Twenty Newsgroups Text Classification

Dimensionality Reduction

Lu Zhao, Yan Liang

Stony Brook University

Machine Learning Project
May 07, 2015

Motivation

A major characteristic of text classification is the high dimensionality of the feature space.

(i) Feature space consists of hundreds of thousands terms;

(ii) Redundant features add more noise than signal, which might hurt the prediction accuracy.

In this project, we explore effectiveness of feature selection using greedy, simulated annealing algorithms and latent feature extraction using Principle Component Analysis.

Feature Selection

General selection strategies:

One step: calculate the scores independently and select the features that have top scores;

Iterative:

(i) Choose single feature that
   (1) Greedy: has highest score;
   (2) Simulated Annealing: the selected feature might not have highest score with certain probability

(ii) Rescore features conditioned on the selected features;

(iii) Repeat.

Simulated Annealing Process

Simulated annealing algorithm for feature selection:

1. Start with an arbitrary feature and calculate the conditional score
2. Repeat N times:
   - (i) Randomly choose a feature and calculate the conditional score;
   - (ii) If it has higher score than the previously accepted score, accept this new feature;
   - (iii) Else accept this feature with probability:

$$\exp(-\frac{\Delta}{K \times T});$$

     where K is Boltzmann constant, T is current temperature and Δ is the difference between score of current feature and previously accepted features;
   - (iv) Cool the system with cooling rate r;
3. return the last accepted feature

Feature Selection

Score of feature X:

(i) Document frequency: number of documents that feature X appears at least once;

(ii) Conditional mutual information of given Z:

$$I(X;Y|Z) = \sum_{x,y,z} p_{X,Y,Z}(x,y,z) \log \frac{p_Z(z)p_{X,Y,Z}(x,y,z)}{p_{X,Z}(x,z)p_{Y,Z}(y,z)};$$

(iii) Conditional χ^2 statistic:

$$\chi^2_{X,Y|Z} = \sum_z \sum_x \sum_y \frac{(O_{x,y|z} - \frac{O_{x|z}O_{y|z}}{N})^2}{\frac{O_{x|z}O_{y|z}}{N}},$$

where $O_{x,y|z}$ is the frequency of $\{X = x, Y = y\}$ given $\{Z = z\}$.

Data Set

Summary of date set:

(i) Number of categories: 20;
(ii) Feature space: Bag of words (only English word is considered);
(iii) Number of features: 129,797;
(iv) Number of documents in training set: 11,314;
(v) Number of documents in test set: 7,532.

Data Set

Training classifiers:

(a) 10-fold cross validation;

(b) knn: candidates of k are 2,3,4,6,10,15,20,30,50

(c) Naive Bayes: prior count are 0.01,0,05,1,2,5

(d) SVM: penalty candidates are 1,10,100,1000,10000

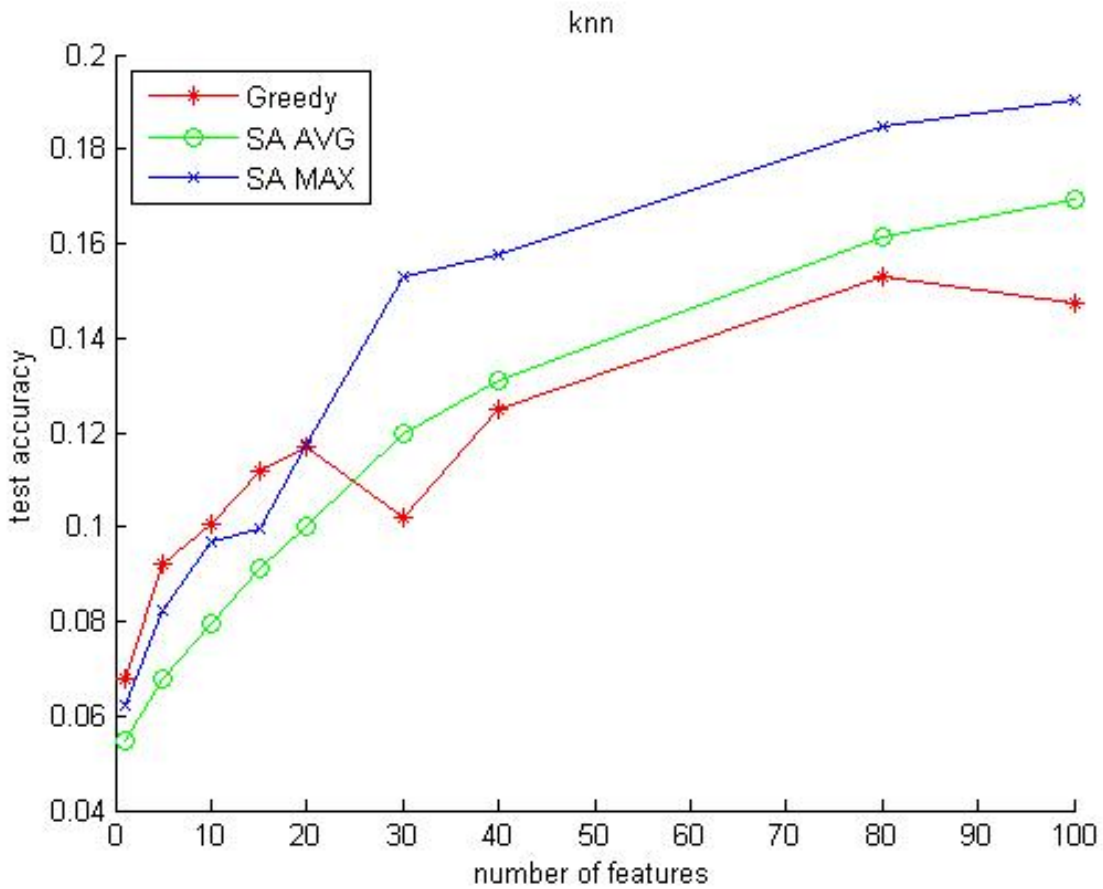
Classification on entire feature space:

