Některé potíže s klasifikačními modely v praxi

Nikola Kaspříková
KMAT FIS VŠE v Praze

Literatura

- J. M. Chambers: Greater or Lesser Statistics: A Choice for Future Research. *Statistics and Computation* 3, 1993.
- L. Breiman: Statistical Modeling: The Two Cultures. *Statistical Science* 16, 2001.
- D. J. Hand: Classifier Technology and the Illusion of Progress. *Statistical Science* 21, 2006.
- R. J. Bolton, D. J. Hand: Statistical Fraud Detection: A Review. *Statistical Science* 17, 2002.
- Ch. Hennig: Measurement of Quality in Cluster Analysis. 59th ISI WSC, 2013.
- T. Hastie, R. Tibshirani, J. Friedman: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2009.
- W.N. Venables, B.D. Ripley: Modern Applied Statistics with S. 2002.

Klasifikace (klasifikačních) modelů

- Model typu „Budeme se na něj dívat“
- Model pro podporu rozhodování, model je návodem k jednání (výkonnost nás bude zajímat)

Zjišťování podvodů

- Metody a výsledky se moc nesdělují, podvodníci se také učí a reagují na vývoj
- Supervised learning - chytáme ty neúspěšné, známé vzory
- a unsupervised - zachycení neobvyklého chování (to se ale v čase mění)
- link analysis, social networks analysis
- nejisté přiřazení třídám, různě velké skupiny a různé ceny chybné klasifikace
- počítáme *suspicion score* (Bolton a Hand), je potřeba prověřit mimo statistickou analýzu
- měly by se uvážit ceny chybné klasifikace, náklady vyšetřování a přínosy z odhalení podvodu

Shlukování

- „Do not assume that 'clustering' methods are the best way to discover interesting groupings in the data; in our experience the visualization methods are often more effective. There are many different clustering methods, often giving different answers, so the **danger of over-interpretation is high**.“ (Venables a Ripley)
- „**Specifying an appropriate dissimilarity measure is far more important** in obtaining success with clustering **than choice of clustering algorithm**. This aspect of the problem is emphasized less in the clustering literature than the algorithms themselves, since it depends on domain knowledge specifics and is less amenable to general research.“ (Hastie et al.)
- Subjektivně stanovené míry odlišnosti často mají překvapivé vlastnosti (nesymetrické, neplatí trojúhelníková nerovnost,...)

Vzdálenosti pro sekvence stavů

Hamming – pro sekvence stejné délky; počet pozic, na kterých se sekvence liší

LCP, LCS

Edit distance – náklady na optimální převod jedné sekvence na druhou pomocí základních operací – vložení, odmazání nebo náhrada stavu v sekvenci

Edit distance

Vstup: dvě sekvence a ceny jednotlivých úprav
(Jak zvolit ceny úprav?)

Řešení se hledá technikou dynamického programování

Ceny: S=2; I,D=1

		1	2	3	4	5	6	7	8	9	10
	0	**M**	**I**	**T**	**U**	**T**	**L**	**G**	**S**	**T**	**G**
1	**T**	2	3	2	3	4	5	6	7	7	8
2	**L**	3	4	3	4	5	4	5	6	7	8
3	**G**	4	5	4	5	6	5	4	5	6	7
4	**S**	5	6	5	6	7	6	5	4	5	6
5	**C**	6	7	6	7	8	7	6	5	6	7
6	**T**	7	8	7	8	9	8	7	6	5	6
7	**G**	8	9	8	9	10	9	8	7	6	5

Hodnocení klasifikátorů

- IFCS - podpora benchmarkingu ve vývoji shlukovacích metod (Hennig)
- Data pro benchmarking
    - Generovaná data
    - Reálná data se známými třídami
    - Reálná data bez známých tříd
- Problémy se srovnáváním (Hand)
    - Na jiných datech může technika fungovat jinak
    - Nebere se v úvahu různá úroveň zkušeností uživatelů

Hodnocení klasifikátorů

- Data při aplikaci klasifikátoru často nejsou výběrem ze stejného rozdělení jako data, nad kterými je klasifikátor vytvářen
- Při tvorbě klasifikátoru se učinily volby a předpoklady, které nemusejí být správné -> zavádějící představa o výkonnosti klasifikátoru
- Obvykle se předpokládá, že třídy jsou objektivně dané, není nejistota o jejich přiřazení
- Parametry/struktura se odhadují podle optimalizačního kritéria, které ne zcela odpovídá skutečnému cíli klasifikace. Hodnocení podle podílu chybně klasifikovaných případů nebývá vhodné
- Apriorní pravděpodobnosti a ceny se mění
- ... Jaký je přínos z vývoje lepších klasifikátorů?