NHDC and PHDC: Non-propagating and propagating heat diffusion classifiers

NHDC and PHDC: Non-propagating and Propagating Heat Diffusion Classifiers

Haixuan Yang
Dept. of Comp. Sci. and Eng.
The Chinese Univ. of HK
Hong Kong
Email: hxyang@cse.cuhk.edu.hk

Irwin King
Dept. of Comp. Sci. and Eng.
The Chinese Univ. of HK
Hong Kong
Email: king@cse.cuhk.edu.hk

Michael R. Lyu
Dept. of Comp. Sci. and Eng.
The Chinese Univ. of HK
Hong Kong
Email: lyu@cse.cuhk.edu.hk

***Abstract*— By imitating the way that heat flows in a medium with a geometric structure, we propose two novel classification algorithms, Non-propagating Heat Diffusion Classifier (NHDC) and Propagating Heat Diffusion Classifier (PHDC). In NHDC, an unlabelled data is classified into the class that diffuses the most heat to the unlabelled data after one local diffusion from time 0 to a small time period, while in PHDC, an unlabelled data is classified into the class that diffuses the most heat to the unlabelled data in the propagating effect of the heat flow from time 0 to time t, which means that in PHDC, the heat diffuses infinitely many times from time 0 and each time period is infinitely small. In other words, we measure the similarity between an unlabelled data and a class by the heat amount that the unlabelled data receives from the set of labelled data in the class, and then classify the unlabelled data into the class with the most similarity. Unlike the traditional method, in which the heat kernel is applied to a kernel-based classifier we employ the heat kernel to construct the classifier directly; moreover, instead of imitating the way that the heat flows along a linear or nonlinear manifold, we let the heat flow along a graph formed by the k-nearest neighbors. An important and special feature in both NHDC and PHDC is that the kernel is not symmetric. We show theoretically that PWA (Parzen Window Approach when the window function is a multivariate normal kernel) and KNN are actually special cases of NHDC model, and that PHDC has the ability to approximate NHDC. Experiments show that NHDC performs better than PWA and KNN in prediction accuracy, and that PHDC performs better than NHDC.**

I. Introduction

The heat flow throughout a geometric manifold with initial conditions can be described by the following second order differential equation :

$$\begin{cases} \frac{\partial f}{\partial t} - \Delta f &= 0, \\ f(x,0) &= f_0(x), \end{cases}$$

where $f(x,t)$ is the heat at location x at time t, beginning with an initial distribution of heat given by $f_0(x)$ at time zero. The heat or diffusion kernel $K_t(x,y)$ [1] is a special solution to the heat equation with a special initial condition called the delta function $\delta(x-y)$, which has the following properties: $\delta(x-y)=0$ for $x \neq y$; $\int_{-\infty}^{+\infty} \delta(x-y)dx = 1$. The delta function $\delta(x-y)$ in the heat diffusion setting has the physical meaning – it describes a unit heat source at position y when there is no heat in other positions. Based on this, the heat kernel $K_t(x,y)$ describes the heat distribution at time t diffusing from the initial unit heat source at position y. Since arbitrary initial conditions can be considered as a combination of heat sources with different intensities at different positions, as a consequence of the linearity of the heat equation, the heat kernel can be used to generate the solution to the heat equation according to the following equation

$$f(x,t) = \int_M K_t(x,y) f_0(y) dy.$$

The heat kernel $K_t(x,y)$ can be considered as a generalization of Gaussian density. This is because that when the underlying manifold is a flat n-dimensional Euclidean space, the heat kernel $K_t(x,y)$ has an explicit form

$$(4\pi t)^{-\frac{m}{2}} \exp(-\frac{||x-y||^2}{4t}), \tag{1}$$

which is the same as the Gaussian density. When the geometric manifold varies, the corresponding heat kernel varies and can be considered as the generalization of Gaussian density from flat Euclidean space to general manifold.

