

(12) 发明专利

(10) 授权公告号 CN 102136087 B

(45) 授权公告日 2013.08.21

(21) 申请号 201110054769.8

(22) 申请日 2011.03.08

(73) 专利权人 湖南大学

地址 410082 湖南省长沙市麓山南路2号

(72) 发明人 张大方 王晓阳

(74) 专利代理机构 长沙正奇专利事务所有限责任公司 43113

代理人 马强

(51) Int. Cl.

G06N 3/08 (2006.01)

H04L 12/26 (2006.01)

(56) 对比文件

CN 101651568 A, 2010.02.17, 全文.

CN 101729323 A, 2010.06.09, 全文.

田津等. 基于三阶段 RBFNN 学习算法的复杂样本分类研究. 《系统工程与电子技术》. 2006, 第 28 卷 (第 1 期), 第 114-118 页.

蒋定德等. 流量矩阵估计研究综述. 《计算机科学》. 2008, 第 35 卷 (第 4 期), 第 5-9、13 页.

蒋定德等. 基于广义回归神经网络的流量矩阵估计. 《计算机应用研究》. 2009, 第 26 卷 (第 7 期), 第 2676-2679 页.

7 期), 第 2676-2679 页.

蒋定德等. IP 骨干网络流量矩阵估计算法研究. 《电子科技大学学报》. 2010, 第 39 卷 (第 3 期), 第 420-424 页.

关卿等. 基于多数据源的网络流量矩阵估计. 《计算机工程》. 2009, 第 35 卷 (第 14 期), 第 122-124 页.

审查员 房琦

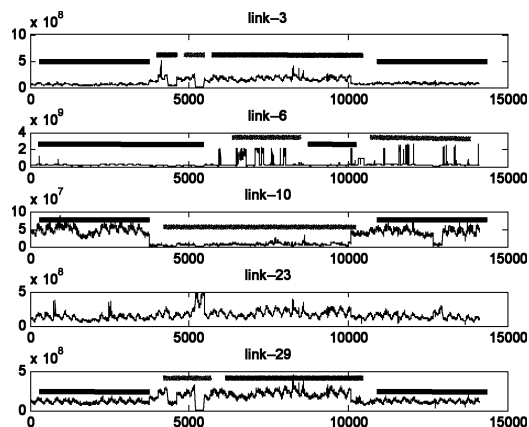
权利要求书2页 说明书6页 附图4页

(54) 发明名称

基于多神经网络的流量矩阵估计方法

(57) 摘要

本发明为一种基于多神经网络的流量矩阵估计方法,可提高现有网络流量矩阵估计的准确性,它通过对样本流量先分类,再分别训练,有效的克服了传统神经网络用于流量矩阵估计时存在记忆消失或变形的问题;本发明的多神经网络估计的误差明显低于传统的神经网络。



1. 一种基于多神经网络的流量矩阵估计方法,其特征在于,它包括样本分类阶段,网络训练阶段和网络估计阶段的三个阶段,其中:

样本分类阶段为:将直接测量得到的链路流量Y按K-means算法进行分类,记录各个分类的中心c和半径d;

网络训练阶段为:将各个分类的链路流量做为输入,对应的OD流量X作为输出,建立神经网络,并训练,记录各个网络的权值;

网络估计阶段为:对于某一时刻的链路测量值,根据各分类中心c和半径d来判断其所属分类,再激活对应的神经网络,计算出OD流量的初步估计值,选取最满足链路约束式(1)的估计值,再使用IPFP算法调整,得到最终的估计值;

$$Y=AX \quad (1)$$

其中,A表示路由矩阵;所述样本分类阶段中的K-means分类算法如下:

给定条件:样本值(X,Y),其中X表示OD流量矩阵,Y表示相应的链路流量矩阵,Y的列向量为m维列向量,X和Y的列数为T

Step1 令初聚类数目K=2,令初始施瓦兹信息准则值SIC为一个极大值;

Step2 从Y的T列中随机选择K列作为初始聚类中心c(c的各个列代表各个相应分类的中心向量);

Step3 根据每个聚类的中心,计算Y的每一列与这些中心的欧几里德距离,根据最小距离对相应的列进行重新划分,即

$$D(y_i)=\{i|\min_{i=1,\dots,K} \|y_j-c_i\|\}, y_j \in Y \quad (2)$$

其中 y_j 表示矩阵Y的第j列, $D(y_j)$ 表示 y_j 的分类编号, c_i 表示矩阵c的第i列, $\|\cdot\|$ 表示欧氏范数,下同;

