DISA tool: discriminative and informative subspace assessment with categorical and numerical outcomes

Leonardo Alexandre^{1,2,3}, Rafael S. Costa^{3,4} and Rui Henriques^{1,2}

¹ INESC-ID, Lisboa, Portugal
² Instituto Superior Técnico, Universidade de Lisboa, Portugal
³ LAQV-REQUIMTE, DQ, NOVA School of Science and Technology, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal
⁴ IDMEC, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais 1, 1049-001, Lisbon, Portugal

*{leonardoalexandre,rmch}@tecnico.ulisboa.pt, rs.costa@fct.unl.pt

Abstract

Motivation: Pattern discovery and subspace clustering play a central role in the biological domain, supporting for instance putative regulatory module discovery from omic data for both descriptive and predictive ends. In the presence of target variables (e.g. phenotypes), regulatory patterns should further satisfy delineate discriminative power properties, well-established in the presence of categorical outcomes, yet largely disregarded for numerical outcomes, such as risk profiles and quantitative phenotypes.

Results: DISA (Discriminative and Informative Subspace Assessment), a Python software package, is proposed to assess patterns in the presence of numerical outcomes using well-established measures together with a novel principle able to statistically assess the correlation gain of the subspace against the overall space. Results confirm the possibility to soundly extend discriminative criteria towards numerical outcomes without the drawbacks well-associated with discretization procedures. A case study is provided to show the properties of the proposed method.

Availability: DISA is freely available at <https://github.com/JupitersMight/DISA> under the MIT license.

Contact: {leonardoalexandre,rmch}@tecnico.ulisboa.pt and {rs.costa}@fct.unl.pt

1 Introduction

The discovery of discriminative patterns have proven essential to support predictive and descriptive tasks in biological and medical data domains [2, 4, 7, 10, 11, 13]. In contrast with classic informative patterns, the ability to discriminate an outcome of interest is assessed and, possibly, incorporated in the discovery process [1]. In this context, statistical significance (probability of pattern occurrence against expectations) and discriminative power views are combined in pattern-centric models to aid diagnostics and study regulatory responses to events of interest [1, 7].

Despite the relevance of extending biclustering and pattern mining tasks with discriminative criteria in the presence of target variables, existing contributions are centered on categorical outcomes. To our knowledge, there are no software packages able to robustly assess subspace rules in the presence of numerical outcomes [5, 9]. We propose DISA (Discriminative and Informative Subspace Analysis), a software package in Python (v3.7) to assess patterns with numerical outputs by statistically testing the correlation gain of the subspace against the overall space.

Fig. 1. Example of class-conditional subspaces with varying homogeneity. The constant subspace has pattern (value expectations) $\varphi_{J_1} = \{c_1=1.1, c_2=0.45, c_3=0.9\}$, while the order-preserving subspace satisfies the $y_2 \leq y_3 \leq y_1$ permutation on 3 observations.

The intersecting points between the two probability density functions is computed to identify the range of values discriminated by the subspace.

DISA supports 53 pattern-centric metrics in total (list in supplementary material) in linear time of observations, returning results in clusters, each ordered by significance to the user.

2 Methods

A list of subspaces and outcome observations (Figure 2) are the minimum necessary input for DISA. If DISA receives a numerical outcome, a range of values in which samples are valid is determined. DISA accomplishes this by approximating two probability density functions (e.g. Gaussians), one for all the observed targets and the other with targets of the target subspace.

Fig. 2. DISA workflow. Input: multivariate data (optional), list of subspaces, outcome variable; Statistical calculus: subspace properties, discriminated ranges from pdf intersection points (numerical outcomes only), and metrics (e.g. statistical significance, gini index, information gain); Output: list of metrics per subspace.

3 Discussion

To illustrate DISA properties, we considered *yeast* dataset [8] and Breast Cancer Wisconsin (diagnostic) dataset [12] (details on preprocessing in supplementary material), both available at the UCI repository [3]. Table 1 provides a synthesis on the DISA assessment over two subspaces produced by BiCPAMS, a state-of-the-art pattern-based biclustering search. P1 corresponds to a subspace in yeast data and P2 to a subspace in breast cancer data (full list is provided in supplementary material).

Table 1. Properties of patterns P1, $\{mcg_{(4)}, gvh_{(4)}, erl_{(0)}, pox_{(0)}\}$, and P2, $\{radius_{0_{(4)}}, perim_{0_{(4)}}, area_{0_{(4)}}, perim_{2_{(4)}}, area_{2_{(4)}}\}$, as well as the span of each property and interestingness threshold.

