An IFS-based Similarity Measure to Index Electroencephalograms

Ghita Berrada and Ander de Keijzer

MIRA - institute for biomedical technology and technical medicine
University of Twente, PO Box 217,7500 AE
Enschede, The Netherlands
{g.berrada,a.dekeijzer}@utwente.nl

Abstract. EEG is a very useful neurological diagnosis tool, inasmuch as the EEG exam is easy to perform and relatively cheap. However, it generates large amounts of data, not easily interpreted by a clinician. Several methods have been tried to automate the interpretation of EEG recordings. However, their results are hard to compare since they are tested on different datasets. This means a benchmark database of EEG data is required. However, for such a database to be useful, we have to solve the problem of retrieving information from the stored EEGs without having to tag each and every EEG sequence stored in the database (which can be a very time-consuming and error-prone process). In this paper, we present a similarity measure, based on iterated function systems, to index EEGs.

Keywords: clustering, indexing, electroencephalograms (EEG), iterated function systems (IFS)

1 Introduction

An electroencephalogram (EEG) captures the brain's electric activity through several electrodes placed on the scalp[1]. The result is a multidimensional time series[2]. An EEG signal can be classified into several types of cerebral waves characterised by their frequencies, amplitudes, morphology, stability, topography and reactivity. The interpretation of the sequence of cerebral waves, their localisation and context of occurrence (eg eyes closed EEG or sleep EEG) leads to a diagnosis. The complexity of the sequences of cerebral waves, the non-specificity of EEG recordings (for example, without any context being given, the EEG recording of a chewing artifact can be mistaken as that of a seizure (see figure 1)) and the amount of data generated make the interpretation process a difficult, time-consuming and error-prone one. Consequently, the interpretation process is being automated, in part at least, through several methods mostly consisting in extracting features from EEGs and applying classification algorithms to the sets

[1] usually 21 in the International 10/20 System
[2] 19 channels in the International 10/20 System

of extracted features to discriminate between two different patient states (usually the "normal" state and a pathological state). For example, empirical mode decomposition and Fourier-Bessel expansion are used in [13] to discriminate between ictal EEGs (i.e EEGs of an epileptic seizure) and seizure-free EEGs. The

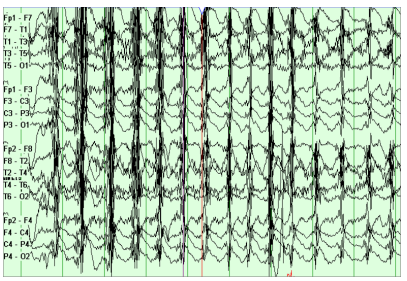


Fig. 1. EEG with a chewing artifact

interpretation methods are usually tested on different datasets. To make them comparable, a benchmark database of EEGs is required. Such a database has to designed so as to be able to handle queries in natural language such as the following sample queries:

1. find EEGs of non-convulsive status epilepticus
2. find EEGs showing rhythms associated with consumption of benzodiazepines and remove all artefacts from them

Obtaining a simple answer to this set of queries would require the EEG dataset to be heavily and precisely annotated and tagged. But what if the annotations are scarce or not available? Furthermore, the whole process of annotating and tagging each and every sequence of the EEG dataset is time-consuming and error-prone. This means that feature extraction techniques are necessary to solve all of these queries since they can help define a set of clinical features representative of a particular pathology (query 1) or detect particular sets of patterns and process the EEG based on them (query 2). EEG recordings correspond to very diverse conditions (eg. "normal" state, seizure episodes, Alzheimer disease). Therefore, a generic method to index EEGs without having to deal with disease-specific features is required[3]. Generic methods to index time series often rely on the definition of a similarity measure. Some of the similarity measures proposed include a function interpolation step, be it piecewise linear interpolation or interpolation with AR (as in [8] to distinguish between normal EEGs and EEGs originating from the injured brain undergoing transient global ischemia) or ARIMA models, that can followed by a feature extraction step (eg. computation of LPC cepstral coefficients from the ARIMA model of the time series as in [9]). However,

