SUPERSET LEARNING AND DATA IMPRECISATION

Eyke Hüllermeier
Intelligent Systems Group
Department of Computer Science
University of Paderborn, Germany

eyke@upb.de
OUTLINE

PART 1
Superset learning

PART 2
Optimistic loss minimization

PART 3
Data imprecisiation

What it is about ....

A general approach to superset learning ....

Using superset learning for weighted learning ...
SUPERSET LEARNING

... is a specific type of **weakly supervised learning**, studied under different names in machine learning:

- *learning from partial labels*
- *multiple label learning*
- *learning from ambiguously labeled examples*
- ...

... also connected to learning from **coarse data** in statistics (Rubin, 1976; Heitjan and Rubin, 1991), missing values, data augmentation (Tanner and Wong, 2012).
Consider a standard setting of **supervised learning** with instance space $\mathcal{X}$, output space $\mathcal{Y}$, and hypothesis space $\mathcal{H}$.

Output values $y_n \in \mathcal{Y}$ associated with training instances $x_n$, $n = 1, \ldots, N$, are not necessarily observed precisely but only characterised in terms of **supersets**

$$Y_n \ni y_n.$$ 

Set of imprecise/ambiguous/coarse observations is denoted

$$\mathcal{O} = \{(x_1, Y_1), \ldots, (x_N, Y_N)\}$$

An **instantiation** of $\mathcal{O}$, denoted $\mathcal{D}$, is obtained by replacing each $Y_n$ with a candidate $y_n \in Y_n$. 

EXAMPLE: CLASSIFICATION

Classes

[Diagram showing scatter plot with classes represented by different colors]
EXAMPLE: CLASSIFICATION

Classes

one of many instantiations
EXAMPLE: REGRESSION

one of infinitely many instantiations

$y$  

$x$
How to learn from (super)set-valued data?

We suggest that successful learning should go hand in hand with data disambiguation, i.e., finding out about the (precise) $y_n$ underlying the imprecise observations $\bar{Y}_n$ ...
DATA DISAMBIGUATION

Classes
DATA DISAMBIGUATION

Classes
DATA DISAMBIGUATION
DATA DISAMBIGUATION
A plausible instantiation that can be fitted reasonably well with a **LINEAR** model!

A less plausible instantiation, because there is no **LINEAR** model with a good fit!
A plausible instantiation that can be fitted quite well with a **QUADRATIC** model!

It all depends on how you look at the data!
\[ \mathcal{O} = \{ \bigcirc, \bigcirc \} \]

`assume both class distributions to be Gaussian`
DATA DISAMBIGUATION

\( \mathbf{O} = \{ \mathbb{O}, \mathbb{O} \} \)

plausible instantiation

quadratic discriminant

assume both class distributions to be Gaussian
\[
\mathcal{O} = \{ \circ, \bigcirc \}
\]

Data disambiguation

Assume both class distributions to be Gaussian

Implausible instantiation
Model identification and data disambiguation should be performed simultaneously:

\[ P(h, D) = P(h) P(D | h) = P(D) P(h | D) \]
OUTLINE

PART 1
Superset learning

PART 2
Optimistic loss minimization

PART 3
Data imprecision
Likelihood of a model \( h \in \mathcal{H} \):

\[
\ell(h) = P(O, D \mid h) = P(D \mid h) P(O \mid D, h) = P(D \mid h) P(O \mid D)
\]

Imprecise observation only depends on true data, not on the model.
Imprecise data is a superset, but no other assumption.
We derive a principle of **generalized empirical risk minimization** with the empirical risk

\[ \mathcal{R}_{emp}(h) = \frac{1}{N} \sum_{n=1}^{N} L^*(Y_n, h(x_n)) \]

and the **optimistic superset loss** (OSL) function

\[ L^*(Y, \hat{y}) = \min \{ L(y, \hat{y}) \mid y \in Y \} . \]

how well the (precise) model fits the imprecise data
The $\epsilon$-insensitive loss $L(y, \hat{y}) = \max(|y - \hat{y}| - \epsilon, 0)$ used in support vector regression corresponds to $L^*$ with $L$ the standard $L_1$ loss $L(y, \hat{y}) = |y - \hat{y}|$ and precise data $y_n$ being replaced by interval-valued data $Y_n = [y_n - \epsilon, y_n + \epsilon]$. 
GENERALIZATION TO FUZZY DATA

