Unfolding feature space for non-supervised learning of a electrocardiogram database

1st Alejandra Ossa Yepes
dept. Mathematical Science
Universidad EAFIT
Medellin, Colombia
aossay@eafit.edu.co

***Abstract*—Study of electrocardiogram features processed by wavelet data by unsupervised learning using different study spaces as well as the implementation of two clustering algorithms associated with three different metrics (Euclidean, Manhattan and Mahalanobis).**

***Index Terms*—Electrocardiograms, unsupervised analysis, Kmeans, Fuzzy Cmeans**

I. Introduction

In this project we want to perform an unsupervised analysis on the data obtained in [3] which are Electrocardiograms (ECGs) provide valuable clinical information about a patient's cardiac status. Since the widespread implementation of electronic health records (EHRs), ECG records and patient data–including laboratory test results and diagnosis of disease and prescribed drug histories–have accumulated in daily clinical practice. These records are an excellent source of practice-based evidence for evaluating electrophysiological changes on ECGs under many clinical circumstances. it is desired to perform 8 visualizations for 3 different spaces (high, medium and low dimensions) with different clustering methods associated to three different metrics.

In previous cases there have already been studies on the non-supervision of learning machines for electrocardiograms such as Rodriguez, Gallego, Mora, Orozco and Bustamante (Rodriguez. C. A, et al., 2014), obtained results with k-means of 97,41% and 92,94% for specificity and sensibility, respectively, although sensibility is higher than the obtained with the methodology used in this work , it is pertinent to explain that the mentioned work takes into account only the ventricular contraction heartbeats. Juie D. Peshave and Rajveer Shastri (Peshave. J. D and Shastri.R, 2014), obtain similar results with the 85 % for sensibility when clustering 3 different types of arrhythmias using Thresholding's method.N.Jannah and S. Hadjiloucas (Jannah. N and S. Hadjiloucas. S, 2014) use supervised classifiers as the Support Vector Machine (MSVM) and Complex Support Vector Machine (CSVM) and obtain results in terms of 94% for accuracy, thus the supervised methods result be also useful in the process of arrhythmias identification [7].

Using unsupervised methods and especially k-means allow to achieve good results. But unlike from other works, in this paper, the segment-bases approach for clustering, the centroid initialization and the feature selection together are used, contributing to the detection of minority classes, reduction of the computational cost and convergence of the k-means algorithm, with the aim to realize a better heartbeat clustering and facilitate the cardiologist gives a diagnosis of a pathology checking 2 or 3 prototype heartbeats of a group and give to the patients trustworthy and timely results of the medical exams , with the aim that the patients get an adequate treatment. The main contributions of this work are the segment-bases approach for clustering, the feature selection and the centroid initialization.

II. Methodology

Clustering or cluster analysis is an unsupervised learning problem.It is often used as a data analysis technique for discovering interesting patterns in data, there are many clustering algorithms to choose from and no single best clustering algorithm for all cases. data clustering is considered an interesting approach for finding similarities in data and putting similar data into groups. Clustering partitions a data set into several groups such that the similarity within a group is larger than that among groups. The idea of data grouping, or clustering, is simple in its nature and is close to the human way of thinking; whenever we are presented with a large amount of data, we usually tend to summarize this huge number of data into a small number of groups or categories in order to further facilitate its analysis. Moreover, most of the data collected in many problems seem to have some inherent properties that lend themselves to natural groupings [2].

For the execution of this work will be supported by [6] where the necessary methods and concepts are developed.

A. Dataset

ECG data of one South Korean tertiary teaching hospital with 1,103 beds. The study protocol was approved by the Ajou University Hospital Institutional Review Board. All ECGs performed from 1 June 1994 to 31 July 2013 were included in the database. There were no restrictions with respect to comorbidities or prescribed drugs. The database contained 979,273 ECGs from 461,178 patients.

