Model Reuse with Subgroup Discovery

Hao Song and Peter Flach

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

Abstract. In this paper we describe a method to reuse models with Model-Based Subgroup Discovery (MBSD), which is a extension of the Subgroup Discovery scheme. The task is to predict the number of bikes at a new rental station 3 hours in advance. Instead of training new models with the limited data from these new stations, our approach first selects a number of pre-trained models from old rental stations according to their mean absolute errors (MAE). For each selected model, we further performed MBSD to locate a number of subgroups that the selected model has a deviated prediction performance. Then another set of pre-trained models are selected only according to their MAE over the subgroup. Finally, the prediction are made by averaging the prediction from the models selected during the previous two steps. The experiments show that our method performances better than selecting trained models with the lowest MAE, and the averaged low-MAE models.

1 Introduction

In this paper we propose a model reuse approach exploiting Model-Based Subgroup Discovery (MBSD). The general idea of model reuse is to use trained models from other operating contexts under a new operating context. Such a strategy has two main benefits. Firstly, it can dramatically reduce model training time on the new operating contexts. Secondly, if the new operating context only has limited data, as model reuse essentially extends the scale of the training data by adding training data from other operating contexts, it can help further improve the prediction's performance.

One major challenge for model reuse is that the patterns in the data can vary through different operating contexts. This makes it difficult to directly apply trained models from the training contexts to a new context. For instance, to predict the activities of daily living (ADLs) from the reading of sensors is one of the leading applications of a smart home. However, as both the households and layout of the house varies from different houses, it is hard to directly use a model trained in one particular house to another house. Therefore, to recognise and deal with such variations through different operating contexts has become a non-trivial research task for model reuse [1].

In this paper, we will use a variation of the Subgroup Discovery (SD) scheme [2–4] to help the reused models to adopt the new context. SD is a data mining technique. It uses a descriptive model to learn the unusual statistic of a target variable in a given data-set. However, traditional SD approaches generally focus on the statistic of a single attribute in a fixed data-set. This makes it less appropriate for model reuse. Therefore, we propose an extended method MBSD in this paper. The main modification is to change

the target variable in the SD task from an attribute into the prediction performance of an attribute from a particular base model. Through this modification, MBSD can be used to discover the prediction pattern of a trained model in a new operating context. This can help locate the potential sub-context where the trained model can be directly applied, or the potential sub-context where other trained models are required.

The experiments are based on a machine learning challenge MoReBikeS, which is organised by the workshop LMCE 2015, within the conference ECML-PKDD 2015. The task is to predict the number of bikes available at a particular rental station 3 hours later, given some history data. In detail, the overall data-set is obtained from 275 bike rental stations located in Valencia, Spain. For the participants, everyone gets access to the data for all the 275 stations during the October of 2014 and use these data as training data. 6 trained linear regression models are also provided for each station from station 1 to station 200. These linear regression models are trained with the data that covers the whole year of 2014. Therefore, the task of the challenge is, by reusing these trained models and limited training data, to predict the number of available bike at some new bike stations (station 201 to station 275).

The method we used can be briefly described as follows. For any station to be predicted, the one-month training data can be applied to select a number of models with good performances (low MAE values), these models are called base-models. The assumption here is that these base-models are only suitable to some unknown sub-context of the context to be predicted (the sub-context is similar to the training context of the base-models), and not suitable to some other sub-context. With such an assumption, we can perform MBSD to discover these sub-contexts and to further select a number of models with good performances only under the sub-contexts. These models are denoted as sub-models. Finally, the overall prediction can be obtained by averaging the prediction from both base-models and sub-models, with some averaging strategy. The experiments show that, with MBSD, the MAE can be further reduced comparing to simple averaging the prediction of the base-models.

This paper is organised as follows. In section 2 some preliminaries of Subgroup Discovery are given and in section 3 the basic concept of MBSD is introduced. The method to reuse models with MBSD is stated in section 4. Section 5 shows some experiments with the MoReBikes data. A conclusion of the whole paper is provided in section 6.

2 Subgroup Discovery

In this section we will give some preliminaries and corresponding notations of SD.

Subgroup Discovery (SD) [2–4] is a data mining technique that learns rules to describe patterns of some attributes in a given data-set. Since the construction of subgroups is driven by some attributes, called target variable, it can be seen as a descriptive model which is learnt in a supervised way. However, SD still differs from predictive models as in SD we are not aiming to predict the target variable, but to discovery some interesting patterns with respect to it. Therefore, the definition of an interesting pattern needs to be given. In existing literature, an interesting pattern often refers to a different class distribution (for binary/nominal target variable), or in general to an unusual statistic (for binary/nominal/numerical target variable). On the other hand, because such

patterns often have a small coverage, some literature also define SD as a model to find patterns that have both large coverage and unusual statistic.

