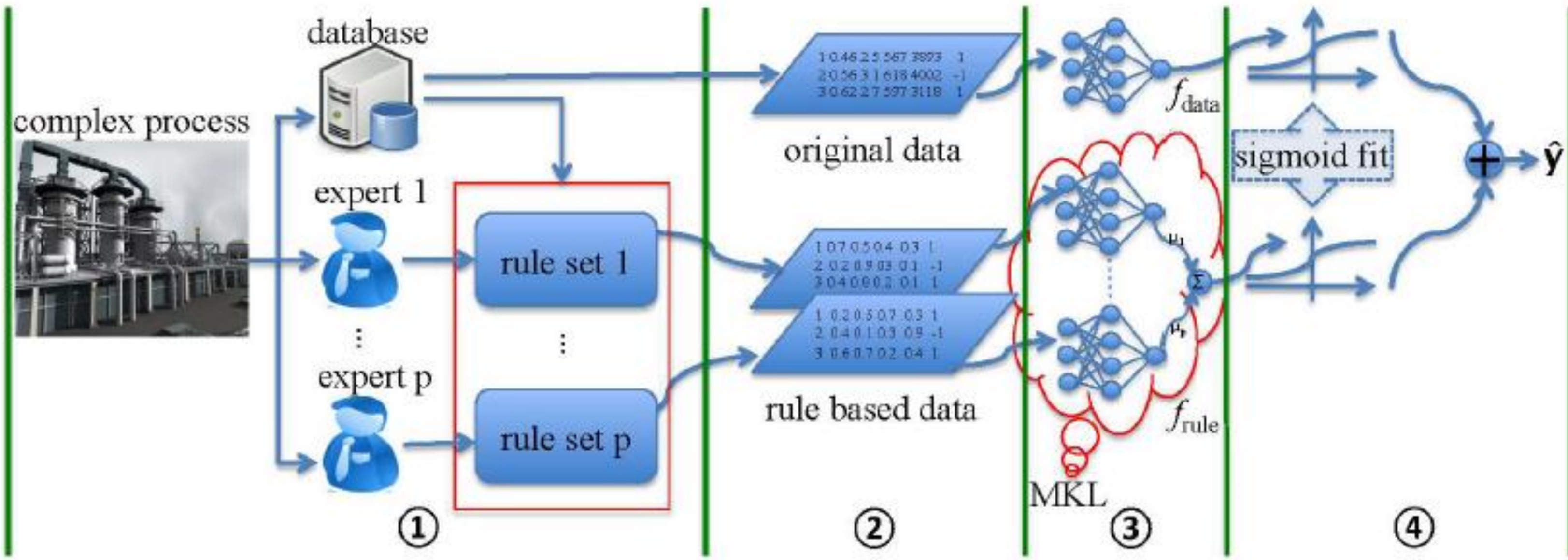
Exploiting Expertise Rules for Statistical Data-Driven Modeling

(Dec 21, 2016)

Cheng Jianlin


database
complex process
expert 1
expert p
rule set 1
rule set p
original data
rule based data
f_{data}
f_{rule}
MKL
sigmoid fit
$\hat{\mathbf{y}}$
①
②
③
④

Outline

- I. Rule based Data Expression
- II. Rule based Data Fusion by l2 Multiple Kernel Learning
- III. Ensemble Data and Rule based Model by Sigmoid Fitting
- IV. Results

• I. Rule based Data Expression

CART is used as a preprocessing step to extract if · · · then · · ·rules.

• I. Rule based Data Expression

For some expert, assume that there are m items **if** $\cdots$ **then** $\cdots$ logical rules composed of the following two kinds

- **if** $x_{i_1} \in V_{i_1}^j$ *and* $\cdots$ *and* $x_{i_t} \in V_{i_t}^j$, **then** $y = y^j$,
- **if** $x_{i_1} \in V_{i_1}^j$ *or* $\cdots$ *or* $x_{i_t} \in V_{i_t}^j$, **then** $y = y^j$.

