Data Set Balancing

David L. Olson[1]

[1] University of Nebraska, Department of Management, Lincoln, NE 68588-0491 USA
dolson3@unl.edu
ait.unl.edu/dolson

Abstract. This paper conducts experiments with three skewed data sets, seeking to demonstrate problems when skewed data is used, and identifying counter problems when data is balanced. The basic data mining algorithms of decision tree, regression-based, and neural network models are considered, using both categorical and continuous data. Two of the data sets have binary outcomes, while the third has a set of four possible outcomes. Key findings are that when the data is highly unbalanced, algorithms tend to degenerate by assigning all cases to the most common out come. When data is balanced, accuracy rates tend to decline. If data is balanced, that reduces the training set size, and can lead to the degeneracy of model failure through omission of cases encountered in the test set. Decision tree algorithms were found to be the most robust with respect to the degree of balancing applied.

1 Introduction

Data mining technology is used increasingly by many companies to analyze large databases in order to discover previously unknown and actionable information that is then used to make crucial business decisions. This is the basis for the term "knowledge discovery". Data mining can be performed through a number of techniques, such as association, classification, clustering, prediction, and sequential patterns. Data mining algorithms are implemented from various fields such as statistics, decision trees, neural networks, fuzzy logic and linear programming. There are many data mining software product suites, to include Enterprise Miner (SAS), Intelligent Miner (IBM), Clementine (SPSS), and Polyanalyst (Megaputer). There are also specialty software products for specific algorithms, such as CART and See5 for decision trees, and other products for various phases of the data mining process.

Data mining has proven valuable in almost every academic discipline. Understanding business application of data mining is necessary to expose business college students to current analytic information technology. Data mining has been instrumental in customer relationship management [1] [2], financial analysis [3], credit card management [4], banking [5], insurance [6], tourism [7], and many other areas of statistical support to business. Business data mining is made possible by the generation of masses of data from computer information systems. Understanding this information generation system and tools available leading to analysis is fundamental for business students in the 21st Century. There are many highly useful applications in practically every field of scientific study. Data mining support is required to make sense of the masses of business data generated by computer technology.

A major problem in many of these applications is that data is often skewed. For instance, insurance companies hope that only a small portion of claims are fraudulent. Physicians hope that only a small portion of tested patients have cancerous tumors. Banks hope that only a small portion of their loans will turn out to have repayment

problems. This paper examines the relative impact of such skewed data sets on common data mining algorithms for two different types of data – categorical and continuous.

2 Data Sets

The paper presents results of experiments on outcome balancing using three simulated data sets representative of common applications of data mining in business. While simulated, these data sets were designed to have realistic correlations across variables. The first model includes loan applicants, the second data set insurance claims, and the third records of job performance.

2.1 Loan Application Data

This data set consists of information on applicants for appliance loans. The full data set involves 650 past observations, of which 400 were used for the full training set, and 250 for testing. Applicant information on age, income, assets, debts, and credit rating (from a credit bureau, with red for bad credit, yellow for some credit problems, and green for clean credit record) is assumed available from loan applications. Variable Want is the amount requested in the appliance loan application. For past observations, variable On-Time is 1 if all payments were received on time, and 0 if not (Late or Default). The majority of past loans were paid on time.

Data was transformed to obtain categorical data for some of the techniques. Age was grouped by less than 30 (young), 60 and over (old), and in between (middle aged). Income was grouped as less than or equal to $30,000 per year and lower (low income), $80,000 per year or more (high income), and average in between. Asset, debt, and loan amount (variable Want) are used by rule to generate categorical variable risk. Risk was categorized as high if debts exceeded assets, as low if assets exceeded the sum of debts plus the borrowing amount requested, and average in between. The categorical data thus consisted of four variables, each with three levels. The continuous data set transformed the original data to a 0-1 scale with 1 representing ideal and 0 the nadir for each variable.

2.2 Insurance Fraud Data

The second data set involves insurance claims. The full data set includes 5000 past claims with known outcomes, of which 4000 were available for training and 1000 reserved for testing. Variables include claimant age, gender, amount of insurance claim, number of traffic tickets currently on record (less than 3 years old), number of prior accident claims of the type insured, and Attorney (if any). Outcome variable Fraud was 0 if fraud was not detected, and 1 if fraud was detected.

