Small and Medium Enterprises Risky Prediction System Based On Cloud Computing

Malin Song
Department of Statistics, Anhui University of Finance and Economics, China

Ben-Chang Shia
Department of Statistics and Information Science & Applied Statistics, Fu Jen Catholic University, Taipei Hsien
Email: stat.nan@gmail.com

Hailiang Yao and Wei Wang
School of Statistics Central University of Finance and Economics ,Beijing, China

Kuangnan Fang
Department of Statistics, School of Economics, Xiamen University, Fujian Province, 361005
Email: xmufkn@xmu.edu.cn

***Abstract*—This paper establishes the system of C^{2}FAST and Improved Cloud-R based on cloud computing and parallel computing. Small and medium enterprises can implement data mining technology on the systems to obtain the relevant financial information and financial risk prediction models.**

***Index Terms*— Data Mining; Cloud Computing; MES Model; Improved Cloud-R; C^{2}FAST**

Ⅰ. INTRODUCTION

Since the second half of 2007, due to the U.S. subprime mortgage crisis, causing the global financial crisis, and coupled with energy and raw materials continued to rise, let the global economy grow up slowly. Taiwan has also been the impact of the environment. There are 68.49% of small and medium enterprises (SMEs) financing from bank, Therefore, financial institutions to reduce collection costs and the risk of bad debts, they much emphasis on "risk management" to ensure sustainable development when corporate lending. But now Taiwan's SMEs to the bank and other financial institutions to apply for loans is more and more difficult compared with past, hence the use of financial risk analysis system will bring need a lot of help for the business or investors who need the financial information. Taiwan's small and medium operators who are usually technical staff and they usually start business by themselves, so often a lack of marketing and financial capacity, if coupled with a lack of sound financial management system, they will often need to take excessive risks or missed opportunities for business growth. And after the attacks of international financial crisis, many companies have poor business outbreak of financial crisis, so the financial position of SMEs also subject to the test. Only through the financial statements, Investors often can't immediately estimate the company's operating condition, so they can't make objective judgments to make further investment or understanding the company's operating condition. So long as the data update instantly, investors will be able to get the latest Financial information and the message of Financial risk forecasting result; Enterprises also can understand whether their partnership have financial crisis to decide whether to continue cooperate or not,... and so on. Because Many investors can't immediately know the condition of the company operating so that losses, for this reason, if it can immediately understands the overall economic situation of companies is very important for investors. With the rapid changes in technology, consumer's demand and rely on computer networks is increasing, so they need be provided to instantly information and the rapid network environment.

Because of the need to improve services to customers demand, the company also needs a more powerful database and faster computational speed and some more effective statistical methods. But the high cost of building a private server, so using cloud technology can reduce the costs of operating, hardware and software for business. "Cloud technology" is very popular in recent years, which with immediateness, high-efficiency features. Through a remote server to create a virtual platform, so users can immediately use the virtual platform to computing from their own computers. Therefore, the construction of a "cloud of Financial Crisis Prediction Model" for investors or interested parties is very important, it will save social costs and to improve their overall performance.

Corresponding author is Kuangnan Fang.

doi:10.4304/jsw.7.10.2236-2240

Ⅱ .ESTABLISHMENT OF SME RISK PREDICTION SYSTEM

In this paper, we reference the significant variables of financial risk models in the literature as basis, then find out those financial variables from TEJ for SME, from year 2007 to 2009, to made reference database, and then combined R and PHP into the server. Finally make the cloud computing and financial analysis systems. SME Risk Prediction System in this paper is divided into two parts, elementary system and advanced system.

Risk Prediction of Small and Medium Enterprises	
Logistic Model，Decision Tree，Random Forests，SVM，Bayesian Network，Artificial Neural network	Other Data Mining Technologies
Elementary System: C²FAST	Advanced System: Improved Cloud-R
distributed processing，grid computing	parallel computing
Infrastructure	

Figure1. The Diagram of SME Risk Prediction System

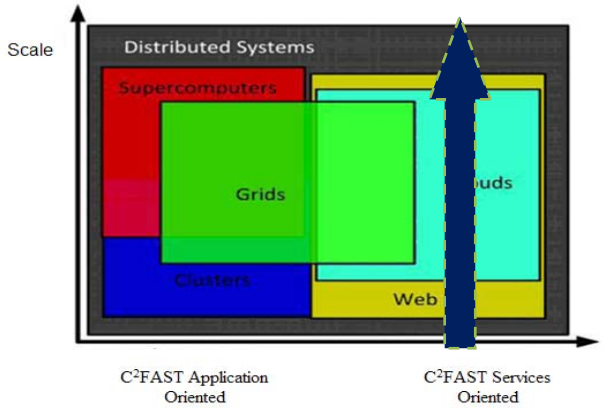

Figure2. Grid Computing and Cloud Computing in C²FAST Server

A. *Elementary System—C²FAST*

C²FAST is an online financial analysis and risk prediction tool which is established from the cloud of platform. It can be used for financial analysis and risk prediction of SME. The cloud technology used in C²FAST is as shown in figure 2..

