Framework for Exploring and Understanding Multivariate Correlations

Louis Kirsch[✉], Niklas Riekenbrauck, Daniel Thevessen, Marcus Pappik,
Axel Stebner, Julius Kunze, Alexander Meissner, Arvind Kumar Shekar,
and Emmanuel Müller

Hasso Plattner Institute, University of Potsdam, Potsdam, Germany
{louis.kirsch,niklas.riekenbrauck,daniel.thevessen,
marcus.pappik,axel.stebner,julius.kunze,
alexander.meissner}@student.hpi.de,
arvind.shekar@guest.hpi.de, emmanuel.mueller@hpi.de

Abstract. Feature selection is an essential step to identify relevant and non-redundant features for target class prediction. In this context, the number of feature combinations grows exponentially with the dimension of the feature space. This hinders the user's understanding of the feature-target relevance and feature-feature redundancy. We propose an interactive Framework for Exploring and Understanding Multivariate Correlations (FEXUM), which embeds these correlations using a force-directed graph. In contrast to existing work, our framework allows the user to explore the correlated feature space and guides in understanding multivariate correlations through interactive visualizations.

1 Introduction

The amount of data collected in various applications such as life-sciences, e-commerce and engineering is ever-growing. A common method used to avoid the curse-of-dimensionality and reduce the cost of collecting data is feature selection. In order to provide a smaller yet predictive subset of features, a large variety of existing approaches [4] such as CFS compute the relevance of each feature to the target class, as well as the redundancy between features.

However, the user does not get an overview of all correlations in the dataset. Furthermore, the selection process is non-transparent, as the reason for a feature's relevance or redundancy is not explained by these algorithms. This non-transparency impairs the user's understanding of the data. A high-dimensional dataset may also contain many redundant features, i.e., features exhibiting linear or non-linear dependency. Hence, the first challenge for explaining the feature selection process is to present relevance and redundancy jointly in an informative layout. The second challenge is to guide the user in understanding how features

are correlated as opposed to merely returning a correlation score. We address these two challenges by contributing FEXUM, a framework that provides:

(1) A visual embedding of feature correlations (relevances and redundancies).
(2) User-reviewable multivariate correlations.

This leads to a more comprehensible selection process in comparison to state-of-the-art tools, reflected in Table 1. While most tools focus on fully-automated statistical selection of features, with FEXUM we aim at explaining traditional black-box algorithms. KNIME is a renowned tool that offers filter-based feature selection using linear correlation and variance measures. However, without customized extensions, it does not address feature redundancy during selection. RapidMiner and Weka take redundancy into account, but do not provide an overview of all feature correlations. Additionally, they do not explain the reason for the relevance of a feature.

Table 1. Comparison of feature selection tools

Tools	Relevance	Redundancy	Correlation overview	Correlation explanation
KNIME	✔	✗	✗	✗
RapidMiner	✔	✔	✗	✗
Weka	✔	✔	✗	✗
FEXUM	✔	✔	✔	✔

2 FEXUM

FEXUM is an application that allows instant access with a web browser. We achieve this by basing our infrastructure on AngularJS and the Django web framework. To ensure scalability for large datasets, we distribute computations to multiple machines with Celery. The entire framework is open source and available online[1].

2.1 Relevance-Redundancy Embedding

As explained in Sect. 1, existing feature ranking methods do not provide a comprehensive overview of correlations that facilitate exploring the dataset. Therefore, our goal is to simultaneously visualize all feature correlations to the target (relevance) and pairwise correlations (redundancy). We allow for arbitrary relevance and redundancy measures. For now, we employ the concept of conditional dependence from [2] to quantify the correlations.

