International Journal of Artificial Intelligence
p-ISSN: 2407-7275, e-ISSN: 2686-3251

Original Research Paper

Privacy and Customization of Clusters with Application Development

Chandana[1], Pallavi[1], Rekha Sahithi[1], Shwetha[1], Sumithra Devi[1]

[1] *Department of Information Science and Engineering, Dayananda Sagar Academy of Technology and Management Bangalore. Karnataka, India.*

Article History
Received:
28.04.2022

Revised:
21.05.2022

Accepted:
29.05.2022

***Corresponding Author:**
Chandana, G. R.
Email
grchandana63@gmail.com

Abstract: This paper presents a high level view of how clusters are being used in large number of domains for preserving and protecting the data. Because of these, clusters are being exposed to many attacks coming from open network. Hence there are many methods to design a privacy preserving clusters. To ensure these preserving clusters, cluster validity measurements are done for different type of Data. Protocols are used to do the privacy preserving. The clusters analysis are used in banking sector for identification of the bank customer profile. Algorithms are used to find the Sensitive data before making the individual data into clusters of data .and then the privacy is applied only on these sensitive data different financial data sets are used and trained. Dealing with the clustering algorithms which suits for Big Data. Data mining tools are used in banking sector to configure large data and communication between customers and the bank. Timely information on fraudulent activities is the strategy for the banking sector. Different fraudulent are used for identification of the frauds. We identify a sets of privacy profiles as way to help users manage their mobile applications privacy preferences. These preferences help in people comfort in handling the options in bank applications and providing the security model for these banking applications. Overall this papers is all about banking applications which will let users customise their data clusters which has to be protected and different level of privacy is introduced in other to access this private data.

Keywords: Data Mining, Fraudulent Activities, Open Network, Preserving Cluster, Sensitive Data.

International Journal of Artificial Intelligence, vol. 9, no. 1, pp. 24-31, June 2022. **DOI: 10.36079/lamintang.ijai-0901.357**

1. Introduction

A cluster is a collection of interconnected computers that share and coordinate the use of resources for a wide variety of users. Mechanisms for effective and flexible security management have become vital, but they are also difficult to implement. Because these clusters are employed in so many different fields, their security and protection are critical. When given a collection of data, cluster algorithms aggregate related things together [1]. Because online browsing, email, and other services leak personal information, privacy is sacrificed in exchange for service benefits. In other domains, though, privacy is required. For various purposes, a variety of validity techniques are introduced. Because clustering is an unsupervised process, it's critical to assess the cluster algorithm. For this investigation, many validity indices are introduced. The term "data mining" refers to the process of collecting usable information from a set of data.

The extracted information could be in any format. These generated privacy worries about the information gathered. The majority of clustering algorithms work with both numerical and categorical input. Researchers in bioinformatics, on the other hand, may employ clustering alphanumeric data. Because of its effectiveness, data mining is utilised to combat fraud. It's a well-defined procedure that starts with data and ends with models or patterns. In our research, we also apply NN, a data mining technique. The design of the credit card detection system's neural network architecture is an unsupervised method that uses transaction data to produce four separate clusters. IT has aided the banking business in dealing with a variety of economic difficulties. Banks have realised that client relationships are a critical component of their success in recent years. Customer relationship management (CRM) is a technique that will assist them in developing long-term ties with customers while also increasing revenues and profits. CRM is very important in the banking industry. In this research, we will use TwoStep Cluster to identify bank clients' profiles, starting with a public dataset provided by a German bank. This method has the advantage of determining the correct number of clusters; thus, the goal is to find the appropriate number of profiles in order to effectively manage current and potential clients

2. Related Work

Let's concentrate on the work that has already been done on privacy-preserving clustering. Machine learning algorithms can be used in two ways to create privacy-preserving designs. The first is to apply transformations to the financial data sets before using the algorithm. Several academics have used this method for building privacy-preserving clustering algorithms. A second method for designing privacy-preserving algorithms is to leverage methods from secure multiparty computation. The advantage of this approach over perturbation is that these algorithms may be guaranteed to be private.

