Feature Importance For Clustering

No Author Given

No Institute Given

1 Formalization

Let us suppose that we have a collection of data points $X \subseteq \mathbb{R}^n$ to be gathered into k clusters using a clustering algorithm such as k-means or fuzzy c-means. Such algorithms produce a partition $C = \{C_1, \ldots, C_j, \ldots, C_k\}$ such that $C_j \subset X$ denotes the j-th cluster with prototype p_j. The problem consists of determining which features are leading the emergence of the cluster structure. This research project proposes a scoring function to determine feature importance when performing cluster analysis of numeric data.

Equation (1) formalizes the proposed method to determine the importance of the i-th feature given a clustering partition,

$$\Phi(i) = \phi(i) / \sum_{l=1}^{n} \phi(l) \tag{1}$$

such that

$$\phi(i) = \sum_{j=1}^{k} \sum_{q=1}^{k} \frac{|p_j(i) - p_q(i)|}{||p_j - p_q||_2} \tag{2}$$

where n gives the number of features describing the problem, k denotes the number of clusters, p_j is the prototype associated with the j-th cluster, while $p_j(i)$ represents the i-th feature value for the j-th prototype. Moreover, $||\cdot||_2$ is the Euclidean norm of a vector. This feature importance measure reports values in the interval $[0, 1]$. Moreover, it holds that $\sum \Gamma(i) = 1$, which is deemed an appealing property of feature importance metrics.

2 Assignment

The project assignment consists of implementing this measure for k-means and spectral clustering algorithms. However, students will be provided with a notebook implementing this measure and an unsupervised variant of SHAP for fuzzy c-means. The main modification to provided implementation is the way in which the membership values are estimated. A selection of datasets taken from the UCI Repository can be downloaded from Canvas. More details about the research methodology as a whole are provided below. The steps in blue are provided by the instructors, the ones in red need to be completed.

- Step 1. The measure is implemented in Python such that it supports k-means and spectral clustering algorithms.
- Step 2. The datasets are described (in terms of number of instances, features and decision classes) and prepared for clustering.
- Step 3. For each dataset, the number of clusters is determined as the number of decision classes in the original dataset.
- Step 4. For each dataset and algorithm, sort the features descending using their importance scores as computed by Equation (1).
- Step 5. For each dataset and algorithm, remove the features progressively in the same order they appear in the ranking. Every time a feature is removed, the clustering algorithms are executed (using the optimal number of clusters found in Step 3) and the respective performance scores are recalculated. The results are visualized as illustrated below.

Example. The figure below shows the clustering performance when removing the sorted features in a dataset progressively. The dataset has 14 numerical features. The y-axis denotes the scores and the x-axis denotes the feature ranks. As such, the first point denotes the Silhouette or Fuzzy Partition Coefficient for k-means and fuzzy c-means, respectively, when using all features. The second point denotes the scores after removing the most important feature in the ranking. Similarly, the third point denotes the scores after removing the first-ranked and second-ranked features together. The fourth point denotes the scores after removing the top-3 features. This procedure is repeated until all features but the two least important ones have been removed.

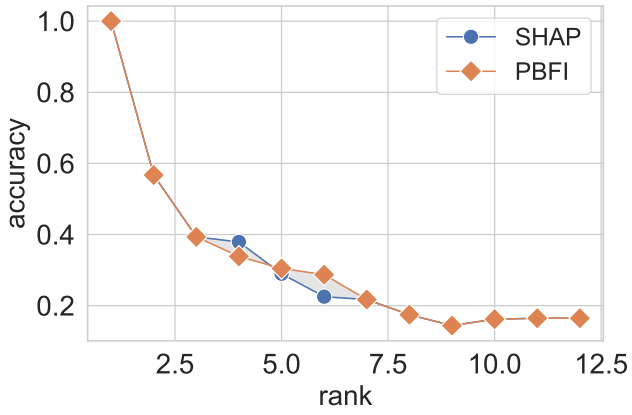

Fig. 1. Clustering performance as we progressively remove features in the same order as determined by the feature relevance score.

Note. A notebook implementing the previous steps using the fuzzy c-means algorithm will be provided. The notebook computes an unsupervised version of SHAP that can be used as a baseline.

- Step 6. For each dataset, compare the feature importance scores produced by fuzzy c-means, k-means and spectral clustering. Notice that you need to calculate the membership of each data point to its cluster since neither k-means nor spectral clustering provide that statistic.
- Step 7. Calculate feature importance using the original SHAP algorithm for supervised learning and the k-nearest neighbors classifier as the estimator. Do the feature importance scores change when compared with the scores obtained from the clustering results in the previous step?
- Step 8. Repeat the previous step for other classifiers and inspect whether the results align with the observed patterns.

3 Submission

The submission consists of a mini-paper with pertinent sections presenting the experiment results. The report will be assessed based on its correctness, completeness, research methodology and discussion of results. The best project will be extended with our help and submitted to an international scientific conference for peer review and potential publication in a book!! Naturally, your name will appear next to ours as co-authors of the proposed method, which will be a great addition to your curriculum. It is worth mentioning that now we know that the proposed measure works for the fuzzy c-means algorithm, but its performance on other clustering algorithms remain unknown.

4 Relevant dates

- 21.10.2022 The project details are released.
- 04.11.2022 The groups are registered on Canvas.
- 25.11.2022 The project is uploaded to Canvas.
- 15.12.2022 The winning team is announced.

Have lots of fun with this!