第 43 卷 第 4 期 计 算 机 工 程 2017 年 4 月
Vol. 43 No. 4 Computer Engineering April 2017

·人工智能及识别技术· 文章编号：1000-3428(2017)04-0194-06 文献标志码：A 中图分类号：TP18

基于模型空间的树形数据处理方法

董亚东，李正宇，汪 阳

（中国科学技术大学 计算机科学与技术学院，合肥 230001）

摘 要： 树形结构数据包括多个结点的属性信息与结点间的连接结构信息，然而传统的机器学习对树形数据的处理方法比较单一。为此，提出一种适用于树形结构的树形回声状态网络方法，使用树形回声状态网络对树形结构数据进行建模，得到固定维数的空间模型，从而将复杂的树形结构数据转换为模型空间中的点。基于模型空间的思想，通过模型空间中点的距离来度量树形结构数据之间的相似度，并将模型与核方法相结合以提高分类器的判别能力。实验结果表明，树的回声状态网络方法与传统方法相比，在相关数据集上有着较好的测试性能。

关键词： 机器学习；回声状态网络；水库；模型空间；树结构数据

中文引用格式： 董亚东，李正宇，汪 阳. 基于模型空间的树形数据处理方法[J]. 计算机工程，2017，43(4)：194-199，206.

英文引用格式： Dong Yadong，Li Zhengyu，Wang Yang. Tree-structured Data Processing Method Based on Model Space[J]. Computer Engineering，2017，43(4)：194-199，206.

Tree-structured Data Processing Method Based on Model Space

DONG Yadong，LI Zhengyu，WANG Yang

（School of Computer Science and Technology，University of Science and Technology of China，Hefei 230001，China）

【Abstract】 Classical machine learning methods are not enough for dealing with tree data because the tree contains not only node information but also structure information. Therefore，this paper proposes an approach of tree echo state network，applicated to tree-structured，it uses the tree-structured echo state network to model tree-structured data，gets a fixed-size space model，and converts the complex tree-structured data into points in the model space. Based on the idea of model space，the similarity between the tree-structured data is measured by the distance between the models. It combines the model with the kernel methods to facilitate classification performance. Experimental results show that，compared with traditional algorithms，the tree echo state network has better performance in related datasets.

【Key words】 machine learning；echo state network；reservoir；model space；tree-structured data

DOI： 10.3969/j.issn.1000-3428.2017.04.033

0 概述

树形结构数据作为一种维数可变化的结构形数据，相比向量数据更适合描述现实世界中事物之间的关系，是一种在工程应用中常见的数据表示方式。在计算机领域中，XML 和 HTML 结构数据就是一种树形结构数据。在生物信息学领域中，蛋白质结构数据可以被看成是一种树形结构数据。在化学领域中，通常将某些结构形的化学物质数据看成是树形结构数据，例如本文实验部分的烷烃化合物、多糖化合物以及最近研究火热的树状大分子[1]材料等。

传统的机器学习算法主要适用于解决一些平面型数据[2]。然而树形结构数据不仅包含结点信息，而且还包含了结点间的结构化信息。所以，使用传统的机器学习算法处理树形结构数据可能导致原始数据中关系信息的丢失，数据的结构信息也难以表达，而且需要相关领域的专家设计相应结构化数据的矢量化方法。由于树形结构数据在现实生活中的普遍存在，使得使用机器学习算法处理此类树形结构数据成为一个被广泛研究的领域，例如使用关系数据挖掘的方法[3]、基于树距离的学习方法[4-5]、统计关系学习方法[6-7]、概率学习方法[8-9]、基于核与神经网络方法[10]等。在基于核与神经网络方法中经常使用方法是基于子树的核方法[2]，它的基本思想是把树映射到子树空间中，如果两棵树的公共子树越多，或映射后在子树空间内越接近，那么这两棵树就越相似。然而，目前针对子树的核方法存在着算法复杂度高、特征重复计算以及应用范围的局限性等问题。

作者简介： 董亚东（1990－），男，硕士研究生，主研方向为机器学习；李正宇、汪 阳，硕士研究生。

收稿日期： 2016-04-26 **修回日期：** 2016-05-27 **E-mail：** guage@mail.ustc.edu.cn

本文在回声状态网络的基础上,提出一种适用于树形结构数据的树形回声状态网络方法,并使用模型空间的思想对通过树形回声状态网络建模得出的向量数据进行处理,从而实现对原始树形结构数据的处理。树形回声状态网络作为一般的回声状态网络来说,是由不需要训练的迭代非线性水库和线性的输出权重组成,所以它是一种线性的训练算法,在算法的时间复杂度上较优。在树形回声网络中,首先对树形结构数据使用自底向上(从叶子结点到最后的根结点)的方法,对每个结点求出其水库状态,然后使用基于根或者基于结点平均的两种状态映射函数将树形结构数据转化为模型空间中的向量数据,最后使用模型空间的思想度量模型的距离达到对树形数据分类的目的。

1　回声状态网络

回声状态网络[11]是一种特殊的循环神经网络,它和一般的神经网络相比具有 3 个特点[11]:1)它的中间层由一个随机生成而且无需训练的循环神经网络,称为水库(reservoir);2)输出权值是唯一需要学习的部分;3)网络的训练是线性的。

