Usages of Selective Regression

Amit Gross
Usages of Selective Regression

Final Paper

Submitted in partial fulfillment of the requirements for the degree of Master of Science in Computer Science

Amit Gross

Submitted to the Senate of the Technion — Israel Institute of Technology
tevet 5778 Haifa March 2018
The final paper was done under the supervision of Prof. Ran El-Yaniv in the Computer Science Department.

First and foremost, I want to thank my advisor, Prof. Ran El-Yaniv. His support and guidance were invaluable to the completion of this work. He taught me how to research and address theoretical and empirical problems, how to establish research questions, and how to keep digging for more with follow up questions. Ran was very constructive and patient through personal hard times I had during the work, and for that I will always be grateful to him. I would like to thank my wife for her support and helping me to push forward. Last but not least, I would like to thanks my children, Guy and Maya, for bringing happiness to my life. This thesis is dedicated to them.

The generous financial support of the Technion is gratefully acknowledged.
## Contents

Abstract .................................................. 1

Abbreviations and Notations ............................ 2

1 Introduction ............................................. 4

2 Problem Model ........................................... 8
  2.1 Regression Model .................................... 9
  2.2 Selective Regression Model ......................... 10

3 Ensemble of Selective Regressors ................... 11
  3.1 Clustering Selective Regression .................... 12
    3.1.1 The number of clusters ........................ 14
    3.1.2 Time complexity considerations ................. 15
    3.1.3 Experiments .................................... 16
  3.2 Selective Ensemble Based on Random Attributes Selection ... 20
    3.2.1 Experiments .................................... 21
  3.3 Conclusions ......................................... 23
List of Figures

3.1 Number of clusters: Error rate ........................................ 14
3.2 Artificial data with 25 clusters ...................................... 15
3.3 Effect of the number of attributes per regressor: Linear regressor 21
3.4 Effect of the number of attributes per regressor: Decision tree regressor ................................................................. 22
List of Tables

3.1 Clustering and no clustering selective regression, coverage=0.3  17
3.2 Clustering and no clustering selective regression, coverage=0.7  25
3.3 Clustering regression trees and random forests, coverage=0.3  26
3.4 MSE: Dimensionality reduction techniques with linear regressor  27
Abstract

Using selective regression, it is possible to increase accuracy of predictions by abstaining from answering when there is insufficient knowledge. This work is about increasing the accuracy of selective regression even further and using simple selective models to create a more complex one, by using an ensemble of selective regressors. We demonstrate how to achieve improved accuracy by using two methods to build our ensemble. In the first approach, we first split the samples in the input dataset into several clusters, and use each such cluster to train a regressor. Then, when given a new instance, we choose a regressor result that did not reject the new instance. In the second approach we train several regressors, where each regressor is using only a subset of the data’s original features. This allows us to create several lower dimensionality regressors that are less prone to overfitting, especially when the training set is fairly small. We then choose which regressor should be used by discarding those that reject an example given for labeling. We empirically tested the two approaches on various datasets, and saw that it can indeed boost accuracy compared to a single regressor or non-selective ensembles, depending on the distribution of the actual data. Finally, we present conclusions drawn from our findings and raise some follow up research questions that arise from this work.
Abbreviations and Notations
$\mathcal{X}$ — Feature space
$\mathcal{Y}$ — Output space
$f : \mathcal{X} \to \mathcal{Y}$ — Learned regression function
$m$ — Number of samples in the training set
$S_m$ — Training set
$\mathcal{H}$ — Hypothesis class
i.i.d. — Independent and identically distributed
$\theta$ — Linear regression coefficients
$\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+^+$ — Loss function
$R(f)$ — True risk of regressor $f$
$\hat{R}(f)$ — Empirical risk of regressor $f$
$\mathbf{X} \in \mathbb{R}^{d \times m}$ — Matrix notations for $S_m$
MSE — Mean squared error
$f^* : \mathcal{X} \to \mathcal{Y}$ — An unknown target function
ERM — Empirical risk minimizer
$\hat{g} : \mathcal{X} \to \{0, 1\}$ — Reject function
RC — Risk–coverage
$\Phi(g)$ — Expectation for not abstaining
$t$ — Number of sub-probability spaces (unknown)
k — Number of regressors used in the ensemble
AB — Attribute bagging
Chapter 1

Introduction
In the field of supervised machine learning, the concept of regression refers to learning how to predict real valued labels associated with feature vectors, using a model learned from a training set. Regression can be used in various ways to solve problems. Some examples where regression is used are predicting the future price of a stock, the speed and altitude of a missile we wish to intercept, our server’s workload on New Year’s Eve, or the weight a certain alloy can sustain under extreme conditions.

An ensemble of classifiers or regressors in supervised machine learning consists of multiple classifiers or regressors that are used together to solve the same task, to increase the accuracy of the solution over a single classifier [8]. The ensemble helps the learner get a better model, and overcome three main difficulties:

- Statistical: Reduce mistakes due to variance, by aggregating results of multiple models.
- Representational: When the target function is not in the hypothesis class, but an ensemble, which is a superset of the hypothesis class, is.
- Computational: When the target function is in the hypothesis class, but getting close enough to it is expensive in terms of computation required or the required training set size. An example, is a hypothesis class with many local minima.

In the approach we adopt, the learner produces an ensemble of classifiers or regressors \( h_1, ..., h_L \), and decisions are made using the weighted or unweighted votes of ensemble members.

In standard regression, given a training set drawn from some population, the learner is expected to learn a hypothesis that allows it to explain other samples drawn from the same population. In selective regression, the learner is given an option to abstain from answering for a subset of the population. This flexibility allows a significant increase in the learners’s accuracy for the samples it does tag. This solution can be used in situations where the cost of abstaining is lower than making an error, or if the decision in these cases can be propagated to more expensive layers [15]. Examples of some cases where the abstain option is applicable are deciding whether to invest in a certain stock, patient automatic diagnosis or classifying a tweet as hate speech.
In this work, we propose to combine the two techniques to improve accuracy even further. We suggest building an ensemble of selective regressors, and using the selective property to decide which of the ensemble members should be used for labeling. We introduce two approaches to building ensembles based on selective regressors. The techniques proposed in this work are general and can be used for any given selective regressor to build a model. We implemented and tested it on two selective regressors, linear regression with a disbelief index [23], and random forests, with a naive nearest neighbor as the selective property.

