Data-Enhanced Predictive Modeling for Sales Targeting

Saharon Rosset* Richard D. Lawrence†

Abstract

We describe and analyze the idea of data-enhanced predictive modeling (DEM). The term "enhanced" here refers to the case that the data used for modeling is sampled not from the true target population, but from an alternative (closely related) population, from which much larger samples are available. This leads to a "bias-variance" tradeoff, which implies that in some cases, DEM can improve predictive performance on the true target population. We theoretically analyze this tradeoff for the case of linear regression. We illustrate DEM on a problem of sales targeting for a set of software products. The "correct" learning problem is to differentiate non-customers from newly acquired customers. The latter, however, are scarce. We illustrate how we can build better prediction models by using more flexible definitions of interesting targets, which give bigger learning samples.

1 Introducion

A common situation in data modeling is when the available learning sample from the population of interest is relatively small, but a much larger sample is available from a similar population. The main question we address in this paper is, how can we leverage this abundant, relevant data towards improving predictive modeling? We have encountered this phenomenon in the context of targeting problems, where we are looking for potential customers among a large population of non-customers. The learning problem we define involves differentiating customers that have recently bought the product for the first time ("positive examples", which are usually quite rare) from non-customers. However, the population of veteran, established customers is being ignored completely in this approach. This population is often significantly larger than that of the recently converted "positive examples" above, and since it represents customers, conceivably carries some information about what separates potential new customers from non customers. Many targeting applications do not make that distinction and simply aim to model the differences between customers and non-customers. While this does not represent the correct targeting task it does take advantage of the abundant pre-existing customers.

We argue here that the decision between solving the correct problem with a small amount of data or the closely related, but different, problem with more examples is a bias-variance issue:

- Solving the correct problem minimizes the bias
- Solving the surrogate DEM problem with more data increases the stability of the resulting solution but incurs a cost of increased bias. Thus, we end up closer to a somewhat wrong solution

In this paper, we discuss this trade-off both theoretically and empirically. In Sec. 2 we define the generic DEM approach, and in Sec. 3 offer a quantification of the bias-variance trade-off involved in the case of linear regression. We demonstrate this on simulation data in Sec. 4, and in Sec. 5 we present a sales targeting case study. On this real-life example, DEM proved beneficial in some, but not all, cases we examined.

Although the idea of DEM is surely one that has been applied in practice by many different data modelers — whether knowingly or unknowingly — we are not aware of any publications discussing it in the same form of this paper. The most closely related work we know of is that on multitask learning [1]. In multitask learning, several related learning tasks are being trained simultaneously. In some respects, our formulation can be viewed as a special case of multitask learning, where we are really only interested in the model for one of the tasks, and sharing information between tasks has the specific goal of solving that one task better. This makes our formulation fundamentally different from the general multi-task learning both statistically and algorithmically.

Semi-supervised learning [3] deals with leveraging unlabeled data, in addition to the labeled data of supervised learning, to learn the structure of models. It shares to some extent the motivation behind DEM, but the lack of labels clearly separates the problem.

In several fields, there has been work on adapting models to changing data distributions (e.g., [5] in Natural Language Processing). The problem formulations

*IBM T. J. Watson Research Center, Yorktown Heights, NY Email: srosset@us.ibm.com

†IBM T. J. Watson Research Center, Yorktown Heights, NY Email: ricklawr@us.ibm.com

have some fundamental similarities, however the adaptation problem is more involved, while the DEM problem is simpler, with all data given in advance.

2 Standard modeling and DEM

In the generic predictive modeling framework, we have a learning sample $\{\mathbf{x}_i, y_i\}_{i=1}^n$ drawn i.i.d. from a population D, and we use it to learn a "model" $m(\mathbf{x})$ describing the relationship between $\mathbf{x}$ and y. We then apply it for prediction, i.e., we get additional examples, where we observe only $\mathbf{x}$, and our model predicts the value of y. The model quality is its expected *future* performance on this prediction task, that is $E_D L(Y, m(\mathbf{x}))$ where L is some loss function, such as misclassification rate. This is the modeling scenario that we will call **standard modeling**.

