International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 11, November 2015
Available at http://internationaljournalofresearch.org

Evaluation of Ensemble methods for uplift modeling

Glacy Elizabeth Jacob
Assistant Professor,
Department of Computer Science,
Malla Reddy Engineering College for Women,
Maisammaguda, Hyderabad.

M. Sunitha
Assistant Professor,
Department of Computer Science,
Malla Reddy Engineering College for Women,
Maisammaguda, Hyderabad.

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

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.

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 11, November 2015
Available at http://internationaljournalofresearch.org

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.

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.

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; Ra´s 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.

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 11, November 2015
Available at http://internationaljournalofresearch.org

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.

a) 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.

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

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 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.

b) Bagging of uplift models:-

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.

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 11, November 2015
Available at http://internationaljournalofresearch.org

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 ra
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

c) 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 DT i taken from the treatment dataset, the other on a sample DC i 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.

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.

FUTURE WORK

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.

REFERENCES

[1] Abe N, Verma N, Apte C, Schroko R (2004) Cross channel optimized marketing by reinforcement learning.

[2] In: Proceedings of the tenth ACM SIGKDD conference on knowledge discovery and data mining (KDD'04), pp 767–772

[3] 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

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

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

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

[7] 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

[8] 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

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

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

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

[12] 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

[13] 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

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

[15] 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

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

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

[18] 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

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

[20] Ja´skowski 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

[21] 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

[22] 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

[23] Lo VSY (2002) The true lift model: a novel data mining approach to response modeling in database marketing. SIGKDD Explor 4(2):78–86

[24] 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

[25] 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

[26] 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.

[27] Credit Research Centre, University of Edinburgh Management School Radcliffe NJ, Surry PD (2011) Real-world uplift modelling with significance-based uplift trees.

[28] Portrait Technical Report TR-2011-1, stochastic solutions Ra´s Z, Wyrzykowska E, Tsay L-S (2009) Action rules mining. In: Encyclopedia of data warehousing and mining, vol 1. IGI Global, pp 1–5

[29] Robins J (1994) Correcting for non-compliance in randomized trials using structural nested mean models. Commun Stat Theory Methods 23(8):2379–2412

[30] 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

[31] 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

[32] Rzepakowski P, Jaroszewicz S (2012) Decision trees for uplift modeling with single and multiple treatments. Knowl Inf Syst 32:303–327

[33] 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

[34] Witten IH, Frank E (2005) Data mining: practical machine learning tools and techniques. Morgan Kaufmann, Ann Arbor