# A Data Leakage Identification System Based on Truncated Skew Symetric Gaussian Mixture Model

**S. Praveen Kumar, Y Srinivas, M. Vamsi Krishna**

***Abstract*: Data transfer from source to destination has become more essential as the organizations working under a frame need to exchange the data for processing the information and solving the necessary tasks. This concept of data transfer has become a most tricky task with the advent of hackers, intruders and other guilt agents who try to steal the sensitive data for unethical means. The present article addresses the issue of identifying such data leakages and also provides a platform for data preventing from such issues.**

***Index Terms*: Truncated skew Gaussian Mixture, Hackers, Intruders, Guilt agent, Data Leakage.**

## I. INTRODUCTION

Data leakage primarily concerns with the loss of data or stealing of data from a source without permission from the authorities. The source of data leakages can be multifold viz., due to negligence, improper storing, malicious attacks, cyber threats, insider attacks this problem indirectly leads to heavy losses for the organizations, in particular while dealing with applications of bidding, financial transactions, sensitive data traversal, official data that needs to be shared between the organizations.

Many models and methodologies have been proposed and are being used in practice to safeguard the sensitive information and to uphold the data within the organization [1] [2] [3] [4].This data protecting or data prevention has become a more challenging task with the migration of the software companies towards the usage of virtual machines are sharing the data to remote places using the concepts of Cloud. The data transfer that is presently in practice between the organization isthrough a distributed framework is getting exaggerated with such violations of software practices and trying to attack the data and to steal the information [5] [6] [7].

This problem has become more demanding in particular scenarios of cloud computing usages. However, the data that is shared using the cloud environment cannot be identified against its storage location, the security concerns with respect to accessing, safeguard of the data and even the cloud user are unaware of the factual about the proper place where it is getting stored. To add, privacy preserving mechanism help to secure the data to some extent where the mechanisms adopted help to preserve the core private information to some extent. Research is also extended using the cryptographic system of message transfer but here the main shortcoming is that it cannot overcome the insiders attack[9] [10]. Methodologies are also proposed to utilize the concepts of watermarking techniques and nevertheless, these methods help to identify the transfer of data from the unethical means and cannot identify the guilt agent. [11] [12].Other technologies based on SVM, Neural network based approaches, classification, clusteing based approaches were also highlighted with the specific objective of protecting the sensitive information and identifying the guilt agent [13] [14] [15].Text based methodologies, semantic based methodologies, keyword based technologies were also highlighted for decline with specific application of protecting the sensitive information with respect to email data.

However, in the above presented methodologies, little light was thrown in the direction of protecting the sensitive information by developing a method that bundles both compression of data and transfer the data, such that with least efforts one can safeguard the data as well as reduce the dimensionality. Therefore, with this background, the present article is focused in this direction by formulating a methodology based on truncated skew Gaussian distribution which serves as both a key generator and a dimensionality reduction procedure. In order to experiment, the present framework, we have considered the KDD cup data set available in the web and the model is applied to prevent the confidentiality of the data.

The rest of the paper is structured as follows. The Section-2 of the article deals with the truncated Gaussian mixture model. The KDD cup data set considered is presented in Section-3. The Section -4 of the article deals with the methodology and the proposed architecture. The results derived are compared to that of the existing methods and are presented in Section-5 of the article. The Section -6 of the article summarizes with a scope for further extension.

## II. TRUNCATED GAUSSIAN MIXTURE MODEL

The probability density function of the truncated skew normal distribution is given by

$$F_{\mu,\sigma,\lambda}(x) = \frac{2}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right).\phi\left(\lambda\,\frac{x-\mu}{\sigma}\right)\text{------} \qquad (1)$$

**Revised Manuscript Received on 30 September 2018.**
***** Correspondence Author

**S.Praveen Kumar,** Research Scholar, Centurion University of Technology and Management, Rajaseetapuram (Orissa), India.

**Dr. Y Srinivas,** Department of Information Technology, GIT Gandhi Institute of Technology and Management, Visakhapatnam (A.P), India.

**Dr. M. Vamsi Krishna,** Department of Computer Science and Engineering, Centurion University, Paralakhemundi (Orissa), India.

Where, $\mu \epsilon R, \sigma > 0$ and $\lambda \ \epsilon R$ represents the location, scale and shape parameters respectively. Where $\phi and \phi$ denote the probability density function and the cumulative density function of the standard normal distribution.

