软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2018,29(12):3716−3732 [doi: 10.13328/j.cnki.jos.005488]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

一种时效感知的动态加权 Web 服务 QoS 监控方法*

何志鹏[1,2], 张鹏程[1], 江 艳[1], 吉顺慧[1], 李雯睿[3]

[1](河海大学 计算机与信息学院,江苏 南京 211100)

[2](中国科学院 计算机网络信息中心,北京 100190)

[3](南京晓庄学院 信息工程学院,江苏 南京 211171)

通讯作者: 张鹏程, E-mail: pchzhang@hhu.edu.cn

摘 要: 服务质量(quality of service,简称 QoS)是衡量 Web 服务好坏的重要标准,也是用户选择 Web 服务的重要依据.能够实时而准确有效地对 Web 服务进行监控,是 Web 服务质量保障的重要基础.为此,提出了一种时效感知的动态 Web 服务 QoS 监控方法.该方法在传统加权监控方法中融入了滑动窗口机制和信息增益原理,简称 IgS-wBSRM (information gain and sliding window based weighted naive Bayes QoS runtime monitoring).该方法以一定的初始训练样本进行环境因素权值初始化,利用信息熵(information entropy,简称 IE)及信息增益(information gain,简称 IG)对样本所处混沌状态的确定作用,依次读取样本数据流,计算样本数据单元出现前后各影响因子组合的信息增益,结合 TF-IDF(term frequency-inverse document frequency)算法对早期的初始化权值进行动态更新,修正传统算法对监控分类的类间分布偏差问题和参数未更新问题.另外,考虑训练样本数据的时效性,结合滑动窗口机制来对影响因子组合权值进行同步更新,以消解长期累积的历史累赘数据对近期服务 QoS 的影响.在模拟数据集和开源数据集上的结果表明:利用滑动窗口机制可以有效摒弃历史数据的过期信息,结合滑动窗口机制实现的基于信息增益的动态权值算法能够更加准确地监控 Web 服务 QoS,总体监控效果明显优先于现有方法.

关键词: 服务质量;时效感知;信息增益;滑动窗口;动态监控

中图法分类号: TP311

中文引用格式: 何志鹏,张鹏程,江艳,吉顺慧,李雯睿.一种时效感知的动态加权 Web 服务 QoS 监控方法.软件学报,2018, 29(12):3716−3732. http://www.jos.org.cn/1000-9825/5488.htm

英文引用格式: He ZP, Zhang PC, Jiang Y, Ji SH, Li WR. Time-Aware Web service QoS monitoring approach under dynamic environments. Ruan Jian Xue Bao/Journal of Software, 2018,29(12):3716−3732 (in Chinese). http://www.jos.org.cn/1000-9825/5488.htm

Time-Aware Web Service QoS Monitoring Approach Under Dynamic Environments

HE Zhi-Peng[1,2], ZHANG Peng-Cheng[1], JIANG Yan[1], JI Shun-Hui[1], LI Wen-Rui[3]

[1](College of Computer and Information, Hohai University, Nanjing 211100, China)

[2](Computer Network Information Center, The Chinese Academy of Sciences, Beijing 100190, China)

[3](School of Information Engineering, Nanjing Xiaozhuang University, Nanjing 211171, China)

Abstract: Quality of service (QoS) is an important criterion to measure the quality of Web services, and it is an important aspect for users to choose Web services. This paper proposes a dynamic weighting Web service QoS monitoring method based on information gain

* 基金项目: 国家自然科学基金(61572171, 61702159); 江苏省自然科学基金(BK20170893); 中央高校基本科研业务费(2018B16014)

Foundation item: National Natural Science Foundation of China (61572171, 61702159); Natural Science Foundation of Jiangsu Province (BK20170893); Fundamental Research Funds for the Central Universities of China (2018B16014)

收稿时间: 2016-12-24; 修改时间: 2017-07-03; 采用时间: 2017-10-31; jos 在线出版时间: 2018-02-08

CNKI 网络优先出版: 2018-02-08 13:25:16, http://kns.cnki.net/kcms/detail/11.2560.TP.20180208.1325.014.html

and sliding window mechanism. IgS-wBSRM initializes the environmental factors' weights with a certain amount of initial training samples. It also employs the theory of information entropy and gain to determine the chaotic state of the samples. IgS-wBSRM reads the sample data flow in sequence, calculates the information gain of each impact factor combination after the arrival of sample data unit. Then it updates the initialized weights with TF-IDF in a dynamic environment. In this way, IgS-wBSRM can correct the uneven classification problem between classes and the off-line constant problem in traditional monitoring approach wBSRM. Moreover, considering the timeliness of the training sample data, IgS-wBSRM combines sliding window mechanism to update the weights of each impact factor combination, so that it can eliminate the impact on the recent service running state that the accumulated historical data bring. The experiment results under a real world QoS Web service data set demonstrate that with the sliding window mechanism, IgS-wBSRM can abandon the expiration information of historical data effectively, and the dynamic weighting method combined with sliding window mechanism and information gain can monitor the QoS more accurately. The overall monitoring effect is markedly better than existing QoS monitoring approaches.

Key words: quality of service; time-aware; information gain; sliding window; dynamic monitoring

随着云计算、大数据等新技术对传统 Web 服务的推动,企业之间甚至于个人对服务的需求正逐步转变为服务间的交互.在复杂多变的环境中,Web 服务的运行环境及其运行质量也无时无刻不发生着改变.为了让服务能够适应动态异构的环境,亟待解决在实时变化的环境中对 Web 服务的服务质量(qulity of service,简称 QoS)进行准确而灵敏地监控[1−3].

服务质量是 Web 服务的一组非功能属性的集合,是衡量第三方服务好坏的重要指标,每个 QoS 属性表示 Web 服务某一方面的质量信息,如响应时间、吞吐量、可靠性等[4].由于每个质量属性都有相应的属性值,如何在不确定的网络和服务本身的弹性下动态而灵敏地监控 Web 服务是否失效,即转变为如何在不确定环境下利用 QoS 属性数据进行有效且实时的监控问题.

大多数的 QoS 属性标准可以用概率质量属性的方式来表达[5],如:响应时间可描述为“某服务对客户请求的响应时间小于 10s 的概率为 50%”,可用性可描述为“某服务 24 小时内离线的概率应小于 0.02”等.这使得传统上对 QoS 属性数据的分析转变为在不确定性环境下基于概率或统计的手段对现有属性数据进行统计分析与计算,也即近年来兴起的概率监控方法(probabilistic monitor)[6].

目前,研究人员已经提出了不少概率监控方法,典型的包括基于估值计算[7]、基于经典假设检验 SPRT[6,8,9]和基于贝叶斯的方法[10−12].现有对 QoS 属性进行的概率监控方法作为一种十分重要的保证措施,仍不完善.大部分方法[7−10]未考虑复杂多变环境下的实时监控问题,一些考虑了环境因素影响的研究方法[11,12]并未对监控的时效性和监控分类带来的类间分布不均衡问题进行深入研究.为了解决这些问题,本文提出了一种新颖的时效感知监控方法,该方法结合滑动窗口机制和信息增益原理来实现 Web 服务 QoS 的动态加权监控,简称 IgS-wBSRM(information gain and sliding window based weighted naive Bayes runtime monitoring).方法着重考虑对环境的影响因素进行时效性加权,同时消除监控分类间影响因子分布不均衡而带来的监控失准问题,以使监控能够适应动态的运行环境和数据环境,让监控更具有实时动态性与准确性.

本文的主要贡献包括:

(1) 从监控时效性出发,构建了细粒度的动态监控算法.引入滑动窗口机制,同时,基于该机制与信息增益的结合,设计了于新旧窗口不同的监控样本数据集下,基于信息增益的改进动态权值更新算法.使得运行时监控始终能够兼顾最新的样本数据块,对运行监控参数实时地进行动态更新;

(2) 将融入滑动窗口机制进而结合信息增益改进的加权算法嵌入运用于 Web 服务 QoS 监控中,同步解决了现有相关研究中利用传统 TF-IDF 算法进行加权所产生的类间分布偏差问题;

(3) 基于给定标准下的模拟数据集和开源数据集下精心设计了相关实验,并与现有 QoS 监控方法进行了比较,实验结果验证了方法的有效性和优越性.

本文第 1 节介绍目前研究相关的动态.第 2 节介绍并引入有关概念.第 3 节是 IgS-wBSRM 方法概述和算法描述.第 4 节采用自定义的模拟数据集和真实数据集来验证方法的有效性.第 5 节是总结与未来工作展望.

