Machine Learning for Survival Analysis: A Case Study on Recurrence of Prostate Cancer

Blaž Zupan[1,2,4], Janez Demšar[1], Michael W. Kattan[3], J. Robert Beck[4], and I. Bratko[1,2]

[1] Faculty of Computer Science, University of Ljubljana, Slovenia,
[2] J. Stefan Institute, Ljubljana, Slovenia
[3] Memorial Sloan Kettering Cancer Center, New York City, NY, USA
[4] Baylor College of Medicine, Houston, TX, USA

Abstract. This paper deals with the problem of learning prognostic models from medical survival data, where the sole prediction of probability of event (and not its probability dependency on time) is of interest. To appropriately consider the follow-up time and censoring — both characteristic for survival data — we propose a weighting technique that lessens the impact of data from patients for which the event did not occur and have short follow-up times. A case study on prostate cancer recurrence shows that by incorporating this weighting technique the machine learning tools stand beside or even outperform modern statistical methods and may, by inducing symbolic recurrence models, provide further insight to relationships within the modeled data.

1 Introduction

Among prognostic modeling techniques that induce the models from medical data, the survival analysis methods are specific in both the modeling and the type of data required. The survival data normally include the *censor* variable that indicates whether some outcome under observation (like death or recurrence of a disease) has occurred within some patient specific *follow-up time*. The modeling technique has then to consider that for some patients the follow-up may end before the event occurs. In other words, it must take into account that for the patients for which the event has not occurred during the follow-up period it might have occurred just after it.

A well-accepted statistical technique that appropriately considers the follow-up time and censoring is the Cox proportional hazards model [3]. Alternative machine learning approaches based on artificial neural networks (ANN) have been investigated by Ripley and Ripley [11]. Since ANNs are primarily developed for classification tasks, the simplest way to employ them for survival analysis is to model the occurrence of event within a specific follow-up time. This requires the omission of patients with shorter follow-up and for whom the event did not occur, thus potentially biasing the probability estimates of event. Another approach proposed by the same authors but potentially suffering from similar biasing problem is to divide survival times into a set of non-overlapping intervals

W. Horn et al. (Eds.): AIMDM'99, LNAI 1620, pp. 346–355, 1999.

in order to model each interval separately. Alternatively, one can employ the statistical techniques to estimate the survival probabilities and model them with some machine learning technique. Using this scheme, Biganzoli et al. [2] estimate probabilities with logistic regression and feed them to ANN. Similarly, Kattan et al. [5] also use ANN, but instead model the patient's null martingale residual.

Typically, given the patient's data, survival models attempt to determine the probability of event to occur within a specific time. Frequently, however, there are cases in survival analysis where the prediction of *whether* the event will eventually occur or not is of primary importance. For example, for the urologist deciding whether to operate on patients with clinically localized prostate cancer the probability of cancer recurrence is a very important decision factor. In such cases, the survival analysis requires purely classification models that classify either to occurrence or non-occurrence of event, optionally model the class probabilities, and appropriately consider the censoring.

In this paper, we propose a framework which uses selected machine learning techniques to construct classification models from survival data. To properly address censoring in the training data, a weighting technique is proposed that lowers the importance of patients with short follow-up time and for whom the event does not occur. We investigate the applicability of this framework to the problem of modeling prostate cancer survival data and compare machine learning methods to standard statistical survival analysis techniques. The potential advantages of the proposed framework stem from the advantages of the selected machine learning methods. Symbolical induction techniques can help to understand underlying relationships in the prostate cancer data. Some machine learning techniques can discover and use non-linearities and variable interactions, thus overcoming the limitations of linear statistical predictors. We use three different statistics to examine the performance of machine learning methods and compare them to statistical approaches.

We begin by describing the prostate cancer dataset used (Section 2). The applied machine learning and statistical methods are described next (Section 3), with an emphasis on computing and employing the appropriate weights for the patient's data. The same Section also describes the experimental design and statistics that were used to compare the performance of resulting models. Section 4 presents the experimental results and discusses the differences and advantages of selected prediction methods. Section 5 summarizes the results and concludes the paper.