1.1　回声状态网络的结构和工作原理

回声状态网络的核心就是一个水库。所谓的水库就是随机生成的、大规模的、稀疏连接(连接率在 1% ~5% 之间)的递归结构,如图 1 所示。

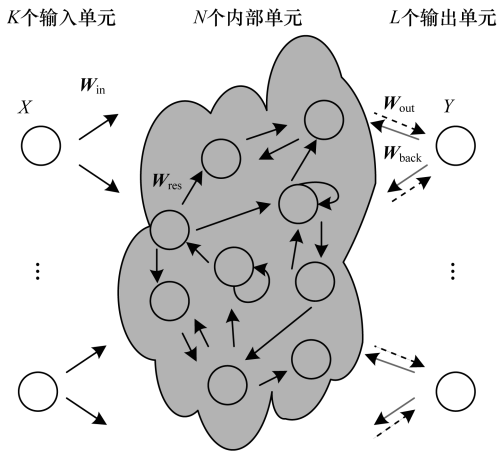

图 1　回声状态网络的结构

从结构上讲,回声状态网络是一种特殊类型的循环神经网络。其思想是用大规模随机连接的循环网络,取代经典神经网络的中间层,从而简化网络的训练过程。基于图的结构,假定系统有 K 个输入单元、N 个内部神经元和 L 个输出单元,那么输入单元 $\boldsymbol{u}(t)$ 、内部状态 $\boldsymbol{x}(t)$以及输出单元 $\boldsymbol{y}(t)$ 在时刻 t 的值分别为[11]:

$$\boldsymbol{u}(t)=[u_1(t),u_2(t),\cdots,u_K(t)]^{\mathrm{T}} \tag{1}$$

$$\boldsymbol{x}(t)=[x_1(t),x_2(t),\cdots,x_N(t)]^{\mathrm{T}} \tag{2}$$

$$\boldsymbol{y}(t)=[y_1(t),y_2(t),\cdots,y_L(t)]^{\mathrm{T}} \tag{3}$$

则回声状态网络的状态方程[12]为:

$$x(t+1)=f(\boldsymbol{W}_{\mathrm{res}}x(t)+\boldsymbol{W}_{\mathrm{in}}u(t)+\boldsymbol{W}_{\mathrm{back}}y(t)) \tag{4}$$

$$y(t+1)=f_{\mathrm{out}}(\boldsymbol{W}_{\mathrm{out}}x(t+1)+\boldsymbol{W}_{\mathrm{bias}}^{\mathrm{out}}) \tag{5}$$

其中,$\boldsymbol{W}_{\mathrm{res}}$,$\boldsymbol{W}_{\mathrm{in}}$,$\boldsymbol{W}_{\mathrm{back}}$分别表示状态变量、输入和输出对状态变量的连接权值矩阵;$\boldsymbol{W}_{\mathrm{out}}$表示输出权值矩阵;$\boldsymbol{W}_{\mathrm{bias}}^{\mathrm{out}}$表示输出偏置;$f$ 表示内部神经元激活函数,通常情况下取双曲正切函数 tanh[11];f_{out}表示输出函数,通常取恒等函数[5]。在网络的训练中 $\boldsymbol{W}_{\mathrm{res}}$,$\boldsymbol{W}_{\mathrm{in}}$,$\boldsymbol{W}_{\mathrm{back}}$随机产生且保持不变。而输出权值矩阵 $\boldsymbol{W}_{\mathrm{out}}$需要经过训练得到。

1.2　回声状态网络训练过程

回声状态网络的训练过程就是通过初始化 $\boldsymbol{W}_{\mathrm{res}}$ 和 $\boldsymbol{W}_{\mathrm{in}}$,然后求 $\boldsymbol{W}_{\mathrm{out}}$的过程。

1.2.1　水库初始化

$\boldsymbol{W}_{\mathrm{in}}$的初始化:$\boldsymbol{W}_{\mathrm{in}}$中的元素可以在 $[-scale, scale]$之间均匀分布,这里 $scale$ 表示输入权重的伸缩尺度[12]。

$\boldsymbol{W}_{\mathrm{res}}$的初始化:1)随机生成矩阵 $\boldsymbol{W}_1$;2)求出其谱半径 ρ;3)令 $\boldsymbol{W}_{\mathrm{res}}=\alpha\dfrac{\boldsymbol{W}_1}{\rho}$,其中 $\alpha\in(0,1]$,最后得到的 $\boldsymbol{W}_{\mathrm{res}}$谱半径就是 α。在 $\boldsymbol{W}_{\mathrm{res}}$的初始化过程中,令 $\boldsymbol{W}_{\mathrm{res}}$谱半径在区间 (0,1]的原因如下:回声状态网络的状态转移方程可以简单地表达为 $x(t+1)=\boldsymbol{W}_{\mathrm{res}}x(t)$。对于矩阵 $\boldsymbol{W}_{\mathrm{res}}$,可以对它进行 SVD 分解,$\boldsymbol{S}$ 和 $\boldsymbol{D}$ 是正交阵,$\boldsymbol{V}$ 是一个对角阵,对角线上的元素是 $\boldsymbol{W}_{\mathrm{res}}$的特征值的绝对值。如果 $\boldsymbol{W}_{\mathrm{res}}$的谱半径,即最大特征值大于 1,随着时间的推移,系统的能量越来越大,会造成系统不稳定。直观地说会使得当前状态对以后持续很长时间的状态值产生很大的影响。