In the first approach for building an ensemble of selective regressors, we examine the case where our probability space favors some subset of families \( \mathcal{X} \times \mathcal{Y} \) over others. For example, in two-dimensional space, predicting the length of a halibut (a fish) is approximately linear to its weight, but the coefficients are dependent on the fish’s gender [10]. Generally, it is not rare that a relatively simple hypothesis class \( \mathcal{H} \) can fit a subset of the samples, but when covering more examples, this hypothesis class underfits.

To address these issues, we introduce Clustering Selective Regression. We enhance selective regression coverage by using several simpler regressors, each applying to some part of the population. We first use an unsupervised clustering method to split the data into smaller subsets. Then, we train a selective regressor for each cluster. When receiving an unlabeled example, we choose the appropriate regressor (if there is such), and answer accordingly. This allows us to use a fairly simple regressor to learn more sophisticated data. This resembles a similar work, where cascading was proposed: Train a selective classifier on all data, check which samples are rejected and use them to train the next classifier; continue to do so until the model has sufficient coverage [15]. The main difference between this approach and ours is when the selective property is used. While the cascading approach uses it both in learning time and for each unlabeled instance, our approach separates the data into regressors based on a different property, and does not favor using cheaper regressors first. This approach is covered in Section 3.1.

Another usage of ensembles is for reducing overfitting, especially in sparse feature space, where the data size is similar to the number of features. In such cases, dimensionality reduction is often used before learning the model. One such common tool to do this is Principal Components Analysis (PCA) [13], which helps choose the most significant principal components.
to be the new features. Another option is to create an ensemble of learners, where each learner is trained using a random selection of features from the data. The features are randomly selected with repetitions, and an ensemble of classifiers is built upon them [2]. We suggest that instead of generating regular classifiers or regressors, we generate selective regressors on each model. In our approach, when presented with an unlabeled example, use a regressor that did not abstain from answering is used. This method allows us to cope with high dimensionality data using a simple approach that can be combined with multiple other dimensionality reduction approaches, e.g., random bagging (as suggested originally), or more sophisticated techniques such as word2vec [14] and salient terms [9] for text classification. This is covered in Section 3.2.
Chapter 2

Problem Model
2.1 Regression Model

Let $\mathcal{X}$ be some feature space, and $\mathcal{Y}$ be some output space. In standard regression, the goal is to learn a regressor $\hat{f} : \mathcal{X} \rightarrow \mathcal{Y}$, using a finite training set of $m$ labeled examples $S_m = \{(x_i, y_i)\}_{i=1}^m$ that are sampled i.i.d. (independent and identically distributed) from an unknown underlying distribution $P$ over $\mathcal{X} \times \mathcal{Y}$. The regressor is to be selected from some hypothesis class $\mathcal{H}$. For example, when we consider linear regression from $d$-dimensional vectors to a scalar, we have $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}$, and $\mathcal{H} = \{f(x) = \theta^T \cdot x : \theta \in \mathbb{R}^d\}$. Let $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}^+ \setminus 0$ be a loss function\(^1\), and define the true risk of the regressor $f$ to be $R(f) = \mathbb{E}_{P(x,y)}[\ell(f(x), y)]$, and the empirical risk $\hat{R}(f) = \sum_{i=1}^m \ell(f(x_i), y_i)$.

In cases where $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}$, we will use the matrix notations $X \in \mathbb{R}^{d \times m}$ as a matrix where each row is an example and $y \in \mathbb{R}^m$ is a column vector of the corresponding labels. A common loss function for linear regression is Mean Squared Error (MSE) and is defined as: $MSE(\hat{f}) = \mathbb{E}_{(x,y) \in P}[\ell(f(x) - y)^2]$.

It is common to assume that for each $x \in \mathcal{X}$, the associated $y \in \mathcal{Y}$ value is $y = f^*(x) + \epsilon$ where $f^* : \mathcal{X} \rightarrow \mathcal{Y}$ is some unknown target function, and $\epsilon$ is a noisy normally distributed random variable with zero mean and some unknown variance. In this case, our function $\hat{f}$ is an approximation for the unknown $f^*$, and is often denoted as the Empirical Risk Minimizer (ERM).

The goal of the learner is to find $\hat{f}$ from some hypothesis class $\mathcal{H}$ such that the true risk $R(\hat{f})$ is minimized using some predefined size for the training set $m$. An alternative is to produce function $\hat{f} \in \mathcal{H}$ such that $R(\hat{f}) < \delta$ using the smallest training set possible, given a target risk $\delta \in \mathbb{R}$.

An ensemble of regressors is a set $\{\hat{f}_i \in \mathcal{H}_i\}$ where each $\hat{f}_i$ is an ERM of its training set. Some common methods to generate ensembles include manipulating the data order, injecting randomness or manipulating the features. Then, given an unlabeled example, the value of $\hat{f}(x)$ is calculated by some weighted or unweighted combination between the ensemble members.

---

\(^1\)Where $\mathbb{R}^+ \triangleq \{x : x \in \mathbb{R}, x \geq 0\}$. 

9
2.2 Selective Regression Model

The concept of selective classification was introduced [4] more than 50 years ago. Chow suggested the concept of abstaining as a safeguard against excessive misrecognition. In selective regression, the learner produces an additional function, \( \hat{g} : \mathcal{X} \to \{0, 1\} \), where \( \hat{g}(x) = 0 \) indicates abstaining from predicting the value.

\[
(f, \hat{g})(x) = \begin{cases} 
\text{UNDEFINED}, & \hat{g}(x) = 0 \\
\hat{f}(x), & \hat{g}(x) = 1.
\end{cases}
\]

It is very easy to build a regressor that is always perfectly accurate, by rejecting any unlabeled example not seen during training. We, however, are interested in a metric that combines both our risk and our coverage, which is often referred to as the risk–coverage (RC) trade off. We define our real risk according to this tradeoff:

\[
R(f, g) = \mathbb{E}_{P(x,y)} \ell(f(x), y) \cdot g(x) \Phi(g),
\]

where \( \Phi(g) = \mathbb{E}_{P(x,y)}[g(x)] \) is the expectation of not abstaining.