Mathematically, suppose the data-set contains N instances and M attributes. Traditional SD assumes that one from the M attributes is selected as the target variables, the corresponding attribute of the i_{th} instance is denoted as $\mathbf{y}_i \in \mathbb{R}^1$, the domain of this attribute is denoted as $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^N$. The rest $M-1$ attributes are used as the description attributes, denoted as $\mathbf{d}_i \in \mathbb{R}^{M-1}$, the domain of this attributes is denoted as $\mathbf{D} = \{\mathbf{d}_i\}_{i=1}^N$.

A subgroup is denoted as a function $g : \mathbf{D} \to \{0,1\}$. Hence $g(\mathbf{d}_i) = 1$ means that the i_{th} instance is covered by this subgroup and vice versa. We use $\mathbf{G} = \{i : g(\mathbf{d}_i) = 1\}$ to denote the set of instances to be covered by the subgroup g.

The task of (top q) subgroup discovery can be defined as, given a set of candidate subgroups $\mathbb{G} \subseteq 2^{\mathbb{D}}$, and a quality measure $\phi : g \to \mathbb{R}$, to find a set of q subgroups $\mathbb{G}_q = \{g_1, ..., g_q\}$, so that $\phi(g_1) \geq \phi(g_2) \geq ... \geq \phi(g_q)$, and $\forall g_i \in \mathbb{G}_q, \forall g_j \in \mathbb{G} \setminus \mathbb{G}_q$: $\phi g_i \geq \phi g_j$.

With respect to the quality measure, since all the quality measures used in this paper can be seen as a extension of the quality measure Continuous Weighted Relative Accuracy (CWRAcc) [5], the definition of CWRAcc is given here:

$$\phi_{CWRAcc}(g) = \frac{|\mathbf{G}|}{N} \cdot \left(\frac{\sum_{i \in \mathbf{G}} \mathbf{y}_i}{|\mathbf{G}|} - \frac{\sum_{i=1}^N \mathbf{y}_i}{N}\right) \tag{1}$$

3 Model-Based Subgroup Discovery

In this section we will briefly introduce the concept of MBSD, together with the quality measures and search strategy applied in the following experiments. An example of MBSD performing with a particular bike station will be given.

3.1 Motivation

The motivation of MBSD is to import models in a SD process, so that the resulted subgroups can contain richer information. Although this concept is similar to the Exceptional Model Mining (EMM) framework [6, 7], but MBSD differs from EMM by the way of using the models. In EMM, a model will be trained under each candidate subgroup, and then the quality of the subgroup is evaluated by the parameter deviation between the model trained under the subgroup and the model trained under the whole data-set (global model). On the other hand, in MBSD only the global model is involved in the discovery process. For each candidate subgroup, the quality of the subgroup will be evaluated either according to the likelihood of the global model under the subgroup (for both non-predictive models and predictive models), or the prediction performance of the global model in the subgroup (for predictive models). Since the purpose of this paper is to reuse models via MBSD, we omit a detailed discussion about the differences between MBSD and EMM. In general, since in MBSD only a global model is required, repeated training through different candidate subgroups is avoided, this makes MBSD more appropriate for model reuse.

3.2 Quality Measure for Regression Models

As in this paper the MBSD task only involves regression, here we only show 4 quality measures for regression models.

Suppose the target attribute to be predicted is denoted as $\mathbf{y}_i$ for the i_{th} instance and the prediction made by the base-model is denoted as $\hat{\mathbf{y}}_\mathbf{i}$. The first proposed quality measure Weighted Relative Mean Absolute Error (WRMAE) is based on the absolute error of the base model, $\mathbf{z}_i^{AE} = |\hat{\mathbf{y}}_\mathbf{i} - \mathbf{y}_i|$. This quality measure is designed to find subgroups with large coverage and relatively higher MAE than the population.

$$\phi_{WRMAE}(f_{base}, \mathbf{G}) = \frac{|\mathbf{G}|}{N} \cdot \left(\frac{\sum_{i \in \mathbf{G}} \mathbf{z}_i^{AE}}{|\mathbf{G}|} - \frac{\sum_{i=1}^{N} \mathbf{z}_i^{AE}}{N}\right) \tag{2}$$

Similarly, if the aim is to find subgroups where the base model tends to have lower MAE than the population, the negative absolute error $\mathbf{z}_i^{NAE} = -|\hat{\mathbf{y}}_\mathbf{i} - \mathbf{y}_i|$ can be applied. The second proposed quality measure Weighted Relative Mean Negative Absolute Error (WRNMAE) is given as:

$$\phi_{WRMNAE}(f_{base}, \mathbf{G}) = \frac{|\mathbf{G}|}{N} \cdot \left(\frac{\sum_{i \in \mathbf{G}} \mathbf{z}_i^{NAE}}{|\mathbf{G}|} - \frac{\sum_{i=1}^{N} \mathbf{z}_i^{NAE}}{N}\right) \tag{3}$$

