Uplift modeling for clinical trial data

Maciej Jaśkowski MACIEJ.JASKOWSKI@GMAIL.COM
Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland

Szymon Jaroszewicz S.JAROSZEWICZ@ITL.WAW.PL
National Institute of Telecommunications, Warsaw, Poland
Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland

Abstract

Traditional classification methods predict the class probability distribution conditional on a set of predictor variables. Uplift modeling, in contrast, tries to predict the *difference* between class probabilities in the treatment group (on which some action has been taken) and the control group (not subjected to the action) such that the model predicts the net *effect* of the action. Such an approach seems to be well suited to analysis of clinical trial data and to allow for discovering groups of patients for which the treatment is most beneficial. One of the purposes of this paper is to verify this claim experimentally.

Additionally, we present an approach to uplift modeling which allows for application of standard probabilistic classification models, such as logistic regression, in the uplift setting. Further, we extend the approach such that standard classification models built on the treatment and control datasets can be incorporated in a manner similar to semi-supervised learning in order to improve prediction accuracy. The usefulness of both approaches has been verified experimentally on publicly available clinical trial data.

1. Introduction

Traditional classification methods predict the class probability distribution in a given dataset. Based on these predictions an action is often taken on the classified individuals. This approach however is often incorrect, especially in the case of controlled medical trials. The purpose of such studies is to determine, whether there is a difference in response between the treatment group (subject to a therapy) and the control group (subject to an alternative treatment or placebo) in order to asses the effect *caused* by the treatment. Standard classification methods do not allow for the use of control groups and are thus of limited use in this setting.

Uplift modeling, in contrast, allows for the inclusion of a control group and aims at explicitly modeling the difference in outcome probabilities between the two groups, thus being much better suited to clinical data analysis. Moreover, uplift models allow for direct identification of patients for whom the treatment is most effective. This is unlike traditional statistical analysis of such trials which usually aims to determine whether the treatment was effective overall or whether the effectiveness of the treatment differed among a small number of subgroups defined *prior* to the study. This can be interpreted as an uplift model based on a single attribute, since the remaining variables are not used to predict differences between groups. In contrast, true uplift modeling allows for the *discovery* of such subgroups.

In the experimental section we give an example how a treatment may give negative results overall, but positive results for a group of patients selected by an uplift model. We believe that this property of uplift modeling could be very useful for developing personalized medicine.

The term *uplift modeling* was coined in the context of direct marketing applications, where the problems encountered are quite similar to clinical trials: a marketing campaign plays the role of the treatment. A campaign may have a negative effect on customers (an annoying campaign may cause a customer abandon the company completely) which can be seen as an analogue of treatment side effects. See (Radcliffe & Surry, 1999)

Appearing in the *ICML 2012 Workshop on Clinical Data Analysis*, Edinburgh, Scotland, UK, 2012. Copyright 2012 by the author(s)/owner(s).

for a detailed description of uplift modeling in direct marketing.

1.1. Previous work

Surprisingly, uplift modeling has received relatively little attention in the literature. The most obvious approach uses two separate probabilistic models, one built on the treatment and the other on the control dataset, and subtracts their predicted probabilities. The advantage of the two-model approach is that it can be applied with any classification model. Moreover, if uplift is strongly correlated with the class attribute itself, or if the amount of training data is sufficient for the models to predict the class probabilities accurately, the two-model approach will perform very well also in the uplift case. The disadvantage is, that when the uplift follows a different pattern than the class distributions, the models will focus on predicting the class, instead of focusing on the weaker 'uplift signal'. See (Radcliffe & Surry, 2011) for an illustrative example.

A few papers addressed decision tree construction for uplift. See e.g. (Hansotia & Rukstales, 2002; Chickering & Heckerman, 2000; Radcliffe & Surry, 1999; 2011). In (Rzepakowski & Jaroszewicz, 2010) uplift decision trees have been presented which are more in line with modern machine learning algorithms. The approach has been extended to the case of multiple treatments in (Rzepakowski & Jaroszewicz, 2011).