Some recent successful applications of heat kernel includes [1], [2] and [3]. In [1], the authors approximate the heat kernel for multinomial family in a closed form, from which great improvements are obtained over the use of Gaussian or linear kernels. In [2], the authors propose the use of discrete diffusion kernel to discrete or categorical data, and show that the simple diffusion kernel on the hypercube can result in good performance for such data. In [3], the authors employ heat kernel to construct weight of a neighborhood graph, and apply it to a non-linear dimensionality reduction algorithm.

Based on the successful applications of the heat kernel on the classification problem, it is natural to explore the use of heat kernel in a wider area where the underlying geometry is unknown or its heat kernel cannot be approximated in the same way as in [1]. To achieve our goal, we represent the underlying geometry by a finite neighborhood graph, instead of approximating the heat kernel in a given geometry. Then we establish a heat diffusion model based on this graph, instead of on the manifold.

The remaining of the paper is organized as follows. In Section II, we establish a heat diffusion model on the graph. In Section III, we present *Non-propagating Heat Diffusion*

Classifier (NHDC) and in Secion IV, we establish *Propagating Heat Diffusion Classifier* (PHDC). In Section V, we interpret the model in more details. Then in Section VI, we describe the connection between NHDC and other models, and the connection between NHDC and PHDC. Moreover, we analyze the difference between the heat kernel proposed in [2] and our heat kernel. In Section VII, we show and discuss our experimental results and conclusions. Section VIII provides the conclusion.

II. Heat Diffusion Model on Graph

First we give our notation for the heat diffusion model on graph. Consider a directed weighted graph $G = (V, E, W)$, where $V = \{v_1, v_2, \ldots, v_n\}$, $E = \{(v_i, v_j) \,|\, \text{there is an edge from } v_i \text{ to } v_j\}$ is the set of all edges, and $W = (w_{ij})$ is the weight matrix. Different from the normal undirected weighed graph, the edge (v_i, v_j) is considered as a pipe that connects to nodes i and j, and the weight w_{ij} is considered as the length of the pipe (v_i, v_j). The value $f_i(t)$ describes the heat at node i at time t, beginning from an initial distribution of heat given by $f_0(i)$ at time zero.

We establish our model as follows. Suppose, at time t, each node i receives $M(i, j, t, \Delta t)$ amount of heat from its neighbor j during a period of Δt. The heat $M(i, j, t, \Delta t)$ should be proportional to the time period Δt and the heat difference $f_j(t) - f_i(t)$. Moreover, the heat flows from node j to node i through the pipe that connects nodes i and j, and therefore the heat diffuses in the pipe in the same way as it does in the m-dimensional Euclidean space as described in Eq. (1). Based on this consideration, we assume that $M(i, j, t, \Delta t) = \alpha \cdot \exp(-\frac{w_{ij}^2}{\beta})(f_j(t) - f_i(t))\Delta t$. As a result, the heat difference at node i between time $t + \Delta t$ and time t will be equal to the sum of the heat that it receives from all its neighbors. This is formulated as

$$f_i(t + \Delta t) - f_i(t) = \sum_{j:(j,i)\in E} \alpha \cdot \exp(-\frac{w_{ij}^2}{\beta})(f_j(t) - f_i(t))\Delta t \tag{2}$$

Note that when $f_i(t) > f_j(t)$, node i receives a negative amount of heat, i.e., it sends out a positive amount of heat.

To find a closed form solution to Eq. (2), we express it as a matrix form:

$$\frac{f(t + \Delta t) - f(t)}{\Delta t} = \alpha H f(t), \tag{3}$$

where $H = (H_{ij})$, and

$$H_{ij} = \begin{cases} -\sum_{k:(k,i)\in E} \exp(-\frac{w_{ik}^2}{\beta}), & j = i; \\ \exp(-\frac{w_{ij}^2}{\beta}), & (j, i) \in E; \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

The matrix H is called as *non-propagating diffusion kernel* in the sense that the heat diffusion process stops after the nodes diffuse their heat to their neighbors. Let $t = 0$, Eq. (3) can be rewritten as

$$f(\Delta t) = (I + \alpha \Delta t H) f(0). \tag{5}$$

Eq. (5) is one closed form solution to Eq. (2) in the setting of non-propagating heat diffusion, where it describes the heat distribution after a time period of Δt from time 0.

