A Novel Approach for Location PrivacyPreservation of Clusters

Nayan Prakash[1] , Rishu Raj[2],Riya Gothi[3] ,Sahil Singh[4],Dr.Rashmi Amardeep[5]
Department of ISE,DSATMBANGALORE,INDIA
nayanprakash2002@gmail.com

Abstract:
Location privacy is one of the burning issues in respect of the LBS and other services, since there are several risks of data leakage happening all over the world as well as the location data leakage is also happening. Withthe increment of these data breaches several advertisements and bad cyber crime happens with different person. To solve these issues and to secure the location privacy as well as location data hiding we have applied k- anonymity algorithm to anonymise the data sets so that the location information of users will hide themselves. The aim of our paper is to demonstrate the novel approach for location privacy preservation using machine learning and deep learning algorithms applicable on various parameters as well as on various data sets. We have applied L diversity and T closeness for the same so that we make sure about the protection of the individual stored data dets as well at the same time.

1 Introduction

By combining their features in group of at least individuals, k-anonymity safeguards the security of individual people. The approach is predicated onthe notion that we have an entry-containing dataset. A set of characteristics that include (non-sensitive) data on a person, such as their age, gender, zipcode, etc., make up each entry. These characteristics are referred to as "quasi-identifiers"[2] because, evenin vast datasets, a person may frequently beuniquely identified by a combination of a few of them into a "super-identifier" e.g. the combination of zip code, age and gender might be so exact that only one particular single user in a dataset has a given combination. The model also presume thatthe dataset has a single sensitive characteristic that we wish to safeguard since it contains details like a person's location. The approach may also be applied to datasets with multiple sensitive attributes or datasets in which there is no obvious separation between sensitive characteristics and quasi-identifiers.

Now that our dataset contains individual rows , k-anonymity mandates that we arrange those rowsinto groups of at least k rows and replace the quasi-identifier properties of those rows with aggregate quantities, thereby rendering it impossible to read the individual values

Adversary with knowledge of all values of a person's quasi-identifier characteristics may only determine which group a person may belong to but not if the person is really present in the dataset, persons are protected.

One significant issue with this method is that it is feasible for everyone in a k-anonymous organization to have the same value for the sensitive characteristic. When an opponent is aware of a person's position in the k-anonymous group, they may still determine with 100% confidence what the value of the person's sensitive attribute is.[2] By ensuring that each k-anonymous group has at least l distinct values of the sensitive property, the "l-diversity" extension of k-anonymity may be used to solve this issue. Therefore, even if an attacker is able to determine a person's group, he or she would still not be able to determine with confidence the value of that person's sensitive attribute.

However, an adversary might still use probabilistic reasoning to gather some knowledge about aperson's vulnerable characteristic even while utilising l-diversity: An attacker can infer that aspecific person who is aware of the group's membership would -with high probability- hold a certain value of the vulnerable characteristic if, for instance, 4 out of 5 members of a group of 5 anonymous individuals possess that value. Again, this issue can be solved by utilising an extension of k-anonymity termed the "t-closeness" criterion: According to t-closeness, each k-anonymous group's statistical distribution of vulnerable attribute values must be "near" to the distribution of that attribute throughout the whole dataset. Typically, the Kullback-Leibler (KL) divergence can be used to quantify the proximity.

1.1 K Anonymity (without row suppresion)

With this method, a dataset with a user identity attribute may be easily anonymized withouthaving to group the rows first.

Unlike the function above, this one does not produce a dataframe containing the count variable. The identical dataframe is instead returned, but k-anonymized. All columns that are not categorical will have strings as their return type.

1.2 L Anonymity (without row suppresion)

A dataset having a user identity characteristic may be easily anonymized using this procedure without clustering the rows beforehand.

Unlike the function above, this one does not produce a dataframe containing the count variable. The identical dataframe, anonymsied for I-diversity, is returned instead. All columns that are notcategorical will have a string return type

1.3 T Closeness (without row suppresion)

With this method, a dataset with a user identity attribute may be easily anonymized without having to group the rows first.

As with the previous function, this one does not produce a dataframe containing the count variable. The identical dataframe, anonymised for t-closeness, is returned instead. All columns that are not categorical will have a string as their returntype.

2. Implementing k-anonymity

Finding the ideal segmentation into k-anonymous group is an NP-hard task, and it is difficult to turn a dataset it in to a k-anonymous. Fortunately, there are a number of useful algorithms that frequently use greedy search methods to obtain "good enough" answers.

We will examine the so-called "Mondrian" technique in this lesson, which partitions the primary data into ever-smaller groups using a greedy search algorithm The procedure makes the assumption that all characteristics have been transformed to numerical values and that we can calculate the "span" of a particular data attribute.

