基于离群点挖掘的工业控制系统异常检测

陈 庄 黄 勇 邹 航
(重庆理工大学计算机科学与工程学院 重庆 400054)

摘 要 目前,工业控制系统广泛应用于我国电力、水利、污水处理、石油天然气、化工、交通运输、制药以及大型制造行业,针对工业控制系统的攻击越来越频繁,而目前市场上工业控制系统的安全产品十分稀少。虽然主流的组态软件具有控制变量报警功能模块,但其只能处理单一变量超过阈值时的报警,不能识别出由多个变量共同引起的异常。为此,针对工业控制系统的变量数据、通信协议、高实时性等特点,提出了基于自适应聚类的离群点挖掘方法——ACBOD方法,该方法包括数据采集、聚类、簇的标识以及簇的离群点检测4个阶段,对工业控制系统OPC Server上的变量数据进行数据分析。实验证明,该方法可以很好地发现工业控制系统中的异常数据,并能够发现未知的异常,能够极大地提高工业控制系统的安全防护能力。

关键词 工业控制系统,聚类,离群点挖掘,自适应聚类,异常检测
中图法分类号 TP393 **文献标识码** A

Anomaly Detection of Industrial Control System Based on Outlier Mining

CHEN Zhuang HUANG Yong ZOU Hang
(College of Computer Science and Engineering, Chongqing University of Technology, Chongqing 400054, China)

Abstract At present, industrial control system is widely used in electric power, transportation, water conservancy, large manufacturing industry and national critical infrastructure. ICS has become the important part of the national security strategy. The attacks against to the industrial control systems are more and more frequent, and there are little security products specifically for the industrial control system. Although most of the configuration software has variable alarm function, it is just sutable for a single variable, rarely from an overall consideration of the overall security. In order to effectively improve the industrial control system information security protection, based on the specific data and protocol and the highly real-time requirement, this paper proposed the Adaptive Clustering-Based Outlier Detection——ACBOD method to analyze the variable data from the OPC Server. This method has 4 parts: data acquisition, clustering, Identification of clusters, and the cluster outlier detection. The testing results show that this method can find abnormal data in industrial control systems effective, also can find an unknown exception, and it can greatly improve the industrial control system safety protection ability.

Keywords Industrial control system, Clustering, Outlier mining, Adaptive clustering, Abnormal behavior detection

1 引言

工业控制系统是由各种自动化控制组件以及对实时数据进行采集、监测的过程控制组件共同构成的确保工业基础设施自动化运行、过程控制与监控的业务流程管控系统[1]。工业控制系统广泛应用于我国电力、水利、污水处理、石油天然气、化工、交通运输、制药以及大型制造行业,其中超过80%的涉及国计民生的关键基础设施依靠工业控制系统来实现自动化作业,工业控制系统已是国家安全战略的重要组成部分[2]。针对工业控制系统的攻击越来越频繁,2010年,"网络超级武器"Stuxnet病毒通过针对性的入侵工业控制系统,损坏其中央离心机,严重威胁到伊朗布什尔核电站核反应堆的安全运营。而目前市场上专门针对工业控制系统的安全产品十分稀少,工业控制系统的安全问题十分严峻,急需针对工业控制系统的数据特点、操作特点有针对性地研究相关的安全技术,以提升工业控制系统的整体安全。

由于入侵行为与正常行为本质上是可区分的(这也是对入侵能够检测的前提),在行为特征空间中,相对于正常行为的分布区域,一个入侵行为的特征向量是离群点,因此,通过离群点检测方法可以有效地实现异常检测。离群点检测是数据挖掘的基本任务之一[3,4],故称为离群点挖掘,其目的是消除噪音或发现潜在的、有意义的知识,在很多情况下,罕见的事件比正常出现的事件更令人感兴趣,例如复杂工业生产过程中的参数异常波动[5],这些异常数据可以用来检测工业控

到稿日期:2013-06-25 返修日期:2013-10-27 本文受科技型中小企业技术创新基金项目(12C26115116106),重庆理工大学研究生创新基金(YCX2012102)资助。

