第46卷第9期 中国科学技术大学学报 Vol. 46, No. 9
2016年9月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Sep. 2016

文章编号:0253-2778(2016)9-0749-08

基于KLDA的图核降维方法

余亚军,潘志松,胡谷雨,莫小勇,薛 胶

(解放军理工大学指挥信息系统学院,江苏南京 210007)

摘要:图结构具有较强的表达能力和较高的灵活性,对图结构数据的识别及分类属于结构模式识别的范畴.对图结构数据的研究思路是将图结构数据通过图核转化为向量空间中的向量,然后采用传统的机器学习算法对其进行分析.基于图结构的数据表示与分析已经成为机器学习领域的研究热点.于是提出对经典的图结构分析方法进行扩充,利用核线性判别分析方法(KLDA)对图核诱导的高维特征空间进行降维,得到与原始图结构特征空间对应的低维度的特征空间,然后采用传统的机器学习算法对这些新的数据进行分析.在标准数据集上的实验效果验证了该方法的有效性.

关键词:图分类;图核;核线性判别分析;降维

中图分类号:TP18 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2016.09.006

引用格式:余亚军,潘志松,胡谷雨,等. 基于KLDA的图核降维方法[J]. 中国科学技术大学学报,2016,46(9):749-756.

YU Yajun, PAN Zhisong, HU Guyu, et al. Dimensionality reduction method of graph kernel based on KLDA[J]. Journal of University of Science and Technology of China, 2016,46(9):749-756.

Dimensionality reduction method of graph kernel based on KLDA

YU Yajun, PAN Zhisong, HU Guyu, MO Xiaoyong, XUE Jiao

(*College of Command Information System, PLA University of Science and Technology, Nanjing* 210007, *China*)

Abstract: Graph structure has strong expression ability and high flexibility. The identification and classification of graph structure data fall into the category of structural pattern recognition. The research idea of the graph structure data is to transform the graph structure data to the vector in the vector space, then the traditional machine learning algorithm is used to analyze the vector. Data representation and analysis based on graph structure has become a hot research topic in the field of machine learning. The classical graph kernel method was extended. The kernel linear discriminant analysis (KLDA) was employed to reduce the dimension of the high dimension feature space, and the low dimensional feature space corresponding to the original graph structure data was obtained. Then the traditional machine learning algorithm was used to analyze these new data. The effectiveness of the proposed method is verified by the experimental results on standard data sets.

Key words: graph classification; graph kernel; kernel linear discriminant analysis; dimensionality reduction

收稿日期:2016-03-01;**修回日期**:2016-09-17

基金项目:国家自然科学基金(61473149)资助.

作者简介:余亚军,男,1990年生,硕士生. 研究方向:机器学习与模式识别、云计算. E-mail: 492675818@qq.com

通讯作者:潘志松,博士/教授. E-mail: hotpzs@hotmail.com

0 引言

模式识别可以分为统计模式识别和结构模式识别[1]. 统计模式识别的对象能够表征为固定长度的特征向量,对固定长度的特征向量进行分析有其数学上的便利,因为我们能够使用成熟的算法对该类问题进行处理;对具有图结构的数据进行处理则属于结构模式识别的范畴. 可以表征为图结构的数据很多,包括基因序列、蛋白质分子结构及社会关系网络[2-4]等. 这些实际应用中的数据呈现结构化的特征,统计模式识别的方法无法准确表达,对于这些需要表达内部结构关系的数据,用图结构表示更为合适. 图 1 是用图方法表示化学分子,其中图中的顶点代表原子,边代表原子之间的键,原子名称为顶点的标签,键的类型为边的标签[5].

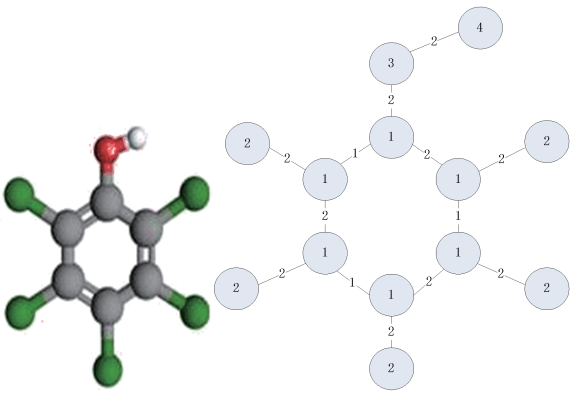

图 1 图方法表示化学分子

Fig. 1 Graph representation of data

图结构数据具有强大的表达能力和较高的灵活性,不但能够反映对象整体的特性,而且没有固定的维数限制,还能够反映对象部分及部分之间的关系;缺点是没有通用的算法对图结构进行分析. 因此将统计模式识别同结构模式识别结合起来,将图结构数据映射到向量空间,从而能够利用统计模式识别中众多成熟的方法对图结构数据进行分析. 近年来,基于图数据表示的分类和聚类算法在机器学习和数据挖掘等相关领域得到了广泛的研究[6].

