K-Anonymous Privacy Preserving Manifold Learning

Sonakshi Garg[a] and Vicenç Torra[b]
Umeå University, Umeå, Sweden

Keywords: K-Anonymity, MDAV, Manifold Learning, Geodesic Distance.

Abstract: In this modern world of digitalization, abundant amount of data is being generated. This often leads to data of high dimension, making data points far-away from each other. Such data may contain confidential information and must be protected from disclosure. Preserving privacy of this high-dimensional data is still a challenging problem. This paper aims to provide a privacy preserving model to anonymize high-dimensional data maintaining the manifold structure of the data. Manifold Learning hypothesize that real-world data lie on a low-dimensional manifold embedded in a higher-dimensional space. This paper proposes a novel approach that uses geodesic distance in manifold learning methods such as ISOMAP and LLE to preserve the manifold structure on low-dimensional embedding. Later on, anonymization of such sensitive data is achieved by M-MDAV, the manifold version of MDAV using geodesic distance. MDAV is a micro-aggregation privacy model. Finally, to evaluate the efficiency of the proposed approach machine learning classification is performed on the anonymized lower-embedding. To emphasize the importance of geodesic-manifold learning, we compared our approach with a baseline method in which we try to anonymise high-dimensional data directly without reducing it onto a lower-dimensional space. We evaluate the proposed approach over natural and synthetic data such as tabular, image and textual data sets, and then empirically evaluate the performance of the proposed approach using different evaluation metrics viz. accuracy, precision, recall and K-Stress. We show that our proposed approach is providing accuracy up to 99% and thus, provides a novel contribution of analysing the effects of K-anonymity in manifold learning.

1 INTRODUCTION

The amount of data produced every day is exponentially increasing. Machine learning algorithms are evolving day-by-day to provide useful information from this data. With the generation of big data, there also exist enormous high-dimensional data in which the number of instances and attributes are relatively very large, such that data-points become very far from each other. This introduces significant challenges in descriptive and exploratory data analysis. The high-dimensional data in today's world exist in many different forms: ranging from tabular data with higher number of rows and columns, to image data, textual data etc. When the data has two or three dimensions, graphical plots helps in visualizing the local geometry of the data. But corresponding high-dimensional graphs are less intuitive. Thus, to help the visualization structure of such data, dimensions of the data must be minimised. We are cursed by dimensionality of the data. As the dimensionality increases, a larger percentage of the training data resides in the corners of the feature space (Spruyt, 2014). To conquer this problem of curse of dimensionality, dimension reduction can be helpful as it creates a reduced set of linear or nonlinear transformations of the input feature space. It also speeds up the computation power by consuming less memory. The data in lower-embedding space would require less trainable parameters, which leads to less chances of over fitting and thus a more generalised model can be obtained.

Manifold Learning (Tenenbaum et al., 2000) states that any real-world high-dimensional data set lie on a low-dimensional manifold embedded in a higher-dimensional space. Manifold learning methods are being commonly applied in various applications including financial markets (Huang et al., 2017) and medical images (Seo et al., 2019) (Kadoury, 2018) to visualize high-dimensional data. However, the main focus of these techniques is on preserving the inherent structure of the data.

Consequently, when dimensionality of a feature space moves towards infinity, distance measures (e.g.

[a] https://orcid.org/0000-0002-7204-8228
[b] https://orcid.org/0000-0002-0368-8037

Garg, S. and Torra, V.
K-Anonymous Privacy Preserving Manifold Learning.
DOI: 10.5220/0012053400003555
In *Proceedings of the 20th International Conference on Security and Cryptography (SECRYPT 2023)*, pages 37-48
ISBN: 978-989-758-666-8; ISSN: 2184-7711

Euclidean distance, Manhattan distance, Mahalanobis distance etc.) lose their effectiveness to measure similarity in high-dimensional spaces. Euclidean distance only considers numerical distance between two points and calculates its shortest linear path. This distance does not take into account where the actual data lies because it contains no information about the shape of the data. Manhattan and Mahalanobis distance have similar properties. In contrast, Geodesics generalize of the concept of distance for curved surfaces. The geodesic distance considers neighbouring points and finds actual graph-distance between them. It measures the shortest path length passing over the entire data set.

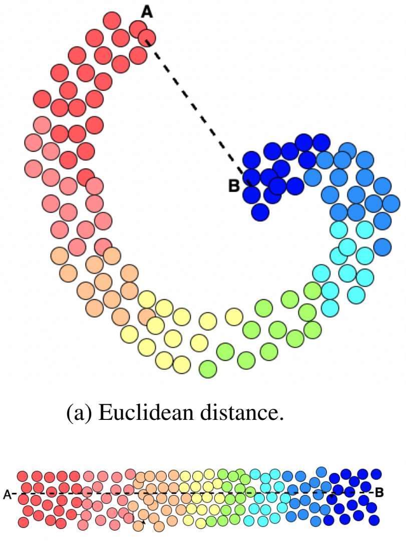
Definition 1.1 (**Geodesic Distance**). Let M denote a d-dimensional manifold. Consider two points m_1, m_2 $\in M$ and a smooth path $\gamma : [0,1] \rightarrow M$ such that γ (0) = m_1 and γ (1) = m_2. The derivative $\gamma'(t)$ depicts the velocity of γ since it passes through the point $\gamma(t)$. The length of the curve $L(\gamma)$ is defined as :

$$L(\gamma) = \int_0^1 \langle \gamma'(t), \gamma'(t) \rangle^{1/2}_{\gamma(t)} dt$$

The distance between points m_1 and m_2, i.e., $\rho(m_1, m_2)$ is infimum over all possible paths connecting the two points m_1 and m_2. If this distance is achieved by a particular path γ, we say that $\rho(m_1, m_2) = \gamma$ is the geodesic distance between two points on a manifold.

