IRELAND: an MILP-based algorithm for learning interpretable input-output relationships from large binary classification data

Marleen Balvert[1][0000−0002−2376−9301]

Department of Econometrics & Operations Research, Tilburg School of Economics and Management, Tilburg University, Tilburg, the Netherlands
m.balvert@tilburguniversity.edu

In the field of machine learning the focus has been largely on the development of methods that can achieve increasingly high classification and prediction accuracies. In the past couple of years the interpretability of classification and prediction methods has gained attention as well. Motivated by applications in bioinformatics, where understanding the input-output relationship is at least as important as obtaining good predictions, this work focuses on the development of a method that learns the relationship between binary input- and output data, where the input-output relationship is described by a Boolean phrase in disjunctive normal form (DNF). Existing methods have been shown to work very well on small datasets and give intuitive results, but run into computational issues for datasets with a large number of samples and input features. With the recent advent of big data, particularly in the field of bioinformatics, there is a need for a faster solution algorithm. The algorithm proposed here allows for the analysis of datasets with tens of thousands of samples and features.

A Boolean phrase in DNF is a truth statement formulated as an OR combination of AND clauses. Let $x_{n,p} = 1$ if sample n has feature p and 0 otherwise, for samples $1, ..., n$ and features $1, ..., p$. An example of a Boolean phrase in DNF is: "IF ($x_{n,12} = 1$ AND $x_{n,35} = 1$ AND $x_{n,6} = 1$) OR ($x_{n,21} = 1$) OR ($x_{n,42} = 1$ AND $x_{n,43} = 1$), THEN sample n is predicted to be in class 1, else in class 0". Note that binary output data implies that a sample is in either of two classes. Boolean phrases in DNF are very easy to comprehend and interpret [7]. They are particularly useful for describing the relationship between genetic information and phenotype, e.g. cell behavior, disease susceptibility or drug response, which is the motivation behind this work.

Mixed integer linear programming (MILP) can be used to learn Boolean phrases from binary data such that the accuracy of the class predictions is maximized [1–3, 5, 6, 8, 9]. The number of binary variables in such an MILP increases with the number of data samples, the number of features per sample and the number of AND clauses included in the final Boolean phrase. As MILPs with a large number of variables are notoriously computationally expensive, these models only work well on relatively small datasets with at most 1,000 samples and 4 or 5 AND clauses.

This work proposes an algorithm called IRELAND - Iterative Rule Extension for the Logical ANalysis of Data - to solve the MILP for datasets of up to at least 10,000 samples and an unlimited number of AND clauses. Similar to [4], [8]

and [2], IRELAND iteratively generates AND clauses by alternating between a master problem that selects the best AND clauses from a pool to be included in the final Boolean phrase, and a sub problem that generates new promising AND clauses. When the number of samples is large, the sub problem is still computationally expensive. However, note that the newly generated AND clause should result in an increased number of true positives when added to the Boolean phrase in the previous iteration, without adding too many false positives. IRELAND makes use of this observation by excluding the class 1 samples that were already classified as class 1 in the previous iteration from the sub problem. This strongly reduces the computational complexity. Given the remaining set of samples, the sub problem maximizes the number of true positives while restricting the number of false positives, and is solved for various upper bounds on the number of false positives. The master problem then selects AND clauses from the pool such that a weighted sum of true positives and true negatives is maximized.

Classification problems are generally bi-objective problems: one aims to maximize the number of true positives, while minimizing the number of false positives. The framework of IRELAND allows for computing the trade-off curve between these conflicting objectives at minimal computational cost. Recall that a pool of AND clauses is generated where the number of false positives varies between AND clauses. The master problem can then be reformulated into a bi-objective problem using the ε-constraint method and use this pool to generate Boolean phrases that form the Pareto curve between true and false positives.

IRELAND was tested on synthetic datasets, where the number of samples, features, AND clauses and the level of noise were varied. The results show that for small datasets, those with $N \leq 1,000$, IRELAND has no benefit over using the original MILP, and is often even slower. However, for $N > 2,000$, directly solving the MILP often takes several hours, where the model was always stopped after four hours, or even runs out of memory. For these large datasets IRELAND found a solution with high classification accuracy within one hour. In addition, IRELAND provided the full Pareto front showing the trade-off between sensitivity and specificity. IRELAND thus enables the abstraction of Boolean phrases in DNF for large complex problems that could not be solved with the MILP formulation.

References

1. Chang, A., Bertsimas, D., Rudin, C.: An integer optimization approach to associative classification. In: Advances in neural information processing systems. pp. 269–277 (2012)
2. Dash, S., Gunluk, O., Wei, D.: Boolean decision rules via column generation. In: Advances in Neural Information Processing Systems. pp. 4655–4665 (2018)
3. Hammer, P.L., Bonates, T.O.: Logical analysis of data—an overview: From combinatorial optimization to medical applications. Annals of Operations Research **148**(1), 203–225 (2006)
4. Hansen, P., Meyer, C.: A new column generation algorithm for logical analysis of data. Annals of Operations Research **188**(1), 215–249 (2011)

5. Hauser, J.R., Toubia, O., Evgeniou, T., Befurt, R., Dzyabura, D.: Disjunctions of conjunctions, cognitive simplicity, and consideration sets. Journal of Marketing Research **47**(3), 485–496 (2010)
6. Knijnenburg, T.A., Klau, G.W., Iorio, F., Garnett, M.J., McDermott, U., Shmulevich, I., Wessels, L.F.: Logic models to predict continuous outputs based on binary inputs with an application to personalized cancer therapy. Scientific reports **6**, 36812 (2016)
7. Lakkaraju, H., Bach, S.H., Leskovec, J.: Interpretable decision sets: A joint framework for description and prediction. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1675–1684 (2016)
8. Malioutov, D., Varshney, K.: Exact rule learning via boolean compressed sensing. In: International Conference on Machine Learning. pp. 765–773. PMLR (2013)
9. Wang, T., Rudin, C.: Learning optimized or's of and's. arXiv preprint arXiv:1511.02210 (2015)