Time to Event Machine Learning - Evaluation

Daniel Saggau

11/9/2020

Graf et al 1999

Introduction

- Point prediction of event free times inevitably gives poor results
- Second approach is predicting the survival of event status at fixed point in time, diminishing the expected misspecification/error.
- expected brier score may be interpreted as a mean squared error of prediction when the estimated prob, which take values in interval [0,1] are viewed as prediction of event status at t,$I(T>t)$ in {0,1}.
- **Advantage**: More sophisticated to use estimated probabilities for prediction: Diagnostic test based on predictive values ; probabilities of positive or negative disease status rather than classification of diseased or not diseased.
- Therefore: brier score, measures average discrepancies between true disease status and predicted value, better than misclassification rate

Conclusion

- Make statement that time to event itself cannot adequately be predicted.
- Best at t=0 is try to estimate the probability that the event of interest will not occur until t*.
- Therefore, measure **needs to compare estimates of event free probabilities**
- **Criticism:** ROC methodology cannot capture feature of prognostic classification: Applies to c-index and index of concordance
- c-index has interpretation of the area under the ROC curve for an artificially constructed diagnostic test in the survival context
- Brier and log score are well investigated and according to paper need to be implemented in SA.
- Summary measures of inaccuracy may be obtained by using loss functions integrated with respect to suitable weight function
- Introduce re-weighting scheme, leading to quantities that don't depend on censoring distribution asymptotically

P.Wang 2017 Machine Learning for Survival Analysis

- Named after Glenn W. Brier and initially used for inaccuracy of probabilistic forecasts
- Can only be used with probabilistic outcomes [0;1, sum for each ind. is 1]
- Extended in 1999 for either binary or categorical outcomes for survival analysis

- When using censored data, we re-weight the censored information
- rare events problematic (need to research)
- uni-dimensional predictions the same as the mean squared error

Kattan, M. W., & Gerds, T. A. (2018). The index of prediction accuracy: an intuitive measure useful for evaluating risk prediction models.

Background- C-index

- Due to interpretability desired by clinicians, concordance statistics such as Harrell's c-index, and ROC curve found popularity
- Interpretable for pairs of subjects, but seldom interest to counsel pair of patients
- Only accounts for discrimination and not calibration
- For modeler good to have measure that isolates discrimination, however isolation requires that calibration also be assed for a comprehensive performance analysis
- Don't separate between useless and harmful model
- They suggest that one would assume a harmful model (incorrectly predicting certainty) having worse score relative to a useless model (a model always predicting prevalence) that predicts with some level of predictive ability
- Further Harell's c index does not provide a value specific to the time of the horizon of prediction
- This paper suggests that performance prediction ought to be specific to the time horizon of the prediction

Background-Brier Score

- Overcomes limitations by distinguishing useless from harmless models
- Brier score reflects calibration and discrimination
- Estimated specifically for time horizon
- **BUT**: less interpretable because requires that the performance of model is compared to the performance of the best of the useless models
- Data dependence of reference value complicates Brier Score interpretation
- useless benchmark depends on overall event risk

Conclusion

- Suggests that one limitation of (index of prediction accuracy; aka calibrated brier score) IPA is that measure using average performance may not reflect improvements that affect a small subset of the population.
- Competing risks are mentioned; claim that two patients may have different clinical decisions even when exact same predicted risk of particular event (competing risks)

Gerds, T. A., & Schumacher, M. (2006). Consistent estimation of the expected Brier score in general survival models with right-censored event times

Introduction

- Traditionally, statisticians quantify how close prediction is to the actual outcome, using the R squared and the Brier score.
- Calibration can be measured by e.g. Hosmer Lemeshow "goodness of fit" test.
- Discrimination can be quantified (do patients who have outcome have higher risk predictions than those who dont) using measures such as sensitivtiy , specificity, and area under the ROC (or concordance statistic)
- New methods to assess prediction have gained attention such as reclassifcation tables, net reclassification improvement (NRI), integrated discrimination improvement (IDI), decision curves based on decision analytic measures(due to novelty not as applied and as common.

