International Journals of Advanced Research in Computer Science and Software Engineering
ISSN: 2277-128X (Volume-7, Issue-7)

Research Article | **July 2017**

An Algorithm to Mitigate the Attacker by Applying Homomorphic Encryption

Manmeet Kaur, Amrit Kaur
Punjabi University, Patiala, Punjab,
India

DOI: 10.23956/ijarcsse/V7I7/0121

Abstract— K-Modes is an eminent algorithm for clustering data set with categorical attributes. This algorithm is famous for its simplicity and speed. The KModes is an extension of the K-Means algorithm for categorical data. Since K-Modes is used for categorical data so 'Simple Matching Dissimilarity' measure is used instead of Euclidean distance and the 'Modes' of clusters are used instead of 'Means'.The major drawback of k-mode is that the user needs to define the centroid points. To overcome this problem, k-mode with entropy based similarity coefficient was introduced in order to find good initial center points and the accurate result of the clustering is to be obtained. The nature-inspired harmonic algorithm is hybridized to optimize the k-mode algorithm. Harmonic K-Mode Algorithm is proposed in this work that reduces the computation time and improves the accuracy for cluster generation. The performance is evaluated using different output parameters such as execution time, space complexity, accuracy, precision and recall on different datasets such as amazon book review, news aggregator, online retail, seed and wholesale consumer.

Keywords - Data Mining,Clustering, K-Mode Algorithm.

I. DATA MINING

Data mining is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.[1] It is an interdisciplinary subfield of computer science.[1] The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.[1] Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.[1] Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD.

II. INSIDER COLLUSION ATTACK IN KERNAL BASED DATA MINING SYSTEM

A type of security attack or threat in which a node intentionally makes a secret agreement with an adversary, or the node is somehow made to have such an agreement. The adversary may collect confidential information from the system, and then conduct sophisticated attacks on the system by exploiting false data injection through one or more compromised nodes. Collusion is a well-known attacking strategy which basically refers to a set of malicious customers who gather their individual knowledge about the protection system, whatever it is, to obtain unprotected multimedia content. It has first been mentioned in cryptography when protocols have been established to part a secret between different individuals. Typical examples include secret sharing, also referred to as threshold cryptography, and conference keying [1]. The idea of secret sharing is to start with a secret and then to divide it into pieces called shares which are distributed amongst users such that the pooled shares of specific users allow reconstruction of the original secret. These schemes can be exploited to enable shared control for critical actions. Vault deposit accounts are a good illustration of such a procedure. Both the customer key and the bank manager key are required to grant access to the account. If any part of the secret (key) is missing, the door of the vault remains closed. This idea can be extended to more than two people. Access to a top secret laboratory can for instance be controlled by access badges: admittance necessitates a security guard badge and a researcher badge.

III. CLUSTERING USING K-MODE

Clustering is the method of segregating the objects into groups such that the objects in a certain cluster have high degree of similarity with each other than the objects in the other clusters. Clustering process gained boom with the advent of various algorithms for numerical data. KMeans is one of the most commonly used partition based algorithm for numerical data. For categorical data however, various other algorithms are available. KModes which is an extension of K Means is one of such popular algorithms. K-Modes algorithm is an extension of K-Means but with the following differences. First is that a Simple matching dissimilarity function suitable for categorical data is used instead of Euclidean distance. Secondly, Modes are used to represent centroids instead of Mean values and finally a frequency based method is used to find centroids in each iteration of the algorithm. Also, as K-Modes is an extension of K-Means algorithm, the limitations of K-Means are also thus carried forward to K-Modes. One of them is the dependency on initial centroids to improve the accuracy of the clusters. A lot of research has been done, and some is still underway to address this.

K-Mode algorithm is an extension of K-Means which is the partitioning based clustering algorithm. It is an eminent algorithm for clustering data set with categorical attributes and is famous for its simplicity and speed.Modes are used to represent centroids instead of Mean values and finally a frequency based method is used to find centroids in each iteration of the algorithm [6].