Some regression techniques related to this paper are available. Most researchers follow the two model approach either explicitly or implicitly. In some cases every variable has two coefficients assigned to it, one for the treatment and another for the control case. In others (typically for linear regression), there is a single set of parameters, which however are obtained by subtracting coefficients of two models built separately on the treatment and control data. To our knowledge the approach presented in this paper (based on the class variable transformation) is the first which allows for constructing directly a *single*, linear model which predicts the difference between class probabilities in the treatment and control groups. Moreover, the approach is fully generic, and can be applied with any probabilistic classification model.

Some approaches to regression based uplift modeling have been investigated by the statistical community (Robins, 1994; Robins & Rotnitzky, 2004; Vansteelandt & Goetghebeur, 2003) under various names such as *nested mean models*. Typically linear regression is considered, where models are built separately on the treatment and control datasets, and the coefficients can then be subtracted from each other to produce a single model. This differs from our approach in two ways. First, our method estimates uplift model's parameters directly, without the intermediate stage of building the two models. Second, we address the problem of classification, and offer a general solution which works for all types of classification models. For example, simple subtraction of parameters does not allow one to obtain a single uplift logistic regression model, which is possible with our approach.

In (Vansteelandt & Goetghebeur, 2003) an approach to nested mean models for logistic regression is analyzed, which is most relevant to our work. Two separate logistic models are built on the treatment and control datasets, and their coefficients subtracted. This however results in the difference of *scores*, not of predicted probabilities. A difference in scores (the log of the odds ratio) is much harder to interpret and to apply in a cost based analysis. Imagine, for example, a situation where the treatment class probability is almost equal to 1. In such a case, the model should be insensitive to variable changes affecting only the treatment class probability, as the uplift depends on the control group alone. If one uses scores instead of probabilities this is however not the case. Again, two separate models are built as an intermediate step, while the approach presented here, based on a class variable transformation, induces a single model based on all data.

Clinical trials analysis often involves models where an indicator of the group (treatment or control) is included as a variable in the model. This is equivalent to adding a fixed, group dependent, offset to a classical model. Performance of such models is evaluated in the experimental section. If interactions between all attributes and the group indicator are included, the method reduces to the two model approach. The idea has also been exploited, with some modifications, in the data mining literature (Lo, 2002; Larsen, 2011).

Recent, thorough literature overviews on uplift modeling can be found in (Rzepakowski & Jaroszewicz, 2010) and (Radcliffe & Surry, 2011).

2. Notation

Let us first formalize the uplift modeling problem and introduce the notation used throughout the paper. Let $X_1, \dots, X_m \in \mathbb{R}$ be *predictor variables* and $Y \in \{0,1\}$ be a *class variable* whose behavior is to be modeled. For the class variable, the value of 1 is assumed to be the positive outcome (success) and the value of 0, negative (failure). Additionally, let us introduce a variable $G \in \{T, C\}$ which represents the fact that a given object has been treated ($G = T$) or is in the control group ($G = C$). The *uplift* is defined as the difference between success probabilities in the treatment and control groups. As a shorthand notation, probabilities conditioned on $G = T$ will we denoted by P^T and probabilities conditioned on $G = C$ by P^C. Our task is now to build a model which predicts

$$\begin{aligned}&P(Y = 1|X_1, \ldots, X_m, G = T)\\ &\qquad - P(Y = 1|X_1, \ldots, X_m, G = C)\\ &\quad = P^T(Y = 1|X_1, \ldots, X_m) - P^C(Y = 1|X_1, \ldots, X_m),\end{aligned}$$

that is the uplift caused by taking the action conditional on $X_1, \ldots, X_m$.

