软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2016,27(8):1978−1992 [doi: 10.13328/j.cnki.jos.004850]

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

一种环境因素敏感的 Web Service QoS 监控方法*

庄 媛[1], 张鹏程[1], 李雯睿[2], 冯 钧[1], 朱跃龙[1]

[1](河海大学 计算机与信息学院,江苏 南京 211100)

[2](可信云计算与大数据分析重点实验室(南京晓庄学院),江苏 南京 211171)

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

摘 要: 面向服务系统的执行能力依赖第三方提供的服务,在复杂多变的网络环境中,这种依赖会带来服务质量(QoS)的不确定性.而 QoS 是衡量第三方服务质量的重要标准,因此,有效监控 QoS 是对 Web 服务实现质量控制的必要过程.现有监控方法都未考虑环境因素的影响,比如服务器位置、用户使用服务的位置和使用时间段负载等,而这些影响在实际监控中是存在的,忽略环境因素会导致监控结果与实际结果有悖.针对这一问题,提出了一种基于加权朴素贝叶斯算法 wBSRM(weightednaive Bayes running monitoring)的 Web Service QoS 监控方法.受机器学习分类方法的启发,通过 TF-IDF(term frequency-inverse document frequency)算法计算环境因素的影响,通过对部分样本进行学习,构建加权朴素贝叶斯分类器.将监控结果分类,满足 QoS 标准为 c_0,不满足 QoS 标准为 c_1,监控时调用分类器得到 c_0 和 c_1 的后验概率之比,对比值进行分析,可得监控结果满足 QoS 属性标准、不满足 QoS 属性标准和不能判断这 3 种情况.在网络开源数据以及随机数据集上的实验结果表明:利用 TF-IDF 算法能够准确地估算环境因子权值,通过加权朴素贝叶斯分类器,能够更好地监控 QoS,效率显著优于现有方法.

关键词: 服务质量;影响因子;TF-IDF 算法;加权朴素贝叶斯分类器;监控

中图法分类号: TP311

中文引用格式: 庄媛,张鹏程,李雯睿,冯钧,朱跃龙.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

英文引用格式: Zhuang Y, Zhang PC, Li WR, Feng J, Zhu YL. Web service QoS monitoring approach sensing to environmental factors. Ruan Jian Xue Bao/Journal of Software, 2016,27(8):1978−1992 (in Chinese). http://www.jos.org.cn/1000-9825/4850.htm

Web Service QoS Monitoring Approach Sensing to Environmental Factors

ZHUANG Yuan[1], ZHANG Peng-Cheng[1], LI Wen-Rui[2], FENG Jun[1], ZHU Yue-Long[1]

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

[2](Key Laboratory of Trusted Cloud Computing and Big Data Analysis (Nanjing Xiaozhuang University), Nanjing 211171, China)

Abstract: The execution capacity of service-oriented system relies on the third-party services. However, such reliance would result in uncertainties in consideration of the complex and changeable network environment. Hence, runtime monitoring technique is required for service-oriented system. Effective monitoring technique towards Web QoS, which is an important measure of third-party service quality, is necessary to ensure quality control on Web service. Several monitoring approaches have been proposed, however none of them consider the influences of environment including the position of server and user usage, and the load at runtime. Ignoring these influences, which

* 基金项目: 国家自然科学基金(61572171, 61202097, 61202136, 61370091); 高等学校博士学科点专项科研基金(20120094120009); 江苏省自然科学基金(BK20130852); 中央高校基本科研业务费(B15020191)

Foundation item: National Natural Science Foundation of China (61572171, 61202097, 61202136, 61370091); Research Fund for the Doctoral Program of Higher Education of China (20120094120009); Natural Science Foundation of Jiangsu Province of China (BK20130852); Fundamental Research Funds for the Central Universities of China (B15020191)

收稿时间: 2014-11-28; 修改时间: 2015-02-16; 采用时间: 2015-04-22; jos 在线出版时间: 2016-03-25

CNKI 网络优先出版: 2016-03-25 16:07:10, http://www.cnki.net/kcms/detail/11.2560.TP.20160325.1607.001.html

exist among the real-time monitoring process, may cause monitoring approaches to produce wrong results. To solve this problem, this paper proposes a new environment sensitive Web QoS monitoring approach, called wBSRM (weighted Bayes runtime monitoring), based on weighted naive Bayes and TF-IDF (Term Frequency-Inverse Document Frequency). The proposed approach is inspired by machine learning classification algorithm, and measures influence of environment factor by TF-IDF algorithm. It constructs weighted naïve Bayes classifier by learning part of samples to classify monitoring results. The results that meet QoS standard are classified as c_0, and those that do not meet is classified as c_1. Classifier can output ratio between posterior probability of c_0 and c_1, and the analysis can lead to three monitoring results including c_0, c_1 or inconclusive. Experiments are conducted based on both public network data set and randomly generated data set. The results demonstrate that this approach is better than previous approaches by accurately calculating environment factor weight with TF-IDF algorithm and weighted naïve Bayes classifier.

Key words: quality of service; impact factor; TF-IDF algorithm; weighted naïve Bayesian classifier; monitor

近年来,无论是企业内部还是企业外部,面向服务的体系结构(service oriented architecture,简称 SOA)都得到了越来越广泛的应用[1].面向服务设计框架作为一种应用最为广泛的软件设计方法,对软件质量的精确度要求越来越高,面向服务的体系结构将应用程序的不同服务通过良好的接口联系起来,优点不言而喻,缺点是单个软件的失效可能会影响上下文调用的软件.因此,为了保证应用程序的顺利运行,方便设计者选择组成程序的服务,要求设计的服务能够达到一定的服务质量(quality of service,简称 QoS)需求,如可靠性、可用性、安全性等[2].这些 QoS 需求在动态的网络运行环境中应达到一定概率阈值,可用概率质量属性来描述,通过监控 Web 服务是否满足该概率阈值在运行时评价服务的 QoS[3,4].故 QoS 监控是确保软件能够及时发现失效的必不可少的步骤,持续监控有益于在面向服务系统中,迅速而有效地寻找更优的 Web 服务,提高服务质量[4].

大部分 QoS 需求可由概率质量属性来表示[2],如服务可靠性需求可描述为“该服务 1 年内的平均无故障运行时间为 95%”,响应时间需求可描述为“对该服务发出调用请求后,在 8s 内响应的概率为 80%”.所以,当前的 QoS 监控方法借助于针对概率质量属性的监控方法,其中,运用比较广泛的是 Grunske 和 Zhang 提出的 ProMo (probabilistic monitor)方法[5,6].该方法的思想是:基于假设检验理论 SPRT(sequential probability ratio test)[7],先对总体的特征作出某种假设;然后,通过抽样研究的统计推理,推断出接受或拒绝该假设.其基于小概率反证法思想,即,小概率事件(P<0.01 或 P<0.05)在一次实验中基本上不会发生,先提出假设(检验假设 H_0),再用适当的统计方法确定假设成立的可能性大小:如果可能性小,则认为假设不成立;如果可能性大,则认为假设成立.虽然经典假设检验是目前 QoS 监控中使用较为广泛的统计学方法,但仍存在若干不能避免的缺陷:首先,对于固定水平检验,需要先给定显著性水平α,计算原假设的拒绝域,但是α究竟多大比较精确,并未给出具体的标准,而根据不同的显著性水平有时会得到相反的结论;其次,通过α给定的拒绝域来检验有时并不有效.有研究发现:一个以 10^{10} 的α拒绝 H_0 的经典结论,当 n 充分大时,此 H_0 的后验概率逐渐趋近于 1,该结论被称为 Lindley 悖论[8].因此,当样本容量不断增大时,假设检验基本失效.

贝叶斯算法则比较直截了当,直接计算出原假设 H_0 和备择假设 H_1 的后验概率α_0 和α_1,并计算后验概率比来判断检测结果.然而,现有的贝叶斯算法没有解决贝叶斯算法本身的缺陷,即贝叶斯的独立假设并不适合所有情况.因为在 QoS 监控中,运行环境和上下文的波动使软件的运行环境产生非常大的不确定性[9],每个样本的采集都具有各自的“身份证”,即样本的监控时间、客户端位置、样本服务器的属性等,这些因素决定了样本对整体决策的影响[10].此时,朴素贝叶斯算法所带来的误差在 QoS 监控中可能会导致实际情况与监控结果不符的误差.使用加权朴素贝叶斯算法并根据环境影响因素计算权值,可以科学地减小误差,因此,结合环境因素影响的监控技术更贴近实际,可看做是保证 Web QoS 的基础[11,12].

