Uplift modeling with survival data

Szymon Jaroszewicz
National Institute of Telecommunications
Warsaw, Poland

Institute of Computer Science
Polish Academy of Sciences
Warsaw, Poland
s.jaroszewicz@ipipan.waw.pl

Piotr Rzepakowski
National Institute of Telecommunications
Warsaw, Poland
p.rzepakowski@itl.waw.pl

ABSTRACT

Uplift modeling is a subfield of machine learning which aims at predicting differences between class probabilities in treatment and control groups. This setting is frequently encountered in medical trials which are thus a natural application domain for the technique. Unfortunately, clinical trials usually involve survival data with censoring and most machine learning methods, including uplift modeling methods, do not directly allow for the use of such data. In this paper we demonstrate that, under reasonable assumptions, uplift modeling can be easily applied to survival data, while maintaining the correctness of decisions made by the model. This is in contrast to standard classification, where the use of censored training examples is more difficult. We test our approach on a publicly available clinical trial dataset and show that using an uplift model it is possible to obtain significant improvements in survival rates.

Categories and Subject Descriptors

I.2.6 [**Computing Methodologies**]: Artificial Intelligence-Learning

Keywords

uplift modeling, personalized medicine, machine learning, survival data, censoring

1. INTRODUCTION

Uplift modeling is a machine learning technique which aims at predicting the differences in class probabilities between two groups: the treatment group, subjected to some action, and the control group not subjected to the action (or subjected to an alternative action). This paper describes how to apply uplift modeling to survival data used in medicine and shows that a specific property of uplift modeling makes it applicable to survival data also in the presence of censoring.

ACM SIGKDD Workshop on Health Informatics (HI-KDD) New York City, August 2014

Clinical trials are a key tool of evidence based medicine. A group of patients (the treatment group) is subjected to a therapy and another group (the control group) is given placebo or an alternative treatment. Statistical tests are then used to decide whether the treatment under consideration brings real benefits. Personalized medicine requires an analysis at a finer level: the effectiveness of the treatment is assessed based on the characteristics of a specific patient. Uplift modeling is directly applicable in this case because it provides medical practitioners with models which are able to predict the impact of a therapy for a given individual.

Clinical trials typically produce survival data. The key target variable is the *survival time* of each patient. Another important feature of such data is *censoring*: survival time of some patients (for example those who haven't been in the study for a sufficiently long period of time) is not known exactly, all that is known is that it is longer than the observed value.

In contrast, classification data used in machine learning contains a single, discrete class variable taking a finite number of values. Censoring is typically not allowed. Since machine learning methods have been extremely successful in predicting discrete outcomes, there has been significant interest in applying them to medical problems. To use those methods with survival data one typically picks a threshold θ and treats all patients who survived at least that amount of time as successes and the remaining ones as failures of the therapy. A binary classification model is then constructed and used to predict for which patients the therapy will be effective.

There are, however, two problems with this approach: the first is that class probabilities, predicted by classification models, are not necessarily the quantities of interest in clinical trials, the second is the presence of censoring.

The first problem stems from the fact that classification models do not take the control group into account. A medical treatment is worthwhile only if, for a given patient it works better than placebo. For example, if the patient recovered after the treatment but would have recovered regardless of it, then we have subjected her to unnecessary risk of side effects and incurred unnecessary costs. We believe that uplift modeling, which explicitly models the difference between class probabilities in the treatment and control groups is the solution to this problem.

In this paper we will show that, additionally, uplift modeling is, to a large degree, immune to the second problem, and can be applied directly even in the presence of censoring.

2. RELATED WORK

We will begin by describing the literature on applying machine learning methods to survival data. Afterwards, we will give a brief overview of uplift modeling and related techniques.

Most machine learning methods (and to the best of our knowledge all current uplift modeling methods) do not directly allow for the use of censored data. Exceptions are survival trees [1] or regularized Cox regression [19]. Another set of approaches is based on transforming survival data into standard classification tasks to which unmodified classifiers can be applied. The problem here is that the true class is unknown for censored cases and needs to be estimated. An overview of early approaches, such as the removal of all censored cases, can be found in [13], where their deficiencies, such as introduction of additional bias, are discussed. A newer approach, which weighs the likelihood of censored data points with an estimate of survival probability can be found in [21]. None of those transformations are perfect as they require, for example, good estimates of all censored survival times.

As mentioned above, even if such transformations were successful, classification models do not take into account the control group which makes them unsuitable for most medical problems, uplift modeling being more appropriate. Moreover, we will show in the next section that transformations of censored data are not necessary when this technique is applied.

We will now briefly describe uplift modeling algorithms and related techniques available in the literature.

The simplest uplift model consists of two classifiers, one built on the treatment training set, another on the control. The probabilities predicted by the control classifier are then subtracted from the probabilities predicted by the treatment classifier to give an estimate of the real effect of the treatment. The double classifier model, while intuitive, has the drawback that both models may focus on predicting the class variable instead of modeling the (usually weaker) difference between class distributions in the two groups. An illustrative example can be found in [11].