In this paper, a *classification model* is defined as a function of $X_1, \ldots, X_m$ returning a value in the range $[0, 1]$ interpreted as the probability that Y takes the value 1. An *uplift model* is a function of $X_1, \ldots, X_m$ returning a value in the range $[-1, 1]$ interpreted as the difference between the probabilities of the event $Y = 1$ in treatment and control distributions. Models will be denoted with uppercase letter M with superscripts; M^T and M^C are classification models for treatment and control data respectively, and M^U an uplift model.

In the case of uplift modeling we now have two training (and testing) datasets, one for the treatment group and one for control. Let us denote the training datasets as D^T and D^C respectively. Further, let the i-th record of the treatment dataset be denoted by $\mathbf{d}_i^T = (\mathbf{x}_i^T, y_i^T)$, where $\mathbf{x}_i^T$ denotes its part corresponding to $X_1, \ldots, X_m$ and y_i^T its Y value. Analogous notation is used for the control dataset.

The most obvious uplift approach using two separate probabilistic models, one built on the treatment and the other on the control dataset, can now be defined formally using our notation. Let M^T and M^C be classification models built on D^T and D^C respectively. They can trivially be combined into an uplift model

$$\begin{aligned}&M^U(X_1, \ldots, X_m)\\ &\qquad = M^T(X_1, \ldots, X_m) - M^C(X_1, \ldots, X_m).\end{aligned}$$

3. Adapting standard classification models to the uplift case using class variable transformation

In this section we present a simple class variable transformation, which allows for the conversion of an arbitrary probabilistic classification model into a model which predicts uplift. Note that this is different from the approach which uses two models, as here a *single* model is created which directly models the uplift, instead of separately modeling treatment and control probabilities.

Let us define a variable $Z \in \{0, 1\}$ such that

$$Z = \begin{cases} 1 & \text{if } G = T \text{ and } Y = 1,\\ 1 & \text{if } G = C \text{ and } Y = 0,\\ 0 & \text{otherwise.}\end{cases} \tag{1}$$

Intuitively, Z equals one if we know that, for a given case, the outcome in the treatment group would have been at least as good (recall that class 1 is considered success) as in the control group, had we known for this case the outcome in both groups.

Let us now look at the probability of the event $Z = 1$. We have, after taking into account the definition of Z,

$$\begin{aligned}&P(Z = 1|X_1, \ldots, X_m)\\ &\quad = P(Z = 1|X_1, \ldots, X_m, G = T)P(G = T|X_1, \ldots, X_m)\\ &\quad + P(Z = 1|X_1, \ldots, X_m, G = C)P(G = C|X_1, \ldots, X_m)\\ &\quad = P(Y = 1|X_1, \ldots, X_m, G = T)P(G = T|X_1, \ldots, X_m)\\ &\quad + P(Y = 0|X_1, \ldots, X_m, G = C)P(G = C|X_1, \ldots, X_m)\end{aligned}$$

Typically, we assume that G is independent of $X_1, \ldots, X_m$, otherwise the study is not well designed. Taking this into account we have $P(G|X_1, \ldots, X_m) = P(G)$ and

$$\begin{aligned}&P(Z = 1|X_1, \ldots, X_m)\\ &\qquad = P^T(Y = 1|X_1, \ldots, X_m)P(G = T)\\ &\qquad\quad + P^C(Y = 0|X_1, \ldots, X_m)P(G = C).\end{aligned}$$

Let us now make an additional assumption (discussed in detail below) that $P(G = T) = P(G = C) = \frac{1}{2}$. We obtain

$$\begin{aligned}&2P(Z = 1|X_1, \ldots, X_m)\\ &\quad = P^T(Y = 1|X_1, \ldots, X_m) + P^C(Y = 0|X_1, \ldots, X_m)\\ &\quad = P^T(Y = 1|X_1, \ldots, X_m)\\ &\qquad\quad + 1 - P^C(Y = 1|X_1, \ldots, X_m),\end{aligned}$$

and finally

$$\begin{aligned}&P^T(Y = 1|X_1, \ldots, X_m) - P^C(Y = 1|X_1, \ldots, X_m)\\ &\qquad\qquad = 2P(Z = 1|X_1, \ldots, X_m) - 1.\end{aligned}$$

So modeling the conditional uplift of Y is equivalent to modeling the conditional distribution of the new variable Z. In other words, we can transform the class variable according to (1), combine the treatment and control training sets, apply *any* standard classification method (capable of predicting class probabilities) to the new dataset, and, as a result obtain an uplift model for Y. We have thus shown a reduction of the uplift modeling problem to standard classification.

