The Classification Model of Affecting the Enterprise Usage on Cloud Computing Based on Decision Tree

Heyong Wang* and Jinjiong Lan

College of E-Business, South China University of Technology, 510006, China

Abstract: with the rapid development of information technology, cloud computing has become one of the most important trends of future development of information technology and gained considerable attention and applications. In order to accelerate the pace of the commercialization process of cloud computing, this paper studies the factors affecting the commercialization of cloud computing and Decision tree C5.0 is adopted to study the classification model for analyzing enterprise usage on cloud computing. Contrasted with the other three typical classification model, Decision tree C5.0 model is verified as the most suitable and stable model to predict whether an enterprise will use cloud computing.

Keywords: Cloud computing, C5.0 model, Decision tree, Index system.

1. INTRODUCTION

Along with the rapid development of IT technology, cloud computing as the future direction of the development of computer technology, is receiving considerable attention. Cloud computing [1] is a model for enabling omnipresent, convenient, on-demand network access to the pool of shared resources. Cloud computing obtains the required hardware, software and platform resources services through the network by on-demand and scalable way. For users, resources in the cloud appear to expand at any time, infinitely scalable, pay-per-use and available all the time if needed, like the use of water and electricity. Now the world's leading IT companies are implementing and promoting cloud computing, the superiority of cloud computing services are being recognized gradually by the enterprises.

Accordance with the type of cloud computing services, cloud computing has the following characteristics [2, 3]:

- Flexible service. The scale of the service can quickly expand or contract to adapt to the dynamic changes of the traffic load automatically.
- On-demand service. Provide users with applications, data storage, infrastructure and other resources as a service, allocate resources automatically according to users' needs.
- Service can be billed. It can monitor user's resource usage and services billings are based on the use of resources.
- Extremely cheap. A large number of enterprises have no burden of increasingly high cost of data center management with automation and centralized management and the versatility can make resource utilization increasingly significantly, etc,.

Cloud computing has developed rapidly in recent years, but its application is not as widely accepted as expected, because cloud computing is still in its early stage of development. Cloud computing is facing many security threats such as the lack of legal norms, which make cloud computing hard to promote to enterprises. This paper studies the factors affecting the application of cloud computing to find what impedes enterprises' deployment of cloud computing and give the corresponding solutions.

2. SELECTION OF THE EVALUATION INDEX SYSTEM

When selecting indicators for the evaluation index system, general principles should not only be followed such as scientific, operability and so on, but also take the specific circumstances of cloud computing technology into account. Therefore, this paper summarizes some factors which may affect the acceptance of cloud computing by an enterprise and a set of study variables are gained [4-12]

According to the Table **1**, the index system is established by the following analytical framework:

Due to the acceptance differences of cloud computing, the nature of the enterprise is considered on cloud computing deployment as a factor. For example, IT enterprise always pays attention to the development of new technologies, their awareness and acceptance of cloud computing is relatively high; while the enterprise of manufacturing is limited to the nature of their enterprise, their awareness and acceptance of cloud computing are relatively low.

- The awareness of enterprise for cloud computing will affect the application and popularity of cloud computing.
- The cloud computing services take on diversified development, which makes the enterprise concern about compatibility and interoperability of the cloud computing services.

Table 1. The integration of study variables.

Factors Considered	Content of Factors	Author/Reference
The profit of cloud computing	Reduce costs, increase flexibility and improve the level of information technology [4-6]	Catteddu D *et al.* 2010 [4] Misra S C *et al.* 2011 Etro F .2011 [5] William Y Chang *et al.* 2010 [6]
The cost of Self-built Cloud computing platform	Including hardware, software, human resources operations and maintenance fees and so on	Zhou qiang 2010 [7] Khajeh Hosseini A *et al.* 2012 [8]
The nature of enterprise	The time enterprise established , industry and size	Yu C S *et al.* 2009 [9] Low C *et al.* 2011 [10] Gary Garrison *et al.* 2012 [11]
Enterprise's attitude for cloud computing	Level of knowledge of the information technology and the intensity of the attention and support toward the information technology	Low C *et al.* 2011 [10] Gary Garrison *et al.* 2012[11]
The data security in cloud computing	The issues of data storage and data privacy	Catteddu D *et al.* 2010 [4] Armbrust M *et al.* 2010 [12]
The standards of cloud computing platform	The migration of data between different cloud platforms, standards of different cloud platforms are inconsistent	Yu C S *et al.* 2009 [9] Armbrust M *et al.* 2010 [12]
The capability of cloud computing service	The stability and reliability of the platforms to meet the needs of enterprises	Catteddu D *et al.* 2010 [4] Low C *et al.* 2011 [10]

- Cloud computing is faced with many challenges, such as the security issues [13], the qualifications of the service provider and network security. The existence of these challenges undoubtedly hampers the commercialization of cloud computing in enterprises.

