Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2004 Proceedings

Americas Conference on Information Systems (AMCIS)

December 2004

An Empirical Study of Cost-sensitive Classification in Campaign Management

Ying Lu
University of Wisconsin--Milwaukee

Atish Sinha
University of Wisconsin--Milwaukee

Huimin Zhao
University of Wisconsin - Milwaukee

Thomas Martin
Journal Sentinel Direct

An Empirical Study of Cost-sensitive Classification in Campaign Management

Ying Lu
University of Wisconsin—Milwaukee
yinglu@uwm.edu

Atish Sinha
University of Wisconsin—Milwaukee
sinha@uwm.edu

Huimin Zhao
University of Wisconsin—Milwaukee
hzhao@uwm.edu

Thomas Martin
Journal Sentinel Direct
tmartin@journalsentinel.com

ABSTRACT

Extremely unbalanced data and unequal costs are key challenges in data mining for campaign management in CRM. This paper presents an empirical study of cost-sensitive classification in a real-world campaign analysis in the newspaper industry, where the response rate is extremely low while the two types of misclassification costs are very different. Incorporating cost information provided by domain experts, the authors investigate the use of three classification methods, i.e., naive Bayes, logistic regression, and decision tree, for campaign response prediction. The data are analyzed under various estimated cost matrices. Cross-validated evaluation results show that cost-sensitive classification techniques are capable of obtaining both higher expected response rates and higher expected profits than the current practice when the estimated misclassification costs are moderately unequal. This empirical study provides useful insights for practice in campaign management, as well as research in classification using extremely unbalanced data and unequal costs.

Keywords

Data mining, cost-sensitive classification, CRM, customer acquisition, campaign management, newspaper industry.

INTRODUCTION

Customer relationship management (CRM) is gaining increasing attentions from many organizations. According to the recent forecast of Forrester research, the global CRM market will grow at more than 11% annually and will hit $73.8 billion in 2007 (http://www.forrester.com). The central theme of CRM is to achieve better marketing communication by providing the right offer to the right customer at the right time through the right channel. Specifically, the objectives of CRM are to improve customer acquisition, customer retention, customer loyalty, and customer profitability (Swift 2001; Winer 2001).

New customer acquisition is a critical process in CRM, usually utilizing multiple approaches such as direct mail, telemarketing, and customer service. In general, a firm first identifies market segments containing prospective customers and then plans and conducts marketing campaigns (Berson, Smith, and Thearling 2000). In many cases, these campaigns involve high volumes of mailings and large amount of data. Thus, data mining applications have been employed to automate the process and identify patterns of response behaviors (Apte, Liu, Pednault, and Smyth 2002; Berson et al. 2000; Brachman et al. 1996). However, because the typical response rates tend to be very low, the data collected are usually extremely unbalanced. Further, the consequences of errors in misclassification of potential customers are costly. Thus, the problems of unbalanced classes and asymmetric misclassification costs have posed challenges to the application of traditional data mining techniques in campaign analysis. Indeed, machine learning from unbalanced data sets is an important issue for both practice and research (Elkan 2001; Japkowicz 2000; Provost and Fawcett 2000). A number of studies have examined various approaches to this problem (e.g., Elkan 2001; Japkowicz 2000; Provost and Fawcett 1997; Provost and Fawcett 1998; Provost and Fawcett 2000; Weiss and Provost 2003). However, there is still a lack of empirical research especially using data from real-world problems (Salzberg 1999).

Therefore, the purpose of this research is to examine the application of cost-sensitive classification to campaign management in CRM and to evaluate the business value of cost-sensitive classification. In this paper, we present an empirical study in the context of newspaper industry. Incorporating misclassification cost information provided by domain experts, we applied three classification methods, i.e., naive Bayes, decision tree, and logistic regression, to analyze the response behavior and profits using data obtained from marketing campaigns by a newspaper company. In order to better reflect the actual effect of the campaign, we adjusted the ratios of the two types of errors (i.e., false positive and false negative) in the cost matrix based on the assumption that some of the responders were actually influenced by the campaign while others were either existing subscribers on different products (terms of subscription) or those subscribed due to factors other than the campaign itself (e.g., news and special events, and recent moving, etc.). Our evaluation results show that cost-sensitive learning using the selected three classification methods can achieve both higher expected response rates and higher expected profits than the current practice (i.e., the random direct mail campaign to target groups that are drawn from commercial databases provided by third-party vendors).