环境因素的影响已经在 QoS 预测[9]、动态 QoS 组合[11]、QoS 度量[12]和 QoS 选择[13]等方法中考虑过.然而,现有的 QoS 监控方法还没有考虑到环境因素.针对这一问题,本文提出一种基于加权朴素贝叶斯[14]的 QoS 监控方法(weighted naive Bayes runtime monitoring,简称 wBSRM).该方法首次提出通过训练的方法量化环境因素的影响,并受网页关键字搜索的启发,创新地运用了在搜索领域应用广泛的 TF-IDF 算法实现环境因素影响的量化,通过加权贝叶斯思想将环境因素的影响与 QoS 监控结合起来,并通过一系列实验证明了该方法紧密结合实际情况,科学地减小了误差.该方法分为训练和监控两个阶段:训练时,将样本满足 QoS 属性标准设为α_0 类,不满

足 QoS 属性标准设为α_1类,通过学习,对部分样本进行计算,得到加权朴素贝叶斯分类器;监控时,对每个样本调用朴素贝叶斯分类器,得到样本满足 QoS 属性标准的 c_0类以及不满足 QoS 属性标准的 c_1类的后验概率之比,对比值进行分析,可得样本集是否满足 QoS 属性或者不能判断.其中,权值的计算引入了 TF-IDF 算法[15].该算法的加入使 wBSRM 可以考虑环境因素的影响,对部分样本调用 TF-IDF 算法,将得到不同环境因素对分类的权值表,在监控时,根据样本所提供的环境因素信息调用权值表,作为该样本调用加权朴素贝叶斯分类器时所加的权值,减小监控误差,提高监控速度.

1 预备知识

1.1 加权朴素贝叶斯分类器

分类是数据挖掘中一个重要的问题,分类算法的核心部分是构造分类器,现已有众多分类方法,朴素贝叶斯因其计算高效、精确度高,并具有坚实的理论基础而得到了广泛的应用.朴素贝叶斯的思想基础[14]是这样的:对于给出的待分类样本集,求解在此样本集出现的条件下各个类别出现的概率,其中最大概率的类别被认为此待分类样本集的类别.令 $C=\{c_0,c_1\}$是预定义的类别集,$X=\{x_1,x_2,x_3,\ldots,x_n\}$是样本向量,根据贝叶斯公式:

$$P(c_i \mid X)=\frac{P(c_i)P(X \mid c_i)}{P(X)} \tag{1}$$

为了简化 $P(X|c_i)$的估计,朴素贝叶斯假定:当 X 属于类 c_i 时,X 中的元素 x_k 的取值和 x_1 的取值是相互独立的,这样对于给定的类 c_i 的条件概率就可以分解为

$$P(X \mid c_i)=\prod_{k=1}^{n}P(x_k \mid c_i) \tag{2}$$

将公式(2)带入公式(1)中,得到:

$$P(c_i \mid X)=\frac{P(c_i)\prod_{k=1}^{n}P(x_k \mid c_i)}{P(X)} \tag{3}$$

实际上,由于 $P(X)$对于所有的类 c_i 都是一样的,所以上式中,分子的最大值的类别就是 X 的分类结果.由于分类过程是基于朴素贝叶斯假设来进行的,所以这种方法称为朴素贝叶斯分类方法.

$$C(X)=\arg\max_{c_i \in C}\{P(c_i)P(X \mid c_i)\} \tag{4}$$

朴素贝叶斯理论认为所有的样本数据对分类的重要性是一致的,然而事实却并非如此,因此,可以根据不同的样本数据的分类重要性赋给样本不同的权值.得到加权朴素贝叶斯公式:

$$C(X)=\arg\max_{c_i \in C}\{P(c_i)P^{w_i}(X \mid c_i)\} \tag{5}$$

整个方法分为准备阶段和分类阶段.准备阶段:对样本数据进行训练,得到先验条件概率 $P(x_k|c_i)$和实际概率 $P(c_i)$以及样本权值;分类阶段:计算后验概率,返回使后验概率最大的类和样本.

1.2 二项分布的经验贝叶斯估计

在成败型实验中,很多参数估计问题都可归结为二项分布的参数估计.在 QoS 监控中,样本满足 QoS 属性与不满足 QoS 属性标准,可看做成败型检验[16].设事件 A 为样本满足 QoS 属性需求,事件 A 出现的概率为 $\theta(0\leqslant\theta\leqslant1)$,那么在 n 次独立实验中,A 出现了 x 次(x=0,1,2,…,n)的概率为

$$P(x \mid \theta)=\binom{n}{x}\theta^x(1-\theta)^{n-x} \tag{6}$$

贝叶斯方法把θ作为随机变量,赋予它一个先验分布 $c(\theta)$,结合实现样本,应用贝叶斯公式来对θ进行估计,本文使用的方法是经验贝叶斯估计(EB 估计),这种方法把经典的方法和贝叶斯方法结合起来对θ进行估计.由于无先验信息,我们将θ看作在$(0,\lambda)$上的均匀分布函数,设θ的先验分布为

$$c(\theta)=\begin{cases}1/\lambda, & 0<\lambda<1\\ 0, & \text{其他}\end{cases} \tag{7}$$

那么 x 的边缘分布:

$$P_G(x)=\int c(\theta)p(x\mid\theta)\mathrm{d}\theta=\int_0^{\lambda}\frac{1}{\lambda}\binom{n}{x}\theta^x(1-\theta)^{n-x}\mathrm{d}\theta \tag{8}$$

$$E(x)=\int xP_G(x)\mathrm{d}x=\int_0^{\lambda}\frac{n\theta}{\lambda}\mathrm{d}\theta=\frac{n}{2}\lambda \tag{9}$$

如果有经验样本 $x_1,x_2,x_3,\ldots,x_n$,可令 $E(x)=\frac{1}{m}\sum_1^m x_i=\bar{x}$,则 $\frac{n}{2}\lambda=\bar{x}$,由于 $0\leqslant\theta\leqslant1$,因此$\lambda\leqslant1$,因此可取:

$$\lambda=\min\left\{1,\frac{2\bar{x}}{n}\right\} \tag{10}$$

λ确定后,在平方损失下求θ的贝叶斯估计 $\hat{\theta}$.根据贝叶斯公式,由公式(6)和公式(7)得θ的后验概率密度为

$$h(\theta\mid x)=P(x\mid\theta)c(\theta)\Big/\int c(\theta)P(x\mid\theta)\mathrm{d}\theta.$$

其核为$\theta^x(1-\theta)^{n-x}$,故

$$\hat{\theta}=E(\theta\mid x)=\int_0^{\lambda}\theta h(\theta\mid x)\mathrm{d}\theta=\int_0^{\lambda}\theta^{x+1}(1-\theta)^{n-x}\mathrm{d}\theta\Big/\int_0^{\lambda}\theta^x(1-\theta)^{h-x}\mathrm{d}\theta \tag{11}$$

由于 x 是正整数,根据公式(11)易算积分,且易证明.该式中,$\hat{\theta}$ 是关于λ的递增函数.这说明,在现实样本相同的情况下,λ越大,得到的 $\hat{\theta}$ 也越大.而λ与经验样本均值成正比,所以,经验样本的均值越大,$\hat{\theta}$ 越大.也就是说,在 EB 估算中,$\hat{\theta}$ 不仅与现实样本 x 有关,而且与先验信息有关.

1.3 TF-IDF算法

TF-IDF(term frequency-inverse document frequency)[15]是一种用于资讯检索与资讯探勘的常用加权技术,是如何度量网页和查询的相关性的关键技术.TF 表示查询的词在单个网页中出现的词频,词频越高,就代表查询和该网页的相关度越高;IDF 表示查询的词在所有网页中出现的词频,该值越高,表示查询越难以得到结果.一些学者们也发现,所谓的 IDF 的概念,就是一个特定条件下关键词的概率分布交叉熵.总的来说,一个词预测主题的能力越强,权重越大;反之,权重越小.