Another scenario is to discover the subgroups where the base-model tends to over-estimate the target attribute. Now the quality measure should be designed according to the over-estimated error:

$$\mathbf{z}_i^{OE} = \begin{cases} \hat{\mathbf{y}}_i - \mathbf{y}_i & \text{if } \hat{\mathbf{y}}_i \geq \mathbf{y}_i \\ 0 & \text{otherwise} \end{cases}$$

Notice here the under-estimations are forced to be zeros, hence the quality of subgroups will not be affected by having both high over-estimated error and high under-estimated error. On the other hand, subgroups with both high errors can be discovered with the quality measure WRMAE. The quality measure Weighted Relative Mean Over-Estimated Error (WRMOE) is given as:

$$\phi_{WRMOE}(f_{base}, \mathbf{G}) = \frac{|\mathbf{G}|}{N} \cdot \left(\frac{\sum_{i \in \mathbf{G}} \mathbf{z}_i^{OE}}{|\mathbf{G}|} - \frac{\sum_{i=1}^{N} \mathbf{z}_i^{OE}}{N}\right) \tag{4}$$

As shown above, the under-estimated error and corresponding quality measure Weighted Relative Mean Under-Estimated Error (WRMUE) can be defined as:

$$\mathbf{z}_i^{UE} = \begin{cases} \mathbf{y}_i - \hat{\mathbf{y}}_i & \text{if } \mathbf{y}_i \geq \hat{\mathbf{y}}_i \\ 0 & \text{otherwise} \end{cases}$$

$$\phi_{WRMUE}(f_{base}, \mathbf{G}) = \frac{|\mathbf{G}|}{N} \cdot \left(\frac{\sum_{i \in \mathbf{G}} \mathbf{z}_i^{UE}}{|\mathbf{G}|} - \frac{\sum_{i=1}^{N} \mathbf{z}_i^{UE}}{N}\right) \tag{5}$$

3.3 Description Language and Search Strategy

In traditional SD, the description of subgroups can be built on any attribute other than the target variable. In EMM with predictive models, the description of subgroups can be built on any attribute except the input and output of the model. This is because the essential aim of SD is to use some other attributes to describe the pattern of some target attributes, hence the description should avoid to use the target attributes. However, for MBSD with predictive models, the description of subgroups can potentially be built on any attribute in the data-set. The reason behind this is that the pattern MBSD (with predictive models) tries to describe is the prediction pattern of the base model, instead the pattern of the attributes.

As many other logical models, there exists many ways to split the hypothesis space to generate the candidate subgroups. This generally involves fixing the operations on each attribute. In this paper we will simply use a conjunction of attribute-value pairs as the description language. For numerical attributes, a pre-processing is performed to divide each numerical attribute into equal size bins and further treat them as nominal attributes. Since in the experiments there are a large amount of SD tasks, we also assume all the subgroups are described by any single attribute from the description attributes. This can help further reduce the search cost. Also, for each attribute, only the best subgroup described by that attribute will be selected. Hence top q subgroups can be seen as subgroups described by q attributes respectively.

As only a single attribute is used to describe each subgroup, the search strategy can be seen as a refinement process with adding different values of the corresponding attribute. Here we further use a greedy covering algorithm to increase the search speed and reduce the memory usage. The algorithm is performed as follows. The bin with the highest mean value of the target variable (e.g. AE) is added to the description at each step. The algorithm terminates once the quality measure is smaller than the previous step. This covering algorithm is generally similar to a beam search algorithm except the beam width is fixed as 1, due to the fact that the refinement is done within the same attribute.

3.4 MBSD with Single Bike Station

In the MoReBikeS challenge, there are totally 25 attributes in the data-set. Table 1 summarises the information for each attribute in the provided one-month data, such as name, type (binary, nominal, numerical), number of values, and number of bins configured in the MBSD task (only for numerical attributes).

Although the 25 attributes can all be used to construct the candidate subgroups, as the attribute **bikes** is the variable to be predicted, it can be removed from the description. Also because the MBSD task is going to be performed for each individual station during Oct 2014, the attribute **station**, **latitude**, **longitude**, **year**, **month**, and **timestamp** can be further excluded.