During the derivation we have made an assumption that $P(G = T) = P(G = C) = \frac{1}{2}$, which needs not hold in practice. In this case we may reweight, or resample, the training datasets such that the assumption becomes valid. Notice that such a transformation does not affect the conditional class distributions. To see this rewrite $P(Y, X_1, \ldots, X_m, G) = P(Y, X_1, \ldots, X_m|G)P(G)$ and note that reweighting only affects $P(G)$. Of course the inner workings of the learning algorithm can be affected by the reweighting, however, as long as the algorithm does a reasonably good job at modeling the conditional class distributions, the results will still be meaningful. Also, the effect on the learning algorithm need not be negative, similar approaches are highly beneficial in learning with imbalanced classes (Batista et al., 2004), so it makes sense to apply them also in the case of imbalanced treatment and control groups.

The transformation introduced in this section allows for the use of a wide variety of classifiers to model uplift directly. The number of parameters is typically smaller for single models than in the two model approach, which means smaller variance of model predictions. Of course it is possible that the two model approach will result in smaller bias. Experiments in Section 5 suggest that single models are capable of outperforming the two-model approach.

Another advantage of the single model approach is that single linear models are much easier to interpret. We get a single set of coefficients, and the direction and strength of the influence of each variable on the uplift can easily be assessed.

4. Augmenting uplift modeling using treatment and control classifiers

We have argued that learning uplift models directly may be better than having two separate models. However, in many cases the two model approach still offers good performance. Moreover, one might suspect that standard models built separately on the treatment and control datasets could help an uplift learning algorithm achieve better accuracy.

In this section we present an approach similar to semi-supervised learning, where an uplift algorithm and standard classifiers built separately on the treatment and control datasets help each other by labeling more and more training examples. More specifically, recall that for each training case we only know its outcome after treatment *or* if no treatment was applied. However, we may try to predict the most probable treatment class for objects in the control group and vice-versa. This can be done based on either the uplift model or based on two separate models built on the treatment and control datasets. In the algorithm proposed in this paper, all three models add new labels, enriching each others training datasets in the style of semi-supervised learning. See Algorithm 1 for an overview, details will be discussed next.

Algorithm 1 An algorithm for uplift modeling using a semi-supervised style learning approach.

```
Input: Training datasets $D^T$ (treatment) and $D^C$
    (control)
Output: Uplift model $M^U$
$D_c^T \leftarrow D^T$; $D_c^C \leftarrow D^C$
$D_u^T \leftarrow D^T$; $D_u^C \leftarrow D^C$
repeat
    $M^U \leftarrow$ build_uplift_model($D_u^T, D_u^C$)   #use the
    class variable transformation
    $M^T \leftarrow$ build_classifier($D_c^T$)
    $M^C \leftarrow$ build_classifier($D_c^C$)
    $y_c^T \leftarrow$ assign treat. class to $\mathbf{x}_c^C \in D^C$ using $M^U$
    $y_c^C \leftarrow$ assign contr. class to $\mathbf{x}_c^T \in D^T$ using $M^U$
    $y_u^T \leftarrow$ assign treat. class to $\mathbf{x}_u^C \in D^C$ using $M^T$
    $y_u^C \leftarrow$ assign contr. class to $\mathbf{x}_u^T \in D^T$ using $M^C$
    $D_c^T \leftarrow D_c^T \cup \{(\mathbf{x}_c^C, y_c^T)\}$
    $D_c^C \leftarrow D_c^C \cup \{(\mathbf{x}_c^T, y_c^C)\}$
    $D_u^T \leftarrow D_u^T \cup \{(\mathbf{x}_u^C, y_u^T)\}$
    $D_u^C \leftarrow D_u^C \cup \{(\mathbf{x}_u^T, y_u^C)\}$
until stopping condition is met
```