Next, we try to find another closed form solution to Eq. (2) in the setting of propagating heat diffusion. In the limit $\Delta t \to 0$, Eq. (3) becomes

$$\frac{d}{dt} f(t) = \alpha H f(t), \tag{6}$$

Solving Eq. (6), we get

$$f(t) = e^{\alpha t H} f(0) = e^{\gamma H} f(0), \tag{7}$$

where $\gamma = \alpha t$, and $e^{\gamma H}$ is defined as

$$e^{\gamma H} = I + \gamma H + \frac{\gamma^2}{2!} H^2 + \frac{\gamma^3}{3!} H^3 + \cdots. \tag{8}$$

The matrix $e^{\gamma H}$ is called as *propagating diffusion kernel* in the sense that the heat diffusion process continues infinitely many times after the nodes diffuse their heat to their neighbors for the first time.

Eq. (7) is the solution to Eq. (2) when we consider propagating heat diffusion. It has a natural property as shown in the following theorem.

Theorem 1: The solution in Eq. (7) has the property of heat preserving.

Based on the two closed form solutions Eq. (5) and Eq. (7), we establish two different classifiers in the next two sections.

III. Non-propagating Heat Diffusion Classifier

Assume that there are c classes, namely, $C_1, C_2, \ldots, C_c$. Let the labelled data set contains M samples, represented by $(\mathbf{x}_i, k_i)$ $(i = 1, 2, \ldots, M)$, which means that the data point $\mathbf{x}_i$ belongs to class C_{k_i}. Suppose the labelled data set contains M_k points in class C_k so that $\sum_k M_k = M$. Let an unlabelled data set contains N unlabelled samples, represented by $\mathbf{x}_i$ $(i = M + 1, M + 2, \ldots, M + N)$.

We first employ the neighborhood construction algorithm commonly used in the literature, for example in [3], [4], [5] and [6], to form a graph for all the data. Then we apply the non-propagating heat diffusion kernel to the graphs. For the purpose of classification, for each class C_k in turn, we set the initial heat at the labelled data in class C_k to be one and all other data to be zero, then calculate the amount of heat that each unlabelled data receives from the labelled data in class C_k. Finally, we assign the unlabelled data to the class from which it receives most heat. More specifically, we describe the resulting non-propagating Heat Diffusion-Based Classifier as follows.

[Step 1: Construct neighborhood graph] Define graph G over all data points both in the training data set and in the unlabelled data set by connecting points $\mathbf{x}_j$ and $\mathbf{x}_i$ from $\mathbf{x}_j$ to $\mathbf{x}_i$ if $\mathbf{x}_j$ is one of the K nearest neighbors of $\mathbf{x}_i$ measured by the Euclidean distance. Let $d(i, j)$ be the Euclidean distance

between point $\mathbf{x}_i$ and point $\mathbf{x}_j$. Set edge weight w_{ij} equal to $d(i,j)$ if $\mathbf{x}_i$ is one of the K nearest neighbors of $\mathbf{x}_j$, and set $n = M + N$.

[Step 2: Compute the Non-propagating Heat Kernel] Using Eq. (4), get the Non-propagating Heat Kernel H.

[Step 3: Compute the Heat Distribution] Let

$$f^k(0) = (x_1^k, x_2^k, \ldots, x_M^k, \underbrace{0, 0, \ldots, 0}_{N})^T,$$

$k = 1, 2, \ldots, c$, where $x_i^k = 1$ if $C_{k_i} = C_k$, $x_i^k = 0$ otherwise. Then we obtain c results for $f(\Delta t)$, namely, $f^k(\Delta t) = Hf^k(0)$, $k = 1, 2, \ldots, c$.

By Eq. (5), $f^k(\Delta t)$ should be equal to $(I + \alpha \Delta t H) f^k(0)$, but the identity matrix I and the constant $\alpha \Delta t$ have no effect on the classifier introduced in Step 4, so we simply let $f^k(\Delta t) = Hf^k(0)$. $f^k(0)$ means that all the data points in class C_k have a unit heat at the initial time while other data points have no heat, and the corresponding result $f^k(\Delta t)$ means that the heat distribution at time Δt is caused by the initial heat distribution $f^k(0)$.

