ELSEVIER Artificial Intelligence in Medicine 20 (2000) 59–75

Artificial Intelligence in Medicine

www.elsevier.com/locate/artmed

Machine learning for survival analysis: a case study on recurrence of prostate cancer

Blaž Zupan [a,b,c,*], Janez Demšar [a], Michael W. Kattan [d], J. Robert Beck [c], I. Bratko [a,b]

[a] *Faculty of Computer and Information Science, University of Ljubljana, Tražaška 25, SI-1000 Ljubljana, Slovenia*

[b] *J. Stefan Institute, Ljubljana, Slovenia*

[c] *Baylor College of Medicine, Houston, TX, USA*

[d] *Memorial Sloan Kettering Cancer Center, New York, NY, USA*

Received 10 November 1999; received in revised form 3 February 2000; accepted 27 March 2000

Abstract

Machine learning techniques have recently received considerable attention, especially when used for the construction of prediction models from data. Despite their potential advantages over standard statistical methods, like their ability to model non-linear relationships and construct symbolic and interpretable models, their applications to survival analysis are at best rare, primarily because of the difficulty to appropriately handle censored data. In this paper we propose a schema that enables the use of classification methods — including machine learning classifiers — for survival analysis. To appropriately consider the follow-up time and censoring, we propose a technique that, for the patients for which the event did not occur and have short follow-up times, estimates their probability of event and assigns them a distribution of outcome accordingly. Since most machine learning techniques do not deal with outcome distributions, the schema is implemented using weighted examples. To show the utility of the proposed technique, we investigate a particular problem of building prognostic models for prostate cancer recurrence, where the sole prediction of the probability of event (and not its probability dependency on time) is of interest. A case study on preoperative and postoperative prostate cancer recurrence prediction shows that by incorporating this weighting technique the machine learning tools stand beside modern statistical methods and may, by inducing symbolic recurrence models, provide further insight to relationships within the modeled data.

* Corresponding author. Tel.: +386-1-4768402; fax: +386-1-4251038.
E-mail address: blaz.zupan@fri.uni-lj.si (B. Zupan).

0933-3657/00/$ - see front matter
PII: S0933-3657(00)00053-1

Keywords: Survival analysis; Censored data; Machine learning; Data weighting; Prostate cancer recurrence; Outcome prediction after radical prostatectomy; Prognostic models in medicine

1. Introduction

Among prognostic modeling techniques that induce models from medical data, the survival analysis methods are specific both in terms of modeling and the type of data required. The survival data normally include the censor variable, which 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 must then consider that for some patients the follow-up may end before the event occurs. In other words, it must take into account that for patients for whom the event has not occurred during the follow-up period, the event may eventually occur.

Typically, given the patient's data, survival models attempt to determine the probability of the 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 the occurrence or to the non-occurrence of event and optionally model the outcome probabilities, and appropriately consider the censoring.

Recently, the machine learning community has developed various tools that have been successfully used in the construction of classification models, including medical prognostic models [15,18]. In this paper, we propose a framework which allows us to use machine learning techniques to construct classification models from survival data. To properly address censoring in the training data, patients for whom the event did not occur and have short follow-up time require special treatment. Note that for them the final outcome is not known with certainty. Trivial solutions to this problem by their removal from the data set or considering them as examples where the event will not occur would bias the modeling [22,12] and should thus be avoided. To properly treat such cases, we propose a technique that assigns a distribution of outcomes instead of a single outcome. The distribution is assessed through the outcome probability estimate based on the Kaplan–Meier method. Since most machine learning techniques do not deal with outcome distributions, the schema is implemented using weighted examples. Although developed independently, the proposed technique is similar to the one used by Ripley and Ripley [22]. The main difference, however, is that they use data weighting only when testing the models, whereas for their construction different approaches are used.

The benefits of the proposed framework stem from the potential advantages of machine learning methods. Symbolic induction techniques can help us to understand underlying relationships in the prostate cancer data. Some machine learning

techniques can discover and use non-linearities and variable interactions [12], thus overcoming the limitations of linear statistical predictors.

We investigate the applicability of the proposed framework to the problem of modeling prostate cancer survival data and use two different machine learning methods. While any machine learning method that induces models from weighted examples may be used, a naive Bayes classifier and induction of decision trees were selected for our study because of their simplicity, acceptance and generally good performance. The two were compared to the Cox proportional hazards model [6], which is a standard statistical survival analysis technique for prediction based on multiple variables.

We use two separate datasets to construct the prostate cancer survival models. The preoperative data set includes data on tests that were administered prior to the prostatectomy (prostate removal), while postoperative dataset also includes data from several routinely performed pathologic tests. Preoperative data are generally fully known at least 2 weeks prior to the operation, while postoperative data generally are complete approximately a month following the operation. Clinically, both prediction models would be very useful. A model based on preoperative data could be used for patient decision making as to whether the ability of prostatectomy is worth the potential treatment complications (impotence and incontinence). If the predicted probability of recurrence were high, patients might choose one of several other treatments which did not have the adverse effects of such an aggressive therapy. Postoperatively, a prediction model is also useful, but for different purposes. If recurrence could be predicted postoperatively, prior to actual recurrence, a second therapy (after the prostatectomy) could be administered quickly, when it is potentially the most effective. Thus, preoperative and postoperative prediction models are both very useful but for different purposes: deciding whether to undergo prostatectomy at all, and then whether to add additional treatment.