In the **DEM** scenario, we are still interested in predicting cases that are drawn according to distribution D, however we are using training data drawn according to a different distribution $\tilde{D}$. In principle, $\tilde{D}$ may differ from D in the marginal distribution of $\mathbf{x}$, the conditional distribution of $Y|\mathbf{x}$, or both. We concentrate here on the situation that the marginal distribution of $\mathbf{x}$ is either identical under D and $\tilde{D}$, or has a negligible role on the modeling process (e.g., because we are building a discriminative model, and the marginals are close enough), and thus the main concern is the different conditional distribution of $Y|\mathbf{x}$ under D and $\tilde{D}$. We assume we have a larger sample, $\{\tilde{\mathbf{x}}_j, \tilde{y}_j\}_{j=1}^N$ available from $\tilde{D}$ and the main question we are asking is, under what circumstances would we be better off building our model using this larger enhanced sample, rather than the smaller one from D.

The bias-variance (or estimation-approximation) tradeoff involved in this decision is intuitively clear. Using more data for solving a DEM problem would generally give a more stable solution, i.e., one that has smaller variance or estimation error; however this stable solution is inherently not the solution we are looking for, since our modeling problem involves D and the underlying relationship between $\mathbf{x}$ and Y.

When we come to apply DEM in practice, we should keep several considerations in mind. First, even if we choose a DEM approach, it does not seem reasonable to discard the sample we have from our true target population D. Thus, the training sample we would actually be using is the union of $\{\mathbf{x}_i, y_i\}_{i=1}^n$ and $\{\tilde{\mathbf{x}}_j, \tilde{y}_j\}_{j=1}^N$, which can be considered as a random sample from a mixture model of D and $\tilde{D}$. Second, our ultimate interest is in the performance of our models on predicting for data from D. Thus, any validation approach has to be applied appropriately, and evaluate the performance on D only. For example, for k-fold cross validation we would hold out a portion of the $n+N$ training data in each fold of the cross validation, but evaluate the performance only on the data that come from the "unbiased" sample, i.e., are drawn from D. This is the procedure we use in the next sections.

3 Statistical analysis: DEM in linear regression

We demonstrate the effect of DEM through a statistical analysis of its effect in linear regression. The rigorous results we derive on linear regression will serve as intuitive guides for the approximation-estimation tradeoff involved in DEM in other modeling situations. The linear regression solution to the standard problem is:

$$\begin{aligned} \hat{\beta} &= \arg\min_{\beta\in\mathbb{R}^p} \|\mathbf{y} - X\beta\|_2^2 \\ &= (X^\intercal X)^{-1} X^\intercal \mathbf{y} \qquad (3.1) \end{aligned}$$

Denote by $Z = (X^\intercal, \tilde{X}^\intercal)^\intercal$ the predictor matrix to be used for fitting the DEM model, and similarly by $\mathbf{v} = (\mathbf{y}^\intercal, \tilde{\mathbf{y}}^\intercal)^\intercal$ the response vector for this model. Then the solution of the DEM linear regression problem is:

$$\hat{\beta}_Z = (Z^\intercal Z)^{-1} Z^\intercal \mathbf{v} \qquad (3.2)$$

For the purpose of this analysis, we assume a general homoscedastic model, i.e., $Y = f(\mathbf{x}) + \epsilon$ when sampling from D, $Y = \tilde{f}(\mathbf{x}) + \epsilon$ when sampling from $\tilde{D}$ with $\epsilon \sim (0, \sigma^2)$ i.i.d.

