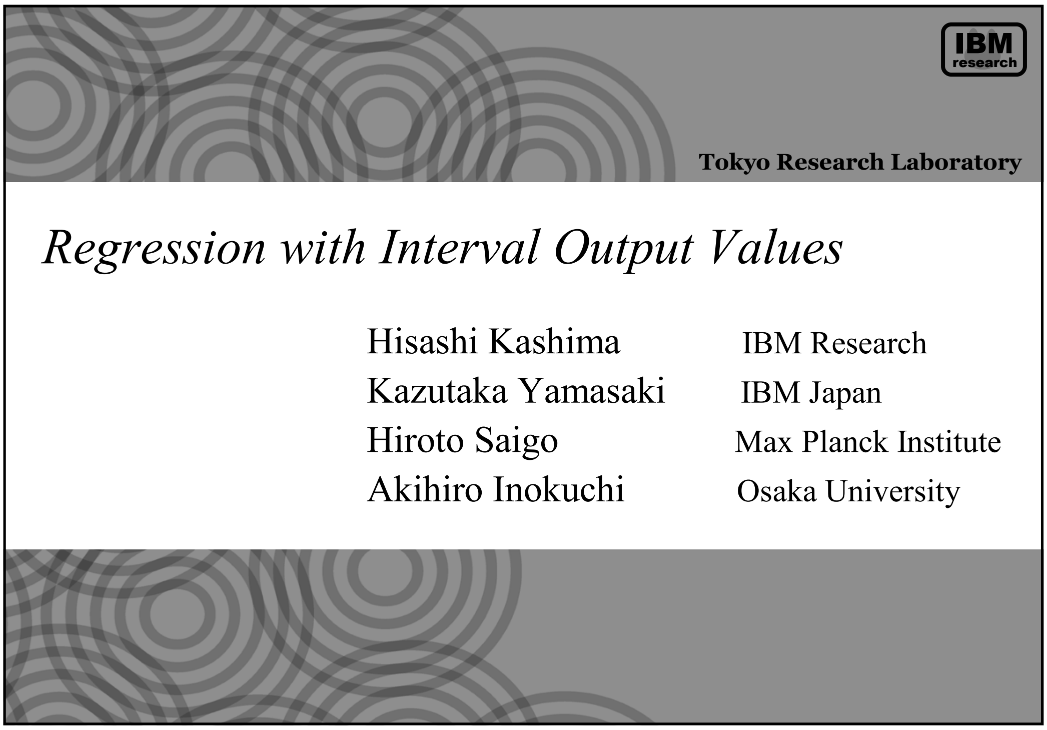
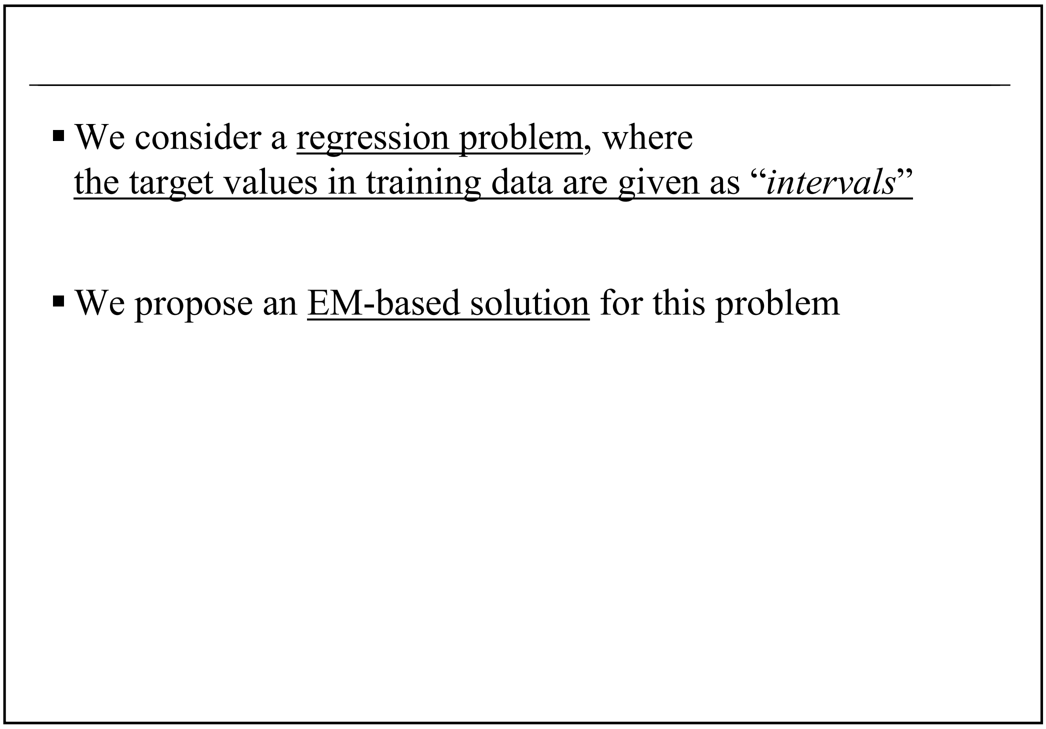


