Slide 1

# IJCNN 2001 Challenge: Generalization Ability and Text Decoding

**Chih-Chung Chang and Chih-Jen Lin**

Department of Computer Science and
Information Engineering
National Taiwan University

IJCNN 2001, Washington D.C., July 17 2001

Slide 2

## Outline

- Support vector machines (SVM) for classification and Regression
- Generalization ability challenge (GAC)
- Text decoding cgallenge (TDC)

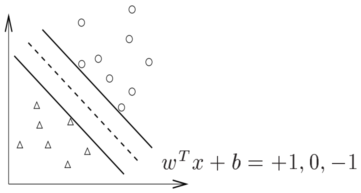




Slide 3

## Support Vector Machines

- Classification: finding maximal margin between two classes with nonlinear mapping

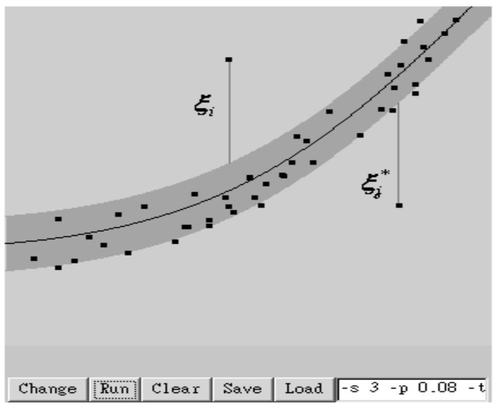



$$\min \quad \frac{1}{2} w^T w + C(\sum_{i=1}^{l} \xi_i)$$

$$y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, i = 1, \ldots, l$$

Slide 4

- Regression: $\epsilon$-insensitive loss function



Slide 5

## Generalization Ability Challenge (GAC)

- Given time series length $T = 50,000$
- The $k$th data

$$x_1(k), x_2(k), x_3(k), x_4(k), x_5(k), y(k)$$

- $y(k)$: output, affected only by $x_1(k), \ldots, x_4(k)$
- $x_5(k) = 1$ considering $y(k)$ for evaluating accuracy
- 50,000 training data, 100,000 testing data (in two sets)
- Past and future information may affect $y(k)$

Slide 6

## GAC: Data Analysis

- $x_1(k)$: periodically nine 0s, one 1, nine 0s, one 1, and so on.
- $x_2(k), \ldots, x_4(k)$: in $[-1.5, 1.5]$
- 90% of 50,000 training $y(k) = -1$
- Guessing all data -1: 90% accuracy ??
- 90% $x_5(k) : 1$





Slide 7

## GAC: Encoding Schemes

- $x_1(k)$: two ways
  - 10 binary attributes
    $x_1(k-5), \ldots, x_1(k+4)$ for the $k$th data
  - $x_1(k)$: an integer in 1 to 10
  - Which one is bettter
  - We think 10 binaries better for SVM
- $x_4(k)$ more important
  - Including $x_4(k-5), \ldots, x_4(k+4)$ for the $k$th data
  - Why these 10 data ? By some cross validation
- Each training data: 22 attributes

Slide 8

## GAC: scaling

- Should we scale each attribute to $[-1, 1]$ or $[0, 1]$ ?
- Usualy needed for SVM
- Now each attribute is in $[-1.5, 1.5]$
- We decided not to do this

Slide 9

## GAC: Training SVM

- Software used: LIBSVM
- 5-fold cross validation
- RBF kernel $K(x_i, x_j) = \phi(x_i)\phi(x_j) = e^{-\gamma \|x_i - x_j\|^2}$
- Two parameters: $\gamma$ and $C$
- Try 3000 data first:
- Best parameter: $C \approx 2^7, \gamma \approx 2^0$, CV rate $\geq 97\%$
- Try 10,000 data

Slide 10

## LIBSVM: an Integrated Library for SVM

- Includes
  - Support vector classification/regression
  - $\nu$-Support vector classification/regression
  - One-class support vector classification
- Features
  - Both C++/Java
  - Unbalanced data
  - Graphic interface
- Interface
  - Matlab: from Ohio State University

