Gul, Asma; **Perperoglou, Aris**; **Khan, Zardad**; **Mahmoud, Osama**; **Miftahuddin, Miftahuddin**; **Adler, Werner**; **Lausen, Berthold**

Ensemble of a subset of kNN classifiers. (English) Zbl 1416.62338

Adv. Data Anal. Classif., ADAC 12, No. 4, 827-840 (2018).

Summary: Combining multiple classifiers, known as ensemble methods, can give substantial improvement in prediction performance of learning algorithms especially in the presence of non-informative features in the data sets. We propose an ensemble of subset of kNN classifiers, ESkNN, for classification task in two steps. Firstly, we choose classifiers based upon their individual performance using the out-of-sample accuracy. The selected classifiers are then combined sequentially starting from the best model and assessed for collective performance on a validation data set. We use bench mark data sets with their original and some added non-informative features for the evaluation of our method. The results are compared with usual kNN, bagged kNN, random kNN, multiple feature subset method, random forest and support vector machines. Our experimental comparisons on benchmark classification problems and simulated data sets reveal that the proposed ensemble gives better classification performance than the usual kNN and its ensembles, and performs comparable to random forest and support vector machines.

MSC:

62H30 Classification and discrimination; cluster analysis (statistical aspects)
68T05 Learning and adaptive systems in artificial intelligence

Keywords:

ensemble methods; bagging; nearest neighbour classifier; non-informative features

Software:

R; ESKNN; e1071; ipred; UCI-ml; Kernlab; mlbench

Full Text: DOI

References:

[1] Bache K, Lichman M (2013) UCI machine learning repository. http://archive.ics.uci.edu/ml. Accessed 3 October 2014

[2] Barandela, R.; Valdovinos, RM; Sánchez, JS, New applications of ensembles of classifiers, Pattern Anal Appl, 6, 245-256, (2013)

[3] Bauer, E.; Kohavi, R., An empirical comparison of voting classification algorithms: bagging, boosting, and variants, Mach Learn, 36, 105-139, (1999)

[4] Bay S (1998) Combining nearest neighbor classifiers through multiple feature subsets. In: Proceedings of the Fifteenth International Conference on Machine Learning, vol 3. Morgan Kaufmann Publishers Inc., pp 37-45

[5] Breiman, L., Bagging predictors, Mach Learn, 24, 123-140, (1996) · Zbl 0858.68080

[6] Breiman L (1996b) Out-of-bag estimation. Tech. rep. http://citeseerx.ist.psu.edu. Accessed 3 October 2014

[7] Cannings T, Samworth R (2015) Random projection ensemble classification. arXiv:1504.04595v1.pdf. Accessed 3 October 2015 · Zbl 1373.62301

[8] Cover, T.; Hart, P., Nearest neighbor pattern classification, IEEE Trans Inf Theory, 13, 21-27, (1967) · Zbl 0154.44505

[9] Domeniconi C, Yan B (2004) Nearest neighbor ensemble. In: IEEE Proceedings of the 17th International Conference on Pattern Recognition (ICPR 2004), vol 1, pp 228-231

[10] Grabowski S (2002) Voting over multiple k-nn classifiers. In: Proceedings of the International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science IEEE, pp 223-225

[11] Gul A, Perperoglou A, Khan Z, Mahmoud O, Adler W, Miftahuddin M, Lausen B (2015) R package: ESKNN: ensemble of subset of K-nearest neighbours classifiers for classification and class membership probability estimation. http://cran.r-project.org/web/packages/ESKNN/index.html. Accessed 30 Sept 2015

[12] Guvenir HA, Akkus A (1997) Weighted k nearest neighbor classification on feature projections. http://www.cs.bilkent.edu.tr/tech-reports/1997/BU-CEIS-9719.pdf. Accessed 3 October 2014

[13] Hall, P.; Samworth, R., Properties of bagged nearest neighbour classifiers, J R Stat Soc Ser B (Statistical Methodology), 67, 363-379, (2005) · Zbl 1069.62051

[14] Hernández-Orallo, J.; Flach, P.; Ferri, C., A unified view of performance metrics: Translating threshold choice into expected classification loss, J Mach Learn Res, 13, 2813-2869, (2012) · Zbl 1436.62260

[15] Hothorn, T.; Lausen, B., Bagging tree classifiers for laser scanning images: a data-and simulation-based strategy, Artif Intell Med, 27, 65-79, (2003)

