Data Min Knowl Disc (2015) 29:1531–1559
DOI 10.1007/s10618-014-0383-9

Ensemble methods for uplift modeling

Michał Sołtys · Szymon Jaroszewicz · Piotr Rzepakowski

Received: 1 October 2013 / Accepted: 3 September 2014 / Published online: 17 September 2014

Abstract Uplift modeling is a branch of machine learning which aims at predicting the *causal* effect of an action such as a marketing campaign or a medical treatment on a given individual by taking into account responses in a treatment group, containing individuals subject to the action, and a control group serving as a background. The resulting model can then be used to select individuals for whom the action will be most profitable. This paper analyzes the use of ensemble methods: bagging and random forests in uplift modeling. We perform an extensive experimental evaluation to demonstrate that the application of those methods often results in spectacular gains in model performance, turning almost useless single models into highly capable uplift ensembles. The gains are much larger than those achieved in case of standard classification. We show that those gains are a result of high ensemble diversity, which in turn is a result of the differences between class probabilities in the treatment and control groups being harder to model than the class probabilities themselves. The feature of uplift modeling which makes it difficult thus also makes it amenable to the application of ensemble methods. As a result, bagging and random forests emerge from our evaluation as key tools in the uplift modeling toolbox.

Keywords Uplift modeling · Ensemble methods · Bagging · Random forests

Responsible editor: Johannes Fürnkranz.

M. Sołtys · S. Jaroszewicz (✉)
Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland
e-mail: s.jaroszewicz@ipipan.waw.pl

S. Jaroszewicz · P. Rzepakowski
National Institute of Telecommunications, Warsaw, Poland
e-mail: p.rzepakowski@gmail.com

1 Introduction

Machine learning is primarily concerned with the problem of classification, where the task is to predict, based on a number of attributes, the class to which an instance belongs, or the conditional probability of it belonging to each of the classes. Unfortunately, classification is not well suited to many problems in marketing or medicine to which it is applied. Consider a direct marketing campaign where potential customers receive a mailing offer. A typical application of machine learning techniques in this context involves selecting a small pilot sample of customers who receive the campaign. Next, a classifier is built based on the pilot campaign outcomes and used to select customers to whom the offer should be mailed. As a result, the customers most likely to buy *after* the campaign will be selected as targets.

Unfortunately this is not what a marketer wants! Some of the customers would have bought regardless of the campaign; targeting them resulted in unnecessary costs. Other customers were actually going to make a purchase but were annoyed by the campaign. The result is a loss of a sale or even a complete loss of the customer (churn). While the second case may seem unlikely, it is a well known phenomenon in the marketing community (Hansotia and Rukstales 2002; Radcliffe and Surry 2011).

In order to run a truly successful campaign, we need, instead, to be able to select customers who will buy *because* of the campaign, i.e., those who are likely to buy if targeted, but unlikely to buy otherwise. Similar problems arise in medicine where some patients may recover without actually being treated and some may be hurt by the therapy's side effects more than by the disease itself.

Uplift modeling provides a solution to this problem. The approach employs two separate training sets: *treatment* and *control*. The objects in the treatment dataset have been subject to some action, such as a medical treatment or a marketing campaign. The control dataset contains objects which have not been subject to the action and serve as a background against which its effect can be assessed. Instead of modeling class probabilities, uplift modeling attempts to model the *difference* between conditional class probabilities in the treatment and control groups. This way, the *causal* influence of the action can be modeled, and the method is able to predict the true gain (with respect to taking no action) from targeting a given individual. To date, uplift modeling has been successfully applied in real life business settings. An American bank used uplift modeling to turn an unsuccessful mailing campaign into a profitable one (Grundhoefer 2009). Applications have also been reported in minimizing churn at mobile telecoms (Radcliffe and Simpson 2008).

Ensemble methods are a class of highly successful machine learning algorithms which combine several different models to obtain an *ensemble* which is, hopefully, more accurate than its individual members. The goal of this paper is to evaluate selected ensemble methods in the context of uplift modeling. Our comparison will be focused on bagging and Random Forests (which is a form of bagging using additional randomization), two very popular ensemble techniques, which, as we demonstrate, offer exceptionally good performance. Boosting, another important technique, is beyond the scope of this paper as adapting it to uplift modeling requires an extensive theoretical treatment and merits a separate investigation. Further, we provide an explanation for good performance of those methods which, in our opinion, is that the nature of

uplift modeling naturally leads to highly diverse ensembles. The 'uplift signal' is weak compared to changes in conditional class probabilities which makes the prediction problems difficult; the members of the ensemble are thus very sensitive to noise introduced by random sampling and/or randomized decision tree splits which makes them very different from each other.

In practice, uplift modeling is frequently applied in the marketing domain which in itself is likely (we do not have access to a large enough collection of real marketing datasets to demonstrate this experimentally) to promote ensemble diversity due to the so called *correlation problem* (Abe et al. 2004), i.e., the fact that predictor variables are usually very weakly correlated with customer behavior.

The contribution of this paper is to provide a thorough analysis of ensemble methods in the uplift modeling domain. First we discuss how various types of uplift decision trees can be combined into ensembles. Then we provide an extensive experimental evaluation on real and artificial datasets showing excellent performance of such methods. We also discuss theoretical properties of uplift ensembles and provide an explanation for their good performance based on the concept of ensemble diversity. Although the use of ensemble methods in uplift modeling has already been mentioned in the literature Radcliffe and Surry (2011) and Guelman et al. (2012), to the best of our knowledge this is the first detailed treatment of the subject including both theoretical analysis and thorough experimental verification.

The remaining part of the paper is organized as follows: Sect. 2.1 gives a literature overview, Sect. 3 describes ensemble methods in the context of uplift modeling, and the experimental Sect. 4 demonstrates excellent performance of those methods. Section 5 then offers an explanation for this good performance through an analysis of model diversity. Finally, Sect. 6 concludes the paper.

2 Uplift modeling

In this section we will discuss the state of the art and introduce the notation used in the paper. We begin, however, by mentioning the biggest challenge one encounters when designing uplift modeling algorithms. The problem has been known in statistical literature (see e.g. Holland (1986)) as the

> *Fundamental Problem of Causal Inference.* For every individual, only one of the outcomes is observed, after the individual has been subject to an action (treated) or when the individual has not been subject to the action (was a control case), *never* both.

Essentially this means that we do not know whether the action was beneficial for a given individual and, therefore, cannot assess model's decisions at the level of individuals. This is different from classification, where the true class of an individual is known, at least in the training set.

2.1 Related work

Despite its practical appeal, uplift modeling has received surprisingly little attention in the literature. In this section we will present the related work. We begin with the

motivation for uplift modeling and related techniques and a brief overview of ensemble methods, then we discuss the available uplift modeling algorithms, and finally present current references on using ensemble methods with uplift models.

The first publication explicitly discussing uplift modeling was Radcliffe and Surry (1999). It presents a thorough motivation including several use cases. General discussions of uplift modeling and its applications can also be found in Hansotia and Rukstales (2002) and Radcliffe and Surry (2011).

Experiments involving control groups are becoming common in website optimization, where they are used with so called A/B tests or multivariate tests (Kohavi et al. 2009). The focus of those methods is, however, different from uplift modeling as their main goal is to verify the *overall* effectiveness of a change in website design, not selecting the right design for each customer (looking into specific subgroups is usually mentioned only in the diagnostic context). Another related technique is action rule discovery (Adomavicius and Tuzhilin 1997; Raś et al. 2009) which is concerned with finding actions which should be taken to achieve a specific goal. This is different from uplift modeling which aims at identifying groups on which a predetermined action will have the most positive effect. Contrast sets introduced by Bay and Pazzani (2001) allow for finding subgroups in two datasets on which a specified quantity differs significantly. This is different from uplift modeling which aims at predicting this difference at the level of single records.

The most popular ensemble methods are bagging (Breiman 1996), boosting (Freund and Schapire 1997) and Random Forests (Breiman 2001). Other ensemble methods exist, such as Extremely Randomized Trees (Geurts et al. 2006) or Random Decision Trees (Fan et al. 2003). Essentially, those methods differ by the way randomness is injected into the tree learning algorithm to ensure that models in the ensemble are diverse. In Liu et al. (2008) a unifying framework is proposed which encompasses many approaches to randomization. As we mentioned in Sect. 1, this paper will only look into bagging and Random Forests.

2.2 Notation

We will now introduce the notation used throughout the paper. The probabilities in the treatment group will be denoted by P^T and the probabilities in the control group by P^C. The convention will be kept for other notations, with the superscript ^{T} denoting quantities related to the treatment group and the superscript ^{C} quantities related to the control group. For example, the treatment training dataset will be denoted with D^T and the control training dataset with D^C.

Both training datasets have the same set of predictor attributes $X_1, \ldots, X_m$ and a class attribute Y. The joint domain of the X's (the sample space) is denoted with $\mathcal{X}$, and the domain of Y is assumed to be $\mathcal{Y} = \{0, 1\}$, with 1 considered the positive or desired outcome, e.g. a customer responds to a marketing offer or a patient survives a specified amount of time. We define a *classification model* as a function

$$m(\mathbf{x}) : \mathcal{X} \to [0, 1]$$

with the assumption that $m(\mathbf{x})$ is an estimator of $P(Y = 1|\mathbf{x})$, i.e., the model estimates the probability of the desired outcome conditional on the values $\mathbf{x}$ of the predictor variables. An *uplift model* is a function

$$m^U(\mathbf{x}) : \mathcal{X} \rightarrow [-1, 1] \tag{1}$$

understood as an estimator of

$$P^T(Y = 1|\mathbf{x}) - P^C(Y = 1|\mathbf{x}), \tag{2}$$

that is of the *difference* between success probabilities in the treatment and control groups or, in other words, the expected *net gain* from performing the action on an individual described by predictor attributes' values $\mathbf{x}$. The quantity given in Eq. 2 is also referred to, by some authors, as *uplift*. For consistency, throughout the paper, we will use the term *net gain*.

