基于微簇的桥梁监测数据流异常识别研究

吴运宏[1,2] 舒 昕[1,2] 王戒躁[1,2] 闫光辉[3]
[1](中铁大桥局武汉桥梁特种技术有限公司 湖北 武汉 430073)
[2](桥梁结构健康与安全国家重点实验室 湖北 武汉 430205)
[3](兰州交通大学电子与信息工程学院 甘肃 兰州 730070)

摘 要 针对桥梁健康监测系统中的数据流异常问题,提出一种基于微簇的数据流异常检测框架。首先对原始采集信号进行数据合并、缺失值填补等预处理;由于监测系统各传感器测点数据间存在一定的关联,利用主成分分析法提取桥梁主要特征参数,去除重叠信息;利用密度聚类算法把数据流转换成微簇,进行微簇的实时生成,并根据微簇更新机制进行微簇维护,对数据流进行分类。通过对湖北某大桥监测数据的实验表明,该方法具有较好的异常识别能力,可以自适应概念漂移现象。

关键词 桥梁健康监测 数据流异常识别 多传感器网络 主成分分析 微簇

中图分类号 TP306.1 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2016.09.011

ON ANOMALY DETECTION OVER DATA STREAM OF BRIDGE HEALTH MONITORING BASED ON MICRO CLUSTER

Wu Yunhong[1,2] Shu Xin[1,2] Wang Jiezao[1,2] Yan Guanghui[3]
[1](*Wuhan Bridge Special Technology Corporation, Wuhan 430073, Hubei, China*)
[2](*State Key Laboratory for Health and Safety of Bridge Structures, Wuhan 430205, Hubei, China*)
[3](*College of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070, Gansu, China*)

Abstract Aiming at the problem of data stream anomaly in bridge health monitoring system, this paper proposes a micro cluster-based data stream anomaly detection framework. First, it pre-processes the primitive acquisition data by data merging and missing data imputation. Since there is certain correlation between the data of measuring points of each sensor in monitoring system, it uses principal component analysis to extract bridge's main feature parameters in order to remove the redundant information. Then it converts the data streams into micro clusters with density clustering algorithm, carries out real-time generation of micro clusters, and maintains the micro clusters according to their updating mechanism, as well as classifies data streams. It is demonstrated through the experiment on monitoring data of a certain bridge in Hubei Province that the proposed method has stronger capability of anomaly detection, and is able to self-adapt for the concept drift phenomenon.

Keywords Bridge health monitoring Data stream anomaly detection Multi-sensor networks Principal component analysis Micro cluster

0 引 言

桥梁结构健康监测系统通过多传感器网络采集桥梁重要部位数据,对桥梁健康进行智能评估[1]。多传感器网络采集的流式数据连续到达、频繁随时间变化、数据量并不确定[2],根据数据的变化来判断桥梁运营情况,是桥梁健康监测中的一项重要研究内容。当桥梁结构出现损伤或产生其他干扰时,如传感器损坏、监测系统不完善或测试环境发生变化等,便会引起数据流分布不平稳,数据的走向和分布随时间不断变化,产生概念漂移现象[3]。因此,在复杂多变的环境中提高异常数据流的识别精度,正确判断桥梁健康状况,尽可能避免漏检和误检成为桥梁监测中研究的关键问题。

针对数据流异常检测问题,国内外学者做了大量的研究。Muthllkrishnan 等人把离群点定义为异常,解决了基于时间序列的特殊数据流模型的异常检测问题[4];Park 等人建立了用户活动的数据流,对用户正常行为模式建模,采用基于数据流的聚类方法对用户异常行为实现异常检测[5];文献[6]中利用 K-means 算法对大量原始数据进行聚类分析,对处理后得到的簇再用 ID3 决策树进行训练以消除因阈值带来的问题,最后综合两种算法的权值实现分类,提高了预测的准确度。这些方法都能有效地处理海量数据,不需要使用训练数据线下训练,空间复杂度较低。针对传感器网络中的数据,许多学者也提出了相应的算法。Song 等人为了把属性集对聚类算法的影响降至最低,利用

收稿日期:2014-11-19。国家自然科学基金项目(61163010);东湖国家自主创新示范区现代服务业试点项目(2011-dhfwy-029)。吴运宏,高级工程师,主研领域:桥梁施工监控,桥梁维修整治。舒昕,助理工程师。王戒躁,教授级高工。闫光辉,教授。

Apriori 算法先计算属性的频繁项集,然后作为 K-means 聚类算法的属性集合进行聚类,取得了较好的效果[7];李娜等基于信息熵理论提出了一种基于层次聚类的无监督异常检测算法[8]。肖政宏等人利用 K 邻近算法对传感器网络节点进行分簇,簇内节点的异常检测采用贝叶斯分类算法,簇头节点的异常检测则采用平均概率方法,该方案有较高的检测率及较低的误检率[9];文献[10]中研究了时间检测的覆盖空洞问题,设计了精确可认证的分布式覆盖方案保证事件检测的准确性。

