Predictive data mining for delinquency modeling

Bharatheesh T.L.
Bigants Consulting
Bangalore, INDIA

Iyengar S.S.
Distinguished Professor
Department of Computer Science
Louisiana State University
Baton Rouge, Louisiana 70803, USA

Predictive data mining is the process of automatically creating a classification model from a set of examples, called the training set, which belongs to a set of classes. Once a model is created, it can be used to automatically predict the class of other unclassified examples. Some datasets encountered in real life applications have skewed class distributions. Many predictive modeling systems are not prepared to induce a classifier that accurately classifies the minority class under such situation. In this work, an attempt has been made to build the predictive model for delinquency in credit cards users, using the state of art methods. The success of the model is defined in different terms than the ones found in literature. Different sampling schemes are evaluated and a modified naïve Bayes classifier is used as classifier. The results are encouraging and it is proposed to compare the prototype with ensemble of models.

Key words: Data mining, Predictive modeling, Delinquency modeling, skewed distributions, naïve Bayes classifier.

1. Introduction

Delinquency Analysis refers to the process in which financial institutions deploy to measure the propensity of customers to become defaulters i.e. to repay the credit extended to them. This is a significant problem in credit card business, where a customer uses the credit card and end up being the defaulter. It is in the interest of the institution to identify the likely defaulters at the earliest and take necessary action to prevent further loss of revenue. In general, the model which is used to measure the propensity of default of an individual customer is called *delinquency model* and the decision support system built on such models are

called *Early Warning Systems (EWS).* One of the important characteristics of such an Early Warning System is to minimize the misclassification of true customers being classified as defaulters and maximizes the predicted defaulters. In implementing the process for this model several other problems were encountered. This paper outlines the implemented solution and the scope for future research

2. Predictive data mining

Predictive data mining is the process of automatically creating a classification model from a set of examples, called the training set, which belongs to a set of classes. Once a model is created, it can be used to automatically predict the class of other unclassified examples.

In other words, in predictive data mining, a set of 'n' training examples is given to an inducer. Each example

Let $X \subset D_1 \times D_2 \times \ldots \times D_m$ where D_j is the domain of the jth feature.

$Y \subset D_o$, Where D_o is the domain of the class attribute which is discrete

The training examples are tuples (X, Y) where Y is the label, output or class. Given a set of training examples, the data mining algorithm or inducer outputs a classifier model such that, given a new example, it accurately predicts the label Y.

For a number of application domains, a huge disproportion in the number of cases belonging to each class is common. For instance, in detection of fraud in telephone calls [1] and credit card transactions [4], the number of legitimate transactions is much higher than the number of fraudulent transactions. In insurance risk modeling [3], only a small percentage of the policyholders file one or more claims in any given time period. Also, in direct marketing [2], it is common to have a small response rate (about 1%) for most marketing campaigns.

Thus, building predictive models with skewed class distributions is an important issue in data mining. Many data mining systems are not prepared to induce a classifier that accurately classifies the minority class under such situation. Frequently, the classifier has good classification accuracy for the majority class, but its accuracy for the minority class is unacceptable. Unfortunately, that is the norm for most applications with imbalanced data sets, since these applications aim to profile a small set of valuable entities that are spread in a large group of "uninteresting" entities.

Credit card delinquency modeling is one such domain where the number of customers who are non-delinquents is far greater than the percentage of delinquents. Hence any predictive model built on the raw data set is bound to produce an inferior model in identifying the delinquents from the unseen transactions.

In this paper we discuss some of the methods we used to solve the problem of building predictive models for delinquency detection with imbalanced data distributions. Later we discuss the best model under the given circumstances and its performance against state of art algorithms.

This rest of the paper is organized as follows: Section 2 discusses various metrics to measure the performance of predictive models data have class imbalance; Section 3 explains the approach used by the authors; Section 4 discusses results of the research and conclusions are given in section 5.

2. Metrics for measuring predictive model performance

The different types of errors and hits performed by a classifier can be summarized in a confusion matrix [5]. Table 1 illustrates a confusion matrix for a two-class problem, with classes labeled positive and negative:

Table 1. Different types of errors and hits for a two classes problem.

Actual ▼ Predicted ▸	***Positive***	***Negative***
Positive	True positive (a)	False positive (b)
Negative	False negative (c)	True negative (d)

From such matrix it is possible to extract a number of metrics to measure the performance of learning systems, such as

Error rate (E) = (c+b) /(a+b+c+d)

Accuracy (Acc) = (a+d) /(a+b+c+d) = 1 - E.

The error rate (E) and the accuracy (Acc) are widely used metrics for measuring the performance of learning systems [6]. However, when the prior probabilities of the classes are very different, such metrics might be misleading. For instance, it is straightforward to create a classifier having 99% accuracy (or 1% error rate) if the data set has a majority class with 99% of the total number of cases, by simply labeling every new case as belonging to the majority class. These are called as stupid classifiers, which are of no interest to the business community [7].

In [5], the authors have proposed the following metrics using Table 1.

– False negative rate: FN = b / (a+b) is the percentage of positive cases misclassified as belonging to the negative class;

– False positive rate: FP = c / (c+d) is the percentage of negative cases misclassified as belonging to the positive class;

– True negative rate: TN = d / (c+d) = 1 - FP is the percentage of negative cases correctly classified as belonging to the negative class;

– True positive rate: TP = a / (a+b) = 1 - FN is the percentage of positive cases correctly classified as belonging to the positive class;

These four class performance measures have the advantage of being independent of class costs and prior probabilities. It is obvious that the main objective of a classifier is to minimize the false positive and negative rates or, similarly, to maximize the true negative and positive rates. Unfortunately, for most “real world” applications, there is a tradeoff between FN and FP and, similarly, between TN and TP.

