Semantic Cluster based Classification for Data Leakage Detection for the Cloud Security

C. Suresh Kumar
Associate Professor, Dept. of Computer Science
M.K.U. College, Madurai, India.

K. Iyakutty, Ph.D.
Professor, Dept of Nanotechnology and Physics
SRM University, Chennai, India

ABSTRACT
A novel approach for the data leak detection in the cloud environment is discussed in this paper. The paper uses the semantic based clustering for the anomaly detection to find the data leak. The Clustering is further used for the classification to add up for the semi supervised classification. After classification the threat patterns are stored in the database for further preventive actions in the data transmission. The necessary theory is discussed and the proposed approach is discussed with the results obtained.

Keywords
Data leak prevention, semantic clustering, and semi supervised classification

1. INTRODUCTION
Data leakage is a problem which occurs intentionally or in an unintentional way. The intentional data leakage is the problem where the hackers try to disclose the data for fun or for any profit motivation. The unintentional damage is based on an accidental happening. Either intentional or unintentional act it involves in the loss of the data which leads to the loss and the reputation of the enterprise. This is common to every enterprise possessing the data and facing the trouble of leaking. It is also a global problem to think through since the data sharing is the mandate of the day for the communication. In the resource sharing enterprise which moves to the cloud computing for the distributed nature also severely suffer from the problem of data leakage.

A survey says that by 2014 public IT services will exceed the traditional IT services by nearly five times [1]. The forecasting based on the cloud computing predicts that worldwide revenue from the public IT cloud services will reach 55.5 billion dollars in 2014 [2]. The cloud is prone to the data leakage because of its operational characteristics and its architecture, the chance is more because of the huge transactions which involves risks and challenges [3].

The preventive mechanism is discussed in this paper which deploys the semantic clustering with the semi supervised classification for the updating of the threat history. The threat repository could be further used as the detection database to avoid the data breaches. This paper addresses the issue of proactive management through the learning from the previous data. The idea fills the gap of semantic based threat identification, which is lacking in the existing methods.

The paper is organized as section 2 discuss about the Data leakage prevention methods in cloud computing and the section 3 deals with anomaly detection techniques. Section 4 opens the methods for classification after clustering. Section 5 discusses the proposed method in detail with the pseudo code. Results are discussed in section 6.

2. DATA LEAKAGE PREVENTION METHODS IN CLOUD COMPUTING
An enterprise's important and the most valuable asset is its data. It is always very important to device the methods to prevent the data leakage. In this section let us discuss on the methods involves with the data leakage prevention. This section concentrates on this problem in cloud environment.

Data is the valuable asset which should be secured when it is at rest, motion and it is in use. The data leak prevention method is the one where the data at any stage should be assured in safe. The main objective of any data leakage prevention method is defined as follows [4]

- To trace and record the perceptive data or information throughout the venture.
- To watch and manage the travel of the perceptive data across the business enterprise.
- To watch and manage the travel of the perceptive data in the area of end user.

The various data leak prevention methods are discussed here. There are three types of data leakage prevention methods are in practice [5]. They are briefed as

a) Network based methods – this type of data leak prevention mechanisms work through the data which are transferred through the network. It normally detects the data which is in motion.

b) Endpoint based methods – This type of mechanism work at the user's point .They normally used to detect the data leakage from the accessing point.

c) Embedded based methods – These methods deploy the customized tools which are embedded within an application.

3. METHODS FOR ANOMALIES DETECTION
Computer security measures normally spotlight on the preventing mechanisms like authentication process, filtering the data or malicious information, and encryption of the messages, but another significant surface is detecting attacks once the preventive measures are violated [6]. In order to proceed in this situation, analysis of huge volume of data to done. Two general methods are used to tackle the above said situation say Signature or misuse detection and anomaly detection. The first method looks for prototype indicating familiar attacks where as the second method deviation from the normal behavior pattern. Signature detection normally suits for the attack which are previously known but not guaranteed for the new detection or unknown attacks. Anomaly detection methods are very well suited for the

unknown or the new attacks. It could provide an alarm about the new detections.

