Work In Progress Talk: Levelwise Data Disambiguation by Cautious Superset Classification

Julian Rodemann [1] Dominik Kreiss [1] Eyke Hüllermeier [2]
Thomas Augustin [1]

[1]Dep. of Statistics, LMU Munich

[2]Dep. of Computer Science, LMU Munich

July 9, 2022

Contents

Contents

Optimistic Superset Learning

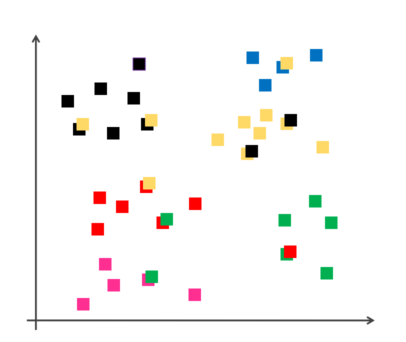


Figure: Partly ambiguous data.

Optimistic Superset Learning

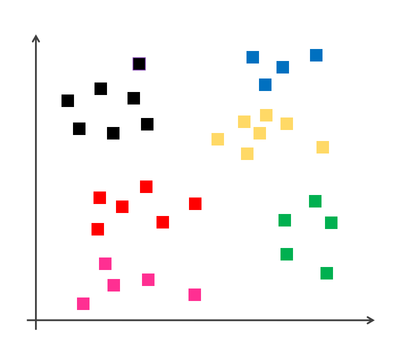


Figure: A "plausible" instantiation...

Image credits: Eyke Hüllermeier

Optimistic Superset Learning

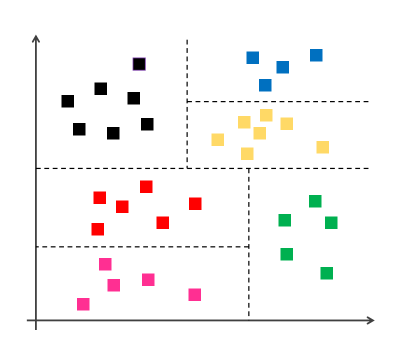


Figure: ...that can be well-explained by a fitted model.

Optimistic Superset Learning

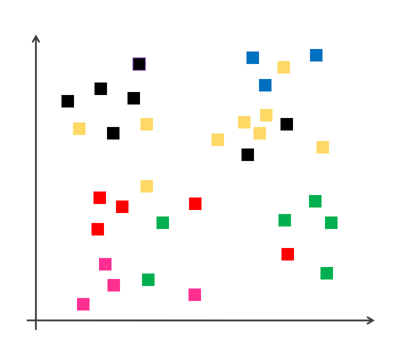


Figure: While a less "plausible" instantiation...

Optimistic Superset Learning

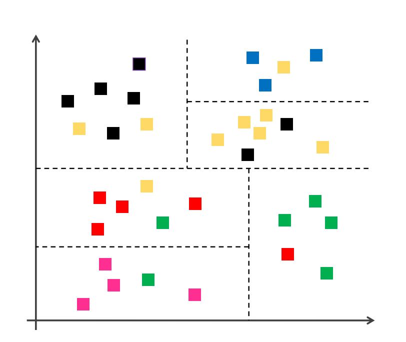


Figure: ...results in a worse performing model.

Optimistic Superset Learning

- [Hüllermeier, 2014] introduced *Optimistic Superset Loss*[1]

$$L_{opt}(\hat{y}_i, Y_i) = \min_{y \in Y_i} L(\hat{y}_i, y), \tag{1}$$

  with $L(\cdot)$ a loss function $L : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$.
- Minimizing the corresponding empirical risk is called Optimistic Superset Learning (OSL).

[1] See also [Hüllermeier and Cheng, 2015], [Hüllermeier et al., 2019] and [Lienen and Hüllermeier, 2021]

Contents

Contents

Setup

- Motivation: Find a singleton representation of set-valued data
- Consider the observations $\mathcal{O} = \{(x_i, Y_i)\}_{i=1}^n \in \left(\mathcal{X} \times 2^{\mathcal{Y}}\right)^n$ with categorical $\mathcal{Y}$.
- Y_i is regarded a superset of a true underlying singleton $y_i \in \mathcal{Y}$.
- Let $\mathbf{Y} = Y_1 \times Y_2 \times \cdots \times Y_n$ be the Cartesian product of the observed supersets; denote the number of different observed categories by q.[2]
- Any singleton vector $\mathbf{y} = (y_1, \ldots, y_i, \ldots, y_n)' \in \mathbf{Y}$ is called an *instantiation* of the observed set-valued data.

