Interesting scatterplot discovery

Background

When dealing with a dataset containing multiple (categorical and/or numerical) variables, a first step for any data-driven knowledge discovery task could be to look at the pairwise scatterplots.
When the number of variables becomes relatively large, it may be difficult to check all the scatterplots manually though, as the number of pairwise scatterplots grows exponentially. We can try to tackle this problem by designing a data mining algorithm that *automatically detects statistically significant patterns in each of the scatterplots*. This is a task closely related to "finding significance in data visualizations" [4].
Moreover, since a pairwise scatterplot will only show the visualisation that indicates the marginal probability distributions and the interaction between merely two variables, interactions among other variables are ignored. To investigate these possible interactions, we can adopt the approach of "subgroup discovery" [2, 3] and "exceptional model mining" [1].
Specifically, for a dataset containing variables (X, Y, H_1, …, H_k), we can create a lot of sub-datasets by setting some of the (H_1, …, H_k) variables to a certain value or to a certain range (depending on whether the variable is categorical or numerical), and compare the scatterplots of (X,Y) based on the full dataset and the sub-datasets, to search for subsets that result in interesting scatter plots.
We therefore propose a Master's thesis project aiming for proposing an explanatory data mining framework by combining these two approaches.

Research directions & challenges

- How to quantify the "significance" of patterns in scatter plots, and how to quantify the differences between the scatter plot of the full dataset and a sub-dataset? Among others, one possible approach is to firstly "summarize" the data points on the scatter plots by an adaptive (two-dimensional) histogram [6, 5], and then quantify the "significance" or the "differences" based on the resulting histogram.
- As the search space will be gargantuan, a smart (heuristic) algorithm is probably needed.

Requirements

- We are looking for a Master's student who is highly self-motivated and interested in data mining research.
- Excellent programming skills and knowledge of pattern mining algorithms are required.
- Knowledge in some of the following area is an advantage: statistical significance testing, the minimum description length (MDL) principle, subgroup discovery, tree-based models, and density estimation methods.

Supervisors

Lincen Yang, Matthijs van Leeuwen

References

[1] Leman, Dennis, Ad Feelders, and Arno Knobbe. "Exceptional model mining." Joint European conference on machine learning and knowledge discovery in databases. Springer, Berlin, Heidelberg, 2008.

[2] Lavrač, Nada, et al. "Subgroup discovery with CN2-SD." Journal of Machine Learning Research 5.Feb (2004): 153-188.
[3] Van Leeuwen, Matthijs, and Arno Knobbe. "Non-redundant subgroup discovery in large and complex data." Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Berlin, Heidelberg, 2011.
[4] Savvides, Rafael, et al. "Significance of patterns in data visualisations." Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2019.
[5] Yang, Lincen, Mitra Baratchi, and Matthijs van Leeuwen. "Unsupervised Discretization by Two-dimensional MDL-based Histogram", to appear.
[6] Kontkanen, Petri, and Petri Myllymäki. "MDL histogram density estimation." Artificial Intelligence and Statistics. 2007.