It is possible to extend the definitions to multiclass problems by including costs or benefits of each outcome $y \in \mathcal{Y}$. For example, let v_y denote the benefit resulting from a given individual ending up in class y (after being subject to the action or when left untreated). Then, the expected net gain of taking the action on an individual described by a feature vector $\mathbf{x}$ is

$$-c + \sum_{y \in \mathcal{Y}} v_y P^T(Y = y|\mathbf{x}) - \sum_{y \in \mathcal{Y}} v_y P^C(Y = y|\mathbf{x}), \tag{3}$$

where c is the cost of taking the action. When $\mathcal{Y} = \{0, 1\}$, $v_1 = 1$, $v_0 = 0$, and $c = 0$, Eq. 3 reduces to Eq. 2. In this paper we will not use the cost model, and Eq. 2 will be our working definition of the net gain. A discussion on using costs in uplift modeling can be found in Hansotia and Rukstales (2002).

2.3 Current uplift modeling algorithms

The most obvious approach to uplift modeling is to build two classification models m^T and m^C on the treatment and control groups respectively and to subtract their predicted probabilities:

$$m^U(\mathbf{x}) = m^T(\mathbf{x}) - m^C(\mathbf{x}).$$

We will call this approach the *double classifier* approach. Its obvious appeal is simplicity; however in many cases the approach may perform poorly. The reason is that both models can focus on predicting the class probabilities themselves, instead of making the best effort to predict the (usually much weaker) 'uplift signal', i.e., the difference between conditional class probabilities in the treatment and control groups. See Radcliffe and Surry (2011) for a detailed discussion and an illustrative exam-

ple.[1] Nevertheless, in some cases the approach is competitive. This is the case when the amount of training data is large enough to accurately estimate conditional class probabilities in both groups or when the net gain is correlated with the class variable, e.g. when people likely to buy a product are also likely to positively respond to a marketing offer related to that product. As we shall see in Sect. 4, ensemble methods are an effective technique for improving the performance also for double classifier models.

Other approaches to uplift modeling try to *directly* model the difference in conditional success probabilities between the treatment and control groups. Most active research follows this direction. Currently such methods are mainly adaptations of two types of machine learning algorithms: decision tree learners and regression models to the uplift case. The first approach to uplift decision tree learning has already been presented by Radcliffe et al. (1999), albeit with very few details given. In a more recent report (Radcliffe and Surry 2011) the authors provide a thorough description of their approach: the decision trees have been specially adapted to the uplift case by using a splitting criterion based on statistical tests of the differences between treatment and control success probabilities introduced by the split.

Another type of uplift decision tree was presented by Hansotia and Rukstales (2002). In the proposed approach, a single uplift decision tree is built which explicitly models the difference between responses in treatment and control groups. The algorithm uses a splitting criterion called $\Delta\Delta P$, which selects tests maximizing the difference between the differences between treatment and control success probabilities in the left and right subtrees, i.e., by maximizing the desired quantity directly.

Another decision tree for uplift modeling is proposed in Chickering and Heckerman (2000). The tree is modified such that every path ends with a split on whether a given person has been treated or not. Otherwise the algorithm is a standard decision tree construction procedure from Buntine (1992), so all remaining splits are selected such that the class (not the net gain) is predicted well.

In Rzepakowski and Jaroszewicz (2010) uplift decision trees have been presented which are more in line with modern tree induction algorithms, the splits are selected based on information theoretical criteria and a pruning method is included. The approach has been extended to the case of multiple treatments in Rzepakowski and Jaroszewicz (2012). This is the uplift model we are going to use as base learner for our ensembles, so we will discuss the approach in more detail in Sect. 3.1.

A few regression techniques for uplift modeling have, under various names, been proposed in medicine, social science and marketing. Most researchers, however, follow the two model approach either explicitly or implicitly. Details can be found in Robins (1994), Robins and Rotnitzky (2004), Vansteelandt and Goetghebeur (2003), Lo (2002) and Larsen (2001). In Jaśkowski and Jaroszewicz (2012) a class variable transformation was presented which allows for converting an arbitrary classification model (the paper used logistic regression) into an uplift model. As a result, a single classifier is built which directly models the difference between success probabilities in the treatment and control groups. Recently Pechyony et al. (2013), the approach has been extended

[1] The example is based on artificial data with two attributes, one strongly affecting the class probabilities independently from the treatment received, the other determining the relatively small sensitivity to the treatment. A model based on two decision trees uses only the first attribute.

to work in the context of online advertising, where it is necessary to not only maximize the net gain, but also to increase advertiser's benefits through maximizing response rate in the treatment group. This type of problems are beyond the scope of this paper.

2.4 Ensemble methods for uplift modeling

We are aware of only two papers using ensemble methods in uplift modeling. In Radcliffe and Surry (2011) the authors mention the successful use of bagging in their uplift modeling practice. Unfortunately, the report contains only a brief note of the technique with no experimental or theoretical evaluation. In this paper we present a thorough experimental evaluation of bagged uplift models as well as an analysis of the theoretical aspects of the technique in the uplift modeling context. Based on it, we present a compelling argument for high utility of bagging in uplift modeling.

Guelman et al. (2012) present an adaptation of the Random Forest algorithm to the uplift case. The adaptation uses splitting criteria defined in Rzepakowski and Jaroszewicz (2010, 2012), but at each node in the tree a random subset of attributes is first selected from which the best test is then picked. For details see Sect. 3.3. Unfortunately the authors do not present an experimental verification of the technique or comparison with other uplift approaches. This gap is filled in this paper, where we compare Random Forests with bagging and single uplift models on several datasets. The experiments and a discussion of the results can be found in Sect. 5.3.

3 Bagging and random forests for uplift modeling

In this section we discuss modifications to ensemble methods needed to apply them to the task of uplift modeling. We begin by describing the base learners we are going to use, then we talk about implementations of uplift bagging and Random Forests.

3.1 Base learners

As our base learners we are going to use both dedicated uplift decision trees and the double classifier models. For the double classifier approach we used pairs of unpruned J4.8 decision trees from the `Weka` package. This is a version of the well known C4.5 learner and is not discussed here in detail, see Quinlan (1992) and Witten and Frank (2005).

As a second type of base learner we are going to use *E-divergence based uplift decision trees* proposed in Rzepakowski and Jaroszewicz (2010, 2012). For the sake of completeness, we will now describe the method briefly. A single tree is built by simultaneously splitting the treatment and control training sets. At each level of the tree the test is selected such that the divergence between class distributions in the treatment and control groups is maximized after the split. Various measures of the divergence lead to different splitting criteria.

Take two probability distributions $P = (p_1, \ldots, p_n)$ and $Q = (q_1, \ldots, q_n)$. There are many ways to measure how far P is from Q, the most common being the Kullback-

Leibler divergence (Csiszar and Shields 2004) and the squared Euclidean distance (which, for symmetry, we call *E-divergence*):

$$KL(P:Q) = \sum_i p_i \log \frac{p_i}{q_i},$$
$$E(P:Q) = \sum_i (p_i - q_i)^2.$$

In Rzepakowski and Jaroszewicz (2010) both divergences have been used, here we focus only on the Euclidean distance as it gave better results in the experiments.[2] The splitting criterion used in Rzepakowski and Jaroszewicz (2010, 2012) is based on the *E-divergence gain* defined as

$$E_{gain}(A) = E\left(P^T(Y) : P^C(Y)|A\right) - E\left(P^T(Y) : P^C(Y)\right), \tag{4}$$

where A is the test being evaluated. This expression measures the increase in divergence after the split. The conditional divergence used in the equation is defined as

$$E\left(P^T(Y) : P^C(Y)|A\right) = \sum_{a \in \mathcal{A}} P(a) E\left(P^T(Y|a) : P^C(Y|a)\right),$$

where $\mathcal{A}$ is the set of possible outcomes of the test A and $P(a)$ is a weighted average of probabilities of outcome a in the treatment and control training sets. It can be shown (Rzepakowski and Jaroszewicz 2010) that the gain possesses several desirable theoretical properties. Instead of using the raw value of the gain, the tree learning algorithm in Rzepakowski and Jaroszewicz (2010, 2012) uses *gain ratio*, which is obtained by dividing (4) by a factor penalizing tests with a large number of outcomes as well as tests which lead to very different splits in the treatment and control training sets. The details are omitted to save space and can be found in Rzepakowski and Jaroszewicz (2010, 2012).

Following Breiman's suggestion Breiman (1996), the trees used as base learners for the ensembles are not pruned. Our experiments confirmed that unpruned trees outperform pruned trees as ensemble members. Single pruned trees are however included in our experiments for comparison. In Rzepakowski and Jaroszewicz (2012) a pruning strategy based on so called *maximum class probability difference* criterion was proposed. Here we use a different approach based on Areas Under the Uplift Curves (AUUCs), which we found to perform better. Uplift curves are used to assess performance of uplift models and are discussed in detail in Sect. 4.2. The approach works by splitting available data into training and validation sets. The tree is built on the training datasets (treatment and control), then, for each node, the validation AUUC of the subtree rooted at that node is compared to the AUUC we would obtain had the

[2] One reason is that KL-divergence tends to infinity when one of the q_i probabilities is very close to zero. This results in numerical instability and estimation problems, negatively affecting lower levels of the trees where little data is available for learning. E-divergence is much better behaved in this respect.

Model construction:

Input: Treatment dataset D^T, control dataset D^C,
the number of trees in the ensemble B
Output: An ensemble $m_1^U, m_2^U, \ldots, m_B^U$ of uplift models

1. **For** $i \leftarrow 1, \ldots, B$:
2. $D_i^T \leftarrow$ draw a bootstrap sample from D^T
3. $D_i^C \leftarrow$ draw a bootstrap sample from D^C
4. Build an uplift model m_i^U based on D_i^T and D_i^C
5. **Return** $m_1^U, m_2^U, \ldots, m_B^U$

Net gain predicted for a new instance x:

$$m^U(\mathbf{x}) = \frac{1}{B}\sum_{i=1}^{B} m_i^U(\mathbf{x})$$

Fig. 1 Bagging algorithm for uplift models

subtree been replaced with a single leaf. If the latter is larger, the subtree is pruned. This is a direct adaptation of classical tree pruning based on validation sets (Breiman et al. 1984).