Seeing from the above picture, in distributed processing, grid computing combines cluster computing and super-computing with the medium software installed in the computer for computing processing, let the computation load can be allocated to all the computers. Grid computing is the former of Cloud computing, it can provide corresponding services based on Web 2.0.

The elementary system of SME can implement six data mining algorithms including: Logistic Model，Decision Tree，Random Forests，SVM，Bayesian Network，Artificial Neural network。

The implementation process of elementary system is as figure 3.

Small and medium enterprises can draw on the C²FAST and make on-line financial analysis and risk prediction easily and quickly using six data mining technology.

B. *Advanced System—Improved Cloud-R*

The basic system uses only six method of Data Mining, This basic system is still not very complete, such as there is no more complete function to check data. So follow-up researchers can increase the system function, for example, exclude outliers, normality test. Future researchers may use other methods, such as factor analysis, MARS, and time series analysis methods; they can use the system of improved Cloud-R to complete their plans.

Cloud-R is different from the other R web interface is Cloud-R is written in php. Users do not need to install R software or java, and other packages. It can be used as a

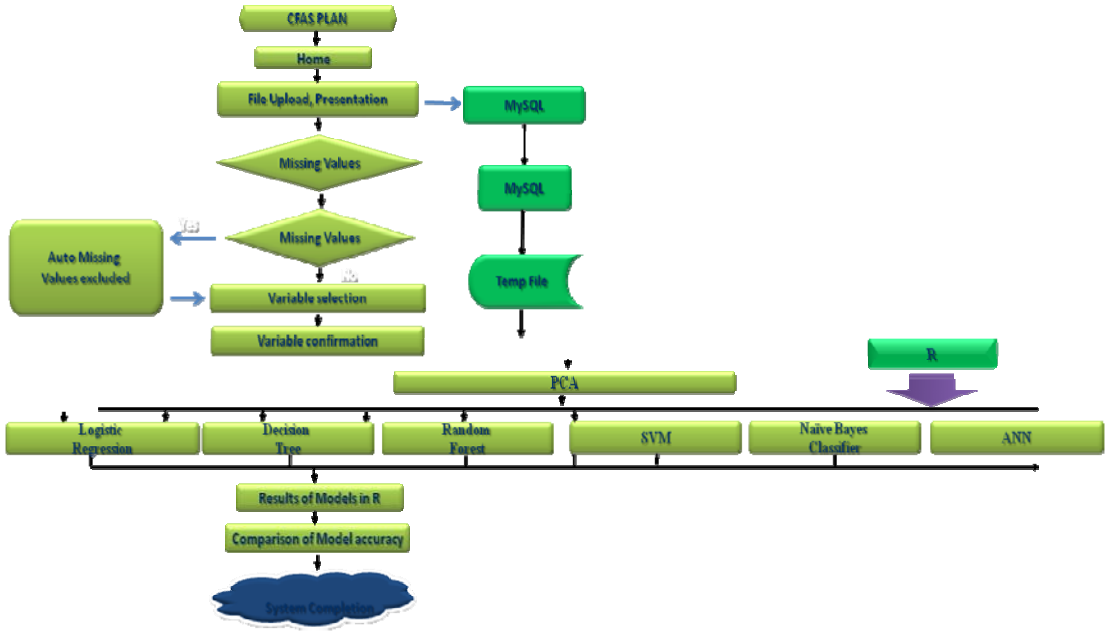

Figure3 The implementation process of elementary system

Platform for cloud computing. Cloud-R offers the convenience of users that users needn't to update software version, not limited to the performance of their computer, operating system version and troubling with file size. Cloud-R can share others the methods of statistical analysis and discuss the results. In the other word, if users can connect internet and use the browser of computer, they can implement the calculation of probability and statistics without constraints of time and space like commercial statistical software.