2 Patient Data

The dataset initially consisted of records from all 1055 patients admitted to The Methodist Hospital (Houston, TX) with the intent to operate on their clinically localized prostate cancer between June 1983 and December 1996. Excluded from analysis were the 55 men initially treated with radiation, and 1 treated with cryotherapy. Sixteen men whose disease status (free of disease versus cancer recurrence) was unknown were also excluded. The mean age was 63 years and 85% of the patients were Caucasian.

We selected the following routinely performed clinical variables as predictors of recurrence: pretreatment serum PSA levels (`prepsa`), primary (`bxgg1`) and secondary Gleason grade (`bxgg2`) in the biopsy specimen, and clinical stage assigned using the TNM system (`uicc`) [9]. Treatment failure was defined as either clinical evidence of cancer recurrence or an abnormal postoperative PSA (0.4 ng/ml and rising) on at least one additional evaluation. Patients who were treated with hormonal therapy (N=8) or radiotherapy (N=25) after surgery but before documented recurrence were treated as failures at the time of second therapy. Patients who had their operation aborted due to positive lymph nodes (N=24) were considered immediate treatment failures. To accommodate for some of the modeling methods used, we additionally excluded 16 men having either primary or secondary or both Gleason grades unknown. The resulting dataset thus included 967 patients, of which 189 (19.5%) recurred. For the methods that only use discrete predictor variables (e.g. naive Bayes and association rules), the PSA level was discretized using 5 intervals by computing the quartiles from the training data.

3 Methods

Several statistical and machine learning modeling methods were used and evaluated. They include classification methods (logistic regression, decision trees, adaptation of association rules, naive Bayesian classifier, and artificial neural networks), statistical survival analysis methods (Cox proportional hazards model) and regression methods (artificial neural networks). The resulting models were compared on the basis of the weighted classification accuracy, weighted average probability assigned to the correct class and concordance index. To use the classification-based techniques, the patient's data was weighted according to follow-up time and censor (recurrence).

3.1 Weight Assignment

For the purpose of learning the classification models, data of each patient was assigned a corresponding weight. The weight of the patient that recurred is 1 (one *knows* that the patient recurred). As the certainty that non-recurrent patients will not recur grows with their follow-up time, they have to be weighted accordingly. Their weights are derived from the the null martingale residual (NMR) [12,5], which is computed from the follow-up time and the censor indicator of whether the patient recurred. Computation of the NMR is completely independent of the predictor variables and simply represents the difference between the observed and expected number of recurrences which should have been observed for that point in time (i.e., the patient's follow-up time).

NMR is interpreted as being proportional to the risk of the recurrence for the patient given only his follow-up time [5]. Intuitively, the lower the risk of recurrence, the more likely it is that the patient that is non-recurrent is also a good example for the patients that never recur. Thus, we weighted the non-recurrent patients with weights that were proportional to 1 – NMR. We also assumed that the non-recurrent patients with follow-up time of more than 5 years never recur. The weights were linearly scaled so that a patient with hypothetical follow-up time of 0 would have a weight equal to 0, and a patient with a follow-up time of 5 years or more would have a weight equal to 1.

3.2 Modeling Techniques

The following modeling techniques were used:

Decision trees: our own implementation of the ID3 recursive partitioning algorithm [10] was used that included pre- and post-pruning as proposed by [8]. Weights were used in the estimation of probabilities. The basic idea of ID3 is to divide the patients into ever smaller groups until creating the groups with all patients corresponding to the same class (recurrent, non-recurrent). The division criteria is a function computed from predictor variables.

Naive Bayesian Classifier: assuming the independence of attributes, the probability that a patient described with values of predictor variables $V = (v_1 ... v_n)$ recurs can be estimated by Bayesian formula

$$P(R|V) = P(R) \prod_{i=1}^{n} \frac{P(R|v_i)}{P(R)}$$

where $P(R)$ is the apriori probability of recurrence and $P(R|v_i)$ is the conditional probability of recurrence if i-th predictor variable has the value v_i; both are estimated from the training set of patients. Note that this formula can be derived from the more common form $P(R|V) = P(R)/P(V) \prod_i P(v_i|R)$ by reusing the Bayesian rule $P(v_i|R) = P(R|v_i)P(v_i)/P(R)$. The probability for non-recurrence is computed in the same way and the resulting probabilities must be normalized to sum to 1.