- $Y_n$ is a **subset** of $\mathcal{Y}$ (with characteristic function $\mathcal{Y} \rightarrow \{0, 1\}$)
- $Y_n$ is a **fuzzy subset** of $\mathcal{Y}$, characterized in terms of a membership function $\mathcal{Y} \rightarrow [0, 1]$

![Interval](image1.png)

**interval**

![Fuzzy Interval](image2.png)

**fuzzy interval**
GENERALIZATION TO FUZZY DATA

\[ L^{**}(Y, \hat{y}) = \int_0^1 L^*([Y]_\alpha, \hat{y}) \, d\alpha \]

\( \alpha \)-cut

LOSS
GENERALIZATION TO FUZZY DATA

\[ L^{**}(Y, \hat{y}) = \int_0^1 L^\star([Y]_\alpha, \hat{y}) \, d\alpha \]

\[ R_{\text{emp}}(h) = \frac{1}{N} \sum_{n=1}^{N} L^{**}(Y_n, h(x_n)) \]
GENERALIZATION TO FUZZY DATA

\( L^{**}(Y, \hat{y}) \)

→ Huber loss!
GENERALIZATION TO FUZZY DATA

→ (generalized) Huber loss!
Superset learning naturally applies to learning problems with *structured outputs*, which are often only *partially specified* and can then be associated with the *set of all consistent completions*. 
... is the problem to learn a model that maps instances to **TOTAL ORDERS**
over a fixed set of alternatives/labels:

\[ \mathcal{Y} = \{ABCD, ABDC, \ldots, DCBA\} \]
LABEL RANKING

... is the problem to learn a model that maps instances to **TOTAL ORDERS** over a fixed set of alternatives/labels:

\[ (0, 37, 46, 325, 1, 0) \]

... *likes more*
... *reads more*
... *recommends more*
...
... is the problem to learn a model that maps instances to **TOTAL ORDERS** over a fixed set of alternatives/labels:

(0, 37, 46, 325, 1, 0)

*Training data is typically incomplete!*
... is the problem to learn a model that maps instances to **TOTAL ORDERS** over a fixed set of alternatives/labels:

\[
\begin{align*}
A & > C > B > D \\
A & > C > D > B \\
A & > B > C > D \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
D & > B > A > C
\end{align*}
\]

(set of linear extensions)

**Training data is typically incomplete!**
**LABEL RANKING LOSSES**

**KENDALL**

\[
L(\pi, \pi^*) = \sum_{1 \leq i < j \leq M} \left[ (\pi(i) - \pi(j))(\pi^*(i) - \pi^*(j)) < 0 \right]
\]

**SPEARMAN**

\[
L(\pi, \pi^*) = \sum_{1 \leq i \leq M} |\pi^*(i) - \pi(i)|
\]
- Two missing label scenarios: missing at random, top-rank
- General conclusion: more robust toward incompleteness
OUTLINE

PART 1
Superset learning

PART 2
Optimistic loss minimization

PART 3
Data imprecisiation
DATA IMPRECISIATION

So far:
*Observations are imprecise/incomplete, and we have to deal with that!*

Now:
*Deliberately turn precise into imprecise data, so as to modulate the influence of an observation on the learning process!*

Motivated by the following observation:

\[(Y \subset Y') \Rightarrow (L^*(Y, \cdot) \geq L^*(Y', \cdot))\]
EXAMPLE WEIGHING
EXAMPLE WEIGHING

minimize \( \sum_{i=1}^{n} w_i \cdot (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 \)
We suggest an alternative way of weighing examples, namely, via „data imprecisiation“ ...
EXAMPLE WEIGHING

weighing through „imprecisiation“

\[
\text{minimize } \sum_{n=1}^{N} L^*_2 (Y_n, w^\top x_n)
\]
Different ways of (individually) discounting the loss function.