In Section 2 we begin by describing two prostate cancer datasets used in our experimental evaluation. The proposed treatment of censored data that uses outcome distributions (data weighting) is described in Section 3, together with a description of machine learning techniques, 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. An overview of related work is given in Section 5. Section 6 summarizes the results and concludes the paper.

2. Patient data

Two prostate cancer datasets were used in this study. They both include patients that were treated with radical prostatectomy, and were followed-up to observe the recurrence of the cancer. While the first dataset includes only preoperative data, the second dataset additionally incorporates data gathered postoperatively. The task in both cases was to construct a model that would, given the corresponding patient's data, predict the probability of recurrence.

2.1. Preoperative data

The preoperative dataset consists 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 55 men initially treated with radiation, and one 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.

Four routinely performed clinical tests were selected as predictors of recurrence (Table 1). 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 (S-Plus implementation of Cox's proportional hazards model), we additionally excluded 16 men having either primary or secondary or both Gleason grades unknown. For the naive Bayes classifier, the PSA level was discretized using five intervals by computing the quintals from the training data.

The resulting dataset thus included 967 patients, of which 189 (19.5%) recurred, and of those that did not recur 68 had follow-up time of equal or longer than 7 years (7.0%) and 710 (73.4%) shorter than 7 years. For the last group, the mean follow-up time was 37.5 months.

2.2. Postoperative data

All 1055 patients mentioned above, plus those additionally treated with radical prostatectomy at the same hospital from December 1996 to June 1997 were

Table 1
Variable names and descriptive statistics for preoperative data[a]

Variable	Abbreviation	Values and distributions
Primary Gleason grade	gg1	1 (14), 2 (260), 3 (594), 4 (95), 5 (4)
Secondary Gleason grade	gg2	1 (12), 2 (164), 3 (543), 4 (237), 5 (11)
Clinical stage	Stage	T1ab (82), T1c (145), T2a (264), T2b (239), T2c (180), T3a (57)
Preoperative PSA (ng/l)	PrePSA	Continuous, min = 0.1, mean = 9.9, max = 100.0

[a] The numbers of corresponding patients are given in brackets.

Table 2
Variable names and descriptive statistics for postoperative data

Variable	Abbreviation	Values and distributions
Gleason sum	Gleason	3 (2), 4 (5), 5 (106), 6 (350), 7 (454), 8 (61), 9 (14), 10 (4)
Prostatic capsular invasion	PCI	None (184), invading capsule (396), focal (152), established (264)
Surgical margins	SurgMarg	Negative (853), positive (143)
Seminal vesicle invasion	SemVesInv	No (862), yes (134)
Lymph nodes	LNodes	Negative (925), positive (71)
Preoperative PSA (ng/l)	PrePSA	Continuous, min = 0.1, mean = 10.4, max = 100.0

candidates for the postoperative dataset. Excluded were patients with positive lymph nodes and aborted operations (for a detailed description see Ref. [14]). The final postoperative dataset includes 996 patients. In addition to pretreatment prostate specific antigen level, five routinely performed pathologic tests were included as variables: Gleason sum in the surgical specimen, prostatic capsular invasion, surgical margin status, seminal vesicle invasion and lymph node status (Table 2). The dataset includes 189 (19.0%) patients that recurred. Of those that did not recur, 107 (10.7%) have follow-up time longer or equal to 7 years and 700 (70.3%) have follow-up time of less than 7 years.

3. Methods

The naive Bayes classifier and the induction of decision tree machine learning methods were used and evaluated. Their performance was compared to a Cox proportional hazards model on the basis of the classification accuracy, specificity and sensitivity, correlation of predicted probability and probability estimated by the Kaplan–Meier method, and concordance index (area under receiver operating characteristic curve). We first explain how we treat censored data, then briefly introduce the machine learning techniques, and finally describe the statistics used for comparison.

3.1. Handling censored data

The particular characteristic of survival data is that for some patients the follow-up is too short to determine a definite outcome. For example, it is assumed that if the prostate cancer patient who has undergone radical prostatectomy remains disease free for at least 7 years [12], the cancer has been successfully cured. However, if a non-recurrent patient has been followed up for less than 7 years, the outcome is not certain.

The above reasoning provides motivation to split the prostate cancer survival data into three groups: the first consists of patients that recurred (the outcome is known), the second of patients that did not recur and were followed for more than 7 years after the operation (for these non-recurrence is assumed), and the third of non-recurring patients with follow-up of less than 7 years. For the last group, we assess the probability of each outcome using the Kaplan–Meier method [16]. In essence, for prostate cancer the Kaplan–Meier method estimates the probability of non-recurrence at a particular follow-up time. Fig. 1 gives the Kaplan–Meier survival curve for the preoperative dataset used in this study. It illustrates the overall proportion of patients who remain free from recurrence over time. Time begins with surgery, so the horizontal axis is months following surgery. As time increases, a smaller proportion of patients are left without recurrence. The Kaplan–Meier survival curve for the postoperative dataset is similar to the one shown for preoperative data.

