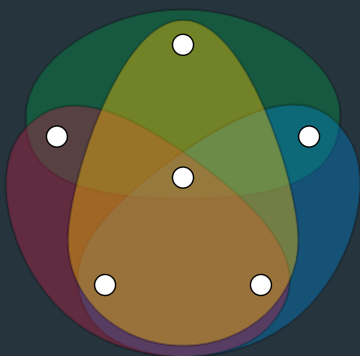


Binary Classification With Hypergraph Case-Based Reasoning

DOLAP 2018

Alexandre Quemy
IBM Analytics
Politechnika Poznańska

AGENDA

- Binary classification problem
- Hypergraph representation
- Model Space and Model Selection
- Experiments & Results
- Improvements, work in progress and future plans

Binary classification problem

IBM

Binary Classification problem

Classical formulation:

Find a mapping h

$$h: \; X \to \{0,1\}$$
$$x \mapsto h(x)$$

such that h minimize the classification error.

Very often $X = \mathbb{R}^N$

In practice, ML algorithm select or build **h** from a model-space **H** made of restrictions or hypothesis on the „shape” of **h** based on the data.

Problem: Given a training set $\mathbf{x} \in X^n$, optimize $\min_{h \in H} \sum_{x \in \mathbf{x}} 1_{\{f(x) \neq h(x)\}}$

Binary Classification problem

Formulation:

Consider an abstract space of information $\mathbb{F}$ and a σ-algebra $\mathcal{F}$ s.t. $(\mathbb{F}, \mathcal{F})$ is measurable.

Work Hypothesis: $\mathbb{F}$countable (finite) space and $\mathcal{F} = \mathcal{P}(\mathbb{F})$

The unknown measurable mapping:

$$\begin{aligned} J : \mathcal{P}(\mathbb{F}) &\to \{0, 1\} \\ x &\mapsto J(x) \end{aligned}$$

Problem: Given a training set $X \in \mathcal{P}(\mathbb{F})^n$, optimize $\min_{\bar{J}} \sum_{x \in \mathcal{P}(\mathbb{F})} 1_{\{J(x) \neq \bar{J}(x)\}}$

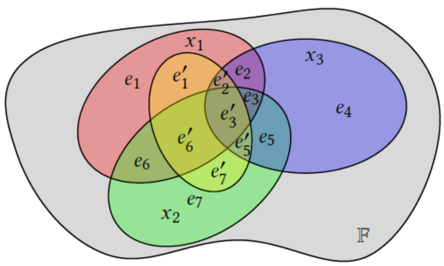
Hypergraph representation

Hypergraph representation

Few definitions:

Hypergraph: $H = (V, X)$ with V a set of vertices,
X the hyperedges such that $\forall x \in X,\ x \subseteq V$.

The projection operator π_H



$\forall x \in \mathcal{P}(\mathbb{F}), D_x$ discretionary features

$d_H(x) = \{x' \in X \mid x \cap x' \neq \emptyset\}$

$d_H^{(l)}(x) = \{x' \in d_H(x) \mid J(x') = l\}$

Partition or Intersection Family:
$\mathcal{E}_H = \{e_i\}_{i=1}^m = \{e \in \bigcup_{x \in X} \pi_H(x)\}$

Model Space and Model Selection

Model Space and Model Selection

Model Space:

Given $H = (\mathbb{F}, X)$ and $\mathcal{E} = \{e_i\}_i^m$:

$$w(e, x) = \frac{|x \cap e|}{|x \setminus D_x|}$$

Support

Importance of e in x

$$\begin{cases} s_{w,\mu}(x) &= \sum_{i=1}^{m} w(e_i, x)\mu(e_i) \\ \sum_{i=1}^{m} w(e_i, x_j) &= 1 \quad \forall 1 \leq i \leq n \\ \sum_{i=1}^{m} \mu(e_i) &= 1 \end{cases}$$

Intrinsic strength of e w.r.t. H

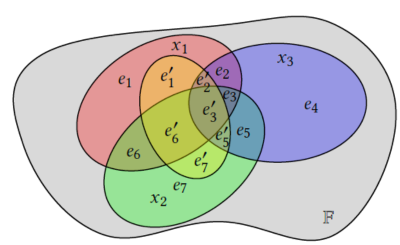
$$w(e_i, x_j) = 0 \text{ if } e_i \cap x_j = \emptyset$$

Decision rule:

$$\forall x \in \mathcal{P}(\mathbb{F}), \; \bar{J}(x) = \begin{cases} 1 & s(x) > 0 \\ 0 & s(x) \leq 0 \end{cases} \qquad \text{(R1)}$$