1 相关工作

1.1 传统Web服务QoS相关监控方法

Zeng 等人[13]根据企业定义的指标及相关的评估公式,采用模型驱动方法设计了 QoS 观测模型,模型能够系统地探测服务 QoS.但该方法仅考虑企业自定义的标准,并没有考虑到用户定义的可靠性标准.Radovanovic 等人[14]在基于 TR-069 远程管理协议的云平台上监控不同用户设备的 QoS,方法提供了不同的能量分布模型以及按照特定 QoS 需求的 IoT(internal of thing)网络.Coppolino 等人[15]所提出的方法集成了 QoSMONaaS(QoS monitoring as a service)及 SocIoS 框架监控 SLA(service level agreement)文件中的相关 QoS 指标,方法的有效性还没得到有力证明.Michlmayr 等人[16]提出了同时监控客户端及服务器端的 QoS,客户端监控是在特殊的时间间隔内发送请求并使用 QUATSCH 工具监控,服务器端监控室持续监测 QoS 属性值,但该方法性能开销很大.Raimond[17]使用时间自动机监控服务提供者与用户之间的 SLA,SLA 需求包含很多非功能属性.总体而言,这些方法并没有将 Web 服务 QoS 监控归纳为一般的概率监控问题.

1.2 概率监控方法

Chan 等人[7]在.Net 应用程序中监控 PCTL(probabilistic computation tree logic)属性,通过观察成功或不成功的监控样本数量和总体样本数量的比例获得统计证据,进而与预期的概率阈值进行比较,得出相应监控决策结论.但在此方法中,由于缺乏对结果的统计学分析,容易因预定义的概率不同于属性的真正概率,导致实际误差很大.Sammapun 等人[18]对概率扩展的 MEDL(meta-event definition language)的 MaC(monitoring and checking)框架进行验证,该方法先估计成功样本占总样本数的概率,再依据假设检验判断系统在给定的置信水平下是否符合概率属性.Grunske 等人[6]提出了使用连续随机逻辑(continuous stochastic logic,简称 CSL)的子集 CSL^{Mon} 的概率监控方法 ProMo,该方法使用 SPRT(sequential probabilistic ratio test),在显著性水平α和 $1-\beta$下验证了 CSL^{Mon} 公式的正确性,但该方法的猜想需要整个监控过程中监控概率不变,不支持持续监控.Zhang 等人[9]提出了 PTPSC(time property sequence chart[19]的概率扩展),结合 TBA 和 SPRT 自动生成概率监控;Grunske 等人[8]提出了基于假设检验测试的改进的通用统计决策程序,该程序通过回退统计分析并复用以前的监控结果,实现了连续概率监控,但 SPRT 在其系统的生命周期内,监控概率需要一直不变,而在现实生活中,概率规范会随着客户的需要而改变,一旦概率改变,以前的监控结果将无法复用,监控需要重新开始;同时,对于 SPRT,当系统的实际概率分布于给定属性标准概率附近时,所得监控结果将大量落入 SPRT 的中立区,出现方法失效.

为解决先前方法存在的连续监控等问题,Zhu 等人[10]提出了一种基于传统贝叶斯的概率监控方法,利用相关 Web 服务运行时 QoS 属性参数统计计算贝叶斯因子,在以假设检验的形式对监控决策进行判断.但该方法受先验分布的影响,而在未知状态下,合理的先验分布概率在实际中很难得以选择.

Zhuang 等人[11,12]针对之前方法都未考虑在监控过程中实际存在的环境因素的影响,如服务器的地理位置、响应负载时间段等,提出了加权朴素贝叶斯监控方法 wBSRM.该方法通过 TF-IDF 算法衡量环境因素的影响,通过学习部分样本对监控结果分类构造加权朴素贝叶斯分类器.然而,此方法还存在一定的缺陷,分析如下.

(1) 考虑了环境因素的影响,使得监控更加贴合实际,然而此方法仅通过训练早期部分样本得到权值表,而该权值表在后期被无限调用,权值的计算缺乏动态性与实时性;

(2) 从早期开始的历史冗余样本所携带的过期分类信息极易对现阶段的实时监控产生影响,致使监控决策结果不再准确;

(3) 现有基于加权朴素贝叶斯的 wBSRM 方法所采用的传统 TF-IDF 加权方法在 Web 服务监控中具有决策分类间分布偏差问题,该问题使得监控结果容易受数据样本的类间分布变动影响,从而将导致服务监控出现监控延迟判断、二分类监控决策间噪声抖动等现象.

本文引入滑动窗口机制,并以其为基础,结合信息增益对现有权值算法进行改进,同步解决了监控方法的实时动态性以及传统权值算法在 Web 服务监控中具有决策分类间分布偏差问题.

2 相关概念引入

2.1 加权朴素贝叶斯分类器

朴素贝叶斯分类模型[20]来源于贝叶斯分类的基础:贝叶斯定理.其思想基础为:对于给出的待分类项,求解在此项出现的条件下各个类别出现的概率,其中,概率最大者就被认为此待分类项所属类别.

令 $U=\{X,C\}$是离散随机变量的有限集,其中 $X=\{x_1,x_2,\ldots,x_n\}$为样本变量,类变量 C 的取值范围$\{c_0,c_1\}$.样本变量 $X=\{x_1,x_2,\ldots,x_n\}$属于类 c_j 的概率,根据贝叶斯定理有:

$$p(c_j \mid x_1,x_2,...,x_n)=\frac{p(x_1,x_2,...,x_n \mid c_j)p(c_j)}{p(x_1,x_2,...,x_n)}=\alpha p(c_j)\prod_{i=1}^{n}p(x_i \mid x_1,x_2,...,x_{i-1},c_j) \tag{1}$$

而根据朴素贝叶斯的独立性假设:默认样本分类项中的各元素相互独立,那么不难看出,对应提出的系数项 $\alpha=\frac{1}{p(x_1,x_2,...,x_n)}$ 为独立于分类的常数.

而对于 $p(c_j \mid x_1,x_2,...,x_n)=\alpha p(c_j)\prod_{i=1}^{n}p(x_i \mid c_j)$,根据最大后验准则(maximum a posteriori,简称 MAP),朴素贝叶斯分类器对于观测样本集 $X=\{x_1,x_2,\ldots,x_n\}$以后验概率 $p(c_j|x_1,x_2,\ldots,x_n)$最大的类作为样本集的类标签.所以,朴素贝叶斯分类器可表述为

$$C(X)=\arg\max_{c_j\in C}\ p(c_j)\prod_{i=1}^{n}p(x_i \mid c_j) \tag{2}$$

从上述表达式可以看出,只需要从训练集中学习两组参数值:先验概率 $p(c_j)$和类条件密度 $p(x_i|c_j)$便可以了.在实际场景下,由于 $p(x_i|c_j)$以及 $p(c_j)$都为较小的概率数值,在运算过程中,若依照公式(2)进行连乘运算,则可能运算到最后,由于计算机精度受限而对计算值小数部分做忽略以及截取处理,甚至有可能最终无法显示.考虑到本身对概率计算中所需的较高精度与方法要求,同时为了方便运算,故最终分类决策判定公式取对数变换如下:

$$C(X)=\arg\max_{c_j\in C}\left\{\log\left(p(c_j)+\sum_{i=1}^{n}p(x_i \mid c_j)\right)\right\} \tag{3}$$

然而,朴素贝叶斯的独立性假设在实际的应用中并不常见.在实际监控分类中,每个样本对分类的影响是不同的,朴素贝叶斯分类器并没有很好地考虑这个问题.那么一种切合实际的想法是:根据不同的样本对分类系统的作用来为每个样本的概率加一个权重,以此平衡其实际影响关系,并称其为加权朴素贝叶斯.相对应的,有如下分类模型如图 1 所示.

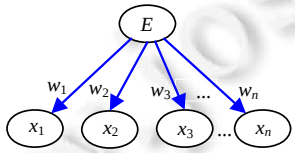

Fig.1 Weighted naive bayes classifer model

图 1 加权朴素贝叶斯分类模型

假定第 i 个样本项 x_i 的权值为 w_i,现定义有加权朴素贝叶斯的决策公式如下:

$$C(X)=\arg\max_{c_j\in C}\left\{\log(1+p(c_j))+\prod_{i=1}^{n}w_i\times\log(1+p(x_i \mid c_j))\right\} \tag{4}$$

该公式中体现了每个样本对分类决策的作用,权重的实际意义是每个样本对分类所产生的影响.其中,取 $\log(1+p(x_i|c_j))$考虑到实际对于概率 $p(x_i|c_j)$其值小于 1,则 $\log(p(x_i|c_j))$小于 0,而权值代表的是样本对分类的重要性,且 log 函数在有限定义域上为单调递增函数,故于此取概率加 1 的对数值,使得加权正确,并且对分类的决策无影响.

2.2 信息熵与信息增益

信息熵与信息增益的概念源于香农的信息论,而对于熵,宏观上反映的是任何一种能量在空间中分布的均匀程度[21].若该能量分布越具有倾向性地密集,则对应熵值越小.对于信息熵,也称作平均自信息量,若在随机事件发生前计算一影响项的熵,那么此时的熵值代表着该影响项对该事件发生的不确定程度的度量;相反,若于之后计算,则其值代表可从该影响项中获取的信息量.

