A Fully Distributed Framework for Cost-sensitive Data Mining

Wei Fan, Haixun Wang, Philip S. Yu
IBM T.J.Watson Research Center
30 Saw Mill River Road, Hawthorne, NY 10532
{weifan, haixun, psyu}@us.ibm.com

Salvatore J. Stolfo
Department of Computer Science
Columbia University, New York, NY 10027
sal@cs.columbia.edu

Abstract

In this paper, we propose a fully distributed system (as compared to centralized and partially distributed systems) for cost-sensitive data mining. Experimental results have shown that this approach achieves higher accuracy than both the centralized and partially distributed learning methods, however, it incurs much less training time, neither communication nor computation overhead.

1. Introduction

During the last two decades, our ability to collect and store data has significantly out-paced our ability to extract "knowledge". *Data Mining* is the process of identifying valid patterns. In a relational database context, a typical task is to explain and predict the value of some attribute given a collection of tuples with known attribute values. One of the main challenges in machine learning and data mining is the development of inductive learning techniques that scale up to large and possibly physically distributed datasets. Many organizations seeking added values from their data are already dealing with overwhelming amounts of information. On the other hand, in many areas of application where different examples carry different benefits, it is not enough to maximize the accuracy based on the simplified *cost-insensitive* assumption that each example has the same benefit and there is no penalty for misclassification. For example, credit card fraud detection seeks to detect frauds with high transaction amount. In this paper, we are interested in studying accurate and efficient frameworks for distributed cost-sensitive learning.

2. Cost-sensitive Learning

Suppose that the cost to investigate a fraud for a credit card transaction x is \$90 and the amount of transaction for x is $Y(x)$. In this case, the *optimal decision-making policy* is to predict x being fraud if and only if $(R(x) = P(x) \cdot Y(x)) > 90$, where $P(x)$ is the estimated probability that x is a fraud. $R(x)$ is called the *risk* to solicit x. This policy has an error-tolerance property in which the exact value of $P(x)$ is not important as long as it does not switch from above to below (or vice versa) a *decision threshold*, $T(x)$. For credit card fraud detection, $T(x) = \frac{90}{Y(x)}$. Re-writing the optimal decision policy using decision threshold, if and only if $P(x) > T(x)$, the optimal decision is to predict fraud; otherwise, the decision is non-fraud. This property makes probability estimate resilient to small errors.

$P(x)$ is easy to estimate. For decision trees, suppose that t is the number of examples and p is the number of positive examples associated with a node, then $P(x) = \frac{p}{t}$. For some problems, such as charity donation, the benefit (donation amount) is not known *a priori*. We employ the multiple linear regression method to estimate the benefit.

3. Fully-distributed Framework

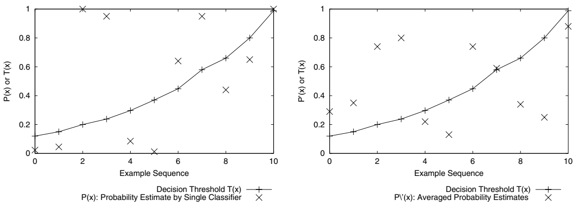
Assume that there are k participating distributed sites and the data subset at each site is denoted as S_i, and a classifier C_i is trained from S_i. The classifier outputs a class label C_i and a member probability $P_i(x)$ ($\in [0, 1]$) for each testing example x. For problems for which the benefit of x is not known in advance, a separate dataset is used to compute a model to estimate this benefit. One straightforward approach is to simply average individual risks to compute combined risk, $\bar{R}(x) = \frac{\sum R_i(x)}{k} = \frac{\sum P_i(x) \cdot Y_i(x)}{k}$. More sophisticated methods are discussed in [2], however these approaches don't necessarily bring higher accuracy.

Desiderata The fully distributed learning is very likely to have higher benefits than the "global" classifier (or "centralized learning") because of its "smoothing effect" as shown in the *cost-sensitive decision plot* in Figure 1. For each data point x, we plot its decision threshold $T(x)$ and probability estimate $P(x)$ in the same figure. The sequence on the x-axis is ordered increasingly by their $T(x)$ values.