The aim of this study was to establish a real-world ECG database that can be used to evaluate the effects of drugs

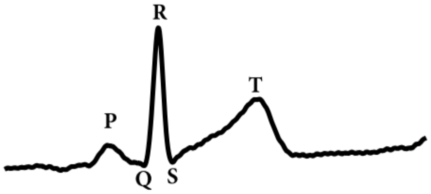


Fig. 1. Electrocardiogram segment [1]

and diseases on ECG changes, by updating and upgrading our previous ECG-ViEW database. This new database will provide an opportunity to evaluate the effects of a drug or combination of drugs, on electrophysiological changes in patients with many diseases and drug treatments.Figure 1 shows a wave segment of an electrocardiogram in which the study points of the electrocardiogram are marked [3].

Characteristics		**Value**
Patients (n)		461,178
Healthy individuals		94,326
Electrocardiogram, n		979,273
Age, years		42.6 ± 19.2
Male sex, n (%)		231,058 (50.1)
ECG parameters	*RR interval, ms*	851.7 ± 197.0
	PR interval, ms	157.1 ± 26.7
	QRS duration, ms	91.1 ± 15.2
	QT interval, ms	390.0 ± 43.5
	QTc interval, ms	425.4 ± 31.5
	P axis (degrees)	48.4 ± 24.8
	QRS axis (degrees)	45.2 ± 38.1
	T axis (degrees)	46.2 ± 38.4
Source, n (%)	*Electronic health records*	48,083 (4.9)
	ECG management system	865,590 (88.2)
	Printouts	67,384 (6.9)
Department, n (%)	*Emergency*	173,356 (17.7)
	Health examination	177,972 (18.1)
	Inpatient	193,851 (19.8)
	Outpatients	435,878 (44.4)

For computational reasons, only data from 2010 onwards will be considered in order to reduce the size of the data and speed up the process of obtaining results, leaving a total of 3,263 processed data so that the missing data are the averages of the variable to which they belong, and the variable categories were renamed as follows:

Emergency	**E**	1	ECG management system	**M**	1
Health examination	**H**	2	Scanned paper ECG	**P**	2
Outpatient	**O**	3	Electronic health records	**E**	3
Inpatient	**I**	4			

B. Different Types of Distance Metrics

Many Supervised and Unsupervised machine learning models such as C-means and K-Means depend upon the distance between two data points to predict the output. Therefore, the metric we use to compute distances plays an important role in these models.

1) Euclidean Distance: The Euclidean distance computes the real straight line distance between two points, i.e. it measures the 'as-thecrow-flies' distance. If $p = p_1, ..., p_n$ and $q = q_1, .., q_n$ [8] the Euclidean distance is defined as:

$$EUD(p,q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

2) Manhattan Distance: The Manhattan distance is also known as the "absolute value" or city block distance. It computes the distance that would be traveled to get from one data point to the other if a grid-like path is followed. It is the sum of the differences of their corresponding components. The Manhattan distance is defined as [8]:

$$MN(p,q) = \sum_{i=1}^{n} |p_i - q_i|$$

3) Mahalanobis Distance: The Mahalanobis distance is based on the correlations between variables. It is defined as [8]:

$$MD(p,q) = \sqrt{(p_i - q_i)^T V^{-1} (p_i - q_i)}$$

where V is the covariance matrix of $A_1..A_m$ and A_j is the vector of values for attribute j occurring in the training set instances $1..n$

C. K-means Clustering

The K-means clustering, or Hard C-means clustering, is an algorithm based on finding data clusters in a data set such that a cost function (or an objection function) of dissimilarity (or distance) measure is minimized. In most cases this dissimilarity measure is chosen as the Euclidean distance. A set of n vectors x_j, $j = 1...n$ are to be partitioned into c groups G_i, $i = 1...c$ The cost function, based on any distance between a vector x_k in group j and the corresponding cluster center c_i , can be defined by [6]:

$$J = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \left(\sum_{k, \mathbf{x}_t \in G_i} \|\mathbf{x}_k - \mathbf{c}_i\|^2 \right)$$

