Predicting with Confidence: A Case-Based Reasoning Framework for PredictingSurvival in Breast Cancer

Christopher L. Bartlett, Isabelle Bichindaritz, Guanghui Liu

Intelligent Bio Systems Laboratory, Biomedical and Health Informatics
State University of New York at Oswego, 7060 NY-104, Oswego, NY 13126
ibichind@oswego.edu

Abstract

There is usually a trade-off between predictive performance and transparency, where the reasoning process behind an algorithm is shielded behind a ”black-box.” In medical domains, experts being responsible for their decisions need to understand the reasons behind machine-generated recommendations. This paper presents a transparent case-based survival analysis framework that automatically retrieves an optimal number of solved survival cases and adapts them to predict the survival of a new case. With this methodology, retrieved and adapted survival cases lend an insight into which cases a prediction is based on. Our framework is capable of integrating DNA methylation, gene expression, and their combination in breast cancer. Additionally, we test our approach with and without feature selection and demonstrate the usefulness of the adaptation phase. We demonstrate that our framework performs at least as effectively as other state-of-the-art methods while affording greater explainability.

Introduction

The predictive performance of machine learning algorithms has progressed tremendously, though many recent works has focused on improving explainability (Lundberg et al. 2018). Machine-learning researchers typically allude to a ”black-box” effect when input and output are understood, but the processing that occurs in-between is obscure. Methods leading to a greater level of explainability are necessary in medicine since users are ultimately responsible for their clinical decisions and thus need to make informed decisions. One of the most explainable algorithms are instance-based learners (IBLs), such as *k*-Nearest Neighbor (*k*NN), for which decisions are made by similarity between a new case and solved retrieved cases, which can serve as explanations for a system recommendations (Lamy et al. 2018).

Taking advantage of the known explainability of IBL systems, this paper presents Case Based Reasoning with Confidence, or CBR-CONF. CBR-CONF is a case-based reasoning system for survival prediction that uses a novel confidence metric to determine an optimal number of similar cases to retrieve for each test case to solve. Specifically, we contribute the following:

1. **Optimal number of retrieved cases:** CBR-CONF locates an optimal number of similar cases for each test case based on an automatically defined level of confidence in each retrieved case. Retrieval stops when we can confidently assign a solution. This method extends upon our prior work in (Bartlett, Liu, and Bichindaritz 2020a) and (Bartlett, Liu, and Bichindaritz 2020b).
2. **Multi-level case elaboration and refinement:** Significantly different DNA methylation levels found at a high-order cluster of probes that serve similar functionswere utilized and compared. We then perform a similar series of tests using mRNA expression levels prior to integrating the two microarray technologies to find enriched motifs and transcription factors.
3. **Novel case adaptation technique:** Our case adaptation technique is tailored to survival analysis, which is an original application domain for CBR.
4. **Explainability:** In addition to the known explainability of IBL, CBR-CONF applies feature selection to determine an optimal biomarker signature for this disease, based on gene expression, methylation, enriched motifs, pathway analysis, and transcription factors.

Research Background

Survival analysis is about predicting how likely an event is to happen over time. In our case, the event we are interested in is the death of the cancer patient. In survival analysis, we do not necessarily know how long has each patient lived as the experiment may have stopped before their death. The individuals in a population who have not been subject to the death event are labeled as right-censored. We observe either the survival time, if we have the

death date, or a censured time if we only have the date of last visit to the doctor. A survival instance is usually represented as (x_i, t_i, δ_i) where x_i is the feature vector, t_i is the observed time, δ_i is the indicator: 1 for an uncensored instance, which means the patient is dead, and 0 for a censored instance, which is a patient being alive.

Existing methods for survival analysis include the random survival forest (Ishwaran et al. 2008), deep learning (Katzman et al. 2018) (Lee et al. 2018) (Hao et al. 2019) (Martini et al. 2019), and statistical methods.

In case-based reasoning, Karmen et al. calculate similarity based on survival functions (Karmen et al. 2019). Our work adopts different CBR strategies for survival analysis and in addition integrates multi-omics data, while Karmen at al.'s data are at the clinical level (phenotypic).

Methods

DNA methylation and mRNA gene expression data for breast cancer (BRCA) was downloaded from The Cancer Genome Atlas (TCGA) Research Network: https://www.cancer.gov/tcga.