1.2.2　线性回归

不失一般性[12],假定 $\boldsymbol{W}_{\mathrm{back}}=0$。$\boldsymbol{W}_{\mathrm{out}}$的计算包括采样和权值计算 2 个步骤。

1)采样阶段:采样阶段首先确定网络的初始状态,通常情况下初始状态 $x(0)=0$。训练样本 $\boldsymbol{u}(t)$,$t=1,2,\cdots,T_1$ 依次输入,按照式(1) ~式(3)完成水库状态 $\boldsymbol{x}(t)$和输出 $\tilde{\boldsymbol{y}}(t)$的收集。从某个时间 T_0 开始系统趋于稳定,以向量 $(x_1(i),x_2(i),\cdots,x_N(i))$ $(i=T_0,T_0+1,\cdots,T_1)$ 为行构成矩阵 $\boldsymbol{B}\in(T_1-T_0+1)\times N$,同时相应的样本数据 $\boldsymbol{y}(t)$也被收集,并构成一个列向量 $\boldsymbol{T}\in(T_1-T_0+1)\times 1$。

2)权值计算阶段:因为水库状态 $\boldsymbol{x}(t)$和输出 $\tilde{\boldsymbol{y}}(t)$ 是线性关系,而需要实现的目标是利用网络实际输出 $\tilde{\boldsymbol{y}}(t)$ 逼近期望输出 $\boldsymbol{y}(t)$。从数学的观点来看,这是一个线性回归问题,所以有 $\boldsymbol{W}_{\mathrm{out}}=\boldsymbol{B}^{-1}\boldsymbol{T}$。

1.3 模型空间

水库确定后,对于不同类型的时序数据 S_1 和 S_2,训练所得到的输出权值矩阵 $\boldsymbol{W}^1_{\text{out}}$ 和 $\boldsymbol{W}^2_{\text{out}}$ 即是该时序数据的模型,通过比较模型可以度量两组时序数据之间的距离,从而对原始数据进行分类。

在传统的机器学习中是将学习的算法直接作用于数据的原始空间,基于数据的统计特性来得到相关的结论。这样的思路在一些特殊的情形下很难应用,而模型空间的思想是在学习算法和原始数据之间加上了一个抽象的模型层,将局部原始数据概括为模型表示,利用模型来代表原有的数据,其后的学习算法直接作用于模型所在的空间。模型空间的方法是使用模型来逼近数据。模型空间就是由逼近模型的全体组成的函数空间。在模型空间中,这些模型就成为模型空间中的点集。这样,考虑构建基于模型空间的学习策略或者算法,可以使得学习策略在动态、不确定性环境中也具有比较高的逼近精度和泛化性能。

2 树的回声状态网络

上面介绍了回声状态网络,它用来处理时序数据。树的回声状态网络的基本思想是:把树看作是一个有限的、多分支的序列,每一个结点相当于其所有孩子结点的后继。用回声状态网络可以得到树的水库状态和树的回声状态网络模型[2]。然后用模型空间的思想,把这些状态看成是树的一个模型,通过度量模型可以对树进行分类或者回归。树的回声状态网络工作过程如图 2 所示。

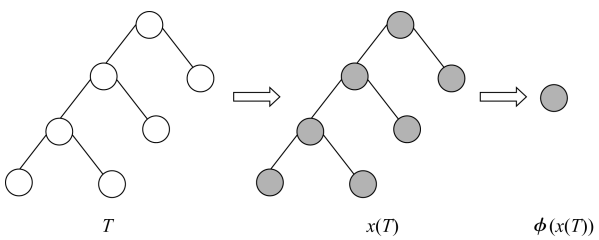

图 2 树在树形回声状态网络中的处理过程

首先是计算出每个结点 n 的状态变量(即水库状态 $x(n)$),得到整个树的水库状态 $x(T)$,然后通过一个状态映射函数 ϕ 得到一个和水库状态维数相同的向量 $\boldsymbol{\phi}(x(T))$,并且这个向量就作为树 T 的模型。

2.1 树的水库状态

假定树 T 的每个结点 n 都有一个标签 $u(n)$ 代表结点的属性,同时 $u(n)$ 也作为水库的输入。把树看作是一个有限的、多分支的序列,每一个结点相当于其所有孩子结点的后继,如图 3 所示。类似于回声状态网络中的式(6)的定义,对于每一个结点 n,有:

$$x(n)=f\left(\boldsymbol{W}_{\text{in}}u(n)+\sum_{i=1}^{d(n)}\boldsymbol{W}_{\text{res}}x(ch_i(n))\right) \tag{6}$$

