Mass Data Analysis and Forecasting Based on Cloud Computing

Tong Yang
School of Finance & Public Administration, Anhui University of Finance and Economics, China
Email: ivytong520@163.com

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

Jinrui Wei and Kuangnan Fang
Department of Statistics, School of Economics, Xiamen University, Fujian Province, 361005

***Abstract*—Owing to the less limitation of Cloud Computing, different fields (such as telecommunications, tour and medical industry etc.) can combine their industrial specialty and professional skills with Internet to construct many small clouds in Cloud Computing. In addition, like many gear wheels in the machines, many small clouds support a large cloud that is composed of different services to provide users with general-purpose applications in Cloud Computing. By the concept of Parallel Computing, we build a small cloud for statistical forecasting service (FaaS, Forecasting as a Service) with integrating Cloud Computing and data mining methods based on R, PHP and MySQL. The greatest advantage on R is that it is open and free. Moreover, R code can combine with PHP code to be a web page. Furthermore, it can solve the installation and extension problems by cloud computing that makes enterprises not that "heavy".**

***Index Terms*—Cloud Computing, Parallel Computing, Data Mining, FaaS, R, SQL**

I. Introduction

"We are drowning in information but starve for knowledge". Fortunately, however, data mining is a technique to understand and convert raw data into knowledge that unknown before. That means the typical task of data mining is to predict based on what we have. As a matter of fact, prediction plays an important part in modern people's life, such as the weather forecast, cost estimate in the financial area, the number of defective products in the productive area and etc. To extract information used for forecasting from dataset, we need to construct the forecasting models with the statistical software such as SPSS and SAS assistance.

However, there are some lacks of statistical software which is provided with predictive function in the market. Speaking of SPSS, it can only use prediction models in internal objects, which make limitations bigger. Although SAS software can use diversely forecast models, the software need to be written in the program language. It is not easy for beginners to use. In addition, the biggest problem to the software is that companies need to pay high rent cost and update the software every year. It is a problem for some small companies to pay the high costs. Furthermore, the statistical software needs a lot of computer space to be installed. As a result, if users pay the bill by the numbers of use on the software, which is needless to install and just need to connect the Internet, it would be a cost-effective choice. The concept will be achieved by Cloud Computing. In regard of advantages and shortcomings for statistical software, we want to have a try to improve these weaknesses and make it better.

R provides a variety of statistical technique. And the greatest advantage on R is that it is open and free. Moreover, R code can combine with PHP code to be a web page. Furthermore, it can solve the installation and extension problems by cloud computing that makes enterprises not that "heavy". So we build a small cloud for statistical forecasting service (FaaS, Forecasting as a Service) with integrating Cloud Computing and data mining methods based on R. Due to FaaS with the characteristics of distributed computing, this system can more quickly handle the huge data and reduce the requirement of the server. It is the point that the FaaS system we set up in this study different from TKU Net-Stat and Cloud-R statistical analysis website.

This study begins by taking a brief review of Cloud Computing and data mining. In section 3, we analysis the pros and cons of the forecasting models we incorporated, such as Regression, Logistic Regression, Time Series, Artificial Neural Network, Random Forest, Support Vector Machine, Multivariate Adaptive Regression Splines and etc. Before that, framework and steps of analysis have been given. We would show the forecasting interface structure and use database to illustrate in section 4. Finally we conclude with some reflections.

II. Cloud Computing

A. Cloud Computing

Corresponding author, Tong Yang,
Email: ivytong520@163.com.

doi:10.4304/jsw.7.10.2189-2195

In 2006, the term Cloud Computing is first mention by Google CEO Eric Schmidt. And maybe we could get some basic idea through following words: “It starts with the premise that the data services and architecture should be on servers. We call it "cloud computing" – they should be in a "cloud" somewhere. And that if you have the right kind of browser or the right kind of access, it doesn't matter whether you have a PC or a Mac or a mobile phone or a BlackBerry or what have you – or new devices still to be developed – you can get access to the cloud. There are a number of companies that have benefited from that. Obviously, Google, Yahoo!, eBay, Amazon come to people mind. The computation and the data and so forth are in the servers.” Cloud Computing is widely used in the lives, like Gmail which is an example of Google application in SaaS[*].

