Naive Bayesian-Based Nomogram for Prediction of Prostate Cancer Recurrence

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

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

Abstract

This paper introduces a schema with naive-Bayesian classifier and patient weighting technique to develop a prostate cancer recurrence prediction model from patient data. We propose the graphical presentation of naive-Bayesian classifier with a nomogram, which can be used both for prediction or can provide means to data analysis. The method was used to construct a predictive model for prostate cancer recurrence; the results were favorable both in terms of interpretability and predictive accuracy.

1 Introduction

In medicine, it is often required to use purposely collected data in order to construct models that either serve for prediction and support in medical decision making, or are used solely for data analysis. Machine learning offers various methods for efficient construction of descriptive models from data. At its best, the modern modeling technique should offer both: accuracy of the developed model and transparency so that the decision maker knows *why* and *how* the model derives the decision. Machine learning methods may support both perspectives, and are as such increasingly used in developing fields of Intelligent Medical Data Analysis [4] and Medical Data Mining [9].

When modeling is used for the survival analysis, specific mechanisms for proper treatment and modeling of the data are required. Namely, 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*. Therefore, the modeling technique has to consider that for some patients the follow-up may end before the event occurs. This paper offers a solution to this problem by introducing a weighting technique that lowers the importance of patients with short follow-up time and for whom the event does not occur, i.e., the patients for which one knows the event did not (yet) occur but does not know whether it will occur in the future.

The paper considers the survival analysis problem for prostate cancer recurrence after prostatectomy. For this problem, the physicians are mainly interested if after the operation the cancer will recur, so that the survival analysis problem actually translates to the pure classification problem. This is in a sense different from classical survival analysis, which essentially estimates the probability of recurrence given the follow-up time, but as a consequence, this approach allows the use of broader range of classification-based machine learning techniques.

We propose the schema that uses naive Bayesian classifier to construct the model for prostate cancer recurrence prediction. Naive Bayes is one of the simplest but, surprisingly, one of the often most accurate machine learning predictors, especially for the medical domains [3]. In order to appropriately use it for derivation of prognostic models for prostate cancer recurrence, we couple it with the above mentioned patient data weighting technique. We present the naive Bayesian predictor as a nomogram, and show how this can be used for prediction as well as for data analysis. To evaluate the quality of naive Bayes, we compare its performance with Cox proportional hazards model, a standard statistical method for survival analysis.

2 Methods

The approach presented in this paper incorporates three techniques. Example weighting is used to appropriately consider survival data when building classification models. We use naive Bayesian technique to induce the model from data, and represent the derived model graphically using the nomogram. Besides the physician's interpretation and validation of the model, three different statistics are used to assess the model's performance.

2.1 Example weighting

Survival analysis normally estimates the hazard for an event to occur or the expected time before the event occurs ("survival time" of an individual). From machine learning point of view, this is more a kind of regression problem, which we can turn into a classification problem by discretizing the class value: setting a limit T_f divides the examples into two classes – to those *surviving* the period T_f and those *failing* before that time.

The main difficulty with this approach is handling the *censoring* which is specific for survival data. There can be examples which are known to be "alive" at time $T_0 < T_f$, after which the trace of them was lost or they have failed for an unrelated reason. Statistical methods for survival analysis are time-aware and can take such examples into account until T_0 and ignore them afterwards. Machine learning methods cannot handle "disappearing examples". Skipping them completely could, on the other hand, significantly change the statistical properties of example set and influence the results [2, 7].

To cope with this, we weight the examples with the weights depending on the follow-up time and the class of the example. Failing examples and examples which are not censored before the time T_f are confident and have a weight of 1. Examples with follow-up times shorter that T_f ("disappearing examples") are assigned a follow-up time dependent weight $W(T)$, which monotonically increases from $W(0) = 0$ to $W(T_f) = 1$. We use the null Martingale residual (NMR) [8, 2], which is computed from the example set itself and is, in terms of domain of prostate cancer prediction, proportional to the risk of deceasing given only the follow-up time (and not the values of attributes). 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 $1 - \text{NMR}$. 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.

2.2 Naive Bayesian nomograms

The naive Bayesian classifier estimates the probabilities of classes for the given example by using the formula

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

where $P(r_k)$ is the priori probability of class r_k and $P(r_k|v_i)$ is the conditional probability of class r_k if i-th attribute has the value v_i; both are estimated from the training set of examples. Since the formula (naively) assumes the independence of attributes, the computed probabilities can be larger than 1 and do not necessarily sum to 1. If the naive Bayesian classifier is to be used as a class probability predictor, the probabilities must be normalized.