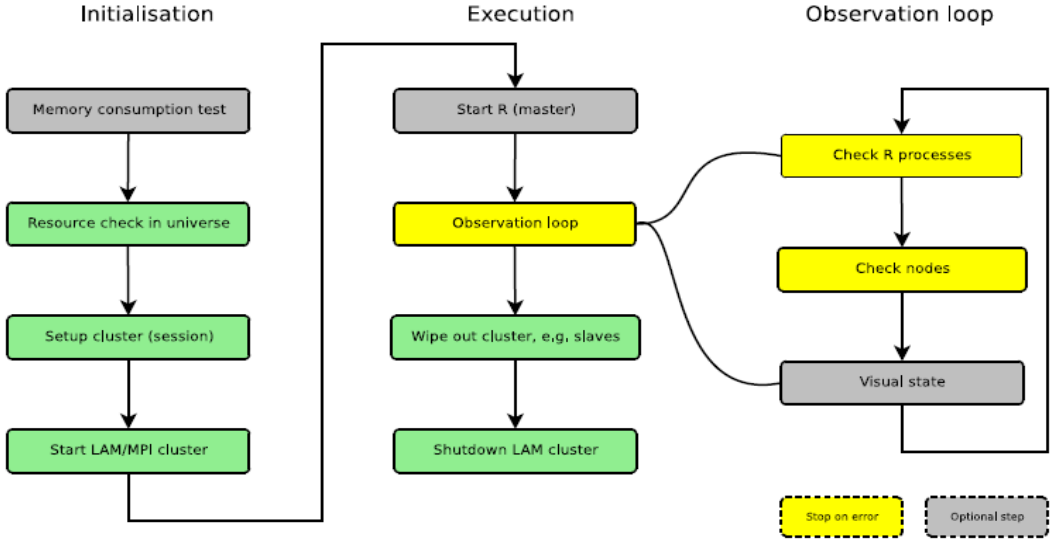

Figure4. Cloud R Site Map

TABLE 2
THE DIFFERENT CORRECT RATE OF PREDICTIVE VALUE OF Y (FINANCIAL CRISIS)

Accuracy / Analysis	Correct Classification		Recall		Precision		F-Measure	
	Train	Test	Train	Test	Train	Test	Train	Test
Logistic Regression	90.53 %	90.16 %	30%	36.36 %	68.18 %	80 %	41.67 %	50%
Decision Tree	88.66 %	87.28 %	0%	0%	NaN	NaN	NaN	NaN
Random Forests	89.55 %	91.16 %	42.11 %	20%	64.86 %	42.86 %	51.06 %	27.27 %
Support Vector Machines	91.81 %	88.08 %	29.79%	16%	100%	80%	45.90%	26.67%
Bayes	69.21 %	67.86 %	31.91%	28%	55.56%	70%	40.54%	40%
ANN	96.24%	85.26%	74.47%	32%	89.74%	47.06%	81.40%	38.10%

At present, the most important function of Cloud-R website is the function suite to support distributed computing. Cloud-R website has been pre-loaded networkspace suite which provides the provision of three independent computer nodes included a total of eight CPU core. So it makes the program of costing a lot of memory and computing power of the program distribute to each CPU core for operation. Cloud-R is expected to be able to make a variety of platforms and hardware architecture computers easily join the group, which allows users easily and quickly joined the operation in the desired node and hopes for more users added and enhanced the efficiency of operations.

The management tool sfCluster and the corresponding R package snowfall were developed. The two packages are designed to make parallel computing easier and more flexible. SfCluster completely hides the setup and handling of clusters from the user and monitors the execution of all parallel programs for problems affecting the cluster and machines. Together with snowfall it allows the use of parallel computing in R without further knowledge of cluster implementation and configuration.

We use parallel computing technology in the R and improve the Cloud-R with application of package snow and sfcluster. The improved Cloud-R makes the computing capacity more convenient, efficient and powerful for users, so that not only rapidly develops various fields required many statistical a analysis and

mathematical operations, but also further narrow the threshold between R software and users. In the course of establishing model, we can respectively input the program code needed to execute to the improved Cloud-R system in order to establish SME Risky Models. We can analyze the results from the model. The Improved Cloud-R platform also is as a basis platform for effective expansion.

III. Establish And Analysis Of Financial Risk Models

This study takes Taiwan Economic Journal (TEJ) financial information to construct a financial crisis risky model. We refer to financial crises literature of domestic and foreign scholars to construct crisis risky models. Then we choose the financial crisis definition, crisis risky model and the important variables of literature as a reference. A lot of scholars use logistic regression model to constructs a warning model that have a good predictability. Furthermore, univariate analysis and multivariate analysis are also used by some scholars. This study will add data mining methods that include decision tree and neural network as modeling reference. However, in the variable aspect, more people select liabilities ratio, return on assets and current ratio as the significant variable .This study will serve this variable as the basis. We explained variables by using the definition and formula. In this study, 27 variables selected are as follows.