对于给定的概率分布 $p=\{p_1,p_2,\ldots,p_n\}$,则该分布所携带的信息量就被称为 P 的熵 $H(P)$,公式如下:

$$H(P)=-(p_1\times+p_2\times\log_2^{p_2}+\ldots+p_n\times\log_2^{p_n})=-\sum_{k=1}^{n}p_k\times\log_2^{p_k} \tag{5}$$

而对于本监控中的信息增益(information gain)[22],顾名思义,其对应为影响因子组合样本项出现前后为决策分类事件所带来的熵的差值.

对应的有信息增益的定义公式:

$$\begin{aligned}IG(s)&=H(C)-H(C\mid s)\\&=-\sum_{c_j\in C}p(c_i)\times\log(p(c_j))+p(s)\times\sum_{c_j\in C}p(c_j\mid s)\times\log(p(c_j\mid s))+p(\overline{s})\times\sum_{c_j\in C}p(c_j\mid\overline{s})\times\log(p(c_j\mid\overline{s}))\end{aligned} \tag{6}$$

对应到本服务监控方法,其中,$C=\{c_0,c_1\}$,对应表示决策分类的结果集,其中,$H(C)$对应在没有出现某个具体的影响因子组合样本项 s 之前监控样本属于某决策类别的熵,即此时对监控分类结果的不确定性度量,$H(C|s)$为影响因子组合样本项 s 出现后样本属于该决策类别的熵,对应样本项 s 对分类的不确定性度量.而对应的二者之差,则表示影响因子组合样本项 s 出现前后对样本分类的影响不确定程度,也即该影响因子组合样本项 s 所对应的信息增益.公式(6)中,$p(s)$表示 s 样本项出现在 C 类别中的概率, $p(\overline{s})$ 表示 s 出现在样本数据中但不出现在 C 类别中的概率.

信息增益值越大,则该影响因子组合数据单元 s 对样本分类的影响程度越大,应该赋予较高的权重;相反,则该影响因子组合数据单元 s 对样本分类的影响程度较小,监控时应该赋予较低权重.

总而言之,信息增益分类的基本思想:某个特征项的信息增益越大,贡献越大,对分类也越重要.

2.3 结合信息增益的改进TF-IDF加权

在综合考虑了词条的词频 TF(term frequency)和反文档频率 IDF(inverse document frequency)[23]后我们发现:传统的 TF-IDF 算法仅将文档集作为整体来考虑,对于 IDF 值的计算并未考虑到特征项在类间分布情况存在的不均衡偏差问题.传统 TF-IDF 算法的主要思想是:如果一个分词分布于各别文档的密度越大,则意味其对该文档内容的区别能力越强;如果一个分词在文档集中分布的密度越均匀,那么说明该分词对该文档内容的区别能力越差.在早期的利用 TF-IDF 算法进行影响因子加权的 Web 服务质量监控中,公式如下:

$$w_{R_i}=TF\times IDF(R_i)=(n_{c_j}^{R_i}/n_{c_j})\times\log\left(\frac{n}{n_{R_i}}\right) \tag{7}$$

其中, w_{R_i} 是表示环境因子组合 R 对分类的权值,R_i 代表第 i 个组合项, $n_{c_j}^{R_i}$ 表示影响因子组合 R_i 中属于类别 c_j 的数量, n_{c_j} 表示类别 c_j 的个数,n 表示样本的整体数, n_{R_i} 则对应监控数据项中为 R_i 的数据项数.

影响因子组合的信息增益是其在类间分类能力的表示,且信息增益值与其分类能力成正比,即:信息增益值的大小反映了影响因子组合对分类的影响强弱.TF-IDF 计算的是影响因子组合的信息量,信息增益计算的是新加入样本的环境因子组合带来的信息量,进行归一化处理后,二者属同一数量级,现定义权值更新公式如下:

$$w_{R_i\_new}=w_{R_i}\times IG(R_i) \tag{8}$$

其中,$IG(R_i)$表示影响因子组合 R_i 的信息增益,具体定义见第 2.2 节.

3 一种时效感知的动态加权 Web 服务 QoS 监控方法 IgS-wBSRM

3.1 IgS-wBSRM方法引入与概述

利用 TF-IDF 算法进行权值计算,在信息检索领域甚至数据挖掘领域中都有着及其重要的意义与作用.早期的相关研究[11,12]曾利用该算法在监控初始化期以有限的一定数量的数据样本对监控服务所涉及的环境影响因子进行权值计算,而后期所有的监控都依据该早期的初始化权值表进行相应计算,算法明显呈现出两大弊端:

(1) 传统的 TF-IDF 算法将分类数据文档集当作一个完整实体来考虑,其中,IDF 的计算并没有考虑到环境影响因子在类间的分布情况,见表 1.

Table 1 Compositional impact factors' frequency table (times)

表 1 影响因子组合在不同分类文档中的频数表(频次)

影响因子组合	类别			
	C_0		C_1	
	$D_{C_0\_old}$	$D_{C_0\_new}$	$D_{C_1\_old}$	$D_{C_1\_new}$
⟨China,China⟩	8	4	0	0
⟨Poland,Turkey⟩	0	6	6	0

其中:影响因子组合以类似⟨China,China⟩的形式给出,分别代表客户端地理位置与服务器端地理位置;$D_{C_0\_old}$ 表示旧时间片窗口范围内 C_0 类分类数据文档; $D_{C_0\_new}$ 表示新时间片窗口范围内 C_0 类分类数据文档;同理,依次有 $D_{C_1\_old}$,$D_{C_1\_new}$.

由表 1 可见,⟨China,China⟩环境影响因子组合在 C_0 类别的中的 2 个分类数据文档集中大量出现,而在其他类别基本未出现,那么这种环境影响因子的分类能力显然是很强的,故理应赋予其较高权值,但同时我们也可以看到,⟨Poland,Turkey⟩环境影响因子组合也同时出现在了 2 个分类数据文档中,其中,C_0 类数据文档集中出现 1 次,C_1 类数据文档集中出现 1 次,且在两分类数据文档中均匀分布.显然,这类环境影响因子组合携带识别两种类别的分类信息很少,应赋予较低权值.现在对上述数据依照现有的监控加权研究中使用的传统 TF-IDF 算法进行加权,结果见表 2.

Table 2 Compositional impact factors' weight table

表 2 影响因子组合在不同分类文档中的权值表

影响因子组合	类别			
	C_0		C_1	
	$D_{C_0\_old}$	$D_{C_0\_new}$	$D_{C_1\_old}$	$D_{C_1\_new}$
⟨China,China⟩	0.089 4	0.060 3	0	0
⟨Poland,Turkey⟩	0	0.090 4	0.088 5	0

由表 2 可以看到:影响因子组合⟨Poland,Turkey⟩反而获得了更高的权重,甚至于超过了本明显对 C_0 类有着很好分类能力的⟨China,China⟩影响因子组合.这是一个明显错误的结果,也是传统 TF-IDF 加权的类间分布偏差的体现.

(2) 仅利用传统的 TF-IDF 加权方法,对早期部分样本数据进行一次初始化权值训练而后期无限期使用,这显然是与复杂多变的 Web 服务环境有悖的.这样的方法不具备灵活性与动态性,很容易将历史积累的错误信息带到当前环境的计算中来;基于概率统计的服务监控,若监控在从开始统计起始一直不断加入动态样本情况下仍然将初始统计的成功率加入统计,即使此时对 QoS 标准进行判断仍然满足,实则此时的服务也已经可能失效很久,而这正是由于历史冗余数据的基数大,统计信息不易改变的特性.而现有监控将已经过期的样本也加入统计计算,服务在发布后短期内一定是正常运行的,在后期未加强维护的原因下,会出现服务失效的情况.而如果我们从开始就对 Web 服务 QoS 的概率质量属性标准进行成功率统计,并且一直加入监控的判断中,如此一来,将会于之后明显导致监控结果的错误.

IgS-wBSRM 方法结合信息增益与滑动窗口机制来改进加权模式,使之成为一种时效感知的动态加权 Web

服务 QoS 监控方法,方法整体结构如图 2 所示.

