Artificial Immune Systems: A Predictive Model for credit scoring

Thabiso Peter Mpofu, Dr .G Venkata Rami Reddy

Abstract— With the advent of the global financial crisis which hit the global economy, credit scoring has become of the essence. The global financial crisis also known as the "credit crunch" was largely attributed to the issuance of credit to individuals with no capacity to return the money. Credit scoring has become a very important task in the credit industry. Various credit scoring methods such in areas as artificial neural networks (ANNs), statistical based methods and decision trees have been proposed to increase the accuracy of credit scoring models. The proposed Artificial Immune Systems (AIS) are an artificial intelligence technique modelled on natural immune system processes have been used to solve various kinds of real life processes with success. In this paper we compare the performance of current classifiers used in credit rating against Artificial Immune Systems. Artificial Immune Systems have various algorithms used to implement them. The algorithm under consideration is the negative selection algorithm. Artificial Immune Systems (AIS) are found to be produce competitive results very close to traditional artificial intelligent systems such as Neural Networks.

Index Terms— credit scoring, Artificial Intelligence, negative selection algorithm, financial, credit rating, artificial immune system, artificial neural networks (ANNs)

———————————— ◆ ————————————

1 INTRODUCTION

Whenever a bank customer applies for a loan, a decision has to be made. A choice on whether to award the loan to the applicant or not has to be made. The institution issuing a loan be it a bank, micro finance institution or in the issuance of credit cards has to assess the risk associated in the form of credit scoring. Credit scoring is one of the most successful applications and operations research techniques used in banking and finance, and is also one of the earliest financial risk management tools developed [1]. The ability to accurately assess the level of the borrower's credit is therefore very important [2]. Enterprises have to be always on their toes if they are to maintain their business as a going concern. Various methods are used in order to assess and measure the default rates of different enterprises. Some of the credit risk ratings are quantitative methods, random probability and some based on classifiers [3]. Credit scoring was developed by Fair and Isaac in the early 1960s and in simple terms corresponds to producing a score that can be used to classify customers into two separate groups: the credit-worthy or"good" group (likely to repay the credit loan), and the non credit-worthy or "bad" group (rejected due to its high probability of defaulting) [4]. Credit scoring can be viewed as a classification problem. In order to solve this problem lenders have been using different techniques in the past years such as classical statistical techniques, discriminant analysis and logistic regression, were acknowledged by the market and employed. Artificial Neural Networks have also been used for example in Credit Rating Analysis [7] and Personal Credit Rating Assessment for the National Student Loans [8], Personal Credit Rating Using Artificial Intelligence Technology for the National Student Loans[9], Research of electronic commercial credit rating based[10]. Artificial immune systems (AIS) have been proposed as an alternative approach to solve computational intelligence problems [4]. The key feature of natural immune systems is its ability to distinguish non self(foreign substances) from self without prior knowledge of all possible non-self variants. AIS are implemented using various algorithms. The AIS algorithm under consideration here is the Negative selection Algorithm (NSA).

2 BACKGROUND

2.1 A Review of Credit Scoring Techniques
Credit Scoring techniques are divided into two parts
  i.   Statistical Based Techniques
 ii.   Artificial Intelligence Based Techniques

2.1.1 Statistical Based Techniques
Various credit scoring statistical based techniques have been researched on and implemented. These statistical based techniques include linear based, discriminant analysis, probit analysis, decision trees and logistic regression. Of all these statistical based techniques, logistic regression and discriminant analysis have proven to be the most popular [20]. Logistic regression and discriminant analysis will therefore be described minor detail.

2.1.1.1 Logistic Regression

Logistic regression (LR) is a probabilistic statistical classification model.LR model is one of the most preferred in solving classification problems. LR model can fit various kinds of distribution functions such as Gamble, Poisson, and normal distributions, unlike other statistical tools (e.g. discriminant analysis or ordinary linear regression). It is more suitable for the fraud detection problems. In addition, in order to increase its accuracy and flexibility, several methods have been proposed to extend the traditional binary logistic regression model including multinomial logistic regression model and logistic regression model for ordered categories [5].