To alleviate this problem, several learning algorithms designed specifically for uplift modeling have been proposed. Many of them build decision trees using splitting criteria designed to select tests which maximize differences between class probabilities in the treatment and control groups. In [11] the authors present decision trees based on statistical significance of differences between treatment and control success probabilities. Uplift decision trees which are based on information theoretical splitting criteria and include a tree pruning step have been proposed in [16, 17].

A generalization of the nearest neighbor method to uplift modeling has been presented by Larsen in [9]. The method proved highly successful in our experiments, and will be discussed in more detail in Section 4. Linear uplift models have been presented in [6] and [23]. The first paper proposes a class variable transformation which allows for application of an arbitrary probabilistic classifier (logistic regression was used in the paper) to the uplift modeling problem. The second paper adapts linear Support Vector Machines to take into account the presence of a control group.

More detailed overviews of uplift modeling approaches can be found in [11], [18] or [17].

In the medical and statistical literature regression methods have been developed, under various names such as *nested mean models*, to address similar problems [3, 22]. Similar techniques have been considered in the context of G-estimation [14, 15]. Most of those methods, however are variants of the double classifier approach (when interactions of the treatment indicator with all variables are included in the model) or assume that the difference in class probabilities is independent of patient's characteristics (the treatment indicator is simply included as an additional regression variable). Uplift models, in contrast, provide a method to directly predict the difference between the treatment and control success probabilities on the level of individuals using a *single* model.

One of the contributions of this paper is to show that uplift modeling is a viable option for selecting the right treatment based on the characteristics of an individual patient.

3. APPLYING UPLIFT MODELING TO SURVIVAL DATA

We will now present the main result of the paper, which shows how uplift models can be applied to survival data commonly occurring in medicine. First, let us introduce some notation.

3.1 Notation

Each object (e.g. patient) is described by a feature vector x which comes from some sample space $\mathcal{X}$. Each object has an associated binary outcome $y \in \{0, 1\}$. The event $y = 1$ is assumed to be the successful outcome, such as recovery from a disease or survival for a specified amount of time. A classification model M is a function

$$M : \mathcal{X} \to [0, 1]$$

whose value, $M(x)$, is interpreted as the probability that x belongs to the positive class. When firm decisions need to be made, some threshold η is chosen, for example $\frac{1}{2}$, and cases for which $M(x) \geq \eta$ are classified as positive.

In a randomized trial, each patient is assigned at random to one of two groups: the *treatment group*, subject to the therapy under consideration, or to the *control group* administered placebo or an alternative treatment. Let P^t denote the probabilities in the treatment group and P^c the probabilities in the control group. An *uplift model* is a function

$$M : \mathcal{X} \to [-1, 1]$$

whose value, $M(x)$, is interpreted as the difference between success probabilities in the treatment and control groups $P^t(y = 1|x) - P^c(y = 1|x)$. If $M(x) > 0$ than the model predicts the treatment to be beneficial for a patient described by feature vector x.

Note that uplift modeling requires two training datasets, D^t and D^c containing cases, respectively, from the treatment and control groups. The problem which makes uplift modeling difficult, known as the *fundamental problem of causal inference* [5] is that, for each training case we know only the outcome after treatment or after placebo, never both. Therefore we cannot determine whether the treatment was really successful for a given individual.

We now briefly discuss survival data and censoring. More thorough discussion can be found e.g. in [10]. Let T^* denote a random variable representing the survival time for individuals in a given population, that is the time until some

event takes place, such as patient death or disease recurrence. Sometimes the survival time is longer than the time the patient has been under observation and is thus unknown. All we know is that it is later than the last time the patient was seen. Such a situation is called *censoring* (right censoring to be precise). There are various reasons for censoring, the patient might have entered late in the study or dropped out early [10].

Survival data usually involves two attributes: the observed survival time T and a binary censoring status variable $\delta \in \{0, 1\}$ indicating whether the event has been observed (typically indicated by $\delta = 1$). If the event has not been observed ($\delta = 0$) the true survival time T^* has been *censored*, i.e., we only know that $T^* > T$ where T is the last observed time.

Let C be the right censoring time, i.e. the last time the patient has been observed in the study. The *observed* survival time is the minimum of the true survival time and the censoring time

$$T = \min(T^*, C). \tag{1}$$

In the most general case, both T^* and C jointly depend on x. Usually, however, stronger assumptions need to be made. Below we describe how, under an appropriate censoring model, uplift modeling is directly applicable to survival data.

3.2 Survival data and uplift modeling

We now present the proposed transformation of survival data into classification data and show that it is indeed appropriate for uplift modeling.

We are going to build a model which will predict how the treatment influences the probability that a given individual, described by a feature vector x, survives at least θ, where θ is some threshold of interest to clinicians. Note that we are interested in the event $T^* \geq \theta$, but the true survival time is not available to us due to censoring. Let us introduce the following class variable:

$$Y = \begin{cases} 1, & \text{if } T \geq \theta, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Note, that Y is based on the observed time given in Equation 1, not the inaccessible survival time T^*. We will now demonstrate that, under reasonable assumptions, an uplift model trained on the class variable given in Equation 2 does in fact make correct predictions about the impact of the treatment on the true survival time T^* for an individual described by a feature vector x. Let us begin with an assumption about censoring.