Step4 使用式(3)计算各分类的元素个数 s_i (向量s的各个元素即为相应各个分类的元素个数),使用式(4)重新计算每个聚类的中心c,根据式(5)重新计算每个聚类的半径 d_i (向量d的各个元素即为相应各个分类的距离),根据式(6)计算此次分类的偏差J;

$$s_i = \sum_{D(y_j)=i} 1 \quad (3)$$

$$c_i = \frac{\sum_{D(y_j)=i} y_j}{s_i} \quad (4)$$

$$d_i = \max_{D(y_j)=i} \|y_j - c_i\| \quad (5)$$

$$J = \sum_{j=1}^T \|y_j - c(D(y_j))\|^2 \quad (6)$$

其中 s_i 表示向量s的第i个元素值, d_i 表示向量d的第i列;

Step5 转Step3直到分类的偏差J不再变小时转Step6,并记录此时分类的中心c;

Step6 转Step2直至分类中心c不再发生变化时转Step7,并将X并入到相应的Y分类中,记录(X,Y)的分类及其中心c和半径d;

Step7 通过(7)计算此次分类的施瓦兹信息准则值SIC,令K=K+1转Step2直至SIC不再变小时中止,输出分类数目K,输出(X,Y)的分类及其中心c和半径d,

$$SIC = J + \lambda m K \log T \quad (7)$$

其中 λ 为权重因子。

2. 根据权利要求 1 所述多神经网络的流量矩阵估计方法,其特征在于,所述网络估计阶段判断所属分类时使用 (8) 式确定

$$\|y - c_i\| < \rho d_i \quad (8)$$

其中 y 是对于某时间的链路测量值, ρ 表示警戒参数, ρ 越小,估计的结果越精确,但 ρ 过小可能会使链路测量值不属于所有的分类,当 ρ 大于 1 时,神经网络的估计结果可靠性大大降低。

3. 根据权利要求 1 所述多神经网络的流量矩阵估计方法,其特征在于,所述网络估计阶段 IPFP 算法如下:

给定条件:(1) OD 初始估计值 x , x 的维数为 n

(2) 观测到某时刻的链路流量 y , y 的维数为 m

(3) 当前时刻路由矩阵 A , A 为 $m \times n$ 的矩阵

步骤一、设定最大迭代步数 K , 收敛的误差 ε , 计算当前链路误差 $error$

$$error = \frac{\|y - Ax\|}{\|y\|} \quad (10)$$

步骤二、重复以下步骤,直至 $error$ 小于 ε 或者迭代步数大于 K

for $j = 1:m$

$$ye = A(j,:) * x;$$

$$x = A(j,:)^T \times (y(j)/ye * x) + (\text{ones}(1,n) - A(j,:))^T \times x;$$

endfor

其中 $A(j,:)$ 表示 A 的第 j 行, $y(j)$ 表示 y 的第 j 个元素,“ T ”表示矩阵的转置运算,“ $*$ ”表示乘运算,“ $/$ ”表示除运算,“ \times ”表示矩阵的点乘运算,即矩阵的各个元素对应相乘, $\text{ones}(1,n)$ 表示 1 行 n 列的元素值全为 1 的矩阵;

步骤三、根据式 (10) 重新计算 $error$, 直至 $error$ 小于 ε 或者迭代步数大于 K 时转步骤四, 否则转步骤二;

步骤四、输出调整后的 x 。

基于多神经网络的流量矩阵估计方法

技术领域

[0001] 本发明涉及网络测量领域和神经网络领域,具体是基于多神经网络的流量矩阵估计方法。

背景技术

[0002] 流量矩阵是全网流量的概览,矩阵中的元素表示网络中始于一个结点(源结点)而终止于另一个结点(目的结点)的流量(OD 流量)。此源、目的结点对又被称为 OD 对。OD 流量的测量与计算在网络结构配置,管理,网络流量工程等研究与工程实践中具有重要的意义,特别是近年来的研究发现网络 OD 流量的测量可以用于网络异常的检测与识别,因而 OD 流量的测量与计算研究受到了国内外理论界和工业界的广泛重视。由于流量矩阵需要捕获网络流量的全局状态,直接监控代价非常高,实际上几乎是不可行的。近年来,由间接观测进行流量矩阵估算已成为一个非常热门的研究领域。