Table : Text Classification Test Accuracy

	k-nearest neighbor	Naive Bayes	SVM
Accuracy	0.41($k = 2$)	0.76($\alpha = 0.01$)	0.77($C = 1$)

Comparison between Greedy and Simulated Annealing

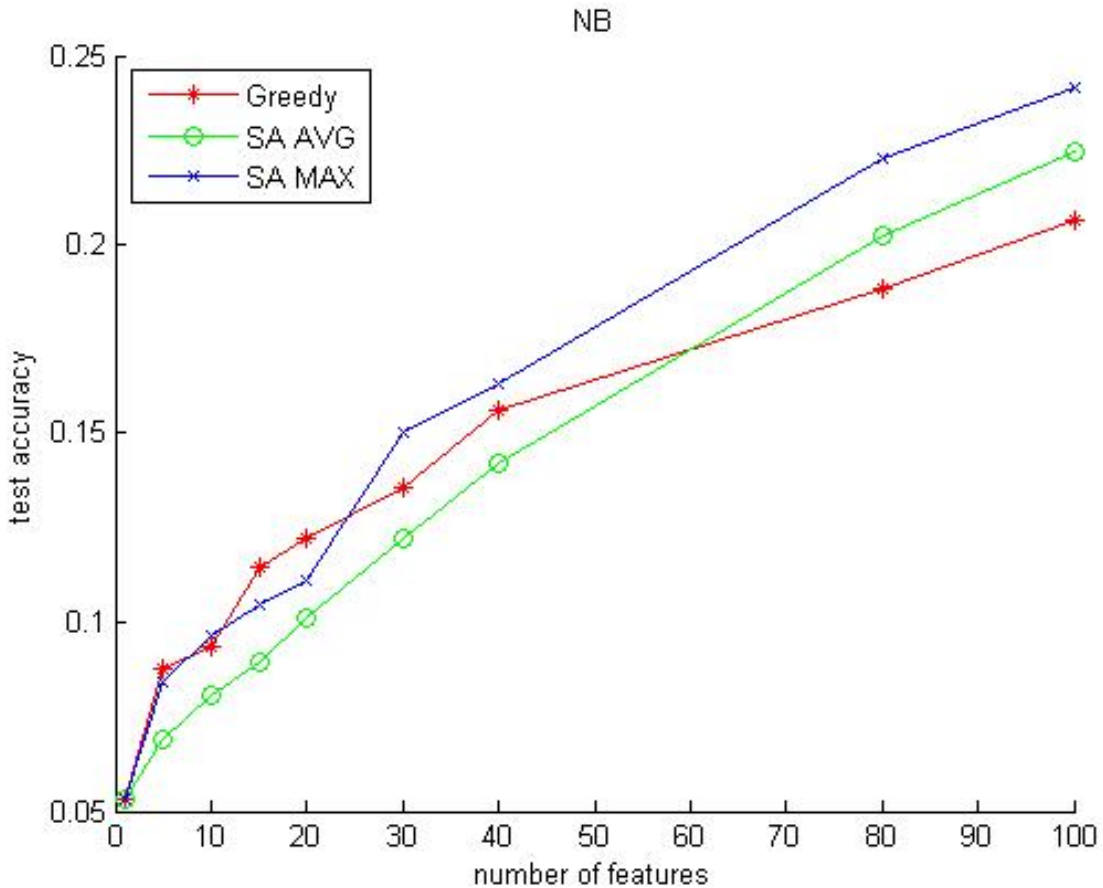
Scoring method: document frequency.
Classifier: k-nearest neighbor.