Using Kaplan–Meier estimates from all groups of patients, we compute the probability P_{rf} of recurrence-free outcome for each of the patients from the third group. Given the patient's follow-up time T_{f}, this probability is equal to:

$$P_{\mathrm{rf}} = P(\text{non-recurrence (7 years)} \mid \text{non-recurrence } (T_f)) = \frac{P(\text{non-recurrence (7 years)})}{P(\text{non-recurrence } (T_{\mathrm{f}}))}$$

The patient's probability of recurrence is then $P_{\mathrm{r}} = 1 - P_{\mathrm{rf}}$. The outcome for this patient is, therefore, a distribution $(P_{\mathrm{rf}}, P_{\mathrm{r}})$. Since most of the machine learning tools do not include mechanisms to handle distributions as class values, instead of a single data record for each patient from the third group, two copies of the patient's data are created, one labeled with an outcome 'not-recurred' and weighted with weight P_{rf}, and the other with an outcome 'recurred' and weighted with weight P_{r}.

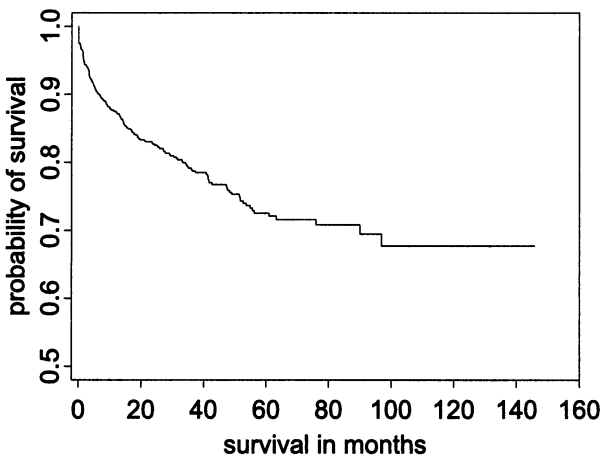

Fig. 1. Plot of the Kaplan–Meier survival probability estimates for the preoperative dataset.

3.2. Modeling techniques

Induction of decision trees and naïve Bayes machine learning method were used to construct prognostic models from data. They were compared to Cox's proportional hazards model, which is a standard statistical technique for modelling censored data.

3.2.1. Decision tree induction

Our own implementation of the ID3 recursive partitioning algorithm [21] was used. The basic idea of ID3 is to divide the patients into ever smaller groups until creating the groups with all or a majority of patients corresponding to the same class (recurrent, non-recurrent). The division criterion is a function computed from predictor variables. The decision tree induction algorithm included pre- and post-pruning. In pre-pruning, the recursive partitioning is stopped if it would build leaves consisting of less than 20 data records. We observed in cross-validation experiments that when this limit was lowered, the performance degraded significantly. For post-prunning, a minimal-error pruning algorithm [20] using the m-probability estimate [5] with $m = 2.0$ was used.

3.2.2. Naive Bayes classifier

Assuming independence of attributes, the probability that a patient described with values of predictor variables $V = (v_1 \ldots v_n)$ recurs can be estimated by the Bayes 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 the ith 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 $P(R|V) = P(R)/P(V)\Pi_i P(v_i|R)$ by reusing the Bayes 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.

3.2.3. Cox's proportional hazards model

Cox's proportional hazards model [6] as implemented in the S-PLUS software (PC Version 4.5; Redmond, WA) was used. Using the Cox's model for prediction, the probability was estimated for the patients to recur within 7 years after the operation.

Decision trees and naive Bayes classifiers are constructed from weighted data (see Section 3.1). The treatment of the weighted examples is fairly straightforward. Both algorithms need to estimate conditional and unconditional probabilities of classes. Instead of the usual formula which divides the number of examples of class C with the number of all considered examples ($P(C) = (\#\text{ examples of } C)/(\#\text{ examples})$), we divide the sum of weights of examples of class C by the sum of weights of all examples considered $\left(P(C) = \left(\Sigma_{E\in C}\text{weight }(E)\right)\Big/\left(\Sigma_{E}\text{weight }(E)\right)\right)$.

3.3. Experimental design and evaluation statistics

To evaluate the proposed weighting schema and the modeling methods, a standard technique of stratified 10-fold cross-validation was used [19]. This divides the patient data set into ten 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 ten experiments. The same training and test data sets were used for all modeling methods. The weights for the non-recurring patients with short follow-ups were estimated by building the Kaplan–Meier survival curve from the patients in the training sets only, and then used to weight the patients in the corresponding test sets. To assess the performance of the model on the test datasets, the following statistics were derived:

— *Classification accuracy*, which is expressed in percent of patients in the test set that were classified correctly. A recurrence probability of higher than 0.5 was considered as a prediction for a patient to recur.
— *Sensitivity* expressed as the probability of correctly predicting recurrence.
— *Specificity* expressed as the probability of correctly predicting non-recurrence.
Correlation of the predicted probability with recurrence-free probabilities estimated from Kaplan–Meier survival curves as described in the previous section.
— *Concordance index*, which is a measure developed by Harrell et al. [10] and is interpreted as the probability that, given two randomly drawn patients where the patient with the shorter follow-up has recurred, the patient who recurs first has a higher predicted probability of recurrence. Notice that this is equivalent to the area under the receiver operating curve [9]. The concordance index is computed from the test data set as a proportion of consistent patient pairs over the number of usable patient pairs. A patient pair is usable if a patient with a shorter follow-up time recurred. A pair is consistent, if the patient with a shorter follow-up time is assigned a higher probability of recurrence.