One approach for creating a selective regressor is based on bounding to a given input the possible values any regressor that is minimizing the risk for the training set can produce and rejecting any of these functions that exceeds our regressor beyond a given threshold [23]. In practice, for linear regression, this is done by building two models for an unlabeled example \( x \). The first uses \( S_{m,\epsilon} = S_m \cup (x, \hat{f}(x) + \epsilon) \), and the second uses \( S_{m,-\epsilon} = S_m \cup (x, \hat{f}(x) - \epsilon) \), where \( \epsilon > 0 \) is a calibrated parameter that controls the coverage. If the two generated models differ significantly—then the selective regressor abstains. This approach is referred to as disbelief index, and in this work, we will use it as our selective regressor, unless otherwise noted.
Chapter 3

Ensemble of Selective Regressors
3.1 Clustering Selective Regression

Algorithm 1 Clustering selective regression

1: procedure Learn($X$, $y$, $k$)
2:     ($idxs_1$, ..., $idxs_k$) ← k-means([X, y])
3:     for $i = 1$, ..., $k$ do
4:         $X_i$ ← $X(idx_i)$
5:         $y_i$ ← $y(idx_i)$
6:         $\hat{f}_i, \hat{g}_i$ ← ERM($X_i, y_i$)
7:     end for
8:     return (($\hat{f}_1, \hat{g}_1$), ..., ($\hat{f}_k, \hat{g}_k$))
9: end procedure

10: procedure Predict($x$, model)
11:     $\hat{Y}_x$ ← $\{ \hat{f}_i(x)|\hat{g}_i(x) = 1 : (\hat{f}_i, \hat{g}_i) \in model\}$
12:     if $\hat{Y}_x = \emptyset$ then
13:         reject
14:     else
15:         return median($\hat{Y}_x$)
16:     end if
17: end procedure

Often, the function we are trying to learn is more complex than our hypothesis class, or alternatively, we do not have enough samples to accurately find it, even if the target function is in the hypothesis class (as in neural networks, for example). One approach to boosting our learner is by using an ensemble of regressors, and combining the results of each individual regressor into one result. There are several approaches for generating ensembles, including injecting noise into the data, reordering it or taking different partitions of it. We are interested in how to boost a selective regressor, and also utilize the selective property of the ensemble members. We also want the ensemble itself to have the selective property.

Let there be $P(x, y) = \bigcup_{i=1,\ldots,t} P_i(x, y)$, for some unknown $t$, and let

$$\forall i = 1,\ldots,t : h_i = (f_i, g_i) \in \mathcal{H} : h_i = argmin_{\mathcal{H}}(R_{P_i}(h_i)),$$

where $R_{P_i} = \frac{\mathbb{E}_{P_i(x,y)}[f(f(x),y)\cdot g(x)]}{\Phi_i(g)}$ is our true risk for the $i$th regressor, and similarly, $\Phi_i(g) = \mathbb{E}_{P_i(x,y)}[g(x)]$ is its coverage. We would like to use it to
create a new model \((\hat{f}, \hat{g})\) such that \(\forall h \in \mathcal{H} : R(\hat{f}, \hat{g}) \leq R(h)\).

We would also like to cluster all training samples into \(t\) clusters and learn a selective model on each of them; however, \(t\) is unknown. Number of clusters to be used is discussed in Section 3.1.1. Henceforth, we refer to the number of clusters used as \(k^1\).

We offer two methods to decide which cluster to use. The first method applies the same clustering technique used during training, and the second method uses the selective property of each regressor.

In the first approach, and assuming clustering was done based on \(\mathcal{X}\) alone, we apply the same clustering mechanism, and find the relevant cluster. For example, if we used \(K\)-Means clustering with \(d : \mathcal{X} \rightarrow \mathbb{R}\) as the distance function, and let \(c_1, ..., c_k\) be our cluster means, then for point \(x \in \mathcal{X}\), we choose \(h_i \in \mathcal{H}\), where \(i = \arg\min\{d(x, c_j) | j = 1, ..., k\}\). Formally, the selective regressor created with this technique is:

\[
(\hat{f}, \hat{g})(x) = \begin{cases} UNDEFINED, & \hat{g}_i(x) = 0 \\ \hat{f}_i(x), & \hat{g}_i(x) = 1, \end{cases}
\]

where \((f_i, g_i) \in \mathcal{H}\) and \(i\) is the chosen cluster.

This solution has a few drawbacks. First, we cannot use \(\mathcal{Y}\) when clustering our data and generating our model. Second, it enforces usage of only one member in the ensemble per unlabeled instance, even if multiple members are valid, which could help reduce the variance error. Third, it cannot be trivially extended by addition of a new regressor without retraining on the complete training set.

To address these issues, we propose using \(\hat{g}_i\) itself to select which cluster to use. We simply invoke \((f_i, g_i)\) for all \(i = 1, ..., k\), and select the clusters that were not rejected. If multiple values were not rejected, we can choose the median value, or according to the underlying metric used for \(g_i\), for example the disbelief index, as described in [23]. Given an ensemble of regressors \(\{f_i : \mathcal{X} \rightarrow \mathcal{Y}, g_i : \mathcal{X} \rightarrow \{0, 1\} | i = 1, ..., k\}\), and an unlabeled sample \(x\), define the set of candidate values \(\hat{Y}_x = \{f_i(x) | g_i(x) = 1\}\). If the set \(\hat{Y}_x\) is empty, we will abstain from labeling \(x\); otherwise, we will label it with \(\text{median}(\hat{Y}_x)\).

Formally: let \(\{f_i : \mathcal{X} \rightarrow \mathcal{Y}, g_i : \mathcal{X} \rightarrow \{0, 1\}\}\) be a set of selective regres-

\(^1\)Ideally, \(k = t\) but since \(t\) is unknown, \(k\) is our approximation to it.
Figure 3.1: Number of clusters: Error rate

...sors, and \( x \) be an unlabeled sample. Define: \( \hat{Y}_x = \{f_i(x)|g_i(x) = 1\} \). Define the ensemble of the members \( \{f_i: \mathcal{X} \rightarrow \mathcal{Y}, g_i: \mathcal{X} \rightarrow \{0,1\}\} \) as:

\[
(f, \hat{g})(x) = \begin{cases} 
\text{UNDEFINED,} & \text{if } \hat{Y}_x = \emptyset \\
\text{median}(\hat{Y}_x), & \text{otherwise}
\end{cases}
\]

The training and labeling procedures are both described in Algorithm 1. The experimental results of this technique are described in Section 3.1.3.