ASSUMPTION 1. *The survival time T^* and the censoring time C are conditionally independent given x.*

This assumption is frequently made in survival analysis [10]. We now have

$$\begin{aligned} 1-P(Y=1|x) &= P(T < \theta|x) = P(T^* < \theta \vee C < \theta|x) \\ &= P(T^* < \theta|x) + P(T^* \geq \theta \wedge C < \theta|x) \end{aligned}$$

which, using Assumption 1

$$\begin{aligned} &= P(T^* < \theta|x) + P(T^* \geq \theta|x)P(C < \theta|x) \\ &= P(T^* < \theta|x) + (1 - P(T^* < \theta|x))P(C < \theta|x) \\ &= P(T^* < \theta|x) - P(T^* < \theta|x)P(C < \theta|x) + P(C < \theta|x) \\ &= P(T^* < \theta|x)[1 - P(C < \theta|x)] + P(C < \theta|x). \end{aligned} \tag{3}$$

We have thus expressed the conditional probability of the class variable Y in terms of survival and censoring probabilities for a given individual.

Note, however, that estimating $P(Y = 1|x)$ does not allow us to draw conclusions about $P(T^* \geq \theta|x)$ since the term $P(C < \theta|x)$ may vary between individuals. Unfortunately, estimation of $P(C < \theta|x)$ is not necessarily easy. We will now introduce an additional assumption which will make the use of uplift models possible, without the need to estimate the censoring time distribution.

ASSUMPTION 2. *The censoring times C are independent of the treatment group assignment.*

That is, the censoring time distributions are identical in the treatment and control groups: $P^t(C < \theta|x) = P^c(C < \theta|x) = P(C < \theta|x)$. To see how this assumption becomes useful in the context of uplift modeling note that the difference in conditional probabilities of the class variable Y given in Equation 2 is

$$\begin{aligned} &P^t(Y=1|x) - P^c(Y=1|x) \\ &\quad= P^c(T^* < \theta|x)[1 - P^c(C < \theta|x)] + P^c(C < \theta|x) \\ &\quad\quad -P^t(T^* < \theta|x)[1 - P^t(C < \theta|x)] - P^t(C < \theta|x) \\ &\quad= P^c(T^* < \theta|x)[1 - P(C < \theta|x)] + P(C < \theta|x) \\ &\quad\quad -P^t(T^* < \theta|x)[1 - P(C < \theta|x)] - P(C < \theta|x) \\ &\quad= [P^t(T^* \geq \theta|x) - P^c(T^* \geq \theta|x)]P(C \geq \theta|x). \end{aligned}$$

Now, the treatment is beneficial (with respect to surviving at least θ) if $P^t(T^* \geq \theta|x) - P^c(T^* \geq \theta|x) > 0$. Notice that, since $P(C \geq \theta|x)$ is non-negative, the sign of this expression is identical to the sign of $P^t(Y = 1|x) - P^c(Y = 1|x)$, i.e., the difference in probabilities of the class variable Y.

Thus, under Assumptions 1 and 2, an uplift model is capable of making *correct* recommendations even though the predicted value of the difference in probabilities may be incorrect. Note, that the dependence of the censoring time C on the features x *is* allowed.

Assumption 2 may seem restrictive; it may be violated e.g. when one of the therapies imposes more burden on the patient, who is then more likely to quit. However, the assumption remains valid in many important cases. One example is the intention to treat analysis, provided that all patients are followed even if they stop receiving the treatment. Note, that for this type of analysis not following some patients leads to biased results [8], so care is usually taken to follow all participants. Of course, the assumption remains true if the burden of both treatments is similar. Specifically, the assumption is true in the frequently occurring case when the only source of censoring is the date of entry into the study, known as type III censoring [10].

Notice that the class transformation given in Equation 2 will not work in the case of classification. Suppose we pick some thresholds θ and $\eta > 0$, and want to classify points with $P(T^* \geq \theta|x) \geq \eta$ as positive. It is then easy to see

from (3) that

$$P(T^* \geq \theta|x) \geq \eta \Leftrightarrow \frac{P(Y=1|x)}{P(C \geq \theta|x)} \geq \eta,$$

so deciding whether the probability of interest is greater or equal to η requires an estimate of $P(C \geq \theta|x)$.

4. EXPERIMENTAL EVALUATION

In this section we are going to apply the transformation of survival data for uplift modeling described in the previous section to a real clinical trial dataset. We begin by describing the dataset used in the experiments and the uplift models we used, then discuss the methods of evaluating uplift models, and finally present the results of our experiments.