这个理论同理可用在我们的加权监控中.度量不同的影响因子权值组合对 Web 服务成功或者失败的影响是困难的,原因在于,监控本身就是一个概率统计的问题,无法将某一单个样本看作标量,这个相关度是很难求得的.实际上,我们只要关心影响因子组合对样本集的分类影响就可以了,通过训练一系列样本集,可以得到每个样本的加入对整个样本集分类的影响.所以,借鉴检索的加权技术方法(TF-IDF),定义如下:

定义 1(影响因子权值). 对分类的影响,随着它在类中出现的频率增大而增大,随着它在总的样本集中出现的概率的增大而减小.

设定 w_R 代表影响因子组合 R 对分类的权值,那么它的值可以由下面的公式求得:

$$w_R=TF\times IDF(R)=(n_{c_i}^R/N_{c_i})\times\log\left(\frac{N}{n_R}\right) \tag{12}$$

$n_{c_i}^R$ 表示影响因子组合 R 中属于类别 c_i 的数量,N_{c_i} 表示类别 c_i 的个数,N 表示样本的整体数,n_R 表示影响因子组合为 R 的样本数.

2 一种考虑环境因素的加权朴素贝叶斯监控方法 wBSRM

2.1 加权贝叶斯监控方法概况

图 1 为结合 TF-IDF 算法和加权朴素贝叶斯分类器方法的监控方法 wBSRM 的总体结构图,在图中可以清晰地看到,用户使用服务的环境各有不同,比如用户使用台式机或者手提电脑、使用无线网络或有线网络、所请求服务的服务器在地球的哪个位置,请求服务的时间段可能有不同的负载变化,经验证,这些不同环境都会对服务的质量有一定的影响.而概率监控过程将监控结果看成(0,1)分布的,并且将每个样本对监控结果的影响看

作是相同的,即如果响应时间要求小于 0.3s,那么样本的响应时间无论是 0.5s 还是 0.7s,其监控结果都是 0.显而易见,其准确性很差,响应时间 0.7s 的样本的监控结果虽然是 0,但是对服务失效的影响更加严重.举一个例子,如果点击鼠标之后,Web 服务 5s 还没有反应,用户很可能直接关闭服务,这对服务来说已经失效了.但是运用原有的监控系统是不能监控出失效的,因此,我们提出一种加权朴素贝叶斯监控方法.

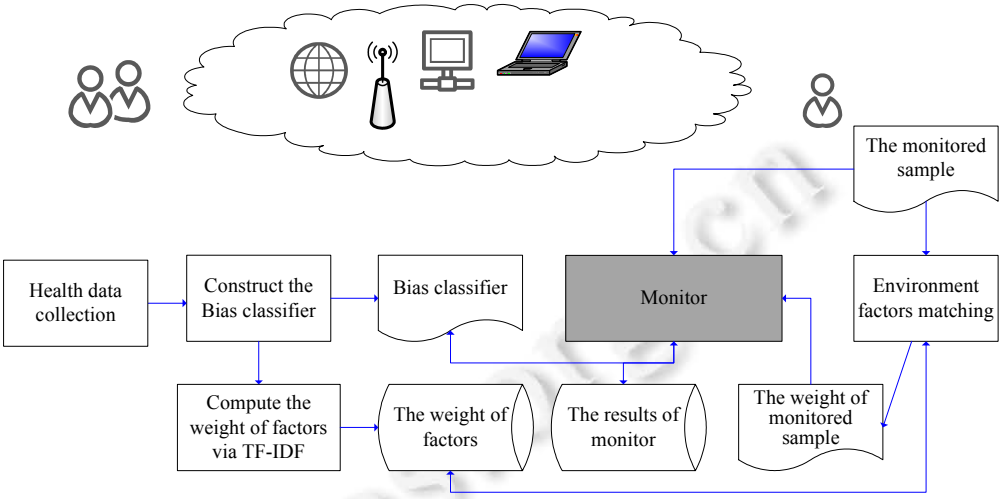


Fig.1 wBSRM architecture overview

图 1 wBSRM 总体结构图

对图 1 主要模块功能解释如下:

- Health data collection:去掉具有缺失数据的样本,对样本进行充分的离散化,合理的离散化能够减少误差,在使用较少样本时得到比较合理的先验信息,提高监控精确度.
- Construct the Bias classifier:通过对二项分布的经验估计可得伯努利分布的概率,用来构造朴素贝叶斯分类器.
- Compute the weight of factors via TF-IDF:TF-IDF 算法被用来构造加权朴素贝叶斯分类器,我们首先通过朴素贝叶斯算法对 Web 服务进行监控,然后运用 TF-IDF 算法计算影响因子权值.
- Environment factors matching:提取样本的环境因子,在权值库中进行匹配,得到样本的权值.
- Monitor:通过样本的 0-1 值和权值,调用朴素贝叶斯分类器,得到监控结果储存数据库中.

由于朴素贝叶斯的独立假设的缺陷,与我们将样本的 0-1 序列中相同的样本值对监控的影响一致这种方法相似,都没有考虑样本的权值,我们采用更精准的分类方法,即加权朴素贝叶斯方法.在我们的监控中,环境因素是影响服务质量的重要因素,也是可以量化样本权值的信息,为了更清晰地解释环境因素,将其分为 3 类:

- 用户的角度:用户所在地、使用服务时间、采用的不同网络、设备参数以及配置文件等.
- 服务器角度:服务器所在位置、计算机复杂度、系统的资源(CPU、RAM、硬盘和 I/O 等).
- 环境的角度:服务器的负载、网路性能等.

现有方法如 *K*-means 在估算环境因素对 QoS 影响时,分别求出各类因素的影响,再根据多层聚类的方式得出 QoS 值[9].类似地,我们也可以使用聚类的方式得到影响因子的权值.然而,计算得到的单个因素的影响因子的值始终有着一定的误差,只能无限接近实际值.因此,为了减小误差,我们把这些影响因素的集合定义为影响因子组合.通过建立数量庞大的数据库,得到较为精准的影响因子权值表.对于不同服务,录入对其影响较大的因子.其中,服务器的负载受各个时间段的影响,如节假日和休息时间会使网络负载过大,所以可以通过合理分割时间段来求个时间段的影响因子对服务的影响.监控的过程记录数据,可以根据不同服务的需求定期对权值表进行修改.

2.2 加权贝叶斯监控方法实现

定义样本满足 QoS 属性标准为 c_0 类,不满足 QoS 属性标准为 c_1 类,对部分样本训练,得到加权朴素贝叶斯分类器,所加权值由样本所在的环境因素决定,通过 TF-IDF 算法得到不同环境对该样本影响分类的重要性,将这一重要性设置为该样本的权值.监控时,对样本集调用加权朴素贝叶斯分类器,得出样本满足 QoS 属性、不满足 QoS 属性、样本不能判断监控结果这 3 种情况.具体方法如下.

令 $C=\{c_0,c_1\}$ 是预定义的类别集,满足 QoS 属性为 c_0 类,不满足 QoS 属性为 c_1 类,$X=\{x_1,x_2,x_3,\ldots,x_n\}$ 是样本向量,$x_k\in(0,1)$,$x_k=1$ 表示该样本满足 QoS 属性,$x_k=0$ 表示该样本不满足 QoS 属性.例如,QoS 标准为服务响应时间小于 0.3s 的概率大于 85%,$x_k=1$ 表示本次监控的样本响应时间小于 0.3s,$x_k=1$ 表示响应时间大于 0.3s,样本 X 属于类别 c_i 的概率可以由后验概率 $P(c_i|X)$ 表示.贝叶斯分类器确定样本集的类别的依据是通过估计后验概率 $P(c_i|X)$ 来实现的,朴素贝叶斯分类器将后验概率 $P(c_i|X)$ 较大的类别作为样本所分到的类别.然而 $P(c_i|X)$ 很难直接求得,必须从训练数据中进行估计,通常直接估计比较难,这里,我们使用贝叶斯公式(3)进行计算.

公式(3)中,$P(c_i)$ 是样本中,某样本的加入使样本集的点估计可靠度属于类 c_i 的概率,对于类别 c_i 的概率通常取样本集的最大似然估计作为它们的估计值,可以由下式表达:

$$P(c_i)=\frac{m_i}{|x|} \tag{13}$$

其中,m_i 表示样本中使样本集点估计可靠度属于类 c_i 的个数,$|X|$ 是样本集中的样本数.