Association rules: introduced in 1993 by Agrawal [1], association rules search for regularities in the data as the rules of the form *precondition → consequence*. The "quality" of a rule is measured by its *support*, i.e. the proportion of patients for which the rule was observed, and the *confidence*, the proportion of patients for which the consequence hold among the patients which satisfy the precondition part of the rule. Only the rules with a reasonable support and confidence level are taken into account. In our implementation, we have restricted the preconditions to include only predictor variables, and the consequence to include only the prediction of recurrence or non-recurrence. For prediction, the voting technique is used with each rule for which the patient's data satisfy the precondition part voting with the weight proportional to its support. The probability for recurrence is then predicted as a normalized number of votes for recurrence.

Artificial neural network: feed-forward neural network with a single hidden layer as available by `nnet` package for S-PLUS [13] was used. The ANN either modeled the recurrence or NMR, i.e., was used either for classification or regression.

Logistic regression: we used a logistic regression available through the command `glm` in S-PLUS.

Cox proportional hazards model as implemented in S-PLUS was used. Using the Cox model for prediction, the probability was estimated for the patients to recur within 5 years after the operation.

3.3 Experimental Design and Evaluation Statistics

To evaluate the modeling methods, a standard technique of stratified 10-fold cross-validation was used [7]. This divides the patient data set to 10 sets of approximately equal size and equal distribution of recurrent and non-recurrent patients. In each experiment, a single set is used for testing the model that has been developed from the remaining nine sets. The evaluation statistics for each method is then assessed as an average of 10 experiments. The same training and testing data sets were used for all modeling methods. To assess the performance of the model from the test datasets, the following statistics were derived:

Classification accuracy (CA), which is expressed in percent of patients in the test set that were classified correctly. Where induced models output probability of recurrence, a probability of higher than 0.5 was considered as a prediction for a patient to recur.

Average probability assigned to the correct class (AP). For the patients in the test set, the probabilities are assigned by the induced model for each of the classes ("recur" and "does not recur"). Knowing the "correct" class, the corresponding probabilities are averaged across the patients in the test set. AP of 1.0 would thus mean that the model always predicted the right class and assigned it the probability of 1.0.

Concordance index (CI), the measure developed by Harell [4], is interpreted as the probability that, given two randomly drawn patients, the patient who recurs first has had predicted a higher probability of recurrence. CI is computed from the testing data set as a proportion of consistent patient pairs over the number of usable patient pair. A patient pair is usable if a patient with a shorter follow-up time recurred. A pair is consistent, if the patients with a shorter follow-up time is assigned a higher probability of recurrence.

The problem with CA and AP occurs when predicting for non-recurrent patients with short follow-up times. Intuitively, if a non-recurrent patient with a short follow-up time is misclassified the error made would be smaller than in the case of a non-recurrent patients with a long follow-up time. For this reason, we weight the patients in the test sets as well (see Section 3.1) and adjust CA and AP scores accordingly. We call the resulting statistics a weighted CA and weighted AP, respectively.

Table 1. Results of performance evaluation. The best two scores for each statistics are printed in bold.

modeling technique	outcome modeled	weighted classification accuracy	weighted prob. assigned to the correct class	concordance index
default		73.1	0.606	0.500
naive Bayes	recurrence	**75.5**	**0.706**	**0.759**
association rules	recurrence	74.8	0.661	0.717
decision tree	recurrence	73.2	**0.662**	0.653
ANN	recurrence	72.5	0.639	0.729
logistic regression	recurrence	73.4	0.626	**0.769**
ANN	NMR	65.6	-	0.734
Cox	recurrence	**75.8**	0.625	0.756

4 Results and Discussion

Table 1 shows the three performance measures when different modeling techniques are applied to prostate cancer survival data. Overall, naive Bayes and Cox proportional hazards model seem to perform best. Logistic regression obtained the highest concordance index, while it performed poorer on the other two statistics used. Note that for most methods the classification accuracy is only slightly above the “default”, which classifies to the majority class in the training set (non-recurrence).