Strictly speaking, Cloud Computing is a model of enabling ubiquitous, convenient and on-demand network access to a shared pool of configurable computing resources (e.g. networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction according to NIST (National Institute of Standards and Technology). In brief, Cloud Computing means to transfer the processing core of work from the computer to the internet; so-called paradigm shift in IT industry. In addition, users have started storing the private data from own computer to the network space, namely the Internet server. Thus, all users can store and take their data through Internet no matter when and where. In other words, the data and applications on the Internet are as vast as the cloud. And anyplace users ignored in the real world could take the data and use the applications for free, that is Cloud Computing. (nikkei BP, 2010)

So the advantages of Cloud Computing is (1) Reducing the request of personal computer; (2) Reducing the cost; (3) Update software efficiently; (4) Sharing the resource among the three types of operation system: Windows, Mac OS and Linux; (5) Convenience access the documents; (6) Computing faster[†].

The idea of “Cloud” in Cloud Computing came from engineers often drew a cloud as a sketch plan of network. Through this idea, each application program on network could be seen as a cloud. And the size of cloud depends on the application scope. Moreover, according to NIST’s definition, “Cloud” could be classified into the following four kinds of subordinate models.

B. Public Cloud

Users could use Public Cloud service thorough Internet or the third party service suppliers. The goodness of Public Cloud is whippy and cost efficiency. But it should be notice that “public” is not equal “free”. The mean of public include free or very cheap charge. Moreover, Public Cloud not denotes user’s data may be looked over by anyone. Public Cloud suppliers often enforce use access control mechanism on users.

C. Private Cloud

To compare with Public Cloud there are more virtues in Private Cloud like flexibility and suitable to provide service. Moreover, Private Cloud makes suppliers and users get a better control on foundation configures, improve safety and flexibility. The difference between Public Cloud and Private Cloud is that all data and procedures in Private Cloud are managed in organization. Furthermore, it would not be affected by the internet bandwidth, safety concerns and legal restrictions.

D. Community Cloud

Community Cloud is controlled and used by organizations which have similar benefits such as particular safety demands and common purpose. Community members use the Cloud Computing data and application programs in community.

E. Hybrid Cloud

Hybrid Cloud is a combination of Public Cloud and Private Cloud. In this kind of mode, users often contract the non-critical information and process on public cloud. In the same time, they could also control the critical business information and data.

F. The Development of Cloud Computing

Recalling the time when computers are just became available in the market, we can find Cloud Computing an inevitable trend. Because the huge physical volume and expensive price, computer was much faraway for general public. Thomas J. Watson who was the IBM early president in United States once said,” I think there is a world market for about five computers.”He is not optimistic about a sales market of computer. Before1970s, computers were primarily for schools or corporate mainframes. That makes lots of users share a CPU resource at the same time. Till the promotion of personal computer in the 1980s and the application of Internet in later 1990s, computers are no longer faraway but closing the life. As a result computers have become an integral part of life.

To review the beginning of computers in early days, super computers was invented to take operating as main.The supercomputer means the host calculation with hign-performance and huge saving. Many people have the misunderstanding to the supercomputer because they think it is a large computer with fast and special CPU, and having thousand times ability than personal computer. However, current super computers are no longer a single shape of core mainframe like Cray supercomputer in 1980s, but converting to Cluster Computer with numerous CPU to constitute.

[*] The Service Applications of Cloud Computing is (1) SaaS (Software as a Service); (2)PaaS (Platform as a Service); (3)IaaS (Infrastructure as a Service)

[†] Because Cloud Computing adopts the mode that scatter type's operation, and this kind of mode would segment the originally huge data into several small data. And then, these partition behind of the data asunder store in several servers. Consequently, can scatter the work to several servers and carry out the work at the same time while carrying on the reading and storing of data. Thus, this kind of method makes the processing speed of file faster. (Michael Miller, 2009)

