Mateo W. Racca

UNIVERSITAT ROVIRA i VIRGILI
Escola Tècnica
Superior d'Enginyeria

NEURAL AND EVOLUTIONARY COMPUTING

(MESIIA): Assignment #3: Unsupervised learning with PCA, t-SNE, k-means, AHC and SOM

Part 1

The chosen dataset is commonly known as the HCV dataset (Hepatitis C Virus) from the UCI Machine Learning Repository. Below is the detailed description of the dataset and a link to the source webpage:

HCV Dataset Description:

Domain: Medical / Hepatology.

Objective: The dataset is typically used for research and analysis in the medical domain, often for the purpose of understanding the factors associated with Hepatitis C Virus infection and its stages.

Features: The dataset includes various medical measurements such as liver enzymes, bilirubin, albumin, and other blood chemistry measurements. It also contains demographic information like age and gender.

Target: The dataset categorizes patients into different categories based on the stage of Hepatitis or other clinical diagnoses related to HCV. The possible values are `'0=Blood Donor'`, `'0s=suspect Blood Donor'`, `'1=Hepatitis'`, `'2=Fibrosis'`, `'3=Cirrhosis'`

Data Points: It initially has 615 instances, each representing different patients or clinical cases.

The dataset can be retrieved from: https://archive.ics.uci.edu/dataset/571/hcv+data

Preprocessing

Firstly, the dataset is loaded, and we separate the features we want to analyze from the target labels. The target labels, which represent the classes in the dataset, are set aside because unsupervised learning algorithms work without labeled output data.

We then categorize the features into numerical and categorical types since they require different treatments. Numerical features are standardized to have a mean of zero and a standard deviation of one. This normalization is crucial because unsupervised algorithms like PCA and k-means can be skewed by features that operate on larger scales.

Categorical features, in this case binary, are encoded to ensure they are represented numerically, making them suitable for mathematical operations performed by the algorithms.

After applying these transformations, the cleaned and transformed data is saved back into a CSV file.

Part 2:
Let's walk through the process for both the NEC and HCV datasets:

PCA (Principal Component Analysis): for both datasets, we performed PCA to reduce the dimensionality of the data and to capture the most variance in two principal components. This step simplifies the datasets and makes them easier to visualize and interpret. We plotted the scatter plot of the first two principal components, with each point colored according to its class label, providing a visual of the data's distribution and any apparent clustering. We also created scree plots to visualize the proportion of variance explained by each principal component, which helps in deciding how many components to keep for further analysis.

t-SNE (t-Distributed Stochastic Neighbor Embedding): we then applied t-SNE, which is a non-linear dimensionality reduction technique, particularly useful for visualizing high-dimensional data in two or three dimensions. We experimented with different perplexity settings and other parameters to find a meaningful representation of the data where similar instances are modeled by nearby points and dissimilar instances are modeled by distant points. Each data point was colored based on its class, providing insight into class separability in the embedded space.

k-Means Clustering: using the k-means clustering algorithm, we attempted to partition the data points into k clusters for different values of k. We visualized these clusters by plotting the data points using the first two principal components from PCA as coordinates. This helped us compare the k-means clusters with the true class labels. For the optimal k (which could be the actual number of classes if known), we used a confusion matrix to assess the accuracy of the clustering against the true labels.

AHC (Agglomerative Hierarchical Clustering): we performed AHC using both UPGMA and complete linkage methods. AHC is a bottom-up clustering method where each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. We used Euclidean distances between original patterns as the input and visualized the results using dendrograms. In these dendrograms, each merge is represented by a horizontal line, and the height of the merge indicates the distance between the two clusters.

SOM (Self-Organizing Map): finally, for the Self-Organizing Map, we trained a network to produce a low-dimensional (typically two-dimensional), discretized representation of the input space of the training samples. We chose a grid size that gave us at least 100 neurons. SOMs help in visualizing complex data with many variables by producing a map of these variables. We visualized the SOM using a U-matrix, which displays the distance between neurons, and overlaid it with the class labels. We also plotted component planes to see how each feature contributes across the map.

Throughout the analysis of both NEC and HCV datasets, we encountered a few challenges, such as handling non-numeric values, choosing the right parameters for t-SNE, and dealing with key errors in color maps. By addressing these issues, we could ensure that the unsupervised learning techniques were applied correctly. Each technique provided a different perspective on the data: PCA for overall structure, t-SNE for local patterns, k-means for potential grouping, AHC for hierarchical structure, and SOM for topology preservation and feature mapping.

All the process can be found on the notebooks of each dataset in the github repository: https://github.com/raccamateo/NEC_A3

The application of unsupervised learning techniques to the NEC and HCV datasets has provided a comprehensive insight into the underlying structure and distribution of the data. Here are some final thoughts on the findings from each method applied:

Results:

PCA: the PCA results suggest that while some variance is captured in the first two components, more components may be necessary to fully understand the data's structure. The scree plots indicated that two components were not sufficient to capture the majority of variance within the datasets, and increasing to four components provided a better representation, particularly for the HCV dataset.

t-SNE: the t-SNE visualizations demonstrated the power of non-linear dimensionality reduction in clustering data points that are similar to each other. This method was particularly useful in revealing data clusters that were not apparent in the PCA results. The t-SNE plots showed distinct groupings, suggesting that the data contains several clusters that could correspond to different subtypes or stages within the HCV and NEC datasets.

k-Means: the k-Means clustering algorithm provided a straightforward approach to partitioning the data into k clusters. When visualized using PCA components, some clusters appeared distinct, while others overlapped, indicating that some classes might be inherently closer to each other or that the cluster centroids did not align perfectly with the true class distributions.

AHC: the dendrograms generated by AHC revealed a hierarchical structure within the data. It was evident that different linkage criteria led to different cluster structures, which could be important for understanding the relationships between data points. However, the full dendrograms were complex, and it was challenging to determine the exact number of clusters without additional information or domain knowledge.

SOM: the SOMs provided a visual representation of the data's topology, revealing how various features and samples are distributed across the map. The U-matrix and component planes helped visualize the high-dimensional data in two dimensions, showcasing regions of similarity and difference within the datasets.

In conclusion, the unsupervised learning techniques applied to the NEC and HCV datasets each contributed uniquely to understanding the datasets' structure. While PCA and t-SNE facilitated the visualization of data distribution, k-Means and AHC offered insights into the natural groupings within the datasets. The SOM further extended these insights by providing a topological map of the data's features. These analyses underscore the multi-faceted nature of unsupervised learning and the importance of using a variety of methods to gain a comprehensive view of the data. However, it is also clear that these methods have their limitations and must be carefully chosen and interpreted within the context of the data and the domain-specific knowledge available.