以上聚类算法部分基于单传感器,部分不需要离线训练模型阶段。由于正常数据和异常数据特征差异较大,很容易就能将所有未知异常划为一个簇,忽略了数据中的概念漂移。多传感器网络中,异常数据的评价不能以单个传感器瞬时异常为标准进行处理,需要综合评价;数据流分布随时间不断变化,要求算法能对数据进行增量式聚类及自适应识别。针对这些问题,本文在现有工作基础上,提出一种把主成分分析与微簇思想相结合的数据流异常检测框架:利用主成分分析法提取桥梁主要特征,对高维空间的属性进行降维;对数据流用密度算法聚类时,引入微簇模型,把数据流转化为微簇,提高处理效率,自适应分类。真实数据集上的实验表明,该算法对传感器数据流有较好的支持,能有效应对概念漂移,提高异常检测的准确率。

1　桥梁监测数据流分析

多传感器网络采集的数据不再像以前是静态的、有限的、平稳的、低速的数据,而是无限产生的、类分布随时间变化的、高速的、动态的、海量的数据[11],这种像水流一样按顺序产生的数据,就叫数据流。在桥梁健康监测数据流中,数据会随着季节更替、桥梁行车环境、监测系统自身因素(如传感器温漂)而发生改变,异常数据会因异常种类不同,如传感器损坏、船撞等,包含的类标签并不一样,从而产生概念漂移。

概念漂移就是数据流的走向和分布随时间不断变化,数据隐含内容改变而导致目标概念的改变[12],形式主要包括渐进式漂移和突变式漂移。如网络入侵检测中,会因入侵行为发生改变,而产生突变式的概念漂移;顾客的购物偏好会随着季节的变化,产生渐进式的概念漂移。桥梁健康监测中的多传感器网络数据流,既存在渐进式概念漂移,也存在突变式概念漂移[13]。

假设有数据流模型 $S = \{s_1, s_2, \cdots, s_n, \cdots\}$,其中每段都由若干个连续的数据点组成。再设数据段 $S_m = \{s_m, \cdots, s_{p-t}\}$ 是分布平稳的,类标签是 m,表示概念 M;数据段 $S_n = \{s_{p+t}, \cdots, s_n\}$ 分布亦平稳,类标签为 n,表示概念 N。

如图 1 所示,数据流 S 隐含的概念从 M 变化到 N,在时间 $2t$ 内进行,说明发生了概念漂移。当时间 t 取值较小时,可认为发生了突变式概念漂移,发生了很快的数据分布变化;反之,t 较大时,概念漂移的发生是缓慢的,概念为 N 的数据组慢慢"渗透"到概念为 M 的数据组中,直至分布平稳[14]。

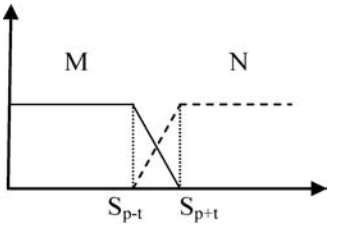

图 1　概念漂移示意图

实际应用中,引起概念漂移的因素是多种多样的。人们并不关心到底发生了哪种漂移、概念漂移对数据流造成的影像有多大,而更加注重于漂移何时发生,如何对当前学习模型进行修正,使更新后的数据模型能够适应新的数据分布。

2　基于微簇的数据流异常检测

对于刚采集来的桥梁监测数据,采集初始时刻不同、不同属性传感器采集频率不一样,包含噪声多,需要对数据进行缺失值填补、时间同步等预处理操作;根据胡顺仁等人在分析桥梁监测系统中各传感器之间关联度提出的观点,不同传感器采集的数据间存在千丝万缕的联系[15],反映信息时有一定重叠,需要把桥梁主要属性提取出来以简化和精炼数据;对提取出来的桥梁主要特征数据用基于微簇的聚类算法对数据进行进一步分析,来判断桥梁健康状况,如图 2 所示。

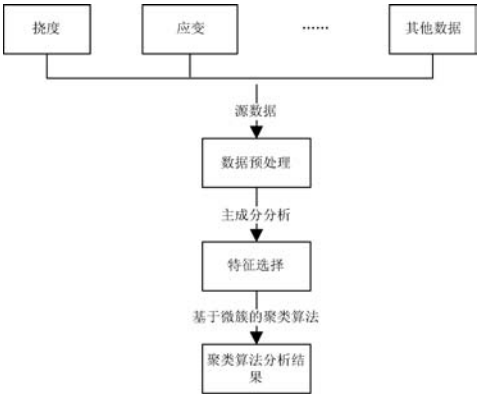

图 2　桥梁监测数据流异常识别流程

2.1　桥梁数据特征提取

主成分分析 PCA 是一种有效的特征提取方法,其基本思想是把原来多个变量通过线性变换得到一组正交基,从而产生新的综合变量,这些新变量的正交性为零[16]。该方法对相关性大的数据进行过滤,在保证原有数据信息丢失最少的情况下选择有代表性的指标,尽量消除属性间的相互影响,对多维空间变量进行降唯处理,达到压缩和简化数据的目的。

