IJCTA, 8(5), 2015, pp. 2301-2306

An Enhanced Automated Data Classification System using Complex Networks

Renu S.* **and S. H. Krishnaveni****

Abstract: Automated data classification systems shrink security disquiet to the user and extend the prospect of organizational data management system. Data classification depends on the nature and type of the organization. File sampling, key sampling, sample comparisons and machine learning techniques are used for automated data classification. Sampling network is used for data sampling and key sampling; it is a complex network which splits files into small blocks. Machine learning techniques are used to collect samples from the file blocks. FK list is used for file key comparisons to identify file type and store it into the corresponding security region.

Index Terms: Machine learning, Sampling network, complex network, Key sampling, File sample, FK List.

1. INTRODUCTION

Data classification has a vital role in the administration of management systems. According to the current scenario we have several management techniques. But reviews show that they are not up to the mark. While considering a cross view we can see that the existence of learning techniques does not support automated mechanism. But through machine learning we can yield the system in to an automated one. Introduction of sampling techniques supports this automated system. The commonly used methods in data classification are decision trees, rule-based methods, probabilistic methods, SVM methods, instance-based methods, and neural networks. Networking is the fundamental technique for all communication. But complex networking structure pull backs the optimization factors.

Data classification can be used in different areas such as Machine learning, Pattern recognition, big data analysis and Data mining etc. Here, we are mainly focusing classification of organizational data. Organizational data can be mainly classified into four categories:-Sensitive data, Private data, Protected data and Public data [1]. Data sampling plays an important role in automated data classification. A random sampling technique can be used for selecting samples from different pieces of the file [2].

We can also use machine learning technique for data sampling. Training in large data set will produce an accurate result. Machine learning algorithms can be learned class information from data sets, but the created classes' meaning isn't always clear [3].

The learning and training approach of machine learning can be mainly classified into three categories:-supervised learning, unsupervised learning and reinforcement learning. Supervised learning is a controlled approach, computer is trained with sample input and corresponding output given by a teacher and the goal is to learn general rule and maps input to output. The goal of unsupervised learning is to discover the hidden patterns in data; there is no labeled training for controlling input output formats. In reinforcement learning, to achieve an assured goal computer program must interact with a dynamic environment without explicitly telling whether it close to its goal or not [4].

* Research Scholar, Dept. of Computer Science Engineering, NIU, Kumaracoil, Thuckley, Tamil Nadu, India

** Assistant Professor, NIU, Kumaracoil, Thuckley, Tamil Nadu, India

File sampling is the method of collecting data from different part of the input file. The selection pattern can be simple and every bit of information has equal priority [5]. The difference between individual results within the sample is a good indicator of variance in the overall population, which makes it relatively easy to estimate the accuracy of results.

Clustering is a process of collecting a set of objects or samples into a group in which the samples are more similar to each other than those in other group [6][7].

2. RELATED WORKS

Sandeep K. Sood et.al.[8] Proposed a model for cloud data security, which is divided into two phases, first phase for data storage and second phase for data retrieval. In the first phase a Sensitive Rating algorithm can be used to classify data into four categories such as sensitive, public, private and protected. Confidentiality, integrity and availability factors were used for calculating sensitive rating.

Parikshit et.al. Proposed [9] a model focusing three Dimensional securities in Cloud Computing. The three dimensions are confidentiality, integrity and availability. This model provides three rings of security based on the sensitivity of data. An algorithm was used to categorize the data. The inner ring offer higher security and the outer ring have lower security and the middle ring offers intermediate security.

Liping jing et.at. proposed[10] a model for a stratified sampling method for generating subspace component data sets in ensemble clustering of high dimensional data. The concept feature stratum is used for data sampling which provide better representations of the clustering structure in the original data set.

Geoff B. Irvine et. at. Proposed[11] a model which focus a variable data rate sampling for the intention of low power data acquisition in a small foot-print microsystems. The system facilitates energy saving by developing dynamic power management techniques and is based on the Adams–Bashforth and Adams–Moulton multistep predictor–corrector methods for ordinary differential equations**.** Newton–Gregory backward difference interpolation formulae and past value substitution are used to facilitate sample rate changes. It is necessary to store only 2m + 1 equispaced past values of t and the corresponding values of y, where y =g (t), and m is the number of steps in the Adams methods. To demonstrate this technique, fourth-order methods are used, but it is possible to use higher orders to improve accuracy if required.

Renu S et.al [12] proposed a binary tree model for data classification. A binary tree is used for data classification. This work has mainly divided into three parts namely Simple binary tree approach, Weighted tree approach, Complex network Approach. A binary simple binary tree, which each node represent fundamental security concerns such as Confidentiality, Integrity and Availability. This method is very effective and time consuming. Single path is activated at a time based on the user requirements. A brain storming technique is used for path selection. In weighted tree approach, a sub-tree is activated at a time. Depending on the root value given by the user left or right sub-tree is activated. User has a facility to activate select a path from eight different paths. This method is less time consuming and compared with simple binary tree approach. In complex network approach user select a suitable path based on the path value. Higher the path value shows higher security and vice versa. This process is less time consuming and accuracy is moderate.