[0003] 流量矩阵的行对应 OD 对,列对应不同时刻的流量需求。令 $y(t) = (y_1(t), y_2(t), \dots, y_m(t))^T$ 表示一个网络中所有链路的流量值, m 表示链路的总数。 $x(t) = (x_1(t), x_2(t), \dots, x_n(t))^T$ 为该网络中所有 OD 对的流量矩阵, n 表示网络中 OD 对的总数, x_j 表示第 j 个 OD 对之间的流量需求。 $A = (a_{ij})$ 是 $m \times n$ 阶的 0-1 矩阵。 A 的列指明了某个 OD 流量需求在网络中所要经过的全部链路的集合,显然, A 是一个包含了实际路由信息的矩阵。令矩阵 $Y = (y(1), y(2), \dots, y(T))$ 表示 T 时间段的链路流量,令矩阵 $X = (x(1), x(2), \dots, x(T))$ 表示 T 时间段的流量矩阵,则链路流量、路由矩阵和流量矩阵三者之间的关系可以表示如下:

$$[0004] \quad Y = AX \quad (1)$$

[0005] 通常,由于网络中 OD 对的数量要远大于链路数,即 $n \gg m$, A 不是一个满秩矩阵,这意味着式 (1) 将有无穷多组可能解,是一种病态的线性逆问题 (ill-posed linear inverse problem)。流量矩阵估算所要解决的就是在已知链路流量 Y 和路由矩阵 A 的情况下从式 (1) 中求出流量矩阵 X 。其中,链路流量 Y 可以通过一般的流量数据采集方法(如 SNMP)得到,路由矩阵 A 可以通过收集内部路由协议 (IGP) 的配置信息或者通过收集路由器之间交互的链路状态信息获得,也可以通过计算最短路径树得到。

[0006] 为求解问题 (1),国内外许多学者进行了大量有意义的研究工作,主要思路是在上述方程的基础上,增加一些假设来获得 (1) 的最优解。如假定流量的泊松模型、高斯模型、重力模型、信息论独立模型、马尔可夫高斯模型、独立连接模型等。

[0007] 当前网络流量被公认最为重要的统计特征是尺度特性(如自相似性)和多尺度行为特性(如长相关、多重分形性),网络流量的复杂特性使得要想获得更高的估计精确度,需要使用更复杂的模型,神经网络是一个较理想的方法。目前国内外有很多学者将目前已有的很多种神经网络用于该领域。然而传统神经网络普遍存在着这种困境:稳定性和可塑性不可调和的矛盾,这使得神经网络在学习新的样本时会影响甚至是遗忘旧有的记忆。流量矩阵的复杂特性和高维特性需要大量的样本才能够充分的训练神经网络的权值。而这些

必将导致神经网络的训练速度非常慢,而对大量的高维样本数据的训练,会加剧传统神经网络记忆性和可塑性的矛盾,即记忆扭曲变形或记忆消失的问题。为了克服这些问题,我们提出一种多神经网络的流量矩阵估计方法。

发明内容

[0008] 本发明要解决的技术问题是,针对传统神经网络用于流量矩阵估计领域所面临的上述困境,提出一种基于多神经网络的流量矩阵估计方法。

[0009] 本发明的技术方案是,所述基于多神经网络的流量矩阵估计方法包括样本分类阶段,网络训练阶段和网络估计阶段共三个阶段,其中:

[0010] 样本分类阶段为:将直接测量得到的链路流量Y按K-means算法进行分类,记录各个分类的中心c和半径d;

[0011] 网络训练阶段为:将各个分类的链路流量做为输入,对应的OD流量X作为输出,建立传统的神经网络,并训练,记录各个网络的权值;

[0012] 网络估计阶段为:对于某一时刻的链路测量值,根据各分类中心c和半径d来判断其所属分类,再激活对应的神经网络,计算出OD流量的初步估计值,选取最满足链路约束式(1)的估计值,再使用IPFP算法调整,得到最终的估计值;

[0013] $Y = AX$ (1)

[0014] 其中,A表示路由矩阵。

[0015] 以下对本发明做出进一步说明。

[0016] 流量矩阵描述其对应的IP网络中所有从源节点传输到目的节点的流量。流量矩阵中的每一行代表着从某一个源节点到另一个目的节点的OD流(Origin-Destination Flow)在不同时刻的流量,每个列表示网络中所有的OD流在某个时间片段(time slot)(如5分钟)的流量。流量矩阵反映了对应网络中所有源节点到目的节点在不同时间的流量需求情况。图1为美国的Abilene IP骨干网络部分链路7周内的流量数据。可看出流量表现出截然不同的类别特征;而OD流量的空间自相似性使得链路的特征在一定程度上反映了OD流量的特征。所以,使用单神经网络进行训练时,会出现已有的训练记忆消失或变形的问题。而多神经网络则分别对不同类别的流量建立网络进行训练,这便解决了记忆性和可塑性的矛盾。