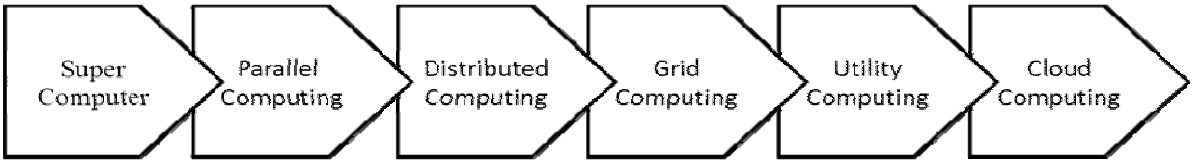

Figure 1. Evolution Process of Cloud Computing

In the early 1990s, because of the GNU of Free Software Foundation (FSF) and Linus's contributions, it raised the first high-speed computing revolution that PC Cluster defeated a super computer by slightly strikes in a big way. It reduced the computational cost so that Parallel computing suddenly became famous. After late 1990s, the network gradually applied in financing institution and the distributed object technology gradually became popular. The project "finding aliens" was raised by SETI@Home in early 2000 that lending the computing resources from all over the world that started the new page of Distributed Computing. Distributed Computing refers to divide the large work into small parts and compute tasks by more than one computer at a same time then integrate the results. Distributed Computing can finish the work which single computer is not able to load.

Immediately, Global National Center followed the essence of Distributed Computing to promote Grid Computing. Grid Computing extended from Distributed Computing means integrating a variety of different platforms, different structure and different computer levels by Distributed Computing. Grid Computing means processing data scattered everywhere with the public basis. Utility Computing mainly promote a kind of ideal enterprise information configures that makes IT service a concept of user charges like water, electricity and gas. Using more and having more charges. In 2006, Amazon announced Elastic Compute Cloud (EC2) that successfully established the on-demand computing service and started the prelude of Cloud Computing. After the process of operations, Cloud Computing have inherited the Cluster technology of Parallel Computing, bent the fault-tolerance feature of Distributed Computing, and then developed the new thinking "Data Center as a Computer."

Our research tried to build an R system forecasting and put it on cloud by the concept of Parallel Computing. As previous mentioned, parallel computing refers to the ability to use multiple CPUs concurrently to perform calculations that would otherwise be carried out sequentially in a single CPU. Moreover, there are two main approaches for parallel processing. One is the sharing memory machines which can offers programming simplicity. The other one is distributing memory machines. For our R system, there are several available R libraries like package "snow" that allow us using multiple CPUs with minimal changes. Snow is a package for simple parallel computing framework. Parallel computing performs local data analysis and sends partial results to other sites. There are small fractions of data moving between sites that saving the communication overhead. (Elio Lozano, 2010) We could use a large number of similar specifications of computers as the server to perform the program operation. Therefore, we would reduce the basic structure of parallel computing. It could coordinate the information exchange between computers easier and distribute processing performance better.

III. IMMEDIATELY ANALYSIS WEB SITES

A. *Cloud-R*

Cloud-R is a website founded by the graduate student of National Central University named Kun-hsien Lin. The purpose to found Cloud-R is that improving the weakness of R software. In the framework of Cloud-R, R software will be expanded into a web service so that the users are able to easily use the R software. Cloud-R website not only provides users with improved web interface for R but also reduces computing time on the statistical analysis.

First of all, Cloud-R website is on the basis of R software. It is able to make a variety of statistical processing of data and analyze the real-time graphics. What's more, users can expend the ability to operate by themselves with clouding computing. Cloud-R website realizes the possibility which operating the R software by Web browser. Second, people who use Cloud-R Website do not constantly update the version, also not be limited to their computers' performance and the file size, and so on. Users only need to use a browser to connect the internet, then enjoying the convenience of R statistical software.Last but not least, the specialty of Cloud-R website is a registration system. When people log on the system, they will enter the exclusive page which belongs to the affiliate. The affiliate can have their store place. Therefore, user can use Cloud-R website to complete the program and to save the project.

B. *R-php*

R-php is a project developed inside the Department of Statistical and Mathematical Sciences "Silvio Vianelli" of the University of Palermo (Italy) and considered as target the realization of web-oriented statistical software. R-php is an open-source project with the code released by the authors and can be freely installed.

The idea to project a statistical software that can be used through Internet comes out from the following considerations: it is an ascertained fact that the growing spread of Internet and the request of new services from its users has changed and still is radically changing the way to access the daily use structures by now the most of information and services goes through the Web and the software philosophy goes to the same direction; In general, the trend on producing software that are installed on a single computer is going down in favor of software that can be used through a connection to the Internet and

a browser.

The main distinction of R-php is use R Language to do statistical analyses; moreover the different between R-php and other R Language web projects consists in the presence of an interactive module (R-php point-and-click) inside R-php. That could let users even do not understand R Language very well also could do statistical analyze.

