K-NN Classifier for Data Confidentiality in Cloud Computing

[1]Munwar Ali Zardari, [2]Low Tang Jung, [3]Nordin Zakaria
CIS Department, Universiti Teknologi PETRONAS
Seri Iskandar, Malaysia
[1]munwar_g02103@utp.edu.my, [2]lowtanjung@petronas.com.my, [3]nordinzakaria@petronas.com.my

***Abstract*—Securing the data in cloud is still a challenging issue. In cloud, many techniques are used to secure the data. Data encryption is a data security technique which is widely used for data security. Deciding data security approach for the data without understanding the security needs of the data is not a technically valid approach. Before applying any security on data in cloud, it is best to know the security needs of the data. What data need security and what data do not need security. In this paper, we propose a data classification approach based on data confidentiality. K-NN data classification technique is modulated in the cloud virtual environment. The aim to use K-NN is to classify the data based on their security needs. The data is classified into two classes, sensitive and non-sensitive (public) data. After data classification we got that what data need security and what data does not need security. The RSA algorithm is used to encrypt the sensitive data to keep it secure. We used CloudSim simulator to find the results of proposed approach in cloud. The proposed approach will easily decide the security needs of the data. After the data classification, it is easy to select an appropriate security for data according to the need of data. The results show that this approach is appropriate as compared to store data in cloud without understanding the security requirements of data.**

Keywords—Cloud Computing; Distributed Computing; K-NN; RSA; Data Classification; Data confidentiality/sensitivity; non-sensitive

I. INTRODUCTION

Cloud Computing is an internet based distributed virtual environment. All computational operations are performed on cloud through the Internet [1]. The cost of the resource management is more than the actual cost of the resources. So, it is often better to get the required resources by renting despite purchasing one's own resources. Basically, the cloud computing provides all IT resources for rent. The simple definition of cloud computing is: "A distributed virtual environment provides virtualisation based IT-as-Services by rent". Beside all of the services like Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS) and Infrastructure-as-a-Service (IaaS), cloud also provides storage as a service, in which distributed database servers are available for rent to consumers. These services are available for all users without any business or data bias.

Consumers nowadays are using cloud services to avoid IT infrastructure purchasing and maintenance cost. A large amount of data can be stored on cloud. Cloud computing poses a number of challenging threats in the distributed storage model [2]. The data security is always the main challenging threat for quality of services and also stops the users to adopt cloud services [3]. In cloud storage, all kinds of data are stored on servers, and they are stored through two storage methods. The first method is to encrypt received data and store them on cloud servers. The second method is to store data on servers without encryption. These data storage methods can face data confidentiality issue. It is known that data are often not of the same type and have different properties and characteristics. In a cloud environment, a consumer's data are stored on remote servers that are not physically known by the consumer and there is high always chance of confidentiality leakage. This paper focuses on the confidentiality threat in the cloud environment. When a dataset is being transferred to cloud, it passes through a security mechanism, such as data encryption (without understanding the features of data) or directly being stored on servers without encryption. All data have different kinds of sensitivity levels. So, it would be non-technical to just send data into a cloud without understanding its security requirements. To address the security requirements of data, we have proposed a data classification model in the cloud environment to classify data according to its sensitivity level.

978-1-4799-0059-6/13/$31.00 ©2014 IEEE.

II. RELATED WORK

To show the importance of data security in cloud computing, the European Network and Information Security Agency (ENISA) published a report titled "Cloud Computing: Benefits, risks and recommendations for information security" in Nov-2009. In the report, ENISA found different cloud risks and their effects on the cloud consumers [4]. A crypto co-processor was suggested in [5] to solve the data security threats in cloud. The crypto co-processor is a tool which provides security-as-a-service on demand controlled by a third party. Crypto co-processor allows the users to select the encryption technique to encrypt the data and divide data into different fixed chunks. This is to make hacker not knowing the starting and ending points of the data. But the limitation with this study is that the end user may not be technically savvy enough to select powerful technique for data encryption.