2.1.1.2 Discriminant analysis

Discriminant analysis is a credit scoring technique developed

to discriminate between two groups. It is widely agreed that the discriminant approach is still one of the most widely established techniques to classify customers as good credit or bad credit.[20]

2.1.2 Artificial Intelligence Based Techniques

Artificial Intelligence based techniques are algorithms inspired by nature. These Artificial Intelligence based techniques include Artificial Nueral Networks(ANN) modeled on the way the brain works, Genetic Algorithm(GA), based on the phenomenon of evolution of species, and the last one is Artificial Immune System, based on the natural immune system [5].

2.1.2.1 Artificial Neural Networks (ANNs)

ANNs are computational models inspired by an animal's central nervous system which is capable of machine learning as well as pattern recognition [17]. ANNs are inspired by the functionality of the nerve cells in the brain. Just like humans, ANNs can learn to recognise patterns by repeated exposure to many different examples. They are non-linear models that can classify based on pattern recognition capabilities. This gives them an advantage over conventional statistical techniques used in industry which are primarily linear. In the field of credit scoring, studies have shown that neural networks perform significantly better than statistical techniques. [1], [5]. ANN have been used in credit rating and credit scoring quite extensively as illustrated in the following papers : "Artificial Neural Networks for Corporation Credit Rating Analysis"[7], "Personal Credit Rating Assessment for the National Student Loans based on Artificial Neural Network"[8], "Personal Credit Rating Using Artificial Intelligence Technology for the National Student Loans" where a Back Propagation neural network was used [9], "Research of electronic commercial credit rating based on Neural Network with Principal Component Analysis" [10]

2.1.2.2 Genetic Algorithms (GAs)

GAs try and replicate the natural selection process. The natural selection process involves the passing in of genes to the next generation [21]. GAs are inspired by biological evolution and offer efficient problem-solving mechanisms. A problem's solution is evolved over many processing cycles, each time producing better solutions. Application of GAs is rapidly expanding with successful applications in finance trading, fraud detection and other areas of credit risk. Desai et al. investigated the use of GAs as a credit scoring model in a credit-union environment while Yobas et al. compared the predictive performances of four techniques, one of which is GAs, GA faired quite well coming in second place[1][5].

2.1.2.3 Artificial Immune Systems (AIS)

AIS are an Artificial Intelligence technique based the natural immune system of the body. AIS have a learning and memory component, and perform pattern recognition. AIS were implemented in a paper titled "An Artificial Immune System for Extracting Fuzzy Rules in Credit Scoring" [11]. Weka data mining software was used to classify and in turn compared with other well known classifiers. They used the clonal selection algorithm to implement the AIS. Competitive results with high accuracy were obtained. In this paper we will be using the AIS as well but using a different algorithm known as the Negative Selection Algorithm

2.2 Summary

Various credit scoring techniques were looked at. Statistical based and Artificial Intelligence based Techniques that were used in credit scoring were described briefly. The first review looked at statistical based methods and then we looked at nature inspired algorithms. Some studies found statistical techniques to perform better than AI techniques, while others concluded just the opposite.

3 METHODOLOGY

3.1 General Design Issues

Three general design issues had to be dealt with before the implementation of each of the algorithms

1. Choice of Design Language
2. The Source of Data Used
3. Choice of Input parameters from the German dataset

3.1.1 Choice of Design Language

Matlab was chosen as the platform to model the system for prediction..Matlab is considered to be amongst the leading platforms for technical projects. Matlab operates in matrice and vector form and is ideally suited for numeric datasets where complex systems can be modelled with relative ease.