The quantity we are interested in is the expected (future) squared error loss for these two models. If we denote a second, independent draw of the response vector by Y_{new}, we are interested in:

$$\begin{aligned} &E_{\mathbf{Y},\mathbf{Y}_{\text{new}}} \frac{1}{n}\sum_i (Y_{\text{new},i} - \mathbf{x}_i\hat{\beta})^2 = \\ &= \sigma^2 + \frac{1}{n}\|f(X) - XE\hat{\beta}\|^2 + \frac{1}{n} tr(X\text{Var}(\hat{\beta})X^\intercal) \qquad (3.3) \end{aligned}$$

which gives us the bias-variance decomposition for the expected prediction MSE of linear regression, with the first term in (3.3) representing the *irreducible error*, the second term is the *Bias*^{2} and the third term is the *Variance*. Using (3.1), we re-write the bias and variance as (see, e.g., [2] (chapter 7)):

$$\begin{aligned} \text{Bias}^2 &= \frac{1}{n}\|(I - X(X^\intercal X)^{-1}X^\intercal) f(X)\|^2 \\ \text{Variance} &= \frac{1}{n}\sigma^2 \sum_i \mathbf{x}_i (X^\intercal X)^{-1} \mathbf{x}_i^\intercal = \frac{p}{n}\sigma^2 \end{aligned}$$

If we analyze the DEM model in the same spirit, we get:

$$\begin{aligned} &E_{\mathbf{V},\mathbf{Y}_{\text{new}}} \frac{1}{n}\sum_i (Y_{\text{new},i} - \mathbf{x}_i\hat{\beta}_Z)^2 = \\ &= \sigma^2 + \frac{1}{n}\left(\|f(X) - XE\hat{\beta}_Z\|^2 + tr(X\text{Var}(\hat{\beta}_Z)X^\intercal)\right) \end{aligned}$$

and plugging in the mean and variance of $\hat{\beta}_Z$, we get:

$$\begin{aligned}\text{Bias}_Z^2 &= \tfrac{\|f(X) - X(Z^\intercal Z)^{-1}\left(X^\intercal f(X) + \tilde{X}^\intercal \tilde{f}(\tilde{X})\right)\|^2}{n}\\ \text{Variance}_Z &= \tfrac{1}{n}\sigma^2 \textstyle\sum_i \mathbf{x}_i (X^\intercal X + \tilde{X}^\intercal \tilde{X})^{-1} \mathbf{x}_i^\intercal\end{aligned}$$

We are now ready to illustrate that DEM is favorable in terms of Variance, while the standard modeling approach minimizes Bias^2.

Theorem 3.1. *The variance is decreased by DEM, i.e.:*

$$\textit{Variance}_Z \leq \textit{Variance}$$

If rank$(Z) = p$*, the inequality is strong.*

Proof. It is easy to see that $\text{tr}(Z(Z^\intercal Z)^{-1}Z^\intercal) = p$, and from the positive semi-definiteness of $Z^\intercal Z$ we thus get:

$$\begin{aligned}\text{Variance}_Z &= \frac{1}{n}\sigma^2 \sum_i \mathbf{x}_i (X^\intercal X + \tilde{X}^\intercal \tilde{X})^{-1}\mathbf{x}_i^\intercal \leq \\ &\leq \frac{1}{n}\sigma^2 \text{tr}(Z(Z^\intercal Z)^{-1}Z^\intercal) = \text{Variance} \qquad (3.4)\end{aligned}$$

If Z is of rank p, and $\tilde{X}$ is not uniformly 0, the inequality clearly becomes strong.

We can approximately quantify this reduction in variance as:

$$\text{Variance}_Z \approx \frac{n}{n+N}\text{Variance}$$

if we assume that $\tilde{X}$ and X come from the same distribution and thus that the $n+N$ terms in (3.4) are of similar magnitude.

