Revisiting and improving semi-supervised learning: a large dimensional approach

Xiaoyi Mai, Romain Couillet

REVISITING AND IMPROVING SEMI-SUPERVISED LEARNING: A LARGE DIMENSIONAL APPROACH

Xiaoyi Mai [1], *Romain Couillet* [1,2]

[1]CentraleSupélec, University Paris-Saclay and [2]GIPSA-lab, University of Grenoble-Alpes

ABSTRACT

The recent work [1] shows that in the big data regime (i.e., numerous high dimensional data), the popular semi-supervised graph regularization, known as semi-supervised Laplacian regularization, fails to effectively extract information from unlabelled data. In response to this problem, we propose in this article an improved approach based on a simple yet fundamental update of the classical method. The effectiveness of the former is supported by both asymptotic results and simulations on finite data sets.

Index Terms— semi-supervised learning, large dimensional statistics, random matrix theory.

1. INTRODUCTION

Semi-supervised learning (SSL) [2, 3] is a learning approach that employs both labelled and unlabelled data. As data labelling process is often quite expensive both in time and human resources, SSL aims particularly to improve learning accuracy by using a large amount of unlabelled data, in conjunction with a small set of labelled data. The SSL problem is however theoretically intriguing. Indeed, while combining both labelled and unlabelled data should logically allow SSL to consistently outperform supervised and unsupervised learning, SSL methods have been repeatedly observed to fail to meet this basic requirement in practice [4, 5, 6, 7]. In this article, as a follow up [1], we claim that one explanation for this important inconsistency is rooted in a fundamentally erroneous "finite dimensional intuition", which is however naturally resolved when taking a large dimensional perspective on the SSL problem.

Among the popular SSL methods known to suffer from the aforementioned impairment, semi-supervised Laplacian regularization [8, 9] is a graph-based approach with underlying connections to label propagation [10], random walk [11] and electrical network analogs [12]. Although driven by a straightforward, natural reasoning to learn from the data graph in a semi-supervised manner (as presented in Subsection 2), the Laplacian regularization was shown in [1] to have a negligible unlabelled data learning rate in high dimensions, as a direct consequence of the *distance concentration* phenomenon in large dimensions [13, 14, 15]. The inefficiency of the Laplacian regularization with respect to unlabelled data causes it to be outperformed by its purely unsupervised counterpart, spectral clustering [16], on the same high dimensional datasets with abundant unlabelled data. On account of this crucial remark, we introduce in this paper a fundamental improvement of the methods via a mere update of the Laplacian regularization matrix. This in turn allows for recovering asymptotic consistency as the number of either labelled or unlabelled data increases. The proposed updated algorithm is shown additionally to surpass spectral clustering, irrespectively of the number of unlabelled data, as opposed to the classical Laplacian regularization.

In this introductory paper, for lack of space and for readability, we focus on presenting the updated algorithm and on proving its advantage under a basic large dimensional data model, along with validating simulations, leaving the detailed discussions and advanced data modelling to a longer version. The remainder of the article is precisely organized as follows: Section 2 starts by introducing the classical graph-based semi-supervised learning approach, where we recall the essential results of [1] demonstrating a negligible contribution from unlabelled data in high dimensional classification with Laplacian regularization. We then introduce the new regularization approach in Section 3, along with the theoretical basis justifying its usage. Experimental results on finite datasets are also provided to confirm the theoretical analysis in Section 4, evidencing the performance gain of the proposed method over Laplacian regularization as well as over spectral clustering.

2. BACKGROUND

2.1. Semi-Supervised Laplacian regularization

Consider a set of n data vectors $X = [x_1, \ldots, x_n] \in \mathbb{R}^{p\times n}$ of dimension p, belonging to one of the similarity classes $\mathcal{C}_1$ or $\mathcal{C}_2$. The dataset is further divided in two subsets $X = \begin{bmatrix} X_{[l]} & X_{[u]} \end{bmatrix}$ with 'l' and 'u' respectively standing for 'labelled' and 'unlabelled'. That is, we dispose of a labeling vector $y_{[l]} \in \{-1, 1\}^{n_{[l]}}$ (-1 if $x_i \in \mathcal{C}_1$, 1 if $x_i \in \mathcal{C}_2$) for the data vectors x_i composing $X_{[l]}$, while the classes of the data

Couillet's work is supported by the GSTATS UGA IDEX DataScience chair and the ANR RMT4GRAPH Project (ANR-14-CE28-0006).

in $X_{[u]}$ remain to be determined. There are $n_{[l]}$ elements in $X_{[l]}$ and $n_{[u]}$ in $X_{[u]}$.

