基于 Seed 集的半监督核聚类

李昆仑,张 超,曹 铮,刘 明
LI Kun-lun,ZHANG Chao,CAO Zheng,LIU Ming

河北大学 电子信息工程学院,河北 保定 071002
College of Electronic and Information Engineering,Hebei University,Baoding,Hebei 071002,China
E-mail:likunlun@hbu.cn

LI Kun-lun,ZHANG Chao,CAO Zheng,et al.Semi-supervised kernel clustering algorithm based on seed set.Computer Engineering and Applications,2009,45(20):154-157.

Abstract: This paper presents a novel semi-supervised kernel clustering algorithm called Seed Kernel K-Means(SKK-Means) algorithm.It uses labeled data to generate initial seed clusters to guide the clustering process and data partition,and uses kernel method to map the input data into a high-dimensional feature space and calculates the distance between data points with a kernel function.The algorithm is compared with the other clustering algorithms such as K-Means and Kernel K-Means,on UCI databases in some numeric experiment.

Key words: semi-supervised clustering;seed set;kernel method;K-means

摘 要:提出了一种新的半监督核聚类算法——SKK-均值算法。算法利用一定数量的标记样本构成 seed 集,作为监督信息来初始化 K-均值算法的聚类中心,引导聚类过程并约束数据划分;同时还采用了核方法把输入数据映射到高维特征空间,并用核函数来实现样本之间的距离计算。在 UCI 数据集上进行了数值实验,并与 K-均值算法和核-K-均值算法进行了比较。

关键词:半监督聚类;seed 集;核方法;K-均值
DOI:10.3778/j.issn.1002-8331.2009.20.046 **文章编号:**1002-8331(2009)20-0154-04 **文献标识码:**A **中图分类号:**TP181

1 引言

在机器学习及相关领域中如何高效地处理海量数据一直是一个公开的难题。传统的监督学习方法通常需要大量的含有类别标记的数据作为训练集来保证算法的泛化能力,但有标记数据往往数量很少而且不易获取[1],无法满足监督学习的要求;无监督学习是一种自动学习方式,不需要有标记数据作为监督信息,但是所得到的结果通常是不够精确的。近年出现的半监督学习引起了越来越多的关注[2]。半监督学习综合利用少量有标记数据和大量无标记数据,把有标记和无标记的数据结合起来进行学习,既有监督学习的特点,又不需要大量的有标记样本[3]。其中,半监督聚类是一种常用的方法。

在现有的半监督聚类算法中,基于约束的方法[4]被广泛采用。该方法使用少量监督信息约束聚类的搜索过程,指导算法向一个比较好的划分进行。其中监督信息可以是样本的类别标记或者一对样本是否属于同一类别的约束关系。算法的实现方式主要有:(1) 修改聚类的目标函数以满足给定的约束;(2)在聚类过程中满足约束条件;(3) 利用有标记数据初始化聚类参数并在聚类过程中约束数据的划分。

基于上述利用有标记数据初始化聚类参数并在聚类过程中约束数据的划分的半监督方案, 受经典的 K-均值算法启发,提出了 Seed-核-K-均值算法(SKK-均值算法)。该算法利用一定数量的有标记样本构成 seed 集来初始化聚类中心, 并引导聚类的过程,约束数据的划分。同时引入了核技巧,将聚类过程扩展到核空间,即把数据从原始的输入空间通过核函数映射到高维的特征空间来进行距离的计算,从而使同簇中的数据点相似程度及不同簇的数据点相异程度加大,并且原输入空间中具有非线性边界的簇也可以被发现。这样,算法既体现了半监督学习在聚类中的优势,又使用了核技巧,准确度大大改善。

2 背景知识

首先介绍算法所需的一些背景知识, 包括经典的 K-均值算法和核方法。

2.1 K-均值算法

K-均值算法是一种基于迭代和再分配的经典聚类方法,它把一个由 N 个无标记样本组成的数据集 $X=\{x_i\}_{i=1}^{N}$,$x_i \in \mathbb{R}$ 划

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60773062,No.60873100);河北省科技支撑计划项目(the Scientific and Technical Supporting Programs of Hebei Province of China under Grant No.072135188);河北省教育厅科研计划项目(the Scientific Research Project of Department of Hebei Education of China under Grant No.2008312)。