4.1 The `colon` dataset

The `colon` dataset available in the `survival` package of the `R` system comes from the clinical trial of an adjuvant chemotherapy for colon cancer. There are two types of treatment: levamisole, a low-toxicity compound, and levamisole combined with 5-FU (Fluorouracil), a moderately toxic chemotherapy agent. A control group is also included, which received no treatment and was subject to observation only. For each patient, two types of events are recorded: recurrence of the disease and death, giving us two possible modeling targets. We thus built separate uplift models for lack of recurrence and patient survival.

Since most uplift models allow for only one type of treatment we created separate datasets for treatment with levamisole only and levamisole combined with 5-FU. The combined therapy was more effective overall and uplift models did not produce any improvement over administering the combined treatment to all patients. In the remaining part of the paper we thus consider only the levamisole and observation arms of the study.

Another issue is how to select the threshold θ in the conversion described in Equation 2. We began by using the median of observed (censored) times of disease recurrence and patient survival in order to obtain a balanced class distribution. To test with a larger number of censored cases we then repeated the experiments with thresholds set to the third quartile of the observed times.

To build uplift models we used 10 predictor variables: patient's sex and age, obstruction of the colon by tumor, perforation of the colon, adherence to nearby organs, number of lymph nodes with detectable cancer, three levels of differentiation of the tumor, the extent of local spread (submucosa, muscle, serosa, contiguous structures), time from surgery to registration, and an indicator of more than 4 positive lymph nodes.

4.2 Methods of evaluating uplift classifiers

In this paper we use two ways of evaluating uplift models, one based on so called uplift curves, the other based on computing simulated survival curves. We describe them in turn.

Similarly to two training datasets, uplift modeling requires two test sets, one containing treatment, the other control cases. The main problem of evaluating uplift classifiers is, that for each test case we only know one of the responses, either after the action was taken, or when no action was taken, never both. Methods of evaluating uplift classifiers are thus based on an assumption that cases which are similarly scored by a model do indeed behave similarly. In other words, we assume that the k percent of the treatment test set which the uplift model scored highest is comparable to the k percent of highest scoring cases in the control test set; gains on the top k percent of cases in both datasets can thus be subtracted from each other to obtain a meaningful estimate of the difference in success probabilities.

In practice it is easier to visualize the performance using an *uplift curve*. The curve is obtained by repeating the above procedure for several values of k and plotting the obtained values on a chart. In medical context, the interpretation of an uplift curve is as follows: on the x axis we pick the percentage of the population, selected according to the model, which should receive the treatment. The corresponding value on the y axis then gives an estimate of the gain in success rate achieved due to the treatment with respect to the case when no patients receive the treatment. The gain is expressed as percentage of the *total* population. Example uplift curves will be shown in Section 4, more details on testing uplift models and on uplift curves can be found in [12, 16, 4].

To summarize the curve using a single numerical value we compute Areas Under the Uplift Curves (AUUC). Notice that, uplift curves can achieve negative values (the results of an action can be worse than doing nothing), and the area under an uplift curve can also be negative.

Uplift curves are a useful diagnostic tool, but they do not take into account survival times and censoring. Moreover, performing statistical tests on such curves is not easy. To bring our experiments closer to medical methodology we used a cross-validation based approach to survival analysis described in [20]. Of course, since in the uplift setting we have two training sets, cross-validation is performed on both of them in parallel. For each train-test split, an uplift model is build on the two training sets and used to predict which cases in the treatment and control test sets should be subjected to the therapy. Cases in the treatment test set selected for treatment by the model as well as cases in the control test set not selected by the model (i.e. cases where model recommendation agrees with the true action taken) are then combined over all cross-validation folds. The resulting data is used to draw a survival curve which can be viewed as a simulated application of an uplift model to decide whether the treatment should have been administered to each patient. More details can be found in [20].

4.3 Uplift models used

We have tested several uplift models: a double logistic regression model, a single logistic model with class variable transformation described in [6], uplift decision trees from [17]. Finally, the best model turned out to be the uplift k nearest neighbor method proposed in [9]. The method is based on selecting, for the point x for which prediction is made, k nearest neighbors from both treatment and control training datasets. The improvement in success rate with respect to the control is then estimated based on those points. The influence of each selected point on the final decision is weighted by the inverse of its distance from x. The best results were obtained for $k = 1$, so that value is used in all subsequent experiments.

