Predicting clinical outcomes of ovarian cancer patients: deep survival models and transfer learning

Elena Spirina Menand, Nisrine Jrad, Jean-Marie Marion, Alain Morel, Pierre Chauvet

Predicting clinical outcomes of ovarian cancer patients: deep survival models and transfer learning

Elena SPIRINA MENAND

LARIS, Université d'Angers, Institut de Cancérologie de l'Ouest, Angers, France. E-mail: elena.menand@etud.univ-angers.fr

Nisrine JRAD

LARIS, Université Catholique de l'Ouest, Angers, France. E-mail: njrad@uco.fr

Jean-Marie MARION

LARIS, Université Catholique de l'Ouest, Angers, France. E-mail: marion@uco.fr

Alain MOREL

CRCINA, Université d'Angers, Institut de Cancérologie de l'Ouest, Angers, France. E-mail: alain.morel@ico.unicancer.fr

Pierre CHAUVET

LARIS, Université Catholique de l'Ouest, Angers, France. E-mail: pierre.chauvet@uco.fr

With the advent of high-throughput sequencing technologies, the genomic platforms generate a vast amount of high dimensional genomic profiles. One of the fundamental challenges of genomic medicine is the accurate prediction of clinical outcomes from these data. Gene expression profiles are established to be associated with overall survival in cancer patients, and this perspective the univariate Cox regression analysis was widely used as primary approach to develop the outcome predictors from high dimensional transcriptomic data for ovarian cancer patient stratification.

Recently, the classical Cox proportional hazards model was adapted to the artificial neural network implementation and was tested with The Cancer Genome Atlas (TCGA) ovarian cancer transcriptomic data but did not result in satisfactory improvement, possibly due to the lack of datasets of sufficient size. Nevertheless, this methodology still outperforms more traditional approaches, like regularized Cox model, moreover, deep survival models could successfully transfer information across diseases to improve prognostic accuracy. We aim to extend the transfer learning framework to "pan-gyn" cancers as these gynecologic and breast cancers share a variety of characteristics being female hormone-driven cancers and could therefore share common mechanisms of progression.

Our first results using transfer learning show that deep survival models could benefit from training with multi-cancer datasets in the high-dimensional transcriptomic profiles.

Keywords: TCGA, transcriptome, survival analysis, Cox model, deep learning, transfer learning.

1. Introduction

The recent development of high-throughput sequencing technology and machine learning methodology resulted in a great progress in the field of oncology research based on genomic profiles. However, while the high-dimensional data generated, such as RNA-seq, keep growing, a real need for appropriate machine learning techniques, capable of dealing with mass data, has appeared.

Ovarian cancer is a complex, heterogeneous genetic disease. Because of the high risk of recurrence in high-grade serous ovarian carcinoma (HGS-OvCa), the development of outcome predictors is important not only for patient stratification but also to recognize categories of patients that are more likely to respond to particular therapies (Verhaak et al. 2012). The lack of successful treatment strategies for ovarian cancer led The Cancer Genome Atlas (TCGA) researchers to gather the HGS-OvCa genomic profiles in order to identify molecular abnormalities that influence pathophysiology, affect outcome and constitute therapeutic targets (Bell et al. 2011).

Gene expression profiles are considered to reflect the cancer progression driven by mutations and epigenetic modifications. These profiles were established to be associated with overall survival and the study (Bell et al. 2011) developed the prognostic signatures for ovarian cancer based on the TCGA microarray gene expression

Proceedings of the 31st European Safety and Reliability Conference
Edited by Bruno Castanier, Marko Cepin, David Bigaud, and Christophe Berenguer
Published by Research Publishing, Singapore.
ISBN: 978-981-18-2016-8; doi:10.3850/978-981-18-2016-8_505-cd

profiles using a univariate Cox regression analysis, and validated them on external datasets.