Therefore, an evaluation index system is constructed for enterprise applications of cloud computing, based on the above analysis of influencing factors. The valuation index system is composed of three major sectors [14]: "the internal type of factors in enterprise", the external type of factors of enterprise" and "the status of deployment of cloud computing in enterprise" constitute the level index; "the type of property of enterprise", "the type of consciousness about cloud computing", the type of cloud construction" and "the type of risk of the platform of cloud computing" form the secondary index and twelve indicators are selected to constitute the tertiary index [15],which are shown as Table **2**.

3. QUANTITATIVE METHODS FOR THE EVALUATION INDEX SYSTEM

In the process of establishing analytical model, some quantitative methods are commonly used to quantify the non-numerical data, such as: Comprehensive Index, numerical methods, the dummy variable method, Likert scale method, fuzzy mathematics method. Taking the actual situation of the research object into account, dummy variables, comprehensive Index and Likert scale method are used to quantify the specific indicators, as shown in Table **3**.

3.1. The Dependent Variable

The status of deployment of cloud computing in enterprise: According to The status of deployment of cloud computing in enterprise, Likert scale method is used to quantify the dependent variable, as shown in Table **4**.

3.2. The Type of Property of Enterprise

According to whether there is an association to the information technology industry or not, dummy variable method is used to quantify variable industry: If the enterprise belongs to the industry of transmission of information, software and information technology, the value of 1 is assigned 1 to it or else it is assigned the 0 value

3.3. The Type of Consciousness About Cloud Computing

The awareness on cloud computing: According to the cognitive level of enterprise on cloud computing, Likert scale method is used to quantify this variable, as shown in Table **5**.

The attitude toward self-built cloud: According to whether the enterprise is in favor of the self-built cloud, Likert scale method is used to quantify this variable: If the enterprise approves self-built cloud, it is assigned a value of 1 or else it is assigned 0.

Table 2. Evaluation index system of affecting the enterprise' usage of cloud computing.

	Secondary Index	*Tertiary index*		*Content of Index*
		Notation	*Name of Index*	
The internal type of factors	The type of property of enterprise	X1	Industry	1.The industry of transmission of information, software and information technology 2. Not included in the industry of transmission of information, software and information technology
		X2	Size of enterprise	1.Less than 100 employees 2.Between 101 and 500 3.Between 501 and 1000 4.More than 1000
		X3	The time enterprise established	1. Before 2000 2. Between 2000 and 2003 3. Between 2004 and 2007 4. Between 2008 and 2011
	The type of consciousness	X4	The awareness for cloud computing	1.Completely not heard of cloud computing 2.Ordinary technology 3.Pay close attention
		X5	The attitude toward self-built cloud	1.Should not be 2.Should be
	The type of cloud construction	X6	The profit of deploying cloud computing	1.Improve efficiency 2.Resources sharing between different devices 3.Safety, do not have to worry about data loss, virus attack 4.Save hardware investment 5.More flexible configuration and management 6.Improve equipment utilization, reduce costs and save energy
		X7	The cost of self-built cloud	1.The initial investment cost 2.The high technical requirements in self-built cloud 3.Very difficult to maintain self-built cloud 4.Using the services from well-known large companies is more cost-effective
		X8	The kind of cloud computing services	1.Only IaaS 2. Only PaaS 3.Only SaaS 4.IaaS and PaaS 5.PaaS and SaaS 6.IaaS、PaaS and SaaS
The external type of factors	The type of risk of the platform	X9	Employee's attitude toward cloud computing	1.Just a gimmick 2.Inevitable trend 3.Epoch-making revolution
		X10	The security issues of cloud computing	1.Availability and accessibility of data 2.The integrity of data 3.Data privacy 4.Recoverability of the data 5.The process of long-term data

Table 2.contd…

	Secondary Index	*Tertiary index*		*Content of Index*
		Notation	*Name of Index*	
		X11	The standards issues of different platforms	1.The issues of data compatibility, using different cloud computing services generates incompatible data 2.The issue of sustainability of cloud computing services, especially the services using the standards of other country 3.Lack of unified guidance for self-built cloud computing center 4.Hard to guarantee the quality of the services of cloud computing
		X12	The technical performance of the cloud platform	1.The maturity ,reliability and safety of technology of cloud computing provider 2.Physical security of Cloud platform 3.Network security of cloud computing provider
The deployment status		Y		1. Not scheduled 2. Under consideration 3. Have been deployed 4. Being used

Table 3. Quantitative methods.