Accuracy, sensitivity and specificity all use weights assigned to the test examples. On the other hand, probability correlations and the concordance index do not need information about weights.

4. Results and discussion

For preoperative data, Table 3 shows the results when applying different modeling techniques. Overall, the naive Bayes and Cox proportional hazards model seem to perform better than decision trees, although the differences are not significant.

The results for the concordance index are very similar to those reported in Kattan et al. [13], 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 the Cox's proportional hazards model and 0.76 for

Table 3
Results for preoperative data

Modeling technique	Classification accuracy	Sensitivity	Specificity	Probability correlation	Concordance index
Default	68.1	0.0	100.0	0.00	0.50
Naive Bayes classifier	70.8	35.8	87.1	0.41	0.75
Decision tree induction	68.8	22.9	90.3	0.37	0.72
Cox	69.7	19.5	93.2	0.39	0.76

ANN using null martingale residual as the outcome. In their later study [11], using the Cox's model only and bootstrapping for validation, they have obtained a concordance index of 0.79.

Further differences among the performance of predictive methods when assessed by 10-fold cross validation may be additionally analyzed by means of calibration curves $P_{KM}(P_r)$ [24]. This is constructed as follows. Say that a patient with an estimated probability of recurrence P_{KM} by the Kaplan–Meier method is presented to one of our prediction models, which predicts a recurrence probability of P_r. For all non-recurrent patients with a follow-up time of more than 7 years P_{KM} is changed to 1.0 (see Section 3.1). For all patients, their corresponding points $P_{KM}(P_r)$ are entered on the graph. Instead of these points, a smoothed curve which best approximates the relationship between two probabilities is computed and presented. Ideally, a calibration curve would be a 45° straight line $P_{KM} = P_r$.

Fig. 2 shows calibration curves for the three modeling methods. The curves for the naive Bayes classifier and Cox are rather similar, with a difference that the naive Bayes classifier seems to become overconfident when predicting recurrence with a probability close to 1. This may be the reason for the higher sensitivity of the naive Bayes classifier. The plot for the decision trees is interesting since it is very close to the ideal curve, but shows that decision trees predict probabilities only within a certain range.

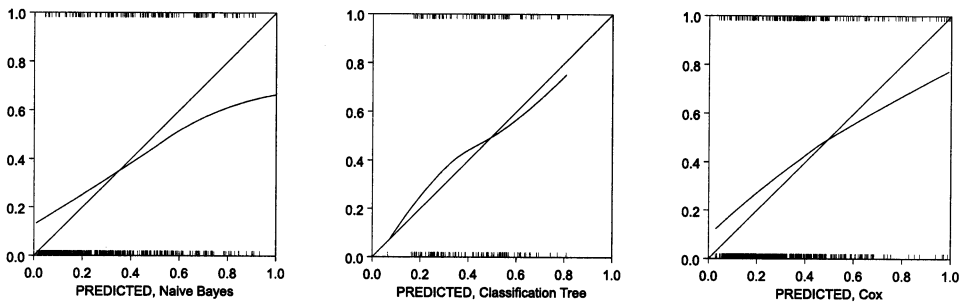

Fig. 2. Calibration curves for the preoperative data.

A classification tree induced from the complete preoperative data set is given in Fig. 3. The tree is relatively small and simple but in concordance with physiological knowledge on this domain, and uses preoperative PSA followed by secondary Gleason grade and clinical stage as the most predictive variables. Our previous study [25] using a different weighting technique developed a decision tree with secondary Gleason grade at its root. The differences between these trees may be attributed to the high similarity in terms of informativity of the predictive variables. Thus for this domain, a relatively small change in data (change in weights) may result in different classification trees.

Next, a naive Bayes classifier was constructed from a complete preoperative data set. To analyze it we here show a graphical device called a nomogram [17] that uses the naive Bayes formula to compute recurrence probability. The nomogram (Fig. 4) shows the impact of individual features on the probability of recurrence (lower labels on feature lines) and non-recurrence (upper labels). The values to the right of zero (bold vertical line) favor (non)recurrence and the values on the left speak against it. For example, observe gg2 and non-recurrence on the nomogram for preoperative data: values of 5 and 4 vote against, and values 3, 2 and 1 vote for non-recurrence. The nomogram can be used to compute the probabilities of outcomes. First, the impact factors for feature values must be summed for recurrence and for non-recurrence. The sums are then converted into probability estimates using the lookup graph below and, finally, normalized to sum to 1. For example, a patient with preoperative data (gg1 = 3, gg2 = 3, PrePSA = 11, Stage = T2c) has the sum $-0.05 + 0.1 + 0.05 - 0.05 = 0.05$ against and $0.1 - 0.2 - 0.1 + 0.1 = -0.1$ for recurrence. Approximation by the lookup table gives about 84% for