陈 庄(1964—),男,博士,教授,主要研究方向为企业信息化管理、网络与信息安全,E-mail:zhuang.ch@gmail.com;黄 勇(1989—),男,硕士,主要研究方向为网络与信息安全;邹 航(1979—),男,硕士,实验师,主要研究方向为网络与信息安全。

制系统是否正常运行。

组态软件中虽然也有相关的变量报警提示，但都只是针对某一个模拟量或者开关量的报警提示，不能针对全局考虑，有可能会出现所有的变量和开关量都在报警阈值之内，但实际上已经是异常操作，而组态软件自带的报警提示对这类异常不能察觉。本文通过对工业控制系统的数据特点与操作特点进行相关研究，针对性地提出了基于自适应聚类的离群点挖掘算法 ACBOD(Adaptive Clustering-Based Outlier Detection)，采用自适应阈值的方法自适应地选取最佳的聚类结果进行离群点检测。该方法首先通过 OPC 协议获取工业控制系统 OPC Server 上的数据，并对采集到的数据进行数据预处理，通过聚类算法对预处理之后的数据进行聚类，然后计算每一个簇的离群因子，将离群因子大的簇判定为离群簇，该离群簇中的所有对象则被判定为异常，实现对工业控制系统异常行为的有效检测。

2 离群点挖掘方法的研究现状

通过对现有的离群点挖掘方法的深入分析、比较，得到如下几个有用的结论：

(1)基于统计的方法与基于距离的离群点挖掘方法[6,7]都是从全局角度考虑的全局一致方法，当数据集含有多种分布或数据集由不同密度子集混合而成时，算法效果不佳；基于聚类的离群点挖掘算法考虑到了数据的局部性质，在很大程度上克服了这一不足。

(2)基于距离的离群点挖掘方法与基于密度的离群点挖掘方法以及基于神经网络的离群点挖掘方法的时间复杂度一般为 O($m \cdot N^{\wedge} 2$)(这里 N 是问题规模，m 是数据维数)或更高[8-10]，其可扩展性差，难以用于大规模数据集和增量更新，更不能满足数据流处理对时效性的要求；而且基于神经网络的离群点挖掘方法一般需要有高质量的样本数据进行训练，而在实际应用中，能获得训练样本的情况不多，绝大多数都是没有标定的正常的训练样本，而基于聚类的离群点挖掘方法则很大程度上弥补了该缺陷。

(3)大部分离群点挖掘方法需要人工设计参数或规则[11]，在环境改变(如网络环境变化、用户群体发生变化等)时必须重新进行人工设计，算法的灵活性、适应性较差[12]。本文提出的自适应阈值选择方法能很好地解决算法灵活性、适应性等问题。

(4)绝大部分的离群点挖掘方法不能处理混合属性的数据，而工业控制系统的数据中不仅有下位机采集到的模拟量还有对应的开关量(也就是布尔量)，采用基于聚类的离群点挖掘算法则能很好地处理这种问题。

鉴于上述分析，本文提出了基于自适应聚类的离群点挖掘算法，并将其应用到工业控制系统异常检测之中。

3 算法的相关定义与计算

3.1 相关定义

假设数据集 D 有 m 个属性，其中有 m_c 个分类属性和 m_N 个数值属性，$m=m_c+m_N$，用 D_i 表示第 i 个属性的集合。

定义 1 如果一个对象不强属于任何簇，则称该对象为基于聚类的离群点。

定义 2 给定簇 C，$a \in D_i$，a 在 C 中关于 D_i 的频度定义为 C 在 D_i 上的投影中包含 a 的次数：

$$Freq_{C|D_i}(a)=|\{object|object \in C, object\ D_i=a\}|$$

定义 3(摘要信息 CSI) 给定簇 C，C 的摘要信息 CSI (Cluster Summary Information)定义为：

$$CSI=\{n, Cluster, Summray\}$$

其中，n 为类 C 的大小，Cluster 为类 C 中对象标识的集合，Summary 由分类属性中不同取值的频度信息和数值属性的质心两部分构成，即：

$$Summary=\{\langle Stat_i\ Cen\rangle \mid Stat_i=(a, Freq_{C|D_i}(a)) \mid a \in D_i, 1 \leqslant i, j \leqslant m_c, Cen=(p_{m_c+1}, p_{m_c+2}, \cdots, p_{m_c+m_N})\}$$

