doi:10.3969/j.issn.1005-3026.2015.06.004

一种资源与服务性能关系的建模方法

张 斌,王 林,赵秀涛,张长胜
(东北大学 信息科学与工程学院,辽宁 沈阳 110819)

摘 要:获取资源与服务性能的关系模型是在云环境中为服务合理分配虚拟资源的关键.然而,训练数据的规模往往显著影响这种非线性关系模型的准确率.针对现有方法不足,提出了将协同过滤推荐(CFR)和支持向量回归(SVR)相结合的服务性能动态建模方法(CSDM).该方法在服务部署与运行时同时训练两种模型,并选择二者中MAE占优的性能模型预测给定资源状态下的服务性能,从而保证预测精度.同时,CSDM引入择优阈值以降低模型训练代价.实验表明,CSDM在不同规模的训练数据上均有较高的预测准确率,且择优阈值对预测精度和建模效率具有显著影响.

关 键 词:云服务;性能模型;资源状态;协同过滤推荐;支持向量回归
中图分类号:TP 311.5 **文献标志码**:A **文章编号**:1005-3026(2015)06-0773-04

A Novel Modeling Method for Relationships Between Resources and Service Performance

ZHANG Bin, *WANG Lin*, *ZHAO Xiu-tao*, *ZHANG Chang-sheng*
(School of Information Science & Engineering, Northeastern University, Shenyang 110819, China. Corresponding author: ZHANG Bin, E-mail: zhangbin@mail.neu.edu.cn)

Abstract: The relationship model between resources and service performance is a key to the proper virtual resource allocation for services in cloud environment. However, the accuracy of these non-linear relationship models is usually significantly influenced by the scale of training data. Aiming at the shortcomings of related work, a dynamic service performance modeling method named CSDM, which combines collaborative filtering recommendation and support vector regression, was proposed. In CSDM, for better accuracy, both performance models were trained at service deployment time and runtime, and the one with lower MAE was selected to estimate the performance under given resource status. In addition, a merit-based threshold was introduced to reduce training costs of performance models. The experimental results showed that CSDM had higher accuracy on different scales of training data, and the merit-based threshold had a significant effect on the prediction accuracy as well as the modeling efficiency.

Key words: cloud service; performance model; resource status; CFR (collaborative filtering recommendation); SVR (support vector regression)

云环境中的资源(CPU个数、内存大小等)是按需付费的,因此往往希望能够尽可能准确地估计在给定资源数量下的服务性能.为此,通常需要建立能够描述资源与服务性能之间关系的服务性能模型.

为了构建服务性能模型,现有研究大多采用回归分析、排队论模型、机器学习等方法.文献[1]采用多重线性回归模型预测不同资源(CPU时间、磁盘、网络)利用率下的组件性能.文献[2]基于M/M/n/PS排队模型建立服务性能模型,并用于预测服务的平均响应时间.文献[3]基于人工神经网络的方法针对不同虚拟化架构下影响应用性能的关键资源类型进行建模.然而,这些方法在获取资源与服务性能关系方面各有不足:应用

收稿日期:2014-04-24
基金项目:宁夏回族自治区自然科学基金资助项目(NZ13265);中央高校基本科研业务费专项资金资助项目(N120804001,N120604003);沈阳市科技基金资助项目(F12-277-1-80);国家科技支撑计划项目(2014BAI17B00).
作者简介:张 斌(1964-),男,辽宁沈阳人,东北大学教授,博士生导师.

性能与分配的资源量之间往往不存在线性关系[4],因此不易采用回归分析建模;现有排队论建模方法多为分析 CPU 资源与服务性能的关系,而实际影响服务性能的资源还可能包括内存、带宽等;机器学习方法通常需要大量训练数据,因此在训练数据有限的服务初始部署阶段,性能模型的预测准确率往往较低.

针对现有方法的不足,本文提出了一种将适用于非线性关系建模的协同过滤推荐(collaborative filtering recommendation, CFR)算法[5]与支持向量回归(support vector regression, SVR)算法[6]相结合的服务性能动态建模方法(CFR & SVR based dynamic modeling of service performance, CSDM).

