Learning from Incomplete and Inaccurate Supervision

Zhen-Yu Zhang, Peng Zhao, Yuan Jiang and Zhi-Hua Zhou

LAMDA Group
Nanjing University, China
2019.08.07

KDD2019
Outline

• Weakly supervised learning
• Learning from Incomplete and one-sided Inaccurate Supervision (LIoIS)
  • Recover the risk (of non-noisy env.)
  • Learn importance from incomplete data
• Experiments
  • Benchmark comparison
  • Bug detection task
• Conclusions
Outline

• Weakly supervised learning

• Learning from Incomplete and one-sided Inaccurate Supervision (LIoIS)
  • Recover the risk (of non-noisy env.)
  • Learn importance from incomplete data

• Experiments
  • Benchmark comparison
  • Bug detection task

• Conclusions
Weakly supervised learning

• An example of weakly supervised learning

Typical type of weak supervision. [Z.-H. Zhou 2018]
Weakly supervised learning

• **One-sided label noise**

- buggy codes
- optimal classifier
-wrong classifier
-one-sided noisy labeled data

: labeled data
: unlabeled data

clean codes
Related works

- Inaccurate supervision learning (noisy label)
  - IW (Liu 2016), Robust NN (Gao 2018), ...

- Incomplete supervision learning (semi-supervised)
  - S4VM (Li 2015), PNU (Sakai 2017), ...

- Robust semi-supervised learning
  - LSSC (Lu 2015), SIIS (Gong 2017), ...

- Challenge
  - little work on inaccurate & incomplete supervision
    - a few noisy labeled data + unlabeled data
  - lack consideration of noise structure
Outline

• Weakly supervised learning

• Learning from Incomplete and one-sided Inaccurate Supervision (LIoIS)
  • Recover the risk (of non-noisy env.)
  • Learn importance from incomplete data

• Experiments
  • Benchmark comparison
  • Bug detection task

• Conclusions
Outline

• Weakly supervised learning

• Learning from Incomplete and one-sided Inaccurate Supervision (LIoIS)
  • Recover the risk (of non-noisy env.)
    • Learn importance from incomplete data

• Experiments
  • Benchmark comparison
  • Bug detection task

• Conclusions
Formulation

• Notations

- \( \bullet \): observed positive data
- \( \bigcirc \times \): observed negative data
- \( \square \): unlabeled data
- \( \text{optimal classifier} \)

\[\text{Learning from Incomplete and Inaccurate Supervision}\]
Formulation

• Notations

\( \tilde{P} = \{(x_1, +1), (x_2, +1), \ldots, (x_{n\tilde{P}}, +1)\} \)
Formulation

• Notations

\[ \tilde{P} = \{(x_1, +1), (x_2, +1), \ldots, (x_{n_{\tilde{P}}}, +1)\} \]

\[ \tilde{N} = \{(x_1, -1), (x_2, -1), \ldots, (x_{n_{\tilde{N}}}, -1)\} \]
Formulation

- Notations

\[ \tilde{P} = \{(x_1, +1), (x_2, +1), \ldots, (x_{n_P}, +1)\} \]

\[ \tilde{N} = \{(x_1, -1), (x_2, -1), \ldots, (x_{n_N}, -1)\} \]

\[ U = \{x_1, x_2, \ldots, x_{n_U}\} \]
Formulation

• Notations

\[ \tilde{P} = \{(x_1, +1), (x_2, +1), \ldots, (x_{n\tilde{P}}, +1)\} \]

\[ \tilde{N} = \{(x_1, -1), (x_2, -1), \ldots, (x_{n\tilde{N}}, -1)\} \]

\[ U = \{x_1, x_2, \ldots, x_{nU}\} \]

instance-dependent one-sided label noise:

\[ h(x) = \Pr[\hat{y} = -1|x, y = +1] \]
Recover the true risk

- **Risk of one-sided Inaccurate Supervision (oIS Risk)**
  - noisy labeled data

\[
R_{oIS}(g) = \theta_p \mathbb{E}_p[\sigma_+(x)\ell(g(x), +1)] + \theta_{\tilde{N}} \mathbb{E}_{\tilde{N}}[\sigma_-(x)\ell(g(x), -1)]
\]

- **Instance-dependent weight**

\[
\sigma_+(x) = \frac{1}{\text{Pr}[\hat{y} = +1|x, y = +1]}, \quad \sigma_-(x) = \text{Pr}[y = -1|x, \hat{y} = -1]
\]