The rest of the paper is organized as follows. First, we provide an overview of CRM and data mining methods, as well as the business problem studied in this paper. In addition, cost-sensitive classification for unbalanced datasets is also discussed briefly in next section. The following section presents the empirical evaluation, which includes an introduction of the background context, a description of the data set as well as the model building procedures. Results are then reported and discussed. Finally, we present the conclusions and a brief discussion of future work.

DATA MINING FOR CRM

CRM is "an enterprise approach to understanding and influencing customer behavior through meaningful communications in order to improve customer acquisition, customer retention, customer loyalty, and customer profitability" (Swift 2001, p12). Data mining, a core technology in CRM, is a process "to identify valid, novel, potentially useful, and understandable correlations and patterns in existing data" (Chung and Gray 1999). It is a multi-step process, consisting of data preprocessing, data analysis, and interpretation of results. In general, data mining techniques have been applied in CRM to solve business problems such as customer acquisition, customer retention, customer segmentation, channel optimization, fraud prevention, and market-basket analysis (Apte et al. 2002; Berson et al. 2000; Swift 2001).

Classification Methods in CRM

Classification, or supervised learning, constructs a hypothetical mapping from a set of attributes (i.e., independent variables) describing instances in a problem domain to a set of predefined classes (i.e., dependent variable) based on a training data set (Witten and Frank 2000). A variety of classification methods have been developed in such fields as statistical pattern recognition, machine learning, and artificial neural nets (Berson et al 2000; Groth 2000; Johnson and Wichern 2002; Swift 2001). In our empirical evaluation, we have chosen three widely used classification methods, i.e., naive Bayes, logistic regression, and decision tree.

Naive Bayes is a special case of the Bayes classification method; it assumes that the attributes are conditional independent given the class membership (Witten and Frank 2000). The method stores a simple probabilistic summary for each class; this summary contains the conditional probability of each attribute value given the class, as well as the prior probability (or base rate) of the class. Each time the algorithm encounters a new instance, it updates the probability stored with the specified class. Neither the order of training instances nor the occurrence of classification errors has any effect on this process.

Logistic regression (Hosmer and Lemeshow 2000) assumes a specific curvilinear relationship between independent variables and the probability of a class outcome. A nonlinear logistic regression function has the form of $\pi(x) = \exp(\alpha + \beta x)/(1 + \exp(\alpha + \beta x))$, where x is a vector of independent variables and $\pi(x)$ is the probability that a given instance belongs to the positive class. The parameters are estimated using the maximum-likelihood method.

Most decision tree methods use a non-backtracking heuristic splitting procedure that recursively partitions a set of training observations into disjointed subsets. The major difference between different methods is in the heuristic measure of goodness used to select the splitting attribute at a tree node. For example, the popular C4.5 uses information gain (also known as mutual information and entropy reduction) and gain ratio to choose a single attribute at a splitting node (Quinlan 1993).

Campaign Management and Cost-sensitive Classification

Customer acquisition is the primary means of growth for many businesses, including the newspaper industry. Identifying and attracting new subscribers is of paramount importance to newspaper companies, considering the significant challenges from decreasing circulation, and an increasing number of alternative sources such as digital media. In practice, new customer acquisition is often complemented with other programs such as customer loyalty and customer retention. However, the focus of this paper is on marketing campaigns for new customer acquisition. In these marketing campaigns, a typical question posed by a newspaper company is: Which prospects are most likely to respond to a particular offer? Accordingly, the response behavior could be analyzed as a simple binary response or categorical response behavior. In this paper, the response behavior is categorized into two classes: responders and non-responders. Responders are those who have purchased subscription following a direct marketing campaign. According to domain experts, the response rates of such “cold contacts” are generally low, often in the range of 1-5%. Hence, the data are heavily unbalanced with non-responders being the predominant group. Further, the misclassification costs for the two classes are very different. A false positive classification, when the direct campaign mail is directed to someone who is not interested in subscription to the promoted product, incurs additional mailing and handling costs. On the other hand, a false negative prediction, when a potential customer could not be reached, leads to loss of business with this valuable customer.