To get an explainable form of the classifier, we turn the product into a sum, by computing $\ln P(r_k|V)$ instead of $P(r_k|V)$:

$$\begin{aligned}\ln P(r_k|V) = \ln P(r_k) \quad &+ \quad \textstyle\sum_{i=1}^{n} \ln \frac{P(r_k|v_i)}{P(r_k)} \\ = B(r_k) \quad &+ \quad \textstyle\sum_{i=1}^{n} w_{k,i,v_i}\end{aligned}$$

In this form, each class r_k has its base probability, abbreviated by $B(r_k)$, which can be increased or decreased by different attributes, abbreviated by w_{k,i,v_i}. The value of w_{k,i,v_i} is positive if the probability of r_k after knowing that i-th attribute's values is v_i is greater than priori probability of r_k, and negative if it is smaller. Therefore each term of the sum can be interpreted as the weighted vote of the i-th predictor for or against the class r_k.

Values of w_{k,i,v_i} can be plotted in a *nomogram* [5]. Nomogram consists of vertical bars, one bar for each attribute-class combination. Values w_{k,i,v_i} for all possible triples (k, i, v_i) are computed and plotted, with class k and attribute i selecting the bar, w_{k,i,v_i} is used as coordinate and v_i is printed as a label. If there are only two possible classes, the nomogram can be further simplified by using the two sides of the same bar for both classes, therefore having only one bar for each attribute.

Nomograms can also be used for predicting class or the probabilities of classes for a given example. The values of w_{k,i,v_i} for the chosen class and example's values of attributes must be summed, added to the base probability and, if probabilities and not just the most probable class are needed, exponentialized and normalized. Except for the summing and normalizing, the calculation can be done graphically, using a lookup graph tabulating the function $f_{r_k}(x) = e^{B(r_k)+x}$ for each r_k. An example of a nomogram is shown in Figure 1. Note that nomograms can be used to present the naive Bayesian classifier in general and not only when used in survival analysis.

2.3 Measuring the quality of models

To compare predictive accuracy of Cox proportional hazards model and naive Bayes, we used a standard technique of stratified 10-fold cross validation [6], which averages over 10 experiments, each time deriving and testing the model on different datasets. Given the patient's data, both models derived the probability of recurrence, where for Cox model the probability was estimated for the patients to recur within 5 years after the prostatectomy.

Three different quality measures are used. *Classification accuracy* (CA) is expressed in percent of patients in the test set that were classified correctly. A probability higher than 0.5 was considered as a prediction for a patient to recur. The examples are weighted, so that, if a surviving example with a short follow-up time is misclassified the error is smaller than in the case of misclassifying a surviving example with a long follow time.

Average probability assigned to the correct class (AP) averages the probability of recurrence assigned to the recurrent patients and probability of non-recurrence for non-recurrent patients in the test dataset. As for CA, the probabilities are weighted.

Another measure, *concordance index* (CI) [1], estimates the probability that, given two randomly drawn examples, the example which fails first is predicted a higher probability of failing. CI is computed from the testing data as the proportion of consistent example pairs over the set of usable example pairs. A pair is usable if the example with a shorter follow-up time fails (while the other example may fail later or not at all), and consistent if the example with a shorter follow-up time is assigned the higher probability of failing.

3 Experiments on prostate cancer recurrence data

The above methods were used to construct a predictive model from the prostate cancer recurrence data. We employed the naive Bayes algorithm from the ML* package, which can handle weighted examples and supports adding new statistics, such as concordance index which has not been included in ML* before.

3.1 Patients data

We have used the data from 967 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. The dataset used consists of the following routinely measured 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`). 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. Of 967 patients, 189 (19.5%) recurred.

3.2 Nomogram

The nomogram in Figure 1, derived from naive Bayesian classifier, shows the impact of individual attributes on probability of recurrence (upper labels) 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, 1 vote *for* non-recurrence. The lookup graph below helps converting the sums of impacts of attributes to probabilities, again using the upper values for computing the recurrence and lower for non-recurrence.