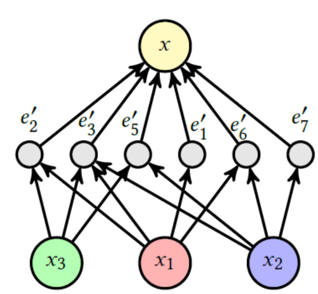
Model Space and Model Selection

Hypergraph Case-Based Reasoning:



Algorithm 1 HCBR (High level view)

Build H and $\mathcal{E}$ from X.
Calculate w and μ on $\mathcal{E}$.
Adjust μ with training algorithm
for each x in test set **do**
    Calculate the projection $\pi(x)$.
    Calculate the support $s(x)$ using the projection.
    Predict using the updated rule (R2).
end for



Model Space and Model Selection

Model Selection:

Intrinsic strength of $e \in \mathcal{E}$ w.r.t. $x \in X$

$$\forall l \in \{0,1\},\ S^{(l)}(e,x) = \frac{|d^{(l)}(e)|^{\frac{|x \cap e|}{|x|}}}{\sum\limits_{e_j \in \mathcal{E}} |d^{(l)}(e_j)|^{\frac{|x \cap e_j|}{|x|}}} \qquad = \text{distribution of support for } l \text{ in } x$$

Instrinsic strength of $e \in \mathcal{E}$ w.r.t. $H = (\mathbb{F}_X, X)$:

$$\forall l \in \{1,0\},\ S^{(l)}(e) = \frac{|e|}{|\mathbb{F}_X|} \sum_{x \in d^{(l)}(e)} S^{(l)}(e,x)$$

$$\forall l \in \{1,0\},\ \mu^{(l)}(e) = \frac{S^{(l)}(e)}{\sum\limits_{e' \in \mathcal{E}} S^{(l)}(e')} \qquad = \text{distribution of support for } l \text{ over } \mathcal{E}$$

$$\mu(e) = \mu^{(1)}(e) - \mu^{(0)}(e)$$

Model Training

Objective: Minimizing a sort of Hinge-loss

Algorithm 2 Model training

Input:
- X: training set
- y: correct labels for X
- k: number of training iterations
- $\mu^{(1)}, \mu^{(0)}$: weights calculated with (4.5)

Output:
- Modified vectors $\mu^{(1)}, \mu^{(0)}$

```
for k iterations do
    for x_i ∈ X do
        ȳ_i ← J̄(x_i)
        if ȳ_i ≠ y_i then
            for e ∈ π(x_i) do
                μ^(y_i)(e) ← μ^(y_i)(x_i) + w(e, x_i)|μ(e)|
                μ^(ȳ_i)(e) ← μ^(ȳ_i)(x_i) − w(e, x_i)|μ(e)|
            end for
        end if
    end for
end for
```

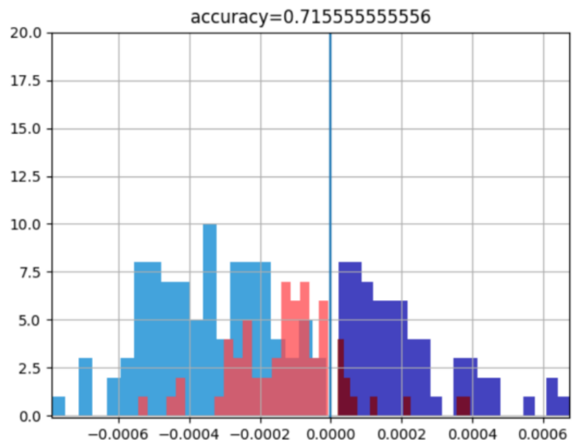
Correcting μ s.t. $J(x) = \bar{J}(x)$ is a bad idea:

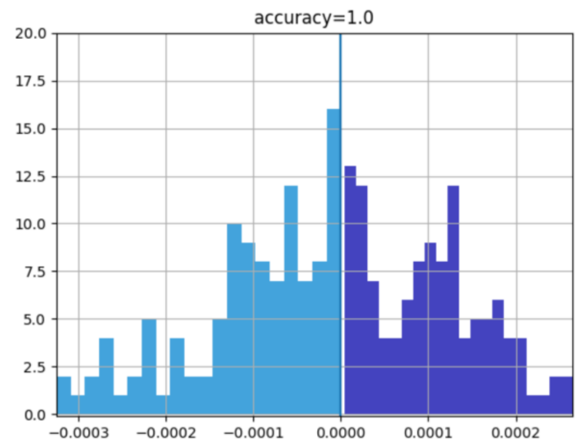
1. The neighbor cases might become wrongly classified,
2. The meaning of s and μ is lost.

Idea: gradually adjust $\mu^{(l)}$ proportionally their contribution in x.

Drawback: order dependant! No convergence proven.