In (Lu and H., 2015), we empirically compared standard locally weighted linear regression with this approach and essentially found no difference.
We suggest an alternative way of weighing examples, namely, via „data imprecisiation“ ...
FUZZY MARGIN LOSSES

GENERALIZED HINGE LOSS

$w = 1$

$w = \frac{3}{4}$

$w = \frac{1}{2}$

$w = \frac{1}{4}$

$w = 0$
FUZZY MARGIN LOSSES

Different ways of (individually) discounting the loss function.
Semi-supervised learning with SVMs: Consider unlabeled data as instances labeled with the superset \([-1, +1]\). The generalized loss \(L^*\) with \(L\) the standard hinge loss then corresponds to the (non-convex) “hat loss”.
DATA DISAMBIGUATION
Robust loss minimization techniques:

- **Robust truncated-hinge-loss support vector machines (RSVM)** trains SVMs with the truncated version of the hinge loss in order to be more robust toward outliers and noisy data (Wu and Liu, 2007).

- **One-step weighted SVM (OWSVM)** first trains a standard SVM. Then, it weights each training example based on its distance to the decision boundary and retracts using the weighted hinge loss (Wu and Liu, 2013).

- **Our approach (FLSVM)** is the same as OWSVM, except for the weighted loss: instead of using a simple weighting of the hinge loss, we use the optimistic fuzzy loss.

Non-convex optimization problem solved by concave-convex procedure (Yuille and Rangaraja, 2002).
Table 1: Experimental results: Average misclassification rate on test data (with standard deviation) for different methods, data sets, and noise levels.

<table>
<thead>
<tr>
<th>perc</th>
<th>data sets</th>
<th>SVM</th>
<th>OWSVM</th>
<th>RSVM</th>
<th>FLSVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>Wdbc</td>
<td>0.0281 (0.0114)</td>
<td>0.0263 (0.0087)</td>
<td><strong>0.0228</strong> (0.0100)</td>
<td>0.0374 (0.0159)</td>
</tr>
<tr>
<td></td>
<td>Bupa</td>
<td>0.3188 (0.0928)</td>
<td>0.3043 (0.0774)</td>
<td><strong>0.3072</strong> (0.0776)</td>
<td>0.3188 (0.0934)</td>
</tr>
<tr>
<td></td>
<td>Banknote</td>
<td>0.0153 (0.0110)</td>
<td><strong>0.0095</strong> (0.0050)</td>
<td>0.0153 (0.0083)</td>
<td>0.0124 (0.0101)</td>
</tr>
<tr>
<td></td>
<td>Parkinsons</td>
<td>0.1333 (0.0334)</td>
<td>0.1128 (0.0292)</td>
<td><strong>0.1077</strong> (0.0215)</td>
<td>0.1077 (0.0215)</td>
</tr>
<tr>
<td>10%</td>
<td>Wdbc</td>
<td>0.0387 (0.0133)</td>
<td><strong>0.0281</strong> (0.0144)</td>
<td>0.0316 (0.0146)</td>
<td>0.0334 (0.0130)</td>
</tr>
<tr>
<td></td>
<td>Bupa</td>
<td>0.3391 (0.0442)</td>
<td><strong>0.3304</strong> (0.0527)</td>
<td>0.3159 (0.0635)</td>
<td>0.3159 (0.0720)</td>
</tr>
<tr>
<td></td>
<td>Banknote</td>
<td>0.0233 (0.0063)</td>
<td>0.0168 (0.0110)</td>
<td><strong>0.0146</strong> (0.0068)</td>
<td>0.0131 (0.0095)</td>
</tr>
<tr>
<td></td>
<td>Parkinsons</td>
<td>0.1385 (0.0229)</td>
<td>0.1231 (0.0215)</td>
<td>0.1231 (0.0215)</td>
<td><strong>0.1179</strong> (0.0215)</td>
</tr>
<tr>
<td>20%</td>
<td>Wdbc</td>
<td>0.0615 (0.0124)</td>
<td>0.0474 (0.0100)</td>
<td><strong>0.0386</strong> (0.0171)</td>
<td>0.0422 (0.0209)</td>
</tr>
<tr>
<td></td>
<td>Bupa</td>
<td>0.3855 (0.0364)</td>
<td>0.3478 (0.0369)</td>
<td><strong>0.3275</strong> (0.0873)</td>
<td>0.3362 (0.0601)</td>
</tr>
<tr>
<td></td>
<td>Banknote</td>
<td>0.0241 (0.0050)</td>
<td>0.0248 (0.0056)</td>
<td>0.0211 (0.0125)</td>
<td><strong>0.0175</strong> (0.0075)</td>
</tr>
<tr>
<td></td>
<td>Parkinsons</td>
<td>0.1385 (0.0466)</td>
<td>0.1333 (0.0493)</td>
<td><strong>0.1279</strong> (0.0429)</td>
<td>0.1436 (0.0493)</td>
</tr>
<tr>
<td>30%</td>
<td>Wdbc</td>
<td>0.0791 (0.0270)</td>
<td>0.0633 (0.0245)</td>
<td>0.0633 (0.0314)</td>
<td><strong>0.0580</strong> (0.0302)</td>
</tr>
<tr>
<td></td>
<td>Bupa</td>
<td>0.3884 (0.0854)</td>
<td><strong>0.3710</strong> (0.0861)</td>
<td>0.3826 (0.1033)</td>
<td>0.3710 (0.1018)</td>
</tr>
<tr>
<td></td>
<td>Banknote</td>
<td>0.0313 (0.0110)</td>
<td>0.0277 (0.0077)</td>
<td>0.0270 (0.0092)</td>
<td><strong>0.0255</strong> (0.0215)</td>
</tr>
<tr>
<td></td>
<td>Parkinsons</td>
<td>0.1846 (0.0459)</td>
<td>0.1897 (0.0693)</td>
<td>0.1846 (0.0712)</td>
<td><strong>0.1692</strong> (0.0716)</td>
</tr>
</tbody>
</table>
Under what conditions is (successful) learning in the superset setting actually possible?
systematic imprecisiation
THEORETICAL FOUNDATIONS

