基于知识粒度的时间序列异常检测研究

杨志勇,朱跃龙,万定生
(河海大学 计算机与信息学院,江苏 南京 210098)

摘 要:时间序列的异常检测多以相似性分析方法来处理,时间代价高昂。为减少异常检测的时间,文中围绕知识粒度方法进行研究与探讨。知识粒度在数据异常检测中应用广泛,但在时间序列的异常检测上应用较少。文中针对时间序列上下文相关异常(点)检测,提出利用知识粒度异常检测方法对于输入属性越多检测粒度越细的特性,来查找时间序列中的异常数据。实验证明,基于知识粒度的方法无需先验信息,在整个处理过程中无需事先分析历史数据,而是通过属性间的组合粒度来划分异常数据与正常数据,提高了异常检测的效率。知识粒度方法在不确定信息处理研究中的表现十分突出,文中将知识粒度在时间序列异常检测中进行应用尝试,为时间序列异常检测提供了一种新的思路。

关键词:时间序列;知识粒度;粗糙集;异常检测
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2016)07-0051-04
doi:10.3969/j.issn.1673-629X.2016.07.011

Research on Time Series Anomaly Detection Based on Knowledge Granularity

YANG Zhi-yong, ZHU Yue-long, WAN Ding-sheng
(College of Computer and Information, Hohai University, Nanjing 210098, China)

Abstract: Most of the time series' anomaly detections are processed with the similarity analysis, and their time complexity is rather high. In order to reduce the time of anomaly detection, it studies and discusses the method of knowledge granularity in this paper. Knowledge granularity is widely applied in the anomaly detection of data, but rarely used in anomaly detection on time series. In view of context dependent anomaly (point) detection in time series, the knowledge-granularity-based anomaly detection is proposed to search the anomalous data in time series, in which the more the attributes are, the finer the detection granularity is. Experiments show that the method based on knowledge granularity does not require a priori information, partition of the abnormal data and normal data through the combination of the attributes without analysis of historical data previously, and the efficiency of anomaly detection has been improved. The knowledge granularity method is very prominent in the research of uncertain information processing. It tries to apply the knowledge granularity in the anomaly detection of time series in this paper, thus to provide a new approach for anomaly detection of time series.

Key words: time series; knowledge granularity; rough set; anomaly detection

0 引 言

异常检测也称为异常挖掘或离群检测[1],是从大量数据中提取隐含在其中的人们事先不知道的但又潜在有用的信息和知识的过程。异常检测中包含时间序列和非时间序列的检测。非时间序列的异常检测主要用于发现异常点集,数据之间无先后顺序,众多学者已经对非时间序列数据进行了深入研究,常用的检测方法有基于距离的方法[2-3]、基于统计的方法[4-5]、基于偏离的方法[6]、基于聚类的方法[7]和基于密度的方法[8]。而时间序列各个点数据有先后顺序,有逻辑关系与递推关系,时间序列的异常检测难度极大,不能单纯通过距离、密度等方法进行处理[9]。时间序列异常的挖掘,需要同时考虑几年甚至几十年跨度的数据[10]。

目前,时间序列异常检测的研究分为三类:一是时间序列上下文相关异常(点)检测;二是时间序列模式(子序列)异常检测;三是时间序列异常周期检测。

文中方法针对时间序列上下文相关异常(点)检测,主要是检测出时间序列中和上下文信息明显偏离

收稿日期:2015-09-27 **修回日期**:2016-01-06 **网络出版时间**:2016-06-22
基金项目:国家科技支撑计划课题(2015BAB07B01);水利部公益性行业科研专项(201501022)
作者简介:杨志勇(1990-),男,硕士研究生,研究方向为数据挖掘;朱跃龙,教授,博士生导师,研究方向为水信息学、智能信息处理;万定生,教授,CCF 会员,研究方向为信息处理与信息系统。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20160621.1701.012.html