与已有研究不同,本文的 CSDM 方法在性能日志有限的服务初始部署阶段采用 CFR 算法建模,从而提高数据有限情形下的预测准确率;随着性能日志增加而逐渐采用 SVR 算法建模,该算法鲁棒性好、计算复杂度低,对于非线性、高维训练数据具有良好的建模能力[7].

1 服务性能模型及其构建过程

根据应用目标不同,可以采用多种属性度量服务性能,如响应时间、吞吐量等,本文选择服务的平均响应时间(单位:s)作为其性能属性,并基于此给出服务性能模型的定义.

1.1 服务性能模型

定义 1 服务性能模型(service performance model, SPM). SPM 是一个描述资源状态与服务性能之间映射关系的模型,表示为 Θ: $\boldsymbol{R}\rightarrow Q$,其中,$\boldsymbol{R}=(r_1,r_2,\cdots,r_u)$,表示资源状态向量,$r_i(i=1,2,\cdots,u)$表示第 i 个资源类型的值;Q 表示服务的平均响应时间;Θ 表示 $\boldsymbol{R}$ 到 Q 的映射函数.

为了说明 CSDM 方法的基本过程且不失一般性,本文为资源状态向量选择 4 种常用的资源类型,即 CPU 个数、内存大小、网络带宽和存储能力,分别以 r_1,r_2,r_3,r_4 表示,其单位:CPU(个)、内存(MB)、网络带宽(MB/s)、存储(GB).

1.2 服务性能模型的构建过程

定义 2 服务性能日志项(service performance log item, SPLI). SPLI 是记录服务在某种资源状态下平均响应时间的一条数据,可用 6 元组 <ServiceID, CPU_Count, Mem_Size, BW_Size, Vol_Size, AResTime> 表示,各项分别表示服务标识号、CPU 个数、内存大小、网络带宽、存储容量,以及服务的平均响应时间.

在预测服务性能时要根据准确率选择性能模型,本文采用平均绝对误差(mean absolute error, MAE)来度量性能模型的准确率:

$$\mathrm{MAE}=\frac{\sum_{i=1}^{n}|\hat{q}-q|}{n}. \tag{1}$$

式中:$\hat{q}$ 表示性能预测值;q 表示性能观测值;n 表示验证数据的条数.

为了减少模型的离线训练时间,当 CFR 性能模型或 SVR 性能模型的预测准确率连续多次占优时,则认为该模型具有更高的准确率,并且在服务的后续运行过程中仅对其更新. 因此,设定一个择优阈值来确定唯一的性能模型,从而避免不必要的模型训练,本文以其在增量性能日志上的 MAE 连续低于另一个性能模型的次数 m 作为选择该模型的阈值.

具体的 CSDM 建模过程如图 1 所示.

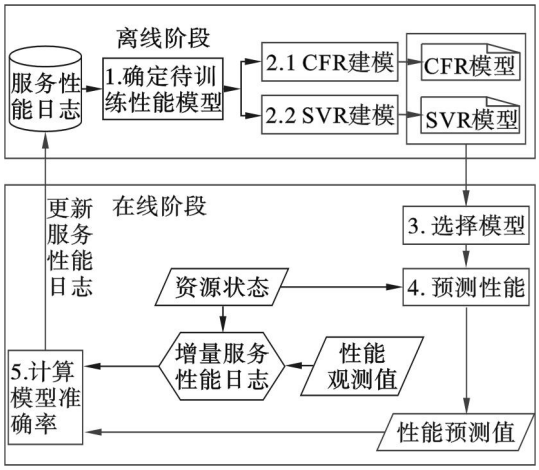

图 1 服务性能模型的构建过程

Fig. 1 Build process of service performance model

如图 1 所示,服务性能模型的构建包括两个阶段:离线阶段和在线阶段,离线阶段负责训练性能模型,而在线阶段负责选择模型来预测给定资源状态下的服务性能.

