Puliparambil, Bhavithry Sen; **Tomal, Jabed**; **Yan, Yan**

Benchmarking penalized regression methods in machine learning for single cell RNA sequencing data. (English) Zbl 1496.92075

Jin, Lingling (ed.) et al., Comparative genomics. 19th international conference, RECOMB-CG 2022, La Jolla, CA, USA, May 20–21, 2022. Proceedings. Cham: Springer. Lect. Notes Comput. Sci. 13234, 295-310 (2022).

Summary: Single cell RNA sequencing (scRNA-seq) technology has enabled the biological research community to explore gene expression at a single-cell resolution. By studying differences in gene expression, it is possible to differentiate cell clusters and types within tissues. One of the major challenges in a scRNA-seq study is feature selection in high dimensional data. Several statistical and machine learning algorithms are available to solve this problem, but their performances across data sets lack systematic comparison. In this research, we benchmark different penalized regression methods, which are suitable for scRNA-seq data. Results from four different scRNA-seq data sets show that sparse group lasso (SGL) implemented by the SGL package in R performs better than other methods in terms of area under the receiver operating curve (AUC). The computation time for different algorithms varies between data sets with SGL having the least average computation time. Based on our findings, we propose a new method that applies SGL on a smaller pre-selected subset of genes to select the differentially expressed genes in scRNA-seq data. The reduction in the number of genes before SGL reduce the computation hardware requirement from 32 GB RAM to 8 GB RAM. The proposed method also demonstrates a consistent improvement in AUC over SGL.

For the entire collection see [Zbl 1492.92002].

MSC:

92D20 Protein sequences, DNA sequences
68T05 Learning and adaptive systems in artificial intelligence
62P10 Applications of statistics to biology and medical sciences; meta analysis

Keywords:

single cell RNA sequencing; machine learning; LASSO; feature selection; high dimensional data; R

Software:

AS 136; msgl; SGL; seagull; DropLasso; HieRFIT; biglasso; R

Full Text: DOI

References:

[1] Slovin, S., et al.: Single-cell RNA sequencing analysis: a step-by-step overview. RNA Bioinform. 343-365 (2021). doi:10.1007/978-1-0716-1307-8_19

[2] Kiselev, VY; Andrews, TS; Hemberg, M., Challenges in unsupervised clustering of single-cell RNA-seq data, Nat. Rev. Genet., 20, 5, 273-282 (2019)· doi:10.1038/s41576-018-0088-9

[3] Kaymaz, Y., Ganglberger, F., Tang, M., Fernandez-Albert, F., Lawless, N., Sackton, T.B.: HieRFIT: Hierarchical Random Forest for Information Transfer. bioRxiv (2020). doi:10.1101/2020.09.16.300822

[4] Pouyan, MB; Kostka, D., Random forest based similarity learning for single cell RNA sequencing data, Bioinformatics, 34, 13, i79-i88 (2018)· doi:10.1093/bioinformatics/bty260

[5] Chen, X.W., Jeong, J.C.: Enhanced recursive feature elimination. In: Sixth International Conference on Machine Learning and Applications (ICMLA 2007), pp. 429-435. IEEE (2007)

[6] Tibshirani, R., Regression shrinkage and selection via the lasso: a retrospective, J. Roy. Stat. Soc. Ser. B (Stat. Methodol.), 73, 3, 273-282 (2011) · Zbl 1411.62212 · doi:10.1111/j.1467-9868.2011.00771.x

[7] Zou, H.; Hastie, T., Regularization and variable selection via the elastic net, J. Roy. Stat. Soc. Ser. B (Stat. Methodol.), 67, 2, 301-320 (2005) · Zbl 1069.62054 · doi:10.1111/j.1467-9868.2005.00503.x

[8] Khalfaoui, B., Vert, J.P.: DropLasso: a robust variant of Lasso for single cell RNA-seq data. arXiv preprint arXiv:1802.09381

(2018)

[9] Yuan, M.; Lin, Y., Model selection and estimation in regression with grouped variables, J. Roy. Stat. Soc. Ser. B (Stat. Methodol.), 68, 1, 49-67 (2006) · Zbl 1141.62030 · doi:10.1111/j.1467-9868.2005.00532.x

[10] Simon, N.; Friedman, J.; Hastie, T.; Tibshirani, R., A sparse-group lasso, J. Comput. Graph. Stat., 22, 2, 231-245 (2013)· doi:10.1080/10618600.2012.681250