[3] as the number of disease-specific classifiers grows exponentially

ARIMA/AR methods assume that the EEG signal is stationary, which is not a valid assumption. In fact, EEG signals can only be considered as stationary during short intervals, especially intervals of normal background activity, but the stationarity assumption does not hold during episodes of physical or mental activity, such as changes in alertness and wakefulness, during eye blinking and during transitions between various ictal states. Therefore, EEG signals are quasi-stationary. In view of that, we propose a similarity measure based on IFS interpolation to index EEGs in this paper, as fractal interpolation does not assume stationarity of the data and can adequately model complex structures. Moreover, using fractal interpolation makes computing features such as the fractal dimension simple (see theorem 21 for the link between fractal interpolation parameters and fractal dimension) and the fractal dimension of EEGs is known to be a relevant marker for some pathologies such as dementia (see [7]).

2 Background

2.1 Fractal interpolation

Fractal dimension Any given time series can be viewed as the observed data generated by an unknown manifold or attractor. One important property of this attractor is its fractal dimension. The fractal dimension of an attractor counts the effective number of degrees of freedom in the dynamical system and therefore quantifies its complexity. It can also be seen as the statistical quantity that gives an indication of how completely a fractal object appears to fill space, as one zooms down to finer and finer scales. Another dimension, called the topological dimension or Lebesgue Covering dimension, is also defined for any object and a fortiori for the attractor. A space has Lebesgue Covering dimension n if for every open cover [4] of that space, there is an open cover that refines it such that the refinement [5] has order at most $n+1$. For example, the topological dimension of the Euclidean space $\mathbb{R}^n$ is n. The attractor of a time series can be fractal (ie its fractal dimension is higher than its topological dimension) and is then called a strange attractor. The fractal dimension is generally a non-integer or fractional number. Typically, for a time series, the fractal dimension is comprised between 1 and 2 since the (topological) dimension of a plane is 2 and that of a line is 1. The fractal dimension has been used to:

- uncover patterns in datasets and cluster data ([10, 2, 15])
- analyse medical time series ([14, 6]) such as EEGs ([1, 7])
- determine the number of features to be selected from a dataset for a similarity search while obviating the "dimensionality curse" ([12])

[4] A covering of a subset S is a collection $\mathcal{C}$ of open subsets in X whose union contains all of S at least. A subset $\mathcal{S} \subset X$ is open if it is an arbitrary union of open balls in X. This means that every point in $\mathcal{S}$ is surrounded by an open ball which is entirely contained in X. An open ball in a in metric space X is defined as a subset of X of the form $B(x_0, \epsilon) = \{x \in X | d(x, x_0) < \epsilon\}$ where x_0 is a point of X and ϵ a radius.

[5] A refinement of a covering $\mathcal{C}$ of S is another covering $\mathcal{C}'$ of S such that each set B in $\mathcal{C}'$ is contained in some set A in $\mathcal{C}$

Iterated function systems We denote as $\mathbf{K}$ a compact metric space for which a distance function d is defined and as $\mathbb{C}(\mathbf{K})$ the space of continuous functions on $\mathbf{K}$. We define over $\mathbf{K}$ a finite collection of mappings $\mathbb{W} = w_{ii\in[1,n]}$ and their associated probabilities $pi_{i\in[1,n]}$ such that

$$p_i \geq 0 \quad \text{and} \quad \sum_{i=1}^{n} p_i = 1$$

We also define an operator T on $\mathbb{C}(\mathbf{K})$ as $(Tf)(x) = \sum_{i=1}^{n} p_i(f \circ w_i)(x)$. If T maps $\mathbb{C}(\mathbf{K})$ into itself, then the pair (w_i, pi) is called an iterated function system on $(\mathbf{K}, d)$. The condition on T is satisfied for any set of probabilities p_i if the transformations w_i are contracting, in other words, if, for any i, there exists a $\delta_i < 1$ such that: $d(w_i(x), w_i(y)) \leq \delta_i d(x, y) \quad \forall x, y \in K$. The IFS is also denoted as hyperbolic in this case.