Here $\{V_{i_q}^j\}_{q=1,\cdots,t}^{j=1,\cdots,m}$ is called basic rule interval including three types, i.e. $(-\infty\ \ a)$, $[b\ \ +\infty)$ and $[c\ \ d)$. $\{i_1, \cdots, i_t\} \subset \{1, \cdots, n\}$ is the index of features used in the jth rule. Next we define the i_qth feature of the kth sample's membership, denoted as $\delta_{V_{i_q}^j}(x_{k_{i_q}}) \in [0\ 1]$, on the rule interval $V_{i_q}^j$ as

$$\delta_{V_{i_q}^j}(x_{k_{i_q}}) = \begin{cases} \dfrac{1}{1+e^{s_{i_q}^j - x_{k_{i_q}}}}, & if\ V_{i_q}^j \triangleq [b, +\infty) \\ e^{-\frac{|x_{k_{i_q}} - m_{i_q}^j|}{\sigma_{i_q}^j}}, & if\ V_{i_q}^j \triangleq [c, d) \\ \dfrac{1}{1+e^{x_{k_{i_q}} - b_{i_q}^j}}, & if\ V_{i_q}^j \triangleq (-\infty, a). \end{cases} \qquad (1)$$

• I. Rule based Data Expression

where $s_{i_q}^j = \min_k\{x_{k_{i_q}} | x_{k_{i_q}} \in V_{i_q}^j, k = 1, \cdots, l\}$ and $b_{i_q}^j = \max_k\{x_{k_{i_q}} | x_{k_{i_q}} \in V_{i_q}^j, k = 1, \cdots, l\}$, $m_{i_q}^j$ and $\sigma_{i_q}^j$ denote the mean and standard deviation of the i_qth feature of samples in the set $\{x_{k_{i_q}} | x_{k_{i_q}} \in V_{i_q}^j, k = 1, \cdots, l\}$. Then, operators $\vee$ and $\wedge$ are used to define the $or-$type and $and-$type rule's density support of input sample $\mathbf{x}_k$ as

$$r_j(\mathbf{x}_k) = \vee_{q=1}^t \delta_{V_{i_q}^j}(x_{k_{i_q}}) = \max_{1 \le q \le t} \delta_{V_{i_q}^j}(x_{k_{i_q}}), \quad (2)$$

and

$$r_j(\mathbf{x}_k) = \wedge_{q=1}^t \delta_{V_{i_q}^j}(x_{k_{i_q}}) = \min_{1 \le q \le t} \delta_{V_{i_q}^j}(x_{k_{i_q}}), \quad (3)$$

respectively. $r_j(\mathbf{x}_k)$ depicts the support degree of the kth sample $\mathbf{x}_k$ to the jth rule. In this way, we generate the rule based data for the specified expert as

$$\mathbb{R} = \{\mathbf{r}_k, y_k\}_{k=1}^l \quad (4)$$

• II. . Rule based Data Fusion by l2 Multiple Learning

$$\min_{\mathbf{w},b,\mathbf{e}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{1}{2\nu}\sum_{k=1}^{l} e_k^2 \tag{5}$$

$$\text{s. t.} \quad y_k = \mathbf{w}^T\Phi(\mathbf{x}_k) + b + e_k, \;\; k = 1, \ldots, l. \tag{6}$$

grangian function of Eqs.(5,6) is $L(\mathbf{w}, b, \mathbf{e}; \boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{1}{2\nu}\sum_{k=1}^{l} e_k^2 - \sum_{k=1}^{l}\alpha_k(\mathbf{w}^T\Phi(\mathbf{x}_k) + b + e_k - y_k)$ where α_k is

$$\min_{\mathbf{w},b,\mathbf{e}} \max_{\boldsymbol{\alpha}} L(\mathbf{w}, b, \mathbf{e}; \boldsymbol{\alpha}) = \max_{\boldsymbol{\alpha}} \min_{\mathbf{w},b,\mathbf{e}} L(\mathbf{w}, b, \mathbf{e}; \boldsymbol{\alpha})$$

$$\max_{\boldsymbol{\alpha}} \quad \boldsymbol{\alpha}^T\mathbf{y} - \frac{1}{2}\boldsymbol{\alpha}^T K\boldsymbol{\alpha} - \frac{\nu}{2}\boldsymbol{\alpha}^T\boldsymbol{\alpha} \tag{7}$$