Recently, artificial neural networks (ANN) caught attention to solve problems with genomic profiles. (Yousefi et al. 2017; Ching, Zhu, and Garmire 2018) used ANN to construct survival models using the TCGA gene expression data. The authors (Ching, Zhu, and Garmire 2018) used the high-throughput transcriptomic data of the different TCGA cancer types and compared survival methods such as regularized Cox model, Random Survival Forests, CoxBoost and the proposed Cox-nnet method. Their approach Cox-nnet gave satisfactory results for some cancer types, especially for TCGA Kidney Renal Cell Carcinoma (KIRC), and insufficient results for other types, for example, in the OV dataset. The study of (Yousefi et al. 2017) applied the ANN to the survival analysis of the TCGA-BRCA transcriptional and integrated features datasets, exploring at the same time the benefits of transfer learning with multi-cancer datasets.

The objective of this work is to experiment the transfer learning strategy in the task of ovarian cancer prognostication with the up-to-date harmonized (aligned to hg38) RNA-sequencing (RNA-seq) data from the TCGA-OV project in order to detect significant prognostic features.

The outline of the paper is as follows: section 2 presents the Cox survival analysis technique based on neural networks and the different aspects of deep learning, section 3 describes materials and methods and section 4 present the results and discusses the future work.

2. Survival analysis and deep learning

2.1. *Cox proportional hazards and neural networks*

Survival analysis, one of the statistics subfields, deals with the time-to-event as outcome. When the outcome is unknown during the observation period, it is called censoring and it is one of the major difficulties in survival analysis. Recently Wang et al (Wang, Li, and Reddy 2017) have created a taxonomy of different approaches in this branch of statistics, distinguishing the traditional statistical and machine learning methods. One of the commonly used statistical method is a semi-parametric Cox regression or Cox proportional hazards. In this model each data instance is described by a triplet (X_i,, t_i,, δ_i), where X_i = (x_{i1}, x_{i2}, …, x_{iP}) is the feature vector for instance i, t_i is the observed time, time of failure if δ_i is 1 or right-censoring if δ_i is 0. We note here the number of observations N and the number of features P. In this framework the rate of event at time t given that no event occurred before time t, i.e. the hazard function is:

$$h(t,X_i) = h_0(t)exp(X_i\beta) \qquad (1)$$

where $h_0(t)$ is the baseline hazard function (an arbitrary nonnegative function of time), and β^T = (β_1, β_2, …, β_P) is the coefficient vector. To note that the features are assumed to have an exponential influence on the outcome but the baseline hazard function, $h_0(t)$, is unspecified, thus resulting in a semi-parametric model. This makes it impossible to fit the model using standard likelihood function, instead the partial likelihood is used:

$$L(\beta) = \prod_{j=1}^{N} \left[\frac{exp(X_j\beta)}{\sum_{i\in R_j} exp(X_i\beta)}\right]^{\delta_j} \quad (2)$$

where R_j is the set of indices, i, with $y_i \geq t_j$ (those at risk at time t_j). The coefficient vector is estimated by maximizing this partial likelihood, or equivalently, minimizing the negative log-partial likelihood for improving efficiency:

$$LL(\beta) = -\sum_{j=1}^{N} \delta_j \left\{X_j\beta - log\left[\sum_{i\in R_j} exp(X_i\beta)\right]\right\} \quad (3)$$

The extension of Cox regression with artificial neural networks was first proposed by (Faraggi and Simon 1995), who replaced the linear predictor of the Cox regression model, by a one hidden layer multilayer perceptron (MLP). This work was further explored by (Yousefi et al. 2017) (SurvivalNet), (Ching, Zhu, and Garmire 2018) (Cox-nnet), (Katzman et al. 2018) (DeepSurv) and who proposed to incorporate the advances of deep learning framework and demonstrated that their methods outperform the classical Cox method. The linear predictors in their models become:

$$\theta_i = G(WX_i + b)^T\beta \quad (4)$$

where W is the coefficient weight matrix between the input and hidden layer of size H x P, H is the number of neurons in the hidden layer, b is the bias vector of size H and G is the nonlinear activation function. The partial log likelihood (3) can be written as:

$$PL(\beta, W) = \sum_{\delta_j=1} \left\{\theta_j - log\left[\sum_{i\in R_j} exp(\theta_i)\right]\right\} \qquad (5)$$

2.2. *Regularization*

When applied to high-dimensional transcriptomic data, the major issue of this model is overfitting which can be overcome with the help of different regularization techniques such as ridge regularization, dropout, early stopping and to a lesser extent batch normalization.

