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

Ensemble of optimal trees, random forest and random projection ensemble classification.

(English) Zbl 1459.62115
Adv. Data Anal. Classif., ADAC 14, No. 1, 97-116 (2020).

Summary: The predictive performance of a random forest ensemble is highly associated with the strength of individual trees and their diversity. Ensemble of a small number of accurate and diverse trees, if prediction accuracy is not compromised, will also reduce computational burden. We investigate the idea of integrating trees that are accurate and diverse. For this purpose, we utilize out-of-bag observations as a validation sample from the training bootstrap samples, to choose the best trees based on their individual performance and then assess these trees for diversity using the Brier score on an independent validation sample. Starting from the first best tree, a tree is selected for the final ensemble if its addition to the forest reduces error of the trees that have already been added. Our approach does not use an implicit dimension reduction for each tree as random project ensemble classification. A total of 35 bench mark problems on classification and regression are used to assess the performance of the proposed method and compare it with random forest, random projection ensemble, node harvest, support vector machine, kNN and classification and regression tree. We compute unexplained variances or classification error rates for all the methods on the corresponding data sets. Our experiments reveal that the size of the ensemble is reduced significantly and better results are obtained in most of the cases. Results of a simulation study are also given where four tree style scenarios are considered to generate data sets with several structures.

MSC:

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

Keywords:

ensemble classification; ensemble regression; random forest; random projection ensemble classification; accuracy and diversity

Software:

gclus; OTE; RPEnsemble; T3C; e1071; nodeHarvest; mlbench; ipred; UCI-ml; propOverlap; R; penalized; ElemStatLearn; Kernlab; C4.5

Full Text: DOI

References:

[1] Adler, W.; Peters, A.; Lausen, B., Comparison of classifiers applied to confocal scanning laser ophthalmoscopy data, Methods Inf Med, 47, 1, 38-46 (2008)· doi:10.3414/ME0348

[2] Adler, W.; Gefeller, O.; Gul, A.; Horn, Fk; Khan, Z.; Lausen, B., Ensemble pruning for glaucoma detection in an unbalanced data set, Methods Inf Med, 55, 6, 557-563 (2016)· doi:10.3414/ME16-01-0055

[3] Ali, K.; Pazzani, M., Error reduction through learning multiple descriptions, Mach Learn, 24, 3, 173-202 (1996)

[4] Bache K, Lichman M (2013) UCI machine learning repository. http://archive.ics.uci.edu/ml

[5] Bachrach, Lk; Hastie, T.; Wang, Mc; Narasimhan, B.; Marcus, R., Bone mineral acquisition in healthy asian, hispanic, black, and caucasian youth: a longitudinal study, J Clin Endocrinol Metab, 84, 12, 4702-4712 (1999)

[6] Bauer, E.; Kohavi, R., An empirical comparison of voting classification algorithms: bagging, boosting, and variants, Mach Learn, 36, 1, 105-139 (1999)· doi:10.1023/A:1007515423169

[7] Bernard S, Heutte L, Adam S (2009) On the selection of decision trees in random forests. In: International joint conference on neural networks, IEEE, pp 302-307

[8] Bhardwaj, M.; Bhatnagar, V.; Sharma, K., Cost-effectiveness of classification ensembles, Pattern Recognit, 57, 84-96 (2016)· doi:10.1016/j.patcog.2016.03.017

[9] Bolón-Canedo, V.; Sánchez-Maroño, N.; Alonso-Betanzos, A., An ensemble of filters and classifiers for microarray data classification, Pattern Recognit, 45, 1, 531-539 (2012)· doi:10.1016/j.patcog.2011.06.006

[10] Brahim, Ab; Limam, M., Ensemble feature selection for high dimensional data: a new method and a comparative study, Adv Data Anal Classif, 12, 1-16 (2017)

[11] Breiman, L., Random forests, Mach Learn, 45, 1, 5-32 (2001) · Zbl 1007.68152 · doi:10.1023/A:1010933404324

[12] Brier, Gw, Verification of forecasts expressed in terms of probability, Mon Weather Rev, 78, 1, 1-3 (1950)· doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

[13] Buta, R., The structure and dynamics of ringed galaxies. iii-surface photometry and kinematics of the ringed nonbarred spiral ngc 7531, Astrophys J Suppl Ser, 64, 1-37 (1987)· doi:10.1086/191190

[14] Cannings TI, Samworth RJ (2016) RPEnsemble: Random Projection Ensemble Classification. https://CRAN.R-project.org/package=RPEnsemble, r package version 0.3

[15] Cannings, Ti; Samworth, Rj, Random-projection ensemble classification, J R Stat Soc Ser B (Stat Methodol), 79, 4, 959-1035 (2017) · Zbl 1373.62301 · doi:10.1111/rssb.12228

[16] Domingos P (1996) Using partitioning to speed up specific-to-general rule induction. In: Proceedings of the AAAI-96 workshop on integrating multiple learned models, Citeseer, pp 29-34

[17] Friedman, Jh, Multivariate adaptive regression splines, Ann Stat, 19, 1-67 (1991) · Zbl 0765.62064 · doi:10.1214/aos/1176347963

[18] Goeman JJ (2012) penalized: Penalized generalized linear models. http://CRAN.R-project.org/package=penalized, penalized R package, version 0.9-42

[19] Gul A, Khan Z, Perperoglou A, Mahmoud O, Miftahuddin M, Adler W, Lausen B (2016a) Ensemble of subset of k-nearest neighbours models for class membership probability estimation. In: Analysis of large and complex data, Springer, pp 411-421 · Zbl 1416.62338

