# THE WONDERS OF THE NORMALIZED COMPRESSION DISSIMILARITY REPRESENTATION

*Emanuele Olivetti*

NeuroInformatics Laboratory (NILab), Bruno Kessler Foundation, Trento, Italy
Center for Mind and Brain Sciences (CIMeC), University of Trento, Italy

## ABSTRACT

We propose a method to effectively embed general objects, like audio samples, into a vectorial feature space, suitable for classification problems. From the practical point of view, the researcher adopting the proposed method is just required to provide two ingredients: an efficient *compressor* for those objects, and a way to *combine* two objects into a new one. The proposed method is based on two main elements: the dissimilarity representation and the normalized compression distance (NCD). The dissimilarity representation is an Euclidean embedding algorithm, i.e. a procedure to map generic objects into a vector space, which requires the definition of a distance function between the objects. The quality of the resulting embedding is strictly dependent on the choice of this distance. The NCD is a distance between objects based on the concept of Kolmogorov complexity. In practice the NCD is based on two building blocks: a compression function and a method to combine two objects into a new one. We claim that, as soon as a good compressor and a meaningful way to combine two objects are available, then it is possible to build an effective feature space in which classification algorithms can be accurate. As our submission to the IEEE AASP Challenge, we show a practical application of the proposed method in the context of acoustic scene classification where the compressor is the free and open source *Vorbis* lossy audio compressor and the combination of two audio samples is their simple *concatenation*.

***Index Terms*—** Dissimilarity Representation; Kolmogorov Complexity; Compression; Classification

## 1. INTRODUCTION

In a given domain of application, e.g. the classification of audio segments, the availability of efficient lossy compression algorithms means that efficient models are known to separate the signal from the noise. An efficient lossy compressor preserves the most of the signal and discards most of the noise.

In this work we show how an efficient compressor can be used as a building block for accurate classification of generic objects. We propose a procedure to create an effective vectorial representation of acoustic scenes recordings that is essentially based on *just on the length* of compressed audio files. We prove that this vectorial representation preserves most of the scene information within each audio recording by showing that a classifier constructed on this feature space accurately predict the scene content of the recording.

This document is licensed under the Creative Commons Attribution 3.0 License (CC BY 3.0).
http://creativecommons.org/licenses/by/3.0/
© 2013 The Authors.

The theoretical framework of this approach lies in the in the concept of Kolmogorov Complexity [1]. The Kolmogorov complexity of an object $x$, is the shortest binary program that outputs $x$. The Kolmogorov complexity is not computable in practice, but recently it has been shown that approximations can be using compression algorithms [2]. From the idea of approximating the Kolmogorov complexity of an object with the length of compressed object, e.g. in bytes in case of files, a parameter-free universal distance called *normalized compression distance* (NCD) has been proposed [2]. This distance has been shown to be an effective one in hierarchical clustering problems across many different domain of application [2].

In the pattern recognition literature it has been shown that, given a (even non-metric) distance function, it possible to create effective vectorial representations of general objects by means of an Euclidean embedding technique called *dissimilarity representation* [3]. The dissimilarity representation maps an object to the vector space where each dimension is the distance of that objects from a given set of objects, called *prototypes* or *landmarks*. This representation has been shown to be effective in many different domain of application [3, 4].

In the following we present a brief formal introduction to the theoretical elements of the proposed approach, namely the NCD and the dissimilarity representation. Then, as support to our claims, we show experimental results on the public dataset of the acoustic scene classification IEEE AASP challenge 2013. This work is also the description of our submission to that challenge.

## 2. METHODS

In the following we briefly describe the dissimilarity representation and the normalized compression distance.

### 2.1. The Dissimilarity Representation

The dissimilarity representation [3] is a lossy Euclidean embedding algorithm that maps general objects $X \in \mathcal{X}$ into $\mathbb{R}^p$. The dissimilarity representation is defined by the transformation function $\phi(X)_{\Pi}^{d}$ s.t.:

$$\phi(X)_{\Pi}^{d} = [d(X, \tilde{X}_1), \ldots, d(X, \tilde{X}_p)] \quad (1)$$

where $d : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is a given distance function and $\Pi = \{\tilde{X}_1, \ldots, \tilde{X}_p)\}$, $\tilde{X}_i \in \mathcal{X}$, is a given set of objects called *prototypes* or *landmarks*.

The quality of the embedding is highly dependent on the choice of the distance function $d$ and the choice of the prototypes $\Pi$ [5]. When constructing the dissimilarity representation for a given dataset $D$, the choice of the prototypes is usually a carefully

selected subset of the available dataset (see [5]). In case of dataset with small sample size, like the one of the acoustic scene classification of the IEEE AASP challenge, the prototypes can be the whole dataset, i.e. $\Pi = D$.

### 2.2. The Normalized Compression Distance

Given two generic objects $x, y \in \mathcal{X}$, the Kolmogorov complexity of $x$ given $y$, denoted as $K(x|y)$, is the length of the shortest binary program, for the reference universal prefix Turing machine, that, on input $y$, outputs $x$ (see [2, 1]). The Kolmogorov complexity of $x$, denoted as $K(x)$, is then the length of the shortest binary program that, with no input, outputs $x$.