作者简介:李昆仑(1962-),男,博士,副教授,硕士生导师,主要研究领域为模式识别,人工智能等;张超(1983-),男,硕士生,主要研究领域为模式识别。

收稿日期:2009-01-09 **修回日期:**2009-04-01

分为 K 个簇 $X_1, X_2, \cdots, X_K$(得到的结果叫做 X 的一个划分 $\{X_k\}_{k=1}^{K}$),使得簇内的相似度较高,而簇间的相似度较低。这里的相似度一般是指簇中样本与簇中心的距离,通常采用欧氏距离;簇中心一般取簇中所有样本的算术平均值;算法一般采用基于欧式距离和类内误差平方和准则。即,K-均值聚类的目标函数为:

$$J=\sum_{k=1}^{K}\sum_{i=1}^{N_k}\| x_i-m_k \|^2,\text{其中 } m_k=\frac{1}{N_k}\sum_{i=1}^{N_k}x_i \tag{1}$$

2.2 核方法

核方法(Kernel Methods)是一种奇妙的方法,它根据分类器的误差率不依赖于空间的维数这一特性,构造一个特征映射,把在输入空间中线性不可分的样本映射到一个高维的特征空间(称为核空间)中,使得核空间中样本是线性可分或近似线性可分的[5]。这里所用到的映射通常是非线性的,具体形式未知,可以通过适当的核函数来实现。由于核空间是一个希尔伯特空间,从而核函数就是一种希尔伯特空间的内积,可以表示为 Mercer 核的形式:$K(x_i,x_j)=\phi(x_i)\cdot\phi(x_j)$,其中 $\phi: X\to\mathbb{F}$ 定义了一个从输入空间 X 到核空间 $\mathbb{F}$ 的映射。

基于核方法的聚类方法主要有三类:(1)度量的核函数化(kernelization of the metric),如 B.Kulis 等提出的对每个数据点引入权值(weight)的加权核-K-均值算法,用于对图像的标准化分割[6]。(2)在特征空间中的聚类,这种方法使用核函数映射各个模式,然后在特征空间计算中心,如 Mark Girolami 提出的 K-均值聚类方法[7]。(3)支持向量(SV)的描述方法,如,Francesco Camastra 提出的核聚类方法[8]。

3 算法描述

3.1 Seed-核-K-均值算法

本文所提出的算法是基于核-K-均值算法,并加入 seed 集作为监督信息来实现的。对于 N 个无标记样本 x_i(i=1,2,…,N)组成的数据集 X,算法的目的是得到 X 的一个划分 $\{X_k\}_{k=1}^{K}$,使得对于采用平方和误差准则形式的目标函数为最小。

取 X 中的 Ns 个样本 x_i 组成 X 的子集 $S(S\subseteq X)$,然后对 S 中的样本进行标记,并且假定 X 的每一个簇 X_k 中,至少有一个 $x_i\in S$ 点。这样就得到 Ns 个有标记样本 x_{Si} 组成的集合 Xs,这些有标记样本 x_{Si} 称为 seed,它们组成的集合 Xs 成为 seed 集,得到 Xs 的一个划分 $\{Xs_k\}_{k=1}^{K}$,其中 Xs 的第 k 类中有 Ns_k 个有标记样本。与之对应,可以得到 X 的子集 S 的一个划分 $\{S_k\}_{k=1}^{K}$,每个类 S_k 中同样有 Ns_k 个样本,若 $x_i\in S_k$,则 $x_i\in X_k$。为了简便,可以把 S 和 Xs 统称为 seed 集。利用这个 S 集的划分可以得到原数据空间的聚类中心为:$m_k=\frac{1}{Ns_k}\sum_{x_i\in S_k}x_i$。在计算均值时使用的是无标记的数据 x_i,因此不会由于维数的不同而影响后续的距离计算和聚类过程。

通过核技巧的运用,可以将原始空间的距离计算转化为特征空间的内积计算,并用核函数来代替内积。即,在核空间中任意一个样本和某类均值之间的距离按

$$\| \phi(x_i)-\phi(m_k) \|^2=\| \phi(x_i)-\frac{1}{Ns_k}\sum_{i=1}^{Ns_k}\phi(x_i) \|^2=$$

$$1-2\cdot\frac{1}{Ns_k}\sum_{j=1}^{Ns_k}K(x_l,x_j)+\frac{1}{Ns_k^2}\sum_{j=1}^{Ns_k}\sum_{l=1}^{Ns_k}K(x_j,x_l) \tag{2}$$