从式(6)中可以看出,当树退化成为一个序列,即任意结点 n,$d(n)=1$ 时,状态方程和传统的回声状态网络相同。这里需要说明的是对于一个树形数据集中的每个树形数据,水库是固定的即水库的维数和稀疏参数都设置相同(只需要生成一次水库)。

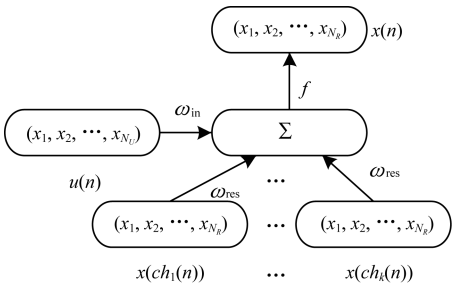

图 3 树结点的水库状态

这是一个递归的定义,利用树的递归结构,可以用 $O(|T|)$ 的时间得到所有结点的水库状态,从而得到树的水库状态 $x(T)$。

2.2 参数的初始化

$\boldsymbol{W}_{\text{in}}$ 的初始化。和传统回声状态网络相同,$\boldsymbol{W}_{\text{in}}$ 中的元素可以在 $[-scale, scale]$ 之间均匀分布。

$\boldsymbol{W}_{\text{res}}$ 的初始化。和传统回声状态网络略有不同。令:

$$k=\max\{d(n)\}(n\in N(T)),\alpha=k\rho(\boldsymbol{W}_{\text{res}}) \tag{7}$$

其中,k 可以大于 1。根据前文的分析,时序数据的长度很长,如果 α 大于 1,则随着时间的推移,系统的能量越来越大,会造成系统不稳定。而这里树的高度和时序数据的长度相比很短,影响并不大。实验结果表明,有时 α 大于 1 效果更好。

2.3 状态映射函数

得到树的水库状态 $x(T)$ 之后,可以根据实际情况选择不同的映射函数,把 $x(T)$ 映射为这棵树的模型 $\boldsymbol{\phi}(x(T))$。本文定义 2 种状态映射函数:

1)根状态映射。直接用根结点的水库状态作为树的模型,即:

$$\boldsymbol{\phi}(x(T))=x(r(T)) \tag{8}$$

2)平均状态映射。用所有结点的水库状态的平均作为树的模型,即:

$$\boldsymbol{\phi}(x(T))=\frac{1}{|T|}\sum_{n\in N(T)}x(n) \tag{9}$$

这 2 种映射函数对应的方法分别称为 TreeESN-R 和 TreeESN-M。在最后的实验部分可以看出,TreeESN-R 适用于根结点比其他结点占更多权重的数据,而 TreeESN-M 适用于所有结点都重要的数据。

2.4 在模型空间中的处理

本文通过回声状态网络得到树的模型,将树结构数据的处理转换到模型空间中。而且由于学习到的模型是向量形式,因此可以用传统的机器学习模型(这里使用 SVM 分类器)来处理它。算法 1 是使

用树形回声状态网络求解树形数据的模型。

算法 1　树形数据的模型

输入　$D=\{T,Y\}=\{(T_n,y_n)\}_{n=1}^{N}$是训练数据集,其中 T_n 是树形数据,包括结点值和孩子信息;初始化信息为scale和∂

//树形数据包括树的结构和结点的值

输出　树 T 的模型 $\phi(x(T))$

//模型用于后期模型空间的计算

1. W_{in}初始化[- scale,scale]

//W_{in}为输入对状态变量的连接权值矩阵

2. W_{res}初始化$\partial=k\rho(W_{res})$,其中

$k=\max\{d(n)\}(n\in N(T))$

//W_{res}为状态变量

3. for i = 1:d(n) do

//计算树的水库状态

4. $x(n)=f(W_{in}u(n)+\sum_{i=1}^{d(n)}W_{res}x(ch_i(n)))$

5. end for

6. if TreeESN-R then

//根状态映射

7. $\phi(x(T))=x(r(T))$

8. else if TreeESN-M then

//平均状态映射

9. $\phi(x(T))=\frac{1}{|T|}\sum_{n\in N(T)}x(n)$

10. end if

由于树形结构数据结点属性值和结构信息的属性值不同,因此对于不同的树形数据 T_1 和 T_2,经过训练出来的模型 $\phi(x(T_1))$和 $\boldsymbol{\phi}(x(T_2))$也不相同,把 $\phi(x(T))$看作该树形数据的模型,通过比较该模型可以度量两组树形数据之间的距离,从而对其分类或者回归。其处理流程如图 4 所示。

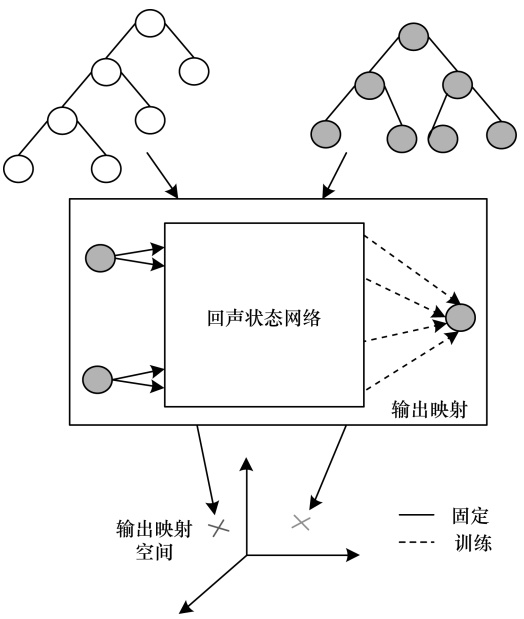

图 4　树形数据在回声状态网络和模型空间中的处理流程

2.5　支持向量机

支持向量机[13](Support Vector Machine,SVM)是一种在统计学习理论之后兴起的机器学习方法,可以用来解决回归问题、分类问题、异常点检测问题等。

