Unsupervised learning for patient classification using electrocardiograms data

1st Santiago Isaza Cadavid
EAFIT University
School of Applied Sciences and Engineering
Medellín, Colombia
sisazac@eafit.edu.co

I. Introduction

This paper addresses elements of supervised learning in an electrocardiogram problem. The data set consists of a sample of more than 12245 fiducial points belonging to both sick (Arrhythmia) and healthy patients and seven features. The subset of sick patients is 2000 while the healthy ones are 10245. The main purpose is to develop models that have an almost always accurate classification for healthy patients. In this order of ideas, if a healthy patient is always classified correctly, any characteristic that is slightly anomalous will allow the cardiac health of a new patient to be questioned. This in turn could lead to further health checks. The most important thing in terms of ethics and health for this project is that if someone is sick they are not misclassified.

There are numerous investigations that use supervised learning in the study of electrocardiograms and their classification. Here we will discuss some of the most recent and significant ones. The authors [1] propose a neural network-based method to predict the mortality risk of ICU patients using unstructured electrocardiogram (ECG) text reports. The proposed model, when compared to four standard ICU severity scoring methods, outperformed all by $10 - 13\%$ in terms of prediction accuracy. Another significant work was developed by [2], they propose an ECG pre-training method that learns both local and global contextual representations for better generalization and performance on subsequent tasks. In addition, we propose randomized lead masking as an ECG-specific augmentation method to make our proposed model robust to an arbitrary set of leads. Experimental results on two subsequent tasks, cardiac arrhythmia classification and patient identification, show that our proposed approach outperforms other state-of-the-art methods. Finally, another work worth mentioning is [3]. In this research, the authors propose an SSL algorithm based on ECG delineation and show its effectiveness for arrhythmia classification. Our experiments demonstrate not only how the proposed algorithm improves DNN performance across various labeled datasets and fractions of datasets, but also how features learned through pre-training on one dataset can be transferred when tuned on a different dataset.

II. Method

A. Mountain clustering

Finding cluster centers based on a density metric known as the mountain function is easy with the mountain clustering approach. This approach can be used as a preprocessing for more complex clustering techniques and is a straightforward way to identify approximated cluster centers [4]. Once the grid V has been constructed with all the potential centers, the mountain function (better known as the density function) is calculated for each of the centers (equation 1). Once we have the center with the highest density, it is assigned as center 1, and the new centers are generated by eliminating the effect of the immediately previous center (equation 2), doing the process for as many centers as convenient.

$$m(v) = \sum_{i=1}^{N} \exp\left(\frac{\|v - x_i\|^2}{2\sigma^2}\right) \tag{1}$$

$$m_j(v) = m(v) - m\left(c_{j-1}\right) \exp\left(\frac{\|v - c_{j-1}\|^2}{2\beta^2}\right) \tag{2}$$

B. Subtractive clustering

Similar to the mountain method, the subtractive method works by calculating the density of the points, but this time, in order to reduce the computational capacity employed, the same points in the database are used as candidate points to be centers. The density for the point x_i is defined in equation 3 [4]. Subsequently, once the point with the highest density has been found and labeled as the center, the following centers are calculated by removing the effect of the previous one, as shown in equation 4.

$$D_i = \sum_{j=1}^{N} \exp\left(\frac{\|x_i - x_j\|^2}{\left(\frac{r_a}{2}\right)^2}\right) \tag{3}$$

$$D_i = D_i - D_{c_1} \exp\left(\frac{\|x_i - x_{c_1}\|^2}{\left(\frac{r_b}{2}\right)^2}\right) \tag{4}$$

C. K-means clustering

The core of K-means clustering, sometimes referred to as Hard C-means clustering, is the discovery of data clusters in a data set such that a cost function (or an objection function) of dissimilarity (or distance) measure is minimized. This dissimilarity measure, which is most frequently used for the Euclidean distance, is the [4]. Equation 5 can be used to create the cost function based on the Euclidean distance between a vector x_k in a group j and the corresponding cluster center c_i.

$$J = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \left(\sum_{k, x_k \in G_i} \|x_k - c_i\|^2 \right) \tag{5}$$

The partitioned groups are defined in the binary membership matrix U, where the element u_{ij} equation 6 is 1 if the if the data x_j belongs to the group i and 0 otherwise.

