

海量数据干扰下的危险 Web 数据挖掘技术研究

王曙霞,熊曾刚

(湖北工程学院 计算机与信息科学学院,湖北 孝感 432000)

摘 要: 提出一种海量数据干扰下基于自组织映射的危险 web 数据挖掘算法,通过海量数据的预测值与实际值之间的误差对其中的干扰数据进行判断和排除,在此基础上,通过自组织特征映射网络对危险 web 数据进行挖掘,介绍了自组织特征映射网络及输出层竞争的详细过程,确定一个可形成映射的网络,将待挖掘危险 web 数据看作是输入向量输入自组织映射网络中,在输出图上产生相应的胜出点,将相似的输入向量汇聚在映射图的相邻区域,与该区域距离较远的胜出点对应的输入向量则可被判断是危险 web 数据. 仿真实验结果表明,采用所提算法对海量数据干扰下的危险 web 数据进行挖掘,不仅具有很高的挖掘效率,而且在挖掘精度上也有很高的性能.

关键词: 海量数据;干扰;危险 web 数据;挖掘

中图分类号: TP311

文献标识码: A

文章编号: 1000-7180(2016)02-0087-05

Huge Amounts of Data Under the Interference of Dangerous Web Data Mining Technology Research

WANG Shu-xia, XIONG Zeng-gang

(School of Computer and Information Science, Hubei Engineering University, Xiaogan 432000, China)

Abstract: Put forward a lot of data under the dangerous web data mining algorithm based on self-organizing mapping, through massive amounts of data error between the predicted values and the actual value of the judgment and exclude the interference of data, on this basis, through self-organization feature mapping networks for dangerous web data mining, self-organizing feature map network was introduced and the detailed process of the output layer competition, determine a network, can form mapping will be dangerous web data mining as a self-organizing map network input vector input, the output wins produced the corresponding points on the graph, similar input vector convergence in adjacent areas of the map, distance and the region's victory points corresponding to the input vector is dangerous web data can be judge. The simulation results show that the proposed algorithm for huge amounts of data under the interference of dangerous web data mining, not only has high efficiency, and also has high performance in mining precision.

Key words: huge amounts of data; interference; dangerous web data; mining

1 引言

近年来,随着计算机的逐渐发展和广泛应用,人们对网络安全的要求越来越高^[1-2]. Web 数据库是在互联网中以 Web 查询接口方式访问的数据资源,是促进 Internet 发展的主要因素之一,因此,有效挖掘危险 Web 数据、提高网络安全性具有重要意义,它已经成为相关学者研究的重点课题^[3-5].

目前,对危险 Web 数据进行挖掘的方法很多,

主要包括基于分布的危险 Web 数据挖掘方法、基于聚类的危险 Web 数据挖掘方法、基于距离的危险 Web 数据挖掘方法和基于密度的危险 Web 数据挖掘方法. 相关研究也取得了一定的成果,其中,文献[6]提出了一种基于局部离群因子的危险 Web 数据挖掘算法,利用局部离群因子描述一个 Web 数据对象的局部离群程度,这个算法适合用于静态情况下的数据库,但在真实运用里,大部分 Web 数据都是随时间动态增高的,需要从头计算所有 Web 数据目

标的少数偏离因子,计算效率很低.文献[7]提出一种基于 IncLOF 算法的危险 Web 数据挖掘算法,当新增数据目标抵达时,只对受干扰的数据目标从头计算它的 LOF,但这个算法只探究了每次插入一条数据目标时的状况,当数据库中同时插入多个数据对象时,其效率将降低得很快.文献[8]提出一种基于统计的危险 Web 数据挖掘算法,假设给定数据集有一个分布或概率模型,然后将和模型不一致的 Web 数据看作是危险数据,大多统计方法都是针对单个属性的,但大部分数据分布是未知的,因此,该算法挖掘精度不高;文献[9]提出一种基于距离的危险 Web 数据挖掘算法,该算法将距离其他大部分对象的距离较远的 Web 数据看作是危险 Web 数据,但该算法不适用于大数据集;文献[10]提出一种基于密度的危险 Web 数据挖掘算法,当数据集包含多个分布或者数据集由不同密度子集混合而成时,Web 数据是否有害,不仅与距离有关系,而且与邻域里的密度情况相连,但该算法需要大量的先验知识,在实际应用中有时会因先验知识不足造成一定的困难.

