

**Ultra High-Dimensional Nonlinear Feature Selection for Big Biological Data**

**Abstract:**

Machine learning methods are used to discover complex nonlinear relationships in biological and medical data. However, sophisticated learning models are computationally unfeasible for data with millions of features. Here we introduce the first feature selection method for nonlinear learning problems that can scale up to large, ultra-high dimensional biological data. More specifically, we scale up the novel Hilbert-Schmidt Independence Criterion Lasso (HSIC Lasso) to handle millions of features with tens of thousand samples. The proposed method is guaranteed to find an optimal subset of maximally predictive features with minimal redundancy, yielding higher predictive power and improved interpretability. Its effectiveness is demonstrated through applications to classify phenotypes based on module expression in human prostate cancer patients and to detect enzymes among protein structures. We achieve high accuracy with as few as 20 out of one million features — a dimensionality reduction of 99.998%. Our algorithm can be implemented on commodity cloud computing platforms. The dramatic reduction of features may lead to the ubiquitous deployment of sophisticated prediction models in mobile health care applications.

**Existing System:**

The scale and complexity of big biological data pose new challenges to existing machine learning algorithms. There is a trade-off between scalability and complexity: linear methods scale better to large data, but cannot model complex patterns. Nonlinear models can handle complex relationships in the data but are not scalable to the size of current datasets. In particular, learning nonlinear models requires a number of observations that grows exponentially with the number of features. Biological data generated by modern technology has as many as millions of features, making the learning of nonlinear models unfeasible with existing techniques. To make matters worse, current nonlinear approaches cannot take advantage of distributed computing platforms.

**Proposed System:**

We propose a novel feature selection framework for big biological data that makes it possible for the first time to identify very few relevant, non-redundant features among millions. The proposed method is based on two components:

Least Angle Regression (LARS), an efficient feature selection method [8], and the Hilbert-Schmidt Independence Criterion (HSIC), which enables the selection of features that are non-linearly related. These properties are combined to obtain a method that can exploit nonlinear feature dependencies efficiently, and furthermore enables distributed implementation on commodity cloud computing platforms. We name our algorithm Least Angle Nonlinear Distributed (LAND) feature selection. Experiments demonstrate that the proposed method can reduce the number of features in real world biological datasets from one million to tens or hundreds, while preserving or increasing prediction accuracy for biological applications.