Data based construction of kernels for classification

Hrushikesh N. Mhaskar, Sergei V. Pereverzyev, Vasyl Yu. Semenov and Evgeniya V. Semenova

Abstract This paper is an announcement for our longer paper in preparation. Traditional kernel based methods utilize either a fixed kernel or a combination of judiciously chosen kernels from a fixed dictionary. In contrast, we construct a data-dependent kernel utilizing the components of the eigen-decompositions of different kernels constructed using ideas from diffusion geometry, and use a regularization technique with this kernel with adaptively chosen parameters. In this paper, we illustrate our method using the two moons dataset, where we obtain a zero test error using only a minimal number of training samples.

1 Introduction

The problem of learning from labeled and unlabeled data (semi-supervised learning) has attracted considerable attention in recent years. A variety of machine learning algorithms use Tikhonov single penalty or multiple penalty schemes for regularizing with different approaches to data analysis. Many of these are kernel based algorithms that provide regularization in Reproducing Kernel Hilbert Spaces (RKHS). The problem of finding a suitable kernel for learning a real-valued function by regu-

Hrushikesh N. Mhaskar
Institute of Mathematical Sciences, Claremont Graduate University, Claremont, CA91711, USA, e-mail: hmhaska@gmail.com

Sergei V. Pereverzyev
Johann Radon Institute for Computational and Applied Mathematics, Linz, Austria, e-mail: sergei.pereverzyev@oeaw.ac.at

Vasyl Yu. Semenov
R&D department, Scientific and Production Enterprise “Delta SPE”, Kiev, Ukraine e-mail: vasyl.delta@gmail.com

Evgeniya V. Semenova
Institute of Mathematics of NASU, Kiev, Ukraine e-mail: senenovaevgen@gmail.com

larization is considered, in particular, in the papers [16], [17] (see also the references therein), where different approaches were proposed. All the methods mentioned in these papers deal with some set of kernels that appear as a result of parametrization of classical kernels or linear combination of some functions. Such approaches lead to the problem of multiple kernel learning. In this way, the kernel choice problem is somehow shifted to the problem of a description of a dictionary of predefined kernels, on which multiple kernel learning is performed.

In the present paper we propose an approach to **construct a kernel** directly from observed data rather than choosing one from a given kernel dictionary in advance. The approach uses ideas from diffusion geometry (see, e.g. [1, 2, 3, 5, 11]), where the eigenvectors of the graph Laplacian associated to the unlabeled data are used to mimic the geometry of the underlying manifold that is usually unknown. The literature on this subject is too large to be cited extensively. The special issue [7] of *Applied and Computational Harmonic Analysis* is devoted to an early review of this subject. Most relevant to the current paper are the papers [5], [6], where the graph Laplacian associated with the data has been used to form additional penalty terms in a multi-parameter regularization functional of Tikhonov type. In contrast to [5], [6], we use eigenvectors and eigenfunctions of the corresponding family of graph Laplacians (rather than a combination of these graph Laplacians) to construct a data-dependent kernel that directly generates an RKHS.

In Section 2, we summarize some known theoretical facts relevant to our paper. Our numerical algorithm is described in Section 3. In Section 4, we present the experimental results with the two moons data set.

2 Background

The subject of diffusion geometry seeks to understand the geometry of the data $\{x_i\}_{i=1}^n \subset \mathbb{R}^D$ drawn randomly from an unknown probability distribution μ, where D is typically a large ambient dimension. It is assumed that the support of μ is a smooth sub-manifold of $\mathbb{R}^D$ having a small manifold dimension d. The theory works with eigenfunctions of the Laplace-Beltrami operator of this manifold. However, since the manifold is unknown, one needs to approximate the Laplace-Beltrami operator. One way to do this is using a graph Laplacian as follows.