The methylation data pertained to 763 primary tumor samples and the 485,577 probes that exist on the Illumina Human Methylation 450 bead chip. Each probe represents a location on the DNA where a methyl group (-CH3) may be found. DNA methylation has been associated with many diseases and disorders, including trauma and aging. Methylation β values, which are an estimation of the methylation levels between 0 and 1, were extracted. 0 indicates that the site is completely non-methylated, while 1 indicates that it is completely methylated. Similarly, 0.8 and above is commonly referred to as being hypermethylated while 0.2 and below is hypomethylated. We discarded samples that had a survival duration less than 0 months. After batch elimination, the remaining probes were used to locate differentially methylated regions, which consist in clusters of probes that are a possible functional region for gene transcriptional regulation. These clusters are based on physical proximity on a chromosome and serve as a feature reduction mechanism based on biology. The number of features was reduced to 8,722.

Gene expressions came from the Illumina HiSeq RNA sequencing data collected from TCGA. Low correlating samples were filtered out and normalized prior to performing differential expression analysis (DEA). A false discovery rate of 0.01and a log fold change cut-off of 1 were used in this analysis.After preprocessing, DEA and reducing samples to just the primary cancer tissue, 1159 features and 763 samples were retained.

In addition to the genomic and transcriptomic data, each case features the overall survival (in months) T and the censoring information (1 if the case is deceased, and 0 if the case is living at the time of the last follow-up).

23 enriched motifs were found through the R package ELMER (Silva et al. 2018), which groups together DNA methylation probes correlated with expression of these genes. Gene-probe pairs where this occurs are set aside and used to locate enriched motifs and upstream regulatory transcription factors.

Prognostic Groups

Prognostic groups for survival were established using a method similar to (Chen et al. 2018). A multivariate cox regression was constructed using either DNA methylation features,or gene expression features as covariates in order to test each feature's contribution to the survival state. To find riskgroups, the beta coefficient for each feature was multiplied by its expression value (beta value for DNA methylation) and these values were then summed together to find a singleprognostic score for each sample (Formula (1)). Samples were ranked by their prognostic scores and divided into equal-sized low, medium and high risk groups (Fig. 1).

$$PrognosticScore = \sum_{i=1}^{NumberOfFeatures}(\beta) * (expression) \quad (1)$$

Confidence Metric

In order to predict the survival length of test samples, a confidence metric was established to indicate when to cease retrieval. For training samples in each risk group, the mean of each feature was calculated and used to construct one prototypical case *P* for each risk group.

$$Conf = 1/2\ (dist(\alpha, P)) + (d(a, uns1) + \cdots + d(a, unsn))/n \quad (2)$$

As shown in Formula (2), the confidence metric is one half of the Euclidean distance *dist* from case *a* to its prototype *P*, added to the average distance of case *a* to each unsolved case. The system's confidence in its solution for any given unsolved case would then be the summation of each retrieved case's confidence value until confidence reaches 100%. The predicted survival time for the unsolved case would then be the meansurvival time of each retrieved case.

Feature Selection

To locate a subset of features that were highly specific to the overall survival of the breast cancer samples, a feature selection process was constructed based on multivariate Cox regression. 100 randomized training and testing splits of the data (with replacement) were used to extract a biomarker panel. After each iteration, features that had a significant log rank less than 0.01 were notated and validated with the test set. Upon conclusion of the 100 runs, only the features that were significant in at least 50% of the training sets were kept. Prognostic scores were then the sum of the average beta coefficients for each of these features multiplied by the feature's methylation beta value (for DNAmethylation) or the expression value (for mRNA expression).

Survival Prediction

Once cases in the case base were retrieved for an unsolved case and their average survival duration was determined and

assigned to the unsolved case, a cox proportional hazards model was built. Assigned survival times and survival status was used to construct the model, with the concordance index being employed to test the accuracy of the model. Concordance index measure evaluates how correct is the ordering of predicted times. It is interpreted as follows: 0.5 is the expected result from random predictions, 1.0 is perfect concordance and, 0.0 is perfect anti-concordance (multiply predictions with -1 to get 1.0). For example, if Case A died after 36 months and Case B diedafter 33 months and the model predicted that Case A woulddie after 24 months and Case B would die after 11 months, this is still a correct prediction.

Case Adaptation

A case adaptation phase followed feature selection and survival prediction to determine if adapting the case increased or decreased the concordance index. Case adaptation is the ability to morph training cases to reflect a test case more closely. Therefore, after retrieving similar training cases, the average difference between the survival of training cases and the test case was computed. This difference was then added or subtracted to the training case's survival duration. Once this was performed,we retested with the adapted cases.

Results

Results were obtained using only DNA methylation, using only gene expressions (RNA), and using motifs combining the two.

