# 面向入侵检测的基于IMGA和MKSVM的特征选择算法

井小沛 汪厚祥 聂 凯 罗志伟
(海军工程大学电子工程学院 武汉 430033)

**摘 要** 入侵检测系统处理的数据具有数据量大、特征维数高等特点,会降低检测算法的处理速度和检测效率。为了提高入侵检测系统的检测速度和准确率,将特征选择应用到入侵检测系统中。首先提出一种基于免疫记忆和遗传算法的高效特征子集生成策略,然后研究基于支持向量机的特征子集评估方法。并针对可能出现的数据集不平衡造成的特征子集评估能力下降,以黎曼几何为依据,利用保角变换对核函数进行修改,以提高支持向量机的分类泛化能力。实验仿真表明,提出的特征选择算法不仅可以提高特征选择的效果,而且在不平衡数据集上具有更好的特征选择能力。还表明,基于该方法构建的入侵检测系统与没有运用特征选择的入侵检测系统相比具有更好的性能。

**关键词** 特征选择,入侵检测,遗传算法,支持向量机,修正核函数
**中图法分类号** TP393 **文献标识码** A

## Feature Selection Algorithm Based on IMGA and MKSVM to Intrusion Detection

JING Xiao-pei WANG Hou-xiang NIE Kai LUO Zhi-wei
(Electrical Engineering Institution, Naval University of Engineering, Wuhan 430033, China)

**Abstract** In order to improve performances of intrusion detection system in terms of detection speed and detection rate, it is necessary to apply feature selection in intrusion detection system. Firstly, an efficient search procedure based on immune memory and genetic algorithm (IMGA) was proposed. Then, support vector machine (SVM) based on wrapper feature evaluation methods was surveyed, in order to improve the feature selection performance of unbalanced datasets. We used the conformal transformation and Riemannian metric to modify kernel function, and reconstructed a new Modified Kernel SVM (MKSVM). Finally, the simulation experimental results show that this approach can improve the process of selecting important features, and has better feature selection ability on the unbalanced data. Furthermore, the experiments indicate that intrusion detection system with this feature selection algorithm has better performances than that without feature selection algorithm.

**Keywords** Feature selection, Intrusion detection, Genetic algorithm, Support vector machine, Modified kernel

## 1 引言

作为网络安全防御体系重要组成部分的入侵检测系统(Intrusion Detection System, IDS),从计算机系统或者网络中收集、分析数据特征,检测任何企图破坏计算机资源的完整性、保密性和可用性的异常行为[1]。由于网络通信的行为是通过特征来刻画的,因此入侵检测系统需要分析大量的数据特征。但实际上,入侵检测要处理的大量数据特征中,存在大量冗余或无关特征,只有很少一部分特征与分类结果有关,这使入侵检测系统耗用大量的计算资源,增加系统训练时间,并降低准确度。此外,大量的计算也会使系统的检测速度受影响,在实时性已经成为入侵检测系统重要指标的今天,这个问题显得尤为突出。因此,通过特征选择方法找出能准确刻画网络行为的特征,对于入侵检测系统显得至关重要。

一个特征选择算法是由"特征子集生成"、"特征子集评估"、"终止条件"和"结果验证" 4部分组成的[2],其中特征子集生成和特征子集评估是最主要的两个阶段,即搜索策略和评估函数。搜索策略主要有完全搜索、启发式搜索、随机搜索和一些混合搜索策略;特征评估算法主要分为过滤式(Filter)和封装式(Wrapper)两种特征选择框架。其中Wrapper模型以机器学习算法的分类正确率作为特征子集的度量指标,因此其特征选择的效果要更好些,更适合用于对检测率要求高的入侵检测。目前搜索策略研究主要集中在随机搜索方面。如文献[3]利用遗传算法来选择最优特征子集;文献[4]中提出了一种基于粒子群优化算法和相关性分析的特征选择方法,它可以较好地得到特征子集;文献[5]提出一种基于遗传算法和禁忌搜索相混合的搜索策略来对特征子集空间进行随机搜索。在面向入侵检测的特征评估算法方面,主要研究集中在Wrapper型特征算法,如支持向量机[5]、神经网络[6]等。