3. PROPOSED WORK

The proposed work focused on automated data classification using sampling, clustering, machine learning and complex network and File–Key list (FK list) comparisons. A complex network structure is used for data classification. Fig1.1 shows an overall structure of an automated data classifier. The input node receives a file F and is divided into small blocks such as f1, f2, f3…fn and send it into file hunks. File hunk is small blocks which is used to store and process small file pieces. A machine learning technique or a random sampling technique is used for collecting file samples from file blocks. Each file hunk is connected with an amass node; a node which collect all samples from the connected file hunks. Amass node check file samples to avoid duplication and also it act as a cluster for collecting and processing file samples. Amass nodes are connected with a Chief node; a node is one which collect and co-ordinate the file samples and key samples together. File-key comparisons are performed at the Chief node. FK List is used for file key comparisons.FK List is a dynamic doubly link list for checking and comparison operations.

After checking and comparison operation chief node send file security code to the input node. Input node stores file into the corresponding security region. The main four security regions are sensitive region for storing sensitive and highly confidential data, protected region for storing intermediate security data, private for moderate security data and public region for common data. All security regions are connected with the input node.

FK List is a dynamic doubly link list which performs all file key comparison. FKList consist of file samples such as $s_{1,}$ $s_{2,\ldots}s_n$ and key samples. The key samples are two types' main key samples and sub-key samples. $Kc_{1,}$ $Kc_{2\ldots\ldots}Kc_n$ are represented as main key samples for confidentiality and $Ki_{1,}$ $Ki_{2,}\ldots Ki_n$ are represented as the main key samples for integrity. $Kc_{11,}$ $Kc_{12\ldots\ldots}Kc_{1n}$ are sub key samples of main key Kc_1 and $Ki_{11,}$ $Ki_{12,}\ldots Ki_{1n}$ are the sub key samples of integrity $Ki_{1.}$

Every file samples compare with each main key values, if their is a match found the sub-key values of the corresponding main key value will activated and take place a more detailed comparison. Result of

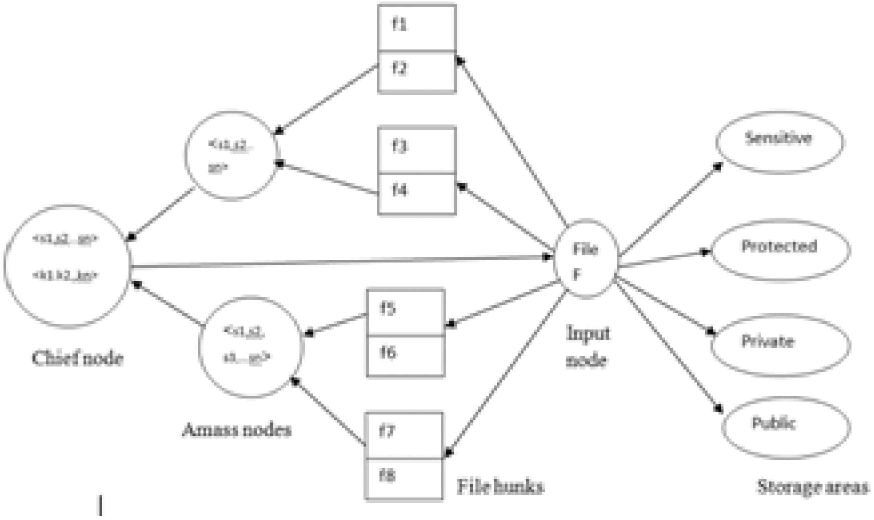
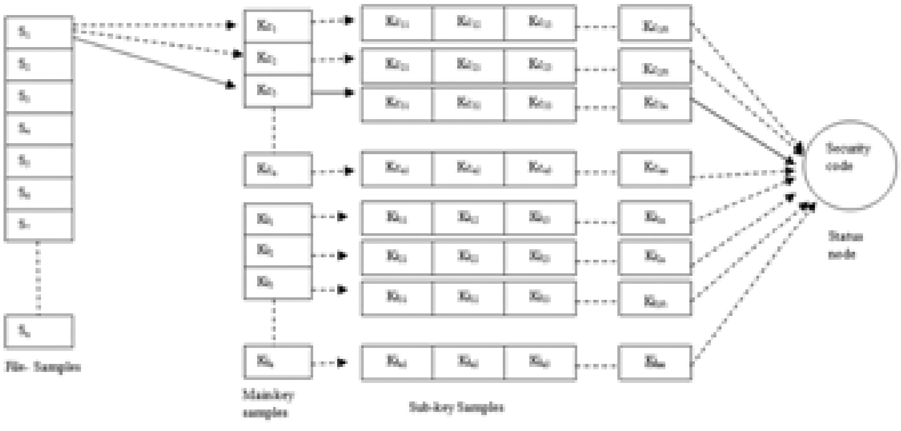


Figure 1: Over view of data classifier



* Key samples (Kc – Key samples confidentiality, Ki – Key samples Integrity)

Figure 2: FKList for file key comparison

comparison is stored at the status node. Status node is a node which keeps and sends security status of the file after detailed comparison.