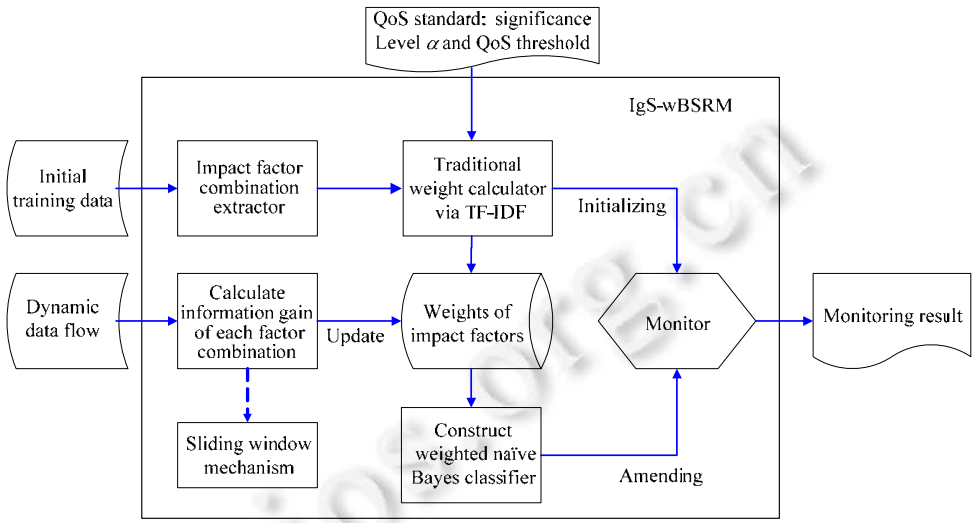

Fig.2 Structural diagrams of IgS-wBSRM system

图 2 IgS-wBSRM 系统结构图

主要模块功能介绍如下:

- Initial training data, Dynamic data flow:利用早期部分历史 QoS 数据样本(即 initial training data)进行初始化权值表训练,而监控器运行时则根据运行时的动态 QoS 数据流(即 dynamic data flow)进行异步权值更新;
- Impact factor combination extractor:对数据集进行影响因子组合分词得到各个独立的影响因子组合并进行频数统计;
- Calculate information gain of each factor combination:依赖结合滑动窗口机制,通过信息增益来改进传统 TF-IDF 算法来对各影响因子组合进行实时动态的权值更新;
- Monitor:通过利用传统方法进行加权的初始化权值等参数来初始化监控器,利用以融入滑动窗口机制的基于信息增益动态加权方法更新过的各影响因子权值等参数来修正监控器.

IgS-wBSRM 方法融入滑动窗口机制进而结合信息增益,很好地解决了上述所提出的问题.同样,以上述列举问题时所提出例子为例,当采用 IgS-wBSRM 模型进行监控后,可以看到结果发生了显著且如预期一致的变化,具体权值见表 3.

Table 3 Compositional impact factors' weight table

表 3 影响因子组合在不同分类文档中的权值表

影响因子组合	类别			
	C_0		C_1	
	$D_{C_0\_old}$	$D_{C_0\_new}$	$D_{C_1\_old}$	$D_{C_1\_new}$
⟨China,China⟩	0.004 8	0.007 2	0	0
⟨Poland,Turkey⟩	0	0	0	0

从表 3 中可以看到:原本未携带识别两种类别的分类信息的⟨Poland,Turkey⟩环境影响因子组合于此时的权值表中 C_0 与 C_1 两类中的权值均被置 0,而⟨China,China⟩影响因子组合的权值则如其预期侧重分布在 C_0 类中.由此可以看出:融入滑动窗口机制进而结合信息增益所改进的加权算法能够有效地避免类间分布偏差现象,可以很好地避免赋予那些在监控分类间分布均匀但实则并无携带识别与决策分类能力的影响因子组合以较高的

权值.

3.2 IgS-wBSRM方法实现

定义 QoS 属性标准β、QoS 属性值要求 *QoS_Value* 以及滑动窗口的大小 *m*,样本分为训练样本和监控样本.首先通过远程过程调用获取早期的训练样本,定义影响因子组合,通过带有环境因素的样本获取影响因子组合(主要是用户的位置及服务位置的组合),并统计各类影响因子组合的样本满足 QoS 属性值要求的样本数,通过统计计算满足 QoS 标准的次数,得到传统 TF-IDF 算法的参数,代入公式(7)计算不同的环境影响因子组合对分类的权值;同时训练出样本的先验信息,监控时,加入新的样本时,计算样本的信息增益更新训练阶段得到的权值,并通过滑动窗口大小保证样本的时效性,更新分类器的各项参数,以后验概率最大的类作为最终样本集的分类结果.详细过程如下.

设 $C=\{c_0,c_1\}$是监控分类的结果集,C_0 表示满足 QoS 属性标准,C_1 表示不满足 QoS 属性标准,例如:QoS 属性标准为 Web 服务响应时间小于 10s 的概率大于 50%,令 $X=\{x_1,x_2,x_3,\ldots,x_n\}$为样本向量,数据预处理时,我们将满足样本属性值要求的样本值赋值为 1,将不满足样本属性值要求的样本赋值为 0,即 $0<x_k<1$,其中,$k\in\{1,2,\ldots,n\}$,每加入一个样本对 QoS 标准进行检验,并记录满足 QoS 标准的次数,记录下每个样本是否满足 QoS 属性值要求以及在此样本处时成败检验,即,是否满足此时的概率属性要求.当样本数量达到滑动窗口的大小时,此时加入一个新的样本的同时,去掉最早期的一个样本,同时根据丢弃的样本调整满足 QoS 属性值要求的样本数以及成功通过概率属性标准要求的次数,同时调整分类器中样本的条件概率,从而实现滑动窗口策略,保证监控样本是现阶段近期未过期的有效样本.

根据加权朴素贝叶斯的决策函数公式(4),首先要求 $p(c_j)$和 $p(x_i|c_j).p(c_j)$为某一类别的先验概率,其计算公式如下:

$$p(c_j)=\frac{n}{n_{c_j}} \tag{9}$$

其中,n_{c_j} 表示 QoS 标准检验可靠度属于类 c_j 的个数,n 则表示样本集中总的样本个数.

$p(x_i|c_j)$表示在 c_j 类中出现样本 x_i 的概率,其计算公式如下:

$$p(x_i\mid c_j)=\frac{n_{c_j}^{x_i}}{n_{c_j}} \tag{10}$$

其中,$n_{c_j}^{x_i}$ 表示 c_j 类中 x_i 出现的次数,n_{c_j} 表示 c_j 类中的样本总数.通过训练样本可以得到 $n_{c_j}^{R_i}$,n_{c_j} ,n 和 n_{R_i} ,通过 TF-IDF 算法可以计算环境影响因子组合对样本集分类的权值.但是此时的权值是通过部分样本训练得到,离线单调,如果后面加入的数据的环境影响因子未在训练集中存在过,这样加权值会不够准备,可能会导致监控结果的错误.在这里,本文利用信息增益值与影响因子组合数据单元在类间分布的密集程度成正比的关系,根据实时传入的具体影响因子组合数据单元的信息增益来优化 TF-IDF 算法,从而对经早期数据样本训练得到的影响因子组合权值表进行更新与修正,获得实时动态的、更加精准的影响因子组合权值,使得监控结果更准确.

根据信息增益的公式得到实际监控时实时传入的具体影响因子组合数据单元的信息增益计算公式如下:

$$IG(R_i)=-\sum_{c_i\in C}\left(\frac{n_{c_i}}{n}\right)\times\log\left(\frac{n_{c_i}}{n}\right)+\frac{n_{R_i}}{n}\times\sum_{c_i\in C}\frac{n_{c_i}^{R_i}}{n_{c_i}}\times\log\left(\frac{n_{c_i}^{R_i}}{n_{c_i}}\right)+\frac{n-n_{R_i}}{n}\times\sum_{c_i\in C}\frac{n_{c_i}-n_{c_i}^{R_i}}{n-n_{R_i}}\times\log\left(\frac{n_{c_i}-n_{c_i}^{R_i}}{n-n_{R_i}}\right) \tag{11}$$

影响因子组合数据单元的信息增益是环境因子在分类间的分布信息和其分类能力的描述,且信息增益值与其分类能力成正比.影响因子的分类能力与其在类间的分布的均匀程度成反比,即:影响因子分布越不均匀,其携带的分类信息越多,分类能力越强,通过信息增益计算的信息增益值也越大.因此,信息增益值的大小反映了影响因子在类间的分类强弱,通过信息增益因子对传统的 TF-IDF 算法改进,实现权值的动态修正.基于信息增益的环境因子构造的权值的更新公式如下:

$$w_{R_i} = TF \times IDF(R_i) + IG(R_i) = (n_{c_j}^{R_i} \mid n_{c_j}) \times \log\left(\frac{n}{n_{R_i}}\right) \times IG(R_i) \tag{12}$$

最后,调用加权朴素贝叶斯分类器(4)得到监控结果.

IgS-wBSRM 相关实现算法描述如下.

算法 1 是对 IgS-wBSRM 方法的总体描述,分为初始化参数训练阶段与监控阶段.