[2] Notably, $q \leq |\mathcal{Y}|$.

Contents

Cautious Superset Classification

- For each $\mathbf{y} \in \mathbf{Y}$, we find $\hat{\mathbf{y}}^{(\mathbf{h},\mathbf{y})}(\mathbf{x})$ by empirical risk minimization.
- We evaluate the so trained model $\hat{\mathbf{y}}^{(\mathbf{h},\mathbf{y})}(\mathbf{x})$ by its empirical risk

$$\mathcal{R}_{emp}(\mathbf{h}, \mathbf{x}, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^{n} L(\hat{y}_i^{(\mathbf{h},\mathbf{y})}(x_i), y_i), \hat{y}_i \in \hat{\mathbf{y}}^{(\mathbf{h},\mathbf{y})}(\mathbf{x}), y_i \in \mathbf{y}, x_i \in \mathbf{x},$$

  $L(\cdot)$ again a loss function $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$.
- We then consider

$$\mathbf{y}^*_{\mathcal{R}_{emp}} = \arg\min_{\mathbf{y} \in \mathbf{Y}} \mathcal{R}_{emp}(\mathbf{h}, \mathbf{x}, \mathbf{y}) \tag{2}$$

  the most plausible instantiation, given a model, a loss function, and the (singleton) covariates.

Cautious Superset Classification

- Note that in contrast to Optimistic Superset Learning [Hüllermeier, 2014], equation (2) requires estimating q^n models.

  $\Longrightarrow$ Restrictions on $\mathbf{Y}$ and/or $2^{\mathcal{Y}}$ needed, e.g. clustering and homogenous treatment of clusters

Contents

Narrowing Down Supersets

- Consider the 0/1-loss

$$L(\hat{y}_i^{(\mathbf{h},\mathbf{y})}(x_i), y_i) = I(\hat{y}_i^{(\mathbf{h},\mathbf{y})}(x_i) \neq y_i), \tag{3}$$

  I the indicator function.
- We can characterize (the model of) an instantiation $\mathbf{y} \in \mathbf{Y}$ by $n \cdot \mathcal{R}_{emp}(\mathbf{h}, \mathbf{x}, \mathbf{y})$, the number of misclassifications, using the 0/1-loss.

Narrowing Down Supersets

Definition ($\mathcal{E}$-Optimistic Subset)

Let $\mathbf{Y}$ be the Cartesian product of the observed supersets as above and $\mathcal{E} \in \mathbb{N}$ a pre-defined upper bound for classification errors. Then

$$\mathbf{Y}_{\mathcal{E}} = \{\mathbf{y} \in \mathbf{Y} \mid n \cdot \mathcal{R}_{emp}(\mathbf{h}, \mathbf{x}, \mathbf{y}) \leq \mathcal{E}\} \subseteq \mathbf{Y},$$

shall be called $\mathcal{E}$*–optimistic subset of* $\mathbf{Y}$.

Narrowing Down Supersets

Definition (i-th Consideration Function)

Let $y_i \in \mathbf{y} \in \mathbf{Y}_{\mathcal{E}}$ be the class of a fixed observation $i \in \{1, \dots, n\}$ in an instantiation $\mathbf{y} \in \mathbf{Y}_{\mathcal{E}}$. For varying $\mathcal{E}$, the function

$$
\begin{aligned}
f_i : \mathbb{N} &\longrightarrow 2^{\mathcal{Y}} \\
\mathcal{E} &\mapsto \{y \in \mathcal{Y} \mid \exists \mathbf{y} \in \mathbf{Y}_{\mathcal{E}} : y = y_i, y_i \in \mathbf{y}\}
\end{aligned}
$$

shall be called *consideration function* of observation i.