For simplicity, from now on we will use model $i-j$ to refer the model j of station i ($j=1$ for **short**, $j=2$ for **short_temp**, $j=3$ for **full**, $j=4$ for **full_temp**, $j=5$ for **short_full**, $j=6$ for **short_full_temp**).

attribute	type	number of values	number of bins
station	nominal	275	NA
latitude	numerical	275	275
longitude	numerical	275	275
numDocks	numerical	19	19
timestamp	numerical	745	745
year	numerical	1	1
month	numerical	1	1
day	numerical	31	31
hour	numerical	24	24
weekday	numerical	7	7
weekhour	numerical	168	168
isHoliday	binary	2	NA
windMaxSpeed.m.s	numerical	28	28
windMeanSpeed.m.s	numerical	16	16
windDirection.grades	numerical	17	17
temperature.C	numerical	142	16
relHumidity.HR	numerical	72	8
airPressure.mb	numerical	283	32
precipitation.l.m2	numerical	1	1
bikes_3h_ago	numerical	41	41
full_profile_3h_diff_bikes	numerical	17304	32
full_profile_bikes	numerical	17632	41
short_profile_3h_diff_bikes	numerical	419	32
short_profile_bikes	numerical	231	32
bikes	numerical	41	41

Table 1: The 25 attributes in the October data-set and their properties.

For instance, Figure 1 (left) shows the prediction of station 201 from the model $1-1$ during Oct 2014, together with the ground truth. Figure 1 (right) gives the empirical distribution of the prediction errors.

If MBSD is performed with the prediction shown above and the quality measure WRMAE is applied, the best (rank 1) subgroup is found with the attribute **weekhour**. The corresponding attribute values are shown in Figure 2 (left). It can be seen that, since we treat this numerical attribute as a nominal attribute (e.g. the candidate subgroups can contain any combination of attribute values), the found attributes values look sparse. However, there are still some patterns can be told from the figure. For instance, most of the attribute values are located around the night of each day. Figure 2 (right) gives the empirical distribution of the prediction errors within the subgroup. Comparing to Figure 1, here the distribution of errors has a significantly higher variance, which indicates a higher MAE. Figure 3 (left) shows the best subgroup found with the quality measure WRMNAE. Since WRMNAE can be seen as a negative version of WRMAE, it can be seen the best subgroup with WRMNAE is the compliment of the best group of WRMAE.

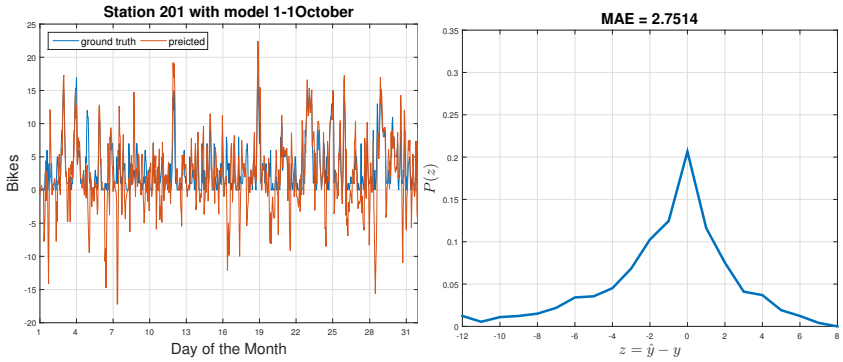

Fig. 1: The prediction of station 1 from the model **short** (left), and the empirical distribution of errors (right).

Rank	WRMAE(WRMNAE)	WRMOE	WRMUE
1	**weekhour**	**weekhour**	**weekhour**
2	**hour**	**full_profile_3h_diff_bikes**	**hour**
3	**full_profile_3h_diff_bikes**	**windMaxSpeed**	**day**
4	**day**	**hour**	**full_profile_3h_diff_bikes**
5	**full_profile_bikes**	**windDirection**	**full_profile_bikes**

Table 2: The description attributes for top-5 subgroups with each quality measure.

Similarly, we can also find subgroups with the quality measure WRMOE and WRMUE. The results (attribute values of the best subgroup and error distribution within the subgroup) of WRMOE and WRMUE are given in Figure 4 and Figure 5 respectively.

It can be seen for all the 4 quality measures the best subgroup is described by the attribute **weekhour**. However, the description attributes for top-q subgroups can vary with different quality measures. The description attributes for top-5 subgroups with each quality measure is given in Table 2.

In general, MBSD can be used to find the deviated prediction patterns in a given data-set. For the regression models, MBSD is set to use one attribute to describe the data points that the base model tends to predict well/not well. Therefore, each attribute used by the subgroups can be seen sharing some non-linear correlation to the model's prediction. This is similar to an attribute selection (e.g. regularisation), but in a non-linear form.