为了将图结构数据映射到向量空间中,需要对图结构数据进行处理. 核方法就是一种能很好的将结构化数据映射到向量空间中的方法. 在向量型数据分析中,核方法已经得到较为广泛的应用,核函数 $k(x,x')$ 可以用来衡量两个向量型数据 x 和 x' 之间的相似度. 核方法在特征向量表示的数据中能够很好的适用,同时也可将其适用于结构化的数据[7]. 定义两图之间核函数的方法称为图核,即用核函数 $k_{ij}(G_i,G_j)$ 来度量图 G_i 和 G_j 的相似性. 在图结构数据的分析中,常用于图数据分析的核函数即图核定义方法可以分为以下三类:①基于路径的图核,如最短路径核[8],随机路径核等;②基于有限规模子图的图核,圈核[9]等;③基于子树模式的图核,如快速子树核[10]等. 经典的图核模式发现方法[11]即利用图核将含有 N 个图的数据集转化成为 $N \times N$ 的核矩阵 $\boldsymbol{K} = (k_{ij}(G_i,G_j))_{N\times N}$,再直接使用支持向量机(SVM)进行分类. 经过图核方法将原始数据映射到高位特征空间中的数据往往存在一些冗余特征,或与数据分析不相关的特征. 属性太多不仅会增加计算的复杂性,使其计算性能明显下降,而且还有可能降低分类精度,甚至可能会导致"维数灾难". 有时,由于数据采集量有限,使得数据训练样本集相对于数据的高维数来说过小,还会导致算法在计算此类数据时出现"过拟合"现象[12]. 吴遐[7]等于 2011 年提出使用 KPCA 对图核诱导的高位空间进行降维,KPCA 是对 PCA 进行核化. 然而由于 KPCA 是一种无监督的降维方法,且在图核函数的定义过程中仅用到图中顶点及边的标签信息,没能利用对图分类更有价值的图数据类别标签等监督信息,而 LDA 由于在机器学习领域的广泛应用,并且在降维的同时能够考虑到数据的类别信息,属于有监督学习算法,因此本文将 LDA 核化,并对传统的经典图核模式发现方法进行扩充,使用核线性判别分析(KLDA)[13]将图核诱导到高维空间中的高维数据进行降维,充分利用图数据自身的类别标签信息;然后对降维后的向量型数据用 SVM 进行分析. KLDA 是将 LDA 进行核化,核化后的 LDA 在新的特征空间中隐式执行,可以使其应用于非线性可分的数据中. 同时 KLDA 继承了 LDA 属性,将图数据的标签信息考虑进去,更有利于图分类,属于有监督学习.

1 图核

本文定义的图 G 为一个具有三元组 (V,E,ξ) 的数据结构. 其中 V 是图中所有顶点的集合,$V = \{v_1,v_2,\cdots,v_n\}$;$E$ 是图中所有无向边的集合,$E = \{e_1,e_2,\cdots,e_m\}$,$\xi$ 是一个函数,用于将图中的顶点及边映射到一个字符表中,即 $\xi:V,E \to \Lambda$,其中 Λ 为字符表构成的集合. 任意节点 v 的邻居节点指同节点 v 直接连接的节点,用 $\zeta(v)$ 表示节点 v 的邻居

节点，则 $\zeta(v)=\{v' \mid (v,v')\in E\}$.

核方法广泛应用于向量型数据的分析中，并且形成了一套成熟的核方法体系. 定义两个图结构之间的核函数即为图核. 通过映射 φ 将原始空间中的图映射到高维甚至无穷维向量空间(特征空间)中，使得

$$k(G_1,G_2)=<\varphi(G_1),\varphi(G_2)> \tag{1}$$

成立. 其中 $\varphi(G)$ 的分量可以是两图中某一公共子路径的条数等. 核 $k:G\times G\to R$ 则可以看成是两个图之间的相似性度量，则含有 N 个图的数据集可以通过以上计算得到 $N\times N$ 的核矩阵 $\boldsymbol{K}=(k_{ij}(G_i,G_j))_{N\times N}$.

本文使用的图核是 Shercashidze 等于 2009 年提出的 Weisfeiler-Lehman 子树核[14]，这种核方法对大规模图数据集有较为理想的计算速度. 给定图 G 及 G'，其第 k 次迭代产生的标号为 $\Lambda_k=\{\delta_{k0},\delta_{k1},\cdots,\delta_{k|\Lambda_k|}\}$，其中 $|\Lambda_k|$ 代表集合 Λ_k 的大小. 定义映射 $\chi_k:\{G,G'\}\times\Lambda_i\to N$，其中 $\chi_k(G,\delta_{kj})$ 是标号 δ_{ij} 在图 G 中出现的个数，N 是自然数集合. 那么，经过 h 次迭代后图 G 和 G' 之间的 Weisfeiler-Lehman 子树图核定义为：

$$kh_{\text{WLsubtree}}(G,G')=<\psi h_{\text{WLsubtree}}(G),\psi h_{\text{WLsubtree}}(G')> \tag{2}$$

式中，

$$\psi^h_{\text{WLsubtree}}(G)=(\chi_0(G,\delta_{01}),\cdots,\chi_0(G,\delta_{0|\Lambda_0|}),\cdots,\chi_h(G,\delta_{h1}),\cdots,\chi_h(G,\delta_{h|\Lambda_h|})),$$
$$\psi^h_{\text{WLsubtree}}(G')=(\chi_0(G',\delta_{01}),\cdots,\chi_0(G',\delta_{0|\Lambda_0|}),\cdots,\chi_h(G',\delta_{h1}),\cdots,\chi_h(G',\delta_{h|\Lambda_h|})).$$