The categorical data set was generated by grouping Claimant Age into three levels and Claim amount into three levels. Gender was binary, while number of tickets and prior claims were both integer (from 0 to 3). The Attorney variable was left as five discrete values. Outcome was binary. The continuous data set transformed the original data to a 0-1 scale with 1 representing ideal and 0 the nadir for each variable.

2.3 Job Application Data

The third data set involves 500 past job applicants, of which 250 were used for the full training set and 250 reserved for testing. This data set varies from the first two in that

there are four possible outcomes (unacceptable, minimal, adequate, and excellent, in order of attractiveness).

Some of these variables were quantitative and others are nominal. State, degree, and major were nominal. There is no information content intended by state or major. State was not expected to have a specific order prior to analysis, nor was major. (The analysis may conclude that there is a relationship between state, major, and outcome, however.) Degree was ordinal, in that MS and MBA are higher degrees than BS. However, as with state and major, the analysis may find a reverse relationship with outcome.

The categorical data set was created by generating three age groups, two state outcomes (binary), five degree categories, three majors, and three experience levels. The continuous data set transformed the original data to a 0-1 scale with 1 representing ideal and 0 the nadir for each variable.

3. Experiments

These data sets represent instances where there can be a high degree of imbalance in the data. Data mining was applied for categorical and continuous forms of all three data sets. For categorical data, decision tree models were obtained using See5, logistic regression from Clementine, and Clementine's neural network model applied. For continuous data sets, See5 was used for a regression tree, and Clementine for regression (discriminant analysis) and neural network. In each case, the training data was sorted so that a controlled experiment could be conducted. First, the full model was run. Then the training set was reduced in size by deleting cases with the most common outcome until the desired imbalance was obtained.

The correct classification rate was obtained by dividing the correctly classified test cases by the total number of test cases. This is not the only useful error metric, especially when there is high differential in the cost by error type. However, other error metrics would yield different solutions. Thus for our purposes, correct classification rate serves the purpose of examining the degradation of accuracy expected from reducing the training set in order to balance the data.

3.1 Loan Data Results

The loan application training set included 45 late cases of 400, for a balance proportion of 0.1125 (45/400). Keeping the 45 late cases for all training sets, the training set size was reduced by deleting cases with on-time outcomes, for late-case proportions of 0.15 (300 total), 0.2 (225 total), 0.25 (180 total), and 0.3 (150 total). The correct classification rates and cost results are shown in Tables 1 through 6.

The first test is shown in Table 1, using a decision tree model on categorical data.

Table 1. Categorical Loan Data, Decision Tree

Train	Proportion Late (0)	Predict 0 = 1	Predict 1 = 0	Correct
400	0.1125	20	0	0.920
300	0.1500	20	0	0.920
225	0.2000	9	39	0.808
180	0.2500	9	39	0.808
150	0.3000	9	32	0.836

Using the full training set had a relatively low proportion of late cases (0.1125). This training set yielded a model predicting all cases to be on-time, which was correct in 0.92 of the 250 test cases. As the training set was balanced, the correct classification rate deteriorated, although some cases were assigned to the late category. Note that this trend was not true throughout the experiment, as when the training set was reduced to 150 cases, the correct classification rate actually increased over the results for training set sizes of 180 and 225.

Table 2 shows the results of the logistic regression model on categorical data.

Table 2. Categorical Loan Data, Logistic Regression

Train	Proportion Late (0)	Predict 0 = 1	Predict 1 = 0	Correct
400	0.1125	17	6	0.908
300	0.1500	9	34	0.828
225	0.2000	9	34	0.828
180	0.2500	9	34	0.828
150	0.3000	9	34	0.828

Here the full training set was again best. Balancing the data yielded the same results from then on.

Table 3 shows the results for a neural network model on categorical data.

