

基于云模型的异常挖掘算法

李 磊¹, 彭 勇²

(1 江南大学 物联网工程学院, 江苏 无锡 214122;

2 江南大学 轻工过程先进控制教育部重点实验室, 江苏 无锡 214122)

摘 要: 针对基于距离的异常挖掘算法对混合属性数据进行异常挖掘时检测率低的问题, 引入概率的思想提出基于云模型的差异度计算公式, 并设计了基于云模型的异常挖掘算法, 实验结果表明该算法检测率高, 误报率低。

关键词: 云模型; 异常挖掘; 入侵检测; 检测率; 差异度量

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2013)08-0082-04

Outlier Mining Algorithm Based on Cloud Model

LI Lei¹, PENG Yong²

(1 School of Internet of Things, Jiangnan University, Wuxi 214122, China;

2 Key Laboratory of Advanced Process Control for Light Industry, Ministry of Education, P. R. China, Jiangnan University, Wuxi 214122, China)

Abstract: According to the problem of low detection rate for mixed value data outlier mining based on the distance of the outlier mining. By introducing probability idea, put forward a difference degree calculation formula and outlier mining algorithm based on cloud model. The experimental results show that this algorithm has a high detect rate and the distorting rate is lower.

Key words: cloud model; outlier mining; intrusion detection; detect rate; measure of dissimilarity

1 引言

异常挖掘是数据挖掘的重要研究内容之一。所谓异常挖掘是在给定的数据集中发现偏离大部分数据的数据^[1]。现有的异常挖掘方法大体分为: 基于统计的方法、基于距离的方法、基于密度的方法和基于聚类的方法, 但上述方法不适用于对混合属性的数据进行数据挖掘。

云模型是文献[2]中提出的定性定量间转换的不确定性转换模型中, 该理论将定性定量转换中的模糊性和随机性集成到一起, 已广泛地应用于挖掘空间广义知识和关联规则, 表达发现的知识等领域。而混合属性数据中包含有大量的定性概念, 若能更准确地表达这些定性概念, 则计算出的对象间的距离更准确, 进而可得到准确的挖掘数据。

本文针对基于距离的异常挖掘算法对混合属性数据进行异常挖掘时检测率低的问题, 提出了基于云模型的差异度计算公式, 并用此作为度量对象间的距离, 设计了基于云模型的异常挖掘算法, 实验结果表明算法在处理混合属性数据时不仅能提高挖掘的准确度, 并能减少挖掘时间。

2 基本概念

2.1 云模型

定义 1 设 U 是一个用精确数值表示的定量论域, C 是 U 上的定性概念, 若定量值 $x \in U$, 且 x 是定性概念 C 的一次随机实现, x 对 C 的确定度 $\mu(x) \in [0, 1]$ 是有稳定倾向的随机数。

$$\mu: U \rightarrow [0, 1] \quad \forall x \in U \quad x \rightarrow \mu(x)$$

则 x 在论域 U 上的分布成称为云, 每一个 x 称为一

收稿日期: 2012-09-30; 修回日期: 2012-11-29

基金项目: 轻工过程先进控制教育部重点实验室开放课题资助(江南大学)项目(APCLI1001)

个云滴^[10].

定义 2 云的数字特征^[2]

期望 Ex : 云滴在论域空间分布的期望.

熵 En : 定性概念的不确定性度量, 由概念的随机性和模糊性决定.

超熵 He : 熵的不确定性度量, 即熵的熵.

定义 3 给定论域中的某个数据属性 X 的频率分布函数 $f(x)$, 根据 X 的属性值频率的实际分布情况, 自动生成若干粒度不同的云 $C(Ex_i, En_i, He_i)$ 的叠加, 每个云代表一个离散的、定性的概念, 这种从连续的数值区间到离散的概念的转换过程, 称为云变换. 其数学表达式为

$$f(x) \rightarrow \sum_{i=1}^n (a_i * C(Ex_i, En_i, He_i)),$$

式中, a_i 为幅度系数, n 为变换后生成离散概念的个数.

2.2 簇的表示模型

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, 假定分类属性位于数值属性之前, 用 D_i 表示第 i 个属性取值的集合.

