Strojové učení se zaměřením na vliv vstupních dat

Irina Perfilieva, Petr Hurtík, Marek Vajgl

Centre of excellence IT4Innovations
Division of the University of Ostrava
Institute for Research and Applications of Fuzzy Modeling
Ostrava, Czech Republic

2014-09-24

IRAFM

Učení

- Obecné učení chápáno jako proces s cílem detekce a uložení významných rysů nad daným zdrojem
- S využitím rysů probíhá rozpoznání nového objektu na základě 'nějak definované' podobnosti
- V našem případě detekce typu vady kamene na základě vzorů daných vad

Historický vývoj učících algoritmů

- 1952: Eliza, simulace rozhovoru s psychoterapeutem
- 1957: Neuronové sítě, rozpoznávání písmen
- 1995: Support vector machines
- 2003: Logistic regression

Neuronové sítě

- Představeno 1943, ispirováno lidským mozkem
- Základní jednotka: neuron
- Problém: XOR. Vyřešení: síť s více vrstvami neuronů
- Problém: Black box
- K učení vyžadováno velké množství vzorků
- Aktuální téma: deep learning. Klasifikace objektů ve scéně za použití desetitisíců neuronů a urychlení na GPU.

Separační stromy

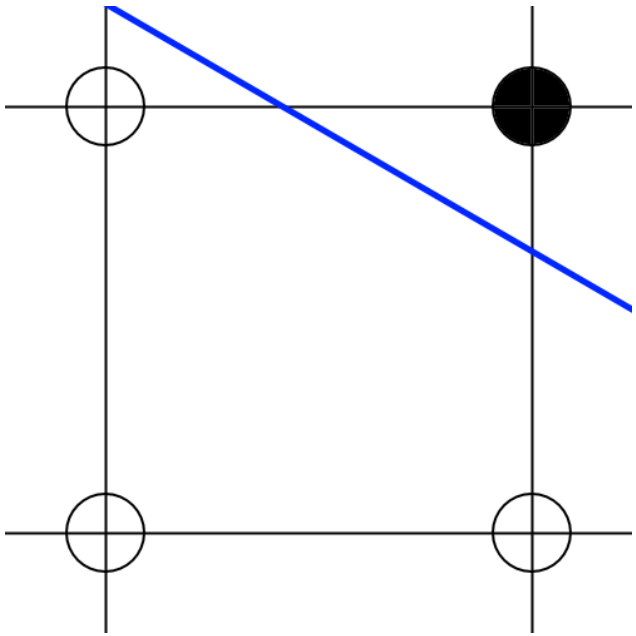
- Rozdělení problému na dva podproblémy (binární strom), nebo n–tici (nární strom)
- Použití především v indexaci dat a vyhledávání v nich (databáze)
- Výhoda: rychlost zpracování
- Nevýhoda: automatizované budování stromu. Při rozdělení se minimalizuje chyba, nemusí ale konvergovat k nule

Učení II

- Více formálně: detekce počtu n a klasifikace charakteristik $k_1, ..., k_n$ množiny objektů s cílem jejich klasifikace na disjunktní třídy $T_1, ..., T_n$
- Snažíme se dosáhnout:
  - úplnosti
  - robustnosti