Table 3. Categorical Loan Data, Neural Network

Train	Proportion Late (0)	Predict 0 = 1	Predict 1 = 0	Correct
400	0.1125	20	0	0.920
300	0.1500	15	17	0.872
225	0.2000	13	26	0.844
180	0.2500	13	26	0.844
150	0.3000	9	43	0.792

The results for this model were consistent with expectations. Reducing the training set to balance the outcomes yielded less and less accurate results.

Tests were also conducted on continuous data with the same three algorithms. Table 4 gives the results for a linear regression model on continuous data. These results were similar to those obtained with categorical data. Here there was an anomaly with the training set of 180 observations, but results were not much different from expectations.

Table 4. Continuous Loan Data, Regression Tree

Train	Proportion Late (0)	Predict 0 = 1	Predict 1 = 0	Correct
400	0.1125	20	0	0.920
300	0.1500	15	15	0.880
225	0.2000	8	46	0.784
180	0.2500	9	41	0.800
150	0.3000	8	46	0.784

Table 5 shows results for a discriminant analysis model applied to the continuous data.

Table 5. Continuous Loan Data, Discriminant Analysis Regression

Train	Proportion Late (0)	Predict 0 = 1	Predict 1 = 0	Correct
400	0.1125	20	0	0.920
300	0.1500	19	1	0.920
225	0.2000	16	7	0.908
180	0.2500	13	20	0.868
150	0.3000	11	28	0.844

These results were slightly better than those obtained for categorical data, exhibiting the expected trend of decreased accuracy with smaller training set.

Table 6 shows the results for the neural network model applied to continuous data.

Table 6. Continuous Loan Data, Neural Network

Train	Proportion Late (0)	Predict 0 = 1	Predict 1 = 0	Correct
400	0.1125	19	2	0.916
300	0.1500	17	10	0.892
225	0.2000	11	28	0.844
180	0.2500	9	33	0.832
150	0.3000	8	46	0.784

The neural network model for continuous data was slightly less accurate than the results obtained from applying a neural network model to categorical data. The trend in accuracy was as expected.

As expected, the full training set yielded the highest correct classification rate, except for two anomalies. Data mining software has the capability of including a cost function that could be used to direct algorithms in the case of decision trees. That was not used in this case, but it is expected to yield parallel results (greater accuracy according to the metric driving the algorithm would be obtained with larger data sets). The best of the six models was the decision tree using categorical data, pruning the training set to only 150 observations.

Continuous data might be expected to provide greater accuracy, as it is more precise than categorical data. However, this was not borne out by the results. Continuous data is more vulnerable to error induced by smaller data sets, which could have been one factor.

3.2 Fraud Data Set

The fraud data set was more severely imbalanced, including only 60 late cases in the full training set of 4000. Training sets of 3000 (0.02 late), 2000 (0.03 late), 1000 (0.06 late), 600 (0.1 late), 300 (0.2 late), and 120 (0.5 late) were generated. Table 7 shows the decision tree model results.

Table 7. Fraud Data Set, Categorical Data, Decision Tree

Train	Proportion Fraud(1)	Predict 0 = 1	Predict 1 = 0	Correct
4000	0.015	22	0	0.978
3000	0.020	22	0	0.978
2000	0.030	22	0	0.978
1000	0.060	17	8	0.975
600	0.100	17	8	0.975
300	0.200	17	8	0.975
120	0.500	17	8	0.975

Only two sets of results were obtained. The outcome based on larger training sets was degenerate – assigning all cases to be OK (not fraudulent). This yielded a very good correct classification rate, as only 22 of 1000 test cases were fraudulent.

Table 8 gives results for the logistic regression model.

Table 8. Fraud Data Set, Categorical Data, Logistic Regression

Train	Proportion Fraud(1)	Predict 0 = 1	Predict 1 = 0	Correct
4000	0.015	20	2	0.978
3000	0.020	19	2	0.979
2000	0.030	19	2	0.979
1000	0.060	17	9	0.974
600	0.100	17	9	0.974
300	0.200	16	34	0.950
120	0.500	11	229	0.760*

* - Model with 120 in training set included 31 null predictions, due to no training case equivalent to test case

Balancing the data from 4000 to 3000 training cases actually yielded an improved correct classification rate. This degenerated when training file size was reduced to 1000, and the model yielded very poor results when the training data set was completely balanced, as only 120 observations were left. For the logistic regression model, this led to a case where the test set contained 31 cases not covered by the training set.

