Self-Supervised Deep Low-Rank Assignment Model for Prototype Selection

Xingxing Zhang, Zhenfeng Zhu, Yao Zhao, Deqiang Kong
Institute of Information Science, Beijing Jiaotong University, Beijing, China
Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing, China
{zhangxing,zhfzhu,yzhao,dqkong}@bjtu.edu.cn

Abstract

Prototype selection is a promising technique for removing redundancy and irrelevance from large-scale data. Here, we consider it as a task assignment problem, which refers to assigning each element of a source set to one representative, i.e., prototype. However, due to the outliers and uncertain distribution on source, the selected prototypes are generally less representative and interesting. To alleviate this issue, we develop in this paper a Self-supervised Deep Low-rank Assignment model (SDLA). By dynamically integrating a low-rank assignment model with deep representation learning, our model effectively ensures the goodness-of-exemplar and goodness-of-discrimination of selected prototypes. Specifically, on the basis of a denoising autoencoder, dissimilarity metrics on source are continuously self-refined in embedding space with weak supervision from selected prototypes, thus preserving categorical similarity. Conversely, working on this metric space, similar samples tend to select the same prototypes by designing a low-rank assignment model. Experimental results on applications like text clustering and image classification (using prototypes) demonstrate our method is considerably superior to the state-of-the-art methods in prototype selection.

1 Introduction

Prototype selection is the task of finding exemplar samples, called prototypes, from a large collection of data points. This is at the center of many problems in data analysis and processing field because it holds several advantages over data storage, compression, synthesis, cleansing and visualization. First, the memory cost for storing information on the data can be significantly reduced using prototypes. Second, prototypes help in clustering of data, and, as the most prototypical samples, can be used for efficient synthesis of new data points. More importantly, the computational efficiency for data modeling, such as classifier training [Garcia et al., 2012; Zhang et al., 2018] and active learning [Lin et al., 2018], can be improved by working on prototypes. In addition, selecting prototypes helps to remove redundant or irrelevant points, such as outliers. Finally, prototype selection has been applied for anomaly detection [Cong et al., 2011], web media summarization and recommendation [Meng et al., 2016; Cong et al., 2017] (see Figure 1), segmentation of dynamic data [Elhamifar et al., 2016], and more.

To characterize the informativeness of prototypes in terms of ability to represent the entire distribution, many selection strategies have been proposed, generally including two categories. The first one involves greedy methods that select next prototype with respect to the previously selected prototypes, and usually maximizes submodular functions, such as graph-cuts and facility location [Wang and Zhang, 2013; Elhamifar and Kaluza, 2017; Krause et al., 2008]. Obviously, they usually result in high computational complexity. The second category is comprised of model-driven methods that aim to maximize global objective based on subspace learning or pairwise relationship. Their resulting solutions are closely related to a variety of optimization algorithms.

Concretely, subspace learning based models mainly focus on the data that lie in one or several low-dimensional subspaces. The Rank Revealing QR algorithm [Chan, 1987] is a representative one. By representing data in a low dimensional space with possibly minimal representation error, various effective models have been formulated to learn a selection matrix, such as the sparse dictionary selection (SDS) method [Cong et al., 2011] and sparse modeling representative selection (SMRS) model [Elhamifar et al., 2012].

Figure 1: Assisted photo albuming (i.e. web media recommendation) by prototype selection. When we want to build a new photo album, the only thing we need is to select exemplar samples related to the initialized topic.
To further exploit data structure, some improved methods have been proposed [Dornaika and Aldine, 2015; Wang et al., 2017]. For instance, structured sparse dictionary selection (SSDS) [Wang et al., 2017] method tried to select prototypes with both representativeness and diversity via three structured regularizers. However, by multi-linear coding criteria, these methods are effective only when the samples from different groups are sufficiently dissimilar.