Another option, with which we did not experiment, to further boost the technique, could be to use fuzzy clustering, and create each selective regressor with a weighted input, similar to the method described for the non-selective ensembles in other works [22].

3.1.1 The number of clusters

In practice, the proposed solution enriches the hypothesis space, and creates a new one \( \mathcal{H}' = \{(h_1, ..., h_k)|h_i \in \mathcal{H}\} \). Note that \( \forall h \in \mathcal{H}, (h, h, ..., h) \in \mathcal{H}' \), which implies that the enhanced hypothesis class is at least as diverse as the original one. This means that the suggested solution is more prone to overfitting, and indeed, the case when \( k = m \) will give a perfect match for all training samples, but will have little value for unknown examples. This makes the selection of the number of clusters important to the quality of the solution. In order to check this hypothesis, we ran an experiment on an
artificial dataset as follows

\[ y = \text{floor}(x)^2 \times x + e, \]

where \( x \in \mathbb{R} \) and \( e \) is 0 mean random variables: \( e \sim \mathcal{N}(0, \sigma) \).

By setting \( x \) to various ranges, we created a dataset where each cluster on its own is linear—yet a simple linear regression model will underfit. Each dataset has a different number of clusters. We ran the proposed method with varying values of \( k \). The error rate on this dataset is shown in Figure 3.1, and the learned functions are shown in Figure 3.2. As we expected, the results show a tendency to overfit when using too large \( k \), but can significantly improve the results when the selection of \( k \) is appropriate for the dataset, and when the data diverges more from the original hypothesis class of the linear selective regression.

Some approaches to choosing the number of clusters automatically include using cross-validation or penalizing models with high numbers of clusters, as an extra regularization factor.

### 3.1.2 Time complexity considerations

The proposed solution introduces \( O(mnk) \) extra time to run K-Means clustering (using Lloyd’s algorithm [12], for example). It, however, then reduces
the time consumed by the learner for each learned component. Assuming the learner's run time is superlinear in the number of examples, which is a common case, and is the case with linear regression, this could benefit the overall runtime. Accordingly, the overall runtime is:

\[ O\left(\sum_{i=1}^{k} T(m_i, n) + mnk\right), \]

where \( T : \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{N} \) is the time complexity of the regressor, and \( m_i \) is the number of samples in cluster \( i \).

If we take \( T(m, n) = n^2 \cdot m \), which is the complexity of the linear regression\(^2\), and add an assumption of uniform distribution over the clusters, we get

\[ O\left(\sum_{i=1}^{k} \left(\frac{n}{k}\right)^2 \cdot m + mnk\right) = O\left(\frac{n^2 \cdot m + mnk^2}{k}\right). \]

Assuming \( n > k \), this gives us a time complexity of \( O\left(\frac{n^2 \cdot m}{k}\right) \), which is better than regular regression by a factor of \( k \).

In addition to the above, the processing of each regressor can be done in parallel, after the clusters are calculated. This allows for an even faster learning time, by a factor of \( k \), for the generation of the regressors.

This faster learning time, however, does not come for free. The technique introduces the extra cost of determining which regressor to use during labeling of each example by the model. When we choose which regressor is used based on the clustering alone, that can be found easily with a \textit{KD-Tree} \[^1\], for example, in \( O(\log(k) \cdot n) \) time. If the regressor is chosen, however, using the selective property, each regressor must be rerun again. For linear regression, this means \( O(n \cdot k) \) extra time.

### 3.1.3 Experiments

We tested the approach using the selective property to determine which ensemble member should be used. For this purpose, we ran tests both on artificial data, and on real data. The artificial data is the same as described in Section 3.1.1, where the generated data is:

\[^2\text{Using close formula solution}\]
<table>
<thead>
<tr>
<th>Dataset Name</th>
<th>MSE clustering</th>
<th>STD clustering</th>
<th>MSE bootstrapping</th>
<th>STD bootstrapping</th>
<th>MSE base</th>
<th>STD base</th>
</tr>
</thead>
<tbody>
<tr>
<td>space-ga-scale</td>
<td>0.0204</td>
<td>0.0013</td>
<td>0.0231</td>
<td>0.0015</td>
<td>0.0225</td>
<td>0.0014</td>
</tr>
<tr>
<td>abalone-scale</td>
<td>3.4054</td>
<td>0.2492</td>
<td>4.0654</td>
<td>0.3115</td>
<td>4.0626</td>
<td>0.3119</td>
</tr>
<tr>
<td>cpusmall-scalea</td>
<td>385.9946</td>
<td>43.44</td>
<td>53.0048</td>
<td>1.1686</td>
<td>54.027</td>
<td>1.1818</td>
</tr>
<tr>
<td>housing-scale</td>
<td>8.5008</td>
<td>1.2863</td>
<td>11.1995</td>
<td>1.4457</td>
<td>11.6711</td>
<td>1.4378</td>
</tr>
<tr>
<td>bank-32-un</td>
<td>0.0063</td>
<td>0.0003</td>
<td>0.0062</td>
<td>0.0003</td>
<td>0.0062</td>
<td>0.0003</td>
</tr>
<tr>
<td>winequalitydata</td>
<td>0.3525</td>
<td>0.0287</td>
<td>0.3516</td>
<td>0.029</td>
<td>0.3632</td>
<td>0.0295</td>
</tr>
<tr>
<td>winequalitye.data</td>
<td>0.4838</td>
<td>0.0241</td>
<td>0.5035</td>
<td>0.0251</td>
<td>0.5057</td>
<td>0.025</td>
</tr>
<tr>
<td>parkinsons-total-updrs</td>
<td>97.0007</td>
<td>3.3851</td>
<td>113.1116</td>
<td>3.8497</td>
<td>111.9387</td>
<td>3.8071</td>
</tr>
<tr>
<td>ships-compressor-decay</td>
<td>1.28E-05</td>
<td>9.49E-07</td>
<td>3.65E-05</td>
<td>1.14E-06</td>
<td>3.78E-05</td>
<td>1.15E-06</td>
</tr>
<tr>
<td>forestfires</td>
<td>2.7985</td>
<td>0.4171</td>
<td>2.2446</td>
<td>0.3453</td>
<td>2.2542</td>
<td>0.3422</td>
</tr>
<tr>
<td>Concrete-streangh</td>
<td>45.3077</td>
<td>4.8775</td>
<td>69.2142</td>
<td>5.562</td>
<td>68.4639</td>
<td>5.5334</td>
</tr>
<tr>
<td>Power-plant-production</td>
<td>21.0107</td>
<td>0.641</td>
<td>22.9282</td>
<td>1.022</td>
<td>22.9295</td>
<td>1.0213</td>
</tr>
<tr>
<td>SkillCraft</td>
<td>0.9789</td>
<td>0.0486</td>
<td>1.1214</td>
<td>0.0548</td>
<td>1.1691</td>
<td>0.0557</td>
</tr>
</tbody>
</table>

