

DP-DILCA: Learning Differentially Private Context-based Distances for Categorical Data

(Discussion Paper)

Elena Battaglia¹, Ruggero G. Pensa¹

¹University of Turin, Italy

Abstract

Distance-based machine learning methods have limited applicability to categorical data, since they do not capture the complexity of the relationships among different values of a categorical attribute. Nonetheless, categorical attributes are common in many application scenarios, including clinical and health records, census and survey data. Although distance learning algorithms exist for categorical data, they may disclose private information about individual records if applied to a secret dataset. To address this problem, we introduce a differentially private algorithm for learning distances between any pair of values of a categorical attribute according to the way they are co-distributed with the values of other categorical attributes forming the so-called context. We show empirically that our approach consumes little privacy budget while providing accurate distances.

Keywords

differential privacy, metric learning, categorical attributes, distance-based methods

1. Introduction

Most machine learning and data analysis methods rely, directly or indirectly, on their ability to compute distances or similarities between data objects. Although different definitions of distance/similarity exist, they are relatively easy to compute, provided that data are given in form of numeric vectors. Additionally, for most of the above-mentioned distance-based methods, differentially private counterparts of them have been proposed as well. Differential privacy [1] is a computational paradigm which guarantees that the output of a statistical query applied to a secret dataset does not allow to understand whether a particular data object is present in the dataset or not. In recent years, many differentially private variants have been proposed for most distance based algorithms, including kNN [2], SVM [3] and k-means [4].

When data are described by categorical features/attributes, instead, distances can only account for the match or mismatch of the values of an attribute between two data objects, leading to poorer and less expressive proximity measures (e.g., the Jaccard similarity). And yet, intuitively, a patient whose disease is “gastritis” should be closer to a patient affected by “ulcer” than to one having “migraine”. An efficient solution consists in using some distance learning algorithm


SEBD 2021: The 29th Italian Symposium on Advanced Database Systems, September 5-9, 2021, Pizzo Calabro (VV), Italy

✉ elena.battaglia@unito.it (E. Battaglia); ruggero.pensa@unito.it (R. G. Pensa)

ORCID 0000-0001-5145-3438 (R. G. Pensa)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

to infer the distance between any pair of different values of the same categorical attribute from data. Among all existing methods, DILCA [5] is one of the most effective. DILCA’s objective is to compute the distance between any pair of values of a categorical attribute by taking into account the way the two values are co-distributed with respect to the values of other categorical attributes forming the so-called context. According to DILCA, if two values of a categorical attribute are similarly distributed w.r.t. the values of the context attributes, then their distance is lower than that computed for two values of the same attribute that are divergently distributed w.r.t. the values of the same context attributes. DILCA has been successfully used in different scenarios including clustering [5], semi-supervised learning [6] and anomaly detection [7]. However, if applied to a secret dataset, it may disclose a lot of private information.

In this paper, we address the problem of learning meaningful distances for categorical data in a differentially private way. To this purpose, we first introduce a differentially-private extension of DILCA adopting the exponential mechanisms. We show experimentally that it provides accurate distances even with relatively small values of privacy budget ϵ . Additionally, we show that our algorithm (which we call DP-DILCA) is effective in two distance-based learning scenarios, including clustering and k-NN classification.

2. Background

In this section, we introduce the necessary background required to understand the theoretical foundations of our method and, contextually, we introduce its related scientific literature.

2.1. Differential Privacy

Differential privacy [1] is a privacy definition that guarantees the outcome of a calculation to be insensitive to any particular record in the data set. More formally, we report the following definition [1]:

Definition 1 (ϵ -differential privacy). *Let $\mathcal{M} : \Omega \rightarrow \mathcal{R}$ be a randomized mechanism (i.e. a stochastic function with values in a generic set \mathcal{R}) and consider a real number $\epsilon > 0$. We say that \mathcal{M} preserves ϵ -differential privacy if for all pair of datasets D, D' differing for only one record and $\forall r \in \mathcal{R}$, $\frac{P(\mathcal{M}(D)=r)}{P(\mathcal{M}(D')=r)} \leq e^\epsilon$.*

Differential privacy satisfies two important properties: composition and post-processing [1]. The composition property states that by combining the results of several differentially private mechanisms, the outcome will be differentially private too, and the overall level ϵ of privacy guaranteed will be the sum of the level of privacy of each mechanism. On the other hand, the post-processing property says that once a quantity r has been computed in a differentially private way, any following transformation of this quantity is still differentially private, with no need to spend part of the privacy budget for it.

Several mechanisms and techniques preserving differential privacy have been proposed in literature. Two of the most famous mechanisms are the Laplace and the Exponential mechanisms [1]. They both calibrate the amount of random noise they inject in the computation by looking at the sensitivity of the function (or utility function) considered:

Definition 2 (Global sensitivity). Let $q : \Omega \rightarrow \mathbb{R}^d$ be a numeric function. The global sensitivity $GS(q)$ is a measure of the maximal variation of function q when computed over two datasets differing for only one record and is defined as $GS(q) = \max_{D \sim D'} \|q(D) - q(D')\|_1$.

2.2. DILCA

Measuring similarities or distances between two data objects is a crucial step for many machine learning and data mining tasks. While the notion of similarity for continuous data is relatively well-understood and extensively studied, for categorical data the similarity computation is not straightforward. The simplest comparison measure for categorical data is *overlap* [8]: given two tuples, it counts the number of attributes whose values in the two tuples are the same. The overlap measure does not distinguish different values of attributes, hence matches and mismatches are treated equally. Among all the proposed methods for distance computation, we focus on DILCA [5], a framework to learn context-based distances between each pair of values of a categorical attribute Y . The main idea behind DILCA is that the distribution of the co-occurrences of the values of Y and the values of the other attributes in the dataset may help define a distance between the values of Y (intuitively, two values that are similarly co-distributed w.r.t. all the other values of all the other attributes are similar and so they should be close in the new distance). However, not all the other attributes in the dataset should be taken in consideration, but only those that are more relevant to Y . We call this set of relevant attributes with respect to Y the *context* of Y . DILCA distance is defined as follow

Definition 3 (DILCA distance). Let y_1, \dots, y_n be the values of attribute Y . For each pair y_i, y_j with $i, j = 1, \dots, n$, the distance between y_i and y_j is computed as

$$d(y_i, y_j) = \sqrt{\frac{\sum_{X \in \text{context}(Y)} \sum_{k=1}^{|X|} (P(y_i|x_k) - P(y_j|x_k))^2}{\sum_{X \in \text{context}(Y)} |X|}}$$

where $\text{context}(Y)$ is the set of the attributes belonging to the context of Y , $|X|$ is the number of values attribute X can assume, and $P(y_i|x_k)$ is the conditional probability that Y takes value y_i given that X has value x_k .

The conditional probabilities $P(y_i|x_k)$ are estimated from the data: the contingency table between attributes X and Y is constructed and this contingency table can be interpreted as the empirical joint distribution of the two variables.

3. DP-DILCA

In this section, we introduce a method whose final goal is to inject some form of randomness in DILCA in order to make the resulting distances among the values of the target attribute Y differentially private. There are two moments when DILCA algorithm accesses the original (secret) dataset: the context and the contingency tables computation phases. If the context selection is made preserving $h \cdot \varepsilon$ -differential privacy (where $h \in [0, 1]$) and the computation of all the contingency tables is made preserving $(1 - h)\varepsilon$ -differential privacy, then the composition

Algorithm 1: $DPContext(D, Y, h \cdot \varepsilon, k)$

Input: The original dataset D with N records and attributes $F = \{X_1, \dots, X_m\}$, the target attribute $Y \in F$, the privacy budget $h\varepsilon$, the number k of attributes in the context

Result: The set $context(Y)$

```
1  $gs \leftarrow \frac{2}{N} \left( \frac{1}{\ln(2)} + \log(N) \right);$ 
2  $\mathcal{F} \leftarrow \{X_1, \dots, X_m\} \setminus \{Y\};$ 
3  $context(Y) \leftarrow \emptyset;$ 
4 for  $t = 1$  to  $k$  do
5   | Select an object  $X \in \mathcal{F}$  with probability proportional to  $exp\left(\frac{h\varepsilon \cdot MI(Y, X)}{2 \cdot k \cdot gs}\right);$ 
6   |  $context(Y) \leftarrow context(Y) \cup \{X\};$ 
7   |  $\mathcal{F} \leftarrow \mathcal{F} \setminus \{X\};$ 
8 end
```

and the post-processing theorems guarantee that the overall algorithm preserves ε -differential privacy.

The context selection procedure used by DILCA is an application of a filter method for supervised feature selection. Indeed, some work has been done on differentially private feature selection. For instance, [9] and [10] present two alternative differentially private implementations of a feature selection method that preserves nearest-neighbor classification capability. In [11], instead, the authors study the sensitivity of several association measures used for feature selection and integrate the noised version of these measures in two differentially private classifiers. Here, we propose a differentially private selection method that measures the connection of two attributes by looking at the (distorted) Mutual Information between them and then extracts the k most relevant attributes. Mutual Information is a widely used measure of association in the supervised feature selection problem and it can be computed as $I(X, Y) = H(X) + H(Y) - H(X, Y)$ (where $H(\cdot)$ is the entropy function). Thus, finding the X which maximizes $I(X, Y)$ is equivalent to finding the X which maximizes $I'(X, Y) = H(X) - H(X, Y)$.

Theorem 3.1 (Sensitivity of $I'(X, Y)$). *Given a dataset D with N records and two attributes X and Y , an upper bound of the sensitivity of $I'(X, Y)$ is $\frac{2}{N} \left(\frac{1}{\ln(2)} + \log(N) \right)$.*

Algorithm 1 describes the differentially private version of context selection. It requires the specification, as input parameter, of the desired number k of attributes in the context of the target attribute. When setting the value of parameter k , one must consider that lower values of k are preferable, from a differentially private point of view. In step 5 of Algorithm 1, the exponential mechanism is applied k times, in order to extract the top k attributes: each application of the exponential mechanism requires part of the overall privacy budget; thus, the smaller k is, the higher the accuracy of the selected context. Algorithm 1 preserves $h\varepsilon$ -differential privacy.

Once the context of target attribute Y has been selected, DILCA algorithm computes the contingency table $CT(X, Y)$ between Y and X , for each X in context. Instead of the exact value $CT(X, Y)$, we compute a distorted contingency table via the Laplace mechanism. This

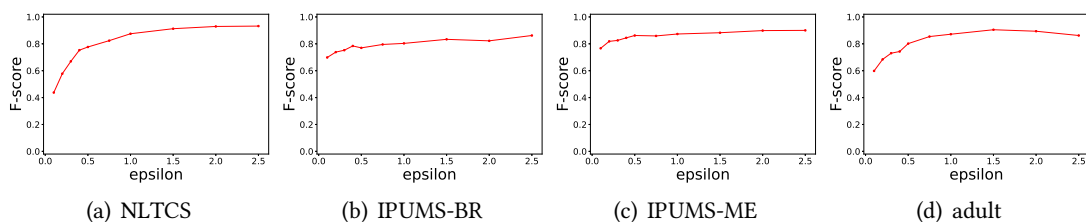


Figure 1: Average F-score of the differentially private context.

mechanism needs the specification of two parameters, the sensitivity of the function CT , that can be proved to be 2, and the privacy budget. Since the total number of needed contingency tables is $k = |\text{context}(Y)|$, the privacy budget we spend for each contingency table is $\frac{(1-h)\epsilon}{k}$. In this way, the overall DP-DILCA algorithm preserves ϵ -differential privacy.

4. Experiments

In this section, we describe the experiments conducted to evaluate the performances of DP-DILCA. For this evaluation, we use four real-world datasets: adult¹, NLTCs², IPUMS-BR and IPUMS-ME³.

4.1. Assessment of context selection

In the first experiment, we run DP-DILCA on the real-world datasets in order to assess the quality of the context they select. For each dataset, we consider one attribute at a time as target attribute and we compute its differentially private context for increasing levels of privacy budget ϵ . Then we compare the context selected by DP-DILCA with the context obtained with the corresponding non-private method. In all the experiments we set $k = 3$. To evaluate the similarity between the private and non-private context for each target attribute, we use the F-score, i.e., the harmonic mean of precision and recall. For each ϵ , we repeat the experiments 30 times and we compute the mean value of all scores. Figure 1 shows the results of our comparison: for each ϵ we report the average value of the F-score over all the attributes of each dataset. In all the datasets, the results achieved by DP-DILCA increase with respect to ϵ .

4.2. Assessment of the distance matrices

In this section we repeat the same experiments on the real-world data presented in Section 4.1, but we focus on the final output of DP-DILCA: the distances between the values of the target attribute. As before, for each dataset we consider one attribute at a time as target and we compute the differentially private distance matrix associated to its values, for increasing levels of privacy budget ϵ . Then we compare the distances obtained with DP-DILCA with those

¹<https://archive.ics.uci.edu/>

²<http://lib.stat.cmu.edu/>

³<https://international.ipums.org/>

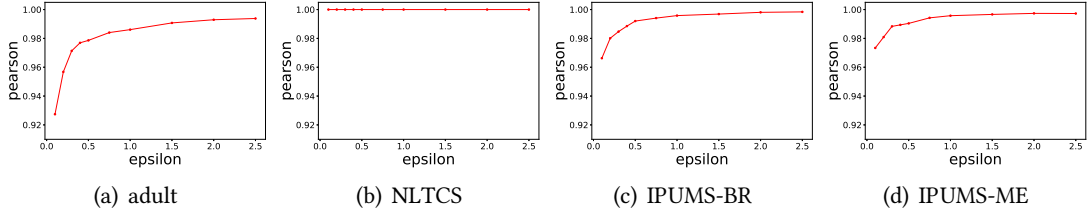


Figure 2: Average Pearson correlation between the differentially private distance matrices and the correspondent non private ones.

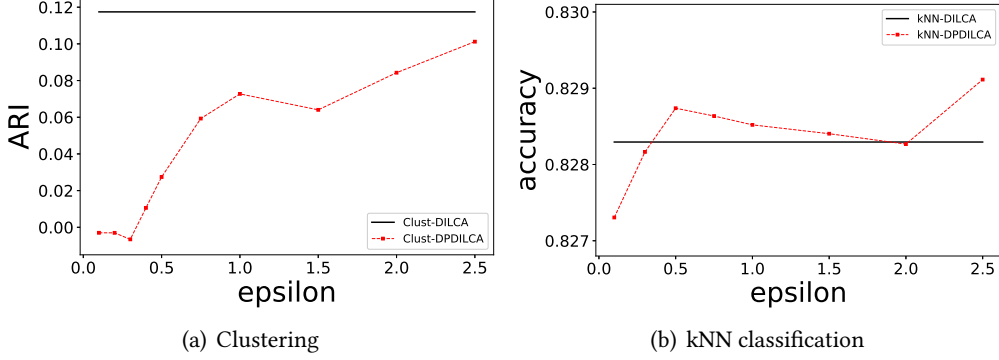


Figure 3: Results on clustering (a) and kNN classification (b).

obtained with the corresponding non-private method. In this experiment we set $k = 3$ and $h = 0.3$. We quantify the linear correlation between the private distance matrix M' , with shape $n \times n$, and its non-private counterpart M through the sample Pearson's ρ correlation coefficient. The ρ coefficient takes values between -1 (perfect negative correlation) and 1 (perfect positive correlation). If the two matrices are not correlated we will have $\rho \cong 0$.

For each ε , we repeat the experiments 30 times and we compute the mean value of the sample Pearson correlation coefficient. Figure 2 shows the results of our computations: for each ε we report the average value of the measure over all the attributes of each dataset. The results show that there is positive correlation between private and non-private distances and the Pearson's coefficient increases as ε grows. Notice that the Pearson coefficient is always 1 when the target attribute has only two values. For this reason, for NLTCS (Figure 2(b)), which consists of binary attributes only, the Pearson's correlation is always maximum.

4.3. Experiments on clustering and classification

In this section, we assess the effectiveness and utility of the distances computed by our differentially private algorithms. To this purpose, we embed DP-DILCA into two distance-based learning algorithms: the Ward's hierarchical clustering algorithm and the kNN classifier. Both the algorithms take as input the matrix of the pairwise distances between the data objects. DP-DILCA's output is the distance between values of a categorical attribute; if it is applied to

all attributes in F , then the distance between any pair of objects o_i, o_j , both described by F can be computed as $objDist(o_i, o_j) = \sqrt{\sum_{X \in F} M_X[o_i.X, o_j.X]^2}$, where M_X is the distance matrix returned by DP-DILCA for attribute X and $o_i.X$ and $o_j.X$ are the values of attribute X on objects o_i and o_j [5]. We will refer to this metric as $objDist_{DPDILCA}$. Similarly, we will call $objDist_{DILCA}$ the metric obtained by the non-private DILCA algorithm.