实际情况中，我们得到的是图数据的集合. 假定图数据集合大小为 N，Weisfeiler-Lehman 子树核一次迭代需要遍历集合中的 N 个图数据，即对于由 N 个图数据构成的节点标号集合中任意节点标号，统计每幅图中含有该标号的个数. 已有的工作提出了多种将图数据映射到特征空间的方法，它们共同的缺陷在于这些方法的计算复杂度高，直接导致这些算法的扩展性较差. 虽然 Weisfeiler-Lehman 子树核一次循环需要遍历所有图数据，但其算法复杂度为 $O(Nhm)$，因此该算法能够扩展到具有上千个节点的情况，图 2 中步骤 1 到 5 为 Weisfeiler-Lehman 子树核的一次迭代过程.

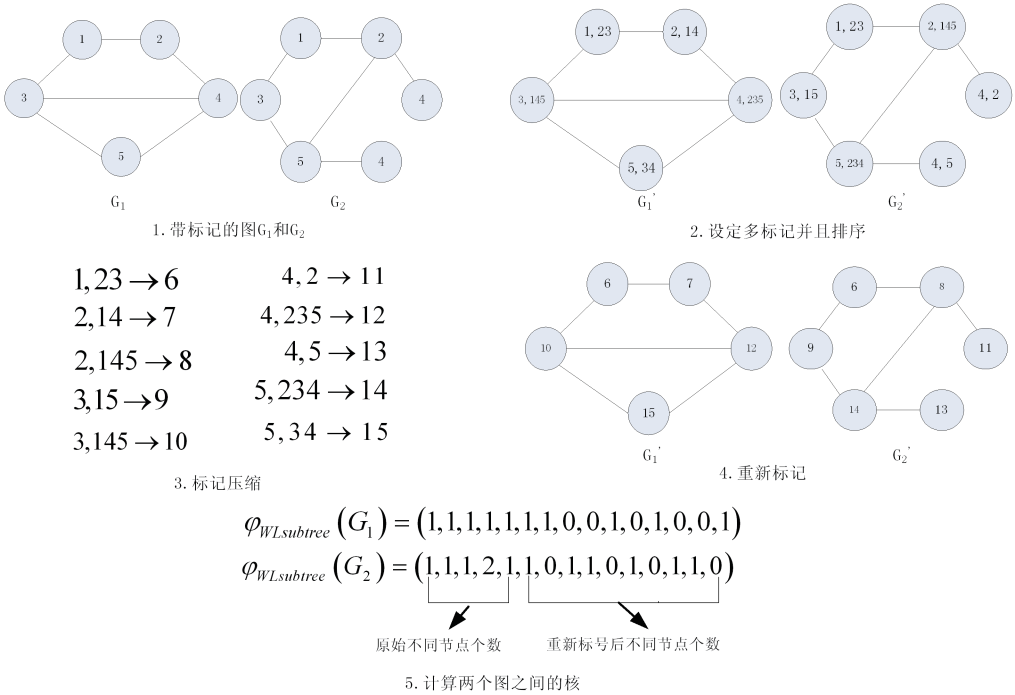

图 2　Weisfeiler-Lehman 子树核核的一次迭代过程

Fig. 2　First Iteration of the Weisfeiler-Lehman subtree kernel

2 基于KLDA的图核降维方法

2.1 KLDA

线性判别分析(LDA)是机器学习、数据挖掘领域经典且热门的一种算法,LDA的原理是将带标签的数据,通过投影映射到维度更低的空间中,使得投影后同类别的数据在投影后的空间中更接近,而不同类别的数据在投影空间中离得更远.直观地说,LDA的思想是寻找一个使得类与类间分离最大化的投影.给定两个带标签的数据集 C_1 和 C_2,均值为:m_1,m_2. $m_i=\frac{1}{l_i}\sum_{n=1}^{l_i} xi_n$. l_i 是 C_i 样本的个数.LDA的目的是投影后类与类之间的中心点尽量分离,同时样本点的方差尽量较小,即最大化

$$\boldsymbol{J}(\boldsymbol{w})=\frac{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_B\boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_w\boldsymbol{w}} \tag{3}$$

式中,$\boldsymbol{S}_B$ 为类间离散度矩阵,$\boldsymbol{S}_w$ 为类内离散度矩阵. $\boldsymbol{S}_B=(m_2-m_1)(m_2-m_1)^{\mathrm{T}}$,$\boldsymbol{S}_w=\sum_{i=1,2}\sum_{n=1}^{l_i}(x_n^i-m_i)(x_n^i-m_i)^{\mathrm{T}}$.

为了求得投影方向,将 $\boldsymbol{J}(\boldsymbol{w})$ 对 $\boldsymbol{w}$ 求导,并使之为0,整理得 $(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_B\boldsymbol{w})\boldsymbol{S}_w\boldsymbol{w}=(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_w\boldsymbol{w})\boldsymbol{S}_B\boldsymbol{w}$.因为我们只关心 $\boldsymbol{w}$ 的方向,而 $\boldsymbol{S}_B\boldsymbol{w}$ 与 (m_2-m_1) 的方向一致,$\boldsymbol{S}_B\boldsymbol{w}$ 可以用 (m_2-m_1) 代替,因此,

$$\boldsymbol{w}\propto\boldsymbol{S}_w^{-1}(m_2-m_1) \tag{4}$$

这就是Fisher提出的线性判别分析[15],故也称为Fisher线性判别分析.

