LITERATURE REVIEW: FINAL REPORT

Fadhil Salih

NOVEMBER 15, 2018

STAT-517
UNIVERSITY OF IDAHO

EFFECTS OF FEATURE CONSTRUCTION ON CLASSIFICATION PERFORMANCE: AN EMPERICAL STUDY EMPERICAL STUDY IN BANK FAILURE PREDICTION

ABSTRACT

Feature construction is crucial in data mining procedures, especially when it comes to evaluating classification performance. This research paper emphasizes the importance of feature construction from existing features, guided by domain knowledge, in increasing the classification performance. In their paper, Zhao, Sinha, & Ge, 2008, have used the example of predicting bank failure using appropriate feature construction to analyze their hypothesis. The paper evaluates classification performance by calculating the expected misclassification costs instead of using performance measuring attributes like error rates or classification accuracy, to better understand the practical effects of feature construction on classification performance. The data mining models used for this research are logistical regression, C4.5 decision tree, neural network, and K-nearest neighbor. From the results it was found that feature construction, guided by domain knowledge, increases the classifier performance and the degree of improvement varies with the data mining method that was employed for the analysis. It was seen that the use of constructed financial ratios was more effective in increasing the classification performance while using logistic regression and back-propagation neural networks than C4.5 decision tree or K-nearest neighbor.

INTRODUCTION

Relatively little research has been conducted to examine the effects of feature construction guided by domain knowledge on classification. In this paper, feature construction is used for generating a classification model to predict bank failures. Financial ratios are the features that are constructed from raw accounting variables that are used to represent various characteristics of a bank. The study compares the performance of classifiers built using raw accounting variables with that of classifiers built using a set of financial ratios constructed from those raw variables under various settings. The performance of these classifiers was measured using expected misclassification cost which was calculated based on equations formulated using domain knowledge. Logistic regression, back propagation neural networks, C 4.5 decision tree, and K-nearest neighbor were the learned classifiers used for this study because they were widely popular and effective in similar situations. Other classification models can be used to test this hypothesis as well.

Feature Construction

To develop better classification algorithms, substantial research has been conducted in data mining. Relatively limited research has been conducted on the importance of feature construction. Developing higher level features from raw features by utilizing domain knowledge may potentially increase the classification performance. For example, the ratio between total income and total assets of the bank, better represents the earning ability of the bank rather than total income or total assets alone.

Few examples of sources which may result in the degradation of classification performance are feature irrelevancy, feature redundancy, and feature interaction. To address these problems, we can use various approaches such as feature selection, feature extraction, and feature construction. Feature selection involves the selection of an appropriate subset of features that contains useful information for a specific task. Feature extraction and feature construction involve formulating composite features which are functions of the raw features.

In this case, 93 raw accounting variables were used to represent a scenario with no feature construction. For the experimental group, 26 financial ratios which were identified from previous studies, was used to represent a scenario with feature construction. The results were compared to test the accuracy of the hypothesis.