**underlying true(+1) and observed as true(+1)**

**observed as false(-1) and underlying false(-1)**

- **Intuitions:**
  - the **harder** of positive data to be observed, the more important it be;
  - the **less confident** of noisy negative data, the less important it be.
Recover the true risk

• The oIS risk **equals to** the true risk (non-noise case)

\[ R_{oIS}(g) = R(g) \]

• Expected oIS risk

\[
R_{oIS}(g) = \theta \tilde{E}_P [\sigma_+(x) \ell(g(x), +1)] + \theta \tilde{E}_N [\sigma_-(x) \ell(g(x), -1)]
\]

• Empirical oIS risk

\[
\hat{R}_{oIS}(g) = \frac{\theta \tilde{P}}{n \tilde{P}} \sum_{i=1}^{n \tilde{P}} \sigma_+(x_i) \ell(g(x_i), +1) + \frac{\theta \tilde{N}}{n \tilde{N}} \sum_{j=1}^{n \tilde{N}} \sigma_-(x_j) \ell(g(x_j), -1),
\]
Theoretical analysis (oIS Risk)

• Excess risk analysis

**Theorem 1** (Excess Risk of learning from oIS). Assume that the loss function $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_+$ is non-negative and $L$-Lipschitz continuous. Given that hardness $h(x) \in [0, h]$, then, for any $\delta > 0$, with probability at least $1 - \delta$, we have

$$R(\widehat{g}_{oIS}) - R(g^*) \leq \frac{4\theta_P (1 - h)^{-1}}{n} \sum L \mathbb{R}_n \widehat{P}(G) + \frac{4\theta_N L \mathbb{R}_n \tilde{N}(G)}{n} + \sqrt{\frac{\ln(4/\delta)}{2n \widehat{P}}} \sqrt{\frac{\ln(4/\delta)}{2n \tilde{N}}} + 0(1/\sqrt{n})$$

introduced by incomplete P data

introduced by noisy N data

• Remarks
  • Recover the result of *instance-independent* label noise setting
  • Reflect the **hardness** ($h$) of extremely noisy examples
Recover the true risk

• Risk of incomplete supervision
  • unlabeled data

\[ R(g) = 2\theta_P \mathbb{E}_P [\ell(g(x), +1)] + \mathbb{E}_U [\ell(g(x), -1)] - \theta_P \]

• The PU risk equals to the true risk (non-noise case)

\[ R_{PU}(g) = R(g) \]

[du Plessis 2014]

• Empirical PU risk

\[ \hat{R}_{PU}(g) = 2\frac{\theta_P}{n_P} \sum_{i=1}^{n_P} \ell(g(x_i), +1) + \frac{1}{n_U} \sum_{k=1}^{n_U} \ell(g(x_k), -1) \]
Recover the true risk

Problem setting

instance-dependent label noise

inaccurate supervision: **oIS risk**

\[
\hat{\theta}_o \frac{1}{n_o} \sum_{i=1}^{n_o} \sigma_+(x_i)\ell(g(x_i), +1) + \frac{1 - \hat{\theta}_o}{n_o} \sum_{j=1}^{n_o} \sigma_-(x_j)\ell(g(x_j), -1)
\]

little labeled data and unlabeled data

incomplete supervision: **PU risk**

\[
\frac{2\hat{\theta}_p}{n_p} \sum_{i=1}^{n_p} \ell(g(x_i), +1) + \frac{1}{n_U} \sum_{k=1}^{n_U} \ell(g(x_k), -1)
\]

Weighted combination

\[
\hat{R}_{LOIS}(g) = y \left\{ \frac{2\hat{\theta}_p}{n_p} \sum_{i=1}^{n_p} \ell(g(x_i), +1) + \frac{1}{n_U} \sum_{k=1}^{n_U} \ell(g(x_k), -1) \right\}
\]

\[
(1 - y) \left\{ \frac{\hat{\theta}_o}{n_o} \sum_{i=1}^{n_o} \sigma_+(x_i)\ell(g(x_i), +1) + \frac{1 - \hat{\theta}_o}{n_o} \sum_{j=1}^{n_o} \sigma_-(x_j)\ell(g(x_j), -1) \right\}
\]

Weights from labeled data? No! limited number!
Outline

• Weakly supervised learning

• Learning from Incomplete and one-sided Inaccurate Supervision (LIoIS)
  • Recover the risk (of non-noisy env.)
  • Learn importance from incomplete data

