用于图学习的主干图核方法

常新功,沈 亮,景丽荣
(山西财经大学 信息管理学院,山西 太原 030006)

摘 要:对于结构化数据的学习是数据挖掘领域一个重要的分支。至今,出现了许多十分优秀的结构化数据学习方法。核方法是其中有效的学习方法之一,文中在Gäertner等人研究的基础上,提出了一种主干图核方法。该方法定义了图中重要程度较高的子结构为主干图,它有效地降低了图学习的规模。利用随机路径核函数来定义主干图核函数并对不同阶的主干图给予不同的权重。通过自适应的离散粒子群算法来对核相似矩阵进行学习。实验结果表明,文中方法能够很好地对图数据进行学习。

关键词:机器学习;核方法;主干图核;子结构;粒子群算法
中图分类号:TP181 **文献标识码**:A **文章编号**:1673-629X(2011)08-0117-04

Backbone Kernels for the Graph Data Learning

CHANG Xin-gong, SHEN Liang, JING Li-rong
(Information Management Faculty, Shanxi University of Finance and Economics, Taiyuan 030006, China)

Abstract: Learning structured data is an important branch of the data mining field. So far, there have been many good methods of structured data learning. Kernel method is one of the effective learning ways. Based on Gäertner and other researchers' study, proposes a backbone graph kernel method. It defines that the sub-structure with higher importance is the backbone graph. It effectively reduces the size of graph learning. It uses random path kernel function to define the main graph kernel functions and gives different weights to backbone graph which have different order. Uses the adaptive discrete particle swarm algorithm to learn the similar kernel matrix. Experimental results shows that the method that the paper proposed can well learn of the graph data.

Key words: machine learning; kernel method; backbone kernel; substructure; particle swarm optimization

0 前 言

Herbert A. Simon [1]和Newman[2]在其文中提到许多真实世界的系统都能用结构化数据表示,比如:函数结构、XML结构化、群体社会网络、国际贸易中的国家关系、化合物分子结构等,在这些结构中可能存在着许多有用的信息,诸如:化合物分子中有毒的分子结构,经济圈的划分等。近年来,众多研究者越来越关注结构化数据的挖掘和学习。

图作为一种典型的结构化数据,对它的学习主要包括:频繁子结构发现,图分类、图聚类等,关于图数据挖掘的方法包括很多。其中,核方法(Kernel Methods)通过将数据映射到高维特征空间,然后在新的特征空间中分析和处理数据。由于该方法只需要构建一个核函数来度量样本之间的相似度,与其数据的表现形式并无联系,并且它能够很好地保证其泛化性能,因此,它能够有效地对图这样的结构化数据进行机器学习。

文中在Gäertner[3]、S. V. N. Vishwanathan[4]等人的基础上,提出了一种主干图核(BackBone Kernels)方法。该方法将图数据结构分层次处理,提取出图中重要程度较高的子结构数据,减少图数据挖掘时的计算复杂程度。另外,引入自适应种群的离散粒子群算法[5]来对核相似矩阵进行学习。通过实验表明,上述方法取得较好的聚类效果,特别是在规模较大的图数据结构集上有着良好的表现。

1 研究背景

1.1 图核函数

至今,已有许多研究者提出了各种图核算法。1999年,David Haussler[6]提出卷积核,它比较2个结构型数据的所有不同的分解,对于不同的分解有着不同的图核函数。2003年,Gäertner, Flach, Wrobel[3]等人指出得到完美图核是一个至少和图同构一样难的问题,计算任何子结构同构也是一个NP难的问题,并说

收稿日期:2011-01-10;**修回日期**:2011-05-09
基金项目:国家自然科学基金资助项目(60873100);山西省高校科技研究与开发项目(20081023);山西省自然科学基金资助项目(2010011022-1)
作者简介:常新功(1968-),男,山西太原人,教授,硕士生导师,CCF会员,研究方向为进化计算、数据挖掘;沈 亮,硕士研究生,研究方向为进化算法、数据挖掘。