However, it is computationally expensive to calculate the redundancy score for all feature pairs. We propose, hence, to infer the redundancy scores heuristically from random subsets. The pseudo-code for this computation is made available online[2]. Our visualization provides a layout in which a smaller distance of a

---

[1] https://github.com/KDD-OpenSource/fexum.
[2] https://github.com/KDD-OpenSource/fexum-hics/blob/master/FPR.pdf.

feature to the target denotes a greater relevance, while a smaller distance between two features denotes a greater redundancy. We interpret this as a graph in which nodes represent features and weighted edges represent distances. These distances do not obey the triangle inequality and therefore cannot be mapped to metric space. We address this challenge by applying force-directed graph drawing [1]. Our algorithm places features randomly and applies forces proportional to the difference between their current distance and their correlation-defined distance. With these forces, we run a simulation until equilibrium is reached. This method is suitable even for datasets with several hundreds of features. If the correlation measure supports it, the view is updated iteratively, minimizing waiting time for the user. This is the case for our current implementation. As shown in Fig. 1, force-directed graph drawing allows soft clustering of features. Serving as a major advantage, this provides not only a relevance ranking of features, but also an understanding of how features interact with each other in terms of redundancy. This enables the user to freely select one feature from each cluster, potentially in accordance with the user's domain knowledge.

Fig. 1. Features drawn using a force-directed graph (right), with the target highlighted in green. An analysis view of two features (left) for inspecting the correlations. (Color figure online)

2.2 Understanding Feature Relevance and Redundancy

Having selected a feature set $S \subset X$, where $X = \{x_1, \cdots, x_d\}$ is a d-dimensional dataset, the second goal of our framework is to provide insight into its correlations with the target y. We propose using the average divergence between the marginal probability of y and the probability of y conditioned on different value ranges of S. For every feature $s \in S$, a value range of interest can be chosen.

If a feature s correlates with the target feature y, there exists a value range of s which changes the distribution of y in contrast to y's marginal distribution [2].

As shown in Fig. 1, our framework allows specifying the respective value range per feature using value sliders. Therefore, both bivariate and multivariate correlations can be detected. Our framework guides the user with four essential components for understanding correlations with the target.

Both the target's marginal probability distribution and the distribution conditioned on the selected value ranges are rendered in Fig. 1: (1). Changing value ranges updates this plot in real-time, allowing the user to test hypotheses evaluated according to the resulting divergence from the marginal distribution. Identifying the right hypotheses becomes challenging with more features to consider. To address this, value ranges that maximally violate the assumption of statistical independence w.r.t. the target feature are highlighted in a histogram above the sliders in (2). This tells the user which ranges strongly contribute to bivariate correlations. Nevertheless, it is still difficult to find multivariate correlations. Therefore, a table in (3) suggests multiple configurations, where each configuration specifies a value range for each $s \in S$. Each configuration is scored based on the divergence of its probability distribution and only the highest scoring configurations are displayed. Selecting one of these suggestions updates the respective value range sliders and the probability distribution plot. Finally, in case y is categorical, we visualize the data points within the value ranges in our two dimensional scatter plot in (4), each data point colorized according to its respective class.

2.3   Demonstration

FEXUM can be used with a wide range of datasets, supplied through upload by the user. While it is currently in use in industry, we will demonstrate our framework on publicly available datasets from medical, social and physical applications. As an example, we now show how our framework enhances feature selection for the Wisconsin Breast Cancer (Diagnostic) dataset [3] in Fig. 1.

In the rendering of our force-directed graph, we observe varied feature relevance scores and clusters of redundant features. In particular, features derived from similar properties such as *radius_mean* and *radius_worst* achieve comparable relevances and are highly redundant to each other. Based on this first impression, we decide to have a closer look at the most relevant feature *perimeter_worst*. We can easily find influential value ranges in the analysis view, because they are highlighted in red in the histogram. The overall relevance score can be corroborated by analyzing several individual value ranges, which can be chosen based on the framework's recommendations or expert knowledge.

Since we support multivariate correlations, the current subset can be iteratively expanded in a similar fashion. As demonstrated, the framework guides in exploration and review of correlations.

References

1. Fruchterman, T.M., Reingold, E.M.: Graph drawing by force-directed placement. Softw. Pract. Exp. **21**(11), 1129–1164 (1991)
2. Keller, F., Müller, E., Bohm, K.: HiCS: high contrast subspaces for density-based outlier ranking. In: ICDE (2012)
3. Lichman, M.: UCI machine learning repository (2013). http://archive.ics.uci.edu/ml. Accessed 17 Apr 2017
4. Molina, L.C., Belanche, L., Nebot, À.: Feature selection algorithms: a survey and experimental evaluation. In: ICDM (2003)