For $\varepsilon > 0$ and $x, y \in \mathbb{R}^D$, let

$$W^{\varepsilon}(x,y) := \exp\left(-\frac{\|x-y\|^2}{4\varepsilon}\right). \tag{1}$$

We consider the points $\{x_i\}_{i=1}^n$ as vertices of an undirected graph with the edge weight between x_i and x_j given by $W^{\varepsilon}(x_i,x_j)$, thereby defining a weighted adjacency matrix, denoted by $\mathbf{W}^{\varepsilon}$. We define $\mathbf{D}^{\varepsilon}$ to be the diagonal matrix with the i-th entry on the diagonal given by $\sum_{j=1}^{n} W^{\varepsilon}(x_i,x_j)$. The graph Laplacian is defined by the matrix

$$\mathbf{L}^{\varepsilon} = \frac{1}{n}\left\{\mathbf{D}^{\varepsilon} - \mathbf{W}^{\varepsilon}\right\}. \tag{2}$$

We note that the eigenvalues of $\mathbf{L}^{\varepsilon}$ are all real and non-negative, and therefore, can be ordered as

$$0 = \lambda_1^{\varepsilon} < \lambda_2^{\varepsilon} \le \cdots \le \lambda_n^{\varepsilon}. \tag{3}$$

It is convenient to consider the eigenvector corresponding to λ_k^{ε} to be a function on $\{x_j\}_{j=1}^n$ rather than a vector in $\mathbb{R}^n$, and denote it by ϕ_k^{ε}, thus,

$$\lambda_k^{\varepsilon}\phi_k^{\varepsilon}(x_i) = \sum_{j=1}^n L_{i,j}^{\varepsilon}\phi_k^{\varepsilon}(x_j) = \frac{1}{n}\left(\phi_k^{\varepsilon}(x_i)\sum_{j=1}^n W^{\varepsilon}(x_i,x_j) - \sum_{j=1}^n W^{\varepsilon}(x_i,x_j)\phi_k^{\varepsilon}(x_j)\right), \tag{4}$$
$$i = 1,\ldots,n.$$

Since the function W^{ε} is defined on the entire ambient space, one can extend the function ϕ_k^{ε} to the entire ambient space using (4) in an obvious way (the Nyström extension). Denoting this extended function by Φ_k^{ε}, we have (cf. (4), [19])

$$\Phi_k^{\varepsilon}(x) = \frac{\sum_{j=1}^n W^{\varepsilon}(x,x_j)\phi_k^{\varepsilon}(x_j)}{\sum_{j=1}^n W^{\varepsilon}(x,x_j) - n\lambda_k^{\varepsilon}}, \tag{5}$$

for all $x \in \mathbb{R}^D$ for which the denominator is not equal to 0. The condition that the denominator of (5) is not equal to 0 for any x can be verified easily for any given ε. The violation of this condition for a particular k can be seen as a sign that for a given amount n of data the approximations of the eigenvalue λ_k of the corresponding Laplace-Beltrami operator by eigenvalues λ_k^{ε} cannot be guaranteed with a reasonable accuracy.

The convergence of the extended eigenfunctions Φ_k^{ε}, restricted to a smooth manifold X, to the actual eigenfunctions of the Laplace-Beltrami operator on X is described in [4, Theorem 2.1].

3 Numerical algorithms for semi-supervised learning

The approximation theory utilizing the eigen-decomposition of the Laplace-Beltrami operator is well developed, even in greater generality than this setting, in [12, 10, 13, 14, 9]. In practice, the correct choice of ε in the approximate construction of these eigenvalues and eigenfunctions is a delicate matter that affects greatly the performance of the kernel based methods based on these quantities. Some heuristic rules for choosing ε have been proposed in [11, 8]. These rules are not applicable universally; they need to be chosen according to the data set and the application under consideration.