[20] Gul, A.; Perperoglou, A.; Khan, Z.; Mahmoud, O.; Miftahuddin, M.; Adler, W.; Lausen, B., Ensemble of a subset of knn classifiers, Adv Data Anal Classif, 12, 1-14 (2016)

[21] Halvorsen K (2012) ElemStatLearn: Data sets, functions and examples. http://CRAN.R-project.org/package=ElemStatLearn, r package version 2012.04-0

[22] Hapfelmeier, A.; Ulm, K., A new variable selection approach using random forests, Comput Stat Data Anal, 60, 50-69 (2013) · Zbl 1365.62417 · doi:10.1016/j.csda.2012.09.020

[23] Hothorn, T.; Lausen, B., Double-bagging: combining classifiers by bootstrap aggregation, Pattern Recognit, 36, 6, 1303-1309 (2003) · Zbl 1028.68144 · doi:10.1016/S0031-3203(02)00169-3

[24] Hurley C (2012) gclus: Clustering Graphics. http://CRAN.R-project.org/package=gclus, r package version 1.3.1

[25] Janitza, S.; Celik, E.; Boulesteix, Al, A computationally fast variable importance test for random forests for high-dimensional data, Adv Data Anal Classif, 12, 1-31 (2015)

[26] Karatzoglou A, Smola A, Hornik K, Zeileis A (2004) kernlab—an S4 package for kernel methods in R. Journal of Statistical Software 11(9):1-20, http://www.jstatsoft.org/v11/i09/

[27] Khan Z, Gul A, Perperoglou A, Mahmoud O, Werner Adler M, Lausen B (2014) OTE: Optimal Trees Ensembles. https://cran.r-project.org/package=OTE, r package version 1.0

[28] Khan Z, Gul A, Mahmoud O, Miftahuddin M, Perperoglou A, Adler W, Lausen B (2016) An ensemble of optimal trees for class membership probability estimation. In: Analysis of large and complex data, Springer, pp 395-409 · Zbl 1416.62338

[29] Latinne P, Debeir O, Decaestecker C (2001a) Limiting the number of trees in random forests. In: Multiple Classifier Systems: Second International Workshop, MCS 2001 Cambridge, UK, July 2-4, 2001 Proceedings, Springer Science \& Business Media, vol 2, p 178 · Zbl 0987.68896

[30] Latinne P, Debeir O, Decaestecker C (2001b) Limiting the number of trees in random forests. Multiple Classifier Systems pp 178-187 · Zbl 0987.68896

[31] Lausser, L.; Schmid, F.; Schirra, Lr; Wilhelm, Af; Kestler, Ha, Rank-based classifiers for extremely high-dimensional gene expression data, Adv Data Anal Classif, 12, 1-20 (2016)

[32] Leisch F, Dimitriadou E (2010) mlbench: Machine learning benchmark problems. R package version 2.1-1

[33] Li HB, Wang W, Ding HW, Dong J (2010) Trees weighting random forest method for classifying high-dimensional noisy data. In: IEEE 7th international conference on e-business engineering (ICEBE), 2010, IEEE, pp 160-163

[34] Liberati, C.; Camillo, F.; Saporta, G., Advances in credit scoring: combining performance and interpretation in kernel discriminant analysis, Adv Data Anal Classif, 11, 1, 121-138 (2017) · Zbl 1414.62421 · doi:10.1007/s11634-015-0213-y

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

[36] Mahmoud O, Harrison A, Perperoglou A, Gul A, Khan Z, Lausen B (2014a) propOverlap: Feature (gene) selection based on the Proportional Overlapping Scores. http://CRAN.R-project.org/package=propOverlap, r package version 1.0

[37] 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 Bioinf, 15, 1, 274 (2014)· doi:10.1186/1471-2105-15-274

[38] Meinshausen, N., Node harvest, Ann Appl Stat, 4, 4, 2049-2072 (2010) · Zbl 1220.62084 · doi:10.1214/10-AOAS367

[39] Meinshausen N (2013) nodeHarvest: Node Harvest for regression and classification. http://CRAN.R-project.org/package=nodeHarvest, r package version 0.6

[40] Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F (2014) e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. http://CRAN.R-project.org/package=e1071, r package version 1.6-4

[41] Mitchell, T., Machine learning (1997), Burr Ridge: McGraw Hill, Burr Ridge · Zbl 0913.68167

[42] Oshiro, Thais Mayumi; Perez, Pedro Santoro; Baranauskas, José Augusto, How Many Trees in a Random Forest?, Machine Learning and Data Mining in Pattern Recognition, 154-168 (2012), Berlin, Heidelberg: Springer Berlin Heidelberg, Berlin, Heidelberg

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

[44] Quinlan J (1996) Bagging, boosting, and c4. 5. In: Proceedings of the national conference on artificial intelligence, pp 725-730

[45] R Core Team (2014) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/

[46] Schapire, R., The strength of weak learnability, Mach Learn, 5, 2, 197-227 (1990)

[47] Tumer, K.; Ghosh, J., Error correlation and error reduction in ensemble classifiers, Connect Sci, 8, 3-4, 385-404 (1996)· doi:10.1080/095400996116839

[48] Tzirakis, P.; Tjortjis, C., T3c: improving a decision tree classification algorithm's interval splits on continuous attributes, Adv Data Anal Classif, 11, 2, 353-370 (2017) · Zbl 1414.68081 · doi:10.1007/s11634-016-0246-x

[49] Zhang, H.; Wang, M., Search for the smallest random forest, Stat Interface, 2, 3, 381-388 (2009) · Zbl 1245.62058 · doi:10.4310/SII.2009.v2.n3.a11