Report II - Experiment Design in Computer Sciences

Matheus Silva de Lima[1]

Abstract

In this second report for the 2023 spring semester of Experiment Design in Computer Sciences, we analyze two different approaches for measuring similarity scores in a subspace-based image classification algorithm. We propose a new similarity function, called *scaled cosine similarity*, and show that it underperforms the classical *cosine similarity* using a paired t-test on various signal energy levels.

1. Introduction

Subspace-based methods for image classification are a class of algorithms that performs image classification based on simple linear algebra concepts. A basic algorithm for classifying images using subspaces can summarized as follows:

Training Phase

1. Take a set of M black-and-white labeled images $x \in \mathbb{R}^{\text{width}\times\text{height}}$. Vectorize then to $x \in \mathbb{R}^d$, $d = \text{width} \times \text{height}$.
2. Group each image from the training set by label. For each group, construct a matrix $X \in \mathbb{R}^{d\times r}$, arranging the r same labeled images in columns.
3. For each group, perform a Singular Value Decomposition (SVD)

$$X = U\Sigma V^T$$

where,

$$\Sigma = \text{diag}(\sigma_1^2\sigma_2^2\sigma_3^2...), \sigma_1^2 \geq \sigma_2^2 \geq \sigma_3^2 \geq ... \geq 0$$

.

4. Select a minimum energy threshold $0 < e_{\min} \leq 1$. Then, calculate the number of necessary columns n that captures an amount of energy $e \geq e_{\min}$ by

$$\min_n \frac{\sum_{i=1}^{n} \sigma_i^2}{\sum_{i=1}^{r} \sigma_i^2} \geq e_{\min}$$

5. Finally, select the n first columns of U as the bases ϕ_i of subspace Φ.

Evaluation phase

[1]Computer Vision Laboratory, University of Tsukuba

1. Take a vectorized image $x_{\text{eval}} \in \mathbb{R}^d$ and calculate the cosine similarity

$$\text{CS} = \sum_{i=1}^{n} \frac{(x, \phi_i)^2}{\|x\| \|\phi_i\|}$$

between the input image and each subspace.
2. Select the image class that yields the highest cosine similarity between subspace Φ and x_{eval} as the candidate.

In this report, we introduce another method for calculating similarities, called *scaled cosine similarity*, defined as

$$\text{SCS} = \sum_{i=1}^{n} \frac{\sigma_i^2}{\sum_{j=1}^{n} \sigma_j^2} \frac{(x, \phi_i)^2}{\|x\| \|\phi_i\|}$$

.

This similarity score uses singular values σ^2 as weights to the sum of squared cosines, the idea being that it automatically prioritizes the most significant eigenvectors.

In this report, we compare classification accuracy of both CS and SCS for various levels of minimum energy $0 > e_{\min} \geq 1$ on the EMNIST dataset. Section 2 describes the methodology, Section 3 presents the analysis results and Section 4 has a concise discussion about the findings. Finally, we finish the report.

2. Methodology

For this experiment, we test the classification accuracy of both methods on the EMNIST[1] dataset. We apply a bootstrapping procedure, randomizing the dataset and selecting a small portion (1/100) of it at each iteration, for both training and evaluation sets. This way, we generate an accuracy random variable that approximately follows a normal distribution, while reducing the processing cost of the analysis.

We calculate the necessary number of different energy levels N in order to secure an experimental result with confidence level $(1 - \alpha) \geq 0.95$ and experimental power $(1 - \beta) \geq 0.8$. In order to do so, we first conduct a pilot study in order to measure the standard deviation within groups std_{CS}, std_{CS}, and the standard deviation of the difference $\text{std}_{\text{CS}} - \text{std}_{\text{SCS}}$. As a minimum interesting effect, we select $\delta^* \geq 0.02$, that is, an accuracy difference smaller than 2% is not considered relevant to our experiment.

We then perform a one-sided paired t-test on the different energy levels, considering the Null Hypothesis H_0 that the mean classification accuracy of $\text{CS} \leq \text{SCS}$, and alternate hypothesis $H_1 = \text{CS} > \text{SCS}$.

Finally, we check the assumption of data normality using a Q-Q plot.