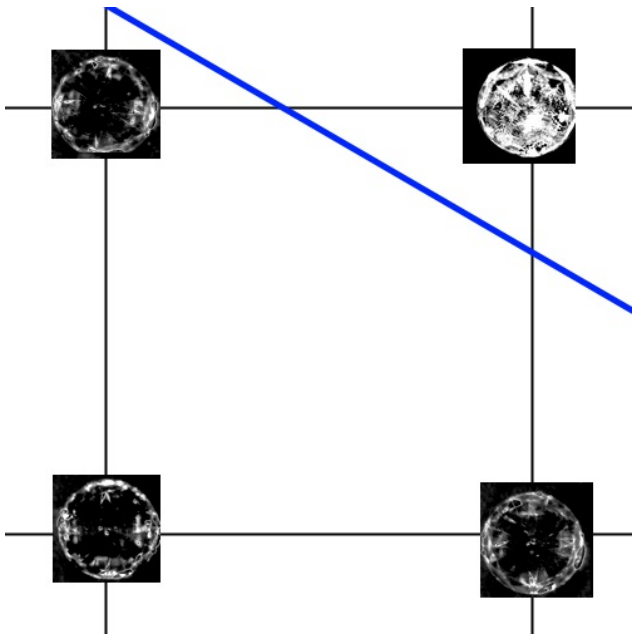
Úplnost

- Klasifikace na 2 třídy
  - Hledání (minimální) množiny charakteristik umožňující nalezení (lineární) separability
  - Pokud můžeme provézt klasifikaci, pak lze říci, že **množina je konzistentní** vzhledem k požadované klasifikaci
  - Při nenalezení řešení je nutné zvětšení dimenze