3.2 相关计算

3.2.1 模拟变量值属性的规范化

为了减小数值属性不同度量单位对结果的影响，需要对模拟变量值属性进行规范化，本文采用文献[3]中的绝对方差标准化对数值属性进行标准化。

(1)计算平均的绝对偏差 S_f

$$S_f=\frac{1}{m}(|x_{1f}-m_f|+|x_{2f}-m_f|+\cdots+|x_{nf}-m_f|) \quad (1)$$

其中，$x_{1f}, x_{2f}, \cdots, x_{nf}$ 是对象 f 的 m 个度量值，m_f 是对象 f 的平均度量值，即：

$$m_f=\frac{1}{m}(x_{1f}+x_{2f}+\cdots+x_{mf})$$

(2)计算标准化的度量值 Z_{if}

$$Z_{if}=\frac{x_{if}-m_f}{S_f} \quad (2)$$

其中，$i=1,2,\cdots,m$。

3.2.2 距离的计算

给定 D 的簇 C、C_1 和 C_2，对象 $p=[p_1, p_2, \cdots, p_m]$ 与 $q=[q_1, q_2, \cdots, q_m]$。

(1)对象 p，q 在属性 i 上的距离 $dif(p_i, q_i)$ 定义为：

对于分类属性或二值属性

$$dif(p_i, q_i)=\begin{cases}1, & p_i \neq q_i\\ 0, & p_i=q_i\end{cases}$$

对于数值属性或顺序属性

$$dif(p_i, q_i)=|p_i-q_i|$$

(2)两个对象 p、q 间的距离 $d(p,q)$ 定义为每个属性上的距离的幂平均值，即

$$d(p,q)=\sqrt{\sum_{i=1}^{m}\frac{dif(p_i, q_i)^2}{m}} \quad (3)$$

(3)对象 p 与簇 C 间的距离 $d(p,C)$ 定义为 p 与簇 C 的摘要之间的距离：

$$d(p,C)=\sqrt{\sum_{i=1}^{m}\frac{dif(p_i, C_i)^2}{m}} \quad (4)$$

其中，$dif(p_i, C_i)$ 为对象 p 与簇 C 在属性 D_i 上的距离。

对于分类属性 D_i，其值定义为 p 与簇 C 中每个对象在属性 D_i 上的距离的算术平均值，即

$$dif(p_i, C_i)=1-\frac{Freq_{C|D_i}(p_i)}{|C|}$$

对于数值属性 D_i，其值为：

$$dif(p_i, C_i)=|p_i-C_i|$$

(4)簇 C_1 和 C_2 间的距离 $d(C_1, C_2)$ 定义为两个摘要之间的距离：

$$d(C_1,C_2)=\sqrt{\frac{\sum_{i=1}^{m} dif(C_i^{(1)},C_i^{(2)})^2}{m}} \tag{5}$$