In contrast, pairwise relationship based models aim at the data that could be naturally grouped under a certain dissimilarity metric. One naive approach is Kmedoids [Kaufman and Rousseeuw, 1987], which finds k medoid centers as prototypes by pursuing the minimum total distance from all samples. Some variants [Nellore and Ward, 2015] were further proposed based on Kmedoids. Unlike Kmeans [Duda et al., 2012] or Kmedoids, Affinity Propagation (AP) algorithm [Frey and Dueck, 2007] does not require any initialization for prototypes, but has suboptimal property. To tackle this issue, dissimilarity-based sparse subset selection (DS3) method [Elhamifar et al., 2016] was recently proposed to select prototypes via a trace minimization model. In general, this kind of methods performs well only under appropriate similarity metric.

Furthermore, some real datasets do not live in a vector space, e.g., social network data or proteomics data. Therefore, model-driven methods based on pairwise relationship have more advantages on prototype selection. However, most existing algorithms suffer from imposing restrictions on the type of pairwise relationship. More importantly, unreliable metric caused by outliers or uncertain data distribution significantly reduces the quality of prototypes. Inspired by representation learning [Song et al., 2017], we consider introducing a unified framework for dissimilarity learning and prototype selection. As a result, the metric is efficiently self-refined with weak supervision from prototypes, and conversely, prototypes are more informative by using refined metrics. These construct the basic motivation of our framework.

In summary, the main contributions of this work are highlighted as follows.

- By considering the prototype selection as a task assignment problem, we develop a Self-supervised Deep Low-rank Assignment model (SDLA), which aims to jointly learn ideal dissimilarity metrics in embedding space and select informative prototypes in metric space.
- Unlike DS3 [Elhamifar et al., 2016], inspired by the good performance of deep representation learning, we propose to learn dissimilarity metrics based on a variant of a denoising autoencoder, in which categorical similarity is preserved with weak supervision from prototypes.
- Working on the learned metric space, similar samples tend to select the same prototypes by designing a low-rank assignment model, thus guaranteeing a diversified selection.
- The quality of prototypes selected by the proposed framework is reasonably evaluated with examples in text clustering and image classification (using prototypes), showing very promising results.

2 The Proposed Framework

2.1 Problem Statement

Let $X = [x_1, ..., x_m] \in \mathbb{R}^{d \times m}$ be the source data matrix of $m$ samples in $\mathbb{R}^d$. As shown in Figure 2, we consider the problem of selecting an interesting subset of $m$ samples, called prototype set, that can efficiently describe and summarize all samples in source data. Inspired by DS3 [Elhamifar et al., 2016], we primarily pursue the minimum assignment cost based on pairwise relationships on source.

2.2 Problem Formulation

It is noteworthy that pairwise dissimilarity matrix $D = \{d_{ij}\}_{i,j=1,...,m}$ between samples is directly given as the input to DS3 [Elhamifar et al., 2016]. Actually, the given $\{d_{ij}\}$, such as Euclidean distance or $\chi^2$ distance, may be not discriminative due to uncertain distribution and outliers in $X$, thus leading to a less representative and diversified selection. In light of the good performance of representation learning [Bengio et al., 2013], we consider circularly refining the dissimilarity metric and the selection indicator. Thereby, available semantic information from selected prototypes can be employed to learn the parameters of metric architecture, while conversely, more discriminative metric can improve the quality of prototypes. Towards this end, we develop a Self-supervised Deep Low-rank Assignment model (SDLA) for prototype selection as follows:

$$\begin{align*}
\min_{\Theta, \{z_{ij}\}} & \sum_{j=1}^{m} \sum_{i=1}^{m} \Phi_{\text{dSim}}(\Phi_{\text{rep}}(\Theta; x_i), \Phi_{\text{rep}}(\Theta; x_j))z_{ij} \\
& + \omega_1 \Omega(\Theta) + \omega_2 \Psi(Z) \\
\text{s.t.} & \sum_{i=1}^{m} z_{ij} = 1, \forall j; \quad z_{ij} \geq 0, \forall i, j,
\end{align*}$$