The partitioned groups are defined by a *cxn* binary membership matrix **U**, where the element u_{ij} is 1 if the jth data point x_j belongs to group i , and 0 otherwise. Once the cluster centers c_i are fixed, the minimizing u_{ij}. for Equation before can be derived as follows:

$$u_{ij} = \begin{cases} 1 \text{ if } \|\mathbf{x}_j - \mathbf{c}_i\|^2 \leq \|\mathbf{x}_j - \mathbf{c}_k\|^2, \text{ for each } k \neq i \\ 0 \text{ otherwise} \end{cases}$$

Restated, $\mathbf{x}_j$ belongs to group i if $\mathbf{c}_i$ is the closest center among all centers. Since a given data point can only be in a group, the membership matrix U has the following properties:

$$\sum_{i=1}^{c} u_{ij} = 1, \forall j = 1, \ldots, n \qquad \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij} = n$$

On the other hand, if u_{ij} is fixed, then the optimal center $\mathbf{c}_i$ that minimize Equation (15.1) is the mean of all vectors in group i :

$$\mathbf{c}_i=\frac{1}{|G_i|}\sum_{k,\mathbf{x}_k\in G_i}\mathbf{x}_k$$

where $|G_i|$ is the size of G_i, or $|G_i|=\sum_{j=1}^{n}u_{ij}$.

D. Fuzzy C-means Clustering

Fuzzy C-means clustering (FCM), relies on the basic idea of Hard C-means clustering (HCM), with the difference that in FCM each data point belongs to a cluster to a degree of membership grade, while in HCM every data point either belongs to a certain cluster or not. So FCM employs fuzzy partitioning such that a given data point can belong to several groups with the degree of belongingness specified by membership grades between 0 and 1. However, FCM still uses a cost function that is to be minimized while trying to partition the data set. The membership matrix U is allowed to have elements with values between 0 and 1. However, the summation of degrees of belongingness of a data point to all clusters is always equal to unity [6]:

$$\sum_{i=1}^{c}u_{ij}=1,\forall j=1,\ldots,n$$

The cost function for FCM is a generalization of Equation

$$J\left(U,\mathbf{c}_1,\ldots,\mathbf{c}_c\right)=\sum_{i=1}^{c}J_i=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}d_{ij}^{2}$$

where u_{ij} is between 0 and 1 and $m\in(1,\infty)$ is a weighting exponent. The necessary conditions for Equation to reach its minimum are

$$\mathbf{c}_i=\frac{\sum_{j=1}^{n}u_{ij}^{m}\mathbf{x}_j}{\sum_{j=1}^{n}u_{ij}^{m}}\qquad\qquad u_{ij}=\frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{kj}}\right)^{2/(m-1)}}$$

The algorithm works iteratively through the preceding two conditions until the no more improvement is noticed. In a batch mode operation, FCM determines the cluster centers c_i and the membership matrix $\mathbf{U}$

E. Mountain Clustering

The mountain clustering approach is a simple way to find cluster centers based on a density measure called the mountain function. This method is a simple way to find approximate cluster centers, and can be used as a preprocessor for other sophisticated clustering methods. The first step in mountain clustering involves forming a grid on the data space, where the intersections of the grid lines constitute the potential cluster centers, denoted as a set $\mathbf{V}$. The second step entails constructing a mountain function representing a data density measure. The height of the mountain function at a point $v\in V$ is equal to

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

where x_i is the ith data point and σ is an application specific constant. This equation states that the data density measure at a point *v* is affected by all the points x_i in the data set, and this density measure is inversely proportional to the distance between the data points xi and the point under consideration *v* . The constant σ determines the height as well as the smoothness of the resultant mountain function. The third step involves selecting the cluster centers by sequentially destructing the mountain function. The first cluster center c_1 is determined by selecting the point with the greatest density measure. Obtaining the next cluster center requires eliminating the effect of the first cluster. This is done by revising the mountain function: a new mountain function is formed by subtracting a scaled Gaussian function centered at

$$m_{\text{new}}\left(\mathbf{v}\right)=m(\mathbf{v})-m\left(\mathbf{c}_1\right)\exp\left(-\frac{\|\mathbf{v}-\mathbf{c}_1\|^2}{2\beta^2}\right)$$

The subtracted amount eliminates the effect of the first cluster. Note that after subtraction, the new mountain function $m_{\text{new}}(v)$ reduces to zero at $v=1$ [6].