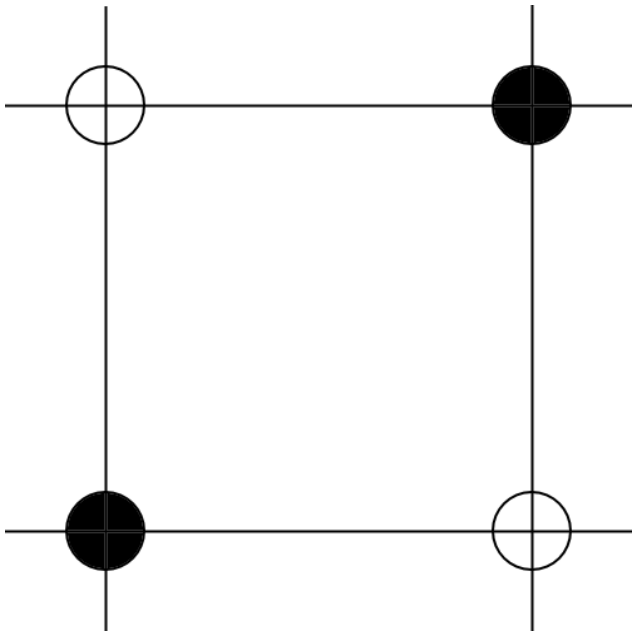
Úplnost II

- $n = 2$, tedy máme dvě charakteristiky
- $\alpha \in \{0, 1\}$



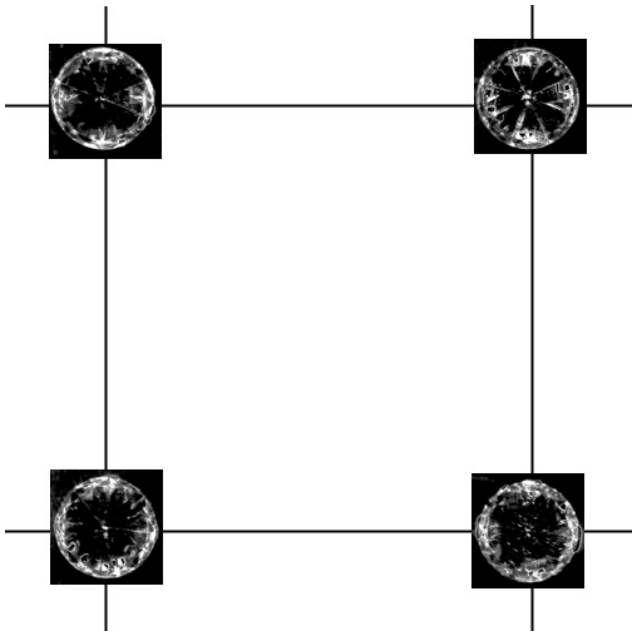
Úplnost III

- Může nastat situace neseparovatelnosti množiny



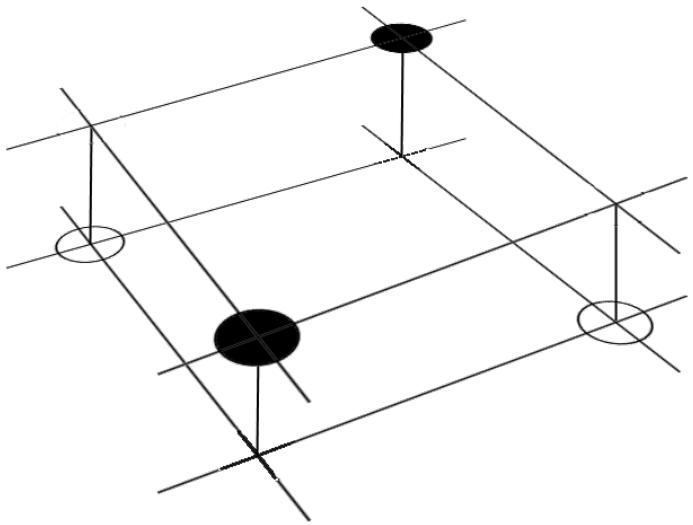
Úplnost III

- Může nastat situace neseparovatelnosti množiny



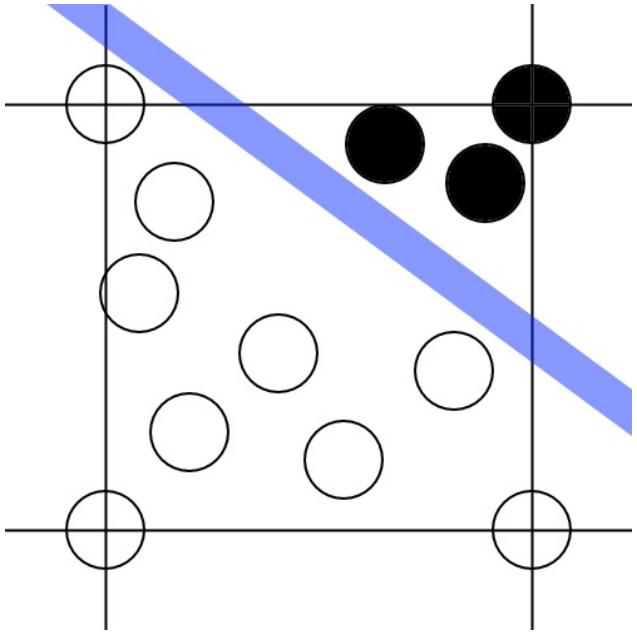
Úplnost IV

- Při nenalezení řešení je nutné zvětšit dimenzi, tj. počet charakteristik n



Robustnost

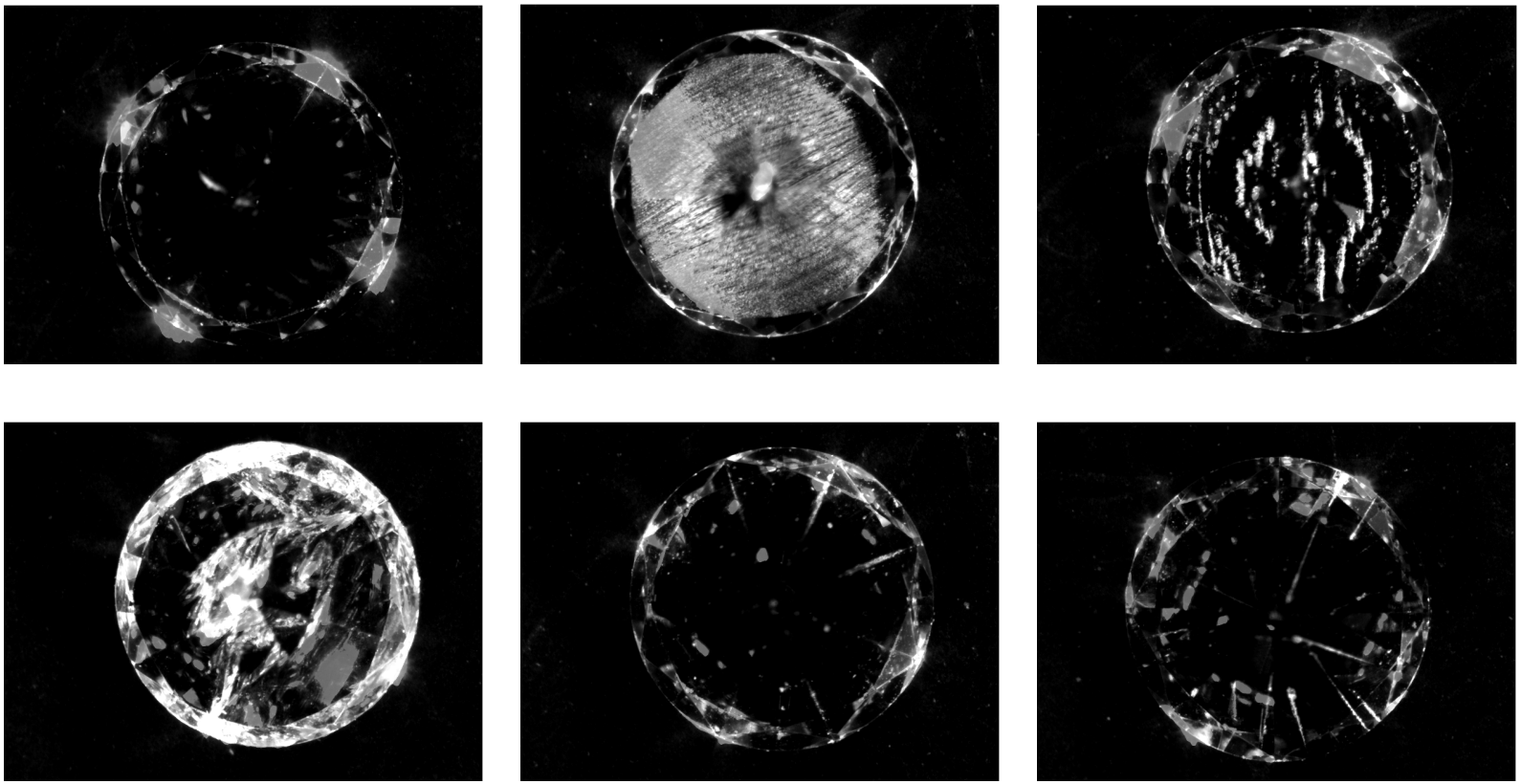
- Odolnost algoritmu vůči vstupním hodnotám
  - Charakteristiky nabývají hodnot $\alpha \in \langle 0, 1 \rangle$



Úplnost a konzistence

- Nalezení separace vstupní množiny je podmíněné její konzistencí
- Vznik nekonzistentní množiny vstupů vůči zvoleným charakteristikám klasifikace vede k nemožnosti provedení klasifikace
- Nekonzistence vzniká rozporem v množině vstupních dat, tj. blízkých objektů z různých klasifikačních tříd
- (V teoretické oblasti) lze pouze zvětšit dimenzi úlohy (tj. zvětšit množinu zvolených charakteristik)