The data vectors x_i are viewed as nodes on a graph, connected by non-negative weights reflecting their closeness (i.e., similarity), thus giving rise to a weight matrix W of the form

$$W = \{w_{ij}\}_{i,j=1}^n = \left\{ h\left(\frac{1}{p}\|x_i - x_j\|^2\right)\right\}_{i,j=1}^n \quad (1)$$

for some decreasing non-negative function h. The (connectivity) degree of the i-th data point is given by $d_i = \sum_{j=1}^n w_{ij}$, and the diagonal matrix $D \in \mathbb{R}^{n\times n}$ having d_i as its diagonal elements is called the degree matrix.

Graph-based learning methods are built upon a smoothness assumption, stating that data points x_i, x_j connected with a large weight w_{ij} tend to belong to the same group. This suggests the existence of a target class function f that varies little between highly connected data points. The smoothness condition over the graph is usually incorporated in a smoothness penalty term:

$$\mathcal{Q}(f) = \frac{1}{2}\sum_{,ij=1}^n w_{ij}(f_i - f_j)^2 = f^\mathsf{T} L f \quad (2)$$

where $L = D - W$. The matrix L is referred to as the graph Laplacian, or more precisely the unnormalized Laplacian, to be distinguished from the normalized Laplacian $L_s = I_n - D^{-1/2}WD^{-1/2}$, the random walk normalized Laplacian $L_r = I_n - D^{-1}W$, etc. Substituting L with L_s or L_r in $\mathcal{Q}(f)$ gives rise to differently normalized smoothness penalties.

Noting that $\mathcal{Q}(f)$ is minimized to zero at the meaningless solution $f = 1_n$, the unsupervised graph learning method, spectral clustering, consists in finding the smoothest f orthogonal to 1_n, leading to the optimization

$$\min_{f\in\mathbb{R}^n} f^\mathsf{T} L f \quad s.t. \quad \|f\| = 1 \text{ and } f^\mathsf{T} 1_n = 0 \quad (3)$$

the solution of which is easily shown to be the eigenvector associated with the second smallest eigenvalue of L. In the semi-supervised setting, part of data samples are pre-labelled, indicating a prior knowledge on the labelled part $f_{[l]}$ of the target function f, if we write $f = \begin{bmatrix} f_{[l]} & f_{[u]}\end{bmatrix}$. Semi-supervised Laplacian regularization consists in filling the unlabelled $f_{[u]}$ by maximizing the overall smoothness in f as

$$\min_{f\in\mathbb{R}^n} f^\mathsf{T} L f \quad s.t. \quad f_{[l]} = y_{[l]}. \quad (4)$$

As the w_{ij}'s are non-negative, the optimization problem is strongly convex with unique solution

$$f_{[u]} = -L_{[uu]}^{-1} L_{[ul]} f_{[l]}. \quad (5)$$

The final decision step determines the class of unlabelled data x_i according to the sign of f_i ($\mathcal{C}_1$ if $f_i < 0$, $\mathcal{C}_2$ otherwise).

2.2. Failure in high dimensions

As shown above, the Laplacian regularization approach seems to be a perfectly natural way of learning from the graph in a semi-supervised manner, although the non-linear form of (1) prevents formal analysis. The latter difficulty is bypassed in the random matrix-based analysis of [1] which reveals that for comparably large n, p, the classification performance is only driven by the labelled data, implying the *inefficiency* of Laplacian regularization in utilizing unlabelled data. To focus on the message of the inefficient unlabelled data learning without distracting notations, we recall the results of [1] for a simplified, yet sufficiently expressive, model and refer the interested readers to [1] for further details.

Assumption 1. *Data samples $x_1, \ldots, x_n$ are i.i.d. observations from a generative model such that, for $k \in \{1, 2\}$, $\mathbb{P}(x_i \in \mathcal{C}_k) = 1/2$, and*

$$x_i \in \mathcal{C}_k \Leftrightarrow x_i \sim \mathcal{N}(\mu_k, I_p).$$

with $\|\mu_2 - \mu_1\| = O(1)$ (with respect to p).