Theorem 3.2. *The bias is increased by DEM, i.e.:*

$$\textit{Bias}_Z^2 \geq \textit{Bias}^2$$

Proof.

$$\begin{aligned}& n \cdot \text{Bias}_Z^2 = \\ &= \|f(X) - X(Z^\intercal Z)^{-1}\left(X^\intercal f(X) + \tilde{X}^\intercal \tilde{f}(\tilde{X})\right)\|^2 = \\ &= \|f(X) - X(X^\intercal X)^{-1}X^\intercal f(X) + \\ &\quad + X(X^\intercal X)^{-1}X^\intercal f(X) - X(Z^\intercal Z)^{-1}E\mathbf{v}\|^2 = \\ &\overset{(*)}{=} \frac{1}{n}\|f(X) - X(X^\intercal X)^{-1}X^\intercal f(X)\|^2 + \\ &+ \frac{1}{n}\|X(X^\intercal X)^{-1}X^\intercal f(X) - X(Z^\intercal Z)^{-1}E\mathbf{v}\|^2 \geq \\ &\geq n \cdot \text{Bias}^2\end{aligned}$$

where the equality in (*) is because the two summands are orthogonal.

If f and $\tilde{f}$ are actually linear in the data, i.e., $f(\mathbf{x}) = \mathbf{x}^\intercal\beta$ and $\tilde{f}(\mathbf{x}) = \mathbf{x}^\intercal\tilde{\beta}$ then Bias$^2 = 0$ and:

$$\text{Bias}_Z^2 = \frac{1}{n}\|X(Z^\intercal Z)^{-1}\tilde{X}^\intercal\tilde{X}(\beta - \tilde{\beta})\|^2$$

That is, the increase in bias depends on the distance between the standard solution β and the DEM one $\tilde{\beta}$.

4 Simulated data study

We first illustrate the properties of Bias^2 and Variance through a simple linear regression example. Our feature vectors $\mathbf{x}$ are in $p = 20$-dimensional space, and the "true" model is defined by $\beta = (10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0, ..., 0)$, i.e., it depends only on the first 10 coefficient:

$$y_i = 10x_{i,1} + 9x_{i,2} + ... + x_{i,10} + \epsilon_i \ , \ \epsilon_i \sim N(0, 10^2)$$

The model for the enhanced data is similar, except that the coefficient vector has random noise added, whose magnitude is smaller than the "true" coefficient β_1:

$$\tilde{\beta}_j = \beta_j + \delta_j \ , \ \delta \sim N(0, 3^2)$$

All X values are drawn i.i.d. $N(0, 1)$.

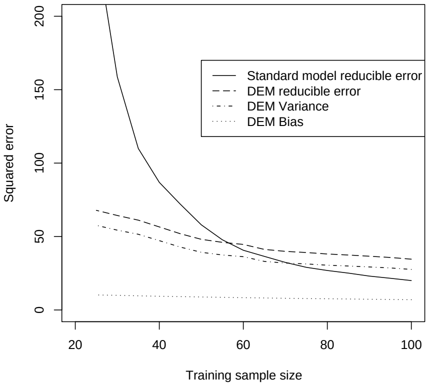

Figure 1: Bias^2 and Variance of regression models. The horizontal axis is the number of training data from the "correct" distribution.

We fix the size of the data enhancement to be $N = 200$ and examine the effect of changing the "correct" data size n between 25 and 100. In each setting, we can analytically calculate Bias^2 and Variance (the Bias^2 of the standard model is 0, since the truth is linear). Fig. 1 shows these quantities as a function of the size n of the correct sample. We observe that as long as n is big enough (more than about 55), we are better off ignoring the DEM sample. When the correct data becomes scarce, though, the importance of variance reduction through inclusion of the DEM data surpasses that of unbiasedness, and the DEM approach gives better solutions.

Next, we create a similar example for classification. The setup is very similar, with $p = 20$ dimensions and a logistic model with the same β:

$$logit(P(Y = 1|\mathbf{x})) = 10x_1 + 9x_2 + ... + x_{10}$$

As before, we create $\tilde{\beta}$ by adding gaussian noise with variance 9 to β. We evaluate model performance by its misclassification rate on a large test sample. In this example, the lines cross, and DEM becomes beneficial, once we are down to about $n = 55$ "correct" examples. The results can be seen in Fig. 2.