In contrast to the traditional literature, where a fixed value of ε is used for all the eigenvalues and eigenfunctions, we propose in this paper the construction of a

Algorithm 1 Algorithm for kernel ridge regression with the constructed kernel (7)

```
Given data $\{x_i\}_{i=1}^n \in X$,   $\{x_i, y_i\}_{i=1}^m$ are the labeled examples; $y = \{y\}_{i=1}^m$.
Introduce the grid for parameter $\alpha$: $\alpha_k = p^k, k = 1, 2, \ldots, N$
Calculate Gram matrix $\widehat{K}_m$ consisting of the sub-matrix $\{K_n(x_i, x_j)\}_{i,j=1}^m$ defined by (7) in labeled points
for k=1:N do
    Calculate $C_{\alpha_k}$ as
        $C_{\alpha_k} = (\alpha_k I + \widehat{K}_m)^{-1} y,$
    Find the $\alpha_{min}$ such that $\|\widehat{K}_m C_{\alpha_k} - y\|$ is minimized.
end for
The decision-making function is
    $f_n^*(x) = \sum_{i=1}^{m} (C_{\alpha_{min}})_i K_n(x, x_i).$
```

kernel of the form

$$K_n(x,t) = \sum_k (n\lambda_k^{\varepsilon_{j_k}})^{-1} \Phi_k^{\varepsilon_{j_k}}(x) \Phi_k^{\varepsilon_{j_k}}(t); \tag{6}$$

i.e., we select the eigenvalues and the corresponding eigenfunctions from different kernels of the form W^{ε} to construct our kernel. We note again that in contrast to the traditional method of combining different kernels from a fixed dictionary, we are constructing a single kernel using eigenvectors and eigenfunctions of different kernels from a dictionary.

Table 1 Results of testing for two moons dataset

n:	m	Error
50	2	0%
50	4	0%
50	6	0%
30	2	17%
30	4	8%
30	6	0%
10	2	38%
10	4	10%
10	6	2%

Our rule for selecting the ε_{j_k}'s is based on the well-known quasi-optimality criterion [18] that is one of the simplest and oldest, but still a quite efficient strategy for choosing a regularization parameter. According to that strategy, one selects a suitable value of ε (i.e. the regularization parameter) from a sequence of admissible values $\{\varepsilon_j\}$, which usually form a geometric sequence, i.e. $\varepsilon_j = \varepsilon_0 q^j, j = 1, 2, \ldots, M$; $q < 1$. We propose to employ the quasi-optimality criterion in the context of the approximation of the eigenvalues of the Laplace-Beltrami operator. Then by analogy to [18] for each particular k we calculate the sequence of approximate eigenvalues $\lambda_k^{\varepsilon_j}, j = 1, 2, \ldots, M$, and select $\varepsilon_{j_k} \in \{\varepsilon_j\}$ such that the differences $|\lambda_k^{\varepsilon_j} - \lambda_k^{\varepsilon_{j-1}}|$ attain their minimal value at $j = j_k$.

Since the size of the grid of ε_j is difficult to be estimated beforehand and, at the same time, has a strong influence on the performance of the method, we propose the following strategy for the selection of the grid size M. We note that the summation in formula (6) has to be done for indices k for which the corresponding eigenvalue $\lambda_k = \lambda_k^{\varepsilon_{j_k}}$ is non-zero. It is also known that the first eigenvalue

$\lambda_1^{\varepsilon_j} = 0$. To prevent the other $\lambda_k^{\varepsilon_j}$ from becoming too close to zero with the decreasing of ε_j, we propose to stop continuation of the sequence ε_j as soon as the value of $\lambda_2^{\varepsilon_M}$ becomes sufficiently small. So, the maximum grid size M is the smallest integer for which $\lambda_2^{\varepsilon_M} < \lambda_2^{(thr)}$, where $\lambda_2^{(thr)}$ is some estimated threshold. Taking the abovementioned into account, we also replace the formula for the kernel calculation (6) by the kernel

$$K_n(x,t) = 1 + \sum_{k=2}^{n} (n\lambda_k^{\varepsilon_{j_k}})^{-1} \Phi_k^{\varepsilon_{j_k}}(x) \Phi_k^{\varepsilon_{j_k}}(t); \tag{7}$$

Algorithm 1 described above uses the constructed kernel (7) in kernel ridge regression from labeled data. The regression is performed in combination with a discrepancy based principle for choosing the regularization parameter α. More details can be found in [15].