本文提出一种海量数据干扰下基于自组织映射的危险 Web 数据挖掘算法,通过海量数据的预测值与实际值之间的误差对其中的干扰数据进行排除,在此基础上,通过自组织映射网络对危险 Web 数据进行挖掘.经实验验证,所提算法具有很高的精度和效率.

2 海量数据干扰下基于自组织映射的危险 Web 数据挖掘算法

2.1 海量数据干扰去除

为了使危险 Web 数据能够被有效地挖掘出来,本文首先通过海量数据的预测值与实际值之间的误差对其中的干扰数据进行判断和排除.

在进行海量数据干扰排除的过程中,随机选择两组数值,分别用 $v_1, v_2, \dots, v_i$ 和 $u_1, u_2, \dots, u_i$ 进行描述,分别对两组数据进行求和,得到的结果分别用 V 和 U 进行描述.依据 V 和 U ,通过下式对二者之间的斜率 K 进行计算:

$$K = V/U \quad (1)$$

通过下式获取 K 的修正频率:

$$pf_i = K * v_i \quad (2)$$

在对海量数据进行干扰排除时, f_i 描述的数据初始频率和 pf_i 描述的修正频率间会产生差值,差值的绝对值被称作偏差,用下式进行描述:

$$pb_i = |pf_i - f_i| \quad (3)$$

选择 pb_i 中的最大值,用 pb_m 进行描述,将其和预先给定的允许偏差 yp 进行比较.

如果 $pb_m \leq yp$,则说明该组数据是稳定正常的,不存在干扰数据;否则,说明该组数据中至少存在一个或几个偏差过大的数据,也就是说该组数据中存在干扰数据,这时选择其中偏差最大的数据重新进行迭代,直至 $pb_m \leq yp$.

通过上述过程可对海量干扰数据进行有效地去除,保证留下数据的纯正性,为危险 Web 数据挖掘提供了有效基础.

2.2 基于自组织特征映射网络的危险 Web 数据挖掘实现

通过上节描述的方法完成海量数据干扰去除后,通过自组织特征映射网络对危险 Web 数据进行挖掘.

自组织特征映射网络模型根据真实神经细胞里的一种特性敏感细胞,在外界讯号干扰下,形生一种敏感特性,这个特性的生成是经过自学习获取的,经常自组织特征映射网络系统由输入层和输出层构成,用图 1 进行描述.

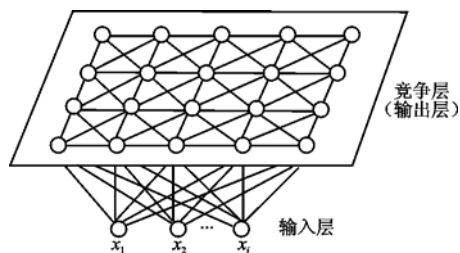


图1 自组织特征映射网络拓扑结构图

可将自组织特征映射网络的输入层当作是外界的信息干扰,输出层看作是大脑神经细胞,输入层的全部神经元,经过权向量与输出层的各神经元相连接,在输出层中竞争的详细过程如下所述.