The ratios $c_0 = \frac{n}{p}$, $c_{[l]} = \frac{n_{[l]}}{p}$ and $c_{[u]} = \frac{n_{[u]}}{p}$ are uniformly bounded in $(0, +\infty)$ for arbitrarily large p.

Theorem 1. *Let h be three-times continuously differentiable in a neighborhood of 2. For $k = \{1, 2\}$, $i > n_{[l]}$ (x_i unlabelled), define $\mathcal{P}_i$ as the probability of correctly classifying x_i with Laplacian regularization[1]. Then, under Assumption 1,*

$$\mathcal{P}_i - \Phi\left(\frac{1}{\sqrt{r_{\text{lap}}}}\right) = o_P(1)$$

where $\Phi(u) = \frac{1}{2\pi}\int_{-\infty}^u e^{-\frac{t^2}{2}} dt$ and

$$r_{\text{lap}} = \frac{4}{\|\mu_1 - \mu_2\|^2} + \frac{1}{\|\mu_1 - \mu_2\|^4 c_{[l]}}. \quad (6)$$

Theorem 1 states that the classification accuracy of the Laplacian regularization decreases with r_{lap}, given by (6). While r_{lap} decreases with $c_{[l]}$, it does not with $c_{[u]}$, implying a vanishing unlabelled data learning in high dimensions.

The reason behind this impairment is that, under Assumption 1 that ensures a non-trivial large dimensional classification problem, all w_{ij} converge to $h(2)$, a manifestation of the well-known distance concentration phenomenon in large data. Consequently, the solution $f_{[u]}$ obtained by minimizing $\sum_{i,j=1}^n w_{ij}(f_i - f_j)^2$ is a vector of constant values, up to small fluctuations. Thanks to the existence of class-structured information in the small fluctuations of $f_{[u]}$, the classification has a non-trivial performance. But the vanishingly small amplitude of the class 'signal' on unlabelled data, insignificant when compared to the labelled data 'signal' $f_{[l]}$, causes the asymptotic performance to depend only on the the labelled data ratio $c_{[l]}$. To cope with this problem, an appropriate correction is proposed in the next section.

[1]The probability of correct classification presented here is for the best choice of Laplcian matrices among L,L_s and L_r.

3. PROPOSED METHOD

3.1. Regularization with centered similarities

The proposed method aims to update the semi-supervised Laplacian regularization. As subsequently discussed in Subsection 3.2, this update enables the use of the full set of labelled and unlabelled (large dimensional) data.

Our approach can be interpreted as following the same reasoning as the Laplacian regularization, however with a centered similarity matrix $\hat{W} \in \mathbb{R}^{n\times n}$ of the form

$$\hat{W} = PWP \quad \text{with} \quad P \equiv I_n - \frac{1}{n}1_n 1_n^\mathsf{T}. \tag{7}$$

Note that the centering operation preserves the distance between the inner- and inter-class similarities in the previous graph, in the sense that the average inner-class similarity minus the average inter-class similarity is unchanged after centering. While the relative information between data points remains intact, using $\hat{W}$ as a similarity measure does *a priori* generate a problem: since $\hat{W}$ has positive and negative elements, the optimization of the smoothness penalty is not necessarily convex and might admit no finite solution. This issue is settled by fixing $\|f\|$, so that the optimization reads

$$\min_{f_{[u]}\in\mathbb{R}^{n_{[u]}}} -f^\mathsf{T}\hat{W}f \quad s.t. \quad \|f_{[u]}\|^2 = n_{[u]}e^2. \tag{8}$$

This can be solved by introducing a Laplacian multiplier α to the norm constraint $\|f_{[u]}\|^2 = n_{[u]}e^2$, leading up to

$$f_{[u]} = \left(\alpha I_{n_{[u]}} - \hat{W}_{[uu]}\right)^{-1}\hat{W}_{[ul]}f_{[l]} \tag{9}$$

where α is determined by $\alpha > \|\hat{W}_{[uu]}\|$ and $\|f_{[u]}\|^2 = n_{[u]}e^2$.

Considering α as the tuning parameter for implementation convenience, the method is summarized in Algorithm 1.