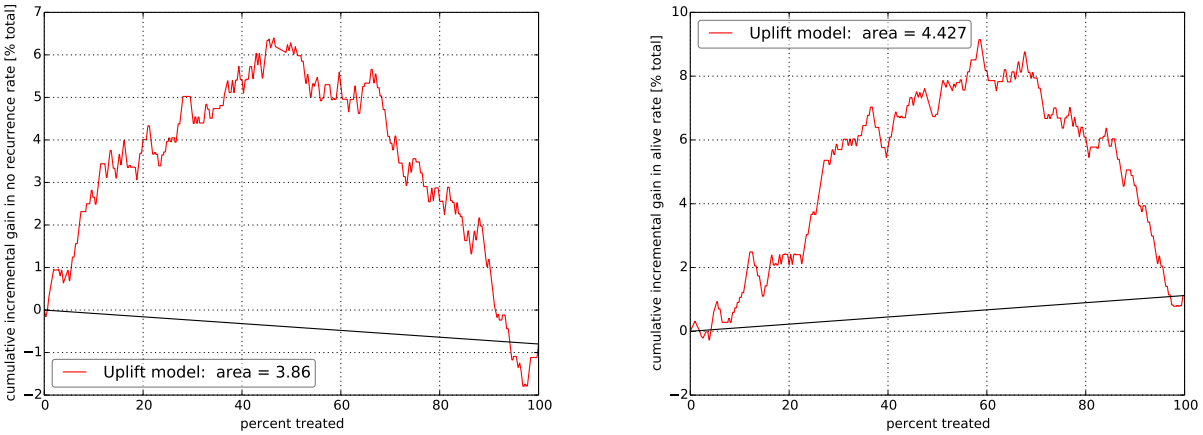

Figure 1: Averaged uplift curves of five fold cross-validated uplift models, built at the median of observed times for each patient. The left panel shows an uplift curve for lack of recurrence of the disease, the right panel for patient survival.

4.4 Results of experiments

We are now ready to present the actual experimental results. Figure 1 shows the uplift curves for lack of recurrence and patient survival targets. The curves were obtained using five fold cross-validation. The time threshold θ was picked to be the median censored time of the respective event, i.e. 1052 days for recurrence and 1875 days for death. It can be seen that by administering levamisole to only half of the patients one can expect 6% more patients to remain recurrence free at the selected time threshold with respect to no patient receiving levamisole. Note that the treatment is, overall, not effective, so an almost identical gain can be obtained with respect to administering levamisole to all patients. The diagonal line in the chart denotes random assignment of patients to the treatment group. The corresponding improvement for patient survival is even larger and amounts to about 8% of the population.

The results presented in the figure are encouraging, but they are expressed in terms of an artificial class variable, not the true survival time, so any conclusions have to be taken with caution. To obtain more credible evaluation we use simulated survival curves obtained using the procedure from [20] described above. The curves are shown in Figure 2. Again, five fold cross-validation was used.

It can be seen that the curves for both levamisole treatment and observation groups are almost identical, indicating that the treatment is not effective overall. Using an uplift model to select patients for the treatment one can, however, obtain much better survival rates and lower recurrence. The dashed vertical lines denote the cutoff thresholds θ. It can be seen that even though the model was built to maximize survival at that specific time point, the black curves, corresponding the uplift models, dominate over the whole range of survival times, with the difference in survival rates increasing with time.

To obtain further verification of the fact that the uplift model is indeed better than using either levamisole or no treatment for all patients we performed statistical tests for survival rates. Survival curves are typically compared using the log-rank test [10], here however we are interested in a specific time point on the curve: the one corresponding to the threshold θ used to build the model.

An overview of statistical tests for comparing survival curves at a specific time point θ can be found in [7]. A 'naive' test is based on the statistic

$$\chi_a^2 = \frac{(\hat{S}_1(\theta) - \hat{S}_2(\theta))^2}{\hat{S}_1(\theta)^2\hat{\sigma}_1(\theta)^2 - \hat{S}_2(\theta)^2\hat{\sigma}_2(\theta)^2},$$

where $\hat{S}_i(\theta)$ are the Kaplan-Meier estimators of the compared survival functions at θ and $\hat{\sigma}_i(\theta)$ are the estimates of the respective standard deviations obtained using the Greenwood's formula [10]. The statistic asymptotically follows the chi-squared distribution with one degree of freedom. In [7], the authors demonstrated that the test is overly optimistic and suggested several alternatives based on transformations of the survival curves. Following their results we have chosen the log-transform which yields the test statistic

$$\chi_b^2 = \frac{(\log \hat{S}_1(\theta) - \log \hat{S}_2(\theta))^2}{\hat{\sigma}_1(\theta)^2 + \hat{\sigma}_2(\theta)^2},$$

which asymptotically also follows the chi-squared distribution with one degree of freedom.

Since we want the treatment based on model's selection to be an improvement over both: treating all patients and treating none of them, we performed separate tests to compare the uplift model survival curve with the curves computed on the treatment and control datasets. Table 1 shows the test results for the curves presented in Figure 2. It can be seen that the results are highly significant and using the uplift model to select patients for treatment does indeed bring improvement over indiscriminate administration of the treatment or subjecting all patients to observation only.

It can be seen that the log-transform based test is indeed more conservative, but the results remain significant.