Under such conditions, traditional learning algorithms may not produce satisfactory classifiers as they generally assume balanced class distributions and uniform misclassification costs (Japkowicz 2000; Provost and Fawcett 1997; Provost and Fawcett 1998; Provost and Fawcett 2000). This is especially true when there are very few cases belonging to a particular class, as is the case in this paper. In such situations, the lack of representation of the minority class would produce unsatisfactory classifiers and may even provide misleading results if the classifiers are not adjusted (Witten and Frank 2000). In practice, the commonly used techniques for dealing with unbalanced data and unequal misclassification costs are based on artificial rebalancing, either through "upsampling" (i.e., replicating cases from the minority class) or "downsampling" (i.e., ignoring cases from the majority class) (Japkowicz 2000; Weiss and Provost 2003). More efficiently, a cost matrix can be incorporated into the learning algorithms to weight the instances accordingly, if the misclassification costs and class distributions are known (Elkan 2001; Witten and Frank 2000). In this paper, we use cost-sensitive classification methods that can weight the instances according to the cost matrix obtained from domain experts.

EMPIRICAL EVALUATION

The newspaper company studied in this work is a dominant player in the news market of a major midwestern city in the USA. Recently, it launched several test campaigns to attract new customers. The campaigns were designed to target non-subscribers, as well as casual readers, who purchase newspapers at retail stores. As in typical “cold contact” campaigns, the response rates were generally low (close to 1%). In order to improve response rates and increase profits for future campaigns, it is critical to understand the patterns of response behavior through campaign analysis. Therefore, we collaborated with the director of the campaigns to explore the response behavior patterns using data mining techniques and to evaluate the business value of data mining in this campaign management practice. Over a period of three months, we conducted unstructured interviews as well as used email and phone call for daily contacts with the director, in order to gain a precise understanding of the practices in general and the data set in particular.

Data Description

The original data set, collected from multiple campaigns, was stored in an Access database containing 13 tables (with hundreds of attributes). Currently, the newspaper company selected the target campaign groups from commercial databases provided by third-party vendors. Specific constraints were imposed with respect to the corresponding purpose of each campaign based largely on intuitive selections of demographics attributes. For example, in a campaign on a weekend special edition, partnering with a gardens & gifts company, home ownership, age and other relevant demographics attributes were used as the primary selection criteria for potential subscribers. In this paper, we pooled all the data from these campaigns together for our analysis.

For the sake of learning efficiency, we randomly sampled about 9,943 cases (representing a 0.68% sampling ratio) from the majority class (i.e., non-responders). Similarly, after the sampling, the minority class had 161 responders (representing a 63.4% sampling ratio). The sampling ratios used for both the majority and the minority group were largely constrained by the “noise” in the original data set (e.g., incomplete records, missing values etc.). The class distribution of the resulted training data was highly skewed, with a total of 10,104 records consisting of 9,943 non-responders (98.4%) and 161 responders (1.6%).

After consulting with the domain experts, we identified and deleted some irrelevant attributes. 109 attributes were included in the final training data set, covering a variety of information about prospective customers, including:

- Demographic variables – age, gender, income, marital status, education, family, etc.

- Psychographics variables – lifestyle, travel, reading interests, exercise, etc.
- Behavioral variables – mail order, travel, sports, computer usage, etc.

All 109 attributes were recoded. In particular, variables such as age, length of residence, income, etc. were treated as continuous variables, while marital status, occupation, and all the psychographics variables were recoded as categorical variables. This resulted in a data set containing 7 continuous variables and 102 categorical variables. The target variable was coded as 1 and 0, representing responders and non-responders, respectively. The data file was then transformed into the format required by the Weka data-mining package (Witten and Frank 2000).