明对于图相似性的问题是找到其近似解。Gäertner 等人在该文章中还提出了一种具有时间复杂度为 $O(n^6)$ 开销的随机路径核算法，通过其实验论证随机路径核能够有效地描述图之间的相似性，但是存在时间开销较大、计算冗余等问题。2005 年，Fröhlich. H[7] 等人提出了最佳完美匹配核函数，它将学习样本提取出特定的子结构再进行相似性计算。Karsten M. Borgwardt、Hans-Peter Kriegel[8] 提出了另外一种基于弗洛伊德算法的最短路径核函数。该核函数具体多项式的时间复杂度，并且能够有效地避免冗余。2006 年，S. V. N. Vishwanathan[4] 等人在随机路径核函数的基础上提出了一种快速的随机路径图核函数，在文中作者指出，该图核函数的理论时间复杂度为 $O(n^3)$，与 Gäertner 等人的随机路径核函数相比，大大提高了核函数的效率。上述几种都是基于路径相关的图核方法，此外，还有一些研究者提出加权分解核[9]、基于子树的图核方法[10] 等。

2010 年，Adam Wo źnica、Alexandros Kalousis[11] 等人提出了一种自适应的图核函数。该图核函数以路径核、子树核为基础，它能够很好地分解不同的图结构，并且对不同的子结构上采用特定类型之间的映射。最后在文中指出一个标准的图核函数应该具备以下 4 个条件：

1）能够很好地计算图之间的相似性。

2）计算的时间复杂度应该在多项式时间内。

3）核矩阵应该满足半正定。

4）能够应用在不同种类的图数据结构上。

1.2 核矩阵学习

核方法学习包括 2 个关键问题。第一个问题是对于核函数的构造，另外一个问题是对于核相似矩阵的学习。研究者针对核相似矩阵提出很多优秀的学习方法。文中采用离散的粒子群算法来对核矩阵进行学习，粒子群算法是一种群体智能算法，它没有很多参数需要调整，十分简洁，易于实现。在学习过程中通过粒子之间信息的交互能够不断地向最优解靠拢。实验表明离散的粒子群算法能够很好地对核矩阵进行学习。

2 文中方法

2.1 相关定义

定义 1 带标签的图（labeled graph）：带标签的图是对一个图（可以是有向图、无向图或混合图，文中所有涉及的实例和实验都在无向图上开展）的节点和边加了标签（label）后的图，它可以表示为一个四元组 $G=(V,E,L,lab)$。其中，V 为节点集合，$E\subseteq V\times V$ 为边集合，L 是标签集合。标签函数 lab：$V\cup E\rightarrow L$ 赋于节点和边以不同的标签以定义对象之间的关系。

定义 2 主干图（backbone graph）：在带标签的图 $G(V,E,L,\text{lab})$ 中，若节点 v_i 的度大于等于某一个阈值 ε（默认 $\varepsilon=2$）时，称该顶点为主干顶点，由主干顶点组成的图称之为主干图。在图 G 中若存在环状子结构，将环状子结构以一节点 V_j 替代，并且该子结构的父节点设为 V_j。然后再进行主干图的查找，在相似性的时候以该子结构来进行计算。示例如图 1：

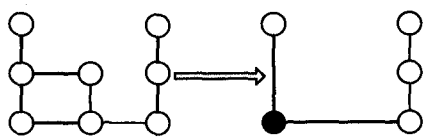


图 1 含有环状子结构的图 G 与其处理后的图 G'

其中，节点 V_j 的子图如图 2：

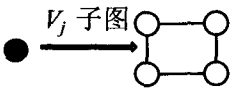


图 2 顶点 V_j 的子图

定义 3 n 阶主干图：对于带标签的图 $G(V,E,L,\text{lab})$，经过 n 次的主干图搜索后，称之为 n 阶主干图，记为 G^n。n 阶主干图的主要算法如表 1 所示。

