用于图分类的组合维核方法

李宇峰[1)] 郭天佑[2)] 周志华[1)]

[1)](南京大学计算机软件新技术国家重点实验室 南京 210093)

[2)](香港科技大学计算机科学与工程系 香港)

摘 要 对图等内含结构信息的数据进行学习,是机器学习领域的一个重要问题.核方法是解决此类问题的一种有效技术.文中针对分子图分类问题,基于 Swamidass 等人的工作,提出用于图分类的组合维核方法.该方法首先构建融合一维信息的二维核来刻画分子化学特征,然后基于分子力学的相关知识,利用几何信息构建三维核来刻画分子物理性质.在此基础上对不同维度的核进行集成,通过求解二次约束二次规划问题来获得最优核组合.试验结果表明,文中方法比现有技术具有更好的性能.

关键词 机器学习;图分类;核方法;结构信息;集成学习

中图法分类号 TP18 **DOI 号**: 10.3724/SP.J.1016.2009.00000

Combo-Dimensional Kernels for Graph Classification

LI Yu-Feng[1)] James Tin-Yau Kwok[2)] ZHOU Zhi-Hua[1)]

[1)](*National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing* 210093)

[2)](*Department of Computer Science and Engineering, Hong Kong University of Science & Technology, Hong Kong*)

Abstract Learning from structured data, such as graphs, is an important problem in machine learning. Kernel method is regarded as a powerful solution to such a problem. This paper focuses on molecular graph classification and, following Swamidass et al.'s work, proposes an improved method using combo-dimensional kernels. The method proposed first constructs 2D kernels combined with 1D information to describe chemical characteristics, and to describe physical characteristics, it then constructs 3D kernels based on geometrical information and related molecular mechanics knowledge. Furthermore, inspired by ensemble learning with multiple dimensions, the method finds the optimal kernel combination by quadratically constrained quadratic programming. Experiments show that the proposed method outperforms existing algorithms.

Keywords machine learning; graph classification; kernel methods; structure information; ensemble learning

1 引 言

常用的许多分类算法都假设样本是由向量表示,然而在许多实际应用问题中,样本通常是带有结构的,例如生物学中 RNA、DNA 的序列表示[1],自然语言处理的树型表示[2],XML 的半结构化(semi-structured)表示[3]以及化学中分子的图表示[4]等等.如何利用这些样本内含的结构信息从而进行有效地学习,是机器学习领域的一个重要问题.这方面

收稿日期:2008- - ;最终修改稿收到日期:2008- - .本课题得到国家自然科学基金(60635030,60721002)资助.**李宇峰**,男,1983 年生,博士研究生,主要研究方向为机器学习、数据挖掘. E-mail: liyf@lamda.nju.edu.cn. **郭天佑**,男,1966 年生,博士,副教授,博士生导师,主要研究领域为机器学习、模式识别等. **周志华**,男,1973 年生,博士,教授,博士生导师,主要研究领域为机器学习、数据挖掘、信息检索、模式识别等.

的成果可以在化学分子的毒性检测、诱变分析、致癌检测[5-6]等许多应用中发挥作用.

核方法(kernel methods)通过将数据映射到高维特征空间,并在特征空间中优化结构风险,在泛化性能上有一定的保证.由于这类方法只需要构建一个核函数用于度量样本之间的相似度,而不受限于样本的具体表示形式,因此能有效地对图(graph)这样的结构型数据进行学习[7].

对给定的学习任务构建合适的核函数,是核方法的核心问题.其难点在于既需要高效且合理地利用样本信息以刻画样本的相似度,又需满足核矩阵的半正定性以得到最优解.就化学分子的图型结构而言,已经有很多研究者提出了多种核函数,如随机路径核[8]、最佳匹配核[9]、间隔化核[10]等等.最近,Swamidass 等人[4]结合领域知识,分别为化学分子的一维(1D)、二维(2D)、三维(3D)表示设计了一族核函数,取得了很好的结果.然而该方法仅考虑分子的化学特征,忽略了分子的物理特征.本文在 Swamidass 等人工作的基础上,提出了组合维核方法(Combo-Combinational Kernel method).该方法首先构建融合了 1D 信息的 2D 核来刻画化学特征,然后结合分子力学中能量的概念,基于距离、角度、二面角等几何信息来构建可以刻画物理特征的 3D 核.在此基础上对不同维度的核进行集成,本文引入二次约束二次规划(Quadratically Constrained Quadratic Programming, QCQP)方法来产生 2D 核和 3D 核的最优组合.实验表明,本文方法在 PTC 和 NCI 的 10 个数据集上比现有方法的性能更好;在 Mutag 数据集上与现有算法性能相当.