IV. FORECASTING MODELS OF DATA MINING

A. Outline of Constructing Forecasting Models

Data Mining means that the huge and complicated data on the past history is analyzed, summarized and integrated by a variety of analytical methods and techniques for extracting useful information. In a word, Data Mining is to find out the message hidden in the data and provide enterprise management with the references when making decisions. The function of the Data Mining can be divided into five types: Classification, Estimation, Prediction, Affinity grouping and Clustering. In this study, the main direction is Prediction.

The different professional fields have the different Data Mining processes even if the same industry also causes the obvious difference because of using the different analytical methods and knowing the specialized knowledge in varying degrees. Therefore, the standardization and systematization of Data Mining processes are particularly important. The scholars propose the different opinions about the flow chart in which CRISP-DM (Cross Industry Standard Process for Data Mining) and KDD Process (Knowledge Discovery in Databases Process) are the most representative. These two processes of data mining are shown in TABLE Ⅰ.

CRISP-DM emphasizes the implementation of Data Mining methods and procedures from the methodology point. So CRISP-DM is often used to deal with business issues in the business community. Its flow chart is following. KKD Process is a process of discovering knowledge that excavates useful, novel and previously undiscovered knowledge which exists originally in the data form a lot of data by some statistical system software or others. This study focuses on the data forecasting and it does not use the step of Business Understanding, so we use Data Mining KDD Process. At First, we should understand the data and relevant knowledge, and then establish the target dataset. After establishing the target dataset, we select the required data in the target dataset. And then we chose the more appropriate algorithm by the data of the features after dealing with the data; and then by Data Mining technology to build the model. Finally, we should determine that the quality of the model through the interpretation and evaluation.

TABLE Ⅰ
THE DATA MINING PROCESSES

Step Name	CRISP-DM	KDD Process
Step Process	1. Business Understanding 2. Data Understanding 3. Data Preparation 4. Modeling 5. Evaluation 6. Deployment	1. Selection 2. Pre-processing 3. Transformation 4. Data Mining 5. Interpretation/Evaluation

B. Notes on Forecasting Models

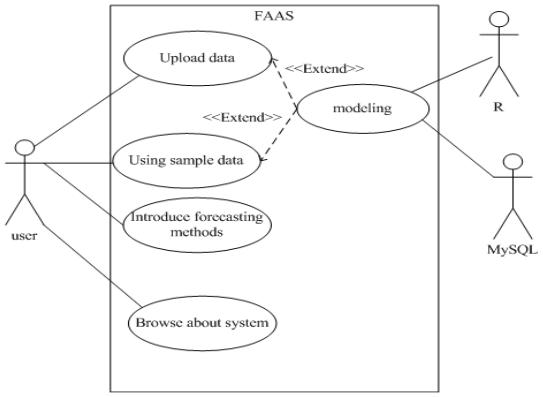

Figure 2. Activity Diagram

We selected seven common forecasting methods. There are Regression, Logistic Regression, Time Series, Artificial Neural Network, Random Forest, Support Vector Machine and Multivariate Adaptive Regression Splines (MARS). Noteworthily, you should keep your eyes open on the flexibility and restriction of these methods. For instance, Regression may encounter the collinearity among variables; Logistic Regression could only deal with the dataset that the dependent variable is nominal. when confronts with large changes in the outside world, a greater deviation may occur in Time Series; Artificial Neural Network has poor explanatory capability; Random Forest with non-uniform training dataset, the accuracy of classification is lower. Using Support Vector Machine, the core function must be consistent with the Mercer's condition; the number of explanation equations of MARS has not the certain criteria.

Ⅳ. INTERFACE STRUCTURE AND ILLUSTRATION

A. Interface Structure

The interface structure is divided into four parts: (1) Forecasting Methods Introduce; (2) Upload Data; (3) Sample Data; (4) Forecasting System Introduce. Forecasting Methods Introduce include simple introduction of seven forecasting methods in this forecasting system. User could know the data type of

these forecasting methods. There are similar structures in those two parts: Upload Data and Sample Data. The different of these two parts is user could upload their data, and then the upload data would be duplicated and stored in distant server. The part of Example Data is set up inside dataset. The purpose is let user know how to use this system without upload any data. Following, user chooses forecasting methods and gets the result. Finally, user could know the forecasting result.