2.1. The k-Means Clustering Algorithm

Algorithm k-means (privacy-preserving k-means clustering) begin initialize nA, c, µ1, • • • , µc do classify nA samples according to nearest µi for i := 1 to c step 1 do Let C A i be the *i*-th cluster.

Compute ai = P xj∈CA i xj and bi =| C A i | recompute µi by invoking the protocol PW
Do until no change in µi return µ1, µ2, • • •, µc
end.

We must build a privacy-preserving protocol to compute the cluster means in order to create a privacy-preserving k-means that does not utilise a Third TP. For the Weighted Average, a Privacy-Preserving Protocol Problem.

In clustering algorithms, the WAP protocol is utilised for privacy. Every iteration of a k-means algorithm recalculates the means and reclassifies the K-means samples. When it discovers "no change" within the k means, the algorithm is terminated. The right interpretation of "no change" is dependent on the metric being used. We'll suppose that the initial cluster means are selected at random [2].

Big data and data mining methods hold the vast data set in e-business. Data mining tasks leads to the identifying the data subjects and also the disclosure of personal data. To overcome these issues, privacy- preserving data mining approaches have been developed to fulfil the first, contradictory condition. It has gained popular as a result of incorporating privacy protections into data mining jobs.

International Journal of Artificial Intelligence, vol. 9, no. 1, pp. 24-31, June 2022. **DOI: 10.36079/lamintang.ijai-0901.357**

SOM, as an unsupervised competitive learning technique, is effective in dividing input data into individual nearest clusters.

The SOM cluster technique improves the scalability of the recommendation process by stabilising the online computational complexity. To securely launch SOM, we're developing a methodology based on the Genetic Algorithm, which has recently become critical for researchers in addressing complex problems. Some methods for preserving privacy for separated data have already been proposed. These studies have aided data owners in cooperating when they have insufficient data and need to combine fragmented data for new facilities [3]. Lindell and Pikans present a privacy-preserving ID3 algorithm based on cryptographic approaches for horizontally partitioned data, which is followed by Clifton. On securing a two-party scalar product protocol, Vaidya and Clifton provided privacy- preserving association rule mining for vertical partitioned data. Privacy-preserving another popular method for resolving privacy issues in partitioned data is the Nave Bayes classifier.

Cluster validity assessment is the process of determining the outcomes of a clustering algorithm. For evaluating and selecting an optimal clustering system, two measuring criteria have been implemented:

Compactness: Each cluster's members should be as near as feasible to one another. The variance is a popular measure of compactness.

Separation: The clusters themselves should be separated by a large distance. The distance between two clusters can be measured in three ways: distance between the clusters' closest members, distance between the clusters' most distant members, and distance between the clusters' centres.

To evaluate the results of clustering algorithms, we must use different criteria, including External Criteria, Internal Criteria, and Relative Criteria. Internal and external criteria are both based on statistical approaches and have a high computational requirement. The clusters are operated by external validity algorithms based on a few user-specific intuitions. Internal criteria are based on a few metrics derived from data sets and a clustering structure. The validity index is the foundation of the comparison. Several validity indexes have been created and introduced.

These indices are used to check the goodness of the clusters.they are:

1. Dunn and Dunn like Indices: if a data set contains well-separated clusters, the distances among the clusters are generally huge and the diameters of the clusters are refered to be small.
2. Davies Bouldin Index: It is based on similarity measure of clusters. It measures the average of similar in each and every cluster and it is most the similar one.
3. RMSSDT and RS Validity Indices: hierarchical clustering algorithms use these indices.
4. SD Validity Index: the average of scattering of clusters is done and the total separation of clusters is done.