Principle of fractal interpolation If we define a set of points $(x_i, F_i) \in \mathbb{R}^2 : i = 0, 1, ..., n$ with $x_0 < x_1 < ... < x_n$, then an interpolation function corresponding to this set of points is a continuous function $f : [x_0, x_n] \to \mathbb{R}$ such that $f(x_i) = F_i \quad$ for $i \in [0, n]$. In fractal interpolation, the interpolation function is often constructed with n affine maps of the form:

$$w_i \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a_i & 0 \\ c_i & d_i \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e_i \\ f_i \end{pmatrix} \quad i = 1, 2, ..., n$$

where d_i is constrained to satisfy: $-1 \leq d_i \leq 1$. Furthermore, we have the following constraints:

$$w_i \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} = \begin{pmatrix} x_{i-1} \\ y_{i-1} \end{pmatrix} \quad \text{and} \quad w_i \begin{pmatrix} x_n \\ y_n \end{pmatrix} = \begin{pmatrix} x_i \\ y_i \end{pmatrix}$$

After determining the contraction parameter d_i, we can estimate the four remaining parameters (namely a_i,c_i,e_i,f_i):

$$a_i = \frac{x_i - x_{i-1}}{x_n - x_0} \tag{1}$$

$$c_i = \frac{x_n x_{i-1} - x_0 x_i}{x_n - x_0} \tag{2}$$

$$e_i = \frac{y_i - y_{i-1}}{x_n - x_0} - d_i \frac{y_n - y_0}{x_n - x_0} \tag{3}$$

$$f_i = \frac{x_n y_{i-1} - x_0 y_i}{x_n - x_0} - d_i \frac{x_n y_0 - x_0 y_n}{x_n - x_0} \tag{4}$$

d_i can be determined using the geometrical approach given in [11]. Let t be a time-series with end-points (x_0, y_0) and (x_n, y_n), and (x_p, y_p) and (x_q, y_q) two consecutive interpolation points so that the map parameters desired are those defined for w_p. We also define α as the maximum height of the entire function measured from the line connecting the end-points (x_0, y_0) and (x_n, y_n) and β as the maximum height of the curve measured from the line connecting (x_p, y_p) and (x_q, y_q). α and β is positive (respectively negative) if the maximum value is reached above the line (respectively below the line). The contraction factor d_p is then defined as $\frac{\beta}{\alpha}$. This procedure is also valid when the contraction factor is computed for an interval instead of for the whole function. The end-points are then taken as being the end-points of the interval. For more details on fractal interpolation, see [3, 11].

Estimation of the fractal dimension from a fractal interpolation The theorem that links the fractal interpolation function and its fractal dimension is given in [3].The theorem is as follows:

Theorem 21 *Let n be a positive integer greater than 1, $\{(x_i, F_i) \in \mathbb{R}^2 : i = 1, 2, ..., n\}$ a set of points and $\{\mathbb{R}^2; w_i, i = 1, 2, .., n\}$ an IFS associated with the set of points where:*

$$w_i \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a_i & 0 \\ c_i & d_i \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e_i \\ f_i \end{pmatrix} \quad for \quad i = 1, 2, .., n.$$

The vertical scaling factors d_i satisfy $0 \leq d_i < 1$ and the constants a_i,c_i,e_i and f_i are defined as in section 2.1 (in equations 1,2,3 and 4) for $i = 1, 2, ..., n$. We denote G the attractor of the IFS such that G is the graph of a fractal interpolation function associated with the set of points.
If $\sum_{i=1}^{n} |d_i| > 1$ and the interpolation points do not lie on a straight line, then the fractal dimension of G is the unique real solution D of $\sum_{i=1} |d_i| a_i^{D-1} = 1$.