Adding the ridge regularization term to the partial log likelihood (5) gives the following cost function:

$$cost(\beta, W) = PL(\beta, W) + \lambda\left(||W||_2 + ||\beta||_2\right) \qquad (6)$$

where $||\ ||_2$ designates L2 norm penalty function and λ is a regularization coefficient leading to a weight decay.

In addition to ridge regularization, when using ANN it is common to employ dropout regularization(Srivastava et al. 2014). During training, this approach randomly zeroes some of the elements of the input with probability p (dropout rate or fraction). This has proven to be an

effective technique for regularization and preventing the co-adaptation of neurons as described in the paper (Hinton et al. 2012).

Early stopping means stopping the training as soon as performance on a validation set starts to get worse. If regularization methods like weight decay that update the loss function to encourage less complex models are considered "explicit" regularization, then early stopping may be thought of as a type of "implicit" regularization, much like using a smaller network that has less capacity (Zhang et al. 2017).

Batch normalization (also known as batch norm) is a method used to accelerate the training of artificial neural networks. It draws its power from normalizing activations, and from incorporating this normalization in the network architecture itself. It was proposed by (Ioffe and Szegedy 2015) and offers small regularization effect as well.

2.3. *More data and transfer learning*

Another possibility to deal with a substantial generalization error is to get more data and apply transfer learning strategy as in the study of (Yousefi et al. 2017).

Indeed, gynecologic cancers share a variety of characteristics, their development is influenced by female hormones, and they are managed by a particular medical specialty, gynecologic oncology as underlined by (Berger et al. 2018). In this study, the authors refer to the following multi-cancer group as "pan-gyn" and focus on five TCGA tumor types: high-grade serous ovarian cystadenocarcinoma (OV), uterine corpus endometrial carcinoma (UCEC), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), uterine carcinosarcoma (UCS), and invasive breast carcinoma (BRCA). They found molecular features that differed in the "pan-gyn" group and the TCGA non-gynecologic tumor types.

This lets us hypothesize that augmenting OV training data with other datasets from the "pan-gyn" group could improve OV prognostication. The transfer learning rule of thumb being that while adding more training data, the validation and training sets should still come from the same target distribution, OV cancer in our case.

2.4 *Automated hyperparameter optimization*

Deep neural networks' prediction accuracy is highly dependable on many hyperparameters (number of layers, number and type of activation functions in each layer, and choices for optimization/regularization techniques). These details of algorithm tuning are crucial to judging whether a given technique is genuinely better, or simply better tuned.

The naïve approach of the exhaustive grid search of the hyperparameters space is time consuming, so other , more intelligent strategies have appeared recently for automated hyperparameter optimization using Bayesian optimization supported by Sequential Model-Based Global Optimization (SMBO) methodology (Bergstra, Yamins, and Cox 2013; Martinez-Cantin 2014).

SMBO algorithms have been used in many applications where evaluation of the fitness function is expensive. In an application where the true fitness function, as *PL* in our case, is costly to evaluate, model-based algorithms approximate it with a surrogate that is cheaper to evaluate. A point that maximizes the surrogate becomes the proposal for where the true function *PL* should be evaluated, thus resulting in a fewer fitness function evaluations and a faster hyperparameter optimization (Bergstra, Yamins, and Cox 2013).

3. Materials and methods

3.1. *Gene expression and clinical data*

TCGA RNA-sequencing data and clinical data were downloaded from Genomics Data Commons (GDC) portal (https://portal.gdc.cancer.gov/) using the pipeline of the R/Bioconductor package TCGAbiolinks (Colaprico et al. 2016). The harmonized RNA-seq data (HTSeq counts) were normalized using the existing TCGAbiolinks normalization function which is recommended for differential expression analysis.

