Interval Data Classification under Partial Information: A Chance-constraint Approach

Sahely Bhadra *J. Saketha Nath* Aharon Ben-Tal
Chiranjib Bhattacharyya

PAKDD 2009

Data Uncertainty

- Real-world data fraught with uncertainties, noise.
  - Measurement errors, non-zero least counts etc.
  - Inherent heterogenity:
    - Bio-medical data e.g. Micro-array, cancer diagnostic data.
  - Computational/Respresentational convinience.

Data Uncertainty

- Real-world data fraught with uncertainties, noise.
    - Measurement errors, non-zero least counts etc.
    - Inherent heterogenity:
        - Bio-medical data e.g. Micro-array, cancer diagnostic data.
    - Computational/Respresentational convinience.
- Many datasets provide partial information regarding noise.
    - e.g., Wisconsin breast cancer datasets (support, mean, std. err.)
    - Micro-array datasets (replicates)

Data Uncertainty

- Real-world data fraught with uncertainties, noise.
  - Measurement errors, non-zero least counts etc.
  - Inherent heterogenity:
    - Bio-medical data e.g. Micro-array, cancer diagnostic data.
  - Computational/Respresentational convinience.
- Many datasets provide partial information regarding noise.
  - e.g., Wisconsin breast cancer datasets (support, mean, std. err.)
  - Micro-array datasets (replicates)
- Classifiers accounting for uncertainty generalize better.

Problem Definition

Problem:

- Assume partial information regarding uncertainties given:
  - bounding intervals (i.e. support) and means of uncertain eg.
- Make no distributional assumptions.
- Construct classifier that generalizes well.

Existing Methodology

[Laurent El Ghaoui *et.al.*, 2003]:

- Utilize support alone; neglect statistical information
  - True datapoint lies somewhere in bounding hyper-rectangle
- Construct regular SVM

Existing Methodology

[Laurent El Ghaoui *et.al.*, 2003]:

- Utilize support alone; neglect statistical information
  - True datapoint lies somewhere in bounding hyper-rectangle
- Construct regular SVM

SVM Formulation:

$$\min_{\mathbf{w},b,\xi_i} \quad \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_i \xi_i$$

$$\text{s.t.} \quad y_i(\mathbf{w}^\top \mathbf{x}_i - b) \geq 1 - \xi_i, \ \xi_i \geq 0$$

Existing Methodology

[Laurent El Ghaoui *et.al.*, 2003]:

- Utilize support alone; neglect statistical information
  - True datapoint lies somewhere in bounding hyper-rectangle
- Construct regular SVM

IC-BH Formulation:

$$\min_{\mathbf{w},b,\xi_i} \quad \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_i \xi_i$$

$$\text{s.t.} \quad y_i(\mathbf{w}^\top \mathbf{x}_i - b) \geq 1 - \xi_i,\ \xi_i \geq 0,\ \forall\ \mathbf{x}_i \in \mathcal{R}_i$$

Limitations of Existing Methodologies

- Neglect useful statistical information regarding uncertainty
- Overly-conservative uncertainty modeling leads to less margin
  - Poor generalization

Limitations of Existing Methodologies

- Neglect useful statistical information regarding uncertainty
- Overly-conservative uncertainty modeling leads to less margin
  - Poor generalization

Proposed Methodology:

- Use both support and statistical information
- Employ Chance-Constraint Program (CCP) approaches
- Relax CCP using Bernstein bounding schemes
  - Not overly-conservative — better margin and generalization
  - Leads to convex Second Order Cone Program (SOCP)

Proposed Formulation

SVM:

$$\min_{\mathbf{w},b,\xi_i} \quad \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_i \xi_i$$
$$\text{s.t.} \quad y_i(\mathbf{w}^\top \mathbf{x}_i - b) \geq 1 - \xi_i \quad , \; \xi_i \geq 0$$