- 训练阶段首先于第 1 行构造一个哈希表用以保存权值表;然后,从第 2 行开始循环遍历初始训练样本;第 4 行、第 5 行判断当前组合项是否已在现有权值表中初始化过;接着,第 6 行、第 7 行调用算法 2(即 *StandardDecision* 方法),利用前期定义好的 QoS 概率质量属性标准β和 QoS 属性值阈值 *QoS_Value* 对初始化训练样本进行判断,进而进行各影响因子组合的分类统计;然后,于第 8 行、第 9 行根据所得统计信息基于传统 TF-IDF 算法计算各影响因子组合对不同分类的权值;同时,于第 10 行、第 11 行计算对应分类的先验概率,从而完成影响因子权值表等参数的初始化;
- 监控阶段则定义好监控样本的计数统计值后,由第 2 行开始进入循环,每次循环完成当前所到的监控样本组合项到来后,对权值表相应影响因子组合的权值更新以及此时监控决策的判断,其中:于第 3 行首先查找已初始化完成的影响因子组合权值表,得到此时到来组合项对应的初始权值;然后,于第 5 行~第 9 行调用 *StandardDecision* 方法实时统计窗口内样本数据量,并结合滑动窗口动态调整监控范围内的样本数据;再于第 10 行、第 11 行根据实时数据,依据公式(12)对权值表进行动态更新;第 12 行计算 $p(x_i|c_j)$,进而于第 13 行依据更新后的权值以及所得 $p(x_i|c_j)$调用算法 3(即 *computeAftPro_*c_j 方法)计算此时各分类对应后验概率,再于第 14 行~第 17 行计算后验概率之比,并依据比值进行服务监控结果的决策分类.

算法 2 与算法 3 是对整体算法中的两个分部功能的实现:算法 2 对应计算实时窗口下的各分类中当前影响因子组合所占比例,以进一步实时调整参数 $n_{c_0}^{x_i}$ 以及 $n_{c_1}^{x_i}$,同时返回分类结果 $C(X)$;对于算法 3,于第 1 行通过简单比例计算 $p(c_j)$,然后于第 2 行、第 3 行计算并记录此时样本组合项对应各分类的条件概率,第 4 行~第 7 行则依据公式(4)首先判断当前窗口是否滑动,然后进一步计算当前窗口下朴素贝叶斯分类器的后验概率并返回.

算法 1. IgS-wBSRM 算法.

训练阶段.

输入:训练样本 $T=\{x_1,x_2,x_3,\ldots,x_n\}$;*QoS* 概率质量属性标准$\beta$;QoS 属性值阈值 *QoS_Value*;初始化样本循环计数值 *n*;

输出: $w_{c_j}^{R_i}$:影响因子组合权值;pro_c_j:先验概率.

1. *Create weightTable*;
2. **while** $(x_k^{R_i} \in T)$;
3. *n*++;
4. **if** $(weightedTable.contains(R_i))$; $n_{R_i}++$;
5. **else** $weightedTable.add(R_i)$ and $n_{R_i}==1$;
6. **if** $(StandardDecision(x_k^{R_i})==C_0)$; $n_{c_0}^{R_i}++$; $n_{c_0}++$;
7. **else** $n_{c_1}^{R_i}++$; $n_{c_1}++$;
8. $w_{c_0}^{R_i} = n_{c_0}^{R_i} \times 1.0/n_{c_0} \times \log(n/n_{R_i})$;
9. $w_{c_1}^{R_i} = n_{c_1}^{R_i} \times 1.0/n_{c_1} \times \log(n/n_{R_i})$;
10. $pro\_c_0 = n_{c_0}/(n_{c_0}+n_{c_1})$;
11. $pro\_c_1 = n_{c_1}/(n_{c_0}+n_{c_1})$;
12. *Initial weightTable*;

监控阶段.

输入:监控样本 $S=\{x_1,x_2,x_3,\ldots,x_n\}$;$n$:对当前监控样本总量的计数统计; $w_{c_j}^{R_i}$:影响因子组合权值表;pro_c_j:先验概率;滑动窗口大小 m;QoS 属性值阈值:QoS_Value;QoS 概率质量属性标准β;

输出:$C(X)$:监控结果;K:后验概率指标; $w_{c_j}^{R_i}$:更新后的影响因子组合权值表.

1. n=0;
2. **while** $(x_k^{R_i} \in S)$;
3. $w_{c_j}^{R_i} = getWeight(x_k^{R_i})$;
4. n++;
5. **if** (n>m);
6. **if** $(StandardDecision(x_k^{R_i}) == C_0)$; $n_{c_0}^{R_i}--$; $n_{c_0}--$;
7. **else** $n_{c_1}^{R_i}--$; $n_{c_1}--$;
8. **if** $(x_{n-m}^{R_i} \Leftarrow QoS\_Value)$; $n_{c_0}^{R_i}++$; $n_{c_0}++$;
9. **else** $n_{c_1}^{R_i}++$; $n_{c_1}++$;
10. $IG_{R_i} = computeIG(R_i)$;
11. $w_0 = updateW_{ic_0}(x_k^{R_i}), w_1 = updateW_{ic_1}(x_k^{R_i})$;
12. $p(x_i \mid c_0) = n_{c_0}^{x_i} / n_{c_0}$, $p(x_i \mid c_1) = n_{c_1}^{x_i} / n_{c_1}$;
13. $AftPro\_c_0 = computeAftPro\_c_0(x_k^{R_i}), AftPro\_c_1 = computeAftPro\_c_1(x_k^{R_i})$;
14. k=$aftPro\_c_0$/$aftPro\_c_1$;
15. **if** (k>1) 接受原假设,即服务处于正常状态
16. **else if** (k<1) 拒绝原假设,即服务失效
17. **else** (k=1) 落入中立区,无法判断服务失效与否

算法 2. $StandardDecision(x_k^{R_i})$ 算法.

输入:监控样本 $S=\{x_1,x_2,\ldots,x_n\}$;QoS 概率质量属性标准β;QoS 属性值阈值:QoS_Value;

输出:$C(X)$:分类结果.

1. **if** $(x_k^{R_i} \Leftarrow QoS\_Value)$
2. $successQoS$++;
3. c=$successQoS$×1.0/n;
4. **if** ($c \geqslant \beta$); $n_{c_0}^{x_i}++$; **return** c_0;
5. **else** $n_{c_1}^{x_i}++$; **return** c_1;

算法 3. $computeAftPro\_c_j(x_k^{R_i})$.

输入:监控样本 $S=\{x_1,x_2,\ldots,x_n\}$;滑动窗口大小 m;满足 QoS 属性值要求的样本数:$successQoS$; $w_{c_j}^{R_i}$:影响因子组合权值表;$recordPreProc_j$:保存先验概率的窗口队列;

输出:后验概率 $afterProc_j$.

1. $double\ pro\_c_j = computePro\_c_j(x_k^{R_i})$;
2. $double\ RPreProc_j = w_{c_j}^{R_i} \times \log(pow(plc0, temp) \times pow(1-plc0, 1-temp))$;
3. $recordPreProc_j.add(RPreProc_j)$;
4. if (n>m), $prePro\_c_j$=$prePro\_c_j$+$RPreProc_j$−$recordPreProc_j.get(n-m-1)$;
5. else $prePro\_c_j$=$prePro\_c_j$+$RpreProc_j$;

6. *double aftProc*$_j$=log(*pro_c*$_j$)+*preProc_c*$_j$;
7. return *afterProc*$_j$;

4 实验及结果分析

为了验证本文提出的 IgS-wBSRM 监控方法的合理性与有效性,实验分别在给定标准下的自定义模拟数据集和来自香港中文大学发布的真实世界 Web 服务质量(quality of Web service,简称 QWS)数据集[24,25]下,将本方法与文献[11,12]中提出的 wBSRM(weighted naive Bayes runtime monitoring)方法以及文献[10]中提出的 iBSRM(improved Bayes runtime monitoring)方法进行比较分析.由于 QWS 真实数据集没有确定的错误率和服务质量属性标准,因此,第 1 组实验中首先采用给定标准下的自定义模拟数据集进行实验,对本方法进行合理性验证;再利用 QWS 真实数据集进行真实数据集下的实验以验证本法的有效性与实用性.

4.1 实验数据集及环境配置

实验基于 Eclipse 开发平台,使用 Java 编程语言设计并实现所提出的方法.所涉及使用的两种不同的实验数据集及相关实验环境参数见表 4.

Table 4 Experimental data set and the main related parameters

表 4 实验环境及实验数据集相关参数

(a) 环境配置

Configuration item	Experimental environment parameter
RAM	DDR4 12G
Hard disk	5400 rpm HD
CPU	Intel Core i5-3337U 1.8GHz

(b) 数据集配置

Relevant parameter	Custom simulated data set	QWS real data set
Client location	2 fixed position of the client	Part of 339 Client information
Server location	2 fixed location on the server	Part of 5825 Server information
Data set	5000 Response time QoS value	6000 Response time QoS value of original data set

4.2 IgS-wBSRM方法的验证

4.2.1 自定义模拟数据集下的实验分析与验证

在模拟实验中,定义 QoS 需求为响应时间小于 10s 的概率不低于 50%,并采用按一定约束随机生成的 5 000 个模拟响应时间 QoS 数据中的前 3 500 个,本实验中涉及影响因子组合⟨United States,Colombia⟩及⟨United States,China⟩,其中,两影响因子组合所对应的初始权值与此后的真实数据集下的实验分析一致,皆通过真实数据集 QWS 中相同数据段的 1 000 个真实数据样本训练获得,于 1200~1700 处注入响应时间 QoS 参数值大于 10s 的错误样本大于 50%,将样本 1000~1800 区间的影响因子组合定义为⟨United States,Colombia⟩,对应有对 c_0 类的权重为 0.006 929 84,对 c_1 类权重为 0.002 031 05.再于 2000~2500 处注入响应时间 QoS 参数值大于 10s 的错误样本大于 50%,将样本 1900~2700 区间的影响因子组合定为⟨United States,China⟩,对应有对 c_0 类的权重为 0.009 516 66,对 c_1 类权重为 0.084 154 64.