Supplemental survival data were downloaded from the standardized dataset named the TCGA Pan-Cancer Clinical Data Resource (TCGA-CDR) (Liu et al. 2018). We merged the OV survival data from TCGA-CDR with the GDC clinical data. We made a choice to perform our tests on OS endpoint. The corresponding TCGA-CDR columns included OS for status and OS.time for time-to-event data. OS column contained the value 0 encoding for alive (censored) status and 1 for deceased (failure) and OS.time contained numbers of days from the date of diagnosis to either the date of last follow up if OS was 0 or time to death if OS was 1.

We downloaded the RNA-seq data for the following TCGA projects: TCGA-OV, TCGA-BRCA, TCGA-UCEC, TCGA-CESC, TCGA-UCS. After merging RNA-seq and clinical data and discarding cases without survival information, we obtained 372 samples for OV, 1076 for BRCA, 541 for UCEC, 291 for CESC, 55 for UCS. All the datasets contained 17,000 + gene expression features in common.

3.2. *Performance metric*

The widely used in survival analysis Concordance-index (C-index) measures the concordance between predicted risk score and observed survival time. This measure is computed for all comparable pairs in the test set and the number of times the predictions are concordant is summarized. The C-index value of 0.5 is equivalent to random guess and 1 is the perfect concordance, so higher C-index means better model performance.

3.3. *Data pre-processing*

For our tests, we (log2+1) transformed the normalized values and split our dataset into 5 folds using R package MTLR (Yu et al. 2011) thus constructing 5 different splits into training and test sets with respectively

80% and 20% of samples for a further 5-fold cross-validation. As the accuracy obtained on one test set could be very different from the accuracy obtained for a different test set, the widely used K-fold cross-validation technique ensures that each fold is used as a test set at some point and provides the solution to the reliability problem. The split was done using the stratification by the OS.time and OS features in order to have similar distributions of survival times and censoring in training and test sets. To compare the survival of training and test set splits, we plotted Kaplan-Meier curves and calculated the log-rank test p-value and concluded that the difference of survival between our generated training and test sets was not significant.In order to facilitate the training procedure, the training data were standardized to zero-mean and unit-variance to comply with best practices for training deep learning algorithms. The training data included the samples from OV-only and different combinations of OV and the datasets among BRCA, UCEC, CESC and UCS. For our tests we used the DeepSurv implementation of the Python package pycox (Kvamme, Borgan, and Scheel 2019)..

3.4. *Bayesian optimization*

For each of the 5 training sets, we performed 4-fold cross-validation for hyperparameter automated search, only the OV dataset samples were used in the validation sets and 16 different combinations of cancer types as optimization sets. We used python library hyperopt (Bergstra, Yamins, and Cox 2013) for Bayesian optimization with adaptive Tree of Parzen Estimators algorithm and the following search space: number of layers (1–8), layer width (8–2048), dropout rate (0–0.6), weight decay (0-0.9), learning rate for Adam optimizer (Kingma and Ba 2017) (0.00001-0.1) and activation function among ReLU (Nair and Hinton 2010), SELU (Klambauer et al. 2017), hyperbolic tangent (tanh), sigmoid function and a maximum of 200 trials.

The best network design was then used to re-train a deep survival model using the optimization and validation samples, and the C-index of this best model is reported using the held-out OV testing samples. We repeated this procedure 10 times for each test dataset. To compare the C-index values in different experiments, we performed Wilcoxon rank-sum tests and report the significant (<0.05) p-values.

4. Results

Transfer learning experiments showed that ANN survival models could benefit from training with multi-cancer datasets in the high-dimensional transcriptional data. The results of our tests are presented in the Fig. 1. Training with four combined datasets OV+BRCA+UCS, OV+CESC+UCS, OV+BRCA+UCEC+UCS, and OV+BRCA+UCEC+CESC+UCS resulted in the small but significant improvements to the ANN survival model (Wilcoxon rank-sum p-values respectively of 0.018, 0.02, 0.0033 and 0.0045). Among these results, the best C-index gain of 2.1% was with OV+BRCA+UCEC+UCS combined dataset.