[0017] 经验和大量的实践表明:欧氏范数能均衡反映向量各分量的差异程度和向量的大小。所以我们选用欧氏范数作为链路流量差异化的度量。分类的目的是为了使各分类的向量差异化大,而分类内的各向量差异小。我们首先选用K-means方法(使用施瓦兹信息准则确定分类数K)对链路流量进行分类。再对各个类分别进行神经网络训练,其中各个类的链路流量作为神经网络的输入,链路流量对应的OD流量作为神经网络的输出。进行流量矩阵估计时,选择合适的分类对应的网络进行估计,一般地,通过神经网络计算的估计流量不满足式(1),因此需要根据实际的链路流量对估计值进行适当调整。本文采用IPFP算法。整个流量样本分类训练过程和流量矩阵的估计过程表示如下。

[0018] 流量样本分类训练算法:

[0019] Step1 采用一段时间的直接测量值(X,Y)作为样本,其中X表示OD流量,Y表示相应的链路流量;

[0020] Step2 使用 K-means 算法或其它算法对样本 Y 进行合理的分类,其中使用贝叶斯信息准则 (BIC 或施瓦兹信息准则 (SIC) 确定或根据经验确定分类个数 K,并将 Y 对应的 X 并入到相同的分类中,使用 (4) (5) 计算并记录各分类中心 c_i 和半径 d_i , $i = 1, 2, \dots, K$;

[0021] Step3 对各个分类分别建立神经网络,以各个分类的 Y 作为输入,X 作为输出,并使用对应的神经网络训练算法进行训练。

[0022] 其中 K-means 分类算法表示如下:

[0023] 给定条件:样本值 (X, Y),其中 X 表示 OD 流量矩阵,Y 表示相应的链路流量矩阵,Y 的列向量为 m 维列向量,X 和 Y 的列数为 T

[0024] Step1 令初聚类数目 $K = 2$,令初始施瓦兹信息准则值 SIC 为一个极大值;

[0025] Step2 从 Y 的 T 列中随机选择 K 列作为初始聚类中心 c (c 的各个列即为各个相应分类的中心向量);

[0026] Step3 根据每个聚类的中心,计算 Y 的每一列与这些中心的欧几里德距离,根据最小距离对相应的列进行重新划分,即

$$[0027] \quad D(y_j) = \{i | \min_{i=1, \dots, K} \|y_j - c_i\|\}, y_j \in Y \quad (2)$$

[0028] 其中 $\|\cdot\|$ 表示欧氏范数,下同;

[0029] 其中 y_j 表示矩阵 Y 的第 j 列, $D(y_j)$ 表示 y_j 的分类编号, c_i 表示矩阵 c 的第 i 列;

[0030] Step4 使用式 (3) 计算各分类的元素个数 s (向量 s 的各个元素即为相应各个分类的元素个数),使用式 (4) 重新计算每个聚类的中心 c,根据式 (5) 重新计算每个聚类的半径 d (向量 d 的各个元素即为相应各个分类的距离),根据式 (6) 计算此次分类的偏差 J;

$$[0031] \quad s_i = \sum_{D(y_j)=i} 1 \quad (3)$$

$$[0032] \quad c_i = \frac{\sum_{D(y_j)=i} y_j}{s_i} \quad (4)$$

$$[0033] \quad d_i = \max_{D(y_j)=i} \|y_j - c_i\| \quad (5)$$

$$[0034] \quad J = \sum_{j=1}^T \|y_j - c(D(y_j))\|^2 \quad (6)$$

[0035] 其中 s_i 表示向量 s 的第 i 个元素值, d_i 表示向量 d 的第 i 列;

[0036] Step5 转 Step3 直到分类的偏差 J 不再变小时转 Step6,并记录此时分类的中心 c;

[0037] Step6 转 Step2 直至分类中心 c 不再发生变化时转 Step7,并将 X 并入到相应的 Y 分类中,记录 (X, Y) 的分类及其中心 c 和半径 d;

[0038] Step7 通过 (7) 计算此次分类的施瓦兹信息准则值 SIC,令 $K = K+1$ 转 Step2 直至 SIC 不再变小时中止,输出分类数目 K,输出 (X, Y) 的分类及其中心 c 和半径 d。

$$[0039] \quad SIC = J + \lambda m K \log T \quad (7)$$

[0040] 其中 λ 为权重因子。

[0041] 流量矩阵的估计算法:

[0042] Step1 对于某时间的链路测量值 y,当 y 满足式 (8) 时激活相应的网络 (ρ 为警戒参数),否则抑制该网络;

$$[0043] \quad \|y - c_i\| < \rho d_i \quad (8)$$

[0044] 其中 ρ 表示警戒参数, ρ 越小, 估计的结果越精确, 但 ρ 过小可能会使链路测量值不属于所有的分类, 当 ρ 大于 1 时, 神经网络的估计结果可靠性大大降低。

