WaveCluster with Differential Privacy

Rajat Raj Singh (CSE 3rd year, IIT Roorkee)

21114079

Abstract:

Clustering is one of the most popular unsupervised learning algorithms that is used to learn information from unlabeled data. There are many different techniques available for clustering but in recent times one of the grid-based clustering algorithms- WaveCluster is emerging as one of the prominent techniques because of its capability of finding clusters of arbitrary shapes. The aim of this report is to study different ways in which differential privacy can be applied to the WaveCluster algorithm such that the utility of the WaveCluster results is preserved. Furthermore, we will also devise a way to quantitatively measure and compare the utility of different algorithms proposed.

Overview and Principles:

To understand the different algorithms proposed to apply differential privacy on WaveCluster, it is important to first understand the preliminary concepts and principles:

1. Differential Privacy:
   In current times, the privacy and confidentiality of individual data is of utmost importance. Hence, in various data analysis scenarios where the data to be analyzed involves personal information of different individuals, it is given that the privacy of each individual must be maintained. Hence, before publishing the data analysis results of such data, we need to make sure that there is no leak of any individual's data. One of the techniques by which it can be achieved is differential privacy. By applying the differential privacy technique, we make sure that the final output result is insensitive to any individual, i.e. there is no way of knowing whether a particular individual was present in the dataset or not by looking at the published result. In this way, the privacy of everyone is maintained. The level of privacy is determined by parameter ϵ, which is also called privacy budget. The smaller the ϵ, the better is the guarantee of privacy.
   There are 2 common ways to ensure ϵ-differential privacy:
   1. Laplace Mechanism: The output result is modified by adding Laplacian noise before publishing.
   2. Exponential Mechanism: All the possible outputs are assigned a score by a quality function, and a randomized algorithm outputs the result with probability proportional to the exponent of the score of the outputs.

   There are also two properties of differential privacy which play an important role in the proposed algorithms:

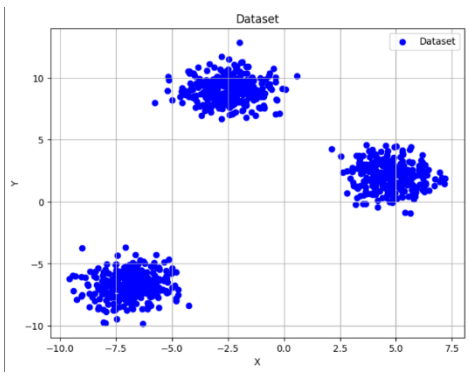
   1. Sequential Composition Property: If n different differentially private algorithms are applied to a dataset each having $\epsilon = \epsilon i$ then we can say that the overall result is ϵ' - differentially private where ϵ' is the sum of each ϵi.

2. Parallel Composition Property: If different ϵ-differentially private algorithms are applied to different disjoint parts of the dataset then the overall result is said to be ϵ-differentially private.

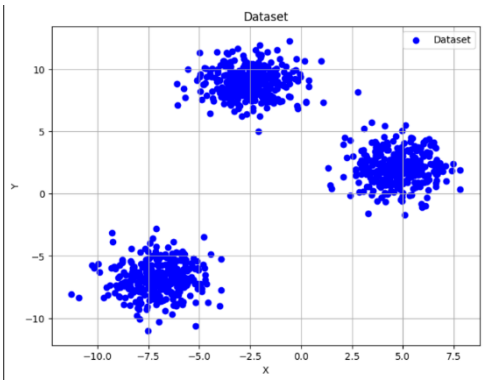
2. WaveCluster:
   WaveCluster is a very specialized clustering algorithm as it works by applying a wavelet transform to differentiate between clusters. By applying a wavelet transform, it distinguishes between high-contrast areas(high-frequency areas)from low-contrast ones. Applying WaveCluster consists of 4 steps:
   1. Quantization: First quantize the feature space into grids and get the count of data points occurring in each grid, thus creating a count matrix M.
   2. Wavelet Transform: Then apply a wavelet transform to it and get the average subband matrix W, which is an approximation of the count matrix M. In this research, we have used Haar Transform as the wavelet transform.
   3. Significant Grid Identification: Now we identify the significant grids from matrix W. We first make a positive sorted list L from the values in W, then take the pth percentile value as the threshold value. Here p is a parameter. Any grid in matrix W with a value greater than the threshold is a significant value and others are non-significant.
   4. Cluster Identification: We then group all the significant grids in the matrix W into clusters by using connected component analysis and then map the clusters back to the original multi-dimensional space, finally labeling the original data points into their relevant clusters.