Using expert knowledge, the average cost of handling each contact and the average revenue of each subscription from the campaigns were estimated. A conservative estimation of the value of a two-year subscription by a responder was used as the basis for the average revenue of each responder. Then, adjusted by the sampling ratios, the average cost of handling each contact and the average revenue of each subscription were estimated at $1.55 and $268.93, respectively. However, among the responders, only some are "true new subscribers", who are actually influenced by the campaigns; while others would subscribe anyway, with or without the campaigns. Let p denote the percentage of responders who subscribed because of the campaigns. In practice, the actual value of p for a particular campaign is often not well understood. Usually, it is assumed that all the responders are new subscribers in analyzing and evaluating a marketing campaign. Consulting with domain experts, we estimated that p would be in the range of 30% to 50%. In fact, a wider ratio range of p was experimented in additional analyses. The response rates appeared to be relatively insensitive to the different p ratios. Conversely, when the ratio was greater than 60%, data mining techniques did not show significant improvement in terms of profits than the current practice. Hence, in this paper, we tested the classification methods under a series of settings: p = 30%, 35%, 40%, 45%, and 50%. The expected response rates and expected profits due to the campaigns were estimated and compared.

Model Building

In this study, we employed three classification methods, including naive Bayes, logistic regression, and J4.8 decision tree (Weka's implementation of C4.5) available in the Weka data mining package (Witten and Frank 2000). First, we derived a set of cost matrices based on the estimated cost/revenue information by domain experts (i.e., corresponding to the false positive and false negative values in the confusion matrices) and the settings of p. Then, we ran the three classification algorithms using each cost matrix corresponding to a specific p value. The training instances were weighted based on the values in the cost matrices. In addition, we used ten-fold cross-validation, a widely recommended estimation method (Witten and Frank 2000), to estimate the performance of the learned classifiers.

RESULTS

The cross-validated performance of a learned classifier can be described using a confusion matrix (see Table 1). The confusion matrices of the classifiers learned by logistic regression, naive Bayes, and J4.8 decision tree under the various settings of p are summarized in Table 2. The estimated response rates are plotted in Figure 1, and the estimated profits due to campaigns are plotted in Figure 2.

		Predicted Class	
		0 (Non-resp)	1 (Resp)
Actual Class	0 (Non-resp)	True Negative (TN)	False Positive (FP)
	1 (Resp)	False Negative (FN)	True Positive (TP)

Table 1. Confusion Matrix of a Learned Classifier.

Classification Method	p(%)									
	30		35		40		45		50	
Logistic Regression	8561	1382	8328	1615	8134	1809	7923	2020	7730	2213
	95	66	92	69	85	76	84	77	81	80
Naïve Bayes	7430	2513	7245	2698	7073	2870	6896	3047	6768	3175
	76	85	74	87	70	91	69	92	69	92
J4.8Tree	9659	284	9613	330	9582	361	9580	363	9532	411
	93	68	95	66	96	65	95	66	89	72

Table 2. Confusion Matrices of Learned Classifiers.