表 1 n 阶主干图及其子图的查找过程

```
输入：无向图 G(V,E,L)，阶数 n，A 为图 G 的邻接矩阵，subGraph 为子图类
输出：关于图 G 的 n 阶主干图，称之为 G^n，相对的阶数产生的子图 subG^n 集合
算法描述：
Get_N_BackBoneGraph(Graph G , int n)
{
  if(n==1) reutrn G;
  n--;
  subGraph subG;
  for(int i=1;i<=G.NumVertical;i++)
   if(G.degree(V(i))>=1)
      {
          G.degree(V(i))--;
          for(int j=1;j<=G.NumVertical;j++)
    if(G.A(i,j)==1 and G.degree(V_j)>=ε)
                  {G.A(i,j)--;subG.father=V_j;subG.A(i)
=G.A(i);}
      }
  Get_N_BackBoneGraph(G,n);
}
```

设 $\varepsilon=2$，则如下无向图的 1，2，3 阶主干图如图 3 所示。

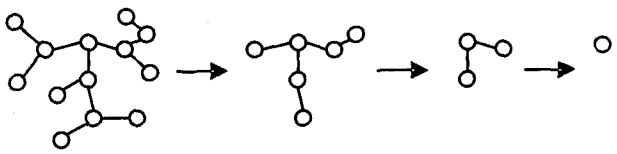
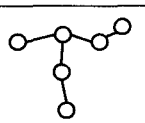
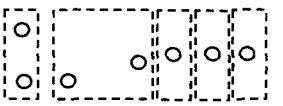
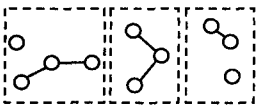
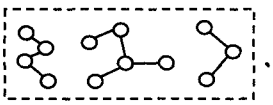


图 3 图 G 的 1、2、3 阶主干图

其对应的主干图的子图如表 2 所示。

表 2　图 G 的 1、2、3 阶主干图及其子图

N	N阶主干图	N阶主干图中各个顶点包含的子图
1		
2		
3		

2.2　主干图核

2.2.1　主干图核

对于 2 个无向图 G 和 G_1，令 $G' = \{G^1, G^2, \cdots, G^{n1}\}$ 为 G 的主干图集合，同理 $G_1' = \{G_1^1, G_1^2, \cdots, G_1^{n2}\}$ 为 G_1 的主干图集合。另外，$\mathrm{sub}G = \{\mathrm{sub}G^1, \mathrm{sub}G^2, \cdots, \mathrm{sub}G^{m1}\}$ 和 $\mathrm{sub}G_1 = \{\mathrm{sub}G_1^1, \mathrm{sub}G_1^2, \cdots, \mathrm{sub}G_1^{m2}\}$ 分别为 G 和 G_1 的主干图对应的顶点的子图。

定义 4（主干图核）　对于一个半正定、对称的图核函数 $K(\,)$，定义主干图核为：

$$K_{\mathrm{backbone}}(G, G_1) = \sum_{i=2}^{m} \partial(i)\left(K(G^i, G_1^i) + \frac{1}{|G^i|} \max_j \sum_{k=1}^{|G^i|} K(\mathrm{sub}G^k, \mathrm{sub}G_1^j)\right)$$

上述表达式的前半部分定义了主干图的相似性，后半部分定义了主干图顶点的各个子图之间的相似性。另外，对于不同阶的主干图赋予不同的权重。其中 $\partial \sim \pi(|\frac{m}{2}|)$，故 $\partial(i) = \frac{\lambda^k e^{-\lambda}}{i!}$，并且 $\lambda = |\frac{m}{2}|$。

2.2.2　复杂度和半正性分析

●复杂度分析。

文中的 $K(\,)$ 采用了快速的随机路径图核函数[4]，并使用深度优先策略来生成图中的随即路径，图的结构采用邻接表存储。其时间复杂度为 $O(n+e)$，其中 n 为图的顶点数，e 为图的边的条数。在文献中指出，最短路径核函数的的时间复杂度为 $O(n^3)$，其中 n 为图的顶点数。令主干图核函数的主干图和各个子图的规模为：$n_1, n_2, n_3 \cdots n_k$，并且 $n_1 + n_2 + n_3 \cdots + n_k = n$，那么有：

$$kO(n_1^3) + kO(n_2^3) + kO(n_3^3) + \cdots + kO(n_k^3) \leqslant k^2 \max_j O(n_j^3) \leqslant k^2 O(n^3)$$