本文在研究上述方法的基础上,以遗传算法为基础,提出一种将免疫记忆(Immune Memory, IM)与遗传算法(Genetic Algorithm, GA)相结合的搜索策略IMGA。特征评估方法选用在小样本和非线性数据上具有很好泛化能力的支持向量机(Support Vector Machine, SVM)。为了提高SVM在不平衡数据上的特征选择能力,提出一种基于修正核函数(Modified Kernel)的支持向量机——MKSVM,用来评估特征子集。

到稿日期:2011-08-10 返修日期:2011-10-18 本文受海军十一五预研项目(4010601010201)资助。
井小沛(1983—),男,博士生,主要研究方向为信息安全、机器学习,E-mail:jingxiaopei@163.com。

## 2 基于 IM 与 GA 的混合搜索策略

遗传算法是一种借鉴生物界自然选择和自然遗传机制的随机化搜索算法，其主要特点是群体搜索策略和群体中个体之间的信息交换。将其用于特征选择具有以下优点[7]：1)遗传算法的搜索过程不是直接作用在特征对应的数据上，而是作用在将特征编码后的字符串上；2)遗传算法用于特征选择，对特征样本集无线性等特殊要求，适用性较广；3)遗传算法的搜索过程是从一个群体到另一个群体，不易陷入局部最优解。但随着问题规模和复杂程度的不断提高，遗传算法在求解中存在“漂移”现象，使得计算、搜索时间过长。

免疫记忆特性是免疫系统中不可或缺的一个重要特征。免疫记忆是指机体在初次接受抗原刺激之后，免疫系统能够将抗体的部分细胞作为记忆细胞而被保存下来，对于今后侵入的同类抗原，相应的记忆细胞会被迅速激发而产生大量的抗体[8]。

本文将 IM 的记忆功能应用到 GA 的搜索过程中，针对 GA 计算搜索时间过长的缺点，构建免疫记忆算子，将种群中适应度高的优秀个体迁移到记忆库。IMGA 算法既克服了 GA 的遗传漂移现象，又保留了 GA 种群多样性的优势。

### 2.1 编码规则和适应度函数

对于给定的含有 $n$ 维特征的数据集 $D$，特征选择的目的就是选择出满足目标函数的最优特征子集 $F_S$。本文采用 0 和 1 的二进制编码方式来代表特征子集。其具体方法如下：对于特征集 $C=\{c_1,c_2,\cdots,c_n\}$，其空间可以影射为二进制编码串，这个编码串的每位对应一个特征。如果某个位置为 1，则表明选择该特征；如果为 0，则表明不选该特征。这样，每个二进制编码串就对应一个特征子集。采用二进制编码的好处还在于其容易进行交叉、变异等遗传操作。

定义特征子集 $F_S$ 的适应度函数为 $f(x)$。适应度函数由分类器的分类准确率和所选特征子集的大小两部分组成，分类器的分类准确率越高，特征子集的维数就越小，则该个体的适应度值越大，因此有较大的机会遗传到下一代。定义 $f(TPR)$ 为特征子集 $F_S$ 对应的检测率，$L$ 为特征子集 $F_S$ 选中的特征数，$n$ 为原始特征集的维数，则适应度函数为：

$$f(x)=e^{f(TPR)}-\frac{L}{n} \tag{1}$$

### 2.2 免疫记忆算子

```
Begin
    Generation=0;Generating original population;
    Fitness computing;
    Memory depot building:move θ% best fitness individuals to memory depot;
    Iteration;
    While (don't meet termination condition)
        Fitness computing:compute the fitness of new generation population,and remove θ% worst fitness individuals
        Comparison:compare the rest new population with the memory population;
        Memory depot update:select θ% best fitness individuals as new memory population
    Generation:=Generation+1;
End
End
```