3.2 Bagging of uplift models

Figure 1 shows the bagging algorithm adapted to the uplift modeling problem. Overall, the algorithm is almost identical to classical bagging used for classification (Breiman 1996). The only difference is that two bootstrap samples are now taken independently from the treatment and control datasets and that members of the ensemble are each built on a pair of samples. Note that we are averaging the predicted net gains, that is the predicted differences between success probabilities in the treatment and control groups (Eq. 2).

Of course one can use any type of uplift model as the base learner, including double classifiers. It turns out that the latter case is equivalent to using a double classifier consisting of two bagged classifier ensembles, one built on the treatment, the other on the control dataset. To see this, denote by m_i^T the classifier built on the bootstrap sample D_i^T, by m_i^C the classifier built on D_i^C, and by $m_i^U = m_i^T - m_i^C$ the i-th double classifier uplift model added to the ensemble. Then, for a given input vector $\mathbf{x}$ the prediction of the bagged uplift model is

$$m^U(\mathbf{x}) = \frac{1}{B}\sum_{i=1}^{B} m_i^U(\mathbf{x}) = \frac{1}{B}\sum_{i=1}^{B}\left[m_i^T(\mathbf{x}) - m_i^C(\mathbf{x})\right] = \frac{1}{B}\sum_{i=1}^{B} m_i^T(\mathbf{x}) - \frac{1}{B}\sum_{i=1}^{B} m_i^C(\mathbf{x}),$$

which is exactly the difference between success probabilities predicted by bagged classifiers trained separately on the treatment and control datasets.

In this paper we will examine ensembles of both double classifiers and dedicated uplift trees, and show bagging to be highly profitable in both cases.

Input: Treatment dataset D^T, control dataset D^C,
the number of randomly selected attributes k
Output: A randomized uplift tree m

1. **If** stopping condition:
2. **Return** a tree consisting of a single leaf
3. Select k attributes at random
4. Pick a test A based on one of the selected attributes using the E-divergence gain ratio
5. **For** each outcome a of A:
6. build a tree m_a recursively on subsets of D^T and D^C
7. **Return** a tree with test A in the root and m_a's as subtrees

Fig. 2 An algorithm for building a member of an Uplift Random Forest

3.3 Random forests for uplift modeling

In case of Random Forest classifiers we tested both the method proposed by Guelman and others in Guelman et al. (2012), which we call *Uplift Random Forests*, and ensembles of double randomized decision trees, which we call *Double Uplift Random Forests*. Uplift Random Forests work the same as bagged E-divergence based uplift decision trees, except that extra randomization is added to the test selection process while building ensemble members: the test for each node in a tree is selected based only on a randomly selected subset of available attributes.

Figure 2 shows the algorithm for building a single member tree of an Uplift Random Forest. The original paper Guelman et al. (2012) used KL-divergence based test selection proposed in Rzepakowski and Jaroszewicz (2010). Here we used the Euclidean distance based criterion (see previous section). The number k of randomly selected attributes was chosen to be the ceiling of the square root of the total number of attributes. Construction of the tree was stopped when either no more than 3 training records remained in the treatment or control training sets or the tree height exceeded 20. Those values were chosen arbitrarily to prevent excessively large trees. Building larger trees had very little impact on the results.

Of course, it is also possible to build a random forest composed of double randomized decision trees, one built on a bootstrap sample D_i^T taken from the treatment dataset, the other on a sample D_i^C taken from the control dataset. We call such models *Double Uplift Random Forests*. Note that this approach involves stronger randomization as each tree constructed on the treatment set is randomized independently of trees constructed on the control. By an argument analogous to the one for bagging, such an uplift model is equivalent to a double classifier model consisting of two Random Forest classifiers.

In our experiments we used `Weka`'s `RandomTree` classifier to construct members of the ensemble. Unfortunately the `RandomTree` class uses a slightly different splitting criterion than J4.8 tree which we use in bagged double classifiers. The former uses raw entropy gain and the latter uses entropy gain ratio, i.e., the gain is divided by the entropy of the test itself. Moreover J4.8 uses heuristics to eliminate tests with very low entropies, see Quinlan (1992) for details. This makes comparison of bagged double classifiers with Double Uplift Random Forests more difficult, but we chose not

to modify the implementations of `Weka` tree learners as they are a standard used by the community, and since neither criterion is uniformly better than the other.

3.4 Theoretical properties

This section discusses theoretical properties of ensemble methods in the uplift setting. We analyze those properties by treating uplift modeling as an instance of classification or regression problems. Essentially, most theoretical properties of classification and regression ensembles almost directly carry over to the uplift case, but are of purely theoretical interest, since the respective quality measures cannot be computed due to the Fundamental Problem of Causal Inference.

The most popular explanation of why bagging works is variance reduction (Breiman 1996). This argument is typically used in the regression context with Mean Squared Error criterion. The error is decomposed into three components: Bayes error, model variance and model bias. In Breiman (1996) Breiman shows that averaging several models decreases the variance component and concludes that for bagging to work, the models in the ensemble must be sufficiently diverse for the reduction in variance to offset to use of bootstrap samples instead of the full training dataset.

Uplift modeling is not, strictly speaking, a regression task, but can be viewed as such when the conditional net gain (defined in Eq. 2) is treated as a numerical quantity to be predicted.[3] Breiman's argument can then be applied directly and is not repeated here. An analogous argument can of course be used when talking about net gain defined in terms of costs or benefits (Eq. 3).

One can also view uplift models as classifiers. The class to be predicted is whether the action will have a positive impact on the given individual. The model given in Eq. 1 then decides that the action should be taken on an individual $\mathbf{x}$ if $m^U(\mathbf{x}) > 0$. Unfortunately, due to the Fundamental Problem of Causal Inference, we never know whether the action was truly beneficial if taken on a given object, so we cannot assess uplift model correctness at the level of individuals.

In the classification context, it can be shown [see Hansen and Salamon (1990) and an essentially equivalent argument in Breiman (1996)] that bagging inflates the predicted probability of the most frequent class thus resulting in improved accuracy. The same argument can directly be applied to uplift models treated as classifiers. Even though the true class values are not available to us, the argument shows that if an ensemble's members correctly decide on taking the action with probability greater than 0.5, we may expect the ensemble to perform better than a single model.

In Breiman (2001) a bound is given on the performance of a classification ensemble in terms of the strength of individual models and correlations between them. However, the definitions of strength and correlation require the knowledge of the true class which is not available in uplift modeling. The bound thus remains true in principle, but the

[3] Note that net gain is bounded to the interval $[-1, 1]$ so the Mean Squared Error criterion may not be appropriate for values close to ± 1 since it does not take into account the error's asymmetry. In practical situations, however, the predicted net gain is rarely close the boundaries of the interval, so the square loss is applicable.

values involved cannot themselves be computed. For this reason we define our own measures of strength and diversity in Sect. 5.

Other explanations for good performance of ensemble methods are presented by Dietterich (2000). They include, for example, the fact that ensembles use a richer model space than single models. All those explanations trivially carry over to the uplift case.

4 Experimental evaluation

In this section we present an experimental evaluation of bagging and Random Forests for uplift modeling. We begin with a general discussion on assessing performance of uplift models, then present the actual experimental results.

4.1 Benchmark datasets for uplift modeling

A significant problem one encounters while working on uplift modeling is the lack of publicly available datasets. Even though control groups are ubiquitous in medicine and their use in marketing is growing, there are relatively few publicly available datasets which include a control group and a reasonable number of predictive attributes. In our experiments we are going to use several publicly available datasets which include true control groups obtained through randomization. Additionally, we also include datasets from the UCI repository artificially split into treatment and control groups. We first describe the datasets coming from randomized trials and later the procedure used to split the UCI benchmarks.

Table 1 summarizes the datasets with real control groups used in our experiments. The first dataset comes from Kevin Hillstrom's MineThatData blog (Hillstrom 2008) and contains results of an e-mail campaign for an Internet based retailer. The dataset contains information about 64,000 customers who have been randomly split into three groups: the first received an e-mail campaign advertising men's merchandise, the second a campaign advertising women's merchandise, and the third was kept as a control. Data is available on whether a person visited the website and/or made a purchase (conversion). We only focus on visits since very few conversions actually occurred. In this paper we use the dataset in two ways: combining both e-mailed groups into a single treatment group (the resulting dataset is called `Hillstrom visit`) and using only the women's merchandise group (dataset called `Hillstrom visit w.`). The women's group was chosen because the campaign on this group was, overall, much more effective.

Additionally, we use several publicly available clinical trial datasets which accompany a book on survival analysis by Pintilie (2006) or are available in the `R` package for statistical computing. The first medical dataset available with Pintilie (2006) is the Bone Marrow Transplant (`BMT`) data on patients who received two types of bone marrow transplant: taken from the pelvic bone (used as the control group since this was the procedure commonly used at the time the data was collected) or from the peripheral blood (a novel approach, used as the treatment group in this paper). The peripheral blood transplant is easier on the donor but may result in a higher rate of rejection

Table 1 Datasets from randomized trials used in the paper

Dataset	Source	#Records			#Attributes
		Treatment	Control	Total	
Hillstrom visit	MineThatData blog (Hillstrom 2008)	42,694	21,306	64,000	8
Hillstrom visit w.	MineThatData blog (Hillstrom 2008)	21,306	21,306	42,612	8
BMT cgvh	Pintilie (2006)	49	51	100	4
BMT agvh	Pintilie (2006)	49	51	100	4
Tamoxifen	Pintilie (2006)	321	320	641	10
Pbc	R, survival package	158	154	312	20
Bladder	R, survival package	38	47	85	6
Cgd	R, survival package	65	63	128	10
Colon death	R, survival package	614	315	929	14
Colon recurrence	R, survival package	614	315	929	14
Veteran	R, survival package	69	68	137	9
Burn	R, KMsurv package	84	70	154	17
Hodg	R, KMsurv package	16	27	43	7

in the recipient. The goal of using an uplift model is to pick a group of patients for whom the alternative therapy is applicable without the increased risk. There are two target variables representing the occurrence of the chronic (cgvh) and acute (agvh) graft versus host disease. We ignore the survival nature of the data and simply treat nonoccurrence as the successful outcome. There are only three randomization time variables: the type and extent of the disease and patient's age.

