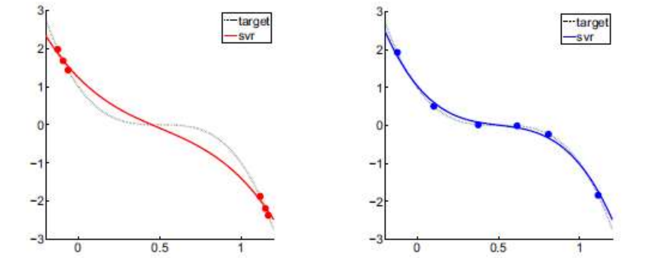
Paper review: Passive Sampling for Regression

Problem statement



(a) SVR with skewed sample (b) SVR with unskewed sample

Figure 1. Example of SVR functions trained from six skewed sample and unskewed sample.

The paper discusses active sampling, which selects informative samples from unlabeled data to maximize the accuracy of the learned function. However, active sampling can be inefficient for regression, as it requires learning and validating regression functions at each iteration. Instead, the paper proposes passive sampling, which identifies informative samples based on their geometric characteristics in the feature space, rather than the learned function. Passive sampling is more efficient and effective for regression than active sampling, which suffers from performance fluctuations due to selecting samples with high regression errors, which are often noisy. Extensive experiments show that passive sampling outperforms even the omniscient active sampling, which knows the labels of the unlabeled data.

Motivation

1. K. Brinker, "Active learning of label ranking functions", Proc. Int. Conf. Machine Learning (ICML'04), 2004.
2. H. Yu, "SVM selective sampling for ranking with application to data retrieval", Proc. Int. Conf. Knowledge Discovery and Data Mining (KDD'05), 2005.

These two referenced articles show us that active sampling for regression is difficult because it is challenging to identify the uncertainties of data points based on the function learned in the previous iteration.

To address this, the paper proposes passive sampling techniques for regression, which select samples based on their geometric characteristics in the feature space, rather than their regression errors. Passive sampling is more efficient, accurate, and stable than active sampling, particularly in real-world data sets, where active sampling suffers from serious performance fluctuations.

key ideas, techniques, and contributions

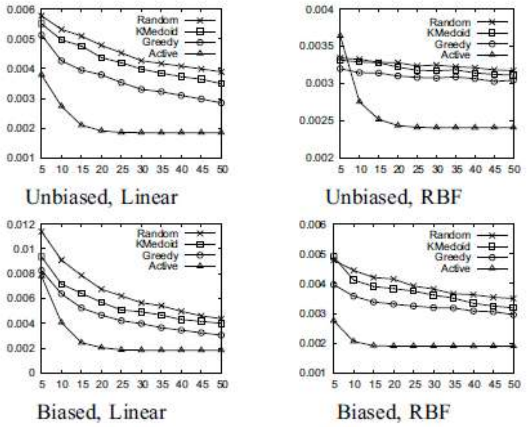
Selective sampling, Passive sampling, Active learning, Active sampling, Regression

Selective sampling, also known as active learning or active sampling, is a method used in classification and rank learning to select the most informative samples from a pool of unlabeled data. It minimizes the cost of labeling while maintaining the quality of the learned function. Active sampling is a variant of selective sampling that selects the most ambiguous or uncertain sample for classification or ranking based on the function learned at the previous iteration. Passive sampling is a new technique proposed for regression that selects the most informative samples based on their geometric characteristics in the feature space rather than their regression errors. It is more efficient, accurate, and stable than active sampling, especially for real-world data sets. Regression is a supervised learning problem that estimates the target value of an input object based on a set of input features.