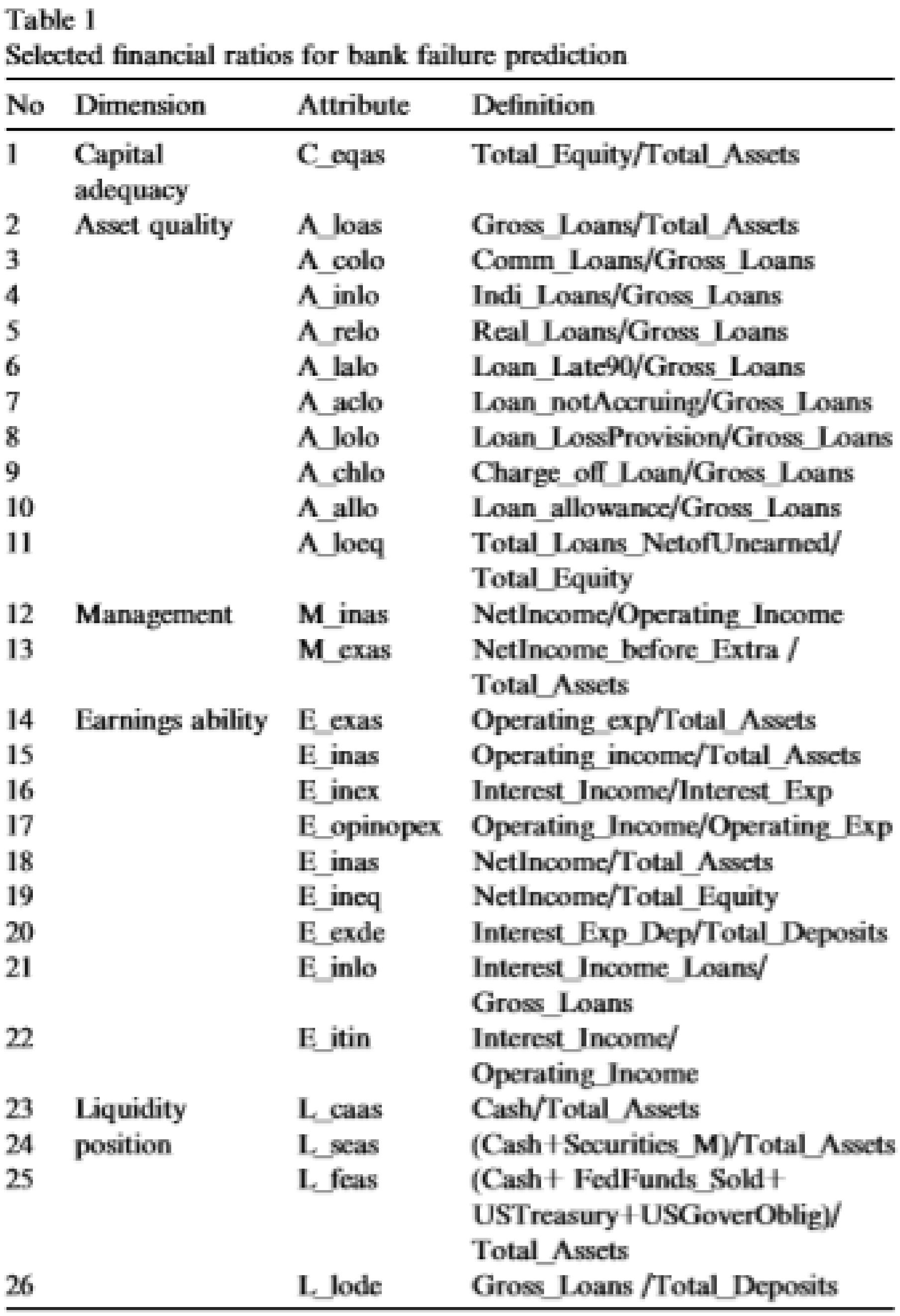


Table 1
Selected financial ratios for bank failure prediction

No	Dimension	Attribute	Definition
1	Capital adequacy	C_eqas	Total_Equity/Total_Assets
2	Asset quality	A_loas	Gross_Loans/Total_Assets
3		A_colo	Comm_Loans/Gross_Loans
4		A_inlo	Indi_Loans/Gross_Loans
5		A_relo	Real_Loans/Gross_Loans
6		A_lalo	Loan_Late90/Gross_Loans
7		A_aclo	Loan_notAccruing/Gross_Loans
8		A_lolo	Loan_LossProvision/Gross_Loans
9		A_chlo	Charge_off_Loan/Gross_Loans
10		A_allo	Loan_allowance/Gross_Loans
11		A_loeq	Total_Loans_NetofUnearned/ Total_Equity
12	Management	M_inas	NetIncome/Operating_Income
13		M_exas	NetIncome_before_Extra / Total_Assets
14	Earnings ability	E_exas	Operating_exp/Total_Assets
15		E_inas	Operating_income/Total_Assets
16		E_inex	Interest_Income/Interest_Exp
17		E_opinopex	Operating_Income/Operating_Exp
18		E_inas	NetIncome/Total_Assets
19		E_ineq	NetIncome/Total_Equity
20		E_exde	Interest_Exp_Dep/Total_Deposits
21		E_inlo	Interest_Income_Loans/ Gross_Loans
22		E_itin	Interest_Income/ Operating_Income
23	Liquidity position	L_caas	Cash/Total_Assets
24		L_seas	(Cash+Securities_M)/Total_Assets
25		L_feas	(Cash+ FedFunds_Sold+ USTreasury+USGoverOblig)/ Total_Assets
26		L_lode	Gross_Loans /Total_Deposits

Banking System

Bank regulatory bodies of the United States such as Office of Comptroller of Currency (OCC), the Federal Reserve (Fed), Federal Deposit Insurance Corporation (FDIC), ensures that the banking system functions efficiently. They conduct off-site monitoring of bank conditions by using classification models that are built to predict bank failures so that appropriate counter measures can be employed. Most of these models use financial ratios constructed from publicly available balance and income data that commercial banks are required to report to the regulatory authorities on a regular basis. These financial ratios are proven to be more effective in predicting and explaining bank failures than raw accounting variables. CAMEL, which is an acronym for Capital Adequacy, Asset Quality, Management Quality, Earnings Ability, and Liquidity Position, represents the five major characteristics of a bank's financial and operational conditions. The financial ratios are constructed to best represent these characteristics to evaluate a bank's financial and operational conditions.

Misclassification Cost