Contents

Motivation: Resolving Ties

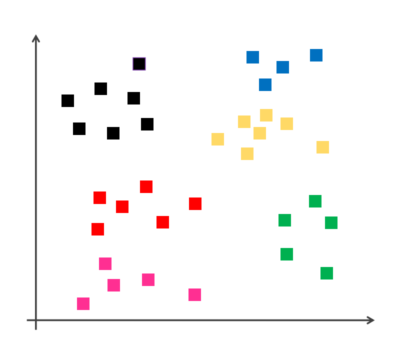


Figure: Recall the “good” instantiation...

Image credits: Eyke Hüllermeier

Motivation: Resolving Ties

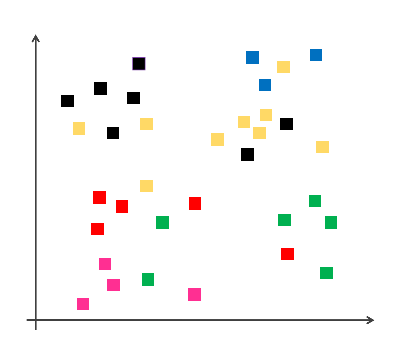


Figure: ... and the “bad” one.

Motivation: Resolving Ties

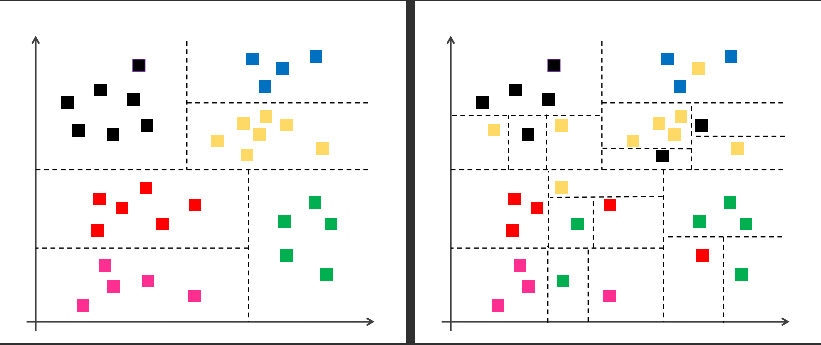


Figure: Notably, both instantiations can be completely separated.

Resolving Ties

- The total order induced by $\mathcal{R}_{emp}(\mathbf{h}, \mathbf{x}, \mathbf{y}^*)$ can include ties:

$$\mathbf{Y}^*_{\mathcal{E}} \stackrel{def}{=} \{\mathbf{y}^* \mid n \cdot \mathcal{R}_{emp}(\mathbf{h}, \mathbf{x}, \mathbf{y}^*) = \mathcal{E}\}$$

- Idea: Use SVMs and relax hyperparameter C that controls model generality in $\mathbf{h} = (C, \mathbf{h}_r^{'})'$.[1]

$$\mathbf{y}^*_C = \arg\min_{\mathbf{y}^*} \underline{\arg\min_{C}} \{\mathcal{R}_{emp}(\mathbf{h}_r, C, \mathbf{x}, \mathbf{y}^*) \mid \mathbf{y}^* \in \mathbf{Y}^*_{\mathcal{E}}\} \tag{4}$$

[1] with remaining hyperparameters $\mathbf{h}_r$.

Resolving Ties

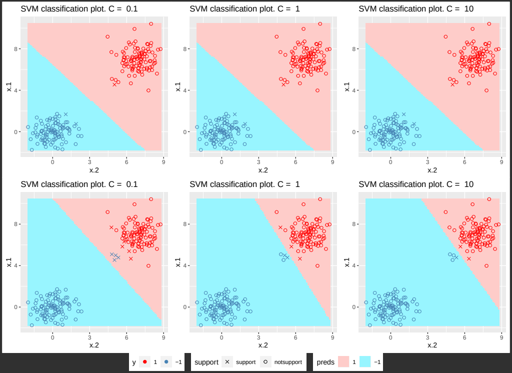
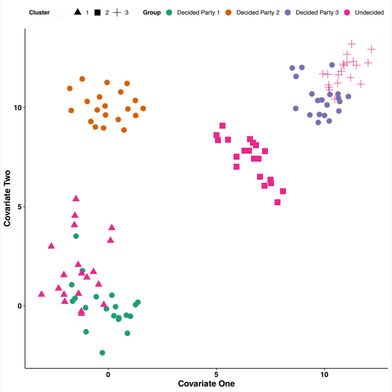


Figure: Different instantiations of set-valued observations require different levels of C in order to be classified correctly.