已知有 p 维随机变量 $X = \{x_1, x_2, \cdots, x_p\}$,它们的线性组合构成新的综合变量。设 $F1$ 表示原变量的第一个线性组合所形成的主成分指标,即:

$$F1 = a_{11}x_1 + a_{21}x_2 + \cdots + a_{p1}x_p \tag{1}$$

每一个主成分所提取的信息量可以用方差来表示,为了让每一个综合变量尽可能多地包含原有变量的信息,主成分的方差应该越大越好,即 Var($F1$) 达到最大。通常第一主成分 $F1$ 是所有主成分中包含信息量最大的,若第一主成分不足以表达原来 p 个变量的信息,再考虑选取第二、第三主成分,并且 $F1$ 与 $F2$ 之间要保持互相独立,即两者的协方差:

$$\mathrm{Cov}(F1, F2) = 0 \tag{2}$$

以此类推,可以构造出 $F1, F2, \cdots, Fm$ 为原变量指标的第一、第二、…、第 m 个主成分。如式(3)所示:

$$\begin{cases} F1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p \\ F2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p \\ \vdots \\ Fm = a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mp}x_p \end{cases} \tag{3}$$

对于每个系数 a,均应满足规范化条件:

$$a_{i1}^2 + a_{i2}^2 + \cdots + a_{ip}^2 = 1 \quad \mathrm{Cov}(F_i, F_j) = 0$$

需要计算原变量 x_i 与 x_j 的相关系数矩阵：

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{np} \end{bmatrix} \tag{4}$$

其中，$r_{ij}(i,j = 1,2,\cdots,p)$ 为原变量 x_i 与 x_j 相关系数且 $r_{ij} = r_{ji}$，计算公式为：

$$r_{ij} = \frac{\sum_{k=1}^{n}(x_{ki} - \overline{x_i})(x_{kj} - \overline{x_j})}{\sqrt{\sum_{k=1}^{n}(x_{ki} - \overline{x_i})^2 \sum_{k=1}^{n}(x_{kj} - \overline{x_j})^2}} \tag{5}$$

计算相关矩阵 R 的特征值与特征向量。相关矩阵的特征方程为：$|R - \lambda \cdot I| = 0$，求出特征值并按大小顺序排列 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p \geqslant 0$。再由多项式求出对应于特征值的特征向量 e_i，且 $\| e_i \| = 1$。

接下来计算主成分的方差贡献率和累计贡献率。各主成分的方差 $\mathrm{D}(Fi)$ 又称为该主成分的方差贡献，$\dfrac{\mathrm{D}(Fi)}{\sum_{i=1}^{m}\mathrm{D}(Fi)}$ 为它的方差贡献率，公式为：

$$\frac{\mathrm{D}(Fi)}{\sum_{i=1}^{m}\mathrm{D}(Fi)} = \frac{\lambda_i}{\sum_{k=1}^{p}\lambda_k} \tag{6}$$

前 p 个主成分的方差贡献率之和为方差累计贡献率，计算公式为：

$$T = \frac{\sum_{k=1}^{i}\lambda_i}{\sum_{k=1}^{p}\lambda_k} \quad i = 1,2,\cdots,p \tag{7}$$

根据前 n 个主成分的累计贡献率之和超过某一阈值（如 90%）来选取主成分的个数。

2.2 基于微簇的异常识别分析

常规的 K-means 算法、密度聚类等算法都是针对静态数据库一次性完成聚类。然而桥梁监测数据是以数据流的方式不断到达的，这就要求聚类算法必须以增量的方式对数据进行聚类，而不是每当数据到达时都需要扫描数据库来确定数据的标签。本文引入基于微簇的数据流处理方法来提高检测的精度。

定义 1 微簇 一组数据点的集合，这个集合由一个中心点与其半径为 ε 邻域中的数据组成。设一个微簇由多维数据流 $x_1, x_2, \cdots, x_n$ 和时间戳 $T_1, T_2, \cdots, T_n$ 组成，时间戳即为数据点到达的时间。微簇可记为这样一个三元组 (c, ε, ω)。其中 c 为微簇中心点，$c = \dfrac{\sum_{i=1}^{n} f(t - T_i)x_i}{\omega}$；$\varepsilon$ 为微簇半径，$\varepsilon = \dfrac{\sum_{i=1}^{n} f(t - T_i)\, dist(x_i, c)}{\omega}$，$dist(x_i, c)$ 为数据点 x_i 到微簇中心点 c 的欧氏距离，$dist(x_i, c) = \sqrt{\sum_{i=1}^{m}(x_i - c)^2}$；$\omega$ 为权重，$\omega = \sum_{i=1}^{n}\omega_i$，$\omega_i$ 是微簇中数据点 x_i 的时间权重。在二维平面上面，微簇实际上就是以中心点为圆心，以 ε 微半径的圆所包含的数据点的集合。

传统聚类算法中，经过对数据集一次性扫描后数据集被分为不同的类别，这些类形状任意，包含的数据点任意，如图 3 所示。在基于微簇的聚类算法中，数据集合均以微簇的形式存在，并且微簇能够覆盖数据流中随时到达的数据，完成聚类。

