

Evaluace fuzzy ILP klasifikátoru na datech o dopravních nehodách

Jan Dědek, Peter Vojtáš, Marta Vomlelová

Katedra softwarového inženýrství, MFF, Univerzita Karlova v Praze
Malostranské náměstí 25, 118 00, Praha 1
`dedek@ksi.mff.cuni.cz`

Abstrakt Motivací pro tento příspěvek je klasifikace zpráv o dopravních a jiných nehodách podle jejich závažnosti. Pro tento účel jsme použili nový druh klasifikátoru, který je založen na induktivním logickém programování a fuzzy logice. V příspěvku nabízíme srovnání tohoto fuzzy ILP klasifikátoru s dalšími alternativami, jmenovitě s rozhodovacími stromy J48, s klasifikátorem SVM a s vícevrstevným perceptronem. Výsledky tohoto srovnání ukazují, že fuzzy ILP klasifikátor je na těchto datech srovnatelný se zmíněnými alternativami.

Klíčová slova: fuzzy, ILP, klasifikace, strojové učení

1 Úvod

V našem výzkumu se zabýváme extrakcí informací z webových reportů. Jednou z datových sad, které používáme, jsou záznamy o hasičských výjezdech publikované na stránkách Hasičských záchranných sborů ČR¹. Klasifikace závažnosti nehody představuje přímočaré využití strukturovaných informací extrahovaných z reportu o ní. V tomto příspěvku se nezabýváme tím, jak získat strukturované informace z textového reportu (o tom viz např. naši práci [1]), ale tím, jak na základě takových informací určit závažnost dané nehody.

V naší předešlé publikaci [2] jsme implementovali netypický druh klasifikátoru, který je založen na induktivním logickém programování a fuzzy logice. Tento klasifikátor jsme použili na stejná data jako v této práci, ale vyhodnocení úspěšnosti tohoto klasifikátoru nebylo úplné. Nedostatky předešlé práce shrneme v následujících dvou bodech a v této práci se je pokusíme napravit a podrobněji vysvětlit výsledky srovnání.

Křížová validace. Aby bylo srovnání metod alespoň trochu objektivní a daly se z něj odvodit statisticky relevantní závěry, je nutné spustit test několikrát na různých trénovacích a testovacích datech. Křížová validace je ideálním prostředkem k naplnění tohoto cíle.

Srovnání s alternativami. Aby čtenář mohl posoudit, jak kvalitních výsledků metoda dosahuje, je nutné dát k dispozici i výsledky alternativních řešení nad shodnými daty.

¹ <http://www.hzscr.cz/>

2 Data

Data, na kterých jsme experiment provedli, pocházejí z padesáti hasičských zpráv. U každé zprávy jsme vyhodnotili celkem třináct ukazatelů, které sloužily jako vstup pro klasifikaci (např. počet mrtvých, vyčíslená škoda, doba potřebná na likvidaci nehody, počet zasahujících jednotek, atp). Závažnost nehody jsme rozdělili do čtyř tříd a stanovovali jsme ji na základě celkového dojmu, který jsme z popisu nehody získali.

3 Fuzzy ILP klasifikátor

Podrobný popis konstrukce fuzzy ILP klasifikátoru, který jsme použili, je možné nalézt v článku [2]. Spojení ILP s fuzzy logikou spočívá v (1) zavedení axiomů monotonie pro numerické atributy – například axiom:

$\text{mrtvých_alespoň}(\text{Nehoda}, \text{Počet}) \leftarrow \text{mrtvých}(\text{Nehoda}, \text{Počet2}), \text{Počet} \leq \text{Počet2}$

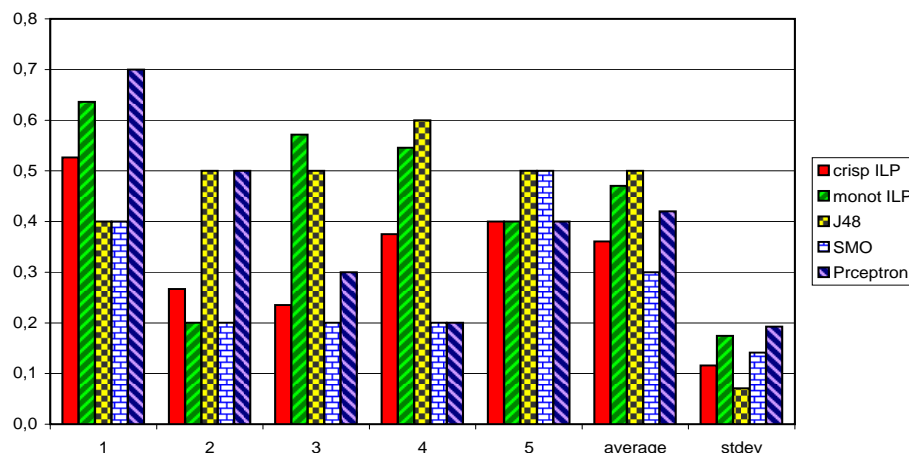
a v takzvaném zmonotonizování učicích příkladů. V případě, že cílový atribut klasifikace je monotónní (např. stupeň závažnosti nehody 1 je menší než stupně 2 a 3), se můžeme pokusit naučit místo crisp predikátu ($\text{závažná}(\text{Nehoda}, \text{Stupeň})$) odpovídající monotonizovaný predikát ($\text{závažná_alespoň}(\text{Nehoda}, \text{Stupeň})$). Induktivní procedura má potom více pozitivních učicích příkladů a monotonizované predikáty lépe vystihují povahu numerických dat. Pro porovnání s klasickými klasifikátory je však nutné výsledky monotónních pravidel převést zpět na crisp. Fuzzy ILP klasifikátor budeme dále označovat *monot* ILP, naproti tomu *crisp* ILP označuje prosté ILP bez monotonizace dat.