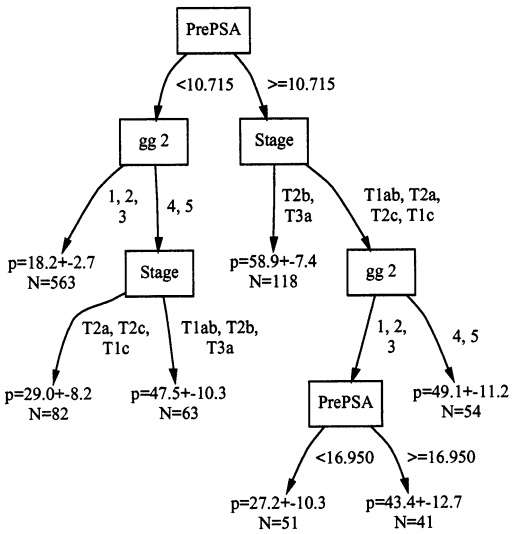

Fig. 3. Decision tree for preoperative prostrate cancer recurrence prediction. Leaf nodes give the probability of recurrence and the number of patient records from which this probability was assessed.

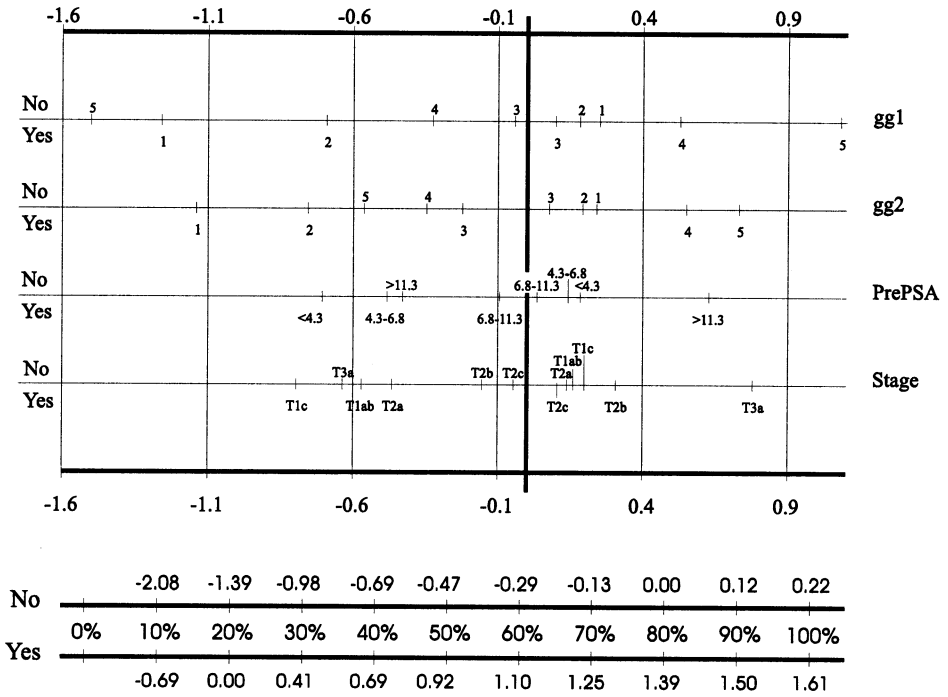

Fig. 4. Preoperative nomogram for predicting probability of recurrence and non-recurrence based on probability estimates by the naive Bayes classifier.

non-recurrence and 18% for recurrence, which multiplied by $(0.84 + 0.18)^{-1}$, gives the probabilities of 82% for and 18% against recurrence.

The nomogram also points out some specifics about the recurrence domain we are modeling. It reveals that the two Gleason scores are very important factors for the decision as their values are most dispersed through the score line that the nomogram provides. It also shows that the other two factors used are informative too, as their values are well dispersed as well. Note, though, that PSA was in our study discretized and under a different discretization the results may not be the same. The potential anomaly concerning the Stage attribute is evident. Namely, the expected order of severity of stages would be T1ab, T1c, T2a, T2b, T2c, T3a, while the nomogram suggests T1c, T1ab, T2b, T2a, T2c, T3a. This indicates that the data may undersample this problem subspace, and further analysis (potentially using additional data) is required to investigate the stage ordering.

For the experiments on postoperative data, the results are summarized in Table 4. As expected, these prognostic models perform, in general, better than with preoperative data. This was expected, since, intuitively, the data gathered during or after the invasive treatment should contain more predictive information than preoperative non-invasive tests. Most importantly, classification accuracy and the concordance index are improved. However, the sensitivity of all modeling tech-

niques investigated is still low, most probably due to the dominance of non-recurrent patients in the data. Interestingly, a similar problem was observed by Ripley and Ripley [22], who proposed to increase the cost of misclassifying recurrent patients. Trivially, this can be done by setting the probability margin for recurrence from 0.5 to some lower value. For both datasets investigated, increasing sensitivity may prove important when constructing prognostic models for clinical use, but investigation of methods that would do so while maintaining other properties of the models was beyond the scope of this paper.