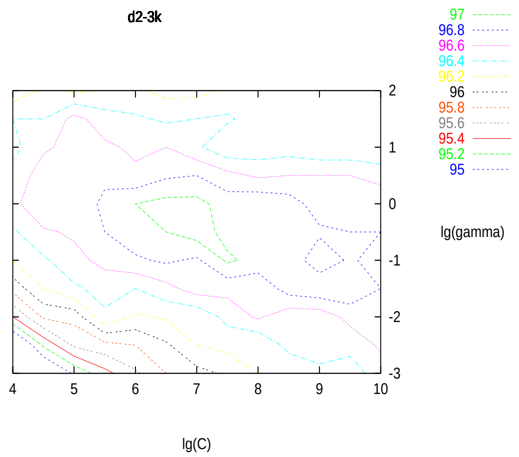




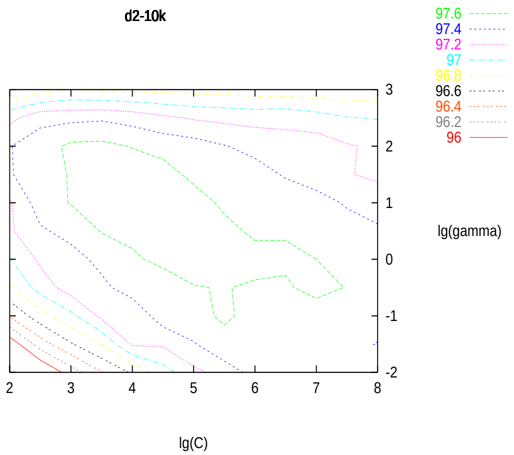
Slide 11

  - R interface from Technische Universität Wien
- Software, document, and FAQ available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

Slide 12



**Slide 13**



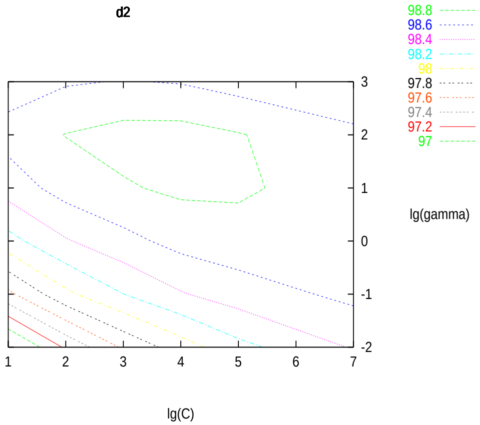

**Slide 14**

- Best parameter $C: 2^4 - 2^5, \gamma: 2^0 - 2^1$
- Best parameters moved
- Try all 50,000 data





**Slide 15**




**Slide 16**

- Moved again
- Not very clear why this happens
- Due to scaling ?? Tried scaled data but similar
- Anyway, the move is not much for larger data sets
- A dimemma: Using $C = 2^4, \gamma = 2^1 - 2^2$
  - Parameter by training 40,000 data (CV)
  - We have 50,000
- Final decision: $C = 2^4, \gamma = 2^2$
- Training SVM: One $(C, \gamma)$, 50,000 data, 5-fold CV: $< 1$ hour on PIII-500 with 384 RAM

**Slide 17**

## GAC: Results and Discussions

- Test set 1: 656 errors, Test set 2: 637 errors
- About 3000 support vectors of 50,000 training data
  A good case for SVM
- We are lucky as this is a problem where model selection works
- Different encodings: 10 $x_2(k)$ and 10 $x_3(k)$ ?

**Slide 18**

## Text Decoding Challenge (TDC)

- English text $\Rightarrow$ binary sequence by 8-bit ASCII
- SPACE $\Rightarrow$ NULL
- Encoding: Slide a window of 2 bits (left $\rightarrow$ right)
- Each two-bit pattern mapped by one of four functions:

$$y(k) = f_i(u_1(k), u_2(k)), i = \mathtt{00}, \mathtt{01}, \mathtt{10}, \mathtt{11},$$

  $u_1(k), u_2(k)$ are iid uniformly in (-1,+1).
- Example: a sequence

$$\mathtt{001011}\ldots$$





**Slide 19**

$$y(1) = f_{\mathtt{00}}(u_1(1), u_2(1))$$
$$y(2) = f_{\mathtt{01}}(u_1(1), u_2(1))$$
$$\vdots$$

- Problem: Given

$$(u_1(k), u_2(k), y(k))$$

- Find (decoding)

$$bit(k)$$

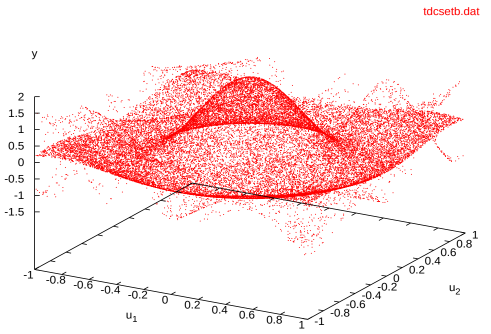
**Slide 20**

## TAC: Data Analysis

- 13.5% bits are one
- Proability of 00: $(1 - 0.135)^2 \approx 0.75$
- Most patterns are 00