(1)

where $\Phi_{\text{rep}}(\Theta; \cdot)$ learns a deep representation of a sample, and $\Theta$ is the set of related parameters. $\Phi_{\text{dSim}}(\cdot; \cdot)$ aims to capture the dissimilarity of two samples in learned embedding space, thus $d_{ij} = \Phi_{\text{dSim}}(\Phi_{\text{rep}}(\Theta; x_i), \Phi_{\text{rep}}(\Theta; x_j))$. $Z = \{z_{ij}\}_{i,j=1,...,m}$ is a selection matrix, $z_{ij} \in [0, 1]$ is the probability of selecting $x_i$ as the prototype of $x_j$, thus enforcing $\sum_{i=1}^{m} z_{ij} = 1$ for $x_j$. The first term in the objective function corresponds to the total assignment cost, the second term corresponds to the constraints in the deep architecture.
and the last term corresponds to the number of prototypes. \( \omega_1 \) and \( \omega_2 \) are nonnegative parameters to balance these terms. Figure 3 provides an illustration of the proposed model in (1).

Concretely, for any \( x_i \) and \( x_j \), if \( d_{ij} = 0 \), then \( z_{ij} = 1 \) due to low assignment cost, which implies that \( x_i \) and \( x_j \) are from the same class. While \( z_{ij} = 0 \) if \( d_{ij} = \infty \), which implies that \( x_i \) and \( x_j \) are from different classes. Since \( Z \) provides additional semantic information for all samples, \( D \) will be learned more discriminatively by preserving categorical similarity with weak supervision from \( Z \).

Considering the separability of both objective and constraints in (1), we further divide (1) into two subproblems \( P_1 \) and \( P_2 \), which are cyclic self-refinement:

- \( P_1 \) corresponding to the pairwise dissimilarity learning module in Figure 3, is written as

\[
\min_{\Theta} \sum_{i,j=1}^{m} \Phi_{dSim}(\Phi_{\text{rep}}(\Theta; x_i), \Phi_{\text{rep}}(\Theta; x_j))z_{ij} + \omega_1 \Omega(\Theta)
\]

(2)

- \( P_2 \) corresponding to the assignment model learning module in Figure 3, is written as

\[
\min_{\{z_{ij}\}} \sum_{i,j=1}^{m} \Phi_{dSim}(\Phi_{\text{rep}}(\Theta; x_i), \Phi_{\text{rep}}(\Theta; x_j))z_{ij} + \omega_2 \Psi(Z)
\]

\[
s.t. \sum_{i=1}^{m} z_{ij} = 1, \forall j; \quad z_{ij} \geq 0, \forall i, j
\]

(3)

The details about learning \( \Theta \) and \( Z \) are introduced in Section 3.1 and Section 3.2, respectively. As a result, the indices of nonzero rows of the solution \( Z^* \) correspond to the indices of those samples that are chosen as the data prototypes. In addition, \( Z^* \) shows the membership of samples in \( X \) to prototypes. That is, \( z^*_j = [z_{1j}^*, \ldots, z_{mj}^*]^T \in \mathbb{R}^m \) corresponds to the probability vector of \( x_j \) being represented by each sample in \( X \). Therefore, we can obtain a partitioning of \( X \) under the rule that, if \( X = \{x_1, \ldots, x_m\} \) denotes the set of selected prototypes, we can assign \( x_j \) to the prototype \( x_{\delta_j} \) by

\[
\delta_j = \arg \min_{i \in \{1, \ldots, m\}} \! z_{ij} \quad \text{(4)}
\]

Consequently, \( X \) is classified into \( \kappa \) groups corresponding to \( \kappa \) prototypes via the selection matrix \( Z^* \).