The results for concordance index are very similar to those reported in Kattan et al. [5], although they have used a different validation technique (a repetitive drawing of 70% cases for training while using the remaining 30% for testing). They obtained 0.74 for Cox model and 0.76 for ANN using NMR as the outcome.

The neural network used three neurons in the hidden layer — an architecture that yielded the best overall performance. Although for other methods their parameters could be tuned for best performance as well, such study exceeds the primary intention of the paper to demonstrate the utility of machine learning tools for survival data analysis. Thus, methods were run with their default parameters instead.

A decision tree as induced from the complete dataset is given in Fig. 1. The tree is in concordance with physiological knowledge on this domain, and interestingly brings up a secondary Gleason score (`bxgg2`) as the most important factor for the recurrence prediction. The tree also pinpoints an anomaly in the prostate recurrence data used: a clinical stage T1ab is expected to be less severe than stages T1c to T2c, yet the tree predicts the opposite. This indicates that the data may undersample this problem subspace, and further analysis (potentially using additional data) is required to investigate this anomaly.

Association rules that predict the recurrence are shown in Fig. 2. The required minimal support was 0.05 and confidence 0.2. Note that the rules mostly involve the conditions on both Gleason scores (`bxgg1` and `bxgg2`) requiring these to be

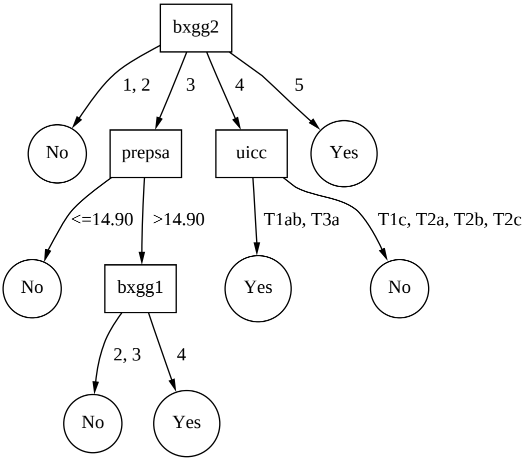

Fig. 1. Decision tree for prostate cancer recurrence prediction.

3 or higher, and PSA level (`prepsa`) being higher than 14.9. There were 25 rules (not shown here) with same requirements on support and confidence found that predict to non-recurrence. We observed that some conditions of rules from both groups overlap, making the interpretation harder but also suggesting that the rules for recurrence with conditions not found in the other groups should be considered important. An example of such rule is `bxgg2=4 -> recur`, the importance of which was also confirmed by physicians.

We additionally analyzed the performance of classifiers by means of calibration curves $q(p)$ [13], where q is the fraction of recurrent patients for which the model predicted the recurrence probability p. Ideally, a calibration curve would be a straight line $q = p$. Fig. 3 shows a calibration curve for naive Bayes and Cox model. Both models are overconfident when predicting recurrence, especially when probability goes toward 1. Similar overconfidence was observed for other classifiers as well. Naive Bayes seems more accurate when predicting lower probabilities of recurrence.

An interesting calibration curve is that for association rules: the curve is close to ideal, but shows also the major weakness of this predictor: its highest predicted probability of recurrence is about 0.7. This also indicates that the method could be improved provided more appropriate voting mechanism that decides for and against the recurrence can be found.

Finally, we show a graphical device called a nomogram [6] that uses the naive Bayesian formula to compute recurrence probability. The nomogram (Fig. 4) shows the impact of individual features on probability of recurrence (upper labels on feature lines) and non-recurrence (lower labels). The values right of zero favor (non)recurrence and the values on the left speak against it. For example observe `bxgg2` and non-recurrence: values of 5 and 4 vote against, and values 3, 2 and 1 vote for non-recurrence. Nomogram can be used to compute the probabilities of outcomes. First, the impact factors for feature values must be summed, once