Table 9 shows results for the neural network model applied to categorical data. The neural network model applied to categorical data was quite stable until the last training set where there were only 120 observations. At that point, model accuracy became very bad.

Table 9. Fraud Data Set, Categorical Data, Neural Network

Train	Proportion Fraud(1)	Predict 0 = 1	Predict 1 = 0	Correct
4000	0.015	20	1	0.979
3000	0.020	20	2	0.978
2000	0.030	20	2	0.978
1000	0.060	19	2	0.979
600	0.100	19	2	0.979
300	0.200	17	17	0.966
120	0.500	10	461	0.529

Table 10 displays results for the regression tree applied to continuous data.

Table 10. Fraud Data Set, Continuous Data, Regression Tree

Train	Proportion Fraud(1)	Predict 0 = 1	Predict 1 = 0	Correct
4000	0.015	22	0	0.978
3000	0.020	22	0	0.978
2000	0.030	22	0	0.978
1000	0.060	20	8	0.972
600	0.100	17	18	0.965
300	0.200	17	18	0.965
120	0.500	15	57	0.928

The regression tree for continuous data had results very similar to those of the decision tree applied to categorical data. For the smaller training sets, the continuous data yielded slightly inferior results.

Table 11 gives results for the discriminant analysis model.

Table 11. Fraud Data Set, Continuous Data, Discriminant Analysis Regression

Train	Proportion Fraud(1)	Predict 0 = 1	Predict 1 = 0	Correct
4000	0.015	22	0	0.978
3000	0.020	22	0	0.978
2000	0.030	22	0	0.978
1000	0.060	17	18	0.965
600	0.100	17	18	0.965
300	0.200	17	18	0.965
120	0.500	13	265	0.722

The discriminant analysis model using continuous data had results with fewer anomalies than logistic regression obtained with categorical data, but was slightly less accurate. It also was not very good when based upon the smallest training set. The neural network model based on continuous data was not as good as the neural network model applied to categorical data, except that the degeneration for the training set of 120 was not as severe.

Table 12 shows relative accuracy for the neural network model applied to the continuous data.

Table 12. Fraud Data Set, Continuous Data, Neural Network

Train	Proportion Fraud(1)	Predict 0 = 1	Predict 1 = 0	Correct
4000	0.015	22	0	0.978
3000	0.020	22	0	0.978
2000	0.030	22	1	0.977
1000	0.060	20	9	0.971
600	0.100	19	10	0.971
300	0.200	17	23	0.960
120	0.500	10	334	0.656

Overall, application of models to the highly imbalanced fraud data set behaved as expected for the most part. The best fit was obtained with logistic regression and neural network models applied to categorical data. Almost all of the models over the original data set were degenerate, in that they called all outcomes OK. The exceptions were logistic regression and neural network models over continuous data. The set of runs demonstrated the reverse problem of having too small a data set. The neural network models for both categorical and continuous data had very high error rates for the equally balanced training set, as did the logistic regression model for categorical data. There was a clear degeneration of correct classification rate as the training set was reduced, along with improved cost results, except for these extreme instances.

3.3 Job Applicant Data Results

This data set was far more complex, with correct classification requiring consideration of a four by four outcome matrix. The original data set was small, with only 250 training observations, only 7 of which were excellent (135 were adequate, 79 minimal, and 29 unacceptable). Training sets of 140 (7 excellent, 66 adequate, 38 minimal, and 29 unacceptable), 70 (7 excellent and 21 for each of the other three categories), 35 (7 excellent, 10 adequate, and 9 for the other two categories), and 28 (all categories 7 cases) were generated. Results are shown in Table 13.