• Experiments
  • Benchmark comparison
  • Bug detection task

• Conclusions
Learn importance from incomplete data

• Review the importance (weight) in oIS Risk

\[ R_{oIS}(g) = \theta_P \mathbb{E}_P[\sigma_+(x) \ell(g(x), +1)] + \theta_N \mathbb{E}_N[\sigma_-(x) \ell(g(x), -1)], \]

\[ \sigma_+(x) = \frac{\Pr[x, y = +1]}{\Pr[x, \hat{y} = +1]} = \frac{\theta_P \Pr[x|y = +1]}{\theta_P \Pr[x|\hat{y} = +1]} \]

• Ratio estimation with the help of unlabeled data
  • empirical Bregman divergence minimization [Sugiyama 2010]

\[ \hat{B}_f(\sigma_+||\hat{\sigma}_+) = \frac{1}{n^\hat{P}} \sum_{i=1}^{n^\hat{P}} \nabla f(\hat{\sigma}_+(x_i)) \hat{\sigma}_+(x_i) - \frac{1}{n^\hat{P}} \sum_{i=1}^{n^\hat{P}} f(\hat{\sigma}_+(x_i)) \]

estimate by noisy positive group

\[ -\frac{1}{m} \sum_{j=1}^{m} \nabla f(\hat{\sigma}_+(x_j)) \]

estimate by true positive group
Learn importance from incomplete data

• Ratio estimation with the help of unlabeled data
  • empirical Bregman divergence minimization [Sugiyama 2010]

\[
\hat{B}_f(\sigma_{+r}||\hat{\sigma}_{+r}) = \frac{1}{n_P} \sum_{i=1}^{n_P} \nabla f(\hat{\sigma}_{+r}(x_i))\hat{\sigma}_{+r}(x_i) - \frac{1}{n_P} \sum_{i=1}^{n_P} f(\hat{\sigma}_{+r}(x_i)) - \frac{1}{m} \sum_{j=1}^{m} \nabla f(\hat{\sigma}_{+r}(x_j)).
\]

estimate by noisy positive group
estimate by true positive group

• Estimate the true positive group
  • pseudo labels provided by pre-trained PU classifier

\[
P_{PU} = \{(x,+1)|g_{PU}(x) = +1\}
\]
\[
N_{PU} = \{(x,-1)|g_{PU}(x) = -1\}
\]

with very limited positive data, the direct classification is risky
Learning from Incomplete and one-sided Inaccurate Supervision (LIoIS)

Problem setting

instance-dependent label noise

inaccurate supervision: oIS risk

\[ \frac{\hat{\theta}_2}{n \tilde{P}_2} \sum_{i=1}^{n \tilde{P}_2} \sigma_+(x_i) \ell(g(x_i), +1) + \frac{1 - \hat{\theta}_2}{n \tilde{N}} \sum_{j=1}^{n \tilde{N}} \sigma_-(x_j) \ell(g(x_j), -1) \]

little labeled data and unlabeled data

incomplete supervision: PU risk

\[ \frac{2\hat{\theta}_1}{n \tilde{P}_1} \sum_{i=1}^{n \tilde{P}_1} \ell(g(x_i), +1) + \frac{1}{n_U} \sum_{k=1}^{n_U} \ell(g(x_k), -1) \]

estimated with incomplete supervision

Weighted combination

\[ \tilde{R}_{LIoIS}(g) = \gamma \left\{ \frac{2\hat{\theta}_1}{n \tilde{P}_1} \sum_{i=1}^{n \tilde{P}_1} \ell(g(x_i), +1) + \frac{1}{n_U} \sum_{k=1}^{n_U} \ell(g(x_k), -1) \right\} + (1 - \gamma) \left\{ \frac{\hat{\theta}_2}{n \tilde{P}_2} \sum_{i=1}^{n \tilde{P}_2} \sigma_+(x_i) \ell(g(x_i), +1) + \frac{1 - \hat{\theta}_2}{n \tilde{N}} \sum_{j=1}^{n \tilde{N}} \sigma_-(x_j) \ell(g(x_j), -1) \right\} \]

Theoretical analysis

\[ R(\hat{g}_{LIoIS}) - R(g^*) = O\left(\frac{1}{\sqrt{n \tilde{P}}} + \frac{1}{\sqrt{n \tilde{N}}} + \frac{1}{\sqrt{n_U}}\right) \]

optimal convergence rate
Outline

• Weakly supervised learning

• Learning from Incomplete and one-sided Inaccurate Supervision (LIoIS)
  • Recover the risk (of non-noisy env.)
  • Learn importance from incomplete data