Procedure:

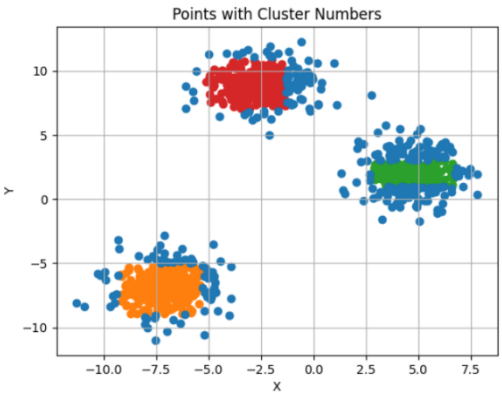
Baseline Approach: The most common way of applying differential privacy to a dataset is to first generate a synthetic dataset from the original dataset by adding Laplacian noise to it, i.e. by adding 2d $Lap(\frac{1}{\epsilon})$ noise to each data point. Now we can apply the normal WaveCluster algorithm to it and obtain an ϵ-differentially private result. We will call this algorithm the Baseline approach. This method is relatively easy to implement but the problem is that the final WaveCluster results would not be of much use as we had already distorted the dataset at the very beginning before even applying the WaveCluster. This change in the dataset can cause very different clusters to form than the desired ones and thus the utility of such algorithm is very low as useful information cannot be inferred form the analysis results. This Baseline approach can be visualised from the figures given below.



Original dataset



Synthetic Dataset



Baseline Result

To address this utility issue, the research paper explores 3 different algorithms, which instead of changing the original dataset, apply differential privacy in the key steps of the WaveCluster algorithm, thus ensuring privacy as well as maintaining the utility of the produced analysis results.

The 3 algorithms are described as follows:

PrivQT (Private Quantization): In this algorithm, differential privacy is applied in the quantization step of the WaveCluster algorithm. In the quantization step, the count matrix M is created after quantizing the feature space of the given dataset. Now this matrix M is perturbed by adding Laplacian noise to it, i.e. each grid of the matrix is added with $Lap(\frac{1}{\epsilon})$, given a privacy budget ϵ. Let's call the resultant matrix W'. Now by the parallel composition property of ϵ-differential privacy, W' is ϵ-differentially private. Now since the subsequent steps of the WaveCluster algorithm are applied on this matrix W', the final output wave clusters are also ϵ-differentially private.

The above discussed algorithm is also pretty simple. But the problem is that adding Laplacian noise to count matrix M can significantly distort the final output clusters. This is because Laplacian noise is zero-mean, hence approximately half the added noise values will be positive and the other half negative. Now in the subsequent steps of the WaveCluster algorithm namely the significant grid identification step, these added positive noise values can cause a grid to become falsely significant, thus wrongly involving in cluster formation. Similarly, the added negative noise values can cause a potential significant grid to have its value less than the threshold and thus being unable to participate in cluster formation. This is why PrivQT algorithm is still not able to give much utility to the output clustering results.

PrivTHR(Private Quantization with refined noisy density threshold): This algorithm is an improvement to PrivQT algorithm. Similar to PrivQT, it first adds Laplacian noise to the count matrix M to obtain noisy count matrix M'. Now we apply the wavelet transform to it to get the average subband matrix W' as usual. In the significant grid identification step, we make the sorted positive list L' from the matrix W' but before we get the pth percentile value of the list L' to be set as threshold, we prune the list L' so as to reduce the distortion of the cluster results. This pruning is done as follows-

Suppose W is matrix obtained by applying wavelet transform to the original count matrix M (without the addition of Laplacian noise), and let Z denote the list of non-positive values in W. Then about half of the values of Z would have become false positives if Laplacian noise would have to been applied to matrix M. Hence we need to prune about |Z|/2 values from L' to get a good value of a threshold that reduces distortion.

Hence, we prune |Z|'/2 starting values in L' and then find the pth percentile of new L' to get a new threshold. Here |Z|' is a noisy estimate of |Z| which is obtained by adding Laplacian noise to it. Now the new threshold can be used to get the wave clusters, which provide much better utility than PrivQT results. In this algorithm, differential privacy is applied in two places, first in obtaining noisy count matrix M' and second in obtaining noisy |Z|' from |Z|. Hence the ϵ needs to be divided into 2 parts $\alpha\epsilon$ and $(1-\alpha)\epsilon$ where α is between 0 and 1. $\alpha\epsilon$ is the privacy budget for noise addition in quantization step and $(1-\alpha)\epsilon$ is the privacy budget for obtaining |Z|'. We can see from the sequential composition property of differential privacy that the overall result is ϵ-differentially private.