Algorithm for K-Mode Clustering:

1. Generate K clusters by arbitrarily selecting data objects and choose K initial cluster centre, one for every of the cluster.

2. Assign data object to the cluster whose cluster centre is near toward it according to equation

$d(X,Y) = \sum_{K=1}^{m} \delta(x_i, y_i)$ ..... *Eq (1)*

$where \delta(x_i, y_i) = \begin{cases} 0, & x_i = y_i \\ 1, & otherwise \end{cases}$ .....*Eq (2)*

3. Update the K cluster base on allocation of data objects. Calculate K latest modes of every one clusters.

4. Repeat step 2 to 3 awaiting no data object has changed cluster relationship otherwise some additional predefined criterion is fulfil.

Further, another major limitation applicable to both KMeans
and K-Modes is to input the number of clusters atthe very beginning based on anticipation. This sometimesmay be off the mark, and hence requires an in-depth priorknowledge of the domain, and good prior idea ofexpected results. This restricts the algorithm to be usedby experienced domain experts only.

IV. LITERATURE SURVEY

Peter Shaojui Wang et al. [1] considered a new insider threat for the privacy preserving work of distributed kernel-based data mining (DKBDM), such as distributed support vector machine. Among several known data breaching problems, those associated with insider attacks have been rising significantly, making this one of the fastest growing types of security breaches. Once considered a negligible concern, insider attacks have risen to be one of the top three central data violations. Insider-related research involving the distribution of kernel-based data mining is limited, resulting in substantial vulnerabilities in designing protection against collaborative organizations. Prior works often fall short by addressing a multifactorial model that is more limited in scope and implementation than addressing insiders within an organization colluding with outsiders. A faulty system allows collusion to go unnoticed when an insider shares data with an outsider, who can then recover the original data from message transmissions (intermediary kernel values) among organizations.

Nilima V. kayarkar [2] aimed to find data of owner which is leaked and detect agent who leaked data. In this paper the agent which is responsible for data leakage is the insider. Means agent is a person who is present within the organization. Insider attacks arise not from system errors but from staff working inside the company's enterprise. The unauthorized transfer of classified information from computer to outside word is called data leakage. For business purpose, it is necessary to send important data to trusted parties. But from this trusted parties this information is reach at unauthorized place such as website or somebody's laptop. It is very challenging and necessary to detect leakage and guilty person responsible for data leakage.

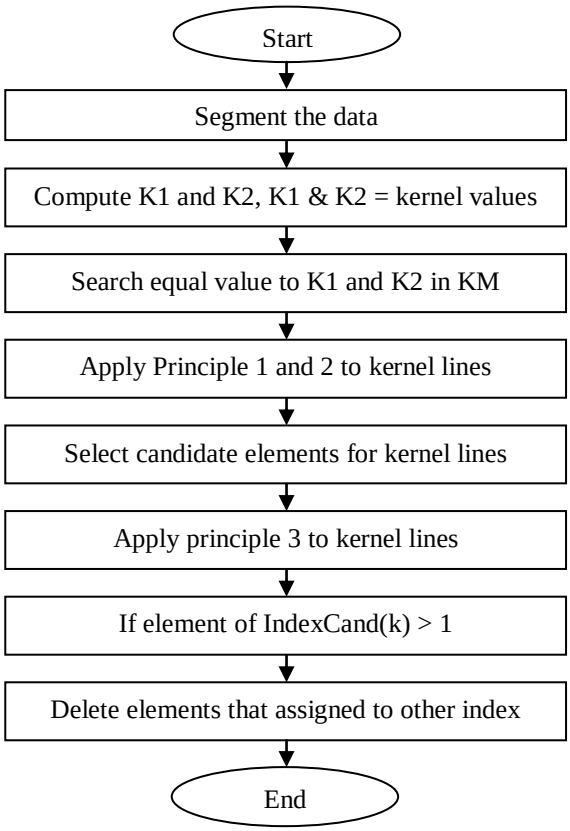
Rupesh Samant and U. V. Wanaskar [3] considered another insider risk for the security protecting work of distributed kernel based data mining (DKBDM), like distributed support vector machine. Among a few known information rupturing issues, those connected with insider attacks have been rising significantly, making this one of the quickest developing sorts of security ruptures. Once considered an immaterial concern, insider attacks have ascended to be one of the main three focal information in fringement. Insider-related research including the appropriation of piece based information mining is constrained, bringing about generous vulnerabilities in outlining insurance against community oriented associations. Earlier works regularly miss the mark by tending to a multi factorial model that is more restricted in degree and execution than tending to insiders inside an association plotting with pariahs. A defective framework permits intrigue to go unnoticed when an insider offers information with an untouchable, who can then recoup the first information from message transmissions (mediator bit values) among associations. This attack required just availability to a couple of information sections inside the associations as opposed to requiring the scrambled authoritative benefits regularly found in the appropriation of information mining situations.

