Essay 2 - Pattern Aided Regression (PXR) and Exceptional Model Mining (EMM)

Najwa Laabid, University of Eastern Finland May 25, 2020

Introduction

Many datasets are made of heterogeneous subgroups. This heterogeneity stands in the way of approximating predictor-response relationships. To overcome this issue, Dong and Taslimitehrani (2015) propose regression models capable of recognizing subpatterns in data. Duivesteijn et al. (2012) suggest a method to look for exceptional regression models explicitly (i.e., looking for subgroups which modify regression coefficient substantially when used as a single training set for the model). Duivesteijn et al. (2015) provide an overview of Exceptional Model Mining in general, with an overview of its methods, benefits, and where it fits in the data mining field in general. This article will provide an overview of these three papers, with a focus on the key innovations in each work and how they relate to one another. The following sections will provide an analysis of each paper in turn before concluding with a general critic and open questions concerning all three papers.

CPXR

State-of-the-art regression models fall short on two fronts: accuracy in prediction and interpretability of results. Dong and Taslimitehrani (2015) reckon that the issue facing these models is their inability to capture predictor-response relationships of distinct logical (pattern-specific) subgroups. The article proposes a new type of regression models, dubbed Pattern Aided Regression (PXR), capable of capturing diverse subgroup relationships, alongside an algorithm to build PXR models systematically. The key idea of PXR is to construct triplets of patterns, local models, and a weight reflecting the error reduction capability of the model (P_i, f_{P_i}, w_i) fitted on a pattern-specific subgroup i. The algorithm starts by splitting the dataset to large and small residual errors' groups with respect to the global model (LE and SE), such that the ratio of the large error group to the entire dataset is greater than or equal to a user-defined value p. The algorithm then looks for patterns contrasting LE to SE, using the GcGrowth algorithm, group them by equivalence classes (EC) based on matching datasets, and save the shortest pattern from each EC in a patternset PS. These patterns are known as contrast patterns because they identify subgroups of different error groups (LE or SE). Each pattern in PS is added as a property to a regression model, which is then evaluated on its ability to reduce the residual errors compared to the global model. Only the models with able to reduce the error rate the most are kept, forming a set PM. We then add a model f_d fitted on all data points not belonging to any of the patterns of PM. This technique of building pattern-aided regression models is known

as contrast pattern-aided regression (CPXR). In practice, PXR returns either 1) the output of a local prediction function when the data point fits a known pattern, 2) a weighted average of the results of multiple models when the data points fits multiple patterns, or 3) the output of its default model f_d otherwise. The distribution of data among these three options is influenced by the user-defined value p, which dictates the number of patterns found in the data.

To test the performance of the algorithm, two CPXR implementations, using linear regression and piecewise linear regression as local models respectively, were evaluated on 50 real datasets, and compared to 4 state-of-the-art regression models. The results show that CPXR has a higher accuracy, is less prone to noise and data sensitivity, and achieves faster results using contrast rather than frequent patterns. It beats most methods in interpretability, and produces a higher accuracy than similarly-interpretable piecewise linear regression models. One difficulty facing this method though is the high cost of its pattern mining step, which remains a computational bottleneck despite reported pruning strategies (such as choosing the smallest pattern from every EC class). Better results can be achieved if we can somehow identify the patterns with the highest residual error reduction without fitting local models to all potential candidates. This investigation can be a good venue for future research. CPXR still remains a valuable approach, showing promise in linear as well as non-linear regression models, among which is logistic regression as reported by other research (Taslimitehrani and Dong (2014)).

EMM

Duivesteijn et al. (2012) investigates a generalization of subgroup discovery (SD), known as Exceptional Model Mining (EMM). Much like CPXR, the focus here is on subsets of data for which a (linear regression) model fitted on the subset differs from the same model fitted on the entire dataset. Identifying these groups is relevant to many applications, particularly when the behavior of the subgroup goes against established laws and/or expectations (e.g., Giffen Behavior: counterintuitively, a population's demand for a product increases with its price under certain circumstances). EMM improves on SD in the following aspects: ability to handle multiple response variables, performing a level-wise heuristic search instead of a depth-wise exhaustive search adopted, and the use of Cook's distance as a quality measure. Cook's distance seems to be a particularly innovative approach, since it takes into consideration the variance of the estimated coefficients, is immune to the difference in scale of the variables, and offers upper bounds useful in pruning the list of candidates without performing explicit regression computations.

