Neural Network Classifers in Arrears Management

Esther Scheurmann and Chris Matthews

Faculty of Science, Technology & Engineering,
La Trobe University, P.O. Box 199 Bendigo 3552, Victoria, Australia
Phone: +61 3 54447998, Fax: +61 3 54447557
c.matthews@latrobe.edu.au
EssScheurmann@gmail.com

Abstract. The literature suggests that an ensemble of classifiers outperforms a single classifier across a range of classification problems. This paper investigates the application of an ensemble of neural network classifiers to the prediction of potential defaults for a set of personal loan accounts drawn from a medium sized Australian financial institution. The imbalanced nature of the data sets necessitates the implementation of strategies to avoid under learning of the minority class and two such approaches (minority over-sampling and majority under-sampling) were adopted here. The ensemble out performed the single networks irrespective of which strategy was used. The results also compared more than favourably with those reported in the literature for a similar application area.

Keywords: neural network ensembles, minority over-sampling, majority under-sampling, loan default, arrears management.

1 Introduction

Authorised Deposit-Taking Institutions (ADIs) are corporations that are authorised under the Australian Banking Act (1959) to invest and lend money. ADIs include banks, building societies and credit unions. ADIs generate a large part of their revenue through new lending or extension of existing credit facilities as well as investment activities. The work described here focuses on lending, in particular the creation and management of customer personal loan accounts. The development of credit scoring models to aid in loan approval is well established. Traditionally these have been statistically based[9,11] although more recently artificial neural network approaches have attracted some research interest[4,13,15]. However there has been less work in the management of existing accounts. Substantial amounts of money are spent on recovery of defaulted loans, which could be significantly decreased by having the option of tracking a high default risk borrowers' repayment performance. This is sometimes referred to as *arrears* or *collections* management.

This is essentially a classification problem. Loan accounts could be classified as high or low risk depending on the risk of the customer not meeting their

W. Duch et al. (Eds.): ICANN 2005, LNCS 3697, pp. 325–330, 2005.

repayment committments. Multi-layer artificial neural networks can be considered as non-linear classifiers and, given their success in credit scoring, may be of use in identifying high risk accounts. A recent study[2] compared a neural network approach to the prediction of early repayment and loan default with more traditional approaches. The results were promising and suggested that a neural network approach outperformed the traditional approaches, particular for the prediction of early repayment.

The research reported here focuses only loan default and applies an ensemble as well as a single classifier approach. The data used is real life data sourced from a medium sized Australian bank and includes a low proportion of bad accounts. The paper is organised as follows: Section 2 provides a brief overview of ensembles and classifiers, section 3 discusses the data used in more detail and the experiments conducted, and section 4 discusses the experimental results. The paper concludes with a discussion of possible areas for future work that arise from the results presented here.

2 Classifiers and Ensembles

In simplest terms, a classifier divides examples into a number of categories. Classifiers may be trained on a data set and then tested on unseen data to determine their generalisation capabilities. Typically training uses a supervised learning approach i.e the target class is known for both the training and testing data. It has been shown that the use of an ensemble, rather than a single classifier, significantly improves classification performance [5,8,16]. Ensembles are particularly useful for classification problems involving large data sets[3] and can be constructed and combined in various ways[5,14].

Each member of the ensemble could be trained and tested on a subset of the total data set. This approach works well for unstable learning algorithms such as those used by artificial neural networks[5]. Several methods are available for the selection of these subsets. They can simply be selected at random (with or without replacement). The data set could be divided into a series of disjoint subsets and the training sets could be formed by leaving out one or more of the subsets, which might be reserved for testing. In these situations the ensemble members are trained independently of each other[10]. Another approach is to use a boosting algorithm such as the ADABOOST algorithm[6] which builds the ensemble by using datasets formed by focusing on misclassified examples. Ensembles can also be constructed using subsets of the input attributes. This approach is particularly useful when there is some redundancy amongst the inputs. In situations where there are many target classes ensemble members can be constructed using a reduced set. The number of target classes can be reduced combining several together. Whatever methods are choosen for ensemble construction the designer should ensure that there is diversity amongst individual ensemble members.

There are several ways of combining or fusing the decision of each individual classifier into one final ensemble decision. The simplest is to use an unweighted voting system where it is assumed that the relative importance of each individual decision is the same. If this is not the case then appropriate weightings could be introduced. A discussion of the possibilities can be found in [5,14] and examples of ensemble application areas in [1,12,17].