Projevy různých typů poškození

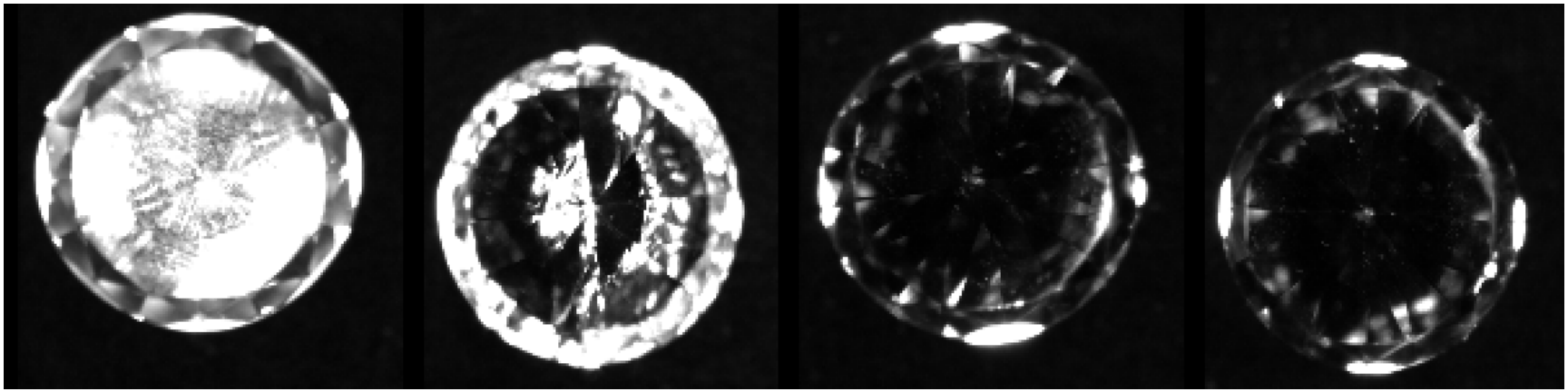


Chyby při učení prakticky

- Neúplnost vstupních dat
- Separovatelnost vad
- Kvalita vstupu (kontrast, rozmazanost)
- Malý počet vstupů

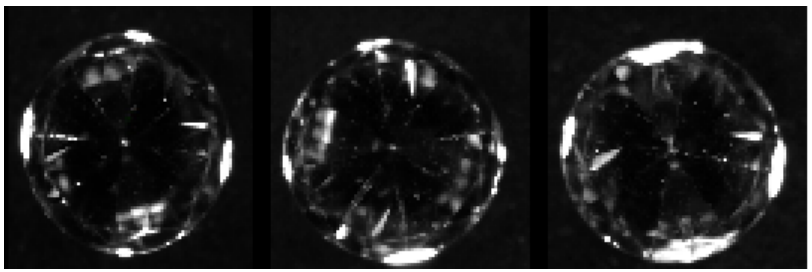
Neúplnost vstupních dat

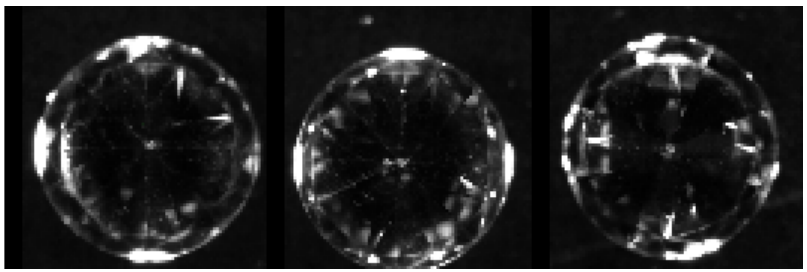
- Množina vstupních dat nepokrývá souvisle celou škálu projevů vady



Separovatelnost vad I

- Odlišné vady se projevují se stejnými artefakty
  - Výsledkem je špatná kategorizace daného kamene
  - Artefakty se nedají popsat triviálním způsobem
  - Hodnota charakteristiky nemá přímou funkční závislost na míře poškození kamene