The single cloud is a central storage place for all consumers' data and the single central storage is easier to hack than compared to multiple storages. IBM proposed a new concept of inner-cloud in 2010. The inner-cloud is the clouds of a cloud. The inner-cloud storage model is more reliable and

trustworthy as compared to a single cloud storage model [6]. In the inner-cloud model, the hash function and digital signature are hybridised to provide data authentication and integrity in a cloud. Whereas the data security key is divided and shared in multiple clouds; but this process of sharing of keys leads to a key issue when one cloud is not available. The integrated Data Protection as a Service (DPaaS) model is also used for data security [7]. DPaaS integrates information flow checking, encryption, and application installation in cloud computing to avoid the implementation of the FDE and FHME techniques which are not affordable by small and medium enterprises and cloud service providers. The public cloud has still security challenges and data outsourcing is a big challenge. In data outsourcing, the user can not be sure about the location, data transaction accuracy and security of the stored data.

Most of these techniques i.e. discussed above work on data encryption for data security. To encrypt complete data, it is very expensive in the context of time and memory. It would be better to separate the sensitive data from the public data first and encrypt only the sensitive data.

Classification of objects is an important area of research and of practical applications in a variety of fields, including pattern recognition and artificial intelligence, statistics, cognitive psychology, vision analysis and medicine [8][9]. There are numerous machine learning techniques that have been developed and investigated for classification. However, in many pattern recognition problems, the classification of an input pattern is based on the data where the respective sample size of each class is small. Moreover, the sample may possibly not be representative of the actual probability distribution, even if it is known [8]. In such cases, there are many techniques that work on similarity and distance in feature space, for instance, clustering and discriminate analysis [10]. In many areas, the K-Nearest Neighbour (K-NN) algorithm is used for classification of instances. K-NN is the simplest clustering technique with low complexity. This decision rule provides a simple nonparametric procedure for the assignment of a class label to the input pattern based on the class labels represented by the k-nearest neighbour of the vector. K-NN classification is more suitable for those problem domains characterised by data that is only partially exposed to the system prior to employment [11] [12].

III. PROPOSED MODEL

Machine learning techniques are mostly used in pattern recognition and data segmentation. In our proposed model we used the K-NN machine learning technique in the cloud computing environment to solve the data confidentiality problem.

To separate sensitive and non-sensitive data, the K-NN classifier is used in a designed simulation environment. The value of k is maintained to one (1) for accuracy. After finding the sensitive and non-sensitive data, the sensitive data is further transferred to the RSA encryption algorithm for data encryption to protect sensitive data from unauthorised users. Therefore, the public data is directly allocated a Virtual Machine (VM) without encryption. The VM will process the data and communicate with storage servers for the data storage on the cloud servers. Most of the clouds implement data encryption techniques to protect data. But it is better to decide the security of the data based on the sensitive level of the data instead imposing encryption on complete data or just sending complete data into the server without any security. The classification technique in cloud will easily decide the security requirements of the data. In this way we can save our data from over-security and under-security situations and also save time and memory resources. In this paper, we classified data into two different classes class1 (non-sensitive data) and class2 (sensitive data). Figure 1 shows the detailed steps to solve the data confidentiality issue in cloud computing.

A. Data Classification

Data classification is a machine learning technique used to predict the class of the unclassified data. Data mining uses different tools to know the unknown, valid patterns and relationships in the dataset. These tools are mathematical algorithms, statistical models and Machine Learning (ML) algorithms [13]. Consequently, data mining consists on management, collection, prediction and analysis of the data. ML algorithms are described in two classes: supervised and unsupervised.

Supervised Learning: In supervised learning, classes are already defined. For supervised learning, first, a test dataset is defined which belongs to different classes. These classes are properly labelled with a specific name. Most of the data mining algorithms are supervised learning with a specific target variable. The supervised algorithm is given many values to compare similarity or find the distance between the test dataset and the input value, so it learns which input value belongs to which class.

Unsupervised Learning: In unsupervised learning classes are not already defined but classification of the data is performed automatically. The unsupervised algorithm looks for similarity between two items in order to find whether they can be characterised as forming a group. These groups are called clusters. In simple words, in unsupervised learning, “no target variable is identified”.

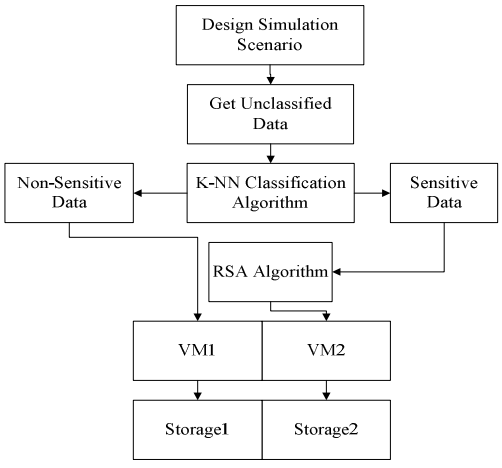

Fig. 1. Proposed Model

The classification of data in the context of confidentiality is the classification of data based on its sensitivity level and the impact to the organization that data be disclosed only authorized users. The data classification helps determine what baseline security requirements/controls are appropriate for safeguarding that data.