One of the most popular semi-supervised learning approaches is co-training (Blum & Mitchell, 1998); the data is assumed to consist of two independent sets of attributes (views), each used to build a separate model. Both models then contribute to a common training set. Having two independent views guarantees that the approach is not caught in an 'overfitting loop', where wrongly classified labels are added to the training set, further amplifying incorrect behavior. Our algorithm is more in line with (Goldman & Zhou, 2000), where a single feature set is used with two different models, sequentially adding new examples to each others training sets. As long as the models' predictions are uncorrelated, the algorithm may converge to a solution better than each model can achieve on its own.

The key steps of Algorithm 1 perform selection of new records for which the unknown class values are assigned based on the classification and uplift models. The training datasets for the uplift and classification models are separate; the records whose new classes have been assigned by the uplift model are added to the training sets of the classification models and vice-versa. This way, if the two approaches model different aspects of the problem, they will contribute to each others improvement without getting caught in an 'overfitting loop'. Similar assumptions are typically made for semi-supervised learning, see (Goldman & Zhou, 2000) for details.

Following the semi-supervised methodology, labels are added to those records of which the classifiers are most certain. To add a treatment class to a control training record we use the classification model built on the treatment set to find the class of all records in the control dataset, pick the one for which the predicted probability of $Y = 1$ is closest to one or zero, and add this record to the treatment set with the predicted class. More formally, in step 9 of the algorithm

$$\begin{aligned} \mathbf{x}_u^C &= \operatorname*{arg\,max}_{\mathbf{x}_u^C \in D^C} \max\{M^T(\mathbf{x}_u^C), 1 - M^T(\mathbf{x}_u^C)\} \\ y_u^T &= \begin{cases} 1 & \text{if } M^T(\mathbf{x}_u^C) > \frac{1}{2}, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Step 10 is analogous. In order to assess the missing treatment class based on the uplift model we need to make it as consistent with the uplift prediction as possible. Thus, if the predicted uplift is greater than zero we predict the treatment class to be 1, otherwise we set it to zero. More formally, step 7 of the algorithm is implemented as

$$\begin{aligned} \mathbf{x}_c^C &= \operatorname*{arg\,max}_{\mathbf{x}_c^C \in D^C} \left| M^U(\mathbf{x}_c^C) \right| \\ y_c^T &= \begin{cases} 1 & \text{if } M^U(\mathbf{x}_c^C) > 0, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Step 8 is analogous.

The last remaining aspect is the stopping criterion. In our case we simply ran the algorithm for 20 iterations, labeling about 1% of records in each iteration.

5. Experiments on clinical trial data

In this section we evaluate the usefulness of uplift modeling for the analysis of clinical trial data. Our aim is to assess if it is possible to select a subgroup of patients for whom the treatment under consideration is most beneficial.

Unfortunately, despite the ubiquity of randomized clinical trials, there are few publicly available datasets containing the results of such trials. We use two publicly available datasets accompanying the book (Pintilie, 2006) and one dataset from the UCI repository.

The data from (Pintilie, 2006), and many other clinical datasets, require the use of survival analysis. Unfortunately, at the moment, our uplift approaches are only applicable to classification problems. We have thus decided to use one of the censoring variables itself as the class value. While not perfect (we ignore the time to event), the approach seems appropriate: the group assignment is independent of the time at which it was done and the censoring is directly correlated with the time to the occurrence of the event.

5.1. Methods of evaluating uplift classifiers

Before presenting experimental results we address the problem of evaluating the performance of uplift models.

In addition to two training datasets, we now also have 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 its 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 uplift estimate.