Uma Angadi and G.F Ali Ahammed [4] analyzed some secure data aggregation mechanisms and introduced a new complicated collision attack with its impact on wireless sensor networks. WSN's are usually unattended, they are highly vulnerable to node compromising attacks. Thus making it necessary to ascertain trustworthiness of data and reputation of sensor nodes is crucial for WSN. Iterative Filtering algorithms were found out to be very helpful in this purpose. Such algorithms perform data aggregation and provide trustworthiness assessment to the nodes in the form of weight factors. These algorithms simultaneously aggregate data from multiple sources and provide a trust estimation of these sources, usually in a form of corresponding weight factors assigned to data provided by each source.

Gulrukh Nazneen and Jayant Adhikari [5] defined that privacy takes an important role to secure the data from various probable attackers. When data need to be shared for public advantage as required for Health care and researches, individual privacy is major concern regarding sensitive information. So while publishing such data, privacy should be conserved .While publishing collaborative data to multiple data provider's two types of problem occurs, first is outsider attack and second is insider attack. Outsider attack is by the people who are not data providers and insider attack is by colluding data provider who may use their own data records to understand the data records shared by other data providers.

This problem can be overcome by combining slicing techniques with m-privacy techniques and addition of protocols as secure multiparty computation and trusted third party will increase the privacy of system effectively.
ZulkefliMohdYusop and JemalAbawajy [6] described that Insider attacks become a severe threat to organizations. The emergence of Cloud computing that provides computing as a utility has attracted organizations to store their sensitive data remotely by subscribing the virtual storage from Cloud service provider. While data outsourcing relieves the data owners from burden of local data storage maintenance and security, the steps of embracing Cloud storage service has led to security problems. With the services provided by Cloud service provider that can be extended from Cloud user to Private Cloud and expanded to Public Cloud, there are many possibilities that malicious insider attacks may occur to exploit the weaknesses of Cloud systems. Until now there are no perfect mitigation strategies that can be relied on to solve the threats.
Clifton Phua et al. [7] categorised, compares, and summarises from almost all published technical and review articles in automated fraud detection within the last 10 years. It defines the professional fraudster, formalises the main types and subtypes of known fraud, and presents the nature of data evidence collected within affected industries. Within the business context of mining the data to achieve higher cost savings, this research presents methods and techniques together with their problems.
Pooja Mishra and Gaurav Gupta [8] defined that Node replication detection is a challenging problem. Though the defending against node replication attacks demands immediate attention as compared to the extensive exploration on the defense against node replication attacks in static networks, only a few solutions in mobile networks have been presented. Additionally, whereas most of the presented schemes in static networks exist on the witness-finding strategy that cannot be applied to mobile networks, the velocity-exceeding strategy used in existing schemes in mobile networks incurs efficiency and security problems. Thus, based on our devised challenge-and response and encounter-number approaches, required algorithms are proposed to resist node replication attacks in mobile sensor networks.
ShikhaAgrawal and JitendraAgrawal [9] reviewed various data mining techniques for anomaly detection to provide better understanding among the existing techniques that may help interested researchers to work future in this direction. The data when transferred or stored is primed exposed to attack. Although various techniques or applications are available to protect data, loopholes exist. Thus to analyze data and to determine various kind of attack data mining techniques have emerged to make it less vulnerable. Anomaly detection uses these data mining techniques to detect the surprising behaviour hidden within data increasing the chances of being intruded or attacked. Various hybrid approaches have also been made in order to detect known and unknown attacks more accurately.