Proposed Formulation

SVM:

$$\min_{\mathbf{w},b,\xi_i} \quad \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_i \xi_i$$

$$\text{s.t.} \quad y_i(\mathbf{w}^\top X_i - b) \geq 1 - \xi_i \quad , \ \xi_i \geq 0$$

Proposed Formulation

Chance-Constrained Program:

$$\min_{\mathbf{w},b,\xi_i} \quad \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_i \xi_i$$
$$\text{s.t.} \quad Prob\left\{y_i(\mathbf{w}^\top X_i - b) \leq 1 - \xi_i\right\} \leq \epsilon,\ \xi_i \geq 0$$

Proposed Formulation

Chance-Constrained Program:

$$\min_{\mathbf{w},b,\xi_i} \quad \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_i \xi_i$$
$$\text{s.t.} \quad Prob\left\{y_i(\mathbf{w}^\top X_i - b) \leq 1 - \xi_i\right\} \leq \epsilon,\ \xi_i \geq 0$$

Assumptions:

- $X_i \in \mathcal{R}_i$.
- $\mathbb{E}[X_i]$ are known.
- $X_{ij}, j = 1, \ldots, n$ are independent random variables.

Convex Relaxation

Comments:

- In general, difficult to solve such CCPs.
- Construct an efficient relaxation:
  - Employ Bernstein schemes to upper bound probability
  - Constrain the upper-bound to be less than ϵ

Convex Relaxation

Comments:

- In general, difficult to solve such CCPs.
- Construct an efficient relaxation:
  - Employ Bernstein schemes to upper bound probability
  - Constrain the upper-bound to be less than ϵ

Key Question:

$$Prob\left\{y_i(\mathbf{w}^\top X_i - b) \leq 1 - \xi_i\right\} \leq \text{?}$$

Convex Relaxation

Comments:

- In general, difficult to solve such CCPs.
- Construct an efficient relaxation:
  - Employ Bernstein schemes to upper bound probability
  - Constrain the upper-bound to be less than ϵ

Key Question:

$$Prob\left\{y_i(\mathbf{w}^\top X_i - b) \leq 1 - \xi_i\right\} \leq \textbf{?} \leq \epsilon$$

Convex Relaxation

Comments:

- In general, difficult to solve such CCPs.
- Construct an efficient relaxation:
  - Employ Bernstein schemes to upper bound probability
  - Constrain the upper-bound to be less than ϵ

Key Question:

$$Prob\left\{\sum_j u_{ij}X_{ij} + u_{i0} \geq 0\right\} \leq \text{?}$$

Bernstein Bounding

Markov Bounding:

$$Prob\,(X \geq 0) \leq \text{?}$$

Bernstein Bounding

Markov Bounding:

$$\mathbb{E}_X\left[1_{X\geq 0}\right] \leq \text{?}$$

Bernstein Bounding

Markov Bounding:

$$\mathbb{E}_X\left[1_{X\geq 0}\right] \leq \mathbb{E}\left[\exp\left\{\alpha X\right\}\right] \;\forall\; \alpha \geq 0$$

Bernstein Bounding

Markov Bounding:

$$\begin{aligned}\mathbb{E}_X\left[1_{X\geq 0}\right] &\leq \mathbb{E}\left[\exp\left\{\alpha X\right\}\right]\ \forall\ \alpha \geq 0\\ &= \mathbb{E}\left[\exp\left\{\alpha\left(\sum_j u_{ij}X_{ij} + u_{i0}\right)\right\}\right]\\ &= \exp\left\{u_{i0}\right\}\Pi_j\mathbb{E}\left[\exp\left\{\alpha u_{ij}X_{ij}\right\}\right]\end{aligned}$$

Bernstein Bounding

Markov Bounding:

$$\mathbb{E}_X\left[1_{X\geq 0}\right] \leq \mathbb{E}\left[\exp\left\{\alpha X\right\}\right]\ \forall\ \alpha \geq 0$$
$$= \mathbb{E}\left[\exp\left\{\alpha\left(\sum_j u_{ij}X_{ij} + u_{i0}\right)\right\}\right]$$
$$= \exp\left\{u_{i0}\right\}\Pi_j\mathbb{E}\left[\exp\left\{\alpha u_{ij}X_{ij}\right\}\right]$$