Proceedings of the 22 nd International Conference on Distributed Computing Systems (ICDCS'02)
1063-6927/02 \$17.00 © 2002 **IEEE**

IEEE
COMPUTER
SOCIETY

Figure 1. Cost-sensitive decision plots



The left plot is conjectured for global classifier, while the right plot is conjectured for averaged probability of multiple classifiers. All data points above the $T(x)$ line (with $P(x) > T(x)$) are predicted positive. Since probability estimates by multiple classifiers are uncorrelated, it is very unlikely for all of them to be close to either 1 or 0 and their resultant average will likely spread more "evenly" between 1 and 0. This is visually illustrated in Figure 1 by comparing the right plot to the left plot. The smoothing effect favors more towards predicting expensive examples to be positive. $T(x)$ of expensive examples are low; these examples are in the left portion of the decision plots. If the estimated probability by global classifier $P(x)$ is close to 0, it is very likely for the averaged probability $P'(x)$ to be bigger than $P(x)$, and consequently bigger than $T(x)$ of expensive examples and predict them to be positive. The two expensive data points in the bottom left corner of the decision plots are predicted to be negative by the global classifier, however predicted to be positive by the multiple model. Due to the smoothing effect, averaging of multiple probabilities biases more towards expensive examples than the global classifier. This is a desirable property since expensive examples contribute greatly towards total benefit.

Overhead Analysis The fully distributed framework does not incur any additional computation or communication overhead. As a comparative study, partially distributed learning refers to "meta-learning" [1] and centralized learning refers to bringing the data from every participating sites to a single site to train a global classifier; both methods incur additional communication and computation overhead.

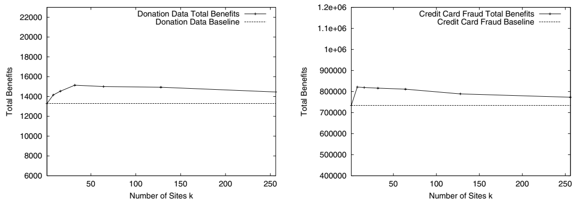
4. Experiments

We have evaluated the distributed learning system on two important types of cost-sensitive problems. In the first problem, neither the probability $P(x)$ nor the benefit $y(x)$ is known. Additionally, only positive example carries profit and there is a penalty for false positives. We use the donation dataset that first appeared in KDD'98 competition. In the second problem, the benefit is known and it is *per instance*. Only positives carry benefit and there is also a penalty for false positives. We use a credit card fraud detection dataset.

Dataset	Total Benefits
Donation	13292.7
Credit Card Fraud	733980

Table 1. Centralized Results (global classifier)

Figure 2. Fully-distributed framework results



Experimental Result The results by the centralized learning (or global classifier) are shown in Table 1, which serve as the baseline to evaluate distributed methods. In Figure 2, we plot changes of total benefits with growing number of sites k using the fully distributed approach. We can clearly see the total benefits are all significantly higher than the baseline for all chosen k for both the donation and credit card fraud datasets. We have also compared with a partially-distributed system using meta-learning. Our experiments have shown that the total benefits by meta-learning are significantly less than the fully-distributed framework.

5. Conclusion

We have proposed a fully distributed framework for cost-sensitive learning using simple averaging. We analyzed the reasons why averaging will also improve accuracy. As expected. experimental results have shown that the accuracy is as good as or even better than the global classifier trained on the all available data from every site, while the fully distributed framework doesn't incur either communication nor computation overhead.

References

[1] P. Chan. *An Extensible Meta-learning Approach for Scalable and Accurate Inductive Learning*. PhD thesis, Columbia University, Oct 1996.

[2] W. Fan, H. Wang, P. S. Yu, and S. Stolfo. A framework for scalable cost-sensitive learning based on combining probabilities and benefits. In *Second SIAM International Conference on Data Mining (SDM2002)*, April 2002.

Proceedings of the 22 nd International Conference on Distributed Computing Systems (ICDCS'02)
1063-6927/02 $17.00 © 2002 **IEEE**