Note that even though the BMT dataset does not, strictly speaking, include a control group, uplift modeling can still be applied. The role of the control group is played by one of the treatments and the method allows for selection of patients to whom an alternative treatment should be applied.

The second clinical trial dataset accompanying (Pintilie 2006), called Tamoxifen, contains data on treatment of breast cancer with a drug tamoxifen. The control group received tamoxifen alone and the treatment group tamoxifen combined with radiotherapy. We model the target variable stat describing whether the patient was alive at the time of the last follow-up. The dataset contains six variables: size of the tumor, histology, hormone receptor level, haemoglobin level, patient's age, and a binary variable set to true if auxiliary node dissection was done. Details can be found in Pintilie (2006).

Additional datasets come from the survival and KMsurv packages from the R statistical computing system. We will discuss them in less detail since full descriptions are easily accessible online. First, datasets available in the survival package. The pbc dataset comes from the Mayo Clinic study of primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984 and includes data on 312 patients who participated in a randomized controlled trial of the drug D-penicillamine (the control group received placebo). We assumed death before the endpoint of the study to be

the negative outcome and a patient receiving a transplant or being censored to be the positive outcome.

The recurrences of bladder cancer dataset (`bladder`) contains information on 85 subjects who received either the thiotepa drug or placebo. For each patient it is reported whether recurrence occurred during four periods of time. We assumed patients for whom there was at least one recurrence to be the negative cases, those without any recurrence, the positive cases.

The dataset `cgd` comes from a placebo controlled trial of gamma interferon in chronic granulotomous disease (CGD) and contains complete information on the time to first serious infection observed through the end of study. Since each patient eventually developed an infection we considered those who did so in less than 180 days to be negative and the remaining ones positive cases.

The `colon` data comes from a trial of adjuvant chemotherapy for colon cancer. There are two types of treatment which we merged together into a single treatment group. The control group received placebo. We analyzed two target attributes: 'death' and 'recurrence or death' with the resulting datasets called respectively `colon recurrence` and `colon death`.

The `veteran` data comes from a randomized trial of two treatment regimens for lung cancer on 137 patients. For uplift analysis, survival time is omitted and patients alive up to the end of the study constitute the positive examples.

Two additional datasets come from the `KMsurv` package. The `burn` dataset has 154 rows describing infections suffered by patients who underwent burns. The treatment group was subject to body cleansing and the control group to routine bathing. Occurrence of staphylococcus aureus infection was the negative outcome. Finally, the `hodg` dataset describes 43 patients who underwent an allogeneic graft or an autologous graft (control group) as a lymphoma treatment. Those who die by the end of the study constitute the negative examples.

In order to increase the number of available datasets we additionally used an approach described in Rzepakowski and Jaroszewicz (2010, 2012) to artificially split standard UCI datasets into treatment and control groups suitable for uplift modeling.

The conversion is performed by first picking one of the data attributes which either has a causal interpretation (this was the case only for the `hepatitis` dataset) or splits the data evenly into two groups. Details are given in Table 2 taken from Rzepakowski and Jaroszewicz (2012). The first column contains the dataset name and the second provides the condition used to select records for the treatment group. The remaining records formed the control. A further postprocessing step removed attributes strongly correlated with the split itself; ideally, the division into treatment and control groups should be independent from all predictive attributes, but this is possible only in a controlled experiment. A simple heuristic was used for this purpose:

1. A numerical attribute was removed if its means in the treatment and control datasets differed by more than 25 %.
2. A categorical attribute was removed if the probability of one of its categories differed between the treatment and control datasets by more than 0.25.

The number of removed attributes versus the total number of attributes is shown in the third column of Table 2.

Table 2 Conversion of UCI datasets into treatment and control groups

Dataset	Treatment/control split condition	#Removed attributes/#original attributes
Australian	a1 = '1'	2/14
Breast-cancer	Menopause = 'PREMENO'	2/9
Credit-a	a7 $\neq$ 'V'	3/15
Dermatology	Exocytosis $\leq$1	16/34
Diabetes	Insu $>$79.8	2/8
Heart-c	Sex = 'MALE'	2/13
Hepatitis	Steroid = 'YES'	1/19
Labor	Education-allowance = 'YES'	4/16
Liver-disorders	Drinks $<$2	2/6
Primary-tumor	Sex = 'MALE'	2/17
Splice	Attribute1 $\in$ {'*A*', '*G*'}	2/61
Winequal-red	Sulfur dioxide $<$46.47	2/11
Winequal-white	Sulfur dioxide $<$138.36	3/11

Further, multiclass problems were converted into binary problems with the majority class assumed to be class 1 (the desired outcome) and the remaining classes merged into class 0. We note that it is possible to use all analyzed uplift methods in the multiclass setting, however, we chose to use binarization in order to make the analysis (e.g. drawing curves) easier.

4.2 Evaluating uplift models

Let us now discuss methods of evaluating uplift models. The task is more challenging than in traditional machine learning because of the Fundamental Problem of Causal Inference mentioned in Sect. 2. For a given individual we know only one of the outcomes, after or without treatment. As a result, we never know whether the action has been truly beneficial for a given individual or not. Therefore, we cannot asses model performance at the level of single data records, this is possible only for groups of similar records.

Recall that building uplift models requires two training sets. Consequently, we also have two test sets: treatment and control. A typical approach to assessing uplift models (Radcliffe and Surry 2011; Hansotia and Rukstales 2002) is to score both test datasets using the same uplift model and assume that objects in the treatment and control groups which have received similar scores are similar and can be compared with each other. In Hansotia and Rukstales (2002) the authors grouped treatment and control test cases by deciles of their scores and estimated net gains by subtracting success rates within each decile.

A more practical modification of this approach is to visualize model performance using *uplift curves* (Rzepakowski and Jaroszewicz 2010; Radcliffe and Surry 2011). Recall that one of the tools for assessing performance of standard classification models

are lift curves,[4] where the x axis corresponds to the number of cases subjected to an action and the y axis to the number of successes captured by the model.

In order to obtain an *uplift curve* we score both test sets using the uplift model and subtract the lift curve generated on the control test set from the lift curve generated on the treatment test set. The number of successes for both curves is expressed as percentage of the total population such that the subtraction is meaningful.

The interpretation of the uplift curve is as follows: on the x axis we select the percentage of the population on which the action is performed, and on the y axis we read the net gain achieved on the targeted group (the net gain on the remaining cases is zero since no action was performed on them). The point at $x = 100\,\%$ gives the gain in success probability we would obtain if the action was applied to the whole population. A diagonal uplift curve corresponds to performing the action on a randomly selected percentage of the population. Examples of uplift curves are given in Sect. 4.4, more details can be found in Rzepakowski and Jaroszewicz (2010) and Radcliffe and Surry (2011).

As with ROC curves, we can use the Area Under the Uplift Curve (AUUC) to summarize model performance with a single number. We subtract the area under the diagonal from this value in order to obtain more meaningful numbers. Note that the area under the uplift curve can be less than zero; this happens when the model gives high scores to cases for which the action has a predominantly negative effect.

4.3 Experimental setup

Our experiments involved four types of uplift ensembles:

Bagged uplift trees. Bagged ensembles of E-divergence based unpruned uplift decision trees (see Sect. 3.1).
Bagged double J4.8 trees. Bagged ensembles of double classifiers based on unpruned J4.8 models from `Weka`.
Uplift Random Forests. Bagged ensembles of randomized E-divergence based uplift decision trees built using the algorithm in Fig. 2.
Double Uplift Random Forests. Bagged ensembles of double classifiers based on randomized trees from `Weka`.

Additionally, for comparison, we included the base models in the experiments: pruned and unpruned E-divergence based uplift trees and double classifier uplift models based on pruned and unpruned J4.8 trees. This choice allowed us to compare the effectiveness of ensembles versus single models, as well as to assess the effect of extra randomness introduced by Random Forests.

All experiments have been performed by randomly splitting each dataset into training (80 % of the data) and test (the remaining 20 %) parts. Each experiment was repeated 128 times, and the resulting uplift curves have been averaged. The reason for this choice was to make the results repeatable and less sensitive to the random seed used. However, the disadvantage of such an approach is that it hides the variance of the predictions. To address this issue we also compute standard deviations of AUUCs

[4] Also known as cumulative gains curves or cumulative accuracy profiles.

computed over the 128 test sets in a manner similar to bootstrap estimates. Moreover, we use statistical tests to asses the results' significance.

4.4 Illustrative examples

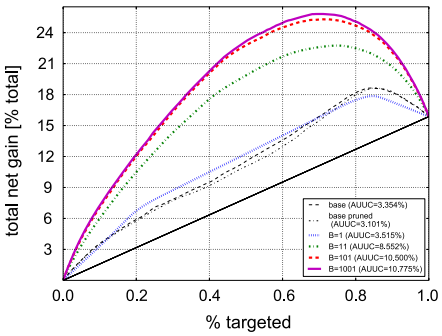
Let us begin by showing some examples of how the use of ensemble methods can dramatically improve the performance of uplift models. Figure 3 shows uplift curves for various bagged and Random Forest ensembles built on three real datasets and three UCI benchmarks artificially split into treatment and control groups. Each chart displays uplift curves for increasing ensemble sizes; Areas Under the Uplift Curves (AUUCs) are given in the legends. For comparison, the charts include uplift curves for base models and pruned base models, i.e., pruned E-divergence based uplift trees or double classifiers based on pruned J4.8 trees, see Sect. 3.1 for details. In case of Random Forest models, the base model is the non-randomized tree of the corresponding type.