Tokyo Research Laboratory

Regression with Interval Output Values

Hisashi Kashima IBM Research
Kazutaka Yamasaki IBM Japan
Hiroto Saigo Max Planck Institute
Akihiro Inokuchi Osaka University



- We consider a regression problem, where the target values in training data are given as "*intervals*"
- We propose an EM-based solution for this problem

Problem definition: Regression with interval output values

In ordinary regression problems,
the output for a training instance is given as a *point*

ID	input **x**				output
	x_1	x_2	x_3	x_4	y
1	2	5	12	6	13
2	...	...	...	...	...
3	5	3	9	10	7

An output is given as a point

In contrast, the output is given as an *interval* **[*l*, *r*]** in our problem

ID	input **x**				output [l, r]	
	x_1	x_2	x_3	x_4	l	r
1	2	5	12	6	11	14
2	...	...	...	...		
3	5	3	9	10	6	∞

An output is given as an *interval*

Examples of regression from interval output values

- Sales prediction from past sales data with occasional losses
- Prediction of number of troubles for quality assurance of projects
- Screening of chemical compounds for drug discovery (QSAR)

Existing methods cannot handle interval outputs appropriately

Existing regression methods can estimate $p(y|\mathbf{x})$ from point outputs

ID	input **x**				output y
	x_1	x_2	x_3	x_4	
1	2	5	12	6	13
2	...	...	...	...	...
3	5	3	9	10	7

model estimation ⇨ $p(y|\mathbf{x})$

The existing regression methods can NOT handle interval outputs

- Naive solution would be "Use only the point outputted instances (and neglect instances with interval outputs)"

ID	input **x**				output **y**		
	x_1	x_2	x_3	x_4	l	r	
1	2	5	12	6	11	14	← neglected
2	...	...	...	...			
3	5	3	9	10	6	6	← used

model estimation ⇨ $p(y|\mathbf{x})$

Our approach: Iterative estimation of the model and "representative values" of interval outputs

Iterate the following two steps:

- Use the current model to give representative values to interval outputs
- Use the representative values to estimate the new model

ID	input **x**				output **y**		representative output
	x_1	x_2	x_3	x_4	l	r	
1	2	5	12	6	11	14	13
2	...	...	...	...		...	...
3	5	3	9	10	6	∞	7

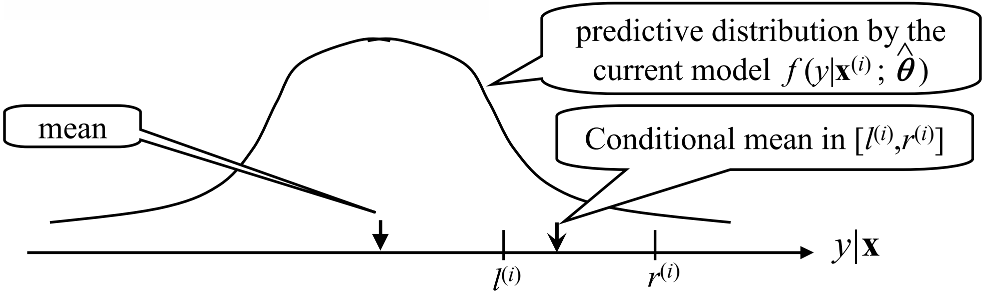
The existing regression methods can be applied if representative values are assigned