其中，$dif(C_i^{(1)},C_i^{(2)})$为簇 C_1 和 C_2 在属性 D_i 上的距离。

3.2.3 半径阈值 r 的计算

聚类算法中参数 r 将影响聚类的结果和算法的时间效率。本文采用基于抽样技术的自适应计算阈值范围的方法，具体描述如下：

(1)在数据集 D 中随机选择 N_0 对对象。

(2)计算每对对象间的距离。

(3)计算式(2)中距离的平均值 EX 和标准差 DX。

(4)半径阈值 $r=EX+\beta\cdot DX$，β 在 0.25 到 −2 之间视情况取值。

3.2.4 半径阈值 r 的评估——DB 指数的计算

DB 指数(Davies-Bouldin Index)是一种衡量聚类质量的方法[13]。当簇间距离增大、簇内距离变小时，DB 指数随之变小，最后指示分簇效果趋好，也就是说 DB 指数越小，聚类所达到的效果越好。本文采用 DB 指数是为了能够选择出最优的半径阈值 r，以期在聚类阶段达到最好的分类结果。本文使用 DB 指数找出最优的 β 取值，每一个数据集的阈值半径都不尽相同，因此，需要在进行大规模数据计算之前，采用抽样技术对该数据集进行估量计算，求出最优的阈值半径 r。

簇间距离是度量簇与簇之间的分离程度，簇内距离是度量簇内的紧凑程度。本文分别将质心链接、质心直径作为度量簇间距离与簇内距离的方法[14]，如式(6)所示：

$$d(C_i,C_j)=d(SC_i,SC_j) \tag{6}$$

其中，SC_i 表示第 i 簇的中心点，该公式指的是第 i 簇的中心点与第 j 簇的中心的距离，即质心链接。

$$\Delta C_i=2\{\frac{\sum_{p=1}^{n} dif(X_P,SC_i)}{|C_i|}\} \tag{7}$$

其中，X_p 表示第 i 簇中的第 p 条数据，$|C_i|$表示第 i 簇中的数据总数 n，该公式是指在一个簇内所有样本与簇中心的平均距离的两倍，即质心距离。

DB 指数计算公式如式(8)所示：

$$DB=\frac{1}{k}\sum_{i=1}^{k}\max_{j,i\neq j}\{\frac{\Delta C_i+\Delta C_j}{d(C_i,C_j)}\} \tag{8}$$

其中，k 为数据集 D 所聚成的簇数。

3.2.5 离群因子的计算

设 $C=\{C_1,C_2,\cdots,C_k\}$是数据集 D 的聚类结果。簇 C_i 的离群因子(Outlier Factor)$OF(C_i)$定义为 C_i 与所有簇间距离的加权平均值：

$$OF(C_i)=\sum_{j=1,j\neq i}^{k}\frac{|C_j|}{|D|}\cdot d(C_i,C_j) \tag{9}$$

其中，$\frac{|C_j|}{|D|}\cdot d(C_i,C_j)$可以看成 C_i 偏离 C_j 的程度，这种偏离程度既体现了簇间相对距离，同时也考虑了参考簇的大小。$OF(C_i)$度量了簇 C_i 偏离整个数据集的程度，其值越大，说明 C_i 偏离整体越远。

4 基于自适应聚类的离群点挖掘算法

4.1 算法思想

离群点是在数据集中偏离大部分数据的数据，而簇的离群因子度量了一个簇(即是簇中所有的对象)偏离整个数据集的程度，自然地将离群因子大的簇看成异常簇，也就是将其中的所有对象看成异常行为。由此提出基于自适应聚类的离群点挖掘方法 ACBOD(Adaptive Clustering-Based Outlier Detection)，而现有的基于聚类的离群点挖掘算法都需要手动地设置阈值，聚类算法中的参数 r 将影响聚类的结果和算法的时间效率。r 越小得到的簇的个数越多，算法时间开销越大；而当 r 达到一定值时只能得到极少的簇甚至一个簇。

ACBOD 方法由 4 个阶段构成，第一阶段：数据采集阶段，是通过 OPC 协议获取工业控制系统 OPC Server 上的数据；第二阶段：聚类阶段，比较每一个半径阈值 r 所对应的聚类 DB 指数，找出最优的聚类半径 r；第三阶段：簇的标识阶段，对该最优半径阈值所聚的簇进行离群因子计算，并标记异常簇；第四阶段：离群点检测阶段，对数据集中的每个对象进行聚类，并计算每个簇的离群因子，输出所有标识为异常的簇中对象。