首先采用不同静态滑动窗口(static sliding window)大小对 IgS-wBSRM 的分类效果进行分析的实验,除实验过程中对部分窗口大小以单独形式进行的实验外,总体上我们以 50 为步长,从 50 窗口大小开始依次(即 50,100,150,200,...)遍历进行实验.实验结果如图 3 所示,其中,纵坐标表示监控结果的分类(1 代表接受原假设,即此时被监控 Web 服务处于正常状态;–1 代表拒绝原假设,即此时被监控 Web 服务出现异常、服务失效),横坐标表示监控数据样本数量,与纵坐标平行的线代表监控决策的改变.

由图 3 可以看出:当滑动窗口越小时,方法监控分类的灵敏度越高;窗口越大时,监控分类结果的断言不变,但会出现相应的判断延迟,窗口愈大,其延迟越高.但同时也可以看到:当窗口大小越小时,监控受近期特殊数据的影响程度越大.例如:当窗口逐渐缩小甚至于缩减为 1 时,此时的分类即完全受当前唯一的特殊数据影响而进

行判断,这样监控结果显然在绝大多数情况下是与事实有悖的.故窗口过小则会导致灵敏度过高,从而降低监控分类的准确性.

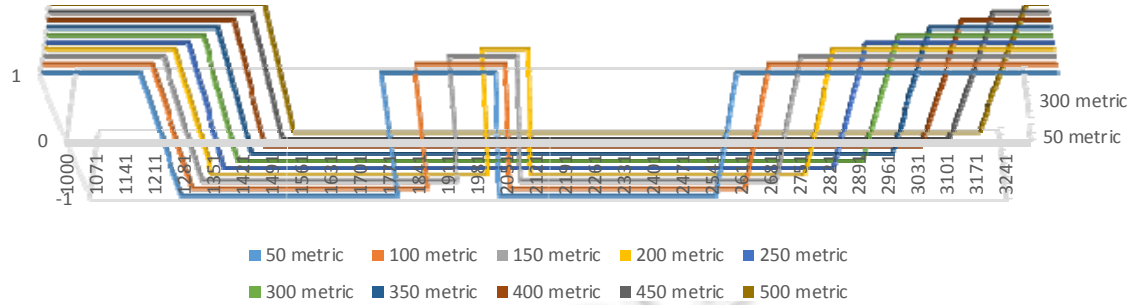

Fig.3 IgS-wBSRM under diffrent sizes of sliding windows

图 3 不同步长窗口大小下的 IgS-wBSRM

同时,观察图 3 可发现:在前一部分窗口大小(50,100,150 等)下的监控决策结果于 1700~2000 处发生了改变,这也正是小窗口下灵敏度高的体现.而在窗口逐渐增大时,该数据段中在之前窗口大小下的监控决策改变消失.这也正是因其延迟性的增大所导致,故我们取最佳窗口大小范围于上述数据段中监控决策有无改变的突变窗口大小附近(即 200~250 周围大小的窗口).而在接下来的实验中,考虑到模拟实验数据集由人为自定义,其数据的稳定性远高于真实数据集中的数据,故本文中将选取 200 步长大小的窗口进行对应真实数据集下的实验,同时,选取 250 步长大小的窗口进行接下来的模拟数据集实验.

接下来的模拟实验中,将 IgS-wBSRM 与现有基于加权朴素贝叶斯算法的分类方法(weighted naïve Bayes running monitoring,简称 wBSRM)以及基于传统贝叶斯的监控方法(improved BSRM,简称 iBSRM)进行定量分析与比较,其中,QoS 标准需求的定义不变.

如图 4 所示:对于相同的模拟数据集,IgS-wBSRM,wBSRM 以及 iBSRM 在监控开始时都表现出了一致的判断,分类结果于 1 类,即 Web 服务质量满足 QoS 属性标准,服务属于正常状态;而紧接着,在少数数据样本过后,iBSRM 便出现了服务失效的判断,而 IgS-wBSRM 方法以及 wBSRM 方法则在监控结果上保持一致,这与我们在定义模拟 QoS 数据样本集时的数据好坏定义一致.当样本数为 1 392 时,IgS-wBSRM 第 1 次检测出了服务时效,此时,可以判断是由样本从 1 200 开始注入的⟨United States,Colombia⟩影响因子组合所携带的失效信息所带来的判断,而此时,wBSRM 以及 iBSRM 皆维持原有监控决策结果未发生变化.而对于 wBSRM 未检测到服务失效,则一方面是因为其受历史累赘数据的失效信息影响,对新到来的活跃数据所携带的分类信息的不敏感性,同时也是由于其未考虑动态的影响因子组合在二分类间出现频次的变化导致其受到监控分类的类间分布偏差性的影响.而在样本数 2 163 附近,wBSRM 方法检测到的服务失效则是其连带在 2 000 样本处开始注入的错误样本的影响下产生的滞后判断现象,这也同时是对上述误判理论的有力证明.对于 1 700 错误样本注入完成后的一段区间,IgS-wBSRM 并未产生二次分类的决策改变,原因则是因为此时中部区间 300 个本身存在一定容错率的数据样本并未使得监控结果达到实际的后验概率比值决策改变标准,而直到 3 013 个样本数据处,IgS-wBSRM 也成功监控检测到了服务从失效状态迁回正常状态,与模拟 QoS 数据样本的总体预期结果一致.

表 5 对应表示的是基于 IgS-wBSRM 方法所构建的监控器在运行监控过程中的其中一次动态更新对二分类影响因子组合权值的影响,表中列举了环境影响因子组合⟨United States,China⟩,⟨United States,Colombia⟩一次更新前后的权值变化.从表中可以看出:此次更新正好对应在目前的窗口范围内的实时状态下调整以前离线不变的初始权值到了完全另外一种预期状态,这便是 IgS-wBSRM 在实际监控过程中对监控状态的调整.

上述实验结果有效地证明了 IgS-wBSRM 监控方法的合理性与优越性,在融入滑动窗口机制下引入信息增益相结合改进的 IgS-wBSRM 方法,切实考虑了二分类决策的类间分布关系的影响以及历史冗余数据对实时动态环境下的监控判断的负面影响,减少了传统基于概率统计的监控所存在的部分误判.

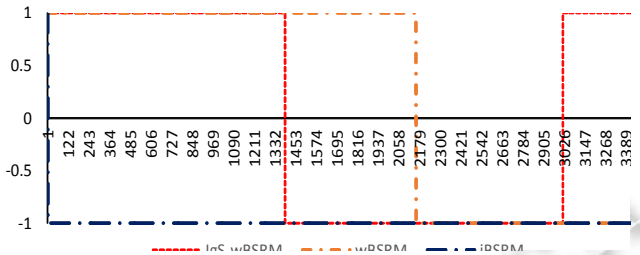

Fig.4 Monitoring results with different approaches

图 4 不同方法间的监控结果

Table 5 Experimental data set and the main related parameters

表 5 一次更新对二分类影响因子组合权值的影响

User location/Server location	⟨United States,China⟩	⟨United States,Colombia⟩
$W_{0\ old}$	0.009 516 65	0.006 929 84
$W_{1\ old}$	0.084 154 64	0.002 031 05
$W_{0\_IgS\text{-}wBSRM}$	5.08E−8	0
$W_{1\_IgS\text{-}wBSRM}$	4.19E−8	0

实验还使用与上述相同的模拟数据集,将滑动窗口机制改进于原 wBSRM,iBSRM 方法中,并与 IgS-wBSRM 进行对比分析,由 200 步长大小的窗口起,拟定 3 种不同规格的窗口大小,分别为 200,350,500,分别进行实验对比.实验监控结果如图 5~图 7 所示,可发现,不同步长窗口下的两种算法的灵敏性都有很大提升.当然,这些结果都是在模拟数据集下所产生具有一定的局限性,但这也仅是对其他基于概率统计的现有监控方法较为简单地引入滑动窗口机制以作探究,而真实情况下的数据将比模拟数据更具实际意义、更为复杂,但这样的实验结果仍是从一些方面部分性说明了有效利用实时数据相较一直使用历史冗余数据更为有效准确.