将LDA核化,即得到核LDA(KLDA),假设 φ 为一个到特征空间 F 的非线性映射,为了得到 F 中的线性判别,我们需要最大化

$$\boldsymbol{J}(\boldsymbol{w})=\frac{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_B^{\varphi}\boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_w^{\varphi}\boldsymbol{w}} \tag{5}$$

式中,$\boldsymbol{S}_B^{\varphi}=(m_2^{\varphi}-m_1^{\varphi})(m_2^{\varphi}-m_1^{\varphi})^{\mathrm{T}}$,$\boldsymbol{S}_w^{\varphi}=\sum_{i=1,2}\sum_{n=1}^{l_i}(\varphi(x_n^i)-m_i^{\varphi})(\varphi(x_n^i)-m_i^{\varphi})^{\mathrm{T}}$;并且,$m_i^{\varphi}=\frac{1}{l_i}\sum_{n=1}^{l_i}\varphi(x_n^i)$.注意,此时 $\boldsymbol{w}\in F$,$\boldsymbol{S}_B^{\varphi}$ 和 $\boldsymbol{S}_w^{\varphi}$ 为在特征空间 F 中的相应的矩阵.LDA可以被重写为点积的形式,并且 $\boldsymbol{w}$ 有一个扩展形式[16]:$w=\sum_{i=1}^{l}\alpha_i\varphi(x_i)$.因此,

$$\boldsymbol{w}^{\mathrm{T}}m_i^{\varphi}=\frac{1}{l_i}\sum_{j=1}^{l}\sum_{k=1}^{l_i}\alpha_j k(x_j,x_k)=\boldsymbol{\alpha}^{\mathrm{T}}M_i \tag{6}$$

式中,$(M_i)_j=\frac{1}{l_i}\sum_{k=1}^{l_i}k(x_j,xj_k)$. $\boldsymbol{J}(\boldsymbol{w})$ 的分子可以被写成 $\boldsymbol{w}^T\boldsymbol{S}_B^{\varphi}\boldsymbol{w}=\boldsymbol{w}^{\mathrm{T}}(m_2^{\varphi}-m_1^{\varphi})(m_2^{\varphi}-m_1^{\varphi})^{\mathrm{T}}\boldsymbol{w}=\boldsymbol{\alpha}^{\mathrm{T}}M\boldsymbol{\alpha}$,其中 $\boldsymbol{M}=(M_2-M_1)(M_2-M_1)^{\mathrm{T}}$.同理,$\boldsymbol{J}(\boldsymbol{w})$ 分母 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_w^{\varphi}\boldsymbol{w}=\sum_{j=1,2}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}_j\boldsymbol{K}^{\mathrm{T}}{}_j\boldsymbol{\alpha}-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}_j1_{l_j}\boldsymbol{K}^{\mathrm{T}}{}_j\boldsymbol{\alpha}=\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{N}\boldsymbol{\alpha}$.其中 $\boldsymbol{N}=\sum_{j=1,2}\boldsymbol{K}_j(\boldsymbol{I}-1_{l_j}\boldsymbol{K}^{\mathrm{T}}{}_j)$.所以,$\boldsymbol{J}$ 可以被重写为:

$$\boldsymbol{J}(\boldsymbol{\alpha})=\frac{\boldsymbol{\alpha}^{\mathrm{T}}M\boldsymbol{\alpha}}{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{N}\boldsymbol{\alpha}} \tag{7}$$

对 $\boldsymbol{J}(\boldsymbol{\alpha})$ 求导并使之等于0,整理得

$$(\boldsymbol{\alpha}^{\mathrm{T}}M\boldsymbol{\alpha})\boldsymbol{N}\boldsymbol{\alpha}=(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{N}\boldsymbol{\alpha})M\boldsymbol{\alpha} \tag{8}$$

所以,

$$\boldsymbol{\alpha}=\boldsymbol{N}^{-1}(M_2-M_1) \tag{9}$$

因此一个新数据 x 在 w 上的投影可通过下式计算:

$$y(x)=(w\cdot\varphi(x))=\sum_{i=1}^{l}\alpha_i k(x_i,x) \tag{10}$$

2.2 基于KLDA的图核降维方法

给定数据集 $\{G_i\}N_{i=1}$,核矩阵 $\boldsymbol{K}=(k_{ij})N_{i,j=1}$,其中,$k_{ij}=k(G_i,G_j)=<\varphi(G_i),\varphi(G_j)>$.这里的核可以是任何现有的图核函数,本文使用的是Weisfeiler-Lehman核函数. $\varphi(G)$ 为图 G 映射到高维特征空间的映射函数.对于一个新的图 G,使用式(10)对数据进行降维:

$$w\cdot\varphi(G)=\sum_{i=1}^{N}\alpha_i k(G_i,G) \tag{11}$$

对数据集中的图均通过上述KLDA过程进行降维得到新的数据集 $\{y_i\}N_{i=1}$(N 为样本数目);然后可以使用传统的机器学习算法,如SVM对数据集进行分析.算法过程如下:

算法 2.1 输入:图数据集 G={G_1,G_2,…G_N};

输出:低维向量数据集 Y={y_1,y_2,…y_N};

步骤:

利用现有的Weisfeiler-Lehman核函数计算核矩阵 K=(k_{ij})$N_{i,j=1}$;

求解(9)式,得到最佳投影方向;

对每个图 G_i,计算低维向量数据 $y_i=w\cdot\varphi(G_i)=\sum_{i=1}^{N}\alpha_i k(G_i,G)$.