It can be seen that applying ensemble methods led to dramatic improvements for those datasets, in some cases more than tripling the Area Under the Uplift Curve and in others turning practically useless single models into highly capable uplift ensembles.

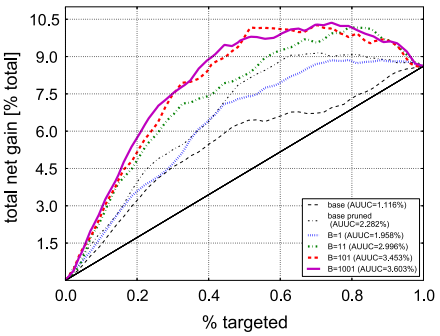
For example, the upper left chart shows uplift curves for the `winequality_white` dataset. The base model, a double classifier based on a J4.8 tree, achieved only a modest improvement over targeting a randomly selected subset of the population. By targeting about 80 % of the database according to base model's selection we are able to obtain the net gain just 3 % higher than if we indiscriminately applied the action to all objects. In contrast, when targeting 70 % of the population selected using a Double Random Forest the difference grows to almost 10 %. The area under the uplift curve for the Random Forest model is more than three times larger than for the base model (regardless, pruned or unpruned)! Similar improvements have been achieved for the `liver_disorders` dataset with the application of bagging to uplift decision trees based on E-divergence test selection criterion.

The `Tamoxifen` dataset is another interesting example; here the base model is practically useless, as its performance is almost identical to random selection of the target group. Applying bagging improved performance significantly: by targeting about 70 % of patients with the drug and radiotherapy and the remaining 30 % with the drug only we would (apart from reducing the number of people subject to radiotherapy and its side effects) achieve, overall, better results than if the combined treatment was administered to all patients. Similar gains are visible for the chronic graft versus host disease in the `BMT` dataset. Using an uplift model, we could target almost 75 % of patients with the alternative, milder therapy while actually achieving lower incidence of side effects. Note that the overall impact of the alternative therapy is negative in this context, but this seems to be due to only about a quarter of the patients for whom it gives particularly bad results.

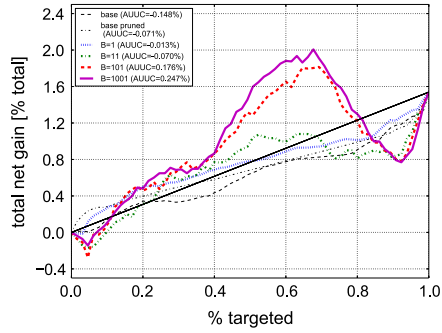
The next subfigure shows the performance of bagged uplift decision trees on the women's merchandise campaign from the `Hillstrom` dataset. The gains are not as spectacular as in the previous cases, but still, the application of bagging resulted in about 10 % increase in the AUUC over a single pruned uplift tree and about 20 % increase over a single unpruned tree.



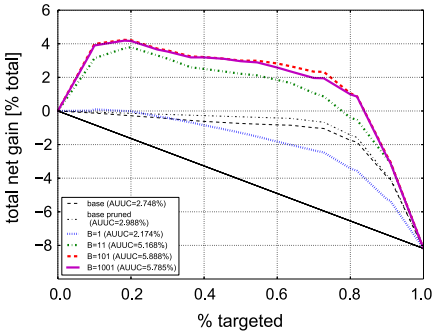
The `winequality_white` dataset. Double Uplift Random Forests.



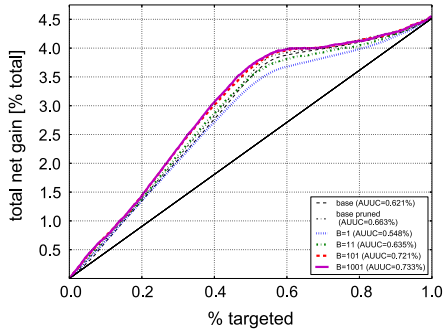
The `liver_disorders` dataset. Bagged E-divergence based uplift trees.



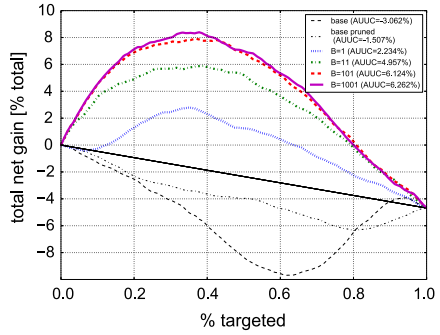
The `Tamoxifen` dataset. Bagged double classifiers based on unpruned J4.8 trees.



The `cgvh` dataset from the BMT trial. Bagged double classifiers based on unpruned J4.8 trees.



The `Hillstrom visit w.` dataset (women's merchandise offer). Bagged E-divergence based uplift trees.



The `credit_a` dataset. E-divergence based Uplift Random Forest.

Fig. 3 Uplift curves for various types of uplift ensembles with increasing number of members B on selected datasets. Areas Under the Uplift Curves (AUUCs) are shown in the legends

The final example is the artificially split `credit_a` dataset, where Uplift Random Forest is seen to perform exceptionally well. The chart requires a comment. It can be seen that the base model makes predictions which are actually worse than random

selection, while an ensemble with just one member performs much better. This is unexpected, since the single member tree was built, due to bootstrapping, on a smaller sample than the full model. The same effect was seen when bagging was applied to E-divergence based uplift decision trees on this dataset. To understand this result we examined the generated trees. When the base model was used, in almost all of the 128 random train/test splits the test in the root of the tree was based on the `A6` attribute which takes 14 different values; this resulted in quick training data fragmentation and poor overall performance. If the same tree construction algorithm was applied to a bootstrap sample taken from the original dataset, tests in the root were almost never based on this attribute resulting in much better trees. The good performance of one member ensembles thus turned out to be a counterintuitive side effect of the test selection criterion proposed in Rzepakowski and Jaroszewicz (2010). As can be seen in the charts presented in the next section this phenomenon occurs (less strongly) also for other datasets as well as for the J4.8 decision trees. To visualize real gains resulting from forming larger ensembles we have included the curves for one model ensembles in all the charts in Fig. 3.

4.5 Performance evaluation of uplift ensembles

The examples shown above were hand-picked to demonstrate the striking benefits ensemble methods can bring to uplift modeling. To provide a more unbiased view, in this section we present a complete analysis comparing the algorithms on all available datasets.

First, we compare the Areas Under the Uplift Curves (AUUCs) of uplift ensembles of increasing sizes with the performance of base models, pruned and unpruned. In case of Random Forests the base models are the corresponding trees *without* randomized attribute selection. The results are shown in Figs. 4 and 5. We begin by discussing the performance of E-divergence based uplift trees, later we move to double classifier models and their ensembles.

Looking at Fig. 4 one can see that for all datasets, except three (`Tamoxifen`, `veteran` and `hepatitis`), forming larger ensembles improves performance for both bagging and Uplift Random Forests, sometimes dramatically so. For the `cgd`, `bladder`, `colon death`, `colon recurrence`, `breast_cancer`, `diabetes`, `heart_c`, `liver_disorders`, `splice`, and both `winequality` datasets the gains over base models were especially large, with Areas Under the Uplift Curves doubling or even tripling. For the `Hillstrom visit` dataset the performance of the ensemble increased steadily as more members were added but fell just short of surpassing the pruned base model. Note that when only women's merchandise offer was considered (see also Fig. 3) bagging brought significant improvement in performance over the base model. The loss of performance on the `Tamoxifen` and `veteran` datasets is most probably due to poor base models.

An interesting observation is good performance of bagging. While in the case of standard classification bagging is considered a simple technique offering only modest improvements in performance, it is very competitive when used for uplift modeling,

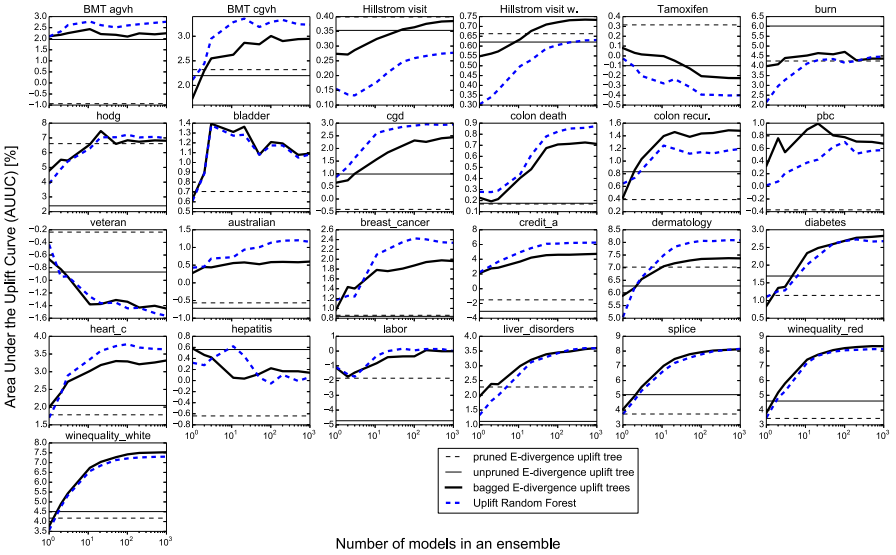


Fig. 4 Areas Under the Uplift Curves versus ensemble size for bagged E-divergence based uplift decision trees and Uplift Random Forests

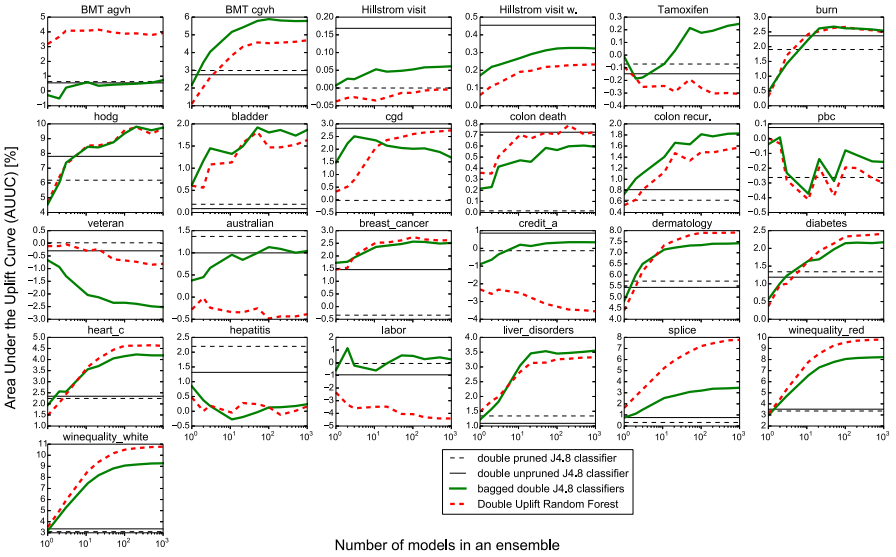


Fig. 5 Areas Under the Uplift Curves versus ensemble size for bagged double classifiers based on J4.8 trees and for Double Uplift Random Forests

comparable to Random Forests. A more thorough discussion of that issue is provided in the next section, where we analyze the correlation between ensemble members.