3.1.2 The Source of Data Used

Because Matlab operates in matrice and vector form, the German.data-numeric dataset with numeric attributes was selected. Categorical attributes are represented in numerical form to able to be manipulated by Matlab. For algorithms that need numerical attributes, Strathclyde University produced the file "german.data-numeric". This file has been edited and several indicator variables added to make it suitable for algorithms which cannot cope with categorical variables. The German.data-numeric dataset consists of credit data which was collected in Germany and is widely used in the credit rating studies. The German.data-numeric dataset has the following attributes shown in table 1 below:

TABLE 1
GERMAN.DATA-NUMERIC DATASET ATTRIBUTES

Dataset	Attribute Type	n	classes	Missing attributes
German numer	24 numeric	1000	700 good 300 bad	nil

The dataset contains 1000 elements and of those 1000, 700 are credit worthy and 300 are credit unworthy. There are a total of 24 numeric attributes within the dataset [18]

3.1.3 Choice of Input parameters from the German.numer dataset

Data is required to create the training and tests data. A variation of attributes was used in this paper.For the training and testing 8, 12, 15, and 24 attributes were used in the preparation of the training and test data. The columns selected were kept consistent so as to be able to find the optimum number of columns which produce the best results.

3.2 Implementation

The implementation is split up into two main sections:

- Training: Creating a set of detectors that will identify 'nonself', or bad loans, by application of the negative-selection process;
- Testing: Using these detectors to distinguish between the people to give the loans to and the ones not to give loans to.

The matching degree of detector and the object to be detected is a main criterion of self non-self recognition ability. The implementation was carried out using the negative selection algorithm.

3.2.1 The Negative Selection Algorithm

The negative selection algorithm can be described as a mathematical representation of the maturation of Tcells in the thymus gland. It uses the principles of self/non-self discrimination to distinguish between two system states of normal and abnormal. The normal in this implementation are the people to give loans to who will return the money back while the abnormal are those that are highly likely to default. The defaulters are identified using detectors trained on a sample set of self (good loans). This process is known as negative selection. The inspiration of the whole algorithm is based on an organ called the thymus. The thymus is responsible for generating T-cells. T cells that react with '*self*'-proteins are rejected and eliminated and only those that do not bind the 'self'-proteins are allowed to remain. This guarantees only foreign or anomalous molecules are recognised. The T-cells(detectors) are then distributed throughout the body, eliminating any foreign bodies(non self) they encounter[19].

3.2.2 Training and Test dataset

3.2.2.1 Training dataset

Training and Test data are cut out of the German.data-numeric dataset. A percentage of the self samples are selected which are then used to train the detectors.The training data was divided into 3 cuts

    i.   **Cut 1 : 25**%

A 25% cut of the self data was used to train the detectors
Under the 25% cut a variation of columns was used to create the test data. Four training datasets with the following properties as illustrated I Table 2 were created

TABLE 2
ATTRIBUTE CHARACTERISTICS FOR CUT 25 %

No of columns	Column numbers
8	1,3,5,6,7,9,11
12	1,3,5,6,7,9,11,13,16,18,20,22
15	1,2,3,5,6,7,8,9,10,12,14,15,16,19,20
24	All columns

    ii.   **Cut 2: 50** %

A 50% cut of the self data was also created and used to train the detectors. Under the 50% cut a variation of columns was used to create the test data. Four training datasets with the following properties were created as shown in table 3 below

TABLE 3
ATTRIBUTE CHARACTERISTICS FOR CUT 50 %

No of columns	Column numbers
8	1,3,5,6,7,9,11
12	1,3,5,6,7,9,11,13,16,18,20,22
15	1,2,3,5,6,7,8,9,10,12,14,15,16,19,20
24	All columns

    iii.   **Cut 3: 65**%

A 65% cut of the self data was used to train the detectors instead of using a 75% cut, this was because using a 75% cut to create training data, there would not be enough data left to create the testset, so a 65% cut was opted for instead and the attribute characteristics are shown in table 4 below

TABLE 4
ATTRIBUTE CHARACTERISTICS FOR CUT 65 %

No of columns	Column numbers
8	1,3,5,6,7,9,11
12	1,3,5,6,7,9,11,13,16,18,20,22
15	1,2,3,5,6,7,8,9,10,12,14,15,16,19,20
24	All columns

The above cuts are created and used to train the detectors using the negative selection algorithm

3.2.2.2 Test dataset

The testset with the corresponding cut size and column numbers was created to test the system trained using the training data created. The testset was used to test the trained system. The number of acceptable loans is already known. The overall accuracy of prediction is then calculated

4   RESULTS AND DATA ANALYSIS

4.1 Results

The system was trained using various cut sizes and varying columns. After the training and testing was done,
The number of self rows is already known (numself), the number of self rows is then predicted using the test data and a comparison is done to come out with the prediction accuracy