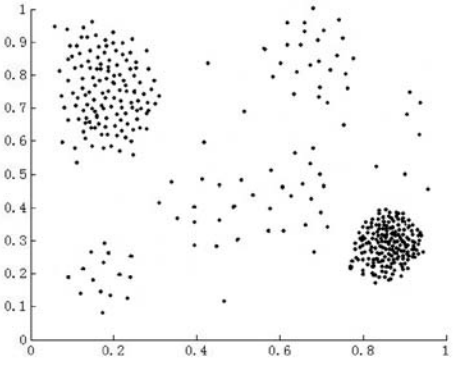

图 3 传统聚类算法

时间窗口分析和 K-means 密度聚类算法将会被用于其中。时间窗口分析就是对某一时刻开始到另一时刻结束时这一段时间内的所有数据进行分析，目的是为了确定需要处理的目标数据流。算法开始工作时，数据集进入内存，当第一个时间窗口数据到达时，微簇并没有形成，需要调用 K-means 密度聚类算法对数据流进行聚类，以形成新的微簇；随着数据流的不断到达，潜在概念漂移可能发生，微簇随着数据流的演化而变化，不断产生新的微簇或者核心微簇组逐渐退化为孤立微簇，即离群点，此时也需要调用密度聚类算法定期更新核心微簇和孤立微簇的集合。核心微簇和孤立微簇定义如下：

定义 2 核心微簇 对于数据流 $p_{i1}, p_{i2}, \cdots, p_{in}$，核心微簇定义为这样一个元组 $(\overline{CF^1}, \overline{CF^2}, \omega, T_s, T_e)$。$\omega = \sum_{j=1}^{n} f(t - T_{ij}) = \sum_{j=1}^{n} \mathrm{e}^{-\lambda(t-T_i)}$ 为核心微簇的权值，其中 $\omega \geqslant \beta\mu, 0 < \beta \leqslant 1$，$\beta$ 为决定孤立点相对于核心微簇的阈值。$\overline{CF^1} = \sum_{j=1}^{n} S(\overline{p}, p_{ij}) \cdot p_{ij}$ 为数据点的加权线性和，其中 $\overline{p} = \sum_{j=1}^{n} \dfrac{p_{ij}}{n}$。$\overline{CF^2} = \sum_{j=1}^{n} S(\overline{p}, p_{ij}) \cdot p_{ij}^2$ 为数据点的加权平方和。核心微簇的中心点为 $c = \dfrac{\overline{CF^1}}{\omega}$，半径为 $r = \sqrt{\dfrac{CF^2}{\omega} - \left(\dfrac{\overline{CF^1}}{\omega}\right)^2}$。$T_s$、$T_e$ 分别表示微簇建立的起始时间和结束时间。

根据核心微簇的定义，核心簇中所包含的数据均为正常数据，最理想的状况就是核心微簇集能覆盖所有的正常数据，且任何一个核心微簇都必不可少。

定义 3 孤立微簇 对于数据流 $p_{i1}, p_{i2}, \cdots, p_{in}$，孤立微簇定义为元组 $(N, \overline{CF^1}, \overline{CF^2}, \omega, T_s, T_e)$。$N$ 表示簇内数据点个数，微簇中心点，半径以及 $\overline{CF^1}$ 与 $\overline{CF^2}$ 的定义都与核心微簇相同，需要注意的是孤立微簇的权值限制为 $\omega < \beta\mu$。孤立微簇对应的是孤立点的集合，即异常数据。

桥梁数据流异常检测的目的就是通过训练数据集建立一个模型，然后监视微簇群的吸收和记录情况并不断更新模型，识别隐含概念漂移的数据流类标签，达到提升识别精度的目的。基

于异常数据在数据流中所占比例较少以及数据特征与正常数据流差别较大的事实[17],可得出:正常数据流产生大量微簇,且不断有新的记录加入而表现活跃;异常数据流产生少量微簇,这些微簇中数据点不如正常微簇群中的数据多,表现较不活跃,且由于产生异常的原因不同,异常微簇之间、异常微簇与正常微簇之间的微簇间距较大[17]。微簇生成步骤如下:

(1) 对训练集 x 采用 K-means 聚类算法;

(2) 设训练集中第一个点生成微簇 C_1,并加入微簇队列;

(3) 取训练集 x 中的点 x_i,计算 x_i 到微簇队列中所有点的欧式距离,若该距离小于微簇半径,则 x_i 加入并更新该微簇;若该距离大于微簇半径,则生成新的核心微簇或孤立微簇;

(4) 重复步骤(2)、步骤(3),直至扫描完训练集中所有点。

根据训练集建立的初始模型,微簇队列包含核心微簇和孤立微簇,接下来对于数据流中新到达的对于数据流中新到达的数据点 p,有如下步骤:

(1) 当该点与核心微簇中心点 c 的欧氏距离满足 $dist(c,p) \leqslant \varepsilon$ 时,p 被核心微簇吸收。如图 4 所示,实线部分的圆表示核心微簇,虚线部分的圆表示孤立微簇, 数据点 P_1 落在核心微簇中直接被吸收,更新核心微簇。