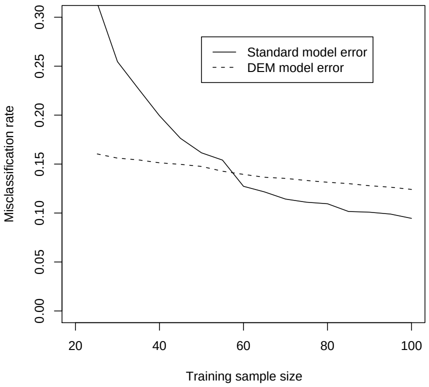

Figure 2: Classification simulation.

To summarize our simulation results, we have shown that DEM is practically useful in these simple examples for both regression and classification. As expected, once the "correct" data sample becomes small enough, prediction error is dominated by estimation error, and the variance control of DEM becomes critical.

5 Case study: sales targeting

At the Data Analytics Research Group at IBM Research, we have been involved in a large sales targeting project, with the objective of helping IBM Software Group (SWG) sales teams in identifying potential customers for different products.

IBM SWG sells five main families or brands of products, defined by areas of Information Technology needs[1]. Our analysis is concentrated on the *DB2* brand, which provides information management solutions. The data we have available for this purpose include historical IBM sales, both of software products and other IBM products (in particular, hardware and services), and external information about companies around the world from the data collected by specialized companies like Dun & Bradstreet (*http://www.dnb.com/us*).

In this case study we concentrate on the problem of modeling *White space* companies, that have not purchased any products from IBM in the past. For these companies, the data we have available to use as features is limited to the external data sources. The variables we consider are from the Dun & Bradstreet database, and include variables like *company size* (expressed as revenue and number of employees), *industry classification*, *company location* etc. Some of these variables are numeric (like company size indicators) and some are categorical (like state), potentially with numerous categories. Variables were transformed as appropriate. For example, company size variables were also represented as rank of company size within the industry, to account for the long tails of company size distribution and the different meaning of "large" in different industries.

We now concentrate on the white-space modeling problem of identifying potential new customers for *DB2* products among non-IBM customers, based on their Dun & Bradstreet information. The analysis we describe here was done using logistic regression. We also experimented with boosted trees and obtained similar results. For confidentiality reasons, some of the details regarding actual numbers of customers involved and model descriptions are omitted.

We would like to understand what characterizes the companies that were not IBM customers before the "current" period (say, the last year), then decided to buy *DB2* within that period. If we accept this definition of "positive" examples, as companies converted within the last year from white-space to *DB2* customers, then our learning problem can be stated as:
Build a model to differentiate companies who were white space (i.e., not IBM customers) on 1/1/2003, then bought DB2 *in 2003, from companies who have never bought from IBM.*

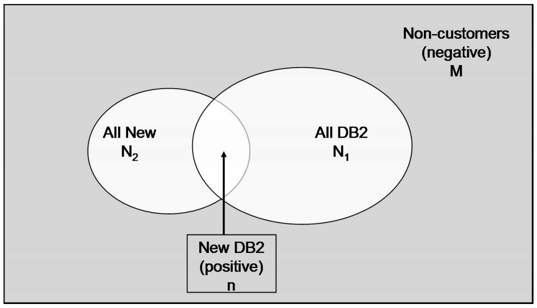

Figure 3: *DB2* sampling groups.

Fig. 3 illustrates the populations that are involved in the learning process. The standard learning problem we have defined is of differentiating the n examples we have of new *DB2* customers, from the M non-IBM customers. n is on the order of 100, while M is large enough for any practical purpose — several tens of thousands. The small size of n limits our ability to learn good models, and there are various ways of increasing the pool of positive examples for learning to create a DEM problem. Here we examine two ways:

1. Consider all new SWG customers in 2003 as positive examples. This is the population of size N_2 in

[1] Please visit *http://www.ibm.com/software* for details

Fig. 3. There are five product groups, so $N_2 \approx 5n$.

2. Consider all *DB2* customers as positive examples. This is the population of size N_1 in Fig. 3, and N_1 is *significantly* larger than n.

We compare the standard model to these two DEM models. We use 20-way cross validation, where only the "correct" data out of the holdout fold is used for evaluation (see Sec. 2). We concentrate our interest on lift at the higher propensity end, since our models are to be used for sales targeting, and only a small percentage of companies can realistically be approached.