3 Optimization

3.1 Pairwise Dissimilarity Learning

As shown in (2), optimizing \( \Theta \) aims to refine pairwise dissimilarities \( \{d_{ij}\} \), and it seamlessly connects the visual content and dissimilarity metric with weak supervision from \( Z \). However, due to the outliers and uncertain distribution on \( X \), we consider generating distributed representation vectors and pairwise dissimilarities on the basis of a denoising autoencoder [Vincent et al., 2008]. The general denoising autoencoder is formulated as:

\[
\begin{align*}
\bar{x}_i &\sim q(\bar{x}_i|x_i); \\
s_i &\sim f(W\bar{x}_i + b); \\
y_i &\sim f(W^Ts_i + b'); \\
L_R(y_i, x_i) &\equiv \|x_i - y_i\|_2^2;
\end{align*}
\]

(5)

\[\Theta = \arg \min_{W, b, b'} \sum_{i=1}^{m} L_R(y_i, x_i).\]

where \( x_i \in X \) is the original input vector, \( i = 1, \ldots, m \), and \( q(\cdot|\cdot) \) is the corrupting distribution. The stochastically corrupted vector, \( \bar{x}_i \), is obtained from \( q(\cdot|x_i) \). Generally, \( \bar{x}_i = (1 - \beta)x_i \), and \( \beta \) is the corruption rate in the training phase. The hidden representation, \( s_i \), is mapped from \( \bar{x}_i \) through the network, which consists of an activation function \( f(\cdot) \), parameter matrix \( W \), and parameter vector \( b \). In the
same way, the reconstructed vector, $y_i$, is also mapped from $s_i$ with parameter matrix $W'$ and parameter vector $b'$. Using a loss function, $L_R(\cdot, \cdot)$, we learn these parameters to minimize the total reconstruction error of $\{y_i\}$ and $\{x_i\}$.

Here, our proposed deep framework contains $L$ layers of nonlinear transformation, and the output $s_i$ of the middle layer is usually used as a representation vector that corresponds to $x_i$ [Okura et al., 2017], thus $s_i = \Phi_{\text{rep}}(\Theta; x_i)$. Then $d_{ij}$ is captured in embedding space. In this work, we define $d_{ij} = \Phi_{\text{Sim}}(s_i, s_j) = 1 - s_i^T s_j$. However, $s_i$ in (5) only holds the information of $x_i$. As shown in Figure 4, it is expected that the distance between two representation vectors $s_i$ and $s_j$, i.e., $d_{ij}$, is smaller, if $x_j$ is more similar to $x_i$ compared with $x_k$ (which is evaluated by $z_{ij} > z_{kj}$). This is just the weak supervision from selected prototypes via $Z$. For this end, we add a triplet loss $L_T$ to the objective as follows:

$$L_T(x_i, x_j, x_k) = \log(1 + \exp(d_{ij} - d_{kj}));$$

In essence, the loss function $L_T$ is a penalty function for sample similarity to correspond to categorical similarity.

In addition, as observed in (2), the assignment cost $L_E$ should be as small as possible.

$$L_E(x_i, x_j) = d_{ij} z_{ij};$$

Consequently, the total loss function $L_A(X)$ is rewritten as follows:

$$L_A(X) = \sum_{i=1}^{m} L_R(y_i, x_i) + \alpha_1 \sum_{j=1}^{m} \sum_{i=1}^{m} L_E(x_i, x_j)$$

Reconstruction loss

$$+ \alpha_2 \sum_{(i,j,k) \in T} L_T(x_i, x_j, x_k);$$

Assignment loss in (2)

$$\text{Self-supervised triplet loss}$$

where $\alpha_1$ and $\alpha_2$ are two hyperparameters to balance three loss terms. $T$ is the set of triplets, which is constructed by associating each positive pair in the minibatch with a semihard negative sample. Specifically, $x_i$ and $x_k$ denote the positive and negative samples of $x_i$ in a triplet, and

$$i^* = \arg \max_{i \in \{1, \ldots, m\}} z_{ij};$$

$$k^* = \arg \min_{k \in \{1, \ldots, m\}} z_{kj}.$$
Algorithm 1 The implementation of SDLA