Table 3.1: Clustering and no clustering selective regression, coverage=0.3
\[ y = \text{floor}(x)^2 \times x + e, \]

where \( x \in \mathbb{R} \) and \( e \) is the 0 mean random variable: \( e \sim N(0, \sigma) \) We used \( x \in [0, 24) \), which generated 24 linear clusters of data, yet the data as a whole was not linear. Then, we ran the clustering solution (with median value selection in case there was more than one matching cluster), with a desired coverage ratio of 0.7. As can be seen in Figure 3.2, the linear regressor underfits, since it generates a model using a single linear line. When we introduce clustering and train several regressors, however, each regressor is more accurate as far as its cluster, which is indeed linear. Since each regressor is trained on a different subset of the training set, the ensemble will not choose regressors that tend to abstain from labeling instances farther away from their cluster. This gives us a much more accurate regressor that resembles the true target function.

In addition, we used selective regression clustering on several real datasets:

- The abalone-scale [11], which is an 8-dimensional dataset where the goal is to predict the age of an abalone.
- The concrete strength dataset [24], which an is 8-dimensional dataset where the goal is to predict the concrete’s compressive strength.
- The wine quality (red and white) datasets [6], which is an 11-dimensional dataset, where the goal is to predict the quality of the wine. “Red” and “white” wine are two separate datasets.
- The skillcraft [18] dataset, which is an 18-dimensional dataset, where the aim is to predict the number of “complex actions” made in a Starcraft 2 video game, per timestamp.
- The bank-32-nm dataset, which is one of the bank family datasets, provided by delve\(^3\), and is used, for example, for assessing the active regression algorithm ALICE [17].
- The Physicochemical Properties of Protein Tertiary Structure dataset, which is a 9-dimensional dataset, where the goal is to predict the size of the residue. The dataset is provided by [11].

\(^3\)http://www.cs.toronto.edu/ delve/data/
• The space, cpusmall and housing datasets, which are taken from the libsvm regression datasets [3].

• The Parkinson UPDRS dataset [19], which is a task where the goal is to predict the Parkinson’s disease symptom score on the UPDRS scale.

• The forestfires, which is a regression task where the aim is to predict the areas burned by forest fires, in the northeast region of Portugal [7].

• The Ships compressor decay dataset, which is a regression task where we aim to predict the compressor decay state coefficient [5].

• The powerplant production dataset is collected from a full load plant, and consists of four features. The goal is to predict the hourly production output [20].

For each dataset, we compared selective regression with and without clustering, and added another baseline when using bootstrapping selective regressors, where we trained several selective regressors, each on random sampling with replacements of the data. We experimented with two coverages: 0.3 and 0.7. The results are shown in Table 3.1 and Table 3.2. The tables show that the method improves baselines in 7/13 datasets and reduces quality in 2/13 datasets (the results of the rest are inconclusive). This shows us that there is tendency for the method to improve the results, but it is not definitive across all datasets. The reasonable explanation for this is similar to what was proposed in the synthetic data experiments, which is usage of this method will improve accuracy should the data itself does not fit the hypothesis class, but subsets of it do.

In addition, we tested the method on random forest regressors, using the same datasets. For this test, we used the simple Nearest Neighbor reject function. Each selective regressor produces a label if there is some sufficiently close (euclidean space) neighbor to the given example in the training set. The results are shown in Table 3.3. The experiment results show that in four datasets, the suggested method improved the learner, while in one dataset it did not help it at all (for the rest, the results are inconclusive).
3.2 Selective Ensemble Based on Random Attributes Selection

Algorithm 2 Attributes bagging using selective regression

1: procedure Learn($X$, $y$, $k$)
2: for $i = 1, ..., k$ do
3: $S_i$ ← Sample(1, 2, ..., $n$)
4: $X_i$ ← Project($X$, $S$)
5: $\hat{f}_i, \hat{g}_i$ ← ERM($X_i, y$)
6: end for
7: return $((S_1, \hat{f}_1, \hat{g}_1), ..., (\hat{f}_k, \hat{g}_k))$
8: end procedure

9: procedure Predict($x$, model)
10: $\hat{Y}_x$ ← $\{\hat{f}_i(\text{Project}(x, S)) | \hat{g}_i(\text{Project}(x, S)) = 1 : (S_i, \hat{f}_i, \hat{g}_i) \in \text{model}\}$
11: if $\hat{Y}_x = \emptyset$ then
12: reject
13: else
14: return median($\hat{Y}_x$)
15: end if
16: end procedure

The approach of using an ensemble based on reduced dimensionality has been suggested before [21], [2]. In this approach, one uses an ensemble of classifiers or regressors where each member of the ensemble is trained on a variant of the training set with reduced dimensionality. This approach has three main advantages: it (1) reduces the “curse of dimensionality” impact, which causes overfitting on sparse data. (2) reduces the correlation between ensemble members. (3) improves the accuracy of the ensemble.

Our approach is similar to attribute bagging (AB) [2], where the learner samples features randomly with repetitions, and trains a classifier or regressor for each subset selected. While in attribute bagging, the ensemble is chosen from among all candidates during training by checking each one’s performance using cross-validation, we suggest using the selective property of each ensemble member to decide which of its members should be used. The suggested technique is to train various ensemble selective members, and for each of them, use a different subset of features during training and label-
Figure 3.3: Effect of the number of attributes per regressor: Linear regressor

ing. When given an unlabeled example, consider only the ensemble members that did not abstain. This is described in Algorithm 2. This approach gives us several advantages over the AB alternative. First, the ensemble itself has the selective property and can abstain when all its members abstain (or there are not enough nonabstaining regressors). Second, this allows a member that is generally not as good as others, but is quite accurate on a small subset of the population, to be used. Third, it simplifies the learning procedure. Fourth, this technique allows very easy extension of the ensemble without the need to rerun cross-validation or other selection methods.