Conclusion

Nancy R Cook 2007, Use and Misuse of the Receiver Operating Characteristic Curve in Risk Prediction

Introduction

- Accuracy of models can be assessed via calibration and discrimination.
- **Calibration**: Measure how well predicted probabilities agree with actual observed risk. (prediction develop disease vs actual of disease)
- **Discrimination**: Measure of how well model can separate those who do and do not have disease of interest.(when predicts values for cases are all higher than for non-cases, we have perfect discrimination even when predicted risk does not match the proportion with disease)
- Discrimination is more interesting when we have a classification problem with patients that have and dont have prevalent disease (e.g. diagnostic testing).
- Discrimination is measured by c-statistics/ROC
- But when dealing with predictive/prognostic modeling, we need measures that calibrate and discriminate such as the brier score/R^2 and likelihood statistics.
- **Problem**: Risk prediction models in cardiovascular literature, use c-statistic, despite working with large prospective cohort studies.

ROC Curve and c Statistic

- c- index is the generalization of the ROC for survival data.
- sensitivity of a test is the probability of a positive test result, or the value above a threshold among those with diseases.
- Specificity is the probability of a negative test result, or a value below a threshold among those without disease
- Sensitivity and specificity can be influenced e.g.
- sensitivity among milder cases of disease
- specificity can depend on characteristics of non-cases, e.g. gender, age or prevalence of concomitant risk factors

- ROC is a plot of sensitivity versus 1- specificity
- The area under the curve or the c statistic ranges from 0.5 (no discrimination) to max of 1 (perfect discrimination)
- Essentially, the c statistic is equivalent to the probability that the measure or predicted risk is higher for a case than for a noncase.
- Further, c-statistic describes how well models rank case and noncase; but not a function of actual predicted probabilities

c statistics and Model Selection

- Because c stat is based on ranks it is less sensitive than e.g. measures based on likelihood

Antolini et al. 2005 A time-dependent discrimination index for survival data

Introduction

- Extension ROC for censored data, with its biggest advantage being the interpretability
- Suggest time dependent discrimination index
- Ability to discriminate is summarized over time
- Prediction is focused on the individual outcome hence argue focus on accuracy of the predictions, rather than merely on the covarate effects and their statistical significance
- Argue that **generalizability** is important and contains two levels:
- reproducability: Patients from same population

- transportability: Patients coming from different plausibly related population
- Assessing predictive accuracy naively model lead to overoptimistic result
- The evaluation of the degree of optimism and subsequent bias corrected predictive accuracy offers insights into true model reproducibility
- Focus on discrimination (see 7;8 & 9)
- c-index is discrimination measure introduced by Harrell et al. which is extension of AUC
- AUC is considered the concordance of the ranking between the predicted probabilities of being diseased for pairs of diseased and non-diseased subjects.
- c-index is concordance between ranking of predicted failure times and that of the observed times for pairs of subjects
- Generally speaking, if there is a one to one correspondence between predicted times and predicted survival functions, the ranking between predicted times can be obtained by opposite ranking of predicted survival prob at any fixed time point
- Using non-proportional hazard model for breast cancer patients

Conclusion

- The core of model development is developing suitable outcome prediction, given the availability of appropriate measures of model predictive accuracy, to evaluate the generealizability to new patients of different prognostic models or to judge the relevance of addtional contributions of new covariates.
- Discrimination measure that is established is the ROC and the c-index is an extension for right censored survival data (subject leaves study before event occurs or ends before event happens),

considering the presence of censoring as a population feature rather than a limitation of the sample
- Grounded in following assumption: A subject who developed event should have less predicted probability of surviving beyond his survival time than any subject who survived longer
- Argue that need separate evaluation for prognostic models

Uno et al 2011 On the C-statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data