Input: X, λ1, λ2, ρ, α. Initial.: k ← 0, D(k).
1: repeat
   2: Initial. t ← 0, Z(t) = ... choose the same dissimilarity metric with DS3.
   3: while not converged do
      4: Update Z(t+1) according to (14);
      5: Update Z1(t+1) according to (15);
      6: Update D2(t+1) according to (16);
      7: Δ1(t+1) ← Δ1(t) + η(Z(t+1) − Z1(t+1));
      8: Δ2(t+1) ← Δ2(t) + η(Z(t+1) − Z2(t+1));
      9: δ3(t+1) ← δ3(t) + η(1T Z(t+1) − 1T);
      10: η(t+1) ← η(t);
      11: t ← t + 1;
   12: end while
   13: Z(k) ← Z(t);
   14: Update Θ in deep architecture (10);
   15: D ← d3 = ΦSim(Φrep(Θ; x1), Φrep(Θ; x2));
   16: D(k+1) = D(k) + αD; // Update with memory
   17: k ← k + 1;
18: until Convergence criterion satisfied

Output: Optimal solution Z* = Z(k).

where Z in (14) is computed by equating the partial derivative of (13) with respect to Z to zero.

\[
Z_1 = \arg \min_{Z_1} \mathcal{L}(Z_1) = \Gamma_{\lambda_1} \eta^{-1}(Z + \frac{1}{\eta} \Delta_1) \tag{15}
\]

\[
Z_2 = \arg \min_{Z_2} \mathcal{L}(Z_2) = \max(\mathcal{S}_{\lambda_2} \eta^{-1}(Z + \frac{1}{\eta} \Delta_2), 0) \tag{16}
\]

where \(\mathcal{S}\) and \(\mathcal{S}\) are singular value soft-thresholding and shrinkage-thresholding operator, respectively. In detail, for any matrix A and parameter γ, the form of analytic solution for \(\mathcal{S}\) is as follows.

\[
\mathcal{S}(A) = \text{sign}(A) \max(|A| - \gamma, 0)
\]

Then, we have the definition of \(\Gamma\) as

\[
\Gamma_{\gamma}(A) = U \mathcal{S}(\Lambda) V^T \tag{18}
\]

where \(A = U \Lambda V^T\) is the singular value decomposition.

4 Implementation Framework

In summary, Alg. 1 shows the steps of detailed implementation of the SDLA model in (1). The algorithm should not be terminated until the change of objective value is smaller than a pre-defined threshold (10^−1 in our experiments). In addition, we initialize \(D\) by Euclidean distance.

5 Experimental Results and Analysis

In this section, we evaluate the performance of SDLA for selecting prototypes on several illustrative problems.
Table 1: Clustering accuracy (AC) (%) of different methods on TDT2 Corpus.