The partitioning of the dataset is taken into account to solve the clustering challenge. The clustering algorithms utilised are as follows: Partitional clustering, hierarchical clustering, density-based clustering, and grid-based clustering are some of the several types of clustering. It's possible that the outcomes will differ. For underlying data, it is the best option. When evaluating high-dimensional data, 2d datasets are typically employed. Visualization and validation are not activities that can be completed by a formal technique. Compactness and separation technique are also taken into account when clustering is ideal. We will give a general overview of the literature that is connected to machine learning algorithms in this paper. A decision tree algorithm is used in conjunction with a classification method. For discrete value datasets, the id3 technique can be utilised. The regularity logistics regression algorithm is built to solve difficult optimization problems of this type. Because the amount of network protocols has exploded, including malware and bonnets, security researchers have attempted to automate protocol reverse engineering. For intrusion detection and packet inspection, automated protocols reverse engineering has been utilised [4] [5]. The goal of program-based analysis was to examine target protocols using a binary programme and some trace inputs. For instance, consider a server and client software [6]. Finally, protocol fuzziness research puts a target system to the test by injecting packets that look like authentic ones [7]. Lindell and Pikans present a privacy-preserving id3 algorithm for horizontally partitioned data based on cryptographic approaches, which is followed by Clifton. Protecting your privacy another common method for resolving privacy issues in partitioned data is the Nave Bayes classifier. Soft computing is another current issue in the modern day that causes privacy concerns in various settings. The goal is to

International Journal of Artificial Intelligence, vol. 9, no. 1, pp. 24-31, June 2022. **DOI: 10.36079/lamintang.ijai-0901.357**

maintain database security while maintaining the highest levels of efficacy and confidence for mined rules [8] [9].

We propose a threat model that guides our efforts in securing clusters in order to prioritise our efforts. The purpose of this threat model is not to cover every possible attack scenario, which is impossible given the constant emergence of new threats and vulnerabilities. Rather, the purpose of this threat model is to comprehend a security posture in light of the fact that threats are many, no protection system is ideal, and protection resources are limited.

The cluster security threat model has three unique features.

- Changing Nature of Clusters
- Shift from Random Reliability Failures to Intentional Attacks
- Security as a Service versus Security as an Obligation

Distributed Security Service: The distributed security functionality needed for a secure service invocation/communication between two objects in different nodes in the cluster can be summarized as follows:

1. Authenticating the source and target objects.
2. Deciding whether the source object can perform this action on the target object.
3. Auditing the action.
4. Protecting the data flow from being modified or eavesdropped during the transit between nodes.

We have given complementary security techniques for clusters ranging from carrier-class to High-Performance Computing clusters in this work. Clusters at the carrier-class end of the cluster environment spectrum must be locked down to the greatest extent possible, with a focus on product. A unified security model with distributed authentication and distributed access control [8] is the corresponding security method we describe for this carrier-class cluster environment.

A lot of previous research on smartphone apps has centred on establishing practical strategies for detecting and managing sensitive personal information leakage [9]. We also provide a summary of the relevant mobile privacy literature, which is divided into three categories [3].

1. Finer-grained Privacy Control: In Android, apps can only be accessible by sensitive resources if they publish permission requests in the manifest files and get permission from users to access these rights at the time of download.
2. Modeling People's Mobile App Privacy Preferences: A second line of research has looked into user concerns and preferences around mobile app privacy.
3. Privacy Preference Learning: Frank et al. reported a first data mining research on mobile app permissions, in which they looked for permission request patterns in Android apps.

The privacy preferences of users in mobile apps are not consistent. This study quantified the fact that mobile app users have a wide range of privacy preferences.

This suggested that crowdsourcing people's average preferences, as suggested by Agarwal and Hall in the PMP privacy settings, could be a good way to go. Despite the diversity, we find that there are only a few small groups of like-minded individuals who have many of the same interests [1].

This report adds to existing mobile app privacy research by quantifying the relationship between app privacy-related behaviours and user privacy preferences. We used static analysis to see how and why third-party libraries use different sensitive resources, and we used crowdsourcing to collect privacy preferences from over 700 people for over 800 apps [10].

Since the last decade, privacy issues with LBSs have been extensively researched. The majority of research efforts are focused on location privacy, or securing user location information before using it. Obfuscation of location is the most common method. The main concept here is to use techniques like cloaking region and dummy location to conceal user position using imprecise locations.

To protect user privacy while publishing user data, current practise primarily relies on policies, such as those governing the employment and storage of published data. However, this strategy does not guarantee that a hostile attacker will not gain access to the user's sensitive information. As a result, privacy-preserving data publishing has been extensively researched in order to provide effective privacy protection when releasing user data.