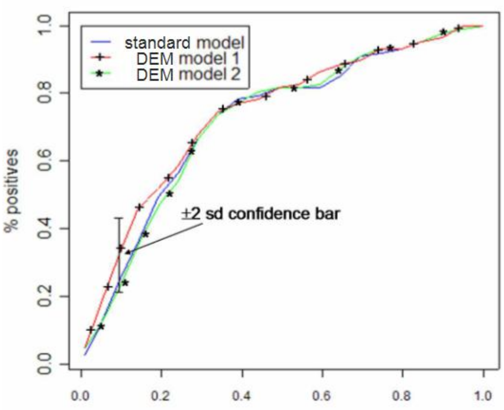

Figure 4: Cross validated lift for the three *DB2* models

Fig. 4 shows the cross-validation lift curves for the three models. We see that on the left side, which represents the highest scored companies, DEM model 1, which is the one using all new SWG customers as positive examples, performs much better than DEM model 2 and the standard model. At 10% of population, which is what a typical targeting effort may be interested in, DEM model 1 successfully recognizes about 35% of actual purchasers, for a lift of 3.5, while the other two models have lift of 2.5 or less. The figure also shows a 2-standard deviation confidence interval for the lift of DEM model 1 at this point, calculated using the method from [4]. Statistical significance is difficult to assert, which is not surprising given the paucity of "true" positive examples, hence high variance of evaluation. However, DEM model 1 is clearly the best choice.

5.1 Performance of the models during field testing

5.1 Performance of the models during field testing As part of their standard process, sales professionals will identify companies likely to purchase a specific IBM software brand. These "opportunities" are logged in a database for further tracking. Since these potential sales have been identified by human experts, it is of interest to compare the propensities predicted for these opportunities with the background propensity distribution. Fig. 5 compares these two distributions for opportunities identified in 2004. The predicted response has been binned in 5 discrete bins, and only 3.2% of opportunities receive a Very High propensity. In contrast, 29.2% of the opportunities identified by sales professionals received this highest score. This analysis suggests that the models are preferentially identifying good candidate buyers, given the nearly 10x enrichment of actual logged opportunities receiving Very High propensity scores.

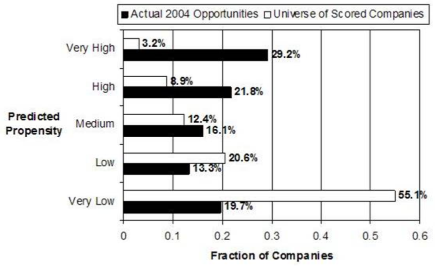

Figure 5: Field performance of DEM model.

6 Summary

We developed the idea of data-enhanced predictive modeling to address the common situation in which only a relatively small number of learning examples are available, but a larger sample is available from a similar population. We showed theoretically that using the standard approach minimizes bias, while using DEM leads to a reduction in variance. Application of the both approaches to a customer targeting problem demonstrated improved accuracy with a DEM model.

The results of DEM model 1 have been embedded during 2005 in a web-based tool, which has been extensively deployed to aid IBM sales professionals in their targeting efforts.

Acknowledgements. We thank G. Atallah, M. Collins, N. Verma and S. Weiss of IBM.

References

[1] R. Caruana. Multitask learning. *Machine Learning*, 28:41–75, 1997.

[2] T. Hastie, T. Tibshirani, and J. Friedman. *Elements of Statistical Learning*. Springer-Verlag, New York, 2001.

[3] T. Mitchell. The role of unlabeled data in supervised learning. In *Proceedings of the Sixth International Colloquium on Cognitive Science*, 1999.

[4] S. Rosset, E. Neumann, U. Eick, N. Vatnik, and S. Idan. Evaluation of prediction models for campaign planning. In *Proceedings of KDD-01*, 2001.

[5] T. Zhang, F. Damerau, and D. Johnson. Updating an nlp system to fit new domains. In *CoNLL 03*, 2003.