本文第 2 节简要介绍基于核的图分类方法和最优核矩阵学习技术;第 3 节给出本文方法和理论分析;第 4 节给出实验结果;最后在第 5 节总结全文.

2 研究背景

2.1 基于核的图分类方法

真实世界中的对象通常包含一定的结构信息,因此,对结构型数据的学习受到了机器学习界越来越多的关注.在早期工作中,研究者们侧重研究图的一些特例,例如字符串、树等,并且提出了多种有效的核算法,如谱核[11]、卷积核[12]等,此类技术被称为是基于语义的核方法[13].2003 年,Gärtner 等人[8]对一般性的图分类进行了研究,指出得到完美图核(即特征映射为双射函数)是一个至少和图同构一样难的问题,并且指出计算子图空间上的内积将是一个 NP 难的问题.这意味着对一般性的图分类问题,研究目标应该是找到比较好的近似解,而不是理想的最优解.Gärtner 等人利用随机路径近似数据分布,提出了一种具有多项式时间开销的随机路径核算法,可以有效地描述了图之间的相似性.此后,很多研究者对一般性的图分类进行了研究,并提出了多种核算法,如间隔化核[10]、Fisher 核[14]、扩散核[15]等,此类技术被称为基于模型的核方法[13].此外,还有一些研究者利用图论中的方法提出了最短路径核[16]、最佳匹配核[9]、圈核[17]等.

值得注意的是,真实问题的图数据通常都具有鲜明的应用背景,如果能有效的利用领域知识,有可能获得更好的性能.注意到化学分子具有三种维度表示,即一维是一个字符串、二维是一幅标记图、三维是一个空间立体模型,Swamidass 等人[4]最近结合领域知识,定义了三个维度下相应的核函数,它以较小的计算开销,取得良好的实验效果.

然而,通过对 Swamidass 等人工作的仔细研究,我们发现 Swamidass 等人的 3D 核虽然拥有最丰富的信息,但其效果并不好.其原因可能在于可能是由于 Swamidass 等人所构建的 1D、2D、3D 核是相互独立的,例如其 3D 核直接对 3D 距离的直方图求高斯核,没有对 2D 信息进行利用,这样就忽略了大量的有用信息.与此类似,他们在构建 2D 核时也没有充分利用重要的 1D 信息.基于这一认识,本文一方面试图在构建高维核时充分利用低维信息;另一方面,注意到 2D 核主要反映分子的化学特征,3D 核主要体现分子的物理特征,本文试图将不同维的核进行集成[18],以获得更好的泛化性能.

2.2 最优核矩阵学习

核函数的选择是核方法的一个关键问题,由于核函数与核矩阵存在着对应关系,一些研究者对核矩阵的选择进行了研究.Cristianini 等人[19]提出了核矩阵之间的相似度度量 Alignment,并指出当核矩阵与目标核矩阵(即由标记构成的核矩阵)之间的 Alignment 值越大时,期望泛化误差将越小.最近,Nguyen 和 Ho[20]指出,Alignment 只是泛化性能的充分条件,而不是必要条件,他们还给出了一种具有充分必要性的相似度度量.

Lanckriet 等人[21]指出,以 SVM 的结构风险为优化目标的核矩阵学习可以等价于求解一个半定规划(SemiDefinite Programming, SDP)问题.SDP 属于凸规划问题,理论上可以保证得到全局最优解.进

一步，当以最大化 Alignment 为目标并且只考虑核的正组合时，核学习问题可以降解为一个 QCQP 问题[21]. 考虑到求解 SDP 的计算开销很大，在问题规模大时难于使用，而 QCQP 虽然速度很快，但是效果却略差，Tsang 和 Kwok[22] 提出了核学习的 SOCP 算法，对精度和复杂度进行了折中.

3　本文方法

本节先介绍所使用的符号和一些基本定义，然后提出基于路径的 2D 核，再提出基于几何信息的 3D 核，最后描述组合维核方法.