的个体(点异常)。

粒计算作为人工智能领域新兴起的一个研究方向,是一种新的不确定信息处理方法[11]。粒计算通过粗糙集模型将论域的划分视作知识,知识粒度则是其重要的度量工具。基于知识粒度的常规异常检测方法是通过单个数据的存在与否对知识粒度的影响来度量数据的异常,如果直接使用此方法处理时间序列异常,则会因为时间序列属性的单一性而变成了类似前面提到的基于距离的异常、基于密度的异常等传统方法,从而无法正确检测出异常。但是,根据知识粒度异常检测属性越多知识粒度越小[12]的特点以及时间序列往往会伴随多条与之相关的时间序列的特点,在多条相关联时间序列的伴随下,可以使用基于知识粒度的异常检测方法。

文中利用知识粒度异常检测方法对于输入属性越多检测粒度越细的特性,查找时间序列中的点异常[13]。根据知识粒度检测方法对输入数据不要求数据以外的任何先验信息的优点,提高时间序列异常检测算法的效率,时间复杂度为 $O(m*n)$[12]。

1 知识粒度

粒计算是一种新型的不确定性信息处理方法,其基本思想是利用不同粒度上的信息进行问题求解。粒计算的一个重要模型为粗糙集,粗糙集理论的要点是将分类与知识联系在一起,并用等价关系形式化表示分类,每个等价类称为一个知识粒[14]。知识粒有粗有细,为了度量等价类的粗细程度,因而有了知识粒度。知识粒度可以描述知识的区分能力,知识粒度越小,它的区分能力越强[15]。应用知识粒度区分知识的能力,可以检测出数据中的异常。下面介绍知识粒度在异常检测应用中的相关定义与定理[16]:

(1)设 IS = (U,A,V,f) 是一个信息系统,其中 U 是非空有限对象集,称为论域,A 是属性集,$V=\bigcup_{a\in A} V_a$,V_a 称为属性 a 的值域,$f_a:U\to V_a$ 是信息函数。每一个属性子集 $B\subseteq A$ 确定了一个二元不可区分关系,IND(B) = $\{(x,y)\in U\times U \mid \forall a\in B, f_a(x)=f_a(y)\}$,IND($B$) 是 U 上的等价关系。所有等价类的集合用 $U/\mathrm{IND}(B)$ 来表示,称之为 U 上的知识,简写为 U/B。U/B 中的每个等价类称为一个知识粒。

(2)设 IS = (U,A,V,f) 是一个信息系统,$P,Q\subseteq A$,若对任意 $u,v\in U$,有 $u(U/P)v\Leftrightarrow u(U/Q)v$,则称 U/P 与 U/Q 相等,记为 $U/P = U/Q$。

(3)设 IS = (U,A,V,f) 是一个信息系统,$P,Q\subseteq A$,若对任意 $u,v\in U$,有 $u(U/P)v\Rightarrow u(U/Q)v$,则称 U/P 比 U/Q 细,记为 $U/P\leqslant U/Q$。

(4)设 IS = (U,A,V,f) 是一个信息系统,$P,Q\subseteq A$,若对任意 $u,v\in U$,有 $U/P\leqslant U/Q$ 且 $U/P\neq U/Q$,则称 U/P 比 U/Q 严格细,记为 $U/P < U/Q$。

(5)设 IS = (U,A,V,f) 是一个信息系统,$U/A = \{X_1,X_2,\cdots,X_m\}$,$A$ 的知识粒度记为 KG(A):

$$\mathrm{KG}(A)=\mathrm{KG}(\mathrm{IND}(A))=\frac{|\mathrm{IND}(A)|}{|U|^2}$$

式中,$|\mathrm{IND}(A)|$ 表示 $\mathrm{IND}(A)\in U*U$ 的基数,展开后如下式:

$$\mathrm{KG}(A)=\sum_{i=1}^{m}|X_i|^2/|U|^2 \tag{1}$$

定理 1:设 IS = (U,A,V,f) 是一个信息系统,$P,Q\subseteq A$,若 $P\Rightarrow A$,则 $\mathrm{KG}(P)\leqslant \mathrm{KG}(Q)$。

证明:由 $P\Rightarrow Q$ 可知,$\mathrm{IND}(P)\subseteq \mathrm{IND}(Q)$,则 $|\mathrm{IND}(P)\leqslant |\mathrm{IND}(Q)||$。有:

$$\mathrm{KG}(P)=\frac{|\mathrm{IND}(P)|}{|U|^2}\leqslant\frac{|\mathrm{IND}(Q)|}{|U|^2}=\mathrm{KG}(Q)$$