3 实验

3.1 数据集

本文使用图数据分析中常用的数据集PTC[17]进行实验,PTC是致癌的分子数据集,它含有417

种化合物，每种化合物均用图结构表示. 这 417 种化合物被标记为{EE，IS，E，CE，SE，P，NE，N}中的一种，根据其对雄性大鼠(PTC_MM)、雌性大鼠(PTC_FM)、雄性小鼠(PTC_MR)和雌性小鼠(PTC_FR)是否具有致癌作用可将其分成两类，具有致癌作用称为正类，不具有致癌性称为负类. 其中标记 CE、SE 和 P 被认为为正类，标记 NE 和 N 被认为为负类，EE、IS、E 表示不能确定. 本文实验针对 PTC 数据集中分别取标记为 P 的化合物为正类和标记为 N 的化合物为负类组成样本集进行实验. PTC_MM、PTC_FM、PTC_MR、PTC_FR 中图的平均节点数分别为 25. 05、25. 25、25. 56、26. 08，平均变数分别为 25. 39、25. 62、25. 96、26. 53. 具体信息见表 1.

表 1　图数据集

Tab. 1　Graph data sets

数据集	正类个数	负类个数	总个数	平均节点数
PTC_MM	66 (37.7%)	109 (62.3%)	175	25.05
PTC_FM	77 (41.4%)	109 (58.6%)	186	25.25
PTC_MR	62 (35.6%)	112 (64.4%)	174	25.56
PTC_FR	61 (32.6%)	126 (67.4%)	187	26.08

3.2　实验设置

本文实验中使用的图核函数为 Shervashidze 等[14]于 2009 年提出的 Weisfeiler-Lehman 快速子树核. Weisfeiler-Lehman 快速子树核在运算速度上具有一定的优势，尤其是在对大规模的数据集分析上优势更加明显. 将本文提出的方法与经典的图核方法模式发现方法[11]、吴霞等[7]提出的基于图核的降维方法进行比较. 其中，模式发现方法是将图数据映射到向量空间中之后直接使用 SVM 分类器进行分类，基于图核的降维方法使用核主成分分析方法 KPCA[18]进行降维. 实验中，子树核迭代次数(h)设置为 5. 由于 KLDA 至多可生成 C-1 维子空间(C 为样本类别)，因此，实验中，KLDA 投影到 1 维空间. KPCA 所降维数均为 5 维[7]. 对 PTC 的四个数据集进行 10 倍交叉验证，取数据集的 9 份作为训练集，1 份为测试集，并使用 SVM 作为分类器，每次实验重复 10 次.

3.3　实验结果及分析

表 2 列出了基于图核的降维方法和基于 KLDA 的图核降维方法在 PTC_MM、PTC_FM、PTC_MR 和 PTC_FR 数据集上的平均分类精度以及方差大小. 基于图核的降维方法采用 KPCA 对图核诱导的特征空间进行降维，KPCA 是对 PCA 进行核化，由于 KPCA 是一种无监督的降维方法，且在图核函数的定义过程中仅用到图中顶点及标签的信息，没有用到对图分类更有价值的图数据类别标签等监督信息. 本文使用的 KLDA 方法属于有监督降维方法，使用 KLDA 对图核诱导的特征空间进行降维时将图数据类别标签也考虑进去，因此效果好于 KPCA. 从表 2 中的结果也可以看出，基于 KLDA 的图核降维方法效果确实好于基于图核的降维. 对比经典的图核方法模式发现方法，基于 KLDA 的图核降维方法的分类精度有较高的提升. 实验中，KPCA 和 KLDA 使用的核函数为径向基函数.

表 2　在 PTC 上的平均分类精度比较

Tab. 2　Comparation of mean classification accuracy in PTC

数据集	基于图核的降维(%)	基于 KLDA 的图核降维(%)
PTC_MM	62.19±0.1205	62.43±0.1412
PTC_FM	62.10±0.2047	62.78±0.4450
PTC_MR	65.32±0.0203	65.62±0.0355
PTC_FR	67.29±0.0313	67.52±0.0419

图 3-6 分别是在 PTC_MM、PTC_FR、PTC_FM 和 PTC_MR 上，将基于图核的降维方法与本文提出的方法进行对比的实验结果，“*”线为基于图核的降维方法，“·”线为基于 KLDA 的图核降维方法. 采用 10 倍交叉验证，取数据集的 9 份作为训练集，1 份为测试集，并使用 SVM 作为分类器，每次实验重复 10 次. 从图中可以看出，基于 KLDA 的图核降维在这四个数据集上的分类精度均高于基于图核的降维方法，且与基于图核的降维方法相比较为稳定. 注意到经典图核模式发现方法在这 4 个数据集上的平均分类精度为 61. 0%、61. 0%、62. 8%、66. 7%，进一步验证了降维的必要性.