4.2 算法描述

算法流程如图 1 所示，该流程主要包括 4 个部分，分别是数据采集部分、聚类部分、簇的标识部分以及离群点检测部分。

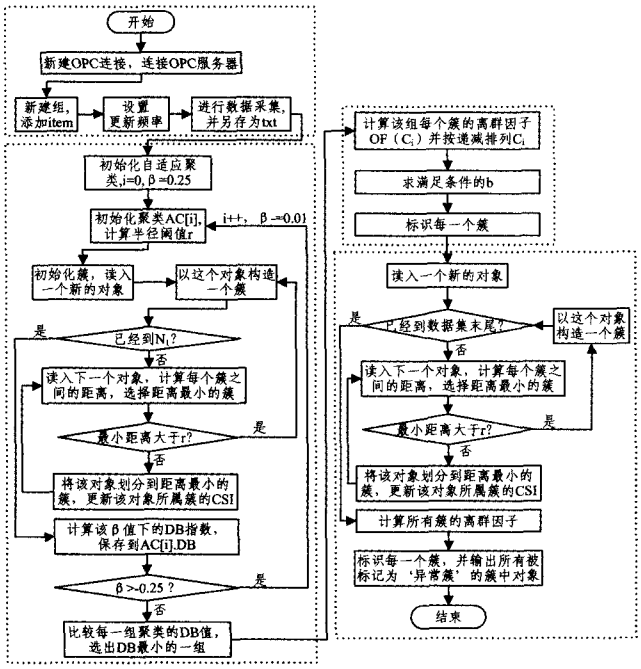

图 1 ACBOD 算法流程图

4.2.1 数据采集阶段

工业控制系统的数据采集与传统的入侵检测数据采集有所不同，传统的入侵检测大部分是采集网络数据，而工业控制系统因其协议的独特性，单单采集网络传输数据很难达到检测异常的效果。而针对工业控制系统的攻击最终都是要通过改变下位机的控制状态来实现对整个工业控制系统的破坏，因此，本文主要是采集工业控制系统下位机的变量数据，包括开关量、模拟量的数据。

具体的采集过程为：输入参数为 OPC Server IP 地址，以及 OPC Server 名称。首先根据输入的 OPC Server IP 以及 OPC Server 名称连接 OPC Server，新建 OPC 组 group，选择需要采集的变量，此处默认采集全部模拟量以及开关量。设置采集周期，并将采集到的数据保存为 txt 形式。算法步骤如下：

Step 1　新建 OPC 连接，连接 OPC 服务器；

Step 2　新建组 group，并添加 item；

Step 3　设置更新速率；

Step 4　进行数据采集，并保存为 txt 格式。

4.2.2　数据聚类阶段

聚类分析能够发现强相关的对象组，而异常检测则需要发现不与其他对象强相关的对象。可见，聚类可应用于异常检测。有些聚类算法，如 DBSCAN、BIRCH、ROCK、Waver-Cluster 等具有一定的异常检测能力，但它们的主要目标是产生有意义的簇，而不是异常检测，异常检测只是其副产品。这些算法在处理过程中通常将离群点当作噪音而对其忽略或剪枝，从而产生不了很好的离群点挖掘结果。本文采用基于最小距离原则的聚类算法对采集到的数据进行聚类。