The results presented so far are encouraging, but they do not convincingly show that our method does indeed work in the presence of censoring. The reason is that, respectively, only 15 and 13 observations are censored (that is censoring occurred *before* the threshold θ) for patient survival and disease recurrence data. The total number of data records

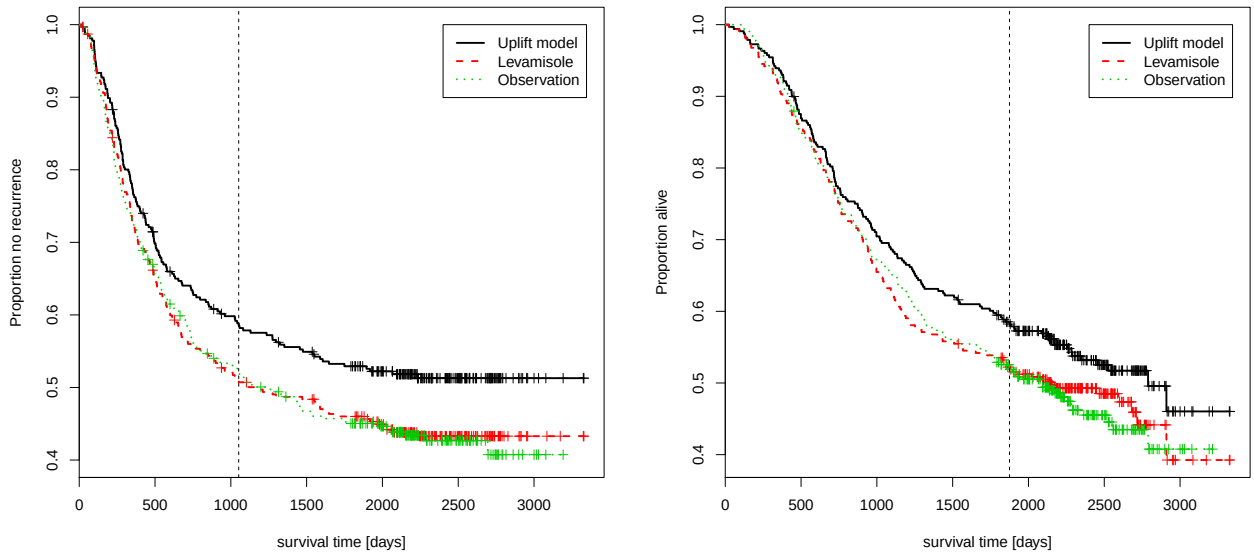

Figure 2: Survival curves for the observation group, the group treated with levamisole and a simulated application of an uplift model to select patients for treatment. The dashed vertical line shows the time point θ used to build the model, chosen to be the median of observed times for each patient. The left panel shows curves for lack of recurrence of the disease, the right panel for patient survival.

Table 1: Test results for difference in survival rates between uplift model's selection and the treatment and control groups at specific points in time. Low amount of censoring.

test type	uplift vs. treatment χ^2 stat.	uplift vs. treatment p-value	uplift vs. control χ^2 stat.	uplift vs. control p-value
recurrence free, $\theta = 1052$ days				
naive	14.486	$1.412 \cdot 10^{-4}$	8.602	$3.357 \cdot 10^{-3}$
log-trans	12.533	$3.999 \cdot 10^{-4}$	7.691	$5.550 \cdot 10^{-3}$
patient survival, $\theta = 1875$ days				
naive	8.184	$4.226 \cdot 10^{-3}$	8.257	$4.059 \cdot 10^{-3}$
log-trans	7.334	$6.767 \cdot 10^{-3}$	7.401	$6.519 \cdot 10^{-3}$

is 625. To increase the number of censored observations, we moved the survival threshold θ to the third quartile of observed times for all patients, i.e., 2324 days for patient survival and 2227 days for disease recurrence. The number of censored observations is, respectively, 149 (23.8%) and 123 (19.7%).

The resulting uplift curves are shown in Figure 3, the survival curves in Figure 4 and statistical tests results are given in Table 2. It can be seen that the improvement for disease recurrence data has become even more pronounced, despite higher censoring. This is likely due to the fact that the effects of the therapy become stronger with time, making it easier for the model to detect which patients respond to the therapy. This effect seems to offset the higher amount of cases lost due to censoring. Statistical tests confirm this visual observation with p-values in the range of 10^{-6}.

In case of patient survival, the results got somewhat worse, probably due to increased censoring. The uplift model's survival still dominates both treatment and control survival curves, but to a lesser degree. Tests comparing model's selection with the control still show significant improvement, but tests comparing with levamisole treatment have lost significance. Note, however, that this is likely due to the fact that the threshold θ, chosen at the third quartile of censored survival times, corresponds to a local bump in treatment effectiveness.

5. CONCLUSIONS AND FUTURE RESEARCH

We have presented a method for converting survival data obtained in clinical trials into classification data to which uplift modeling can be applied. We have demonstrated that, under reasonable assumptions, the predictions of an uplift model trained on such data remain correct even though the predicted probabilities need not be correct in the presence of censoring. We have applied the approach to data from a clinical trial of colon cancer treatment and shown that using an uplift model to select patients for treatment does indeed improve the outcome of the therapy in terms of both patient survival and disease recurrence.

An important direction of future research is medical verification of the results. The authors suspect that applying the treatment only to some of the patients gives better results due to side effects of levamisole which can be quite severe. For example, up to 5 percent of patients develop agranulocytosis [2] which negatively affects their immune system. Those conclusions, however, require further verification.