4 Experimental results

In this section we consider classification of the two moons dataset that can be seen as the case $D = 2$, $d = 1$. The software and data were borrowed from bit.ly/2D3uUCk. For the two moons dataset we take $\{x_i\}_{i=1}^n$ with $n = 50, 30, 10$ and subsets $\{x_i\}_{i=1}^m \subset \{x_i\}_{i=1}^n$ with $m = 2, 4, 6$ labeled points. The goal of semi-supervised data classification problems is to assign correct labels for the remaining points $\{x_i\}_{i=1}^n \setminus \{x_i\}_{i=1}^m$.

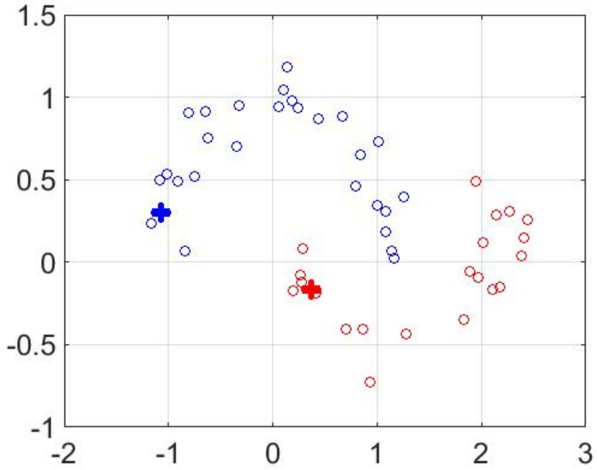

Fig. 1 Classification of "two moons" dataset with extrapolation region. The values of parameters are m=2, $\varepsilon_0 = 1$, q=0.9, $\lambda_2^{(thr)} = 10^{-6}$.

For every dataset (defined by the pair (n,m)) we performed 10 trials with randomly chosen labeled examples.

As follows from the experiments, the accuracy of the classification is improving with the growth of the number of unlabeled points. In particular, for $n \geq 50$, to label all points without error, it is enough to take only one labeled point for each of two classes ($m = 2$). At the same time, if the set of unlabeled points is not big enough, then for increasing the accuracy of prediction we should take more labeled points. The result of the classification for the two moons dataset with $m = 2$ as well as the corresponding plot of selected ε are shown in Figures 1 and 2. The big crosses correspond to the labeled data and other points are colored according to the constructed predictors. The parameters' values were m=2, $\varepsilon_0 = 1$, q=0.9, $\lambda_2^{(thr)} = 10^{-6}$.

The application of the proposed method to other classification problems including automatic gender identification can be found in [15].

Acknowledgements Sergei V. Pereverzyev and Evgeniya Semenova gratefully acknowledge the support of the consortium AMMODIT funded within EU H2020-MSCA-RICE. The research of Hrushikesh Mhaskar is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via 2018-18032000002. The paper has been finalized while the first two co-authors took part in the workshop "On the frontiers of high-dimensional computation" at the MATRIX Research Institute, Creswick, June 2018. The support of MATRIX is gratefully acknowledged.