[Step 4: Classify the data] For $l = 1, 2, \ldots, N$, compare the p-th ($p = M + l$) components of $f^1(\Delta t), f^2(\Delta t), \ldots, f^c(\Delta t)$, and choose class C_k such that $f_p^k(\Delta t) = \max_{q=1}^{c} f_p^q(\Delta t)$, i.e., choose the class that distributes the most heat to the unlabelled data $\mathbf{x}_p$, then classify the unlabelled data $\mathbf{x}_p$ to class C_k.

In Figure 1, we illustrate a neighborhood graph, in which three cases are represented by circle and labelled as class 1, two cases are represented by square and labelled as class 2, and one case is represented by a triangle and is unlabelled. According to Step 1, there is an edge from $\mathbf{x}_j$ to $\mathbf{x}_i$ if $\mathbf{x}_j$ is one of the K nearest neighbors of $\mathbf{x}_i$, and hence the in-degree of each node is K. In the graph in Figure 1, K is set to be 2.

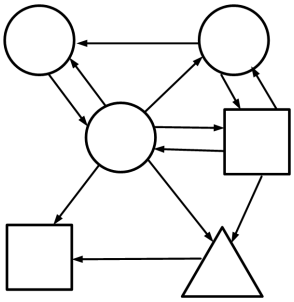

Fig. 1. Neighborhood Graph

Figure 2 shows how heat flows from one node to another node when the initial heat is 1 at nodes in class 1 and 0 at other nodes. A node diffuses heat only to its successors through the directed edge. As a result of the non-propagating heat diffusion, one square receives heat, represented by two small circles, from its two circle predecessors; one square receives heat, represented by one small circle, from its one circle predecessor; the unlabelled data (triangle) receives heat, represented by one small circle, from its one circle predecessor.

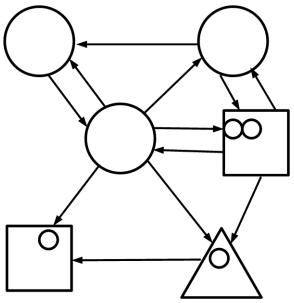

Fig. 2. Non-propagating Heat Diffusion Result on the Neighborhood Graph

Similarly Figure 3 shows the result of non-propagating heat flow when the initial heat is 1 at nodes in class 2 and 0 at other nodes.

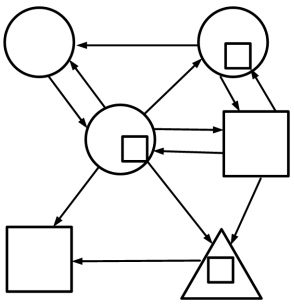

Fig. 3. Non-propagating Heat Diffusion Result on the Neighborhood Graph

The unlabelled data (triangle) receives heat both from nodes in class 1 and nodes in class 2. According to Step 4, we classify the unlabelled data as the class from which it receives the most heat. Through comparison the amount of heat in the triangle in Figure 2 and Figure 3, we classify the unlabelled data to class 2.

In this non-propagating heat diffusion classifier (NHDC), we only consider the heat flow in a small time period, and heat diffuses only once during such a period. We have two free parameters in NHDC: K and β. In the next section, we consider the propagating effect of infinitely many times of heat flow: The heat diffuses to its neighbors first, then these neighbors diffuse the heat further to their own neighbors. This process continues until an appropriate time t is reached.

IV. Propagating Heat Diffusion Classifier

In this classifier, we replace the non-propagating heat diffusion kernel H with the propagating heat diffusion kernel $e^{\gamma H}$. Consequently, the algorithm in Section III changes to the following.

[Step 1: Construct neighborhood graph] The same as Step 1 in Section III.

[Step 2: Compute the Propagating Heat Kernel] Using Eq. (4) and Eq. (8), get the Heat Kernel $e^{\gamma H}$.

[Step 3: Compute the Heat Distribution] $f^k(0)$ is the same as Step 3 in Section III. Using Eq. (7), we obtain c results for $f(t)$, namely, $f^k(t) = e^{\gamma H} f^k(0)$, $k = 1, 2, \ldots, c$.