Following figure 4-1 shows the case diagram. The main purpose of using case diagram is to show what system functions are performed. This research setting 1 actor as "user" and our research define 4 main use cases: Upload data, Using sample data, Introduce forecasting methods and Browse about system. Upload data and Using sample data extend another function- Modeling. Modeling makes R language and MySQL as a support actor, that users can choose it to analysis and predict data what they want.

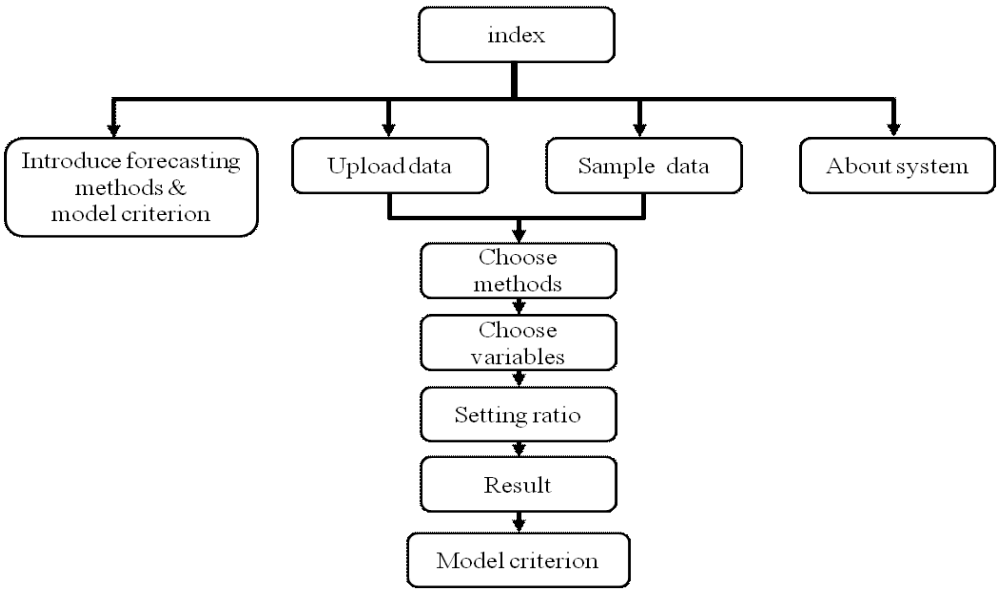

Figure 3. Flow Chart

Figure2 is an activity diagram which shows the relationship to using cases. Generally speaking, action diagram contains two points of view: System and Function. This research does not study in-depth because Function viewpoint focuses on more detailed features of the internal operation of the computer.

B. R Language

R Language is an integrated data processing and statistical computing software which can be downloaded and installed from official website of R Language (http://www.r-project.org/). This software was jointly developed by Ross Ihaka and Robert Gentleman in the University of Auckland, New Zealand. That is why it origin name called "R". Now the "R Development Core Team" takes the charge of development. R Language is mainly followed by creation of the S Language which has a common nature with S-plus Language. The different is that the output of R Language shows at least message. The operation of R is similar to MATLAB, Visual Basic, Java Script, and the syntax is also similar to C Language. Above all, R is free and different from statistical software such as SAS and SPSS that cost money. R can also save the output to an object and provide the follow-up calculations. The significant features of R Language is open-source, excellent statistical analysis and mapping capability. Because it is open-source, users can set necessary instructions and they also can write in C language, PHP and other languages. Thus, its program writing is considerable flexibility. There are complete extensions (packages) allow people to install on the website. These extensions are not only in the field of statistical analysis also in the field of financial analysis, ecology and dynamic computing, etc. Therefore, the reason of our research using R Language is free and excellent statistical functions. Through packages and custom orders of R, our research could integrate forecasting models required.

C. PHP

The official name of PHP is Hypertext Preprocessor. PHP is also called ''Personal Homepage Program'' by some people. PHP was established in 1995 by Rasmus Lerdorf and it was originally used to track the individual resume are viewed. Now, PHP has promoted from the Personal Homepage to the Hypertext Preprocessor because the use and performance of PHP is constantly upgrading.PHP is a CGI program and it uses to handle interaction between Internet users and server. PHP is the embedded language of HTML so it is particularly suitable for developing web application. PHP is the server-side and there is the cross-platform technology in PHP. This means that every action of PHP makes on the Server and PHP can perform on most of the work platform. There are three major advantages in PHP: (1) Powerful integration of the database; (2) Free system; (3) Open source code.