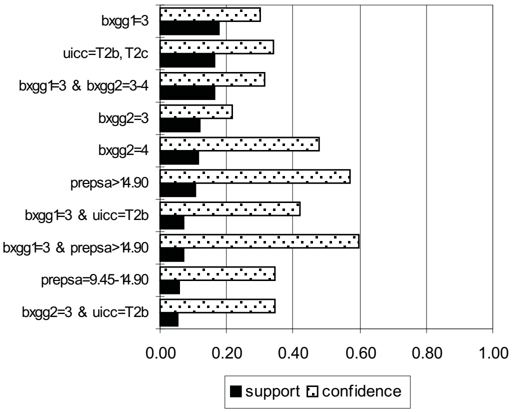

Fig. 2. Association rules for prediction to recurrence of prostate cancer.

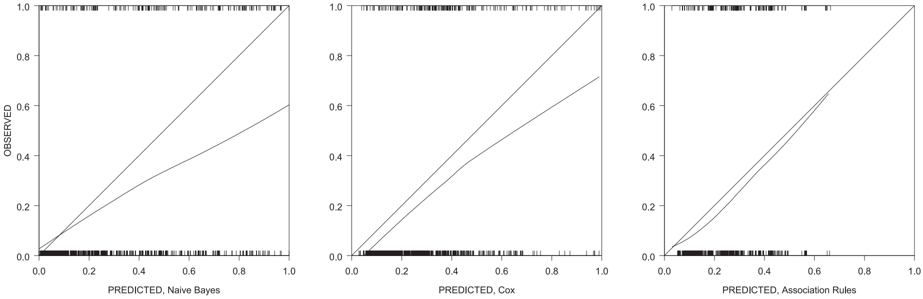

Fig. 3. Calibration curves for naive Bayes (left), Cox model (middle) and association rules (right).

for recurrence and once for non-recurrence, using the scale above (below) the table. The sums are then converted into probability estimation using the lookup graph below and, finally, normalized to sum to 1. For example, patient (`bxgg1=1`, `bxgg2=3`, `prepsa=11`, `uicc=T2a`) has the sum $-0.54-0.09+0.11-0.21=-0.73$ for recurrence and $0.1+0.03-0.05+0.05=0.13$ against. Approximation by the lookup table gives 0.06 for recurrence and 1.0 against which, multiplied by $(0.06+1.0)^{-1}$, gives the probabilities of 0.057 for and 0.943 against recurrence.

The nomogram also points out some specifics about the recurrence domain we are modeling. It reveals that the two Gleason scores are the most important factors for the decision as their values are most dispersed through the score line that nomogram provides — an observation which is in accordance with findings by association rules and decision tree. Furthermore, the anomaly concerning the stage T1ab also pointed out by decision tree is also evident. Namely, it would be expected for T1ab to appear before T1c for recurrence ("Yes" side of `uicc` line) and after T1c for non-recurrence ("No" side of `uicc` line).

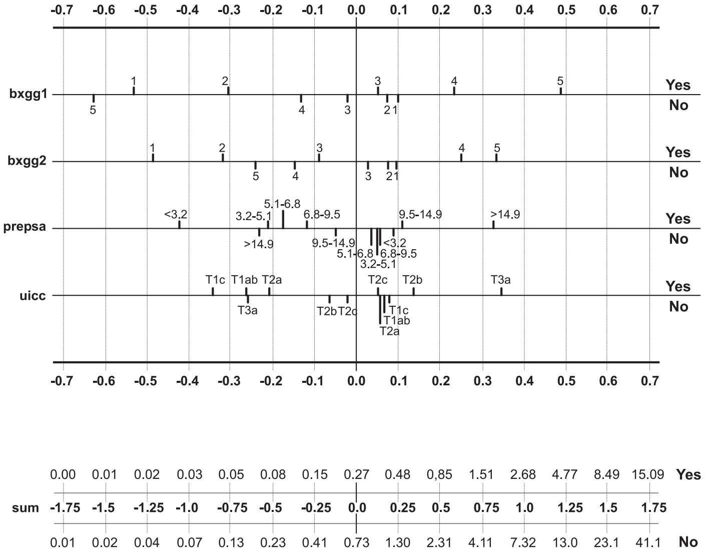

Fig. 4. Nomogram for predicting probability of recurrence and non-recurrence based on probability estimates by Naive Bayes.