Anomaly detection is inspired by the biological immunology. Forrest et al. [7] observe that the immune system of our body works by the identification of the foreign bodies which is previously unknown and molest them. Anomaly detection could be classified under the following categories such as [8]

- Point Anomaly Detection
- Contextual Anomaly Detection
- Collective Anomaly Detection
- Online Anomaly Detection
- Distributed Anomaly Detection

Clustering based anomaly detection is placed under the point anomaly detection where the idea for the detection is that the data with normal behaviour belongs to the cluster which is large and dense, but the anomalous data doesn't belong to the cluster represented previously. This could be further classified into semi supervised method and unsupervised method. The advantage of using the clustering based methods are that it is need not to be supervised which is application in many of the applications in the real world domains. At the same time the drawbacks of this method is it doesn't expected to be working well if the normal patterns do not create any clusters, When the data belongs to be high dimensional distance couldn't be an effective metric.

Contextual anomaly detection works with the base idea that to Identify a context around a data instance and then to identify whether the data instance is anomalous by a group of behavioral characteristics. The advantage in this method is this detection technique is useful when the anomalies are tough to be detected in the global perspective.

4. METHODS FOR CLASSIFICATION AFTER CLUSTERING

Classification is a supervised learning where the part of the data set is used for the training and the rest of the dataset is used for the verification and classification. Semi supervised classification is the learning where the classifier is built using the labeled and unlabeled training dataset [9]. In this technique the unlabeled samples enhances the accuracy of the classifier. The idea used in [10-12] is they first use the clustering concept to cluster the dataset and the cluster results are used for the classification. The semi-supervised approach used in [13] applies the hierarchical clustering and then classify the dataset.

The approach used in [14] is with the basic assumption for the documents in a collection are that each class is composed of a number of mixture components. By identifying the components in the document collection, the classes of documents can thereby be identified from each other. A semi-supervised clustering method is proposed to identify the components (clusters), and further, unlabeled data is used to produce more accurate clusters in document collection to correspond the components of document classes.

Clustering has been used in the literature of text classification either as an approach for dimensionality reduction or as a technique to enhance the training set. In the second case, the enhancement is achieved either by extending the feature vectors of the training examples with new features derived from clustering or by expanding the training set with new examples derived from a large set of unlabeled data[15].

5. PROPOSED METHOD

In this paper we discuss a framework for the proactive security mechanism in cloud computing to solve the data leakage problem. This framework works with the semantic analysis and then the clustering of the dataset is done based on the similarities and the anomaly is detected for the leakage. The next step is to classify the data leakage threat. After classification the history of the threats is being update. This history of the threats serves as the basic repository for the avoidance of the data leakage in the future.

Reactive mechanisms are discussed in the literature but the novel approach for the proactive mechanism is devised in this paper. This is done is two steps, first is the anomaly detection is done for the data leak detection. Contextual clustering process is employed for the anomaly detection. Since the clustering process solely disturbed by the dissimilarities, this is tackled by the contextual part. The classification step is used for the identification or grouping of the patterns which could cause the data loss.

Pseudo code

Step 1: Semantic clustering for the anomalous detection

INPUT: Semantically analyzed dataset, database for the threat patterns

BEGIN

Mapping of Semantic component with the dataset

Grouping and ranking based on the dataset

END

OUTPUT: Semantic clusters with top ranks and the irrelevant clusters with lower rank

Step 2: Semi supervised classification

INPUT: Semantic ranked clusters

BEGIN

Training labels construction based on the semantic scoring

Constructing the classifier with the semi supervised learning

END

OUTPUT: Anomalous behavior patterns

Step 3: Updating of the history of threats

INPUT: Anomalous behavior patterns

BEGIN

Database update with the anomalous behavior patterns

END

OUTPUT: Updated Database for threats

6. RESULTS AND DISCUSSION

The environment is simulated and the following screenshots shows the semantic clustering and the anomaly is shown in the different color