Model Training





Text classification

Complexity

Model Building:

- Constructing $\mathcal{E}_H$: $\mathcal{O}(\sum_{x\in X}|x|)$ (Partition Refinement data structure)
- Calculating $S(x,e)$: $\mathcal{O}(\sum_{x\in X}|x|)$
- On M-uniform hypergraphs: $\mathcal{O}(Mn)$.
- Calculating μ: $\mathcal{O}(|\mathcal{E}_H|)$
- Very pessimistic bound: $|\mathcal{E}_H| \leq \min(2^n-1, |\mathbb{F}_X|)$

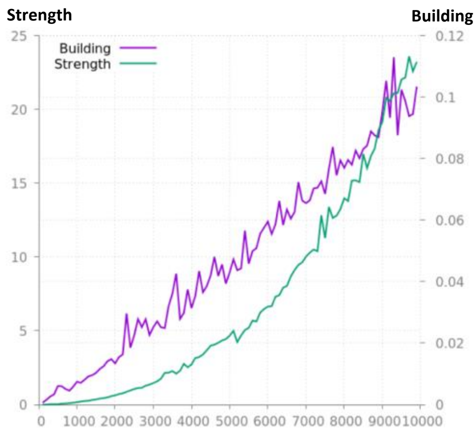
Learning Phase:

- $x\in X$: $\mathcal{O}(|x|)$ steps per x (maximal cardinal for $\pi(x)$)
- dataset X: $\mathcal{O}(k\sum_{x\in X}|x|)$
- M-uniform hypergraphs: $\mathcal{O}(kMn)$.

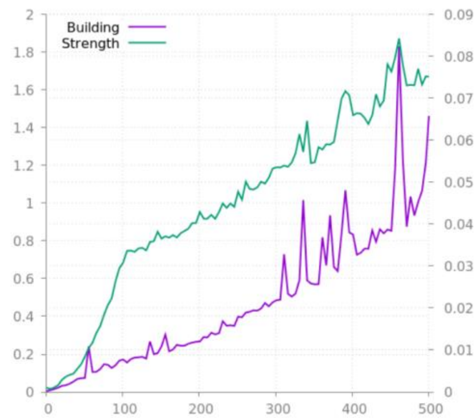
Model Query: $\mathcal{O}(|x|)$ (maximal cardinal for $\pi(x)$).

Complexity

In practice:



Depending on n (m = 10)



Depending on m (n = 100)

Experiments and Results

Code and experiment: github.com/aquemy/hcbr

IBM

Experiments and Results

	Cases	Total Features	Unique	Min. Size	Max. Size	Average Size	Real
adult	32561	418913	118	10	13	12.87	No
audiology	200	13624	376	70	70	70	No
breasts	699	5512	80	8	8	8	No
heart	270	3165	344	12	13	12.99	Yes
mushrooms	8124	162374	106	20	20	20	No
phishing	11055	319787	808	29	29	29	No
skin	245057	734403	768	3	3	3	Yes
splice	3175	190263	237	60	60	60	No

	Accuracy (standard dev.)	Recall	Specificity	Precision	Neg. Pred. Value	F_1 score	Matthews corr. coef.
adult	0.8206 (0.0094)	0.8832	0.6233	0.8808	0.6290	0.8820	0.5081
audiology	0.9947 (0.0166)	1.0000	0.9875	0.9917	1.0000	0.9957	0.9896
breasts	0.9696 (0.0345)	0.9691	0.9676	0.9479	0.9844	0.9575	0.9344
heart	0.8577 (0.0943)	0.8695	0.8437	0.8699	0.8531	0.8653	0.7178
mushrooms	1.0000 (0.0000)	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
phishing	0.9605 (0.0081)	0.9680	0.9514	0.9615	0.9590	0.9647	0.9199
skin	0.9865 (0.0069)	0.9608	0.9932	0.9736	0.9898	0.9672	0.9587
splice	0.9443 (0.0124)	0.9478	0.9398	0.9450	0.9441	0.9463	0.8884

Experiments and Results

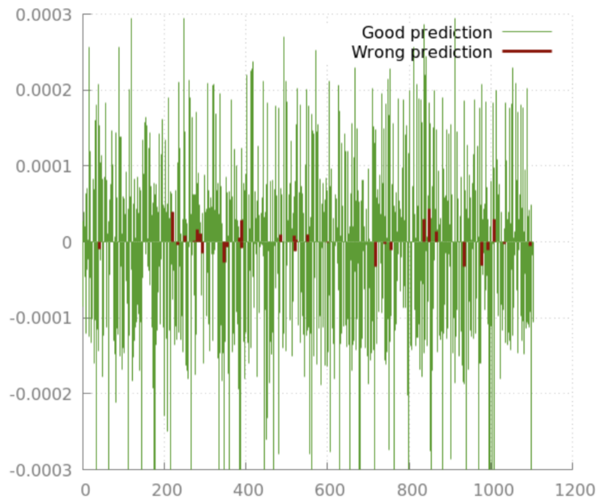
Protocol: 10 fold cross-validation,
no metaparameter tuning (only training)

Contrary to the state-of-art, no assumption, no ad-hoc feature selection or transformation.

