Metody pro kombinování modelů a klasifikátorů

Diplomová práce

Autor
Jan Černý
cernyjn@seznam.cz
České vysoké učení technické v Praze
Fakulta elektrotechnická
Katedra počítačů

Cíle práce

- Zlepšení kvality výstupu regresních a klasifikačních modelů za pomoci ensemblů.
- Automatizace data mining procesu.

Základní modely

Třídy problémů

1) Regresní

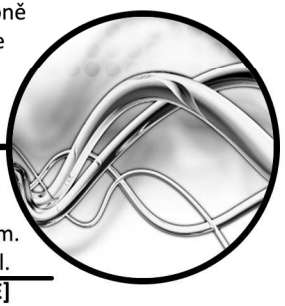
- Pro každý vstupní vektor je výstupního hodnota jedno reálné číslo, kterému se model snaží co nejvíce přiblížit.

1) Klasifikační

- Výstupem je zařazení vstupního vektoru do jedné třídy z konečné množiny tříd.

- Obecně základní model může být cokoliv co nám poskytne na daná vstupní data nějaký výstup.
- V našem softwaru (FAKE-GAME) jsou základní modely reprezentovány jednoduchými funkcemi (exponenciální, polynomiání, sinusová, gaussovská...), ale lze využít libovoně složité modely i modely z jiných aplikací a vylepšit jejich výstupy ensemble algoritmy.

Ensemble



Princip ensemble algoritmů

- Kombinace modelů takovým způsobem, aby výsledek byl lepší než nejlepší z modelů.
- Využívá toho, že každý model dělá různé chyby.

Implementované algoritmy

- Bagging
- Boosting
- Stacking
- Cascade Generalization
- Cascading
- Delegating

Area Specialization (AS)

- Nově navržený algoritmus využívající model v oblasti kde je nejlepší.
- Využití modelů relevantních pouze v některé oblasti dat.

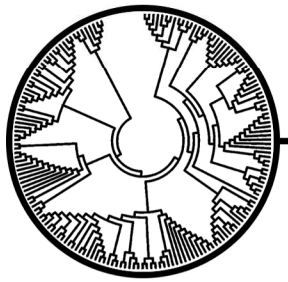
Divide ensemble (DIV)

- Nově navržený algoritmus, který přiděluje každému modelu pouze tu oblast dat na které se optimálně naučí a zvládne ji aproximovat.

Experimenty

- Ens. = sloupec s nejlepším základním ensemble algoritmem.
- Zakl. = nejlepší základní model.

	modely [RMSE]			
data	**zakl.**	**ens.**	**AS**	**DIV**
block	6.6	4.02	**1.21**	3.2
bump	6.85	6.7	**1.39**	4.8
doppler	6.32	5.24	**1.24**	3.21
hevysine	5.1	2.19	0.52	**0.4**



Hierarchický ensemble

- Zobecnění principu ensemblů takovým způsobem, že jako základní modely můžeme využít i ensemble algoritmy.
- Vznik stromových struktur s nekonečným stavovým prostorem => nutnost neúplného prohledávání (evoluce).
- Experimenty jak s ensembly tak s hierarchickými ensembly ukazují na velikou datovou závislost všech algoritmů pro tvorbu modelů.
- Neexistuje tedy jediný algoritmus (nebo jejich kombinace), který by byl nejlepší na všechna data.

=> Potvrzení správnosti našeho přístupu sestavování optimálního modelu výběrem kombinací algoritmů pomocí evoluce.

Evoluční algoritmus

- Algoritmus pro evoluci stromových struktur a optimalizaci proměnných.
- Algoritmus je problémově nezávislý a využívá funkcí problémově závislého kontextu.

Evoluční algoritmus

- Problémově nezávislý (pracuje s obecnými stromy).
- Provádí evoluční operace:
  - mutace uzlu,
  - přidání uzlu,
  - mutace proměnné.

Kontext

- Problémově závislá část algoritmu.
- Načítání a předzpracování dat.
- Výpočet fitness.

Experimenty

	modely [RMSE]				modely [RMSE]		
data	**zakl.**	**ens.**	**evol.**	**data**	**zakl.**	**ens.**	**evol.**
block	6.6	1.21	**0.0015**	buildings 2	0.525	0.517	**0.46**
bump	6.85	1.39	**0.678**	buildings 3	0.63	0.597	**0.52**
doppler	6.32	1.24	**0.65**	flare 1	0.094	0.095	**0.085**
hevysine	5.1	0.4	**0.0116**	flare 2	0.033	0.032	**0.018**
bosthouse	3.31	3.33	**3.23**	flare 3	0.026	0.024	**0.013**
buildings 1	121.2	115.4	**112.4**	mandarin	0.103	0.097	**0.081**

Experimenty ukazují, že přístup popsaný v mé práci dokáže zlepšit kvalitu výstupu základních modelů o jednotky až desítky procent a zároveň automatizovat proces tvorby modelů.