2 CFR 与 SVR 性能建模

2.1 基于 CFR 构建性能模型

类似推荐系统中用户(User)和项目(Item)的概念,定义由服务性能日志中的 M 个资源状态向量 $R_i(i=1,2,\cdots,M)$ 和 N 个服务项目 $S_j(j=1,2,\cdots,N)$ 构成的用户 - 项目矩阵 $\boldsymbol{Q}$,简称 RSV - SI 矩阵. 矩阵 $\boldsymbol{Q}$ 的每一项 $q_{i,j}$ 表示服务项目 S_j 在资源状态向量 $\boldsymbol{R}_i$ 下的平均响应时间. 如果 S_j 未

在 $\boldsymbol{R}_i$ 下执行过，则 $q_{i,j}=\text{null}$.

2.1.1　计算相似度

与文献[5]类似，为了降低在少量服务上具有相同平均响应时间但实际并不相似的资源状态向量的影响，修正皮尔逊相关系数相似度，得到资源状态向量 $\boldsymbol{e}$ 与 $\boldsymbol{r}$ 的相似度：

$$\text{Sim}(\boldsymbol{e},\boldsymbol{r})=\frac{2\times|I_e\cap I_r|}{|I_e|+|I_r|}\times\frac{\sum_{i\in I}(q_{e,i}-\overline{q_e})(q_{r,i}-\overline{q_r})}{\sqrt{\sum_{i\in I}(q_{e,i}-\overline{q_e})^2}\sqrt{\sum_{i\in I}(q_{r,i}-\overline{q_r})^2}}.\quad(2)$$

服务项目 f 与 s 的相似度：

$$\text{Sim}(f,s)=\frac{2\times|V_f\cap V_s|}{|V_f|+|V_s|}\times\frac{\sum_{v\in V}(q_{v,f}-\overline{q_f})(q_{v,s}-\overline{q_s})}{\sqrt{\sum_{v\in V}(q_{v,f}-\overline{q_f})^2}\sqrt{\sum_{v\in V}(q_{v,s}-\overline{q_s})^2}}.\quad(3)$$

2.1.2　预测 RSV－SI 矩阵中的缺失值

根据文献[5]中的 CFR 算法，选取相似邻居，并基于协同过滤方法补齐矩阵中的缺失值.

1）选取相似邻居. 与 $\boldsymbol{r}$ 相似的资源状态向量集合：

$$S(\boldsymbol{r})=\{r_a\mid r_a\in T(\boldsymbol{r}),\text{Sim}(r_a,\boldsymbol{r})>0,r_a\neq\boldsymbol{r}\}.\quad(4)$$

与 f 相似的服务项目集合：

$$S(f)=\{f_k\mid f_k\in T(f),\text{Sim}(f_k,f)>0,f_k\neq f\}.\quad(5)$$

2）计算缺失值.

$$P(q_{r,f})=w_r(\lambda)\times\left\{+\frac{\sum_{r_a\in S(r)}\text{Sim}(r_a,\boldsymbol{r})(q_{r_a,i}-\overline{r_a})}{\sum_{r_a\in S(r)}\text{Sim}(r_a,\boldsymbol{r})}\right\}$$

$$w_f(\lambda)\times\left\{+\frac{\sum_{f_k\in S(f)}\text{Sim}(f_k,f)(q_{r,f_k}-\overline{f_k})}{\sum_{f_k\in S(f)}\text{Sim}(f_k,f)}\right\}\quad(6)$$

2.1.3　预测服务性能

利用式(6)预测新的资源状态向量 $\boldsymbol{r}$ 下服务 f 的性能. 如果 $\boldsymbol{r}$ 与 f 在矩阵中不存在相似邻居，则令

$$P(q_{r,f})=w_r\times\overline{q_r}+w_f\times\overline{q_f}.\quad(7)$$

2.2　基于 SVR 构建性能模型

SVR 是支持向量机在函数回归领域的推广，是支持向量机在实函数域的研究内容.