定义 4 给定簇 C , C 的摘要信息 CSI 定义为

$$CSI = \{n, Summary\}$$

式中, n 为 C 的大小, $Summary$ 由分类属性中不同取值的频度信息和数值型属性的质心两部分组成, 即

$$Summary = \{ \langle Stat_i, Cen \rangle \mid Stat_i = \{ \langle a, Freq_{C|D_i}(a) \rangle \} \mid a \in D_i, Cen = (c_{m+1}, c_{m+2}, \dots, c_{m+m_N}) \}$$

式中, $Freq_{C|D_i}(a)$ 为分类属性含有 D_i 的次数, c_{m+1} 为数值型属性的质心.

3 基于云模型的异常挖掘算法

3.1 基于云模型的差异度度量方法

传统的基于距离的异常挖掘算法^[3-4]进行差异度度量时, 将含连续数值属性或顺序属性的对象间的差异程度描述为两个属性差的绝对值, 而分类属性的对象间的差异程度描述为 0 或 1, 据此计算出两个对象间的差异程度可能由于连续数值属性或顺序属性的对象间的差异程度数值相互抵消, 从而不能准确地描述差异程度, 导致异常挖掘效率较低. 本文引入云模型计算出各连续数值属性或顺序属性发生的概率, 提出了一种含概率的差异度度量方法, 该方法能够将含连续数值属性或顺序属性的对象间的距离定义在区间 $[0, 1]$ 之间, 能够更准确地表现出混合属性数据的差异程度, 从而为进一步的异常挖掘

提供更准确的数据, 提高异常挖掘的检测率.

为准确的计算各连续数值属性或顺序属性发生的概率, 本文结合定义 3 设计了云变换器, 模型如图 1 所示.

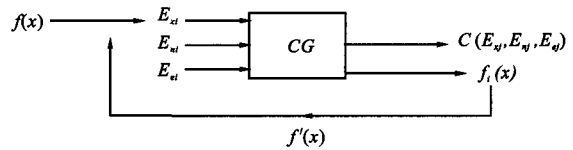


图 1 云变换器模型图

云变换器的工作过程如下:

Step 1: 输入分布函数 $f(x)$;

Step 2: 寻找 $f(x)$ 的波峰值所在位置, 将该处对应的数值定为云的期望 Ex_i , 求出以 Ex_i 为期望的云的熵 En_i 与超熵 He_i , 计算此云的分布函数 $f_i(x)$;

Step 3: 从 $f(x)$ 中减去已知云模型的分布函数 $f_i(x)$, 得到新的数据分布函数 $f'(x)$;

Step 4: 重复步骤 (2), (3), 得到多个云 $C(Ex_j, En_j, He_j) (j \leq n)$.

结合云变换器算法, 设计了基于云模型的概率算法, 算法如下所示.

算法 1 基于云模型的概率计算算法

假设有 n 个数值 $a_1, a_2, \dots, a_n$ (数值定义域为 $[s, t]$).

Step 1: 对 n 个数值中的每一个 $a_i (a_i \in [s, t])$, 计算各个可能值的记录个数 y , 得到 n 个数值的频率分布函数 $f(x)$;

Step 2: 分布函数 $f(x)$ 作为输入通过云变换器得到多个云 $C(Ex_j, En_j, He_j)$;

Step 3: 比较 a_i 与 Ex_j , 找到与 a_i 值最接近的 Ex_j^* , 定义数值 a_i 的云为 $C(Ex_j^*, En_j^*, He_j^*)$, a_i 的期望为 Ex_j^* , 即可计算出 a_i 概率 $p_{ai}(a_i)$ 同时得到云滴 $(a_i, p_{ai}(a_i))$.

定义 5 基于云模型差异度计算公式

给定数据集 D 的簇 C , C_1 和 C_2 , 对象 $p = [p_1, p_2, \dots, p_m]$ 与 $q = [q_1, q_2, \dots, q_m]$.

(1) 对象 p, q 在属性 i 上的差异程度 (或距离) $dif(p_i, q_i)$ 定义如下:

① 对于分类属性或二值属性

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

② 对于连续数值属性或顺序属性

$$dif(p_i, q_i) = \frac{|p_i * p_{\mu}(p_i) - q_i * p_{\mu}(q_i)|}{\max(Ex_{\mu}, Ex_{\mu})},$$

式中, $p_{\mu}(p_i), p_q(q_i)$ 为 p_i, q_i 的概率; Ex_{μ}, Ex_q 分别为 p_i, q_i 的期望, 该处概率和期望由算法 1 给出.

