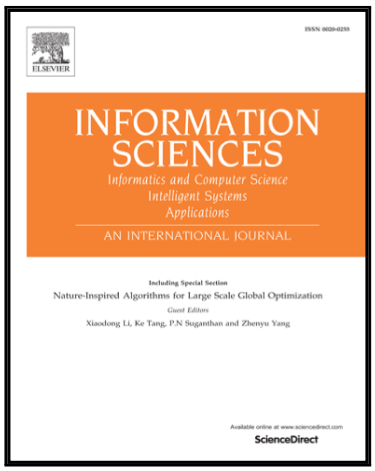

Active Learning for Regression Using Greedy Sampling

Dongrui Wu, Chin-Teng Lin, Jian Huang

PII: S0020-0255(18)30768-0
DOI: https://doi.org/10.1016/j.ins.2018.09.060
Reference: INS 13971

To appear in: *Information Sciences*

Received date: 26 June 2018
Revised date: 22 September 2018
Accepted date: 26 September 2018

Please cite this article as: Dongrui Wu, Chin-Teng Lin, Jian Huang, Active Learning for Regression Using Greedy Sampling, *Information Sciences* (2018), doi: https://doi.org/10.1016/j.ins.2018.09.060

Active Learning for Regression Using Greedy Sampling

Dongrui Wu[a,*], Chin-Teng Lin[b], Jian Huang[a,*]

[a]*Key Laboratory of the Ministry of Education for Image Processing and Intelligent Control, School of Automation, Huazhong University of Science and Technology, Wuhan 430074, China*
[b]*Faculty of Engineering and Information Technology, University of Technology, Sydney, Australia*

Abstract

Regression problems are pervasive in real-world applications. Generally a substantial amount of labeled samples are needed to build a regression model with good generalization ability. However, many times it is relatively easy to collect a large number of unlabeled samples, but time-consuming or expensive to label them. Active learning for regression (ALR) is a methodology to reduce the number of labeled samples, by selecting the most beneficial ones to label, instead of random selection. This paper proposes two new ALR approaches based on greedy sampling (GS). The first approach (GSy) selects new samples to increase the diversity in the output space, and the second (iGS) selects new samples to increase the diversity in both input and output spaces. Extensive experiments on 10 UCI and CMU StatLib datasets from various domains, and on 15 subjects on EEG-based driver drowsiness estimation, verified their effectiveness and robustness.

Keywords: Active learning, regression, greedy sampling, driver drowsiness estimation

1. Introduction

Regression, which estimates the value of a dependent variable (output) from one or more independent variables (predictors, features, inputs), is a common problem in machine learning. To build an accurate regression model, one needs to have some labeled training samples, whose dependent and independent variable values are both known. Generally the more the labeled training samples are, the better the regression performance is. However, in real-world many times it is relatively easy to obtain the values of the independent variables, but time-consuming or expensive to label them. For example, in speech emotion estimation [30, 31] in the 3-dimensional space of valance, arousal and dominance [15], it is easy to record a large number of utterances, but time-consuming to evaluate their emotions [12, 2]. Another example is driver drowsiness estimation from physiological signals such as the electroencephalogram (EEG)

*Corresponding author
Email addresses: drwu@hust.edu.cn (Dongrui Wu), Chin-Teng.Lin@uts.edu.au (Chin-Teng Lin), huang_jan@mail.hust.edu.cn (Jian Huang)