3.2.1 Experiments

For this approach, we are mostly interested in how the method performs when the size of the training set is close to the number of features. For this, we used a 280-dimensional regression task called Blog Feedback [16]. This dataset consists of several features, including a 200 bag-of-words of the most popular words, the weekday, the length of the text, and some time series data regarding number of comments. The task is to predict the number of comments in the next 24 hours. We are interested in the effect of the size of the training set on the various regressors, and the number of attributes used for each regressor. For this purpose, we conducted two experiments
First, we were interested in the effect of the training set size on the results. We provided training sets of varying sizes and observed how the alternatives perform. The results are given in Table 3.4. As we can see, when the training set is very small, especially when it is smaller than the number of features, the selective regressors ensemble has an advantage in reducing the dimensionality in order to get more accurate results. A likely explanation for this phenomenon is that each ensemble member tends to overfit when the number of attributes is very close to the number of training samples; however, each new sample could have a large impact on the generated model, if it is very different from the already-seen samples. This is exactly what the selective regressors do. While each ensemble member is likely to overfit, it is also just as likely to reject examples that do not fit its model. When aggregating several selective regressors, each providing its on answer when it is in its “comfort zone”, the results are enhanced.

In the second experiment, we used a training set of 500 samples, and checked how the number of regressors affects the results of each method. The results can be seen in Figure 3.3. Clearly, both methods perform better when the number of features is kept relatively small. The selective regressor approach, however, is significantly less vulnerable to the amount of features,
and results are only mildly worse than the optimal choice. Similar to what
it was before, a likely explanation is because when the number of features
gets close to the number of training samples, each regressor overfits. The
selective property helps us ignore some regressors, and use only those that
are best for each test sample.

We repeated the experiment using decision tree regressors instead of
linear regression. The results are shown in Figure 3.4. While here the
optimal number of features is higher than when using linear regression, the
proposed method does show improvement over the baseline. In addition, the
stability of the proposed method, using selective regressors over a random
ensemble, can be seen in the results.

3.3 Conclusions

This work introduces an approach to using a selective regressor as a building
block in an ensemble. The ensemble can be built by combining the regressors
in various ways, and using the selective property to determine which member
of the ensemble should be preferred. From this work, we can see that one
can aggregate several simple learners into a complex one by using a selective
property. This approach lets us use a large number of regressors, where
each specializes in a relatively narrow subset of the probability space to
get better results with higher coverage. We see that using clustering as
a preprocessing to create several regressors can be beneficial as it widens
the hypothesis class compared to a single learner, but with it comes the
danger of overfitting when too many clusters are used. We have also seen
that an ensemble of regressors can produce a model with lower variance
compared to one based on non-selective regressors, especially when the train
set is relatively small. We also see that the approach can be utilized when
the feature space is large compared to the number of available examples,
by random sampling the attributes into several regressors, and letting the
selective property decide which regressor should be used per sample.

Some follow-up questions that arise from this work are:

- In our experiments, our ensemble members all had the same cover-
age. Will the quality be improved if we somehow adjust the coverage?
  Specifically, does it make sense to give a regressor that was trained
  on a “wider” probability space (in terms of distance between training
samples) a higher coverage than one that was trained on a very narrow probability space?

• How to optimally choose the number of clusters when building ensemble members based on clustering? We used cross-validation to choose, but can it be done better?

• One can use fuzzy clustering to split the training set, and build the regressors where each sample has a weight proportional to its probability to belong to each cluster. Will this approach improve the results?

• How can unlabeled data help to boost the performance of ensembles made of selective regressors even further?

• Will an ensemble that contains different hypothesis classes yield better results than an ensemble containing only one hypothesis class?
<table>
<thead>
<tr>
<th>Dataset Name</th>
<th>MSE clustering</th>
<th>STD clustering</th>
<th>MSE bootstrapping</th>
<th>STD bootstrapping</th>
<th>MSE base</th>
<th>STD base</th>
</tr>
</thead>
<tbody>
<tr>
<td>space-ga-scale</td>
<td><strong>0.0265</strong></td>
<td>0.0011</td>
<td>0.0305</td>
<td>0.0013</td>
<td>0.0306</td>
<td>0.0013</td>
</tr>
<tr>
<td>abalone-scale</td>
<td>4.0577</td>
<td>0.1981</td>
<td>4.0944</td>
<td>0.1994</td>
<td>4.0601</td>
<td>0.1993</td>
</tr>
<tr>
<td>cpusmall-scale</td>
<td><strong>13.1334</strong></td>
<td>0.3715</td>
<td>43.4125</td>
<td>0.7727</td>
<td>43.4746</td>
<td>0.7751</td>
</tr>
<tr>
<td>bankm.data</td>
<td>0.0067</td>
<td>0.0002</td>
<td>0.0068</td>
<td>0.0002</td>
<td>0.0068</td>
<td>0.0002</td>
</tr>
<tr>
<td>winequality-white</td>
<td>0.3626</td>
<td>0.0194</td>
<td>0.3759</td>
<td>0.02</td>
<td>0.3799</td>
<td>0.0201</td>
</tr>
<tr>
<td>winequalitye-red</td>
<td>0.5022</td>
<td>0.0168</td>
<td>0.5284</td>
<td>0.0179</td>
<td>0.529</td>
<td>0.0179</td>
</tr>
<tr>
<td>parkinsons-total-updrs</td>
<td><strong>111.808</strong></td>
<td>2.492</td>
<td>118.249</td>
<td>2.7925</td>
<td>118.2232</td>
<td>2.7932</td>
</tr>
<tr>
<td>ships-compressor-decay</td>
<td><strong>6.63E-06</strong></td>
<td>1.98E-07</td>
<td>2.90E-05</td>
<td>6.04E-07</td>
<td>MISSING</td>
<td>6.06E-07</td>
</tr>
<tr>
<td>forestfires</td>
<td>2.1201</td>
<td>0.2066</td>
<td>2.1759</td>
<td>0.2108</td>
<td>2.0361</td>
<td>0.1855</td>
</tr>
<tr>
<td>Concrete-streangh</td>
<td><strong>71.5399</strong></td>
<td>5.3175</td>
<td>90.0966</td>
<td>5.5592</td>
<td>92.8821</td>
<td>5.6613</td>
</tr>
<tr>
<td>Power-plant-production</td>
<td><strong>21.6586</strong></td>
<td>0.557</td>
<td>24.4929</td>
<td>0.5975</td>
<td>24.5414</td>
<td>0.5987</td>
</tr>
<tr>
<td>SkillCraft</td>
<td><strong>1.0077</strong></td>
<td>0.032</td>
<td>1.0914</td>
<td>0.0344</td>
<td>1.1704</td>
<td>0.0369</td>
</tr>
</tbody>
</table>