图 1 免疫记忆算子操作过程

记忆算子的操作过程如图 1 所示。设算法的种群个体数为 $P_g$，记忆库个体数为 $P_i$（$P_i=P_g*\theta\%$，$\theta$ 为给定值），且 $P_i\in P_g$，其余个体属于训练种群。

从免疫记忆算子的操作过程可以看出，适应度高的个体得以保留并进行遗传操作，产生新个体，使优良基因得以传承；此外，记忆库的建立保证了新一代种群的最优适应度不低于前一代，加快了 GA 搜索的过程，有效地降低了 GA 存在的“漂移”现象。

### 2.3 IMGA 特征选择搜索算法

将 IMGA 算法应用到特征选择过程中，可以加快最优特征的寻找过程。IMGA 特征选择搜索算法描述如下。

输入：含有 $n$ 位特征的数据集 $D$，种群数量 $P_g$，记忆库规模 $P_i$，最大进化代数 max$G$。

输出：最优特征子集。

初始阶段：

Step 1 初始化：产生初始种群，由特征集合 $F$ 随机生成 $P_g$ 个特征子集作为初始种群 $G(0)$，特征子集位数为 $n$；设置终止条件（达到最大进化代数 max$G$）；进化代数 $t=0$；

Step 2 记忆库建立：采用适应度函数 $f(x)$ 在数据集 $D$ 上计算初始种群 $G(0)$ 中每一个个体的适应度；将种群 $G(0)$ 中适应度高的 $\theta\%$ 个体迁移到记忆库，产生初始记忆种群 $I(0)$；剩余个体保留在种群 $G(0)$ 中；

迭代阶段：

Step 3 遗传操作：采用轮盘赌方法在第 $t$ 代种群 $G(t)$ 和记忆库 $I(t)$ 中按比例选择适应度较高的个体；对选择的个体进行交叉、变异操作，产生新一代种群 $G(t+1)$；

Step 4 适应度评价：采用适应度函数 $f(s)$ 在数据集 $D$ 上计算第 $t+1$ 代种群 $G(t+1)$ 中每一个个体的适应度；

Step 5 种群和记忆库更新：移除种群 $G(t+1)$ 中适应度低的 $\theta\%$ 个体，将剩余个体与记忆库 $I(t)$ 中的个体进行适应度比较，选择最好的 $\theta\%$ 个体添加到记忆库，产生新一代记忆种群 $I(t+1)$；将记忆库中适应度较低的个体迁移到种群 $G(t+1)$ 中；

Step 6 终止检验：如果达到终止条件，则终止计算，输出记忆库中具有最高适应度的特征子集；否则置 $t=t+1$，转向 Step 3。

## 3 基于修正核函数 SVM 的特征评估方法

将 SVM 应用到特征选择中，主要是因为其在非线性数据和小样本数据上具有很好的分类能力，因此选出的特征更能代表类的区分。但 SVM 一般是基于平衡数据（两类样本数相等）进行分类建模的，对于不平衡数据（两类样本数不等），则会出现分类面向样本少的一方偏移，造成分类精度不高，因此选出的特征不能很好地代表类的区分。为了提高 SVM 在不平衡数据上的特征选择能力，提出 MKSVM 模型来提高分类能力。

对于给定分类问题，设定其训练样本集为 $\{x_i,y_i\}$，$i=1,2,\cdots,l$，$x_i\in X$，$y_i\in\{-1,+1\}$，其中 $X$ 为 $d_n$ 维空间。SVM 即为线性分类器，通过构造最优分类面，使得类别间的分类间隔最大[9]。但是这样的分类面不是唯一的，SVM 训练算法就是求出与两类样本间隔尽可能大的分类面。因此，将求最优分类面的问题转化为优化问题：

$$\left.\begin{aligned}&\min\ \phi(\omega)=\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{l}\xi_i\\&\text{s.t.}\ \ y_i(\omega K(x\cdot x_i)+b)\geqslant 1-\xi_i\end{aligned}\right\}\tag{2}$$