PrivTHRem (Private Quantization with Noisy Threshold using Exponential Mechanism): This algorithm is another improvement to the PrivQT algorithm, but unlike PrivTHR, it uses a different method to calculate the threshold by using the exponential mechanism. Similar to the above 2 algorithms, we first get the noisy count matrix M' from the count matrix M in the Quantization step. Now the threshold is calculated as follows:

Suppose W is matrix obtained by applying wavelet transform to the original count matrix M (without the addition of Laplacian noise), and let L be the sorted list of positive values in W, and let k be the number of significant values selected from the threshold received by calculating the pth percentile of the values in L. We now use the values of L to define different intervals (0, W_m], (W_m , W_{m-1}], ….., (W_2, W_1] where W_1 , W_2 , etc are the values in L. The ranks of these partitions are m, m-1, ….., 2, 1 respectively. Now by using the quality function $q(L, X) = -|rank(x) - k|$, we apply the exponential mechanism to determine the partition and then find a uniform random value from that partition which will act as the threshold to determine the significant grids.

Here also the ϵ needs to be divided into 2 parts $\alpha\epsilon$ and $(1-\alpha)\epsilon$ where α is between 0 and 1. $\alpha\epsilon$ is the privacy budget for noise addition in the quantization step and $(1-\alpha)\epsilon$ is the privacy budget for calculating threshold using exponential mechanism . We can see from the sequential composition property of differential privacy that the overall result it ϵ-differentially private.

Also to quantitatively assess the utility of the WaveCluster results derived from these algorithms, we define a Dissimilarity Index. The Dissimilarity Index is a measure that allows us to calculate the dissimilarity level of the true WaveCluster results with the results of a particular differentially private WaveCluster algorithm. The Dissimilarity Index is calculated as follows:

We first define the distance between 2 clusters Ci and Cj as $\max\{|Ci - Cj|\ ,\ |Cj - Ci|\}$. Now for calculating the dissimilarity between true WaveCluster results and differentially private WaveCluster results, we first map each true wave cluster to its differentially private wave cluster and then sum the distances of all these cluster pairs and refer to it as M_{cost}.

Now, the Dissimilarity Index is defined as (M_{cost} / |T|). Here T is the set of significant grids in true WaveCluster results. The lower the dissimilarity index, the better is the utility of the cluster results.

Calculations and Results:

To apply the WaveCluster algorithm, we need to specify 4 parameters –

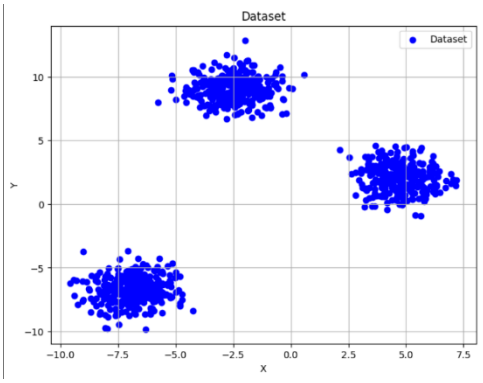
1. The grid size for partitioning the feature space into grids (quantization step). In our implementation, we have taken it to be 1.0.

2. The density threshold parameter p which is used in the calculation of the threshold. In our implementation, we have taken it to be 0.5.
3. The level of decomposition, which is the number of times wavelet transform is applied. We have taken it to be 1.
4. The wavelet transform algorithm to be applied. We have taken Haar transform for this.

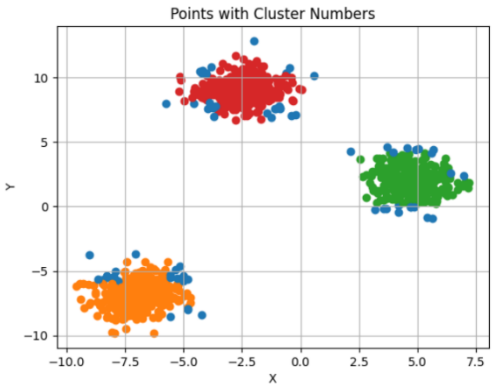
All these parameters are kept the same in all the 3 algorithms discussed above to calculate differentially private wave clusters. The value of ϵ is taken to be 2.0. Furthermore, it was found out that taking $\alpha = 0.9$ works best for PrivTHR algorithm and taking $\alpha = 0.7$ works best for PrivTHRem algorithm.

The dataset on which the algorithms are applied is the standard dataset for clustering downloaded from the site provided by the research paper.