D. MySQL

Database system is the main program to store data. It is

mainly used to manage the information list. Information may come from many different sources, such as research projects, transaction data, customer demands, and so on. Database is like a large filing cabinet which uses electronic ways to maintain the data records. MySQL database system as the main database system in the market, you can easily solve the problem as following: (1) Reducing the records modifying time; (2) Reducing the records filtering time; (3) Getting accessible to save records simultaneously; (4) Setting records order flexible; (5) Output formats are flexible; (6) Remote accessing and transferring records electronically.

E.The Implementation of the Database

Weather the ideal way mentioned can achieve the actual functions or not. People usually want the actual job easy to attain, and MySQL actually reach a demand. Weather MySQL is available to use on the web or not. MySQL does not have the function to save in directly, but users can use other tools to accomplish this function, such as Perl program language, PHP program language. Two languages can be used on MySQL database interface.

Perl has powerful text processing capabilities. It uses the "RTF" format, which is Rich Text Format. The format can handle by all word processors. PHP is used to writeweb programs and allow the website and database to interact. Moreover, PHP has a good interaction with Apache which is the most widely used web server in the world. Therefore, users can easily to display the search results.

The last important issue to be considered is the cost of MySQL. Generally to speak, the price of database system is expensive while MySQL the same as other tools(Perl、DBI、PHP、Apache)usually be free. Only in some case, MySQL needs to have permission. Therefore, establishing a database such as MySQL does not require much money.

F. Illustration

Our research establishes a statistical predict platform in cloud using php, MySQL and R language. The following figure 3 is the system interior flow chart. In the following, we use some examples to introduce FaaS system in detail. Manufacturing data is the industrial index information in Taiwan from January in 1982 to December in 2001. The variables are time, total, food, can, beer and drink. In terms of Continuous Data Analysis, “total” is a dependent variable. “Food, can, beer, and drink” are independent variables. Therefore, it means that using the values which contain the industries of food, can, beer and drink to predict total industrial index in Taiwan.

By following the system procedures, users can gain the forecasting value with different forecasting methods. TABLE Ⅱ list models we chose. In terms of model criteria in the Manufacturing data, we can learn that the smaller the value in MAE, MAPE, and MSE is, the better the model is. Clearly, according to the result, it shows that MARS model is the best. Linear Regression model is second-best. And then, ANN model is third, which better than SVM model.

TABLE Ⅱ
THE COMPARISON OF CONTINUOUS MODELS

Model	MAE	MAPE	MSE
Regression	10.815	16.0379	175.4288
NN	14.2332	17.4032	337.0135
SVM	15.2476	19.9448	361.3946
MARS	9.2158	11.9961	125.8398

Housing credit is the data records housing credit default information. The variables are NO, RiskLevel, overdue, Default, area, Annual income, Education, sex, Maritalstatus, and job.In terms of nominal Data Analysis, “Default” is a dependent variable. “RiskLevel, overdue, Default, area, Annual income, Education, sex, Maritalstatus, and job” are independent variables. Therefore, it means that using the values in the RiskLevel, overdue, Default, area, Annual income, Education, sex, Maritalstatus, and job to predict whether default in the housing credict.

TABLE Ⅲ
THE COMPARISON OF NOMINAL MODELS

Model	Accuracy	Recall	Precision	F-measure
LR	99.0196	0.9968	0.9905	0.9936
NN	83.7975	0.8342	0.9969	0.9083
RF	97.3913	0.9805	0.986	0.9832
SVM	97.7169	0.9857	0.9857	0.9857

TABLEⅢ lists the comparison of nominal models we chose. In terms of model criteria in the Housing credit data, we can learn that the bigger the value in Accuracy, Recall, Precision, and F-measure is, the better the model is. Clearly, according to the result, it shows that Logistic Regression model is the best. SVM model is second-best. And then, Random Forest model is the third, which is better than ANN model.