non-systematic imprecisiation
Liu and Dietterich (2014) consider the **ambiguity degree**, which is defined as the largest probability that a particular **distractor** label co-occurs with the true label in multi-class classification:

\[
\gamma = \sup \left\{ P_{Y \sim D^s(x,y)}(\ell \in Y) \mid (x, y) \in \mathcal{X} \times \mathcal{Y}, \ell \in \mathcal{Y}, p(x, y) > 0, \ell \neq y \right\}
\]
THEORETICAL FOUNDATIONS

Let $\theta = \log(2/(1 + \gamma))$ and $d_\mathcal{H}$ the Natarajan dimension of $\mathcal{H}$. Define

$$n_0(\mathcal{H}, \epsilon, \delta) = \frac{4}{\theta \epsilon} \left( d_\mathcal{H} \left( \log(4d_\mathcal{H} + 2 \log L + \log \left( \frac{1}{\theta \epsilon} \right)) + \log \left( \frac{1}{\delta} \right) + 1 \right) \right).$$

Then, in the realizable case, with probability at least $1 - \delta$, the model with the smallest empirical superset loss on a set of training data of size $n > n_0(\mathcal{H}, \epsilon, \delta)$ has a generalisation error of at most $\epsilon$. 
The **balanced benefit condition**: 

\[
0 \leq \eta_1 \leq \inf_{h \in \mathcal{H}} \frac{\mathcal{R}^S(h)}{\mathcal{R}(h)} \leq \sup_{h \in \mathcal{H}} \frac{\mathcal{R}^S(h)}{\mathcal{R}(h)} \leq \eta_2 \leq 1,
\]

where \( \mathcal{R}^S(h) \) is the expected superset loss of \( h \).

For sufficiently large sample size, 

\[
\mathcal{R}(\hat{h}) \leq \mathcal{R}(h^*) + \Delta(d_{\mathcal{H}}, \epsilon, \delta, \eta_1, \eta_2),
\]

with probability \( 1 - \delta \), where \( h^* \) is the Bayes predictor and \( \hat{h} \) the empirical (superset) risk minimizer; in general, \( \Delta \) cannot be made arbitrarily small.
SUMMARY AND OUTLOOK

- Method for **superset learning** based on **optimistic loss minimization**, performing simultaneous model identification and data disambiguation.
- Our framework covers several **existing methods** as special cases but also supports the systematic development of **new methods**.
- **Completely generic principle** (classification, regression, structured output prediction, ...)
- Example weighing via **data imprecisiation** (→ „modeling data“)
- Works for regression and classification, but seems to be even more interesting for other problems, including ranking, transfer learning, ...
- **More future work**: Algorithmic solutions for specific instantiations of our framework, theoretical foundations.


S. Lu and E. Hüllermeier. Locally Weighted Regression through Data Imprecisiation. Workshop Computational Intelligence, Dortmund, 2015.