Experimental evaluation

Background

In this section, the authors evaluate four different sampling techniques for regression: (1) greedy passive sampling, (2) incremental k-medoids passive sampling, (3) random sampling, and (4) omniscient active sampling. The evaluation is based on the mean absolute error (MAE) between the target outputs and the regression outputs on the testing set. The training set is used to select samples for training the regression model. The authors used the LIBSVM library for support vector regression (SVR)



Unbiased, Linear Unbiased, RBF

Biased, Linear Biased, RBF

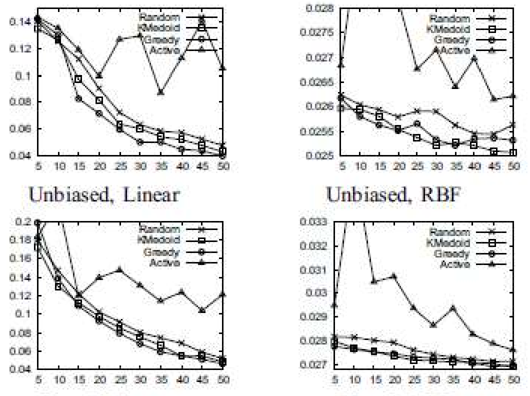
Figure 2. MAE on *noiseless* data. X-axis: # of instances; Y-axis: MAE; "Unbiased": unbiased labeled set; "Biased": biased labeled set; "Linear": linear kernel; "RBF": RBF kernel.

linear function: $f^*(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$

RBF function: $f^*(\mathbf{x}) = exp\left(-\frac{||\mathbf{x}-\mathbf{w}||^2}{2\sigma^2}\right)$

implementation. The omniscient active sampling method assumes the labels of unlabeled data are known, which is not realistic in practice.

Synthetic Data



Unbiased, Linear

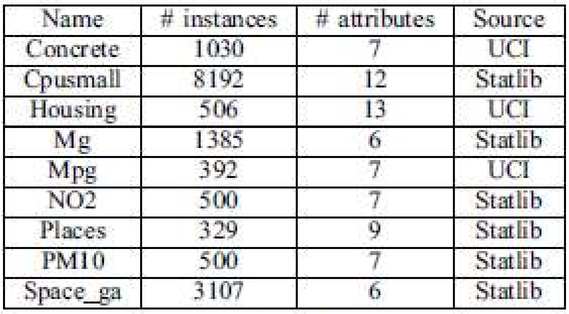
Unbiased, RBF

Biased, Linear

Biased, RBF

Figure 3. MAE on *noisy* data. X-axis: # of instances; Y-axis: MAE; "Unbiased": unbiased labeled set; "Biased": biased labeled set; "Linear": linear kernel; "RBF": RBF kernel.

The authors generated two sets of synthetic data, one noiseless and one noisy, with randomly generated vectors of 10 dimensions. They divided each set into training and testing sets and used the mean absolute error (MAE) as an evaluation metric. They then sampled two sets of 50 labeled points, one randomly and one with a bias, and experimented with four sampling methods: random, active, greedy, and k-medoids.

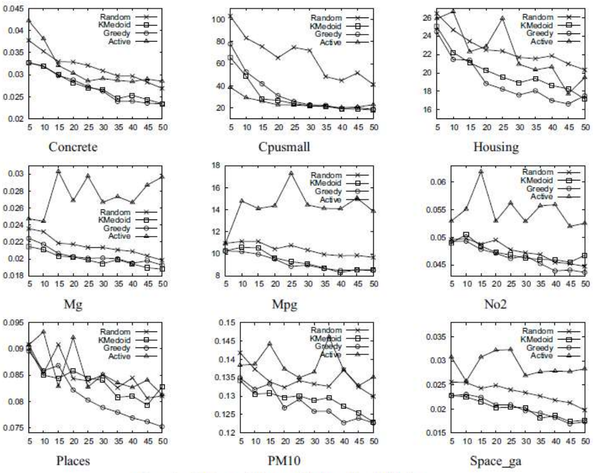


Name	# instances	# attributes	Source
Concrete	1030	7	UCI
Cpusmall	8192	12	Statlib
Housing	506	13	UCI
Mg	1385	6	Statlib
Mpg	392	7	UCI
NO2	500	7	Statlib
Places	329	9	Statlib
PM10	500	7	Statlib
Space_ga	3107	6	Statlib