Fig. 5 shows the postoperative calibration curves. Again, the curves are rather similar, with the predictions of decision trees being limited to a narrower range (specifically, very few patients are assigned probabilities of recurrence of higher than 0.8). The naive Bayes classifier seems to overemphasize recurrence more than Cox, hence higher sensitivity but lower specificity. Interestingly, although calibration curves would suggest larger differences between the two methods, Cox and Bayes perform very similarly in respect to classification accuracy and the concordance index. We should not forget, though, that calibration curves compare predicted probabilities with those estimated from data, so they should only be used for informative purposes and interpreted with caution.

Table 4
Results for postoperative data

Modeling technique	Classification accuracy	Sensitivity	Specificity	Probability correlation	Concordance index
Default	70.8	0.0	100.0	0.00	0.50
Naive Bayes classifier	78.4	26.5	89.1	0.62	0.87
Decision tree induction	77.0	18.7	92.8	0.54	0.84
Cox	79.0	18.2	94.2	0.64	0.88

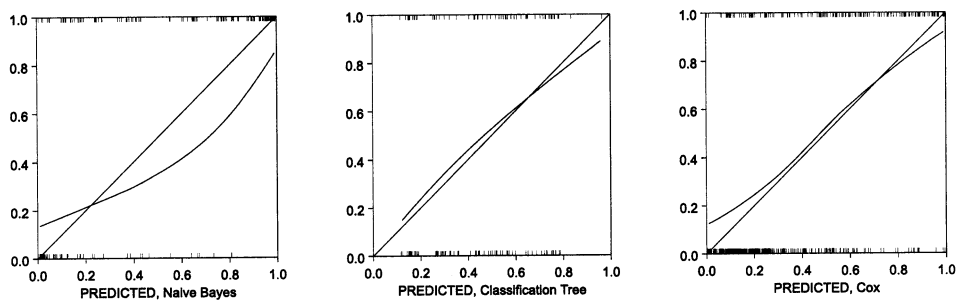

Fig. 5. Calibration curves for the postoperative data.

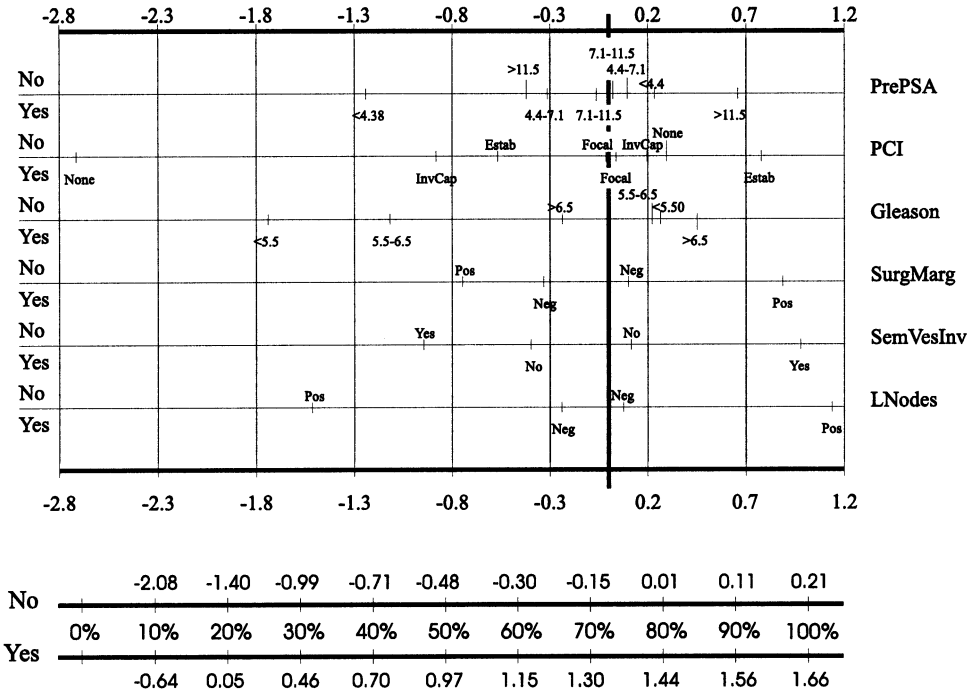

Fig. 6. Postoperative nomogram for predicting probability of recurrence and non-recurrence based on probability estimates by the naive Bayes classifier.

The naive Bayes postoperative nomogram built from a complete postoperative data set is presented in Fig. 6. It shows that while prostatic capsular invasion may be the most predictive variable, others follow closely and for a good predictor all predictive variables may need to be considered. Since decision trees take a minimalistic approach (the tree built from the whole dataset is rather small and does not contain all predictive variables; see Fig. 7), this may be the reason for the slightly poorer performance. The nomogram also shows that if any of the surgical margins, seminal vesicle invasion or lymph nodes involvement are positive, this is strong evidence for recurrence; if they are negative, they do not influence the result that much.

Overall, for both prostate cancer recurrence domains, the collaborating domain experts preferred nomograms over decision trees. The main reason was that the nomograms considered all predictive variables, while when following different paths in the decision trees only a selection of variables is used. Since all predictive variables included in both studies were carefully selected and considered relevant, the latter was viewed as a deficiency.