The authors (Yousefi et al. 2017) noticed that prediction accuracy generally decreases as the proportion of right-censored samples in a dataset increases. We measured the censoring proportion in our datasets: OV (38.44%), BRCA (85.97%), UCEC (83.18%), CESC (75.26%) and UCS (38.18%). Interestingly, the UCS dataset with the smallest right-censoring proportion being present in all the four combined datasets with improved C-index, the best or the most significant gains are still obtained with the datasets with bigger censoring proportions than the target OV dataset itself. We hypothesize that although genetic alterations and expression patterns are often strongly associated with primary disease site, the "pan-gyn" group is likely to share common mechanisms of progression and the improved performance of the deep survival models with augmented datasets could provide additional evidence of these mechanisms.

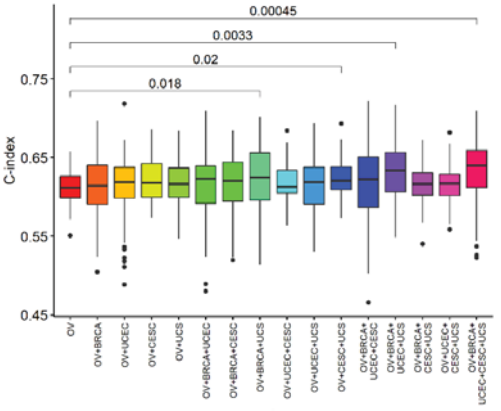

Fig. 1. From left to right are the boxplots of the obtained 5-fold cross validation C-index on the OV test datasets. Higher C-index means better model performance. The horizontal bars in the boxes represent the median values, the boundaries of the boxes delimit lower and upper quartiles, the values outside the boxes are the lowest and the highest observations. The brackets show the significant Wilcoxon rank-sum test p-values.

As a future work, there is as a strong need to interpret the biological meaning of the transcriptional features contributing to the survival patient stratification. However, it is important to understand, as underlined by (Berger et al. 2018), that the "pan-gyn" project possibilities should be considered as hypothesis-generators and are to be tested and validated in the follow-up studies.

5. Conclusion

In this paper, we have presented the Cox proportional hazards methodology and its implementation with the

artificial neural networks. We have discussed the different deep learning techniques such as regularization, automated optimization, meant to overcome the obstacles when dealing with the high-dimensional gene expression data and survival analysis. Since more data is another option to prevent the neural networks from overfitting, we have explored the transfer learning framework applied to the deep survival analysis with the TCGA ovarian RNA-seq data. According to our experiments, the deep survival models could benefit from training with the augmented multi-cancer datatsets, and more data could further improve the survival network performance.

Acknowledgement

Our work is supported by Cancéropôle du Grand Ouest (CGO), France and is part of the program "Emergence 2018".

References

Bell, D., A. Berchuck, M. Birrer, J. Chien, D. W. Cramer, F. Dao, R. Dhir, et al. 2011. "Integrated Genomic Analyses of Ovarian Carcinoma." *Nature* 474 (7353): 609–15. https://doi.org/10.1038/nature10166.

Berger, Ashton C., Anil Korkut, Rupa S. Kanchi, Apurva M. Hegde, Walter Lenoir, Wenbin Liu, Yuexin Liu, et al. 2018. "A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers." *Cancer Cell* 33 (4): 690-705.e9. https://doi.org/10.1016/j.ccell.2018.03.014.

Bergstra, J, D Yamins, and D D Cox. 2013. "Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures," 9.

Ching, Travers, Xun Zhu, and Lana X. Garmire. 2018. "Cox-Nnet: An Artificial Neural Network Method for Prognosis Prediction of High-Throughput Omics Data." Edited by Florian Markowetz. *PLOS Computational Biology* 14 (4): e1006076. https://doi.org/10.1371/journal.pcbi.1006076.