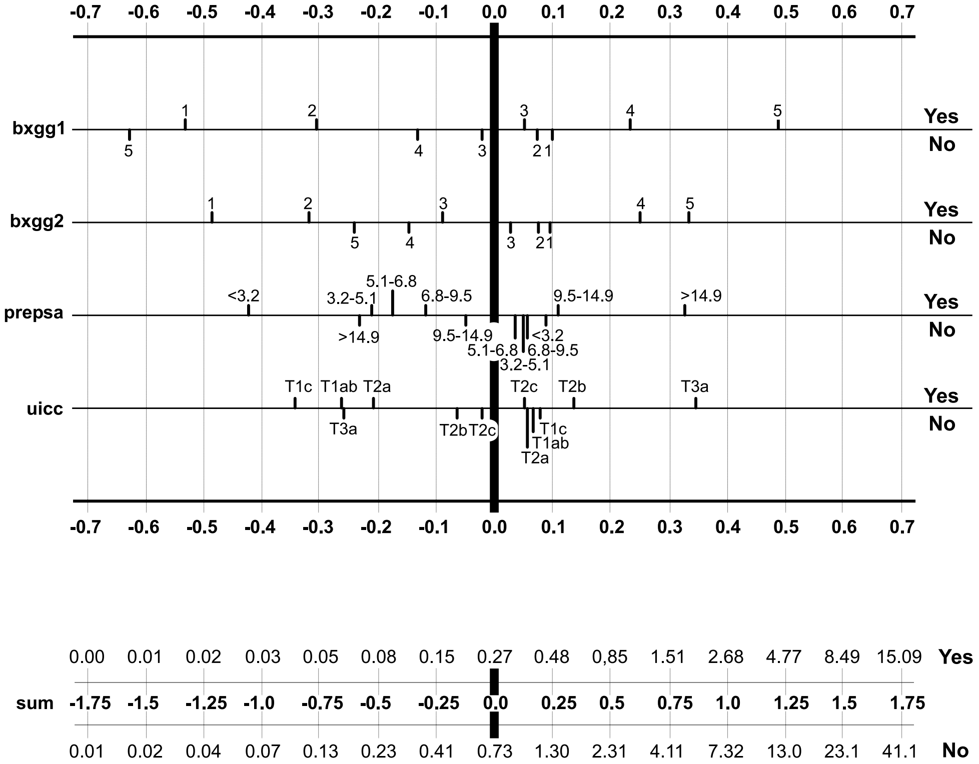

Figure 1: Nomogram showing the impacts of different values of attributes, derived from the probability estimates by naive Bayes

Figure 2 shows the use of the nomogram on a patient `(bxgg1=1, bxgg2=3, prepsa=11, uicc=T2a)`. The solid lines show the calculation for recurrence and the dotted for non-recurrence. Summing the values of impacts gives -0.54-0.09+0.11-0.21=-0.71 for recurrence and 0.1+0.03-0.05+0.05=0.13 against, from which we can already predict that the patient will not recur. Approximation by the lookup table gives "probabilities" of 0.06 for recurrence and 1 against, predicting the patient will not recur. Obtained probabilities should be normalized by multiplying by $(0.06 + 1)^{-1}$, which gives probabilities 0.057 for and 0.943 against recurrence.

3.3 Evaluation of the model

The performance of naive Bayesian classifier was compared with performance of the standard Cox model of proportional hazards. Results for different evaluation statistics are shown in Table 1. Overall, the performance of both modeling techniques is similar, with exception of weighted average probability assigned to the correct class, where Cox performance is significantly lower.

modeling technique	weighted classification accuracy	weighted avg. prob. assigned to correct class	concordance index
naive Bayes	75.5	0.706	0.759
Cox	75.8	0.625	0.756
default	73.1	0.606	0.5

Table 1: Results of performance evaluation

The "default" line shows the performance of a classifier which always predicts the majority class with its priori probability (in our case, it predicts non-recurrence with probability 0.731). An interesting point to be mentioned here is the fact that, judging only by classification accuracy, both methods seem to be