因此,$\mathrm{KG}(P)\leqslant \mathrm{KG}(Q)$ 得证。

从定理 1 可知,信息系统中,属性越多,知识粒度就越小,区分能力就越强。

2 基于知识粒度的时间序列异常检测

2.1 检测方法定义

知识粒度是测量粗糙集理论中不确定信息的方法,在许多领域中都得到了应用,特别是机器学习与数据挖掘等,在非时间序列的异常检测中国内外也有过研究,但对于时间序列的异常检测研究却极少关注。

文中通过时间序列数据的连续性、关联性等特性,将时间序列异常检测问题转化为非时间序列异常检测问题。主要思想为将时间序列中不该在某个时间点出现的数据异常通过调整等价函数转化为非时间序列的离群异常。下面给出加入调整时间序列等价函数定义以及异常结果判定定义:

定义 1:设 IS = (U,A,V,F) 是一个信息系统,$A = \{a_1,a_2,\cdots,a_m\}$,其中,$a_1,a_2,\cdots,a_m$ 分别为多条随时间变化且相互有关联的时间序列属性。

定义 2:设 IS = (U,A,V,F) 是一个信息系统,$A = \{a_1,a_2,\cdots,a_m\}$,$F=\{f_1,f_2,\cdots,f_m\}$,在一个等价关系 IND($B$)中,$x,y\in \mathrm{IND}(B)$,有 $\mathrm{fa}_1(x)=\mathrm{fa}_1(y)$,$\mathrm{fa}_2(x)=\mathrm{fa}_2(y)$,$\cdots$,$\mathrm{fa}_m(x)=\mathrm{fa}_m(y)$。

定义 3:设 IS = (U,A,V,F) 是一个信息系统,$U/A = \{X_1,X_2,\cdots,X_m\}$,若 $\forall x\in U$,$\{U-\{x\}\}/A=\{X_1,X_2,\cdots,X_m\}$,$\mathrm{KG}_x(A)=\sum_{i=0}^{m}\frac{|X_i|^2}{|U|^2}$ 表示把对象 x 从 U 中移除后 A 的知识粒度,则 A 中对象 x 的相对知识粒度 $\mathrm{RG}_A(x)=\mathrm{KG}_x(A)/\mathrm{KG}(A)$。

定义 4:设 IS = (U,A,V,F) 是一个信息系统,$A =$

$\{a_1,a_2,\cdots,a_m\}$,按知识粒度从大到小排列,排列结果为单属性递减序列 $S=\{a_1',a_2',\cdots,a_m'\}$。

定义 5:设 IS = (U,A,V,F) 是一个信息系统,$A=\{a_1,a_2,\cdots,a_m\}$,$S=\{a_1',a_2',\cdots,a_m'\}$,序列 A 中逐一减去序列 S 中属性的所有子集 AS 称为属性子集递减序列。AS= $\{A_1',A_2',\cdots,A_m'\}$,$A_1'=A$,$A_m'=\{a_m'\}$,且 $A_{i+1}'=A_i'-\{a_i'\}$。

定义 6:设 IS = (U,A,V,F) 是一个信息系统,$\forall x\in U$,$W_A(x)=1-|[x]_A|/|U|$ 为 x 在 A 中的权重。

定义 7:设 IS = (U,A,V,F) 是一个信息系统,单属性递减序列 $S=\{a_1',a_2',\cdots,a_m'\}$,属性子集递减序列 AS= $\{A_1',A_2',\cdots,A_m'\}$,对象 x 的异常度为:

$$\mathrm{ROF}(x)=\Big(\sum_{i=1}^{m}\mathrm{RG}_{|a_i|}(x)*w_{\{a_i\}}(x)+\sum_{i=1}^{m}\mathrm{RG}_{|A_i|}(x)*w_{\{A_i\}}(x)\Big)/|U|$$

定义 8:设 IS = (U,A,V,F) 是一个信息系统,v 为设定的一个阈值,对 $\forall x\in U$,若 KOF(x)> v,则 x 为一个异常对象。

2.2　算法描述

根据上述检测方法中的相关定义,基于知识粒度的时间序列异常检测算法可以分为以下几个步骤:

(1)将多条相关的时间序列属性按照定义 1、2 构造出信息系统,设为 IS = (U,A,V,F)。

(2)根据定义 2 中的等价函数,循环 A 中属性,设定时间序列多条属性等价区间。同时划分出知识 $U/\mathrm{IND}(\{A_i\})$,并计算出知识粒度 $\mathrm{KG}(\{A_i\})$。

(3)根据定义 4,按步骤 2 中知识粒度排序,构造单属性递减序列 $S=\{a_1',a_2',\cdots,a_m'\}$。

(4)令 $x=0\sim m$ 循环 S 中属性 a,$j=0\sim n$ 嵌套循环 U 中对象 x,划分出去除 x_j 后的知识 $U/\mathrm{IND}_{x_j}(\{a_i\})$,并计算出知识粒度 $\mathrm{KG}_{x_j}(\{a_i\})$。

(5)计算 S 序列权重 $W_{\{a_i\}}(x_j)$。

(6)根据定义 5 构造属性子集递减序列 AS= $\{A_1',A_2',\cdots,A_m'\}$。

(7)令 $x=0\sim m$ 循环 AS 中属性 A',$j=0\sim n$ 嵌套循环 U 中对象 x,划分出去除 x_j 后的知识 $U/\mathrm{IND}_{x_j}(\{A_i'\})$,并计算出知识粒度 $\mathrm{KG}_{x_j}(\{A_i'\})$。

(8)令 $x=0\sim m$ 循环 AS 中属性 A',划分出去除知识 $U/\mathrm{IND}(\{A_i'\})$,并计算出知识粒度 $\mathrm{KG}(\{A_i'\})$。

(9)计算 AS 序列权重 $W_{\{A_i\}}(x_j)$。

(10)根据定义 8 计算对象 x_j 的异常度 KOF(x_j),查找大于阈值 v 的异常对象。

(11)输出结果集。

算法将时间序列异常问题描述为非时间序列异常问题,重点是时间序列属性的搭配和等价函数定义。

3　实例分析

文中以水文时间序列片段样本数据为例,取同一水利设施同一时间段内水库水位 a、降雨量 b、河道水位 c 的时间序列数据,如表 1 所示。

表 1　时间序列信息系统

U	a	b	c
…	…	…	…
x_1	80	20	12
x_2	109	20	13
x_3	83	22	13
x_4	88	40	15
x_5	95	43	15
x_6	104	42	15
x_7	106	40	15
…	…	…	…

其中,设定检测阈值 $v=0.5$,等价函数分别为:

$f_a(x)=\lfloor x/15\rfloor$,$f_b(x)=\lfloor x/5\rfloor$,$f_c(x)=\lfloor x/2\rfloor$

接下来根据算法步骤,进行单属性知识的划分:

$U/\mathrm{IND}(a)=\{\{x_1,x_3,x_4\},\{x_2,x_7\},\{x_5,x_6\}\}$

$U/\mathrm{IND}(b)=\{\{x_1,x_2,x_3\},\{x_4,x_5,x_6,x_7\}\}$

$U/\mathrm{IND}(c)=\{\{x_1,x_2,x_3\},\{x_4,x_5,x_6,x_7\}\}$

知识粒度为:

$\mathrm{KG}(\{a\})=(3\times3+2\times2+2\times2)/(7\times7)=0.347$

$\mathrm{KG}(\{b\})=(3\times3+4\times4)/(7\times7)=0.510$

$\mathrm{KG}(\{c\})=(3\times3+4\times4)/(7\times7)=0.510$

得到 $S=\{c,b,a\}$。

再计算划分去除 x_j 后的知识粒度:

$\mathrm{KG}_{x_1}(c)=\mathrm{KG}_{x_2}(c)=\mathrm{KG}_{x_3}(c)=0.408$

$\mathrm{KG}_{x_4}(c)=\mathrm{KG}_{x_5}(c)=\mathrm{KG}_{x_6}(c)=\mathrm{KG}_{x_7}(c)=0.367$

$\mathrm{KG}_{x_1}(b)=\mathrm{KG}_{x_2}(b)=\mathrm{KG}_{x_3}(b)=0.408$

$\mathrm{KG}_{x_4}(b)=\mathrm{KG}_{x_5}(b)=\mathrm{KG}_{x_6}(b)=\mathrm{KG}_{x_7}(b)=0.367$