[0045] Step2 计算所有激活网络的流量估计值 x_{neti} , 只激活最接近链路约束 (1) 的估计值的 IPFP 算法模块, 即

$$[0046] \quad x = \{x_{neti} | \min_{neti} ||y - Ax_{neti}||\} \quad (9)$$

[0047] Step3 输出估计值 x , 转 Step1 估计下一个时刻的流量。

[0048] 其中, IPFP 算法如下:

[0049] 给定条件: (1) OD 初始估计值 x , x 的维数为 n

[0050] (2) 观测到某时刻的链路流量 y , y 的维数为 m

[0051] (3) 当前时刻路由矩阵 A , A 为 $m \times n$ 的矩阵

[0052] Step1 设定最大迭代步数 K , 收敛的误差 ε , 计算当前链路误差 $error$

$$[0053] \quad error = \frac{||y - Ax||}{||y||} \quad (10)$$

[0054] Step2 重复以下步骤, 直至 $error$ 小于 ε 或者迭代步数大于 K

[0055] for $j = 1:m$

[0056] $ye = A(j, :) * x;$

[0057] $x = A(j, :)^T \cdot (y(j) / ye * x) + (ones(1, n) - A(j, :))^T \cdot x;$

[0058] endfor

[0059] 其中 $A(j, :)$ 表示 A 的第 j 行, $y(j)$ 表示 y 的第 j 个元素, “ T ” 表示矩阵的转置运算, “ $*$ ” 表示乘运算, “ $/$ ” 表示除运算, “ \cdot ” 表示矩阵的点乘运算, 即矩阵的各个元素对应相乘, $ones(1, n)$ 表示 1 行 n 列的元素值全为 1 的矩阵;

[0060] Step3 根据式 (10) 重新计算 $error$, 直至 $error$ 小于 ε 或者迭代步数大于 K 时转 Step4, 否则转 Step2。

[0061] Step4 输出调整后的 x 。

[0062] 流量的分类训练的整个流程如图 2 所示, 流量的整个估计流程如图 3 所示, 其中实箭头表示正常输入或输出, 虚箭头表示发送抑制信号 (抑制目的模块运行), 只有当椭圆内约束条件满足时, 才不发送抑制信号。

[0063] 本发明使用时间二范数相对误差来衡量该算法, 时间二范数相对误差 $ReLL2T(t)$ 代表在时刻 t , 所有 OD 流量的构造的向量的二范数误差, 如下:

$$[0064] \quad ReLL2_T(t) = \frac{\sqrt{\sum_{n=1}^N (x_t(n) - \hat{x}_t(n))^2}}{\sqrt{\sum_{n=1}^N x_t(n)^2}} \quad (11)$$

[0065] 其中 $x_t(n)$ 表示在 t 时刻 (time slot) 第 n 个 OD 流的流量, $\hat{x}_t(n)$ 表示 $x_t(n)$ 的估计值。

[0066] 图 4 为多 BP 神经网络同单 BP 神经网络的对比, 其中上图表示时间误差, 下图表示时间误差分布, 整体来看, 采用多 BP 神经网络整体上的误差要优于单 BP 神经网络。这说明采用分类后各个网络的记忆要比一个网络的记忆精确。上图中, 500-750, 1000-1300 time units 中, 多神经网络的误差远远低于单神经网络, 这说明多神经网络能够准确记忆那些和大部分流量差异大的流量。

[0067] 图 5 为多 RBF 神经网络（径向基神经网络）与 RBF 神经网络的对比，其中上图表示时间误差，下图表示时间误差分布。从上图可以看出，由于分类估计，多神经网络在估计与普通流量差异大的流量时，误差较低，或是能判断出某时刻误差会偏高。从下图可以看出，整体上，多 RBF 神经网络的误差要略优于单 RBF 神经网络。

[0068] 径向基神经网络中的径向基函数只在局部区域有强烈的反应，所以使用单 RBF 神经网络估计时能在一定的程度上缓解训练记忆消失或变形的问题，通过图 4 和图 5 的对比也可以发现单 RBF 神经网络的估计误差要低于单 BP 神经网络。通过图 4 中的多 BP 神经网络与图 5 中的单 RBF 网络的误差分布对比可知，多 BP 神经网络要优于单 RBF 网络，这也反应了网络流量往往表现出不同的类别特征，使用径向基函数部分抑制神经元并不能有效缓解记忆消失和变形的问题。所以，对网络流量使用分类估计是一个不错的选择。

[0069] 由以上可知，本发明为一种基于多神经网络的流量矩阵估计方法，它通过对样本流量先分类，再分别训练，有效的克服了传统神经网络用于流量矩阵估计时存在记忆消失或变形的问题，该多神经网络估计的误差明显低于传统的神经网络。