(2) 对象 p_i, q_i 间的距离 $d(p, q)$ 表示为

$$d(p, q) = \frac{d_C + d_D}{m} = \sum_{i=1}^m dif(p_i, q_i)$$

上式为对象 p, q 的各属性差异程度之和.

(3) 对象 p 与簇 C 的距离 $d(p, C)$ 定义为 p 与簇 C 的摘要之间的距离, 由两部分构成:

$$d(p, C) = \frac{d_C + d_N}{m},$$

其中, 分类属性部分的距离 d_C 定义为

$$d_C = (\sum_{i=1}^{m_C} dif(p_i, C|D_i))^{\frac{1}{\gamma}},$$

此处

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

数值属性部分的距离 d_N 定义为 p 与 C 的质心间的距离:

$$d_N = (\sum_{i=m_C+1}^{m_C+m_M} |p_i - c_i|^z)^{\frac{1}{z}}.$$

3.2 基于云模型的异常挖掘算法

云模型理论中, 论域 X 在不同的区间内对定性概念的贡献是不相同的, 位于区间 $[Ex - En, Ex + En]$ 的那些元素, 对定性概念的贡献占总贡献的 68.26%; 同样位于区间 $[Ex - 2En, Ex - En]$ 和 $[Ex + En, Ex + 2En]$ 的元素贡献占总贡献的 27.18%; 位于区间 $[Ex - 3En, Ex - 2En]$ 和 $[Ex + 2En, Ex + 3En]$ 的那些元素贡献占总贡献的 4.3%^[5]. 本文将该理论应用于异常挖掘, 设计了基于云模型的异常挖掘算法, 认为对定性概念贡献大的元素一定是正常数据, 而贡献小的元素则可能含有异常数据.

该算法分两个阶段, 第一阶段计算训练集各对象间的差异度, 用正态云表示正常对象集; 第二阶段根据云滴的贡献理论判断数据是否异常. 算法描述如下所示.

算法 2 基于云模型的异常挖掘算法

Step 1: 计算训练集各对象间的差异度, 用正态云表示正常对象集.

(1) 随机抽取对象集 D 中的若干个对象组成训练集 D_1 (训练集大小与对象集大小有关);

(2) 根据差异度计算公式(1)计算所有对象间的距离 $s_1, s_2, \dots, s_n$;

(3) $s_1, s_2, \dots, s_n$ 作为输入通过云变换器得到多个云 $C(Ex_j, En_j, He_j)$;

(4) 计算各距离 $s_1, s_2, \dots, s_n$ 的云滴 $C(s_i, u_i)$,

利用逆向云发生器生成正常对象集的正态云 $C(Ex, En, He)$.

Step 2: 计算对象 p 与训练集组成的簇的距离, 比较判断该对象是否为异常.

(1) 计算对象 p 与训练集组成的簇的距离 s' ;

(2) 计算出云 $C(Ex, En, He)$ 的 $[Ex - 3En, Ex + 3En]$ (称 $3En$ 规则) 或 $[Ex - 2En, Ex + 2En]$ (称 $2En$ 规则) 区间;

(3) 比较 s' 是否落在此区间内, 若不在, 则此对象是异常.

4 实验结果

为了评估本文提出的异常挖掘算法的有效性, 在几个真实数据集上进行实验. 并使用检测率和误报率来度量异常挖掘的准确性.

(1) KDDCUP99 数据集(KDDCUP99 dataset)

KDDCUP99 包含约 4 900 000 条记录, 由 7 个分类属性和 34 个数值型属性刻画, 本文从这个数据集中随机选取一个包含 3 455 条记录的子集, 其中攻击记录 86 条 (占 2.49%). 从检测率和误报率两个方面与蒋盛益等提出的 GLOF^[4] 进行比较.

表 1 中的实验结果表明, 本文算法在 $3En, 2En$ 规则下检测率比 GLOF 算法高, 误报率与 GLOF 算法相当.