Table I
DATA SETS

Real data

According to Figure 4 and 5, the authors proposed a novel passive sampling method called incremental k-medoids sampling (**KMedoid**) and compared its performance with other sampling methods such as Greedy, Random, and Active on both synthetic and real datasets. They used mean absolute error (MAE) as the evaluation metric and used SVR with linear and RBF kernels for the regression tasks. The results showed that KMedoid and Greedy methods outperformed Random and Active methods in most cases. KMedoid was found to perform particularly well in noisy datasets. In terms of objective scores, KMedoid and Greedy methods generated higher scores than Active and Random. In terms of sampling time, KMedoid was found to be faster



Concrete

Cpusmall

Housing

Mg

Mpg

No2

Places

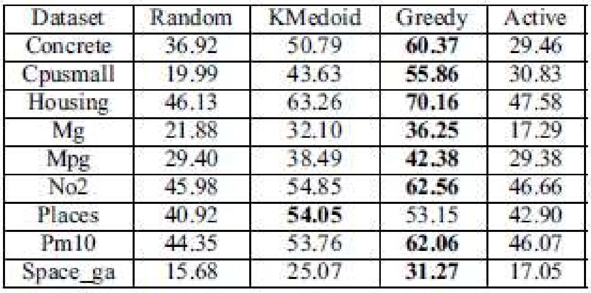
PM10

Space_ga

Figure 4. MAE on real data sets with the unbiased labeled set.

than Active but slower than Random. Overall, the results suggested that the proposed KMedoid method could be a useful and efficient sampling method for passive learning tasks.

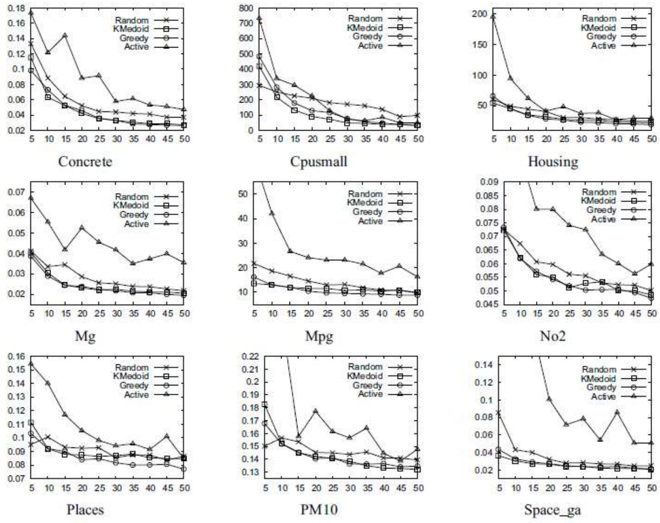
Conclusion



Dataset	Random	KMedoid	Greedy	Active
Concrete	36.92	50.79	**60.37**	29.46
Cpusmall	19.99	43.63	**55.86**	30.83
Housing	46.13	63.26	**70.16**	47.58
Mg	21.88	32.10	**36.25**	17.29
Mpg	29.40	38.49	**42.38**	29.38
No2	45.98	54.85	**62.56**	46.66
Places	40.92	**54.05**	53.15	42.90
Pm10	44.35	53.76	**62.06**	46.07
Space_ga	15.68	25.07	**31.27**	17.05

Table II
THE OBJECTIVE SCORES IN EQ. (2) OF EACH SAMPLING METHOD AFTER SAMPLING 50 DATA POINTS

Conduct experiments on both synthetic and real datasets, comparing the results with active and random sampling methods. The experiments show that passive sampling, particularly the Greedy and KMedoid techniques, outperform active and random sampling methods in terms of accuracy and objective scores. While active sampling is effective when the labeled set is biased, it struggles when the set is unbiased and noisy. Finally, the authors note that passive sampling is faster than active sampling,