To test the statistical significance of the results we will take two approaches. First we are going to look at how far the model's AUUC differs from zero (i.e., random

Table 3 Areas under the uplift curves for base models, bagged ensembles and Random Forests

Dataset	Unpruned E-div. tree	1001 bagged E-div. trees	Uplift Rand. Forest (1001)	Double J4.8 classif.	1001 bagged double J4.8	Double uplift Rand. forest
BMT agvh	1.97 ± 4.76	2.25 ± 4.79	2.77 ± 4.58	0.55 ± 2.82	0.74 ± 5.08	3.92 ± 4.78
BMT cgvh	2.20 ± 4.50	2.95 ± 4.17	3.24 ± 4.29	2.75 ± 3.23	$\mathbf{5.78 \pm 4.39}$*	$\mathbf{4.69 \pm 4.57}$*
Hillstrom visit	$\mathbf{0.35 \pm 0.17}$**	$\mathbf{0.38 \pm 0.17}$**	$\mathbf{0.28 \pm 0.16}$*	0.17 ± 0.18	0.06 ± 0.16	-0.00 ± 0.17
Hillstrom visit w.	$\mathbf{0.62 \pm 0.19}$**	$\mathbf{0.73 \pm 0.18}$**	$\mathbf{0.63 \pm 0.19}$**	$\mathbf{0.45 \pm 0.22}$**	$\mathbf{0.32 \pm 0.21}$*	$\mathbf{0.23 \pm 0.22}$*
Tamoxifen	-0.10 ± 1.46	-0.23 ± 1.17	-0.40 ± 1.12	-0.15 ± 1.42	0.25 ± 1.27	-0.31 ± 1.27
Burn	$\mathbf{6.02 \pm 3.17}$*	4.36 ± 4.38	4.47 ± 4.91	2.37 ± 4.10	2.54 ± 4.64	2.49 ± 4.64
Hodg	2.39 ± 7.67	6.81 ± 8.88	6.98 ± 8.66	7.80 ± 9.24	$\mathbf{9.75 \pm 8.60}$*	$\mathbf{9.72 \pm 8.67}$*
Bladder	0.53 ± 5.21	1.09 ± 5.69	1.08 ± 5.69	0.09 ± 4.88	1.86 ± 6.08	1.63 ± 6.10
Cgd	0.99 ± 2.23	2.44 ± 2.73	$\mathbf{2.95 \pm 2.65}$*	$\mathbf{2.83 \pm 2.07}$*	1.67 ± 2.40	$\mathbf{2.74 \pm 2.17}$*
Colon death	0.18 ± 1.50	0.71 ± 1.28	0.88 ± 1.27	0.72 ± 1.30	0.59 ± 1.46	0.73 ± 1.08
Colon recur.	0.83 ± 2.11	1.48 ± 1.78	1.19 ± 1.73	0.81 ± 2.05	1.83 ± 2.12	1.57 ± 2.19
Pbc	0.82 ± 3.42	0.68 ± 2.92	0.57 ± 2.90	0.08 ± 3.34	-0.16 ± 2.93	-0.30 ± 3.00
Veteran	-0.87 ± 2.90	-1.45 ± 3.00	-1.56 ± 2.93	-0.30 ± 1.97	-2.52 ± 2.15	-0.81 ± 2.31
Australian	-0.72 ± 2.60	0.60 ± 2.31	1.16 ± 2.17	1.00 ± 2.65	1.04 ± 2.23	-0.39 ± 2.18
Breast_cancer	0.84 ± 2.82	1.96 ± 2.76	2.33 ± 2.74	1.46 ± 3.35	2.51 ± 3.09	2.62 ± 2.72
Credit_a	-3.06 ± 2.39	$\mathbf{4.73 \pm 2.23}$**	$\mathbf{6.26 \pm 1.93}$**	0.86 ± 2.50	0.34 ± 2.09	-3.55 ± 1.91
Dermatology	$\mathbf{6.28 \pm 1.97}$**	$\mathbf{7.37 \pm 1.41}$**	$\mathbf{8.09 \pm 1.01}$**	$\mathbf{5.44 \pm 2.31}$**	$\mathbf{7.43 \pm 1.54}$**	$\mathbf{7.92 \pm 1.29}$**
Diabetes	1.69 ± 2.36	$\mathbf{2.83 \pm 2.15}$*	$\mathbf{2.68 \pm 2.14}$*	1.19 ± 2.57	2.17 ± 2.34	$\mathbf{2.41 \pm 2.33}$*
Heart_c	2.05 ± 3.22	3.32 ± 3.39	$\mathbf{3.64 \pm 3.33}$*	2.34 ± 3.50	$\mathbf{4.19 \pm 3.29}$*	$\mathbf{4.62 \pm 3.42}$*
Hepatitis	0.56 ± 4.28	0.14 ± 3.74	0.06 ± 3.56	1.32 ± 4.87	0.24 ± 4.10	0.16 ± 3.98
Labor	-4.72 ± 6.47	-0.01 ± 8.69	0.00 ± 8.39	-0.96 ± 8.13	0.27 ± 8.29	-4.40 ± 5.72
Liver_disorders	1.12 ± 3.48	$\mathbf{3.60 \pm 3.10}$*	$\mathbf{3.60 \pm 3.06}$*	1.09 ± 3.40	$\mathbf{3.55 \pm 3.22}$*	$\mathbf{3.32 \pm 3.07}$*
Splice	$\mathbf{5.04 \pm 0.95}$**	$\mathbf{8.13 \pm 0.88}$**	$\mathbf{8.15 \pm 0.79}$**	0.76 ± 0.79	$\mathbf{3.45 \pm 1.25}$**	$\mathbf{7.78 \pm 0.95}$**
Winequality_red	$\mathbf{4.61 \pm 1.58}$**	$\mathbf{8.33 \pm 1.38}$**	$\mathbf{8.14 \pm 1.38}$**	$\mathbf{3.52 \pm 1.65}$**	$\mathbf{8.22 \pm 1.55}$**	$\mathbf{9.81 \pm 1.49}$**
Winequality_white	$\mathbf{4.51 \pm 0.95}$**	$\mathbf{7.53 \pm 0.72}$**	$\mathbf{7.30 \pm 0.74}$**	$\mathbf{3.35 \pm 1.05}$**	$\mathbf{9.28 \pm 0.76}$**	$\mathbf{10.77 \pm 0.76}$**

Standard deviations are indicated in the table. Results more than one (respectively two) standard deviation above zero are bolded and marked with a star (respectively two stars)

prediction) to determine whether we can expect useful predictions on future data. The results are presented in Table 3. Such an approach, however, does not necessarily show relative model performance since one model can consistently outperform another on most datasets even though both models' AUUCs are within one standard deviation from each other. To address such cases we use the procedure described in Demšar (2006) which ranks models on each dataset and performs a Nemenyi rank test across all datasets. The results are given in Fig. 6. The average rank of each model is marked on a scale. Above the scale, the width of the *critical difference (CD) interval* is marked. AUUCs of models which are more than the length of this interval apart differ significantly. Models which are not significantly different are connected with thick black lines. See Demšar (2006) for details.

The left side of Table 3 shows that in four cases Random Forest ensembles were more than one standard deviation above random predictions, while the base models

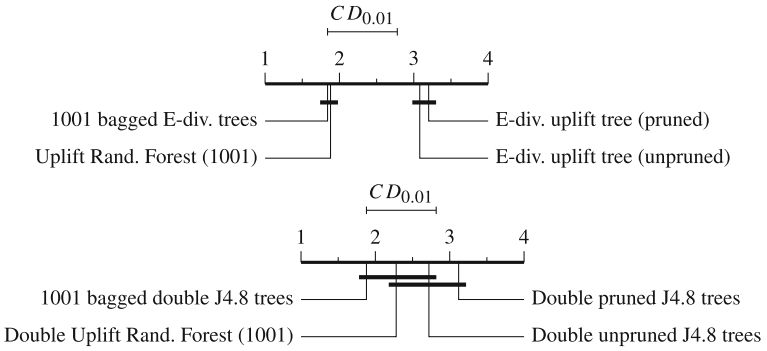


Fig. 6 Critical difference plots for E-divergence and double model based ensembles obtained using the Nemenyi test at the 0.01 significance level

were not. The reverse was true in one case. The upper part of Fig. 6 shows that both types of ensembles outperform pruned as well as unpruned base models at the 0.01 significance level.

The situation is similar for double classifiers based on J4.8 trees as shown in Fig. 5. Here, again, ensemble methods in most cases behave better than base models, often dramatically so. Overall, the differences between the Random Forest approach and bagging are larger in this case. The most probable reason is the slightly different splitting criterion, as discussed in Sect. 3.