图 7-10 中，对 KPCA 和 KLDA 使用多项式函数，然后在 PTC_MM、PTC_FR、PTC_FM 和 PTC_MR 上进行的对比实验结果，实验设置同上. “*”线为基于图核的降维方法，“·”线为基于 KLDA 的图

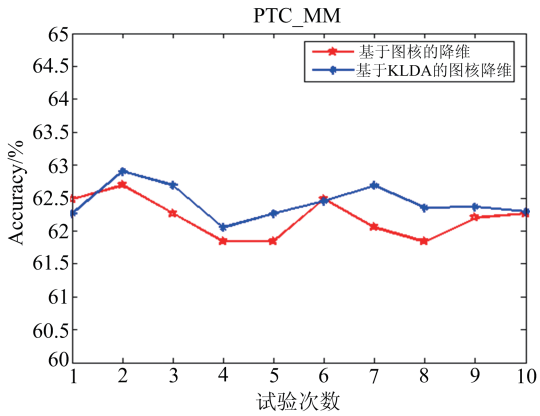

图 3　PTC_MM 上的平均准确率

Fig. 3　Average accuracy on PTC_MM

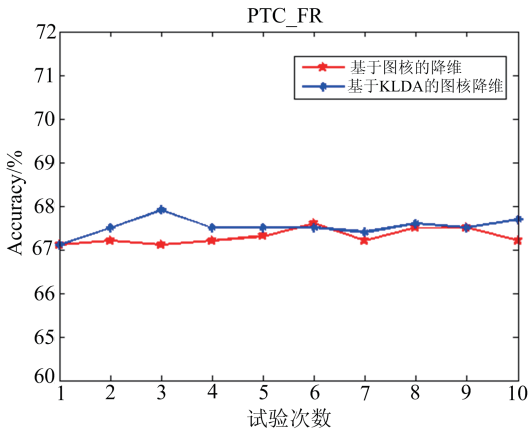

图 4　PTC_FR 上的平均准确率

Fig. 4　Average accuracy on PTC_FR

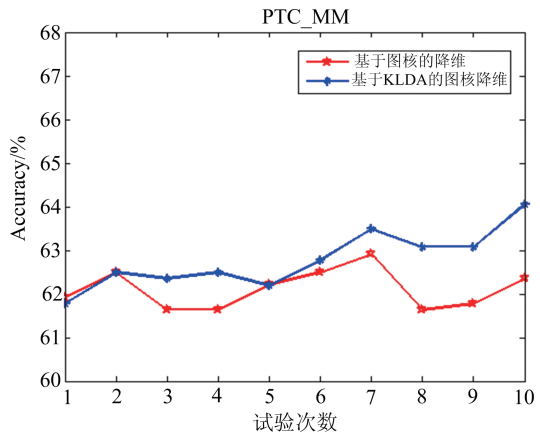

图 5　PTC_FM 上的平均准确率

Fig. 5　Average accuracy on PTC_FM

核降维方法. 由实验结果可以看出，使用多项式核作为核函数的情况下，基于 KLDA 的图核降维的效果也比基于图核的降维方法好.