• Experiments
  • Benchmark comparison
  • Bug detection task

• Conclusions
Experiments

• Compared method
  • non-robust/robust semi-supervised learning
  • noise robust learning (w/o unlabeled data)
  • PU learning (w/o noisy data)

• Datasets
  • UCI benchmark datasets
  • Bug detection datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th># instance</th>
<th># dim</th>
<th>Dataset</th>
<th># instance</th>
<th># dim</th>
</tr>
</thead>
<tbody>
<tr>
<td>house</td>
<td>232</td>
<td>16</td>
<td>australian</td>
<td>690</td>
<td>42</td>
</tr>
<tr>
<td>ionosphere</td>
<td>351</td>
<td>33</td>
<td>diabetes</td>
<td>768</td>
<td>8</td>
</tr>
<tr>
<td>clean1</td>
<td>476</td>
<td>166</td>
<td>german</td>
<td>1000</td>
<td>59</td>
</tr>
<tr>
<td>wdbc</td>
<td>569</td>
<td>14</td>
<td>letter7vs9</td>
<td>1528</td>
<td>16</td>
</tr>
<tr>
<td>islet</td>
<td>600</td>
<td>51</td>
<td>a5a</td>
<td>6414</td>
<td>122</td>
</tr>
<tr>
<td>breastw</td>
<td>683</td>
<td>9</td>
<td>mnist7vs9</td>
<td>14251</td>
<td>600</td>
</tr>
</tbody>
</table>

Brief statistics of benchmark datasets
Outline

- Weakly supervised learning
- Learning from Incomplete and one-sided Inaccurate Supervision (LIoIS)
  - Recover the risk (of non-noisy env.)
  - Learn importance from incomplete data
- Experiments
  - Benchmark comparison
    - Bug detection task
- Conclusions
## Benchmark comparison

- Comparison with baseline algorithms

<table>
<thead>
<tr>
<th>Dataset</th>
<th>LIBSVM</th>
<th>IW</th>
<th>S4VM</th>
<th>LSSC</th>
<th>ROSSEL</th>
<th>SIIS</th>
<th>PUIW</th>
<th>LioIS (ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>house</td>
<td>91.90 ± 1.72</td>
<td>96.29 ± 1.27</td>
<td>93.33 ± 1.37</td>
<td>93.49 ± 2.17</td>
<td>93.26 ± 1.88</td>
<td>88.84 ± 2.70</td>
<td>94.53 ± 2.80</td>
<td>96.03 ± 0.97</td>
</tr>
<tr>
<td>ionosphere</td>
<td>81.54 ± 3.19</td>
<td>83.28 ± 6.51</td>
<td>79.34 ± 7.07</td>
<td>79.26 ± 6.88</td>
<td>88.23 ± 4.64</td>
<td>72.11 ± 16.6</td>
<td>85.69 ± 2.56</td>
<td>90.23 ± 7.43</td>
</tr>
<tr>
<td>clean1</td>
<td>72.84 ± 3.81</td>
<td>64.52 ± 4.15</td>
<td>78.01 ± 3.62</td>
<td>61.03 ± 1.00</td>
<td>77.40 ± 2.37</td>
<td>60.89 ± 5.98</td>
<td>71.07 ± 3.92</td>
<td>86.16 ± 3.42</td>
</tr>
<tr>
<td>wdbc</td>
<td>89.65 ± 2.75</td>
<td>77.47 ± 19.3</td>
<td>80.76 ± 7.28</td>
<td>91.97 ± 2.01</td>
<td>90.87 ± 2.07</td>
<td>92.95 ± 1.32</td>
<td>77.64 ± 12.1</td>
<td>95.52 ± 1.08</td>
</tr>
<tr>
<td>isolei</td>
<td>86.50 ± 2.16</td>
<td>91.63 ± 2.44</td>
<td>87.34 ± 2.99</td>
<td>96.48 ± 1.25</td>
<td>80.13 ± 2.06</td>
<td>98.82 ± 0.48</td>
<td>91.74 ± 2.38</td>
<td>98.61 ± 1.21</td>
</tr>
<tr>
<td>breastw</td>
<td>93.65 ± 1.98</td>
<td>94.71 ± 1.23</td>
<td>91.54 ± 1.69</td>
<td>96.21 ± 1.29</td>
<td>96.53 ± 0.83</td>
<td>96.49 ± 0.68</td>
<td>95.53 ± 1.52</td>
<td>94.85 ± 4.32</td>
</tr>
<tr>
<td>australian</td>
<td>80.20 ± 3.24</td>
<td>80.65 ± 12.9</td>
<td>82.81 ± 3.19</td>
<td>81.87 ± 2.81</td>
<td>79.89 ± 6.92</td>
<td>72.96 ± 3.50</td>
<td>84.47 ± 3.90</td>
<td>86.19 ± 1.05</td>
</tr>
<tr>
<td>diabetes</td>
<td>74.91 ± 1.50</td>
<td>60.79 ± 14.1</td>
<td>69.69 ± 3.71</td>
<td>68.45 ± 2.35</td>
<td>75.69 ± 2.48</td>
<td>67.92 ± 1.37</td>
<td>75.93 ± 2.35</td>
<td>76.26 ± 1.03</td>
</tr>
<tr>
<td>german</td>
<td>64.52 ± 3.89</td>
<td>67.45 ± 4.81</td>
<td>65.81 ± 2.30</td>
<td>62.53 ± 1.86</td>
<td>73.03 ± 0.96</td>
<td>72.24 ± 1.19</td>
<td>68.65 ± 2.21</td>
<td>74.37 ± 2.58</td>
</tr>
<tr>
<td>letter7vs9</td>
<td>90.04 ± 3.88</td>
<td>95.21 ± 1.72</td>
<td>92.45 ± 4.65</td>
<td>94.23 ± 1.68</td>
<td>94.94 ± 1.43</td>
<td>78.47 ± 1.39</td>
<td>95.04 ± 1.34</td>
<td>98.82 ± 0.95</td>
</tr>
<tr>
<td>a5a</td>
<td>70.91 ± 2.42</td>
<td>73.82 ± 4.35</td>
<td>72.29 ± 2.71</td>
<td>68.45 ± 1.69</td>
<td>79.36 ± 1.66</td>
<td>76.36 ± 0.82</td>
<td>74.13 ± 2.47</td>
<td>83.29 ± 0.47</td>
</tr>
<tr>
<td>mnist7vs9</td>
<td>85.63 ± 2.29</td>
<td>90.18 ± 1.62</td>
<td>86.69 ± 1.43</td>
<td>88.76 ± 1.43</td>
<td>81.41 ± 1.18</td>
<td>– ± –</td>
<td>91.82 ± 1.53</td>
<td>96.19 ± 0.33</td>
</tr>
</tbody>
</table>