In this paper, the data is classified into two classes, confidential and public (non-confidential) data. The number of classes depends on the requirements of data owner. Michigan Technological University published a report titled "data classification and handling policy" in which they categorized data based on security into confidential data, internal/privacy data and public data [21]. Another report published by University of Texas Health Science Centre at San Antonio (UTHSCSA) titled "protection by data classification security standard", in which data is classified public, internal, confidential and confidential/high risk [22]. The California State University published a document titled "information security data classification standards". This document describes the three classes of data regarding the data security placed on the particular types of information assets. These three classes are confidential, internal use and general [23]. In this paper we treat all data as confidential except public data because all other data need to be secure at different stages of data process with different methods. These protection methods can be secure area storage, lockable enclosure, and reasonable precautions to prevent access by non-staff and employees, password protection, encryption and so on.

Sensitive (Confidential) data: Data that is classified as confidential contains data that can contain very important data of individuals or organizations. The unauthenticated access to confidential data will catastrophic the owner. Such information might include:

- Personal data: includes personal identifiable information such as social security number, national identification number, passport number, credit card number, driver's license number, medical records, and health insurance policy ID number but not limited.
- Financial Records: includes financial account number, transaction information.
- Business Material: Such a document or data that is unique or specific intellectual property.
- Legal Data: includes potential attorney-privileged material.
- Medical/Health Data: Includes sample code number, address, date of admission, medical record number, date of discharge, health plan beneficiary number.
- Government Data: Includes government intellectual documents, government agency documents, government future plan information.

Non-Sensitive (Non-Confidential/public) data: Public data also considered as unrestricted data. Information which is classified as non-confidential includes data and files that are not critical to business needs or operations. That class also includes data that is deliberately been released to the public for their use, such as press announcements, marketing material or introductory information of any organization.

B. K-NN Classifier

The K-Nearest Neighbour (K-NN) is a supervised machine learning technique; it is widely used for classification, pattern recognition, prediction and estimation. K-NN totally depends on instance-based learning, where a set of training data is stored for the classification of new unclassified datasets [14]. It is the simplest iterative technique to classify unclassified datasets into user specified classes, k. This algorithm discovered by several researchers across different disciplines, most notably Lloyd (1957, 1982), Forgey (1965), Friedman and Rubin (1967), and McQueen (1967) [15].

The KNN Classifier is a very simple classifier that works well on basic recognition problems. Using a relatively larger k may include some not so similar pixels; on the other hand, using a smaller k may exclude some potential candidate pixels. In both cases the classification accuracy will be decrease [16].The Computational complexity of K-NN is high due to the usage of all of the training samples for classification [17]. The algorithm must compute the distance and sort all of the training data distance at each prediction. Another issue is the approach to combining the class labels. The K-N uses the simplest method to take a majority vote, but this can be a problem if the nearest neighbours vary widely in their distance and the closer neighbours more reliably indicate the class of the object [18].

The K-NN algorithm has a set of n labelled samples; n is the number of data items in the set. This can be represented as:

$$D = \{d_1, d_2, d_3, ..., d_n\},$$

where D is the set of total samples and $d_1, d_2, d_3, ..., d_n$ are different samples. The D must be assigned n labels. The set of n labelled samples can be represented as:

$$D = \{d_1, d_2, d_3, ..., d_3 \mid c\}$$

Where c_1 is a class for the targeted values. In algorithm we specified only one class for confidential attributes.

How the K-NN Algorithm Works:

Step 1: Determine the set of n labelled samples: D
Step 2: Determine value of K
Step 3: Calculate the distance between the new input and all of the training data
Step 4: Sort the distance and determine the K-nearest neighbours based on the K-th minimum distance Step 5: Find the classes of those neighboursStep 6: Determine the class of the new input based on a majority vote.