**Slide 21**

## All $(u_1(k), u_2(k), y(k))$




**Slide 22**

## TDC: Guessing $f_{00}$

- 75% data on a special surface
- The same $\sqrt{u_1(k)^2 + u_2(k)^2}$, the same $y(k)$
- Define

$$d(k) \equiv \sqrt{u_1(k)^2 + u_2(k)^2}$$

- Guessing

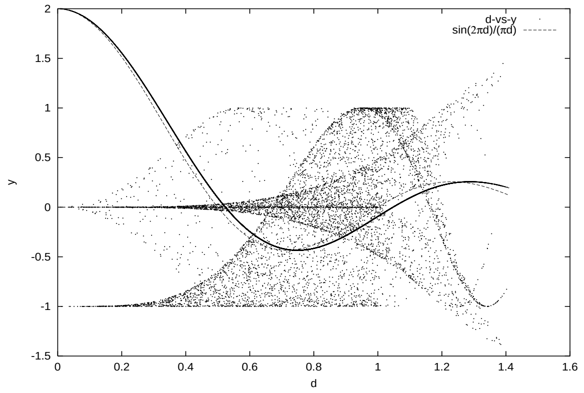
$$f_{00}(u_1, u_2) = \frac{\sin(2\pi d)}{\pi d} + g(u_1, u_2),$$





**Slide 23**

## TDC: $d(k)$ versus $y(k)$




**Slide 24**

## Extract Points on $f_{00}$

- $g$: unknown smooth function
- Linear interpolation
  - Starting from $d = 0$. At $k$:
  - 5 previous points already considered 00: choose medium of y
  - 5 next points not processed yet: medium of $y$
  - A line passing two mediums

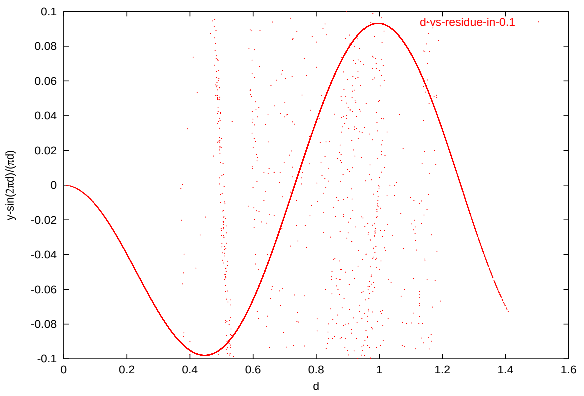
    $$|\text{vertical distance between the line and } y(k)| < 0.001$$

  - $k$th pattern considered 00
- Heuristic confirming 0 bits:
  English text: first bit of each 8-bit character should be 0

Dept. of CS, National Taiwan Univ. July, 2001

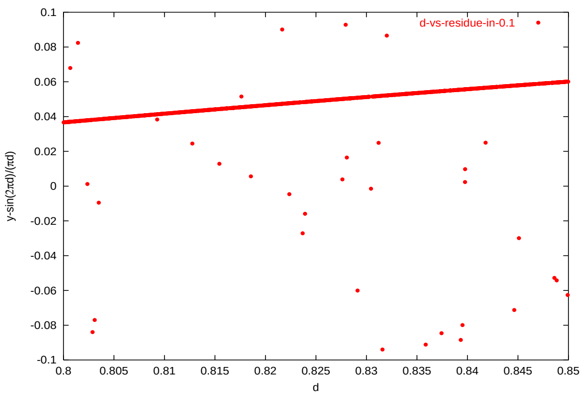
**Slide 25**

## TDC: $d$ versus $g(u_1, u_2)$




**Slide 26**

## TDC: $g$ in a small region








**Slide 27**

## Finding $f_{01}$ and $f_{10}$

- Now we have sequences

  ...00????00??0000?00...

- We have

  0? not 00 $\Rightarrow$ 01
  ?0 not 00 $\Rightarrow$ 10

- The sequence becomes

  ...001??100110000100...