4 Model Reuse with MBSD

In this section we will introduce how to reuse trained models with MBSD. The general idea is, for each deploy context, we can select a bunch of trained models according to their performance. Then with MBSD, we can further detect the (pattern of) data points that the previous models predict well / not well, which can be seen as a sub-context. A

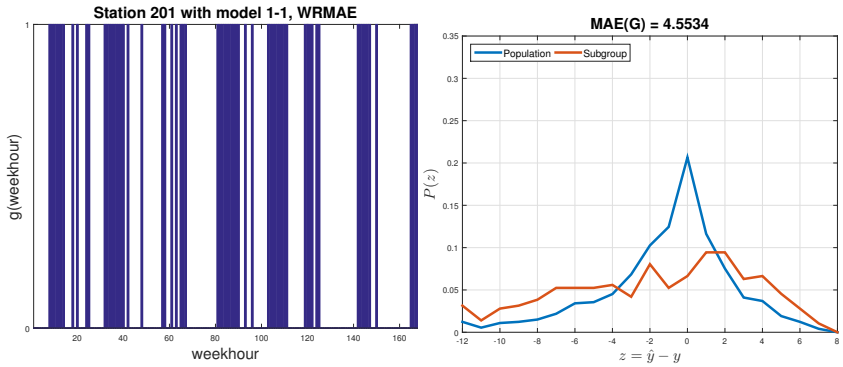

Fig. 2: The best subgroup found with the quality measure WRMAE (left), and the empirical distribution of errors within the subgroup (right).

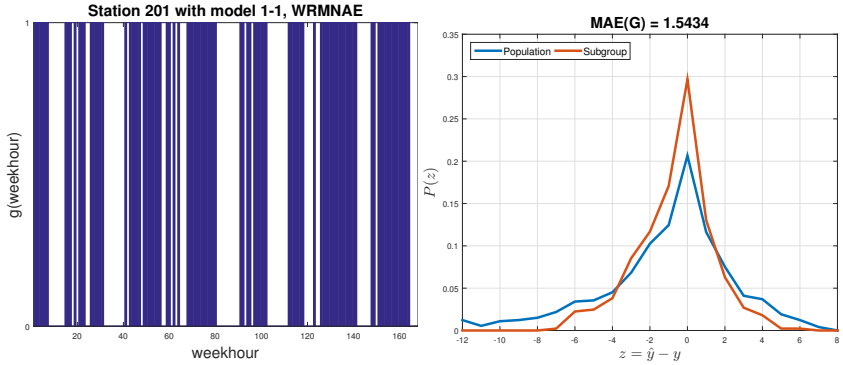

Fig. 3: The best subgroup found with the quality measure WRMNAE (left), and the empirical distribution of errors within the subgroup (right).

number of sub-models are then selected just for these data points. The final prediction is hence estimated by averaging the prediction from the base models and sub-models.

4.1 Baseline Method 1

The first base line method is, for each deploy context, to simply select one model from the 1200 trained models (200 stations, 6 models per station) that has the lowest MAE on the test station.

4.2 Baseline Method 2

The second base line method is, for each test station, to rank the 1200 trained models according to their MAE on the test station. The final prediction is hence the average of the prediction of the top-n models (the selected models are referred as base models):

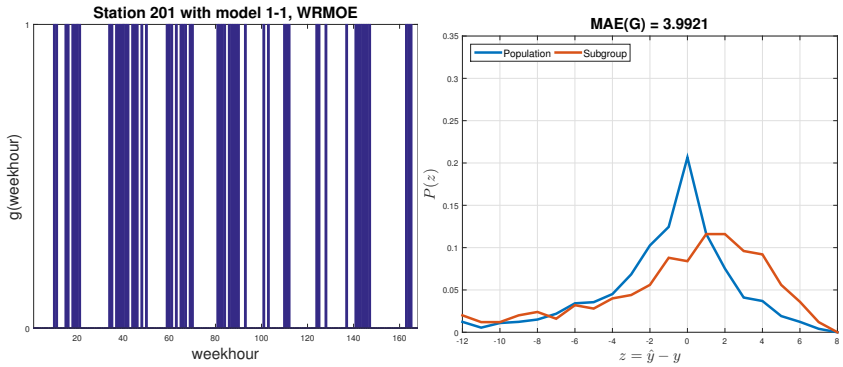

Fig. 4: The best subgroup found with the quality measure WROE (left), and the empirical distribution of errors within the subgroup (right).