为了增加方法的有效性验证，使得结果更具有说服性，我们比较了 KLDA 与 KPCA 维度相同时

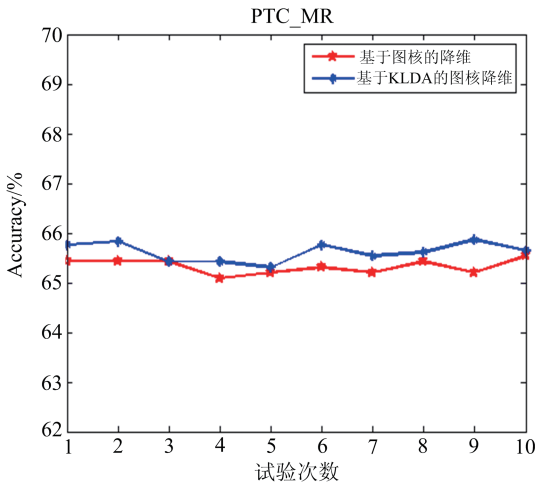

图 6　PTC_MR 上的平均准确率

Fig. 6　Average accuracy on PTC_MR

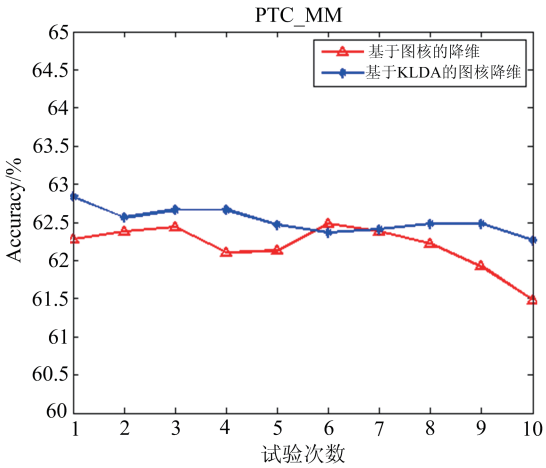

图 7　PTC_MM 上的平均准确率

Fig. 7　Average accuracy on PTC_MM

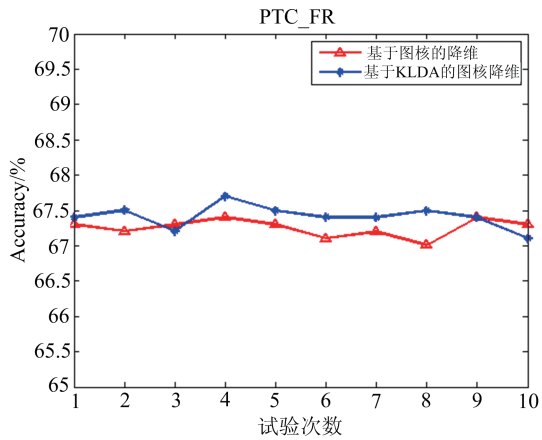

图 8　PTC_FR 上的平均准确率

Fig. 8　Average accuracy on PTC_FR

的性能，由于 KLDA 在二分类时至多可生成 1 维子空间，因此在试验中，将 KLDA 与 KPCA 均投影到

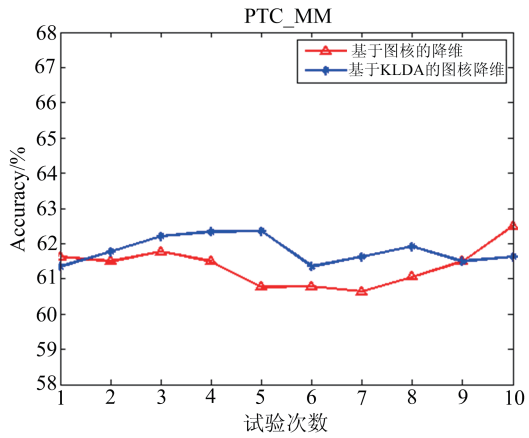

图 9　PTC_FM 上的平均准确率

Fig. 9　Average accuracy on PTC_FM

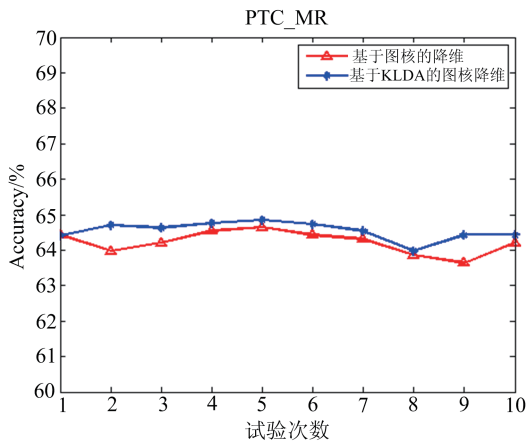

图 10　PTC_MR 上的平均准确率

Fig. 10　Average accuracy on PTC_MR

1 维空间中进行比较. 图 11-14 显示了 KPCA 和 KLDA 使用径向基核函数，将图数据映射到 1 维空间中的精度对比，从图中可以看出，在同维度下，经过 KLDA 降维的分类精度明显高于 KLDA.

