Utility Promises of *Self-Organising Maps* in Privacy Preserving Data Mining

Kabiru Mohammed[1 [0000−0001−8728−2896]], Aladdin Ayesh[1 [0000−0002−5883−6113]], and Eerke Boiten[1 [0000−0002−9184−8968]]

De Montfort University, Cyber Technology Institute, The Gateway, Leicester, LE1 9BH, UK

Abstract. Data mining techniques are highly efficient in sifting through big data to extract hidden knowledge and assist evidence-based decisions. However, it poses severe threats to individuals' privacy because it can be exploited to allow inferences to be made on sensitive data. Researchers have proposed several privacy-preserving data mining techniques to address this challenge. One unique method is by extending anonymisation privacy models in data mining processes to enhance privacy and utility. Several published works in this area have utilised clustering techniques to enforce anonymisation models on private data, which work by grouping the data into clusters using a quality measure and then generalise the data in each group separately to achieve an anonymisation threshold. Although they are highly efficient and practical, however guaranteeing adequate balance between data utility and privacy protection remains a challenge. In addition to this, existing approaches do not work well with high-dimensional data, since it is difficult to develop good groupings without incurring excessive information loss. Our work aims to overcome these challenges by proposing a hybrid approach, combining self organising maps with conventional privacy based clustering algorithms. The main contribution of this paper is to show that, dimensionality reduction techniques can improve the anonymisation process by incurring less information loss, thus enhancing a more desirable balance between privacy and utility properties.

Keywords: k-anonymity · Clustering · Self Organising Map · Privacy Preserving Data Mining.

1 Introduction

Data mining techniques allow the extraction of implicit and useful information from big data. They are programmed to sift through data automatically, seeking patterns that will likely generalise to make evidence-based decisions or accurate predictions that hold in data collections [1]. Although this emerging technology enjoys intense commercial attention, there is a growing concern that data mining results could potentially be exploited to infer sensitive information, therefore potentially breaching individual privacy in a variety of ways [2]. In response to these privacy concerns, Privacy Preserving Data Mining (PPDM) has been proposed

by a number of studies [3–7] as an effective method for accommodating privacy concerns during mining processes to address the risk of re-identification. PPDM aims to provide a trade-off between data utility on one side and data privacy on the other side, by enforcing a certain degree of privacy without relinquishing the purposefulness of the data. It has remained a successful approach, particularly when applied to satisfy anonymisation protection models such as *k-anonymity, l-diversity and t-closeness*. Several published works [8–12] in this area have proposed different clustering techniques to enforce anonymisation models on private data. These techniques work by grouping the data into clusters using a quality measure and then generalising the data in each group separately to achieve an anonymisation threshold [13]. These studies have illustrated that clustering-based methods are able to produce high-quality anonymisation while allowing data mining to take place with less concern about privacy violations. Despite this breakthrough, conventional approaches cannot achieve a good balance between data utility and privacy protection [9, 14, 15]. They mostly optimise privacy, and as a result cannot guarantee a minimum level of data utility [16]. Our aim is to enhance data utility in PPDM processes, which can further guarantee a greater degree of balance between the two properties.

This paper proposes a hybrid strategy for improving data utility in PPDM techniques. In this approach, conventional clustering algorithms such as *OKA* and *K-member* are applied in combination with a Self-Organising Map (SOM). To illustrate this point, we apply proposed method to the Adult data set [17]. In the first step, the aforementioned clustering algorithms are used to anonymise specific features of the dataset using a selected anonymity threshold. Secondly, a SOM is used to map other data features to a 1-dimensional space of a set of neurons. These data features are otherwise dropped by traditional clustering strategies because they are mostly classified as sensitive attributes, thus they may increase the chances of attribute or membership disclosures. This unsupervised neural network model preserves the topological relationship and increases the correlation between features of the primary data space. Thirdly, the results of the anonymisation methods are fused with the results of the 1-dimensional SOM strategy as a single data domain. Lastly, the newly produced dataset is subjected to several classification techniques that are typically employed on the original Adult data.

The main contribution of this work is to show that the proposed strategy is a more productive approach in scenarios where the need for higher data utility supersedes the need for higher data privacy, particularly if there are no privacy costs associated with the desire for more data utility. Therefore, the results obtained with the application of this strategy justify its use. The remainder of the paper is organised as follows: section 2 presents a brief bibliographical review about privacy preserving data mining algorithms, and section 3 describes the main aspects of the SOM algorithm, detailing its advantages to other unsupervised neural networks. Section 4 describes the Adult data set and its properties, while section 5 presents the strategy for this experiment. In section 6, the methodology for conducting the experiments is expressed, while section 7

presents the outcomes of the proposed strategy, comparing it with results obtained from conventional approaches. Finally, section 8 presents conclusions and the direction of future PPDM research.

2 A review of PPDM

k-anonymity was first introduced by Samarati and Sweeney [18] in an attempt to prevent possible re-identification of user information from published microdata. This concept requires that each combination of quasi-identifier values in a released table must be indistinctly matched to at least k respondents [19]. For example, a table $D(S_1, S_2, ..., S_m)$ is said to satisfy k-anonymity if each quasi-identifier QI associated with the values maps to at least k records in a transformed version of the table D^k.[1] More formally, k is the largest number such that the magnitude of each equivalence class in table D is at least k.

The k-anonymity requirement is typically enforced through generalisation, suppression and deletion techniques. Generalisation replaces real values with "less specific but semantically consistent values" [20]. Numerical values are typically specified by a range of values, while categorical values are combined into a set of distinct values based on a hierarchical tree of the data attribute domain. Suppression replaces attribute values with a special symbol, and deletion removes an entire attribute from a dataset. Algorithms based on these techniques are conceptually straightforward; however, there are limitations: the computational complexity of finding an optimal solution for the k-anonymity problem has been shown to be NP-hard [21], possible generalisations are limited by the imposed hierarchical tree [22], also suppression and deletion techniques often compromise data utility by producing results that are unsuitable for further analysis [14]. To overcome these challenges, several PPDM approaches have viewed anonymisation as a clustering problem [6, 8, 10–12]. Clustering-based anonymisation works by partitioning datasets into clusters using a quality measure and generalising the data for each cluster to ensure that they contain at least k records [23]. This method produces high data quality because it reduces data distortion, making the results suitable for further analysis, mining, or publishing purposes. In addition, it is a unified approach, which gives it the benefit of simplicity, unlike the combination of suppression and generalisation techniques in traditional k-anonymity approaches [13]. Several published works in this area have proposed different clustering techniques to enforce anonymisation models on private data.

For instance, Byun et al. in [8] proposed a greedy algorithm for K-member clustering where each cluster must contain at least k records and the sum of all intra-cluster distances is minimised. Although this is shown to be efficient, it is impractical in cases involving categorical attributes which cannot be enumerated in any specific order. Loukides and Shao in [13] improve the greedy clustering algorithm by introducing measures that capture usefulness and protection in k-anonymisation. Thus, they are able to produce better clusters by

[1] D^k denotes a k-anonymised version of the original table D.

ensuring a balance between usefulness and protection. However, this approach suffers from the same drawbacks as its predecessor. In [6], Lin et al propose a new clustering anonymisation method known as *OKA* (One-pass K-means Algorithm). Unlike the conventional K-means clustering, this only runs for one iteration and proceeds in two phases. In the first phase, all records are sorted by their quasi-identifier. Then K records are randomly selected as the seeds (centroids) to build clusters. The nearest records are assigned to a cluster, and the centroid is subsequently updated. In the second phase, formed clusters are adjusted by removing records from clusters with more than k records and adding them to ones with less than k records. Although this algorithm outperforms the K-member algorithm, however the whole process is restricted to one iteration, thus prohibiting the possibility of finding more optimal clustering solutions. The most current modification of the K-member clustering is an improved approach proposed in [24], referred to as K-member Co-clustering. This approach is adjusted to work in conjunction with maximising the aggregate degree of clustering so that each cluster is composed of records which are mutually related. Despite this, it only performs better than the conventional K-member clustering for high anonymity levels ($k > 30$), and has so far been only applied on numerical attributes. Thus, its true performance against other clustering approaches is yet to be fully determined.

So far, clustering approaches have proven to be successful in providing a trade-off between data utility on one side and data privacy on the other side by enforcing a certain degree of privacy without relinquishing the purposefulness of the data. However, there still remain myriad ways of improving the state of the art through hybrid approaches that can sufficiently reduce the risks of inferences while still maintaining maximal data utility with reasonable computational costs.

3 Self-Organising Map

SOM is a variation of the competitive-learning approach in which the goal is to generate a low-dimensional discretized representation of high-dimensional data while preserving its topological and metric relationships [25]. The basic idea in competitive learning is not to map inputs to outputs in order to correct errors or to have output and input layers with the same dimensionality, as in autoencoders. Rather an input layer and output layer are connected to adjacent neurons based on predefined neighbourhood relationships, forming a topographic map [26]. Neurons are tuned to various input patterns until a winning neuron is determined, where the neuron best matches the input vector, more commonly known as the Best Matching Unit (BMU). The BMU (c) for one input pattern (x) can be formally defined by:

$$||x - x_c|| = min||x - x_i||^2 \quad (1)$$

The closer a node is to the BMU, the more its weights get altered, and the farther away the neighbour is from the BMU, the less it learns. The broad

[2] where $||.||$ is the measure of distance.

idea of the training occurs in a similar manner to K-means clustering where a winning centroid moves by a small distance towards the training instance once a point is assigned to it at the end of each iteration. SOM allows some variation of this framework, albeit in a different way because it cannot guarantee assigning the same number of instances to each class. Despite this, SOM is an excellent tool that can be used for unsupervised applications like clustering and information compression. A number of frameworks for combining SOM with clustering techniques to improve the solutions of data mining have been proposed in [27–29].

4 Adult Dataset

The Adult dataset [17] is used in a variety of studies on data privacy [8,30–32] and is considered the de facto benchmark for experimenting and evaluating anonymisation techniques and PPDM algorithms. It is an extract of the 1994 U.S. census database and is generally applied to predict whether an individual's annual income exceeds $50,000 using traditional statistical modeling and machine learning techniques. The data comprises 48,842 entries with 15 different attributes, of which 8 are categorical and 7 numerical.
Table 1 presents all features of the dataset, categorised by their attribute types and their attribute set. 9 features of the dataset have been classified as quasi-identifiers, 5 other features as sensitive attributes, and 1 feature as a non-sensitive attribute. The dataset does not contain any value which can directly identify an individual on its own, thus the lack of an identifier category. This classification of the Adult dataset can be defined as follows:

- **Identifiers:** a data attribute that explicitly declares the identity of an individual e.g. name, social security number, ID number, biometric record.
- **Quasi-Identifiers:** a data attribute that is inadequate to reveal individual identities independently, however, if combined with other publicly available information (quasi-identifiers), they can explicitly reveal the identity of a data subject e.g. date of birth, postcode, gender, address, phone number.
- **Sensitive Attributes:** a data attribute that reveals personal information about an individual that they may be unwilling to share publicly. These attributes can implicitly reveal confidential information about individuals when combined with quasi-identifiers and are likely to cause harm e.g. medical diagnosis, financial records, criminal records.
- **Non-Sensitive Attributes:** a data attribute that may not explicitly or implicitly declare any sensitive information about individuals. These records need to be associated with identifiers, quasi-identifiers or sensitive attributes to determine a respondent's behaviour or action e.g. shopping cart items, cookie IDs, advertising IDs.

Table 1 also presents the quality of each feature in the Adult dataset using 3 measures, *correlation, id-ness, and stability.*

Table 1: Adult Dataset

TYPE	FEATURES	ATTRIBUTE		CORR.	ID-NESS %	STABILITY %
CATEGORICAL	Workclass	Ⓠ		0.047	0.03	69.70
	Education	Ⓠ		-0.046	0.05	32.5
	Marital-status		Ⓢ	0.003	0.02	45.99
	Occupation	Ⓠ		-0.105	0.05	12.71
	Relationship	Ⓠ		-0.171	0.02	40.52
	Race	Ⓢ		-0.068	0.02	85.43
	Native-country	Ⓠ		0.034	0.13	89.59
	Gender	Ⓢ		-0.216	0.01	66.92
NUMERICAL	Age	Ⓠ		0.234	0.22	2.76
	Fnlwgt		Ⓝ	-0.009	66.48	0.04
	Education-num	Ⓠ		0.335	0.05	32.25
	Capital-gain		Ⓢ	0.266	0.37	91.67
	Capital-loss		Ⓢ	0.139	0.28	95.33
	Hours-per-week		Ⓢ	0.229	0.29	46.73
	Income		Ⓢ	1.000	0.01	75.92

Ⓠ = Quasi-Identifier, Ⓢ = Sensitive, Ⓝ = Non-Sensitive

- Correlation: measures the linear correlation between each feature and the label feature *(Income)*. (A value between 1 and -1)
- ID-ness: measures the fraction of unique values.
- Stability: measures the fraction of constant non-missing values.

All of these measures are essential in identifying patterns in the dataset which can help determine which features to select or deselect when applying a machine learning model for a specific task.

5 Proposed Strategy

This section presents a hybrid strategy for improving data quality and efficiency in PPDM using clustering-based approaches such as *OKA* and K-member. The proposed strategy works in the following stages:

1. Initially, the dataset is analysed and vertically partitioned based on the attribute set type: categorical or numerical.
2. A traditional k-anonymity clustering algorithm is applied to a local dataset containing the categorical attribute set to produce a k-anonymised result.
3. SOM is applied to compress the local dataset containing the numerical attribute set that are dropped by the clustering-based algorithms and generate a 1-dimensional representation of all input spaces.
4. The partial results are unified in a combined dataset based on their index and reference vectors, ensuring that objects are in the same order as the original dataset.

5. Classification techniques are applied on the combined results for generic data mining tasks.

An overview of the complete architecture is illustrated in Fig.1.

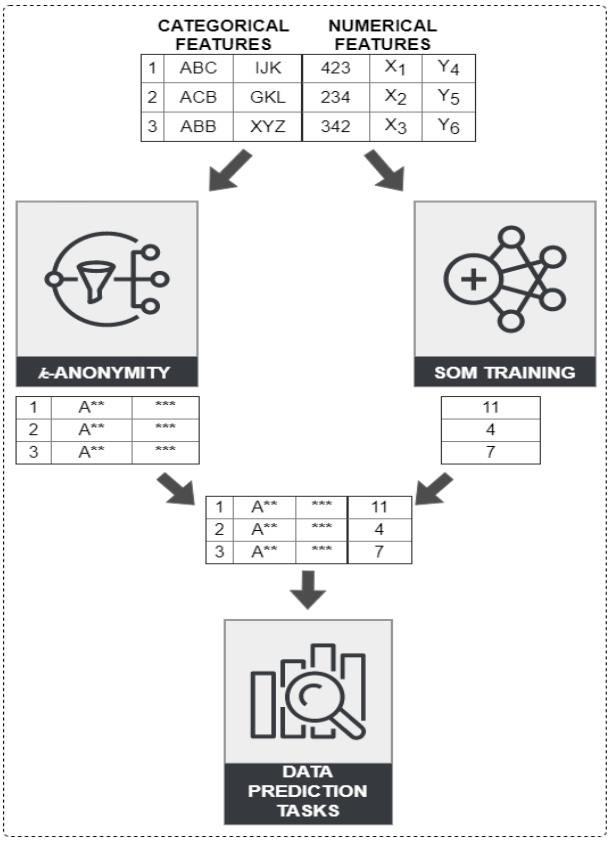


Fig. 1: Architecture of Proposed Strategy

6 Methodology

In order to verify the precision of the proposed strategy, results from this approach are compared with those of conventional clustering-based strategies *(OKA and K-member)*. The implementation of these algorithms available from [33] is specifically designed with the purpose of anonymising the Adult dataset, thus making it suitable for this experiment.

In the aforementioned implementation, a distance function is used to measure dissimilarities among data points for both categorical and numerical attributes.

For numerical attributes, the difference between two values v_i and v_j of a finite numeric domain D is defined as:

$$\delta_N(v_1, v_2) = |v_1 - v_2|/|D| \tag{2}$$

where the domain size $|D|$ is the difference between the maximum and minimum values in D.

However, this is not applicable to categorical attributes as they cannot be enumerated in any specific order. Therefore, for categorical attributes with no semantic relationship amongst their values, every value in such domain is treated as a different entity to its neighbours. For attributes with semantic relationships as is the case in Fig.2 and Fig.3, a taxonomy tree is applied to define the dissimilarity (i.e., distance). Therefore, the distance between two values v_i and v_j of a categorical domain D is defined as:

$$\delta_C(v_1, v_2) = H(\Lambda(v_i, v_j))/H(T_D), \tag{3}$$

where $\Lambda(v_i, v_j)$ is the subtree rooted at the lowest common ancestor of x and y, and $H(T)$ represents the height of tree T.

Example 1. Consider attribute *Workclass* and its taxonomy tree in Fig.3. The distance between *Federal-gov* and *Never-worked* is $2/2 = 1$, while the distance between *Federal-gov* and *Private* is $1/2 = 0.5$. On the other hand, for attribute *Race* as defined in Fig.4, where the taxonomy tree has only one level, the distance between all values is always 1.

It is important to note that only the *Marital-status* and *Workclass* attributes have a predefined taxonomy tree in the clustering implementations published in [33].

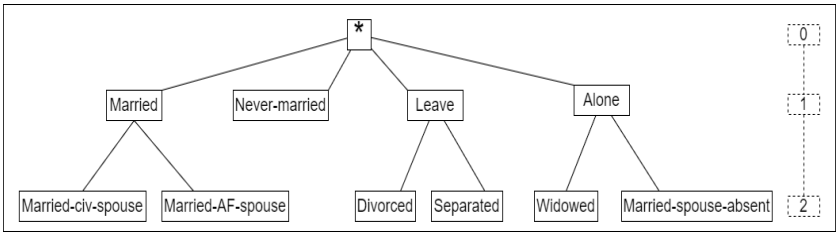


Fig. 2: Taxonomy Tree of *Marital-status*

In our SOM architecture, we use cosine similarity as a distance metric, which ensures the smallest distance between points from the same class and a large margin of separation of points from different classes. This is a particularly useful approach because the Adult dataset has a combination of categorical and numerical data and other more common measures do not translate the distance

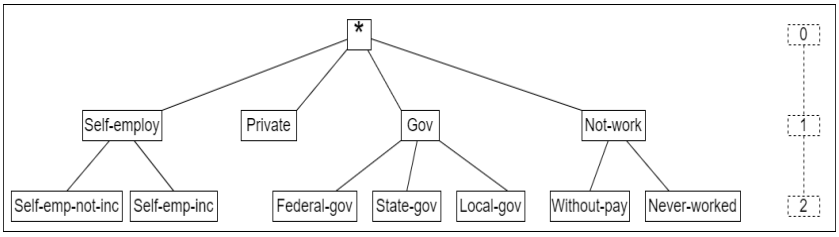


Fig. 3: Taxonomy Tree of *Workclass*

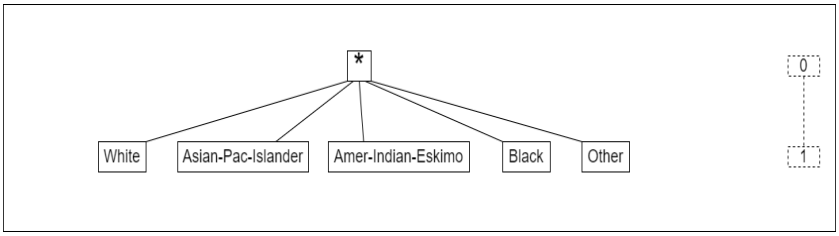


Fig. 4: Taxonomy Tree of *Race*

well between vectors with categorical data. The cosine similarity of two vectors of attributes, a and b, can be formally defined as:

$$C_{(a,b)} = \frac{a * b}{|a| * |b|} \tag{4}$$

Herein, we used a one-dimensional set of 150 neurons. For each sample i, we search for a neuron which is closest to it. The neuron with the smallest distance to the i-th sample is classified as the BMU, and the weight update is executed until all samples are mapped to an output neuron in the set.

We utilise a method known as hyper-parameter optimization for choosing an optimal set of neurons for our SOM based on their correlation with the *Income* attribute. The ultimate goal of the prediction task with the Adult dataset is to identify who earns a certain type of income, thus any set of neurons with the highest correlation with the label feature can enhance this task with the anonymised version of the dataset. With this method, we perform an exhaustive search of all possible neurons within a range of manually set bounds. Following this, the set of neurons with the highest correlation with the label feature (i.e., winning neurons or BMUs) is selected as the optimal neuron set.

Finally, we unify our anonymised features with the SOM feature in a central dataset based on their index and reference vectors. Then, we subject this output to 7 classification models for performing the prediction tasks the Adult dataset is intended for (i.e., income prediction). The seven classification models applied are *Naive Bayes, Generalised Linear Model, Logistic Regression, Deep Learning, Decision Tree, Random Forest* and *Gradient Boosted Trees*.

To validate the experiment, several quality measures were used to evaluate and compare the results of our proposed strategy with the two traditional clustering approaches highlighted earlier *(OKA and K-member)*. The quality measures are as follows:

1. Normalised Certainty Penalty (NCP): measures information loss of all formed equivalence classes.

   (a) For attributes that are numerical, the NCP score of an equivalence class T is defined as:

   $$NCP_{A_{num}}(T) = \frac{max^{T}_{A_{num}} - min^{T}_{A_{num}}}{max_{A_{num}} - min_{A_{num}}} \tag{5}$$

   Where the denominator and numerator represent attribute ranges, say A_{num} for the whole table and T class, respectively.

   (b) For attributes that are categorical, in which no distance function or complete order is present, NCP is described w.r.t the attribute's taxonomy tree:

   $$NCP_{A_{cat}}(T) = \begin{cases} 0, & card(u) = 1 \\ card(u)/|A_{cat}|, & otherwise \end{cases} \tag{6}$$

   where u represents lowermost common predecessor of all values in A_{cat} that are involved in T, $card(u)$ is the number of leaves (i.e., values of attribute) in $u's$ subtree, and $|A_{cat}|$ represents the total count of discrete values of A_{cat}.

   (c) The NCP of class T complete attributes quasi-identifier is:

   $$NCP(T) = \sum_{i=1}^{n} w_i \cdot NCP_{A_i}(T) \tag{7}$$

   where n represents attributes in the dataset. A_i is one of categorical or numerical attribute and have w_i weight, where $\sum w_i = 1$.

2. Accuracy: measures the percentage of correctly classified instances by the classification model used, which is calculated using the number of *(true positives {TP}, true negatives {TN}, false positives {FP} and false negatives {FN})* [34]. Classification accuracy is defined mathematically as:

   $$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

3. FMeasure: another classification-based metric used to measure the accuracy of a classifier model. The metric score computes the harmonic mean between

precision and recall. Precision p denotes the number of true positives divided by all positive results returned by the classification model, whereas recall r denotes the number of true positives by the number of all samples which should have been returned as positive [34].

$$F_1 = 2 * \left(\frac{p * r}{p + r} \right) \tag{9}$$

4. Time: indicates the length of time it takes to execute an algorithm based on an input data size and a k parameter.

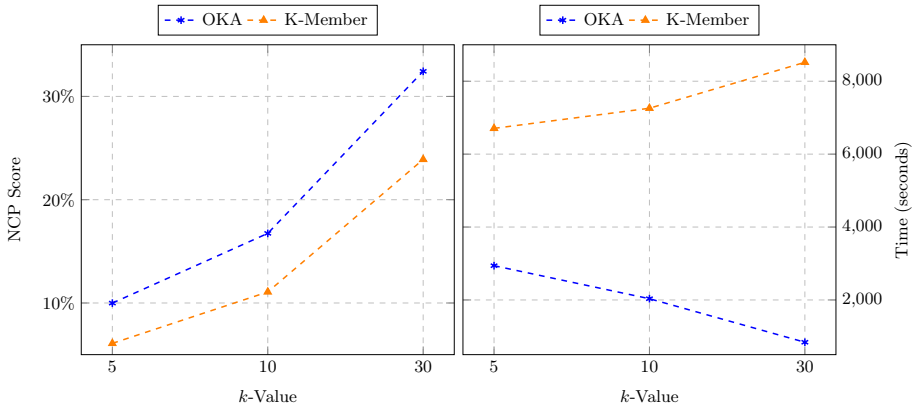
7 Experiments

In this section, we discuss our environment and evaluation methods, which include both information loss and privacy preservation. We have evaluated the accuracy of the proposed approach with conventional classification models on the original and anonymised datasets. The test environment used for our experiment is a Windows platform with an Intel(R) i5-7500T 2.7GHz 4-core processor and 16GB of memory. We have also used another platform, a MacBook with a 2.5 GHz dual-core processor and 8GB memory.

Table 2: NCP score and running time of *OKA* and *K-member* algorithms with 3 different k thresholds.

k-VALUE	**ALGORITHM**	**NCP %**	**TIME(sec)**
5-anonymity	*OKA*	9.99	2939.93
	K-member	6.09	6706.59
10-anonymity	*OKA*	16.74	2034.22
	K-member	11.07	7258.76
30-anonymity	*OKA*	32.43	840.12
	K-member	23.90	8518.48

We have evaluated the performance of the proposed approach with respect to privacy, execution time, accuracy, and F1 score, where F1 score is calculated on the original and updated anonymised version of the Adult dataset.
First, we have evaluated the NCP score of the *K-member* and *OKA* algorithms, considering 3 different k thresholds for anonymity as illustrated in table 2. We have used the *OKA* and *K-member* algorithms for the purpose of anonymity. It is observed that the NCP score using *OKA* is always higher as compared to *K-member*, and by increasing k threshold for anonymity, the difference in loss also increases. The reason behind this is that OKA only uses one iteration for clustering, which leads to higher information loss; however, its one-pass nature makes it more time efficient than *K-member* clustering, thus its execution time is significantly less than that of *K-member*.



(a) NCP Score **(b)** Time

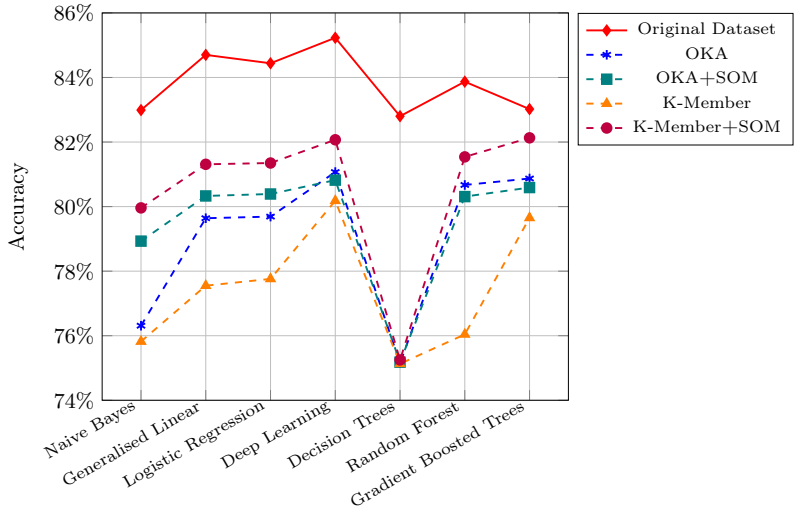
Fig. 5: Information loss & Total time

Following this, we have applied SOM clustering on all the numerical features that are dropped by the two algorithms. These features include *(capital-gain, capital-loss, hours-per-week, and fnlwgt)*. Due to categorical features in the Adult dataset, we have used the cosine similarity metric because Euclidean distance-based results are biased. The bias arises because L1 and L2 distances are not applicable for vectors with text. We have used hyper-parameter tuning to identify the correct number of neurons for SOM, setting the stride size equal to 10 and iterating 100-300 times. After this, we determine the co-relation between the results and the actual income group. We have selected the number of neurons with which we got a higher co-relation.

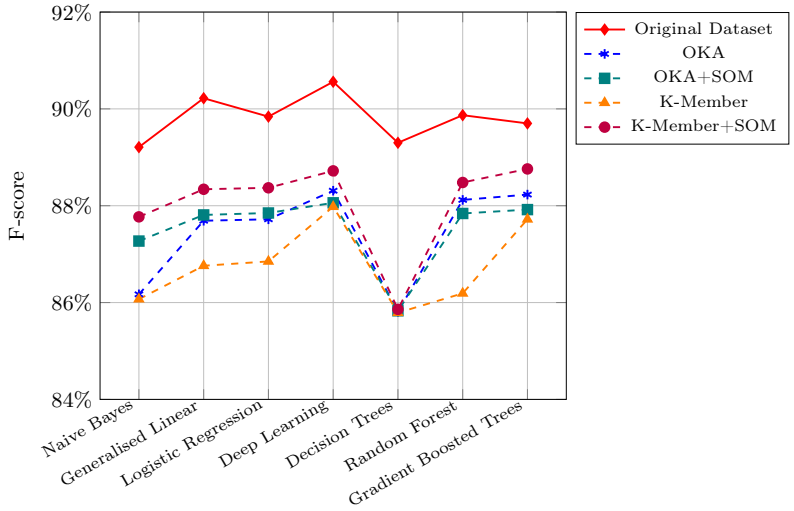
We have modified the Adult dataset in four versions: one is *OKA+SOM*, in which we have applied *OKA* and *SOM* on the original dataset and another version is *K-member+SOM*. The other two versions are obtained by just applying *OKA* and *K-member* methods. After this, we have evaluated the performance of the different resulting datasets on the general classification models. We have categorized the performance by different thresholds of k-anonymity: 5, 10, and 30.

In Fig.6 we have considered 5 members in a cluster. After applying the naive Bayes classification model, we have observed that accuracy of the *K-member+SOM* version of the Adult dataset provides around 80% accuracy whereas on the original dataset it was around 83%. The *OKA+SOM* dataset accuracy is bit lower than that of *K-member+SOM*. Even on the original dataset, the lowest accuracy was given by the decision tree method, and the same applies to our versions. The highest accuracy achieved is around 82% using a deep learning classification model, whereas on the original dataset it is around 85%. The same trend is observed for the F1 score as well.

In Fig.7 we have considered the same variations of data with similar models as we used in previous experiment but with a k-anonymity threshold of 10. It



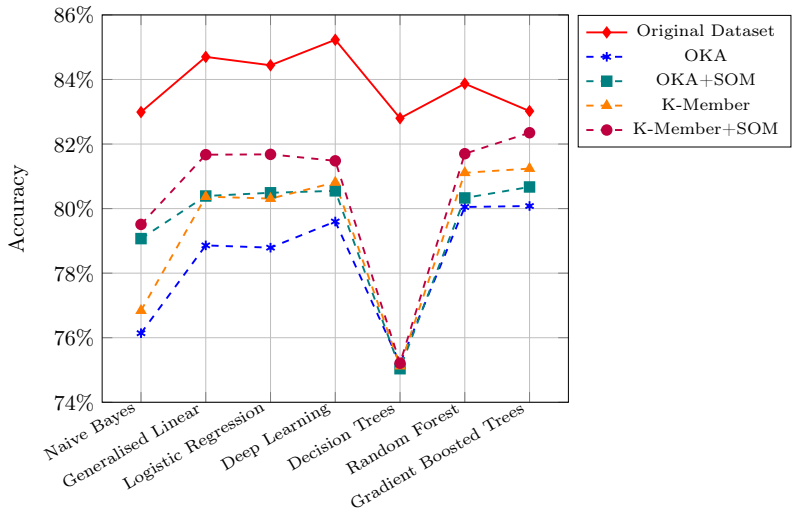
(a) Accuracy of results



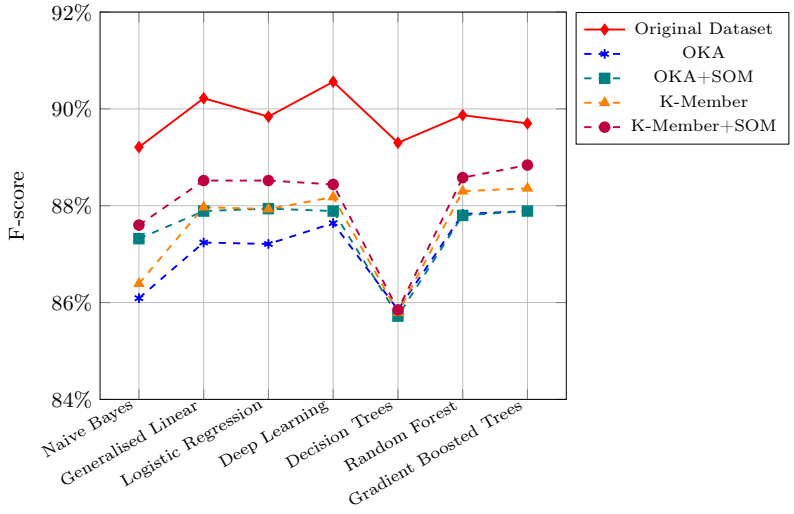
(b) F-score of results

Fig. 6: Income prediction task with *5-anonymity* Adult dataset using seven classification models

is observed that the overall accuracy is lower for all variations of the dataset except for the original one. Still, the dataset generated with *K-member+SOM* gave higher accuracy than other variations on all of the models. In Fig.8 we have evaluated our datasets with anonymity threshold of 30, and we have found that the accuracy of our datasets are slightly lower compared to 10-anonymity classification results, but the *K-member+SOM* dataset still has higher accuracy



(a) Accuracy of results

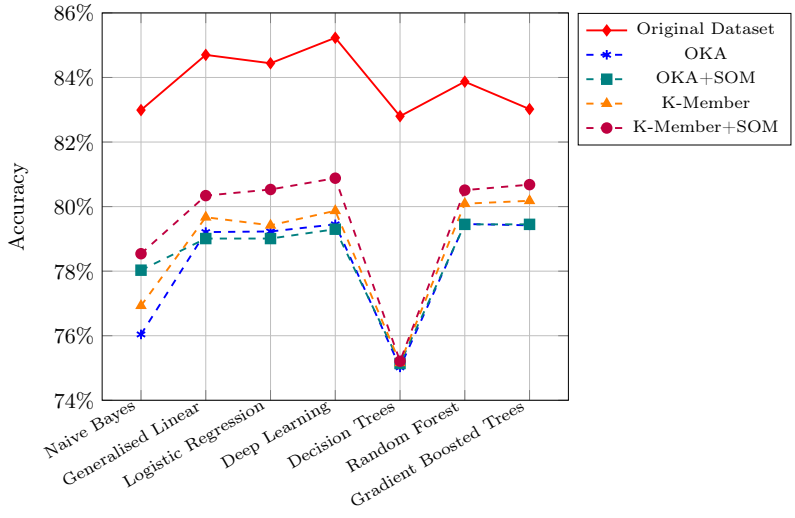


(b) F-score of results

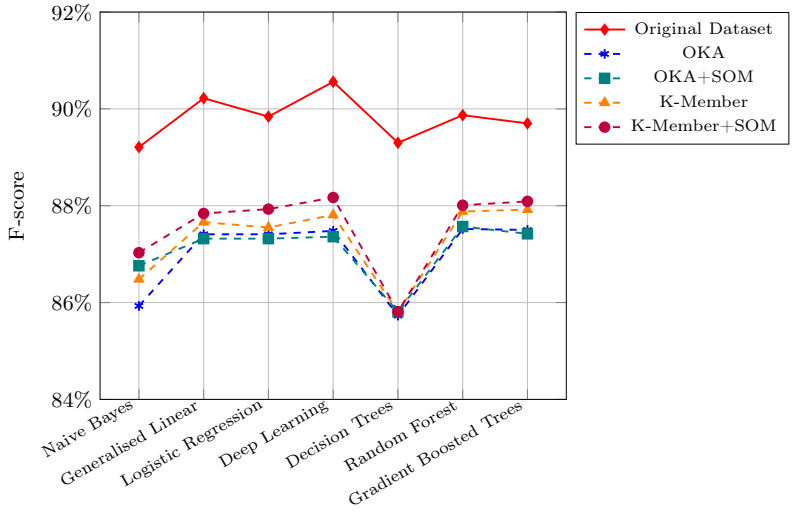
Fig. 7: Income prediction task with *10-anonymity* Adult dataset using seven classification models.

than other variations of original dataset. The *OKA+SOM* dataset accuracy has decreased more significantly than the others.

In evaluations, we have observed that, other than the original dataset, accuracy lowered on all other datasets when the cluster size increased from 5 to 10 to 30. *K-member+SOM* information loss is quite low, which is why its dataset accuracy improved, however, neither the *K-member* nor *OKA* based dataset



(a) Accuracy of results



(b) F-score of results

Fig. 8: Income prediction task with *30-anonymity* Adult dataset using seven classification models.

performed better. Another aspect to consider is that *OKA+SOM* accuracy is lower than *K-member+SOM* because *OKA* uses only one iteration for clustering, which leads to greater time efficiency but also greater information loss compared to other methods. *K-member+SOM* has a trade-off of data loss with time efficiency. This experiment shows that dimensionality reduction is an effective method for preserving the topological and metric relationships of data features,

while anonymising its sensitive content. In addition, results obtained from this process improve the utility of data in classification tasks, as shown in Fig.6, Fig.7 & Fig.8.

8 Conclusion

In this work, we proposed an effective strategy for improving data utility in PPDM approaches by utilising other useful attributes that are otherwise dropped by conventional clustering approaches. By considering these additional attributes, we allow a revised balance between usefulness and protection. We have demonstrated that our approach can efficiently reduce information loss and provide better data for subsequent data mining. Our experiment shows that our strategy has a higher accuracy and F-score in classification tasks for 3 varying k thresholds. Future work will consider applying other data sets to verify the generality of our approach. We will also attempt to optimise our SOM algorithm in order to increase data utility in dimensionality reduction problems and minimise instances of divergence in our results. This will ensure more optimal neurons so that the goal of usefulness can be satisfied in privacy preserving data mining.

References

1. Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, Amsterdam, 3 edition, 2011.
2. M. Narwaria and S. Arya. Privacy preserving data mining — 'a state of the art'. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 2108–2112, March 2016.
3. S. Sharma and D. Shukla. Efficient multi-party privacy preserving data mining for vertically partitioned data. In *2016 International Conference on Inventive Computation Technologies (ICICT)*, volume 2, pages 1–7, Aug 2016.
4. A. Kaur. A hybrid approach of privacy preserving data mining using suppression and perturbation techniques. In *2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, pages 306–311, Feb 2017.
5. W. Liu, S. Luo, Y. Wang, and Z. Jiang. A protocol of secure multi-party multi-data ranking and its application in privacy preserving sequential pattern mining. In *2011 Fourth International Joint Conference on Computational Sciences and Optimization*, pages 272–275, April 2011.
6. Jun-Lin Lin and Meng-Cheng Wei. An efficient clustering method for k-anonymization. In *Proceedings of the 2008 international workshop on Privacy and anonymity in information society - PAIS '08*. ACM Press, 2008.
7. Keng-Pei Lin and Ming-Syan Chen. On the design and analysis of the privacy-preserving SVM classifier. *IEEE Transactions on Knowledge and Data Engineering*, 23(11):1704–1717, nov 2011.
8. Ji-Won Byun, Ashish Kamra, Elisa Bertino, and Ninghui Li. Efficient k-anonymization using clustering techniques. *Advances in Databases: Concepts, Systems and Applications*, pages 188–200, 2007.

9. Stanley Oliveira and Osmar Zaïane. Privacy preserving clustering by data transformation. *Journal of Information and Data Management*, 1(1), 05 2010.
10. Enamul Kabir, Hua Wang, and Elisa Bertino. Efficient systematic clustering method for k-anonymization. *Acta Informatica*, 48(1):51–66, 2011.
11. Xiaoshuang Xu and Masayuki Numao. An efficient generalized clustering method for achieving k-anonymization. In *2015 Third International Symposium on Computing and Networking (CANDAR)*. IEEE, dec 2015.
12. Wantong Zheng, Zhongyue Wang, Tongtong Lv, Yong Ma, and Chunfu Jia. K-anonymity algorithm based on improved clustering. *Algorithms and Architectures for Parallel Processing*, pages 462–476, 2018.
13. Grigorios Loukides and Jianhua Shao. Clustering-based k-anonymisation algorithms. *Database and Expert Systems Applications*, pages 761–771, 2007.
14. Valentina Ciriani, Sabrina De Capitani di Vimercati, Sara Foresti, and Pierangela Samarati. k-anonymous data mining: A survey. In *Privacy-Preserving Data Mining*, pages 105–136. Springer, 2008.
15. Aris Gkoulalas-Divanis and Grigorios Loukides. *A Survey of Anonymization Algorithms for Electronic Health Records*, pages 17–34. Springer International Publishing, Cham, 2015.
16. L. Pin, Y. Wen-bing, and C. Nian-sheng. A unified metric method of information loss in privacy preserving data publishing. In *2010 Second International Conference on Networks Security, Wireless Communications and Trusted Computing*, volume 2, pages 502–505, April 2010.
17. Dheeru Dua and Casey Graff. Adult data set UCI machine learning repository. `http://archive.ics.uci.edu/ml`, 2017.
18. P. Samarati. Protecting respondents identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010–1027, Nov 2001.
19. Valentina Ciriani, S De Capitani Di Vimercati, Sara Foresti, and Pierangela Samarati. k-anonymity. In *Secure data management in decentralized systems*, pages 323–353. Springer, 2007.
20. Pierangela Samarati and Latanya Sweeney. Generalizing data to provide anonymity when disclosing information (abstract). In *Proceedings of the seventeenth symposium on Principles of database systems*. ACM Press, 1998.
21. Adam Meyerson and Ryan Williams. On the complexity of optimal k-anonymity. In *Proceedings of the Twenty-Third ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '04, page 223–228, New York, NY, USA, 2004. Association for Computing Machinery.
22. Balakrushna Tripathy. Database anonymization techniques with focus on uncertainty and multi-sensitive attributes. In *Handbook of Research on Computational Intelligence for Engineering, Science, and Business*, pages 364–383. IGI Global, 2013.
23. Arik Friedman, Ran Wolff, and Assaf Schuster. Providing k-anonymity in data mining. *The VLDB Journal*, 17(4):789–804, Jul 2008.
24. Arina Kawano, Katsuhiro Honda, Hirohide Kasugai, and Akira Notsu. A greedy algorithm for k-member co-clustering and its applicability to collaborative filtering. *Procedia Computer Science*, 22:477–484, 2013.
25. Teuvo Kohonen. *Self-Organizing Maps*. Springer Berlin Heidelberg, 2001.
26. Charu C. Aggarwal. *Neural Networks and Deep Learning*. Springer International Publishing, 2018.
27. Yunus Dogan, Derya Birant, and Alp Kut. SOM++: Integration of self-organizing map and k-means++ algorithms. In *Machine Learning and Data Mining in Pattern Recognition*, pages 246–259. Springer Berlin Heidelberg, 2013.

28. Gorgonio Flavius and Cost Jose Alfredo. PartSOM: A framework for distributed data clustering using SOM and k-means. In *Self-Organizing Maps*. IntechOpen, Apr 2010.
29. S. Tsiafoulis, V. C. Zorkadis, and D. A. Karras. A neural-network clustering-based algorithm for privacy preserving data mining. In *Communications in Computer and Information Science*, pages 269–276. Springer Berlin Heidelberg, 2010.
30. Ji-Won Byun, Yonglak Sohn, Elisa Bertino, and Ninghui Li. Secure anonymization for incremental datasets. In Willem Jonker and Milan Petković, editors, *Secure Data Management*, pages 48–63, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.
31. Mohammad-Reza Zare-Mirakabad, Aman Jantan, and Stéphane Bressan. Privacy risk diagnosis: Mining l-diversity. In Lei Chen, Chengfei Liu, Qing Liu, and Ke Deng, editors, *Database Systems for Advanced Applications*, pages 216–230, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.
32. Ke Wang and Benjamin C. M. Fung. Anonymizing sequential releases. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 414–423, New York, NY, USA, 2006. ACM.
33. Qiyuan Gong. Clustering based k-anonymization. MIT License, January 2016.
34. Aditya Mishra. Metrics to evaluate your machine learning algorithm. `https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234`, 2018.