$P(X|c_i)$ 若已知,我们就可以方便地得到一个最优的分类结果.但是我们并不知道确切的分布,$P(X|c_i)$ 的估计比较困难,因为 X 是一个 n 维向量,而 n 的取值数量级很大,为了简化 $P(X|c_i)$ 的估计,假设当 X 属于类 c_i 时,X 中的元素 x_k 与 x_1 的取值是相互独立的,这样,样本 X 对于给定类 c_i 的条件概率就可以分解为

$$P(X\mid c_i)=\prod_{k=1}^{n}P(x_k\mid c_i) \tag{14}$$

由于样本的先验概率 $P(X|c_i)$ 未知,先验概率又是得到最优分类的必要信息,但同时,由于不知道样本的确切分布,这种方法不能直接运用.伯努利模型经常用在朴素贝叶斯方法的实现中,在 QoS 属性监控中,将样本集看做一个二值向量,$X=\{x_1,x_2,x_3,\ldots,x_n\}$,$x_k\in(0,1)$,$k\in\{0,1,\ldots,n\}$,$x_k=1$ 表示样本满足 QoS 属性;反之表示样本不满足 QoS 属性,符合伯努利分布.此外,贝塔分布可作为贝叶斯分布的共轭先验分布函数,取适当的值,可以令贝塔函数无限接近伯努利分布[16].因此,我们把伯努利分布作为 $P(x_k|c_i)$ 的分布.令 θ_i 表示 $P(x_k=1|c_i)$,表示满足 QoS 属性并属于类 c_i 的样本概率,则样本 x_k 的先验条件概率为

$$P(x_k\mid c_i)=\theta_i^{x_k}(1-\theta_i)^{1-x_k}=\left(\frac{\theta_i}{1-\theta_i}\right)^{x_k}(1-\theta_i) \tag{15}$$

二项独立模型假定:对于给定的类 c_i,样本是否满足 QoS 属性是相互独立的.所以,样本 X 可以看做是 n 重独立的伯努利实验.对于给定的类别 c_i,样本集的先验概率计算可得:

$$P(X\mid c_i)=\prod_{k=1}^{n}\left(\frac{\theta_i}{1-\theta_i}\right)^{x_k}(1-\theta_i)=\arg\left\{\sum_{k=1}^{n}\log(1-\theta_i)+\sum_{k=1}^{n}x_k\log\left(\frac{\theta_i}{1-\theta_i}\right)\right\} \tag{16}$$

由以上公式可以看出,尽管模型中我们考虑了样本满足 QoS 属性和不满足 QoS 属性的情况,但是对分类起作用的实际上还是 $x_k=1$ 的样本.也就是说,二项独立模型是通过样本集 X 中出现满足 QoS 属性的样本来判断它的类别的.然而 θ 未知,因此我们采用经验贝叶斯估计(EB 估计)[17].训练阶段,我们将训练样本集均匀划分若干阶段,每一个阶段都对样本的成功率做了一次测试,每次测试 n 个样本,成功的样本数分别为 $y_1,y_2,y_3,\ldots,y_m,y$.其中,$y_1,y_2,y_3,\ldots,y_m$ 看成是经验样本,y 看成是现实样本.应用公式(11),就可以得到最后阶段的成功率 $\hat{\theta}$.其中,成功的样本数的意义并不是响应时间小于 3s,对 $P(X|c_0)$ 而言,是指阶段样本属于 c_0 的样本子集中出现的响应时间小于 0.3s 的概率;对 $P(X|c_1)$ 而言,是指阶段样本属于 c_1 的样本子集中出现的响应时间小于 0.3s 的概率.这样,通过训练,我们就可以得到用来二项分布中的 θ 值,$\theta\in(\theta_0,\theta_1)$,$\theta_0$ 为关于 c_0 的二项分布估计值,θ_1 同理为 c_0 的二项分布估计值.

$c(x)$表示朴素贝叶斯分类器,其值域表示监控结果的集合,$c(x)=\{c_0,c_1\}$,c_0 表示样本集满足 QoS 属性,c_1 表示样本集不满足 QoS 属性,R 表示样本的影响因子组合,$w(R)$表示影响因子组合使样本对分类影响的权值,T 表示训练集,S 表示测试集.训练集 T 与测试集 S 都可以看做是从一个未知分布 D 中独立同分布采样得到.影响因子加权的朴素贝叶斯监控模型实现的任务是:根据训练集 T 得到一个加权朴素贝叶斯分类器 $c(x)$,$c(x)$在测试集 S 上进行 QoS 属性监控.

wBSRM 受贝叶斯分类算法的启发,在贝叶斯算法的基础上引入了权值,该权值通过 TF-IDF 算法计算得到,更适合实际的 Web 服务质量监控.取 x_k 表示第 k 个样本的值,$y(x_k)$判断样本值是否满足 QoS 属性:若满足,则 $y(x_k)$取值为 1;不满足,则 $y(x_k)$取值为 0.对于每个测试示例 x_k,具有示例的属性 R, $w_{c_i}^R$ 表示影响因子组合 R 对类别 c_i 的权重,c_0 表示接受假设一类,c_1 表示拒绝假设一类.训练阶段,训练样本的先验概率 θ_i 和权值表 $w(R)$,得到先验概率函数 $P(X|c_i)$,进而可得贝叶斯分类器 $C(X)=\arg\max_{c_i\in C}\{P(c_i)P^{w_i}(X\mid c_i)\}$,其中,$w_i$ 表示根据样本的环境因子查表 $w(R)$得到的对 c_0 类和 c_1 类的权值,带入公式(10),得到:

$$C(X)=\arg\max_{c_i\in C}\left\{P(c_i)+\sum_{k=1}^{n}w_{c_i}^R\times\left\{\log(1-\theta_i)+temp\times\log\left(\frac{\theta_i}{1-\theta_i}\right)\right\}\right\} \tag{17}$$

- 训练阶段

将训练样本 T 均匀分成 e 等份进行成败型检验,每份样本数量为 d,由公式(10)得到λ值.λ确定后,调用公式(12)得到每类的先验概率分布中的θ_i 值,调用公式(17)可得先验概率函数 $P(X|c_i)$;同时,对训练样本应用 TF-IDF 算法,得到影响因子权值表.到此,加权朴素贝叶斯分类器所需要的先验条件都得到了解答,即,加权朴素贝叶斯分类器(18)可以在监控中得到调用.

- 监控阶段:

读取测试集 S 内的样本 $x_k^{R_i}$,R_i 表示样本 x_k 所具有的影响因子组合.调用影响因子权值表 $W_{R\_c_0}(x_k^{R_i})$ 和 $W_{R\_c_1}(x_k^{R_i})$,分别得到 R_i 对两类后验概率的权值.调用朴素贝叶斯分类器得到两类的后验概率中的分子 P_c_0 以及 P_c_1.由于后验概率分母一直可变,可得后验概率之比 k,判断 k 值得监控结果.

算法 1. wBSRM.

- 训练阶段

输入:T:训练数据;R:影响因子组合.

输出: $w_{c_i}^R$:影响因子组合的权值;θ_i:先验概率;$C(X)$朴素贝叶斯分类器.

//首先从训练数据中计算出先验概率θ_i

(1) for $x_R\in T$

(2) $\theta_0=E(\theta_0\mid x)=\int_0^{\lambda}\theta_0 h(\theta_0\mid x)\mathrm{d}\theta_0=\int_0^{\lambda}\theta_0^{x+1}(1-\theta_0)^{n-x}\mathrm{d}\theta_0\Big/\int_0^{\lambda}\theta_0^{x}(1-\theta_0)^{h-x}\mathrm{d}\theta_0$

$\theta_1=E(\theta_1\mid x)=\int_0^{\lambda}\theta_1 h(\theta_1\mid x)\mathrm{d}\theta_1=\int_0^{\lambda}\theta_1^{x+1}(1-\theta_1)^{n-x}\mathrm{d}\theta_1\Big/\int_0^{\lambda}\theta_1^{x}(1-\theta_1)^{h-x}\mathrm{d}\theta_1$

//计算影响因子组合对不同假设的权值

(3) for $x_k^{R_i}\in T$

(4) if ($check(R_i)$==1) then n_{R_i} ++;

(5) else creat R_i and n_{R_i} =1;

(6) if ($standard(x_k^{R_i})$==1) then $n_{c_0}^{R_i}$ ++, n_{c_0} ++;

(7) else $n_{c_1}^{R_i}$ ++, n_{c_1} ++;

(8) $w_{c_0}^R=\sum_{k=1}^{n}y_{c_0\&\&R}(x_k)\times 1.0\Big/\sum_{k=1}^{n}y_{c_0}(x_k)\times Math.\log(n/n_R)$

$w_{c_1}^R=\sum_{k=1}^{n}y_{c_1\&\&R}(x_k)\times 1.0\Big/\sum_{k=1}^{n}y_{c_1}(x_k)\times Math.\log(n/n_R)$

- 监控阶段

输入:S:训练数据;W_R:影响因子组合权值表;θ_i:先验概率.