这里定义一个线性的模型:

$$y(x)=\boldsymbol{w}^{\mathrm{T}}\phi(x)+b \tag{10}$$

其中,$\phi(x)$是关于 x 的一个特征空间变换;b 为偏置参数。对于所有的 n 都有 $t_n y(x_n)>0$,t_n为目标值。因此,点 x_n 到决策平面的距离可以表示为:

$$\frac{t_n y(x_n)}{\|\boldsymbol{w}\|}=\frac{t_n(\boldsymbol{w}^{\mathrm{T}}\phi(x_n)+b)}{\|\boldsymbol{w}\|} \tag{11}$$

接下来目标就是优化参数 $\boldsymbol{w}$ 和 b,使式(11)中距离最大化,这里使用二次规划和拉格朗日方法求解问题[14]。最后得到关于参数 a 的最大化的拉格朗日函数如式(12)所示。

$$L(a)=\sum_{n=1}^{N}a_n-\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}a_n a_m t_n t_m k(x_n,x_m) \tag{12}$$

其中,$\boldsymbol{a}=(a_1,a_2,\cdots,a_N)^{\mathrm{T}}$,$a_n$ 称为拉格朗日乘数,核函数表示为 $k(x,x')=\boldsymbol{\phi}(x)^{\mathrm{T}}\phi(x')$。式(12)的限制条件为式(13)、式(14)。

$$a_n\geqslant 0,n=1,2,\cdots,N \tag{13}$$

$$\sum_{n=1}^{N}a_n t_n=0 \tag{14}$$

2.6　回归和分类任务

得到树形结构数据的模型 $\boldsymbol{\phi}(x(T))$ 后,可以把 $\boldsymbol{\phi}(x(T))$当成传统机器学习模型中的数据进行回归或分类任务的训练。

1)树的回归

假定树 T 的目标值是 $y(T)$,把数据的模型构成列向量 $\boldsymbol{X}=(\phi(x(T_1)),\phi(x(T_2)),\cdots)$,目标值也构成列向量 $\boldsymbol{Y}=(y(T_1),y(T_2),\cdots)$,有[11]:

$$\boldsymbol{W}_{\mathrm{out}}\boldsymbol{X}=\boldsymbol{Y} \tag{15}$$

和回声状态网络类似的结论有:

$$\boldsymbol{W}_{\mathrm{out}}=\boldsymbol{Y}\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1} \tag{16}$$

为了避免过拟合,使用岭回归训练输出权值:

$$\boldsymbol{W}_{\mathrm{out}}=\boldsymbol{Y}\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{Y}\boldsymbol{X}^{\mathrm{T}}+\lambda_r\boldsymbol{I}_N)^{-1} \tag{17}$$

其中,λ_r是正则化参数;$\boldsymbol{I}_N$ 是 $N\times N$ 维的单位阵。

2)树的分类

本文利用 SVM 进行分类,采用广泛使用的 libsvm 实现 SVM 的超参数通过 5 重交叉验证得到。

3　实验结果与分析

实验数据包括回归和分类的数据集,选用平均绝对误差、错误率和 Area Under the Cure(AUC)作为评价指标,通过一些树形数据常用的算法和树的回声状态网络算法作用于数据集上的实验结果来比较算法的性能。

3.1 烷烃化合物沸点的回归

烷烃化合物是一种简单的有机物,它由碳链骨架和氢原子组成,可以用树来表示。实验所用的数据集来自于 http://www.di.unipi.it/~micheli/dataset。数据集中包含 150 个不同的烷烃分子式和相应的沸点。实验中为了方便,忽略氢原子的影响,仅考虑碳原子形成的树[15],且每个结点的标签都设为 1,如图 5 所示。

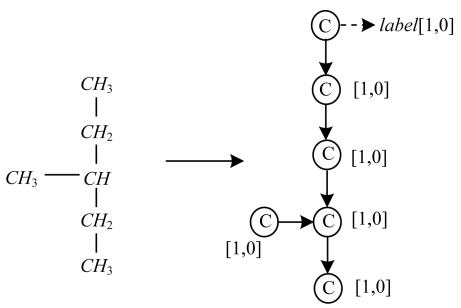

图 5 烷烃化合物的树结构表示

实验中设置树形回声状态网络的参数如下:水库的内部单元矩阵的稀疏设置为 5%,输入权重参数中的 scale 设置为 0.5,并将水库的参数 α 分别设置成 1 和 2 作对比。图 6 和图 7 是参数 α 分别取 1 和 2 时对于烷烃化合物数据集在不同维数下的平均绝对误差和方差。