Introduction

- Risk score system needed for survival analysis
- Key component is distinguishing between subjects that develop event (cases e.g. death in many survival studies) and subjects that do not (control group)
- For binary outcomes(survival/death) ROC (Receiver Operating Characteristics curve) or area under curve (AUC) have been developed
- Authors call these methods c-statistic
- Estimation based on cond. probability for any pair of "case" & "control" the predicted risk
- When response variable is time, we can use above
- If not interested in time point, standard concordance measure used to evaluate overall performance of scoring system
- Vaguely speaking, one can distinguish between methods that use loss functions (between risk score and survival time) and other based on rank correlations between these variables
- C-statistic by Harell at its core is a rank correlation measure, using Kendall's tau for censored surv. data
- Problem with rank correlation: **how to order survival times in the presence of censoring**
- Brown et al.: all observations, giving prob. scores based on Kaplan Meier estimate for T
- Problem: KM not good when **covariates dependent on T**
- Alternative: **"usable pairs"**, excluding rest- Problem: **Dependency on censuring distribution**
- Solution : **modified c-statistic which is consistent for population concordance measure, free of censoring**

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., ... & Kattan, M. W. (2010). Assessing the performance of prediction models: a framework for some traditional and novel measures.

Introduction

- The task of diagnostic and prognostic constitute similar challenges
- The clinician has some information and wants to know how this relates to true patient state currently(diagnostic) or in the future (prognostic)
- Traditional to evaluate statistical prediction we can quantify explained variation via R squared or brier score
- Calibration can be quantified via Hosmer Lemeshow goodness of fit test

- Discrimination can be quantified via measure of sensitivity and specificity via concordance statistics
- New measures include reclassification tables, re-classification improvement and integrated discrimination improvement
- The concept of risk re-classification has caused substantial discussion in methodology and clinical literature
- Decision analytics tools have also found audience for studies focusing on clinical usefulness

Predictions models in medicine

- Model extension with a marker: key interest of clinicians is whether to add a marker to an existing model
- Problem: Overoptimistic weight on marker performance
- Only interested in the incremental added value of the marker
- Usefulness: For a model to be useful it has to be well calibrated
- Decision support important

Traditional performance measures

- Difference between goodness of fit and predictive performance is evaluation on same data while latter requires new data or cordds validation
- **Brier Score**: Quadratic scoring rule, where squared differences between actual binary outcome Y and prediction r are calculated
- model ranges from 0 to 0.25 for a non-informative model with a 50% incidence of the outcome
- For survival data we used weight function, considering conditional probability of being uncensored during time
- Calculate brier score at fixed time points and create time dependent curve
- Score measures discrimination and calibration each which can be assessed separately

Discrimination

- Concordance statistic is most used for discrimination measure with generalized linear regression models.
- As a rank order statistic, ** it is insensitive to systemic errors in calibration such as differences in average outcome**
- popular extension: c statistic with censored data can be obtained by ignoring the pairs that cannot be ordered
- Gonen und Heller propose method where c statistic is independent of censoring
- Time dependent c-statistics have proposed (see above)
- Could also use the discrimination slope to see how weel subject with and without outcome are separated

Calibration

-
-

Novel performance measure

Reclassification

Decision Analysis Curve

Conclusion

Integrated Area under the Curve

$$\text{AUC} = \Pr\left\{z\left(\mathbf{X}_i\right) > z\left(\mathbf{X}_j\right) \mid D_i = 1 \& D_j = 0\right\}$$

Prognostic Index Curve

Prediction error Curves

Steyerberg, E. W., & Vickers, A. J. (2008). Decision curve analysis: a discussion.

Implementation

Gerds, T. A., Cai, T., & Schumacher, M. (2008). The performance of risk prediction models.

Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: a novel method for evaluating prediction models.

- Contribute to existing models by adding weight of clinical consequences
- Considered valuable from a clinical usefulness perspective
- Clearly, a linear weighting of false negative and false positives does not make any sense from the perspective of a clinician
- Clearly a false negative is way more detrimental than a false positive result (e.g. cancer patient , disease spread and when false positive we would go for further testing and detect error)
- Decision curve analysis allows one to identify the range of threshold probabilities in which model was of value, the magnitude of benefit and which of several methods was optimal
- Compare their method to AUC method claiming that:
- AUC metric focuses solely on predictive accuracy of model
- Cannot tell us whether a model is worth using at all or which of two models is preferable
- AUC does not provide insight into usefulness aka does not account in their model that clinician may have other interests
- As a solution the decision analytic framework accounts for consequences and in theory tell us if a model is worth it or not
- Two general problems: require data such as on cost or quality adjusted life years, not found in the validation data set.
- Cannot be evaluated without further information
- Secondly: decision analysis typically requires test or prediction model evaluated give a binary result s.t. the true and false positive and negative results can be estimated (but prediciton is frequently continuously expressed)
- All in all, decision curve analysis is a complement not a substitue for AUC
- Interpretation requires understanding of the liking of the patient *
- The proposed metho does not require obtaining information regarding treatment preferences but need theoretical relation for the threshold probabilty of diease and the realtive value of false positive and false negative results