In practice it is easier to visualize the performance using *uplift curves*. One of the tools for assessing performance of standard classification models are *lift curves*[1]. In a lift curve, the x axis corresponds to the number of cases targeted and the y axis to the number of successes. In our case both numbers are expressed as the percentage of the total population.

The uplift curve is computed by subtracting the lift curve obtained on the control test set from the lift curve obtained on the treatment test set. Both curves are generated using the same uplift model. Recall the number of successes on the y axis was expressed as a percentage of the total population. This guarantees that the y axes of the two subtracted curves have comparable scales. The interpretation of the uplift curve is as follows: on the x axis we select the percentage of patients to receive the treatment (the remaining ones receive placebo or an alternative treatment) and on the y axis we read the difference between

[1] Also known as cumulative gains curves or cumulative accuracy plots

Table 1. Randomization time variables for the Bone Marrow Transplant data.

variable	description
`dx`	diagnosis: acute myeloid leukemia (AML) or chronic myeloid leukemia (CML)
`extent`	the extent of the disease: limited (L) or extensive (E)
`age`	patient's age in years

the success rates in the treatment and control groups. A point at $x = 100\%$ gives the gain in success probability we would obtain if the whole population was treated. A diagonal line connecting points corresponding to $x = 0\%$ and $x = 100\%$ depicts the random selection of $x\%$ of patients for the treatment. The Area Under the Uplift Curve (AUUC) can be used as a single number summarizing model performance. In this paper we subtract the area under the diagonal line from this value in order to get more meaningful numbers. More details on evaluating uplift models and on uplift curves can be found in (Rzepakowski & Jaroszewicz, 2010; Radcliffe & Surry, 2011).

The proposed methods can be used to convert arbitrary classifier into an uplift model, but in this section we only use logistic regression as the base model. The reason is that logistic regression gave better results than other classifiers and including them would unnecessarily clutter the plots.

We compare the uplift model based on the class variable transformation with the approach based on using two separate classification models whose predicted probabilities are subtracted. Also included is the approach based on two logistic models whose scores are subtracted, as described in (Vansteelandt & Goetghebeur, 2003) as well as an approach based on adding a treatment/control indicator variable to the model.

In all cases ten fold cross-validation has been used to assess model performance. The whole procedure has been repeated an additional ten times and the results averaged such that smoother and clearer uplift curves were produced.

5.2. The Bone Marrow Transplant Data

We first apply uplift modeling the Bone Marrow Transplant data available from (Pintilie, 2006). The original study is described in (Couban et al., 2002). The data covers patients who received two types of bone marrow transplant: taken from the pelvic bone (which we used as the control group) or from the peripheral blood (a novel approach, used as the treatment group in this paper). There are only three randomization time variables given in the Table 1.

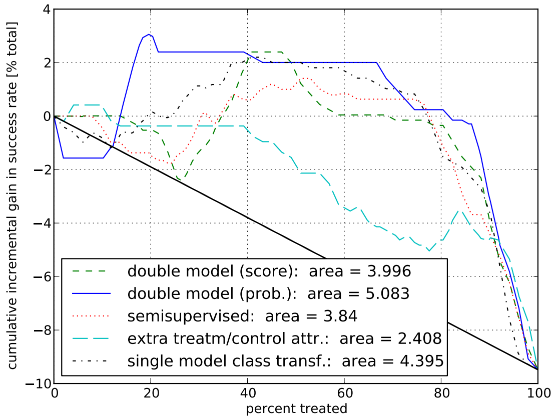

Figure 1. The uplift curve for the bone marrow trial data. The plot illustrates the difference between the nonincidence of chronic GVHD in the two arms of the study.

As the target value we chose the occurrence of the chronic graft versus host disease (GVHD). The nonoccurrence is the successful outcome. As noted in (Couban et al., 2002), the peripheral blood transplant is generally the preferred treatment, so minimizing its side effects is highly desirable.