The main difference between this paper and Dong and Taslimitehrani (2015) is the change of quality measure from total reduction of residual errors (TRR) to Cook's distance. In addition to taking variables' variance into consideration, the new method opened the possibility for pruning the search space with minimal computations. However, the upper bounds used in this pruning strategy require the subgroup size to make up around 50% of the records, which is a rather restrictive condition. Improving on this result requires future work. Another way in which Cook's distance can be improved is through estimating the distance of removing multiple records of subgroups simultaneously by summing the distances of removing single records, which have a

definite simplified formula. This approach has been shown to be flawed theoretically, but may still provide practically relevant results. Even without these improvements, it would be interesting to see how PXR reacts to using Cook's distance and its estimates in place of reduction of residual errors. Cook's distance can be used both to identify the patterns in the LE set and select the most 'different' ones to make up the final model. This merge is likely to show at least an improvement in the speed of execution of PXR.

EMM - extended

Duivesteijn et al. (2015) extends the ideas of Duivesteijn et al. (2012) by providing an overview of EMM with different response models (such as correlation, association, bayesian networks, etc). The paper also proposes various quality measures (taking into account the model choice and group size) falling in three paradigms: statistical tests, entropy functions, and difference quantification. The question of whether to use the entire dataset or a complement of the chosen model as a reference in the comparison of local models has also been addressed, with an overview of the considerations of either choice. Since search strategies are a key aspect to the practicality of this framework, various ideas were scraped from different subfields and gauged for their merits and shortcomings. Finally, the bulk of the article was dedicated to presenting 6 different model families and evaluating the quality measures appropriate to each one of them. These families were tested experimentally, revealing interesting insights in at least two datasets: intriguing animal dependencies in the Europe's mammals' data, and a real-life example of the Giffen behavior in the economical data of the Chinese province Hunan. The paper also argued for the usefulness of EMM beyond finding interesting patterns. Notably, the framework can be used to aid meta-learning tasks.

The article proposed a comprehensive study of the EMM framework and many facets of research relating to it. It was in essence a compilation of the authors' work in the field for almost a decade, with more attention given to formally defining the problem of exceptional model mining and a few related concepts. The article failed to point out shortcomings of the framework explicitly. Complex computations appear to be a constant bottleneck of the method. Defining a universal description language for the mined subgroups is also eluding efforts. As with all search-based approaches, identifying new ways to prune the search space is an area of active research. Since EMMs show promise in their current form, extending their application to as many domains as possible is likely to benefit said domains as well as the mining framework by proposing new problem-specific insights.

Conclusion

The three articles explored the topic of identifying subgroups through special influence on the fitted model. Dong and Taslimitehrani (2015) considered identifying these subgroups using contrast patterns, then using these patterns to build a pattern-aided regression model with a reported accuracy and interpretability higher than state-of-the-art models. Duivesteijn et al.

(2012) looked at exceptional model mining using Cook's distance as a quality measure, and investigated the various upper bounds and computational alternatives it offers. Duivesteijn et al. (2015) provided an overview of the EMM framework in relation to models beyond linear regression and proposed adequate quality measures for each one of them. One difference between EMM and XPR, as pointed out by Duivesteijn et al. (2015), lies in the fact that the subgroups revealed by XPR do not seek to explore exceptionality, but rather improve model fitting, and may therefore be hard to interpret as coherent data subsets. The topic in general fits with the topics explored in class in that it presents alternative ways to identify subsets of the data displaying interesting behavior, with an attempt to formally define and quantify interest in this context. As with most data mining fields, this topic remains an open area of active research, facing the permanent weight of high computational complexity.

References

G. Dong and V. Taslimitehrani. Pattern-aided regression modeling and prediction model analysis. *IEEE Transactions on Knowledge and Data Engineering*, 27(9):2452–2465, 2015.

Wouter Duivesteijn, Ad Feelders, and Arno Knobbe. Different slopes for different folks: Mining for exceptional regression models with cook's distance. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, page 868–876, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450314626. doi: 10.1145/2339530.2339668. URL https://doi.org/10.1145/2339530.2339668.

Wouter Duivesteijn, Ad Feelders, and Arno Knobbe. Exceptional model mining: Supervised descriptive local pattern mining with complex target concepts. 30:47–98, 01 2015. doi: 10.1007/s10618-015-0403-4.

V. Taslimitehrani and G. Dong. A new cpxr based logistic regression method and clinical prognostic modeling results using the method on traumatic brain injury. In *2014 IEEE International Conference on Bioinformatics and Bioengineering*, pages 283–290, 2014.