Contents

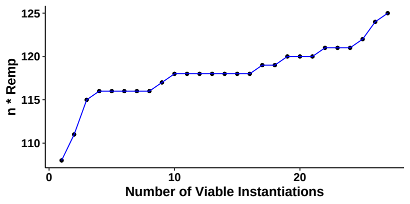
Application: Simulation



Instant.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
Cluster I	1	1	1	1	1	1	1	1	1	2	3	2	3	3	2	3	2	2	3	2	3	3	2	2	3	2	3
Cluster II	2	3	1	3	3	1	2	2	1	1	1	2	3	2	3	1	1	1	3	2	1	2	3	2	3	3	2
Cluster III	3	3	3	2	1	2	1	2	1	3	3	3	3	3	3	1	1	2	1	1	2	1	1	2	2	2	2
n * R_emp	0			2						6						8							9				
Low C	0,01	0,05	0,05	0,08	0,19	0,26	0,32	0,37	0,40	0,13	0,17	0,17	0,18	0,18	0,19	0,43	0,52	0,83	0,85	0,87	0,93	0,94	0,57	0,63	0,76	0,77	0,77

Figure: Simulation setting: 120 observations in a two-dimensional covariate space.

Application: Polling Data provided by Civey



Instant.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
Cluster I	Lef	SPD	SPD	Gre	Gre	Gre	Gre	SPD	SPD	SPD	SPD	Lef	Lef	SPD	Lef	Gre	Lef	Gre	Gre	Lef	SPD	Lef	Lef	Lef	Gre	Gre	SPD
Cluster II	SPD	SPD	Gre	Gre	Lef	Gre	Lef	Lef	SPD	Lef	Gre	Lef	Gre	Gre	Lef	SPD	SPD	Gre	Lef	Lef	Lef	SPD	Gre	Gre	SPD	SPD	SPD
Cluster III	Lef	SPD	Lef	Gre	Gre	Lef	Lef	Gre	Lef	Lef	Gre	Gre	SPD	SPD	SPD	Lef	SPD	SPD	SPD	Lef	SPD	Gre	Lef	Gre	Gre	SPD	Gre
n * R_emp	108	111	115	116					117	118							119		120			121			122	124	125
Low C	990	1000	1000	889	960	960	970	1000	929	828	919	950	950	960	960	990	960	970	899	919	939	828	980	990	950	818	939

Figure: Results from application on polling data.

Contents

Discussion

- Future work:
  - general approaches of "data selection"
    - e.g. integrate the restrictions on $\mathcal{Y}$ and/or $\mathbf{Y}$ as side-constraints for classical OSL.
  - decision criteria for selecting instantiations
    - currently lexicographic order
    - alternatives:
    - multi-objective optimization $\implies$ Pareto front
    - scalarized objective: weighted sum of $\mathcal{E}$ and C
- Questions:
  - Have you heard of superset learning before?
  - Anyone working with set-valued observations?

Contents

Hierarchy on Instantiations

Definition ($\mathcal{E}$-Optimistic Subset)

Let $\mathbf{Y}$ be the Cartesian product of the observed supersets as above and $\mathcal{E} \in \mathbb{N}$ a pre-defined upper bound for classification errors. Then

$$\mathbf{Y}_{\mathcal{E}} = \{\mathbf{y} \in \mathbf{Y} \mid n \cdot \mathcal{R}_{emp}(\mathbf{h}, \mathbf{x}, \mathbf{y}) \leq \mathcal{E}\} \subseteq \mathbf{Y},$$

shall be called $\mathcal{E}$*−optimistic subset of* $\mathbf{Y}$.