In Figure 1(a) Euclidean distance between points A and B is calculated linearly in which the path represented by the Euclidean distance moves away from the data points. But geodesic distance between A and B is computed by joining all adjacent points in the data set, preserving the inherent structure of the data, as depicted in Figure 1(b). This situation may arise more often in high-dimensional data when the data points will be far-away from each other and not all regions of the space are uniformly dense. Therefore, to preserve the local geometry of the data, our approach uses *geodesic distance* instead of Euclidean distance.

In recent years, the availability of personal data has become an important concern with respect to privacy-preserving data mining. We are intended in producing valid data mining results without disclosing the underlying private information. Data anonymisation is one of the privacy models, fundamentally used for Statistical Disclosure Control (SDC) (Cox, 1980) that aims to minimise the identity disclosure. This aids the data controllers to release and process the public data without violating the General Data Protection Regulation (GDPR) policies. A number of techniques have been proposed in order to achieve data anonymisation with respect to multidimensional records (Aggarwal, 2005). However, it is still a challenging task, as obtaining highly accurate results requires looking at original values. When dimensionality of data is high, it becomes even more challenging to preserve privacy of records while maintaining the local geometry of data. (Wang et al., 2020) proposed privacy preserving solutions for high-dimensional data, but it resulted in huge information loss. Therefore, our work aims to bridge this gap and provide privacy preserving framework to anonymize high-dimensional data so that high utility is preserved.



(a) Euclidean distance.

(b) Geodesic Distance.

Figure 1: Euclidean vs Geodesic Distance.

This paper proposes a privacy preserving model to anonymize high-dimensional data while maintaining its manifold structure. Thus, we propose three different approaches viz, M-MDAV, M-ISOMDAV and M-LLEMDAV to transform high-dimensional space to its low-dimensional embedding while preserving the privacy. Algorithm 1 anonymises high-dimensional data directly using M-MDAV. In contrast, Algorithms 2 and 3 use a modified version of manifold learning techniques such as: ISOMAP and LLE, and later on anonymize the obtained lower embedding using M-MDAV micro aggregation method. This is performed to study the effects of K-anonymity privacy model on manifold learning, to provide a comparative analysis and to emphasize the importance of manifold learning. As we will see below, the direct anonymization of high-dimensional data does not lead to good performance of machine learning algorithms. Because of that, it is relevant to study approaches that are appropriate for manifold learning structures. This is a problem that is not considered in the literature, that we tackle here. Thus, we propose a hypothesis that data points should really be in high-dimensions and possess a manifold-like structure in order to analyse the effects of our proposed approaches.

Every type of data requires different preprocessing and a unique approach to handle it in order to generate lower-dimensional embedding. Therefore, to validate the performance of our approach, we have worked on broad areas of applications and used few benchmark data sets involving tabular, textual and medical image data sets with instances ranging from 800 ~ 48000 and number of attributes ranging from 14 ~ 20000. To evaluate the utility of the proposed approaches, we used various machine learning classification models such as: SVM, Naive Bayes, Decision Tree, Random Forest, Gradient Boosting, XGB and, KNN, and we used the best resulting model for testing purposes. Further, we used k-fold cross validation to validate the performance of the models. To estimate the performance of our approach, we compute accuracy, precision, recall and K-Stress values.

To examine the privacy analysis of this paper, we found that K-Anonymity is safe from re-identification and avoids identity disclosure risks. Whereas, differential-privacy requires the selection of a machine learning model before perturbing, to build an appropriate protection method. Thus, this paper provides novel contribution in investigating the effects of K-Anonymity on a high-dimensional data that possess manifold structure.

The main contributions of the paper are as follows.

1. A novel K-Anonymity based privacy-preserving model to anonymize high-dimensional data considering the manifold structure of data.
2. A novel M-MDAV privacy method that is a generalised version of MDAV. It can be applied in any real-world data to protect the data records, and at the same time preserve the data's inherent structure.
3. A study on the importance of geodesic distance in manifold learning approaches such as M-ISOMAP and M-LLEMDAV using natural and synthetic datasets including tabular, image and textual data.
4. An analysis of trade-off between privacy and utility of the samples is provided, in terms of anonymising the records while preserving the neighbourhood of samples during transformations.
5. An analysis on the importance of preserving the manifold structures in anonymization process, as direct anonymization of high-dimensional data leads to poor machine learning performance.
6. We show that the proposed approach provides 99% accuracy in machine learning classification tasks.

The remaining part of the paper is organized in the following manner. Section 2 describes about the definitions of the concepts which are being used in this paper. Section 3 presents step-by-step explanations of the proposed approach. Section 4 discusses about the data sets involved, empirical results and some discussion about them. Finally conclusion and future works are presented in Section 5.

2 PRELIMINARIES

This section provides a brief overview of the necessary concepts that are involved in this paper.

2.1 Manifold Learning

In mathematics, a manifold is a topological space which locally resembles the Euclidean space. Thus, each record in a n-dimensional manifold has a neighbourhood that is homeomorphic to the Euclidean space of dimension n. Manifold learning assumes that sample points lie on a low-dimensional manifold M embedded in a high-dimensional ambient space. The aim of manifold learning is to map sample points from M to a low-dimensional space that preserves its local geometry.

Definition 2.1. Manifold learning considers a finite set of data points $x_1,...x_n \in \mathbb{R}^D$ that exists in a D-dimensional space, and optimize to find low-dimensional points $y_1,...y_n \in \mathbb{R}^d$ when $d \ll D$ such that Euclidean relationship between (y_i, y_j) reflects the intrinsic non-linear relationships between (x_i, x_j).

There are some widely used nonlinear manifold learning approaches including ISOMAP, Locally Linear Embedding (LLE), Laplacian Eigenmaps (LE), t-stochastic Neighbour Embedding(t-SNE), Local Tangent Space Analysis (LTSA), Diffusion Map, and Uniform Manifold Approximation and Projection (UMAP) etc. These techniques help in generating lower- dimensional embeddings of the data while preserving the manifold structure of the data.

Linear Manifold learning techniques assume that the high-dimensional data lies on a linear subspace and as a result linear manifold learning techniques can be successfully applied to linear data. There exists state-of the art approaches for determining the lower-embedding space such as: Principal Component Analysis (Hotelling, 1933), Multi-Dimensional Scaling (Kruskal, 1964), and Linear-Discriminant

Analysis (Fisher, 1936). They help in preserving the linear relationship of the data set. However, when the high-dimensional data lies on a non-linear space, these methods can not capture the inherent structure of the data, thus unable to preserve the pairwise-distance between data points in lower-embedding of the high-dimensional space. Therefore, non-linear manifold learning techniques seek to preserve the non-linear manifolds in high-dimensional space.

2.2 ISOMAP

Isometric Mapping is a non-linear dimensionality reduction approach, that projects the data onto a lower-dimensional space (Tenenbaum et al., 2000). It uses the concept of geodesic distance to find the distance between two points, rather than using the Euclidean distance. The Euclidean distance computes only the distance between two points, completely ignoring the shape of the dataset. In contrast, the geodesic distance generalises the concept of distance for smooth curved surfaces, and calculates the shortest path distance between two points considering the adjacent data points. It can be computed by the construction of an adjacency graph, and then approximate geodesic distance by any shortest path algorithm through the graph. The main steps of the ISOMAP method are described as follows.

- Construct a neighbourhood Graph over high-dimensional space to find the N-nearest-neighbours of each data point. This can be performed in two ways.
  - K-nearest neighbour
  - Selecting neighbours that lie within a fixed radius (epsilon-ball)
- Compute geodesic distance between all data points in a fully-connected neighbourhood graph using any shortest-path algorithm. E.g. Dijkstra's algorithm and Floyd Warshall algorithm. The resulting matrix will also be D(N× N) matrix.
- Construct centering matrix $H = I_N - 1/Ne_Ne_N^T$ where N: size of matrix, I: an identity matrix, $e_N = [1....1]^T \in \mathbb{R}^N$.
- Construct the Kernel matrix $K = -1/2HD^2H$.
- Perform eigenvalue-decomposition on K to obtain the embedded top d-dimensional data points.

The intuition behind working with ISOMAP is, unlike other linear techniques, ISOMAP can compute the non-linear degrees of freedom that underlie complex actual observations. It is also able to obtain a global optimal solution, and is guaranteed to converge asymptotically to the actual structure.

2.3 Locally Linear Embedding (LLE)

LLE is a non-linear dimensionality reduction method which favors the preservation of local data structures because it requires that every data point and its neighbours lie on a linear manifold (Roweis and Saul, 2000). It reconstructs each point as a linear combination of its nearest neighbours, typically using Euclidean distance. Later on, it embeds these points onto a lower-dimensional space while preserving the neighbourhood. LLE falls in a general category of local linear transformation and should be able to perform well for open planar manifolds, with a smooth surface curve. The main steps of the LLE method are explained as follows.

- Find nearest neighbours of data points using Euclidean distance.
- Calculate the reconstruction error by minimising the cost function and obtain W such that:

  $$minW = \sum_{i=1}^{n} |X_i - \sum_{j=1}^{n} W_{ij}X_j|^2$$

  where X=(n×D) i.e., data points in high-dimensions. Every point X_{ij} is a linear combination of its neighbours and weights W_{ij} are computed such that X_i is close to $\sum_{j=1}^{k} W_{ij}X_j$
- Map the data points on low-dimensional space while preserving the weights and obtain Y such that:

  $$minY = \sum_{i=1}^{n} |Y_i - \sum_{j=1}^{n} W_{ij}Y_j|^2$$

  weights W_{ij} between each points gets preserved and low-dimensional embedding Y of dimension (n× d) where $d \ll D$ is obtained.
- Finally, low-dimensional embedding of data set is obtained.

2.4 K-Anonymity

K-Anonymity is a privacy model that limits the risk of re-identification by ensuring the property that each record is indistinguishable from at least another k-1 records, that share identical values for quasi-identifiers (QIDs). These are known as equivalence groups/classes (Samarati and Sweeney, 1998) (Samarati, 2001). It is generally known as power of hiding in the crowd. K-Anonymization in a given way minimises the sum of the squared error (SSE) by solving an objective function having number of parameters more than two. This makes the problem to be NP-Hard. Thus, some heuristic methods are used. K-Anonymity can be implemented using generalization (publishing more general values of the samples), suppression (removal of some samples) and micro aggregation.

Micro aggregation creates some micro-clusters from the entire data set and then replaces the original data set in each cluster by their cluster representatives. In this manner privacy is achieved because now the perturbed data, the cluster representative, is not a single record anymore, instead it is representation of the entire cluster. Each cluster should have a minimum number of records to assure privacy, which is equal to k to satisfy k-anonymity. k is a parameter which determines ”how much” the information is protected, intuitively, the higher the value of k, the more is the protection of information. It decreases the probability of a successful record linkage by generating large equivalence classes.

(De Capitani di Vimercati et al., 2023) illustrates k-anonymity and its main extensions in different applications. In this paper, we have developed a manifold version of Maximum Distance to Average Vector(MDAV) algorithm (Domingo-Ferrer and Mateo-Sanz, 2002) for k-anonymisation based on micro aggregation. It constructs homogeneous clusters from the data set while minimizing the sum of squared errors (SSE) i.e., the distance between each record and its centroid.

$$SSE = \sum_{j=1}^{n}\sum_{i=1}^{n}(x_{ij} - \bar{x_j})^2$$

Differential Privacy (Dwork, 2006) is another mechanism for privacy protection in machine learning. It aims to obfuscate the presence or absence of a particular record in a given dataset, by limiting its effect on the final result. However in real-world applications, data analysis and model construction are just one of the steps of a complex process. One needs to perform exploratory data analysis and test the data on several models before selecting an optimal machine learning model and apply privacy-preserving solutions to it (Torra, 2022), (Domingo-Ferrer et al., 2021).

There are some recent works which includes differential privacy on manifold learning (Vepakomma et al., 2021) and on riemannian manifolds (Reimherr et al., 2021). But according to our knowledge, there are no studies that uses K-anonymity privacy model on the manifolds. Thus this paper provides a novel contribution.

3 METHODOLOGY

This section provides a description of our three different approaches that are developed in this paper to achieve a privacy preserving model that anonymize high-dimensional data considering the manifold structure. Nobody investigated this field of analysing the effects of K-Anonymity privacy model on manifold learning. So, we studied this by proposing three different approaches. Each approach is described in a different algorithm.

Algorithm 1 is the M-MDAV approach, that directly tries to anonymize high-dimensional data using geodesic-MDAV. Algorithm 2 is M-ISOMDAV, which uses ISOMAP for preserving the manifold structure and then uses M-MDAV for anonymization. Algorithm 3 is M-LLEMDAV method. It uses geodesic-LLE and M-MDAV. We have developed three different approaches, since the Algorithm 1 is a manifold version of MDAV that directly anonymises high dimensional data. While the later algorithms use different manifold learning techniques to preserve the inherent structure of data and then anonymize using M-MDAV. A comparative analysis is also conducted between these three algorithms which is described in the later sections of the paper.

The intuition behind developing three different approaches is to analyse the effect of privacy model on manifold learning techniques. To do so, firstly we need a metric that preserves the information of the high-dimensional space. The information should be preserved and not lost while transforming to low-dimensional space, as manifold learning computes distance between points in high-dimensional space and then aims to preserve these distances while transforming to its low-dimensional embedding.

This is achieved by utilising geodesic distance as a metric in manifold learning approaches. Once, the information is transformed in a low-dimensional space, M-MDAV a newly developed manifold version of K-anonymity model is used to protect the information from intruders. We have considered two different manifold learning techniques for the good properties they have, and provided a comparative analysis between them.

Algorithm-1 M-MDAV is a manifold version of state of the art MDAV. Initially, pairwise-geodesic distance between each data points are computed as defined in 1. Then, median of all data points is obtained by minimising the geodesic distance between the data points, as mentioned in the objective function of the algorithm 3. After that, clusters are formed around the data points that are furthest from the median. This process is repeated until all points get clustered. Finally, the clustered data points are replaced by the median of that cluster. The Algotihm-2 M-ISOMDAV is a manifold combination of two different approaches i.e., ISOMAP manifold learning and

Algorithm 1: M-MDAV.

```
Require: Y: original data set and k: integer
Ensure: $Y'$: protected data set
while $|Y| \neq 0$ do
    if $|Y| \geq 3k$ then
        Identify median of all the records denoted by $y_{median}$ such that :
            $$y_{median} = \arg\min_{y \in \mathbb{Y}} \sum_{i=1}^{n} d(y, y_i)^2$$
        where d is the geodesic distance.
        Find the furthest record from centroid $y_{median}$ called as $y_r$ , and furthest record from $y_r$, called as $y_s$. This is also computed using geodesic distance.
        Create Cluster $C_r$ around $y_r$ which consists of $y_r$ and k-1 records closest to it. Cluster $C_s$ which includes $y_s$ and k-1 closets records. k is micro-aggregation parameters which denotes the number of times each combination of values appears in a dataset.
        The dataset gets updated $Y = Y - (C_r, C_s)$
        The clusters get updated as well: $C = C \cup (C_r, C_s)$
    else if $|Y| \geq 2k$ then
        Find $y_{median}$ with all the records in Y.
        Find most distant record $y_r$ from $y_{median}$.
        Create Cluster $C_r$ with $y_r$ and k-1 closest records. Cluster $C_s$ with remaining records.
        The clusters get updated: $C = C \cup (C_r, C_s)$
    else
        $C = C \cup (Y)$
    end if
end while
Produce k-anonymized matrix $Y'$ from clusters C.
```

MDAV micro-aggregation model.

Similarly, the Algorithm-3 M-LLEMDAV is a manifold version of combination of two different algorithms i.e., LLE manifold learning method and MDAV privacy model. The algorithm starts by the construction of neighbourhood graph as it is performed in LLE method. After that, geodesic distance between each point and it's neighbours are computed instead of measuring euclidean distance. Since geodesic distance is a generalisation of distance for curved surfaces, and it is more suitable in high-dimensional space. Later on, the data points are transformed to low-dimensional space while preserving the weights and minimising the objective function as depicted in algorithm 3 1. Finally, the low-dimensional data points are protected using M-MDAV a manifold version of MDAV privacy model.

Algorithm 2: M-ISOMDAV.

```
Require: $D(n, p); p \geq n$ or $p \sim n$
Ensure: $D'(n, d)$ ; $d < n$
Construct the weighted neighbourhood graph M by connecting points $N_i$ and $N_j$ such that if they are closer to ε,then edge length becomes $d_E(N_i, N_j)$.
Compute pairwise-geodesic distance matrix $M'$ : $N * N \in \mathbb{R}$ with all the data-points of matrix M using Dijkstra shortest path algorithm.
Construct a centering matrix H where $H = I_n - 1/Ne_N e_N^T$ and $e_N = [1....1]^T \in \mathbb{R}$.
Compute kernel matrix $K = -1/2HM'^2H$.
Determine eigenvalue decomposition of matrix K of size D into d using any built-in function and decompose it.
Record top-d eigenvalues of K in λ and their corresponding eigenvectors in ν.
Obtain $Y = \sqrt{\lambda}\nu$ the lower d-dimensional vector of dimension (n×d).
Apply M-MDAV approach to Y to perform k-anonymisation as discussed in Algorithm 1.
```

4 EXPERIMENTATION AND RESULTS

In this section we initially present the data sets that are considered for evaluation of the proposed approach. Later on, we describe the computational requirements that are necessary to conduct this experimentation. Finally, we discuss the obtained results and our analysis using the proposed approach.

4.1 Data Set Description

In this sub-section, we describe the different data sets that are involved for this experimentation. A wide-variety of high-dimensional data sets are available with us. Thus, we intended to consider real as well as synthetic data sets. The three-different types of real data set are tabular, image and textual data sets. The number of instances ranges from 800 to 48000, and the number of attributes ranges from 14 to 20000, so that a broad experimentation can be performed to analyze that the proposed approach is suitable on various types of data. The description of data sets used are depicted as follows.

RNA Data. It is a classification data set, that consists of random extraction of gene expression of patients having five-different types of cancerous tumor: KIRC, PRAD, BRCA, LUAD and COAD (Fiorini, 2013). The dimensions of this data set is

```
Algorithm 3: M-LLEMDAV.
Require: $D(n,p); p \geq n$ or $p \sim n$
Ensure: $D'(n,d)$ where $d < n$
1: Construct the weighted neighbourhood graph M by connecting points $N_i$ and $N_j$ such that if they are closer to $\varepsilon$,then edge length becomes $d_E(N_i,N_j)$.
2: Calculate geodesic distance between points $N_i$ and it's neighbors that are selected in above step using Dijkstra shortest path algorithm.
3: Construct each point from its neighbours. Reconstruction errors are calculated by minimising the cost function
       $\varepsilon(W) = \sum_i |N_i - \sum_j W_{ij} N_j|^2$
   subject to constraint $\sum_{j=1}^{n} W_{ij} = 1$. Thus, weights $W_{ij}$ are obtained that reconstructs each data point from its neighbours.
4: Compute the low-dimensional data Y that best preserves the manifold structure, represented by weights $W_{ij}$.
       $\phi(Y) = \sum_i |Y_i - \sum_j W_{ij} Y_j|^2$
   subject to constraint $\sum_{i=1}^{n} Y_i = 0$. Thus, lower-dimensional matrix Y(n*d) is resulted.
5: Apply M-MDAV approach to Y to perform k-anonymisation as discussed in Algorithm 1.
```

(801*20531). The number of attributes are significantly more than the number of instances. High-dimensional visualization of this data is difficult, but the proposed approach makes this easier.

GISETTE Data. It is a handwritten digit recognition problem(Guyon, 2003). The task is to differentiate between highly confusible digits '4' and '9'. This data set is one of five data sets of the NIPS 2003 feature selection challenge. It is also a classification data set having dimensions of (6000*5000).

SPAM Data. It is a textual data set that classifies emails as Spam or Non-Spam (Hopkins, 2002). It consists of 4457 instances which are pre-processed using TF-IDF method that quantifies the relevance of a text using statistical measures. Therefore, when TF-IDF approach is applied on SPAM data set the resultant data has (4457*5055) dimensions. This data set is widely used in natural language processing (NLP) task.

ADULT Data. It is a census income dataset, which consists of numerical and categorical values and predicts whether income of a person exceeds 50K/ yr . It is a classification data set which consists of 48000 instances and 14 attributes.

MADELON. It is an artifically created dataset that consists of two-class classification problem with continuous input variables. It was a part of NIPS 2003 feature challenge having dimension of (4400*500).

4.2 System Description

The experimentation is performed on mac OS with 8-core M1 Pro chip, 16 GB RAM, 500 GB Memory, Python version 3.10 with a steady internet connection was used.

4.3 Results and Analysis

This sub-section describes the visualisation and detailed explanations of the empirical results which were obtained using the proposed approach. The experiments were conducted using three different approaches. They mainly correspond to the application of the three Algorithms described in the previous section. The proposed three approaches provides a way for micro-aggregation to avoid identity disclosure risk using K-anonymity privacy model,since it is safe from re-identification.

The first approach consists of applying M-MDAV directly on the high-dimensional data set and obtain k-anonymous data records in the higher embedding itself. Afterwards, to empirically evaluate the performance, state-of-the art ML classification algorithms such as SVM, Naive Bayes, Gradient Boosting, Decision Tree, Random Forest, Extreme Gradient Boosting and K-nearest neighbours are implemented. The best resulting model is further used for testing purposes. To obtain a more generalised model with less bias, k-fold cross validation technique is also utilised. Finally, to validate the utility of the approach evaluation metrics such as accuracy, precision and recall are recorded.

Following Algorithm 2 M-ISOMDAV, we begin with ISOMAP manifold learning technique to preserve the manifold structure of the data and obtain lower-embedding of the data set. Later on, anonymisation on the low-embedding are performed using M-MDAV algorithm. Finally, the anonymity data sets are classified using all the above mentioned ML models and validated using k-fold cross validation. To examine the utility of perturbed lower-dimensional embedding, we describe another metric known as K-Stress.

Following Algorithm 3 M-LLEMDAV, we use geodesic version of LLE manifold learning technique that tries to preserve the local neighbourhood structure. Then anonymization using M-MDAV is applied and classified using different ML models. A comparative analysis is performed with all three different approaches and the best resulting approach for each data sets are highlighted. Note that, the K-Stress metric cannot be used with Algorithm 1 because K-Stress preserves pairwise distances between high-dimensional and their low-dimensional embedding, and any kind of transformation from high-dimensional space to low-dimensional space is not performed in Algorithm 1.

The performance of the proposed approach is evaluated using four evaluation measures which are described as follows.

Accuracy: is a metric for evaluating classification models, that measures the ratio of number of correct predictions with respect to total number of predictions. Numerical representation of accuracy is depicted as follows.

$$Accuracy = Correct predictions/Total predictions.$$

Precision is a measure of quality that calculates the fraction of correct positive results out of total positive outcomes obtained by the model. Mathematically, it is presented as follows.

$$Precision = TP/TP + FP.$$

where TP is True positives and FP is False Positives.

Recall: is a measure of quantity that computes the fraction of correct positive results out of all relevant samples that should have been classified as positive by the model. Its algebraic representation is

$$Recall = TP/TP + FN$$

where TP is True positives and FN is False Negatives.

K-Stress: is a weighted sum of differences between distance in original space, and the corresponding lower-dimensional space (Kargupta et al., 2005). It is a measure of goodness of fit that requires that distance between two points in perturbed lower-dimensional embedding are well preserved with respect to distance between those points in original higher-dimensional space. The stress indicates the amount of information loss before and after transformation, and expressed as a percentage with 0% stress being equivalent to perfect transformation. Mathematically, it is calculated as follows.

$$\sqrt{\sum(d_{ij} - \delta_{ij}{}^2)/\sum d_{ij}^2}$$

where d_{ij} is pairwise distance between points in higher-dimensional embedding, whereas δ_{ij} is the pairwise distance between points in lower-dimensional space.

Table 1: Evaluation of the proposed approach.

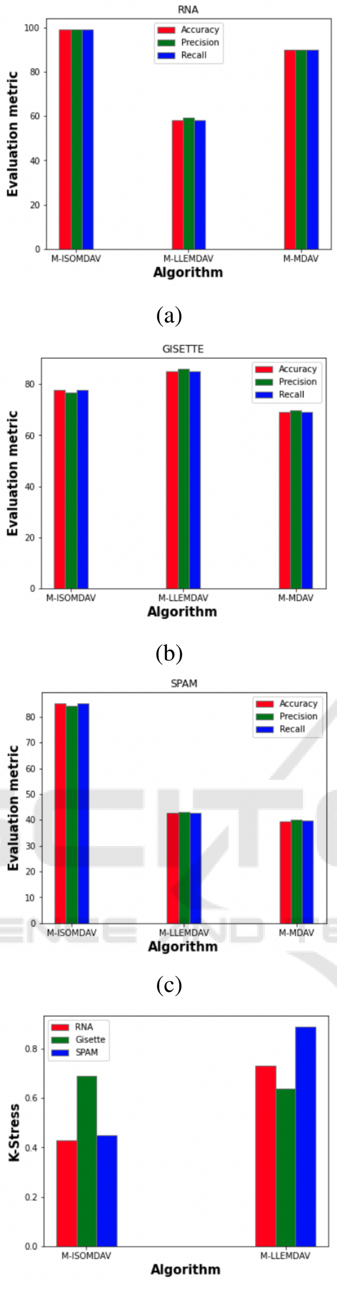
Dataset (D)	D(n,p)	Algorithm	Accuracy	Precision	Recall	K-Stress
RNA	800 × 20531	M-ISOMDAV	**99.17**	**99.18**	**99.17**	**0.43**
		M-LLEMDAV	58.12	59.3	58.13	0.73
		M-MDAV	90.10	90.12	90.11	–
Gisette	6000 × 5000	M-ISOMDAV	77.79	76.82	77.78	0.69
		M-LLEMDAV	**85.13**	**86.10**	**85.14**	**0.64**
		M-MDAV	69.21	69.87	69.18	–
SPAM	5272 × 5055	M-ISOMDAV	**85.20**	**84.34**	**85.21**	**0.45**
		M-LLEMDAV	42.61	43.13	42.59	0.89
		M-MDAV	39.56	40.10	39.81	–

4.4 Discussion

A detailed analysis has been performed for the selection of hyper-parameters. The hyper parameter *k* for k-anonymity is chosen after performing several iterations over different values of *k*. When we used k in the range of (5-10), similar outcomes were resulted in terms of accuracy. When k value was increased to (15-20), the performance of our approach started decreasing. Because a value of *k* larger than 5 is often considered as acceptable for k-anonymity and micro aggregation. So, we chose k=10 as a generalised value for our experiments.

We implemented our approach using seven different machine learning classification models as described above. Upon analysis we found that, the resulting best classification model for RNA data set is K-nearest neighbour classifier. This model is then further used for testing purposes and evaluating the performance using accuracy, precision and recall. The hyper parameters used for tuning the K-nearest neighbour classifier are: number of neighbours to be 5, and weight distribution to be uniform. In contrast, for Gisette and SPAM data set Gradient Boosting classifier turns out to be the best performing model and further used for evaluation purposes. The hyper parameters used for Gradient Boosting classifier are: number of estimators to be 100, learning rate to be 0.1 and maximum depth of the tree to be 5. Rest of the hyper parameters are kept by default as provided by scikit library in python.

Table. 1 presents the tabular representation of results which are obtained on data sets using the three different proposed approach which are involved in this paper, it provides a comparative analysis between our approaches. The first column presents the name of data set used, the second column describes the size of dataset in terms of number of instances and number of attributes, the third column depicts the name of the algorithm, whereas the remaining columns describe the evaluation metrics. They are accuracy, precision, recall, and K-Stress. For each data set, the best performing approach is marked in bold.



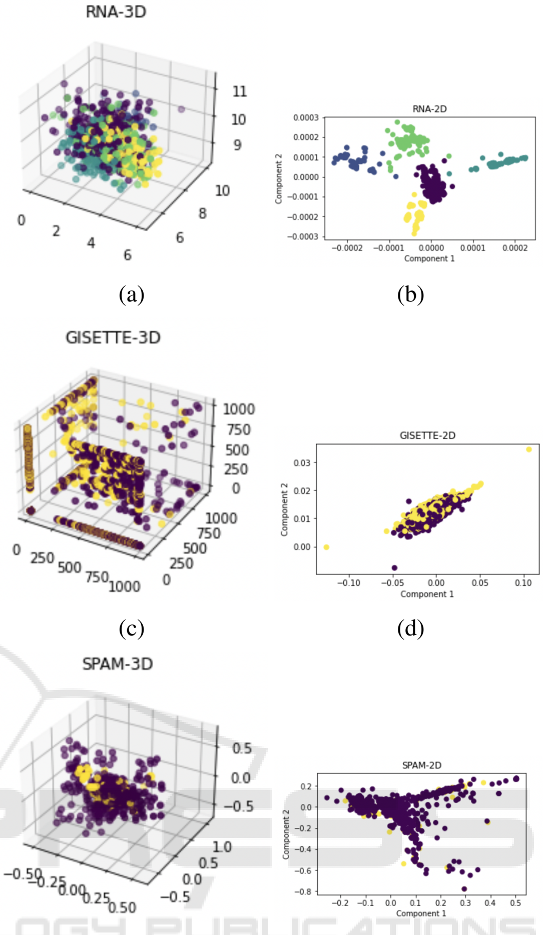
(a)

(b)

(c)

(d)

Figure 2: Performance analysis on (a) RNA (b) Gisette and (c) SPAM Data sets (d) K-Stress values comparison using M-ISOMDAV and M-LLEMDAV method.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 3: Visualization of Data sets in high and low-dimensions.

Upon analysis it is found that for RNA and SPAM data set, M-ISOMDAV approach is providing best accuracy of 99.17% and 85.20 %. K-Stress value is 0.43 which is better than 0.73 that is obtained using M-LLEMDAV for RNA data set. Contrary, for GISETTE data set M-LLEMDAV approach is providing the highest accuracy of 85.13 % and K-Stress of 0.64. However, it can be observed that M-MDAV didn't provide the best results on any data set. Performance of M-MDAV is significantly less than the other approaches. Therefore, it can be analysed that M-MDAV alone is not able to anonymise and preserve the manifold structure of high-dimensional data and it emphasize the importance of manifold learning. It becomes relevant to use approaches that preserves the inherent structure of the data for any machine learning algorithms.

M-LLEMDAV's performance on RNA and SPAM data sets was relatively poor. We think the reason behind this is that these data sets consist of multiple manifolds, and LLE manifold learning algorithm use a variety of tangent linear patches to model a manifold. It represents one function as several small linear functions, thus it is designed to work on slightly simpler datasets (like the Gisette data set).

The Figure 2 depicts the visual representation of three different proposed approaches on the selected data sets and provides a comparative analysis among them. The Figure 2(a) records performance of three proposed approaches on RNA data set in terms of accuracy, precision and recall. It can be clearly seen that, bar plots of M-ISOMDAV are highest in length since they are resulting in most precise outcomes. Similar trends are observed in Figure 2(c), where M-ISOMDAV approach are providing most optimal outcomes in SPAM data set. The Figure 2(d) depicts the K-Stress values on three selected data sets using M-ISOMDAV and M-LLEMDAV approach. It is found that for two data sets i.e., RNA and SPAM the M-ISOMDAV are providing minimal K-Stress values as compared to the other approach.

The visual analysis of original high-dimensional data sets are not possible. So, the impact of the transformation from high-dimensional ambient space to lower-dimensional embedding are represented in Figure 3. We present both 3D graphs and 2D graphs. In 3D graph plots it can be seen that all the data points are immensely overlapped. These representations can be seen for RNA, Gisette and SPAM data set in Figure 3 (a), (c) and (e). The data points in 2D are much easier to be classified and visualisation also becomes better. This is clearly depicted in Figure 3(b) and (d). This analysis provides the visual importance of the proposed approach.

To estimate the accuracy of ML classification models, we use 10 fold cross-validation method. We evaluated the impact of k for k-fold cross validation on the classifier's performance. The value of k is varied from 3 to 10, and misclassification errors are recorded. For all data sets, as the value of k increases, the misclassification error also increased. The minimum error values were obtained for k value 5 majorly. Thus, k is set to be 5 for validating purposes and we used 5-fold cross validation. This effect of varying k with respect to misclassification error is displayed in Figure 4.

We also computed the complexity of the proposed approaches. We found it to be $O(m \times n^3)$ where m is the dimensions of data points and n is the number of data points.

We performed our experiments on a wide variety of datasets including Adult, Madelon and few other image data sets such as MNIST, CIFAR-10 etc. We implemented the above three proposed approaches and recorded accuracy, precision, recall and K-Stress values. The results are presented in Table 2. We propose the following hypothesis based on the analysis of our results.

Hypothesis 1. The data-points should really be in high-dimensions and must possess manifold structure, then only the proposed approaches will be able learn the intrinsic structure of the manifold and anonymize data-points efficiently.

Table 2: Limitation of the proposed approach.

Dataset (D)	D(n,p)	Algorithm	Accuracy	Precision	Recall	K-Stress
Adult	48842*14	M-ISOMDAV	50.12	50.13	50.11	0.35
		M-LLEMDAV	43.32	43.30	42.29	0.32
		M-MDAV	41.19	42.90	40.12	-
Madelon	4400*500	M-ISOMDAV	62.18	62.25	62.19	0.28
		M-LLEMDAV	59.23	59.21	59.23	0.25
		M-MDAV	60.38	60.30	61.21	-

We consistently test the performance of our approach on Adult, Madelon, MNIST etc, datasets by using ablation studies on different hyper-parameters, We found that in the case of Adult and Madelon data set, the data points are not really in high-dimensions, as it should be for the manifold learning techniques. Also, the data-distribution for these datasets is not similar to the manifold structure. Thus, poor performance in terms of accuracy and neighbourhood preservation (K-Stress) is obtained.

