Cooperative Game Theory for Unsupervised Feature Selection in Categorical Data

Chiara Balestra
Technische Universität Dortmund
Dortmund, Germany

Emmanuel Müller
Technische Universität Dortmund
Dortmund, Germany

Andreas Mayr
IMBIE, University Hospital Bonn
Bonn, Germany

Unsupervised feature selection enables the detection of data patterns and the description of these patterns with a concise set of relevant features [8], [10]. For different unsupervised applications, the curse of dimensionality poses a major challenge. Information-theoretic, statistical, and correlation measures [2], [4], [6] are often used to quantify the interaction within a set of features. However, interpretation of higher-order interactions is often non-straight-forward and requires the introduction of human-understandable scores.

In the recent years, a crescent necessity of raising trust over the results of machine learning models has appeared and the utility of accessing to individual features' contribution scores within the machine learning procedure is high. Feature importance scores are very common in supervised learning, however, for unsupervised tasks, the literature is either limited to higher-order interactions [2], [4], [6] that are not easily interpretable, or in contrast, to traditional feature scores [8], [10], not sensitive to higher-order interactions.

Coalitional Game Theory CGT gained success both in interpretable machine learning as well as in supervised feature selection methods [3], [5]. To the best of our knowledge, CGT has not yet being applied to unsupervised feature selection. We propose to compute feature importance scores based on the decomposition of the information contained in a discrete data set by axiomatic Game Theory properties while non forgetting the need of a non-redundant selection of features. In particular, we make use of Shapley values [7] to assess to the feature importance scores. Due to the high flexibility of the value function, the method does not rely on a fixed notion of clustering, anomalies, etc. Our scores optimize towards the features containing the most information on the data set itself and they are simultaneously penalized to get rid of the redundancy among features.

Methods

We consider an N-dimensional data set $\mathcal{DB}$ containing D instances. Each dimension of the data set is the realization set of a random variable. We refer to the set of variables as $\mathcal{F} = \{X_1, \ldots, X_N\}$ and to each dimension X_i as ith feature or variable.

Analyzing the contribution of single features to any possible subset of features, we can get a ranking of importance scores for the features. The higher is the *average contribution*, i.e., the average additional value brought from the feature to any subset of features (cf. Figure 1), the more convenient is to keep it in an eventual set of selected features. After integrating a mechanism of redundancy elimination, we end up with a feature importance ranking which contains information both on the importance of the feature w.r.t. its average contribution and on the redundancy among ranked features.

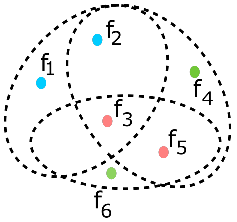

Fig. 1. Each subset of features f_is is considered to compute Shapley values. Correlated features are color-coded.

The Shapley values rely on the definition of a value function that assigns a real number to each subset of features

$$v : \mathcal{A} \subseteq \mathcal{F} \mapsto v(\mathcal{A}) \in \mathbb{R} \tag{1}$$

and satisfies the following properties

1) it assigns zero to the empty set;
2) it is a non-negative function;
3) it is a monotone function over the sets.

The value function for supervised feature selection methods was defined respectively as the accuracy of the model in Cohen et al. [3] and as the performance error in Pfannschmidt et al. [5]. Indeed, in unsupervised feature selection, we have no access to any label information. We need to choose a value function reflecting the structure of the data set and we opt for a measure of the independence of the elements in $\mathcal{A} \subseteq \mathcal{F}$. The initialization of $v(\mathcal{A})$ that we choose to adopt is the *total correlation* $C(\mathcal{A})$ of the subset $\mathcal{A}$.

Definition 1. *The* total correlation $C(\mathcal{A})$ *is defined as*

$$C(\mathcal{A}) = \sum_{X \in \mathcal{A}} H(X) - H(\mathcal{A}) \tag{2}$$

where $\mathcal{A}$ is a set of variables $\mathcal{A} \subseteq \mathcal{F}$ and $H(\mathcal{A})$ is the Shannon entropy of the subset of discrete random variables $\mathcal{A}$.

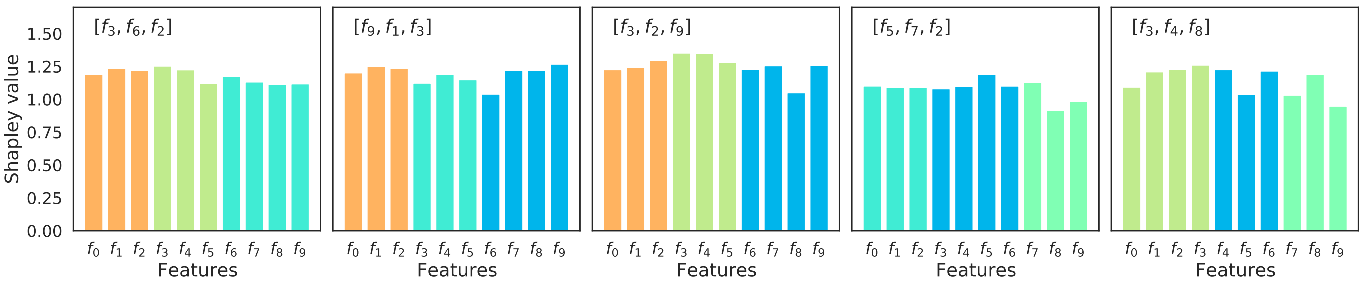

Fig. 2. Shapley values for five different samples of features from the Big Five Personality Test data set. In each plot, correlated features are color-coded. On the top, we see the first three ranked features by the algorithm proposed.