式中，$\xi_i$ 为松弛变量，$\xi_i\geqslant 0$；$C$ 为惩罚系数；$K(x\cdot x_i)$ 为核函数。

SVM 中核函数的作用是将低维非线性空间 $X$ 的数据通过核函数映射 $\Phi(x)$ 到高维特征空间 $H$，将非线性变换转换为某个高维空间中的线性问题，这样就将非线性问题转化成线性问题来解决[10]。对于空间 $X$ 上的连续数据，它们在特征空间 $H$ 的映射 $S$ 是一个可微流形[11]。对于 $X$ 上的两点 $x_i$ 和 $x_j$，它们在面 $S$ 上的距离，即黎曼距离定义为：

$$ds^2=\sum_{i,j}g_{ij}(x)dx_i dx_j\tag{3}$$

式中，$g_{ij}(x)=(\partial\Phi(x)/\partial x_i)\cdot(\partial\Phi(x)/\partial x_j)$ 为空间 $H$ 上的黎曼度量。特征空间 $H$ 成为黎曼空间，其体积为：

$$dv=\sqrt{g(x)}\,dx_1\cdots dx_n\tag{4}$$

式中，$g(x)=\det(g_{ij}(x))$。直观地说，$g(x)$ 反映了特征空间中点 $\Phi(x)$ 附近局部区域被缩放的程度，因此也称 $g(x)$ 为缩放因子。因此，对于不平衡数据，通过引入修正函数来改进核函数的黎曼度量，在样本的稠密区域，减小核数值，在稀疏区域，增加核数值，从而达到间接影响黎曼度量的目的，提高模型分类能力，进而提高特征选择效果。

**定义 1** 对于一个正的可微标量函数 $C(x)$，定义

$$\tilde{K}(x,x')=C(x)C(x')K(x,x')\tag{5}$$

称之为核函数通过因子 $C(x)$ 的保角变换，则 $\tilde{K}(x,x')$ 成为支持向量机的修正核函数。根据定义 1，非线性映射 $\Phi$ 被修正为 $\tilde{\Phi}(x)=C(x)\Phi(x)$。

下面介绍修正函数的构建。设 $d_n$ 维空间 $X$ 中存在两类样本集 $S_+$、$S_-$，样本数目分别是 $N_1$、$N_2$。构造修正函数之前先用 $d(x)$ 表示一个样本与同类其它样本的欧式距离平均值，可以断定在稠密区域样本的 $d(x)$ 值会小于在稀疏区域样本的 $d(x)$ 值。因此 $d(x)$ 是空间分布变化的一个近似量化指示。此外，那些距离样本类中心较远的样本的 $d(x)$ 也比距离样本类中心较近的样本大。$d(x)$ 通过下式计算：

$$d(x)=\begin{cases}(\sum_{i=1}^{N_1}\sqrt{\|x-x_i^+\|^2})/N_1, & x\in S_+\\(\sum_{i=1}^{N_2}\sqrt{\|x-x_i^-\|^2})/N_2, & x\in S_-\end{cases}\tag{6}$$

根据式(6)，本文构造的修正函数为：

$$C(x)=\exp(d(x)/r_i)\tag{7}$$

式中，$r_i$ 为样本类的聚类半径。将式(7)代入式(5)，就得到修正核函数。由此可见，采用本文构建的修正核函数，稀疏样本和边界样本都能获得较大的体积扩张，有利于提高分类精度。

## 4 基于 MKSVM 的 Wrapper 型轻量级入侵检测系统

基于 MKSVM 的 Wrapper 型轻量级入侵检测系统的总体结构如图 2 所示。在对训练集数据进行预处理后，系统采用修正核函数来训练 SVM 模型。建立的 MKSVM 模型对特征子集进行验证评估，选出最优特征并对训练集和测试集进行特征约简。完成特征约简后，训练集和测试集被送到入侵检测系统中进行训练和测试。