输出:$C(X)$:监控结果.

//通过查找 W_R 库得到每个样本的权值

(1) $R_i=get(x_k^{R_i})$;

(2) $w_0=computeW_0(x_k^{R_i})$, $w_1=computeW_1(x_k^{R_i})$;

//调用朴素贝叶斯分类器

(3) for $x_k^{R_i} \in S$;

(4) $P\_c_0=computeAftPro\_c_0(x_k^{R_i})$, $P\_c_1=computeAftPro\_c_1(x_k^{R_i})$;

(5) $K=\dfrac{P\_c_0}{P\_c_1}$;

//根据 k 值得出监控结论

(6) int $m=decision(k)$;

(7) if m>1 返回接受假设结论;

else if m<–1 返回拒接假设结论;

else m=1 返回无法判断继续监控的结论;

3 实 验

3.1 实验环境配置

本文通过实验来模拟监控环境并验证 wBSRM 的有效性,实验环境为一台宏基 Intel Pentium G2030 CPU/4G RAM,采用 Java 语言实现本文所提出的方法,并在两种不同的数据集上进行实验.数据集见表 1.

Table 1 Main experimental parameters under different data sets

表 1 不同数据集下的主要实验参数

实验参数	随机数据集	QWS 数据集	备注
样本数量	1 600	2 100	–
影响因子组合数	2	23	前者影响因子位置固定
错误率	大于 15%	未知	–

数据集 1 采用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(http://www.datatang.com/data/15939)[18],该数据集包括 339 个用户以及 5 825 个真实世界的服务,提供了响应时间以及吞吐量的量化数据.数据含有服务器位置以及用户位置,满足我们的实验需求.真实的数据有助于我们观察环境因素对监控结果的影响,为设计数据集 2 的影响因子权值提供参考.

数据集 2 是按照一定约束随机生成的数据集,该数据集采用注入错误的方式,宏观控制监控的实际结果和环境因素的位置,对 wBSRM 进行有效性检测,在 1 000~1 200 个样本之间注入 15%的响应时间大于 3s 的错误样本,在 iSPRT(improved sequential probability ratio test)[19]无法做出判断的样本处,分别标记若干个样本对两类权值不同的影响因子,测试环境因素对真实结果的影响.

对数据进行预处理,预处理包括离散化以及过滤数据,离散化体现在提取影响因子权值过程中采用随机提取的方式,训练影响因子权值使用 TF-IDF 算法,删去影响因子权值小于 0.001 的影响因子组合,过滤数据即过滤掉数据集中无效的数据,如响应时间为负值等.

3.2 实验结果与分析

本文主要针对的是动态 QoS 监控下,采集的监控样本具有非软件本身的一些影响因素,这些因素会导致同一个样本对监控结果的影响不一致.与相关算法比较,定性地分析了本文方法 wBSRM 的优势.为了进一步验证

wBSRM 的有效性,本实验将 wBSRM 与改进的基于 SPRT 的监控方法(iSPRT)[19]以及具有代表性的基于传统贝叶斯的监控方法(improved BSRM,简称 iBSRM)[20]进行定量的实验比较.由于现有监控方法没有考虑环境因子的影响,因此我们对数据集 1 采用实际数据具体分析的方式进行比较,对数据集 2 采用控制变量法定性分析 3 种方法的结果.

3.2.1 证明环境因子对监控影响

为了证明不同环境因素对 QoS 监控的影响巨大,我们对具有用户 IP 地址、服务器地址和不同时间段的数据进行观察和分析,发现除了服务本身的因素外,这些环境因素也会对监控数据产生影响.数据集通过将一天划分 141 个时间段,分别在同一时间段使用 142 台不同分布的计算机测试 4 532 个网络开源服务站点的响应时间,通过对数据的提取和分析,得到同时段不同用户对不同地点的同一个服务进行请求的响应时间.如图 2 所示,水平坐标面代表具有 50 台不同分布的计算机以及 80 个服务站点的对应点,纵坐标表示对应组合在同一时间段的响应时间.而同一服务地点和同一用户地点在不同时间段请求的响应时间,如图 3 所示,横坐标轴表示时间段,纵坐标表示对同用户位置同服务位置不同时间段的响应时间.

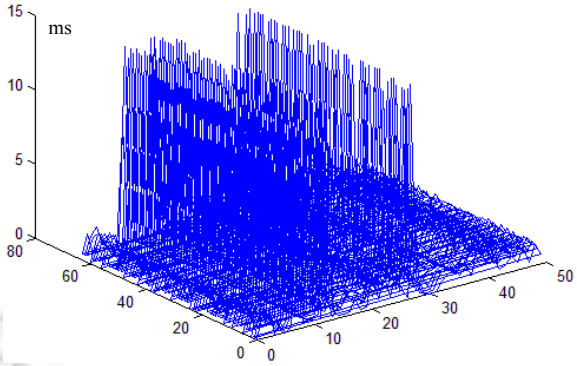


Fig.2 Response time on condition of different user position, server position and same period

图 2 同时段,不同用户位置、不同服务位置的响应时间

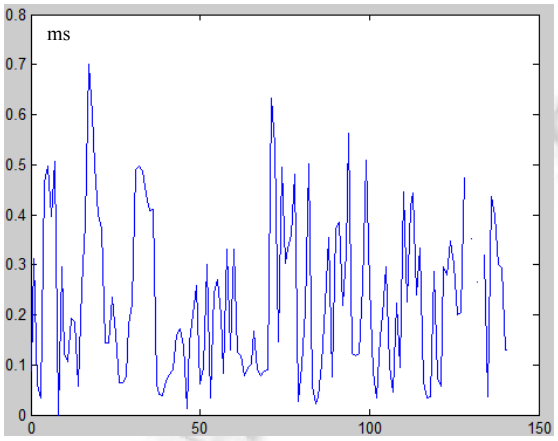


Fig.3 Response time on condition of same user position, server position and different period

图 3 不同时段,同用户位置、同服务位置的响应时间

在同一时段,服务器所在位置的序列号在 10,14,30 时响应时间特别长,几乎达到 15s;而不同时段同用户位置同服务位置的响应时间的极值达到 0.7s.此种情况可能是服务器接近崩溃,用户所在地的网络负载过大,或者服务器和用户所在地通信出现问题,种种原因都可能导致 Web 服务失效.而 0-1 分布的原则使此时监控的样本

仅仅为 0,无论是基于 SPRT 的算法还是基于传统朴素贝叶斯的方法都不能迅速地判断出失效,只能靠在该环境影响下 0 样本的数量增大来做出正确判断.这不仅增大了样本量并延迟了监控时间,更重要的是,如果动态监控不能在用户发现失效之前监控到错误,那么很可能来不及弥补,造成损失.为了更好地度量环境因子的权值,我们把这些影响服务 QoS 属性的因素的组合定义为影响因子组合,这些影响因子组合对监控结果的影响可能是真实结果不满足标准,而监控结果却满足标准.

3.2.2 wBSRM 和其他监控方法的结果比较

第 1 组实验采用真实数据集,测试了本文模型 wBSRM 和基于传统贝叶斯的 iBSRM 方法以及基于传统假设检验 SPRT 方法在不同 QoS 属性标准下的监控结果.

从数据集中提取 2 000 个数据进行训练,得到朴素贝叶斯分类器以及影响因子权值表,提取剩下的 3 000 个数据构建检验数据集,后验概率比大于 1 代表监控结果落在 c_0 类,软件符合 QoS 概率标准;后验概率比小于 1 代表监控结果落在 c_1 类,软件不符合 QoS 概率标准.