Secondary Index	Name of Tertiary Index	Quantitative Methods
The dependent variable	The status of deployment of cloud computing in enterprise	Likert scale method
The type of property of enterprise	Industry	dummy variable method
The type of consciousness	The awareness for cloud computing	Likert scale method
	The attitude toward self-built cloud	dummy variable method
The type of cloud construction	The kind of cloud computing services	Likert scale method
	The profit of deploying cloud computing	comprehensive Index method
	The cost of self-built cloud	comprehensive Index method
The type of risk of the platform of cloud computing	Employee’s attitude toward cloud computing	Likert scale method
	The security issues of cloud computing	comprehensive Index method
	The standards issues of different platforms	comprehensive Index method
	The technical performance of the cloud platform	comprehensive Index method

Table 4. Quantitative methods of the dependent variable.

The State of the Variable	Quantification of State
Not scheduled	0
Under consideration	1
Have been deployed	2
Being used	3

Table 5. Quantitative methods of the type of consciousness about cloud computing.

The State of the Variable	Quantification of State
Completely not heard of cloud computing	1
Ordinary technology	2
Pay close attention	3

Table 6. Quantitative methods of the type of cloud construction.

The State of the Variable	Quantification of State
IaaS	1
PaaS	2
SaaS	3
IaaS and PaaS	4
PaaS and SaaS	5
IaaS、PaaS and SaaS	6

Table 7. Quantitative methods of the type of risk of the platform of cloud computing.

The State of the Variable	Quantification of State
Just a gimmick	0
Inevitable trend	1
Epoch-making revolution	2

3.4. The Type of Cloud Construction

The kind of cloud computing services: According to the mix of service types of enterprises in adopting cloud computing, Likert scale method is used to quantify this variable, as shown in Table **6**.

Comprehensive Index method is used to quantify two variables: The profit of deploying cloud computing and the cost of self-built cloud. Every specific indicator accounts for one point. According to the specific content selection of variable, if a specific indicator has been chosen then the index will add one point.

3.5. The Type of Risk of the Platform of Cloud Computing

Employee's attitude toward cloud computing: according to the varying degrees of status of employees' cloud identity, Likert scale method is used to quantify this variable, as shown in Table **7**.

Finally, comprehensive Index method is used to quantify three variables: the security issues of cloud computing, the standards issues of different platforms and the technical performance of the cloud platform. Similarly, every specific indicator accounts for one point. If a specific indicator has been chosen, the index will add one point.

4. DECISION TREE C5.0

Decision Tree is the learning from class labeled training sample, a decision tree is a flow chart like structure. Essentially, decision tree is generated by utilizing Information theory to induct and analyze the attributes of a large number of samples. Decision Tree consists of three major components: root node, internal node and leaf node. The root node is the topmost node in tree, includes all the information of the samples; each internal (non-leaf) node denotes a test on an attribute, each branch represents an outcome of the test and each leaf (or terminal) node holds a class label [16].

Decision tree is a form of knowledge representation and a high-level overview of all sample data. The goal of generating a decision tree is to create a model that predicts the value of a target variable based on several input variables. When constructing decision tree, the algorithms usually works top-down by choosing a variable at each step that is the next best variable to use in splitting the set of items. "Best" is defined by how well the variable splits the set into homogeneous subsets that have the same value as the target variable. Different algorithms use different formulae for measuring "best". C5.0 algorithm takes advantage of the principle of maximizing gain ratio to chose characteristic attributes. C5.0 algorithm makes improvements on two main defects of ID3 algorithm: (1) Overcoming the inadequacies of ID3 algorithm, which tends to select the attribute of multi-value prop-

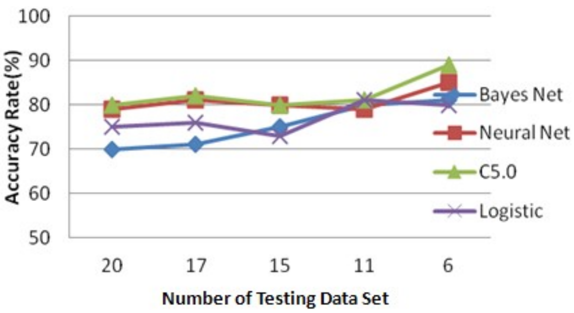

Fig. (1). Four classification models' forecast accuracy comparisons with the reduction of the number of test data.