Dataset	Ref.	Type	Accuracy
adult	[14]	Many classifiers	86.25%
	[15]	SVM	85.35%
		HCBR	**82.06%**
breast	[1]	SVM	99.51%
	[17]	Neural Network	99.26%
	[19]	SVM	98.53%
	[10]	Bayes	98.1%
	[24]	Neural Network	97.36%
	[13]	Bayes	97.35%
		HCBR	**96.96%**
	[7]	SVM	96.87%
	[9]	Rule-based	95.85%
	[11]	Rule-based	95.84%
	[20]	Decision Tree	94.74%
heart	[21]	Neural Network + Rule-based	87.78%
		HCBR	**85.77%**
	[13]	Bayes	83.00%
	[9]	Rule-based	82.96%
mushrooms	[11]	Rule-Based	100.00%
		HCBR	**100.00%**
	[8]	k-NN	99.96%
phishing	[22]	Ensemble	97.75%
	[22]	Random-Forest	97.58%
		HCBR	**96.05%**
	[23]	Neural Network	94.90%
skin	[2]	Generalized Linear Model	99.92%
	[6]	Decision Tree	99.68%
	[5]	Neural Network + Boosting	98.94%
		HCBR	**98.65%**
splice	[5]	Neural Network + Boosting	97.54%
		HCBR	**94.43%**
	[4]	(fuzzy) Decision Tree	94.10%

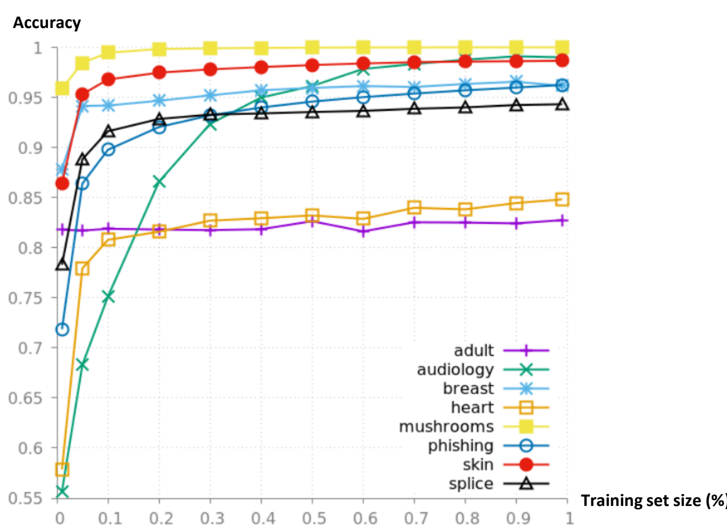
Experiments and Results

Confidence measure:



Classification problem

Very few examples needed + does not overfit:



Improvements, WIP and Future plans

Improvements, WIP and future work

Multiclass and multilabel support:

Straigthforward time-linear extension of mu

Fully online and scalable version:

Online:

Semi-online: training after each decision but the input vector not added to the hypergraph
Fully online: new hyperedge, then weights adjustment

Vertical and horizontal scalability:

Vertical: adding more cases (i.e. fully online)
Horizontal: add more atoms to some cases without starting from scratch

Improvements, WIP and future work

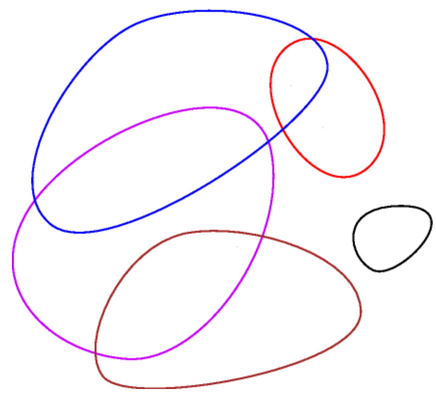
Model Space extension:

Given $H = (\mathbb{F}, X)$ and $\mathcal{E} = \{e_i\}_i^m$:

$$\begin{cases} s_{w,\mu}(x) & = \sum_{k=1}^{K}\sum_{i=1}^{m} w_k(e_i, x)\mu(e_i) \\ \sum_{i=1}^{m} w_0(e_i, x_j) & = 1 \qquad \forall 1 \le i \le n \\ \sum_{i=1}^{m} \mu(e_i) & = 1 \end{cases}$$

$\cancel{w(e_i, x_j) = 0 \text{ if } e_i \cap x_j = 0}$

$w_k(e_i, x_j) \neq 0$ if $\exists$ k-path $x_j \to e_i$



Thank you