Number of self identified =μ
Number of self rows = α

Accuracy = α / μ

the following accuracy of prediction were found as illustrated in table below

TABLE 5
RESULTS FOR ALL THE CUTS

Cut Size (%)	Columns			
	8	**12**	**15**	**24**
25	74.86	80.57	77.71	21.71
50	76.29	78.00	78.57	16.86
65	74.14	80.00	80.00	19.34
	75.14	**79.52**	**78.76**	**19.30**

Average (%)

A balance had to be struck between under training and over-training the prediction system created. The optimum number of columns was found to be 12 columns which had the highest average accuracy of 79.52 across all the cuts created. Using all 24 columns led to overtraining the system resulting in drastically lower accuracy rates. Taking the average of the best 3, we get an average accuracy of 77.81%

4.2 Comparative Analysis

The following are results obtained from similar studies using the German dataset

TABLE 6
COMPARITIVE RESULTS FOR GERMAN DATASET

Rank	Model	Accuracy(%)	Source
1	Neural Networks	78	[22]
2	Negative Selection Algorithm	77.81	This study
3	SAIS	75.4	[1]
4	Naïve Bayes	74.7	[23]

SAIS-Simple Artificial Immune System
NN-Nueral Network

The prediction accuracies for credit rating using the negative selection algorithm are extremely competitive as shown in Table 6. These findings are quite encouraging as the algorithm is extremely portable, with implementation requiring few, if any, modifications for adaptation to other data sets and classification problems.

5   CONCLUSION

Artificial Immune Systems implemented using the Negative selection algorithms were found to produce very competitive results. Artificial Immune Systems (AIS) are found to be produce competitive results very close to traditional artificial intelligent systems such as Neural Networks [14]. As evidenced in this study, very reasonable results were obtained. The use of Artificial Immune Systems will increase and more and more research is done on AIS. AIS are producing good results in the field of prediction and classification. Other potential uses are AIS are still being taken under consideration.

6 FUTURE WORK

The field of Artificial Immune Systems (AIS) possesses a lot of potential. Uses of AIS have been increasing and varied in range from credit evaluation for Mobile Customers [6], Performance Evaluation of a Fraud Detection System [12], Credit Cards Fraud Detection [13] to anomaly detection [15]. Results obtained are very promising. A look at other AIS techniques

such as using the clonal selection algorithm in credit scoring will be looked at in the future work. Other potential uses of the negative selection algorithm will also be considered. Uses in areas such as predicting weather situations such as drought/no drought, rainy/sunny.

Acknowledgments

8. REFERENCES

[1] Kevin Leung, France Cheong, Christopher Cheong "Consumer Credit Scoring using an Artificial Immune System Algorithm," IEEE, pp. 3377-3384, 2007

[2] WANG Wei , NIU WeiHong "Research and Implementation of the Credit Rating System for Bank Customers" IEEE, 2011

[3] Qi Fei , Chi Guotai, Sui Cong "The Small Sample Credit Risk Rating and its Empirical Study" IEEE, pp 632-635, 2011

[4] Antonio I. S. Nascimento and Germano C. Vasconcelos ,"An Experimental Investigation of Artificial Immune System Algorithms for Credit Risk Assessment Applications" In Proceedings of the WCCI IEEE World Congress on Computational Intelligence, pp 1-8, 2012

[5] Antariksha Bhaduri "Credit Scoring using Artificial Immune System Algorithms: A Comparative Study," World Congress on Nature & Biologically Inspired Computing, pp. 1540-1540, 2009

[6] Yang Zong-chang, Kuang Hong "Credit Evaluation for Mobile Customers Using Artificial Immune Algorithms" Proceedings of the 7th World Congress on Intelligent Control and Automation pp.7021-7025, 2008

[7] Liu Yijun, Cai Qiuru, Luo Ye, Qian Jin, Ye Feiyue "Artificial Neural Networks for Corporation Credit Rating Analysis," International Conference on Networking and Digital Society,pp. 81-84, 2009