LioIS w/ t/ l: 11/ 1/ 0 | 10/ 1/ 1 | 12/ 0/ 0 | 10/ 1/ 1 | 9/ 2/ 1 | 9/ 1/ 2 | 8/ 3/ 1 | rank first 9/ 12

**robust supervised**

**robust semi-supervised**

Rank first 9/12

Average accuracy on 12 UCI benchmark datasets.
Benchmark comparison

- Outperforms the Incomplete Supervision algorithms
  - four semi-supervised learning algorithms

Comparison with semi-supervised learning algorithms.
Benchmark comparison

- Outperforms the Inaccurate Supervision algorithms
  - noisy label learning algorithm and direct combination

Performance curve (in accuracy) of proposed approach with an increasing noise rate of negative data.
Outline

• Weakly supervised learning
• Learning from Incomplete and one-sided Inaccurate Supervision (LlIoIS)
  • Recover the risk (of non-noisy env.)
  • Learn importance from incomplete data
• Experiments
  • Benchmark comparison
  • Bug detection task
• Conclusions
Bug detection task

• Training data:
  • buggy code
  • checked clean code
  • unlabeled code

• Main purpose:
  • identify the buggy code

• Performance measure
  • recall@k:
    \[ |S@k \cap B| \]
    top k detected bugs \hspace{1cm} ground-truth bugs
Bug detection task

(a) OrientDB (File) 
(b) OrientDB (class)

Performance curve w.r.t. an increasing number of detected bugs.
Bug detection task

Performance curve w.r.t. an increasing number of detected bugs.

(c) Elasticsearch (File)

(d) Elasticsearch (class)
Bug detection task

• Furthermore: identify potential buggy code
  • two identified noisy negative data

Figure 5: Potential bugs detected by LIoIS in OrientDB and Elasticsearch datasets.
Outline

• Weakly supervised learning

• Learning from Incomplete and one-sided Inaccurate Supervision (LIoIS)
  • Recover the risk (of non-noisy env.)
  • Learn importance from incomplete data

• Experiments
  • Benchmark comparison
  • Bug detection task

• Conclusions
Conclusions

• We propose LIoIS to learn from inaccurate and incomplete supervision.

• We provide the excess risk bound analysis for proposed method LIoIS.

• Experiments show the effectiveness of LIoIS.
Q & A

Thanks!