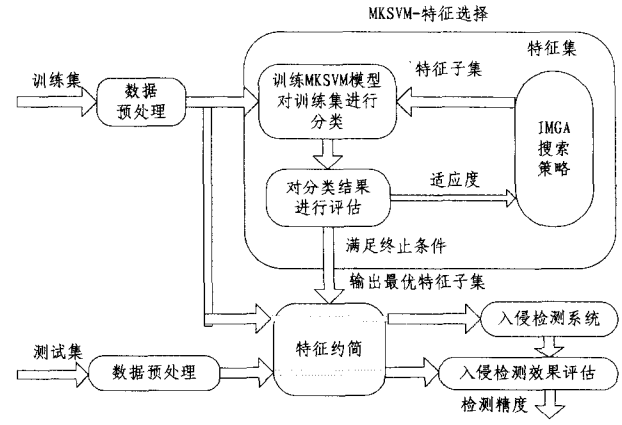


图 2 基于 MKSVM 轻量级入侵检测系统的总体结构

## 5 实验仿真分析

### 5.1 实验数据集

实验数据选用 KDD CUP 99 数据集。这个数据集是网络入侵检测的标准测试集，为 IDS 研究人员提供训练和测试数据集，以比较不同入侵检测方法的优劣，是现在研究最多的 IDS 数据集。数据中包含 41 维特征，34 个数值型字段和 7 个符号型字段[12]。这 41 维特征可以分为 4 部分：TCP 连接的基本特征(1～9 号特征)；TCP 连接的内容特征(10～22 号特征)；基于时间的网络流量统计特征(23～31 号特征)；基于主机的网络流量统计特征(32～41 号特征)。

### 5.2 仿真实验

实验共分两组：实验 1 验证采用本文提出的 MKSVM 特征选择方法的入侵检测系统的建模时间和检测率是否有改进，并将 MKSVM 特征选择方法与其它特征选择方法进行比较分析；实验 2 在不平衡训练集上，将 MKSVM 特征选择方法与 SVM 特征选择方法进行比较。

实验从 KDD CUP 99 data_10_percent 数据集中随机抽取两份数据做训练集。其中，训练集 1 中的正常样本和异常样本分别为 6760 和 5836，基本相等；训练集 2 中的正常样本和异常样本分别为 2152 和 9698，为不平衡数据集。从 corrected 中随机抽出 5 份数据做测试集：每份样本数均为 11272。实验采用的高斯 RBF 核函数($K(x,x')=e^{-\|x-x'\|^2/(2\sigma^2)}$)为 SVM 核函数。采用检测率来判断模型分类效果，定义检测率(True Positive Rate，TPR)为：

TPR＝正确分类的样本个数/总的样本个数＊100%

实验 1 利用 IMGA 搜索策略和 MKSVM 评估方法，在训练集 1 上进行特征选择。为了比较本文采用的基于 IMGA-MCSVM 特征选择的效果，实验同时采用其它特征选择方法来进行特征的选择，这些方法见表 1。

表 1 特征选择方法和搜索策略

| | 特征评估类型 | | | |
|---|---|---|---|---|
| | Filter 型 | | Wrapper 型 | |
| 评估方法 | CFS | CSE | SVM | MKSVM |
| 搜索策略 | GA | GA | GA | IMGA |

基于特征相关性的特征子集选择方法(Correlation Feature Selection，CFS)综合考虑了单一特征的预测值与类间的相关度；基于一致性的特征子集选择方法(Consistency Subset

Evaluation,CSE)将训练数据集映射到特征集上来检测类值的一致性;基于 SVM 的特征子集选择方法采用 C-SVM 来评估特征子集。这些方法的详细说明见文献[13]。特征选择的结果见表 2。

表 2 各种特征选择算法选择的特征子集

| 评估函数 | 选择的特征 |
|---|---|
| CFS | 3,5,6,23,32:service,src_bytes,dst_bytes,count,dst_host_count |
| CSE | 1,3,5,23,34,36,40:duration,service,src_bytes,count,dst_host_same_srv_rate,dst_host_same_src_port_rate,dst_host_rerror_rate |
| SVM | 3,5,34,38:service,src_bytes,dst_host_same_srv_rate,dst_host_serror_rate |
| MKSVM | 5,6,20,28,34:src_bytes,dst_bytes,num_outbound_cmds,srv_rerror_rate,dst_host_same_srv_rate |