[8] Xiao_ jie Zhang, Jian Hu "Personal Credit Rating Assessment for the National Student Loans based on Artificial Neural Network," International Conference on Business Intelligence and Financial Engineering, pp. 53-56, 2009

[9] Jian HU "Personal Credit Rating Using Artificial Intelligence Technology for the National Student Loans," In Proceedings of the 4th International Conference on Computer Science & Education, pp. 103-106, 2009

[10] XUE Xiang-hong, XUE Xiao-feng "Research of electronic commercial credit rating based on Neural Network with Principal Component Analysis," IEEE, pp. 1-4, 2010

[11] Ehsan Kamalloo ,Mohammad Saniee Abadeh "An Artificial Immune System for Extracting Fuzzy Rules in Credit Scoring;" IEEE, pp 1-8, 2010

[12] Elham Hormozi , Mohammad Kazem Akbari "Performance Evaluation of a Fraud Detection System based Artificial Immune System on the Cloud," In Proceedings of the The 8th International Conference on Computer Science & Education, pp 819-823, 2013

[13] Hadi Hormozi , Elham Hormozi "Credit Cards Fraud Detection by Negative Selection Algorithm on Hadoop" In Proceedings of the 5th Conference on Information and Knowledge Technology (IKT), pp 40-43, 2013

[14] M. Gunasekaran and K.S. Ramaswami, "Evaluation of Artificial Immune System with Artificial Neural Network for Predicting Bombay Stock Exchange Trends," Journal of Computer Science 7, pp 967-972, 2011

[15] Yao-Guang We1 , De-Ling Zheng , Ying Wang , "Research of a negative selection algorithm and its application in anomaly detection" In the Proceedings of the Third International Conference on Machine Learning and Cybemetics, Shanghai, , pp.2910-2913, 26-29 August 2004

[16] Statlog, "machine-learning -database," https://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/german.data-numeric, 1993

[17] Wikipidea, "Artificial neural network" http://en.wikipedia.org/wiki/Artificial_neural_network, 2013

[18] Statlog, "machine-learning -database," https://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/. 1993

[19] Alice Delahunty , Denis O Callaghan ,"Artificial Immune Systems for the Prediction of Corporate Failure and Classification of Corporate Bond Ratings," mis.ucd.ie, http://mis.ucd.ie/research/theses/Delahunty-OCallaghan.pdf, 2003

[20] Abdou, H. & Pointon, J. (2011) 'Credit scoring, statistical techniques and evaluation criteria: a review of the literature ', Intelligent Systems in Accounting, Finance & Management, 18 (2-3), pp. 59-88.

[21] David J. Fogarty, "Using genetic algorithms for credit scoring system maintenance functions," International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.6, pp 1-8, November 2012

[22] Y. S. Kim and S. Y. Sohn, "Managing loan customers using misclassification patterns of credit scoring model," Expert Systems with Applications, vol. 26, pp. 567–573, 2004.

[23] Y. Lan, D. Janssens, G. Chen, and G. Wets, "Improving associative classification by incorporating novel interestingness measures," Expert Systems with Applications, vol. 31, pp. 184–192, 2006

_____

- *Thabiso Peter Mpofu received B. Tech degree in Computer Science at Harare Institute of Technology (HIT), Zimbabwe in 2010. He is currently pursuing a masters of technology degree program in computer science at JNTU, Hyderabad, India. He is a HIT staff development research fellow. His research interests are in the area of Artificial Intelligence, Network Security and Mobile Computing.PH-+918179823780. E-mail: thabiso.mpofu@gmail.com.*
- *Dr G Venkata Rami Reddy received the M.Tech. (CSE) degree from JNT University Hyderabad in 1998. He received his Ph.D. degree in Computer Science and Engineering from Jawaharlal Nehru Technological University (JNTU) in 2013. He has been working in JNT University since 2000. Currently he is working as an Associate Professor in Dept of CSE in School of Information Technology, JNT University Hyderabad. He has more than 14 years of experience in teaching and Soft-ware Development. . He has presented more than 15 National and International journal and conference. His research interests include Image Processing, Pattern Recognition, Network Security, Digital Watermarking, Image retrieval, and computer networks.*