5 Conclusion

Deciding whether to operate on patients with clinically localized prostate cancer frequently requires the urologist to classify patients into expected groups such as "remission" or "recur". In this paper we show that models for prostate cancer recurrence that may potentially support the urologist's decision making can be induced from data using standard machine learning techniques, provided that follow-up and censoring has been appropriately considered. For the latter, we propose a weighting technique that lessens the importance of non-recurrent patients with short follow-up times.

The case study on prostate cancer survival data shows that machine learning techniques with proposed weighting schema can, in terms of performance, stand beside or even outweigh standard statistical techniques. The additional feature of inducing interpretable models (like those of decision trees and association rules) was also found beneficial. The best models were obtained by naive Bayesian method, also indicating that for our dataset the potential discovery of non-linearities and variable interaction seems not to play a crucial part (naive Bayesian method does not include them but still outperforms, for example, artificial neural networks).

By inspecting the induced models we can conclude that, for the observed set of patients, the Gleason scores and PSA level are more powerful predictors than clinical stage. In case of Gleason grades 4 and 5 these seem to contribute most to the high probability of recurrence, which is also in accordance to their physiological meaning of "high grade".

The non-recurring patients were weighted by the null martingale residuals, i.e., proportional to their risk of recurrence. We have preliminary tested other weighting techniques (e.g., with weights as a linear or exponential function of follow-up time) and obtained poorer results. Further experimental and theoretical work is needed to gain deeper understanding of the weighting effects.

The authors strongly believe that, although tested only on prostate cancer recurrence data, the proposed methods can be applicable to general survival analysis where the sole prediction of probability of event (and not its probability dependency on time) is of interest.

Acknowledgment

This work was generously supported by the Slovene Ministry of Science and Technology and the Office of Information Technology at Baylor College of Medicine.

References

1. R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proc. ACM SIGMOD Conference on Management of Data*, pages 207–216, Washington, D. C., 1993.
2. E. Biganzoli, P. Boracchi, and L. Mariani, et al. Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. *Statis Med*, 1998.
3. D. R. Cox. Regression models and life-tables. *J R Statist Soc B*, 34:187–220, 1972.
4. F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati. Evaluating the yield of medical tests. *Journal of American Medical Association*, 247(18):2543–2546, 1982.
5. M. W. Kattan, H. Ishida, P. T. Scardino, and J. R. Beck. Applying a neural network to prostate cancer survival data. In N. Lavrač, E. Keravnou, and B. Zupan, editors, *Intelligent data analysis in medicine and pharmacology*, pages 295–306. Kluwer, Boston, 1997.
6. J. Lubsen, J. Pool, and E. van der Does. A practical device for the application of a diagnostic or prognostic function. *Methods of Information in Medicine*, 17:127–129, 1978.
7. D. Michie, D. J. Spiegelhalter, and C. C. Taylor, editors. *Machine learning, neural and statistical classification.* Ellis Horwood, 1994.
8. T. Niblett and I. Bratko. Learning decision rules in noisy domains. In *Expert Systems 86*, pages 15–18. Cambridge University Press, 1986. (Proc. EWSL 1986, Brighton).
9. M. Ohori, T. M. Wheeler, and P. T. Scardino. The new american joint committee on cancer and international union against cancer tnm classification of prostate cancer: Clinicopathologic correlations. *Cancer*, 74:104–114, 94.
10. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.
11. B. D. Ripley and R. M. Ripley. Neural networks as statistical methods in survival analysis. In R. Dybowski and V. Gant, editors, *Artificial Neural Networks: Prospects for Medicine.* Landes Biosciences Publishers, 1998.
12. T. M. Therneau, P. M. Grambsch, and T. R. Fleming. Martingale-based residuals for survival models. *Biometrika*, 77(1):147–160, 1990.
13. W. N. Venables and B. D. Ripley. *Modern applied statistics with S-PLUS.* Springer, New York, second edition edition, 1997.