2.2 K-medoid clustering

An $m \times m$ symmetric similarity matrix S can be associated to the EEGs to be indexed (with m being the number of EEGs to index):

$$S = \begin{pmatrix} d_{11} & d_{12} & \dots & d_{1m} \\ d_{12} & d_{22} & \dots & d_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ d_{1m} & d_{2m} & \dots & d_{mm} \end{pmatrix} \quad \text{where } d_{nm} \text{ is the distance between EEGs } n \text{ and } m \tag{5}$$

Given the computed similarity matrix S (defined by equation 5), we can use the k-medoids algorithm to cluster the EEGs. This algorithm requires the number of clusters k to be known. We describe our choice of the number of clusters below, in section 2.3. The k-medoids algorithm is similar to k-means and can be applied through the use of the EM algorithm. k random elements are, initially, chosen as representatives of the k clusters. At each iteration, a representative element of a cluster is replaced by a randomly chosen nonrepresentative element of the cluster if the selected criterion (e.g. mean-squared error) is improved by this choice. The data points are then reassigned to their closest cluster, given the new cluster representative elements. The iterations are stopped when no reassignments is possible. We use the PyCluster function `kmedoids` described in [5] to make our k-medoids clustering.

2.3 Choice of number of clusters

The number of clusters in the dataset is estimated based on the similarity matrix obtained following the steps in section 3 and using the method described in [4]. The method described in [4] takes the similarity matrix and outputs a vector called envelope intensity associated to the similarity matrix. The number of distinct regions in the plot of the envelope intensity versus the index gives an estimation of the number of clusters. For details on how the envelope intensity vector is computed, see [4].

3 An IFS-based similarity measure

3.1 Fractal interpolation step

We interpolate each channel of each EEG (except the annotations channel) using piecewise fractal interpolation. For this purpose, we split each EEG channel into windows and then estimate the IFS for each window. The previous description implies that a few parameters, namely the window size and therefore the embedding dimension, have to be determined before estimating the piecewise fractal interpolation function for each channel. The embedding dimension is determined thanks to Takens' theorem which states that, for the attractor of a time series to be reconstructed correctly (i.e the same information content is found in the state (latent) and observation spaces), the embedding dimension denoted m satisfies : $m > 2D + 1$ where D is the dimension of the attractor, in other words its fractal dimension. Since the fractal dimension of a time series is between 1 and 2, we can get a satisfactory embedding dimension as long as $m > 2 * 2 + 1$ i.e $m > 5$. We therefore choose an embedding dimension equal to 6. And we choose the lag τ between different elements of the delay vector to be equal to the average duration of an EEG data record i.e 1s. Therefore, we split our EEGs in (non-overlapping) windows of 6 seconds. A standard 20-minutes EEG (which therefore contains about 1200 data records of 1 second) would then be split in about 200 windows of 6 seconds. Each window is subdivided into intervals of one second each and the end-points of these intervals are taken as interpolation points. This means there are 7 interpolation points per interval: the starting point p_0 of the window, the point one second away from p_0, the point two seconds from p_0, the point three seconds away from p_0, the point four seconds away from p_0, the point five seconds away from p_0 and the last point of the window. The algorithm[6] to compute the fractal interpolation function per window is as follows:

1. Choose, as an initial point, the starting point of the interval considered (the first interval considered is the interval corresponding to the first second of the window).
2. Choose, as the end point of the interval considered, the next interpolation point.
3. Compute the contraction factor d for the interval considered.
4. If $|d| > 1$ go to 2, otherwise go to 5.
5. Form the map w_i associated with the interval considered. In other words, compute the a, c, e and f parameters associated to the interval (see equations). Apply the map to the entire window (i.e six seconds window) to yield $w_i \begin{pmatrix} x \\ y \end{pmatrix}$ for all x in the window. 6. Compute and store the distance between the original values of the time series on the interval considered (i.e the interval constructed in steps 2 and 3) and the values given by w_i on that interval. A possible distance is the Euclidean distance.

[6] inspired from [11]

6. Go to 2 until the end of the window is reached.
7. Store the interpolation points and contraction factor which yield the minimum distance between the original values on the interval and the values yielded by the computed map under the influence of each individual map in steps 5 and 6.
8. Repeat steps from 1 to 8 for each window of the EEG channel.
9. Apply steps 1 to 9 to all EEG channels.

