SPARSE RULE GENERATING FOLD-CHANGE CLASSIFICATION FOR MOLECULAR HIGH-THROUGHPUT PROFILES

Annika MTU Kestler[1], Nensi Ikonomi[1] , Silke D Werle[1], Julian D Schwab[1], Friedhelm Schwenker[2] and Hans A Kestler[1,3]

[1] Medical Systems Biology, and

[2] Neural Information Processing,

Ulm University, Albert-Einstein-Allee 11, Ulm, 89077, Germany

(e-mail: {annika.kestler, nensi.ikonomi, silke.kuehlwein, julian.schwab, friedhelm.schwenker, hans.kestler}@uni-ulm.de)

[3] Corresponding author

ABSTRACT: Classifying gene expression profiles can be challenging due to their low sample size and high dimensionality. Existing methods employed often lack interpretability or sparsity, and require extensive data preprocessing. Ensemble methods, such as the Set Covering Machine, enable the construction of classifiers depending only on base classifiers. We propose two novel base classifiers that consider relations between features for constructing interpretable decision functions, denoted fold change classifiers. Here, an intrinsic feature selection and a straightforward semantic and syntactic interpretation can be achieved. The proposed classifier no longer depends on equally scaled data since relative measurements within a sample are considered. The applicability of the proposed method is shown in a case study evaluating neuroendocrine tumors.

KEYWORDS: ensemble method, molecular high-dimensional data, set covering machine, fold changes, neuroendocrine tumors

1 Introduction

Classification in light of potentially formulating biological hypotheses often entails classifying high-dimensional data, where the number of samples is greatly outnumbered by the number of dimensions of each sample (Lausser & Kestler, 2013; Marchand & Shah, 2004). Each dimension of a sample is referred to as a feature, and finding distinctions between samples, that only depend on a subset of features, might lead to the formulation of novel biological hypotheses (Lausser & Kestler, 2013; Marchand & Shah, 2004). Moreover,

discarding irrelevant features can be considered essential in order to obtain sparse and interpretable classifiers, with high generalization abilities (Marchand & Shah, 2004).
The *Set Covering Machine* (SCM), enables the construction of a sparse conjunction of base classifiers (Marchand & Shawe-Taylor, 2003), performing an intrinsic feature selection when using a single threshold on one feature as a base classifier. Previous work in this context and these types of classifiers has been done mainly by (Marchand & Shawe-Taylor, 2003, Valiant, 1984, Haussler, 1988), and more recently by others (Drouin *et al.*, 2016; Drouin *et al.*, 2019; Lausser & Kestler, 2013; Schmid *et al.*, 2013; Kestler *et al.*, 2006; Lausser *et al.*, 2020). A resulting interpretable decision function can be of the form "**IF** $f_1 \geq 5$ **AND** $f_2 < 8$ **THEN** the sample belongs to class ...", with f_1 and f_2 being features / genes.
When analyzing high-throughput expression profiles the mere over- or under-expression of single genes might not suffice to identify biologically relevant genes (Shi *et al.*, 2005). Therefore, considering relations between different gene expressions, by pairwise comparing expressions could lead to the identification of global behaviors and point to biological processes involved (Shi *et al.*, 2005). This motivates base classifiers of type $f_1 < f_2$ or $f_1/f_2 \geq t$ where t is a threshold, relating the two features considered. These base classifiers may be less susceptible to noise, as well as exhibit invariance properties (Lausser & Kestler, 2013). This allows the discovery of similar tendencies among different samples, without depending on identical nomalization of the data.

2 Results

Contrary to the originally published SCM, which constructs a classifier depending on a subset of provided samples (Marchand & Shawe-Taylor, 2003), we are able to construct a sparse classifier, depending only on a subset of features, while eliminating concerns about normalization and data-preprocessing. Due to the interpretable decision functions, learnt by our proposed method, a genotype-to-phenotype relation can be established, potentially revealing novel biological mechanisms.
We employed the proposed method in a case study dealing with pancreatic neuroendocrine tumours (PanNETs). PanNETs are rare but quite heterogeneous tumour entities lacking specific biomarkers for disease progression. The resulting decision function, an ensemble of order relations, is sparse and yields perfect reclassification contrary to other classification methods employed on this data. Here, the gene relations involved in the decision functions could be

validated via a literature search, suggesting mechanistic interactions to be further investigated. The restriction to the evaluation of order relations reduces the gained flexibility of the presented base classifiers. This can be further validated by the sparsity of the decision function, implying that the base classifier involved carry much information.