Algorithm 1 Semi-Supervised Graph Regularization with Centered Similarities

1: **Input:** Labelled dataset $X_{[l]}$ with labels $y_{[l]}$ and unlabelled dataset $X_{[u]}$; Parameter $\alpha > \|\hat{W}_{[uu]}\|$.
2: **Output:** Classification of unlabelled dataset $X_{[u]}$.
3: Compute the similarity matrix W by (1).
4: Compute the centered similarity matrix $\hat{W}$ by (7).
5: Set $f_{[l]} = y_{[l]}$ and compute the target class scores $f_{[u]}$ of unlabelled data by (9).
6: Classify unlabelled dataset $X_{[u]}$ by the signs of $f_{[u]}$.

3.2. Performance Analysis

To compare the learning efficiency of the proposed algorithm to the classical Laplacian approach, we now study the large dimensional performance of the *centered similarities regularization*, analogous to the results given in Theorem 1 on Laplacian regularization.

Theorem 2. *Let h be three-times continuously differentiable in a neighborhood of 2. For $k = \{1, 2\}$, $i > n_{[l]}$ (x_i unlabelled), let $\mathcal{P}_i$ be the probability of correctly classifying x_i with centered similarities regularization. Then, under Assumption 1,*

$$\mathcal{P}_i - \Phi\left(\frac{1}{\sqrt{r_{\rm ctr}}}\right) = o_P(1)$$

where $\Phi(u) = \frac{1}{2\pi}\int_{-\infty}^{u} e^{-\frac{t^2}{2}}dt$ and $r_{\rm ctr} > 0$ is the solution of the fixed-point equation

$$r_{\rm ctr} = \left(1 - \frac{g^2}{\|\mu_1-\mu_2\|^4 c_{[u]}}\right)^{-1}\left[\frac{4}{\|\mu_1-\mu_2\|^2} + \frac{g^2}{\|\mu_1-\mu_2\|^4 c_{[u]}} + \frac{(1-g)^2}{\|\mu_1-\mu_2\|^4 c_{[l]}}\right] \tag{10}$$

with

$$g = \frac{c_{[u]}e\sqrt{1/(1+r_{\rm ctr})}}{c_{[u]}e\sqrt{1/(1+r_{\rm ctr})} + c_{[l]}}. \tag{11}$$

As observed from (11), $g \in (0, 1)$. Note in particular that $g \to 0$ as $e \to 0$, and we have $r_{\rm ctr} = r_{\rm lap}$ by taking $g = 0$ in (10), indicating that the performance of Laplacian regularization is retrieved by the centered similarities regularization in the limit $e \to 0$.

Observe now from (10) that g reflects the sensibility of the classification performance with respect to $c_{[l]}, c_{[u]}$: a large g implying an emphasized impact of the unlabelled data and a reduced effect of the labelled data. It can then be shown that g is an increasing function of e; the influence of labelled and unlabelled data is thus adjustable through the tuning of e, at the two extremes of which the purely supervised and purely unsupervised learning performances are respectively recovered. Using α as a direct parameter as in Algorithm 1, note that large values of α correspond to small e.

Computing the derivative of $r_{\rm ctr}$ with respect to g in the limits of $g \to +\infty$ and $g \to 0$, we next find that $r'_{\rm ctr} > 0$ as $g \to 0$ and $r'_{\rm ctr} < 0$ in the limit $g \to +\infty$. The optimal performance is therefore achieved at a bounded value of g, where the algorithm realizes a genuine semi-supervised learning, allowing for a performance gain over the Laplacian regularization (which we recall is asymptotically equivalent to supervised learning in the large dimensional regime) and over spectral clustering (i.e., unsupervised learning).

Deducing from (10) that a well defined $r_{\rm ctr} > 0$ exists for any $g \in [0, g_{\rm sup})$ where $g_{\rm sup} = \min\{\|\mu_1-\mu_2\|^4 c_{[u]}, 1\}$, the admissible set of g can only enlarge as the number of data samples grows. Adding this to the obvious fact that $r_{\rm ctr}$ is a strictly decreasing function of both $c_{[l]}, c_{[u]}$ at any fixed value of g, we conclude that with an appropriately chosen e, the performance of the centered similarities regularization is consistently improved as more data sample, whether labelled or unlabelled, are included in the learning process.