We refer to the quantity

$$C(\mathcal{A} \cup X_i) - C(\mathcal{A}) = H(\mathcal{A}) + H(X_i) - H(\mathcal{A} \cup X_i)$$

as the *marginal contribution of* X_i *to the subset* $\mathcal{A}$. Our feature importance scores $\phi(X_i)$ are defined summing up all marginal contributions and averaging on the number of possible subsets $\mathcal{A}$, i.e.,

$$\phi(X_i) = \sum_{\mathcal{A} \subseteq \mathcal{F} \setminus X_i} k_{\mathcal{A}} \cdot [H(\mathcal{A}) + H(X_i) - H(\mathcal{A} \cup X_i)] \quad (3)$$

where $k_{\mathcal{A}} = \left(N\binom{N-1}{|\mathcal{A}|}\right)^{-1}$. This is the definition of *Shapley value* [7] in the case where the value function equals the total correlation.

Our algorithm is a greedy algorithm that takes as input the data set $\mathcal{DB}$ without the need of any additional parameter. It works automatically with an included notion of redundancy. At each step, it selects the highest-ranked feature among the ones left, where the ranking is based on Shapley values and correlation with other features. We use as correlation measure the total correlation: a feature $X_j \in \mathcal{F}$ is correlated with $\mathcal{A} \subseteq \mathcal{F} \setminus \{X_j\}$ if

$$H(\mathcal{A}) + H(X_j) - H(\mathcal{A} \cup X_j) > 0. \quad (4)$$

From Information Theory we know that $H(\mathcal{A}) + H(X_j) - H(\mathcal{A} \cup X_j)$ is a non-negative real number and that it equals zero if and only if X_j and $\mathcal{A}$ are independent. We punish the features' scores whenever Equation (4) holds where $\mathcal{A}$ is the set of selected features and X_j the new feature to be ranked.

The algorithm output's is the uncorrelated feature ranking. The ranking is aware of correlations as each of the Shapley values $\Phi(X_i)$ is penalized using the correlation measure $H(X_i) + H(\mathcal{A}) - H(X_i \cup \mathcal{A})$ where $\mathcal{A}$ is the set of already ranked features and X_i is a new feature to be ranked. This algorithm provides a full ranking of features and can be prematurely stopped including an upper bound of features we are willing to rank. We underline the advantage of not having any additional parameter that requires tuning.

Results

From Game Theory and the definition of Shapley Values [7], we know that our feature importance score is fairly allocating the information contained in the data set to each feature w.r.t. the total correlation. We performed experiments showing that our algorithm select uncorrelated features both on synthetic as on real data sets outperforming state-of-the-art methods.

In particular, in the Big Five Personality Test data set, we run multiple time the algorithm on several different subsets of the features. Results are represented in Figure 2; The subsets of correlated features are color-coded while our algorithm aware of redundancy, is selecting features from uncorrelated subsets of features.

Limitations and future work

Shapley values are computationally expensive as their computation involves an exponential evaluation (in the number of features in the data set) of the value function. We keep momentarily the focus on the methodological advantages of our method in unsupervised feature ranking and show experiments on small data sets. However, several approximation exists to compute Shapley values in fairly reasonable time [1], [9].

Moreover, the choice of total correlation as value function restrict the current approach to discrete and categorical data.

References

[1] J. Castro, D. Gómez, and J. Tejada, *Polynomial calculation of the shapley value based on sampling*, Computers & Operations Research, (2009).

[2] C.-H. Cheng, A. Fu, and F. Zhang, *Entropy-based subspace clustering for mining numerical data*, KDD, (1999).

[3] S. Cohen, G. Dror, and E. Ruppin, *Feature selection via coalitional game theory*, Neural computation, (2007).

[4] H. Nguyen, E. Müller, J. Vreeken, F. Keller, and K. Böhm, *Cmi: An information-theoretic contrast measure for enhancing subspace cluster and outlier detection*, SDM, (2013).

[5] K. Pfannschmidt, E. Hüllermeier, S. Held, and R. Neiger, *Evaluating tests in medical diagnosis: Combining machine learning with game-theoretical concepts*, in IPMU, 2016.

[6] D. Reshef, Y. Reshef, H. Finucane, S. Grossman, G. McVean, P. Turnbaugh, E. Lander, M. Mitzenmacher, and P. Sabeti, *Detecting novel associations in large data sets*, Science, (2011).

[7] L. S. Shapley, *A value for n-person games*, Contributions to the Theory of Games, (1953).

[8] S. Solorio-Fernández, J. Carrasco-Ochoa, and J. F. Martínez-Trinidad, *A review of unsupervised feature selection methods*, Artificial Intelligence Review, (2019).

[9] T. van Campen, H. Hamers, B. Husslage, and R. Lindelauf, *A new approximation method for the shapley value applied to the wtc 9/11 terrorist attack*, Social Network Analysis and Mining, (2018).

[10] S. Wang, J. Tang, and H. Liu, *Embedded unsupervised feature selection*, in AAAI, 2015.