References

DROUIN, ALEXANDRE, GIGUÈRE, SÉBASTIEN, DÉRASPE, MAXIME, MARCHAND, MARIO, TYERS, MICHAEL, LOO, VIVIAN G., BOURGAULT, ANNE-MARIE, LAVIOLETTE, FRANÇOIS, & CORBEIL, JACQUES. 2016. Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. *BMC Genomics*, **17**(1), 754.

DROUIN, ALEXANDRE, LETARTE, GAËL, RAYMOND, FRÉDÉRIC, MARCHAND, MARIO, CORBEIL, JACQUES, & LAVIOLETTE, FRANÇOIS. 2019. Interpretable genotype-to-phenotype classifiers with performance guarantees. *Scientific Reports*, **9**(1), 4071.

HAUSSLER, DAVID. 1988. Quantifying Inductive Bias: AI Learning Algorithms and Valiant's Learning Framework. *Artificial Intelligence*, **36**(2), 177–221.

KESTLER, HANS A., LINDNER, WOLFGANG, & MÜLLER, ANDRÉ. 2006. Learning and Feature Selection Using the Set Covering Machine with Data-Dependent Rays on Gene Expression Profiles. *Pages 286–297 of:* SCHWENKER, FRIEDHELM, & MARINAI, SIMONE (eds), *Artificial Neural Networks in Pattern Recognition*. Berlin, Heidelberg: Springer Berlin Heidelberg.

LAUSSER, LUDWIG, & KESTLER, HANS A. 2013. Fold Change Classifiers for the Analysis of Gene Expression Profiles. *Pages 193–202 of:* GAUL, WOLFGANG, VICHI, MAURIZIO, & WEIHS, CLAUS (eds), *Studies in Classification, Data Analysis, and Knowledge Organization*. Heidelberg: Springer.

LAUSSER, LUDWIG, SZEKELY, ROBIN, KLIMMEK, ATTILA, SCHMID, FLORIAN, & KESTLER, HANS A. 2020. Constraining classifiers in molecular analysis: invariance and robustness. *Journal of The Royal Society Interface*, **17**(163), 20190612.

MARCHAND, MARIO, & SHAH, MOHAK. 2004. PAC-Bayes Learning of Conjunctions and Classification of Gene-Expression Data. *Pages 881–*

888 of: SAUL, L., WEISS, Y., & BOTTOU, L. (eds), *Advances in Neural Information Processing Systems*, vol. 17. Boston, MA: MIT Press.

MARCHAND, MARIO, & SHAWE-TAYLOR, JOHN. 2003. The Set Covering Machine. *Journal of Machine Learning Research*, **3**, 723–746.

SCHMID, FLORIAN, LAUSSER, LUDWIG, & KESTLER, HANS A. 2013. Three Transductive Set Covering Machines. *Pages 303–311 of:* GAUL, WOLFGANG, VICHI, MAURIZIO, & WEIHS, CLAUS (eds), *Studies in Classification, Data Analysis, and Knowledge Organization*. Berlin, Heidelberg: Springer International Publishing.

SHI, LEMING, TONG, WEIDA, FANG, HONG, SCHERF, UWE, HAN, JING, PURI, RAJ K., FRUEH, FELIX W., GOODSAID, FEDERICO M., GUO, LEI, SU, ZHENQIANG, HAN, TAO, FUSCOE, JAMES C., XU, Z. AALEX, PATTERSON, TUCKER A., HONG, HUIXIAO, XIE, QIAN, PERKINS, ROGER G., CHEN, JAMES J., & CASCIANO, DANIEL A. 2005. Cross-platform comparability of microarray technology: Intra-platform consistency and appropriate data analysis procedures are essential. *BMC Bioinformatics*, **6**(2), S12.

VALIANT, LESLIE G. 1984. A Theory of the Learnable. *Communications of the ACM*, **27**(11), 1134–1142.