计算,其中 $K(\cdot,\cdot)$ 就是核函数,在本文的讨论中,使用的是 RBF 核函数。

接下来用核-K-均值的方法进行聚类,此时待分样本仍然是 $\phi(x_i)$,但是簇中心变为 $\phi(m_k)=\frac{1}{Ns_k}\sum_{x_i\in S_k}\phi(x_i)$。将其代入核-K-均值算法的目标函数

$$J^{\phi}=\sum_{k=1}^{K}\sum_{i=1}^{N_k}\| \phi(x_i)-\phi(m_k) \|^2 \tag{3}$$

求出每个数据点 x_i 与 k 个簇中心在特征空间中的距离,然后将 k 从 1 取到 K 时得到的距离值相比较,把 x_i 归到距离为最小时的 k 对应的簇中,这样就得到一个对数据集 X 的划分。然后将这次的结果作为下一次的初始划分,进行迭代,并计算前后两次的目标函数的差的绝对值。当这个差值小于一个给定的阈值(通常设为 1^{-6})时,就认为目标函数达到最小,输出结果,否则就重复上面的聚类步骤迭代下去。

3.2 EM 策略

为了优化目标函数,使用基于 EM 算法的策略[9]。EM 算法由 E-步和 M-步组成,通过交替使用这两个步骤,提供一个简单的迭代过程来计算目标函数,逐步改进模型的参数,最后终止于一个极大值点。本文提出的 SKK-均值算法首先使用 seed 样本来计算簇中心,然后以这些簇中心为初始条件计算样本与之距离,得到初始的 K 个簇。然后进行 E-步和 M-步的迭代过程。

E-步:算法将全部数据点分配到不同的簇使目标函数 J^{ϕ} 最小。J^{ϕ} 的值在初始时是一个常数,当一个数据点所属的簇发生变化时其值改变,所有数据点被重新划分后生成新的 J^{ϕ} 值。该步重复上述过程直到对数据点的划分没有变化时为止。

M-步:算法重新计算簇中心。根据 E-步所得出的结果,在每个簇中寻找一个特定的数据点作为该簇的中心。数据点的选取采用了在原空间计算它们的算术平均值的方式。得到了簇中心以后再将 E-步中得到的划分更新。

SKK-均值算法的实现步骤描述如下:

算法 Seed-核-K-均值(SKK-均值)算法。

输入:数据集合 $X=\{x_i\}_{i=1}^{N}$,seed 集 $Xs=\{x_{Si}\}_{i=1}^{Ns}$,类别数 k,核参数 σ,最大迭代次数 max_ite,阈值 p。

输出:使目标函数 J^{ϕ} 为最小的 X 的一个划分 $\{X_k\}_{k=1}^{K}$,错分样本数和错误率.

步骤 1 初始化:利用 seed 集 Xs 中的有标记数据 x_{Si} 确定每个簇的初始聚类中心,得到一个初始的聚类初始划分 $\{C_k^{(0)}\}_{k=1}^{K}$,并设 ite=0。

步骤 2 E-步(1):分别计算每个数据点 x_i 到中心 k 的距离,k=1,2,…,K,采用 RBF 核函数来计算。

步骤 3 E-步(2):把 x_i 分配到与之距离最小的中心所属的簇 $C_k^{(ite)}$ 中,并计算 J^{ϕ}。

步骤 4 M-步:按照新生成的划分,对每个簇重新计算聚类中心,ite+1。

步骤 5 重复步骤 2 到步骤 4 直到前后两次目标函数之差的绝对值小于给定阈值 p,即准则函数 J^{ϕ} 收敛,输出最终聚类结果.

SKK-均值算法在初始阶段利用带有标记的部分数据生成 seed 集,并且使用 seed 集进行初始中心的选取。使得 K-均值初始聚类中心的选取不再是随机的,而是可以控制的。下面的实验表明,采用上述方法可以提高聚类算法的准确性和泛化性。

4 实验及结果分析

利用数值实验对 SKK-均值算法进行验证,并与普通的 K-均值算法和核-K-均值算法进行比较。在算法的实现过程中不同的算法使用了相同的迭代策略。实验中,随机抽取其中一定比例的标记样本作为 seed,然后再对原数据集进行聚类。所有算法都用 MATLAB 编程实现,并且选用了 3 个 UCI 数据集:IRIS 数据集,Crab 数据集和 New-Thyroid 数据集进行实验。实验环境:CPU:AMD Sempron2800+,内存:512 MB,操作系统:Windows XP,编程平台:MATLAB7.1。