3. Partitioning

[3]The data is then divided into k-anonymous sections that use the following algorithm:

1. Set the completed collection of partition to a blank set P(finished)={}
2. Create a partition containing the full dataset in the working set of partitions.
3. Pop a partition from the working set when there are still partitions in it.
    - Determine the relative span of each of the partition's columns.
    - Iterate through the sorted columns after ordering them by span (in decreasing order).
    - Use the midpoint of the columns as the point of separation and attempt to divide the segment along that column.
    - Verify that the generated partitions satisfy our k-anonymity (and perhaps other) requirements.
4. If the answer is true, include the two new segments in the working set as well as end the loop.
5. Include the initial partition in the collection of completed partitions if no column provided a split that was legitimate.
6. Give back the complete set of partitions.

4. System Architecture

The Task Provider Module is made up of several tasks that have been established by various Task Providers in a Network Open Call[1]. When a user requests certain resources, the task provider is responsible for storing such resources and the necessary data in a database. Various data analytic approaches are used to display these data. Usersequipped with different smart devices who are prepared to utilise their resources to perform tasks for task providers make up the user module. The user uses a variety of sensors in their smart devices to perform sensing activities. With the aid of the k-

Means Clustering Algorithm, the users of similarities are divided into k-number of clusters (k-MCA). These clusters have a centroid node, which might be a real node or a virtual one. The huge number of people engaged in the task completion process may be managed with the use of clustering. The MCS network is very dynamic, making managing users a laborious job. K-MCAis in charge of this procedure. Based on the Euclidean distance between each cluster, fake locations are assigned to each user in the cluster. The users are quite active, thus managing the fakelocations is an essential duty. A time synchronization-based fake location allocation mechanism is currently being developed to handlethis. Performance indicators including entropy, cost-benefit analysis, and threshold between users are used to measure privacy. When moving a user from one place to another, performance is assessed. The user establishes a connection to the MCS network using any smart device he possesses that can be utilised to perceive the data required by the MCS network's job provider. Any authentication method is used to verify the user. To get any resources from the job provider, the user must give the task provider his ID. The dummy, which is constructed based on the user location using a fake location generator that is attached, makes the user anonymous in order to safeguard the user's identity. Finding the precise position is challenging since the location is transmitted along with a number of false places.

4.1 K Means

With the aim of assigning each piece of information to a distinct cluster, the K-Means method is a repeating computation that isolates a given dataset into pre-defined specific non- overlapping clusters. At the same time, it makesan effort to maintain the cluster as diverse as is reasonable given the conditions and to keep intra- cluster information concentration as uniform as is reasonable.[2] It ensures that the basis of the whole squared distance between information foci and cluster centroid. The homogeneity within a single cluster is higher the less variability there is within the bunches.[1]The k-means clustering technique for unsupervised learning Using the cue provided by the name, an algebraic formula is employed to determine the data set length, which has a fixed value of k. Users of the same feature's data are gathered into a cluster with a centroid at its centre.The centre point of any cluster, whether actual or virtual, is known as the centroid. In the k-means clustering process, this centroid aids in cluster definition. In other words, the centroid is at the centre of the k-means cluster, which is a specific collection of data points. The close-by data points decide the cluster's size. The k number of centroids are first defined at random by k-means, which may be done repeatedly. By utilising the mean of the average Euclidean distance of all the data points, the centroid is allocated to the data points that are close to it. Benefits of the K-means computation include the ability to define clusters (k), shuffle the dataset, and choose new k data points without replacing any of the existing ones. K-means thenThe expectation-maximization approach to problem solving. Data are appointed to the closest group by the E-step, and each cluster's centroid is recorded bythe M-step.

We initially treat muk fixed and decrease J w.r.t. Then we treat I fixed and minimise J relative to muk. Technically, we update cluster assignmentsand distinguish J w.r.t. I first (E-step). After recalculating the centroids we distinguish J w.r.t. muk (M-step). In other words, group the points withthe cluster that is closest to its centroid. Which equates to recalculating each cluster's centroid to account for the new assignments.

4.2 TSDLA Algorithm

A fake location allocation mechanism is proposed to ensure location privacy in the context of crowdsensing. Since users move around a lot, unscrupulous users may take advantage of the designated fake location to pretend to be another user.[1] The work provided to the device will expire if it is not finished within the allotted time in this article because a correct timing system is built so that when the user changes his or her position, the user will communicate the change. In order for the system to know that the user is leaving, it must receive a surrender message before it can react to any more requests coming from that user location.

This method aids in keeping track of all jobs and the time required to finish each one. It aids in cutting down on communication expenses. Additionally, it shortens the window of opportunity for the hacker to assault the system. It has been demonstrated that the dummy location allocation method offers location privacy under specific service quality and energy restrictions. The given fake location falls inside the range R. To make sure that the location assigned is in the clustered region, the range R is used.

The real location of the user is chosen to carry out the location allocation procedure.[1] The random number K is used to create a random fake location.