[11] Zeng, Y., Breheny, P.: The biglasso package: a memory-and computation-efficient solver for lasso model fitting with big data in R. arXiv preprint arXiv:1701.05936 (2017)

[12] Tibshirani, R.; Saunders, M.; Rosset, S.; Zhu, J.; Knight, K., Sparsity and smoothness via the fused lasso, J. Roy. Stat. Soc. Ser. B (Stat. Methodol.), 67, 1, 91-108 (2005) · Zbl 1060.62049 · doi:10.1111/j.1467-9868.2005.00490.x

[13] Zou, H., The adaptive lasso and its oracle properties, J. Am. Stat. Assoc., 101, 476, 1418-1429 (2006) · Zbl 1171.62326 · doi:10.1198/016214506000000735

[14] Jiang, Y.; He, Y.; Zhang, H., Variable selection with prior information for generalized linear models via the prior lasso method, J. Am. Stat. Assoc., 111, 513, 355-376 (2016)· doi:10.1080/01621459.2015.1008363

[15] Scialdone, A., Computational assignment of cell-cycle stage from single-cell transcriptome data, Methods, 85, 54-61 (2015)· doi:10.1016/j.ymeth.2015.06.021

[16] Cao, X., Xing, L., Majd, E., He, H., Gu, J., Zhang, X.: A systematic evaluation of methods for cell phenotype classification using single-cell RNA sequencing data. arXiv preprint arXiv:2110.00681 (2021)

[17] Zou, H.; Hastie, T., Regression shrinkage and selection via the elastic net, with applications to microarrays, JR Stat. Soc. Ser. B, 67, 301-20 (2003) · Zbl 1069.62054 · doi:10.1111/j.1467-9868.2005.00503.x

[18] Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R., Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res., 15, 1, 1929-1958 (2014) · Zbl 1318.68153

[19] Rani, Y., Rohil, H.: A study of hierarchical clustering algorithm. ter S on Te SIT 2, 113 (2013)

[20] Hartigan, J.A., Wong, M.A.: Algorithm AS 136: a K-means clustering algorithm. J. Roy. Stat. Soc. Ser. C (Appl. Stat.) 28(1), 100-108 (1979). doi:10.2307/2346830 · Zbl 0447.62062

[21] Hua, J.; Liu, H.; Zhang, B.; Jin, S., Lak: lasso and K-means based single-cell RNA-seq data clustering analysis, IEEE Access, 8, 129679-129688 (2020)· doi:10.1109/ACCESS.2020.3008681

[22] Bates, S., Hastie, T., Tibshirani, R.: Cross-validation: what does it estimate and how well does it do it? arXiv preprint arXiv:2104.00673 (2021)

[23] Park, SH; Goo, JM; Jo, CH, Receiver operating characteristic (ROC) curve: practical review for radiologists, Korean J. Radiol., 5, 1, 11-18 (2004)· doi:10.3348/kjr.2004.5.1.11

[24] Hossin, M.; Sulaiman, MN, A review on evaluation metrics for data classification evaluations, Int. J. Data Mining Knowl. Manag. Process, 5, 2, 1 (2015)· doi:10.5121/ijdkp.2015.5201

[25] Soneson, C., Robinson, M.D.: Bias, robustness and scalability in differential expression analysis of single-cell RNA-Seq data. bioRxiv, 143289 (2017)

[26] Kumar, RM, Deconstructing transcriptional heterogeneity in pluripotent stem cells, Nature, 516, 7529, 56-61 (2014)· doi:10.1038/nature13920

[27] Tasic, B., Adult mouse cortical cell taxonomy revealed by single cell transcriptomics, Nat. Neurosci., 19, 2, 335-346 (2016)· doi:10.1038/nn.4216

[28] Li, H., Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors, Nat. Genet., 49, 5, 708-718 (2017)· doi:10.1038/ng.3818

[29] Denyer, T.; Ma, X.; Klesen, S.; Scacchi, E.; Nieselt, K.; Timmermans, MC, Spatiotemporal developmental trajectories in the Arabidopsis root revealed using high-throughput single-cell RNA sequencing, Dev. Cell, 48, 6, 840-852 (2019)· doi:10.1016/j.devcel.2019.02.022