3 Experimental Work

The networks were developed using personal loan accounts created in May 2003. The observation point was 12 months later i.e May 2004. This was considered sufficient time before a realistic assessment of their performance could be made (Fig. 1). The networks were trained to classify whether an account was likely to

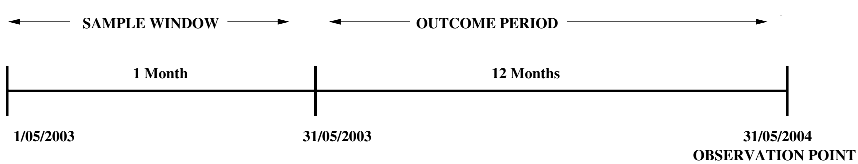

Fig. 1. Data Selection

lapse into arrears or remain healthy. An account was considered *in arrears* (i.e. 'bad') if, at the observation point, the contractual repayment obligations had not been met. Otherwise it was considered *not in arrears*, or 'good'. The data set totalled 1534 accounts consisting of 1471 'good' examples and 63 'bad' examples. The imbalanced nature of the data set was typical across the unsecured loan accounts of the financial institution involved.

23 input attributes were used of which 22 were collected at the time of loan approval and one during the outcome period, reflecting the loan performance. Of these 17 were continuous and 6 discrete. There was no significant correlation between any of the input attributes and the target class except for that collected during the outcome period and even in this case it was weak. The continuous attributes were linearly scaled from 0 to 1 and the discrete attributes were widened and represented as a suitable vector. There was little missing data. There were two target classes. All networks used 46 input neurons and one output neuron. The number of hidden layers and hidden layer neurons varied, depending on the experimental results. The networks were developed using the publically available neural network software *NevProp* and trained using the *quickprop* learning algorithm.

The literature suggests that networks trained on imbalanced data sets of the type used here tend to learn the majority class at the expense of the minority one[7]. A series of preliminary experiments using single networks trained, tested and validated on sets containing a ratio of good to bad examples equal to that in the original data set confirmed this. In arrears management it is important that the classifiers predict well the minority class (i.e. the 'bad' accounts). Several strategies have been suggested to overcome the data imbalance[7] and two (a single *minority over-sampled* network and an ensemble of *majority under-sampled networks*) were used here.

For the minority over-sampled network all majority class examples were retained and the data set was enlarged by sampling each minority class example five times. For each ensemble member all minority class examples were retained and a subset of the majority class, drawn at random, was added. Seven such data sets were created.

In all cases the data sets were subdivided into a training, a testing and a validation set. The proportion of 'good' to 'bad' accounts was 2:1 in each set. Multiple experiments were run to determine the best performing network based on testing set performance, particularly on the classification of 'bad' examples. A validation set was used to provide an estimation of performance on unseen data in the development data set.

4 Experimental Results and Discussion

The training, testing and validation performance of each individual network on the May 2003 − 2004 data is shown in table 1. The minority-oversampled network out performed all individual ensemble members, particularly in the classification of the 'bad' accounts. This is not surprising as the proportion of training and testing examples to the total available examples used during the development of this network was greater than that for the development of each ensemble member.

Table 1. Individual network performance on development (May 2004) data

Ensemble member	Testing % good	Testing % bad	Validation % good	Validation % bad
#1	95	85	88.5	84.6
#2	92.5	80	96	61.5
#3	85	85	57.7	84.6
#4	60	85	88.5	69
#5	85	80	80.8	80
#6	90	80	88.5	80
#7	80	90	65.4	61.5
minority-oversampled network	95	100	93	100

The trained ensemble and minority-oversampled networks were then applied to unseen data viz: personal accounts from June, Nov and Dec 2003 − 2004 (table 2). The proportion of 'good' to 'bad' accounts in these sets was similar to that in the development data set. A simple non-weighted majority voting system was used to determine ensemble performance. The ensemble clearly outperformed the minority-oversampled network in the classification of both 'good' and 'bad' accounts across the three data sets. It also outperformed the average performance of each individual ensemble member. These averages are also shown in the table. These results support the literature observation that the classification performance of an ensemble is superior to that of a single network (in this case that of both a minority-oversampled and a majority-undersampled network)[5,8,16].

The ensemble results also compare more than favourably with those in the analagous part of the study reported in [2]. In this case single networks were used to predict personal loan default after 12 months for a set of accounts from a U.K. financial institution. The minority class (loan default) was over-sampled and the input attributes, although less numerous, were similar to ones used here. The trained network yielded a classification accuracy of 78.8% overall (87.4 % on the good accounts, but only 33 % on the default accounts).