在自组织特性映射网络里,假如外界输入各种不同样本时,输入样本会使输出兴奋细胞的位置产生变化,但自组织后输出的神经元将汇聚成一些神经元组,表示输入样本的特征性.假如输出神经元处于一个二维空间里,那么是一个平面范围,样本自学

习后,在输出神经里秉承一张二维的映射图,性质相同的神经元靠得比较近,特征不同的神经元分得比较散,这个映射进程是用一个竞争算法结束的,其结果能够让一些规则的输入自动排序,在连接权的调节里,能够让各神经元之间的权值分布和输入样本的概率密度分布相类似。

依据上述自组织特征映射网络的介绍及输出层竞争的详细过程,对危险 Web 数据进行挖掘。

如果输入节点的总量是 N , 输出节点的总量为 M , 所有输入节点及输出节点全部相连接, 共存有 $N \times M$ 个连接, 每个连接都有一个权重。

在实行练习前, 首先需确定一个可以变成映射的网络, 这个映射能够表示危险 Web 数据的特性, 把待挖掘 Web 数据当作是输入向量输入自组织映射网络里, 在输出图上生成相对应的胜出点, 详细计算步骤如下所示。

(1) 对映射拓扑及输出映射图的尺寸实行初始化处理, 输入节点是一个 N 维向量, 输出节点是一个二维阵列, 从而产生“映射图”, 当中输出节点之间相互连接的方式被称作“映射拓扑”。

假设一组经海量数据干扰排除后的输入节点为 $X = \{x_1, x_2, \dots, x_N\}$, 一组输出节点为 $C = \{c_1, c_2, \dots, c_N\}$ 。

(2) 初始化权重参数, 用区间 $[0, 1]$ 之间的随机值或者顺序值对连接权参数实行初始化, 同时对它的输入值与权重参数进行规格化处理。得到一组权重参数用 $W = \{w_{11}, w_{12}, \dots, w_{ij}, \dots, w_{NM}\}$ 实行表示, $1 \leq i \leq N, 1 \leq j \leq M, 0 \leq w_{ij} \leq 1$;

(3) 把一个已知自组织映射中的一个输入样本 X 及输出图上的一个胜出点 c_j (胜出点就是和输入样本间隔最短的输出神经元) 相连接, 当中, $\forall i, \|X - c_j\| \leq \|X - c_i\|$;

(4) 对输入样本 X 进行更新;

(5) 选择最小距离。求出输入向量 X 和各输出节点间的最小距离:

$$\|X - c_j\| = \min \left\{ \sqrt{\sum_{k=1}^N (x_k - w_{jk})^2} \right\} \quad (4)$$

式中, $w_{jk} (k = 1, 2, \dots, N)$ 用于描述输入节点 X 和输出节点之间的 N 个连接权值, 则和最小距离相应的输出节点即为胜出点 c_j 。

(6) 修正权重参数。通过下式对胜出点的权重

参数和其相邻节点进行修正:

$$\frac{dw_i}{dt} = \alpha(t) \gamma(i, t) (X - w_i) \quad (5)$$

式中, α 用来表示 t 的单调递增函数, 也就是学习率, 其最小值不低于 0.01。 γ 用来表示邻域函数, 它表现在了相邻输出节点相关的权重参数受干扰的水平。

$$\frac{d\alpha(t)}{dt} = -a \cdot b^{-bt} \quad (6)$$

$$\gamma(i, t) = c \cdot e^{-\frac{\|c_i - c_j\|^2}{k\sigma^2(t)}} \quad (7)$$

式中, $\|c_i - c_j\|$ 用于描述学习节点 c_i 和输出图上的胜出节点 c_j 之间的距离; a, b, c, k 为正常数; $\sigma(t)$ 用于描述邻域半径。

相近的输入向量应映射成输出图里一样或者相邻的节点, 经过拓扑积对其实行平衡。在对权重参数实行更新的进程里, 邻域集的大小慢慢下降至一个既定的阈值, 直至单个节点。上述表现对危险 Web 数据的挖掘至关重要, 拓扑积可通过下式求出:

$$TP = \frac{1}{2M(M-1)} \sum_{j=1}^M \sum_{k=1}^{M-1} \frac{1}{k} \log \left(\prod_{l=1}^k R_1(j, l) R_2(j, l) \right) c_j \quad (8)$$