TABLE 1
VARIABLES OF FINANCIAL RISK MODEL

Y : Major Categories Crisis	X15：Return on Equity –often continued interest
X1 : Current Ratio	X16：Debt Ratio%
X2 ： Accounts Receivable Turnover	X17：Growth Rate of Total Assets
X3：Gross Profit Ratio	X18：Net Growth Rate
X4：Operating Expense Ratio	X19：Net Turnover Times
X5：Sales Growth Rate	X20：Growth Rate of Return on Total Assets
X6：Quick Ratio	X21：The Net Asset Value of Each Share
X7：Inventory Turnover Times	X22：Turnover Per Share
X8：Operating Margin Growth Rate	X23：Total Liabilities / Total Net Worth
X9：Operating Profit Margin	X24：Cash Flow Ratio
X10：total assets turnover	X25 ： Pre-Tax Net Profit Margin
X11：Operating Profit Growth Rate	X26：After- Tax Net Profit Margin
X12：Debt/Eqity	
X13：Fixed Asset Turnover	
X14：Pre-tax Profit Growth Rate	

Independent variables X1 ~ X26, as continuous variables; dependent variable for the variables X27, for that matter whether this company crises have occurred, this variable is a discrete variable.

We can apply the variables in Table above to establish SME Risky Model. There are six methods of data mining to choose, including Logistic Model，Decision Tree，Random Forests，SVM，Bayesian Network，Artificial Neural network. For example, the results of the six models in the condition of financial crisis are as Table2.

Comparison of the correct rate of modeling the prediction value N used various methods (no financial crisis). ANN is the highest accuracy model of training data set for classification. Random forest comes in second; Bayesian network is the lowest. Random forest is the highest accuracy model of test data set for classification. Logistic regression comes in second; Bayesian network is also the lowest. ANN is the highest accuracy model of training data set for recall. Logistic regression comes in second; Bayesian network is the lowest. SVM is the highest accuracy model of test data set for recall. Logistic regression comes in second; Bayesian network is also the lowest. ANN is the highest precision model of training data set. Bayesian network come in second; the decision tree is the lowest. Random forest is the highest precision model of test data set. Logistic regression comes in second; SVM is the lowest. ANN is the highest F-Measure model of training data set. SVM come in second; Bayesian network is the lowest. Random forest is the highest F-Measure model of test data set. Logistic regression comes in second; Bayesian network is also the lowest.

IV. Conclusion

In the system, SME-related risk models are combined with cloud computing technologies. To the clouds in the Software as a Service (SaaS) approach to make a platform and then offered to the public a load is low, convenient and real-time analysis system as a reference. Before the financial crisis, the company can find early signs of the financial crisis and avoid the occurrence of collapse; companies, banks, investors and government agencies can be an early alert and prevention in order to reduce the probability of financial crises; investors also understand the risks of the company's future as a reference for investment decisions. Through the above description, we can know that many important applications of cloud computing in management, finance, information, and statistics, the previous researchers and developers have provided us with many ideas and direction, so we can develop our own systems based on their previous foundations. To provide the general public to use, the system is easy to understand how the system operation. This system provides summary description of each model, the advantages and disadvantages of each model, and analysis model can be used in what conditions. It also provides the explanation and formula of financial variables, which provided by this study. So users do not need be a statistical professional, and without download the statistical analysis software, as long as the link the Internet and then you will obtain the relevant financial information and financial risk prediction model.

ACKNOWLEDGEMENT

This work was Supported by Fundamental Research Funds for the Central Universities (2010221040) , Fujian Social Science Funds (2011C042) , National Bureau of Statistics Funds (2011LD002) and National Natural Science Foundation (71171001) from China.We would like to thank the editor, associate editor, and referees for careful review and insightful comments, which have led to significant improvement of the article.

REFERENCE

[1] A. Rossini, L. Tierney, and N. Li. Simple parallel statistical computing in R. Journal of Computational and Graphical Statistics, 16(2):399–420, 2007.

[2] Jochen Knaus, Christine Porzelius, Harald Binder, Guido Schwarzer. Easier Parallel Computing in R with snowfall and sfCluster. The R Journal Vol. 1/1, May 2009.

[3] M. J. Ray. Rcgi 4: Making web statistics even easier. R News, 1(1):20-21, January 2001.

[4] Richard Newton and Lorenz Wernisch. Rwui: A web application to create user friendly web interfaces for R scripts. R News, 7(2):32-35, October 2007.

[5] Uwe Ligges and John Fox. R Help Desk: How can I avoid this loop or make it faster. R News, 8(1):46-50, May 2008.

[6] Cloud-R: http://epigenomics.ncu.edu.tw/Cloud-R/index_tw.php

[7] Don Chamberlin. Encyclopedia of Database Systems 2009, Part 19, 2753-2760, DOI: 10.1007/978-0-387-39940-9_1091

[8] Elio Lozano . Parallel and Distributed Data Mining－R Wrappers for Message Passing. University of Puerto Rico Mayagüez Campus. 2010

[9] Jose R. Rios Viqueira and Nikos A. Lorentzos.SQL extension for spatio-temporal data. The VLDB Journal, 2007, Volume 16, Number 2, Pages 179-200