附图说明

[0070] 图 1 是链路流量图；

[0071] 图 2 是分类训练流程图；

[0072] 图 3 是多神经网络估计流程图；

[0073] 图 4 是 BP 神经网络误差对比；

[0074] 图 5 是 RBF 神经网络误差对比。

具体实施方式

[0075] 本实施例提供一种基于 BP 神经网络的多神经网络的估计方法。采用美国的 Abilene IP 骨干网络 04 年 3 月最后一周和 4 月前二周的 OD 流量和链路流量作为样本，估计接下来的三周的 OD 流量。包括样本分类，样本训练和流量估计三个步骤。

[0076] 1、样本分类

[0077] 使用 K-means 算法对样本分类，分类个数 K 使用施原则确定，此处 $K = 20$ ，分类过程中采用欧氏距离表征向量的差异，重复多次使用 K-means 算法直至找到稳定的分类。根据 (4) (5) 计算并记录各个分类的中心 c_i 和半径 d_i ， $i = 1, 2, \dots, K$ 。

[0078] 2、样本训练：

[0079] 对这 K 个分类分别使用 BP 神经网络训练，各分类 BP 网络输入层结点数为 30，中间隐层结点数为 81，输出层结点数为 132（本例中不考虑自己流向自己的流量，Abilene 中有 12 个 POP 结点，故为 11×12 个 OD 流），输入层对应 30 维的链路流量值，输入层对应 132 个 OD 流量值，训练前先对流量做线性归一化处理。使用共轭梯度法对样本进行训练，训练完成后，记录此 K 个神经网络的权值。

[0080] 3、流量估计

[0081] 首先获得某一时刻链路流量的测量值，再做归一化处理，对于满足 (8) 式（警戒参数 ρ 设为 1）的神经网络，分别计算估计值，再使用 (9) 式筛选出最满足链路约束 (1) 的估计值。对该估计值使用 IPFP 算法调整。最后对调整后的流量值进行归一化逆处理，输出最