Representative output values are given as conditional means by the current model

- Using the current model $f(y \mid \mathbf{x}; \hat{\boldsymbol{\theta}})$, the representative output for the i -th instance is given as the conditional mean in $[l^{(i)}, r^{(i)}]$
  - When the conditional mean is analytical intractable, use sampling
- Interpreted as an EM-algorithm when f is in the exponential family

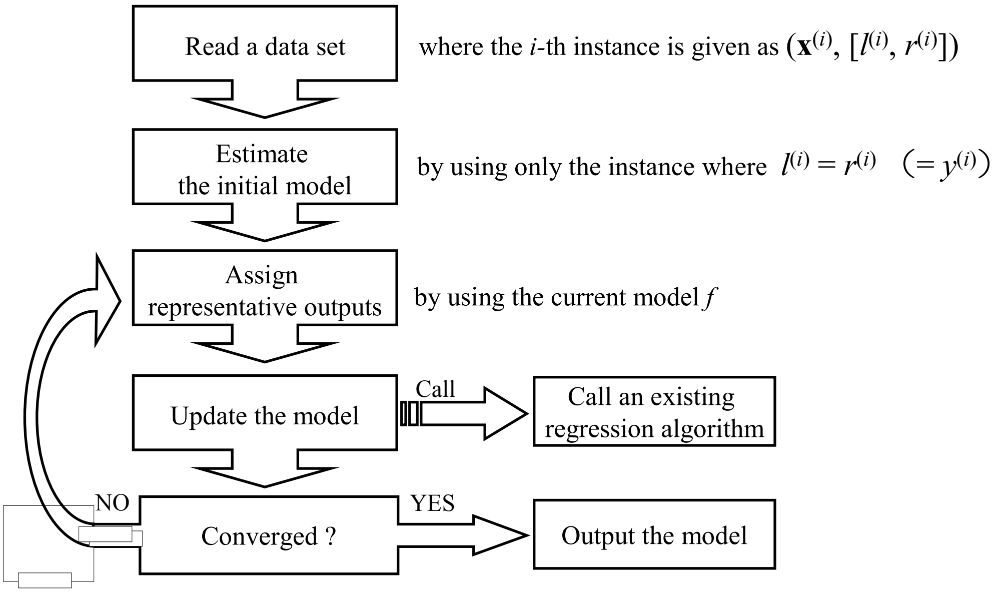
$$y^{(i)} = \frac{\int_{\ell^{(i)}}^{r^{(i)}} y f(y|\mathbf{x}^{(i)};\hat{\theta})dy}{\int_{\ell^{(i)}}^{r^{(i)}} f(y|\mathbf{x}^{(i)};\hat{\theta})dy}$$

Conditional mean in $[l^{(i)},r^{(i)}]$



predictive distribution by the current model $f(y|\mathbf{x}^{(i)}; \hat{\boldsymbol{\theta}})$
mean
Conditional mean in $[l^{(i)},r^{(i)}]$
$l^{(i)}$
$r^{(i)}$
$y|\mathbf{x}$

Procedure



Read a data set — where the i-th instance is given as $(\mathbf{x}^{(i)}, [l^{(i)}, r^{(i)}])$

Estimate the initial model — by using only the instance where $l^{(i)} = r^{(i)} \quad (= y^{(i)})$

Assign representative outputs — by using the current model f

Update the model — Call → Call an existing regression algorithm

Converged ? — NO → Assign representative outputs; YES → Output the model

Experiments with two benchmark data sets

- We used two data set
  - “Boston housing” data set: House price prediction
  - “EDKB” data set: Drug activity prediction of chemical compounds
- We compared the proposed method with two approaches
  - Method 1: Neglect instances with interval outputs
  - Method 2: Use (non-conditional) means as representative outputs
- We used a linear Gaussian model as the base regression method