Cook, N. R., & Ridker, P. M. (2009). Advances in measuring the effect of individual predictors of cardiovascular risk:

- Net reclassification improvement (NRI)
- Integrated discrimination improvement
- IDI is equivalent to testing whether the regression coefficient in a model is equl to zero (similar to R^2 or the proportion of variance explained)
- The NRI and the IDI both condition on the case-control or later disease status
- don't provide information on calibration of the estimated risk
- A limitation of NRI and other reclassification measures is that they depend on the particular categories used
- The calibration test seems to depend somewhat less on the number of categories since the degree of freedom adjust for the number of cateogries
- Suggest that reclassification calibration statistic and NRI may be useful in demonstrating the ability of new models and markers to change risk strata and alter treatment decision

Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: a novel method for evaluating prediction models.

- Compare their method to AUC method claiming that:
- AUC metric focuses solely on predictive accuracy of model
- Cannot tell us whether a model is worth using at all or which of two models is preferable
- AUC does not provide insight into usefulness aka does not account in their model that clinician may have other interests
- Two general problems: require data such as on cost or quality adjusted life years, not found in the validation data set. Cannot be evaluated without further information
- Secondly: decision analysis typically requires test or prediction model evaluated give a binary result s.t. the true and false positive and negative results can be estimated (but prediciton is frequently continuously expressed)
- The proposed method does not require obtaining information regarding treatment preferences but need theoretical relation for the threshold probabilty of diease and the realtive value of false positive and false negative results

Vickers, A. J., Van Calster, B., & Steyerberg, E. W. (2016). Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. bmj, 352, i6.

- In contrast to measuring sensitivity or specificity, here measuring net benefit of using model, marker or test in clinical decision would do more good than harm
- By incorporating clinical consequences, s.a. benfit of finding disease early or the harm of unnecessary further testing
- Net benefit is simply type: we rescale harm by exchange rate to put on same scale as benefit
- plotting net benfit against range of exchange rates in what is called a decision curve
- This exchange rate can be dervied by asking questions such as max. number of patient a doctor would recommend for biopsy to find one cancer. (subjective)
- plot for a lot of reasonable exchange rates in a decision curve

Evaluation of Performance Steyerberg

Zhang, Z., Cortese, G., Combescure, C., Marshall, R., Lee, M., Lim, H. J., & Haller, B. (2018). Overview of model validation for survival regression model with competing risks using melanoma study data. Annals of translational medicine, 6(16).

Schoop, R., Beyersmann, J., Schumacher, M., & Binder, H. (2011).

Cook, N. R. (2008). Statistical evaluation of prognostic versus diagnostic models: Beyond the ROC Curve

- Prognostic models account for dimenison of time and introduce a stochastic element
- ROC used to evaluate separation ability between cases and non-cases (discrimination)
- **Core hypothesis:** Risk classification can aid in comparing clinical impact of two models on risk for both individual and population
- Outcome not only unknown but might not exist yet
- Prognostic studies used for risk stratification for assessing the level of risk *Example: Framingham risk score for cardiovascular disease* Screen for early detection can be done for diagnostic such as cancer screening or colonsoscopy Screening results can subsequently be used in prognostic models for later events
- discrimination and calibration both important for prognostic studies
- Area under the curve is also known as he c-statistic or the c index and can rage fromm 0.5 (no predictive ability, bascially random) or 1 (perfecr discrimination)
- c statistic is based on rank of predicted probabilites and compares ranks in indivudals with and without disease
- Similar to wilcoxon rank sum statistic (see 9)
- Cam be computed parametic and nonparametic
- The ROC curve and c statistic are insenstive in assessing the impact of adding new predictiors to a score or predictive model (see 14)
- The change in the ROC curve depends on both the predictive ability of the orginal set and the strength of the new marker , as well as the correlation between them
- For calibration, the most popular measure is the hosmer lemeshow goodness of fi test (16)
- Problem here: sensitive to how he subgroups are formed but groups need to be formed for evaluation (17)
- When predicting true probability, perfect calibration
- Clinical reclassification measures only discrimination
- Advantage over ROC is that categories based on clinical important risk estimations
- For clinical studies ofen the intermediate risk categories are most important where treatment is questionable
- In summary, NRI and calibration test for cross-classified categories can be used to foramlly assessed clinical utility