2.2.1　SVR 方法构建回归模型

由于资源与服务性能通常是非线性关系，因此采用非线性的 SVR 模型[8]. 通过引入核函数 $k(\boldsymbol{r},\boldsymbol{r}')=[\varphi(\boldsymbol{r}),\varphi(\boldsymbol{r}')]$，SVR 性能模型可由求解式(8)的二次规划问题得到.

$$\min\frac{1}{2}\sum_{i,j=1}^{k}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)k(r_i,r_j)+\sum_{i=1}^{k}\alpha_i(\varepsilon-q_i)+\sum_{i=1}^{k}\alpha_i^*(\varepsilon+q_i).$$

$$\text{s.t.}\begin{cases}\sum_{i=1}^{k}(\alpha_i-\alpha_i^*)=0,\\ \alpha_i,\alpha_i^*\in[0,C].\end{cases}\quad(8)$$

求解上式，可得资源状态向量与服务平均响应时间的非线性模型：

$$\hat{\theta}(\boldsymbol{r})=\sum_{i=1}^{k}(\alpha_i-\alpha_i^*)k(r_i,\boldsymbol{r})+b.\quad(9)$$

2.2.2　核函数选取与 SVR 模型求解

本文选择 RBF 核函数进行非线性转换，即

$$k(x,y)=\exp(-\|x-y\|^2/(2\sigma^2)).\quad(10)$$

为了求解式(8)规划问题，采用序列最小优化(sequential minimal optimization，SMO)算法[9].

3　实验与结果分析

以不同类型的 Java 应用作为服务性能建模对象，从 3 个方面验证本文提出的 CSDM 性能建模方法的有效性：CSDM 与 CFR 模型、SVR 模型的 MAE 对比，择优阈值大小对 CSDM 的 MAE 的影响，以及性能建模的时间开销.

3.1　实验设置

为了深入验证 CSDM 的准确率，实验分析了 4 类服务：CPU 密集型、通信密集型、I/O 密集型以及其他类型.

3.2　实验结果及分析

3.2.1　不同类型服务模型 MAE 对比

从不同类型服务中随机选择一个作为性能建模对象，得到 CFR，SVR 和 CSDM 三种方法的 MAE 随性能日志规模的变化曲线，如图 2 所示.

从图 2 可知，3 种方法从总体上都是随着性能日志项的增多而表现出越来越低的预测误差. 当服务的性能日志项较少时，CFR 模型的 MAE 低于 SVR，此时 CSDM 选择的模型为 CFR；当服务性能日志项较多时，CFR 模型的 MAE 高于 SVR，此时 CSDM 选择的模型为 SVR.

3.2.2　择优阈值 m 对 CSDM 的 MAE 的影响

对于不同类型的服务，表 1 统计了在不同择优阈值的后续 100 次预测中，CSDM 的 MAE 等于 CFR 和 SVR 中最小 MAE 的次数.

由表 1 可知，随着择优阈值的增大，CSDM 方法在后续预测中 MAE 达到较小值的概率越来越大，表明择优阈值对于改善 CSDM 方法的预测准确率具有积极作用.

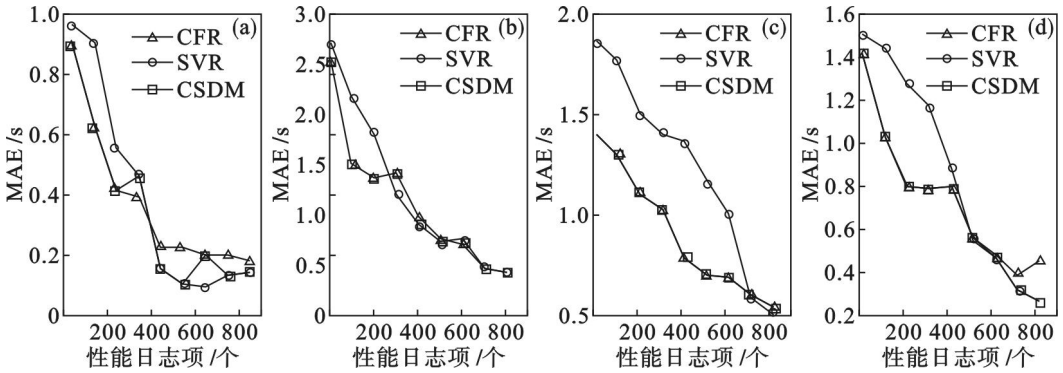

图 2 不同类型服务性能模型的 MAE

Fig. 2 MAE of performance models of different services

(a)—CPU 密集型;(b)—通信密集型;(c)—I/O 密集型;(d)—其他类型.