F. Subtractive Clustering

The problem with the previous clustering method, mountain clustering, is that its computation grows exponentially with the dimension of the problem; that is because the mountain function has to be evaluated at each grid point. Subtractive clustering solves this problem by using data points as the candidates for cluster centers, instead of grid points as in mountain clustering. This means that the computation is now proportional to the problem size instead of the problem dimension. However, the actual cluster centers are not necessarily located at one of the data points, but in most cases it is a good approximation, especially with the reduced computation this approach introduces.

Since each data point is a candidate for cluster centers, a density measure at data point x_i is defined as

$$D_i=\sum_{j=1}^{n}\exp\left(-\frac{\|\mathbf{x}_i-\mathbf{x}_j\|^2}{\left(r_a/2\right)^2}\right)$$

where r_a is a positive constant representing a neighborhood radius. Hence, a data point will have a high density value if it has many neighboring data points. The first cluster center x_{c_1} is chosen as the point having the largest density value D_{c_1} . Next, the density measure of each data point x_i is revised as follows:

$$D_i=D_i-D_{c_1}\exp\left(-\frac{\|\mathbf{x}_i-\mathbf{x}_{c_1}\|^2}{\left(r_b/2\right)^2}\right)$$

where r_b is a positive constant which defines a neighborhood that has measurable reductions in density measure. Therefore, the data points near the first cluster center $\mathbf{x}_{c_1}$ will have significantly reduced density measure.

After revising the density function, the next cluster center is selected as the point having the greatest density value. This process continues until a sufficient number of clusters is attainted [6].

G. Uniform Manifold Approximation and Projection

UMAP (Uniform Manifold Approximation and Projection) is a novel manifold learning technique for dimension reduction. UMAP is constructed from a theoretical framework based in Riemannian geometry and algebraic topology. The result is a practical scalable algorithm that is applicable to real world data. The UMAP algorithm is competitive with t-SNE for visualization quality, and arguably preserves more of the global structure with superior run time performance. Furthermore, UMAP has no computational restrictions on embedding dimension, making it viable as a general purpose dimension reduction technique for machine learning [5].

H. T-SEN

The technique is a variation of Stochastic Neighbor Embedding (Hinton and Roweis, 2002) that is much easier to optimize, and produces significantly better visualizations by reducing the tendency to crowd points together in the center of the map. t-SNE is better than existing techniques at creating a single map that reveals structure at many different scales. This is particularly important for high-dimensional data that lie on several different, but related, low-dimensional manifolds, such as images of objects from multiple classes seen from multiple viewpoints. For visualizing the structure of very large data sets, we show how t-SNE can use random walks on neighborhood graphs to allow the implicit structure of all of the data to influence the way in which a subset of the data is displayed [4].

The methods and elements mentioned above are the pillar of the implementation of this document which has its computational component used in Pyhton, in addition to the algorithms and mathematical basis of the methods are based on [3] being this document the main pillar of the implementation of the methods.

III. Implementation

The structure of the project is based on the construction and experimentation of three parts:

1) For the creation of the clusters in the real space of the data (High dimensions) it is necessary to know an optimal number of centers to use, so the Mountain and Subtractive method was implemented to find these centers, which require different metrics and parameter variations for its implementation, after obtaining this information, the construction of the clusters is carried out and an embedding process is performed for the graphing of these as evidenced in the tables III and II which show the clusters made by the Kmeans and Fuzzy Cmeans algorithms.
2) In the second part two different embedding techniques were generated which focus on performing a dimensionality reduction of the problem from the beginning to compare the clustering process with this type of methods in the same way the implementation process of Kmeans and Fuzzy Cmeans was performed for this new dataset considering the three types of metrics required. This can be evidenced in the tables IV, V, VI, VII which show the clustering process in low dimensions using TSEN and UMAP.