终的流量估计值。计算下一时刻的流量值。后三周的估计误差如图 3 所示。

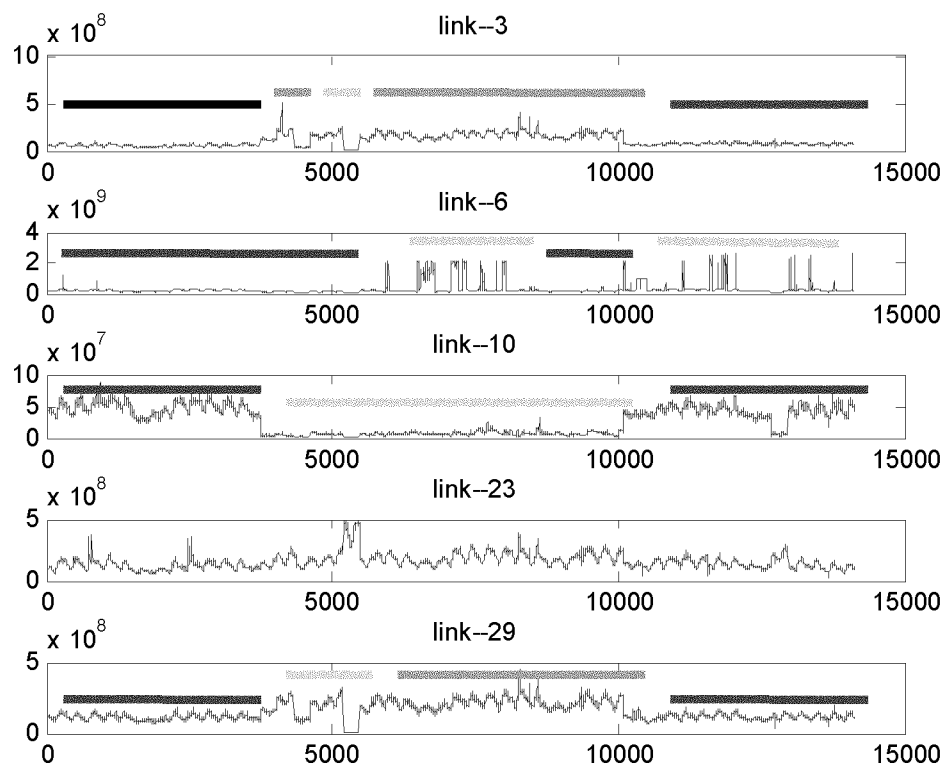


图 1

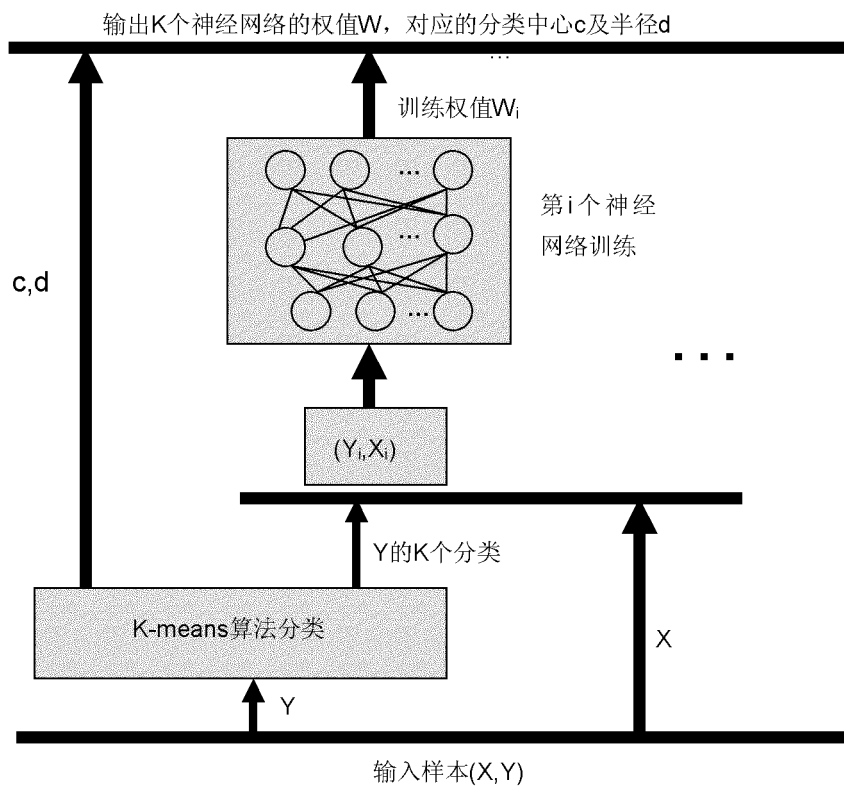


图 2