3. Results

For the pilot study, we select an energy level of $e_{\text{min}} = 0.5$ and 10 repetitions of the bootstrapping procedure. We get the following results:

1. $\text{std}_{\text{CS}} = 0.013$
2. $\text{std}_{\text{SCS}} = 0.017$
3. $\text{std}_{(\text{CS}-\text{SCS})} = 0.015$
4. $\text{std}_{\frac{\text{SCS}}{\text{CS}}} = 1.33$

With this preliminary result, we calculate the total number of experimental levels N for a one-sided t-test

$$N \geq 5.12$$

We select $N = 10$ levels, those being

$$e_{\text{min}} = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$$

With $N = 10$, the test has a power of $0.98 > 0.8$. Next, we run the experiment on those energy levels. A graphic representation of the obtained results can be seen on Figure 1.

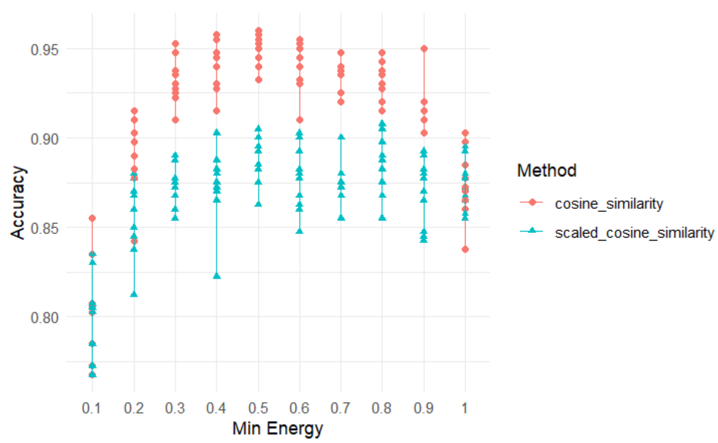

Figure 1: Experimental data by energy levels.

We perform a paired t-test by energy level on the mean of experimental data. Results can be seen in Table 1.

These results lead us to reject the H_0: the true accuracy mean is not equal for both methods. The SCS method under-performs the classical CS method with a difference mean of $4.3\% > \delta^*$. However, we still need to check the assumption of data normality.

For this, we run a Q-Q plot normality test, which can be seen in Figure 2:

The Q-Q normality test shows evidence that our data slightly deviates from a normal distribution. However, as the t-test is relatively robust against small deviations of non-normality, we choose to proceed with the analysis.

Parameter	Value
t	5.8048
df	9
p-value	0.000129
95% confidence interval	$(0.02998529, \infty)$
mean difference	0.043825

Table 1: Results from t-test.

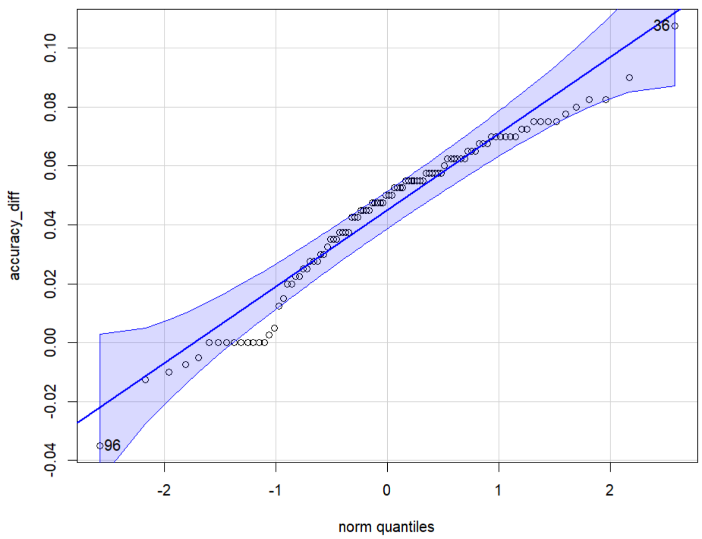

Figure 2: Q-Q plot.

4. Discussion

These results can be seen at first as surprising, as intuitively, it should make sense to give a higher priority to the most important eigenvectors in a similarity score calculation. However, upon a closer look, lower scored eigenvectors also contain important information about the data that is less frequently seen, and can be extremely important for the correct classification. An advantage seen by the SCS compared to CS score is that it quickly plateaus after a given energy level, while the CS score improves to a certain point, and later declines. This happens because the SCS scores tends to remove the noisy data generated by eigenvectors of very-small singular values, and we can expect it to perform better when considering an energy level of $e_{\min} \to 1$.

References

[1] G. Cohen, S. Afshar, J. Tapson, A. van Schaik, EMNIST: an extension of MNIST to handwritten letters, CoRR abs/1702.05373. arXiv:1702.05373. URL http://arxiv.org/abs/1702.05373