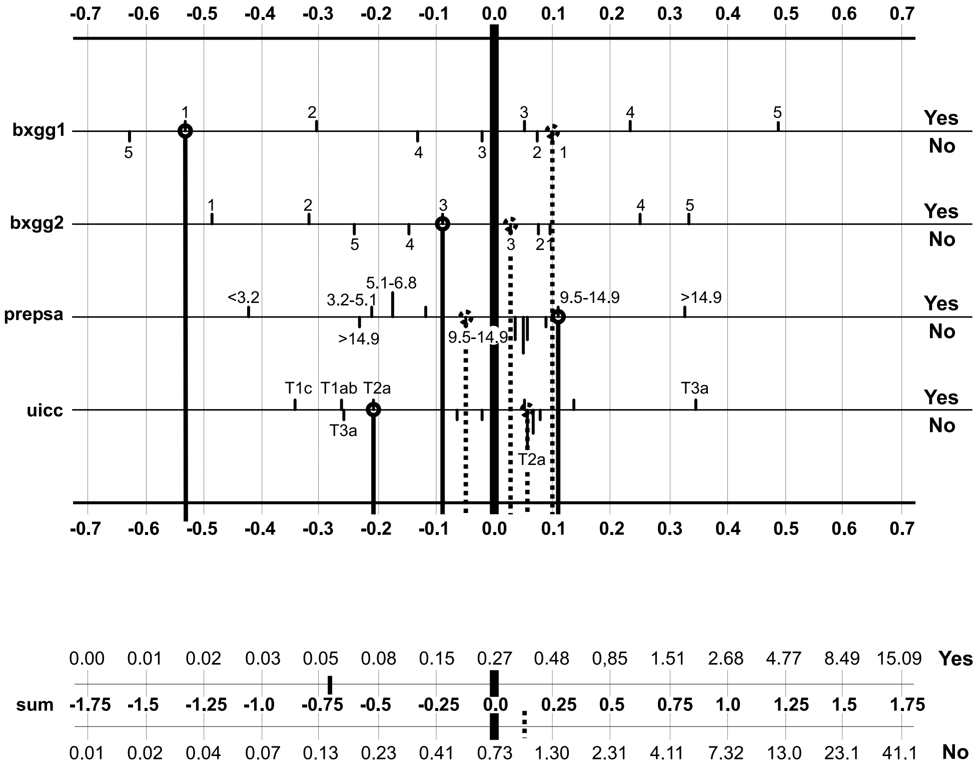

Figure 2: An example of computing probabilities of recurrence and non-recurrence from the nomogram. (Certain labels have been erased to increase the clarity of the figure.)

useless since they do not predict much better than the default classifier. The other two criteria, especially the concordance index, reveal a much better picture: although most of the patients are still classified as non-recurrent, both methods can successfully identify the patients with higher possibility of recurrence.

Another criteria for evaluation of a modeling technique is interpretability of the derived models and their concordance to the general knowledge. In case of naive Bayesian classifier nomogram reveals that the two Gleason scores are the most important factors for the decision as their values are most dispersed through the score line. In general, the impacts of attributes are as expected. The only anomaly occurs in `uicc`: `T1ab` would be expected to appear before `T1c` for recurrence and after `T1c` for non-recurrence. We believe that, while this anomaly is due to specific set of patients that were considered (sampling problem), it requires additional investigation on (preferably) different data set.

4 Conclusion

We have proposed the method to construct naive Bayesian-based nomogram from preoperative patient data for predicting the prostate cancer recurrence after prostatectomy. The experimental evaluation which used a patient data from The Methodist Hospital (Houston, TX) showed that the nomogram can be used both for data analysis and recurrence prediction. For the later, naive Bayes was favorably compared to Cox proportional hazards model.

The implications of this work may be two two-fold. First, the proposed schema can be used for general survival analysis problems when only the prediction of event, and not its time-dependent probability, is in question. Note that, besides naive Bayes, any classifier induction tool may be used that supports the inclusion of example weights. Second, the nomogram as a graphical device for presentation and utility can be used beyond survival analysis for any classification setup where naive Bayes is applicable.

Acknowledgment

Support for this work was provided by the Ministry of Science and Technology of Slovenia as a USA-Slovenia Cooperation in Science and Technology Grant, and by Developmental Support from Information Technology Program at Baylor College of Medicine.

References

[1] 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.

[2] 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.

[3] I. Kononenko, I. Bratko, and M. Kukar. Application of machine learning to medical diagnosis. In *Machine Learning and Data Mining: Methods and Applications*, pages 389–408. John Wiley & Sons, Chichester, 1998.

[4] N. Lavrač, E. Keravnou, and B. Zupan, editors. *Intelligent data analysis in medicine and pharmacology*, Boston, 1997. Kluwer.

[5] 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.

[6] D. Michie, D. J. Spiegelhalter, and C. C. Taylor, editors. *Machine learning, neural and statistical classification*. Ellis Horwood, 1994.

[7] 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.

[8] T. M. Therneau, P. M. Grambsch, and T. R. Fleming. Martingale-based residuals for survival models. *Biometrika*, 77(1):147–160, 1990.

[9] B. Zupan, N. Lavrač, and E. Keravnou. Data mining techniques and applications in medicine (editorial). *Artificial Intelligence in Medicine*, 15, 1999 (in press).