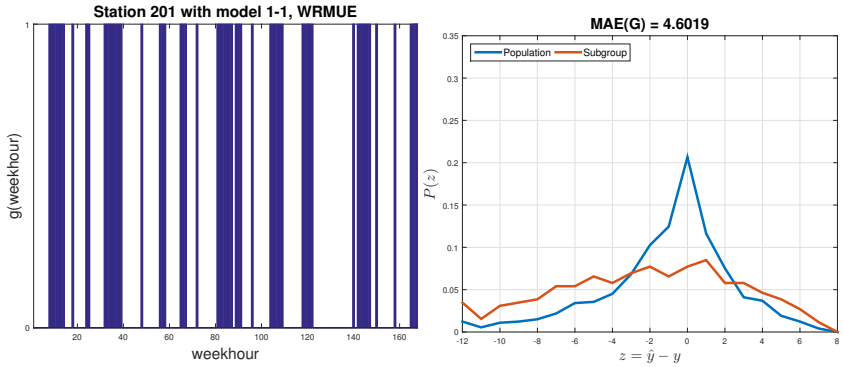

Fig. 5: The best subgroup found with the quality measure WRUE (left), and the empirical distribution of errors within the subgroup (right).

$$\hat{\mathbf{y}}_i^n = \sum_{k=1}^{n} f_{base}^j(\mathbf{x}) \tag{6}$$

Baseline method 2 can be seen as a special case of Bootstrap aggregating (bagging), as each station can be treated as a bootstrap of a mixed context (the data-set of all bike stations).

4.3 MBSD-reuse method

The proposed method is to use MBSD to find the top q subgroups (subgroups described by q attributes) for each base model in the previous method. Then a sub-model is selected according to the MAE within the subgroups:

$$f_{sub}^j = argmin_f \frac{g_j(\mathbf{d}_i) \cdot |\mathbf{y}_i - f(\mathbf{x}_i)|}{|G_j|} \tag{7}$$

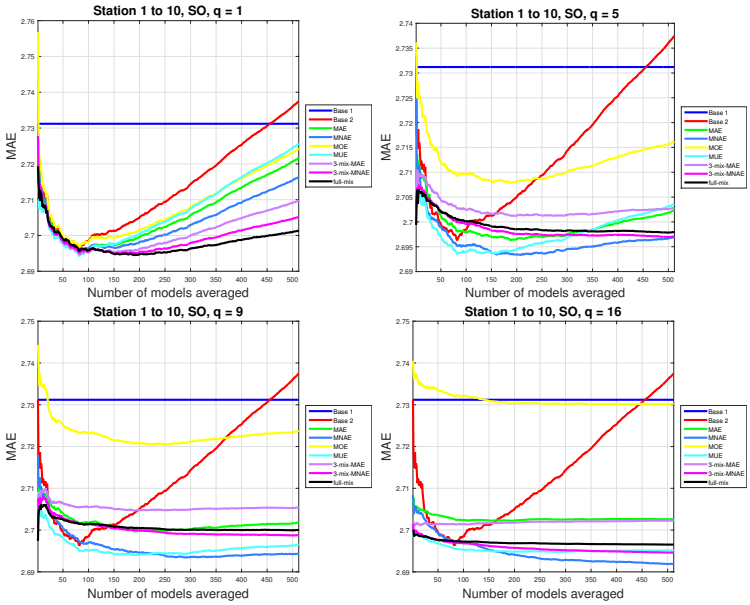

Fig. 6: The error curve for station 1 to 10 (Station-orientated), with top q subgroups adopted for each quality measure.

To combine the predictions from both the base models and sub-models, here the strategy is to use the base models for the data points that are not covered by the subgroups, and use the average of base models and sub-models for the data points within the subgroups. For the j_{th} base model, the mixture model can be given as:

$$f^j_{mix}(\mathbf{x}_i) = \frac{f^j_{base}(\mathbf{x}_i) + g_j \cdot f^j_{sub}(\mathbf{x}_i)}{1 + g_j(\mathbf{d}_i)} \tag{8}$$

For the case that there are multiple subgroups (hence multiple sub-models) for each base model (with different rank or different quality measures), the mixture model (with K different subgroups) can be given as:

$$f^j_{mix}(\mathbf{x}_i) = \frac{f^j_{base}(\mathbf{x}_i) + \sum_{k=1}^{K} g_{j,k} \cdot f^{j,k}_{sub}(\mathbf{x}_i)}{1 + \sum_{k=1}^{K} g_{j,k}(\mathbf{d}_i)} \tag{9}$$

Again, we can get the final prediction by averaging the top-n mixture models:

$$\hat{\mathbf{y}}^n_i = \sum_{j=1}^{n} f^j_{mix}(\mathbf{x}_i) \tag{10}$$

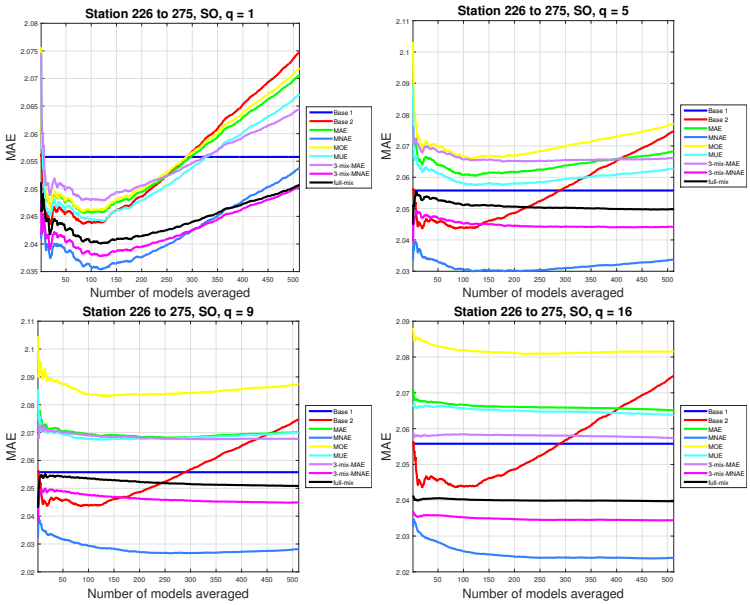

Fig. 7: The error curve for station 226 to 275 (Station-orientated), with top q subgroups adopted for each quality measure.

5 Experiments

With respect to the experiments, the training data is fixed to be the data of 275 stations during Oct 2014. For testing data, the full year data of station 1 to station 10 and the 3-month data of station 226 to station 275 will be used. In the first experiment (Station-oriented), each station is seen as a deploy context. In the second experiment (Non-station-oriented), each group of station (1 to 10, 226 to 275) is seen as a deploy context.

In both experiments, the performances will be compared among 9 methods: 1) base method 1, base method 2, MBSD-WRMAE reuse, MBSD-WRMNAE reuse, MBSD-WRMOE reuse, MBSD-WRMUE reuse, MBSD-3-mixuture reuse (WRMAE, WRMOE, WRMUE), MBSD-3-mixture reuse (WRMNAE, WRMOE, WRMUE), MBSD-4-mixture reuse. A number of up to top 16 subgroups will be used in the prediction, and up to 512 base models are selected and averaged for each deploy context.

The station-oriented error curves for station 1 to station 10 and station 226 to station 275 are given in Fig 6 and Fig 7 respectively. The non-station oriented error curves for the two groups of stations are shown in Figure 8 and Fig 9 respectively.

With respect to the station-oriented approach, it can be seen that the baseline method 2 generally beats baseline method 1. This indicates that, when the training data of the deployment context is limited, to select a bunch of trained models to get the average can potentially help reduce the prediction error. As previously discussed, in this scenario each station can be treated as a bootstrap, the baseline method is hence similar to using

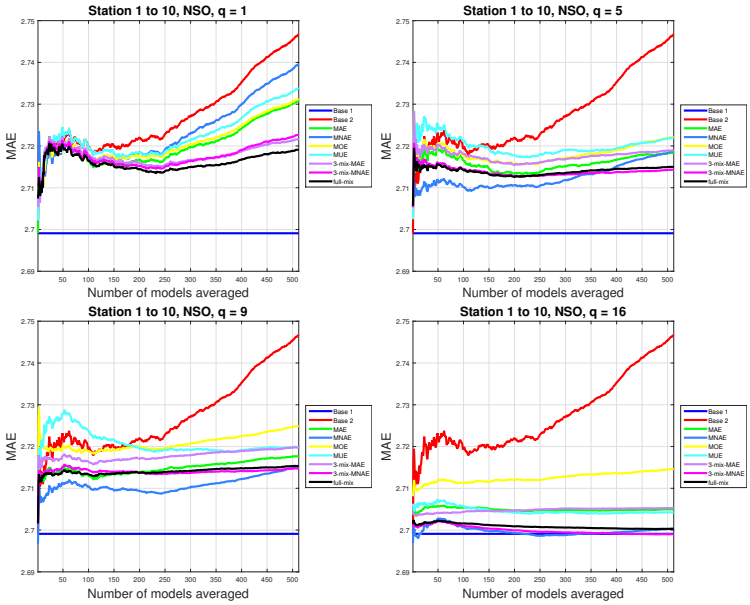

Fig. 8: The error curve for station 1 to 10 (Non-station-orientated), with top q subgroups adopted for each quality measure.

the bagging strategy. However, for this approach an important issue is to decide the number of models to be averaged, as it tends to over-fit quickly this number gets larger.