Another issue is choosing the right threshold θ. Of course, the value may be determined a-priori based on medical con-

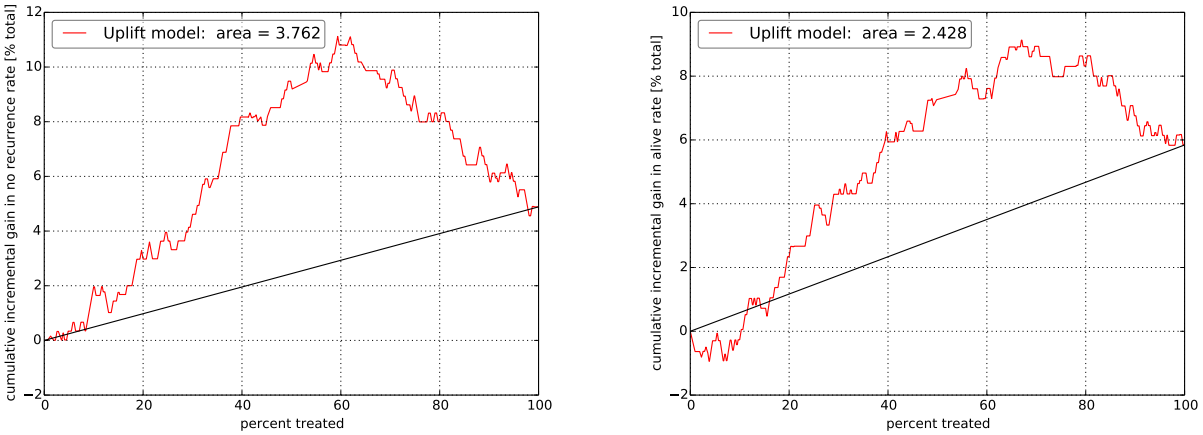

Figure 3: Averaged uplift curves of five fold cross-validated uplift models, built at the third quartile of observed times for each patient. The left panel shows an uplift curve for lack of recurrence of the disease, the right panel for patient survival.

Table 2: Test results for difference in survival rates between uplift model's selection and treatment and control groups at specific points in time. High amount of censoring.

test type	uplift vs. treatment χ^2 stat.	uplift vs. treatment p-value	uplift vs. control χ^2 stat.	uplift vs. control p-value
recurrence free, $\theta = 2227$ days				
naive	22.514	$2.086 \cdot 10^{-6}$	25.450	$4.540 \cdot 10^{-7}$
log-trans.	18.841	$1.420 \cdot 10^{-5}$	21.094	$4.373 \cdot 10^{-6}$
patient survival, $\theta = 2324$ days				
naive	1.605	$2.051 \cdot 10^{-1}$	8.421	$3.709 \cdot 10^{-3}$
log-trans.	1.527	$2.166 \cdot 10^{-1}$	7.504	$6.156 \cdot 10^{-3}$

cerns. It is however possible that individuals with high survival rate at time θ_1 will also have high survival rate at time θ_2. In such a case, the threshold should be chosen to maximize model performance. The therapeutic effect often increases with time (as can be seen in our experiments) making modeling easier, but so does censoring, which in turn makes it more difficult. Finding guidelines for picking the threshold is important for practical applications.

6. ACKNOWLEDGMENTS

This work was supported by Research Grant no. N N516 414938 of the Polish Ministry of Science and Higher Education (Ministerstwo Nauki i Szkolnictwa Wyższego) from research funds for the period 2010–2014.

7. REFERENCES

[1] I. Bou-Hamad, D. Larocque, and H. Ben-Ameur. A review of survival trees. *Statistics Surveys*, 5:44–71, 2011.

[2] DEA. Levamisole. `http://www.deadiversion.usdoj.gov/drug_chem_info/levamisole.pdf`, Apr. 2013.

[3] E. Goetghebeur and K. Lapp. The effect of treatment compliance in a placebo-controlled trial: Regression with unpaired data. *Applied Statistics*, 46(3):351–364, 1997.

[4] B. Hansotia and B. Rukstales. Incremental value modeling. *Journal of Interactive Marketing*, 16(3):35–46, 2002.

[5] P. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, Dec. 1986.

[6] M. Jaśkowski and S. Jaroszewicz. Uplift modeling for clinical trial data. In *ICML 2012 Workshop on Machine Learning for Clinical Data Analysis*, Edinburgh, Scotland, June 2012.

[7] J. Klein, B. Logan, M. Harhoff, and P. Andersen. Analyzing survival curves at a fixed point in time. *Statistics in Medicine*, 26(24):4505–4519, Oct. 2007.

[8] J. Lachin. Statistical considerations in the intent-to-treat principle. *Controlled Clinical Trials*, 21(3):167–189, 2000.

[9] K. Larsen. Net lift models: Optimizing the impact of your marketing. In *Predictive Analytics World*, 2011. workshop presentation.

[10] E. T. Lee and J. W. Wang. *Statistical Methods for Survival Data Analysis*. John Wiley & Sons, Hoboken, New Jersey, 2003.