$$\text{s. t.} \quad \sum_{i=1}^{l}\alpha_i = 0, \tag{8}$$

• II. . Rule based Data Fusion by l2 Multiple Learning

$$\begin{aligned}\omega(K) &= \max\{\boldsymbol{\alpha}^T\mathbf{y} - \frac{1}{2}\boldsymbol{\alpha}^T K\boldsymbol{\alpha} - \frac{\nu}{2}\boldsymbol{\alpha}^T\boldsymbol{\alpha} | \boldsymbol{\alpha}^T\mathbf{1} = 0\} \\ &= -\min\{\frac{1}{2}\boldsymbol{\alpha}^T K\boldsymbol{\alpha} + \frac{\nu}{2}\boldsymbol{\alpha}^T\boldsymbol{\alpha} - \boldsymbol{\alpha}^T\mathbf{y} | \boldsymbol{\alpha}^T\mathbf{1} = 0\}.\end{aligned}$$

$$\min_{\boldsymbol{\mu}} 2\omega(\sum_{i=1}^{p} \mu_i K_i)$$

$$\max_{\boldsymbol{\mu}} -2\omega(\sum_{i=1}^{p+1} \mu_i K_i) \tag{10}$$

$$= \max_{\boldsymbol{\mu}\geq\mathbf{0},\|\boldsymbol{\mu}\|=1} \min_{\boldsymbol{\alpha}^T\mathbf{1}=0} \{\sum_{i=1}^{p+1} \mu_i \boldsymbol{\alpha}^T K_i \boldsymbol{\alpha} - 2\boldsymbol{\alpha}^T\mathbf{y}\} \tag{11}$$

$$\stackrel{1}{=} \max_{\boldsymbol{\mu}\geq\mathbf{0},\|\boldsymbol{\mu}\|\leq 1} \min_{\boldsymbol{\alpha}^T\mathbf{1}=0} \{\sum_{i=1}^{p+1} \mu_i \boldsymbol{\alpha}^T K_i \boldsymbol{\alpha} - 2\boldsymbol{\alpha}^T\mathbf{y}\}. \tag{12}$$

• II. . Rule based Data Fusion by l2 Multiple Learning

$$\max_{\boldsymbol{\mu},\theta} \quad \theta \tag{13}$$

$$\text{s.t.} \quad \|\boldsymbol{\mu}\| \leq 1, \tag{14}$$

$$\mu_i \geq 0, i = 1, \cdots, p+1, \tag{15}$$

$$\sum_{i=1}^{p+1} \mu_i f_i(\alpha) - 2\sum_{k=1}^{l} \alpha_k y_k \geq \theta, \tag{16}$$

$$\sum_{k=1}^{l} \alpha_k = 0, \tag{17}$$

where $f_i(\alpha) = \alpha^T K_i \alpha$, $i = 1, \cdots, p+1$.

• III. Ensemble Data and Rule based Model by Sigmoid Fitting

$$\begin{aligned} P(y=1|\mathbf{x}) &\approx P(y=1|f(\mathbf{x})) \\ &= \frac{1}{1+\exp(\varepsilon f(\mathbf{x})+\gamma)}. \end{aligned} \tag{18}$$

$$\min_{\varepsilon,\gamma} \; -\sum_{k=1}^{l} t_k \log(P_k) + (1-t_k)\log(1-P_k), \tag{19}$$

where

$$\begin{cases} t_k = \begin{cases} \frac{N_+ + 1}{N_+ + 2}, & \text{if } y_k = 1 \\ \frac{1}{N_- + 2}, & \text{if } y_k = -1 \end{cases} \\ P_k = \frac{1}{1+\exp(\varepsilon f(\mathbf{x}_k)+\gamma)}. \end{cases} \tag{20}$$

• III. Ensemble Data and Rule based Model by Sigmoid Fitting

$$\hat{y} = \begin{cases} 1, \text{if } \frac{P_{data}+P_{rule}}{2} \in [0.5\ 1] \\ -1, \text{if } \frac{P_{data}+P_{rule}}{2} \in [0\ 0.5). \end{cases} \tag{21}$$

Algorithm 1 Rule Aided Statistical Data-driven Modeling

Input: data set $\mathbb{D} = \{\mathbf{x}_k, y_k\}_{k=1}^{l}$, rule sets $\mathbb{E}_1, \mathbb{E}_2, \cdots, \mathbb{E}_p$
Output: decision value $\hat{y}$