Table 1: Predicting survival of breast cancer samples from 8,722 DNA methylation probes and from 80 selected DNA methylation probes

Method	Concordance Index	
	8,722 probes	80 probes
CBR-CONF (Adapted)		**0.745**
CBR-CONF	**0.727**	**0.739**
*k*NN (*k*5)	0.671	0.675
Gradient Boosting Tree	0.661	0.715
*k*NN (*k*15)	0.638	0.681
GLMnet	0.637	0.500
*k*NN (*k*10)	0.624	0.658
Random Survival Forest	0.623	0.594
*k*NN (*k*20)	0.597	0.679

Results on DNA Methylation

The first stage was to view DNA methylation alone using the probes located in the differentially methylated regions. This constructed a dataset of 763 cancer samples and 8,724 features (8,722 of which were DNA methylation probes, plus the survival length and the censoring information 0/1). Tests were performed:

- **Before feature selection**: The initial tests used all 8,722 features and were performed using 10-fold cross validation. A *K*-Nearest Neighbor algorithm tailored to survival analysis with 4 different levels of *k*, Random Survival Forest, and Lasso and Elastic-Net Regularized Generalized Linear Models (GLMnet) were also tested and compared. As shownin Table 1, CBR-CONF obtained the highest concordance index.
- **After feature selection**: We then performed a similar test after feature selection. 80 features remained after the 100 iterations. CBR-CONF retrieved an average of 7 cases with a maximum of 20 and a minimum of 2. Table 1 shows that once again CBR-CONF held strong results, especially when case adaptation is applied.

Table 2: Predicting survival of breast cancer samples from 1,149 differentially expressed genes and from 22 selected genes using mRNA expression.

Method	Concordance Index	
	1,149 genes	22 genes
CBR-CONF (Adapted)		**0.699**
CBR-CONF	**0.672**	**0.696**
*k*NN (*k*20)	0.657	0.701
*k*NN (*k*15)	0.644	0.709
*k*NN (*k*5)	0.641	0.654
*k*NN (*k*10)	0.637	0.690
Gradient Boosting Tree	0.634	0.722
GLMnet	0.500	0.503
Random Survival Forest	0.466	0.589

Results on Gene Expression

After normalizing and determining, which genes were differentially expressed from a normal control group, 761 cancer samples and 1,149 genes were kept.

- **Before Feature Selection:** As with methylation, Table 2 shows that CBR-CONF was the highest performer.
- **After Feature Selection:** Only 22 genes remained after feature selection. These features were tested both before and after applying case adaptation. While CBR-CONF still performed well, Table 2 shows that Gradient Boosting Tree was the most performant. During retrieval, the average number of retrieved cases was 2 with a minimum of 2 and a maxi- mum of 3.

Table 3: Predicting survival of breast cancer samples from the weighted average of DNA methylation probes on 23 enriched motifs.

Method	Concordance Index
CBR-CONF (Adapted)	**0.668**
*k*NN (*k*5)	0.667
*k*NN (*k*10)	0.655
CBR-CONF	**0.652**
*k*NN (*k*15)	0.644
*k*NN (*k*20)	0.641
Gradient Boosting Tree	0.550
GLMnet	0.500
Random Survival Forest	0.445

Enriched Motifs

Using the 23 enriched motifs found, we first selected all methylation probes associated with these enriched motifs (and subsequently, their transcription factors). We tested each of these probes using a multivariate cox regression and extracted their log ranking. Probes were then weighted by their inverse log ranking and a weighted average of all probes for an enriched motif was calculated for each sample. As we had a small set of 23 motifs, we opted out of performing feature selection. The results are available in Table 3 and show again that CBR-CONF adapted performed the best.

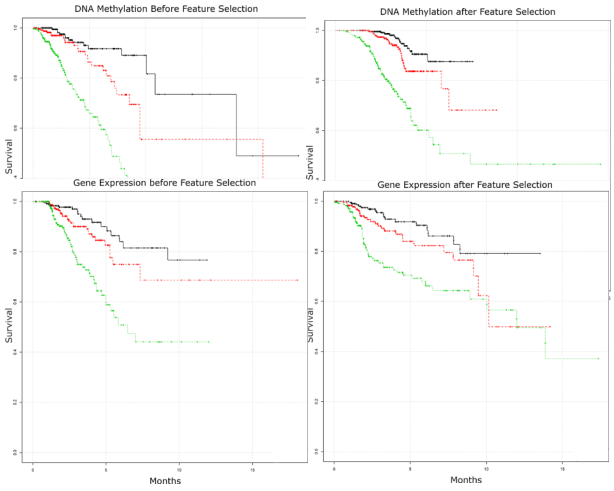

Figure 1. Kaplan-Meier plots of the low, medium and high risk groups using CBR-CONFmodel. The top plots are for DNA methylation before and after feature selection, and the bottom plots are for gene expression beforeand after feature selection. **Legend:** The black line is low risk, red line is medium risk, green line is high risk.