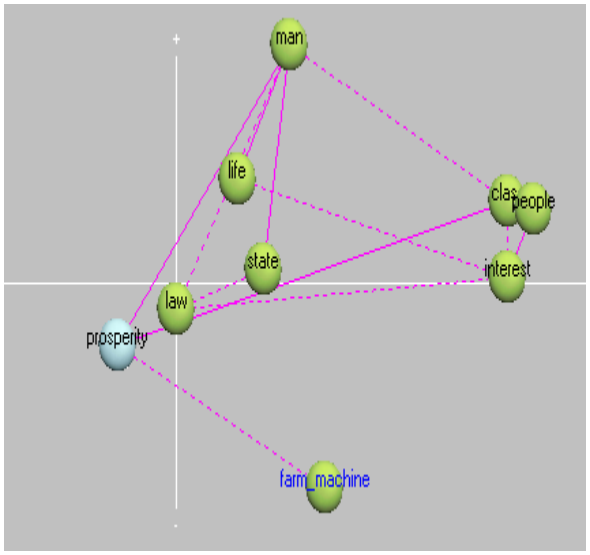

Fig 1 : Semantic clustering when cluster center = 9

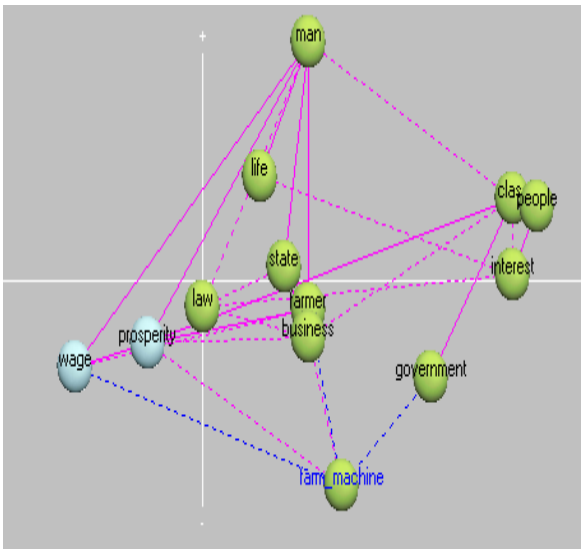

Fig 2 : Semantic clustering when cluster center = 13

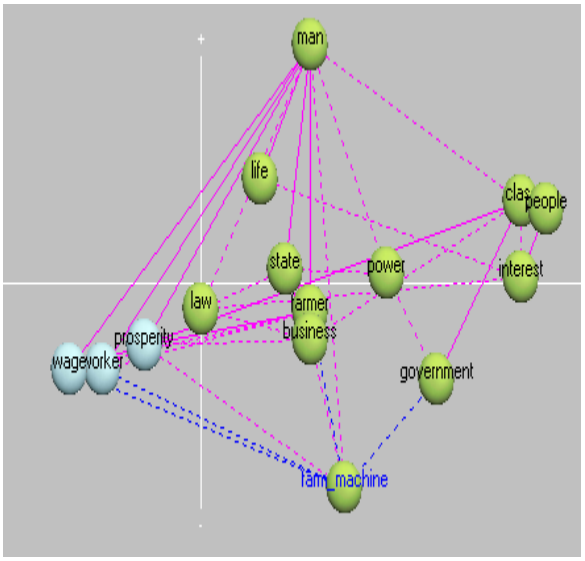

Fig 3: Semantic clustering when cluster center = 15

Table 1: Comparison based on the Micro F1 measure

Number of data in each class	Co-training	TSVM	CBC	Proposed Method
1	0.23	0.41	0.48	0.56
2	0.46	0.483	0.58	0.62
4	0.556	0.54	0.62	0.73
8	0.62	0.6	0.66	0.76
16	0.67	0.66	0.69	0.81

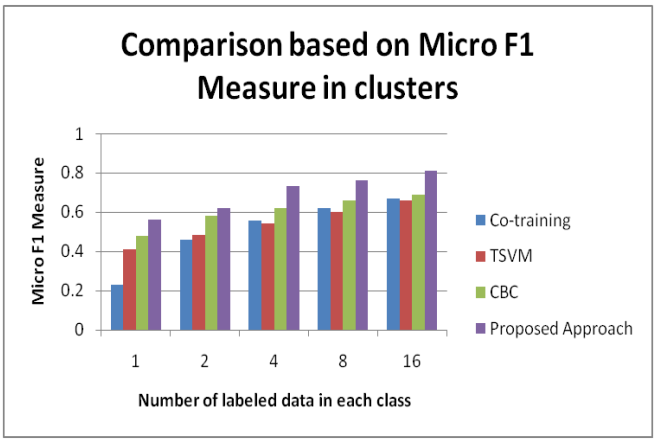

Fig 4: Comparison based on the Micro F1 measure

Table 2: Comparisons of Classification Error in %.