故，主干图核函数的时间复杂度为 $O(n^3)$。

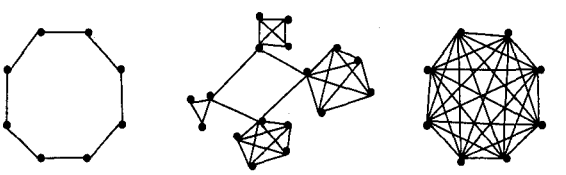
●半正定分析。

文中采用的快速的最短路径核函数[4]在参考文献已经证明了为一个半正定、对称的图核函数。完全类似地，可以证明主干图核为半正定核。

2.3　离散的粒子群算法

在 Eberhart、Kennedy[12] 等人的基础上，采用自适应种群的离散粒子群聚类算法[5]来对核相似矩阵进行学习。其中，聚类结果表示一个解，通过不断地迭代查找从而找到最优的聚类结果。在查找过程中，为了扩大粒子群的搜索范围，在搜索前期使用环拓扑结构，加大对全局范围内的搜索以避免过早陷入局部最优。随着迭代次数的增加，粒子群簇数逐渐减少。在寻找最优解的中期，为了既兼顾粒子群的全局搜索能力，亦考虑到其局部搜索最优解的能力，采用自适应种群的拓扑结构模型，在小种群内采用全连接，种群之间采用环形连接。当粒子群的簇数到一定程度时，其拓扑结构趋向于全连接型拓扑结构结构，着重在一个最可能出现最优解的区域进行搜索。

拓扑结构如图 4 所示。



环型拓扑结构　　自适应种群拓扑结构　　全连接型拓扑结构

图 4　拓扑结构模型

3　实　验

3.1　实验数据

Mutagenesis 是诱变分子数据集，它包含 188 个学习样本，分为 2 类：一类为诱变分子，其余的为非诱变分子。

PTC 是致癌分子数据集，它包含 470 个样本集，分为 4 类：Male Mouse（MM），Female Mouse（FM），Male Rat（MR）和 Female Rat（FR）。每一个样本分子共有 8 种标记：EE、IS、E、CE、SE、P、NE、N。

HIA 是人体小肠吸收分子数据集，该数据集包含 164 个样本集。分成 2 组分类：一类是 high oral bio-availability，另一类是 low oral bio-availability。该数据集分子规模较大。

3.2　实验设置

文中选用了最短路径核函数、最佳匹配核函数，以及自适应的图核函数进行对比测试。获取主干图时阈值 $\varepsilon=2$。在各个相似矩阵后均采用自适应种群的离散粒子群算法对其进行聚类。其中离散粒子群算法的参数设置为：$W1 = 0.729$、$W2 = 2.187$，种群规模为 30，迭代次数为 1000 次。考察独立运行 100 次实验结果的最优值，均值，最差值。以及计算核相似矩阵所花费的时间。

表 3 在各个数据集上的聚类实验结果

DataSet	Algorithm	Worst(%)	Mean(%)	Best(%)
Mutagenesis	Shortest-path kernels	77%	82%	85%
	Optimal Assignment Kernels	74%	80%	83 %
	Adaptive Matching Based Kernels	80%	83%	85%
	文中方法	78%	82%	86%
PTC	Shortest-path kernels	70%	79%	82%
	Optimal Assignment Kernels	69%	78%	82%
	Adaptive Matching Based Kernels	75%	80%	83%
	文中方法	76%	81%	85%
HIA	Shortest-path kernels	80%	83%	85%
	Optimal Assignment Kernels	82%	84%	87%
	Adaptive Matching Based Kernels	82%	84%	88%
	文中方法	83%	86%	88%