[Step 4: Classify the data] For $l = 1, 2, \ldots, N$, compare the p-th ($p = M + l$) components of $f^1(t), f^2(t), \ldots, f^c(t)$, and choose class C_k such that $f_p^k(t) = \max_{q=1}^c f_p^q(t)$, i.e., choose the class that distributes the most heat to the unlabelled data $\mathbf{x}_p$ from time 0 to time t, then classify the unlabelled data $\mathbf{x}_p$ to class C_k.

Since we consider the propagating effect of heat diffusion, this classifier is called Propagating Heat Diffusion Classifier (PHDC). We have three free parameters in AHDBC: K, β and γ.

Different from NHDC, after the first heat diffusion, the heat will continue to diffuse in PHDC. The second heat diffusion is based on the result of the first diffusion, which is roughly illustrated by Figure 4 and Figure 5. The tiny circles mean less amount of heat transmitted in the second diffusion, which may directly come from data (circle) in class 1 or indirectly from data (square) in class 2. The tiny squares have similar meaning. For example, there are two tiny circles in the left-lowest large square. They are the results of the second diffusion: One tiny circle is transmitted indirectly from the small circle in the right large triangle, and the other tiny circle is directly from the large circle in the middle. When the time period Δt tends to zero and in fact our model acts this way, there is infinitely many times $t/\Delta t$ of heat diffusion from time 0 to time t.

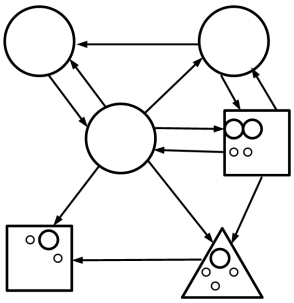

Fig. 4. Second Heat Diffusion Result on the Neighborhood Graph

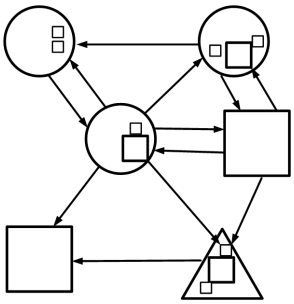

Fig. 5. Second Heat Diffusion Result on the Neighborhood Graph

Remark In Step 1, we construct only one graph over both labelled data and unlabelled data by the method of K nearest neighbors. There are many variants in this step:

1) We can construct the graph by other methods such as ϵ-neighborhood.
2) We can construct c graphs: For each class C_k in turn, construct graph by connecting all the unlabelled data points and data points with label k. In such case, Step 3 and Step 4 need to be changed correspondingly.

V. Interpretation

In Section II, we assume that the heat diffuses in the pipe in the same way as it does in the m-dimensional Euclidean space. Next we will justify this assumption.

It turns out [3] that in an appropriate coordinate system $K_t(x, y)$ on a manifold is approximately the Gaussian:

$$K_t(x, y) = (4\pi t)^{-\frac{m}{2}} \exp(-\frac{||x - y||^2}{4t})(\phi(x, y) + O(t)),$$

where $\phi(x, y)$ is a smooth function with $\phi(x, x) = 1$ and $O(t)$ represents an ignorable term when t is small. Therefore when x and y are close and t is small, we have

$$K_t(x, y) \approx (4\pi t)^{-\frac{m}{2}} \exp(-\frac{||x - y||^2}{4t}).$$

For more details, see [3] and [7].

In our graph heat diffusion model in Section II, we first consider the heat flow in a small time period Δt, and the pipe length between node i and node j is small (recall that only when j is one of the K nearest neighbors, we create an edge from j to i). So the above approximation can be used in our model, and we rewrite it as follows:

$$K_{\Delta t}(i, j) \approx (4\pi\Delta t)^{-\frac{m}{2}} \exp(-\frac{w_{ij}^2}{4\Delta t}). \tag{9}$$