Table 3.2: Clustering and no clustering selective regression, coverage=0.7
<table>
<thead>
<tr>
<th>Dataset Name</th>
<th>MSE clustering STD clustering MSE Forest STD Forest MSE tree std tree</th>
</tr>
</thead>
<tbody>
<tr>
<td>space-ga-scale</td>
<td>0.0187 0.0024</td>
</tr>
<tr>
<td>abalone-scale</td>
<td>3.6923 0.4609 3.9274 0.4973</td>
</tr>
<tr>
<td>cpusmall-scale</td>
<td>7.0326 0.8627 7.9217 0.9351</td>
</tr>
<tr>
<td>housing-scale</td>
<td>5.8340 1.2983 7.2813 2.0095</td>
</tr>
<tr>
<td>bankm.data</td>
<td>0.0088 0.0010 0.0082 0.0009</td>
</tr>
<tr>
<td>winequality-white</td>
<td>0.1889 0.0567 0.3333 0.0921</td>
</tr>
<tr>
<td>winequality-red</td>
<td>0.0691 0.0198 0.1091 0.0289</td>
</tr>
<tr>
<td>parkinsons-total-updrs</td>
<td>130.8205 12.0533 160.1150 13.8606</td>
</tr>
<tr>
<td>ships-compressor-decay</td>
<td>9.84E-07 4.06E-08 1.06E-06 4.98E-08</td>
</tr>
<tr>
<td>forestfires</td>
<td>3.6831 1.0865 3.3450 1.0387</td>
</tr>
<tr>
<td>Concrete-strength</td>
<td>48.5804 17.9772 41.2153 16.3944</td>
</tr>
<tr>
<td>Power-plant-production</td>
<td>15.8519 1.3453 15.0959 1.3615</td>
</tr>
<tr>
<td>SkillCraft</td>
<td>1.4005 0.1364 1.7540 0.1756</td>
</tr>
<tr>
<td>ships-compressor-decay</td>
<td>9.84E-07 4.06E-08 1.06E-06 4.98E-08</td>
</tr>
<tr>
<td>parkinsons-total-updrs</td>
<td>130.8205 12.0533 160.1150 13.8606</td>
</tr>
<tr>
<td>ships-compressor-decay</td>
<td>9.84E-07 4.06E-08 1.06E-06 4.98E-08</td>
</tr>
</tbody>
</table>

Table 3.3: Clustering regression trees and random forests, coverage=0.3
<table>
<thead>
<tr>
<th>Train size</th>
<th>Regular regression</th>
<th>PCA dimensionality reduction</th>
<th>AB ensemble</th>
<th>Selective regressors</th>
<th>PValue</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>7763</td>
<td>48613</td>
<td>9079</td>
<td>6160</td>
<td>0.0002</td>
</tr>
<tr>
<td>200</td>
<td>92412</td>
<td>5855</td>
<td>3067</td>
<td>2734</td>
<td>0.0022</td>
</tr>
<tr>
<td>300</td>
<td>13522</td>
<td>9703</td>
<td>4363</td>
<td>3332</td>
<td>0.0102</td>
</tr>
<tr>
<td>400</td>
<td>8425</td>
<td>4568</td>
<td>3457</td>
<td>3024</td>
<td>0.0412</td>
</tr>
<tr>
<td>500</td>
<td>4901</td>
<td>3082</td>
<td>2328</td>
<td>1937</td>
<td>0.055</td>
</tr>
<tr>
<td>600</td>
<td>4038</td>
<td>3768</td>
<td>2724</td>
<td>2642</td>
<td>0.0024</td>
</tr>
<tr>
<td>700</td>
<td>3217</td>
<td>3230</td>
<td>2804</td>
<td>2905</td>
<td>0.0471</td>
</tr>
<tr>
<td>800</td>
<td>3453</td>
<td>3177</td>
<td>2637</td>
<td>2704</td>
<td>0.4224</td>
</tr>
<tr>
<td>900</td>
<td>3403</td>
<td>2955</td>
<td>2257</td>
<td>2254</td>
<td>0.0189</td>
</tr>
<tr>
<td>1000</td>
<td>3479</td>
<td>3166</td>
<td>2658</td>
<td>2599</td>
<td>0.7976</td>
</tr>
</tbody>
</table>

Table 3.4: MSE: Dimensionality reduction techniques with linear regressor

*Attributes Bagging compared to Selective regression ensemble
Bibliography


על דוגמאות לא מוכחות. הטכניקה המוצאת בשימוש במודלי סלקטיביים הצגתנית יוצרת משאר האלגוריתמים והיבוא של יונים טובים.

לבסש, אין סכום של שבבים זה מראה התכונה של פאונה מצטיינת של סלקטיביים הזריזים הרכבות המופקפות תכניות טובות יוצרות תור ישימש בריכבוניות ובתכונות הפוסטות שלتمثل. הוא יותר שלמות. גם מעולים מפסר של התוכן לה rekl רון Bueno זה לא התמוקדה ולא החריט ו_CURRCTION נchuremos סיני לlicence הגרות מעניין מוספנות.
ביאות, אשר עליהן בישמהו בחרתца לשון A-2 הסינון על עליון בוברキュט מוד. העברות
מתמשכות הבדイラ על-إجراء, התנות.ALIGNاتهمouver בריכות ימייפיצים שלהן.
כותרת את זה, נא מקווים כמו,ximo את הלאבכנר בケット ת İzית טיטי ajust pelix יברユニיט
והם כן, שכר בולט בין S-20A-14-2018

טכנית הגאולוגיה, занו הגאולוגיה של אצל, מיריצי סמר של, יאש, שאר כל
מודל ומקים על תמייניו בחרים. הגאולוגיה של אחרים סירוב השם, בלבד ה
זכות את זה, שהגאולוגיה של אדום קיימאכסตก שולח, אך