5. Related work

While there exist various statistical techniques to model survival-type data (e.g. Kaplan–Meier modeling and Cox's regression [16]), machine learning techniques that would appropriately consider censored data are rare. Most notable exceptions come from the area of artificial neural networks, but even there the techniques vary from ignoring censored patients to treating them properly through modeling the hazard function. For instance, Snow et al. [23] developed a neural network to predict recurrence after radical prostatectomy, but treated censored patients as non-recurrent, thus disregarding follow-up time and potentially biasing the predictions towards non-recurrence. A similar approach was used by Burke et al. [4], where patients with short disease-free follow-up were excluded from the model. Their approach would also bias the resulting model, this time towards predicting higher probabilities for recurrence, since for the patients with short follow-up time only those that recurred would be considered. To appropriately consider censoring, Faraggi and Simon [8] used a similar schema as with Cox's proportional hazards model, but instead of using a log-linear relationship between the independent variables and the underlying hazard function, they use an artificial neural network. Biganzoli et al. [3] split follow-up time into non-overlapping intervals, used interval information as an additional input to a neural network and used it to model the probability of failure. A different approach is presented by Kattan et al. [13] that also used a neural network, but instead they modeled the null martingale residual, e.g. the difference between the observed and expected number of recurrences for the given follow-up time. Their analysis shows that neural network may be superior

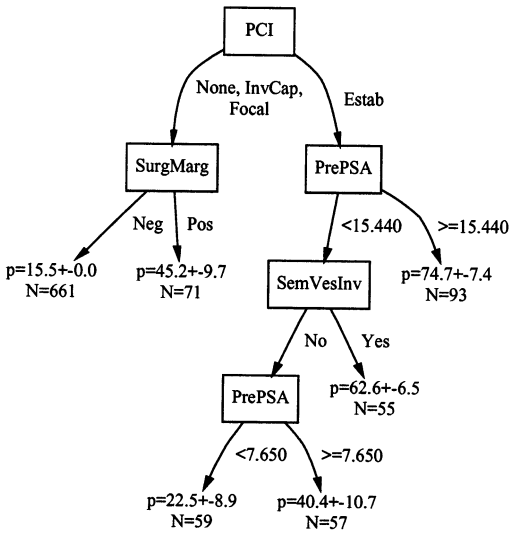

Fig. 7. Decision tree for postoperative prostrate cancer recurrence prediction. Leaf nodes give the probability of recurrence and the number of patient records from which this probability was assessed.

when compared to traditional statistical models, which seems to be attributed to nonlinearities incorporated in the neural network model [12].

A comprehensive overview and classification of existing neural network-based techniques for survival analysis is provided by Ripley and Ripley [22], while D'Amico et al. [7] provide a list of artificial neural networks techniques that specifically target the prostate cancer recurrence prediction.

Anand et al. [2] stress the need to develop the prognostic models that would, instead of hazard or survival functions, explicitly provide a prognostic estimate for an individual patient. They compare regression trees, k-nearest neighbors (k-NN) and a regression variant of an artificial neural network to model the patient's survival time after being diagnosed with colorectal cancer. While they treat the follow-up time for censored patient as the survival time, they propose and subsequently implement an extension of the k-NN method that appropriately treats censoring both in the learning and the prediction phase [1].

Our aim to predict an overall probability of non-recurrence for an individual patient is in line with the suggestion of Anand et al. [2] to construct prognostic models with a directly useful prognostic estimate for a single patient. The idea of weighting the patients is similar to the one proposed by Ripley and Ripley [22]. In their experimental evaluation Ripley and Ripley used a heavy censored melanoma dataset. To incorporate censored data in the test sets, they fitted Kaplan–Meier survival curves and estimated their probabilities P_s of survival to the end of the observed follow-up time. These patients then entered the test set with both possible outcomes and probabilities P_s and $1 - P_s$, respectively. This schema was only used for testing, while for model induction either linear modeling or artificial neural nets were used to learn proportional odds and hazards, or Weibull and log-logistic survival. Interestingly, they have also used a neural network to directly predict the outcome, but for this they omitted the censored patients in the learning set. The work presented in this paper can thus be viewed as an extension to Ripley and Ripley's weighting schema from test to training dataset, thus enabling the use of general type machine learning algorithms that handle weighed data and induce classification models.

6. Conclusions

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 schema that allows us to include data records of non-recurrent patients with short follow-up times in the dataset for modeling.

The main contribution of the work described should be viewed as an enabling technology. Within our schema, any machine learning technique that induces

classification models from weighted examples can be used for building prognostic models from censored data. We exemplify this through two case studies of prostate cancer recurrence and show that a very simple and basic machine learning tool, the naive Bayes classifier, can stand beside a mature and often used statistical method of Cox proportional hazards model which was crafted specifically for survival analysis.

There are various other techniques for handling censored data that we mention in the section on related work. In comparison with these techniques, the advantages of the approach proposed in this paper are simplicity, straightforward integration with standard machine learning techniques, and a comparable performance with the well established statistical technique of Cox proportional hazards model.

The machine learning community has developed techniques that are more elaborate than the two used in our study, and that may be better at discovering nonlinearities and complex predictive variable combinations. With the recent introduction of medical and laboratory information systems, we believe that as the volume of clinical data grows both in the number of records and number of variables stored, machine learning tools may become increasingly important in mining censored data. In this respect, the schema proposed in this paper should further be tested on bigger datasets, where variable selection, combination and construction together with the interpretation of the resulting models may be crucial.