4.1 SKK-均值算法用于对 IRIS 数据集聚类

本实验以标准的 IRIS 数据集作为测试样本。IRIS 数据集是在数据挖掘实验中被广泛采用的数据集,由 150 个四维样本组成,每个样本的 4 个分量分别表示 IRIS 的 SepalLength、SepalWidth、PetalLength 和 PetalWidth 指标。整个样本集包含 3 个 IRIS 种类,分别是 Setosa、Versicolor 和 Virginica,每类各有 50 个样本,其中第一类与后两类是线性可分的,后两类之间线性不可分。实验将 SKK-均值算法所得结果与 Mark Girolami 提出的算法[7]和 Francesco Camastra、Alessandro Verri 提出的算法[8]在 IRIS 数据集上的结果进行比较。

实验中分别从这 150 条数据中按不同比例取出样本组成子集 S(比例分别为 10%、20%、30%和 50%,称为标记率),然后对 S 集中的样本进行标记,得到 seed 集 X_S。再应用 SKK-均值算法用得到的 seed 集初始化聚类中心,指导聚类过程。对于每一种标记率进行 20 次重复实验,每次取的 seed 样本都不同。然后以 20 次实验的平均结果作为最终结果。实验中的核函数一律采用 RBF 核,并且合理调整 RBF 核的参数 σ。得到的实验结果如表 1 与图 1。

表 1 SKK-均值算法对 IRIS 数据集进行聚类的实验结果

		RBF 参数 σ	平均错分点数	平均错误率/(%)
普通 K-均值		–	16.45	10.97
Girolami 核-K-均值		0.5	13.50	9.00
Camastra 核-K-均值		1.1	8.00	5.33
SKK-均值	标记率 10%	0.6	11.65	7.77
	标记率 20%	0.6	7.20	4.80
	标记率 30%	0.6	6.60	4.40
	标记率 50%	0.5	5.55	3.70

图 1 中,图(a)为取 IRIS 数据集中 10%的样本作为 seed,然后用 SKK-均值算法进行聚类,得到的划分结果。实验共进行 20 次,取其中一次的实验结果进行绘图。同样的,图(b)为 20%、图(c)为 30%、图(d)为 50%的样本作为 seed 的聚类结果。

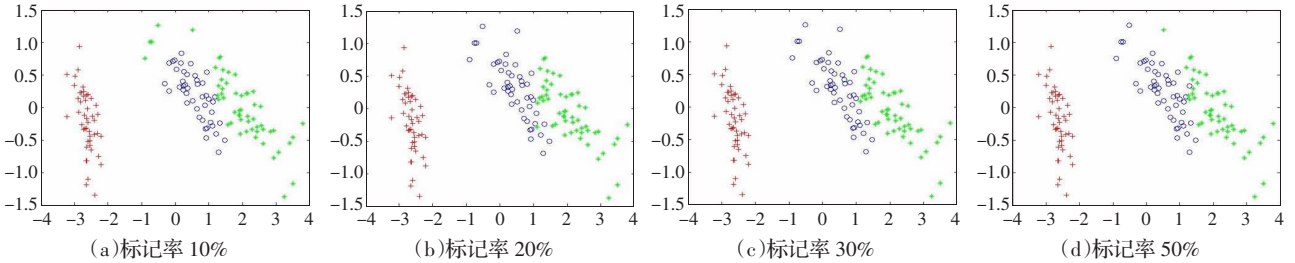
对于 IRIS 数据集,即使是在低标记率(10%,20%)的情况下,SKK-均值算法在准确性方面也优于 K-均值算法,错分点的个数在 7~12 个左右。如果标记率较高(30%,50%),那么算法得到的错分点数为 5~6 个,错分率是很低的。

4.2 SKK-均值算法用于对 Crabs 数据集聚类

Leptograpsus Crabs 数据集包括对两种蟹类动物的雌雄个体之间的 5 个方面的物理测度。该数据集具有 200 条 5 维数据,平均分为 4 类,每类 50 个样本,各类之间全都是线性不可分的。本次实验也按照前面对 IRIS 数据集的做法进行 seed 集的选取,同样对每种标记率都用不同的 seed 样本集进行 20 次实验。并且与 K-均值算法和 Girolami[7]提出的核-K-均值算法进行比较。实验结果见表 2 与图 2。