3.1　符号和特征映射

给定标记图(Labeled Graph) $G(V,E,L)$，其中 V 是顶点 $\{v_i\}_{i=1}^{n}$ 的非空有限集，E 是边 $\{e_j\}_{j=1}^{h}$ 的非空有限集，L 是顶点或边对应的标记集. 则可定义标记图的标记路(Labeled Path) p 为 $p=l_v(v_{i_1})l_e(e_{j_1})\cdots l_v(v_{i_k})l_e(e_{j_k})l_v(v_{i_{k+1}})$，这里 l_v 表示顶点到标记的映射，l_e 表示边到标记的映射，k 为标记路的长度. 简单起见，下文简称标记图为图、标记路为路.

记 $P(l)$ 为数据集中所有长度不大于 l 的路径集合，则对于图 g 及给定路径长度 l，可定义特征映射：

$$\phi_l(g)=(\phi_p(g))_{p\in P(l)} \tag{1}$$

这里 $\phi_p(g)$ 为 1 当且仅当 g 中存在路 p，否则为 0. 类似地，若定义 $\psi_p(g)$ 为 g 中路 p 的条数，则可得到相应的特征映射 $\psi_l(g)$.

任意两幅图 x,y 的核函数可定义为

$$\boldsymbol{K}_l(x,y)=\langle\phi_l(x),\phi_l(y)\rangle=\sum_{p\in P(l)}\phi_p(x)\phi_p(y) \tag{2}$$

3.2　基于路径的 2D 核

3.2.1　Tanimoto 核及 Minmax 核

为了消除图的规模差异性，需要对核函数进行规范化. 本文借鉴 Swamidass 等人的工作[4]，引入两种化学信息学领域经典的规范化技术——Tanimoto 和 Minmax[23]. 由此产生式(3)所示的 Tanimoto 核和式(4)所示的 Minmax 核.

$$\boldsymbol{K}_l^t(x,y)=\frac{\boldsymbol{K}_l(x,y)}{\boldsymbol{K}_l(x,x)+\boldsymbol{K}_l(y,y)-\boldsymbol{K}_l(x,y)} \tag{3}$$

$$\boldsymbol{K}_l^m(x,y)=\frac{\sum_{p\in P(l)}\min(\phi_p(x),\phi_p(y))}{\sum_{p\in P(l)}\max(\phi_p(x),\phi_p(y))} \tag{4}$$

式(3)亦等价于：

$$\boldsymbol{K}_l^t(x,y)=\frac{|\phi_l(x)\cap\phi_l(y)|}{|\phi_l(x)\cup\phi_l(y)|} \tag{5}$$

其中 $|\phi_l(x)\cap\phi_l(y)|$ 与 $|\phi_l(x)\cup\phi_l(y)|$ 分别表示 $\phi_l(x)$，$\phi_l(y)$ 交集与并集的势.

Tanimoto 核及 Minmax 核的半正定性证明可参见文献[23].

3.2.2　路径的生成和复杂度分析

一般使用深度优先策略(DFS)来生成图中的路径，这一策略有很多变种. 例如可以选择分歧式(divergent)，由根结点开始访问同一条边多次，也可以无分歧式，即不允许多次访问同一条边. 前者遍历了所有路径，其时间复杂度为 $O(n\alpha^l)$，这里 n 为顶点个数，α 是顶点的最大度数，l 是路径的最大长度；而后者减少了大量的开销，其时间复杂度仅为 $O(\ln h)$，这里 h 是边数. DFS 还可以选择是否允许生成简单圈. 采用何种生成方式，取决于路径是否高效且贴切地反映了分子的化学性质. Swamidass 等人选用了无分歧式的 DFS，路径的长度最短为 1，最长为 10. 注意到，当路径长度为零时，Tanimoto 核和 Minmax 核比较的是分子间相同原子的个数，这体现了分子的 1D 信息. 此外，环状结构具有丰富的化学性质，例如苯环. 为此，本文使用无分歧式、带简单圈的 DFS，并且设置路径长度为 0 到 10. 根据上文的分析，计算两幅图的 2D 核所需时间开销为 $O(\ln_1 h_1+\ln_2 h_2)$. 总的来说，本节定义的 2D 核不仅考虑了 1D 信息和领域知识，而且还可以快速地计算得到.

3.3　基于几何信息的 3D 核

3.3.1　距离核、角度核和二面角核

分子信息学的相似性原则(similar property principle)指出，结构相近的分子具有相似的性质，而且小型化学分子的 3D 结构信息具有显著的物理意义，如能量、量子力学性能(quantum mechanical property)等[24]. 本文设计了 3 种基于几何信息的 3D 核，它们不仅可以在一定程度上反映分子能量的物理意义，还不需要大的计算开销.