לייצר הספר שיסבול "ולא哈利 מחות של".
וכם גם, בין הפרספקטי של אקולוקס יאש, והם כן, קיימאכסתק שולח, אחר

ההכרים סטרב או. כאשר היא להכרבה שבנו, הוא ביצוע סלקטיביות בכותרת قول
יפור בשירות, יácilון להכרבב בחוש במדרג מהורדה, מימש את הלאנוגרמה
באמצאות אגי הלחתן ומקרבס ילארו, בבגרה ואת החוקרת סكنيسة לעתונות מוסף
בדוי, וכילון ישימוע אקראי. יאש בשpaniesיה, והניק לשפר את אוכז obra, שאר כל
המשטחות, וכל באופן. בפרט, ככרו המימד קורסמס לקורס מסקור hend מה dieta
והגאולוגיה של אגי, והניק לעתונות מסקר, ביןימה הדגש, הוא א新西ון לברבע
והגאולוגיה של אגי,DEL סמייצי הќ"וניה, בצורת עשרים, אカラー
差异化 בין המיקרים וירש, אקזון במאבטעת מודים, ישכר המרכז השיתות
אינו מקרב בעינו מסקר את פונקציית המטרה, סכניקיה ולכות עלת לברבע
ולא הסיכום بصورة מתאימה.

טקニック השיתות, занו הגאולוגיה של אחר, מתחפשים בכ-2 camov ממקימאכסתק שולח, ערכו
באות אוזן הגאולוגיה, שאר הספּו"ב הצות. ערכו לכל
אומדן כי מלא, וממידין את נתוני ההגאולוגיה של אוזן
vvm של מהשקהดน, בחרת את הפק鳡ים של המפלגה של אלכסון, וחזור את הגאולוגיה של מ
דרח המודלים של שולח. יחיד את A שאר, יאש אっこ_שימוע产品研发, שלל

המכנים שיתות, אשר גאולוגיה של אחרים, מיריצי סמר של, יאש, שאר כל
מודל ומקים על תמייניו בחרים. הגאולוגיה של אחרים סירוב השם, أكدron
זכות את זה, שהגאולוגיה של אדום קיימאכסתק שולח, אחר

לייצר הספר שיסבול "ולא哈利 מחות של".
תקציר

בתוך הבינה המלאכותית,нные הנה פעולות שללים של או רמר פליט של שיוויו בoined לולו, שפ פتلكצבה
הלהמאת הזק ערך דייך. ירבע, הוסיף שמעש פסקיץ-מוקים ומסונים וא
רגרסור, על מנט כתב חיאל תחיתון (סוייב וא ררסה). למלש, ירבע אוכסות
הנה עפיי אקראיים - הב מספי למידה מימיית עין התלבה ממקבך דוגמאות
השיכום לאותת התפלורה, ירח הפלאות אקוים מוסיים, וכל חד מחלדמיס יער עין
החלצה, הבתא לזרוגאא שתחפכה על יידי. בירבע עינ עפעשי על工业大学.
בכמנה רשתות הלילה - על ידי הפלאות פוקיצים וה участник עניפי. בירבע לחב
przedsiębior התושב ממקוני אקוים יבור להפכת התחיתigmat הא בראש האקרית (המשטח המבנה
משטח), ואת הרצועה חונכעה מתכנית - ביא איז פוקיצים המדרים את נמצאת mvbro.
האפוריזיון של הלמד בים הפוקיצים מנסת, א מצלאת העיר מסרפק רב של
דוגמאות לגאל לקימיה.

ברגרסיה סלקטיבית, הלמד, בנסס על הפוקיצים המלאכות, מימייר בפונקטיז בחרה.

פונקציה זו הלמדה פוקיצי סרプリ, ומאנפרטרים למול "לתור" "לגר" לכלום אק.
שסר עם מודי ההנאה האצת פוקיצים הפוקיצהי המלאכות, סสยาม ליהיוו רושא
מורע קרוא. זא Wohn ציור. בירבע זה misdemeanor על לייזר "לול" "לי"" שמסים מחר
�ית בחר החצבות בוק מימיי איז מסרפק, הלפס מודי מדרי חיסת הבת הקצבות
ב美國 מימד מסרפק. ברגרסיה סלקטיבי, וא מפענפה תממוד שנקרא "זוכי-箴יכ",
בוק מימד מסרפק. ברגרסיה סלקטיבי, וא מפענפה תממוד שנקרא "זוכי-箴יכ",
שק כל מחזאו ליהיוו ליהיוו מדרי על מחרב קון מואר בים. לון, מדר "זוכי-箴יכ".
מאפרשר של מרדיליצייה של הסוכן לשלושה ביבס להזרה, או מסמי למזרה. בנסוס
שליש, או נסף שמצותנה הנסות של כ"המס"לת, אוכסות, "לגר" כ"סוכי" מופתת
בומבליי שומימ.
ה commerc תוצרת הדפסה בלתי ידועה של פרו. אך לא, מכונת לעמדת המכשבל breadcrumbs.

תודה

בארוש ובראשונה, ברצוני להודות לענשה של, פרו', ואר-ייב. תומיית וחדותינו היא אל גם שמאפשרת את השלמת חיבור זה. אנו למדנו את ייחודה, אך לaneously נאנסו תואריים. אנו מתים, כדי לשלש לפרסום, או כי לשנייתו לפתח שאלות מספונות. ויהי תום, וסבלנו מים בamework קשיח מביתו יאני. על כ- באמצעות או, אני רצון להודות לאשתי על תמיכתו וערתו הלمدن ינשוך. אחותו הבכורה, איני רצון להודות

ליידיי, לאו מאיה, שחכמים אויר לי. עבדו זה ומקדש החל.

אני מודה לעכבני על החמימה הנגדית במשם השיתופיות.
היווים של חיזוי סלקטיבי

忽悠 על עבודה גמר

לשם مليיה חלקי של הדירישה ל מגיע התואר

מניסיון להרשים בנימין המ면서

עמי בורו

הנהלת הטכניון – מרכז טכנולוגי לישראל

אדר הנשען חיפע מרת 2018
שיטות של חיזוי סלקטיבי

ברגרסייה

עמית גורס