- Beginning of the English text
  A C ? ? ? ? ? H ? ? ? ? ? ? ? ? H ? D ? ? ? ? F A B ? I ? H ? L ? L I ? ? ? ? ? ? ? L A ? ? ? D

**Slide 28**

## TDC: Finding 01 and 10 Points

- Now we have some 01 and 10 points
- Support vector regression finding two surfaces
- $\epsilon = 0, \gamma = 20, C = 1$
- Large $\gamma$ of the RBF: $e^{-\gamma \|x_i - x_j\|^2}$ with $\epsilon = 0$: surface fitting
- Points in $\pm 0.005 \Rightarrow$ considered 01(or 10)
- Cannot conflict with existing patterns:
  - ?1 but its $y$ close surface of 10:not change it to 10
  - Some points may be on the intersections 4 functions

**Slide 29**

- Now the first sentence becomes:
  A C R ? S S T H E S T R E E T T H E D ? ? R ? F A B R I ? H T L Y L I T S T ? R E S L A M M E D
- Remaining points are likely to be 11
  A C R O S S T H E S T R E E T T H E D O O R O F A B R I G H T L Y L I T S T O R E S L A M M E D

**Slide 30**

## TDC: Final Corrections

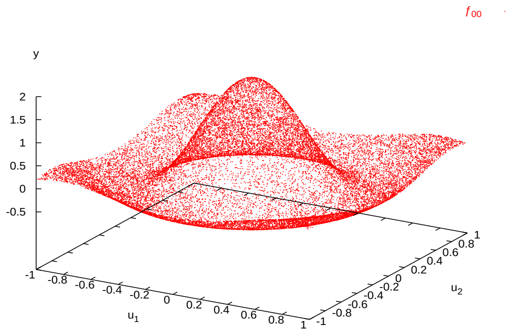
- Redo support vector regression on 4 sets using $\gamma = 20, \epsilon = 0, C = 100$
- More fitted surfaces
- $\ldots axb \ldots$ but $y$ value at $ax$ or $xb$ not very close to $f_{ax}$ or $f_{xb}$ ($\pm 0.01$)
- If $y$ in the range of both $f_{a\bar{x}}$ and $f_{\bar{x}b}$, $\bar{x}$ better than $x$
- Six bits updated
- Final draft
  ACROSS THE STREET THE DOOR OF A BRIGHTLY LIT STORE SLAMMED
- "The heart of a dog" by Mikhail Bulgakov





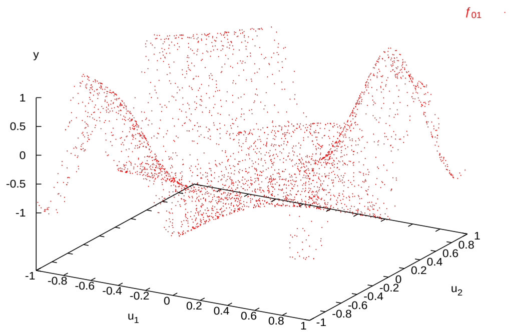
**Slide 31**

## TDC: Final $f_{00}$



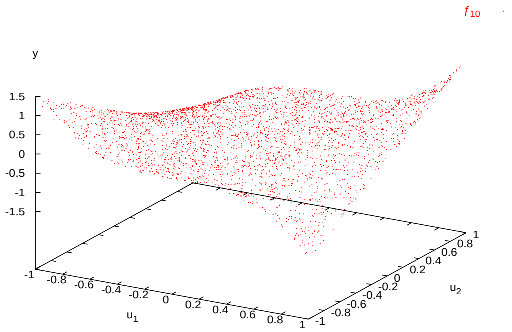

**Slide 32**

## TDC: Final $f_{01}$



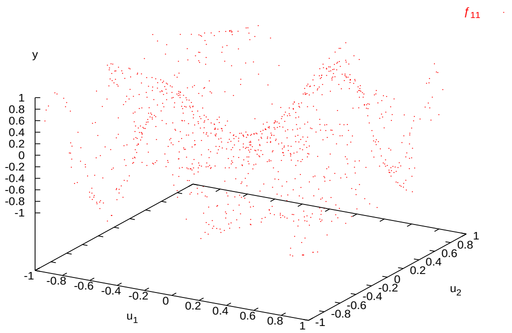
**Slide 33**

## TDC: Final $f_{10}$




**Slide 34**

## TDC: Final $f_{11}$








**Slide 35**

## TDC: Discussions

- Extracting 00: directly use SVM on all data
- $\epsilon = 0$, $\gamma$ large but not too large
- Nearly the same 00 points are obtained
- Support vector regression for separating functions

**Slide 36**

A Note on the Decomposition Methods for Support Vector Regression
Shuo-Peng Liao, Hsuan-Tien Lin, and Chih-Jen Lin

- SVR problem: $\alpha$ and $\alpha^*$
- Decomposition methods: each iteration a working set is selected
- Previous approaches: working set contains only pairs
  e.g. $\{\alpha_i, \alpha_j^*\}$ selected first $\Rightarrow$ $\{\alpha_i^*, \alpha_j\}$ added
- Reason: to keep $\alpha_i \alpha_i^* = 0$ during iterations
- Main result: it is not necessary to use the pair approach
- Both theoretical and experimental results