According to the Mean-Value Theorem and the fact that $K_0(i, j) = 0$, we have

$$\begin{aligned} K_{\Delta t}(i, j) &= K_{\Delta t}(i, j) - K_0(i, j) \\ &= \frac{dK_{\Delta t}(i,j)}{d\Delta t}\Big|_{\Delta t=\beta} \Delta t \\ &\approx \alpha \cdot \exp(-\tfrac{w_{ij}^2}{4\beta})\Delta t, \end{aligned}$$

where the last approximation is based on Eq. (9), β is a parameter that depends on Δt, and $\alpha = \frac{1}{4} w_{ij}^2 \beta^{-m/2-2} - \frac{1}{2} m \beta^{-m/2-1}$. To make our model concise, α and β simply serve as free parameters that unrelated to Δt and w_{ij}. This explains why we assume that the at time t, each node i receives $M(i, j, t, \Delta t) = \alpha \cdot \exp(-\frac{w_{ij}^2}{\beta})(f_j(t) - f_i(t))\Delta t$ amount of heat from its neighbor j.

VI. Connections with Other Models and Related Work

In this section, we establish connections between NHDC and other models, and connection between NHDC and PHDC. We show that PWA (Parzen Window Approach [8] when the window function is a multivariate normal kernel) and KNN (K-Nearest-Neighbors) are actually special cases of NHDC, and that PHDC can approximate NHDC. Finally, we compare our heat kernel with those in the related work.

A. NHDC and Parzen Window Approach

First we review the Parzen Windows non-parametric method for density estimation, using Gaussian kernels. When the kernel function $H(u)$ is a multivariate normal kernel, a common choice for the window function, the estimate of the density at the point x is

$$\widetilde{p}(\mathbf{x}) = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{(2\pi h^2)^{d/2}}\exp(-\frac{||\mathbf{x}-\mathbf{x}_i||^2}{2h^2}). \quad (10)$$

When applying it for classification, we need to construct the classifier through the use of Bayes's theorem. This involves modelling the class-conditional densities for each class separately, and then combining them with priors to give models for the posterior probabilities which can then be engaged to make classification decisions [8]. The class-conditional densities for class C_k can be obtained by extending Eq. (10):

$$\widetilde{p}(\mathbf{x}|C_k) = \frac{1}{M_k}\sum_{i:C_{k_i}=C_k}\frac{1}{(2\pi h^2)^{d/2}}\exp(-\frac{||\mathbf{x}-\mathbf{x}_i||^2}{2h^2}), \quad (11)$$

while the priors can be estimated using $\widetilde{p}(C_k) = \frac{M_k}{M}$. Using Bayes' theorem, we get

$$\widetilde{p}(C_k|\mathbf{x}) = \frac{1}{Mp(\mathbf{x})(2\pi h^2)^{d/2}}\sum_{i:C_{k_i}=C_k}\exp(-\frac{||\mathbf{x}-\mathbf{x}_i||^2}{2h^2}). \quad (12)$$

If $K = n-1$, then the graph constructed in Step 1 will be a complete graph, and the matrix H in Eq. (4) becomes

$$H_{ij} = \begin{cases} -\sum_{k\neq i}\exp(-\frac{w_{ik}^2}{\beta}), & j = i; \\ \exp(-\frac{w_{ij}^2}{\beta}), & j \neq i. \end{cases} \quad (13)$$

Then, in NHDC, the heat $f_p^k(\Delta t)$ that unlabelled data $\mathbf{x}_p$ receives from the data points in class C_k will be equal to $\sum_{i:C_{k_i}=C_k}\exp(-||\mathbf{x}_p-\mathbf{x}_i||^2/\beta)$, which is the Eq. (12) if we let $\gamma = 1/Mp(\mathbf{x})(2\pi h^2)^{d/2}$, and $\beta = 2h^2$. This means that Parzen Window Approach when the window function is a multivariate normal kernel can be considered as a special case of NHDC (when we let $K = n-1$ in NHDC).