Wu, Y. C., & Lee, W. C. (2014). Alternative performance measures for prediction models. PLoS One, 9(3), e91249.

- sbrier upon addition of new markers i equal to the IDI index
- Gini, pietra nd sbrier can be expressed as ration, comparing the resolution power
- Further work needed to fully develop the statistical inference procedures
-

van Houwelingen, H. C. (2000). Validation, calibration, revision and combination of prognostic survival models. Statistics in medicine, 19(24), 3401-3415.

References

Assel, M., Sjoberg, D. D., & Vickers, A. J. (2017). The Brier score does not evaluate the clinical utility of diagnostic tests or prediction models. Diagnostic and prognostic research, 1(1), 1-7.

Antolini, L., Boracchi, P., & Biganzoli, E. (2005). A time-dependent discrimination index for survival data. Statistics in medicine, 24(24), 3927-3944.

Bender, A., Rügamer, D., Scheipl, F., & Bischl, B. (2020). A General Machine Learning Framework for Survival Analysis. arXiv preprint arXiv:2006.15442.

Cook, N. R. (2007). Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation, 115(7), 928-935.

Cook, N. R. (2008). Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. Clinical chemistry, 54(1), 17-23.

Gerds, T. A., & Schumacher, M. (2006). Consistent estimation of the expected Brier score in general survival models with right-censored event times. Biometrical Journal, 48(6), 1029-1040.

Gerds, T. A., Cai, T., & Schumacher, M. (2008). The performance of risk prediction models. Biometrical Journal: Journal of Mathematical Methods in Biosciences, 50(4), 457-479.

Graf, E., Schmoor, C., Sauerbrei, W., & Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. Statistics in medicine, 18(17-18), 2529-2545.

Harrell Jr, F. E., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in medicine, 15(4), 361-387. Chicago

Heagerty, P. J., Lumley, T., & Pepe, M. S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics, 56(2), 337-344.

Kattan, M. W., & Gerds, T. A. (2018). The index of prediction accuracy: an intuitive measure useful for evaluating risk prediction models. Diagnostic and prognostic research, 2(1), 7.

Pencina, M. J., D'Agostino Sr, R. B., D'Agostino Jr, R. B., & Vasan, R. S. (2008). Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Statistics in medicine, 27(2), 157-172

Potapov, S., Adler, W., & Schmid, M. (2012). survAUC: Estimators of prediction accuracy for time-to-event data. R package version, 1-0.

Steyerberg, E. W., & Vickers, A. J. (2008). Decision curve analysis: a discussion. Medical Decision Making, 28(1), 146-149.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., ... & Kattan, M. W. (2010). Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology (Cambridge, Mass.), 21(1), 128.

Steyerberg, E. W. (2019). Clinical prediction models. Springer International Publishing.

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., & Wei, L. J. (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Statistics in medicine, 30(10), 1105-1117.

van Houwelingen, H. C. (2000). Validation, calibration, revision and combination of prognostic survival models. Statistics in medicine, 19(24), 3401-3415.

Vickers, A. J., Van Calster, B., & Steyerberg, E. W. (2016). Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. bmj, 352, i6.

Wang, P., Li, Y., & Reddy, C. K. (2019). Machine learning for survival analysis: A survey. ACM Computing Surveys (CSUR), 51(6), 1-36. Chicago

Wu, Y. C., & Lee, W. C. (2014). Alternative performance measures for prediction models. PLoS One, 9(3), e91249.

Zhang, Z., Cortese, G., Combescure, C., Marshall, R., Lee, M., Lim, H. J., & Haller, B. (2018). Overview of model validation for survival regression model with competing risks using melanoma study data. Annals of translational medicine, 6(16).