Simulated annealing algorithm on average outperforms greedy feature selection.

Comparison between Greedy and Simulated Annealing

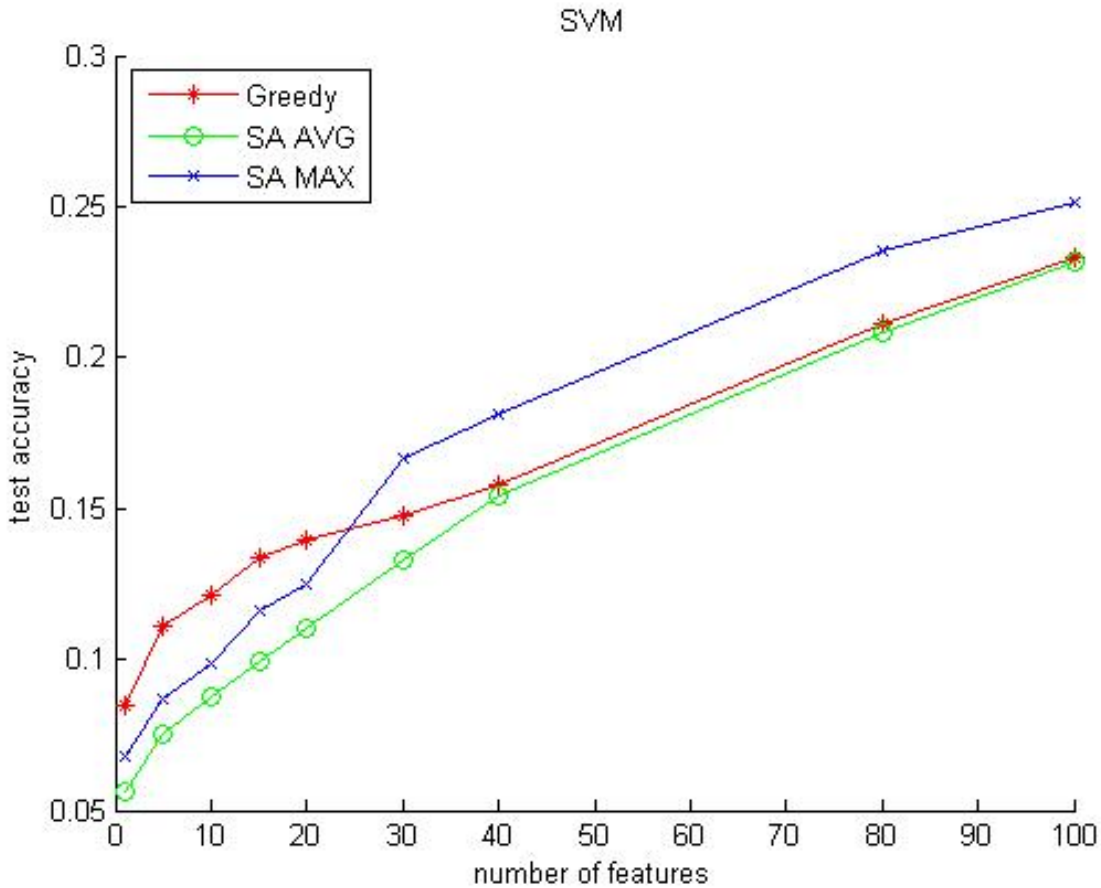
Scoring method: document frequency.
Classifier: Naive Bayes.



Simulated annealing algorithm on average outperforms greedy feature selection.