基于最小距离原则的聚类算法采用摘要信息 CSI 表示一个簇，将数据集分割为半径几乎相同的超球体(簇)。具体的过程如下：

Step 1　计算半径阈值 r，$\beta=0.25$，随机选 N_1 行数据作为训练集。

Step 2　初始化簇，簇集合为空，读入一个新的对象。

Step 3　以这个对象构造一个新的簇。

Step 4　若已到训练集末尾，则转 Step 7，否则读入新对象，利用给定的距离定义，计算它与每个已有簇间的距离，并选择最小的距离。

Step 5　若最小距离超过给定的半径阈值 r，转 Step 3。

Step 6　否则将该对象并入具有最小距离的簇中并更新该簇的各分类属性值的统计频度及数值属性的质心，转 Step 4。

Step 7　计算该半径阈值下的 DB 指数。若 $\beta>-2$，则 $\beta=\beta-0.01$，重新计算半径阈值 r，转 Step 2。

Step 8　比较每一组半径阈值 r 所对应的 DB 值，找出 DB 值最小的一组，并以此时的聚类为最终的聚类结果。

4.2.3　簇的标识阶段

根据第二阶段得到的最佳聚类结果 $C=\{C_1,C_2,\cdots,C_k\}$ 确定离群点。检测结果的准确性与 ε 密切相关。ε 实际上是离群数据所占比例的近似值，ε 越小，检测率越低，同时误报率也越低，根据统计经验，一个数据集中的异常数据通常小于 5%，因此本文的 ε 采用 0.05，具体过程如下：

Step 1　计算每个簇 $C_i(1\leqslant i\leqslant k)$ 的离群因子 $OF(C_i)$。

Step 2　将每个簇按 $OF(C_i)$ 递减的顺序重新排列 C_i $(1\leqslant i\leqslant k)$。

Step 3　求满足 $\dfrac{\sum_{i=1}^{b}|C_i|}{|D|}\geqslant\varepsilon(0<\varepsilon<1)$ 的最小 b。

Step 4　将簇 $C_1,C_2,\cdots,C_b$ 标识为'outlier'类(即其中每个对象均看成异常)，而将 $C_{b+1},C_{b+2},\cdots,C_k$ 标识为'*normal*'类(即其中每个对象均看成正常)。

4.2.4　对象的离群检测阶段

根据第二阶段所得到的聚类结果、阈值半径 r 以及第三阶段的标识，对整个数据集进行离群检测。

Step 1　初始化聚类，簇集合为第二阶段得到的聚类结果，最优阈值半径为 r。

Step 2　读入一个新对象，若到数据集末尾，则转 Step 5。否则根据最近邻方法计算该对象的离群程度。

Step 3　若最近邻距离小于半径阈值 r，则将该对象划分到该簇中，并更新簇的摘要信息。若该对象与每个簇的距离均大于半径阈值 r，则转 Step 4。

Step 4　以该对象构建一个新的簇。

Step 5　计算每个簇的离群因子，输出标识为'Outlier'的簇中对象。

Step 6　结束。

5　实验研究

本文通过两个实验来分析算法的有效性与算法性能。所有的实验采用平台均为 Intel Core 2 Duo 2.00GHz，内存为 3G 的笔记本电脑，操作系统为 Windows 7。

5.1　算法的有效性

该部分实验对本算法进行了相关删减，剪掉了数据采集阶段，使用网络入侵检测数据集 KDD-CUP-99，对本文所提出的自适应方法以及 DB 指数的有效性进行了相关实验。

由于整个 KDD-CUP-99 数据集太大，且正常记录仅占 20%左右，不符合离群点挖掘方法的前提假设，为检测算法 ACBOD 的有效性，从该数据集中随机抽取一个子集 P，该子集包含 40000 条数据，其中包括 38500 条正常记录和 1500 条攻击数据(占 3.75%)。

对于该数据集，首先随机选取其中 500 条记录进行模型的初步建立，并通过自适应算法，得出最优的阈值半径为 0.315，聚类个数为 25，接下来对整个数据集进行聚类计算，结果与文献[15]进行对比，如表 1 所列。

表 1　ACBOD 与 CBOD 算法比较

聚类算法	阈值范围	聚类个数	检测率	误报率
ACBOD	0.315	23	98.69%	5.32%
CBOD	0.10	359	98.39%	6.31%
	0.17	158	98.33%	6.32%
	0.21	88	98.39%	5.36%
	0.23	60	98.39%	5.30%
	0.27	31	98.39%	5.32%
	0.30	25	98.58%	5.32%
	0.33	13	98.39%	5.54%

根据实验结果可以发现，文献[15]中得到的最优阈值半径为 0.30，对比发现本文所提出的自适应半径阈值的计算效果与文献[15]中数次实验得出的最优结果是一致的，甚至精度更高，因此本文所提出的自适应半径阈值的计算方法是有效的。