设 $\boldsymbol{X}^a$、$\boldsymbol{X}^b$、$\boldsymbol{X}^c$、$\boldsymbol{X}^d$ 为分子 m 中原子 a、b、c、d 的三维坐标的向量表示，则 a 和 b 之间的距离 r_m^{ab} 定义为

$$r_m^{ab}=\|\boldsymbol{X}^a-\boldsymbol{X}^b\|_2=\Big(\sum_{i=1}^{3}(\boldsymbol{X}_i^a-\boldsymbol{X}_i^b)^2\Big)^{\frac{1}{2}} \tag{6}$$

由 a、b、c 张成的角度 θ_m^{abc} 定义为

$$\theta_m^{abc}=\arccos\Big(\frac{(\boldsymbol{X}^a-\boldsymbol{X}^b)\cdot(\boldsymbol{X}^b-\boldsymbol{X}^c)}{r_m^{ab}\times r_m^{bc}}\Big) \tag{7}$$

其中“·”表示向量之间的内积.

由 a、b、c、d 张成的二面角 ϑ_m^{abcd} 定义为

$$\boldsymbol{Y}^{abc}=(\boldsymbol{X}^a-\boldsymbol{X}^b)\otimes(\boldsymbol{X}^b-\boldsymbol{X}^c) \tag{8}$$

$$\boldsymbol{Y}^{bcd}=(\boldsymbol{X}^b-\boldsymbol{X}^c)\otimes(\boldsymbol{X}^c-\boldsymbol{X}^d) \tag{9}$$

$$\vartheta_m^{abcd}=\arccos\left(\frac{\boldsymbol{Y}^{abc}\cdot\boldsymbol{Y}^{bcd}}{\|\boldsymbol{Y}^{abc}\|_2\|\boldsymbol{Y}^{bcd}\|_2}\right) \tag{10}$$

其中“$\otimes$”表示向量之间的外积(或叉乘).

对数据集 $\boldsymbol{M}$ 中任意的两个分子 m 与 m',假设它们分别拥有原子 $a_1,a_2,\cdots,a_{|m|}$ 与 $a'_1,a'_2,\cdots,a'_{|m'|}$. 记 a_i^p 为路径 p 中的第 i 个顶点标记,本文定义距离核为式(11),角度核为式(12),二面角核为式(13).

$$\boldsymbol{K}_d(m,m')=\sum_{p\in\phi_1(m)\cap\phi_1(m')}\boldsymbol{k}_d(r_m^{a_1^p a_2^p},r_{m'}^{a_1^p a_2^p}) \tag{11}$$

$$\boldsymbol{K}_a(m,m')=\sum_{p\in\phi_2(m)\cap\phi_2(m')}\boldsymbol{k}_a(\theta_m^{a_1^p a_2^p a_3^p},\theta_{m'}^{a_1^p a_2^p a_3^p}) \tag{12}$$

$$\boldsymbol{K}_t(m,m')=\sum_{\substack{p\in\phi_3(m)\\p\in\phi_3(m')}}\boldsymbol{k}_t(\vartheta_m^{a_1^p a_2^p a_3^p a_4^p},\vartheta_{m'}^{a_1^p a_2^p a_3^p a_4^p}) \tag{13}$$

这里 $\boldsymbol{k}_d$、$\boldsymbol{k}_a$、$\boldsymbol{k}_t$ 均为半正定核.

3.3.2 半正定性和复杂度分析

由式(11)~(13)可看出距离核体现了相同边之间距离的相似度,角度核体现的是相同的连续三元组之间角度的相似度,二面角核则体现了相同的连续四元组之间二面角的相似度.

定理 1. 距离核、角度核、二面角核均为半正定核.

证明. 首先考虑距离核. 记 $u(p,m)=r_m^{a_1^p a_2^p}$,因 $\boldsymbol{k}_d$ 为半正定核,由 Mercer 定理[7]可知,存在映射 φ 满足

$$\boldsymbol{k}_d(u(p,m),u(p,m'))=\langle\varphi(u(p,m)),\varphi(u(p,m'))\rangle \tag{14}$$

考虑式(1)中 ϕ_l 的定义,可以定义特征映射 τ:

$$\tau(m)=\begin{cases}\varphi(u(p,m)), & p\in\phi_1(m)\\ 0, & \text{其它}\end{cases} \tag{15}$$

从而有

$$\boldsymbol{K}_d(m,m')=\langle\tau(m),\tau(m')\rangle \tag{16}$$

成立. 因此,距离核为半正定核得证.