Individual Hierarchy

Definition (i-th Consideration Function)

Let $y_i \in \mathbf{y} \in \mathbf{Y}_{\mathcal{E}}$ be the class of a fixed observation $i \in \{1, \dots, n\}$ in an instantiation $\mathbf{y} \in \mathbf{Y}_{\mathcal{E}}$. For varying $\mathcal{E}$, the function

$$\begin{aligned} f_i : \mathbb{N} &\longrightarrow 2^{\mathcal{Y}} \\ \mathcal{E} &\mapsto \{y \in \mathcal{Y} \mid \exists \mathbf{y} \in \mathbf{Y}_{\mathcal{E}} : y = y_i, y_i \in \mathbf{y}\} \end{aligned}$$

shall be called *consideration function* of observation i.

Individual Hierarchy

Proposition

Function $g_i(\mathcal{E}) = |f_i(\mathcal{E})|$ is monotonically non-decreasing.

Proof.

Let $\tilde{\mathbf{y}} \in \mathbf{Y}_{\mathcal{E}_1}$. Definition 3 directly delivers that $n \cdot \mathcal{R}_{emp}(\mathbf{h}, \mathbf{x}, \tilde{\mathbf{y}}) \leq \mathcal{E}_1$. With $\mathcal{E}_1 < \mathcal{E}_2$ by assumption, we trivially have $n \cdot \mathcal{R}_{emp}(\mathbf{h}, \mathbf{x}, \tilde{\mathbf{y}}) \leq \mathcal{E}_2$ $\implies \tilde{\mathbf{y}} \in \mathbf{Y}_{\mathcal{E}_2}$. Thus, for any two $\mathcal{E}_1, \mathcal{E}_2 \in \mathbb{R}$ with $\mathcal{E}_1 < \mathcal{E}_2$ it holds $\mathbf{Y}_{\mathcal{E}_1} \subseteq \mathbf{Y}_{\mathcal{E}_2}$. Since $f_i(\mathcal{E})$ only contains classes of instantiations in $\mathbf{Y}_{\mathcal{E}}$, the assertion follows. $\square$

Individual Hierarchy

Definition (i-th Preference Function for level $\mathcal{E}$)

Let $y_i \in \mathbf{y}^* \in \mathbf{Y}^*_{\mathcal{E}}$ be the class of a fixed observation $i \in \{1, ..., n\}$ in an instantiation $\mathbf{y}^* \in \mathbf{Y}^*_{\mathcal{E}}$. For a given $\mathcal{E}$, the function

$$p_i^{(\mathcal{E})} : \mathcal{Y} \to \mathbb{R}$$

$$y \mapsto \min\{C \mid C = \underline{\arg\min_{C}}\{\mathcal{R}_{emp}(\mathbf{h}_r, C, \mathbf{x}, \mathbf{y}^*) \mid \mathbf{y}^* \in \mathbf{Y}^*_{\mathcal{E}} \wedge y = y_i \in \mathbf{y}^*\}\}$$

shall be called *preference function* of observation i for subset $\mathbf{Y}^*_{\mathcal{E}}$.

Individual Hierarchy

Proposition

For any fixed i, the element-wise composition $p_i^{(\mathcal{E})} \odot f_i$ induces a total order.

Proof.

Since $p_i^{(\mathcal{E})}$ maps to $\mathbb{R}$, we have $p_i^{(\mathcal{E})} \odot f_i(\mathcal{E}) \in \mathbb{R}^d$, where $d \leq |\mathcal{Y}|$ is the dimension of the output of $p_i^{(\mathcal{E})}$. Since any subset of the total order $(\mathbb{R}, \leq)$ is a total order with the restriction of the total order on the subset, one single output vector $p_i^{(\mathcal{E})} \odot f_i(\mathcal{E}) \in \mathbb{R}^d$ has elements that are totally ordered. $\square$

Contents

References I

Hüllermeier, E. (2014).

Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization.

International Journal of Approximate Reasoning, 55:1519–1534.

Hüllermeier, E. and Cheng, W. (2015).

Superset learning based on generalized loss minimization.

In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 260–275. Springer.

References II

Hüllermeier, E., Destercke, S., and Couso, I. (2019).
Learning from imprecise data: adjustments of optimistic and pessimistic variants.
In *International Conference on Scalable Uncertainty Management*, pages 266–279. Springer.

Lienen, J. and Hüllermeier, E. (2021).
Credal self-supervised learning.
Advances in Neural Information Processing Systems, 34:14370–14382.