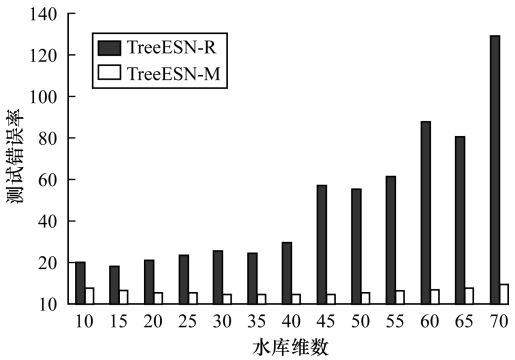

图 6 α = 1 时,TreeESN-R 和 TreeESN-M 在不同水库维数下的平均绝对误差和方差

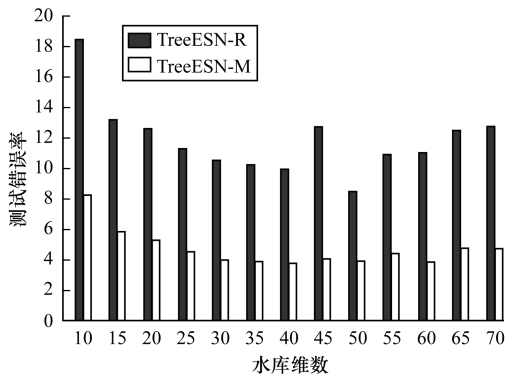

图 7 α = 2 时,TreeESN-R 和 TreeESN-M 在烷烃化合物不同水库维数下的平均绝对误差和方差

从图 6 和图 7 中可以发现,对于烷烃化合物数据集,TreeESN-M 的性能要明显好于 TreeESN-R。α 取值为 2 比取值为 1 时误差要小,与上文讨论相符。水库的维数过大或者过小都会影响性能。且当 $\alpha=2$、水库维数 $N=40$、状态映射函数是平均状态映射时平均绝对误差最小。这里是因为烷烃分子上的每一个碳原子都很重要,所以平均状态映射有更好的性能。

烷烃化合物的数据集经常使用递归神经网络以及其他一些先进的机器学习方法。这些方法主要包括 RCC[2](Recursive Cascade Correlation),CRCC[16](Contextual Recursive Cascade Correlation),SST kernel[17],NN4G[18](Neural Networks for Graphs)。将它们与本文的 TreeESN 的 2 种状态映射方法在相似条件下作对比。这里参数 α 取 2。各种模型在烷烃化合物数据集上测试结果的平均绝对误差对比如表 1 所示。

表 1 不同算法作用在烷烃数据集上的平均绝对误差

算法	平均绝对误差
TreeESN-R	8.09
TreeESN-M	2.78
RCC	10.03
CRCC	8.29
SST	2.93
NN4G	2.34

从表 1 中可以看出,作为线性处理的 TreeESN-M 和其他处理烷烃化合物算法相比有较好的性能,虽然 NN4G 的平均绝对误差稍稍低于 TreeESN-M,但是 TreeESN 的算法复杂度 $O(|T|)$ 远比 NN4G 的算法复杂度 $O(|T|^2)$[18] 具有更优良的性能。所以,在烷烃化合物的处理问题上,本文提出的算法要优于本节中对比的算法。

3.2 多糖化合物的分类

对多糖化合物的分子结构和它表现的生物学特征之间关系的分析,成为机器学习在计算生物学领域的一个有趣课题[5]。现在对多糖化合物进行分类。这是一个二分类问题,一类和白血病有关,标记为 +1;另一类和白血病无关,标记为 -1。多糖化合物的分类包含 leukemia 和 cystic 分类任务,数据集都来自于 KEGG/Glycan[19] 的生物学数据库。leukemia 数据集包含 442 个多糖分子和它的标签,cystic 数据集包含 160 个多糖分子和它的标签,多糖可以看作是由糖链形成的树,树的根结点取决于生物学定义[5]。对于不同的糖分子结点,它的标签用不同的 0-1 向量表示,如图 8 所示。

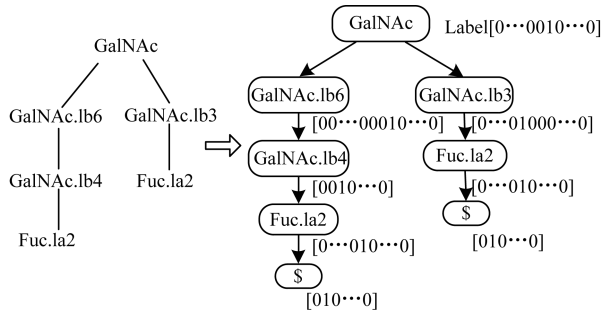

图 8　多糖分子的树结构表示

实验中设置树形回声状态网络的参数如下:水库的内部单元矩阵的稀疏设置为 5%,输入权重参数中的 *scale* 设置为 0.5,水库收敛参数 α 设置为 2。

对于多糖化合物的处理问题,现在比较流行的方法是使用一些基于树形数据核的方法。下面将 TreeESN 的 2 种状态映射方法与这些基于树核的方法(其中某些算法专门为多糖化合物设计)在多糖化合物分类问题上进行对比[20]。这些基于树核的方法有:Yamanishi Kernel[21], the ST and subpath kernels[22], baseline Kailing kernel[23], q-gram and weighted q-gram linkage (LK), KCaM (KM), linkageKCam (LKM) kernels[24]。将这些方法与 TreeESN 在多糖化合物分类问题上作对比。计算应用各自算法结果的 AUC,如表 2、表 3 所示。