(2) 当该点与孤立微簇中心点 c 的欧氏距离满足 $dist(c,p) \leqslant \varepsilon$ 时,p 被孤立微簇吸收,同时计算微簇的权值 ω,若 $\omega \geqslant \beta\mu$,达到核心微簇的条件,则把该微簇从孤立微簇队列中删除,加入到核心微簇中;$\omega < \beta\mu$,更新该孤立微簇。如图 4 中的 P_2。

(3) 当该点到所有微簇(核心微簇和孤立微簇)中心点 c 的欧式距离满足 $dist(c,p) > \varepsilon$ 时,该点不属于任何类,以该点为中心构造孤立微簇,同时加入孤立微簇队列。如图 4 中的点 P_3。

(4) 重复步骤(1)至步骤(3)。

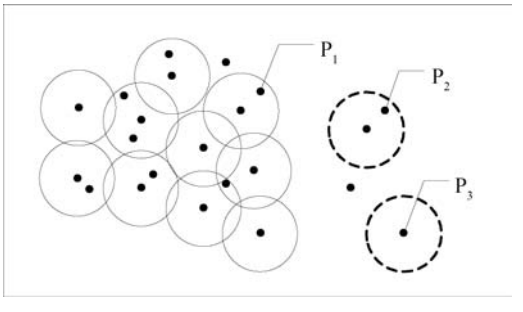

图 4　微簇进化示意图

3　实验结果及评价

3.1　实验数据采集及预处理

本文采用湖北省某长江公路大桥的监测数据作为实验数据。该桥作为沪蓉高速公路主干线湖北省东段和国家高速公路网规划中的大庆至广州高速公路湖北段的共用过江通道,是跨径组合为(3×67.5+72.5+926+72.5+3×67.5) m 的 9 跨半漂浮体系双塔混合梁斜拉桥。主桥桥面最大纵坡 2.0%,双向 6 车道;钢箱梁外形与混凝土箱梁一致,均为分离式双箱单室结构,主梁全宽为 38 m,梁高 3.8 m。根据桥梁的力学特性分析,大桥监控的主要参数为结构温度、环境温湿、斜拉索索力、主梁挠度和振动、结构应变等。每种传感器均主要安装在大桥斜拉索、主梁及主塔上。

实际桥梁监测中,相对来说比较容易获得大量的正常数据样本,结构损伤样本不太容易获取。由于传感器数量较多,本文主要选取 4 号墩到 5 号墩的部分传感器数据作为样本,包括:4 个索力传感器、2 个挠度仪、2 个应变计、2 个加速度传感器及 2 个温湿度传感器,并把它们重新编号。有选择性地选取了不同时间段的桥梁监测数据样本,共 50 000 条数据,人工为其标记,正常数据 45 920 条,传感器损坏异常样本集 1980 条、因大风引起的信号异常样本 850 条、因船撞引起的结构异常样本集 320 条和疲劳损伤样本集 930 条。图 5 是索力监测数据的采集界面。

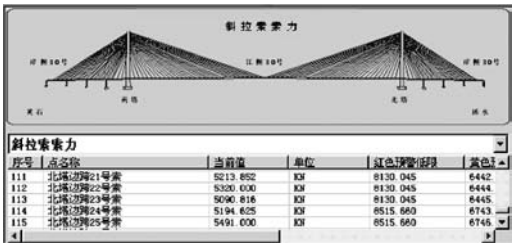

图 5　大桥索力监测界面

传感器采集数据时,不同类型的传感器采集的初始时间不同,采集频率不一样,采集时间间隔不同,还会有缺失值,因而需要对数据进行预处理。对于频率不同的数据,把时间最接近的各个参数记录合并到一起,使之在相同的时间段内。当出现单点缺失值时,用该时刻前序 2 个数据和后续 2 个数据的平均值代替;当出现连续缺失值时,很大可能是传感器出现异常,直接归为异常数据样本集即可。表 1 是对某一时间段内的数据进行采集,然后经过预处理后得到的值。

表 1　某段时间段内数据统计

属性	最大值	最小值	平均值	阀值
索力 1(kn)	5912	5560	5771	7204
索力 2(kn)	5864	5520	5700	7204
索力 3(kn)	5712	5490	5610	7204
挠度 1(mm)	28	21	26	60.2
挠度 2(mm)	26	18	24	60.2
应力 1(mpa)	72	69	71	100
应力 2(mpa)	61	57	60	100
加速度 1(m/s^2)	2.51	1.82	2.12	3.50
温度(c)	36.2	20.5	32.3	55
湿度(%)	90	60	70	100