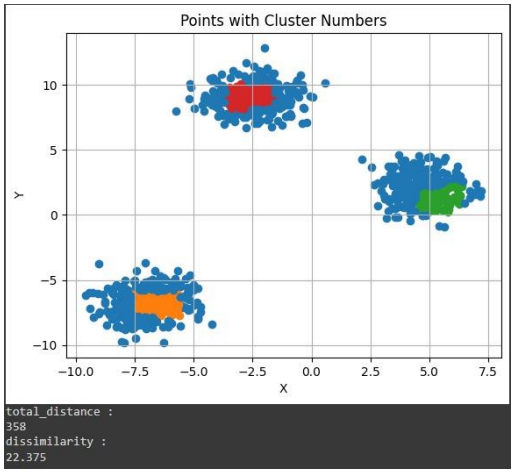
The Results of the algorithms can be seen in the following figure –



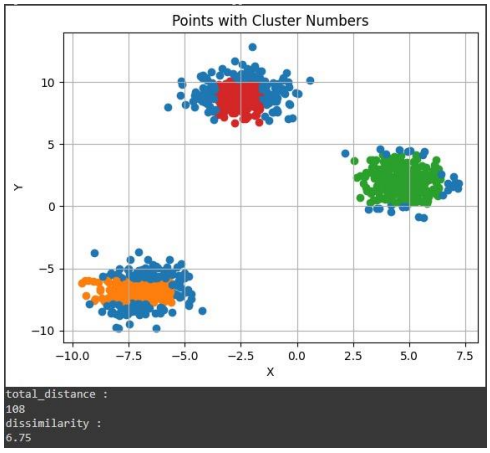
Input Dataset



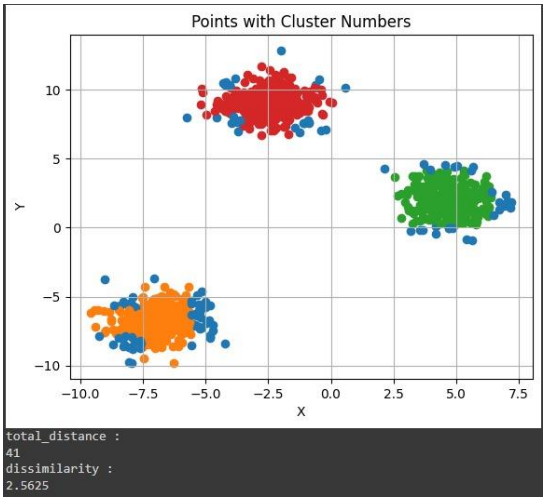
True WaveCluster Result



PrivQT Result and its Dissimilarity Index



PrivTHR Result and its Dissimilarity Index



PrivTHRem Results and its Dissimilarity Index

As we can see from the figure PrivQT provides the least utility as the private WaveCluster results are much different than true WaveCluster results as evident visually and also can be seen from its high dissimilarity. PrivTHR on the other hand gives better results as the produced wave clusters are more closer to the true wave clusters and also the dissimilarity is lower than the dissimilarity of PrivQT. This is because of the reasons already discussed above. PrivTHRem however gives the best results as the private wave clusters produced by it are very much closer to the true wave clusters. Its Dissimilarity Index is also very low. Hence we can see that given a sufficient privacy budget, PrivTHRem gives the best results out of the three algorithms. Another point to note is that even if PrivQT and PrivTHR are not very close to true wave cluster results, these algorithms are still better than the Baseline algorithm as they are not changing the original dataset provided to them, whereas the Baseline algorithm actually makes a synthetic dataset out of the original dataset and then applies WaveCluster algorithm to it.

Conclusion:

In this report, we discussed how to apply differential privacy on one of the popular clustering algorithms - WaveCluster and also described a quantitative measure for measuring the utility of the produced differentially private results. We explored 4 different algorithms to apply differential privacy, first one of which was the very basic Baseline algorithm which does not provide much utility as it involves changing the original dataset. The other 3 algorithms - PrivQT, PrivTHR and PrivTHRem ensure privacy by applying differential privacy at key steps of the WaveCluster algorithm and thus provide better utility than Baseline. Both the PrivTHR and PrivTHRem algorithms apply PrivQT as their first step and then subsequently handle its shortcomings. Out of these algorithms, PrivTHRem gives the best results, followed by PrivTHR and then PrivQT. This is also backed by the Dissimilarity Index readings of these algorithms, which is the quantitative measure that we developed to assess the similarity of private wave cluster results to true wave cluster results.

References:

L. Chen, T. Yu, and R. Chirkova. Wavecluster with differential privacy. CoRR, abs/1508.00192, 2015.