4 Vyhodnocení

Pro vyhodnocení úspěšnosti fuzzy ILP klasifikátoru jsme použili pětinasobnou křížovou validaci a výsledky porovnali s dalšími čtyřmi klasifikátory. Prvním klasifikátorem je prosté ILP bez monotonizace dat (*crisp* ILP). Pro srovnání jsme použili ještě další tři známé klasifikátory, které jsou součástí softwarového balíku Weka². Jmenovitě rozhodovací stromy J48 [3], Support vector machine klasifikátor SMO [4] a vícevrstvý perceptron [5].

	precision		recall		f-measure	
	average	stdev	average	stdev	average	stdev
crisp ILP	0,428	0,104	0,320	0,130	0,361	0,116
monot ILP	0,446	0,154	0,500	0,200	0,471	0,174
J48	0,500	0,071	0,500	0,071	0,500	0,071
SMO	0,300	0,141	0,300	0,141	0,300	0,141
perceptron	0,420	0,192	0,420	0,192	0,420	0,192

Obrázek 1. Úspěšnosti jednotlivých klasifikátorů (průměr a směrodatná odchylka)



Obrázek 2. Hodnoty f-míry během křížové validace, průměr a směrodatná odchylka.

Naměřené úspěšnosti shrnuje tabulka na obrázku 1. Pro vyhodnocení jsme použili standardní míry používané v information retrieval³. Podrobnější analýza hodnot f-míry (harmonického průměru přesnosti a úplnosti) v jednotlivých fázích křížové validace je znázorněna na obrázcích 2 a 3. Výsledky jednotlivých klasifikátorů jsou si velmi blízké. Z průměrných hodnot f-míry a jejich směrodatných odchylek zjistíme, že pro statisticky významný závěr bychom potřebovali víc dat⁴. Nicméně s pomocí kvantilové analýzy (obr. 3) je možné odhadnout, že *monot* ILP dává lepší výsledky než *crisp* ILP a SMO. Relativně špatné výsledky SMO klasifikátoru jsou však pravděpodobně způsobeny vysokým množstvím vstupních atributů, které by bylo vhodné před SMO analýzou zredukovat. S tímto problémem si naopak velmi dobře poradily rozhodovací stromy J48, jejichž výsledky byly nejen vysoké, ale i stabilní.

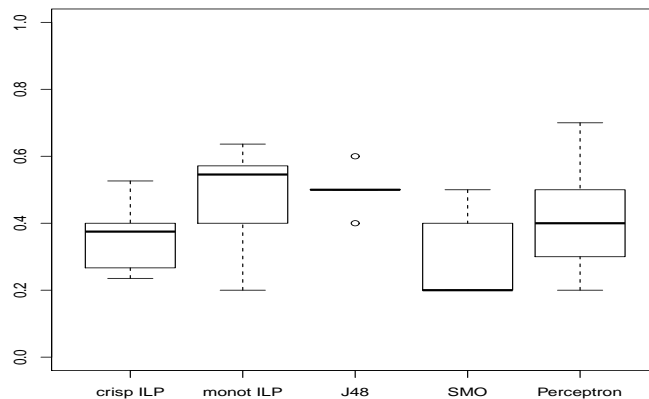
Poděkování

Tato práce byla částečně podpořena českými projekty GAČR-201/09/H057, GAUK 31009 a MSM-0021620838.

² <http://www.cs.waikato.ac.nz/ml/weka/>

³ http://en.wikipedia.org/wiki/Information_retrieval

⁴ Pro jednoduchost uvažujeme normální rozdělení. Pro konfidenci 0,95 bude hranice konfidenčního intervalu vzdálena $1,65 * stdev$ od průměrné hodnoty. Pro f-míru a *monot* ILP dostáváme 0,1839 jako spodní hranici konfidenčního intervalu, což je daleko pod hodnotami ostatních klasifikátorů.



Obrázek 3. Kvantilová analýza f-míry (minimum, $Q_{0,25}$, medián, $Q_{0,75}$, maximum)

Reference

1. Dědek, J., Vojtáš, P.: Linguistic extraction for semantic annotation. In Badica, C., Mangioni, G., Carchiolo, V., Burdescu, D., eds.: 2nd International Symposium on Intelligent Distributed Computing. Volume 162 of Studies in Computational Intelligence., Catania, Italy, Springer-Verlag (2008) 85–94
2. Dědek, J., Vojtáš, P.: Fuzzy classification of web reports with linguistic text mining. In Boldi, P., Vizzari, G., Pasi, G., Baeza-Yates, R., eds.: Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference on. Volume 3., Los Alamitos, CA, USA, IEEE Computer Society (2009) 167–170
3. Quinlan, J.R.: C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1993)
4. Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., Murthy, K.R.K.: Improvements to platt's smo algorithm for svm classifier design. Neural Computation **13**(3) (2001) 637–649
5. Bishop, C.M.: Neural Networks for Pattern Recognition. 1 edn. Oxford University Press (January 1996)

Annotation:

Evaluation of fuzzy ILP classifier on data about car accidents

Our motivation is to classify web reports about car and other accidents according to its seriousness. For this purpose we have used a new classification method based on inductive logic programming and fuzzy logic. In this paper we present a comparison of this method with other classifiers like J48 decision trees, SVM and multilayer perceptron. Our results show that on our dataset the fuzzy ILP method is fully comparable with the alternatives.