The authors strongly believe that, although tested only on prostate cancer recurrence data, the proposed methods are applicable to general survival analysis where the sole prediction of probability of event (and not its probability dependency on time) is of interest. Furthermore, the proposed weighting technique may be extended in a straightforward manner to predict the outcome at a given time interval, thus making it applicable beyond the scope considered in this paper.

Acknowledgements

This work was generously supported by the Slovene Ministry of Science and Technology and the Office of Information Technology at Baylor College of Medicine. The authors are grateful to Peter T. Scardino, M.D., of the Memorial Sloan Kettering Cancer Center for the sharing of his data.

References

[1] Anand SS, Hughes JG, Bell DA, Hamilton PW. Using censored neighbours in prognostication. In: Working Notes of the AIMDM-99 Workshop on Prognostic Models in Medicine: Artificial Intelligence and Decision Analytic Approaches, Aalborg, Denmark, June 1999, pp. 15–19.

[2] Anand SS, Smith AE, Hamilton PW, Anand JS, Hughes JG, Bartels PH. An evaluation of intelligent prognostic systems for colorectal cancer. Artif Intell Med 1999;15(2):193–214.

[3] Biganzoli E, Boracchi P, Mariani L, et al. Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. Stat Med 1998.

[4] Burke HB, Goodman PH, Rosen DB, et al. Artificial neural networks improve the accuracy of cancer survival prediction. Cancer 1997;97(4):857–62.
[5] Cestnik B, Bratko I. On estimating probabilities in tree pruning. In: Proc. European Working Session on Learning EWSL-91, 1991.
[6] Cox DR. Regression models and life-tables. J R Stat Soc B 1972;34:187–220.
[7] D'Amico AV, Moul J, Kattan MW. Prognostic factors and outcomes. In: Vogelzang NJ, Scardino PT, Shipley WU, Coffey DS, editors. Comprehensive Textbook of Genitourinary Oncology. Lippincott William & Wilkins, Baltimore, MD, 2000 (in press).
[8] Faraggi D, Simon R. A neural network model for survival data. Stat Med 1995;14(1):73–82.
[9] Hanley JA, McNeil BJ. The meaning and use of the area under receiver operating characteristic curve. Radiology 1982;143:29–36.
[10] Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. J Am Med Assoc 1982;247(18):2543–6.
[11] Kattan MW, Eastham JA, Stapleton AM, Wheeler TM, Scardino PT. A preoperative nomogram for disease recurrence following radical prostatectomy for prostate cancer. J Natl Cancer Inst 1998;90(10):766–71.
[12] Kattan MW, Hess KR, Beck JR. Experiments to determine whether recursive partitioning (cart) or an artificial neural network overcomes theoretical limitations of Cox proportional hazards regression. Comput Biomed Res 1998;31(5):363–73.
[13] Kattan MW, Ishida H, Scardino PT, Beck JR. Applying a neural network to prostate cancer survival data. In: Lavrac N, Keravnou E, Zupan B, editors. Intelligent Data Analysis in Medicine and Pharmacology. Boston: Kluwer, 1997:295–306.
[14] Kattan MW, Wheeler TM, Scardino PT. Postoperative nomogram for disease recurrence alter radical prostatectomy for prostate cancer. J Clin Oncol 1999;17(5):1499–507.
[15] Lavrac N, Keravnou E, Zupan B, editors. Intelligent Data Analysis in Medicine and Pharmacology. Boston: Kluwer, 1997.
[16] Le CT. Applied Survival Analysis. New York: Wiley, 1997.
[17] Lubsen J, Pool J, van der Does E. A practical device for the application of a diagnostic or prognostic function. Methods Inf Med 1978;17:127–9.
[18] Lucas PJF, Abu-Hanna A. Prognostic methods in medicine. Artif Intell Med 1999;15(2):105–19 (editorial).
[19] Michie D, Spiegelhalter DJ, Taylor CC, editors. Machine Learning, Neural and Statistical Classification. Chichester, UK: Ellis Horwood, 1994.
[20] Niblett T, Bratko I. Learning decision rules in noisy domains. In: Expert Systems 86 (ProcEWSL Brighton). Cambridge University Press, Cambridge, 1986, pp. 15–18.
[21] Quinlan R. Induction of decision trees. Mach Learn 1986;1(1):81–106.
[22] Ripley BD, Ripley RM. Neural networks as statistical methods in survival analysis. In: Dybowski R, Gant V, editors. Artificial Neural Networks: Prospects for Medicine. Landes Biosciences, 1998.
[23] Snow PB, Smith DS, Catalona WJ. Artificial neural networks in the diagnosis and prognosis of prostate cancer: a pilot study. J Urol 1994;152(5 Pt 2):1923–6.
[24] Venables WN, Ripley BD. Modern Applied Statistics with S-PLUS, 2nd edn. New York: Springer, 1997.
[25] Zupan B, Demšar J, Kattan MW, Beck JR, Bratko I. Machine learning for survival analysis: a case study on recurrence of prostate cancer. In: Horn W et al.. editors, AIMDM-99. Springer, 1999, pp. 346–355.