NIPS 2003 Feature Selection Competition

Yi-Wei Chen and Chih-Jen Lin

Department of Computer Science

National Taiwan University

NIPS 2003, December, 2003

Choosing Tools in the Beginning

- Simple statistical measures

  F scores

- Classification methods:
  - Support vector machines (SVM)
  - Random forest

- Reasons:

  We are more familiar with these two methods

  They are rather simple

SVM Feature Selection

- Direct use without feature selection

  Sometimes good enough

- SVM with linear kernel, choose larger primal coefficients

  Not considered here

- Radius margin bound with RBF kernel:

  Modified RBF kernel

$$K(x, y) = \exp(-g_1(x_1 - y_1)^2 - \ldots - g_n(x_n - y_n)^2)$$

Minimize leave-one-out (loo) bound:

$$\text{loo} \leq f(C, g_1, \ldots, g_n)$$

- g_i close to zero, less important

  Two-level minimization:

  $C, g_1, \ldots, g_n$ fixed: SVM optimization problem

  if f carefully constructed, it is differentiable

  But still difficult non-convex problems, n cannot be too large

Random Forest Feature Selection

- 500 trees

  Each tree: using a fixed number of random features

- Each tree: out of bag validation

  Feature importance

SVM and Random Forest

- Our experience:

  Same data, with full parameter selection

  SVM slightly better than RM

- But SVM requires higher cost on training+parameter selection

  SVM more sensitive to parameters

- Random Forest directly gives feature importance

  Mainly used here for selecting features

  i.e., after features selected, still use SVM for prediction

Things We Have Tried

- Validation error:

	arcene	dexter	dorothea	gisette	madelon
simple SVM	0.1331	0.1167	0.3398	0.0210	0.4017
F + SVM	0.2143	0.0800	0.2138	0.0180	0.1300
F + RF + SVM	0.3295	0.0867	0.1251	0.0400	0.0767
RF + RM					0.0750
F+RF+RM			0.1430		0.0850

- F: F score; RF: Random Forest

  SVM: Support vector machines

  RM: radius margin bound

- We focus more on the first three approaches
- Each attribute scaled to [0,1] first
- F score: threshold determined by either CV or human eyes

	arcene	dexter	dorothea	gisette	madelon
threshold	0.1	0.015	0.05	0.01	0.005

- After selecting features, parameter selection on training set conducted (with RBF kernel)

Final Submission

- Using those with the smallest validation error

	train error	valid error	test error	#features
arcene	0.0000	0.1331	0.1527	10000 (100%)
dexter	0.0033	0.0800	N/A	209 (1.04%)
dorothea	0.0256	0.1251	N/A	445 (0.45%)
gisette	0.0000	0.0180	0.0137	913 (18.26%)
madelon	0.0370	0.0750	0.0661	24 (4.8%)

- test error: December 1
- final1 and final2: the same thing except arcene

  a mistake in final1 for arcene

Discussion: SVM and gisette

- gisette: modified from MNIST digit recognition

  Simple SVM works well for this problem

	simple SVM	F + SVM
validation error	0.0210	0.0180

- SVM's problem when # features large:

  RBF kernel

  $$K(x, y) = e^{-g||x-y||^2}$$

  Same g for relevant and irrelevant features

- My experience on MNIST (784 features) and USPS (256 features):

  Features from the same kind of "sources": this issue less serious

  larger #features can be handled.

- Additional features generated from “products of pairs of variables”

  Probes: similar distribution

  This may be why SVM without feature selection works well
- Another problem simple SVM works well is arcene

  Reason ?

Discussion: Radius Margin Bound and Madlon

- The only problem that we find RM bound useful
- Good results by Wei Chu

  I guess they use Bayesian SVM [Chu, Keerthi, Ong]

  Under Bayesian framework,

$$\min f(C, g_1, \ldots, g_n)$$

- Though two different derivations

  Formula a little bit related to the RM loo bound
- In practice: once Keerthi told me that when testing some UCI problems, Bayesian SVM works similar to using one single g, but improve 5% on `splice`

  We then checked the RM bound

The same result

- Looks like this problem is another `splice`
- Issue: Can we know from the generation of this data why the two formulas work ?

Conclusions

- The whole procedure a bit ad hoc

  More systematic procedures ?

- Domain knowledge not used

- We thank organizers for this interesting competition