表 1 CSDM 的 MAE 占优比例

Table 1 Lower MAE ratio of CSDM

m	2	4	7	11	16
CPU 密集型	5	12	20	50	83
通信密集型	6	20	41	63	75
I/O 密集型	15	27	40	49	68
其他类型	10	20	48	56	72

3.2.3 性能建模的时间开销对比

图 3 给出了 CFR 和 SVR 两种算法在不同性能日志规模下的平均消耗时间. 由图 3 可知,对于不同性能日志规模,训练 CFR 算法的平均消耗时间均多于训练 SVR 算法的平均消耗时间,并且 CFR 算法随性能日志项数目的增加,变化幅度较大,表明择优阈值对于改善 CSDM 方法的效率具有积极作用.

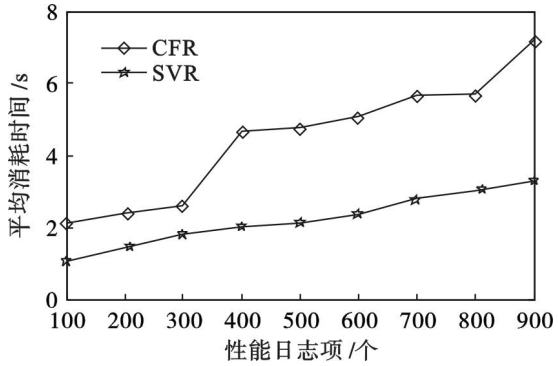

图 3 CFR 与 SVR 的平均消耗时间比较

Fig. 3 Average consumption time of CFR and SVR

4 结 论

本文针对云环境中资源与服务性能关系的建模问题,提出了一种服务动态建模方法 CSDM. 对比实验表明,该方法对于不同类型的服务,在不同规模的服务性能日志上均具有较高的预测准确率. 同时,研究了 CSDM 方法中的择优阈值对于提高性能模型准确率和降低模型训练时间开销的重要性. 本文服务性能模型目前仅考虑了不同资源类型对服务性能的影响,未来工作中将会增加一些其他因素,使模型应用场景更加广泛.

参考文献:

[1] Lloyd W, Pallickara S, David O, et al. Service isolation vs. consolidation: implications for iaas cloud application deployment[C]//IEEE International Conference on Cloud Engineering. San Francisco, 2013: 21 – 30.

[2] Dejun J, Pierre G, Chi C H. Autonomous resource provisioning for multi-service web applications[C]//Proceedings of the 19th International World Wide Web Conference. New York, 2010: 471 – 480.

[3] Kundu S, Rangaswami R, Dutta K, et al. Application performance modeling in a virtualized environment[C]// IEEE 16th International Symposium on High Performance Computer Architecture. Bangalore, 2010: 1 – 10.

[4] Rao J, Wei Y D, Gong J Y, et al. QoS guarantees and service differentiation for dynamic cloud applications[J]. *IEEE Transactions on Network and Service Management*, 2013, 10(1): 43 – 55.

[5] Zheng Z B, Ma H, Irwin K, et al. QoS-aware web service recommendation by collaborative filtering[J]. *IEEE Transactions on Services Computing*, 2011, 4(2): 140 – 152.

[6] Drucker H, Burges C, Kaufman L, et al. Support vector regression machines[C]//Advances in Neural Information Processing System 9. Cambridge, 1997: 155 – 161.

[7] 王宏宇,糜仲春,梁晓艳,等. 一种基于支持向量机回归的推荐算法[J]. 中国科学院研究生院学报, 2007, 24(6): 742 – 748. (Wang Hong-yu, Mi Zhong-chun, Liang Xiao-yan, et al. A recommendation algorithm based on support vector regression[J]. *Journal of University of Chinese Academy of Science*, 2007, 24(6): 742 – 748.)

[8] Lorenzi L, Mercier G, Melgani F. Support vector regression with kernel combination for missing data reconstruction[J]. *IEEE on Geoscience and Remote Sensing Letters*, 2012, 10(2): 367 – 372.

[9] Zhou Q, Zhai Y J, Han P. Sequential minimal optimization algorithm applied in short-term load forecasting[C]//IEEE International Conference on Machine Learning and Cybernetics. Hong Kong, 2007: 2479 – 2483.