$\mathrm{KG}_{x_1}(a)=\mathrm{KG}_{x_3}(a)=\mathrm{KG}_{x_4}(a)=0.286$

$\mathrm{KG}_{x_2}(a)=\mathrm{KG}_{x_7}(a)=0.286$

$\mathrm{KG}_{x_5}(a)=\mathrm{KG}_{x_6}(a)=0.286$

并计算权重 $W_{x_j}(\{A_i\})$:

$W_{x_1}(\{c\})=W_{x_2}(\{c\})=W_{x_3}(\{c\})=0.571$

$W_{x_4}(\{c\})=W_{x_5}(\{c\})=W_{x_6}(\{c\})=W_{x_7}(\{c\})=0.429$

$W_{x_1}(\{b\})=W_{x_2}(\{b\})=W_{x_3}(\{b\})=0.571$

$W_{x_4}(\{b\})=W_{x_5}(\{b\})=W_{x_6}(\{b\})=W_{x_7}(\{b\})=0.429$

$W_{x_1}(\{a\})=W_{x_3}(\{a\})=W_{x_4}(\{a\})=0.571$

$W_{x_2}(\{a\})=W_{x_7}(\{a\})=W_{x_5}(\{a\})=W_{x_6}(\{a\})=0.714$

进而根据步骤 6 构造 AS 序列,如下:

AS = $\{\{a,b,c\},\{a,b\},\cdots,\{a\}\}$

使用上一步同样方法划分并计算知识粒度,得:

$U/\mathrm{IND}(A_1) = \{\{x_1,x_3\},\{x_2\},\{x_4\},\{x_5,x_6\},\{x_7\}\}$

$U/\mathrm{IND}(A_2) = \{\{x_1,x_3\},\{x_2\},\{x_4\},\{x_5,x_6\},\{x_7\}\}$

$U/\mathrm{IND}(A_3) = \{\{x_1,x_3,x_4\},\{x_2,x_7\},\{x_5,x_6\}\}$

$\mathrm{KG}(\{A_1^{'}\}) = 0.224$

$\mathrm{KG}(\{A_2^{'}\}) = 0.510$

$\mathrm{KG}(\{A_3^{'}\}) = 0.510$

然后计算划分去除 x_j 后的知识粒度：

$\mathrm{KG}_{x_1}(A_1) = \mathrm{KG}_{x_3}(A_1) = 0.163$

$\mathrm{KG}_{x_2}(A_1) = \mathrm{KG}_{x_4}(A_1) = \mathrm{KG}_{x_7}(A_1) = 0.204$

$\mathrm{KG}_{x_5}(A_1) = \mathrm{KG}_{x_6}(A_1) = 0.163$

$\mathrm{KG}_{x_1}(A_2) = \mathrm{KG}_{x_3}(A_2) = 0.163$

$\mathrm{KG}_{x_2}(A_2) = \mathrm{KG}_{x_4}(A_2) = \mathrm{KG}_{x_7}(A_2) = 0.204$

$\mathrm{KG}_{x_5}(A_2) = \mathrm{KG}_{x_6}(A_2) = 0.163$

$\mathrm{KG}_{x_1}(A_3) = \mathrm{KG}_{x_3}(A_3) = \mathrm{KG}_{x_4}(A_3) = 0.245$