The uplift curves are shown in Figure 1. Overall, the treatment based on peripheral blood cells has a much higher incidence of chronic GVHD – the success rate is almost 10% lower. However, if one uses the uplift model to select the patients for the peripheral blood based treatment, one can apply it to almost 70% of the population and actually achieve a 2% lower occurrence of GVHD than in the control group.

It can thus be seen that uplift modeling is indeed capable of selecting groups of patients which can benefit from an alternative treatment, even if, overall, the treatment is worse than the standard one.

To understand the effect of each variable on the difference in success probabilities between the two arms of the study it is most convenient to look at the coefficients of the single linear model obtained using the class variable transformation described in Section 3. The formula below gives the linear part of the model:

$$-0.011\text{age}-0.421\{\text{dx=CML}\}+1.135\{\text{extent=L}\}-0.505.$$

The extent of the disease has the decisive impact: patients with local disease have a much lower probability of chronic GVHD if given the alternative treatment

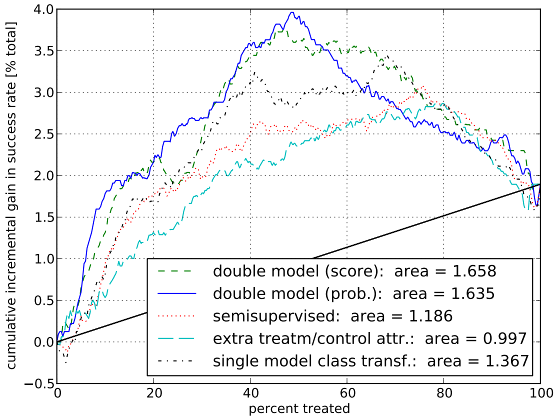

Figure 2. The uplift curve for the tamoxifen trial data.

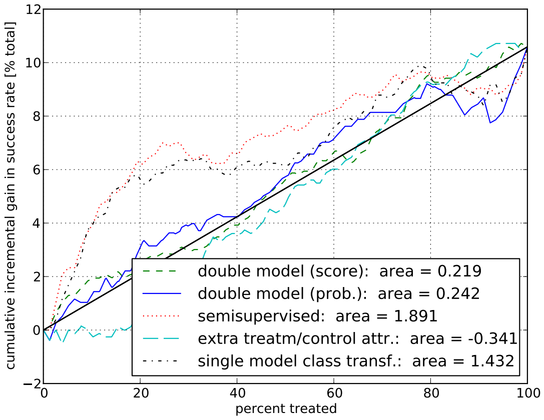

Figure 3. The uplift curve for the UCI hepatitis dataset.

(using bone marrow extracted from peripheral blood). Similarly, if the diagnosis of acute myeloid leukemia (AML) was made, the alternative treatment carries less risk of chronic GVHD.

It can be seen that the double model approach based on subtracting predicted probabilities performed best, with the class variable transformation based approach being second. The approach proposed in (Vansteelandt & Goetghebeur, 2003) lags behind those two approaches. The semisupervised approach fares poorly on this data as does the approach based on adding a group indicator variable.

5.3. The Tamoxifen Data

The second clinical trial dataset we analyze is the data coming from the study of treatment of breast cancer patients with tamoxifen. The study had two arms: tamoxifen-alone (used as control group in this paper) and tamoxifen + radio therapy (used as treatment group), see (Pintilie, 2006) for details. The study analyzed several different outcomes, here we only model the variable `stat` describing whether the patient was alive at the time of the last follow-up. The ‘alive’ value was of course considered success.

Figure 2 shows the uplift curve for the tamoxifen trial. It can be seen that, overall, using tamoxifen with radiotherapy gave a 1.9% increase in success rate over the control group. However, with the uplift model it is possible to select a group comprising about half of the patients, which, when treated with tamoxifen only rises the overall difference in the success rates to almost 4%, double that of applying the alternative treatment to the whole population.