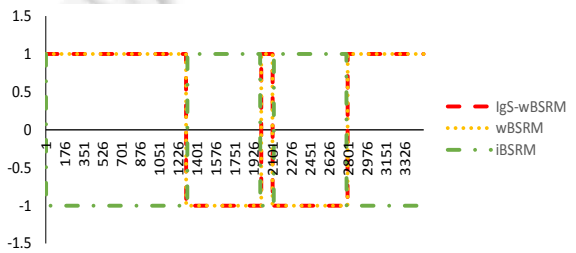

Fig.5 Result under 200 metric sliding window

图 5 以 200 为步长的窗口下监控结果

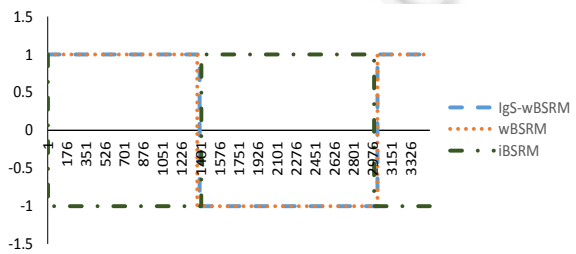

Fig.6 Result under 350 metric sliding window

图 6 以 350 为步长的窗口下监控结果

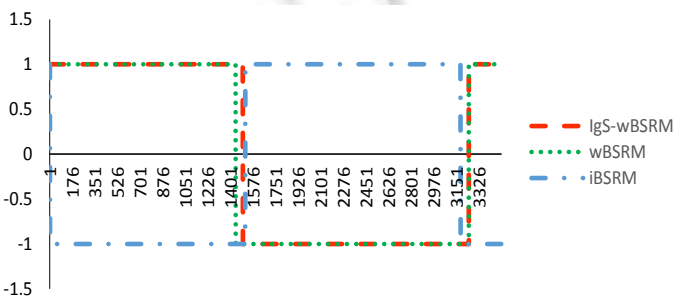

Fig.7 Monitoring result under 500 metric Sliding Window

图 7 以 500 为步长的窗口下监控结果

4.2.2 真实数据集下的实验分析

在上述模拟实验对本方法合理性的验证基础上,下面将通过 QWS 真实数据集来验证本方法的有效性与实用性.

从原始 QWS 真实数据集中提取 6 000 个样本,取前 1 000 个 QWS 样本数据作为初始化训练样本集,以此来训练得出一个初始权值表与加权分类监控器,取之后的 5 000 个数据作为监控数据集,根据之前的实验,其中,IgS-wBSRM 方法的滑动窗口大小设定为 200.

如图 8 描述的是 IgS-wBSRM,wBSRM,iBSRM 在 QWS 真实数据集下的监控结果,同样,纵坐标表示监控结果的分类(1 代表接受原假设,即此时被监控 Web 服务处于正常状态;–1 代表拒绝原假设,即此时被监控 Web 服务出现异常、服务失效),横坐标表示监控数据样本数量.其 QoS 标准为响应时间小于 10s 的概率不低于 50%.图 9 为 1 480~1 720 样本数据段的监控结果,更为细致地展现 3 种监控方法在此数据段中的分类监控结果的变化.图 10 为在上述 QoS 标准下,Web 服务 QoS 参数满足标准与不满足标准的后验概率之间的比值,当其大于 1 时,表示接受原假设,即此时被监控 Web 服务处于正常状态;–1 代表拒绝原假设,即此时被监控 Web 服务出现异常、服务失效,其变化是连续而非离散的值,故可更直观有效地分析 3 种方法其监控结果的变化趋势.

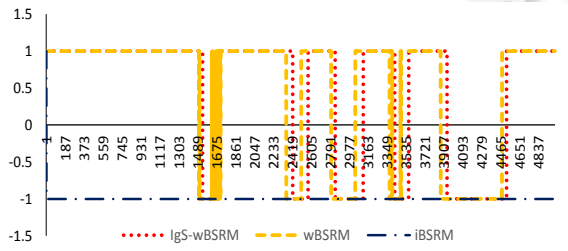

Fig.8 Monitoring results under QWS real data set

图 8 QWS 真实数据集下的监控结果

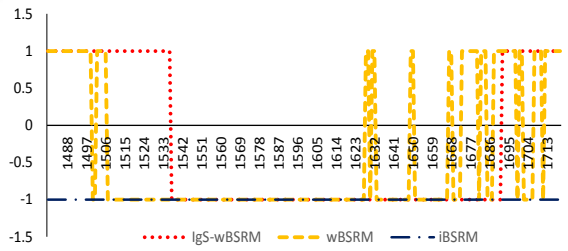

Fig.9 Results during data segment from 1 480 to 1 720

图 9 1 480~1 720 数据段监控结果

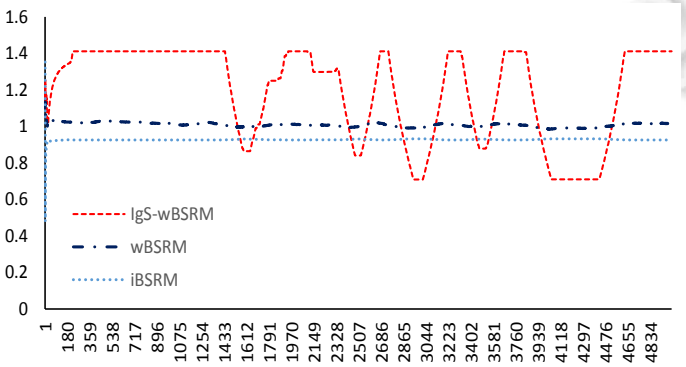

Fig.10 Ratio between posterior probabilities under QWS real data set

图 10 QWS 真实数据集下的后验概率比值

从图 8 中可以看到:IgS-wBSRM 与 wBSRM 在整体的对监控正确性的断言上保持一致;若只考虑能否有效检测出服务失效,IgS-wBSRM 与 wBSRM 都比较准确有效地检测出处于错误数据节点处的服务时效;而 iBSRM 在此时的 QoS 标准(响应时间 10s,概率标准 50%)的情况下,出现了几乎全局相逆的错误判断.同时可以宏观地看到:wBSRM 在真实数据环境中某些数据结点不断地发生着变化,因其监控结果的改变频率过快致使产生了很多甚至于重叠的噪声波段.从图 9 中可以清楚细致地看到:在 1 496,1 502,1 505 短短 10 个监控数据间,wBSRM 竟跳跃了 3 次;同样,在 1 628~1 634 的 7 个监控数据间,也是快速跳跃了 3 次判定.这样的监控分类结果显然是

与事实相悖的.为了进一步探究这种差异影响的来源,同时也是对 IgS-wBSRM 方法有效性与实用的验证,现对二次决策之间各方法满足标准与不满足标准的后验概率之间的比值作图 10.从图中可以看出:wBSRM 方法在监控过程中若前期数据使其决策结果游离于标准左右时,若遇到部分影响类间分布的数据单元,则会收到这些少数数据单元的影响,从而不断且频繁地更变其监控决策.同时,从图中可以清楚地看见:融入滑动窗口机制进而结合信息增益的动态加权算法 IgS-wBSRM 方法,因信息增益对权值不断实时动态地调整,使得其监控的后验概率之比能在维持与标准适当距离的同时,而且能在正确处有效地检测出服务的失效进行决策迅速跳转改变,很好地克服了这种缺陷,总体监控效果与模拟实验所验证的合理性保持一致.

4.2.3 时间效率分析

效率分析分为两部分,初始化权值训练效率分析和对不同算法在实时运行状态下的监控效率分析.

- 对于第 1 部分效率分析,由于 IgS-wBSRM 的权值初始化与 wBSRM 的权值训练方式相同,都采用传统的 TF-IDF 算法进行训练,故初始化权值训练阶段二者效率相同,按照文献[11]的报道,权值的训练时间是在可接受范围;
- 对于第 2 部分效率分析,下面将取真实世界数据集前 3 500 个数据样本,记录对不同 QoS 需求标准下的各监控方法完成一次全数据监控所需时间,并以此求得对应单位数据下的平均监控时间来进行对比分析.而单位数据条件下的监控运行时间,可以有效地反映监控算法的运行效率.

本实验将在表 4(a)所描述的硬件环境下运行,故所得实验结果参数仅代表本实验环境下的特定结果.不同机器间的配置性能差距将会导致实验结果的绝对数值有所偏差,但方法间的相对运行效率状况是不变的.在上述前提下,可得各方法监控运行时间在不同 QoS 需求标准下的具体情况如图 11 所示.

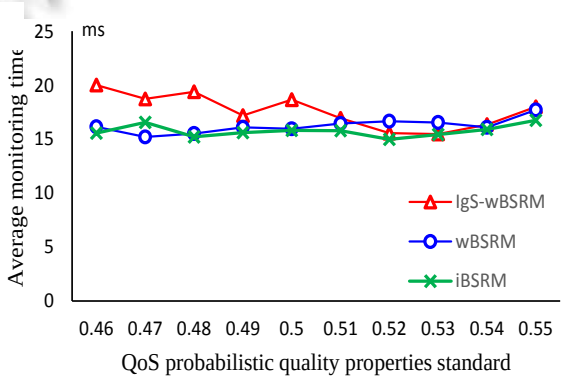

Fig.11 Average monitoring time

图 11 平均监控时间

从图 11 中可以看出:整体上,IgS-wBSRM 监控方法在运行时平均监控时间上略高于 wBSRM 以及 iBSRM 方法.这是因为 IgS-wBSRM 方法在监控的同时还会对现有权值表进行动态更新,而由于同时我们采用了滑动窗口机制,这使得方法在数据的更新统计方面一定程度地减少了样本量,缩短了部分时间,故总体来说时间复杂度并未增加太多,整体监控效果仍在理想范围内.