$\mathrm{KG}_{x_2}(A_3) = \mathrm{KG}_{x_7}(A_3) = 0.286$

$\mathrm{KG}_{x_5}(A_3) = \mathrm{KG}_{x_6}(A_3) = 0.286$

根据以上知识粒度计算去除 x_j 的权重 $W_{x_j}(\{A_i^{'}\})$：

$W_{x_1}(\{A_1^{'}\}) = W_{x_3}(\{A_1^{'}\}) = 0.714$

$W_{x_2}(\{A_1^{'}\}) = W_{x_4}(\{A_1^{'}\}) = W_{x_7}(\{A_1^{'}\}) = 0.857$

$W_{x_5}(\{A_1^{'}\}) = W_{x_6}(\{A_1^{'}\}) = 0.714$

$W_{x_1}(\{A_2^{'}\}) = W_{x_3}(\{A_2^{'}\}) = 0.714$

$W_{x_2}(\{A_2^{'}\}) = W_{x_4}(\{A_2^{'}\}) = W_{x_7}(\{A_2^{'}\}) = 0.857$

$W_{x_5}(\{A_2^{'}\}) = W_{x_6}(\{A_2^{'}\}) = 0.714$

$W_{x_1}(\{A_3^{'}\}) = W_{x_3}(\{A_3^{'}\}) = W_{x_4}(\{A_3^{'}\}) = 0.571$

$W_{x_2}(\{A_3^{'}\}) = W_{x_7}(\{A_3^{'}\}) = W_{x_5}(\{A_3^{'}\}) = W_{x_6}(\{A_3^{'}\}) = 0.714$

最后根据以上结果计算 KOF(x_j)，如表 2 所示。

表 2 KOF 计算结果

对象	KOF	是否异常
x_1	0.394	否
x_2	0.521	是
x_3	0.394	否
x_4	0.426	否
x_5	0.405	否
x_6	0.405	否
x_7	0.479	否

通过表 2 可以看出，对象 x_2 被发现为异常，其水位 109 在时间序列中是不应该发生的，在整个处理过程中并没有分析其周围的数据，而是通过属性间的组合粒度来划分异常数据与正常数据。

4 结束语

文中研究内容为时间序列的异常检测提供了一种新的思路，将知识粒度异常检测方法应用在时间序列的点异常检测上。然而时间序列的异常检测还包括子序列异常与周期异常，在下一步的研究中，应更加充分利用知识粒度在确定性数据处理上的优势，解决更多时间序列异常检测的问题。

参考文献：

[1] Chandola V, Banerjee A, Kumar V. Anomaly detection: a survey[J]. ACM Computing Surveys, 2009, 41(3): 1-15.

[2] Ramaswamy S, Rastogi R, Shim K. Efficient algorithms for mining outliers from large data sets[J]. ACM Sigmod Record, 2000, 29(2): 427-438.

[3] Hao Y, Wang B, Gang X, et al. Distance-based outlier detection on uncertain data[C]//Proc of IEEE international conference on computer and information technology. [s. l.]: IEEE, 2009: 293-298.

[4] Rousseeuw P J, Leroy A M. Robust regression and outlier detection[M]. [s. l.]: John Wiley & Sons, 2005.

[5] Johnson T, Kwok I. Fast computation of 2-dimensional depth contours[C]//Proc of the 4th international conference on knowledge discovery and data mining. New York: ACM Press, 1998: 224-228.

[6] Jagadish H V, Koudas N, Muthukrishnan S. Mining deviants in a time series database[C]//Proceedings of international conference on very large data bases. Edinburgh, Scotland, UK: [s. n.], 1999: 102-113.

[7] Budalakoti S, Srivastava A, Akella R, et al. Anomaly detection in large sets of high-dimensional symbol sequences[R]. Moffett Field: NASA Ames Research Center, 2006.

[8] Breunig M, Kriegel H, Ng R T, et al. LOF: identifying density-based local outliers[C]//Proceedings of the ACM SIGMOD international conference on management of Data. [s. l.]: ACM, 2000: 93-104.

[9] 陈运文，吴 飞，吴庐山，等. 基于异常检测的时间序列研究[J]. 计算机技术与发展，2015，25(4)：166-170.

[10] 林 森. 时间序列异常检测的研究与应用[D]. 南京：河海大学，2008.

[11] 苗夺谦，王国胤，刘 清，等. 粒计算：过去、现在与展望[M]. 北京：科学出版社，2007.

[12] 陈玉明，吴克寿，孙金华. 基于知识粒度的异常数据挖掘算法[J]. 计算机工程与应用，2012，48(4)：118-120.

[13] 肖 辉. 时间序列的相似性查询与异常检测[D]. 上海：复旦大学，2005.

[14] 王国胤. Rough 集理论与知识获取[M]. 西安：西安交通大学出版社，2001.

[15] 李道国，苗夺谦，张红云. 粒度计算的理论、模型与方法[J]. 复旦学报：自然科学版，2004，43(5)：837-841.

[16] 苗夺谦，范世栋. 知识的粒度计算及其应用[J]. 系统工程理论与实践，2002，22(1)：48-56.