erty using information gained by GAIN (D, S). (2) Being able to complete the process of discretization of continuous attributes, overcoming the inadequacies of ID3 algorithm, which can only deal with discrete variables. The formulae of C5.0 algorithm is described below [17]:

Assuming S as a data set, category sets as {C_1, C_2,…,C_c}, select an attribute A to make data set S divided into a plurality of subsets and then assuming attribute A including k non-overlap value {A_1, A_2, …A_k} ,and then:

$$E(S) = -\sum_{i=1}^{c} p_i \log_2 p_i \qquad (3.1)$$

$$E(S,A) = \sum_{v \in Values(A)} \frac{S_v}{S} E(S_v) \qquad (3.2)$$

$$G(A) \ldots Gain(S,A) = E(S) - E(S,A) \qquad (3.3)$$

$$\mathrm{I}(\mathrm{A} = \mathrm{a}_j) = -\sum_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{p}_{\mathrm{ij}} \log_2 p_{ij} \qquad (3.4)$$

$$MAX(Gain - ratio) = MAX(Gain(A) / I(A)) \qquad (3.5)$$

As it can be seen from the above formula, C5.0 algorithm will calculate the amount of information generated by the branch in each step of selecting the split node and then selects the optimal partition node to maximize gain ratio, which enhances the readability of the results and greatly enhances the accuracy of predictions of the model at the same time.

5. EXPERIMENT

5.1. Data Sources

The data was from questionnaires. The data of the survey indicators is obtained by mail invitation, mutual-filling questionnaires and publishing questionnaires on the professional cloud computing forums. 111 questionnaires are received of which 100 questionnaires were valid

5.2. Experiment

The data obtained by the questionnaire were preprocessed in accordance with the quantitative methods as described above. In the experimental part, in order to verify the robustness of the model, the classification test on the data set were carried out several times. The data is divided into 80 training data and 20 test data sets. As shown in the Fig. (**1**), the horizontal axis represents the number of the test sets, and the vertical axis is the accuracy of the test data.

In order to verify the test results of the decision tree model in this paper, four classification models are selected: Neural network model, Logistic regression model, Bayes net model and Bayesian network model. The same training data is used to generate the four models and the same test data for the model evaluation. The effect trend of the prediction accuracy of each model is shown in Fig. (**1**):

Gain Chart is used to further assess the stability of the four models. Specifically, the chart summarizes the utility that one can expect by using the respective predictive models, as compared to using baseline information only, The formulae of Gain Chart is described as below [18]:

$$\frac{\text{The number of matches in the quantile}}{\text{All matching number}} \times 100\%$$

In the assessment of different model, the core idea of Gains Chart is to compare the predictive model with the baseline (the random probability model) to calculate the corresponding gain of predictive model in each percentile data. In general, the steeper the curve, the higher gain predictive model will get, compared with stochastic model, which indicates the more effective predictive model will be in practice. The result of Gain Chart to assess the sound effects of four predictive models is shown as Fig. (**2**):

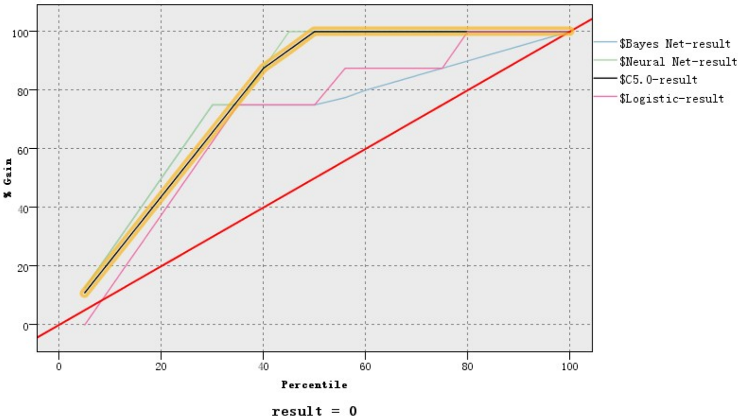

Fig. (2). Four classification models' comparison of gain chart.