The limits of the truncated normal distribution are $Z_1$=a and $Z_m$=b. Where $Z_1$ and $Z_m$ denotes the truncated limits. Truncating equation between these limits we have

$$F_{\mu,\sigma,\lambda}(x)\int_a^b = F_{\mu,\sigma,\lambda}(b) - F_{\mu,\sigma,\lambda}(a) \quad \text{-----} \ (2)$$

Where,

$$F_{\mu,\sigma,\lambda}(a) = \int_{-\infty}^{a} F_{\mu,\sigma,\lambda}(x)dx \quad \text{-----} \ (3)$$

And

$$F_{\mu,\sigma,\lambda}(b) = \int_{-\infty}^{b} F_{\mu,\sigma,\lambda}(x)dx \quad \text{------} \ (4)$$

## III. KDD CUP DATA SET

This is the data set used for The Third International Knowledge Discovery and Data Mining Tools Competition. The Moto of task was to shape a network intrusion detector, an analytical model skilled of distinctive classification characteristics between ``bad'' connections, called intrusions or attacks, and ``good'' normal connections. This database contains a typical set of data to be audited, which includes an extensive diversity of intrusion simulated indifferent network environment. It has been the most wildly used data set for the evaluation of anomaly detection methods. The URL to the data set is http://kdd.ics.uci.edu/databases.

## IV. METHODOLOGY

Unsupervised anomaly detection on multi- or high-dimensional data is of great importance in both fundamental machine learning research and industrial applications, for which density estimation lies at the core. Although previous approaches based on dimensionality reduction followed by density estimation have made fruitful progress, they mainly suffer from decoupled model learning with inconsistent optimization goals and incapability of preserving essential information in the low-dimensional space. In this chapter of the thesis, we present a truncated skew Gaussian Mixture Model (TSGMM) for unsupervised anomaly detection.

Truncated Skew Gaussian Mixture Model (TSGMM) consists of two major components: a compression network and an estimation network. The compression network performs dimensionality reduction for input samples, prepares their low-dimensional representations from both the reduced space and the reconstruction error features, and feeds the representations to the subsequent estimation network. The estimation network takes the feed, and predicts their likelihood/energy in the framework of Truncated Skew Gaussian Mixture Model (TSGMM).

Given the low-dimensional representations for input samples, and then density estimation under the framework of TSGMM. In the training phase with unknown mixture-component distribution φ, mixture means μ, and mixture covariance Σ, estimates the parameters of TSGMM are utilized for the estimation of the likelihood.