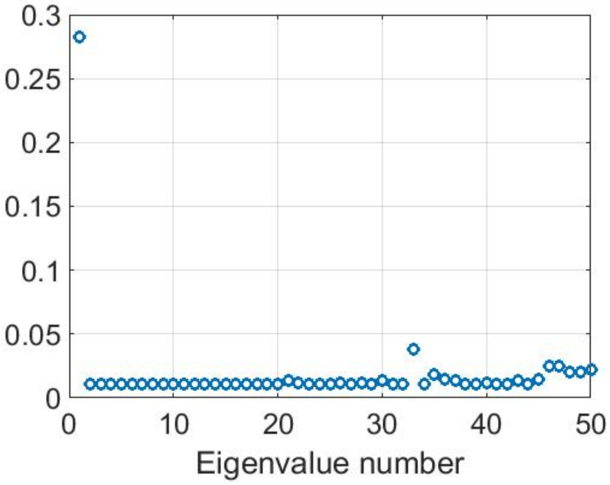

Fig. 2 Plot of adaptively chosen ε for two-moon dataset. The values of parameters are m=2, $\varepsilon_0 = 1$, q=0.9, $\lambda_2^{(thr)} = 10^{-6}$.

References

1. Belkin, M., Matveeva, I., Niyogi P.: Regularization and semi-supervised learning on large graphs. In: Learning theory, pp. 624-638. Springer (2004)
2. Belkin, M., Niyogi P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural computation **15**, 1373–1396 (2003)
3. Belkin, M., Niyogi P.: Semi-supervised learning on Riemannian manifolds. Machine learning **56**, 209–239 (2004)
4. Belkin, M., Niyogi P.: Convergence of Laplacian eigenmaps. Adv. neur. inform. process. 19 **129**, (2007)
5. Belkin, M., Niyogi P., Sindhwani V.: Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. J. Mach. Learn. Res. **7**, 2399–2434 (2006)
6. Bertozzi A.L., Luo Xi., Stuart A.M., Zygalakis K.C.: Uncertainty Quantification in the Classification of High Dimensional Data. In: CoRR (2017) http://arxiv.org/abs/1703.08816
7. Chui, C.K., Donoho D.L.: Special issue: Diffusion maps and wavelets. Appl. Comput. Harm. Anal. **21** (2006)
8. Coifman, R. R. and Hirn, M. J.: Diffusion maps for changing data. Appl. Comput. Harmon. Anal. **36**, 79–107 (2014)
9. Ehler, M., Filbir, F., Mhaskar H.N.: Locally Learning Biomedical Data Using Diffusion Frames. J. Comput. Biol. **19**, 1251–1264 (2012)
10. Filbir, F., Mhaskar H.N.: Marcinkiewicz–Zygmund measures on manifolds. J. Complexity. **27**, 568–596 (2011)
11. Lafon, S.S: Diffusion maps and geometric harmonics. Yale University (2004)
12. Maggioni, M., Mhaskar H.N.: Diffusion polynomial frames on metric measure space. Appl. Comput. Harmon. Anal. **24**, 329–353 (2008)
13. Mhaskar, H.N.: Eignets for function approximation on manifolds. Appl. Comput. Harm. Anal. **29**, 63–87 (2010)
14. Mhaskar, H.N.: A generalized diffusion frame for parsimonious representation of functions on data defined manifolds. Neural Networks **24**, 345–359 (2011)
15. Mhaskar, H.N., Pereverzyev S.V., Semenov, V.Yu., Semenova E.V.: Data based construction of kernels for semi-supervised learning with less labels. RICAM Preprint (2018). https://www.ricam.oeaw.ac.at/files/reports/18/rep18-25.pdf
16. Micchelli, C.A., Mhaskar H.N.: Learning the kernel function via regularization. J.Mach.Learn.Res. **6**, 10127–10134 (2005)
17. Pereverzyev, S.V., Tkachenko, P.: Regularization by the Linear Functional Strategy with Multiple Kernels. Frontiers Appl. Math. Stat. **3**, 1 (2017)
18. Tikhonov, A.N., Glasko, V.B.: Use of regularization method in non-linear problems. Zh. Vychisl. Mat. Mat. Fiz. **5**, 463–473 (1965)
19. von Luxburg U., Belkin, M., Bousquet O.: Consistency of spectral clustering. Ann. Statist. **36**, 555–586 (2008)