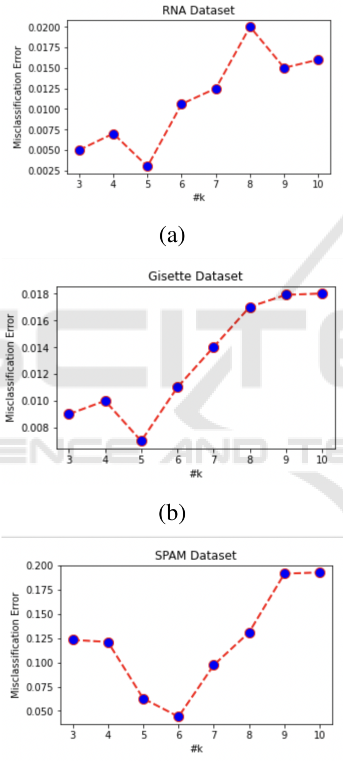
5 CONCLUSION AND FUTURE WORKS

In this paper, we proposed a privacy preserving framework that uses K-Anonymity to anonymize high-dimensional data maintaining its manifold structure. In particular, we proposed three different approaches out of which two use manifold learning techniques to preserve the inherent structure of data during anonymising, while the third one involves only a manifold version of MDAV micro aggregation method to achieve privacy. Later on, machine learning classification models were used to evaluate the performance of the proposed approach.

We evaluated the results in terms of statistical measures such as machine learning classification accuracy and good neighbourhood preservation such as K-Stress values. The results show that the non-linear transformations of data into lower-embedding can preserve the privacy of the data. This paper provides a trade-off between utility and privacy of the records.

We have also shown that anonymising high-dimensional data directly i.e., using M-MDAV alone is not able to preserve the underlying structure of the data and leads to poor machine learning performance. Thus, the proposed approaches are relevant for preservation of manifold structure of the data. We investigated with different types of real and synthetic data sets. We proposed a hypothesis about the need of data-points to be in high-dimensions and possess the manifold structure to be efficiently anonymized using the proposed approach.

In future, we will provide a proof for this hypthesis. Also, we will analyse our approach using geodesic version of different manifold learning approaches such as: t-SNE, LTSA etc. Additional experiments are considered for future work. Our K-anonymity privacy model avoids the risk of identity disclosure. However, it is unable to safeguard against attibute disclosure risk. Thus, in future we would like to formulate k-anonymity privacy model taking into account the attribute disclosure risk in the manifold structure. Also, we would like to analyse the behaviour of differential privacy on our proposed approach.



(a)

(b)

(c)

Figure 4: No. of k vs misclassification error for (a) RNA (b) Gisette and (c) SPAM Data sets.

ACKNOWLEDGEMENTS

This work was partially supported by the Wallenberg Al, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

REFERENCES

Aggarwal, C. C. (2005). On k-anonymity and the curse of dimensionality. In *VLDB*, volume 5, pages 901–909.

Cox, L. H. (1980). Suppression methodology and statistical disclosure control. *Journal of the American Statistical Association*, 75(370):377–385.

De Capitani di Vimercati, S., Foresti, S., Livraga, G., Samarati, P., et al. (2023). k-anonymity: From theory to applications. *TRANSACTIONS ON DATA PRIVACY*, 16(1):25–49.

Domingo-Ferrer, J. and Mateo-Sanz, J. M. (2002). Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and data Engineering*, 14(1):189–201.

Domingo-Ferrer, J., Sánchez, D., and Blanco-Justicia, A. (2021). The limits of differential privacy (and its misuse in data release and machine learning). *Communications of the ACM*, 64(7):33–35.

Dwork, C. (2006). Differential privacy. In *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 33*, pages 1–12. Springer.

Fiorini, S. (2013). https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+rna-seq. *UCI Machine learning repository*.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188.

Guyon, I. (2003). https://archive.ics.uci.edu/ml/datasets/gisette. *UCI Machine learning repository*.

Hopkins, M. (2002). https://archive.ics.uci.edu/ml/datasets/spambase. *UCI Machine learning repository*.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417.

Huang, Y., Kou, G., and Peng, Y. (2017). Nonlinear manifold learning for early warnings in financial markets. *European Journal of Operational Research*, 258(2):692–702.

Kadoury, S. (2018). Manifold learning in medical imaging. In *Manifolds II-Theory and Applications*. IntechOpen.

Kargupta, H., Datta, S., Wang, Q., and Sivakumar, K. (2005). Random-data perturbation techniques and privacy-preserving data mining. *Knowledge and Information Systems*, 7(4):387–414.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27.

Reimherr, M., Bharath, K., and Soto, C. (2021). Differential privacy over riemannian manifolds. *Advances in Neural Information Processing Systems*, 34:12292–12303.

Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326.

Samarati, P. (2001). Protecting respondents identities in microdata release. *IEEE transactions on Knowledge and Data Engineering*, 13(6):1010–1027.

Samarati, P. and Sweeney, L. (1998). Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression.

Seo, K., Pan, R., Lee, D., Thiyyagura, P., Chen, K., Initiative, A. D. N., et al. (2019). Visualizing alzheimer's disease progression in low dimensional manifolds. *Heliyon*, 5(8):e02216.

Spruyt, V. (2014). The curse of dimensionality in classification. *Computer vision for dummies*, 21(3):35–40.

Tenenbaum, J. B., Silva, V. d., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323.

Torra, V. (2022). *Guide to Data Privacy: Models, Technologies, Solutions*. Springer Nature.

Vepakomma, P., Balla, J., and Raskar, R. (2021). Privatemail: Supervised manifold learning of deep features with differential privacy for image retrieval. *arXiv preprint arXiv:2102.10802*.

Wang, R., Zhu, Y., Chang, C.-C., and Peng, Q. (2020). Privacy-preserving high-dimensional data publishing for classification. *Computers & Security*, 93:101785.