V. FLOW CHART



Flow Chart

Principle 1: Because there is a symmetrical property in the kernel matrix, consider only vertical and horizontal kernel lines.
Principle 2: Merge the kernel lines for the same axis of the index, because they all represent the same index.

Principle 3: Remove the kernel lines representing the indices of the other insider's data. Some indices may have been labeled as the insiders' as a result of past search results; therefore, they should not be considered again. Apply homomorphic encryption to the data for protection of insider attack.

VI. EXPERIMENTAL STUDY

A. Datasets

The datasets used in this work is taken from UCI Repository dataset i.e. amazon book review, news aggregator, online retail, seed and wholesale consumer. The total number of instances is 1500 and attributes is 10000 in amazon book review dataset. The total number of instances is 42293 and attributes is 5 in news aggregator dataset. The total number of instances is 54190 and attributes is 8 in online retail dataset. The total number of instances is 210 and attributes is 7 in seed dataset.The total number of instances is 440 and attributes is 8 in wholesale consumer dataset.

B. Output Parameters

The result is compared using five parameters execution time, space complexity, accuracy, precision and recall.

i. **Execution Time:** The execution time is defined as the time spent by the system executing the task.

ii. **Space Complexity:** Space complexity is a measure of the amount of working storage an algorithm needs means how much memoryis needed at any point in the algorithm.

iii. **Accuracy:** The Accuracy is the total number of module that is predicted correctly.

$$\text{Accuracy}=\frac{(TP+TN)}{(TP+FP+TN+FN)} \quad \text{…….. Eq (1)}$$

iv. **Precision:** Precision is the measure of exactness i.e. what percentage of tuples labeled as positive that are actually such.

$$\text{Precision}=\frac{TP}{TP+FP} \quad \text{…….. Eq (2)}$$

v. **Recall:** Recall is the measure of completeness i.e. what percentage of positive tuples did the classifier label as positive.

$$\text{Recall}=\frac{TP}{TP+FN} \quad \text{…….. Eq (3)}$$

C. Results

The performance of the algorithms can be evaluated by using output parameters for the number of cluster 10. Following are the comparison of the results among the algorithms:

i. Execution Time

The graph of the execution time is computed in nanoseconds for the algorithms applied on different datasets are shown in the figure 1.

The figure 1 shows that the proposed algorithmhas better execution time than the existing algorithms.

ii. Space Complexity

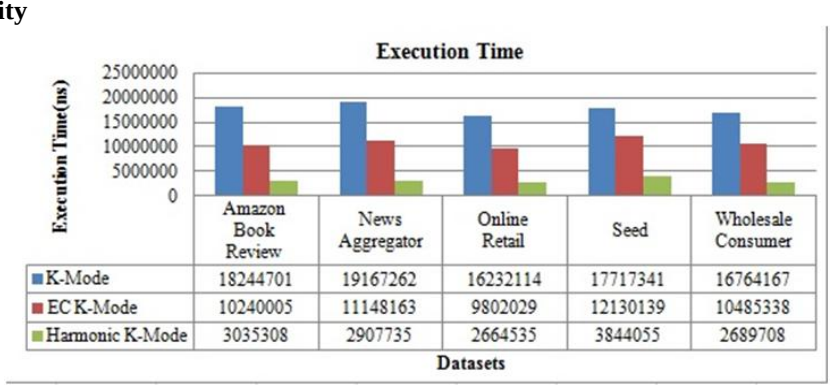


Fig 1: Execution Time

The space complexity is computed in bytes for the algorithms applied on different datasets are shown as below:

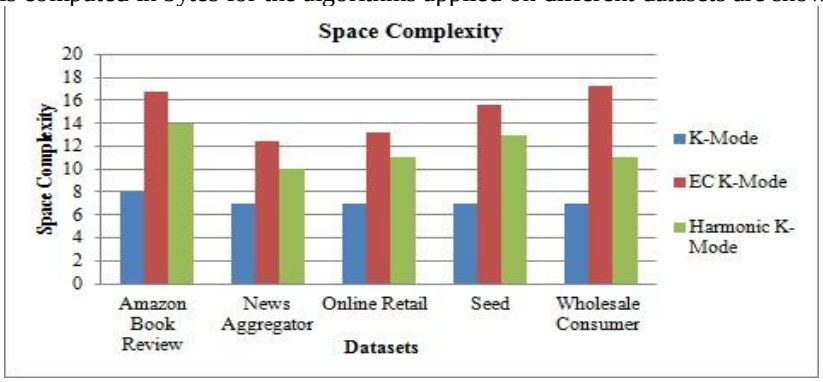


Fig 2: Space Complexity

The figure 2 shows that the space complexity for proposed algorithm is less than the EC K-Mode algorithm.

iii. Accuracy

The graph of the accuracy of the algorithms for different datasets is shown as below:

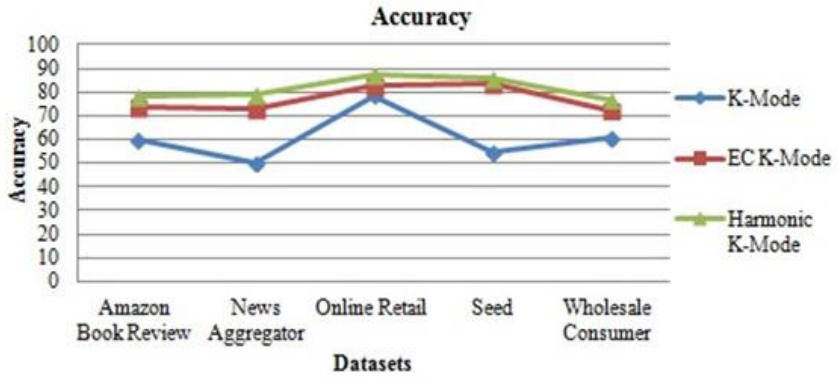


Fig 3: Accuracy

The figure 3 shows that the proposed algorithmhas the better accuracy than K-Mode and EC K-Mode Algorithms.

iv. Precision

The comparison of the precision of the algorithms for different datasets is shown as below:

Table 1: Precision

Dataset	K-Mode	EC K-Mode	Proposed
Amazon book review	51	56	57.6
News aggregator	51.75	67.8	68.6
Online retail	54.3	59	64
Seed	46.5	56.2	73.52
Wholesale Consumer	57.7	65.4	60.87

The table 1 shows that the precision for proposed algorithmis better than the existing algorithm but EC K-Mode has the more precision than K-Mode and Harmonic K-Mode due to the random selection of initial centroids in the wholesale consumer dataset.

v. Recall

The comparison of the recall of the algorithms for different datasets is shown as below:

Table 2: Recall

Dataset	K-Mode	EC K-Mode	Proposed
Amazon book review	66.3	70.2	77
News aggregator	54.3	59.4	62
Online retail	66.75	75.64	80.64
Seed	64.56	69.7	74
Wholesale Consumer	55.2	62.1	53

The table 2 shows that the recall for proposed algorithm is better than the existing algorithm but EC K-Mode has the more recall than K-Mode and Harmonic K-Mode due to the random selection of initial centroids in the wholesale consumer dataset.