Integrated analysis of Fig. (**1**) and Fig. (**2**): Fig. (**2**) shows that Bayesian network model and decision tree C5.0 model are more accurate and more stable, compared with the neural network and logistic regression models; Fig. (**2**) shows the curve of decision tree C5.0 model is relatively steeper: 40% percentile of the data sets corresponds to a gain of 90% of the data sets, indicating the significant practical utility of this model. The decision tree C5.0 model is most stable, while the other classification model has emerged as varying degrees of unstable state, the actual effect is poor. Therefore, taking the assessment of accuracy and utility into account, the conclusion is drawn that the decision tree C5.0 is the best classification model to predict whether the enterprise will deploy cloud computing.

CONCLUSION

In this paper, the influencing factors of affecting enterprise usage on cloud computing is given and C5.0 algorithm is used in order to confirm the classification model of affecting enterprise' usage of cloud computing. Compared with the other three typical classification model, Decision tree C5.0 model is verified as the most suitable and stable model to predict whether an enterprise will use cloud computing.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This research was supported by Project of National Social Sciences Foundation, Grant No. 13BTJ005, the Fundamental Research Funds for the Central Universities, Grant No. 2013XZD01, supported by the Guangdong Province Science and Technology Fund, Grant No. 2012B091100309, and 2012B040500010.

REFREENCES

[1] P. Mell, and T. Grance, "The NIST definition of cloud computing", *National Institute of Standards and Technology*, pp. 1-3,September 2011.

[2] W. Heyong, H. Wu, and W. Feng-Kwei, "Enterprise cloud service architectures", *Information Technology and Management*, vol. 13, pp. 445-454, April 2012.

[3] W. Heyong, "Information services paradigm for Small and Medium Enterprises Based on Cloud Computing", *Journal of Computers*, pp. 1240-1246, May 2013.

[4] D. Catteddu, "Cloud computing: Benefits, risks and recommendations for information security", *Springer Berlin Heidelberg*, p. 17, December 2010.

[5] S. C. Misra, and A. Mondal, "Identification of a company's suitability for the adoption of cloud computing and modelling its corresponding return on investment", *Mathematical and Computer Modelling*, vol. 53, pp. 504-521, March 2011.

[6] W.Y. Chang, H. Abu-Amara, and J. F. Sanford, "Challenges of Enterprise Cloud Services", *Transforming Enterprise Cloud Services*, pp. 133-187, 2010.

[7] Z. Qiang, "Development of chinese small and medium-sized enterprises", *Journal of Small Business and Enterprise Development*, pp. 140-147, February 2006.

[8] F. Etro, "The economics of cloud computing", *The IUP Journal of Managerial Economics*, vol. 9, pp. 2-7, February 2011.

[9] S. Yu, and Y. H. Tao, "Understanding business-level innovation technology adoption", *Technovation*, pp. 92-109, February 2009.

[10] Low, Y. Chen, and M. Wu, "Understanding the determinants of cloud computing adoption", *Industrial management & data systems*, vol. 111, pp. 1006-1023, July 2011.

[11] A. Garrison, S. Kim, and L. Robin, "Success factors for deploying cloud computing", *Communications of the ACM*, vol.55, pp. 62-68, September 2012.

[12] M. Armbrust, A. Fox, R.Griffith, A. D. Joseph, R .Katz, A. Konwinski, and M. Zaharia, "A view of cloud computing," *Communications of the ACM*, vol. 53, pp. 50-58, April 2010.

[13] A. D. Guo, Z. Min, Z. Yan, and X. Zhen, "Study on cloud computing security", *Journal of Software*, pp. 71-83, January 2011.

[14] J.S. Liang, "The establishment of evaluation system of Saving institution", *Statistics and Decision*, pp. 75-78, July 2012.

[15] Q. Xin, and M. Song, "The establishment and application of the evaluation index system of harmonious society of metropolitan", *Statistical Research*, pp. 17-21, July 2007.

[16] M. J. Aitkenhead, "A co-evolving decision tree classification method", *Expert Systems with Applications*, vol. 34, pp. 18-25, January 2008.

[17] S. Thomassey, and A. Fiordaliso, "A hybrid sales forecasting system based on clustering and decision trees", *Decision Support Systems*, vol. 42, pp. 408-421, January 2006.

[18] T. Burez, and J. Van den Poel, "Handling class imbalance in customer churn prediction", *Expert Systems with Applications*, , vol. 36, pp. 4626-4636, March 2009.

Received: September 22, 2014 Revised: November 05, 2014 Accepted: November 06, 2014