C. Cloud Simulation Environment

The CloudSim simulator was used for simulation purposes. Figure 3 shows the proposed simulation environment for cloud service providers to solve data sensitivity/confidentiality issue in cloud computing. At the bottom, we used a CloudSim engine to run the simulation. The Virtual Machine Manager (VMM) was used to manage and allocate VMs to cloudlets (cloud tasks). The number of cloud items used for the simulation is given in table I.

TABLE I. CLOUD ITEMS AND QUANTITY

Items	Quantity
Data Centre	2
Host	1
Broker	1
VM	2
VMM	1
Cloudlet	*2*

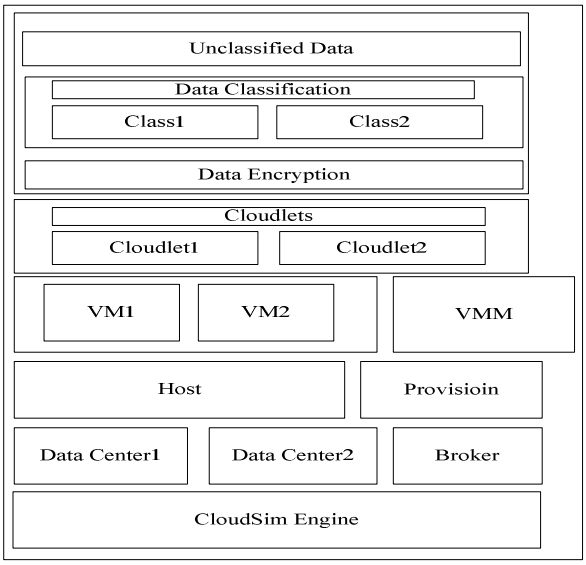

Fig. 2. Simulation Environment

D. Cloud Service Properties and Description

Before simulation it is important to set the properties of all three service models Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS) and Infrastructure-as-a-Service (IaaS). Table II shows the properties of the SaaS modeller which was deployed on a VM in CloudSim. In the SaaS modeller every cloudlet has a specific identification for the VM. Here "ID" represents to a specific cloudlet, and length is the size of the cloudlet. The size of the input and output file is given in Bytes.

Length: The length or size (in MI) of this cloudlet is to be executed in a datacentre.

Input File Size: The file size (in Bytes) of the current cloudlet BEFORE being submitted to a datacentre.

Output File Size: This is the file size (in Bytes) of the current cloudlet AFTER execution.

TABLE II. SAAS PROPERTIES

Cloudlet ID	Length	Input file size (bytes)	Output file size (bytes)
0	4000	158	158
1	3000	139	139
0	4000	158	158

Table III shows the properties of PaaS for the application deployment which contains VM properties. It shows the processing power of the physical computing node which is assigned at the virtual machine level, VM image size (in MB), amount of bandwidth, and the number of cores in which the MIPS' power is shared at the VM level to run the cloudlet. The VMs are managed by VMM.

Machine Instructions Per Second (MIPS): This is the processing power assigned to the VM to execute the instructions according to the specified MIPS.

Image Size: The Image Size (in MB) is the VM image size that is represented to a virtual hard disk file that is used as a template for creating a VM.

Pes Number: The Pes number is the number of processors used at the VM level.

TABLE III. PAAS PROPERTIES FOR VIRTUALISATION MANAGEMENT

VM ID	MIPS	Image size (MB)	Band width	Pes No.	VMM
0	100	10000	1000	1	Xen
1	100	10000	1000	1	Xen

It is also important to use better and stronger infrastructure resources in cloud for better computation and response time. The available resources at this level put a limit on the SaaS modeller requirement, i.e., resources allocated at the VM level can't exceed this limit. Table IV shows the IaaS properties and their values, where "DC ID" is the datacentre identity which is assigned to the VM.

TABLE IV. IAAS PROPERTIES FOR CLOUD SIMULATION

DC ID	RAM (Mb)	Storage	Data Architecture	OS	Bandwidth
2	2048	10000000	X86	Linux	10000
3	2048	10000000	X86	Linux	10000