VII. CONCLUSION

The various clustering techniques are studied and the k-modes algorithm is a popular clustering algorithm and is linearly scalable with respect to the dataset size. K-Modes algorithm is an extension of K-Means and uses simple matching dissimilarity function instead of Euclidean distance. It is especially sensitive to the selection of initial cluster centers and choosing the proper initial cluster centers is a key step for k-mode clustering. It does not guarantee for the optimal solution. To overcome this problem, k-mode with entropy based similarity coefficient was introduced in order to find good initial center points and the accurate result of the clustering is to be obtained. The nature-inspired harmonic algorithm is hybridized to optimize the k-mode algorithm. Harmonic K-Mode algorithm is proposed in this work.

- Harmonic K-Mode Algorithm is the proposed algorithm that may reduce the computation time and improves the accuracy for cluster generation.
- Different parameters are used to analyze the result of optimized k-mode algorithm.
- The performance of the harmonic k-mode is evaluated in terms of execution time, space complexity, accuracy, precision and recall on the different dataset.

- The performance of the proposed algorithm with the existing algorithm has compared using various output parameters for the number of clusters from 2 to 10.
- There is also the comparison of the results of different datasets for harmonic k-mode algorithm for the number of clusters from 2 to 10.

The experimental results show that the proposed algorithm has better results than the existing algorithm.

REFERENCES

[1] Peter Shaojui Wang, Feipei Lai, Hsu-Chun Hsiao, Ja-Ling Wu, "Insider Collusion Attack on Privacy-Preserving Kernel-Based Data Mining Systems", Special Section on Latest Advances And Emerging Applications of Data Hiding, IEEE Access, ISSN: 2169-3536, vol: 4, 2016, pp: 2244-2255

[2] Nilima V. kayarkar, "A Review on insider Collusion Attack on Text, Video and Image File using Privacy-Preserving Kernel based System", International Conference on Recent Trends in Engineering Science and Technology, ISSN: 2321-8169, vol: 5, Issue: 1, 2017, pp: 156-158

[3] RupeshSamant, U. V. Wanaskar, "Survey on Privacy Preserving of Kernel-Based Data Mining Systems from Insider Collusion Attack", International Journal of Innovative Research in Computer and Communication Engineering, ISSN(Online): 2320-9801, ISSN (Print): 2320-9798, vol: 4, Issue: 12, 2016, pp: 21055-21059

[4] Uma Angadi, G.F Ali Ahammed, "Analysis of Secure Data Aggregation Technique for Wireless Sensor Network in the Presence of Adversary Environment based on IF algorithm", International Journal of Advanced Research in Computer and Communication Engineering, ISSN (Online): 2278-1021, ISSN (Print): 2319 5940 vol. 5, Issue 4, 2016

[5] GulrukhNazneen, JayantAdhikari, "Research Paper on m-Privacy for Collaborative Data Publishing", Imperial Journal of Interdisciplinary Research, ISSN: 2454-1362, vol: 2, Issue: 7, 2016, pp: 1407-1409

[6] ZulkefliMohdYusop, JemalAbawajy, "Analysis of Insiders Attack Mitigation Strategies", International Conference on Innovation, Management and Technology Research, vol: 129, 2014, pp: 581-591

[7] Clifton Phua, Vincent Lee, Kate Smith, Ross Gayler, "A Comprehensive Survey of Data Mining-based Fraud Detection Research", Intelligent Computation Technology and Automation, IEEE, DOI: 10.1109/ICICTA.2010.831, 2010

[8] Pooja Mishra, Gaurav Gupta, "Detection of Node Replication Attacks in MSN Using EDD Algorithms", International Research Journal of Engineering and Technology, e-ISSN: 2395 -0056, p-ISSN: 2395-0072, vol: 3, Issue: 3, 2016, pp: 655-659

[9] ShikhaAgrawal, JitendraAgrawal, "Survey on Anomaly Detection using Data Mining Techniques", 19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems, vol: 60, 2015, pp: 708-713