[30] Girard, A.; Sachidanandam, R.; Hannon, G., A germline-specific class of small RNAs binds mammalian Piwi proteins, Nature, 442, 199-202 (2006)· doi:10.1038/nature04917

[31] Calm2 calmodulin 2 [Mus musculus (house mouse)] [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information (2022). https://www.ncbi.nlm.nih.gov/gene/12314. Accessed 17 Jan 2022

[32] Snap25 synaptosomal-associated protein 25 [Mus musculus (house mouse)] [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information (2022). https://www.ncbi.nlm.nih.gov/gene/20614. Accessed 17 Jan 2022

[33] Fabp1 fatty acid binding protein 1, liver [Mus musculus (house mouse)] [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information (2022). https://www.ncbi.nlm.nih.gov/gene/14080. Accessed 17 Jan 2022

[34] SAT1 spermidine/spermine N1-acetyltransferase 1 [Homo sapiens (human)] [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information (2022). https://www.ncbi.nlm.nih.gov/gene/6303. Accessed 17 Jan 2022

[35] LGALS4 galectin 4 [Homo sapiens (human)] [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information (2022). https://www.ncbi.nlm.nih.gov/gene/3960. Accessed 17 Jan 2022

[36] HSP90AA1 heat shock protein 90 alpha family class A member 1 [Homo sapiens (human)] [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information (2022). https://www.ncbi.nlm.nih.gov/gene/3320. Accessed 17 Jan 2022

[37] HNRNPH1 heterogeneous nuclear ribonucleoprotein H1 [Homo sapiens (human)] [Internet]. Bethesda (MD): National Library

of Medicine (US), National Center for Biotechnology Information (2022). https://www.ncbi.nlm.nih.gov/gene/3187. Accessed 17 Jan 2022

[38] König, R., Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication, Cell, 135, 1, 49-60 (2008)· doi:10.1016/j.cell.2008.07.032

[39] Nunnari, G.; Smith, JA; Daniel, R., HIV-1 Tat and AIDS-associated cancer: targeting the cellular anti-cancer barrier?, J. Exp. Clin. Cancer Res., 27, 1, 1-8 (2008)· doi:10.1186/1756-9966-27-3

[40] Corbeil, J., Productive in vitro infection of human umbilical vein endothelial cells and three colon carcinoma cell lines with HIV-1, Immunol. Cell Biol., 73, 2, 140-145 (1995)· doi:10.1038/icb.1995.22

[41] Alfano, M.; Graziano, F.; Genovese, L.; Poli, G., Macrophage polarization at the crossroad between HIV-1 infection and cancer development, Arterioscler. Thromb. Vasc. Biol., 33, 6, 1145-1152 (2013)· doi:10.1161/ATVBAHA.112.300171

[42] The Arabidopsis Information Resource (TAIR). https://www.arabidopsis.org/servlets/TairObject?type=locus&name=At2g43610. www.arabidopsis.org. Accessed 17 Jan 2022

[43] The Arabidopsis Information Resource (TAIR). https://www.arabidopsis.org/servlets/TairObject?type=locus&id=126703. www.arabidopsis.org. Accessed 17 Jan 2022

[44] The Arabidopsis Information Resource (TAIR). https://www.arabidopsis.org/servlets/TairObject?type=locus&name=At2g07698. www.arabidopsis.org. Accessed 17 Jan 2022

[45] The Arabidopsis Information Resource (TAIR). https://www.arabidopsis.org/servlets/TairObject?type=locus&name=At3g51750. www.arabidopsis.org. Accessed 17 Jan 2022

[46] Sun, Q.; Zhang, H., Targeted inference involving high-dimensional data using nuisance penalized regression, J. Am. Stat. Assoc., 116, 535, 1472-1486 (2021)· doi:10.1080/01621459.2020.1737079

[47] Klosa, J.; Simon, N.; Westermark, PO; Liebscher, V.; Wittenburg, D., Seagull: lasso, group lasso and sparse-group lasso regularization for linear regression models via proximal gradient descent, BMC Bioinform., 21, 1, 1-8 (2020)· doi:10.1186/s12859-020-03725-w

[48] Vincent, M.; Hansen, NR, Sparse group lasso and high dimensional multinomial classification, Computat. Stat. Data Anal., 71, 771-786 (2014) · Zbl 1471.62200 · doi:10.1016/j.csda.2013.06.004