式中, M 用于描述输出节点的数量; c_j 用于描述经修正后的胜出点;

$$R_1(j, k) = \frac{d^v(w_j, w_{nk}^A(j))}{d^v(w_j, w_{nk}^v(j))} \quad (9)$$

$$R_2(j, k) = \frac{d^A(j, n_k^A(j))}{d^A(j, n_k^v(j))} \quad (10)$$

式中, $n_k^v(j)$ 用来表示输出空间里 j 的第 k 个最近的邻居; $n_k^A(j)$ 用来表示输入空间里 j 的第 k 个最近的邻居; $d^v(w_j, w_{nk}^A(j))$ 用于描述输入空间中 w_j 和 $w_{nk}^A(j)$ 之间的距离。

通过上述过程产生网络形成映射图后, 对其参数实行解析, 从而挖掘危险 Web 数据。经过求得节点的相对位置和节点之间的相对距离获得输出图上占有最小拓扑面积的最大节点集。对函数实行评价, 获得胜出点的一个子集 C_s , 从而让 $F(C_s)$ 的值抵达最小, 当中 γ 用来表示胜出节点的总数; A 用来表示映射图里被胜出节点的一个子集占领的拓扑范围; $C_s(k)$ 用来表示存有 k 个成员的一个胜出节点的

子集:

$$F(C_S) = \frac{A(C_S(k))}{k^2} \quad (11)$$

经过上述解析进程,把类似的输入向量聚集在映射图的相邻范围,而和这个范围间隔较远的胜出点对应的输入向量则是危险 Web 数据,对其进行挖掘。

3 仿真实验结果

为了验证本文提出的基于自组织映射的危险 Web 数据挖掘算法的有效性,需要进行相关的实验分析。实验将传统聚类算法作为对比进行分析,从挖掘精度和挖掘效率两个方面进行验证。实验对象如表 1 所示。

表 1 实验对象

数据集	总样本集	初始样本集	增量样本集
ASL	700	150	50
Breast-cancer	5 500	1 400	400
Flare-solar	38 000	6 200	3 100
Spambase	3 000	300	100

3.1 运行效率比较

随机选择 2 000 个数据当作初始数据集,新添加的数据目标数量是 1、10、50、100、200、400、600,采用本文算法和传统算法的运行时间结果用表 2 进行描述。

表 2 两种算法运行时间比较结果

新增数据对象数量	本文算法	传统算法
1	0.317	0.538
10	0.323	0.625
50	0.417	0.831
100	0.511	1.254
200	0.932	1.544
400	1.547	1.831
600	1.931	2.105

分析表 2 可知,当新增数据对象数量较少时,本文算法和传统算法所需的时间均比较少,但随着新增数据的逐渐增加,本文算法的运用时间显著低于传统算法,说明本文算法具备很高的使用效率,这也验证了本文算法的高效性。

为了进一步验证本文算法的高效性,将两种算

法挖掘过程的运行代价进行比较,两种算法平均挖掘过程运行代价用图 2 进行描述。

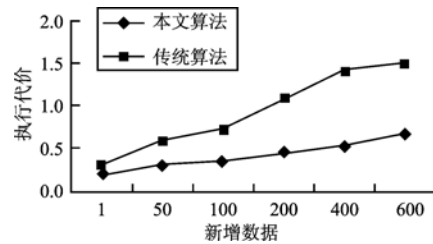


图 2 两种算法执行代价比较结果

分析图 2 可以得出,当新增数据量较小时,本文算法和传统算法的执行代价均较小,随着新增数据量的逐渐增加,两种算法的执行代价均逐渐增加,但本文算法的执行代价增多的幅度显著低于传统算法,且一直低于传统算法,这证明本文算法解决了传统算法的不利因素,提高了危险 Web 数据挖掘的效率,得到较为满意的结果。

3.2 挖掘精度比较

随机选择 2 000 个数据目标当作初始数据集,把 2 000 个数据目标当作新增加的数据目标,在初始数据集和新增加的数据条件里,真实的危险 Web 数据数量为 10、20、30、40、50。分别采用本文算法和传统算法对危险 Web 数据进行挖掘,得到的准确挖掘危险 Web 数据的个数如表 3 所示。