It is used to represent the error in the model in terms of expected monetary loss (normalized) to the bank due to wrong classification. These are classified into two: false negative errors and false positive errors. A false negative error takes place when the classifier misclassifies a bank that is likely to fail, as a survivor. A false positive error takes place when the classifier misclassifies a healthy bank as a bank that is likely to fail. False negative errors incur higher monetary loss than false positive errors because the consequence of the former is bank failure whereas the consequence of the latter is just an onsite examination.

Misclassification cost:

$$\text{Cost}(f) = C_{10} \cdot p \cdot p_{10} + C_{01} \cdot (1-p) \cdot p_{01}$$

Normalized misclassification cost:

$$\text{Cost}'(f) = \frac{\text{Cost}(f)}{C_{10} \cdot p + C_{01} \cdot (1-p)} = \frac{r \cdot p \cdot p_{10} + (1-p) \cdot p_{01}}{r \cdot p + (1-p)}$$

- p: Prior probability of failure
- p10 and p01: False negative and false positive error rates
- C10 and C01: Unit costs of false negative and false positive errors
- r: C10/C01
- Cost’(f): Normalized misclassification cost

RESEARCH DESIGN

Zhao et al., 2008, uses two different methods of feature construction to predict bank failure. The first method uses raw accounting variables whereas, the second method uses constructed financial ratios for bank failure prediction under a wide range of possible settings. The dataset used for this research was acquired from the FDIC website (www.fdic.gov). Information on 121 banks that failed in 1991 and 119 banks that failed in 1992 were collected. Based on 3 characteristics: 1.) geographic location (i.e., state), 2.) size of assets and 3.) charter type (federal chartered or state chartered), every failed bank was matched with non-failed bank as an experimental control. The first sample of 480 (240 failed, 240 non-failed) banks was generated for prediction one year prior to failure. The second sample had generated 468 (234 failed, 234 non-failed) banks among the 480 banks that did not fail for two years, were used for prediction two years prior to failure. Using the first method, 93 raw accounting variables were taken from internal call reports in the Federal Reserve database, which represents the scenario where there is no feature construction. In the second method, 26 financial ratios were chosen to represent the scenario with feature construction. Most of these ratios are of the simplest form, i.e. the numerator and the denominator are raw accounting variables. This is done to evaluate the increase in classification performance even with the simplest feature construction.

