Model Reuse with Subgroup Discovery

Hao Song

Intelligent Systems Laboratory, University of Bristol, United Kingdom
Hao.Song@bristol.ac.uk

Abstract. In this paper we describe a method to reuse models via a subgroup discovery approach. The task is to predict the number of bikes at a rental station three hours later. Instead of training new models, our approach perform a subgroup discovery based on a number of pre-trained models from other bike stations. The experiments show that our approach can significantly reduce the mean absolute error comparing to a simple bagging strategy.

1 Introduction

In this paper we address the problem of model reuse based on a machine learning challenge MoReBikeS hold with ECML-PKDD 2015. The applied method is a extension of the Subgroup Discovery frame work, together with a bagging strategy.

2 Subgroup Discovery

Subgroup discovery [1–4] is a data mining technique that uses rules to describe statistical deviations in large subsets of the data. It can be seen as a descriptive model learnt in a supervised way [5]. Supervised learning means that the construction of subgroups is driven by a particular target variable. However, rather than to predict this variable, the aim is to detect different subsets of the dataset where the selected variable has a significant statistical deviation. As such deviations are likely to occur in small subsets of the data, subgroup discovery also takes the size of the found subsets into consideration. Therefore the evaluation of a subgroup would normally depend on a quality measure based on the unusual distribution along with the size of the subgroup.

In this paper, a variation of subgroup discovery, called Model-Based Subgroup Discovery (MBSD) is proposed to reuse the trained models. The approach is to perform subgroup discovery with the prediction error (e.g. mean absolute error) for a particular trained model. The obtained subgroups than represent a subset of the data that the base model doesn't predict very well. The modified quality measure is:

$$\phi_{CWRAcc}(\mathbf{g}) = \frac{n(\mathbf{g})}{N} \cdot (E(\mathbf{z} \mid \mathbf{g}) - E(\mathbf{z})) \tag{1}$$

Where $\mathbf{g}$ is a particular subgroup, $E()$ is the expectation function, $\mathbf{z} = |\hat{\mathbf{y}} - \mathbf{y}|$ represents the mean absolute error. $\hat{\mathbf{y}} = f(\mathbf{x})$ is the prediction made by a particular model and $\mathbf{y}$ is the ground truth. To reduce the search complexity, in this paper each subgroup is expressed by a single rule on one attribute. A example could be $weekhour \leq 120$.

3 Experiments

The experiments is designed as following. As totally 6 models (rlm full, rlm full temp, rlm short, rlm short full, rlm short full temp, rlm short temp) were trained for station 1 to 200, the MAE for $200 * 6 = 1200$ trained models are calculated iteratively for all the 275 stations. The model with lowest MAE for each station is selected as the base line model. For instance, the base model for station 201 is the model 3 (rlm short) from station 1.

Since a single model can potentially lead to over-fitting, a simple bagging (with models ordered by their MAE on the station to be predicted) is adopted. Then a MBSD is performed to all the models in the bagging, a sub-model is selected for the best subgroup associated with each trained model. The prediction with the subgroup is hence a average of the original model and the sub-model. The error curve for base model, simple bagging approach, and MBSD bagging approach is given in Fig 1.

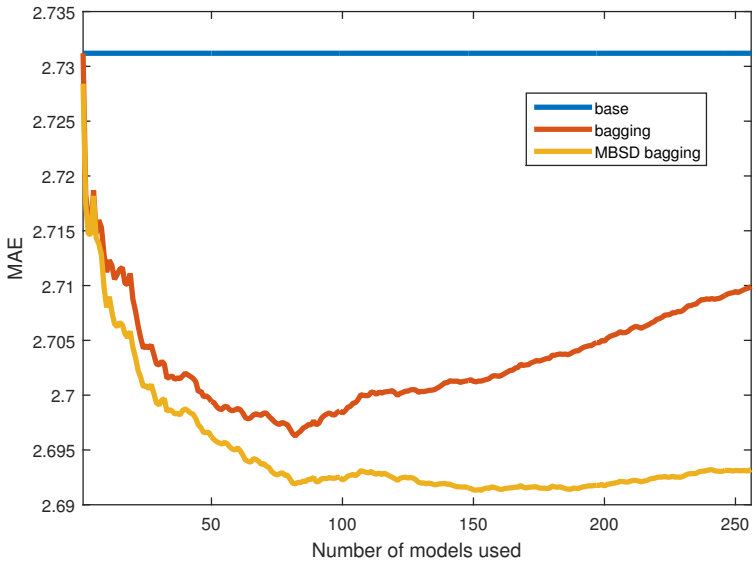

Fig. 1: The error curve for the 3 approaches: base (blue), simple bagging (red), MBSD bagging (orange).

As illustrated by the figure, while bagging is capable to reduce the MAE, it tends to over-fit as the number of model increases. On the other hand, our MBSD approach is robust to over-fitting and can help further reduce the MAE from a simple bagging strategy.

Another issue here would to select the number of models for the final prediction. One strategy we adopted during the leader broad submission was to select the best

number for full year's data of the 10 stations. However it didn't give the best results among all the submissions. One potential reason could be the test data in the leader broad is of small scale, therefore bagging could lead to over-fitting.

References

1. Willi Klösgen. Advances in knowledge discovery and data mining. chapter Explora: A Multipattern and Multistrategy Discovery Assistant, pages 249–271. American Association for Artificial Intelligence, Menlo Park, CA, USA, 1996.
2. Stefan Wrobel. An algorithm for multi-relational discovery of subgroups. In *Principles of Data Mining and Knowledge Discovery*, pages 78–87. Springer, 1997.
3. Nada Lavrač, Branko Kavšek, Peter Flach, and Ljupčo Todorovski. Subgroup discovery with CN2-SD. *The Journal of Machine Learning Research*, 5:153–188, 2004.
4. Franciso Herrera, Cristóbal José Carmona, Pedro González, and María José del Jesus. An overview on subgroup discovery: foundations and applications. *Knowledge and information systems*, 29(3):495–525, 2011.
5. Petra Kralj Novak, Nada Lavrač, and Geoffrey I Webb. Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *The Journal of Machine Learning Research*, 10:377–403, 2009.