Comparison between Greedy and Simulated Annealing

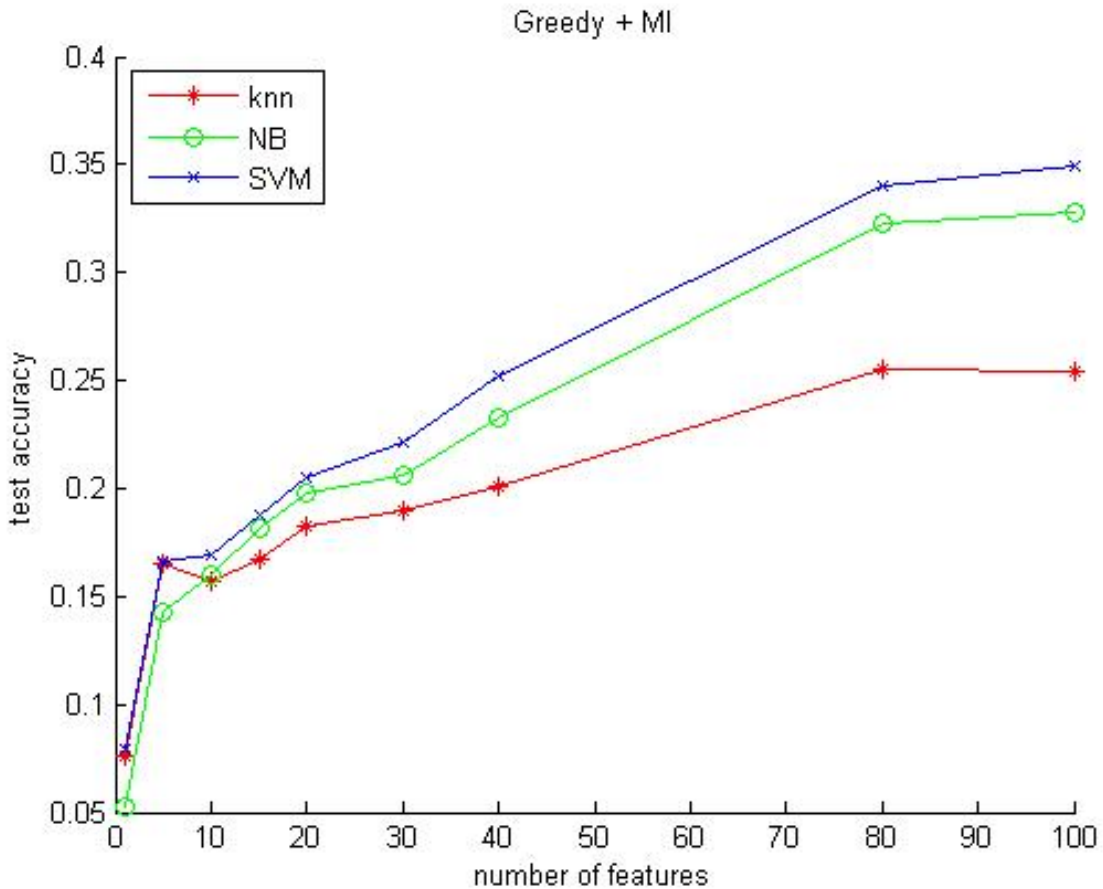
Scoring method: document frequency.
Classifier: SVM.



Simulated annealing algorithm on average performs similar to greedy feature selection.