$$u_{ij} = \begin{cases} 1 \text{ if } \|x_j - c_i\| \leq \underset{x}{\leq} \|x_j - c_k\|^2, \text{ for each k} k \neq i \\ 0 \text{ otherwise} \end{cases} \tag{6}$$

On the other hand if the membership matrix, the new center is calculated with

$$c_i = \frac{1}{|G_i|} \sum_{k, x_k \in G_i} X_k \tag{7}$$

D. C-means fuzzy clustering

Unlike the C-hard means method, the membership matrix is allowed to have a membership grade to a certain cluster or not between 0 and 1. However, the sum of degrees of belongingness of a data point to all clusters is always equal to unity, equation 8

$$\sum_{i=1}^{c} u_{ij} = 1 \quad \forall j = 1, \cdots, n \tag{8}$$

While the cost function is a generalization of equation 5:

$$J(U, c_1, \cdots, c_c) = \sum_{i=1}^{c} J_i = \sum_{j=1}^{c} \sum_{j=1}^{n} u_{ij}^m d_{ij}^2 \tag{9}$$

where $d_{ij} = |c_i - x_j|$ is the Euclidean distance between the ith cluster center and the jth data point, and $m \in [1, \infty)$ is a weighting exponent. The necessary conditions for equation 9 are

$$c_i = \frac{\sum_{j=1}^{n} u_{ij}^m X_j}{\sum_{j=1}^{n} u_{ij}^m} \tag{10}$$

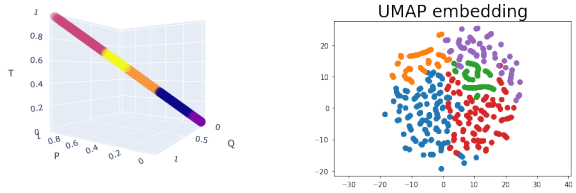
$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{kj}} \right)^{2/(m-1)}} \tag{11}$$

III. Results

A. Arrhythmia dataset

In this section we present the results using a dataset with arrhythmia patients with five different types of the disease.

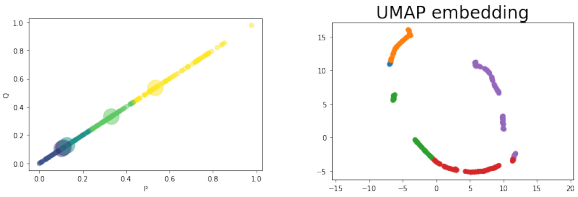
- **K-means clustering**



(a) PQT 3D plot (b) UMAP embedding

Fig. 1. H-means clustering for the arrhythmias dataset

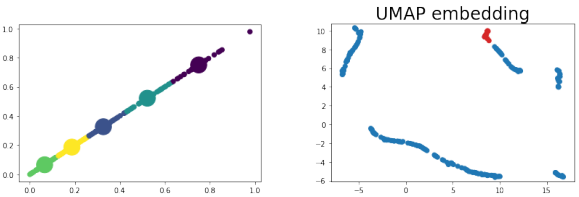
- **Subtractive clustering**



(a) PQ plot (b) UMAP embedding

Fig. 2. Subtractive clustering for the arrhytmias dataset

- **C-means fuzzy clustering**



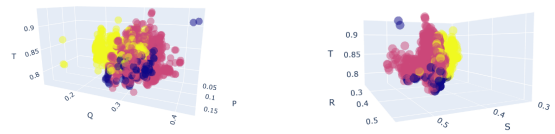
(a) PQ plot (b) UMAP embedding

Fig. 3. C-means clustering for the arrhythmias dataset

B. Healthy dataset

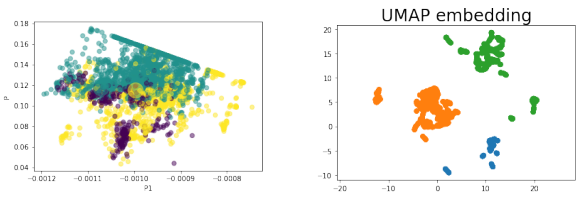
This section presents the results using a dataset of electrocardiograms belonging to healthy patients. However, these patients are divided into different groups according to their personality.