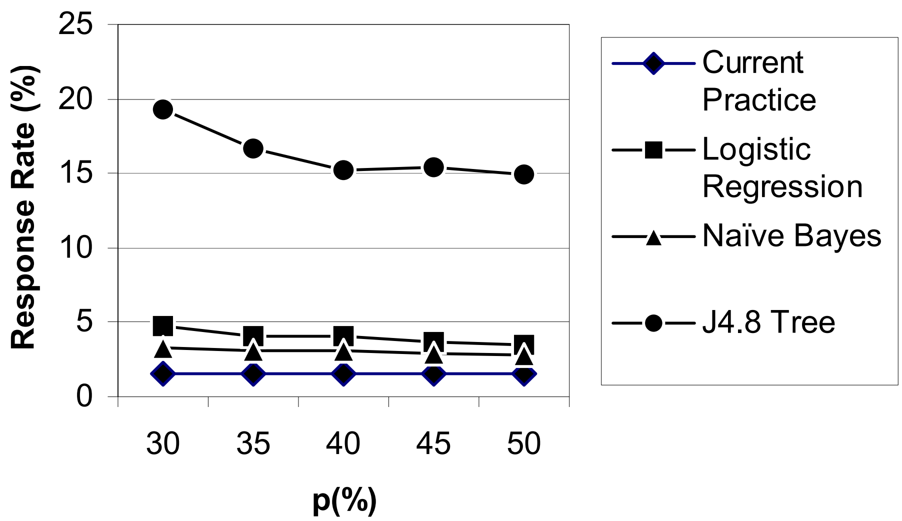

Figure 1: A Comparison of Estimated Response Rates Provided by Different Learned Classifiers.

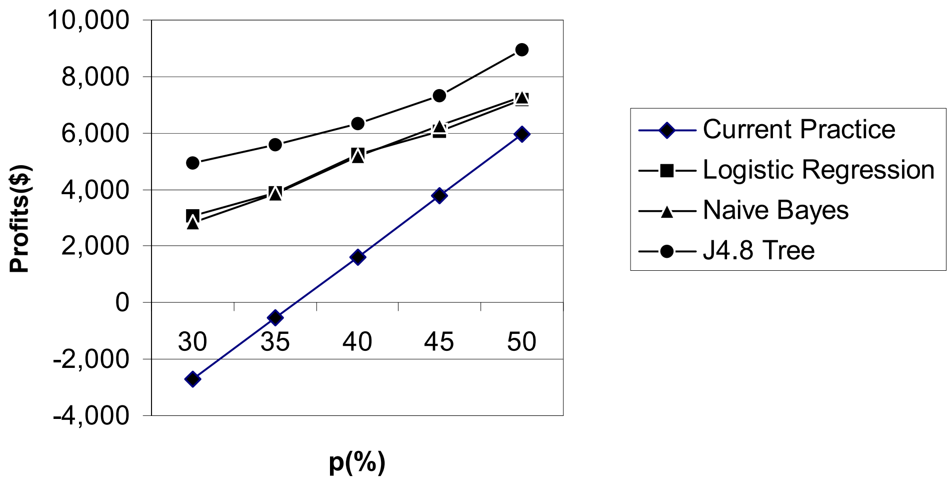

Figure 2. A Comparison of Estimated Profits Provided by Different Learned Classifiers.

The results show that under the various settings of *p*, the classifiers learned by the three classification methods outperform the current practice of random direct campaign with respect to both response rates and estimated profits due to the campaigns. When *p* is below 35%, the current practice (i.e., sending mails to every person) even loses money. In such situations, it becomes more important to apply data mining techniques in planning a direct marketing campaign in new customer acquisition. In addition, J4.8 appears to generate better results than logistic regression and naive Bayes for this particular application.

The findings have several implications for planning and managing marketing campaign. First, it is essential to recognize the nature of a direct marketing campaign, which typically involves extremely unbalanced data and unequal misclassification costs. An appropriate cost matrix should be employed in planning a prospective campaign with data mining classifiers in customer acquisition. In general, an accurate estimation of the misclassification costs of the two predicted classes would be difficult to obtain. Nonetheless, domain knowledge from experts and information of past campaigns would be incorporated as is shown in this study. Secondly, a better understanding of the underlying characteristics of the potential responders is important in planning and managing an effective campaign. As demonstrated in this study, the profits gains are significantly greater using data mining classifiers than the current practice when the *p* ratios are relatively low. However, the gaps in profit gains become narrower as the *p* ratios are increasing. A pilot test with a small sample of the target groups of a prospective campaign could be conducted to obtain information about the possible range of the *p* ratios. Furthermore, as the cost-sensitive classifiers employ the cost matrices that are derived from the various *p* ratios, the results also reflect the effects of misclassification costs. In other words, when the misclassification costs are moderately unequal as is the case in this paper, applying data mining techniques (e.g., cost-sensitive classifiers) in campaign management could contribute to higher response rates and to reap significant profits.