表 3 两种算法精度比较

危险 Web 数据	准确发现危险 Web 数据的数量	
	本文算法	传统算法
10	8	5
20	16	10
30	27	18
40	36	27
50	45	34

分析表 3 可知,采用本文算法对危险 Web 数据进行挖掘,挖掘准确性明显高于传统算法,说明本文算法具有很高的挖掘精度,验证了本文算法的有效性。

为了进一步验证本文方法的精度,针对 Breast-cancer 数据集,对本文算法和传统学习算法的误报率进行综合,获得的后果用图 3 来表述。

分析图 3 可以得出,使用本文方法获得的误报率曲线显著低于传统学习手段,且更加平稳,表明本

文方法能够满足实际应用的需要,而传统学习方法误报率较高,无法满足实际需求。

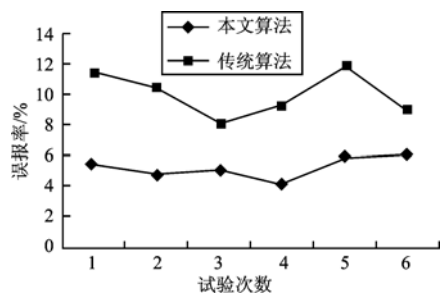


图3 本文方法与传统方法误报率比较结果

4 结束语

本文提出一种海量数据干扰下基于自组织映射的危险 Web 数据挖掘算法,通过海量数据的预测值与实际值之间的误差对其中的干扰数据进行判断和排除,在此基础上,通过自组织特征映射网络对危险 Web 数据进行挖掘,介绍了自组织特征映射网络及输出层竞争的详细过程,确定一个可形成映射的网络,将待挖掘危险 Web 数据当作是输入向量输入自组织映射网络里,在输出图上形成相对应的胜出点,把类似的输入向量聚集在映射图的相邻范围,和这个范围间隔较远的胜出点对应的输入向量则可以被辨别成危险 Web 数据。仿真实验结果表明,采用所提算法对海量数据干扰下的危险 Web 数据进行挖掘,不仅具有很高的挖掘效率,而且在挖掘精度上也有很高的性能。

参考文献:

- [1] 方少卿,周剑,张明新. 基于 Map/Reduce 的改进选择算法在云计算的 Web 数据挖掘中的研究[J]. 计算机应用研究, 2013, 30(2):377-379.
- [2] 李悦,高晶,雷鸣. 基于云计算技术的 Web 数据挖掘的算法研究[J]. 科技资讯, 2014, 12(18):17-17.
- [3] 阮梦黎. 基于半结构化分割的 Web 热点数据挖掘算法[J]. 科技通报, 2015, 31(4):115-117.
- [4] 刘啸,刘玉龙. 基于改进型迭代算法的 Web 数据关联规则挖掘[J]. 科技导报, 2015, 33(3):90-94.
- [5] 刘奎. 基于 Web 数据挖掘的网络课程设计研究[J]. 赤峰学院学报:自然科学版, 2014, 26(21):206-207.
- [6] 吴锐,孙银香. 云计算平台下的 Web 数据挖掘研究[J]. 网络安全技术与应用, 2014, 14(8):67-68.
- [7] 张鑫. Web 数据挖掘在云计算环境下的研究[J]. 数字技术与应用, 2013, 31(3):92-92.
- [8] 唐利超. 基于 Web 的数据挖掘技术研究及其应用[J]. 科技致富向导, 2013, 29(30):25-25.
- [9] 莫学值,韦必忠. Web 数据兴趣区域数据挖掘过程分析[J]. 科技通报, 2013, 29(4):202-203.
- [10] 刘佳. 基于 Web 服务的数据挖掘系统算法的设计[J]. 电脑知识与技术, 2013, 13(5):2951-2954.

作者简介:

王曙霞 女, (1975-), 硕士, 副教授. 研究方向为智能计算与网络安全. E-mail: wsxwsxxia@163.com

熊曾刚 男, (1974-), 博士, 教授. 研究方向为对等计算、网格计算、云计算和信息系统分析与集成.