	P1	P2	Metric interval	Interestingness reference threshold
Information Gain [13]	0.05	0.02	[0,1]	> 0.6
Gini Index [13]	0.01	8×10^{-3}	[0,1]	> 0.6
Chi-Squared [2]	56.10	3.3	$[0, \infty]$	> 3.84
Standardise Lift [10]	0.73	0.71	[0,1]	> 0.6
Stat. Significance [6]	2×10^{-28}	1×10^{-54}	[0,1]	< 0.05

Figure 3 shows the approximated curves and intersection points associated with P1 and P2 patterns. P1 and P2 approximately discriminate $[0.21, 0.5]$ and $[-0.17, 0.44]$ outcome ranges, respectively.

Fig. 3. Gaussian intersections between the outcome variable and the subspace of the outcome variable. Blue line represents the Gaussian of the subspace, orange line represents the Gaussian of the original outcome space. P2 intersections, -0.17 and 0.44, correspond to -20 and 55 on the original scale.

4 Conclusion

DISA is an open source package capable of robustly assessing the statistical significance and discriminative power of association rules in the presence of numerical and categorical outcomes. DISA implements over 50 metrics,

heuristics that can be used to guide the discovery process of discriminative patterns and subspace clusters in biomedical data domains. DISA can be easily embed, therefore aiding the scientific community ability along pattern-centric descriptive and predictive tasks with numerical outcomes.

Funding

This work is supported by Portuguese Foundation for Science and Technology (FCT) under LAETA project (UIDB/50022/2020), IPOscore with reference (DSAIPA/DS/0042/2018), and ILU (DSAIPA/DS/0111/2018). This work was further supported by LAQV, financed by national funds from FCT/MCTES (UIDB/50006/2020 and UIDP/50006/2020), INESC-ID plurianual (UIDB/50021/2020), the contract CEECIND/01399/2017 to RSC and the FCT individual PhD grant 2021.07759.BD to L.A.

References

[1] L. Alexandre, R. S. Costa, L. L. Santos, and R. Henriques. Mining pre-surgical patterns able to discriminate post-surgical outcomes in the oncological domain. *IEEE Journal of Biomedical and Health Informatics*, 2021.

[2] S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets: Generalizing association rules to correlations. In *Proceedings of the 1997 ACM SIGMOD international conference on Management of data*, pages 265–276, 1997.

[3] D. Dua and C. Graff. UCI machine learning repository, 2017.

[4] G. Fang, G. Pandey, W. Wang, M. Gupta, M. Steinbach, and V. Kumar. Mining low-support discriminative patterns from dense and high-dimensional data. *IEEE Transactions on Knowledge and Data Engineering*, 24(2):279–294, 2010.

[5] M. Hahsler, S. Chelluboina, K. Hornik, and C. Buchta. The arules r-package ecosystem: analyzing interesting patterns from large transaction data sets. *The Journal of Machine Learning Research*, 12:2021–2025, 2011.

[6] R. Henriques and S. C. Madeira. Bsig: evaluating the statistical significance of biclustering solutions. *Data Mining and Knowledge Discovery*, 32(1):124–161, 2018.

[7] R. Henriques and S. C. Madeira. Flebic: Learning classifiers from high-dimensional biomedical data using discriminative biclusters with non-constant patterns. *Pattern Recognition*, 115:107900, 2021.

[8] P. Horton and K. Nakai. A probabilistic classification system for predicting the cellular localization sites of proteins. In *Ismb*, volume 4, pages 109–115, 1996.

[9] S. Kaiser, R. Santamaria, T. Khamiakova, M. Sill, R. Theron, L. Quintales, F. Leisch, E. De Troyer, and M. S. Kaiser. Package ‘biclust’. *The Comprehensive R Archive Network*, 2015.

[10] P. D. McNicholas, T. B. Murphy, and M. O’Regan. Standardising the lift of an association rule. *Computational Statistics & Data Analysis*, 52(10):4712–4721, 2008.

[11] E. R. Omiecinski. Alternative interest measures for mining associations in databases. *IEEE Transactions on Knowledge and Data Engineering*, 15(1):57–69, 2003.

[12] W. N. Street, W. H. Wolberg, and O. L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. In *Biomedical image processing and biomedical visualization*, volume 1905, pages 861–870. International Society for Optics and Photonics, 1993.

[13] P.-N. Tan, V. Kumar, and J. Srivastava. Selecting the right interestingness measure for association patterns. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 32–41, 2002.