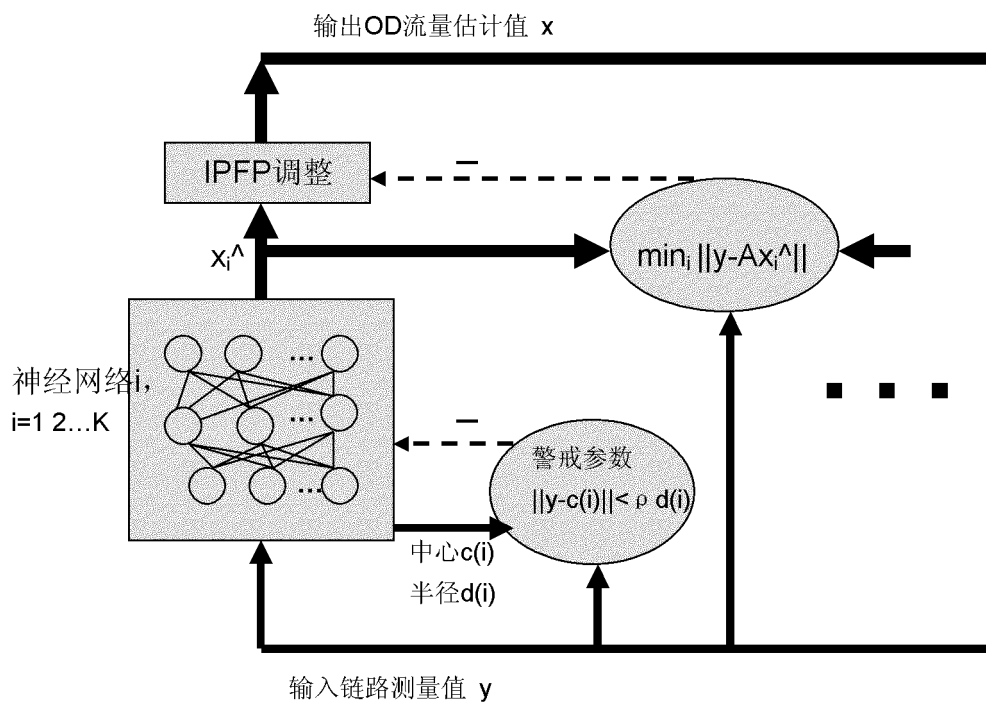


图 3

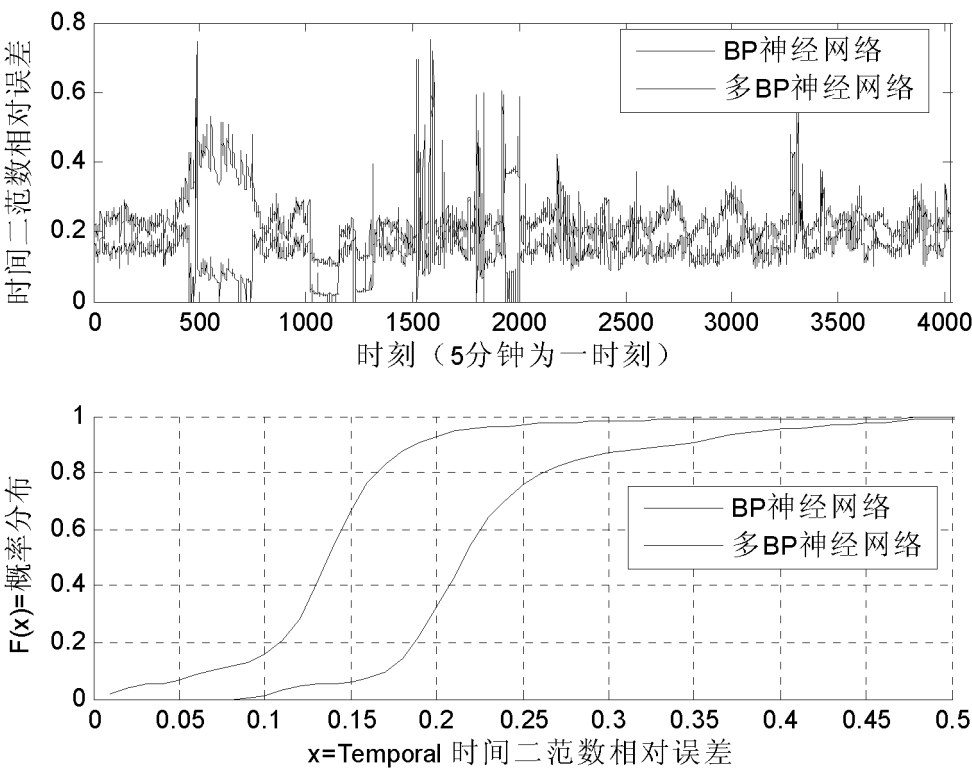


图 4

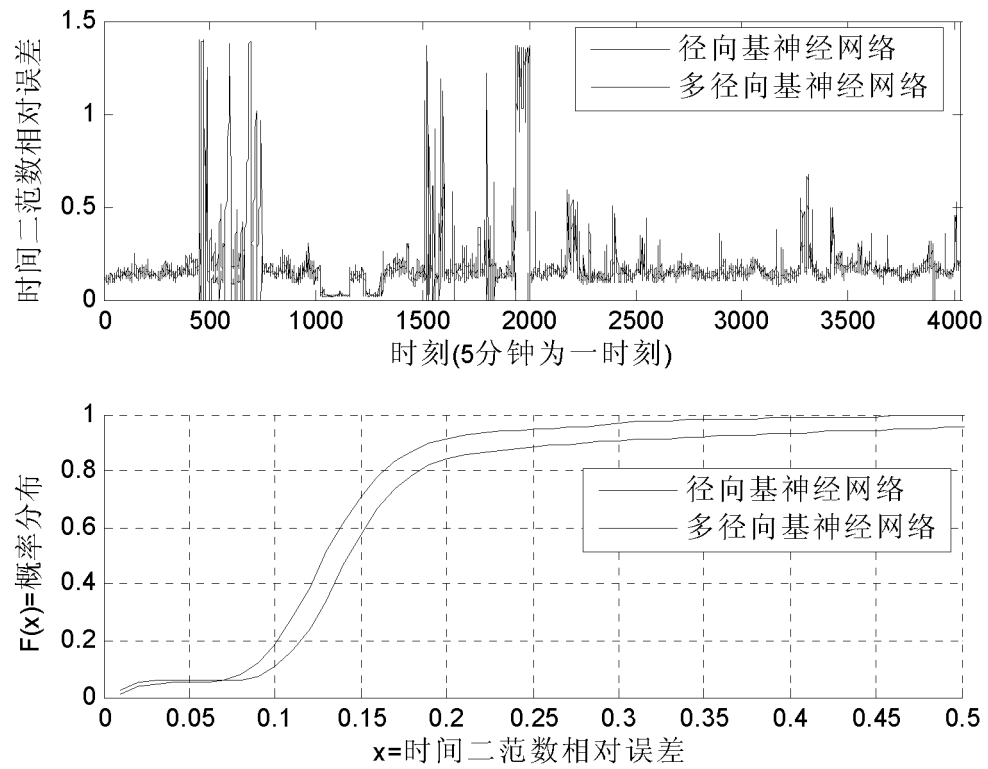


图 5