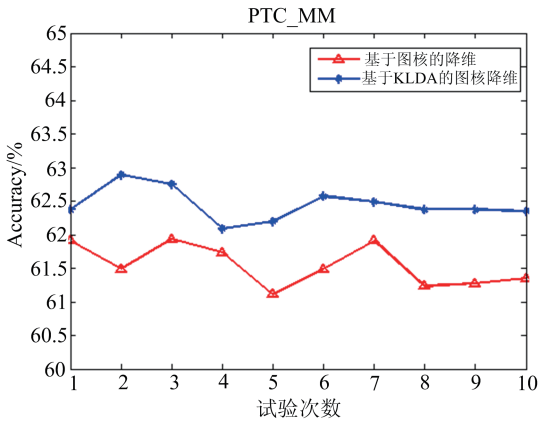

图 11　PTC_MM 上的平均准确率

Fig. 11　Average accuracy on PTC_MM

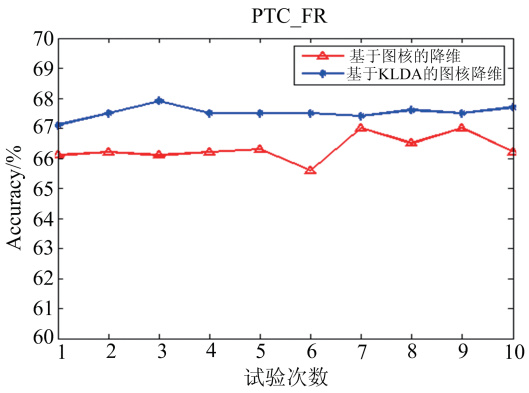

图 12　PTC_FR 上的平均准确率

Fig. 12　Average accuracy on PTC_FR

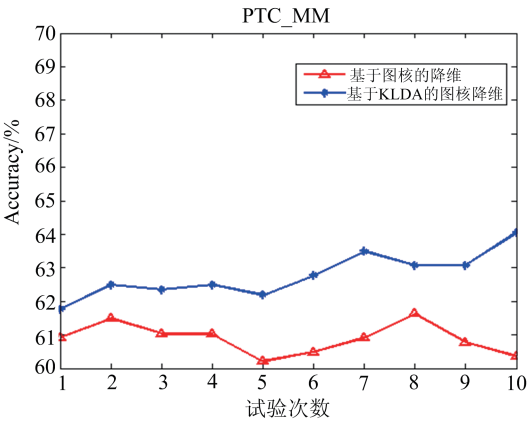

图 13　PTC_FM 上的平均准确率

Fig. 13　Average accuracy on PTC_FM

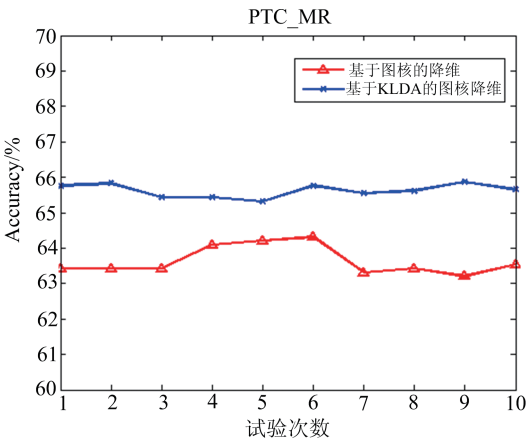

图 14　PTC_MR 上的平均准确率

Fig. 14　Average accuracy on PTC_MR

4　结论

本文将图数据利用 KLDA 的方法进行降维，得到低维向量数据，使之能够使用传统的机器学习方法进行分析. 并在图数据分析中常用的 PTC 数据集上对该方法的分类精度进行验证. 实验结果显示，降

维之后的分类效果好于直接将特征空间数据进行分类的结果,且本文使用的KLDA降维的分类精度好于使用KPCA降维.在今后的工作中,将尝试使用稀疏方法对特征进行特征选择,使之在保持较高分类精度的同时,降低训练和测试时间,以便能够扩展到具有大规模图数据的应用中.

参考文献(References)

[1] 蒋强荣. 图核及其在模式识别中应用的研究[D]. 北京工业大学, 2011.

[2] SHARAN R, IDEKER T. Modeling cellular machinery through biological network comparison [J]. Nature Biotechnology, 2006, 24(4): 427-433.

[3] KUBINYI H. Drug research: Myths, hype and reality [J]. Nature Reviews: Drug Discovery, 2003, 2(8): 665-668.

[4] KUMAR R, NOVAK J, TOMKINS A. Structure and evolution of online social networks[A]// Link Mining: Models, Algorithms, and Applications. New York: Springer, 2010: 337-357.

[5] VISHWANATHAN S V N, SCHRAUDOLPH NN, KONDOR R, et al. Graph kernels [J]. Journal of Machine Learning Research, 2010, 11(2): 1201-1242.

[6] CONTE D, FOGGIA P, SANSONE C, et al. Thirty years of graph matching in pattern recognition [J]. International Journal of Pattern Recognition and Artificial Intelligence, 2004, 18(3): 265-298.

[7] 吴遐, 张道强. 半监督图核降维方法[J]. 计算机科学与探索, 2010, 4(7): 629-636.

[8] BORGWARDT K M, KRIEGEL H P. Shortest-path kernels on graphs [C]// Proceedings of the 5th International Conference on Data Mining. Houston: IEEE Computer Society, 2005:74-81.

[9] HORVÁTH T, GÄRTNER T, WROBEL S. Cyclic pattern kernels for predictive graph mining [C]// Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle: ACM Press, 2004: 158-167.

[10] SHERVASHIDZE N, BORGWARDT K. Fast subtree kernels on graphs [A]// Advances in Neural Information Processing Systems. Vancouver, 2009: 1660-1668.

[11] KRAMER S, DE RAEDT L. Feature construction with version spaces for biochemical applications[C]// Proceedings of the 18th International Conference on Machine Learning. Williamstown: Morgan Kaufmann, 2001: 258-265.

[12] 吴遐. 基于约束的图核方法的研究[D]. 南京航空航天大学, 2011.

[13] MIKA S, RÄTSCH G, WESTON J, et al. Fisher discriminant analysis with kernels[C]// Proceedings of the Neural Networks for Signal Processing. IEEE Press, 1999: 41-48.

[14] SHERVASHIDZE N, SCHWEITZER P, VAN LEEUWEN E J, et al. Weisfeiler-lehman graph kernels[J]. Journal of Machine Learning Research, 2011, 12(9): 2539-2561.

[15] DUDA R O, HART P E, STORK D G. Pattern Classification [M]. 2ed, New York: Wiley-Interscience, 2000.

[16] SCHÖLKOPF B, HERBRICH R, SMOLA A J. A Generalized Representer Theorem[A]// Computational Learning Theory. Berlin Heidelberg: Springer, 2001: 416-426.

[17] HELMA C, KING R D, KRAMER S, et al. The predictive toxicology challenge [J]. Bioinformatics, 2001, 17(1): 107-108.

[18] SCHÖLKOPF B, SMOLA A J. Learning with Kernels [M]. Cambridge, MA: MIT Press, 2002.