International Journal of Artificial Intelligence, vol. 9, no. 1, pp. 24-31, June 2022. **DOI: 10.36079/lamintang.ijai-0901.357**

In this paper, we look at the inference attack. In an LBSN, a user has two sorts of data: public check-in data (public data) that she is ready to share in order to enable customised LBSs, and private data (private data) that she wishes to keep secret, such as gender.

Learning Cluster-wise Obfuscation Functions: We train the best obfuscation function based on user clusters, which means we strive to disguise users' "lifestyles" for privacy protection [7].

The low, high, dangerous, and high risk clusters were employed in the development of the credit card fraud detection system. If a transaction is valid, it is processed; however, if it falls into one of these clusters, it is considered fraudulent. The alarm sounds, and the reason is explained. The fraudulent transaction will be recorded in the database rather being processed.

The main functions of the credit card detection system created with an artificial neural network (ANN) are to allow real-time transaction input and to react to a suspicious transaction that could lead to fraud. The architecture was created using an unsupervised neural network method that was used to transaction data to construct four clusters: low, high, dangerous, and high-risk clusters [11].

The withdrawal and deposit units are the two units that make up the fraud detection system in this study. Each of the two units is made up of the subunits listed below: the database interface, the neural network classification, and the visualization.

The research yielded a model that could detect rapid changes in known patterns and recognise common fraud usage patterns. The CCF detection system was created to run in the background of existing banking software in order to detect unauthorised transactions as they occurred in real time. This method of detecting fraudulent transactions proved to be quite effective and efficient [12].

2.2. Big Data Analysis

Data collection and storage has become simple and convenient in the age of big data. The scale of big data sets is rapidly increasing, posing a significant barrier to data processing. As a result, massive data set clustering research is constantly evolving. Clustering techniques for several sorts of data sets have achieved new levels of accuracy. These algorithms, however, have a number of flaws while working with enormous amounts of data High computational complexity and a long computation time are the key flaws [13] [14].

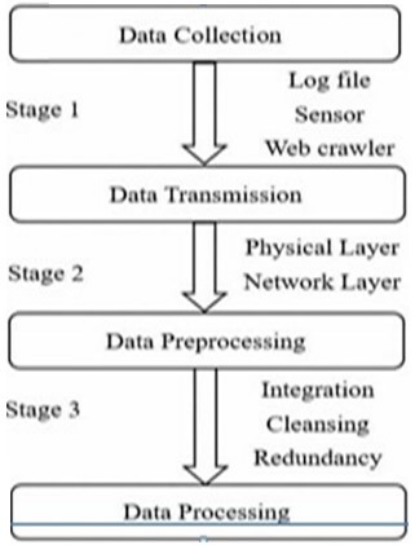

Figure 1. Stage in Big Data Analysis

2.2.1. Clustering Algorithm of Big Data Set

Serial and parallel processing are the two methods for executing the algorithm. The term "serial" refers to the software being executed in order on a single actuator. The term "parallel" refers to the simultaneous execution of a programme on many processors [8].

International Journal of Artificial Intelligence, vol. 9, no. 1, pp. 24-31, June 2022. **DOI: 10.36079/lamintang.ijai-0901.357**

2.2.2. Clustering Algorithm on Mapreduce Computing Framework

Google introduced the MapReduce parallel computing model. It has something to do with huge data sets. The Hadoop distributed file system, which is at the heart of the computing paradigm, divides data and schedules processing activities.

2.2.3. Clustering Algorithm on Spark Computing

Apache Spark is an open source memory computing cluster computing framework. It has greater benefits than MapReduce. There is no need to read or write HDFS because the instantaneous output may be saved in memory. As a result, Spark is better suited to iterative data mining and machine learning techniques.