Inspection of the coefficients of the model obtained through class variable transformation, reveals that the histology of the tumor has in general the highest influence on the outcome, with some types giving positive uplift (i.e. sensitivity to treatment) and others negative. The hormone receptor level is another important variable with the ‘positive’ value negatively affecting the rate of success in the treatment branch. The full model has been omitted to save space.

It can be seen that the double model based approaches perform best on this dataset, and the proposed class variable transformation is worse. The semisupervised approach ranks in-between and the model with an indicator variable did worst.

5.4. The Hepatitis Data

The last dataset we analyze comes from the UCI Machine Learning repository. It contains data of patients suffering from hepatitis together with the information on whether they survived or not. While this is not a controlled clinical trial data, we decided to use it due to scarcity of publicly available data from real clinical trials. It contains an attribute describing whether the patients received steroids and we used it to split the data into the treatment and control groups.

The results are shown in Figure 3. In contrast to the other two datasets, the approaches based on two separate classifiers perform extremely poorly and fail to detect any uplift at all. The class variable transformation based approach preforms much better, and the semisupervised learning style approach improves the results even further. The model with an indicator variable was, again, the worst.

The conclusion is that different uplift approaches should be tried on each new dataset.

6. Conclusions and future research

We have shown that uplift modeling has a potential to select patients who will most benefit from a given treatment. It has been demonstrated that even if the treatment overall is not beneficial (or has higher incidence of side effects), uplift modeling may still be capable of selecting a subgroup of patients for which the treatment is successful (or has a low incidence of side effects).

Moreover, we have presented two new approaches to construction of probabilistic uplift models. The first approach uses a class variable transformation which is capable of transforming any *single* classification model into an uplift model, the other uses a semi-supervised learning style approach to augment an uplift model with extra information coming from two separate classifiers built on the treatment and control datasets. Experimental evaluation has shown that the proposed methods have the potential to outperform other uplift modeling approaches but the performance of various modeling techniques depends to a very large extent (larger than in the case of standard classification) on a specific dataset.

Future research will involve extending the models to survival analysis and further experiments on more datasets.

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 2010–2012.

References

Batista, G. E., Prati, R. C., and Monard, M. C. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explorations*, 6(1):20–29, 2004.

Blum, A. and Mitchell, T. Combining labeled and unlabeled data with co-training. In *Proceedings of the Workshop on Computational Learning Theory (COLT)*, pp. 92–100. Morgan Kaufmann, 1998.

Chickering, D. M. and Heckerman, D. A decision theoretic approach to targeted advertising. In *UAI*, pp. 82–88, Stanford, CA, 2000.

Couban, S. et al. A randomized multicenter comparison of bone marrow and peripheral blood in receipients of matched sibling allogeneic transplants for myeloid malignancies. *Blood*, 200:1525–1531, 2002.

Goldman, S. and Zhou, Y. Enhancing supervised learning with unlabeled data. In *ICML*, pp. 327–334, San Francisco, CA, 2000. Morgan Kaufmann.

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

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

Lo, V. S. Y. The true lift model - a novel data mining approach to response modeling in database marketing. *SIGKDD Explorations*, 4(2):78–86, 2002.

Pintilie, Melania. *Competing risks : a practical perspective.* John Wiley & Sons Inc., 2006.

Radcliffe, N. J. and Surry, P. D. Differential response analysis: Modeling true response by isolating the effect of a single action. In *Proceedings of Credit Scoring and Credit Control VI*. Credit Research Centre, University of Edinburgh Management School, 1999.

Radcliffe, N.J. and Surry, P.D. Real-world uplift modelling with significance-based uplift trees. Technical Report TR-2011-1, Stochastic Solutions, 2011.

Robins, J. Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics - Theory and Methods*, 23(8):2379–2412, 1994.

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

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

Rzepakowski, P. and Jaroszewicz, S. Decision trees for uplift modeling with single and multiple treatments. *Knowledge and Information Systems*, 2011. online: http://www.springerlink.com/content/f45pw0171234524j.

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