5.2　算法的性能分析

该部分实验采用的原型工业控制系统是 SIMATIC WinCC 提供的测试系统 DemoProjectV7 中的一个子系统 Waste gas System。实验数据采集部分由 MATLAB 平台的 OPC Client 工具箱进行采集，主要采集的变量为：Cooling_Water、waste_gas_cool、Evaporation cool、Coarse_dust_tr、El_precipitator、Fine_dust_tr、CO、O2、CO2、H2、SP、PV、PSAL，共 13 个属性，其中前 6 个为开关量，后 7 个为模拟量。采集该系统连续运行 10 个小时的数据，采集间隔为 2s/次，将采集之后的数据保存为 opcdatalog.osf，一共 12 万条数据。数据预处理阶段以及之后的算法实现阶段均采用 Visual C++ 6.0 实现。部分采集的数据如表 2 所列。

(下转第 203 页)

UML-RT diagrams[J]. Electronic Notes in Theoretical Computer Science,2008,195:57-74
[8] Li Tong. An Approach to Modelling Software Evolution Processes[M]. Berlin:Springer-Verlag,2008
[9] 徐洪珍,曾国荪,陈波. 软件体系结构动态演化的条件超图文法及分析[J]. 软件学报,2011,22(6):1210-1223
[10] 李长云. 基于体系结构的软件动态演化研究[D]. 杭州:浙江大学,2005
[11] 谢仲文,李彤,代飞,等. 面向软件动态演化的需求建模及其模型规范化[J]. 计算机科学与探索,2012,6(6):557-576
[12] 梅宏,申峻嵘. 软件体系结构研究进展[J]. 软件学报,2006,17(6):1257-1275
[13] 谭云杰. 大象:Thinking in UML[M]. 北京:中国水利水电出版社,2009

(上接第 181 页)

表 2 经过 MATLAB 数据采集之后的部分数据

ID \ 属性	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0	0	0	1	0	1	6.22	5.27	4	10.13	1368	34.67	89.41
2	0	0	0	1	1	1	7.89	5.64	5.15	9.33	1260	44.67	82.35
3	0	0	0	1	1	1	7.56	4.55	2.46	3.6	486	21.33	31.76
4	1	0	1	1	1	1	14	5.64	8.54	1.4	189	74	12.35
5	1	0	1	1	1	1	2.67	2.36	6.69	1.27	171	58	11.18
…	…	…	…	…	…	…	…	…	…	…	…	…	…
119996	1	1	1	1	1	1	7	10	6.15	5.33	720	53.33	47.06
119997	1	1	1	1	1	1	13.56	8.18	5.30	5.53	747	46	48.82
119998	1	1	1	1	1	1	5.56	5.36	1.54	5.13	693	13.33	45.29
119999	1	1	1	1	1	1	7.22	6.09	1.92	1.4	189	16.67	12.35
120000	1	1	1	1	1	1	1.67	7.27	1.77	2.67	360	15.33	23.53

通过之前算法中的数值属性的数据预处理之后的数据如表 3 所列。

表 3 数值属性数据预处理之后的部分数据

ID \ 属性	7	8	9	10	11	12	13
1	−0.6402	−0.6432	−0.6470	−0.6284	3.5000	−0.5538	−0.3874
2	−0.6427	−0.6502	−0.6518	−0.6379	3.5000	−0.5210	−0.3963
3	−0.6205	−0.6465	−0.6644	−0.6546	3.5000	−0.5019	−0.4120
4	−0.5883	−0.7547	−0.6970	−0.8390	2.8942	0.6057	−0.6201
5	−0.7484	−0.7551	−0.6584	−0.7796	3.0122	0.4877	−0.5582
…	…	…	…	…	…	…	…
119996	−0.6680	−0.6504	−0.6729	−0.6777	3.5000	−0.3971	−0.4338
119997	−0.6265	−0.6568	−0.6729	−0.6717	3.5000	−0.4440	−0.4281
119998	−0.6262	−0.6274	−0.6503	−0.6288	3.4999	−0.5796	−0.3877
119999	−0.5920	−0.6175	−0.7113	−0.7231	3.5000	−0.3794	−0.4765
120000	−0.6651	−0.5999	−0.6640	−0.6535	3.4999	−0.5063	−0.4110

