Povzetek

V diplomskem delu je opisan problem napovednega strojnega učenja in izločanja šuma iz množice učnih primerov z namenom doseganja večje klasifikacijske točnosti učnih algoritmov na novih, še ne videnih primerih. V ta namen je predstavljen in implementiran saturacijski filter, ki temelji na teoretični podlagi izhajajoči iz principa Occamove britve. Predstavljene so tudi teoretične osnove potrebne za razumevanje omenjene problematike, od osnovnih pojmov podatkovnega rudarjenja preko napovednega strojnega učenja in učnih algoritmov do določanja klasifikacijke točnosti in testov za statistično primerjavo uspešnosti napovedovanja učnih algoritmov. Po obravnavi rezultatov dobljenih s testiranjem saturacijskega filtra so predstavljeni še predlogi za nadaljnje delo, ki bi koristili večji in širši uporabnosti saturacijskega filtra.

Na priloženi zgoščenki je programska koda algoritma saturacijskega filtra v programskem jeziku Python, datoteke vseh učnih množic, na katerih je bil algoritem testiran, in datoteke z rezultati vseh opravljenih testiranj. Izdelali smo tudi spletno aplikacijo [24], preko katere je saturacijski filter splošno dostopen na internetu.

Math. Subj. Class. (MSC 2000): `62-07`, `68T30`, `68T37`

Ključne besede:
podatkovno rudarjenje, napovedno strojno učenje, klasifikacijska točnost, odločitvena drevesa, odločitvena pravila, naivni Bayesov klasifikator, princip Occamove britve, filtriranje šuma, saturacijski filter

Keywords:
data mining, predictive machine learning, classification accuracy, decision trees, decision rules, naive Bayes classifier, Occam's razor principle, noise filtering, saturation filter

Literatura

[1] A. Asuncion, D.J. Newman, *UCI Machine Learning Repository*, `http://www.ics.uci.edu/~mlearn/MLRepository.html`, 2007.

[2] L. Breiman, J. Friedman, R. Olshen, C. Stone, *Classification and Regression Trees*, Wadsworth & Brooks, Pacific Grove, CA, 1984.

[3] C. E. Brodley, M. A. Friedl, *Identifying and eliminating mislabeled training instances*, Proceedings of the Thirteenth National Conference on Artificial Intelligence, 1996.

[4] C. A. Brunk, M. J. Pazzani, *An investigation of noise-tolerant relational concept learning algorithms*, Proceedings of the 8th International Workshop on Machine Learning (ML-91), Morgan Kaufmann, 1991, str. 389–393.

[5] P. Clark, R. Boswell, *Rule induction with CN2: Some recent improvements.*, Proceedings of the 5th European Working Session on Learning (EWSL-91), Springer-Verlag, 1991, str. 151–163.

[6] P. Clark, T. Niblett, *The CN2 induction algorithm*, Machine Learning **3** (1989), 261–283.

[7] W. W. Cohen, *Efficient pruning methods for separate-and-conquer rule learning systems*, Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI-93), Morgan Kaufmann, 1993, str. 988–994.

[8] J. Demšar, *Statistical Comparisons of Classifiers over Multiple Data Sets*, Journal of Machine Learning Research **7** (2006), 1–30.

[9] J. Demšar, B. Zupan, G. Leban, *Orange: From Experimental Machine Learning to Interactive Data Mining*, `http://www.ailab.si/orange`, 2004, Fakulteta za računalništvo in informatiko, Univerza v Ljubljani.

[10] J. Fürnkranz, *FOSSIL: A robust relational learner*, Lecture Notes in Artificial Intelligence: Machine Learning: ECML-94 (F. Bergadano, L. De Raedt, ur.), vol. 784, Springer-Verlag, 1994, str. 122–137.

[11] J. Fürnkranz, *Pruning algorithms for rule learning*, Machine Learning **27** (1997), 139–171.

[12] J. Fürnkranz, G. Widmer, *Incremental Reduced Error Pruning*, Proceedings of the 11th International Conference on Machine Learning (ML-94) (W. W. Cohen, H. Hirsh, ur.), Morgan Kaufmann, 1994, str. 70–77.

[13] D. Gamberger, N. Lavrač, *Conditions for Occam's Razor Applicability and Noise Elimination*, Lecture Notes in Artificial Intelligence: Machine Learning: ECML-97 (M. Van Someren, G. Widmer, ur.), vol. 1224, Springer-Verlag, 1997, str. 108–123.

[14] D. Gamberger, N. Lavrač, S. Džeroski, *Noise elimination in inductive concept learning: A case study in medical diagnosis*, Proceedings of the Seventh International Workshop on Algorithmic Learning Theory, 1996, str. 199–212.

[15] S. Gelfand, C. Ravishankar, E. Delp, *An iterative growing and pruning algorithm for classification tree design*, IEEE Transactions on Pattern Analysis and Machine Intelligence **13** (1991), 163–174.

[16] G. H. John, *Robust decision trees: Removing outliers from databases*, Proceedings of the First International Conference on Knowledge Discovery and Data Mining, 1995, str. 174–179.

[17] I. Kononenko, *Strojno učenje*, Založba FE in FRI, Ljubljana, 1997.

[18] I. Kononenko, M. Kukar, *Machine Learning and Data Mining*, Horwood Publishing Limited, Chichester, 2007.

[19] J. Mingers, *An empirical comparison of pruning methods for decision tree induction*, Machine Learning **4** (1989), 227–243.

[20] T. Niblett, I. Bratko, *Learning decision rules in noisy domains*, Research and Development in Expert Systems (M. Bramer, ur.), Cambridge University Press, 1987.

[21] J. R. Quilan, *Induction of Decision Trees*, Machine Learning **1** (1986), 81–106.

[22] J. R. Quilan, *Simplifying decision trees*, International Journal of Man-Machine Studies **27** (1987), 221–234.

[23] J. R. Quilan, *Learning logical definitions from relations*, Machine Learning **5** (1990), 239–266.

[24] B. Sluban, V. Podpečan, *Saturation Filter: Noise Filtering Algorithm*, `http://zulu.ijs.si/web/SaturationFilter/`, maj 2009, Spletna aplikacija.

[25] C. M. Teng, *Correcting noisy data*, Proceedings of the Sixteenth International Conference on Machine Learning, 1999, str. 239–248.

[26] C. M. Teng, *Evaluation noise correction*, Lecture Notes in Artificial Intelligence: Proceedings of the Sixth Pacific Rim International Conference on Artificial Intelligence, Springer-Verlag, 2000, str. 269–273.

[27] C. M. Teng, *A Comparison of Noise Handling Techniques*, FLAIRS-01 Proceedings, AAAI Press, 2001, str. 269–273.

[28] H. Theron, I. Cloete, *BEXA: A covering algorithm for learning propositional concept descriptions*, Machine Learning **24** (1996), 5–40.