For the proposed methods, the figures show that the method MBSD-WRMNAE generally gets the best performance except in one case, where only the top 1 subgroup is used to predict the group station 1 to 10. The reason behind the good performance of MBSD-WRMNAE can be linked to the error distributions given in the previous section. As Fig 3 (right) shows, only with the quality measure WRMNAE, the error distribution is still close to a Gaussian distribution with 0 mean, but with less variance than in the population. The subgroup can hence be seen as a less noisy context, which helps the regression model to capture better parameters. On the other hand, it can be seen, especially with large q, the proposed methods tend to reduce the effect of over-fitting from baseline method 2. This is mainly because these methods are designed to fit a better model for the data points that are not well predicted by the base models. Therefore, the effect will become more significant when the number of q gets larger, as more sub-models are involved in the prediction. This makes the choice of number of averaged models less problematic.

With respect to the non-station-orientated method, the first interesting observation is that, for both groups, the MAE of baseline method is significantly lower than in the station-orientated approach. This indicates that to treat a set of stations as the deploy context can potentially help to get better performance. This also means that the attribute **station** might not be the best attribute to separate (describe) the deploy context. The second observation is that the baseline method 2 generally has a higher MAE than the

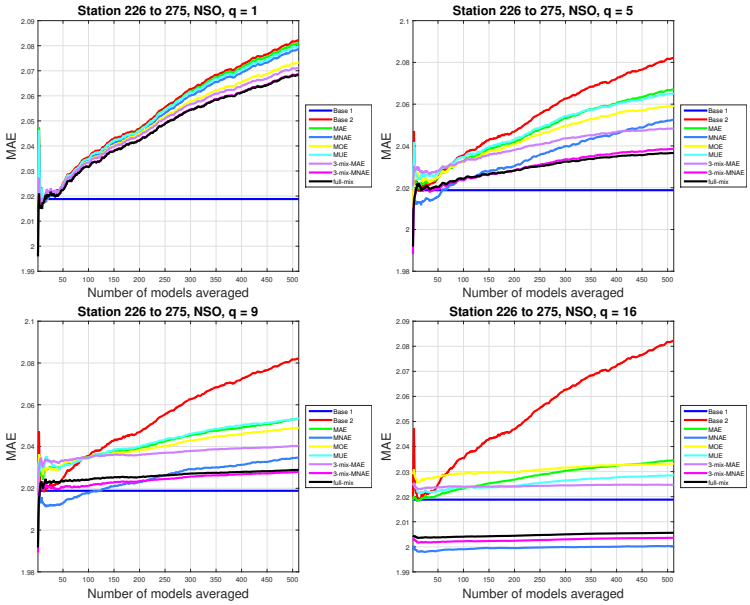

Fig. 9: The error curve for station 226 to 275 (Non-station-orientated), with top q subgroups adopted for each quality measure.

baseline method 1 in the non-station-orientated approach. One possible reason could be that, since now the training data is mixed with different stations, simply select base models according to MAE can cause a significant over-fitting and hence lower down the performance of the averaged prediction.

Since all the proposed methods are essentially based on the baseline method 2, although they generally perform better than the baseline method 2, their MAE is still higher than baseline method 1. However, with the case that $q = 16$, it can be seen both MBSD-WRMNAE and MBSD-3-mixture (WRMNAE, WRMOE, WRMUE) can still reach a MAE lower than the baseline method 1.

6 Conclusion

This paper investigates how SD can be adopted for model reuse. A variation of SD, called Model-Based Subgroup Discovery is used to detect the predictive patterns (subgroups) of the trained models in the new context. A set of sub-models are then selected for these subgroups to construct a mixture model. The experiments show that our proposed method can reduce the MAE of regression models and potentially stop the over-fitting of averaged models.

One further research direction is to develop a model ensemble algorithm with MBSD. Since in this paper some trained models are provided, a more interesting research task is hence to start from preparing the base models that can be further reused. So that the algorithm can finish the whole model reuse procedure.

References

1. Niall Twomey and Peter A Flach. Context modulation of sensor data applied to activity recognition in smart homes. In *LMCE 2014, First International Workshop on Learning over Multiple Contexts*, 2014.
2. 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.
3. Stefan Wrobel. An algorithm for multi-relational discovery of subgroups. In *Principles of Data Mining and Knowledge Discovery*, pages 78–87. Springer, 1997.
4. Nada Lavrač, Branko Kavšek, Peter Flach, and Ljupčo Todorovski. Subgroup discovery with CN2-SD. *The Journal of Machine Learning Research*, 5:153–188, 2004.
5. Martin Atzmueller and Florian Lemmerich. Fast subgroup discovery for continuous target concepts. In *Foundations of Intelligent Systems*, pages 35–44. Springer, 2009.
6. Dennis Leman, Ad Feelders, and Arno Knobbe. Exceptional model mining. In *Machine Learning and Knowledge Discovery in Databases*, pages 1–16. Springer, 2008.
7. Wouter Duivesteijn, Ad J Feelders, and Arno Knobbe. Exceptional model mining. *Data Mining and Knowledge Discovery*, pages 1–52, 2013.