In each dataset a bank was described by a dependent variable (y=0 if there is no failure, y=1 if there is failure) and independent variables (x1, x2…, xn). A factorial design with 2 factors were used for this study. The first factor includes two levels: with and without feature construction. The second factor involves 4 classification methods- logistical regression, C4.5 decision tree, neural network, and K-nearest neighbor. The performance of each classifier was measured using expected miscalculation costs, which were calculated using prior probabilities (0.01 and 0.02) for failure, false negative and false positive error rates, and 10 cost ratios (10, 20, 30…, 100). A 2 X 4 factorial ANOVA was used under each setting of prior probability of failure and cost ratio for each prediction period. Stratified ten-fold cross validation was used to estimate the performance of learned classifiers.

RESULTS

Results indicated that the misclassification cost is lower for the experimental group (with feature construction) than the control group (without feature construction), under all settings. The ANOVA results show that the use of the financial ratios, instead of raw accounting variables, significantly improves the performance ($p < 0.01$), with respect to expected misclassification cost, of classifiers learned using several widely used classification methods. The use of financial ratios improves logistic regression and back-propagation neural networks more than C4.5

decision tree and k-nearest neighbor. The results hold across all settings of prediction period, prior probability of failure, and cost ratio.

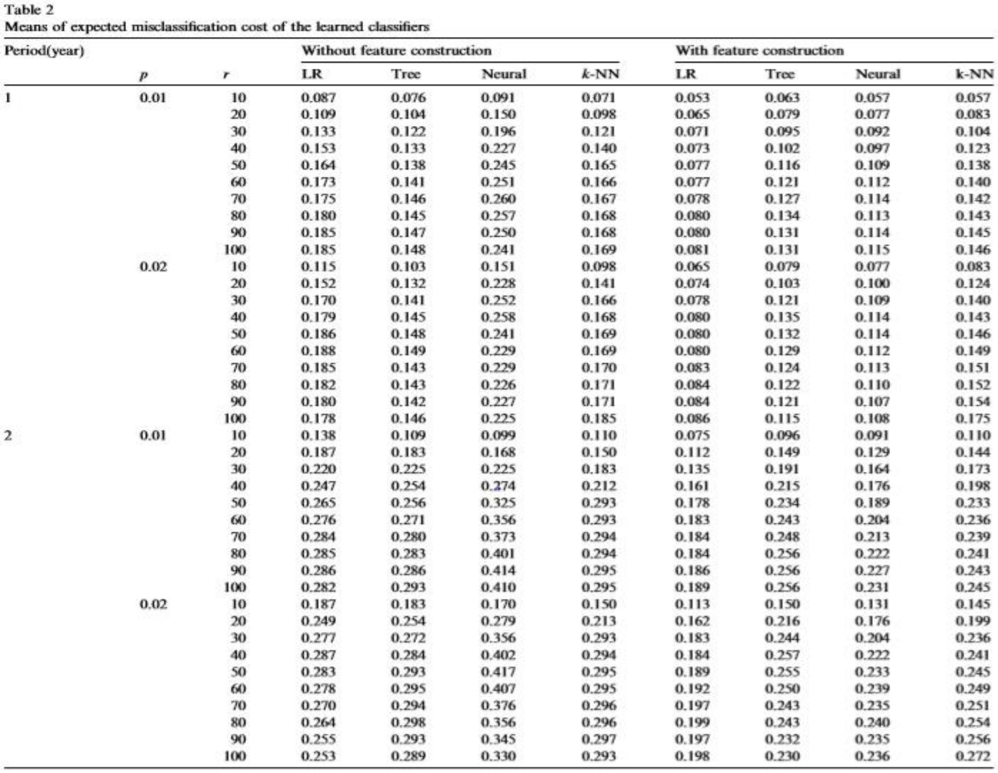
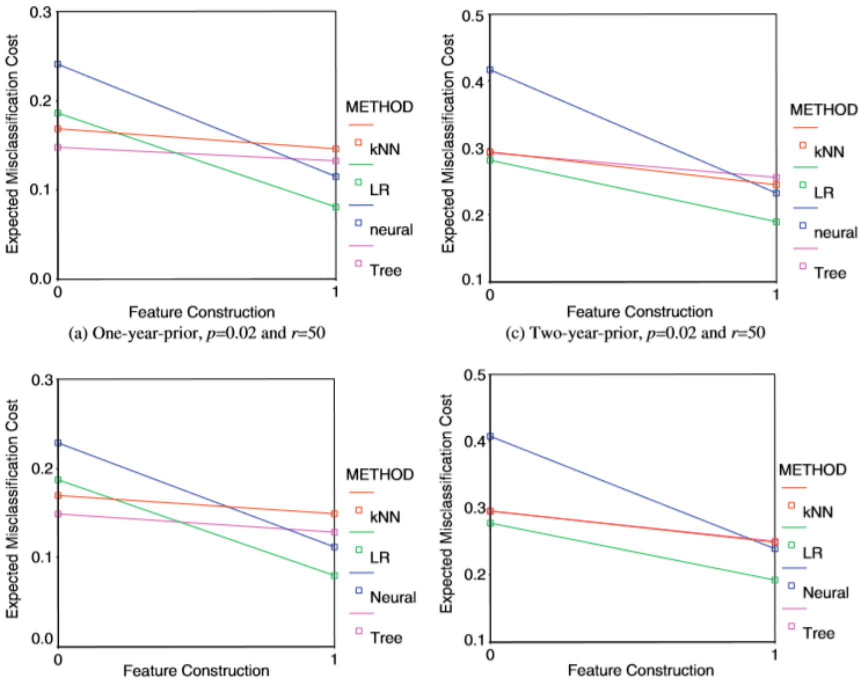