The right side of Table 3 shows that in six cases Random Forest ensembles were more than one standard deviation above random predictions, while the base model was not. The reverse was never true. The lower part of Fig. 6 demonstrates, however, that bagged trees were actually the best model, although their superiority over double unpruned trees cannot be rigorously demonstrated.

Figure 7 presents the performance of all uplift ensemble methods on a single plot. It is clear that using ensemble methods is usually very beneficial. Which of the methods produces best results is very much case dependent. It is also interesting that bagging—the simplest of the ensemble methods—performs very well and is usually comparable to, or better than, Random Forest methods.

In the next section we will offer an explanation for the good performance of ensemble methods for uplift modeling by analyzing correlations between predictions of ensemble members.

5 Analysis of ensemble diversity

In case of ensemble methods used for classification it has long been established that the performance of an ensemble depends on the *diversity* of its members (Breiman 1996, 2001; Liu et al. 2008). In this section we will analyze the diversity of uplift ensembles and, based on the analysis, offer an explanation for their good performance.

The first question we need to answer is how to measure the diversity of uplift models. For ensembles of classifiers, measures of member strength and correlation have been proposed by Breiman (2001) and improved by Buttrey and Kobayashi

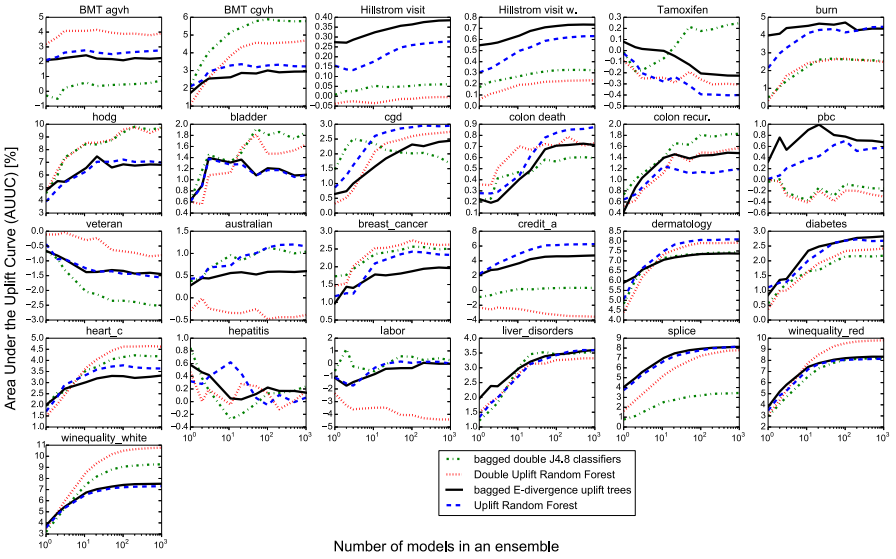


Fig. 7 Areas under the uplift curves versus ensemble size for four types of uplift ensembles

(2003). Unfortunately, those measures are not suitable for uplift models since they require the true class of each instance to be known. In uplift modeling, due to the Fundamental Problem of Causal Inference, only the outcome after treatment or the outcome without treatment is known for a given individual, never both. Therefore, we never know if the action was really beneficial and, consequently, cannot adapt those measures to the uplift case.

We thus had to devise our own measures of model strength and diversity. Strength was measured by simply looking at individual Areas Under the Uplift Curves for all ensemble members. This area can then be averaged over all ensemble members; alternatively, one can simply look at the performance of ensembles containing just a single member which can be read from the initial points of the charts in Figs. 4, 5 and 7.

To measure the diversity of ensemble members we calculate averaged Pearson correlation coefficient of their predicted values of the net gain. Let m_i^U be the i-th member of an uplift ensemble and $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ a dataset. Denote the vector of all net gain values (see Eqs. 1 and 2) predicted by m_i^U on the records of D by

$$\mathbf{u}_i(D) = (m_i^U(\mathbf{x}_j) : j = 1, \ldots, N).$$

The average correlation of predictions made by members $m_1^U, m_2^U, \ldots, m_B^U$ of the ensemble on a dataset D is then measured using

$$\rho(D) = \frac{2}{B(B-1)} \sum_{1 \le i < i' \le B} |\rho(\mathbf{u}_i(D), \mathbf{u}_{i'}(D))|, \tag{5}$$

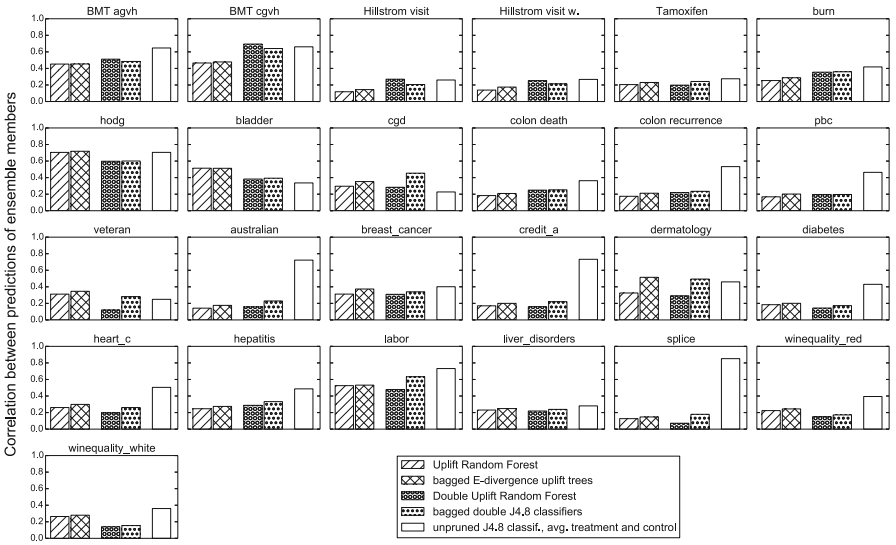


Fig. 8 Ensemble diversity: correlation between predictions of ensemble members

where $\rho(\mathbf{u}, \mathbf{v})$ is the Pearson correlation coefficient between vectors $\mathbf{u}$ and $\mathbf{v}$. That is, we measure ensemble's diversity on a single dataset using averaged absolute values of correlations between predictions of its members. Recall that assessing uplift models requires two test sets: treatment and control. Correlations of member's predictions $\rho(D)$ are averaged over treatment and control test sets and over all random train/test splits to produce the final measure of *ensemble diversity*.

Below we analyze the diversity of various types of ensembles used in our experiments.

5.1 Bagged double classifiers

Figure 8 shows the diversity of various types of uplift ensembles for all benchmark datasets used in our experiments. For comparison, we also include diversity of standard classifiers: unpruned J4.8 trees built separately on the treatment and control datasets. The reported correlation is the average over all 128 treatment and control test sets. Such trees are members of bagged uplift models based on double classifiers, so the comparison of the two rightmost bars in each chart is especially illustrative.

In all but four cases uplift ensembles of double J4.8 trees are more diverse than the classifiers of which they consist. Sometimes the difference is huge, e.g. in case of the `splice` dataset the correlation of predictions made by the uplift ensemble members is just 0.179 even though individual J4.8 trees make highly correlated predictions (coefficient equal to 0.852). Very large differences are also visible for the `australian` and `credit_a` datasets, and large ones for `colon recurrence`, `diabetes`, `heart_c`, `winequality_red`, and `winequality_white`. Note (see Fig. 5) that for all of those datasets adding more members dramatically improved

performance of bagged double J4.8 classifiers, eventually doubling or tripling the AUUC of the ensemble.

One of the main claims of this paper is that this higher diversity is natural for uplift models and that it underlies the good performance of ensemble methods when applied to this problem.

We will now give a more formal explanation of the phenomenon. While correlations are easier to interpret and for this reason are used in the charts, the argument will use covariance of model predictions which is more amenable to calculations. Let $\mathbf{x}$ be a random variable distributed according to the population distribution of X's in the treatment or control group (note that in a properly designed study the predictor variables are identically distributed in both groups). Now, $m_i^T(\mathbf{x})$ and $m_i^C(\mathbf{x})$ are random variables corresponding to predictions of the components of the i-th double classifier in the ensemble. The covariance between the predictions of the i-th and i'-th member of the ensemble can be expressed as

$$\begin{aligned}
&\operatorname{cov}\left(m_i^T(\mathbf{x}) - m_i^C(\mathbf{x}), m_{i'}^T(\mathbf{x}) - m_{i'}^C(\mathbf{x})\right) \\
&\quad = \operatorname{cov}\left(m_i^T(\mathbf{x}), m_{i'}^T(\mathbf{x})\right) + \operatorname{cov}\left(m_i^C(\mathbf{x}), m_{i'}^C(\mathbf{x})\right) \\
&\quad - \operatorname{cov}\left(m_i^T(\mathbf{x}), m_{i'}^C(\mathbf{x})\right) - \operatorname{cov}\left(m_i^C(\mathbf{x}), m_{i'}^T(\mathbf{x})\right). \qquad (6)
\end{aligned}$$

Consider the case which is most difficult for uplift modeling: the success probability varies strongly with $\mathbf{x}$ but the differences between the treatment and control groups are small. This case is difficult, because the uplift model has to pick up the weak 'uplift signal' masked by high variability in class probabilities. The double classifier approach is known to work especially poorly in this case (Radcliffe and Surry 2011). Note however, that since the net gain is small, class probabilities in the treatment and control datasets are close to each other and, therefore, (provided the models m^T and m^C do a reasonably good job) the last two covariances in (6) are likely to be high, decreasing the covariance between uplift ensemble members. Moreover, if the base learners are similar to each other, all four covariances on the right hand side of (6) are likely to have similar values, resulting in the correlation between ensemble members being close to zero.

The scenario described above is frequently encountered in real life applications (Radcliffe and Surry 2011). It is thus a lucky coincidence that the peculiarities of uplift modeling which make the task difficult also make it most suitable for ensemble methods.