V. CONCLUSIONS AND FUTURE WORK

This study sets up a web site combined cloud computing with data mining analysis in the forecasting service, called Forecasting as a Service (FaaS), which provides forecasting services for users. The FaaS system, written by the PHP and R language, is based on Information & Intelligence as a Service, called IIaaS or I2aaS, which is Software as a Service (SaaS) extension. The FaaS system can cut the original huge data into several data, and then let these segmented data spread stored in multiple servers by parallel computing in cloud computing. Next, it uses data mining methods to do data processing, analyzing, modeling and model evaluating. Afterwards, it allows the information to come back the

user's computer though the Internet. With this system, users not only do not need to install the statistical software, but also do not need to think about the level of computer performance, the compatibility of operating systems, and the size of data file. After uploading the data or using the sample data file, users can take advantage of a lot of Data Mining methods to predict.

The FaaS site, set up in this study, is different from the previous mentioned of TKU Net-Stat and Cloud-R statistical analysis website. It improves a weakness that the TKU Net-Stat cannot handle large volumes of data. In term to the part, the study improves the weakness and increases one analyzing methods. For one thing, analysis can be used for a variety of methods to deal with data at the same time. For another, MARS analysis is increased and the value of multiple model criterion forms organized into user-friendly interface to find the best model. Moreover, the study also improves the shortcomings of Cloud-R, users have to learn R language and know how to use R. The study hopes to provide a more convenient way to users to predict the data.

FaaS established by this study still exists something to be improved. The directions which can be improved are following: (1) At present, the system has been able to stored data in the remote database server but the membership system has not been established. (2) FaaS provides the users with the interface of Chinese, so the users may be only those who know Chinese. (3) The seven forecasting methods could be developed and expanded.

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. Sulaiman, S. Mukkamala and A. Sung. SQL infections through RFID. Journal in Computer Virology, 2008, Vol. 4, Number 4, Pages 347-356

[2] B.Cao and A.Badia. Exploiting maximal redundancy to optimize SQL queries. Knowledge and Information Systems, 2009, Vol. 20, No.2, Pages 187-220

[3] Cloud Computing Use Case Discussion Group. The latest version (4.0) of the white paper. NIST. 2010

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

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

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

[7] J. R. Viqueira and N. A. Lorentzos.SQL extension for spatio-temporal data. The VLDB Journal, 2007, Vol. 16, No. 2, Pages 179-200

[8] L., Gordon. Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management. JOHN WILEY & SONS INC. 2004

[9] M. F. Hornick、E. MarcadÃ 、S. Venkayala. Java Data Mining: Strategy, Standard, and Practice: A Practical Guide for architecture, design, and implementation Morgan Kaufmann 2006

[10] M. Zandstra. PHP Objects, Patterns, and Practice. APRESS. 2007

[11] M.Miller. CloudComputi：Web-Based Applications That Change the Way You Work and Collaborate Online. QUE Publishing. 2009

[12] P. Lakhani, E. D. Menschik, A. F. Goldszal, J.P. Murray and M. G. Weiner, et al. Development and Validation of Queries Using Structured Query Language (SQL) to Determine the Utilization of Comparison Imaging in Radiology Reports Stored on PACS. Journal of Digital Imaging, 2006, Vol. 19, No. 1, Pages 52-68

[13] P. Perner. Advances in Data Mining: Applications in E-Commerce, Medicine, and Knowledge Management Springer; 1 edition. 2002

[14] R language: http://www.r-project.org/index.html

[15] R. Meo, G. Psaila and S. Ceri. An Extension to SQL for Mining Association Rules. Data Mining and Knowledge Discovery, 1998, Vol. 2, No. 2, Pages 195-224

[16] R. Ando, K. Byung and Y. Kadobayashi. Log Analysis of Exploitation in Cloud Computing Environment Using Automated Reasoning. Lecture Notes in Computer Science, 2010, Vol. 6444/2010, 337-343, DOI: 10.1007/978-3-642-17534-3_41

[17] S. Frischbier and I.Petrov.Aspects of Data-Intensive Cloud Computing. Lecture Notes in Computer Science, 2010, Vol. 6462, From Active Data Management to Event-Based Systems and More, Pages 57-77

[18] S., Claude. Data Mining with Microsoft SQL Server 2000 Technical Reference. MICROSOFT PR. 2005

[19] StatSoft The creators of STATISTICA: http://www.statsoft.com.