- **K-means clustering**



(a) PQT 3D plot (b) PSR 3D plot

Fig. 4. K-means clustering for the healthy dataset



(a) PQ plot (b) UMAP embedding

Fig. 5. K-means clustering for the healthy dataset

- **Mountain clustering**

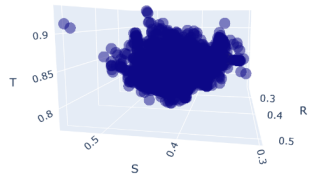


Fig. 6. Mountain clustering for the healthy dataset

- **C-mean fuzzy clustering**

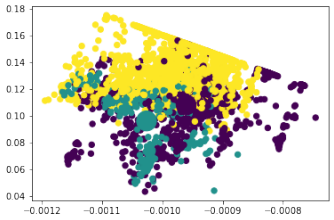
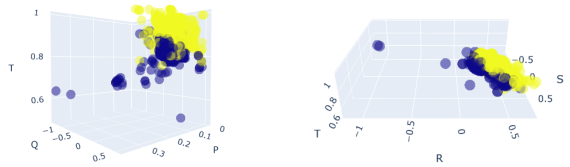


Fig. 7. Fuzzy C-means clustering for the healthy dataset

C. Ill dataset

In this section we present the results in a combined dataset of electrocardiograms. In this case, we use information from healthy patients and patients with arrhythmia. Thus, a classification into two different groups is performed.

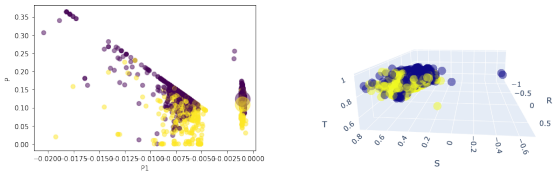
- **K-means clustering**



(a) QRP 3D plot (b) TSR 3D plot

Fig. 8. K-means clustering for the ill dataset

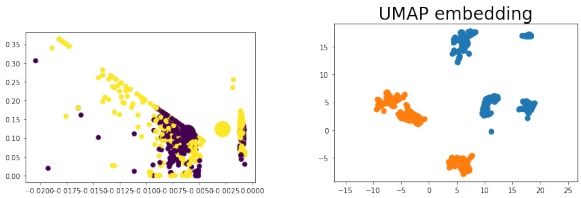
- **Subtractive clustering**



(a) 2D plot (b) TSR 3D plot

Fig. 9. Subtractive clustering for the ill dataset

- **Fuzzy C-means clustering**



(a) 2D plot (b) UMAP plot

Fig. 10. C-means fuzzy clustering for the ill dataset

IV. Conclusions

Given the dimensionality restrictions, algorithms like the mountain and subtractive algorithms did not significantly advance the investigation of the relevant space. Although good results are generally produced in learning phase 2, there is a significant flaw in the interpretability that is lost in the dimensionality reduction. We suggest creating better indexes to evaluate the effectiveness of the algorithms as well as a better tool to analyze and contribute the results produced by knowing which records are closest to each cluster and evaluating how well those records represent the group they form. Nevertheless, the obtained results look quite promising because they correctly divided the spaces according to the set of data that was used. For instance, despite being an unsupervised exercise, the sick group, which was made up of healthy and unwell patients, was correctly classified.

References

[1] G. S. Krishnan and S. Sowmya Kamath, "A supervised learning approach for icu mortality prediction based on unstructured electrocardiogram text reports," in *International Conference on Applications of Natural Language to Information Systems*. Springer, 2018, pp. 126–134.

[2] J. Oh, H. Chung, J.-m. Kwon, D.-g. Hong, and E. Choi, "Lead-agnostic self-supervised learning for local and global representations of electrocardiogram," in *Conference on Health, Inference, and Learning*. PMLR, 2022, pp. 338–353.

[3] C. Luo, G. Wang, Z. Ding, H. Chen, and F. Yang, "Segment origin prediction: A self-supervised learning method for electrocardiogram arrhythmia classification," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 1132–1135.

[4] H. Mishra and S. Tripathi, "A comparative study of data clustering techniques," *Int. Res. J. Eng. Technol.(IRJET)*, vol. 4, no. 5, pp. 1392–1398, 2017.