B. NHDC and KNN

If β tends to infinity, then $\exp(-\frac{w_{ij}^2}{\beta})$ will tend to one, and the matrix H in Eq. (4) becomes

$$H_{ij} = \begin{cases} -K_i, & j = i; \\ 1, & \mathbf{x}_j \text{ is one of the } K \text{ nearest neighbors of } \mathbf{x}_i; \\ 0, & \text{otherwise.} \end{cases} \quad (14)$$

Here K_i is the outdegree of the point $\mathbf{x}_i$ (note that the indegree of the point $\mathbf{x}_i$ is K). Then, in NHDC, the heat $f_p^q(\Delta t)$ that unlabelled data x_p receives from the data points in class C_q will be equal to

$$f_p^q(\Delta t) = \sum_{i:l_i=C_q} 1 \ = K_q,$$

where K_q is the number of the labelled data points from class C_q, which are the K nearest neighbors of the unlabelled data point $\mathbf{x}_p$. Note that when $N = 1$, i.e., when the number of unlabelled data is equal to one, $\sum_{q=1}^{c} K_q = K$. According to Step 4, we will classify the unlabelled data $\mathbf{x}_p$ to the class C_k such that $f_p^k(\Delta t) = K_k$ is the maximal among all $f_p^q(\Delta t) = K_q$. This is exactly what KNN does, and so KNN can be considered as a special case of NHDC (when β tends to infinity and $N = 1$).

C. NHDC and PHDC

When the parameter γ is small, we can approximate $e^{\gamma H}$ in Eq. (8) by its first two items, i.e.,

$$e^{\gamma H} \approx I + \gamma H, \quad (15)$$

then in PHDC, $f^k(t) = e^{\gamma H}f^k(0) \approx f^k(0) + \gamma H f^k(0)$. As the constant γ and the first item $f^k(0)$ impose no effect on the classifier, PHDC possesses a similar classification ability in this case as NHDC, in which $f^k(\Delta t) = Hf^k(0)$. This denotes the relation between NHDC and PHDC.

D. Related Work

The success in [1] is achieved partly because of the speciality of the geometry in the problem. For most geometries, however, there is no closed form solution for the heat kernel. Even worse, in most cases, the underlying geometry structure is unknown. In such cases, it is impossible to construct the heat kernel for the geometry in a closed form. In contrast, there is always a closed form solution – a heat kernel for the graph that approximates the geometry in our model. In [1] and [2], heat kernel is applied to a large margin classifier; in contrast, our kernel is employed directly to construct a classifier.

It is worthy to make a theoretical comparison between the heat kernel in our model and that in [2] because it is impossible to make an empirical comparison between them (as shown below in the second item, their applications are different), and because our heat kernel shows the same appearance $e^{\gamma H}$ as that in [2]. We list below the major differences between them:

1) When the graph is symmetric and β tends to infinity, the matrix H and the heat kernel $e^{\gamma H}$ in our model take the same form as that in [2].
2) Our classifier is mainly concerned with the real-valued data, while the proposed classifier in [2] aims at categorical data in their experiments.
3) Our graph is constructed by the K nearest neighbors in order to approximate the discrete structure of the unknown manifold, while in [2], for each attribute, a graph is constructed by a hypercube, and then the final diffusion kernel is the product of each individual diffusion kernel.
4) Our model is created by the imitation of the non-propagating heat diffusion and the propagating effect of the local heat diffusion. The heat flow in the pipe behaves in the way of locality, and thus it can approximate the heat kernel in the Euclidean space because the time period and the pipe length are small. However, in [2], there is no such consideration.

5) Limited to narrow applications, the kernel in [2] must satisfy two mathematical requirements to be able to serve as a kernel: It must be symmetric and positive semi-definite. In contrast, without the limitation of being applied to a kernel-based classifier, our heat kernel is not necessarily symmetric and positive semi-definite.

Nevertheless, it is interesting to combine these two models by considering the cases when there are both continuous attributes and categorical attributes in the data set. Besides, it is a challenge to apply our heat kernel to a kernel-based classifier when the kernel is not symmetric. These deserve further investigations, but are outside the scope of this paper.

VII. Experiments

The Parzen Window Approach (PWA), KNN, NHDC and PHDC are applied to six datasets from the UCI Repository. Table I describes the datasets we use. The first column refers to the names of the datasets, the second column refers to the number of cases in each dataset, the third column refers to the number of classes, and the fourth column is the number is the number of attributes. In the dataset Credit-g, we only consider the seven continuous attributes while the thirteen discrete attributes are ignored.