When Table 1 is applied to the case under study it will be transformed to as shown in Table 2.

Table 2. Different types of errors and hits for delinquency modeling.

Actual ▼ Predicted ▸	Delinquent	Good
Delinquent	True positive (a)	False positive (b)
Good	False negative (c)	True negative (d)

3. Delinquency modeling using a modified naïve Bayes classifier

The data set used in the case study was from a proprietary source, which cannot be disclosed for the reasons of confidentiality. The original data set had about 45 fields about customer demographics, transactions and payments. The preprocessed data after feature selection [Modified naïve Bayes] has resulted in a compact data set of only 13 variables. The original data contained about 50,000 records of which 6% were identified as delinquents. The problem was to build a classifier from this skewed data, which maximizes the true positives and minimizes false negatives. The results of the predictive data mining experiments on this data set are discussed in the next section.

In order to build the delinquency model the following approaches were tried.

1. Assign misclassification costs. In a general way, misclassified examples of the minority class are more costly than misclassified examples of the majority class.
2. Under-sampling. One very direct way to solve the problem of modeling from imbalanced data sets is to artificially balance the class distributions. Under-sampling aim to balance a data set by eliminating examples of the majority class;
3. Over-sampling. This method is similar to under-sampling. But it aims to achieve more balanced class distributions by replicating examples of the minority class.

A modified naïve Bayes classifier was chosen because of its simplicity and comprehensibility [7]. Other state of art methods like k-nn, C4.5 and logistic regression [6] were also tested for the same data. The results of the experiment are discussed in the next section.

4. Results

In this study an analysis has been done to identify, the class distributions best for training a delinquency prediction inducer. Six combinations of class distributions from 35% minority to 65% minority at 5% steps are used in the study. The proportions studied were – 35% delinquents and 65% normal customers, 40% delinquents and 60% normal customers, 45% delinquents and 55% normal customers, 50% delinquents and 50% normal customers, 55% delinquents and 45% normal customers, 60% delinquents and 4% normal customers and 65% delinquents and 35% normal customers. Error rate, Accuracy of classification and percentage of true positives were used as metrics for sample size selection as well as selection of classifiers. Using these metrics for effectiveness of class distribution it was observed that the optimal distribution generally contains between 50% and 60% of minority class examples. It has also been observed that under sampling of the majority class seems to be better than other methods and will generally lead to results which are no worse than, and often superior to, those which use the natural class distributions. The results of the experiment are shown in Table 3.

Table 3. Results of the predictive model performance for different samples

	50.00%	***55.00%***	***60.00%***	***65.00%***
Error rate	30%	47%	34.96%	35%
Accuracy	70%	53%	65%	65%
True positives	40%	23%	34%	38%

Even though no explicit explanation was possible for why 50% balanced data set was optimal, it was observed that the noisy patterns were minimum at this proportion.

Table4. Results of the predictive model performance for different classifiers

	MNB	***k-nn***	***C 4.5***	***Logistic regression***
Error rate	30%	41%	31%	29%
Accuracy	70%	59%	69%	71%
True positives	40%	22%	35%	37%

It can also be observed that a generative method like modified naïve Bayes performs surprisingly well compared to discriminative methods like decision trees. The success of modified naïve Bayes can be attributed to the mild relaxation of its attribute independence assumption using feature bundling [7].

Even though logistic regression has the highest accuracy rate, modified naïve Bayes has better true positive detection capabilities. The true positive rate has a cost implication, which is directly measurable as well as minimizes the customer dissatisfaction, which is a secondary objective of the bank [7]. Finally it was decided to use the early warning system based on modified naïve Bayes classifier.

5.Conclusions

Delinquency modeling is one of the potential cases for data mining due to its necessity in a credit card industry as well as the complexity of the problem due to skewness in the data. In this research work an attempt has been made to test the effective sample size and classifier algorithm for the delinquency prediction modeling. It has been observed that a 50% balanced sampling with a modified naïve Bayes classifier is a good choice. Further research is in progress to enhance the effectiveness of the classifier, using ensemble of models.

This research is partially funded by National Science Foundation grant (2003 on data mining) for Prof. S.S.Iyengar.

References

1. Tom Fawcett and Foster J. Provost, “Adaptive Fraud Detection”, Data Mining and Knowledge Discovery, 1(3), 1997.
2. Charles X. Ling and Chenghui Li.,” Data Mining for Direct Mining: Problems and Solutions”, In Proceedings of The Forth International Conference on Knowledge Discovery and Data Mining, 1998.
3. Edwin P. D. Pednault, Barry K. Rosen, and Apte C., “Handling Imbalanced Data Sets in Insurance Risk Modeling”, Technical Report RC-21731, IBM Research Report, March 2000.
4. S. J. Stolfo, D. W. Fan, W. Lee, A. L. Prodromidis, and P. K. Chan. Credit Card Fraud Detection Using Meta-Learning: Issues and Initial Results. In AAAI-97 Workshop on AI Methods in Fraud and Risk Management, 1997.
5. Learning with Skewed Class Distributions, Maria Carolina Monard and Gustavo E.A.P.A. Batista, CADERNOS DE COMPUTAC¸ ˜AO XX, 2003
6. Ian H. Witten, Eibe Frank, “Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations”, Morgan Kaufmann, 1999
7. Bharatheesh T.L, “Predictive data mining using a modified naïve Bayes classifier”, Unpublished M.Phil. thesis, MS University, 2003.