IV. RESULTS AND DISCUSSION

In this section, we discuss the results taken after the implementation K-NN and RSA algorithms to improve and mange the confidentiality level of data in a cloud environment. The data selected for this study is the employees' records of an organisation. This data was taken from [20], which contains different types of datasets mostly used by the research community. Table V shows the details of the file before and after classification. The total size of the file was 512KB and total records in file were 5094. The K- NN classifier was used to classify the data based on the confidentiality. The employees' dataset is classified into two classes sensitive and non-sensitive. The selection of confidential and non-confidential data is based on the rules for employees' record which are listed in table VI. After classification, the non-sensitive data was labeled as "Class1" and the sensitive data was labeled as "Class2". The time taken by the K-NN classifier to classify 5094 records was 1075ms as shown in table V. After classification, both classes are assigned to two cloudlets as new tasks for VMs as shown below.

$$(cloudlet\_ID = 0) \leftarrow class\ 1\_data$$

$$(cloudlet\_ID = 1) \leftarrow class\ 2\_data$$

The VMs are assigned cloudlets for further process. The virtual machine processed the data to the selected data centres. The both cloudlets are simply processed when users want to store sensitive data into a secure place without encryption. This situation can be risky for sensitive data. The total simulation time taken by both cloudlets is shown in table VII. Here, cloudlet ID=0 represents the Class1 data and cloudlet ID=1 represents the Class2 data.

If the status of the simulation was "SUCCESS", it means that the simulation was performed successfully. For this simulation each class data was assigned to a different VM and datacentre.

TABLE V. CLASSIFICATION OF DATA

Before Classification		After Classification				
		Class1		Class2		K-NN Time (ms)
Total size of file (KB)	Total Records in file	Size (KB)	Records	Size (KB)	Records	
512	5094	352	3450	160	1644	1075

TABLE VI. DATA SECURITY REQUIREMENTS

S. No	Parameters	Class
1	Employee salary	Sensitive
2	Address	Sensitive
3	Phone number	Sensitive
4	Email address	Sensitive
5	Payment history	Sensitive
6	Mother's maiden name	Sensitive
7	Race	Sensitive
8	Ethnicity	Sensitive
9	Parents' and other family members information	Sensitive
10	Birthplace	Sensitive
11	Gender	Sensitive
12	Marital status	Sensitive
13	Physical description	Sensitive
14	Other parameters' of employee records	Non-sensitive

Encryption of Class2 (sensitive data):

In this case, after the classification the sensitive data (class2) was encrypted using the RSA algorithm to make secure. The simulation time and the size of overall data were increased as compared with the original size of data and total simulation time taken by both cloudlets as mentioned in table VII. Table VIII shows the time taken by K-NN to classify the data, the time taken by RSA to encrypt data, total simulation time and total size of data. After encryption the size of data and the simulation time is increased due to the data encryption. Class1 data took the same time during simulation as it has taken in table V but the simulation time of class2 data changed after the encryption. After encryption, the size of the data increased up to 50166.86KB, which was assumed the size of 167883 normal records (non-encrypted records). The total time taken by the RSA to encrypt 160KB was 2796237 ms. The total time of the simulation with classification and encryption time is 7953112 ms. The total simulation time was calculated using the equation no.1.

$$TST = \left(CT + ET + TC_i\right) \quad (1)$$

where, TST is the total simulation time, CT is the classification time, ET is the encryption time and TC_i is the time that was taken by both cloudlets after classification and encryption. Here $i = 1,2$.

V. CONCLUSION

In this paper, we have proposed a confidential based data classification model for cloud computing. The focus of this study was to classify the data based on data security needs that what kind of data need to be secure and what kind of data keep public. The basic contribution of this model is data confidentiality with machine learning technique (data classification). For data classification, the K-NN classifier is used to classify data based on the security requirements of the data. Based on data security requirements, the data is classified in to two classes, sensitive and non-sensitive data. The sensitive data required more secure and encrypted using the RSA algorithm whereas the non-sensitive data is directly stored on the cloud servers.

The proposed model has been implemented in a designed simulation environment using a CloudSim simulator Furthermore; to our best knowledge this proposed model is the first model in cloud computing with a data classification technique to improve the security of data.

REFERENCES

[1] Smith, P. S. Rawat, G. P. Saroha, and V. Barthwal, "Quality of Service Evaluation of SaaS Modeler (Cloudlet) Running on Virtual Cloud Computing Environment using CloudSim", International Journal of Computer Applications (0975 – 8887), Vol. 53– No.13, September 2012.

[2] D. Purushothaman, and S. Abburu, "An Approach for Data Storage Security in Cloud Computing", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, March 2012.

[3] J. W. Rittinghouse, and J. F. Ransome, "Cloud Computing Implementation, Management, Security", CRC Press 2009 by Taylor and Francis Group, Journal of high technology law, 2010.