完全类似地,可以证明角度核、二面角核为半正定核. 证毕.

基于分子力学中键长、键角、二面角能量等领域知识[25],本文将 $\boldsymbol{k}_d$、$\boldsymbol{k}_a$、$\boldsymbol{k}_t$ 定义为

$$\boldsymbol{k}_d(r_m^p,r_{m'}^p)=\exp\left(-\frac{\|(r_m^p-r_p)^2-(r_{m'}^p-r_p)^2\|^2}{\sigma_{r,p}^2}\right) \tag{17}$$

$$\boldsymbol{k}_a(\theta_m^p,\theta_{m'}^p)=\exp\left(-\frac{\|(\theta_m^p-\theta_p)^2-(\theta_{m'}^p-\theta_p)^2\|^2}{\sigma_{\theta,p}^2}\right) \tag{18}$$

$$\boldsymbol{k}_t(\vartheta_m^p,\vartheta_{m'}^p)=\exp\left(-\frac{\|\cos\vartheta_m^p-\cos\vartheta_{m'}^p\|^2}{\sigma_{\vartheta,p}^2}\right) \tag{19}$$

这里用 r_m^p、θ_m^p、ϑ_m^p 简记 $r_m^{a_1^p a_2^p}$,$\theta_m^{a_1^p a_2^p a_3^p}$,$\vartheta_m^{a_1^p a_2^p a_3^p a_4^p}$. 参见分子力学的领域知识,$r_p$,$\theta_p$ 为路径 p 相应的距离或角度的能量最低值. 为简化计算,这里令 r_p,θ_p 为相同路径集合中的距离或角度的最小值. 最后,$\sigma_{r,p}$,$\sigma_{\theta,p}$,$\sigma_{\vartheta,p}$ 为高斯核的半径.

使用分歧式的 DFS 带来的开销为 $O(n\alpha^l)$. 注意到分子中每一个原子的度数不会大于 4,即 $\alpha\leq4$,并且计算距离、角度、平面角所需的路径长度分别为 $l=1,2,3$,从而 α^l 为有界常数. 为此,本文使用分歧式的 DFS,3D 核的计算开销仅为 $O(n_1+n_2)$. 通过引入平衡二叉树,可以使每个图的存储开销仅为 $O(nh)$. 因此,本文所定义的 3D 核在时间和空间复杂度上都比较小.

3.4 最优核组合

3.2 节定义的 2D 核利用了分子的化学特征,3.3 节定义的 3D 核利用了分子的物理特征,两者显然有较大的差异,而且有一定的互补性. 由集成学习领域[7]的研究可知,对它们进行结合将可望产生更强的泛化能力. 借鉴 Lanckriet 等人[21]的工作,本文以最大化 Alignment 为目标,使用 QCQP 核学习算法来产生本文的组合维核.

设 $\boldsymbol{K}_1,\cdots,\boldsymbol{K}_d$ 为定义在数据集 $\boldsymbol{M}$ 上不同的半正定核,$\boldsymbol{y}$ 是数据集的标记列向量. 根据 Lanckriet 等人的结果[21],求解最大化 Alignment 的半正定核 $\boldsymbol{K}$ 等价于求解以下优化问题:

$$\begin{aligned}&\max_{K}\ \langle\boldsymbol{K},\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}\rangle\\ &\text{w. s. t.}\ \ \boldsymbol{K}=\sum_{i=1}^{d}\boldsymbol{\mu}_i\boldsymbol{K}_i,\quad \boldsymbol{K}\geq0\\ &\qquad\ \langle\boldsymbol{K},\boldsymbol{K}\rangle_F\leq1,\quad \boldsymbol{\mu}\geq0\end{aligned} \tag{20}$$

其中,$\langle\boldsymbol{K}_i,\boldsymbol{K}_j\rangle_F=tr(\boldsymbol{K}_i^{\mathrm{T}}\boldsymbol{K}_j)$ 表示矩阵的内积,$tr(\boldsymbol{M}_{c\times c})=\sum_{i=1}^{c}\boldsymbol{M}_{i,i}$ 表示矩阵的迹.