Feature Selection

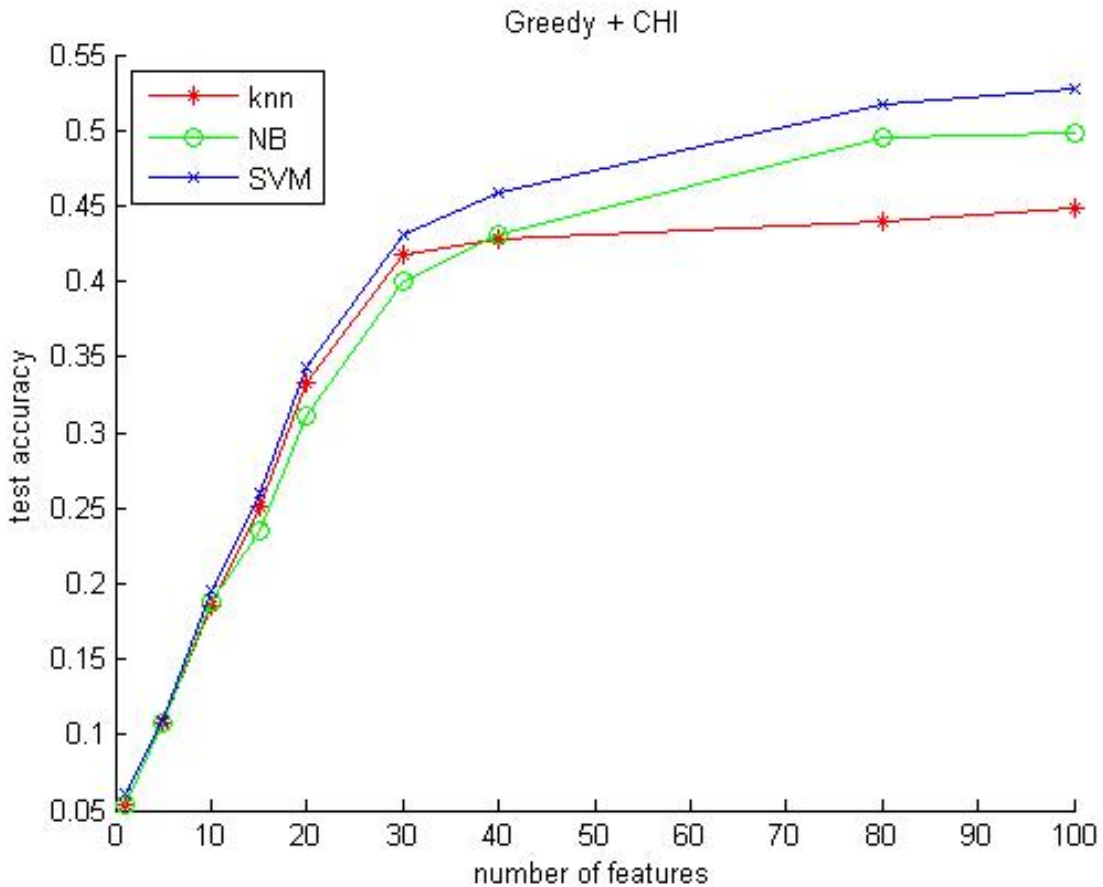
Scoring method: Mutual Information.
Selection strategy: Greedy.



0.1% features achieves 50% test accuracy using entire feature space.

Feature Selection

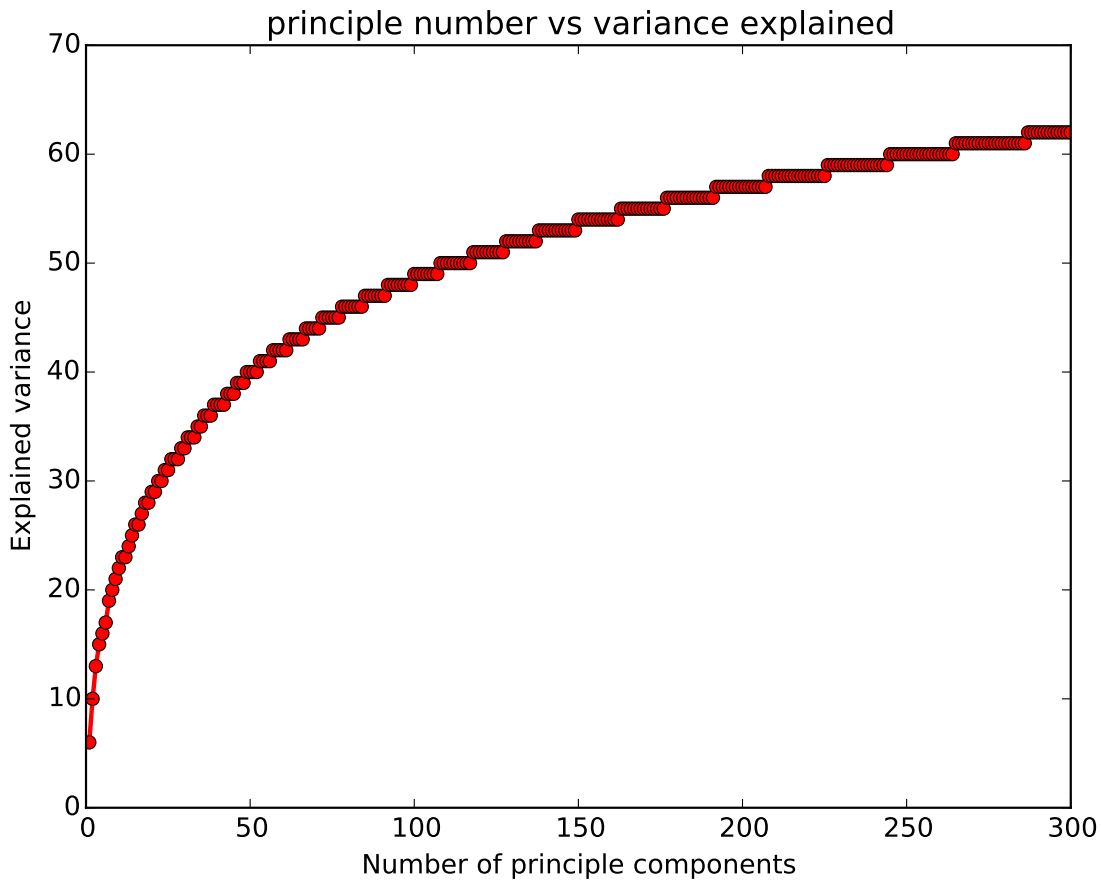
Scoring method: χ^2 statistic.
Selection strategy: Greedy.