Colaprico, Antonio, Tiago C. Silva, Catharina Olsen, Luciano Garofano, Claudia Cava, Davide Garolini, Thais S. Sabedot, et al. 2016. "*TCGAbiolinks*: An R/Bioconductor Package for Integrative Analysis of TCGA Data." *Nucleic Acids Research* 44 (8): e71–e71. https://doi.org/10.1093/nar/gkv1507.

Faraggi, David, and Richard Simon. 1995. "A Neural Network Model for Survival Data." *Statistics in Medicine* 14 (1): 73–82. https://doi.org/10.1002/sim.4780140108.

Hinton, Geoffrey E., Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. 2012. "Improving Neural Networks by Preventing Co-Adaptation of Feature Detectors." *ArXiv:1207.0580 [Cs]*, July. http://arxiv.org/abs/1207.0580.

Ioffe, Sergey, and Christian Szegedy. 2015. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift." *ArXiv:1502.03167 [Cs]*, March. http://arxiv.org/abs/1502.03167.

Katzman, Jared L., Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. 2018. "DeepSurv: Personalized Treatment Recommender System Using a Cox Proportional Hazards Deep Neural Network." *BMC Medical Research Methodology* 18 (1). https://doi.org/10.1186/s12874-018-0482-1.

Kingma, Diederik P., and Jimmy Ba. 2017. "Adam: A Method for Stochastic Optimization." *ArXiv:1412.6980 [Cs]*, January. http://arxiv.org/abs/1412.6980.

Klambauer, Günter, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. 2017. "Self-Normalizing Neural Networks." *ArXiv:1706.02515 [Cs, Stat]*, September. http://arxiv.org/abs/1706.02515.

Kvamme, Håvard, Ørnulf Borgan, and Ida Scheel. 2019. "Time-to-Event Prediction with Neural Networks and Cox Regression." *ArXiv:1907.00825 [Cs, Stat]*, September. http://arxiv.org/abs/1907.00825.

Liu, Jianfang, Tara Lichtenberg, Katherine A. Hoadley, Laila M. Poisson, Alexander J. Lazar, Andrew D. Cherniack, Albert J. Kovatich, et al. 2018. "An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics." *Cell* 173 (2): 400-416.e11. https://doi.org/10.1016/j.cell.2018.02.052.

Martinez-Cantin, Ruben. 2014. "BayesOpt: A Bayesian Optimization Library for Nonlinear Optimization, Experimental Design and Bandits," November, 5.

Nair, Vinod, and Geoffrey E Hinton. 2010. "Rectified Linear Units Improve Restricted Boltzmann Machines," 8.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," June, 30.

Verhaak, Roel G.W., Pablo Tamayo, Ji-Yeon Yang, Diana Hubbard, Hailei Zhang, Chad J. Creighton, Sian Fereday, et al. 2012. "Prognostically Relevant Gene Signatures of High-Grade Serous Ovarian Carcinoma." *Journal of Clinical Investigation*, December. https://doi.org/10.1172/JCI65833.

Wang, Ping, Yan Li, and Chandan K. Reddy. 2017. "Machine Learning for Survival Analysis: A Survey." *ArXiv:1708.04649 [Cs, Stat]*, August. http://arxiv.org/abs/1708.04649.

Yousefi, Safoora, Fatemeh Amrollahi, Mohamed Amgad, Chengliang Dong, Joshua E. Lewis, Congzheng Song, David A. Gutman, et al. 2017. "Predicting Clinical Outcomes from Large Scale Cancer Genomic Profiles with Deep Survival Models." *Scientific Reports* 7 (1). https://doi.org/10.1038/s41598-017-11817-6.

Yu, Chun-Nam, Russell Greiner, Hsiu-Chin Lin, and Vickie Baracos. 2011. "Learning Patient-Specific Cancer Survival Distributions as a Sequence of Dependent Regressors," 10.

Zhang, Chiyuan, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. "Understanding Deep Learning Requires Rethinking Generalization." *ArXiv:1611.03530 [Cs]*, February. http://arxiv.org/abs/1611.03530.