表 2 的右栏是每一种特征选择方法选出的特征子集。如 SVM 选出的特征子集为 3,5,34,38,这些数字表示该特征在 KDD99 数据集的 41 个特征中的排序序号。冒号右边的单词是冒号左边的数字对应的特征名称,如 3 对应 service。

然后将选择的特征在训练集 1 上进行 MKSVM 分类建模,并分别对测试集进行分类测试。为了对比特征选择对分类精度的影响,实验同时将所有 41 维特征在训练集 1 上进行 MKSVM 分类建模,并分别对测试集进行分类测试。实验结果如图 3 所示。

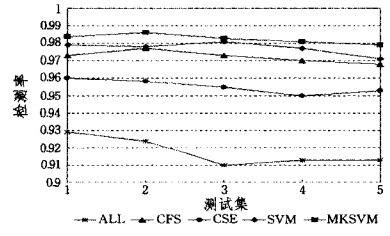


图 3 基于全部特征和不同特征子集的入侵检测系统的检测率

分析图 3 的数据,可以得到以下结果:

1)加入特征选择的入侵检测系统,检测率明显高于没有运用特征选择的入侵检测系统。对于所抽取的测试集而言,平均检测率要高出 5 个百分点。这是因为特征选择能有效地保留关键特征,剔除冗余和不重要的特征,提高检测率。而且对原始特征进行特征选择后,入侵检测系统的建模时间和测试时间都明显缩短。我们记录了应用不同特征子集的入侵检测系统的建模时间和在 5 个测试集上的平均测试时间,结果见表 3。

表 3 不同入侵检测系统的建模时间和平均测试时间

| | 入侵检测系统 | | | | |
|---|---|---|---|---|---|
| | ALL-IDS | CFS-IDS | CSE-IDS | SVM-IDS | MKSVM-IDS |
| 建模时间 | 154.34s | 35.16s | 57.91s | 9.48s | 37.83s |
| 平均测试时间 | 70.36s | 17.25s | 25.09s | 3.52s | 17.58s |

2)基于 Wrapper 型特征选择的入侵检测系统,检测效果要好于基于 Filter 型特征选择的入侵检测系统。这是因为 Filter 型特征选择算法是利用数据的内在特性来评价选取的特征子集,独立于学习算法;而 Wrapper 型特征选择算法则将后续学习算法的结果作为特征子集评价准则的一部分,根据算法生成规则的分类精度选择特征子集。

3)实验数据还表明,基于 IMGA-MKSVM 的入侵检测系统的检测效果优于基于 GA-SVM 的入侵检测系统。这归结于免疫记忆机制的加入,在相同计算代数的情况下,这有助于得到更好的较优解,也就是得到较优的特征子集。

实验 2 利用 IMGA-MKSVM 和 IMGA-SVM 评估方法,在训练集 2 上进行特征选择。特征选择的结果见表 4。

表 4 不平衡数据上两种特征选择方法选择的特征子集

| 特征选择方法 | 选择的特征 |
|---|---|
| IMGA-MKSVM | 4,5,6,12,30:flag,src_bytes,dst_bytes,logged_in,diff_srv_rate |
| IMGA-SVM | 2,6,10,14,22,36:protocol_type,dst_bytes,hot,root_shell,is_guest_login,dst_host_same_src_port_rate |