4. EXPERIMENTAL VALIDATION

This section provides experimental evidence supporting the proposition of the centered similarities regularization. All versions of Laplacian matrices in Subsection 2.1 are tested for Laplacian regularization and spectral clustering, the optimal α for the centered similarities regularization is searched within the admissible range, so as to report the best results. We first verify the validity of the asymptotic results on finite, not-so-large ($p = 80$), datasets. Figure 1 shows that the empirical accuracy closely matches the theoretical one. As predicted by the theoretical analysis, the performance of the Laplacian regularization barely moves when more unlabelled data are used, demonstrating an inconsequential unlabelled data learning rate. For sufficient unlabelled data, the Laplacian regularization is even surpassed by spectral clustering, which, being unsupervised, treats labelled data as if unlabelled. Meanwhile, the accuracy of the proposed algorithm consistently improves as the number of unlabelled data increases, resulting in a growing performance gap over the classical Laplacian approach, with a maintained advantage over spectral clustering as a benefit of utilizing labelled data.

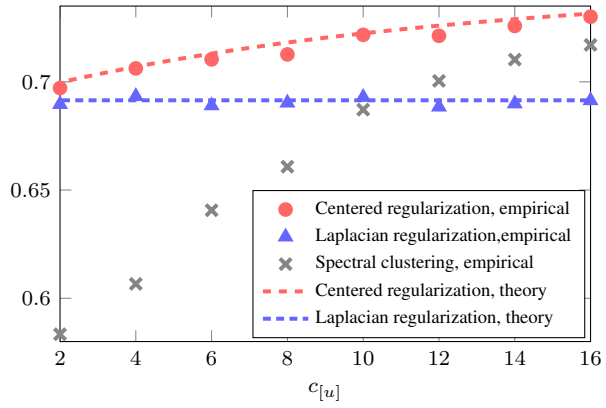

Fig. 1. Empirical and theoretical classification accuracy as a function of $c_{[u]}$ with $c_{[l]} = 2$, $\|\mu_1 - \mu_2\|^2 = 2$, $p = 80$. Graph constructed with $w_{ij} = e^{-\|x_i - x_j\|^2/p}$. Averaged over $50000/n_{[u]}$ iterations.

To compare the proposed method and the traditional approach beyond synthetic data, simulations on real world datasets, here the MNIST database [17], are provided. For a more comprehensive comparison, we perform experiments on commonly used graphs including KNN graphs with various numbers of neighbors $k = \{2^1, \cdots, 2^q\}$, for q the largest integer such that $2^q < n$, and graphs constructed by RBF kernels, i.e., $w_{ij} = e^{-\|x_i - x_j\|^2/\sigma^2}$, with bandwidth σ set to the average data vectors distance. The algorithms accuracies are respectively given for their best performing graph in Figure 2, where a clear advantage is observed for the proposed method. However, the simulation on original MNIST data (top of Figure 2) seems to contradict the statement of negligible unlabelled data contribution for the Laplacian regularization, as the performance of both methods grows with the number of unlabelled data. We conjecture that it is due to the MNIST data being quite easy to separate, without the problem of distance concentration which, we recall, is the main cause behind the unlabelled data learning inefficiency of Laplacian regularization. When the distance concentration phenomenon is noticeable, as in the presence of additional white noise, the predicted insignificance of unlabelled data learning is once again observed (bottom of Figure 2).