Table 2. Performance of the *ensemble* and the *minority-oversampled* network on unseen data

Observation point	June 2004		Nov 2004		Dec 2004	
	good	bad	good	bad	good	bad
ensemble	97.6	100	89	85	94.3	91.3
minority-oversampled network	83.7	91.7	72.5	63.8	75.7	78.8
ensemble member (average)	(84.8)	(89.9)	(77.7)	(71.8)	(80.7)	(78.8)

5 Conclusion and Future Work

Arrears management involves identifying and tracking high risk customer loan accounts. An ensemble of neural network classifiers shows promise as an accurate classifier for predicting potential personal loan defaults. The results reported here illustrate that ensembles outperform single networks, even when the data set is under or over-sampled. Future work includes the application of these approaches to the construction of systems that investigate the effectiveness of the loan approval process. This may include the identification of rejected loan applications that would possibly not default. Finally the development of single and ensembles of rule based classifiers, in an effort to supply a classification explanation for unsecured lending such as personal loan and credit card accounts, is another possible area for future research.

References

1. M.H.L.B. Abdullah and V. Ganapathy. Neural Network Ensemble for Financial Trend Prediction. In *Proceedings TENCON 2000*, volume 3, pages 157—161, Kuala Lumpar, Malaysia, 2000.
2. Bart Baesens, Tony Van Gestel, Maria Stepanova, and Jan Vanthienen. Neural Network Survival Analysis for Personal Loan Data. In *Proceedings of the Eighth Conference on Credit Scoring and Credit Control (CSCC VIII'2003)*, Edinburgh, Scotland, 2003.

3. Nitesh V. Chawla, Lawrence O. Hall, Kevin Bowyer, and W. Philip Kegelmeyer. Learning Ensembles from Bites: A Scalable and Accurate Approach. *Journal of Machine Learning*, 5:421—451, 2004.
4. V.S. Desai, J.N. Crook, and G.A. Overstreet Jr. A Comparison of Neural Networks and Linear Scoring Models in the credit union environment. *European Journal of Operational Research*, 95:24—37, 1995.
5. Thomas G. Dietterich. Machine-Learning Research: Four Current Directions. *AI Magazine*, 18(4):97—136, 1997.
6. Y. Freund and R. Schapire. Experiments with a new boosting algorithm. In *Proceedings Thirteenth International Conference on Machine Learning*, Bari, Italy, 1996.
7. Hongyu Guo and Herna L. Viktor. Learning from Imbalanced Data Sets with Boosting and Data Generation: The Databoost-IM Approach. *ACM SIGKDD Explorations Newsletter: Special Issue on Learning from Imbalanced Datasets*, 6(1):30—39, June 2004.
8. L.K. Hansen and P. Salamon. Neural Network Ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993—1001, 1990.
9. E.M Lewis. *An Introduction to Credit Scoring*. The Athena Press, San Rafael, California, 1992.
10. Y. Liu, X. Yao, Q. Zhao, and T. Higuchi. An Experimental Comparison of Ensemble Learning Methods on Decision Boundaries. In *Proceedings 2002 International Joint Conference on Neural Networks, IJCNN '02*, volume 1, pages 221—226, Honolulu, HI USA, 2002.
11. E Mays. *Handbook of Credit Scoring*. Glenlake Publishing, Chicago, 2001.
12. Yair Shimshoni and Nathan Intrator. Classification of Seismic Signals by Integrating Ensembles of Neural Networks. *IEEE Transactions on Signal Processing*, 46(5):1194—1201, 1998.
13. R.P. Srivastva. Automating judgemental decisions using neural network: a model for processing business loan applications. In *Proceedings of the 1992 ACM Conference on Communications*, Kansas City, Missouri, 1992.
14. Nayer M. Wanas and Mohamed S. Kamel. Decision Fusion in Neural Network Ensembles. In *Proceedings 2001 International Joint Conference on Neural Networks, IJCNN '01*, pages 2952—2957, Washington, DC USA, 2001.
15. D. West. Neural Network Credit Scoring Models. *Computer & Operations Research*, 27:1131—1152, 2000.
16. Yunfeng Wu and Juan I. Arribas. Fusing Output Information in Neural Networks: Ensemble Performs Better. In *Proceedings of the 25th Annual Conference of the IEEE EMBS*, pages 2265—2268, Cancum, Mexico, 2003.
17. Xin Yao, Manfred Fischer, and Gavin Brown. Neural Network Ensembles and Their Application to Traffic Flow Prediction in Telecommunications Networks. In *Proceedings 2001 International Joint Conference on Neural Networks, IJCNN '01*, pages 693—698, Washington, DC USA, 2001.