1: With $\mathbb{E}_i$, transform $\mathbb{D} = \{\mathbf{x}_k, y_k\}_{k=1}^{l}$ into rule based data $\mathbb{R}_i = \{\mathbf{r}_k^i, y_k\}_{k=1}^{l}$ $(i = 1, \cdots, p)$ through Eqs.(1-4);
2: Generate kernel matrix K_i according to $\mathbb{R}_i$ $(i = 1, \cdots, p)$;
3: Employ ℓ_2 MKL algorithm to learn μ_i $(i = 1, \cdots, p)$ through Eqs.(13-17) and derive the rule based classifier f_{rule};
4: Train LS-SVMs classifier on $\mathbb{D} = \{\mathbf{x}_k, y_k\}_{k=1}^{l}$, and get the data based classifier f_{data};
5: Optimize ε and γ in Eq.(18), and transform the decision values of f_{rule} and f_{data} into posterior probabilities;
6: Ensemble rule based model and data based model, and output the decision value $\hat{y}$ through Eq.(21).

• IV. Results

·A Toy Experiment

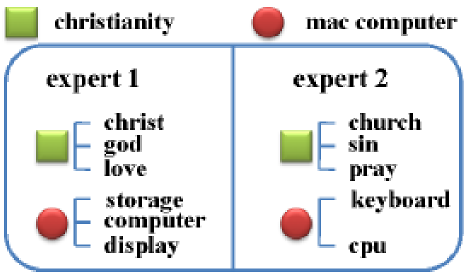


Fig. 2. Green square indicates that a document is of "Christianity" and red circle stands for "Mac Computer".

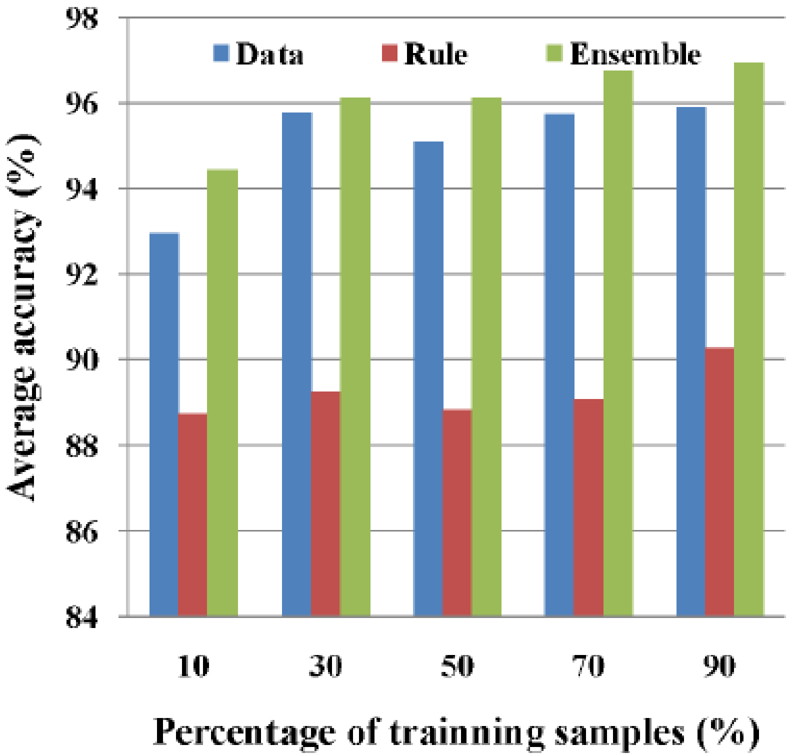


Fig. 3. A toy experiment about two-class document classificaiton.

• IV. Results

• Evaluating the Performance on Benchmark Datasets

TABLE III
10-FOLD CROSS-VALIDATION RESULTS OF DIFFERENT MODELS

Dataset	$f_{Bagging}$	f_{Data}	f_{Rule}	$f_{Ensemble}$
Cancer	97.08±3.39	**97.50**±2.76	96.46±3.09	**97.50**±3.06
Codrna	85.67±9.43	88.67±8.59	85.00±6.87	**89.00**±5.39
German	75.75±5.01	76.75±6.03	74.50±3.92	**77.25**±6.22
Heart	79.00±14.45	77.00±12.69	75.00±10.24	**90.00**±7.75
Ijcnn1	**95.37**±2.74	89.13±2.31	88.50±2.42	92.25±1.75
Ionosphere	93.43±4.25	93.14±4.81	90.00±6.03	**94.86**±4.39

