Klasifikacija mina i kamenja iz sonar dataseta

Alen Andrašek, Monika Majstorović, Luka Valenta

Opis problema

- **Dataset: sonar** (UCI repozitorij za SU)
  - 208 podataka (111 mina, 97 kamenja)
  - 60 featurea (vrijednost između 0.0 i 1.0)
  - Svaki od tih 60 brojeva predstavlja energiju unutar određenog frekvencijskog pojasa, integriranu tijekom određenog vremenskog razdoblja.
  - Često korišten dataset za razne edukacijsko-edukativne svrha i benchmark novih modela u SU.

- **Cilj:**

  Želimo trenirati što bolji model za klasifikaciju sonar dataseta. Konkretno, uspoređivati performanse SVM-a i k-NN-a međusobno te u usporedbi s performansama u dosadašnjim istraživanjima.

Prethodna istraživanja

- Gray wolf classifier (GWC) (2016.)
  - GWO je meta-heuristika inspirirana hijerarhijskom organizacijom čopora i načinom lova sivog vuka.
  - Smanjenje dimenzionalnosti: PCA (sa 60 na 9).
  - Uspoređivali su rezultate s PSO, GSA, PSOGSA. GWC je bio najbolji.
- Online multiple kernel klasifikacija (OMKC) (2013.)
  - Klasificiraju sonar dataset koristeći online multiple kernel klasifikaciju s deterministčkim/stohastičkim ažuriranjem i deterministčkom/stohastičkom kombinacijom algoritmama.
  - Sve kombinacije koristeći OMKC imale su grešku od 24% do 26%.
- Ansambli neuronskih mreža (2010.)
  - *Ansambl* je skup nezavisno treniranih klasifikatora čija predviđanja se kombiniraju nekim statističkim metodama. Ovdje se koristi *bagging* (bootstrap aggregating).
  - Promatrane su 3-bagged i 25-bagged back propagation neuralne mreže i uspoređeno je s „običnim" neuralnim mrežama. 25-bagged se pokazao kao najbolji.

Plan istraživanja

- Smanjenje dimenzionalnosti: PCA
- Podijelit ćemo dataset na 13 disjunktnih podskupova (+stratifikacija).
- Cross-validacija: nested cross-validacija
  - vanjska: odabir modela
  - unutarnja: optimizacija hiperparametara (10-fold cross-validacija + random search)
- O modelima:
  - Koristit ćemo **SVM** s kernelom RBF (Gaussian Radial Basis Function).
  - U **k-NN** modelu biramo između Euklidske, Manhattan, Čebiševljeve i chi-kvadrat udaljenosti.
- Mjera performansa: classification error
- Programski jezik: Python

Literatura

- [1] M.R. Mosavi, M. Khishe, A. Ghamgosar; *Classification of Sonar Dataset Using Neural Network Trained by Grey Wolf Optimization*; Iran University of Science and Technology, Teheran; Iran, 2016.
- [2] R.P. Gorman, T. J. Stejnowski; *Analysis of Hidden Units in a Layered Network Trained to Classify Sonar Targets*; 1987.
- [3] R. K. Jade, L. K. Verma, K. Verma; *Classification using Neural Network and Support Vector Machine for Sonar dataset*; International Journal of Computer Trends and Technology; Vol. 4, Issue 2; pg 116-119; 2013.
- [4] H. T. Hassan, M. U. Khalid, K. Imran; Intelligent Object and Pattern Recognition using Ensembles in Back Propagation Neural Network; International Journal of Electrical & Computer Sciences IJECS-IJENS; Vol. 10, No. 6; pg 52-59; 2010.
- [5] D. Meyer, F. Leisch, K. Hornik; Benchmarking Support Vector Machines; 2002.
- [6] https://medium.com/rants-on-machine-learning/what-to-do-with-small-data d253254d1a89 (Zadnje pristupljeno: 30. travnja 2018.)
- [6] https://pdfs.semanticscholar.org/d6dc/df86df3ece94c2c5effe205d105c561ed5eb.pdf (Zadnje pristupljeno: 30. travnja 2018.)
- [7] https://stats.stackexchange.com/questions/117643/why-use-stratified-cross-validation-why-does-this-not-damage-variance-related-b (Zadnje pristupljeno: 30. travnja 2018.)
- [8] http://cs229.stanford.edu/notes/cs229-notes5.pdf (Zadnje pristupljeno: 30. travnja 2018.)
- [9] J. Bergstra, Y. Bengio; *Random Search for Hyper-Parameter Optimization;* Journal of Machine Learning Research 13, pg. 281-305; 2012.
- [10] Kilian Q. et al., "Distance Metric Learning for Large Margin Nearest Neighbor Classification", Journal of Machine Learning Research 10, pg. 207-244., 2009.
- [11] https://archive.ics.uci.edu/ml/datasets/connectionist+bench+(sonar,+mines+vs.+rocks) (Zadnje pristupljeno: 30. travnja 2018.)
- [12] https://stats.stackexchange.com/questions/61546/optimal-number-of-folds-in-k-fold-cross-validation-is-leave-one-out-cv-always (Zadnje pristupljeno: 30. travnja 2018.)