对式(20)进行对偶变换,问题可转化为如下等价的优化问题:

$$\begin{aligned}&\max_{\boldsymbol{\mu}}\ \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{q}\\ &\text{s. t.}\ \ \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{S}\boldsymbol{\mu}\leq1\\ &\qquad\ \boldsymbol{\mu}\geq0\end{aligned} \tag{21}$$

其中,$\boldsymbol{q}_i=\langle\boldsymbol{K}_i,\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}\rangle_F$,$\boldsymbol{S}_{i,j}=\langle\boldsymbol{K}_i,\boldsymbol{K}_j\rangle_F$.

式(21)是一个二次约束的凸规划问题,即 QCQP

问题，问题中的优化变量只与核矩阵的个数有关，因此计算开销很小. 本文就是使用式(21)对 2D 和 3D 核的最优正组合进行学习.

4　实验测试

4.1　实验数据

本文在 Mutag[26] 和 PTC[27] 以及 10 个 NCI 数据集上进行了实验. 其中前两个数据集规模较小，后 10 个数据集规模较大.

Mutag 是诱变分子数据集，包含 188 个可用于学习的样本，其中 125 个为正例，63 个为反例. 数据集本身并没有提供分子坐标数据，实验中坐标数据采用化学软件 ChemOffice① 生成.

PTC 是致癌分子数据集，包含 470 个样本，分成 4 组分类问题，即 male mice(MM)、female mice(FM)、male rats(MR)、female rats(FR). 每个分子样本共有 8 个标记，即{EE，IS，E，CE，SE，P，NE，N}，其中 CE，SE 和 P 被认为是正例，NE 和 N 被认为是反例，EE，IE 和 E 被认为是不可判别，不参与分类.

NCI 是大型分子数据库，共有 73 个数据集，数据集的规模大致在 3500 到 4200 之间. 本文选用 Swamidass 等人[4] 汇报的 5 个精度最高的数据集，即 NCI_H522、MOLT_4、MCF_7、HCT_116、SK_MEL_5 以及 5 个精度最低的数据集，即 NCI_H226、UACC_257、HCC_2998、OVCAR_5、SNB_75.

4.2　实验设置

本文采用了与 Swamidass 等人[4] 相同的测试方式，即在 Mutag 和 PTC 上使用留一法(leave-one-out)，而在每个 NCI 数据集上进行 20 次 hold-out 测试，每次随机划分 80%的数据为训练数据，剩下的 20%数据为测试数据，记录平均精度和 ROC 值. 本文方法将与 Swamidass 等人的方法进行比较.

本文方法共涉及 2 个参数，即高斯核函数的带宽参数 σ 与 SVM 的正则化(regularization)系数 C. 本文实验中将 σ 设置为路径相应 3D 信息的最大最小规范化值的差. 参数 C 的设置与 Swamidass 等人[4] 使用的设置一致，即 $C=1$.

本文的 DFS 路径生成算法在 VC++集成环境下实现，QCQP 采用优化工具软件 mosek4.0 实现，最后采用软件包 libsvm 实现学习算法.

4.3　实验结果

4.3.1　Mutag 和 PTC

表 1 给出了在 Mutag 和 PTC 4 个分类问题上的实验结果. 其中报告的 Swamidass 等人的 2D 核的结果(表 1 中记为 Sw Minmax 和 Sw Tanimoto)来自文献[4]. 加黑的数字表示最高精度.

表 1　本文方法与 Swamidass 等人方法的性能比较

Kernel methods	Mutag	PTC			
		MM	FM	MR	FR
Sw Minmax[4]	86.2	64.0	64.5	64.5	66.4
Sw Tanimoto[4]	**87.8**	**66.4**	64.2	63.7	66.7
本文方法	84.0	65.8	**65.0**	**64.8**	**68.7**

从表 1 可以看出，本文方法在 PTC 的 3 个分类问题上优于现有算法，在另一个分类问题上和现有算法性能相当，在 Mutag 上本文方法不如 Swamidass 等人的方法. 注意到 3D 核的计算涉及原子的坐标数据，实验中 Mutag 采用化学软件生成近似的坐标数据，然而化学软件的拟合程度有限，势必会在一定程度上加入了噪音数据，这可能会对算法性能造成较大的影响[4]. 此外，Mutag 样本数目较小，最大化 alignment 的求解过程容易导致过配(overfitting)现象，这也可能对算法的性能构成影响. 需要说明的是，表 1 中 Swamidass 等人方法的结果是 Swamidass 等人在论文中报告的最好结果[4]，由于我们未能重新实现该方法且 Swamidass 等人不提供代码，我们只能与其报告的最好结果进行比较.