Table I shows that the search process for the number of clusters was carried out by means of the two different methods and with variation of parameters, so that by averaging the results obtained it is shown that the optimum number of clusters to be generated is 5.

TABLE I
Parameter variation with different metrics

Methods	**Metrics**	**Parameters**		
		(0.4, 0.6)	(0.6, 0.9)	(0.9, 1.2)
Mountain	Euclidean	4	7	5
	Manhattan	7	5	6
	Mahalanobis	5	10	14
Subtractive	Euclidean	4	5	3
	Manhattan	5	7	4
	Mahalanobis	5	5	11

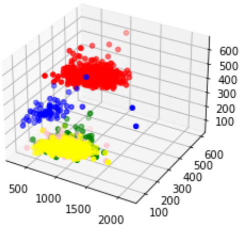
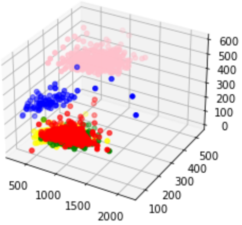
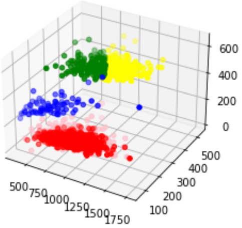
Clusters graphics	Metrics
	Euclidean Distance
	Manhattan Distance
	Mahalanobis Distance

TABLE II
High-dimensional with K-means Clustering

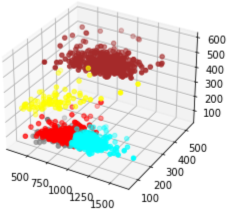
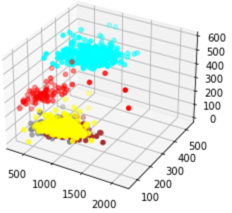
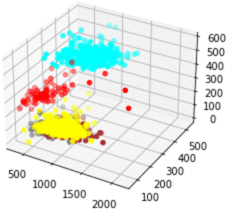
Clusters graphics	Metrics
	Euclidean Distance
	Manhattan Distance
	Mahalanobis Distance

TABLE III
HIGH-DIMENSIONAL WITH FUZZY C-MEANS CLUSTERING

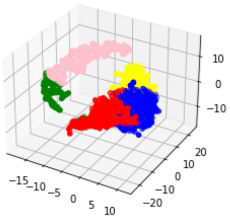
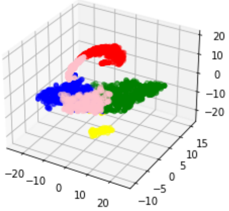
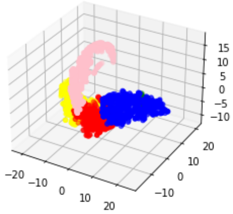
Clusters graphics	Metrics
	Euclidean Distance
	Manhattan Distance
	Mahalanobis Distance

TABLE IV
HALF DIMENSIONS AND BY APPLYING TSEN EMBEDDING AND K-MEANS CLUSTERING

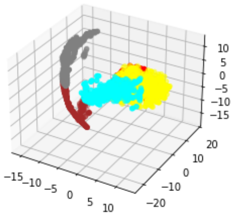
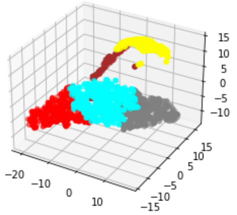
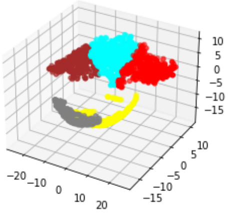
Clusters graphics	Metrics
	Euclidean Distance
	Manhattan Distance
	Mahalanobis Distance

TABLE V
HALF DIMENSIONS AND BY APPLYING TSEN EMBEDDING AND FUZZY C-MEANS CLUSTERING

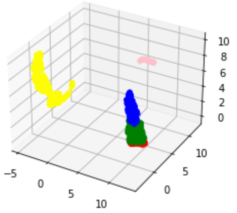
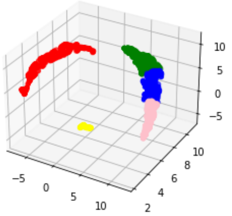
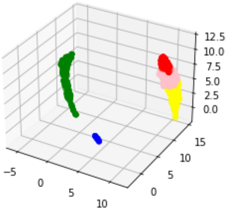
Clusters graphics	Metrics
	Euclidean Distance
	Manhattan Distance
	Mahalanobis Distance