图 4 表示在 QoS 需求描述为“响应时间小于 8s 的概率是 0.36 和 0.37”时,wBSRM,iBSRM,iSPRT 的监控结果.因为该实验采用实际的数据集,因此 QoS 实际值较低.经过实验,选择 0.36 和 0.37 两个标准,能够较直观地得到监控结果曲线.若监控结果满足 QoS 属性标准,则监控结果为 1;否则为–1;无法判断为 0.垂直线表示监控状态的改变.图 5 为 QoS 属性标准为 0.37 时的软件的服务质量满足 QoS 属性标准与不满足 QoS 属性标准的后验概率之比,比值大于 1 表示满足 QoS 属性标准,小于 1 表示不满足 QoS 属性标准,等于 1 表示不能判断.该图对应图 5 中 QoS 属性标准为 0.37 时的 wBSRM 监控结果曲线.

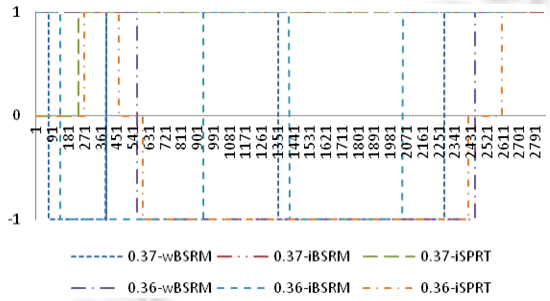


Fig.4 Results of monitoring

图 4 监控结果

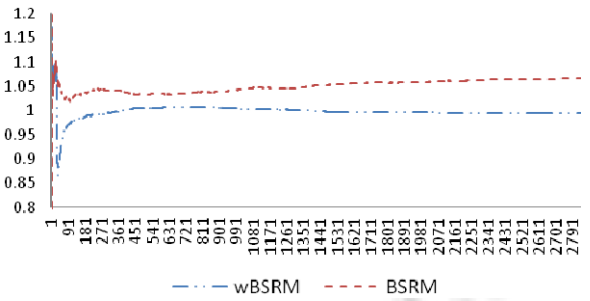
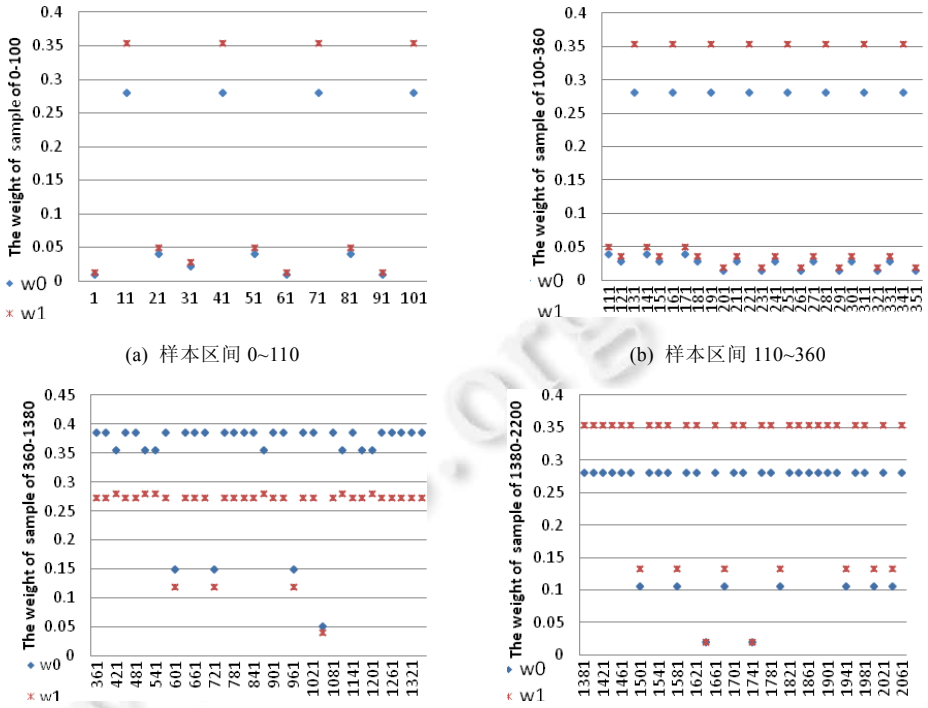


Fig.5 Ratio of posterior probability of wBSRM and iBSRM

图 5 wBSRM 和 iBSRM 的后验概率比

从图 4 可以看出,wBSRM 与 iBSRM 在监控开始的时候结果一致,而 wBSRM 能够更快地检测到服务的失效;iSPRT 在开始时无法监控出结果,并且监控结果所用样本较 iBSRM 多.这是因为贝叶斯相对于传统的假设检验更适合小样本的检测.当 QoS 属性为 0.36 时,wBSRM 在 137 检测到服务失效,而 iBSRM 在 569 检测到服务失效,且 wBSRM 在 iBSRM 判断出服务失效时,能够检测出服务可被接受.当 QoS 属性标准为 0.37 时,iBSRM 不能判断出服务出现错误,wBSRM 在样本数为 72~192 时检测到失效.而 iSPRT 的监控结果基本与 iBSRM 结果一致,iBSRM 对服务质量的改变与之相比更加敏感.为了进一步证明影响因子权值对监控结果的影响,我们对 wBSRM 方法中二次决策改变之间的样本对两种类别的权值进行研究,结果如图 6、图 7 和表 2 所示,图 6 为样本对应的权值,w_0 表示该样本对 c_0 类的权值,w_1 表示样本对 c_1 类的权值.图 7 为样本在不同阶段的影响因子数量.表 2 表示图 7 对应的影响因子对两种分类的具体权值.监控初始时,后验概率比基本一致,但随着样本量的增加,wBSRM 考虑到环境因子的影响,在样本数为 1~100(监控结果一致)以及样本数在 100~350(wBSRM 检测到服务出现错误)时,如图 6(a)和图 6(b)所示,样本影响因子对服务失效类别的权值更大的数量较多.在监控开始阶段,wBSRM 与 iBSRM 的监控结果基本一致.可见,此时权值所占的比重并没有完全影响监控结果,需要一定的累计才能够影响监控的决策.可得权值与决策基本相符,虽然也有影响因子的权值大小与决策有悖,但是在图 6 中

可以看到,与决策相符的影响因子组合的数量较多,例如在 450~1 300 的样本之间,监控结果为 1.然而,这期间的样本有对服务失效类别的权值更大的影响因子组合,见表 2,Finland/Italy 服务成功类别的权值为 0.006 112,对失败类别的权值为 0.009 021 0.然而在图 7 中可以看到,具有 Finland/Italy 影响因子组合的样本仅为 71,占该样本区间的样本总量约 0.076.



(a) 样本区间 0~110 (b) 样本区间 110~360

(c) 样本区间 361~1 380 (d) 样本区间 1 381~2 100

Fig.6 Weights of different sample for classification

图 6 不同样本对分类的权值

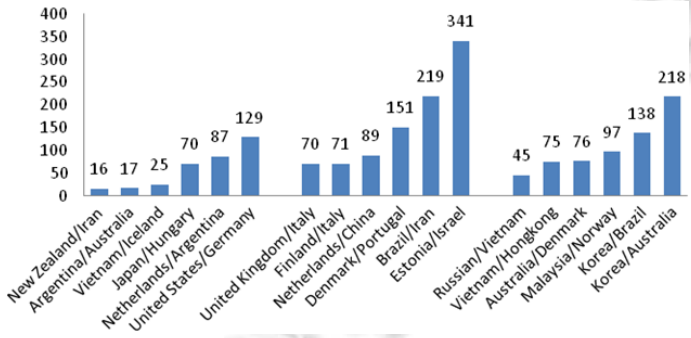


Fig.7 Number of impact factors between the decisions change

图 7 决策改变间的影响因子组合数量

Table 2 Compositional impact factors on the weights of two classification

表 2 影响因子组合对二分类的影响权值

ws/us	New Zealand/ Iran	Argentina/ Australia	Vietnam/ Iceland	Japan/ Hungary	Netherlands/ Argentina	United States/ Germany
w_0	0.014 354	0.009 826	0.028 011	0.021 273	0.039 338	0.280 364
w_1	0.017 844	0.012 233	0.035 926	0.027 093	0.049 761	0.354 168
ws/us	UnitedKingdom/ Italy	Finland/ Italy	Netherlands/ China	Denmark/ Portugal	Brazil/ Iran	Estonia/ Israel
w_0	0.149 391	0.006 112	0.049 761	0.105 012	0.354 377	0.385 129
w_1	0.118 108	0.009 021 0	0.039 338	0.132 547	0.286 714	0.271 963
ws/us	Russian/ Vietnam	Vietnam/ Hongkong	Australia/ Denmark	Malaysia/ Norway	Korea/ Brazil	Korea/ Australia
w_0	0.020 207	0.003 843 1	0.028 061	0.015 768	0.240 394	0.105 012
w_1	0.019 634	0.004 778 2	0.034 939	0.019 634	0.344 168	0.357 172