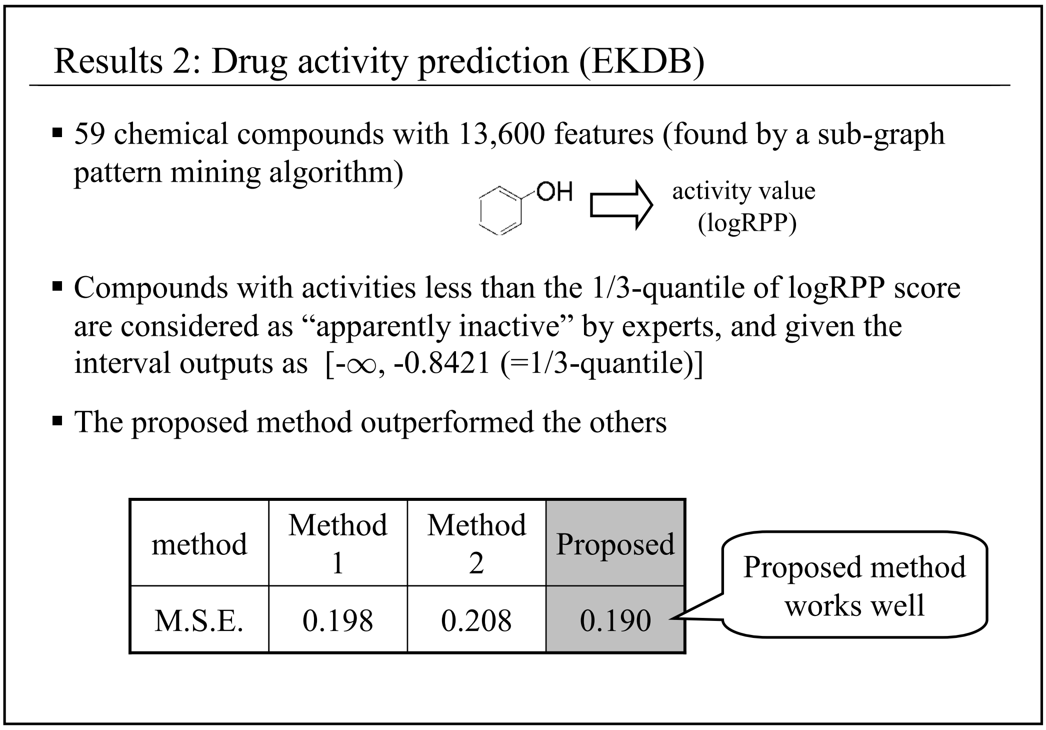
$$f(y|\mathrm{x}, \boldsymbol{\theta}, \sigma) = \mathcal{N}(y|\boldsymbol{\theta}\boldsymbol{\phi}^{\mathsf{T}}(\mathrm{x}), \sigma\mathrm{I})$$

Results 1: Price prediction (Boston housing)

- 506 instances (houses) with 12 input features (#rooms, crime rate, ...)
  - We used a Gaussian kernel as features $\phi_i(\mathbf{x}) = \exp(-\alpha \parallel \mathbf{x} - \mathbf{x}^{(i)} \parallel^2)$
- Since this data set is originally point-outputted, half of the all instances are converted to interval-outputted data by using uniformly random variables Δ_l and $\Delta_r \in 0.1 \cdot y^{(i)}$ for the *i*-th instance
  - Proposed method L uses $[y^{(i)} - \Delta_l, \infty]$ as the *i*-th interval
  - Proposed method LR uses $[y - \Delta_l, y + \Delta_r]$ as the *i*-th interval
- The proposed method outperformed the others

method	Method 1	Method 2	Proposed L	Proposed LR
M.S.E.	14.18	13.84	12.03	11.37

Proposed method works well



Results 2: Drug activity prediction (EKDB)

- 59 chemical compounds with 13,600 features (found by a sub-graph pattern mining algorithm)

OH
activity value
(logRPP)

- Compounds with activities less than the 1/3-quantile of logRPP score are considered as “apparently inactive” by experts, and given the interval outputs as [-∞, -0.8421 (=1/3-quantile)]
- The proposed method outperformed the others

method	Method 1	Method 2	Proposed
M.S.E.	0.198	0.208	0.190

Proposed method works well