The normalized information distance (see [6])

$$NID(x, y) = \frac{\max\{K(x|y), K(y|x)\}}{\max\{K(x), K(y)\}} \quad (2)$$

represents a metric distance between $x$ and $y$ in terms of the Kolmogorov complexity. This distance is not computable but an approximation of it, called normalized compression distance (NCD) was proposed in [2]:

$$NCD(x, y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max(C(x), C(y))} \quad (3)$$

where $C : \mathcal{X} \mapsto \mathbb{N}^+$ is a compression function that takes an object and returns its code-word length, e.g. its size in bytes.

## 3. EXPERIMENTS

We tested the proposed approach of creating a vectorial feature space with the dissimilarity representation and the NCD on the public dataset of the acoustic scene classification IEEE AASP challenge [1] which comprises 10 different scenes (bus, busystreet, office, openairmarket, park, quietstreet, restaurant, supermarket, tube, tubestation) with 10 audio recordings each.

We created a vectorial representation of each audio track with the dissimilarity representation using the whole dataset of 100 audio tracks as prototypes and the NCD as the distance $d$. The NCD was based on the open source Vorbis lossy audio compression codec [2] set at the highest quality [3]. The combination of two audio tracks was defined as the simple concatenation of them [4]. The final dataset resulted in a $100 \times 100$ matrix where each row was the vector of the NCD distance of a given audio track to each of the 100 audio tracks of the dataset. The computation required approximately 4 hours on a standard desktop computer.

In order to test the efficacy of the proposed feature space we estimated the classification accuracy of a classifier with a 10-fold cross-validation procedure. As the classification algorithm we adopted the Random Forest as proposed in [7] and implemented in `scikit-learn` [5] [8] on the whitened data matrix. The Random Forest classifier is an ensemble method which combines classification trees and it is suitable for non-linear problems. We set the number of the base learners, i.e. the size of the ensemble, to 1000 and all the other parameters as to their default values. The 10-fold stratified cross-validated accuracy was $80\%$, where the chance accuracy was $10\%$.

[1] `http://c4dm.eecs.qmul.ac.uk/sceneseventschallenge/`

[2] `http://www.vorbis.com`

[3] `oggenc -q 10 <file.wav> -o compressed.ogg`

[4] `sox <file1.wav> <file2.wav> combined.wav`

[5] `http://scikit-learn.org`

## 4. DISCUSSION

The high accuracy of the classifier, i.e. $80\%$ in a 10-class problem, shows that the proposed vectorial feature space, created with the dissimilarity representation and the NCD, is effective for discriminating the scene content from audio tracks. A striking aspect of the experimental results is that, with the exception of the choice of the Vorbis compressor and the choice of concatenating audio tracks, no domain knowledge was used to create the feature space and the classification system. These facts support the claim that the proposed approach is general and they show that efficient compressors can be used as inferential tools in supervised problems, similarly to what was previously shown for unsupervised problems (see [2]). A systematic analysis of other lossy audio compression algorithms and other ways to combine two tracks into a new one will be investigated as future work.

## 5. REFERENCES


[1] M. Li and P. M. B. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications (Texts in Computer Science)*, 3rd ed. Springer, Nov. 2008. [Online]. Available: http://www.worldcat.org/isbn/0387339981

[2] R. Cilibrasi and P. M. B. Vitanyi, "Clustering by Compression," *IEEE Transactions on Information Theory*, vol. 51, no. 4, pp. 1523–1545, Apr. 2005. [Online]. Available: http://dx.doi.org/10.1109/tit.2005.844059

[3] E. Pekalska, P. Paclik, and R. P. W. Duin, "A generalized kernel approach to dissimilarity-based classification," *J. Mach. Learn. Res.*, vol. 2, pp. 175–211, 2002. [Online]. Available: http://portal.acm.org/citation.cfm?id=944810

[4] E. Olivetti, T. B. Nguyen, and E. Garyfallidis, "The Approximation of the Dissimilarity Projection," in *IEEE International Workshop on Pattern Recognition in NeuroImaging*. IEEE, 2012.

[5] E. Pekalska, R. Duin, and P. Paclik, "Prototype selection for dissimilarity-based classifiers," *Pattern Recognition*, vol. 39, no. 2, pp. 189–208, Feb. 2006. [Online]. Available: http://dx.doi.org/10.1016/j.patcog.2005.06.012

[6] M. Li, X. Chen, X. Li, B. Ma, and P. M. B. Vitanyi, "The similarity metric," *Information Theory, IEEE Transactions on*, vol. 50, no. 12, pp. 3250–3264, Dec. 2004. [Online]. Available: http://dx.doi.org/10.1109/tit.2004.838101

[7] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001. [Online]. Available: http://dx.doi.org/10.1023/a:1010933404324

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, Oct. 2011. [Online]. Available: http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf