Unsupervised Representation Learning by Predicting Random Distances

Hu Wang*, Guansong Pang*, Chunhua Shen†, Congbo Ma
The University of Adelaide, Australia

Abstract

Deep neural networks have gained great success in a broad range of tasks due to its remarkable capability to learn semantically rich features from high-dimensional data. However, they often require large-scale labelled data to successfully learn such features, which significantly hinders their adaption in unsupervised learning tasks, such as anomaly detection and clustering, and limits their applications to critical domains where obtaining massive labelled data is prohibitively expensive. To enable unsupervised learning on those domains, in this work we propose to learn features without using any labelled data by training neural networks to predict data distances in a randomly projected space. Random mapping is a theoretically proven approach to obtain approximately preserved distances. To well predict these distances, the representation learner is optimised to learn genuine class structures that are implicitly embedded in the randomly projected space. Empirical results on 19 real-world datasets show that our learned representations substantially outperform a few state-of-the-art methods for both anomaly detection and clustering tasks. Code is available at: https://git.io/RDP

1 Introduction

Unsupervised representation learning aims to automatically extract expressive feature representations from data without any labelled data. Due to the remarkable capability to learn semantically rich features, deep neural networks have been becoming one widely-used technique to empower a broad range of machine learning tasks. One main issue with these deep learning techniques is that a massive amount of labelled data is typically required to successfully learn these expressive features. As a result, their transformation power is largely reduced for tasks that are unsupervised in nature, such as anomaly detection and clustering. This is also true to critical domains, such as healthcare and fintech, where collecting massive labelled data is prohibitively expensive and/or is impossible to scale. To bridge this gap, in this work we explore fully unsupervised representation learning techniques to enable downstream unsupervised learning methods on those critical domains.

In recent years, many unsupervised representation learning methods [Mikolov et al., 2013; Le and Mikolov, 2014; Misra et al., 2016; Lee et al., 2017; Gidaris et al., 2018] have been introduced, of which most are self-supervised approaches that formulate the problem as an annotation free pretext task. These methods explore easily accessible information, such as temporal or spatial neighbourhood, to design a surrogate supervisory signal to empower the feature learning. These methods have achieved significantly improved feature representations of text/image/video data. But they are often inapplicable to tabular data since it does not contain the required temporal or spatial supervisory information. We therefore focus on unsupervised representation learning of high-dimensional tabular data. Although many approaches, such as random projection [Li et al., 2006], manifold learning [Donoho and Grimes, 2003; Hinton and Roweis, 2003] and autoencoder [Vincent et al., 2010], are readily available for handling those data, many of them [Donoho and Grimes, 2003; Hinton and Roweis, 2003; Rahmani and Atia, 2017] are often too computationally costly to scale up to large or high-dimensional data. Approaches like random projection and autoencoder are very efficient but they often fail to capture complex class structures due to its underlying data assumption or weak supervisory signal.

In this paper, we introduce a Random Distance Prediction (RDP) model which trains neural networks to predict data distances in a randomly projected space. Particularly, as distances generally carry intrinsic class structure information in the data, the representation learner is optimised to learn the class structure to minimise the prediction error. We seek to obtain distances preserved in a projected space to be the supervisory signal because distances are concentrated and become meaningless in high dimensional spaces [Beyer et al., 1999]. Random mapping is a highly efficient yet theoretical proven approach to obtain such approximately preserved distances. Therefore, we leverage the distances in the randomly projected space to learn the desired features. Intuitively, random mapping preserves rich local proximity information but

---

*Equal contribution.
†Corresponding author.
may also keep misleading proximity when its underlying data distribution assumption is inexact; by minimising the random distance prediction error, RDP essentially leverages the preserved data proximity and the power of neural networks to learn globally consistent proximity and rectify the inconsistent proximity information, resulting in a substantially better representation space than the original space. We show that this simple random distance prediction enables us to achieve expressive representations without using labelled data. In addition, some task-dependent auxiliary losses can be optionally added as a complementary supervisory source to the random distance prediction, so as to learn the feature representations that are more tailored to a specific task. In summary, we make the following three main contributions.

- We propose a random distance prediction formulation, which is very simple yet offers a highly effective supervisory signal for learning expressive feature representations that optimise the distance preserving in random projection. The learned features are sufficiently generic and work well for downstream prediction tasks.

- Our formulation is flexible to incorporate task-dependent auxiliary losses that are complementary to random distance prediction to further enhance the learned features, i.e., features that are specifically optimised for a downstream task while at the same time preserving the generic proximity as much as possible.

- As a result, we show that our instantiated model termed RDP achieves substantially better performance than state-of-the-art competing methods in two key unsupervised tasks, anomaly detection and clustering.

2 Random Distance Prediction Model

The Proposed Formulation and The Instantiated Model

We propose to learn representations by training neural networks to predict distances in a randomly projected space without manually labelled data. The key intuition is that, given some distance information that faithfully encapsulates the underlying class structure in the data, the representation learner is forced to learn the class structure in order to yield distances that are as close as the given distances. Our proposed framework is illustrated in Figure 1. Specifically, given data points \( x_i, x_j \in \mathbb{R}^D \), we first feed them into a weight-shared Siamese-style neural network \( \phi(\cdot; \Theta) \), \( \phi : \mathbb{R}^D \rightarrow \mathbb{R}^M \) is a representation learner with the parameters \( \Theta \) to map the data onto a \( M \)-dimensional new space. Then we formulate the subsequent step as a distance prediction task and define a loss function as:

\[
L_{\text{rdp}}(x_i, x_j) = l(\langle \phi(x_i; \Theta), \phi(x_j; \Theta) \rangle, \langle \eta(x_i), \eta(x_j) \rangle),
\]

where \( \eta \) is an existing projection method and \( l \) is a function of the difference between its two inputs.

Here one key ingredient is how to obtain trustworthy distances via \( \eta \). Also, to efficiently optimise the model, the distance derivation needs to be computationally efficient. In this work, we use the inner products in a randomly projected space as the source of distance/similarity since it is very efficient and there is strong theoretical support of its capacity in preserving the genuine distance information. Thus, our instantiated model RDP specifies \( L_{\text{rdp}}(x_i, x_j) \) as follows:

\[
L_{\text{rdp}}(x_i, x_j) = (\phi(x_i; \Theta) \cdot \phi(x_j; \Theta) - \eta(x_i) \cdot \eta(x_j))^2,
\]

where \( \phi \) is implemented by multilayer perceptron for dealing with tabular data and \( \eta : \mathbb{R}^D \rightarrow \mathbb{R}^k \) is an off-the-shelf random data mapping function (see Sections 3.1 and 3.2 for detail).

Despite its simplicity, this loss offers a powerful supervisory signal to learn semantically rich feature representations that substantially optimise the underlying distance preserving in \( \eta \) (see Section 3.3 for detail).

Flexibility to Incorporate Task-dependent Complementary Auxiliary Loss

Minimising \( L_{\text{rdp}} \) learns to preserve pairwise distances that are critical to different learning tasks. Moreover, our formulation is flexible to incorporate a task-dependent auxiliary loss \( L_{\text{aux}} \) such as reconstruction loss [Hinton and Salakhutdinov, 2006] for clustering or novelty loss [Burda et al., 2019] for anomaly detection, to complement the proximity information and enhance the feature learning.

For clustering, an auxiliary reconstruction loss is used:

\[
L_{\text{clu}}(x_i) = (x - \phi'(\phi(x; \Theta); \Theta'))^2,
\]

where \( \phi' : \mathbb{R}^M \rightarrow \mathbb{R}^D \) is a decoder. This loss may be optionally added into RDP to better capture global feature representations.

Similarly, in anomaly detection a novelty loss may be optionally added, which is defined as:

\[
L_{\text{ad}}(x_i) = (\phi(x_i; \Theta) - \eta(x_i))^2.
\]

\footnote{Since we operate on real-valued vector space, the inner product is implemented by the dot product. The dot product is used hereafter to simplify the notation.}
By using a fixed $\eta$, minimising $L_{\text{ad}}$ helps learn the frequency of underlying patterns in the data [Burda et al., 2019], which is an important complementary supervisory source for the sake of anomaly detection. As a result, anomalies or novel points are expected to have substantially larger $(\phi(x; \Theta) - \eta(x))^2$ than normal points, so Eqn. (4) is used to define anomaly score for the anomaly detection task.

Note since $L_{\text{ad}}$ involves a mean squared error between two vectors, the dimension of the projected space resulted by $\phi$ and $\eta$ is required to be equal in this case. Therefore, when this loss is added into RDP, the $M$ in $\phi$ and $K$ in $\eta$ need to be the same. We do not have this constraint in other cases.

Overall, our loss function will then be:

$$L = L_{\text{rdp}} + L_{\text{aux}},$$ (5)

where $L_{\text{aux}}$ is $L_{\text{aux}}^{clu}$ for clustering and $L_{\text{aux}}^{ad}$ for anomaly detection.

3 Theory Analysis of RDP

This section shows the proximity information can be well approximated using inner products in two types of random projection spaces. This is a key theoretical foundation to RDP. Also, to accurately predict these distances, RDP is forced to learn the genuine class structure in the data.

3.1 When Linear Projection Is Used

Random projection is a simple yet very effective linear feature mapping technique which has proven the capability of distance preservation. Let $\mathcal{X} \subset \mathbb{R}^{N \times D}$ be a set of $N$ data points, random projection uses a random matrix $A \subset \mathbb{R}^{K \times D}$ to project the data onto a lower $K$-dimensional space by $\mathcal{X}' = A \mathcal{X}$. The Johnson-Lindenstrauss lemma [Johnson and Lindenstrauss, 1984] guarantees the data points can be mapped to a randomly selected space of suitably lower dimension with the distances between the points are approximately preserved. More specifically, let $\varepsilon \in (0, \frac{1}{2})$ and $K = \frac{20 \log N}{\varepsilon^2}$. There exists a linear mapping $f : \mathbb{R}^D \rightarrow \mathbb{R}^K$ such that for all $x_i, x_j \in \mathcal{X}$:

$$(1 - \varepsilon)||x_i - x_j||^2 \leq ||f(x_i) - f(x_j)||^2 \leq (1 + \varepsilon)||x_i - x_j||^2.$$ (6)

Furthermore, assume the entries of the matrix $A$ are sampled independently from a Gaussian distribution $\mathcal{N}(0,1)$. Then, the norm of $x \in \mathbb{R}^D$ can be preserved as:

$$\Pr\left((1 - \varepsilon)||x||^2 \leq \frac{1}{\sqrt{K}}||Ax||^2 \leq (1 + \varepsilon)||x||^2\right) \geq 1 - 2e^{-\frac{(1-\varepsilon-K)||x||^2}{4}}.$$ (7)

Under such random projections, the norm preservation helps well preserve the inner products:

$$\Pr\left(||\hat{x}_i - \hat{x}_j - f(\hat{x}_i) \cdot f(\hat{x}_j)|| \geq \varepsilon\right) \leq 4e^{-\frac{||x_i - x_j||^2}{4}},$$ (8)

where $\hat{x}$ is a normalised $x$ such that $||\hat{x}|| \leq 1$.

The proofs of Eqns. (6-8) can be found in [Vempala, 1998].

Eqn. (8) states that the inner products in the randomly projected space can largely preserve the inner products in the original space, particularly when the dimension $K$ is large.

3.2 When Non-linear Projection Is Used

Here we show that some non-linear random mapping methods are approximate to kernel functions which are a well-established approach to obtain reliable distance/similarity information. The key to this approach is the kernel function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, which is defined as $k(x_i, x_j) = \langle \psi(x_i), \psi(x_j) \rangle$, where $\psi$ is a feature mapping function but needs not to be explicitly defined and $\langle \cdot, \cdot \rangle$ denotes a suitable inner product. A non-linear kernel function such as polynomial or radial basis function (RBF) kernel is typically used to project linear-inseparable data onto a linear-separable space.

The relation between non-linear random mapping and kernel methods is justified in [Rahimi and Recht, 2008], which shows that an explicit randomised mapping function $g : \mathbb{R}^D \rightarrow \mathbb{R}^K$ can be defined to project the data points onto a low-dimensional Euclidean inner product space such that the inner products in the projected space approximate the kernel evaluation:

$$k(x_i, x_j) = \langle \psi(x_i), \psi(x_j) \rangle \approx g(x_i) \cdot g(x_j).$$ (9)

Let $A$ be the mapping matrix. Then to achieve the above approximation, $A$ is required to be drawn from Fourier transform and shift-invariant functions such as cosine function are finally applied to $Ax$ to yield a real-valued output. By transforming $x_i$ and $x_j$ in this manner, their inner product $g(x_i) \cdot g(x_j)$ is an unbiased estimator of $k(x_i, x_j)$.

3.3 Learning Class Structure By Random Distance Prediction

Our model using only the random distances as the supervisory signal can be formulated as:

$$\arg \min_{\Theta} \sum_{x_i, x_j \in \mathcal{X}} (\phi(x_i; \Theta) \cdot \phi(x_j; \Theta) - y_{ij})^2,$$ (10)

where $y_{ij} = \eta(x_i) \cdot \eta(x_j)$. Let $Y_\eta \in \mathbb{R}^{N \times N}$ be the distance/similarity matrix of the $N$ data points resulted by $\eta$. Then to minimise the prediction error in Eqn. (10), $\phi$ is optimised to learn the underlying class structure embedded in $Y$. As shown in the properties in Eqns. (8) and (9), $Y_\eta$ can effectively preserve local proximity information when $\eta$ is set to be either the random projection-based $f$ function or the kernel method-based $g$ function. However, those proven $\eta$ is often built upon some underlying data distribution assumption, e.g., Gaussian distribution in random projection or Gaussian RBF kernel. Thus, the $\eta$-projected features can preserve misleading proximity when the distribution assumption is inexact. In this case, $Y_\eta$ is equivalent to the imperfect ground truth with partial noise. Then optimisation with Eqn. (10) is to leverage the power of neural networks to learn consistent local proximity information and rectify inconsistent proximity, resulting in a significantly optimised distance preserving space. The resulting space conveys substantially richer semantics than the $\eta$ projected space when $Y_\eta$ contains sufficient genuine supervision information.
4 Experiments
This section evaluates our method through two typical unsupervised tasks: anomaly detection and clustering. The non-linear random projection is used in RDP by default throughout all our experiments.

4.1 Performance Evaluation in Anomaly Detection
Experimental Settings
Our RDP model is compared with five state-of-the-art methods, including iForest [Liu et al., 2008], autoencoder (AE) [Hinton and Salakhutdinov, 2006], REPEN [Pang et al., 2018], DAGMM [Zong et al., 2018] and RND [Burda et al., 2019]. iForest and AE are two of the most popular baselines. The other three methods learn representations specifically for anomaly detection. The RDP consists of one fully connected layer with 50 hidden units, followed by a leaky-RELU layer. It is trained using Stochastic Gradient Descent (SGD) as its optimiser for 200 epochs, with 192 samples per batch. The learning rate is fixed to 0.1.

Our RDP model uses the optional novelty loss for anomaly detection task by default. Note that the representation dimension $M$ in the $\phi$ function and the projection dimension $K$ in the $\eta$ function are set to be the same to alleviate parameter tuning. This means that $M = K = 50$ is used here. Similar to RND, given a data point $x$, its anomaly score in RDP is defined as the mean squared error between the two projections resulted by $\phi(x; \Theta^*)$ and $\eta(x)$. Also, a boosting process is used to filter out 5% likely anomalies per iteration to iteratively improve the modelling of RDP. This is because the modelling is otherwise largely biased when anomalies are present. We repeated the boosting process 30 times to obtain statistically stable results. In order to have fair comparisons, we also adapt the competing methods AE, REPEN, DAGMM and RND into ensemble methods and perform the experiments using an ensemble size of 30.

As shown in Table 1, 14 publicly available datasets taken from the literature [Liu et al., 2008; Pang et al., 2018; Zong et al., 2018], are used, which are from various domains, including network intrusion, credit card fraud detection, and disease detection. Many of the datasets contain real anomalies, including DDoS, Donors, Backdoor, Creditcard, Lung, Probe and U2R. Following [Liu et al., 2008; Pang et al., 2018; Zong et al., 2018], the rare class(es) is treated as anomalies in the other datasets to create semantically real anomalies. The Area Under Receiver Operating Characteristic Curve (AUC-ROC) and the Area Under Precision-Recall Curve (AUC-PR) are used as our performance metrics. The reported performance is averaged over 10 independent runs.

Comparison to the State-of-the-art Competing Methods
The AUC-ROC and AUC-PR results are respectively shown in Tables 1 and 2. RDP outperforms all the five competing methods in both of AUC-ROC and AUC-PR in at least 12 out of 14 datasets. This improvement is statistically significant at the 95% confidence level according to the two-tailed sign test [Demšar, 2006]. Remarkably, RDP obtains more than 10% AUC-ROC/AUC-PR improvement over the best competing method on six datasets, including Donors, Ad, Bank, Celeba, Lung and U2R. RDP can be thought as a high-level synthesis of REPEN and RND, because REPEN leverages a pairwise distance-based ranking loss to learn representations for anomaly detection while RND is built using $L_{adv}$. In nearly all the datasets, RDP well leverages both $L_{rdp}$ and $L_{adv}$ to achieve significant improvement over both REPEN and RND. In very limited cases, such as on datasets Backdoor and Census where RND performs very well while REPEN performs less effectively, RDP is slightly downgraded due to the use of $L_{rdp}$. In the opposite case, such as Probe, on which REPEN performs much better than RND, the use of $L_{adv}$ may drag down the performance of RDP a bit.

Ablation Study
This section examines the contribution of $L_{rdp}$ and $L_{adv}$ and the boosting process to the performance of RDP. The experimental results in AUC-ROC are given in Table 3, where RDP$^{X}$ means the RDP variant that removes the ‘X’ module from RDP. Similar observations can be also derived from AUC-PR results that are omitted due to page limits. In the last two columns, $\text{Org}_{SS}$ indicates that we directly use the distance information calculated in the original space as the supervisory signal, while $\text{SRP}_{SS}$ indicates that we use SRP to obtain the distances as the supervisory signal. It is clear that the full RDP model is the best performer. Using the $L_{rdp}$ loss only, i.e., RDP$^{X,L_{adv}}$, can achieve performance substantially better than, or comparably well to, the five competing methods in Table 1. This is mainly because the $L_{rdp}$ loss alone can effectively force our representation learner to learn the under
lying class structure on most datasets so as to minimise its prediction error. The use of $L_{adv}$ and boosting process well complement the $L_{dp}$ loss on the other datasets.

In terms of supervisory source, RDP and SRP_SS perform substantially better than Org_SS on most datasets. This is because the distances in both the non-linear projection in RDP and the linear projection in SRP_SS is well preserved, enabling RDP to effectively learn much more faithful class structure than that working on the original space.

4.2 Performance Evaluation in Clustering

Experimental Settings
For clustering, RDP is compared with four state-of-the-art unsupervised representation learning methods in four different areas, including HLLE [Donoho and Grimes, 2003] in manifold learning, Sparse Random Projection (SRP) [Li et al., 2006] in random projection, autoencoder (AE) [Hinton and Salakhutdinov, 2006] in data reconstruction-based neural network methods and Coherence Pursuit (COP) [Rahmani and Atia, 2017] in robust PCA. These representation learning methods are first used to yield the new representations, and K-means [Hartigan and Wong, 1979] is then applied to the representations to perform clustering. In this section RDP adds the reconstruction loss $L_{aux}$ by default.

RDP uses a similar network architecture and optimisation settings as the one used in anomaly detection. Compared to anomaly detection, more semantic information is required for clustering algorithms to work well, so the network consists of 1,024 hidden units and is trained for 1,000 epochs. AE is built and trained with the same settings as RDP. Similar to anomaly detection, $M = K$ is also used in clustering.

As shown in Table 4, five high-dimensional real-world datasets are used. Some of the datasets are image/text data. Since here we focus on the performance on tabular data, they are converted into tabular data using simple methods, i.e., by treating each pixel as a feature unit for image data or using bag-of-words representation for text data. Two widely-used clustering performance metrics, Normalised Mutual Info (NMI) score and F-score, are used. Larger NMI/F-score indicates better performance. The performance in the original feature space, denoted as Org, is used as a baseline. The reported NMI score and F-score are averaged over 30 times to address the randomisation issue in K-means clustering.

Comparison to the State-of-the-art Competing Methods
Table 4 shows the NMI and F-score performance of K-means clustering. Our method RDP enables K-means to achieve the best performance on three datasets and ranks second in the other two datasets. RDP-enabled clustering performs substantially and consistently better than that based on AE in terms of both NMI and F-score. This demonstrates that the random distance loss enables RDP to effectively capture some class structure in the data which cannot be captured by using the reconstruction loss. RDP also consistently outperforms the random projection method, SRP, and the robust PCA method, COP. It is interesting that K-means clustering performs best in the original space on Sector. This may be due to that this data contains many relevant features, resulting in no obvious curse of dimensionality issue. Olivetti may contain complex manifolds which require extensive neighbourhood information to find them, so only HLLE can achieve this goal in such cases. Nevertheless, RDP performs much more stably than HLLE across the five datasets.

Ablation Study
Similar to anomaly detection, this section examines the contribution of the two loss functions $L_{dp}$ and $L_{aux}$ to the performance of RDP, as well as the impact of different supervisory sources on the performance. The F-score results of this experiment are shown in Table 5, in which the notations have exactly the same meaning as in Table 3. Similar NMI results can also be observed but omitted due to page limits. The full RDP model that uses both $L_{dp}$ and $L_{aux}$ performs more favourably than its two variants, RDP$^{SL}$ and RDP$^{CLA}$, but it is clear that using $L_{dp}$ only performs very comparably to the full RDP. However, using $L_{aux}$ only may result in large performance drops in some datasets, such as R8, 20news and Olivetti. This indicates $L_{aux}$ is a more important loss function to the overall performance of the full RDP model. In terms of supervisory source, distances obtained by the non-linear random projection in RDP are much more effective than the other two sources on some datasets such as Olivetti and RCV1. Three different supervisory sources are very comparable on the other three datasets.

Table 3: AUC-ROC ablation study results of anomaly detection.

Table 4: NMI and F-score performance of K-means.

Table 5: F-score ablation study performance of K-means clustering.
have demonstrated powerful capability to learn semantic representations. However, most of them use the supervisor signals available in image/video data only, which limits their application to other types of data, such as tabular data. Though our method may also work on image/video data, we focus on handling high-dimensional tabular data to bridge this gap.

**Other Approaches.** There have been several well-established unsupervised representation learning approaches for handling tabular data, such as random projection [Bingham and Mannila, 2001; Li et al., 2006], PCA [Schölkopf et al., 1997; Rahmani and Atia, 2017], manifold learning [Donoho and Grimes, 2003; Hinton and Roweis, 2003] and autoencoder [Hinton and Salakhutdinov, 2006; Vincent et al., 2010]. One notorious issue of PCA or manifold learning approaches is their prohibitive computational cost in dealing with large-scale high-dimensional data due to the costly neighbourhood search and/or eigen decomposition. Random projection is a computationally efficient approach, supported by proven distance preservation theories such as the Johnson-Lindenstrauss lemma [Johnson and Lindenstrauss, 1984]. We show that the preserved distances by random projection can be harvested to effectively supervise the representation learning. Autoencoder networks are another widely-used efficient feature learning approach which learns low-dimensional representations by minimising reconstruction errors. One main issue with autoencoders is that they focus on preserving global information only, which may result in loss of local structure information. Some feature learning methods are specifically designed for anomaly detection [Pang et al., 2018; Zong et al., 2018; Burda et al., 2019]. By contrast, we aim at generic representations learning while being flexible to incorporate optionally task-dependent losses to learn task-specific semantically rich representations.

**6 Conclusion**

We have introduced a novel Random Distance Prediction (RDP) model which learns features in a fully unsupervised fashion by predicting data distances in a randomly projected space. The key insight is that random mapping is a theoretically proven approach to obtain approximately preserved distances, and to well predict these random distances, the representation learner is optimised to learn preserved proximity information while at the same time rectifying inconsistent proximity, resulting in representations with optimised distance preserving. Our idea is justified by thorough experiments in two unsupervised tasks, anomaly detection and clustering, demonstrating that RDP-based anomaly detectors and clustering substantially outperform their counterparts on real-world datasets. Our empirical results also demonstrate that RDP is flexible and very effective to incorporate task-dependent complementary auxiliary losses and learns more expressive representations. We are extending RDP to other types of data, such as image/text data.

**References**