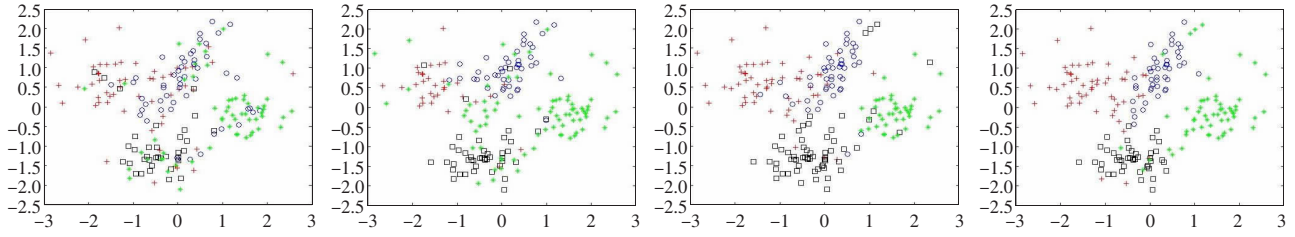
图 2 中,图(a)为取 Crabs 数据集中 10%的样本作为 seed,然后用 SKK-均值算法进行聚类,得到的划分结果。实验共进行 20 次,取其中一次的实验结果进行绘图。同样的,图(b)为 20%、图(c)为 30%、图(d)为 50%的样本作为 seed 的聚类结果。

对于 Crabs 数据集,普通 K-均值算法和核-K-均值算法都难以得到令人满意的划分。而 SKK-均值算法却可以比较有效地划分该数据集。在标记率较低(10%~20%)时,错分点数在 68~91 个左右;在标记率比较高(30%~50%)时,错分点数为



(a)标记率 10% (b)标记率 20% (c)标记率 30% (d)标记率 50%

图 1 SKK-均值算法对 IRIS 数据集进行聚类的结果



(a)标记率 10% (b)标记率 20% (c)标记率 30% (d)标记率 50%

图 2 SKK-均值算法对 Crabs 数据集进行聚类的结果

表 2　SKK–均值算法对 Crab 数据集进行聚类的实验结果

		RBF 参数 σ	平均错分点数	平均错误率/(%)
普通 K–均值		–	140.00	70.00
Girolami 核–K–均值		0.001	138.00	69.00
SKK–均值	标记率 10%	1.40	90.95	45.48
	标记率 20%	1.45	68.45	34.23
	标记率 30%	1.55	44.85	22.43
	标记率 50%	1.50	24.10	12.05

12~22 个,聚类效果比较理想。并且,在对该数据集进行的实验中,随着标记率的提高,聚类性能的提高是明显的。

4.3　用 SKK–均值算法对 New–Thyroid 数据集聚类

下面采用另一个 UCI 数据集——New–Thyroid 数据集来进一步验证 SKK–均值算法。New–Thyroid 数据集包含 215 条 5 维数据,分为 3 类,但是该数据集中每类的样本点数是不平均的。在第一类中,有 150 个样本点,第二类中有 35 个,第三类中有 30 个。对于这种不平均的数据,在 SKK–均值算法中不能仍按从 3 类中取相同百分比的样本点进行标记,这样会使 seed 集的第一类中有较多的样本,在初始化过程中有可能发生过学习现象;同样也不能按照 seed 集中各类别的样本数为 1:1:1 的比率取点,这样会造成第一类中的 seed 样本相对于无标记样本数目过少,起不到应有的监督作用。经过实验比较,采用 2:1:1 的比率进行取值,对于 New–Thyroid 数据集,这种取点比率得到的效果最好。

设抽取样本的基数为 a,用上述比率进行实验。即在第一类中取 $2a$ 个样本,第二、三类中取 a 个样本,得到种子集。再以同样的方式进行 20 次实验,求出平均值作为结果,并且与普通 K–均值算法和 Girolami[7]提出的核–K–均值算法进行比较。实验结果见表 3。

表 3　SKK–均值算法对 New–Thyroid 数据集进行聚类的实验结果

		RBF 参数 σ	平均正确划分点数	平均错误率/(%)
普通 K–均值		–	68.80	68.00
Girolami 核–K–均值		0.001	88.25	58.95
SKK–均值	a=5	0.40	155.20	27.81
	a=10	0.50	160.45	25.37
	a=15	0.40	165.50	23.02
	a=20	0.50	170.70	20.60