3.2　评价方法

本文采用查全率(recall)和精度率(precision)对实验结果进行评价。

算法判断数据流 X 是否属于类 C 时,输出的结果集可以分为 C_0、C_1 和 C_2 三种情况。其中 C_0 表示分类正确的数据集,即识别出数据流 P 属于类 C,并且被验证确实属于类 C 的数据集合;C_1 表示这些数据不属于类 C 但算法把它识别错误识别为类 C 的数据;C_2 表示这些数据属于类 C 但算法却没有识别出来的数据集合。因此查全率公式定义为:

$$Recall = \frac{C_0}{C_0 + C_1} \tag{8}$$

式(8)用来度量算法发生数据漏检的情况。精度率定义为:

$$Precision = \frac{C_0}{C_0 + C_2} \tag{9}$$

式(9)用来度量算法发生数据误判的情况。

3.3 实验结果

首先利用主成分分析方法对收集的数据进行特征提取与降维,由于收集的数据集中湿度变化不大,排除湿度对结果的影响。在此 $P=9, x=\{x_1,\cdots,x_9\}$。通过 Matlab 计算得到前5个主成分的贡献率分别为: $per=[49.90, 15.35, 13.52, 10.85, 4.77]$,所对应的特征值为 $egenvalue=[4.49, 1.38, 1.21, 0.98, 0.43]$。前5个主成分的累计贡献率已达到94%,若按90%以上的信息量来设定阈值,则可以选取前5个新的特征元素,其中第一个新因子 $F1$ 所含信息量最大。

式(10)给出了前5个主成分的系数矩阵,则之前的9个属性 $\{x_1,\cdots,x_9\}$ 可以用5个新的变量 $F1$、$F2$、$F3$、$F4$、$F5$ 表示,其各属性之前的系数就对应着矩阵里面的列值。

$$F=\begin{bmatrix} -0.03 & +0.65 & +0.45 & -0.28 & +0.38 \\ +0.36 & +0.37 & +0.08 & +0.37 & -0.32 \\ +0.35 & -0.19 & -0.28 & +0.39 & +0.52 \\ +0.41 & -026 & +0.24 & -0.03 & -0.21 \\ +0.36 & -0.27 & +0.08 & -0.36 & +0.41 \\ +0.32 & -0.04 & +0.58 & +0.34 & +0.12 \\ +0.26 & +0.47 & +0.58 & +0.04 & +0.22 \\ +0.40 & +0.17 & -0.14 & -0.13 & -0.45 \\ +0.35 & -0.06 & -0.09 & -0.60 & -0.09 \end{bmatrix} \quad (10)$$

新得到的变量也是数据流形式,用基于微簇的数据流异常检测算法对其进行分类,选取50%的数据样本作为训练数据集,剩下的数据作为测试集。需要设定的实验参数主要包括微簇半径 r,孤立点相对于核心微簇的阈值 β,衰减因子 λ。微簇半径 r 应大小适中,太大则同一个簇中会出现不同标签的数据点,太小则增加了算法空间复杂度,达不到聚类效果,该数据需要在实验中凭经验控制;衰减因子决定了历史数据对当前数据流影响的重要程度,可根据实际需要进行设定;阈值 β 则按照桥梁监测数据流的特性来判定。根据文献[18],初始化参数进行如下设置, $InitN=1000$,数据流速 $V=1000$,衰减因子 $\lambda=0.25$,阈值 $\beta=0.01$。

图6中左边的柱形代表查全率,右边柱形代表精度率。可以看出算法对每一类异常的识别效果,包括查全率和精度率。算法对四种异常情况的分类结果准确率大多都在80%以上,船撞事件的精度识别率在75%以上,精度略有降低。因桥梁传感器系统使用环境复杂并且多变,并且往往包含许多噪声数据,易引起数据流的分布不平稳,即发生概念漂移现象,需要测试在发生概念漂移时算法能否自适应处理这些数据流。用精度表示识别正确的数据流与总数据流的比值,建立了精度随数据流的变化曲线坐标轴,如图7所示。

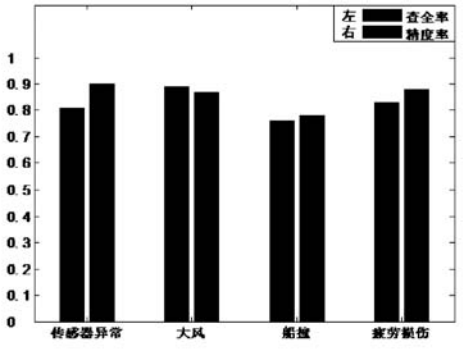

图6 分类查全率及精度率

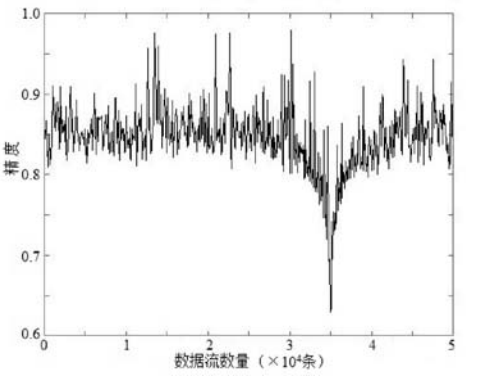

图7 算法识别精度变化

从图7中可以看出,算法识别精度主要保持在80% ~90%之间,在第32 130条数据流出现时识别精度开始呈下降趋势,第34 870条流到达后识别精度达到最低。通过对数据流标签的监测,发现出现概念漂移的数据流是传感器异常数据流,传感器出现异常时,连续缺失值和无效值明显增多,造成数据分布紊乱,算法识别精度明显下降。随着数据流的继续达到,识别精度逐渐升高,证明了该算法的有效性。