4.3.2　NCI

图 1 与图 2 给出了本文方法在 NCI 10 个数据集上分类精度与 ROC 值的实验结果，图中给出 Swamidass 等人的结果[4] 作为比较.

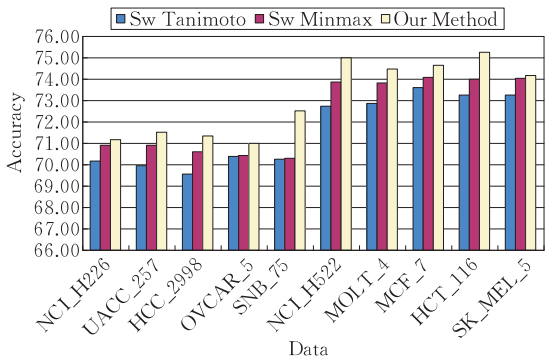

图 1　在 10 个 NCI 数据集上的测试精度比较结果

① ChemOffice Ultra 2004 for Microsoft Windows 2000 or Windows XP, CambridgeSoft Corporation, 100 CambridgePark Drive, Cambridge, MA 02140-9802, U. S. A. Web site: www.cambridgesoft.com

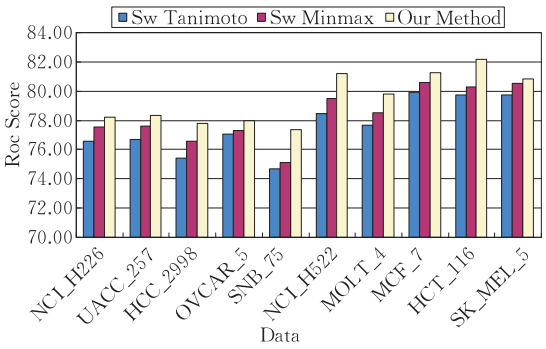

图 2 在 10 个 NCI 数据集上的 ROC 比较结果

从图 1、图 2 可以看出，在 NCI 10 个数据集上，本文方法不管是精度还是 ROC 值，都要优于现有算法. 尤其在 HCC_2998、SNB_75、NCI_H522、HCT_116 等 4 个数据集上，本文方法的优势更加明显.

5 结束语

本文在 Swamidass 等人工作的基础上，针对分子图分类问题，提出了一种组合维核方法. 该方法利用分子化学特征来构建融合了 1D 信息的 2D 核，利用分子物理特征来构建 3D 核，并通过 QCQP 来获得组合维核. 实验表明，本文方法比现有方法具有更好的性能. 本文方法的成功说明，在对内含结构信息的样本进行学习时，如果能充分利用领域知识，可望能获得更好的效果.

参 考 文 献

[1] Durbin R, Eddy S, Krogh A, Mitchison G. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge, UK: Cambridge University Press, 1998

[2] Manning C D, Schütze H. Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press, 1999

[3] Abiteboul S, Buneman P, Suciu D. Data on the Web: From Relations to Semistructured Data and XML. San Francisco, CA: Morgan Kaufmann, 2000

[4] Swamidass S J, Chen J J, Bruand J, Phung P, Ralaivola L, Baldi P. Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity//Proceedings of the 13th International Conference on Intelligent Systems for Molecular Biology. Detroit, MI, 2005: 25-29

[5] Kramer S, De Raedt L. Feature construction with version spaces for biochemical application//Brodley C E, Danyluk A P eds. Proceedings of the 18th International Conference on Machine Learning. Williamstown, MA, 2001: 258-265

[6] Inokuchi A, Washio T, Motoda H. An apriori-based algrithm for mining frequent substructures from graph data//Proceedings of the 4th European Conference on Principles and Practice of Knowledge Discovery in Databases. Lyon, France, 2000: 13-23

[7] Schölkopf B, Smola A J. Learning with Kernels. Cambridge, MA: MIT Press, 2002

[8] Gärtner T, Flach P, Wrobel S. On graph kernels: Hardness results and efficient alternatives//Proceedings of the 16th Annual Conference on Computational Learning Theory and the 7th Kernel Workshop. Washington, DC, 2003: 129-143

[9] Fröhlich H, Wegner J K, Sieker F, Zell A. Optimal assignment kernels for attributed molecular graphs//Proceedings of the 22nd International Conference on Machine Learning. Bonn, Germany, 2005: 225-232

[10] Kashima H, Tsuda K, Inokuchi A. marginalized kernels between labeled graphs//Proceedings of the 20th International Conference on Machine Learning. Washington, DC, 2003: 321-328