• IV. Results

•Tendency Prediction of Thermal State of Blast Furnace

TABLE VI
10-FOLD CROSS-VALIDATION RESULTS OF BLAST FURNACE DATASETS

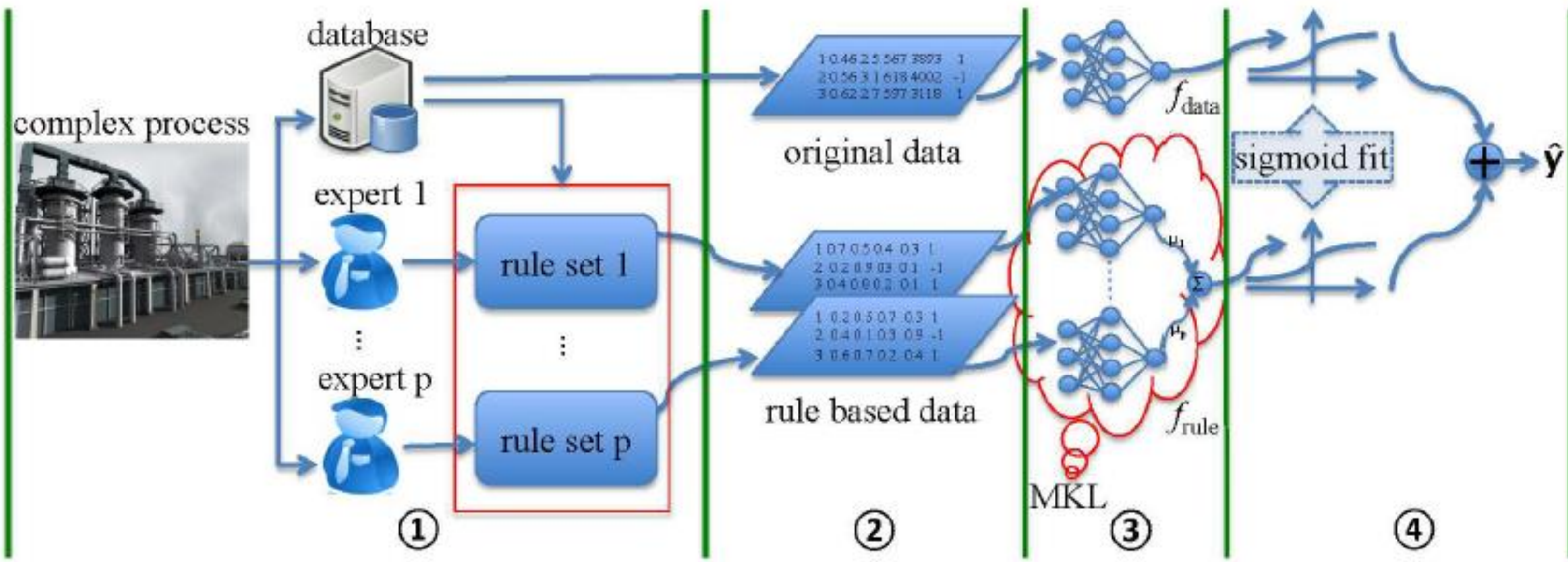
BF	$f_{Bagging}$	f_{Data}	f_{Rule}	$f_{Ensemble}$
(a)	68.33±8.27	77.83±6.58	67.67±6.33	**78.56**±5.94
(b)	77.33±5.49	77.83±4.54	78.83±3.88	**79.17**±4.96

Bibliography

- [1] Exploiting Expertise Rules for Statistical Data-Driven Modeling

Thanks！

模型结构图



database
complex process
expert 1
expert p
rule set 1
rule set p
original data
rule based data
f_{data}
f_{rule}
MKL
sigmoid fit
$\hat{y}$
①
②
③
④

创新点

- I. Rule Extraction through Decision Tree and how to transform the origin data into rule based data according to the expertise rules
- II. Rule based Data Fusion by Multiple Kernel Learning

意义

- I. Improve the model
- II. Make the model more understandable

•方法的有待提高的点

- I. The weights of two models are equal