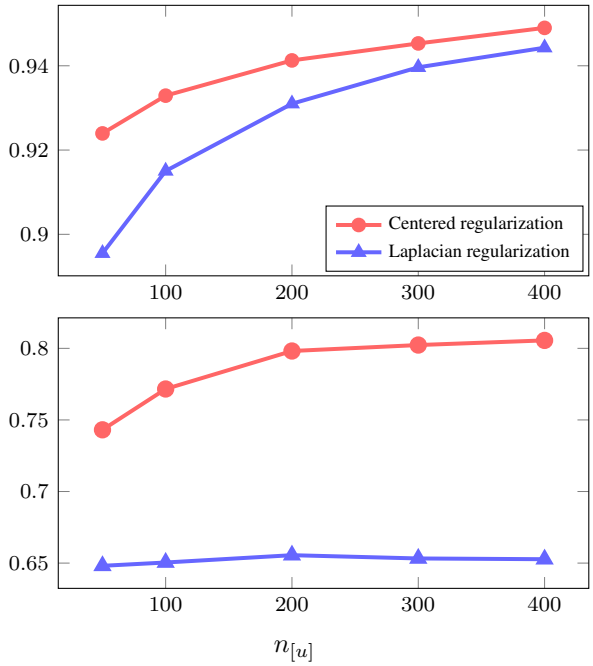

Fig. 2. Classification accuracy on MNIST data (6,9) as a function of $n_{[u]}$ with $n_{[l]} = 10$. Top: pure data. Bottom: noisy data with SNR$= -10$dB. Averaged over 1000 iterations

5. CONCLUDING REMARKS

By identifying and carefully analysing a fundamental 'curse of dimensionality phenomenon' arising on realistic (not necessarily so large) data – here due to the convergence of all data distances –, this work provides for the first time an explanation as well as an elementary corrective answer to the long-standing problem of unlabelled data inefficiency in graph-based semi-supervised learning methods, with compelling simulation evidence on real-world data.

This, along with parallel contributions on large dimensional spectral clustering, simple random neural nets and other supervised learning tools, forcefully suggests that large dimensional statistics, and notably random matrix theory, are major key enablers to future machine learning understanding and design in the big data era.

6. REFERENCES

[1] Xiaoyi Mai and Romain Couillet, "A random matrix analysis and improvement of semi-supervised learning for large dimensional data," *arXiv preprint arXiv:1711.03404*, 2017.

[2] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, *Semi-supervised learning*, MIT press, 2006.

[3] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling, "Semi-supervised learning with deep generative models," in *Advances in neural information processing systems*, 2014, pp. 3581–3589.

[4] Behzad M Shahshahani and David A Landgrebe, "The effect of unlabeled samples in reducing the small sample size problem and mitigating the hughes phenomenon," *IEEE Transactions on Geoscience and remote sensing*, vol. 32, no. 5, pp. 1087–1095, 1994.

[5] Fabio Gagliardi Cozman, Ira Cohen, and M Cirelo, "Unlabeled data can degrade classification performance of generative classifiers.," in *Flairs conference*, 2002, pp. 327–331.

[6] Shai Ben-David, Tyler Lu, and Dávid Pál, "Does unlabeled data provably help? worst-case analysis of the sample complexity of semi-supervised learning.," in *COLT*, 2008, pp. 33–44.

[7] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow, "Realistic evaluation of deep semi-supervised learning algorithms," in *Advances in Neural Information Processing Systems*, 2018, pp. 3239–3250.

[8] Xiaojin Zhu, Zoubin Ghahramani, John Lafferty, et al., "Semi-supervised learning using gaussian fields and harmonic functions," in *ICML*, 2003, vol. 3, pp. 912–919.

[9] Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf, "Learning with local and global consistency," *Advances in neural information processing systems*, vol. 16, no. 16, pp. 321–328, 2004.

[10] Xiaojin Zhu and Zoubin Ghahramani, "Learning from labeled and unlabeled data with label propagation," Tech. Rep., Citeseer, 2002.

[11] Martin Szummer Tommi Jaakkola and Martin Szummer, "Partially labeled classification with markov random walks," *Advances in neural information processing systems (NIPS)*, vol. 14, pp. 945–952, 2002.

[12] Morteza Alamgir and Ulrike V Luxburg, "Phase transition in the family of p-resistances," in *Advances in Neural Information Processing Systems*, 2011, pp. 379–387.

[13] Charu C Aggarwal, Alexander Hinneburg, and Daniel A Keim, "On the surprising behavior of distance metrics in high dimensional space," in *International conference on database theory*. Springer, 2001, pp. 420–434.

[14] Damien Francois, Vincent Wertz, and Michel Verleysen, "The concentration of fractional distances," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 7, pp. 873–886, 2007.

[15] Fabrizio Angiulli, "On the behavior of intrinsically high-dimensional spaces: Distances, direct and reverse nearest neighbors, and hubness," *Journal of Machine Learning Research*, vol. 18, no. 170, pp. 1–60, 2018.

[16] Ulrike Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[17] Yann LeCun, Corinna Cortes, and Christopher JC Burges, "The mnist database of handwritten digits," 1998.