Calculated is the Euclidean separation between the real location and the artificially manufactured fake location. The distance and the value R are compared. If the value is less than R, the user is given the location. It generates a new dummy location if it is higher than the value R. The fake location is turned over to the location allocator after the task is finished. The user gives the fake location back to the location allocator if he or she must leave the clustered area before finishing the task. This surrendering process is necessary to make sure the user is a reliable user. Within the clustered region, no other malicious users are using the user credentials.

5. Conclusion and Future Work

In this paper, we studied the data and location privacy preservation technique by the implementation of k anonymity algorithm as well as further using L diversity and T closeness to anonymize any data or to suppress any data attribute so that we can be protected from data as well as location information based attacks. First of all we will visualize the data elements around us and then we apply data segregation and segmentation so that the sets can be in implementable form then applying k-anonymity algorithm to reduce or to suppress the data attributes. This will provide us different clusters of data and then we will map it accordingly by using python. Under the further development this can be implemented on the abstraction of the live data elements as well as we can calculate the entropy and information gain of that data set to determine the nearby location of the main object. Thus, creating the authentication page to specify the sharing of the exact data frame if the auth is disabled then the data sharing will be blocked on the other hand if auth goes successful we canshare the data objects to get the nearby location ofthe object. If any other third party wants to track any particular object location then they will only get the clustered data. Moreover we have kept the use of legendary algorithm to ensure no anomality or loopholes in the project. The further use of this can be in any sector specially in defence intelligence systems where the enemy can't able topredict the exact location of the target object, making the target difficult to detect**.**

6. References

[1] Domi Evangeline S,Dr.G.Usha."TSDLA:Algorithm for Location Privacy in Clustered LB- MCS network".Proceedings of the Third International Conference on Smart Systems and Inventive Technology (ICSSIT 2020)

[2] M. Ghaffari, N. Ghadiri, M. H. Manshaei, and M. S. Lahijani, ‘‘P4QS: A peer-to-peer privacy preserving query service for location-based mobile applications,’’ IEEE Trans. Veh. Technol., vol. 66, no. 10, pp. 9458–9469, Oct. 2017.

[3]M. Gruteser and D. Grunwald, ‘‘Anonymous usage of location-based services through spatial and temporal cloaking,’’ in Proc. 1st Int. Conf. Mobile Syst. Appl. Services, 2003, pp. 31–42.

[4]B. Gedik and L. Liu, ‘‘Protecting location privacy with personalized k-anonymity: Architecture and algorithms,’’ IEEE Trans. Mobile Comput., vol. 7, no. 1, pp. 1–18, Jan. 2008.

[5] S. Gang, S. Liangjun, L. Dan, Y. Hongfang, and C. Victor, ‘‘Towards privacy preservation for ‘check-in’ services in locationbased social networks,’’ Inf. Sci., vol. 481, pp. 616–634, May 2019.

[6]J. Shao, R. Lu, and X. Lin, ‘‘FINE: A fine-grained privacy-preserving location-based service framework for mobile devices,’’ in Proc. IEEE INFOCOM, Apr./May 2014, pp. 244–252.

[7]X. Zhao, H. Gao, L. Li, H. Liu, and G. Xue, ‘‘An efficient privacy preserving location based service system,’’ in Proc. IEEE GLOBECOM, Dec. 2014, pp. 576–581.

[8]B. Niu, Q. Li, X. Zhu, G. Cao, and H. Li, ‘‘Achieving k-anonymity in privacy-aware location-based services,’’ in Proc. IEEE INFOCOM, Apr./May 2014, pp. 754–762

[9] B. Niu, Z. Zhang, X. Li, and H. Li, ‘‘Privacy-area aware dummy generation algorithms for location-based services,’’ in Proc. IEEE ICC, Jun. 2014, pp. 957–962.

[10] Latanya Sweeney. k-anonymity: a model for protecting privacy. Int. J. Uncertain. Fuzziness Knowl.-Based Syst., 10:557–570, October 2002.

[11] Arvind Narayanan, Narendran Thiagarajan, Michael Hamburg, Mugdha Lakhani, and Dan Boneh. Location privacy via private proximity testing. NDSS’10, 2011

[12] Gabriel Ghinita, Keliang Zhao, Dimitris Papadias, and Panos Kalnis. A reciprocal framework for spatial k-anonymity. Inf. Syst., 35:299–314, May 2010.

[13]] K. Vu, R. Zheng, and J. Gao, "Efficient algorithms for kanonymous location privacy in participatory sensing," in INFOCOM, 2012 Proceedings IEEE. IEEE, 2012, pp. 2399–2407.

[14]B. Niu, Q. Li, X. Zhu, G. Cao, and H. Li, "Achieving k-anonymity in privacy-aware location-based services," Proceedings - IEEEINFOCOM, pp. 754–762, 2014

[15]Y. Zhang, W. Tong, and S. Zhong, "On designing satisfaction-ratioaware truthful incentive mechanisms for k-anonymity location privacy," IEEE Transactions on Information Forensics and Security, vol. 11, no. 11, pp. 2528–
2541,2016.