3.3 实验结果

由于实验环境以及代码的编辑工作与参考文献有差别，故文中的实验结果与参考文献中的实验结果有一定的出入。表 3 给出了在数据集上采用自适应种群的离散粒子群聚类算法的实验结果，图 5 给出了各个图核函数在不同数据集上的运行时间。

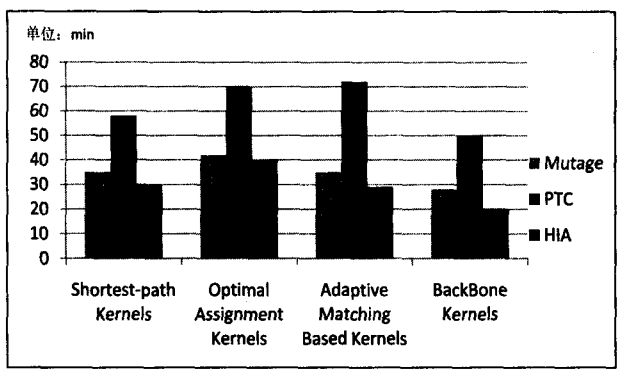


图 5 图核函数在数据集上的运行时间

从上述实验结果可以得出主干图函数在 PTC、HIA 两个数据集上均优于其他图核函数，在 Mutagenesis 数据集上也与自适应的图核函数相差不大。另外，在计算时间方面，主干图核函数明显好于其他图核函数。

4 结束语

在文献[3,4]的基础上，提出了主干图函数，其主要工作在于提取图结构数据中重要程度较高的子结构进行运算，它能有效地减少图学习过程中的复杂程度，并且减少计算过程中的大量冗余。通过实验验证，在不同数据集都有着较优的表现。

参考文献：

[1] Simon H A. The architecture of complexity[J]. Proceedings of the American philosophical Society, 1962,106(6):467-482.

[2] Newman M E J. The structure and function of complex networks[J]. SIAM Review,2003,45(2):167-256.

[3] Gäertner,Flach P,Wrobel S. On graph kernels: Hardness results and efficient alternatives[C]//The 16th Annual Conference on Computational Learning Theory. [s. l.]:[s. n.], 2003:129-143.

[4] Vishwanathan S V N, Borgwardt K, Schraudolph N N. Fast Computation of Graph Kernels [R]. NICTA, 2006:1-9.

[5] 沈 亮，常新功，景丽荣. 自适应种群的高斯动态粒子群聚类算法[J]. 计算机系统应用，2010,19(8):112-116.

[6] Haussler. Convolutional kernels on discrete structures[R]. [s. l.]:[s. n.], 1999.

[7] Fröhlich H, Wegner J K, Siker F, et al. Optimal Assignment Kernel functions for attributed Molecular Graphs[C]//ICML '05 Proceedings of the 22nd International Conference on Machine Learning. [s. l.]:[s. n.],2005:225-232.

[8] Borgwardt K M, Hans-Peter K. Shortest-path kernels on graphs[C]//Proceedings of 5th IEEE International Conference on Data Mining (ICDM'05). Houston:[s. n.],2005:74-81.

[9] Menchetti S, Costa F, Frasconi P. Weighted Decomposition Kernels[C]//Proceedings of the 22 nd International Conference on Machine Learning. [s. l.]:[s. n.],2005.

[10] Shervashidze N, Borgwardt K M. Fast subtree kernels on graphs[J]. Advances in Neural Information Processing Systems,2009(22):1-9.

[11] Woznica A, Kalousis A, Hilario M. Adaptive Matching Based Kernels for Labeled Graphs[J]. Lecture Notes in Computer Science,2010(6119),374-385.

[12] Kenndy J, Eberhart R C. Particle swarm optimization[C]//In: Proc. of the IEEE Int'l Conf. on Neural Networks. Perth: IEEE Press,1995:1942-1948.