3.2 Fractal dimensions estimation

After this fractal interpolation step, each window of each signal is represented by 5 parameters instead of by `signal frequency.window duration` points. The dimension of the analysed time series is therefore reduced in this step. For a standard 20-minutes EEG containing 23 signals of frequency 250 Hz, this amounts to representing each signal with 1000 values instead 50000 and the whole EEG with 23000 values instead of 1150000, thus to reducing the number of signal values by almost 98%. This dimension reduction may be exploited in future work to compress EGGs and store compressed representations of EEGs in the database instead of raw EEGs as the whole EEGs can be reconstructed from their fractal interpolations. Further work needs to be done on the compression of EEG data using fractal interpolation and the loss of information that may result from this compression. Then, for each EEG channel and for each window, we compute the fractal dimension thanks to theorem 21. The equation of theorem 21 is solved heuristically for each 6-second interval of each EEG signal using a bisection algorithm. As we know that the fractal dimension for a time series is between 1 and 2, we search a root of the equation of theorem 21 in the interval [1,2] and split the search interval by half at each iteration until the value of the root is approached by an ϵ-margin (ϵ[7] being the admissible error on the desired root). Therefore, for each EEG channel, we have the same number of computed fractal dimensions as the number of windows. This feature extraction extraction step (fractal dimension computations) further reduces the dimensionality of the analysed time series. In fact, the number of values representing the time series is divided by 5 in this step. This leads to representing a standard 20-minute EEG containing 23 signals of frequency 250 Hz by 4600 values instead of the initial 1150000 points.

3.3 Similarity matrix computation

We only compare EEGs that have at least a subset of identical channels (i.e having the same labels). When two EEGs don't have any channels (except the annotations channel) in common, the similarity measure between them is set to 1 (as the farther (resp. closer) the distance between two EEGs, the higher (resp. lower) and the closer to 1 (resp. closer to 0) the similarity measure). If, for the two EEGs compared, the matching pairs of feature vectors (i.e vectors made of

[7] We choose $\epsilon = 0.0001$ in our experiments

the fractal dimensions computed for each signal) do not have the same dimension then the vector of highest dimension is approximated by a histogram and the M most frequent values according to the histogram (M being the dimension of the shortest vector) are taken as representatives of that vector and the distance between the two feature vectors is approximated by the distance between the shortest feature vector and the vector formed with the M most frequent values of the longest vector. The similarity measure between two EEGs is given by: $\sum_{i=1}^{N} \frac{1}{N} \frac{d(ch_i^{EEG_1}, ch_i^{EEG_2}) - d_{min}}{d_{max} - d_{min}}$
where N is the number of EEG channels, $d(ch_i^{EEG_1}, ch_i^{EEG_2})$ the distance between the fractal dimensions extracted from channels with the same label in the two EEGs compared and d_{min} and d_{max} respectively the minimum and maximum distances between two EEGs in the analysed set. We choose as metrics (d) the Euclidean distance and the normalized mutual information.