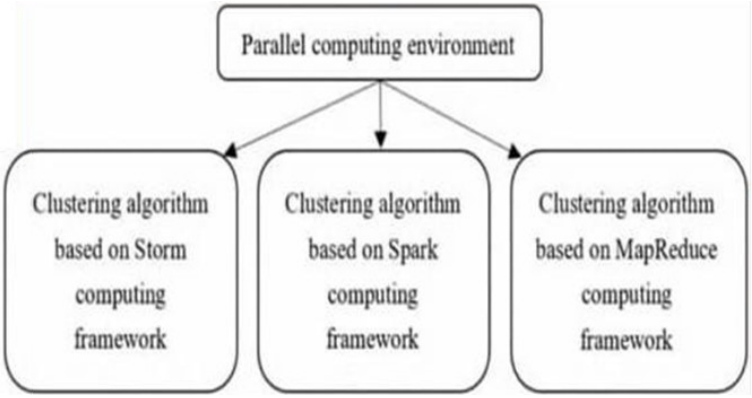

Figure 2. Pararel Computing Environment

2.3. Mobile Application

Mobile computing has caught attention of research community for quite some time. Reached the people smart phones. These and devices are running rich stand-alone chent server applications accessing information six web gateway.ways to develop [9] [14] [15]:

- Mobile network operater
- Phone manufactures
- Mobile application & content providers.

Created impact for independent & freelance developers. Defining the platform providers as provider of OS and development tools to enable creation of high level applications.5 big platform providers.

2.4. Current Practices

The current practice are:

Nokia: Symbian 05 (46-6%)
Apple: iphone 05 (19.3%)
RIM: Blackberry 05 (15.20%)
Microsoft: Willows CE DS (13.6%)
LiMo foundation: Linux OS (5.1%)

Firstly, classify the platform in 3 main components:

1. Development tools
2. Portal
3. Mobile devices.

Development tools to build the mobile application portal the developed application, mobile application for consumer to downland applications.

International Journal of Artificial Intelligence, vol. 9, no. 1, pp. 24-31, June 2022. **DOI: 10.36079/lamintang.ijai-0901.357**

3. Development Tools

Every platform has an SDK that allows third-party developers to create applications that operate on the platform. Many libraries and tools are included in this SDK. When it comes to sharing SDKs that enable and with developers, several ways are used, and some even restrict access as much as possible. The cathedral model entails half of the actors and half of the clients following the proprietory path (Apple, Microsoft). The other half of the users prefer open course technology (linux, google, nokia).

Open platforms use the bazaar model, which gives developers access to all components of their SDK and OS source code. As we found, the Linux, Google, and Nokia are the three most popular open source operating systems [14] [15].

Passes applications from developers to consumers an application is created.

1. Decentralised portal
   Developers can freely upload their app into any third party portal Disadvantage costumes have great variety of portal don't provide comprehensive oversight existing applications! Example: Linux, Microsoft, Limo
2. Centralised portal
   There is a main portal on which all apps are proposed.Example Apple & Google.

4. Result and Discussion

Complementary security measures for clusters are discussed in this study, ranging from carrier-class clusters to High-Performance Computing clusters. Clusters at the carrier-class end of the cluster environment spectrum must be locked down to the greatest extent possible, with a focus on production dependability. Most clustering algorithms' final results are sensitive to accuracy factors in data mining, leading to algorithms being referred to as mature and practically intelligent machine learning algorithms. The approach detects issues in applications, including both unknown and known harmful programmes. The security of applications is ensured by the protection technique. Two privacy- preserving k-means algorithms were presented. We also developed these algorithms and conducted extensive evaluations of them.

Few platforms fours their business providing as into developers and others integrate entire process. There are four types of integration's.

1. Full integration
2. Portal Integration
3. Device integration
4. No integration

5. Conclusion

Complementary security measures for clusters are discussed in this study, ranging from carrier-class clusters to High-Performance Computing clusters. Clusters at the carrier-class end of the cluster environment spectrum must be locked down to the greatest extent possible, with a focus on production dependability. Most clustering algorithms' final results are sensitive to accuracy factors in data mining, leading to algorithms being referred to as mature and practically intelligent machine learning algorithms. The approach detects issues in applications, including both unknown and known harmful programmes. The security of applications is ensured by the protection technique. Two privacy- preserving k-means algorithms were presented. We also developed these algorithms and conducted extensive evaluations of them. There are a few places where more research is needed. To reduce execution and bandwidth overheads, we need to make more improvements to our tools. Other clustering techniques with privacy-preserving variants are something we'd want to investigate. Hierarchical clustering methods are of great importance to us. Data warehousing is used to combine data from diverse data sources into a usable format so that the data may be mined. The data has been analysed, and it is now being used throughout the organisation to aid decision-making.