Explainability

Using Gene Ontology knowledge-base, the 80 DNA methylation probes selected were annotated to their nearest genes.These genes were found to be significantly associated with the positive regulation of response to DNA damage (q-value = 0.021, coverage = 5/47). For the 22 gene expressions selected, we found a significant association to the regulation of complement activation (q-value = 0.002, coverage = 4/27). For transcription factors found within theenriched motifs, the pattern specification process was significant (q-value = 1.4e-06, coverage = 8/184).

Conclusion

In this paper we discussed a case based reasoning framework that assigns a novel confidence metric to each solved case depicting how well that case can be used to solve a new case. We also developed novel retrieval and adaptation steps for survival analysis. Using DNA methylation, gene expression and enriched motifs, we tested our framework to predict survival of breast cancer. We found that our model performed at least as well as several renowned methodologies for survival prediction with the advantage of interpretability. We further tested the presence of overfitting by introducing randomness in the dataset and found only minor differences in results, which suggests that the model does not suffer from overfitting. In the future, we wish to further validate our results using independent datasets and expand the scope to other cancers and diseases.

References

Bartlett, C. L.; Liu, G.; and Bichindaritz, I. 2020a. Case- based reasoning for the analysis of methylation data in oncology. In Barták, R., and Bell, E., eds., *Proceedings of the Thirty-Third International Florida Artificial Intelligence Research Society Conference, 2020*, 401–406. AAAI Press.

Bartlett, C. L.; Liu, G.; and Bichindaritz, I. 2020b. Classifying breast cancer tissue through DNA methylation and clinical covariate based retrieval. In Watson, I., and Weber, R. O., eds., *28th International Conference, ICCBR 2020, Lecture Notes in Computer Science*, 82–96. Springer.

Chen, E.G.; Wang, P.; Lou, H.; Wang, Y.; Yan, H.; Bi, L.; Liu, L.; Li, B.; Snijders, A.M.; Mao, J.H.; and Hang, B.2018. A robust gene expression-based prognostic risk score predicts overall survival of lung adenocarcinoma patients. *Oncotarget* 9(6):6862–6871.

Hao, J.; Kim, Y.; Mallavarapu, T.; Oh, J. H.; and Kang, M.2019. Interpretable deep neural network for cancer survival analysis by integrating genomic and clinical data. *BMC Medical Genomics* 12(Suppl 10):1–13.

Ishwaran, H.; Kogalur, U. B.; Blackstone, E. H.; and Lauer, M. S. 2008. Random survival forests. *Annals of Applied Statistics* 2(3):841–860.

Karmen, C.; Gietzelt, M.; Knaup-Gregori, P.; and Ganzinger, M. 2019. Methods for a similarity measure for clinical attributes based on survival data analysis. *BMC Medical Informatics and Decision Making* 19(1):1–14.

Katzman, J. L.; Shaham, U.; Cloninger, A.; Bates, J.; Jiang,T.; and Kluger, Y. 2018. DeepSurv: Personalized treat- ment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*18(1):1–15.

Lamy, J. B.; Sekar, B.; Guezennec, G.; Bouaud, J.; and Séroussi, B. 2019. Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. *Artificial intelligence in medicine*, *94*, 42-53.

Lee, C.; Zame, W. R.; Yoon, J.; and Van Der Schaar, M. 2018. DeepHit: A deep learning approach to survival analysis with competing risks., *AAAI 2018* 2314–2321.

Lundberg, S. M.; Nair, B.; Vavilala, M. S.; Horibe, M.; Eisses, M. J.; Adams, T.; ... and Lee, S. I. 2018. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering*, *2*(10), 749-760.

Martini, P.; Chiogna, M.; Calura, E.; and Romualdi, C. 2019. MOSClip: multiomic and survival pathway analysis for theidentification of survival associated gene and modules. *Nucleic acids research* 47(14):e80.

Silva, T. C.; Coetzee, S. G.; Gull, N.; Yao, L.; Hazelett, D. J.; Noushmehr, H.; Lin, D.-C.; and Berman, B. P. 2018. Elmerv.2: An r bioconductor package to reconstruct gene regulatory networks from dna methylation and transcriptome pro-files. *Bioinformatics*.