4 结 语

本文对桥梁健康监测系统中的多传感器网络环境进行了分析,提出了一种基于微簇的多属性数据流异常监测框架。监测桥梁的传感器往往数量繁多,存在着复杂多样的关联关系,为了降低时间和空间复杂度,并没必要分析所有数据。首先运用主成分分析原理,对判别桥梁健康状态意义不大、包含重复信息的属性进行降维,以减少数据的时空开销;然后利用微簇的思想对新的主成分进行微簇分类,识别异常数据。真实数据集上的实验表明,在应对桥梁监测数据随时间变化分布不平稳而产生概念漂移的问题时,有较好的分类效果。本文也存在不足之处,如考虑到传感器众多,只是针对桥梁进行局部分析,若要做全全桥分析,算法的时间复杂度和空间复杂度还需要进一步降低。

参 考 文 献

[1] Jingqiu Huang, Ogai H, Chen Shao, et al. On vibration signal analysis in Bridge Health Monitoring System by using Independent Component Analysis[C]//SICE Annual Conference, IEEE, 2010: 2122 - 2125.

[2] 陈斌,陈松灿,潘志松,等. 异常检测综述[J]. 山东大学学报:工学版,2009,39(6):13 - 21.

[3] Elwell R Polikar. Incremental learning of concept drift in nonstationary environments[J]. IEEE Trans on Neural Networks, 2011, 22(10): 1517 - 1531.

[4] Muthukrishnan S, Shah R, Vitter J S. Mining deviants in Time Series Data Streams[C]//Proceedings of the 16th International Conference on Scientific and Statistical Database Management, 2004: 41 - 50.

[5] Park N H, Oh S H, Lee W S. Anomaly intrusion detection by clustering transactional audit streams in a host computer[J]. Information Sciences, 2010, 180(12): 2375 - 2389.

[6] Yasami Y, Mozaffari S P. A novel unsupervised classification approach for network anomaly detection by k-means clustering and ID3 decision tree learning methods[J]. Journal of Supercomputing, 2010, 53(1): 231 - 245.

(下转第96页)

0.0170,0.0715,0.0890,0.0052,0.0054,…),同理可得到其他四个城市配电网的评价指标综合权重向量 ww_2、ww_3、ww_4 以及 ww_5。

3.2 综合评价及结果分析

由五座城市的指标测度评价矩阵 $\boldsymbol{M}_1$、$\boldsymbol{M}_2$、$\boldsymbol{M}_3$、$\boldsymbol{M}_4$ 以及 $\boldsymbol{M}_5$ 和每座城市相对应的综合权重 ww_1、ww_2、ww_3、ww_4 以及 ww_5,由式(5)即可求得综合测度评价矩阵:

$$\boldsymbol{H}=\begin{bmatrix} 0.4404 & 0.4380 & 0.0289 & 0.0781 & 0.0145 & 0 \\ 0.4964 & 0.3322 & 0.0782 & 0.0236 & 0.0426 & 0.0269 \\ 0.4901 & 0.2871 & 0.0695 & 0.0522 & 0.0175 & 0.0837 \\ 0.1835 & 0.1809 & 0.1831 & 0.0804 & 0.0505 & 0.3217 \\ 0.4768 & 0.1382 & 0.0634 & 0.0995 & 0.1424 & 0.0797 \end{bmatrix}$$

在本文中,置信度 λ 取 0.7,由式(7)、式(8)可计算出五个城市配电网的综合得分,A 城市得分 5.21,B 城市得分 5.14,C 城市得分 4.93,D 城市得分 3.40,E 城市得分 4.47,还可以得到 A 城市配电网的等级最高,程序运行结果为等级 2,B 城市和 C 城市的配电网为等级 2,D 城市的配电网为等级 6,E 城市的配电网为等级 4,即配电网发展水平由高到低分别为:A、B、C、E、D。符合配电网 A 是一线城市,配电网 B 和 C 是两个省会级城市,配电网 E 和 D 是一个普通地市级城市的实际情况。此外,该结果与韩震焘在文献[20]中得到的实验结果也相互吻合,进一步印证了方法的科学性与合理性。

4 结 语

本文提出了一种对于配电网多指标体系的综合评价方法,通过对中间值指标进行数值偏移率处理,基于属性识别理论,建立了样本空间矩阵与指标测度评价矩阵,弥补属性识别不能直接处理中间值指标的缺陷。结合主观 G1 赋权法以及客观熵权赋权法,进行目标配电网的综合评价。与相关研究结果的实例对比分析表明,本文提出的综合评价方法科学、合理。同时,本文所提出的基于属性识别的综合评价方法也为其他领域多指标复杂系统的综合评价提供了参考。

参 考 文 献

[1] 魏祖宽,熊娅秋,金在弘. 组件技术在配电网专业分析功能中的研究与应用[J]. 计算机应用与软件,2010,27(6):175-177,241.

[2] 刘志刚,谢志林,徐敏锐,等. 智能电网互动终端系统设计与实现[J]. 计算机应用与软件,2012,29(1):276-279.

[3] 王成山,罗凤章. 配电系统综合评价理论与方法[M]. 北京:科学出版社,2012:191-240.