TABLE VI
HALF DIMENSIONS AND BY APPLYING UMAP EMBEDDING AND K-MEANS CLUSTERING

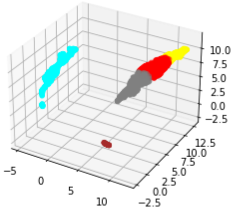
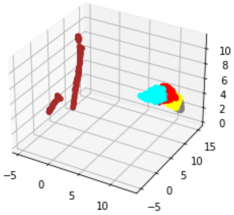
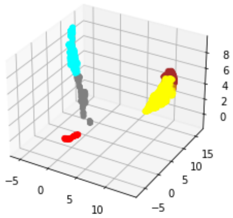
Clusters graphics	Metrics
	Euclidean Distance
	Manhattan Distance
	Mahalanobis Distance

TABLE VII
HALF DIMENSIONS AND BY APPLYING UMAP EMBEDDING AND FUZZY C-MEANS CLUSTERING

IV. Discussion

In particular, this paper has argued that the results of clustering clustering results are highly dependent on the chosen exploration or learning configuration. learning configuration chosen. An initial insight is that the mountain clustering algorithm has significant limitations in terms of the computational space computational space required to construct the n-dimensional network in addition to its susceptibility to parameters. These limitations make the above method impractical in many real-life situations where a large set of features must be explored. Subsequently, the tests showed in the scenarios evaluated that not all of the standards examined, namely Mahalonobis and Manhattan,can form a suitable clustering arrangement. In particular, the ill-conditioned property of the covariance matrix of the covariance matrix is undoubtedly a setback when employing the first part. A surprising fact is that the Fuzzy C-means algorithm had a low average performance compared to the other two. compared to k-means. Therefore, it can be said that each clustering algorithm may have a particular set of conditions in which in which its performance is the best.

In conclusion, the evidence shows that the proposed polynomial transformation does not provide the expected benefits. On the contrary, the embedding procedure went the other way around of the data and induced new partitions in the dataset that might not be related to the truth. that might not be related to the ground truth. There is no doubt that the use of of clustering in different dimensions helps to derive new insights into the structure new insights into the structure, shape, relationships and complexity of the source data. and complexity of the source data. Although the proposed transformations did not transformations did not provide optimal results, the proposed transformations did not provide optimal results.

References

[1] Anup Das, Paruthi Pradhapan, W. G. P. A.-R. T. R.-F. C. S. S. J. L. K. N. D.-C. V. H. Unsupervised heart-rate estimation in wearables with liquid states and a probabilistic readout. *Elsevier* (2018).

[2] Hong He, Yonghong Tan, J. f. X. Unsupervised classification of 12-lead ecg signals using wavelet tensor decomposition and two-dimensional gaussian spectral clustering. *Elsevier* (2019).

[3] Kim YG, Shin D, P. M. Ecg-view ii, a freely accessible electrocardiogram database.

[4] Laurens van der Maaten, G. H. Visualizing data using t-sne. *Journal of Machine Learning Research 9* (2008).

[5] Leland McInnes, John Healy, J. M. Umap: Uniform manifold approximation and projection for dimension reduction.

[6] Mishra, H., S., and Tripathi. A comparative study of data clustering techniques. *University of Waterloo, Ontario, Canada* (2017).

[7] Monica Moreno-Revelo, Sandra Patascoy-Botina, A. P.-B. J. R.-F. J. R.-S. S. M.-R. D. P.-O. Unsupervised analysis applied to the detection cardiac arrhythmias. *Universidad Tecnológica Equinoccial* (2017).

[8] Walters-Williams, J., and Li, Y. Comparative study of distance functions for nearest neighbors. *Springer Science+Business Media B.V.* (2010).