New–Thyroid 数据集具有各簇的样本分布不平均,数量相差较为悬殊的特点,用普通 K–均值算法和核–K–均值算法较难得到合理的划分。SKK–均值算法通过采用抽取各簇之间适当比例的有标记样本,指导聚类过程,得到了较好的聚类结果。在 seed 集样本数较低(20~40 个样本)时,错分率为 25%~28%;seed 集样本数较高(60~80 个样本)时,错分率为 20%~23%。

5　结语

本文提出 SKK–均值算法提取一定比率的数据进行标记,得到 seed 集,然后将 seed 集中的样本作为监督信息初始化聚类中心的选取,指导算法对无标记数据的划分。文章中进行的 3 个实验表明,与普通的 K–均值算法和核–K–均值算法相比,SKK–均值算法在正确率方面明显优于这些无监督算法,提高了聚类算法的准确性;并且随着标记样本数的增加,SKK–均值算法的聚类效果提高越来越明显。在对不平衡的数据集中的标记点选取以及如何处理 seed 集中的“噪声”等方面,还有待进一步研究。

参考文献:

[1] Zhu X J.Semi–supervised learning literature survey,Technical Report 1530[R].Department of Computer Sciences,University of Wisconsin at Madison,Madison,WI,December,2007.

[2] 李昆仑,张伟,代运娜.基于 Tri–training 的半监督 SVM[J].计算机工程与应用,2009,45(22).

[3] Li Kun–lun,Zhang Wei,Ma Xiao–tao,et al.A novel semi–supervised SVM based on tri–training[C]//IITA 2008.

[4] Basu S,Banerjee A,Mooney R.Semi–supervised clustering by seeding[C]//International Conference on Machine Learning,2002:19–26.

[5] Filippone M,Camastra F,Masulli F,et al.A survey of kernel and spectral methods for clustering[J].Pattern Recognition,2008,41:176–190.

[6] Kulis B,Basu S,Dhillon I S,et al.Semi–supervised graph clustering:A kernel approach[C]//Proceedings of the 22nd International Conference on Machine Learning,ICML'05.New York,NY,USA ACM Press,2005:457–464.

[7] Girolami M.Mercer kernel–based clustering in feature space[J].IEEE Transactions on Neural Networks,2002,13(3):780–784.

[8] Camastra F,Verri A.A novel kernel method for clustering[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2005,27(5):801–805.

[9] Dempster A P,Laird N M,Rubin D B.Maximum likelihood from incomplete data via the EM algorithm[J].J Royal Statistical Soc,1977,39(1):1–38.

[10] UCI repository of machine learning databases.http://www.ics.uci.edu/mlearn/MLRepository.html.

(上接 98 页)

参考文献:

[1] Azzedin F,Maheswaran M.Evolving and managing trust in grid computing systems[C]//IEEE Canadian Conference on Electrical & Computer Engineering(CCECE'02),May 2002:1424–1429.

[2] 李小勇,桂小林.大规模分布式环境下动态信任模型研究[J].软件学报,2007,18(6):1510–1521.

[3] Varalakshmi P,Thamarai S,Kanchana P,et al.A robust trust model with rated genuine feedbacks in a grid environment[C]//International Conference on Computational Intelligence and Multimedia Applications,2007:412–419.

[4] Yi Chen,Luo Jun–zhou,Ni Xu–dong.A fuzzy trust evaluation based access control in grid environment [C]//The Third China Grid Annual Conference,IEEE 2008:190–196.

[5] 马满福,姚军.网格环境的一种跨域信任模型[J].计算机应用,2008,28(9):2357–2359.

[6] Hasan O,Brunie L,Pierson J,et al.Elimination of subjectivity from trust recommendation(R),TR–08–008.Department of Computer Science,Purdue University,USA,March 17,2008.

[7] Woodas W K Lai,Kam–Wing Ng.A time–frame based trust model for grids[OL].Grid and Cooperative Computing–GCC,2005.http://www.springerlink.com/content/y85r620810101275/.

[8] 刘兵,汪卫,施伯乐.基于小波变换的序列间距离严格估算[J].计算机研究与发展,2006,43(10):1732–1737.

[9] Azzedin F,Maheswaran M,Mitra A.Trust brokering and its use for resource matchmaking in public resource grids[J].Grid Computing,2006(4):247–263.