然后将选择的特征在训练集 2 上进行 MKSVM 分类建模,并分别对测试集进行分类测试。实验结果如图 4 所示。

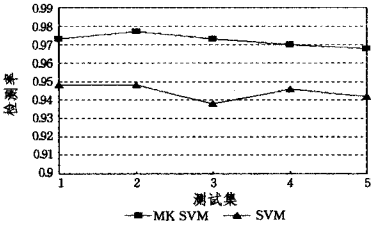


图 4 基于两种不同特征子集的入侵检测系统的检测率

图 4 中的数据显示,利用 IMGA-MKSVM 选择的特征训练的检测器比利用 IMGA-SVM 选择的特征训练的检测器检测率更高。因为在不平衡数据集上,一般 C-SVM 会出现分类面向样本少的一方偏移,所以选出的特征就不能很好地代表类的特点。而 IMGA-MKSVM 通过修正核函数,对不同类样本实施不同的黎曼度量,同时对类边界样本进行体积扩张,即尽量放大黎曼度量,因此具有更好的泛化能力,在不平衡数据集上选出的特征也就更具针对性。

**结束语** 网络数据存在数据量大、数据维度高、连续型特征数目繁多的特点。特征选择在降低数据维度、数据分类、数据聚类中都起着重要的作用。本文利用免疫记忆遗传算法和修正核函数 SVM 模型对入侵检测数据的特征选择问题进行研究,可以有效地提高寻找最优特征子集的能力。并且针对数据集不平衡现象,建立对两类样本进行不同映射的修正核函数,可以很好地解决由此造成的分类面偏移,提高特征选择精度。最后,实验结果表明,本文方法不仅能够得到较少的特征数,有效地降低网络数据的维度,而且能够提高在不平衡数据上的特征选择能力,因此本文构建的轻量级入侵检测系统对行为特征经常变化和攻击行为层出不穷的实时入侵检测来说非常适用。

## 参 考 文 献


[1] 唐正军,李建华. 入侵检测技术[M]. 北京:清华大学出版社,2004

[2] Langley P. Selection of Relevant Features in Machine Learning [C]//Proceedings of the AAAI Fall Symposium on Relevance. 1994:140-144

[3] Tan Feng,Fu Xue-zheng,Zhang Yan-qing,et al. A genetic algorithm based method for feature subset selection [J]. Soft Computing,2008,12(2):111-120

[4] 郭文忠,陈国龙,陈庆良,等. 基于粒子群优化算法和相关性分析的特征子集选择[J]. 计算机科学,2008,35(2):144-146


(下转第 111 页)

### 4.2 服务请求响应机制

当收到 QoS 服务需求后，经过以下 4 个步骤就可获得响应结果。①候选服务集经 QoSMM QoS 预处理后，生成新的 QoS 矩阵；②QEM 模块完成对 QoS 的评估；③按照服务消费者的需求 QoS，TDEM 模块找出目标消费群，并计算用户评价评估值，完成可信度评估；④由 SSM 模块负责服务选择，并将其返回结果提供给服务消费者。在该机制中，由于准则是基于同一目标消费群的评价，具有可比性，因此可为用户选择出最适合自身需求的服务。

## 5 仿真实验

为便于比较与描述，将传统的基于 QoS 的 Web 服务选择算法简称为 SMQ，基于用户反馈的 Web 服务选择机制简称为 SMF，基于 QoS 和用户反馈的服务选择机制简称为 WSMQF。与 SMQ 算法相比，WSMQF 由于增加了反馈评估步骤，因此时间开销有所增加；与 SMF 算法相比，由于 WSMQF 增加了评估 QoS 和查找目标消费群两个步骤，因此时间开销也会稍有增加。但定位所属目标消费群后，由于用户评价数量与原来相比有所减少，因此时间开销有所减少。现以查准率为例，在不同无效评价下对服务查准率进行仿真。查准率表示服务选择系统按照服务消费者需求 QoS 选择出来的服务为理想服务的次数在服务请求中所占的比例。针对服务查准率对提出算法的仿真统计结果如图 6 所示。