表 1 在 KDDCUP99 数据集中检测率与误报率比较

比较项		检测率	误报率
本文算法	GLOF 算法	89.53%	0.87%
	$3En$ 规则	92.39%	0.89%
	$2En$ 规则	93.85%	0.83%

(2) 森林覆盖类型分布数据集

UCI Machine Learning Reporting 提供的森林覆盖类型分布数据集^[6] (Forest Cover Type) 包含了 581 012 条数据, 每条数据包含 55 个属性字段, 本文从中选择了海拔、方向、坡度和树种四种属性进行试验. 使用 VC6.0 软件实现算法, 从检测率和误报率两个方面与李昕等提出的 MMC-MMP^[7] 进行比较.

由表 2 可见, 在 $3En, 2En$ 规则下本文提出的算法检测率均比 MMC-MMP 算法高, 误报率均比 MMC-MMP 算法低. $2En$ 规则下本文提出的算法检测率更高, 误报率比 $3En$ 规则下稍高, 这是因为 $2En$ 规则下的标记的异常数据较多, 在这些异常数据中有些是正常的, 故误报率较高; $3En$ 规则下则产生相反的情况. 因此在应用中可根据对检测率和误

报率的要求,来恰当选择采用 $2En$ 或 $3En$ 规则。

表 2 Forest Cover Type 数据集中检测率与误报率比较

比较项		检测率	误报率
MMC-MMP 算法		91.37%	1.98%
本文算法	$3En$ 规则	93.26%	0.86%
	$2En$ 规则	96.63%	0.90%

由表 3 可看出,本文算法在测试集上的检测运行时间低于 MMC-MMP 算法。主要原因在于本文算法两次利用到云变换算法,减少了一次的计算时间,而且在异常检测中通过云模型大大缩小了对数据的搜索空间,因此本文提出的异常数据挖掘算法取得了较快的运行速度。

表 3 Forest Cover Type 数据集运行时间比较

比较项		运行时间/s
MMC-MMP 算法		1.34
本文算法	$3En$ 规则	0.65
	$2En$ 规则	0.72

5 结束语

本文对基于距离的方法进行了改进,引入了云模型,提出了基于云模型的差异度计算公式,并设计了基于云模型的异常挖掘算法。该算法能够提高混合属性数据异常挖掘的准确度,并能减少计算的时间,通过实验证明了该算法的有效性。本文下一步工作是将该算法用于网络入侵检测,工业控制远程数据监测、故障诊断及计划调度分析等领域。

参考文献:

- [1] E Hong, D W Cheung. Parallel mining of outliers in large database[J]. Distributed and Parallel Databases, 2002,12(1):5-26.
- [2] 李德毅,刘常昱,杜鹃,等. 不确定性人工智能[J]. 软件学报,2004,15(11):1583-1594.
- [3] Eslin Eleazar. Sparse sequence modeling with applications to computational biology and intrusion detection [D]. USA: Colee MBIA University, 2003.
- [4] 蒋盛益,李庆华. 一种增强的局部异常挖掘方法[J]. 计算机研究与发展,2005,42(2):210-216.
- [5] Li Deyi, Han Jiawei, Shi Xuemei, et al. Knowledge representation and discovery based on linguistic atoms [J]. Knowledge-based Systems, 1998, 10(7): 431-440.
- [6] Blake C L, Merz C J. UCI repository of machine learning databases[EB/OL]. (1998-05-01)[2012-09-20]. <http://www.ics.uci.edu/mllearn/MLRepository.html>.
- [7] 李昕,钱旭,王自强. 一种高效的高维异常数据挖掘算法[J]. 计算机工程,2010,36(21):33-36.

作者简介:



李磊 男,(1987-),硕士研究生. 研究方向为数据挖掘、网络入侵检测。

(上接第 81 页)

- [7] Chen Dexi, Yan Li, Wang Haiying. Emulational application by the software FEKO in the radar cross section[J]. Ship Electronic Engineering, 2008(9): 125-129.
- [8] Kusano S, Sato M. Evaluation of trihedral corner reflector for SAR polarimetric calibration [J]. IEICE Transaction on Electronics, 2009,29(1):112-115.
- [9] Polycarpou A C, Balanis C A. A new approach to improve RCS patterns of trihedral corner reflectors[J].

IEEE International Symposium Digest Antennas and Propagation, 1994(3):2298-2301.

- [10] 吴先良. 三角反射器多次散射图形电磁计算[J]. 安徽大学学报:自然科学版,2000,24(3):85-91.

作者简介:

赵俊娟 女,(1978-),博士,工程师. 研究方向为遥感与空间信息处理、数值模拟。