4 Description of the dataset and experiments

We interpolate (with fractal interpolation, as described in section 3) 476 EEGs[8] whose durations range from 1 minute 50 seconds to 5 hours 21 minutes and whose sizes are between 1133KB and 138 MB. All signals in all these files have a frequency of 250Hz. Of the files used, 260 have a duration between 15 and 30 minutes (54.6%)-which is the most frequent duration range for EEGs-, 40 files (8.4%) a duration below 15 minutes and 176 files (37%) a duration higher than 30 minutes. Moreover, 386 files contain 23 signals (81.1 %), 63 20 signals (13.2 %), 13 19 signals (2.7 %), 7 25 signals (1.5 %), 3 28 signals (0.6 %), 1 12 signals (0.2 %), 2 13 signals (0.4 %) and 1 2 signals (0.2 %). The experiments were run on an openSuSe 10.3(x86-64) (kernel version 2.6.22.5-31) server (RAM 32GB, Intel® Quad-Core Xeon® E5420@2.50GHz processor). The files for which the diagnosis conclusion is either unknown or known to be abnormal without any further details are not considered in the distance computation and clustering steps described in section 3. This means that the distance computation and clustering steps are performed on a subset of 328 files of the original 476 files. The similarity matrice obtained is a 328×328 matrix . The files contained in the subset chosen for clustering can be separated in 4 classes: normal EEG (195 files i.e 59.5%), EEG of epilepsy(64 files i.e 19.5%), EEG of encephalopathy(31 files i.e 9.5%) and EEG of brain damage (vascular damage, infarct, or ischemia)(34 files i.e 10.4%). Figure 2 shows the plot of the envelope intensity versus the index for the euclidean-distance-based similarity measure and the plot of the envelope intensity versus the index for the mutual-information-based similarity measure. The plot for the Euclidean-distance based similarity matrix exhibits 2 distinct regions whereas the plot for the mutual-information based similarity matrix exhibits 4 distinct regions. We therefore cluster the data first in 2 different clusters using the Euclidean-based similarity matrix and then in 4 clusters using

[8] unprocessed and unnormalised

the mutual-information based matrix. As we can see, the mutual information-based measure yields the correct number of clusters while the Euclidean distance-based similarity measure isn't spread enough to yield the correct number of clusters. We compare the performance of the IFS-based similarity measure with an autoregressive (AR)-based similarity measure inspired from [9]:

- An AR model is fitted to each of the signals of each of the EEG files considered (at this stage 476). The order of the AR model fitted is selected using the AIC criterion. The order is equal to 4 for our dataset.
- The LPC cepstrum coefficients are computed based on the AR model fitted to each signal using the formulas given in [9]. The number of coefficients selected is the PGCD of the number of points for all signals from all files.
- The Euclidean distance, as well as the mutual information between the computed cepstral coefficients are computed in the same way as with the fractal dimension-based distances for the subset of 328 files for which the diagnosis are known. The resulting similarity matrices (328×328 matrices) are used to perform k-medoid clustering.

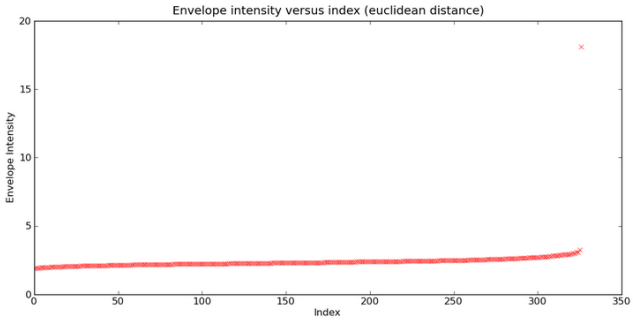
Finally, we use the similarity matrices to cluster the EEGs (see Section 3.3).

5 Results

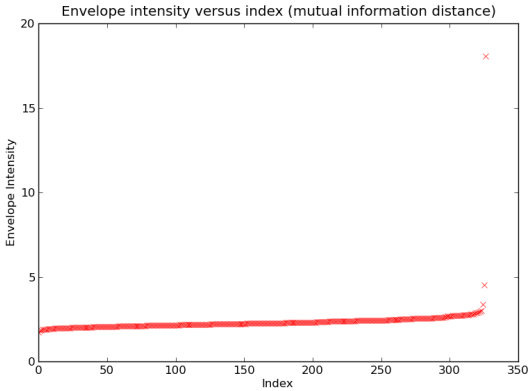
Figure 3 illustrates the relation between the duration of the EEG and the time it takes to interpolate EEGs. It shows that the increase of the fractal interpolation time with respect to the interpolated EEG's duration is less than linear. In comparison, AR modelling execution times increase almost linearly with the EEG duration. Therefore, fractal interpolation is a scalable method and is more scalable than AR modelling. In particular, the execution times for files of durations between 15 and 30 minutes are between 8.8 seconds and 131.7 seconds, that is execution times between 6.8 to 204.5 times lower than the duration of the original EEGs. Furthermore, the method doesn't impose any condition on the signals to be compared as it handles the cases where EEGs to be compared have no or limited common channels and have signals of different lengths. Moreover, fractal interpolation doesn't require model selection as AR modelling does, which considerably speeds up EEG interpolation. Moreover, with our dataset, the computation of the Euclidean distance between the cepstrum coefficients calculated based on the EEGs AR models leads to a matrix of NaN[9]: the AR modelling method is therefore less stable than the fractal interpolation-based method. Table 1 summarises the clustering results for all similarity matrices. The low sensitivity obtained for the abnormal EEGs (epilepsy,encephalopathy,brain damage) can be be explained through the following reasons:

- most of the misclassified abnormal EEGs are EEGs representing mild forms of the pathology represented therefore their deviation from a normal EEG is minimal

[9] The same happens when the mutual information is used instead of the Euclidean distance (all programs are written in Python 2.6)



(a) Euclidean distance-based matrix



(b) Mutual information-based matrix

Fig. 2. Envelope intensity of the dissimilarity matrices

– most of the misclassified abnormal EEGs (in particular for epilepsy and brain damage) exhibit abnormalities on only a restricted number of channels (localised version of the pathologies considered). The similarity measures, giving equal weights to all channels, are not sensitive enough to abnormalities affecting one channel. In future work, we will explore the influence of weights on the clustering performance. About 76% of the normal EEGs are well classified. The remaining misclassified EEGs are misclassified because they exhibit artifacts, age-specific patterns and/or sleep-specific patterns that distort the EEGs significantly enough to make the EEGs seem abnormal. Filtering artifacts before computing the similarity measures and incorporating metadata knowledge in the similarity measure would improve the clustering results.

6 Conclusion

In this paper, we considered the problem of defining a similarity measure for EEGs that would be generic enough to cluster EEGs without having to build

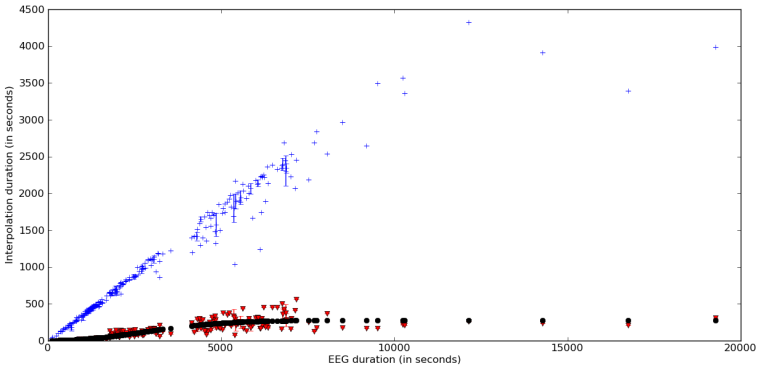


Fig. 3. Execution times of the fractal interpolation in function of the EEG duration compared to the AR modelling of the EEGs. The red triangles represent the fractal interpolation execution times and the blue crosses the AR modelling execution times. the black stars the fitting of the fractal interpolation measured execution times with function $1.14145161064 * (1 - \exp(-(0.5 * x)^{2.0})) + 275.735500586 * (1 - \exp(-(0.000274218988011 * (x))^{2.12063087537}))$ using the Levenberg-Marquardt algorithm

Table 1. Specificity and sensivity of the EEG clusterings

	Specificity	Sensitivity
normal EEG	0.312	0.770833333333
abnormal EEG	0.770833333333	0.312