We run the experiment about clustering as follows: for each real-world dataset, we compute the object distance matrix using the different private and non private metrics, then we run Ward’s hierarchical clustering with these matrices as input. Since the hierarchical algorithm returns a dendrogram which, at each level, contains a different number of clusters, we consider the level corresponding to the number of clusters equal to the number of classes. We call the overall clustering models $Clust_{DPDILCA}$ and $Clust_{DILCA}$, depending on the distance metric adopted. We evaluate the quality of the results through the adjusted rand index (ARI) computed w.r.t. the actual classes [12]. For this reason we run this experiment on dataset adult only. Figure 3(a) shows the mean ARI results over 30 experiments. The value of ε on the x axis of the plot is the overall privacy budget used for the learning of the metric, while the privacy budget spent for computing the distances among values of a single attribute is $\frac{\varepsilon}{m}$. The ARI values of the clustering model with private distance computation grow with respect to the privacy budget, and for high values of ε , they get results close to those of the clustering with non-private distances.

As last experiment, we run the kNN classification algorithm, with $k = 5$. We perform a 4-fold cross-validation: one fold is retained as test set, then the metrics $objDist_{DPDILCA}$, and $objDist_{DILCA}$ are learned on the remaining 3 folds and the classification model is trained on the same set. We call the overall models $kNN_{DPDILCA}$ and kNN_{DILCA} , depending on the distance learning algorithm used. For each dataset, we apply the four kNN models 30 times and compute the mean accuracy of the classification on the the test set. The process is repeated four times and the results are further averaged on the four test sets. In Figure 3(b) we report the mean accuracy of all the models for increasing levels of privacy budget ε . The results of $kNN_{DPDILCA}$ are always very close to those of kNN_{DILCA} , even for very low levels of ε .

5. Conclusion

We have introduced a new family of differentially private algorithms for the data-driven computation of meaningful and expressive distances between any two values of a categorical attribute. Our approach is built upon an effective context-based distance learning framework whose output, however, may reveal private information if applied to a secret dataset. For this reason, we have proposed a randomized algorithm, based on the Laplace and exponential mechanisms, that satisfies ε -differential privacy and returns accurate distance measures even with relatively small privacy budget consumption. Additionally, the metric learnt by our approach can be used profitably in distance-based machine learning algorithms, such as hierarchical clustering and kNN classification.

Acknowledgments

This work is supported by Fondazione CRT (grant number 2019-0450).

References

- [1] C. Dwork, A. Roth, The algorithmic foundations of differential privacy, *Found. Trends Theor. Comput. Sci.* 9 (2014) 211–407.
- [2] M. E. Gursoy, A. Inan, M. E. Nergiz, Y. Saygin, Differentially private nearest neighbor classification, *Data Min. Knowl. Discov.* 31 (2017) 1544–1575.
- [3] K. Chaudhuri, C. Monteleoni, A. D. Sarwate, Differentially private empirical risk minimization, *J. Mach. Learn. Res.* 12 (2011) 1069–1109.
- [4] D. Su, J. Cao, N. Li, E. Bertino, M. Lyu, H. Jin, Differentially private k-means clustering and a hybrid approach to private optimization, *ACM Trans. Priv. Secur.* 20 (2017) 16:1–16:33.
- [5] D. Ienco, R. G. Pensa, R. Meo, From context to distance: Learning dissimilarity for categorical data clustering, *ACM Trans. Knowl. Discov. Data* 6 (2012) 1:1–1:25.
- [6] D. Ienco, R. G. Pensa, Positive and unlabeled learning in categorical data, *Neurocomputing* 196 (2016) 113–124.
- [7] D. Ienco, R. G. Pensa, R. Meo, A semisupervised approach to the detection and characterization of outliers in categorical data, *IEEE Trans. Neural Networks Learn. Syst.* 28 (2017) 1017–1029.
- [8] S. Kasif, S. Salzberg, D. L. Waltz, J. Rachlin, D. W. Aha, A probabilistic framework for memory-based reasoning, *Artif. Intell.* 104 (1998) 287–311.
- [9] J. Yang, Y. Li, Differentially private feature selection, in: *Proceedings of IJCNN 2014*, IEEE, 2014, pp. 4182–4189.
- [10] Y. Li, J. Yang, W. Ji, Local learning-based feature weighting with privacy preservation, *Neurocomputing* 174 (2016) 1107–1115.
- [11] B. Anandan, C. Clifton, Differentially private feature selection for data mining, in: *Proceedings of ACM IWSPA@CODASPY 2018*, 2018, pp. 43–53.
- [12] L. Hubert, P. Arabie, Comparing partitions, *Journal of Classification* 2 (1985) 193–218.