ID de aportación : **15** Tipo: **Oral**

Exploring class separability patterns in high-dimensional datasets with the CSVIZ tool

In a Data Science project, data visualization stands as an essential asset in several phases of its lifecycle. In particular, it is crucial during the Exploratory Data Analysis, supporting the discovery of anomalies, structure, and relationships within the data to fully understand the underlying problems. In classification problems, data visualization is helpful to reveal class separability patterns within the dataset by visually exploring the class distributions and topology. It enables the identification of regions within the feature space where classes are well-separated or overlapped. This is very valuable information that can be later used when building a Machine Learning model, helping to choose an appropriate one and to improve the model performance.

In this context, high-dimensional data arise as a challenge. Traditional visualization techniques such as the scatterplot matrix struggle with their representation due to space limitations to correctly display them. With d variables in a dataset, the scatterplot matrix must include $\frac{d \cdot (d-1)}{2}$ pairwise scatterplots in a single display. As d increases, the graph becomes overplotted. Additionally, humans lack the cognitive ability to discover structures and patterns within a huge scatterplot matrix. Alternative visualization methods often rely on dimensionality reduction techniques, yet this can compromise interpretability as patterns are explored in a transformed space of the original variables. Indeed, in many applications, it's critical to keep the original variables rather than a transformation of them to make informed decisions.

Acknowledging the previously discussed issue, the authors have proposed the Class Separability Visualization (CSViz) method as a new Visual Analytics tool to deal with the visualization of labeled high-dimensional data in their original variables. CSViz addresses this challenge with a subspace approach. It offers a set of 2-Dimensional subspace visualizations, each containing exclusive subsets of points from the original variables with the most valuable and significant separable patterns within the dataset. Thus, CSViz offers an overview of the class separability in the dataset, reducing the number of scatterplots to be inspected compared to the scatterplot matrix. CSViz significantly eases the visual exploration in the EDA, and thus, reduces the amount of time invested in it.

The CSViz method has been implemented using the R software, and the open-source code can be found at https://github.com/URJCDSLab/CSViz.

¿Presentas la comunicación a premio?

Premio estudiante (grado, máster, doctoral)

Afiliación (del autor)

Universidad Rey Juan Carlos

Autor primario: CUESTA SANTA TERESA, Marina (Universidad Rey Juan Carlos)

Clasificación de la sesión: Sesión premio predoctoral

Clasificación de pistas: Estadística