*System Architecture*

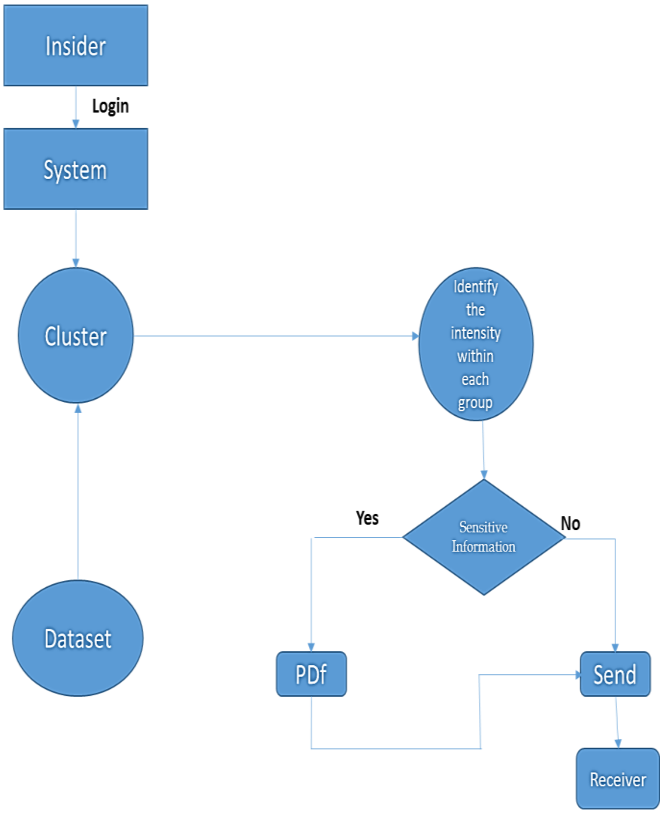

**Fig1: Training Phase**

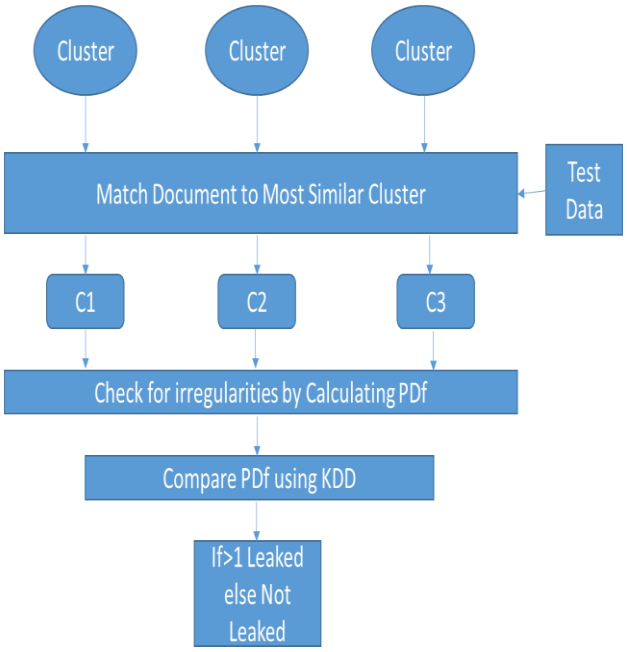
Cluster
Cluster
Cluster
Match Document to Most Similar Cluster
Test Data
C1
C2
C3
Check for irregularities by Calculating PDf
Compare PDf using KDD
If>1 Leaked else Not Leaked

**Fig2: Testing Phase**

The proposed architecture is twofold, which includes both training phase and testing phase. During the training phase, the database considered is clustered and against each segment, the corresponding probabilities are calculated, and these PDF are noted. During the testing phase, each of these clusters are considered and again their PDF values are compared with that of the training data. If there is a difference in these values, it is considered to be leaked else not leaked.

## V.RESULTS

**Table1: Output Derived by Comparing with that of the Existing Methods**

| Method | Identification of leakage | Mean absolute Error | Precision |
|---|---|---|---|
| C4.5 | 72.4 | 0.62 | 0.89 |
| Support Vector Machines | 77.2 | 0.73 | 0.87 |
| Neural Network | 79.8 | 0.7 | 0.78 |
| Gaussian Mixture Model | 83.2 | 0.033 | 0.97 |
| Skew Gaussian Mixture Model | 86.72 | 0.024 | 0.943 |
| Truncated Skew Gaussian Mixture Model | 89.5 | 0.021 | 0.982 |

In order to test the authenticity of the proposed model we have considered the metrics mean absolute error (MAE) and precision. The outputs derived are compared with that of the existing techniques such as C4.5, SVM, Neural Networks, Gaussian Mixture Model, Skew Gaussian Mixture Modeland Truncated Skew Gaussian Mixture Model.

## VI. CONCLUSION

In the proposed method we proposed a methodology to protect the confidentiality of the data. The method proposed helps to identify whether the data has been leaked or not and also helps to safeguard the data. The results derived are compared against efficiency using the metrics Mean absolute error and precision. The results derived showcase that the proposed model outperforms the existing models.

## REFERENCES

1. http://www.istf.jucc.edu.hk/newsletter/IT_03/IT3_Cloud_Computing.pdf
2. http://www.buyya.com/papers/AnekaMagazineArticle1.pdf
3. https://cloudsecurityalliance.org/topthreats/csathreat s.v1.0.pdf
4. http://www.isaca.org/Groups/ProfessionalEnglish/security-trend/Group Documents/DLP-WP14Sept2010 Research.pdf
5. http://www.istf.jucc.edu.hk/newsletter/General_01/Gen2_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. IJCA International Journal of Computer Applications (0975 – 8887) .Volume 110 – No. 6, January 2015
8. Varun Chandola, Arindam Banerjee, and Vipin Kumar, Outlier Detection – A Survey, Technical Report TR0717, 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, and 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., and 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/PKDD08 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.