CSE 546 Final Project Proposal
Feature Selection and Extraction for Noisy Data

Yuqing Ai, Jingchen Hu

Project idea

Feature selection and extraction are important preprocessing steps for machine learning tasks. These steps help reduce overfitting and enhance generalization by filtering / reconstructing the features to exclude the effects of irrelevant variables. In the meantime, they also reduce the feature space and make large scale training more efficient. In this project, we plan to study a variety of feature selection and extraction algorithms. We will implement them and evaluate their performances on datasets.

Data set

We will start with the Dexter dataset (http://archive.ics.uci.edu/ml/datasets/Dexter) which contains a lot of noisy data and irrelevent features added to make training harder. The features are bag-of-word counts, which has a significantly larger dimension than the number of training samples. The test accuracy of the binary classification task provides a measure of how well the models generalize. After running experiments on the Dexter dataset we would like to switch to larger datasets with more samples and feature dimensions if time permits. We will focus on the the algorithm's efficiency as well as accuracy in this part.

Software Implementations

Specifically we plan to implement several of the following algorithms:

- Filter approaches based on metrics like correlations and information gain.
- RELIEF and RELIEFF algorithms.
- Dimension reduction with PCA.
- Wrapper approaches: greedy hill climbing, genetic algorithm, simulated annealing.
- Learning algorithms that perform feature selection as part of the operation:
  - Autoencoder neural network
  - Decision tree
  - Classifiers with L1 regularization (Lasso, hw1)

We will evaluate the performances of the first few algorithms using the same actual classifier (e.g. using logistic regression). We plan to implement most of the above without using libraries. For more complex models such as autoencoder, we might consider using frameworks like TensorFlow.

Papers to read

- Review and Evaluation of Feature Selection Algorithms in Synthetic Problems
- Evaluating feature selection methods for learning in data mining applications
- The Feature Selection Problem: Traditional Methods and a New Algorithm

Milestone

We plan to finish implementing two or three of the algorithms listed above and evaluate their performances on the Dexter dataset.