[16] Hothorn, T.; Lausen, B., Double-bagging: combining classifiers by bootstrap aggregation, Pattern Recognit, 36, 1303-1309, (2003) · Zbl 1028.68144

[17] Hothorn, T.; Lausen, B., Bundling classifiers by bagging trees, Comput Stat Data Anal, 49, 1068-1078, (2005) · Zbl 1429.62246

[18] Hothorn, T.; Lausen, B.; Benner, A.; Radespiel-TrÂĺoger, M., Bagging survival trees, Stat Med, 23, 77-91, (2004)

[19] Karatzoglou, A.; Smola, A.; Hornik, K.; Zeileis, A., kernlab—an S4 Package for Kernel Methods in R, J Stat Softw, 11, 1-20, (2004)

[20] Khoshgoftaar, T.; Hulse, J.; Napolitano, A., Comparing boosting and bagging techniques with noisy and imbalanced data, IEEE Trans Syst Man Cybern Part A Syst Hum, 41, 552-568, (2011)

[21] Kruppa, J.; Liu, Y.; Diener, HC; Holste, T.; Weimar, C.; König, IR; Ziegler, A., Probability estimation with machine learning methods for dichotomous and multicategory outcome: Applications, Biom J, 56, 564-583, (2014) · Zbl 1441.62405

[22] Lausser, L.; Müssel, C.; Melkozerov, A.; Kestler, HA, Identifying predictive hubs to condense the training set of k-nearest neighbour classifiers, Comput Stat, 29, 81-95, (2014) · Zbl 1306.65085

[23] Leisch F, Dimitriadou E (2010) mlbench: Machine Learning Benchmark Problems. R package version 2.1-1

[24] Li, S.; Harner, EJ; Adjeroh, D., Random knn feature selection-a fast and stable alternative to random forests, BMC Bioinform, 12, 450, (2011)

[25] Liu, Z.; Zhao, X.; Zuo, MJ; Xu, H., Feature selection for fault level diagnosis of planetary gearboxes, Adv Data Anal Classif, 8, 377-401, (2014)

[26] Maclin, R.; Opitz, D., Popular ensemble methods: an empirical study, J Artif Res, 11, 169-189, (2011) · Zbl 0924.68159

[27] Mahmoud, O.; Harrison, A.; Perperoglou, A.; Gul, A.; Khan, Z.; Metodiev, MV; Lausen, B., A feature selection method for classification within functional genomics experiments based on the proportional overlapping score, BMC Bioinform, 15, 274, (2014)

[28] Mease, D.; Wyner, AJ; Buja, A., Boosted classification trees and class probability/quantile estimation, J Mach Learn Res, 8, 409-439, (2007) · Zbl 1222.68261

[29] Melville, P.; Shah, N.; Mihalkova, L.; Mooney, R.; Roli, F. (ed.); Kittler, J. (ed.); Windeatt, T. (ed.), Experiments on ensembles with missing and noisy data, 293-302, (2004), Heidelberg

[30] Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F (2012) e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.6-1

[31] Müssel, C.; Lausser, L.; Kestler, HA; Lausen, B. (ed.); Krolak-Schwerdt, S. (ed.); Böhmer, M. (ed.), Ensembles of representative prototype sets for classification and data set analysis, 329-339, (2015), Heidelberg

[32] Nettleton, DF; Orriols-Puig, A.; Fornells, A., A study of the effect of different types of noise on the precision of supervised learning techniques, Artif Intell Rev, 33, 275-306, (2010)

[33] Peters A, Hothorn T (2012) ipred: Improved Predictors. http://CRAN.R-project.org/package=ipred. R package version 0.9-1

[34] R Core Team (2013) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

[35] Samworth, RJ, Optimal weighted nearest neighbour classifiers, Ann Stat, 40, 2733-2763, (2012) · Zbl 1373.62317

[36] Steyerberg, EW; Vickers, AJ; Cook, NR; Gerds, T.; Gonen, M.; Obuchowski, N.; Pencina, MJ; Kattan, MW, Assessing the performance of prediction models: a framework for some traditional and novel measures, Epidemiology, 21, 128-138, (2010)

[37] Zhou, ZH; Yu, Y., Adapt bagging to nearest neighbor classifiers, J Comput Sci Technol, 20, 48-54, (2005)