5 结束语

现有的 Web 服务 QoS 概率监控方法大多对动态环境下的实时监控缺乏考虑,少数考虑了多变环境因素影响的研究方法却未对监控的时效准确性以及监控分类的类间分布偏差等问题进行考虑,而诸如这些,正会导致服务监控出现监控延迟判断、二分类监控决策间噪声抖动等现象.本文给出了一种融入滑动窗口机制进而结合信息增益实现动态加权的 Web 服务监控方法 IgS-wBSRM.方法考虑到现有方法未对历史冗余数据进行处理从而导致实时数据面对历史数据基数大而不易改变决策的现状,在构造监控器时仅以初期数据进行权值等参数

训练,而后期无限期使用导致的参数过期无效性以及在利用传统 TF-IDF 算法对影响因子加权时未曾考虑过的类间分布不均现象等,并在自定义模拟数据集与真实数据集上分别与基于加权朴素贝叶斯的 wBSRM 以及基于传统贝叶斯的 iBSRM 方法进行对比实验.实验结果表明,IgS-wBSRM 在监控稳定性和准确性两个方面都优于其他两种方法.

对于未来的工作,将进一步深入探究滑动窗口大小对监控方法的影响,具体研究是否存在每个固定环境下的理想窗口大小,在此基础上,进而可以考虑一种自适应的动态监控窗口,并通过进一步的实验验证与数据分析以期达到所预期的效果.此外,由于当 QoS 需求标准达到一个极高的要求值时,无论 IgS-wBSRM 或其他方法都无法十分准确地满足监控需求.比如,某服务对于用户的请求访问响应时间在 0.1s 内的概率应该大于 99.99%,这是一个极大的概率值,通过目前现有的监控手段很难监控出结果,值得将来进一步探索,使得方法能够对更为极限的 QoS 需求标准做出准确有效地监控.最后,也计划将 IgS-wBSRM 应用到服务组合、服务动态选择等领域中[26],以提升相应领域方法的稳定性和准确性.

References:

[1] Gunter D, Tierney B, Jackson K, *et al*. Dynamic monitoring of high-performance distributed applications. In: Proc. of the 11th IEEE Int'l Symp. on High Performance Distributed Computing (HPDC-11). IEEE, 2002. 163−170.

[2] Menascé DA. QoS issues in Web services. IEEE Internet Computing, 2002,6(6):72−75.

[3] Baresi L, Guinea S. Towards dynamic monitoring of WS-BPEL processes. In: Proc. of the Int'l Conf. on Service-Oriented Computing. Springer Berlin Heidelberg, 2005. 269−282.

[4] Ran S. A model for Web services discovery with QoS. ACM SIGECOM Exchanges, 2003,4(1):1−10.

[5] Grunske L. Specification patterns for probabilistic quality properties. In: Proc. of ACM/IEEE the 30th Int'l Conf. on Software Engineering (ICSE 2008). IEEE, 2008. 31−40.

[6] Grunske L, Zhang P. Monitoring probabilistic properties. In: Proc. of the Joint Meeting of the European Software Engineering Conf. and the ACM Sigsoft Int'l Symp. on Foundations of Software Engineering. Amsterdam, 2009. 183−192.

[7] Chan K, Poernomo I, Schmidt H, *et al*. A model-oriented framework for runtime monitoring of nonfunctional properties. In: Proc. of the Int'l Conf. on Quality of Software Architectures and Software Quality, and 2nd Int'l Conf. on Software Quality. Springer-Verlag, 2005. 38−52.

[8] Grunske L. An effective sequential statistical test for probabilistic monitoring. Information & Software Technology, 2011,53(3): 190−199.

[9] Zhang P, Li W, Wan D, *et al*. Monitoring of probabilistic timed property sequence charts. Software Practice & Experience, 2011, 41(7):841−866.

[10] Zhu Y, Xu M, Zhang P, *et al*. Bayesian probabilistic monitor: A new and efficient probabilistic monitoring approach based on Bayesian statistics. In: Proc. of the Int'l Conf. on Quality Software. 2013. 45−54.

[11] Zhuang Y, Zhang PC, Li WR, *et al*. Web service QoS monitoring approach sensing to environmental factors. Ruan Jian Xue Bao/Journal of Software, 2016,27(8):1978−1992 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4850.htm [doi: 10.13328/j.cnki.jos.004850]

[12] Zhang P, Zhuang Y, Leung H, *et al*. A novel QoS monitoring approach sensitive to environmental factors. In: Proc. of the IEEE Int'l Conf. on Web Services. 2015. 145−152.

[13] Zeng L, Lei H, Chang H. Monitoring the QoS for Web services. In: Proc. of the Int'l Conf. on Service-Oriented Computing. Springer-Verlag, 2007. 132−144.

[14] Radovanovic S, Nemet N, Cetkovic M, *et al*. Cloud-Based framework for QoS monitoring and provisioning in consumer devices. 2013.

[15] Coppolino L, D'Antonio S, Romano L, *et al*. Effective QoS monitoring in large scale social networks. In: Zavoral F, *et al*. eds. Proc. of the Intelligent Distributed Computing VII, Studies in Computational Intelligence 511. Springer International Publishing, 2014. 249−259.

[16] Michlmayr A, Rosenberg F, Leitner P, *et al*. Comprehensive QoS monitoring of Web services and event-based SLA violation detection. 2009.

[17] Raimondi F, Skene J, Emmerich W. Efficient online monitoring of Web-service SLAs. In: Proc. of the ACM Sigsoft Int'l Symp. on Foundations of Software Engineering. Atlanta, 2008. 170−180.

[18] Sammapun U, Lee I, Sokolsky O, *et al*. Statistical runtime checking of probabilistic properties. In: Proc. of the Runtime Verification. Berlin, Heidelberg: Springer-Verlag, 2007. 164−175.

[19] Zhang P, Li B, Grunske L. Timed property sequence chart. Journal of Systems & Software, 2010,83(3):371−390.

[20] Lewis DD. Naive (Bayes) at forty: The independence assumption in information retrieval. In: Proc. of the European Conf. on Machine Learning. Berlin, Heidelberg: Springer-Verlag, 1998. 4−15.

[21] Wang GY, Yu H, Yang DC. Decision table reduction based on conditional information entropy. Chinese Journal of Computers, 2002,25(7):759−766 (in Chinese with English abstract).

[22] Kent JT. Information gain and a general measure of correlation. Biometrika, 1983,70(1):163−173.

[23] Roelleke T, Wang J. TF-IDF uncovered: A study of theories and probabilities. In: Proc. of the 31st Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. ACM Press, 2008. 435−442.

[24] Zheng ZB, Zhang YL, Lyu MR. Distributed QoS evaluation for real-world Web services. In: Proc. of the 8th Int'l Conf. on Web Services (ICWS 2010). Miami, 2010. 83−90.

[25] Zhang YL, Zheng ZB, Lyu MR. Exploring latent features for memory-based QoS prediction in cloud computing. In: Proc. of the 30th IEEE Symp. on Reliable Distributed Systems (SRDS 2011). Madrid, 2011.

[26] Wang SG, Sun QB, Yang FC. Web service dynamic selection by the decomposition of global QoS constraints. Ruan Jian Xue Bao/ Journal of Software, 2011,22(7):1426−1439 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3842.htm [doi: 10.3724/SP.J.1001.2011.03842]

附中文参考文献:

[11] 庄媛,张鹏程,李雯睿,等.一种环境因素敏感的 Web Service QoS 监控方法.软件学报,2016,27(8):1978−1992. http://www.jos.org.cn/1000-9825/4850.htm [doi: 10.13328/j.cnki.jos.004850]

[21] 王国胤,于洪,杨大春.基于条件信息熵的决策表约简.计算机学报,2002,25(7):759−766.

[26] 王尚广,孙其博,杨放春.基于全局 QoS 约束分解的 Web 服务动态选择.软件学报,2011,22(7):1426−1439. http://www.jos.org.cn/1000-9825/3842.htm [doi: 10.3724/SP.J.1001.2011.03842]

何志鹏(1995－),男,湖北仙桃人,硕士生,主要研究领域为 Web 服务监控.

吉顺慧(1987－),女,博士,讲师,CCF 专业会员,主要研究领域为软件建模、分析、测试与验证.

张鹏程(1981－),男,博士,副教授,CCF 高级会员,主要研究领域为软件建模、分析和验证技术.

李雯睿(1981－),女,博士,副教授,CCF 高级会员,主要研究领域为服务计算.

江艳(1992－),女,硕士,主要研究领域为服务质量监控.