CONCLUSIONS AND FUTURE WORK

Customer acquisition using direct mail is a critical process in CRM. However, in most cases, unbalanced data, a result of the typically low response rates, as well as asymmetric misclassification costs, pose challenges to effective application of data mining techniques. In this paper, we have evaluated three cost-sensitive classification methods in analyzing the campaign data from a large newspaper company.

The use of data mining methods has demonstrated higher response rates and higher profit gains than the current practices. The data set in the newspaper industry may possess unique characteristics that could limit the generalizability of the results to other business domains. Nonetheless, because unbalanced data and asymmetric misclassification costs are common challenges to direct marketing campaigns in other industries, our research provides useful insights to practitioners in campaign management and CRM. Likewise, the empirical evaluation and findings would shed light on research in studying cost-sensitive learning with unbalanced data and unequal costs.

Our current study can be further extended in several directions. First, the research issue warrants investigations in other business domains. Second, J4.8 Decision Tree appears to outperform logistic regression and naive Bayes in this particular application. However, more elaborate evaluation is needed to reliably compare the performance of different methods. For example, statistical evaluations of the classification methods could be conducted. Finally, empirical testing of ensemble classifiers, such as bagging, boosting, and stacking, in such extremely unbalanced situations, is worth investigating.

REFERENCES:

1. Apte, C., B. Liu, E. Pednault, and P. Smyth (2002). "Business applications of data mining." *Communications of the ACM* 45(8): 49-53.
2. Berson, A., S. Smith, and K. Thearling(2000). Building Data Mining Applications for CRM. McGraw-Hill, New York.
3. Brachman, R., T. Khavaza, W. Kloesgen, G. Piatetsky-Shapiro, and E. Simoudis (1996). "Mining business databases." *Coummuications of the ACM* 39(11): 42-48.
4. Chung, H. M. and P. Gray (1999). "Special Section: data mining." *Journal of Management Information Systems* 16(1): 11-16.
5. Elkan, C. (2001). "The foundations of cost-sensitive learning." In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI'01)*: 973-978.
6. Groth, R. 2000. Data Mining: Building Competitive Advantage. Prentice Hall, New Jersey.
7. Hosmer, D.W., and S. Lemeshow (2000). Applied Logistic Regression, second edition. John Wiley & Sons, Inc.
8. Japkowicz, N. (2000). "Learning from imbalanced data set: a comparison of various strategies." In Papers from the AAAI Workshop on Learning from Imbalanced Data Set. Tech. Rep. WS-00-05, Menlo Park, CA: AAAI Press.
9. Johnson, R. A. and D. W. Wichern. 2002. Applied Multivariate Statistical Analysis. Fifth Edition. Prentice Hall, New Jersey.
10. Provost, F. and T. Fawcett (1997). "Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions." In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*.
11. Provost, F. and T. Fawcett (1998). "Robust Classification Systems for Imprecise Environments." In *Proceedings of theFifteenth National Conference on Artificial Intelligence*.
12. Provost, F. and T. Fawcett (2000). "Robust Classification for Imprecise Environments." *Machine Learning* 42: 203-231.
13. Quinlan, J.R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann.
14. Salzberg, S. (1999). "On comparing classifiers: a critique of current research and methods." *Data Mining and Knowledge Discovery* (1): 1-12.
15. Swift, R. S. (2001). Accelerating Customer Relationships: using CRM and Relationship Technologies.Prentice Hall, New Jersey.
16. Weiss, G. and F. Provost (2003). "Learning when Training Data are Costly: The Effect of Class Distribution on Tree Induction." to appear, *Journal of Artificial Intelligence Research*.
17. Winer, R. S. (2001). "A framework for customer relationship management." *California Management Review* 43(4): 89-105.
18. Witten, I. and E. Frank (2000). Data Mining: practical machine learning tools and techniques with JAVA implementations. Morgan Kaufmann Publishers, San Francisco.