表 2　不同算法作用于 leukemia 数据集上的 AUC

算法	AUC 值
TreeESN-R	0.949 3
TreeESN-M	0.971 0
Yamanishi	0.920 1
q-gram	0.934 0
KM	0.935 2
LKM	0.935 5
LK	0.962 7
Subpath	0.970 2
ST	0.960 7
Kailing	0.927 7

表 3　不同算法作用于 cystic 数据集上的 AUC

算法	AUC 值
TreeESN-R	0.772 0
TreeESN-M	0.751 5
Yamanishi	0.343 1
q-gram	0.699 0
KM	0.698 2
LKM	0.696 3
LK	0.762 2
Subpath	0.743 0
ST	0.762 2
Kailing	0.712 7

从表 2、表 3 中可以看出,TreeESN 在 leukmia 数据集上的 AUC 值分别为 0.949 3(TreeESN-R)和 0.971 0(TreeESN-M),在 cystic 数据集上的 AUC 值分别为 0.772 0(TreeESN-R)和 0.751 5(TreeESN-M)。和其他方法相比,TreeESN 的 2 种映射函数都有很好的性能。在 cystic 数据集上,Yamanishi 算法因为数据集个数较少,它的性能变得极差。更为重要的是与其他算法相比,TreeESN 算法的时间复杂度要低很多,因为它是线性分类器。而其他算法比如性能较好的 LK 算法[25],它的时间复杂度呈指数分布。所以,TreeESN 算法与其他处理多糖化合物的树核算法相比具有明显的优势。

4　结束语

本文把树形结构数据看作是一个有限的、多分支的序列,每一个结点相当于其所有孩子结点的后继。通过树形回声状态网络,可以得到一个树的水库状态。然后根据不同的水库映射,选择 TreeESN-R 或 TreeESN-M 映射函数,把水库状态下的一个树形结构数据映射为一个模型。将模型之间的距离作为树之间的距离。由于模型是向量形式,因此可以用传统的机器学习模型(如 SVM 分类器)来处理。实验结果表明,树的回声状态网络和很多已有方法相比,在数据集上有着较好的效果。在 2 个实验中,TreeESN-M 和 TreeESN-R 的性能各有高低,这表明状态映射函数应该根据不同的数据特点进行选择。由于本文提出的方法是一个比较泛化的方法,针对性不足,如何针对不同的数据情况加强树的回声状态网络模型的针对性是下一步要研究的问题。

参考文献

[1] 叶　玲,顾　微,周玉兰. 生物材料聚酰胺胺树状大分子在医学领域研究进展[J]. 高分子通报,2002(4):1-5.

[2] Gallicchio C, Micheli A. Tree Echo State Networks[J]. Neurocomputing, 2013, 101(3): 319-337.

[3] Relational D S, Dathio C, Micheli A. Tree Echo State Networks[J]. Neurocomputing, 2013, 101(3): 319-337.

[4] Džeroski S. Rela Mining [J]. American Journal of International Law, 1935, 29(2): 248-279.

[5] Bille P. A Survey on Tree Edit Distance and Related Problems [J]. Theoretical Computer Science, 2005, 337(1-3): 217-239.

[6] Akutsu T, Fukagawa D, Takasu A, et al. Exact Algorithms for Computing the Tree Edit Distance Between Unordered Trees [J]. Theoretical Computer Science, 2011, 412(4/5): 352-364.

[7] Dietterich T G. Adaptive Computation and Machine Learning[M]. [S.1.]: MIT Press, 1998.

(下转第 206 页)

参考文献

[1] Liu L N, Mackin S. Fault-tolerant Peer-to-Peer Search on Small-world Networks [J]. Future Generation Computer Systems, 2007, 23(8): 921-931.

[2] Xiao W, Parhami B. Cayley Graphs as Models of Deterministic Small-world Networks [J]. Information Processing Letters, 2006, 97(3): 115-117.

[3] Cooper C, Radzik T, Siantos Y. Estimating Network Parameters Using Random Walks[C]//Proceedings of the 4th International Conference on Computational Aspects of Social Networks. Washington D. C., USA: IEEE Press, 2012: 33-40.

[4] 何 静,郭进利,徐雪娟. 微博关系网络模型研究[J]. 计算机工程,2013,39(11):105-108.

[5] Symeonidis P, Ntempos D, Manolopoulos Y. Online Social Networks[J]. IEEE Network, 2010, 24(5): 4-5.

[6] 崔颖安,李 雪,王志晓,等. 在线社交媒体数据抽样方法的比较研究[J]. 计算机学报,2014,37(8):1859-1876.