[11] N. Radcliffe and P. Surry. Real-world uplift modelling with significance-based uplift trees. Portrait Technical Report TR-2011-1, Stochastic Solutions, 2011.

[12] N. J. Radcliffe. Using control groups to target on predicted lift: Building and assessing uplift models. *Direct Marketing Journal, Direct Marketing Association Analytics Council*, 1:14–21, 2007.

[13] B. Ripley and R. Ripley. Neural networks as statistical methods in survival analysis. In *Clinical applications of artificial neural networks*, pages 237–255. Cambridge University Press, 2001.

[14] J. Robins. Correcting for non-compliance in randomized trials using structural nested mean

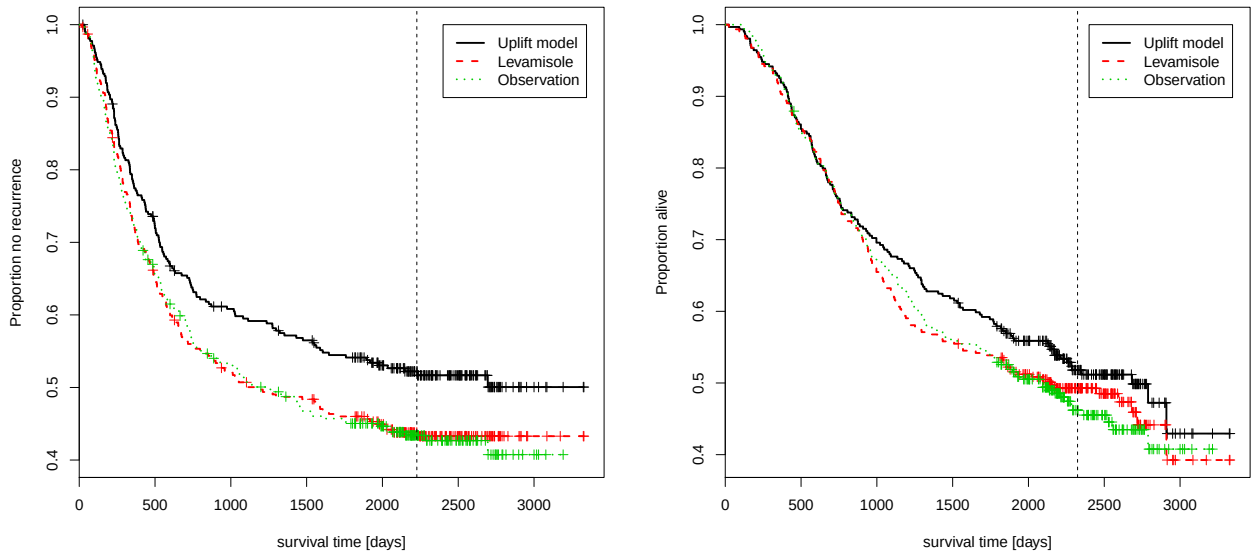

Figure 4: Survival curves for the observation group, the group treated with levamisole and a simulated application of an uplift model to select patients for treatment. The dashed vertical line shows the time point θ used to build the model, chosen to be the third quartile of observed times for each patient. The left panel shows curves for lack of recurrence of the disease, the right panel for patient survival.

models. *Communications in Statistics - Theory and Methods*, 23(8):2379–2412, 1994.

[15] J. Robins and A. Rotnitzky. Estimation of treatment effects in randomised trials with non-compliance and a dichotomous outcome using structural mean models. *Biometrika*, 91(4):763–783, 2004.

[16] P. Rzepakowski and S. Jaroszewicz. Decision trees for uplift modeling. In *Proc. of the 10th IEEE International Conference on Data Mining (ICDM)*, pages 441–450, Sydney, Australia, Dec. 2010.

[17] P. Rzepakowski and S. Jaroszewicz. Decision trees for uplift modeling with single and multiple treatments. *Knowledge and Information Systems*, 32:303–327, Aug. 2012.

[18] P. Rzepakowski and S. Jaroszewicz. Uplift modeling in direct marketing. *Journal of Telecommunications and Information Technology*, 2:43–50, 2012.

[19] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1–13, 2011.

[20] R. M. Simon, J. Subramanian, M.-C. Li, and S. Menezes. Using cross-validation to evaluate predictive accuracy of survival risk classifiers based on high-dimensional data. *Briefings in Bioinformatics*, 12(3):203–214, May 2011.

[21] I. Stajduhar and B. Dalbelo-Basić. Uncensoring censored data for machine learning: A likelihood-based approach. *Expert Systems with Applications*, 39(8):7226–7234, 2012.

[22] S. Vansteelandt and E. Goetghebeur. Causal inference with generalized structural mean models. *Journal of the Royal Statistical Society B*, 65(4):817–835, 2003.

[23] Ł. Zaniewicz and S. Jaroszewicz. Support vector machines for uplift modeling. In *Proc. of The First IEEE ICDM Workshop on Causal Discovery (CD 2013) at the 12th International Conference on Data Mining (ICDM)*, pages 131–138, Dec. 2013.