Bernstein Bounding

Markov Bounding:

$$
\begin{aligned}
\mathbb{E}_X\left[1_{X\geq 0}\right] &\leq \mathbb{E}\left[\exp\left\{\alpha X\right\}\right]\ \forall\ \alpha \geq 0 \\
&= \mathbb{E}\left[\exp\left\{\alpha\left(\sum_j u_{ij}X_{ij} + u_{i0}\right)\right\}\right] \\
&= \exp\left\{u_{i0}\right\}\Pi_j\mathbb{E}\left[\exp\left\{\alpha u_{ij}X_{ij}\right\}\right]
\end{aligned}
$$

Bounding Expectation:

- Given $X \in \mathcal{R}$, $\mathbb{E}[X]$, tightly bound: $\mathbb{E}[\exp\{t\mathbf{X}\}],\ \forall\ t \in \mathbb{R}$

Bernstein Bounding — Contd.

Known Result:

$$\mathbb{E}\left[\exp\{tX_{ij}\}\right] \leq \exp\left\{\frac{\mu_{ij}t + \sigma(\hat{\mu}_{ij})^2 l_{ij}^2}{2}t^2\right\} \ \forall \ t \ \in \ \mathbb{R} \qquad (1)$$

Bernstein Bounding — Contd.

Known Result:

$$\mathbb{E}\left[\exp\{tX_{ij}\}\right] \leq \exp\left\{\frac{\mu_{ij}t + \sigma(\hat{\mu}_{ij})^2 l_{ij}^2}{2}t^2\right\} \;\forall\; t \;\in\; \mathbb{R} \tag{1}$$

- Analogous with Gaussian mgf
  - Variance term varies with relative position of mean!

Bernstein Bounding — Contd.

Known Result:

$$\mathbb{E}\left[\exp\{tX_{ij}\}\right] \leq \exp\left\{\frac{\mu_{ij}t + \sigma(\hat{\mu}_{ij})^2 l_{ij}^2}{2}t^2\right\} \; \forall\; t \;\in\; \mathbb{R} \qquad (1)$$

- Analogous with Gaussian mgf
  - Variance term varies with relative position of mean!

Proof Sketch:

- Support ($a \leq X \leq b$), mean are known.
- $\exp\{tX\} \leq \frac{b-X}{b-a}\exp\{ta\} + \frac{X-a}{b-a}\exp\{tb\}$
- Taking expectations on both sides leads to:

$$\mathbb{E}\left[\exp\{tX\}\right] \leq \exp\left\{\frac{a+b}{2}t + h(lt)\right\},\; h(z) \equiv \log\left(\cosh(z) + \hat{\mu}\sinh(z)\right)$$

$$\leq \exp\left\{\mu t + \frac{\sigma\left(\hat{\mu}\right)^2 l^2}{2}t^2\right\}$$

Main Result — A Convex Formulation

IC-MBH Formulation:

$$\min_{\mathbf{w},b,\mathbf{z}_i,\xi_i\geq 0} \quad \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_i \xi_i$$

$$\text{s.t.} \quad y_i(\mathbf{w}^\top\mu_i - b) + \mathbf{z}_i{}^\top\hat{\mu}_i \geq 1 - \xi_i + \|\mathbf{z}_i\|_1 + \kappa\left\|\mathbf{\Sigma}_i(y_i\mathbf{L}_i\mathbf{w} + \mathbf{z}_i)\right\|_2$$