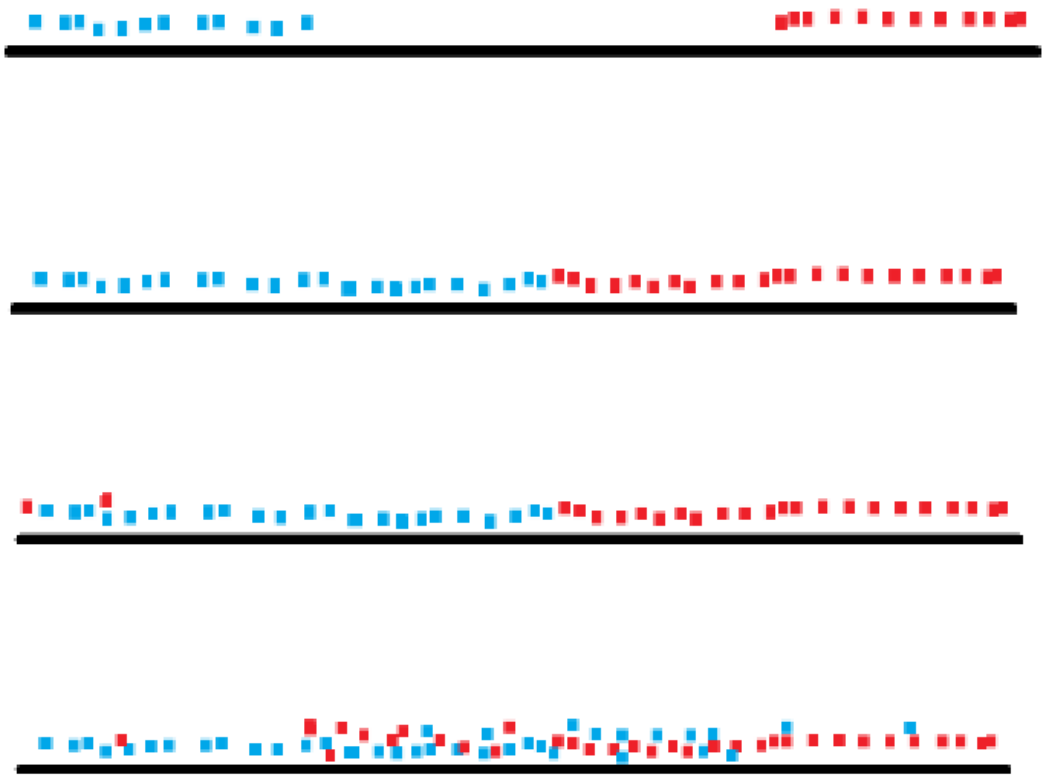
Separovatelnost vad II

- Separace na základě 'vhodně zvolené' charakteristiky
  - Máme n-dimenzionální prostor
  - Objekt je popsán n charakteristikami
  - Cílem je nalezení techniky založené na matematickém modelu, která provede 'nejlepší možné' rozdělení prostoru

Separovatelnost vad III

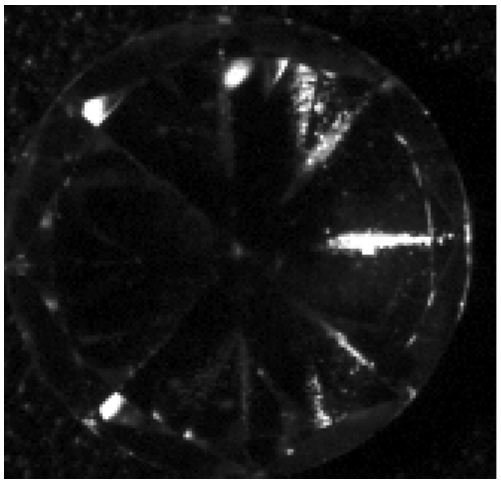
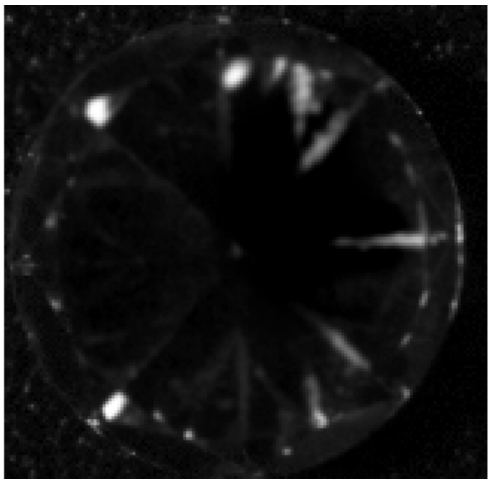
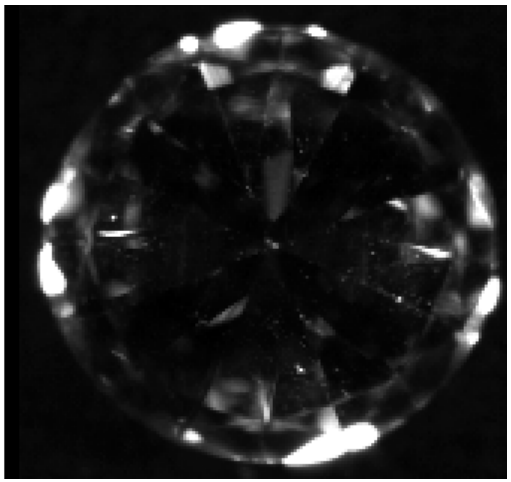
- Implementace spočívá v rozdělení na podprostory
  - Následně je každý podprostor rozdělen zvlášť
- Problém: **Jak rozdělit podprostor**
  - s ohledem na přesnost
  - s ohledem na časovou náročnost
  - v našem případě nalezení prahových hodnot