	Specificity	Sensitivity
normal EEG	0.297752808989	0.657534246575
epilepsy	0.65564738292	0.183006535948
encephalopathy	0.838709677419	0.051724137931
brain damage	0.818713450292	0.114285714286

an exponential number of disease-specific classifiers. We use fractal interpolation followed by fractal dimension computation to define a similarity measure. Not only does the fractal interpolation provide a very compact representation of EEGs (which may be used later on to compress EEGs) but it also yields execution times that grow less than linearly with the EEG duration and is therefore a highly scalable method. It is a method that can compare EEGs of different lengths containing at least a common subset of channels. It also overcomes several of the shortcomings of an AR modelling-based measure as it doesn't require model selection and is more stable and scalable than AR modelling-based measures. Furthermore, the mutual-information based measure is more sensitive to the correct number of clusters than the Euclidean distance-based one. In future work, we will explore other entropy-based measures. It was also shown that the shortcomings of the similarity measure when it comes to clustering abnormal EEGs can be overcome through pre-processing the EEGs before interpolation to remove artifacts, tuning the weight parameters in the measure to account for small localised abnormalities and incorporating qualitative metadata knowledge to the measure. All those solutions constitute future work.

References

1. Accardo, A., Affinito, M., Carrozzi, M., Bouquet, F.: Use of the fractal dimension for the analysis of electroencephalographic time series. Biological Cybernetics 77(5), 339–350 (1997)
2. Barbará, D., Chen, P.: Using the fractal dimension to cluster datasets. In: KDD. pp. 260–264 (2000)
3. Barnsley, M.: Fractals everywhere. Academic Press Professional, Inc., San Diego, CA, USA, second edn. (1988)
4. Climescu-Haulica, A.: How to Choose the Number of Clusters: The Cramer Multiplicity Solution. In: Decker, R., Lenz, H.J. (eds.) Advances in Data Analysis, Proceedings of the 30th Annual Conference of the Gesellschaft für Klassifikation e.V. pp. 15–22. Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Freie Universität Berlin (March 8-10 2006)
5. De Hoon, M., Imoto, S., Nolan, J., Miyano, S.: Open source clustering software. Bioinformatics 20, 1453–1454 (June 2004), `http://portal.acm.org/citation.cfm?id=1092875.1092876`
6. Eke, A., Herman, P., Kocsis, L., Kozak, L.: Fractal characterization of complexity in temporal physiological signals. Physiological measurement 23(1), R–R38 (2002)
7. Goh, C., Hamadicharef, B., Henderson, G.T., Ifeachor, E.C.: Comparison of Fractal Dimension Algorithms for the Computation of EEG Biomarkers for Dementia. In: Proceedings of the 2nd International Conference on Computational Intelligence in Medicine and Healthcare (CIMED2005). Costa da Caparica, Lisbon, Portugal (June 29- July 1 2005)
8. Hao, L., Ghodadra, R., Thakor, N.V.: Quantification of Brain Injury by EEG Cepstral Distance during Transient Global Ischemia. In: Proceedings - 19th International Conference - IEEE/EMBS. Chicago, IL., USA (Oct 30 - Nov 2 1997)
9. Kalpakis, K., Gada, D., Puttagunta, V.: Distance Measures for Effective Clustering of ARIMA Time-Series. In: ICDM '01: Proceedings of the 2001 IEEE International Conference on Data Mining. pp. 273–280. IEEE Computer Society, Washington, DC, USA (2001)
10. Lin, G., Chen, L.: A Grid and Fractal Dimension-Based Data Stream Clustering Algorithm. Information Science and Engieering, International Symposium on 1, 66–70 (2008)
11. Mazel, D.S., Hayes, M.H.: Fractal modeling of time-series data. In: Conference Record of the Twenty-Third Asilomar Conference of Signals, Systems and Computers. pp. 182–186 (1989)
12. Mehmet Malcok and Y. Alp Aslandogan and Ayin Yesildirek: Fractal dimension and similarity search in high-dimensional spatial databases. In: IRI. pp. 380–384 (2006)
13. Pachori, R.B.: Discrimination between ictal and seizure-free EEG signals using empirical mode decomposition. Res. Let. Signal Proc. 2008, 1–5 (2008)
14. Sarkar, M., Leong, T.Y.: Characterization of medical time series using fuzzy similarity-based fractal dimensions. Artificial Intelligence in Medicine 27(2), 201–222 (2003)
15. Yan, G., Li, Z.: Using cluster similarity to detect natural cluster hierarchies. In: FSKD (2). pp. 291–295 (2007)