessing was done over them. The sample consists of 5000 balanced training images, and the test sample contains 2000 images to be classified.
In the following we will explain the adopted methodology that consisted on multiclass classification algorithms that were developed without the use of an external machine learning library and with a progressing complexity.

1 Kernel functions

Along all the tests, the following kernel functions were used each time in a different setting that will be specified:

- **Linear kernel**

$$K(x,y) = x^T y \tag{1}$$

- **RBF kernel**

$$K(x,y) = exp\left(-\gamma ||x-y||^2\right), \gamma > 0 \tag{2}$$

- **Polynomial kernel**

$$K(x,y) = \left(x^T y\right)^{\alpha}, \alpha > 0 \tag{3}$$

- **GHI kernel**

$$K(x,y) = \sum_i \min\left(x[i]^{\beta}, y[i]^{\beta}\right), \beta > 0 \tag{4}$$

- χ^2 **kernel**

$$K(x,y) = exp\left(-\gamma \sum_i \frac{(x[i]-y[i])^2}{(x[i]+y[i])}\right), \gamma > 0 \tag{5}$$

2 Methodology

2.1 Brutforce multicass classification

This is based on a multiclass kernel support vector machine problem, formulated as follows:

$$\begin{array}{ll} \min\limits_{f_1,\dots,f_c \in \mathscr{H}_1 x \dots x \mathscr{H}_c} & \frac{1}{2}\sum\limits_{y=1}^{c} ||f_y||^2 \\ \text{s. t.} & f_{y_i}(x_i) - f_y(x_i) \geq 1, i = 1,\dots,n, \forall y \neq y_i \end{array}$$

Whose dual problem is of the form:

$$\begin{array}{ll} \min\limits_{w_1,\dots,w_c,\xi \in \mathscr{H}_1 x \dots x \mathscr{H}_c} & \frac{1}{2}\sum\limits_{y=1}^{c} ||f_y||^2 + C\sum\limits_{i=1}^{n} \xi_i \\ \text{s. t.} & f_{y_i}(x_i) - f_y(x_i) \geq \mathbb{1}_{y \neq y_i} - \xi_i, i = 1,\dots,n, \forall y \end{array}$$

Thus,

$$\hat{y} = \underset{y=1,\dots,c}{\text{argmax}}\, f_y(\mathbf{x})$$

Where **x** is the test data.
Without any feature extraction of the images, this approach gave a low score (≈ 0.1) either with the RBF kernel or the polynomial kernel on the leaderboard (tested over 50% of the test data).

2.2 Feature extraction: Scale-invariant feature transform

SIFT (Scale-invariant feature transform) is an algorithm presented by David G. LOWE used in computer vision to compare two images. The aim is to construct descriptors that enable to compare images between each others.
The first step is to find the key points of the image that mostly corre- spond to the corners and find the best scale of representation. One possible approach is to create a grid on the image. Each element of the grid serves as a reference point for a SIFT descriptor. Around this point, one begins by modifying the local

coordinate to guarantee the rotation invariance. Then, the retrieved histograms are concate- nated. In order to reduce the sensitivity of the descriptor to changes in brightness, the values are capped and the his- tograms are normalized.
Thus, **starting from the dataset of images we construct an alternative dataset of features that can be inputed in the multiclass classification algorithm**.

2.3 Alternative multiclass classification algorithm

The brutforce approach developed in the first subsection has two main caveats:

- Using the representer theorem, we end up with very big kernel matrices even for relatively small train datasets
- It's very time consuming

Those caveats are reflected in the score leaderboard. That's why a new approach is compulsory. One of the best ones are based on voting principles such as **one versus one (OVO) and one versus all (OVA)** classifiers. We have chosen to solve the underlying dual optimization problem using the sequential minimal optimization algorithm (**SMO**) developed by J. PLATT.

2.4 Parameters fine tunning

2.4.1 SIFT

After several trials, we ended up with an expressive set of SIFT parameters that were used to construct the alternative dataset (see code).

2.4.2 Kernel choice

Using the approach in 2.3, the first step was to test it with a vanilla kernel such as the linear kernel, using the OVO approach we obtain a leaderboard score of ≈ 0.50. This gave an insight about the performance of the approach adopted. The GHI kernel didn't contribute a lot in enhancing the score, whereas the χ^2 kernel with $\gamma = 0.6$ gave an accuracy of 0.992 on the train data and a leaderboard score of order ≈ 0.57. As a result, we ended up considering the χ^2 kernel with $\gamma = 0.6$.

2.4.3 Multiple kernel learning

Using the fact that the OVO χ^2 kernel with $\gamma = 0.6$ has enhanced the leaderboard score, we tried equally ponderated χ^2 kernels with respectively $\gamma = 0.6, 0.7, 1, 2$, which led to a leaderboard score of ≈ 0.603.

2.5 Data augmentation

We extended the size of the training dataset by applying a transformation to the images and keeping the same classification target for those images.
By considering a MKL compounded of the same χ^2 kernels as in the previous section we have slightly enhanced the score leaderboard to ≈ 0.6103.

You can find the code in the following repo:
https://github.com/othmanezarhalimaster/KernelMethodsDataChallenge

References

[1] J. Platt. 1998. Sequential minimal optimization: A fast algorithm for training support vector machines

[2] David G. Lowe 2003. Distinctive image features from scale-invariant key- points