4. IMPLEMENTATION

4.1. Algorithm for file sampling

```
Input: A file F
Output: sensitive || protected || private || Public
 1. File F is spilt into $f_1$, $f_2$, $f_3$…..$f_n$
 2. Set count = 0;
 3. Push file splits into file hunks.
 4. Collect samples $S_{i,\ Si+1}$…..$S_n$ from file splits.
 5. Send samples into the Amaas node.
 6. Collect samples from file hunks
 7. Delete duplicate samples
 8. Group similar samples
 9. Send to chief node
10. Collect samples from all Amaas nodes
11. Check duplication
12. Combine similar groups.
13. Check sample groups with key groups
14. If any match found
 a. Send security value to the file node
15. File node send files to the corresponding category.
16. Stop
```

ii) Algorithm for Key Comparison

```
Input: File samples $S_{i,\ Si+1,}$….$S_n$ and key samples $K_{i,}$ $K_{i+1,}$….$K_n$
Output: Security code of a file
 1. Check each file sample with main key samples
 2. If any match found
        a.  Activate the sub-key values
            and send security code to the security status node.
 3. Stop.
```

5. SECURITY ANALYSIS

At present we used to consider sampling techniques as an optimizing factor. But the available techniques are not in an automatic manner. So here we used to introduce an automated approach which supports sampling and clustering techniques.

Here for the purpose of sampling, we used to consider a sampling network and hence complexity of network can be further reduced into small blocks. Through that we can have a module based structure and simplified approach.

While considering a new system, security is a crucial factor. Here for ensuring security concerns we concentrates on comparisons and hence identified files are being saved in to the specified security regions.

Management of security is a difficult task. So we used to consider various levels of security abstractions and multi level approaches.

Deep analysis shows that clustering, sampling and machine learning techniques are effective and their combined approach can result a superior automated management model.

	Sandeep K. Sood et.al	Parikshit et.al	Liping jing et.at	Geoff B. Irvine et. at	Renu et.al	Proposed model
Time Consuming	Moderate	Moderate	Moderate	Moderate	Moderate	Less
Accuracy	Moderate	Moderate	Moderate	Moderate	Moderate	High
Automatic	Less	Less	Moderate	Moderate	Less	High
Training	No	No	No	No	No	Yes
Testing	No	No	No	No	No	Yes

5.1 Functionality Analysis

The functionality analysis shows that time consuming of proposed system is less compared to the other available systems. Proposed model using both sampling and machine learning technique, so accuracy of the proposed system is high. Training and Testing helps to increase the accuracy of sampling.

6. CONCLUSION

Automated data classification system classifies organizational data in an efficient and secure way. It is a combination of machine learning and complex network. More effective learning produces more accurate output.

References

[1] The California State University (CSU) 8065.S02 Information Security Data Classification.

[2] “The Removal of Spurious Spectral Peaks From Autoregressive Models for Irregularly Sampled Data” Broersen, P.M.T, Instrumentation and Measurement, IEEE Transactions, Volume: 59, Issue: 1, 2009.

[3] “Visual Classification: Expert Knowledge Guides Machine Learning” MacInnes, J. ; Santosa, S. ; Wright, W., Computer Graphics and Applications, IEEE Volume:30, Issue: 1 , pages 8-14, 2010.

[4] Russell, Stuart; Norvig, Peter (2003) [1995]. Artificial Intelligence: A Modern Approach (2nd ed.). Prentice Hall. ISBN 978-0137903955.

[5] Shahrokh Esfahani, Mohammad; Dougherty, Edward . ”Effect of separate sampling on classification accuracy”. Bioinformatics 30 (2):. doi:10.1093/bioinformatics/btt662, pages 242-250, 2014.

[6] Estivill-Castro, Vladimir (20 June 2002). “Why so many clustering algorithms — A Position Paper”. ACM SIGKDDExplorations Newsletter 4 (1): doi:10.1145/568574.568575, pages: 65–75.

[7] Everitt, Brian “ Cluster analysis.” Chichester, West Sussex, U.K: Wiley. ISBN 9780470749913, 2011.

[8] SANDEEP K SOOD,”A combined approach to ensure data security in cloud computing “, Journal of Network and Computer Applications 35 (2012) 1831–1838, 2012.

[9] Dimensional Security in Cloud Computing parikshit Prasad badarinath oja IEEE 2011.

[10] Liping Jing[a, ,], Kuang Tian[a], Joshua Z. Huang[b,] “Stratified feature sampling method for ensemble clustering of high dimensional data”, Springer, Volume 48, Issue 11, November 2015, Pages 3688–3702, 2015.

[11] Geoff B. Irvine, Lei Wang, Peter Dickman, and David R. S. Cumming, “Variable-Rate Data Sampling for Low-Power Microsystems Using Modified Adams Methods” Ieee Transactions On Signal Processing, Vol. 51, No. 12, December 2003.

[12] Renu S. “A Novel Method to Classify Organizational Data Using CIA Tree approach”, Accepted for IEEE sponsored 2015 Online International Conference on Green Engineering and Technologies (IC-GET 2015).