为了测试算法的性能,本文将其与组态软件 WinCC 自带的 Alarm 模块的报警信息进行比较。实验结果显示当 $\beta=-0.03$时,阈值半径 $r=2.60391$,此时聚类个数为 5 个,DB 指数为 14.6865 的聚类结果最优。将 WinCC 自带的 Alarm 变量报警信息对该算法得到的结果进行检验,结果显示已知的异常检验率为 98.57%,未知的离群数据为 19 条。对未知的异常数据进行分析发现这些离群数据确实属于异常行为,例如:(1、0、1、1、1、1、14、5.64、8.54、1.4、189、74、12.35),该数据被挖掘出属于未知异常,经分析,该数据的 CO、O2、CO2、H2 比例明显不协调,但每一个属性值又都没有达到 WinCC 变量报警的阈值,因此 WinCC 变量报警没有该条记录。

实验证明,该方法在较好的空间复杂度与时间复杂度下,能有效地发现在高维属性的工业控制系统中的异常属性。

结束语 本文提出了一种基于自适应聚类的离群点挖掘方法,并将其应用到工业控制系统异常检测中,通过对聚类时半径阈值的自适应选择策略的改进来提高算法的聚类效果,并通过两个实验分别验证了该方法的有效性与性能。实验结果表明,该方法能够很好地发现工业控制系统中已知的、未知的异常,通过发现这些异常操作可以极大地提高对工业控制系统安全性的掌控。

由于工业控制系统的数据具有时间序列的特性,今后的研究工作主要是完善该算法,对工业控制系统数据所特有的时间属性进行研究,在时序性上对算法进行优化。

参 考 文 献

[1] IEC 62443-2-1 ED. 1.0 EN:2010, "Industrial communication networks-Network and system security-Part 2-1:Establishing an industrial automation and control system security program" [R]. International Electrotechnical Commission,2010
[2] 张帅. 工业控制系统安全现状与风险分析[J]. 计算机安全,2012(01):15-19
[3] Han Jia-wei, Micheline K. Data Mining: Concepts and Techniques (2nd Edition)[M]. San Francisco: Morgan Kauffmann Publishers,2006
[4] Haw kins D. Identification of Outliers[M]. London: Chapman and Hall,1980
[5] 唐成龙,王石刚. 基于数据间内在关联性的自适应模糊聚类模型[J]. 自动化学报,2010,36(11):1544-1556
[6] 薛安荣,姚林,鞠时光. 离群点挖掘方法综述[J]. 计算机科学,2008,35(11):13-17
[7] 徐翔,刘建伟,罗雄麟. 离群点挖掘研究[J]. 计算机应用研究,2009,26(1):34-39
[8] 王欣. 基于聚类和距离的大数据集离群点检测算法[J]. 制造业自动化,2010,33(4):101-104
[9] 王茜,唐锐. 基于频繁模式的离群点挖掘在入侵检测中的应用[J]. 计算机应用研究,2013,30(4):1208-1211
[10] 唐成龙,王石刚,徐威. 基于数据加权策略的模糊聚类改进算法[J]. 电子与信息学报,2010,32(6):1277-1283
[11] 杨鹏. 离群检测及其优化算法研究[D]. 重庆:重庆大学,2010
[12] 王茜,杨正宽. 一种基于加权 KNN 的大数据集下离群检测算法[J]. 计算机科学,2011,38(10):177-180
[13] Davies, David L, Bouldin, et al. A Cluster Separation Measure [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,1979,PAMI-1 (2):224-227
[14] 杨斌. 基于聚类的异常检测技术的研究[D]. 长沙:中南大学,2008
[15] 蒋盛益. 基于聚类的入侵检测算法研究[M]. 北京:科学出版社,2008:152-159