5.2 Bagged E-divergence based uplift decision trees

Similar experimental results have been obtained for bagged E-divergence based uplift decision trees. Here again, in all but four cases ensembles of E-divergence based uplift trees are more diverse than individual J4.8 classifiers built on treatment and control datasets, sometimes dramatically so. Although a mathematical justification similar to Eq. 6 is not possible in this case, we believe the reason for the increased diversity is

Table 4 Strength of individual ensemble members for bagged E-divergence based uplift trees and random forests

Dataset	Bagging forest	Random forest
Real		
BMT agvh	**2.11**	2.10
BMT cgvh	1.74	**2.11**
Hillstrom visit	**0.27**	0.16
Hillstrom visit w.	**0.55**	0.30
Tamoxifen	**0.08**	−0.03
Burn	**6.02**	3.26
Hodg	2.39	**3.37**
Bladder	**0.53**	0.49
Cgd	0.99	**1.31**
Colon death	**0.18**	0.12
Colon recurrence	**0.83**	0.55
Pbc	**0.82**	0.04
Veteran	**−0.86**	−0.90
Artificial		
Australian	0.28	**0.43**
Breast_cancer	0.97	**1.17**
Credit_a	2.11	**2.23**
Dermatology	**5.89**	5.05
Diabetes	0.84	**1.10**
Heart_c	**2.00**	1.70
Hepatitis	**0.58**	0.32
Labor	−1.13	**−1.08**
Liver_disorders	**1.96**	1.33
Splice	**4.04**	3.73
Winequality_red	**3.81**	3.47
Winequality_white	**3.76**	3.56

Strength is measured as AUUC of single member ensembles expressed in percent
Bold values indicate better performance of ensemble members

the same: the weaker uplift signal is more difficult to predict than conditional class distribution, making uplift trees very unstable and sensitive to changes in the training sets, such as those caused by taking bootstrap samples. This instability, in turn, results in high diversity and good performance of uplift ensembles.

5.3 Bagging versus random forests

As can be seen by comparing two leftmost bars in the charts in Fig. 8, the diversity of E-divergence based Uplift Random Forests was in *all* cases higher than that of the corresponding bagged uplift tree ensembles (although the difference is sometimes small). This is to be expected since Random Forests use identical tree construction methodology but introduce extra randomness.

It can be seen (Figs. 4, 7) that Uplift Random Forests, overall, perform very well. However, they are not always superior to bagging despite higher model diversity. In some cases bagging performs better and in some other cases both methods are comparable. We believe that the reason is that frequently bootstrapping already provides sufficient diversity to the ensemble, while the randomized test selection restricts the set of attributes available at each tree level which makes individual trees perform worse. To illustrate this phenomenon, Table 4 gives the strength of ensemble members for E-divergence based bagging and Random Forests. In 17 out of 25 cases members of bagged ensembles were indeed stronger.

Increased model diversity is not always able to offset this decrease in strength. This is most clearly visible on the `Hillstrom visit w.` dataset, where adding more members to the Random Forest produced higher gains than it did for bagging (due to higher diversity), but since the individual randomized trees were significantly worse, the overall performance of bagging was better. Similar results have been obtained by Segal for classical regression (Segal 2004).

In case of Double Uplift Random Forests we can also see that their diversity is in general higher than that of bagged double classifiers and the conclusions of the previous paragraph continue to hold. However, since bagging uses a different splitting criterion (see Sect. 3), the conclusions are less clear-cut.

6 Conclusions

The paper presented a theoretical and experimental investigation of the effectiveness of ensemble methods in uplift modeling. The analysis includes two practically important types of uplift models: the double classifier approach and trees which model the net gain directly. Although uplift ensembles have been mentioned before in the literature, this paper is the first to provide a thorough analysis and evaluation, and the first to point out that uplift modeling is especially well suited to the application of such methods. Our experiments on real and artificial data demonstrate that ensemble methods often bring dramatic improvements in performance, turning useless single trees into highly capable ensembles. In some cases the Area Under the Uplift Curve of an ensemble was triple that of the base learner. We demonstrate that features specific to uplift modeling naturally promote high of diversity of ensemble members. Interestingly, this is especially true in cases where uplift modeling itself is difficult.

Further, we compare bagging and Random Forests in the uplift modeling context. We show that Random Forests provide more diverse ensembles at the expense of their members being slightly weaker. In practice both methods perform very well; which one is better is very much case dependent. Random Forests outperform bagging only if increased diversity is able to offset the decrease in individual members' strength.

The most important conclusion of the paper is that ensemble methods come out from the analysis as key uplift modeling tools capable of achieving excellent results. The improvements are typically much bigger than in the case of classification where ensembles are most commonly applied.

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. M.S. was co-funded by the European Union from resources of the European Social Fund. Project POKL 'Information technologies: Research and their interdisciplinary applications', Agreement UDA-POKL.04.01.01-00-051/10-00.

References

Abe N, Verma N, Apte C, Schroko R (2004) Cross channel optimized marketing by reinforcement learning. In: Proceedings of the tenth ACM SIGKDD conference on knowledge discovery and data mining (KDD'04), pp 767–772

Adomavicius G, Tuzhilin A (1997) Discovery of actionable patterns in databases: the action hierarchy approach. In: Proceedings of the third international conference on knowledge discovery and data mining (KDD'97), pp 111–114

Bay S, Pazzani M (2001) Detecting group differences: mining contrast sets. Data Min Knowl Discov 5(3):213–246

Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth, Belmont

Buntine W (1992) Learning classification trees. Stat Comput 2(2):63–73

Buttrey SE, Kobayashi I (2003) On strength and correlation in random forests. In: Proceedings of the joint statistical meetings. Section on statistical computing, San Francisco

Chickering DM, Heckerman D (2000) A decision theoretic approach to targeted advertising. In: Proceedings of the 16th conference in uncertainty in artificial intelligence (UAI'00). Stanford, pp 82–88

Csiszar I, Shields P (2004) Information theory and statistics: a tutorial. Found Trends Commun Inf Theory 1(4):417–528

Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 7:1–30

Dietterich T (2000) Ensemble methods in machine learning. In: First international workshop on multiple classifier systems, pp. 1–15

Fan W, Wang H, Yu PS, Ma Sheng S (2003) Is random model better? On its accuracy and efficiency. In: Proceedings of the third IEEE international conference on data mining (ICDM'03), pp 51–59

Freund Y, Schapire RE (1997) A decision–theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 55(1):119–139

Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. Mach Learn 63(1):3–42

Grundhoefer MD (2009) Raising the bar in cross-sell marketing with uplift modeling. In: Predictive analytics world conference

Guelman L, Guillén M, Pérez-Marín AM (2012) Random forests for uplift modeling: an insurance customer retention case. In: Modeling and simulation in engineering, economics and management. Lecture notes in business information processing (LNBIP), vol 115. Springer, Berlin, pp 123–133

Hansen LK, Salamon P (October 1990) Neural network ensembles. IEEE Trans Pattern Anal Mach Intell 12(10):993–1001

Hansotia B, Rukstales B (2002) Incremental value modeling. J Interact Mark 16(3):35–46

Hillstrom K (2008) The MineThatData e-mail analytics and data mining challenge. MineThatData blog, http://blog.minethatdata.com/2008/03/minethatdata-e-mail-analytics-and-data.html. Accessed 2 April 2012

Holland PW (December 1986) Statistics and causal inference. J Am Stat Assoc 81(396):945–960

Jaśkowski M, Jaroszewicz S (2012) Uplift modeling for clinical trial data. In: ICML, 2012 workshop on machine learning for clinical data analysis. Edinburgh, Scotland, June 2012

Kohavi R, Longbotham R, Sommerfield D, Henne RM (February 2009) Controlled experiments on the web: survey and practical guide. Data Min Knowl Discov 18(1):140–181

Larsen K (2011) Net lift models: optimizing the impact of your marketing. In: Predictive analytics world, 2011. Workshop presentation

Liu FT, Ting KM, Yu Y, Zhou Z-H (2008) Spectrum of variable-random trees. J Artif Intell Res 32(1):355–384

Lo VSY (2002) The true lift model: a novel data mining approach to response modeling in database marketing. SIGKDD Explor 4(2):78–86
Pechyony D, Jones R, Li X (2013) A joint optimization of incrementality and revenue to satisfy both advertiser and publisher. In WWW 2013 Companion Publication, pp 123–124
Pintilie M (2006) Competing risks: a practical perspective. Wiley, Hoboken
Quinlan J (1992) C4.5: programs for machine learning. Morgan Kauffman, Ann Arbor
Radcliffe N, Simpson R (April 2008) Identifying who can be saved and who will be driven away by retention activity. J Telecommun Manag 1(2):168
Radcliffe NJ, Surry PD (1999) 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
Radcliffe NJ, Surry PD (2011) Real-world uplift modelling with significance-based uplift trees. Portrait Technical Report TR-2011-1, stochastic solutions
Raś Z, Wyrzykowska E, Tsay L-S (2009) Action rules mining. In: Encyclopedia of data warehousing and mining, vol 1. IGI Global, pp 1–5
Robins J (1994) Correcting for non-compliance in randomized trials using structural nested mean models. Commun Stat Theory Methods 23(8):2379–2412
Robins J, Rotnitzky A (2004) Estimation of treatment effects in randomised trials with non-compliance and a dichotomous outcome using structural mean models. Biometrika 91(4):763–783
Rzepakowski P, Jaroszewicz S (2010) Decision trees for uplift modeling. In: Proceedings of the 10th IEEE international conference on data mining (ICDM). Sydney, Australia, pp 441–450
Rzepakowski P, Jaroszewicz S (2012) Decision trees for uplift modeling with single and multiple treatments. Knowl Inf Syst 32:303–327
Segal MR (2004) Machine learning benchmarks and random forest regression. Technical report, Center for Bioinformatics & Molecular Biostatistics, University of California, San Francisco
Vansteelandt S, Goetghebeur E (2003) Causal inference with generalized structural mean models. J R Stat Soc B 65(4):817–835
Witten IH, Frank E (2005) Data mining: practical machine learning tools and techniques. Morgan Kaufmann, Ann Arbor