[4] D. Catteddu, and G. Hogben, "Cloud Computing: Benefits, risks and recommendations for information security", ENISA, 2009.

[5] C. P. Ram, and G. Sreenivaasan, "Security as a Service (SasS): Securing user data by coprocessor and distributing the data," Trendz in Information Sciences & Computing (TISC2010), pp. 152-155, Dec. 2010.

[6] C. Cachin, and R. Haas, "Dependable Storage in the Intercloud," IBM Research Report RZ 3783, 2010.

[7] D. Song, E. Shi, I. Fischer, and U. Shankar, "Cloud Data Protection for the Masses," IEEE Computer Society, pp. 39-45, Jan. 2012.

[8] J. M. Keller, M. R. Gray, and J. A. Givens, JR, "A Fussy-K-Nearest Neighbor Algorithm", IEEE Trans. System, Man, and Cybernetics, vol.SMC-15, No. 4, pp 580-585, july/ aug 1985.

TABLE VII. CLOUD SIMULATION

Cloudlet ID	VM ID	Datacentre ID	Status	Start Time	Finish Time	Total Time	Total Time (taken by both cloudlets)
0	0	1	SUCCESS	0 ms	3400 ms	3400 ms	5040 ms
1	1	2	SUCCESS	0 ms	1640 ms	1640 ms	

TABLE VIII. TOTAL TIME TAKEN BY DATA

Classification Time	Time taken by RSA	Simulation time	Total Time (both cloudlels)	Total Size of data (KB)
1075ms	2796237 ms	5155800 ms	7953112 ms	50166.86

[9] E. Hunt, Artificial Intelligence. New York: Acadimic, 1975.

[10] R.O. Duba, and P.E Hart, Pattern Classification and Scene Analysis. New York: Wily, 1973.

[11] A. Whitney, and S.J Dwyer, II, "performance and implementation of K-nearest neibbor decision rule with incorrectly identified training samples", in Proc. 4th Ann. Allerton Conf. On Circuits Band System Theory, 1966.

[12] B. V. Dasarathy, "Nosing aroung the neighbourhood: A new system structure and classification rule for recognition in partially exposed environments" IEEE Trans. Pattern Anal. Machine Intell. Vol.PAMI-2, pp 67-71, 1980.

[13] T. N. Phyu, "Survey of Classification Techniques in Data Mining", Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I IMECS 2009, March 18 – 2009.

[14] D. T. Larose, "Discovering Knowledge in Data: An Introduction to Data Mining", John Wiley & Sons, Inc., ISBN 0-471-66657-2, pp 90-106, 2005.

[15] Lloyd SP (1957) Least squares quantization in PCM. Unpublished Bell Lab. Tech. Note, portions presented at the Institute of Mathematical Statistics Meeting Atlantic City, NJ, September 1957. Also, IEEE Trans Inform Theory (Special Issue on Quantization), vol IT-28, pp 129–137, March 1982.

[16] M. Khan, Q. Ding, and W. Perrizo, "K-Nearest Neighbor Classification on Spatial Data Streams Using P-Trees1, 2", PAKDD 2002, LNAI 2336, pp. 517-528, 2002.

[17] N. Suguna, and K. Thanushkodi, "An improved K-nearest neighbor classification using genetic algorithm", IJCSI, Vol. 7, Issue-4, No 2, July 2010.

[18] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z. H. Zhou, M. Steinbach, D. J. Hand, D. Steinberg, "Top 10 algorithms in data mining", Knowl Inf Syst (2008) 14:1–37, 2008.

[19] R. N. Calheiros, R. Ranjan, A. Beloglazov, A. F. De Rose, and R.r Buyya, "CloudSim: A Toolkit for the Modeling and Simulation of Cloud Resource Management and Application Provisioning Techniques", Software: Practice and Experience, 41(1): 23-50, Wiley, January 2011, which has been published in final form at http://dx.doi.org/10.1002/spe.995.

[20] http://www.sgi.com/tech/mlc/db/, referred: Aug-2013.

[21] Michigan Tech Information Technology Services & Security, "Data Classification and handling Policy", Rev: 2-7-11.

[22] UTHSCSA Data Classification report, "Protection by Data Classification Security Standard", Aug 2006.

[23] The California State University, "Information Security Data Classification", 9-28-2011.