[7] Ahn Y Y, Han S, Kwak H, et al. Analysis of Topological Characteristics of Huge Online Social Networking Services [C]//Proceedings of the 16th International Conference on World Wide Web. New York, USA: ACM Press, 2007: 835-844.

[8] Kurant M, Markopoulou A, Thiran P. On the Bias of BFS (Breadth First Search) [C]//Proceedings of the 22nd International Teletraffic Congress. Washington D. C., USA: IEEE Press, 2010: 49-56.

[9] Beamer S, Asanovic K, Patterson D. Direction-optimizing Breadth-first Search [J]. Scientific Programming, 2013, 21(5): 137-148.

[10] Lu Jianguo, Li Dingding. Sampling Online Social Networks by Random Walk[C]//Proceedings of the 1st ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research. New York, USA: ACM Press, 2012: 33-40.

[11] Stutzbach D, Rejaie R, Duffield N, et al. On Unbiased Sampling for Unstructured Peer-to-Peer Networks [J]. IEEE/ACM Transactions on Networking, 2009, 17(2): 377-390.

[12] Ahmed S E. Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference [J]. Technometrics, 2008, 50(1): 497-537.

[13] Chaim N. Sampling Knowledge: The Hermeneutics of Snowball Sampling in Qualitative Research [J]. International Journal of Social Research Methodology, 2008, 11(4): 327-344.

[14] Gjoka M, Kurant M, Butts C T, et al. Practical Recommendations on Crawling Online Social Networks[J]. IEEE Journal on Selected Areas in Communications, 2011, 29(9): 1872-1892.

[15] Gjoka M, Kurant M, Butts C T, et al. Walking in Facebook: A Case Study of Unbiased Sampling of OSNs[C]//Proceedings of the 29th Conference on Information Communications. Washington D. C., USA: IEEE Press, 2010: 1-9.

编辑 刘 冰

(上接第 199 页)

[8] Raedt L D. Statistical Relational Learning: An Inductive Logic Programming Perspective[M]. Berlin, Germany: Springer, 2005: 3-5.

[9] 陈 鸿,金培权,岳丽华,等. 基于上下文特征分类的评论长句切分方法[J]. 计算机工程,2015,41(9):233-237,244.

[10] Diligenti M, Frasconi P, Gori M. Hidden Tree Markov Models for Document Image Classification [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2003, 25(4): 519-523.

[11] Passerini A. Kernel Methods for Structured Data[J]. Intelligent Systems Reference Library, 2013, 49(1): 283-333.

[12] Jaeger H. Echo State Network [J]. Scholarpedia, 2007, 2(9): 1479-1482.

[13] 杨庆海,卢 波,颜子夜,等. 基于马尔科夫随机场的粘连字符串切分算法[J]. 计算机工程,2013,39(4):258-262.

[14] 谢赛琴,沈福明,邱雪娜. 基于支持向量机的人脸识别方法[J]. 计算机工程,2009,35(16):186-188.

[15] 刘 蓉,刘 明. 基于三轴加速度传感器的手势识别[J]. 计算机工程,2011,37(24):141-143.

[16] Bianucci A M, Micheli A, Sperduti A, et al. Application of Cascade Correlation Networks for Structures to Chemistry [J]. Applied Intelligence, 2000, 12(1): 115-145.

[17] Micheli A, Sona D, Sperduti A. Contextual Processing of Structured Data by Recursive Ccascade Correlation[J]. IEEE Transactions on Neural Networks, 2004, 15(6): 1396-1410.

[18] Micheli A, Portera F, Sperduti A. A Preliminary Empirical Comparison of Recursive Neural Networks and Tree Kernel Methods on Regression Tasks for Tree Structured Domains [J]. Neurocomputing, 2005, 64(2): 73-92.

[19] Micheli A. Neural Network for Graphs: A Contextual Constructive Approach[J]. IEEE Transactions on Neural Networks, 2009, 20(3): 498-511.

[20] Hashimoto K, Goto S, Kawano S, et al. Kegg as a Glycome Informatics Resource[J]. Glycobiology, 2006, 16(5): 63-70.

[21] 郑丽贤,何小海,吴 炜,等. 基于学习的超分辨率技术[J]. 计算机工程,2008,34(5):193-195.

[22] Yamanishi Y, Bach F, Vert J P. Glycan Classification with Tree Kernels[J]. Bioinformatics, 2007, 23(10): 1211-1216.

[23] Kimura D, Kuboyama T, Shibuya T, et al. A Subpath Kernel for Rooted Unordered Trees [M]. Berlin, Germany: Springer, 2011.

[24] Kailing K, Kriegel H P, Schönauer S, et al. Efficient Similarity Search for Hierarchical Data in Large Databases[M]. Berlin, Germany: Springer, 2004.

[25] Li L, Ching W K, Yamaguchi T, et al. A Weighted Q-gram Method for Glycan Structure Classification [J]. BMC Bioinformatics, 2010, 1(1): 293-300.

[26] 李 军,李艳辉,彭存银. 基于自适应遗传算法的路径测试数据生成[J]. 计算机工程,2009,35(2):203-205.

编辑 索书志