[4] 吴开贵,王韶. 基于 RBF 神经网络的电网可靠性评估模型研究[J]. 中国电机工程学报,2000,20(6):9-12.

[5] 谢莹华,王成山. 基于馈线分区的中压配电系统可靠性评估[J]. 中国电机工程学报,2004,24(5):35-39.

[6] Li W,Wang P,Li Z,et al. Reliability evaluation of complex radial distribution systems considering restoration sequence and network constraints[J]. IEEE Transactions on Power Delivery,2004,19(2):753-758.

[7] 郭志忠,刘伟. 配电网安全性指标的研究[J]. 中国电机工程学报,2003,23(8):85-90.

[8] 顾伟,蒋平,蔡桂龙,等. 地区电能质量评估及综合治理研究[J]. 电力需求侧管理,2004,5(2):20-23.

[9] 刘颖英,徐永海,肖湘宁. 地区电网电能质量综合评估新方法[J]. 中国电机工程学报,2008,28(22):130-136.

[10] 张蔓,林涛,曹健,等. 理想区间法在电能质量综合评估中的应用[J]. 电网技术,2009,33(3):33-38.

[11] 李欣然,刘杨华,朱湘有,等. 高压配电网建设规模的评估指标体系及其应用研究[J]. 中国电机工程学报,2006,26(17):18-24.

[12] 李晓辉,张来,李小宇,等. 基于层次分析法的现状电网评估方法研究[J]. 电力系统保护与控制,2008,36(14):57-61.

[13] 张心洁,葛少云,刘洪,等. 智能配电网综合评估体系与方法[J]. 电网技术,2014,38(1):41-45.

[14] 黄志成. 基于模糊聚类的 CSCL 学习者混合属性分组[J]. 计算机应用与软件,2011,28(2):118-121.

[15] 李军,李继光,姚建刚,等. 属性识别和 G1 熵权法在电能质量评价中的应用[J]. 电网技术,2009,33(14):56-61.

[16] 郭奇,曹洪洋. 大气环境质量评价的属性识别法[J]. 环境监测管理与技术,2004,16(3):41-44.

[17] 郭延永,刘攀,吴瑶,等. 基于属性识别的高速公路交通安全设施系统评价[J]. 东南大学学报,2013,43(6):1306-1310.

[18] 张龙云,曹升乐,杨尚阳. 属性识别理论在水安全评价中的应用研究[J]. 山东大学学报,2006,36(5):70-72.

[19] 程乾生. 属性识别理论模型及应用[J]. 北京大学学报:自然科学版,1997,33(1):12-20.

[20] 韩震焘. 城市配电网综合评价体系研究[D]. 天津:天津大学,2012,35-43.

[21] 罗毅,李昱龙. 基于熵权法和灰色关联分析法的输电网规划方案综合决策[J]. 电网技术,2013,37(1):78-80.

[22] 高建明,龚亮亮,吕涛. 基于信息熵的目标平台识别方法[J]. 计算机应用与软件,2013,30(9):224-227.

[23] 张璇,廖鸿志,李彤,等. 基于信息熵和攻击面的软件安全度量[J]. 计算机应用,2013,33(1):19-22,48.

(上接第 50 页)

[7] Song N,Zhou G. A study on intrusion detection based on data mining[C]//International Conference of Information Science and Management Engineering,2010:135-138.

[8] 李娜,钟诚. 基于划分和凝聚层次聚类的无监督异常检测[J]. 计算机工程,2008,34(2):120-123.

[9] 肖政宏,陈志刚,李庆华. WSN 中基于分布式机器学习的异常检测数据研究[J]. 系统仿真学报,2011,23(1):121-127.

[10] Bejerano Y. Coverage verification without location information[J]. IEEE Trans on Mobile Computing,2012,11(4):631-643.

[11] Knorr E M,Ng R T. Algorithms for mining distance-based outliers in large datasets[C]//Proceedings of VLDB 1998 C J. NewYork,USA,Morgan Kaufmann,1998:392-403.

[12] Widmer G,Kubat M. Learning in the Presence of Concept Drift and Hidden Contexts[J]. Machine Learning,1996,23(1):69-101.

[13] Elwell R,Polika R. Incremental learning of concept drift in non-stationary environments[J]. IEEE Trans on Neural Networks,2011,22(10):1517-1531.

[14] Tsymbal A. The problem of concept drift:definitions and related work[J]. Computer Science Department,Trinity College Dublin,2004.

[15] 胡顺仁,陈伟民,章鹏. 桥梁监测系统多传感器测点之间的关联分析[J]. 土木工程学报,2009,42(3):81-86.

[16] 梁栋,张宇峰,袁慎芳,等. 桥梁数据异常诊断方法在伸缩缝中的应用[J]. 数据采集与处理,2011,26(5):579-584.

[17] 肖三,杨雅辉,沈晴霓. 基于微簇的在线网络异常检测方法[J]. 计算机工程与应用,2013,49(6):86-90.

[18] 任培花. 基于微簇进化学习的数据流快速聚类算法研究[J]. 计算机仿真,2013,30(3):343-346.