[11] Leslie C E, Eskin E, Noble W S. The spectrum kernel: A string kernel for SVM protein classification//Proceedings of the Pacific Biocomputing Symposium. Lihue, HI, 2002: 566-575

[12] Collins M, Duffy N. Convolution kernels for natural language//Dietterich T G, Becker S, Ghahramani Z eds. Advances in Neural Information Processing Systems 14. Cambridge MA: MIT Press, 2002: 625-633

[13] Gärtner T. A survey of kernels for structured data. SIGKDD Exploration, 2003, 5(1): 49-58

[14] Jaakkola T S, Haussler D. Exploiting generative models in discriminative classifiers//Kearns M S, Solla S A, Cohn D A eds. Advanced in Neural Information Processing Systems 11. Cambridge, MA: MIT Press, 1999: 487-493

[15] Kondor R I, Lafferty J. Diffusion kernels on graphs and other discrete input spaces//Proceedings of the 19th International Conference on Machine Learning. Sydney, Australia, 2002: 315-322

[16] Borgwardt K M, Kriegel H P. Shortest-path kernels on graphs//Proceedings of the 5th IEEE International Conference on Data Mining. New Orleans, LO, 2005: 74-81

[17] Horvath T, Gärtner T, Wrobel S. Cyclic pattern kernels for predictive graph mining//Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, WA, 2004: 158-167

[18] Dietterich T G. Ensemble methods in machine learning//Kittler J, Roli F eds. Lecture Notes in Computer Science 1867. Berlin: Springer-Verlag, 2000: 1-15

[19] Cristianini N, Shawe-Taylor J, Elisseeff A, Kandola J. On kernel target alignment//Dietterich T G, Becker S, Ghahramani Z eds. Advances in Neural Information Processing Systems 14. Cambridge, MA: MIT Press, 2002: 367-374

[20] Nguyen C H, Ho T B. Kernel matrix evaluation//Proceedings of the International Joint Conferences on Artificial Intelligence. Hyderabad, India, 2007: 987-992

[21] Lanckriet G R G, Cristianini N, Bartlett P, Ghaoui L E, Jordan M I. Learning the kernel matrix with semidefinite

programming. Journal of Machine Learning Research, 2004, 5: 27-72

[22] Tsang W H, Kwok T Y. Efficient hyperkernel learning using second-order cone programming. IEEE Transactions on Neural Networks, 2006, 17(1): 48-58

[23] Gower J C. A general coefficient of similarity and some of its properties. Biometrics, 1971, 27(4): 857-871

[24] Johnson M A, Maggiora G M eds. Concepts and Applications of Molecular Similarity. New York, NY: Wiley, 1990

[25] Höltje H D, Sippl W, Rognan D, Folkers G. Molecular Modeling: Basic Principles and Applications. 2nd Edition. New York, NY: Wiley, 2003

[26] Debnath A K, Compadre R L, Debnath G, Shusterman A J, Hansch C. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. Correlation with molecular orbital energies and hydrophobicity. Journal of Medicinal Chemistry, 1991, 34(2): 786-797

[27] Helma C, King R D, Kramer S, Srinivasan A. The predictive toxicology challenge. Bioinformatics, 2001, 17(1): 107-108

LI Yu-Feng, born in 1983, Ph. D. candidate. His main research interests include machine learning and data mining.

James Tin-Yau Kwok, born in 1966, Ph. D., associate professor, Ph. D. supervisor. His research interests include kernel methods, machine learning, pattern recognition, and artificial neural network.

ZHOU Zhi-Hua, born in 1973, Ph. D., professor, Ph. D. supervisor. His main interests include artificial intelligence, machine learning, data mining, information retrieval, pattern recognition, evolutionary computation and neural computation.

Background

Learning from structured data, such as graphs, is an important problem in machine learning. Kernel method is a powerful solution to such problem. This paper focuses on molecular graph classification. Following Swamidass et al.'s work, we propose an improved method using combo-dimensional kernels. The proposed method first constructs 2D kernels combined with 1D information to describe chemical characteristics. Then, in order to describe physical characteristics, we construct 3D kernels based on geometrical information and related molecular mechanics knowledge. Furthermore, inspired by ensemble learning, we try to combine the kernels to achieve a better performance through quadratically constrained quadratic programming. Experiments show that the proposed method outperforms several existing algorithms. This work is supported by the National Science Foundation of China under grant Nos. 60635030 and 60721002.