Table 13. Job Applicant Data Set, Categorical Data, Decision Tree

	28 training	35 training	70 training	140 training	250 training
Proportion excellent	0.25	0.20	0.10	0.05	0.028
Decision tree	0.580	0.584	0.444	0.508	0.508
Logistic regression	degenerate	degenerate	0.400	0.588	0.608
Neural net – categorical	0.416	0.448	0.392	0.604	.0604
Regression tree	0.484	0.484	0.444	0.556	0.600
Discriminant analysis	0.508	0.544	0.520	0.572	0.604
Neural net - continuous	0.432	0.516	0.496	0.592	0.588

The proportion correct increased as the training set size increased. This was because there were three ways for the forecast to be wrong. A naïve forecast would be expected to be correct 0.25 of the time. The correct classification rate was more erratic in this case. Smaller training sets tended to have lower correct classification rates, but the extreme small size of the smaller sets led to anomalies in results from the decision tree model applied to categorical data. The results from the logistic regression model were

superior to that of the decision tree for the training sets of size 250 and 140. The other results, however, were far inferior, and for the very small training sets were degenerate with no results reported. Neural network model results over categorical data were quite good, and relatively stable for smaller data sets. There was, however, an anomaly for the training data set of 70 observations.

Results for the regression tree model applied to continuous data was inferior to that of the decision tree applied to categorical data except for the largest training set (which was very close in result). Discriminant analysis applied to continuous data also performed quite well, and did not degenerate when applied to the smaller data sets. The neural network model applied to continuous data was again erratic. Neural network models worked better for the data sets with more training observations.

4 Results

The logistic regression model had the best overall fit, using the full training set. However, this model failed when the data set was reduced to the point where the training set did not include cases that appeared in the test set. The categorical decision tree model was very good when 140 or 250 observations were used for training, but when the training set was reduced to 70, it was very bad (as were all categorical models. The decision tree model again seemed the most robust. Models based upon continuous data did not have results as good as those based on categorical data for most training sets. Table 14 provides a comparison of data set features based upon these results.

Table 14. Comparison

Factor	Positive Features	Negative Features
Large data sets (unbalanced)	Greater accuracy	Often degenerate (decision tree, regression tree, discriminant model)
Smaller data sets (balanced)	No degeneracy	Can miss test instances (logistic) May yield poor fit (categorical neural network model)
Categorical data	Slightly greater accuracy (but mixed results)	Less stable (small data set performance often the worst)

5 Conclusions

Key findings are that when the data is highly unbalanced, algorithms tend to degenerate by assigning all cases to the most common outcome. When data is balanced, accuracy rates tend to decline. If data is balanced, that reduces the training set size, and can lead to the degeneracy of model failure through omission of cases encountered in the test set. Decision tree algorithms were found to be the most robust with respect to the degree of balancing applied.

Simulated data sets representing important data mining applications in business were used. The positive feature of this approach is that expected data characteristics were controlled (no correlation of outcome with gender or state, for instance; positive correlations for educational level and major). However, it obviously would be better to use real data. Given access to such real data, similar testing is attractive. For now,

however, this set of experiments has identified some characteristics data mining tools with respect to the issue of balancing data sets.

References

1. Drew, J.H., Mani, D.R., Betz, A.L., Datta, P.: Targeting customers with statistical and data-mining techniques, *Journal of Service Research* **3**:3 (2001) 205-219.
2. Garver, M.S.: Using data mining for customer satisfaction research, *Marketing Research* **14**:1 (2002) 8-17.
3. Cowan, A.M.: Data mining in finance: Advances in relational and hybrid methods, *International Journal of Forecasting* **18**:1 (2002) 155-156.
4. Adams, N.M., Hand, D.J., Till, R.J.: Mining for classes and patterns in behavioural data, *The Journal of the Operational Research Society* **52**:9 (2001) 1017-1024.
5. Sung, T.K., Chang, N., Lee, G.: Dynamics of modeling in data mining: Interpretive approach to bankruptcy prediction, *Journal of Management Information Systems* **16**:1 (1999) 63-85.
6. Smith, K.A., Willis, R.J., Brooks, M.: An analysis of customer retention and insurance claim patterns using data mining: A case study, *The Journal of the Operational Research Society* **51**:5 (2000) 532-541.
7. Petropoulos, C., Patelis, A., Metaxiotis, K., Nikolopoulos, K., Assimakopoulos, V.: SFTIS: A decision support system for tourism demand analysis and forecasting, *Journal of Computer Information Systems* **44**:1 (2003), 21-32.