Concrete

Cpusmall

Housing

Mg

Mpg

No2

Places

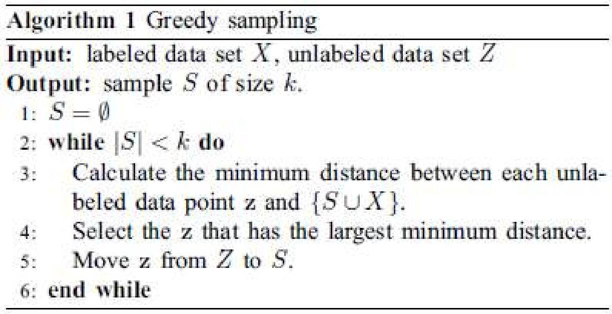
PM10

Space_ga

Figure 5. MAE on real data sets with the biased labeled set.

making it a more efficient approach to regression problems.

Presentation perspective



```
Algorithm 1 Greedy sampling
Input: labeled data set $X$, unlabeled data set $Z$
Output: sample $S$ of size $k$.
$S = \emptyset$
while $|S| < k$ do
    Calculate the minimum distance between each unlabeled data point z and $\{S \cup X\}$.
    Select the z that has the largest minimum distance.
    Move z from $Z$ to $S$.
end while
```

The authors generated noiseless and noisy datasets for two target regression functions and created two sets of 50 labeled points each: one randomly sampled and one biased. They then used four sampling methods (Random, Active, Greedy, and KMedoid) to select five additional points at each iteration, learned regression functions using SVR with linear and RBF kernels, and evaluated their performance on a testing set. The results, averaged over 30 runs, showed that on noiseless data, Active performed the best, while on noisy data, the passive sampling methods (Greedy and KMedoid) outperformed the others, with Active suffering from accuracy fluctuations due to focusing on noisy samples. The experiments on real datasets \showed similar results.

Valuation based on my own criteria

strong points:

1. Proposes a novel approach to sampling called "passive sampling" for regression that is more efficient and effective than traditional "active sampling" methods.
2. Passive sampling selects samples based on geometric characteristics in the feature space rather than the learned function, making it more efficient than active sampling methods that require learning and validating the regression function and evaluating the unlabeled data using the function.
3. The paper presents experimental results that show that the proposed passive sampling method performs even better than the “omniscient” active sampling that knows the labels of unlabeled data, making it a promising approach for regression problems.

weak points:

1. The paper does not provide a detailed explanation of the passive sampling heuristics used to optimize the objective, making it difficult for readers to fully understand and replicate the approach.
2. The paper mainly focuses on the comparison between passive and active sampling methods, but does not provide a comparison with other sampling techniques, which may also be effective for regression problems.
3. The paper only presents experimental results for a limited number of datasets, which may not be representative of all regression problems, thus limiting the generalizability of the proposed approach.

Extended Discussion

If you are the authors, what will you plan to do to address the weaknesses or to improve/extend any part of the solution?

1. Investigate the performance of the proposed passive sampling techniques on a wider range of regression tasks and datasets to verify its generalization capability.
2. Develop more advanced passive sampling techniques that can select samples with more diverse and representative characteristics to improve the performance and generalization capability of the learned regression models.
3. Explore the possibility of combining passive and active sampling techniques to leverage the strengths of both approaches and achieve better performance with fewer labeled data.

If you are to continue working on this paper, what is the next topic you would like to pursue to enrich this research?

One topic I would like to pursue to enrich this research is the application of passive sampling techniques in other types of regression models, such as neural networks or decision trees. This would involve investigating how the geometric characteristics of the feature space can be used to select informative samples for these models and how the performance of passive sampling compares to active sampling for these models.

What other new research topics, or applications, will this proposed solution impact?

To further improve the performance and generalization capability of the learned regression models, I suggest investigating the proposed technique on a wider range of regression tasks and datasets, and developing more advanced sampling techniques that can select more diverse and representative samples.