Error Criterion	REL-ML	ABS SIG	Proposed method
Omission error	3.04	4.64	2.98
Commission error	1.15	2.79	1.04
Class Averaged Error	2.095	3.715	2.01
Total error	1.95	3.58	1.82

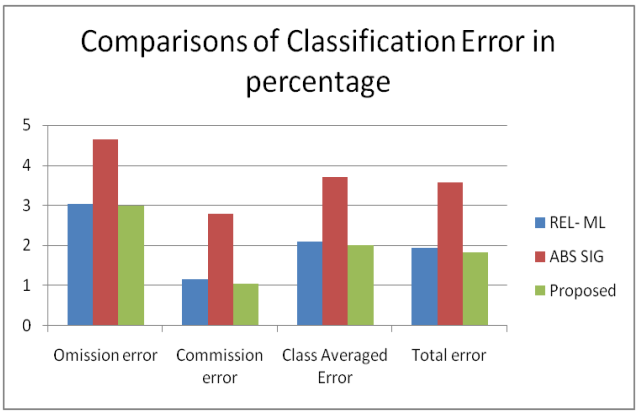

Fig 5: Comparisons of Classification Error in %.

7. CONCLUSION

The paper discusses the preventive mechanism that could be employed in the cloud environment for the detection of the data loss. This approach is quite useful since the semantic concept of the data is included for the clustering process which is often neglected in the Data loss prevention techniques. The idea of semi supervised classification is introduced and the results are shown. The future work could be concentrated on extending the work to the network intrusion detection.

8. REFERENCES

[1] http://www.istf.jucc.edu.hk/newsletter/IT_03/IT-3_Cloud_Computing.pdf

[2] http://www.buyya.com/papers/AnekaMagazineArticle1.pdf

[3] https://cloudsecurityalliance.org/topthreats/csathreats.v1.0.pdf

[4] http://www.isaca.org/Groups/Professional-English/security-trend/GroupDocuments/DLP-WP-14Sept2010-Research.pdf

[5] http://www.istf.jucc.edu.hk/newsletter/General_01/Gen-2_Data_leakage.pdf

[6] Philip K. Chan, Matthew V. Mahoney, Muhammad H. Arshad," Learning Rules and Clusters for Anomaly Detection in Network Traffic", Managing Cyber Threats Massive Computing Volume 5, 2005, pp 81-99

[7] S. Forrest, S. Hofmeyr, and A. Somayaji. Computer immunology. Comm. ACM, 4(10):88-96, 1997.

[8] Varun Chandola, Arindam Banerjee, and Vipin Kumar ,Outlier Detection – A Survey, Technical Report TR07-17, University of Minnesota

[9] Xuan-Hui Wang; Zheng Chen; Hongjun Lu; Wei-Ying Ma, "CBC: Clustering Based Text Classification Requiring Minimal Labeled Data", Proceedings of the Third IEEE International Conference on Data Mining, ICDM '03.

[10] Kyriakopoulou, A., Kalamboukis, T.: Using clustering to enhance text classification. In: 30th annual international ACM SIGIR conference on Research and development in information retrieval (2007)

[11] Raskutti, B., Ferr, H., Kowalczyk, A.: Using unlabeled data for text classification through addition of cluster parameters. In: 9th International Conference on Machine Learning (2002)

[12] Zeng, H. J., Wang, X.H., Chen, Z., Lu, H., Ma, W. Y.: CBC: Clustering based text classification requiring minimal labeled data. In: Third IEEE International Conference on Data Mining (2003)

[13] Hassan H. Malik,John R. Kender, "Classification by Pattern-Based Hierarchical Clustering", ECML/PKDD-08 Workshop 15 September 2008, Antwerp, Belgium

[14] Yoshida, T.; Xijin Tang, "Text Classification Using Semi-supervised Clustering", International Conference on Business Intelligence and Financial Engineering, 2009.

[15] R. Bekkerman, R. El-Yaniv, and Y. Winter,"Distributional word clusters vs. words for text categorization. Journal of Machine Learning Research",3:1183-1208, 2003.