0.1% features achieves 70% test accuracy using entire feature space.

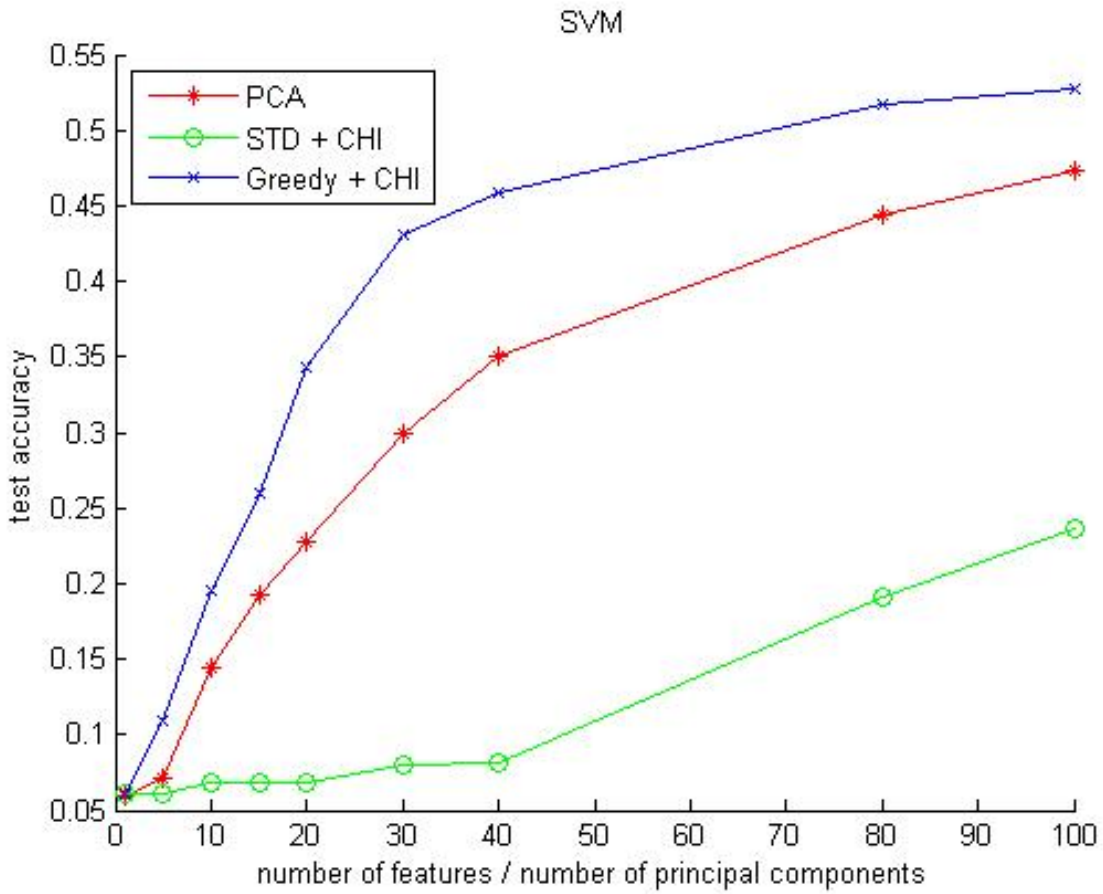
Latent Feature: PCA

Variance explained by first 300 principle components:



Comparison between Different Reduction Methods

Feature Selection: χ^2 statistic (one step and Greedy).
Latent Feature: PCA.
Classifier: SVM.



Conclusions

This project covered:

(i) Feature selection using simulated annealing algorithm outperforms feature selection using greedy algorithm on average;

(ii) Feature selection can significantly reduce dimension of feature space without significantly reduce test accuracy.