References

[1] S. Jha, L. Kruger, P. Mcdaniel, “Privacy Preserving Clustering,” in *Proceedings of the 10th European Symposium on Research in Computer Security*, 2019.

[2] T. Tian, D. Hua and H. Guoping, "Privacy-Preserving Classification on Horizontally Partitioned Data," *2010 International Conference on Computational Intelligence and Security*, pp. 230-233, 2010.

[3] F. Amiri, G. Quirchmayr and P. Kieseberg, "Sensitive Data Anonnymization Using Genetic Algorithms for SOM-based Clustering," *SECURWARE 2018, The Twelfth International Conference on Emerging Security Information, Systems and Technologies, Venice, Italy*, 2018.

[4] R. Alexander, F. Guggenmos, J. Lockl, Jannik G. Fridgen and N. Urbach, "Building a Blockchain Application that Complies with the EU General Data Protection Regulation," in *MIS Quarterly Executive*, vol. 18, pp. 263-279, 2019.

[5] A. Holzer and J. Ondrus, "Trends in Mobile Application Development", *Lecture Notes of the Institute for Computer Sciences, Social-Informaticsand Telecommunications Engineering*, vol. 12, pp. 55-64, 2019.

[6] I. Y. Panessai, M. M. Lakulu, M. H. Abdul, N. A. Z. Rahman, N. Iksan, R. M. Rasli, M. R. Husin, H. Ahmad, S. A. Ariffin, A. Alias, S. K. Subramaniam, S. Majid, M. A. Bora, A. A. Bilong, N. M. Mazli, "Predicting Students' Academic Performance: A Review for the Attribute Used," *International Journal of Academic Research in Business and Social Sciences*, vol. 11, no. 4, pp. 595-603, 2021.

[7] I. Y. Panessai, M. M. Lakulu, M. H. Abdul Rahman, N. A. Z. M. Noor, N. S. Mat Salleh, and A. A. Bilong, "PSAP: Improving Accuracy of Students' Final Grade Prediction using ID3 and C4.5", *International Journal of Artificial Intelligence*, vol. 6, no. 2, pp. 125-133, Dec. 2019.

[8] Dr. K. Chitra, B. Subashini, "Data Mining Techniques and Its Applications in Banking Sector", *International Journal of Emerging Technology and Advanced Engineeringi*, vol. 3, pp. 8, 2019.

[9] M. Pourzandi, D. Gordon, W. Yurcik and G. A. Koenig, Clusters and Security: *Distributed Security for Distributed Systems*. University of Illinois: USA, 2019.

[10] W. Chunqiong, "Research on Clustering Algorithm Based on Big Data Background", *Journal of Physics: Conference Series*, vol. 12, no. 2, 2018.

[11] L. Jialiu, B. Liu, N. Sadeh, Modeling Users' Mobile App Privacy Preferences: *Restoring Usability in a Sea of Permission Settings*. Carnegie Mellon University: United States, 2018.

[12] M. S. Baba, I. Y. Panessai, and N. Iksan, "Solving Rich Vehicle Routing Problem Using Three Steps Heuristic", *International Journal of Artificial Intelligence*, vol. 1, no. 1, pp. 1-19, Jun. 2019.

[13] I. Y. Panessai, M. S. Baba, and N. Iksan,"Applied genetic algorithm for solving rich VRP", *Applied Artificial Intelligence*, vol. 28, no. 10, pp. 957-991, 2014.

[14] F. Ogwueleka, "Data Mining Application in Credit Card Fraud Detection System", *Journal of Engineering Science and Technology*, vol. 6, no. 3, pp. 311-322, 2019.

[15] F. Kovács, C. Legány and A. Babos, "Cluster Validity Measurement Techniques", *Proceedings of the 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases*, 2017.