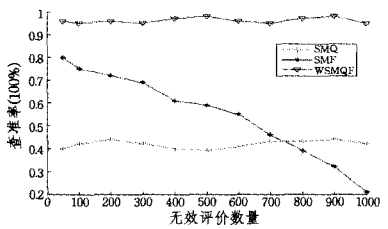


图 6 不同无效评价数量下服务查准率

比较仿真结果可知，SMQ 的查准率比较低，SMF 随着无效评价的增加而降低，WSMQF 查准率一直维持在一个比较高的水平，不会随无效评价的增加而降低。由仿真实验可知，SMQ 时间开销比较小，但是查准率很低；SMF 时间开销比 SMQ 要高，查准率相对于 SMQ 也高；WSMQF 时间开销相对于 SMF 要小，比 SMQ 稍大，但其查准率是 3 种机制中最稳定和最高的。由此，可以看出 WSMQF 在可以接受的时间开销下，查准率方面有明显改进。

**结束语** 上述针对功能相似的 Web 服务选择问题，探讨了基于 QoS 与可信度融合的 Web 服务选择机制，借助仿真实验，初步验证了该服务选择机制的合理性与有效性。鉴于服务选择机制的复杂性，有些问题还有待进一步研究解决，例如如何根据实际需求对目标消费群进行合理划分以及如何确定服务消费者的初始可信度值来抑制服务消费者的恶意评价等，这正是后续的研究课题。

## 参 考 文 献


[1] Raj R J R. Web service recommendation framework using QoS based discovery and ranking process [C]// Advanced Computing (ICoAC), 2011, Third International Conference. 2011: 371-377

[2] 高亚春，张为群. 基于 QoS 本体的 Web 服务描述和选择机制[J]. 计算机科学，2008，35(12)：273-276

[3] 李建楠，胡健生. 基于面向对象 FPN 的 QoS 系统建模分析[J]. 重庆理工大学学报：自然科学版，2011，25(11)：73-77

[4] 刘志中，王志坚，周晓峰，等. 基于 C－MMAS 算法的组合服务动态选择研究[J]. 计算机科学，2010，37(11)：135-140

[5] 边小凡，代艳红. 基于 QoS 的服务发现改进模型[J]. 计算机应用，2008，28(9)：2398-2400

[6] 杨墨，王丽娜. 基于信任容错的 Web 服务可靠性增强方法研究[J]. 通信学报，2010，31(9)：131-138

[7] Raj R J R. Web service selection based on QoS Constraints[C]// Trend in Information Sciences & Computing (TISC). 2010: 156-162

[8] 徐建国，罗永亮，王德才. 基于信誉模型的 Web 服务优化[J]. 电子科技大学学报，2008(28)：322-325


(上接第 99 页)


[5] 陈友，沈华伟，李洋. 一种高效的面向入侵检测系统的特征选择算法[J]. 计算机学报，2007，30(8)：1398-1408

[6] 陈友，程学旗，李洋，等. 基于特征选择的轻量级入侵检测系统[J]. 软件学报，2007，18(7)：1639-1651

[7] 杨孔雨，王秀峰. 免疫记忆遗传算法及其完全收敛性研究[J]. 计算机工程与应用，2005(12)：47-50

[8] Alonso O, Gonzalez F A, Niño F, et al. Search and Optimization: A Solution Concept for Artificial Immune Networks: A Coevolutionary Perspective[C]// Proceedings of 6th International Conference on Artificial Immune Systems. Santos/SP, Brazil, August, 2007

[9] Horng S-J, Su M-Y, Chen Y-H, et al. A novel intrusion detection system based on hierarchical clustering and support vector machines[J]. Expert Systems with Applications, 2011(38): 306-313

[10] Khan L, Awad M, Thuraisingham B. A new intrusion detection system using support vector machines and hierarchical clustering [J]. The VLDB Journal, 2007(16): 507-521

[11] Amari S, Wu S. Improving Support Vector Machine Classifiers by Modifying Kernel Function[J]. Journal of Neural Networks, 1999, 6(12): 783-789

[12] KDD cup 1999 data [EB/OL]. http://kdd.ics.uci.edu/databases/ kddcup99/ kddcup99.html

[13] Witten L H, Frank E. Data Mining: Practical Machine Learning Tools and Techniques (Third Edition)[M]. San Francisco: Morgan Kaufmann, 2011