<table>
<thead>
<tr>
<th># cluster</th>
<th>AP</th>
<th>Kmeans</th>
<th>Kmedoids</th>
<th>SC</th>
<th>LSC</th>
<th>NSHLRR</th>
<th>DS3</th>
<th>Dis-SDLA</th>
<th>SDLA</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>77.18±1.10</td>
<td>57.74±1.39</td>
<td>96.72±0.90</td>
<td>79.34±2.12</td>
<td>65.31±2.09</td>
<td>86.65±1.28</td>
<td>96.72±1.55</td>
<td>95.39±1.95</td>
<td>97.06±0.83</td>
</tr>
<tr>
<td>3</td>
<td>68.28±1.91</td>
<td>62.74±2.80</td>
<td>50.96±2.14</td>
<td>78.11±4.12</td>
<td>94.36±1.48</td>
<td>89.35±2.33</td>
<td>97.45±0.96</td>
<td>96.23±1.15</td>
<td>97.72±1.01</td>
</tr>
<tr>
<td>4</td>
<td>63.95±2.79</td>
<td>64.47±1.95</td>
<td>65.79±1.65</td>
<td>86.65±1.03</td>
<td>71.84±2.84</td>
<td>82.62±1.93</td>
<td>99.47±0.67</td>
<td>97.91±0.75</td>
<td>98.12±0.74</td>
</tr>
<tr>
<td>5</td>
<td>60.41±1.39</td>
<td>45.11±3.65</td>
<td>60.27±2.17</td>
<td>68.59±3.70</td>
<td>71.95±1.03</td>
<td>78.35±1.27</td>
<td>71.80±2.04</td>
<td>72.82±3.09</td>
<td>76.55±2.82</td>
</tr>
<tr>
<td>6</td>
<td>63.52±1.69</td>
<td>66.05±2.31</td>
<td>69.48±3.95</td>
<td>74.73±1.03</td>
<td>87.68±1.43</td>
<td>74.09±2.38</td>
<td>74.93±1.76</td>
<td>73.21±1.79</td>
<td>80.36±1.18</td>
</tr>
<tr>
<td>7</td>
<td>65.89±1.48</td>
<td>56.65±1.40</td>
<td>53.34±3.63</td>
<td>62.05±2.70</td>
<td>65.78±3.75</td>
<td>70.77±2.27</td>
<td>70.65±1.66</td>
<td>70.23±1.65</td>
<td>76.22±1.16</td>
</tr>
<tr>
<td>8</td>
<td>52.37±1.11</td>
<td>68.49±1.49</td>
<td>58.06±3.95</td>
<td>71.66±2.34</td>
<td>70.02±2.58</td>
<td>71.69±1.22</td>
<td>70.92±3.75</td>
<td>71.29±2.24</td>
<td>78.18±1.51</td>
</tr>
<tr>
<td>9</td>
<td>59.78±3.70</td>
<td>68.51±2.89</td>
<td>65.95±2.96</td>
<td>69.30±2.55</td>
<td>69.66±1.74</td>
<td>70.13±1.14</td>
<td>69.45±3.26</td>
<td>69.10±3.84</td>
<td>75.50±2.25</td>
</tr>
<tr>
<td>10</td>
<td>60.11±1.81</td>
<td>52.83±3.24</td>
<td>58.41±2.92</td>
<td>76.81±1.35</td>
<td>72.50±1.19</td>
<td>79.17±1.25</td>
<td>82.51±1.61</td>
<td>81.33±1.47</td>
<td>86.68±1.35</td>
</tr>
</tbody>
</table>

Figure 5: 1-NN classification results of different prototype selection methods on different datasets. (a) Fifteen Scene Categories dataset. (b) USPS digit dataset.

Figure 6: Percentage of outliers among the prototypes selected by different methods as a function of the fraction of outliers.

Figure 7: Percentage of outliers among the prototypes selected by Dis-SDLA as a function of hyperparameter $\nu$.

5.3 Robustness to Outliers

To evaluate the performance of SDLA for rejecting outlier(s), we form a dataset of 9000 images, of which $1 - \theta$ fraction are randomly selected from the Extended YaleB face database, and the remaining, corresponding to outliers, are random images downloaded from the internet. For $\theta \in \{20\%, 30\%, 40\%, 50\%\}$, we run SDLA as well as Kmedoids, AP, SDS, SMRS, DS3, SSDS and Dis-SDLA to select roughly 300 prototypes from the dataset. Figure 6 shows the percentage of outliers among the selected prototypes as we increase the number of outliers in the dataset. Obviously, SDLA results in less outliers compared with other methods.

6 Conclusion

In this work, we introduced a SDLA framework to select representative and discriminative prototypes. Promising experimental results show the effectiveness of SDLA.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant 2016YFB0800404 and the National Natural Science Foundation of China under Grants 61532005, 61332012, and 61572068, and in part by the Fundamental Research Funds for the Central Universities under Grant 2017JJS056.
References


[Cai et al., 2005] Deng Cai, Xiaofei He, and Jiawei Han. Document clustering using locality preserving indexing. TPAMI, 17(12):1624–1637, 2005.