Table 2
Means of expected misclassification cost of the learned classifiers

Period(year)			Without feature construction				With feature construction			
	p	*r*	LR	Tree	Neural	*k*-NN	LR	Tree	Neural	k-NN
1	0.01	10	0.087	0.076	0.091	0.071	0.053	0.063	0.057	0.057
		20	0.109	0.104	0.150	0.098	0.065	0.079	0.077	0.083
		30	0.133	0.122	0.196	0.121	0.071	0.095	0.092	0.104
		40	0.153	0.133	0.227	0.140	0.073	0.102	0.097	0.123
		50	0.164	0.138	0.245	0.165	0.077	0.116	0.109	0.138
		60	0.173	0.141	0.251	0.166	0.077	0.121	0.112	0.140
		70	0.175	0.146	0.260	0.167	0.078	0.127	0.114	0.142
		80	0.180	0.145	0.257	0.168	0.080	0.134	0.113	0.143
		90	0.185	0.147	0.250	0.168	0.080	0.131	0.114	0.145
		100	0.185	0.148	0.241	0.169	0.081	0.131	0.115	0.146
	0.02	10	0.115	0.103	0.151	0.098	0.065	0.079	0.077	0.083
		20	0.152	0.132	0.228	0.141	0.074	0.103	0.100	0.124
		30	0.170	0.141	0.252	0.166	0.078	0.121	0.109	0.140
		40	0.179	0.145	0.258	0.168	0.080	0.135	0.114	0.143
		50	0.186	0.148	0.241	0.169	0.080	0.132	0.114	0.146
		60	0.188	0.149	0.229	0.169	0.080	0.129	0.112	0.149
		70	0.185	0.143	0.229	0.170	0.083	0.124	0.113	0.151
		80	0.182	0.143	0.226	0.171	0.084	0.122	0.110	0.152
		90	0.180	0.142	0.227	0.171	0.084	0.121	0.107	0.154
		100	0.178	0.146	0.225	0.185	0.086	0.115	0.108	0.175
2	0.01	10	0.138	0.109	0.099	0.110	0.075	0.096	0.091	0.110
		20	0.187	0.183	0.168	0.150	0.112	0.149	0.129	0.144
		30	0.220	0.225	0.225	0.183	0.135	0.191	0.164	0.173
		40	0.247	0.254	0.274	0.212	0.161	0.215	0.176	0.198
		50	0.265	0.256	0.325	0.293	0.178	0.234	0.189	0.233
		60	0.276	0.271	0.356	0.293	0.183	0.243	0.204	0.236
		70	0.284	0.280	0.373	0.294	0.184	0.248	0.213	0.239
		80	0.285	0.283	0.401	0.294	0.184	0.256	0.222	0.241
		90	0.286	0.286	0.414	0.295	0.186	0.256	0.227	0.243
		100	0.282	0.293	0.410	0.295	0.189	0.256	0.231	0.245
	0.02	10	0.187	0.183	0.170	0.150	0.113	0.150	0.131	0.145
		20	0.249	0.254	0.279	0.213	0.162	0.216	0.176	0.199
		30	0.277	0.272	0.356	0.293	0.183	0.244	0.204	0.236
		40	0.287	0.284	0.402	0.294	0.184	0.257	0.222	0.241
		50	0.283	0.293	0.417	0.295	0.189	0.255	0.233	0.245
		60	0.278	0.295	0.407	0.295	0.192	0.250	0.239	0.249
		70	0.270	0.294	0.376	0.296	0.197	0.243	0.235	0.251
		80	0.264	0.298	0.356	0.296	0.199	0.243	0.240	0.254
		90	0.255	0.293	0.345	0.297	0.197	0.232	0.235	0.256
		100	0.253	0.289	0.330	0.293	0.198	0.230	0.236	0.272



(a) One-year-prior, *p*=0.02 and *r*=50

(c) Two-year-prior, *p*=0.02 and *r*=50

(b) One-year-prior, *p*=0.02 and *r*=60

(d) Two-year-prior, *p*=0.02, and *r*=60

CONCLUSION

For each setting under different classification methods, the classifier learned with feature construction had a lower expected misclassification cost than the classifier learned without feature construction. The ANOVA results showed that the use of financial ratios, rather than raw accounting variables significantly improved the performance, with respect to expected misclassification cost. The results also showed that the performance varied for different classification methods. It was seen that the use of financial ratios significantly improves logistic regression and back-propagation neural networks more than C4.5 decision tree or k-nearest neighbor under all settings of prediction period, prior probability of failure, and cost ratio. Interestingly, the expected misclassification cost was higher for 2-years-ahead classifiers than corresponding one-year-ahead classifiers. The study provides empirical evidence that feature construction guided by domain knowledge can lead to significant performance improvement of data mining classifiers within the context of a cost-sensitive business problem. A classification method that performs below par in a pure data mining situation should not be excluded from consideration when higher-level features can be constructed based on domain knowledge. To conclude, the main results from this study are: 1.) Feature construction reduces expected misclassification cost for every setting, and 2.) the degree of cost reduction depends on the classification method.

FUTURE WORK

The generalizability of their findings can be validated in other business domains, where both domain knowledge and historical data are available for building prediction models. Future studies could examine if the results extend to multiple-class prediction tasks and regression tasks, such as sales forecasting. More studies are required in different domains to understand what other types of domain knowledge could influence performance. Different ways for incorporating domain knowledge need to be explored depending on specific applications.

REFERENCES

Huimin Zhao, Atish P. Sinha, & Wei Ge. (2008). Effects of feature construction on classification performance: An empirical study in bank failure prediction. *Expert Systems with Applications*, *36*, 2633-2644.