第 2 组实验采用一定约束生成的随机数据集进行测试.在第 1 组实验中,由于数据是真实值,导致数据的实际监控结果没有一个确凿的答案,无论是 iSPRT 还是 iBSRM 都存在一些误差.将本文所述模型与本身存在误差的方法进行比较虽然能够在一定程度上证明方法的可行性,但还是不够完善的,因此,我们设计具有注入错误的数据集进一步证明 wBSRM 的可用性,并在数据集中加入环境因素,进一步证明 wBSRM 的合理性.具体数据集特征如下:

QoS 需求描述为“响应时间小于 3.8s 的概率大于 0.85”,在 900~1200 个样本之间注入响应时间大于 3.8s 的错误样本数大于 15%,在 160 个样本处注入若干 0,使 iSPRT 第 1 次出现无法判断区间,将样本 60~180 区间的影响因子组合定为 United States/Germany,此时,对类 c_1 的权值为 0.354 168,大于对 c_0 类的权值 0.280 364.在 450 个样本处注入若干 0,使 iSPRT 第 2 次出现无法判断区域,将样本 300~520 区间的影响因子组合定义为 Brazil /Iran,此时,对 c_0 类的权值大于 c_1 类的权值.在样本 900~1200 处将影响因子组合再一次定义为 United States/ Germany,其余样本的影响因子组合认为对二类权值一致为 1.

实验结果如图 8 所示,横坐标代表样本个数,纵坐标代表监控结果.

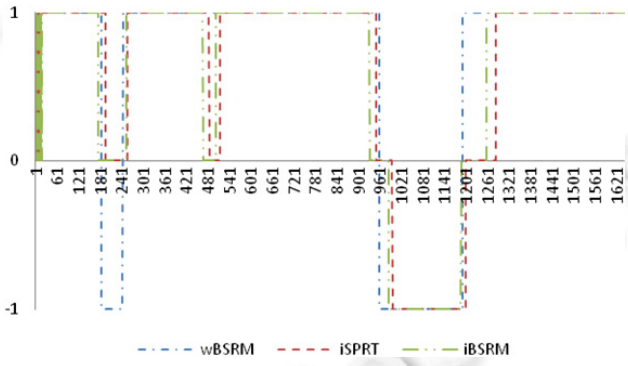


Fig.8 Result of monitoring run

图 8 监控结果

在样本数为 150~240 时,iSPRT 和 iBSRM 第 1 次出现无法判断时,wBSRM 在 186~244 样本之间得到服务失效判断.此时,可以推断是从样本量 60 开始注入的 United States/Germany 影响因子带来判断.在样本数为 470~520 时,iSPRT 和 iBSRM 又一次出现无法判断的情况,wBSRM 判断满足 QoS 属性标准,导致此判断有两种可能:一是实际上 wBSRM 无法判断的几率很小,因为后验概率比约等于 1 的概率就很小;二是影响因子组合对结果的影响,无论哪种可能,wBSRM 至少给出一个判断.事实上,在动态监控中,即使监控结果为无法判断,服务也不会被弃用,需要更多样本来判断也就意味着需要继续使用 Web 服务.所以,此时 wBSRM 的监控结果为满足 QoS 属性虽然与 iSPRT 和 iBSRM 不同,也不影响服务在实际中的使用.在样本数在 900~1200 之间时,wBSRM 在样

本量为 960 时最早检测出服务出错,iBSRM 以样本量 986 次之,iSPRT 在样本量为 997 时最后检测出错误.在样本恢复满足 QoS 属性标准后,wBSRM 率先得到满足 QoS 属性标准的结论.我们分析,源于 wBSRM 无法判断的几率很小的原因,这不影响服务在实际中的使用.但是由于在 900~1200 的样本数之间影响因子组合对 c_1 的权值更大,wBSRM 判断改变的速度较 iSPRT 和 iBSRM 判断改变的速度慢.

3.2.3 计算时间(computing time,简称 CT)

计算时间是指算法生成分类器所需的时间以及监控的平均计算时间,该指标反映了算法的效率.值得一提的是,由于训练分类器的算法运行时间是其他方法所没有的,训练主要目的是为了计算所有影响因子组合的权值.表 3 给出了实验使用数据集训练的时间.分析表 3 可知,训练时间较短可以接受.本实验会进一步与高效的 iSPRT 方法以及基于传统贝叶斯算法的 iBSRM 比较实际监控的平均时间.

Table 3 Time of all the impact factors of c_0 and c_1 under different QoS standards (ms)

表 3 不同 QoS 标准下计算全部影响因子对 c_0 以及 c_1 所需时间 (ms)

计算时间	QoS 标准					
	0.37	0.38	0.39	0.40	0.41	0.42
计算对 c_0 的权值	2.13	3.21	23.08	5.77	7.21	5.03
计算对 c_1 的权值	4.18	2.55	3.15	18.13	1.84	2.28
总计算时间	6.31	5.76	26.23	23.89	9.50	7.31

计算时间的测量采取实际数据集中 QoS 属性小于 8sec 的概率大于 0.37 这一标准进行测量,运行 2 000 个数据记录其总体时间,取样本平均时间为计算时间 CT.如图 9 所示,横坐标代表不同的 QoS 属性标准,纵坐标代表每个样本的平均计算时间.从图中可以看到,iSPRT 所需时间最多,wBSRM 所需的时间略高于 iBSRM.这是因为 wBSRM 需要调用影响因子组合权值库查找每个样本对分类的权值,但是由于在本文中的实际影响因子组合数略小,并采取了哈希表的方式存储,所以 wBSRM 的计算时间高出 iBSRM 并不多.

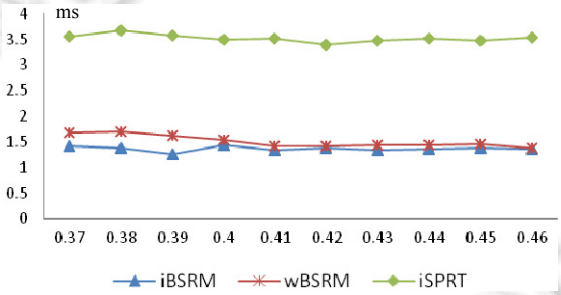


Fig.9 Monitoring time of actual data set

图 9 实际数据集监控时间

4 相关工作

监控系统用来观察系统运行时软件的服务质量,用一系列算法判断服务是否满足服务质量标准.现如今已提出一些针对概率质量属性的监控方法.如果将这些方法应用在 Web 服务领域,不仅都没有考虑环境对监控的影响,而且也存在一些问题.Chan 等人[21]基于.net 应用程序提供了一个平台监控 PCTL 属性,通过计算成功样本与总样本的比值来直接计算概率,然后与预定的概率标准比较,得出结论.这种基于估算的方法缺乏统计分析论证,与真实结果可能出现较大误差.Lee 等人[22]在 Mac(monitoring and checking)框架上进行运行时监控,该框架对 MEDL(meta-event definition language)[23]进行了概率扩展.该方法首先统计成功样本与总体样本的比值,然后利用假设检验根据一定的置信水平来进行评估.Grunske 等人[5]提出了概率属性监控框架 ProMo,针对运行时监控引入概率逻辑 CSL^{Mon},CSL^{Mon} 是 CSL(continuous stochastic logic)的子集,用来定义概率属性.ProMo 使用假设检验技术,在显著性水平 α 和 $1-\beta$ 下,验证 CSL^{Mon} 公式正确性.Zhang 等人扩展了 PSC(property sequence chart)[24]

为 PTPSC(probabilistic timed property sequence chart)[6]来表达概率属性,并定义相应的形式语义以及语义翻译器来自动生成结合 TBA(timed Büchi automata)和 SPRT 程序的概率属性监控器,所有以 PTPSC 语言定义的特性都可以由该监控器监控分析.但是上述假设检验方法不支持连续监控.Grunske[19]改进了 SPRT 方法,采用回退的方法并且复用之前的监控信息实现了动态监控.但是 SPRT 方法在当实际概率值与属性需求概率值相近时,监控结果大量落入中立区,使得方法失效;且 SPRT 要求系统在整个生命周期中,属性需求的概率必须为常量,但实际上该值常常根据客户端需求变更,一旦变更,以前的监控结果就不可以再复用,必须重新开始.如果客户端需求频繁变更,那么该方法执行效率低下,且无法实现连续监控.Zhu 等人[20]提出了一种基于贝叶斯统计的概率监控方法 BaProMon,该方法利用两个算法 BSRM 和 iBSRM,根据 Web 服务 QoS 运行时监控信息和贝叶斯统计原理计算贝叶斯因子,进行假设检验.该方法受先验分布影响,选择一个合适的先验概率是难题.另外,以上方法中都没有考虑环境因素对 QoS 属性监控的影响,而在实际的监控中,环境因素是存在的,可能会造成错误的监控结果,这是本文方法 wBSRM 与之前方法的最大区别和改进之处.