Geometric Interpretation

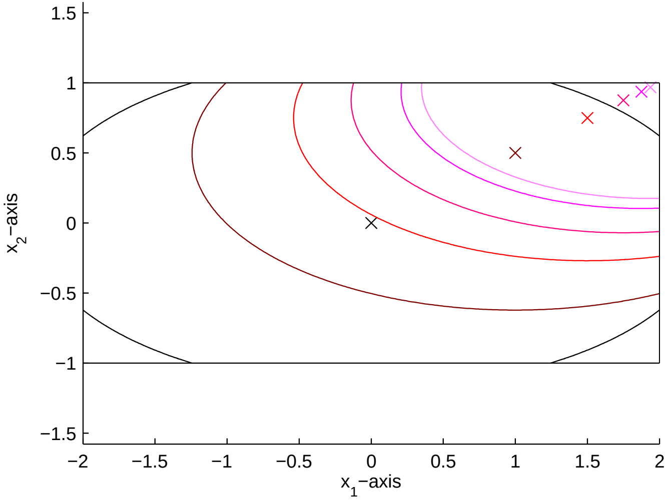

Figure: Figure showing bounding hyper-rectangle and uncertainty sets for different positions of mean. Mean and boundary of uncertainty set marked with same color.

Classification of Uncertain Datapoints

Labeling:

- Support — $y^{pr} = \text{sign}(\mathbf{w}^\top \mathbf{m}_i - b)$
- Mean — $y^{pr} = \text{sign}(\mathbf{w}^\top \mu_i - b)$
- Replicates — y^{pr} is majority label of replicates

Classification of Uncertain Datapoints

Labeling:

- Support — $y^{pr} = \text{sign}(\mathbf{w}^\top \mathbf{m}_i - b)$
- Mean — $y^{pr} = \text{sign}(\mathbf{w}^\top \mu_i - b)$
- Replicates — y^{pr} is majority label of replicates

Error Measures:

- Nominal Error
- Calculate ϵ_{opt} from Bernstein bounding

$$\mathbf{OptErr}_i = \begin{cases} 1 & \text{if } y_i \neq y_i^{pr} \\ \epsilon_{opt} & \text{if } y_i = y_i^{pr} \text{ and } \mathcal{R}(\mathbf{a}_i, \mathbf{b}_i) \text{ cuts opt. hyp.} \\ 0 & \text{else} \end{cases} \quad (2)$$

Numerical Experiments

Table: Table comparing **NomErr** (NE) and **OptErr** (OE) obtained with **IC-M, IC-R, IC-BH** and **IC-MBH**.

Data	IC-M		IC-R		IC-BH		IC-MBH	
	NE	**OE**	**NE**	**OE**	**NE**	**OE**	**NE**	**OE**
$\mathbf{10_U}$	32.07	59.90	44.80	65.70	51.05	53.62	**20.36**	**52.68**
$\mathbf{10}_\beta$	46.46	54.78	48.02	53.52	46.67	49.50	**46.18**	**49.38**
$\mathcal{A}$-$\mathcal{F}$	00.75	46.47	00.08	46.41	55.29	58.14	**00.07**	**39.68**
$\mathcal{A}$-$\mathcal{S}$	09.02	64.64	08.65	68.56	61.69	61.69	**06.10**	**39.63**
$\mathcal{A}$-$\mathcal{T}$	12.92	73.88	**07.92**	81.16	58.33	58.33	11.25	**40.84**
$\mathcal{F}$-$\mathcal{S}$	01.03	34.86	00.95	38.73	28.21	49.25	**00.05**	**27.40**
$\mathcal{F}$-$\mathcal{T}$	06.55	55.02	05.81	58.25	51.19	60.04	**05.28**	**35.07**
$\mathcal{S}$-$\mathcal{T}$	10.95	64.71	**05.00**	70.76	69.29	69.29	**05.00**	**30.71**
WDBC	55.67	**37.26**	$\times$	$\times$	**37.26**	45.82	**37.26**	**37.26**

Conclusions

- Novel methodology for interval-valued data classification under partial information.
    - Employs support as well as statistical information
    - Idea — pose the problem as CCP and relax using Bernstein bounds
- Bernstein bounds lead to less conservative noise modeling
    - Better classification margin and generalization ability
    - Empirical results show $\sim 50\%$ decrease in generalization error
- Exploitation of Bernstein bounding techniques in learning has a promise.

THANK YOU