Separovatelnost vad IV



Kvalita vstupu (kontrast, rozmazanost)

- Kvalita vstupu má výrazný dopad na výpočet charakteristik

Malý počet vstupů

- Lidský prvek řešení vyžaduje minimalizaci velikosti množiny vstupních dat učící sady
  - nutnost ‘vyrobit‘ data reprezentující vady
  - časová náročnost nového ‘učení‘

Srovnání a výsledky

method	D_1		D_2	
	6 classes	2 classes	6 classes	2 classes
proposed	**0.862**	**0.983**	0.674	0.943
rf	0.833	**0.983**	**0.740**	**0.960**
parRF	0.833	**0.983**	0.734	**0.960**
RRFglobal	0.833	**0.983**	0.723	**0.960**
nnet	0.367	0.850	0.441	0.955
cforest	0.817	0.933	0.706	0.944
gamSpline	—	0.833	—	0.944
ctree2	0.467	0.883	0.531	0.938
avNNet	0.417	0.900	0.514	0.938
svmRadial	0.767	0.950	0.508	0.938
pls	0.533	0.817	0.503	0.938
svmRadialCost	0.767	0.967	0.497	0.938
bstSm	—	0.817	—	0.938
bstTree	—	0.817	—	0.938
glmboost	—	0.950	—	0.938
knn	0.583	0.933	0.508	0.932
rpart	0.667	**0.983**	0.633	0.870
rpart2	0.667	**0.983**	0.718	0.864
svmLinear	0.833	0.950	0.480	0.802

IRAFM

Srovnání a výsledky

method	time for model training				time for prediction			
	D_1		D_2		D_1		D_2	
	6 classes	2 classes	6 classes	2 classes	6 classes	2 classes	6 classes	2 classes
proposed	< 1	< 1	< 1	< 1	< 1	< 1	< 1	< 1
gbm	—	370.0	—	396.0	—	2.0	—	2.0
knn	261.2	306.0	258.5	278.4	2.0	2.4	2.0	2.0
pls	263.2	289.6	264.0	284.4	2.8	2.8	2.8	2.4
avNNet	631.2	554.0	1026.1	803.6	2.4	2.8	2.4	2.8
RRFglobal	329.6	332.1	702.4	481.3	2.4	2.8	3.2	2.8
glm	—	275.2	—	281.2	—	3.2	—	2.8
svmLinear	11896.3	864.5	197700.8	26154.8	7.2	3.2	7.2	2.8
svmRadial	1605.7	686.0	3618.2	1338.5	9.2	10.0	8.8	2.8
rf	380.0	340.5	689.2	446.0	2.4	2.8	4.0	3.2
parRF	339.2	358.8	490.4	387.6	2.8	2.8	2.8	3.2
rpart	286.4	320.8	293.3	303.2	3.2	3.2	4.0	3.2
glmboost	—	362.8	—	350.5	—	3.6	—	3.6
nnet	425.6	442.9	500.0	466.4	3.2	4.0	3.6	3.6
rpart2	283.2	310.0	292.0	293.7	3.6	4.4	3.6	3.6
bstLs	—	1037.6	—	1304.5	—	4.8	—	4.4
bstSm	—	1045.7	—	1316.9	—	4.0	—	4.4
bstTree	—	2278.5	—	3026.1	—	4.4	—	4.4
glmnet	714.8	415.3	1507.3	687.6	22.0	6.4	21.6	5.6
ctree2	334.8	406.0	348.9	414.1	6.8	7.6	7.2	7.6
blackboost	—	706.1	—	756.9	—	9.6	—	9.2
cforest	1215.3	775.7	2912.5	1320.1	15.2	10.8	22.0	12.0
svmRadialCost	1607.7	802.0	3615.8	1280.5	8.8	2.8	10.0	12.4
gamSpline	—	1896.1	—	3025.0	—	22.0	—	28.4

Závěr

- Řešena netriviální úloha separace kamenů
- Zjištěn různý projev totožné vady
- Navržen učící algoritmus s ohledem na rychlost a přesnost
- Provedeno srovnání s existujícími algoritmy