5 结束语

现有的监控方法没有考虑环境因素的影响,导致监控结果与事实相违背,出现误差、延误判断时间等问题.针对这一问题,本文给出了基于 TF-IDF 算法和加权朴素贝叶斯分类器算法的环境因子敏感的 Web 服务 QoS 方法 wBSRM.该方法考虑监控的样本所属的不同环境对二类标准的权值,在真实数据集和模拟数据集上,分别对比基于传统贝叶斯的 iBSRM 方法以及基于经典假设检验的 iSPRT 方法,实验结果表明,wBSRM 在性能没有明显降低的情况下,效率明显优于其他 2 种方法.

在未来的工作中,将重点考虑以下几个问题:一是监控时影响因子权值表的修正,将通过进一步的数据分析和实验得到合理的更新影响因子权值表的时机区间,使更新在消耗较少的资源情况下能保证数据正确;二是对于监控样本与训练样本的环境因子不同时的加权方法,设想使用相似性算法将未知影响因子和已知影响因子进行匹配;三是根据影响因素定义不同的服务质量标准进行监控.以上几个方面皆可使 wBSRM 得到更加广泛的运用,使监控结果更加准确.

References:

[1] Arsanjani A, Endre, M, Ang J, Chua S, Comte P, Krogdahl P. Patterns: Service-Oriented Architecture and Web Services. IBM Corporation, Int'l Technical Support Organization, 2004.

[2] 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. [doi: 10.1145/1368088.1368094]

[3] Shao J, Deng F, Wang QX. A model-based software system monitoring approach. Journal of Computer Research and Development, 2010,47(7):1176−1183 (in Chinese with English abstract).

[4] Zeng L, Lei H, Chang H. Monitoring the QoS for Web Services. Berlin, Heidelberg: Springer-Verlag, 2007. [doi: 10.1007/978-3-540-74974-5_11]

[5] Grunske L, Zhang P. Monitoring probabilistic properties. In: Proc. of the 7th Joint Meeting of the European Software Engineering Conf. and the ACM SIGSOFT Symp. on the Foundations of Software Engineering. ACM Press, 2009. 183−192. [doi: 10.1145/1595696.1595724]

[6] Zhang P, Li W, Wan D, Grunske L. Monitoring of probabilistic timed property sequence charts. Software: Practice and Experience, 2011,41(7):841−866. [doi: 10.1002/spe.1038]

[7] Wald A. Sequential tests of statistical hypotheses. The Annals of Mathematical Statistics, 1945,16(2):117−186. [doi: 10.1214/aoms/1177731118]

[8] Breitung K. The Lindley paradox, information and generalized functions. In: Proc. of the 3rd Int'l Symp. on Uncertainty Modeling and Analysis and Annual Conf. of the North American Fuzzy Information Processing Society (ISUMA-NAFIPS'95). IEEE, 1995. 720−723. [doi: 10.1109/ISUMA.1995.527783]

[9] Silic M, Delac G, Srbljic S. Prediction of atomic Web services reliability based on *k*-means clustering. In: Proc. of 2013 the 9th Joint Meeting on Foundations of Software Engineering. ACM Press, 2013. 70−80. [doi: 10.1145/2491411.2491424]

[10] Hossain MS. QoS in Web service-based collaborative multimedia environment. In: Proc. of 2014 the 16th Int'l Conf. on Advanced Communication Technology (ICACT). IEEE, 2014. 881−884. [doi: 10.1109/ICACT.2014.6779087]

[11] Mabrouk NB, Beauche S, Kuznetsova E, Georgantas N, Issarny V. QoS-Aware service composition in dynamic service oriented environments. In: Proc. of the Middleware 2009. Berlin, Heidelberg: Springer-Verlag, 2009. 123−142. [doi: 10.1007/978-3-642-10445-9_7]

[12] Ma Y, Wang SG, Sun QB, Yang FC. Web service quality metric algorithm employing objective and subjective weight. Ruan Jian Xue Bao/Journal of Software, 2014,25(11):2473−2485 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4508.htm [doi: 10.13328/j.cnki.jos.004508]

[13] Xin M, Jiang T, Zhang R. A QoS constraints location-based services selection model and algorithm under mobile internet environment. Int'l Journal of Grid & Distributed Computing, 2014,7(2):124−129. [doi: 10.14257/ijgdc.2014.7.2.12]

[14] Box GEP, Tiao GC. Bayesian Inference in Statistical Analysis. John Wiley & Sons, 2011.

[15] Aizawa A. An information-theoretic perspective of tf-idf measures. Information Processing & Management, 2003,39(1):45−65. [doi: 10.1016/S0306-4573(02)00021-3]

[16] Jeffreys H. The Theory of Probability. Oxford University Press, 1998.

[17] Feder M, Weinstein E. Parameter estimation of superimposed signals using the EM algorithm. IEEE Trans. on Acoustics, Speech and Signal Processing, 1988,36(4):477−489. [doi: 10.1109/29.1552]

[18] 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. [doi: 10.1109/ICWS.2010.10]

[19] Grunske L. An effective sequential statistical test for probabilistic monitoring. Information and Software Technology, 2011,53(3):190−199. [doi: 10.1016/j.infsof.2010.10.003]

[20] Zhu Y, Xu M, Zhang P, Li W, Leung H. Bayesian probabilistic monitor: A new and efficient probabilistic monitoring approach based on Bayesian statistics. In: Proc. of 2013 the 13th Int'l Conf. on Quality Software (QSIC). IEEE, 2013. 45−54. [doi: 10.1109/QSIC.2013.55]

[21] Chan K, Poernomo I, Schmidt H, Jayaputera J. A model-oriented framework for runtime monitoring of nonfunctional properties. In: Proc. of the Quality of Software Architectures and Software Quality. Berlin, Heidelberg: Springer-Verlag, 2005. 38−52. [doi: 10.1007/11558569_5]

[22] Lee I, Sokolsky O, Regehr J. Statistical runtime checking of probabilistic properties. In: Proc. of the Runtime Verification. Berlin, Heidelberg: Springer-Verlag, 2007. 164−175. [doi: 10.1007/978-3-540-77395-5_14]

[23] Kim MZ, Viswanathan M, Kannan S, Lee I, Sokolsky O. Java-MaC: A run-time assurance approach for Java programs. Formal Approachs in System Design, 2004,24(2):129−155. [doi: 10.1023/B:FORM.0000017719.43755.7c]

[24] Zhang P, Li B, Grunske L. Timed property sequence chart. Journal of Systems and Software, 2010,83(3):371−390. [doi: 10.1016/j.jss.2009.09.013]

附中文参考文献:

[3] 邵津,邓芳,王千祥.一种基于模型的软件系统监测方法.计算机研究与发展,2010,47(7):1176−1183.

[12] 马友,王尚广,孙其博,杨放春.一种综合考虑主客观权重的 Web 服务 QoS 度量算法.软件学报,2014,25(11):2473−2485. http://www.jos.org.cn/1000-9825/4508.htm [doi: 10.13328/j.cnki.jos.004508]

庄媛(1990－),女,辽宁营口人,硕士,CCF 学生会员,主要研究领域为服务计算.

冯钧(1969－),女,博士,教授,博士生导师,CCF 专业会员,主要研究领域为时空间数据管理,智能数据处理,数据挖掘.

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.

朱跃龙(1959－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为数据库技术,智能信息处理与数据挖掘,水利信息化.

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.