TABLE I
Description of the Datasets

dataset	**Cases**	**Classes**	**Attributes**
Credit-g	1000	2	7
Diabetes	768	2	8
Glass	214	6	9
Iris	150	3	4
Sonar	208	2	60
Vehicle	846	4	18

In order to make each attribute in the same scale, we preprocess the datasets by transforming the domain of each attribute to the interval [0,1]. Specifically, for each attribute i, we transform the value x for attribute i by $(x-\min(i))/(\max(i)-\min(i))$, where $\min(i)$ and $\max(i)$ are the minimum and maximum value of attribute i, respectively.

The parameter setting is shown in Table II. The figures shown in Table III are the mean error rates of ten-fold cross-validations, and the last row in Table III shows the average results.

The experimental results show that NHDC uniformly outperforms PWA and KNN in accuracy, indicating the superiority of our approach. Furthermore, PHDC improves over NHDC.

VIII. Conclusion

We have presented two classifiers NHDC and PHDC by imitating the way that heat flows in a medium with a geometric structure. By approximating the manifold by the K nearest neighbors graph, we can avoid the difficulty of finding the explicit expression for the unknown geometry in most cases. By establishing the heat diffusion equation on the graph, we avoid the difficulty of finding a closed form heat kernel for some complicated geometries. Moreover, our solution to heat equation has the property of heat preserving, but our heat kernel is not symmetric and positive definite.

TABLE II
Parameters setting of PWA KNN NHDC and PHDC

dataset	**PWA**	**KNN**	**NHDC**		**PHDC**		
	$1/\beta$	K	K	$1/\beta$	K	$1/\beta$	γ
Credit-g	50	31	13	0	11	0	0.02
Diabetes	300	34	33	50	34	150	0.05
Glass	7500	3	40	1750	38	1500	0.27
Iris	350	7	15	0	13	50	0.47
Sonar	1150	3	24	1650	24	1200	0.41
Vehicle	650	10	8	350	10	600	0.11

TABLE III
Mean error rates of PWA KNN NHDC and PHDC

dataset	**PWA**(%)	**KNN**(%)	**NHDC**(%)	**PHDC** (%)
Credit-g	27.65	24.41	**23.90**	23.94
Diabetes	25.04	24.22	**23.70**	23.78
Glass	28.44	29.36	27.01	**26.88**
Iris	2.93	2.64	2.64	**2.21**
Sonar	11.72	17.14	11.25	**10.93**
Vehicle	27.55	28.59	27.10	**27.07**
Average	20.56	21.06	19.26	**19.14**

While NHDC is a generalization of both Parzen Window Approach (when the window function is a multivariate normal kernel) and KNN, PHDC can approximate NHDC if parameter γ is small. Both NHDC and PHDC are proven to be efficient in our experiments.

Acknowledgments

This work is fully supported by two grants from the Research Grants Councils of the Hong Kong Special administrative Region, China (Project No. CUHK4205/04E and Project No. CUHK4235/04E). The UCI Data Repository owes its existence to David Aha and Patrick Murphy.

References

[1] J. Lafferty and G. Lebanon, "Diffusion kernels on statistical manifolds," *Journal of Machine Learning Research*, vol. 6, pp. 129–163, Jan 2005.

[2] R. Kondor and J. Lafferty, "Diffusion kernels on graphs and other discrete input spaces," in *Proceedings of the International Conference on Machine Learning (ICML), 2002*, 2002.

[3] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, Jun 2003.

[4] J. B. Tenenbaum, V. d. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 22, pp. 2319–2323, Dec 2000.

[5] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 22, pp. 2323–2326, Dec 2000.

[6] L. K. Saul and S. T. Roweis, "Think globally, fit locally: unsupervised learning of low dimensinal manifolds," *Journal of Machine Learning Research*, vol. 4, pp. 119–155, Jun 2003.

[7] S. Rosenberg, *The Laplacian on a Riemmannian Manifold*. Cambridge University Press, 1997.

[8] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.