doi:103969/j. issn. 0490-6756. 2016. 03. 011

一种新的 WSN 故障数据挖掘算法

李晓晨，宋正江
(浙江工业职业技术学院，绍兴 312000)

摘　要：为了有效提高无线传感器网络故障数据的判别能力，在以往的研究基础上，本文结合菌群优化算法提出了一种新的挖掘方法 FDMBFO(Fault Data Mining algorithm based on Bacteria Foraging Optimization). 该算法首先通过小波变换和关联系数给出了故障数据分布区间的划分方法，建立了目标挖掘函数，同时利用菌群优化算法实现对目标函数的求解. 最后，通过实际样本数据进行仿真实验，深入分析了影响 FDMBFO 算法的关键因素，并对比研究了 FDMBFO 算法与其它算法之间的性能状况，结果发现 FDMBFO 算法具有较好的适应性.

关键词：无线传感器网络；故障；数据挖掘；分布区间；菌群优化；小波变换
中图分类号：TP309　　**文献标识码**：A　　**文章编号**：0490-6756(2016)02-0305-06

A new fault data mining algorithm of WSN

LI Xiao-Chen, *SONG Zheng-Jiang*
(Zhejiang Industry Polytechnic College, Shaoxing 312000, China)

Abstract: In order to effectively improve the identification ability for fault data of wireless sensor network, a novel mining algorithm FDMBFO (Fault Data Mining algorithm based on Bacteria Foraging Optimization) is proposed by bacteria foraging optimization. In this algorithm, the division method of distribution range is given with wavelet transform and correlation coefficient, and the objective mining function is built. Then, the solving of function is presented by bacteria foraging optimization. Finally, a simulation with actual sample data was conducted to study the key factors of FDMBFO. Compared to performance of other algorithm, the results show that, FDMBFO has better adaptability.

Key words: Wireless sensor network; Fault; Data mining; Distribution range; Bacteria foraging optimization; Wavelet transform

1　引　言

自上世纪九十年代数据挖掘技术被提出后，便在无线传感器网络(Wireless Sensor Network, WSN)领域中得到迅速应用[1-4]. 对无线传感器网络进行数据挖掘的目的是在于从海量网络数据中寻找有意义的模式，模式可以是一组规则，聚类，决策树，依赖网络或其他信息[5-10]. 目前，数据挖掘主要包括四种类型：相关依赖关系的发现、类别判定、类别描述、异常数据挖掘. 前三种是针对绝大部分数据均服从的数据模式，而异常数据挖掘在于找出海量数据中相对孤立的异常数据模式，主要应用于故障和噪声的检测，特别是依据其数据状态及参数的变化来定位和判断故障源.

数据挖掘常用方法有：模糊方法、粗糙集理论、云模型，证据理论、归纳学习、遗传算法、人工神经

收稿日期：2014-12-03
基金项目：浙江省自然科学基金(y1080023)
作者简介：李晓晨(1981－)，女，讲师，硕士，研究方向为数据挖掘.
通讯作者：宋正江. E-mail: songzj1982@163. com.

网络及其它人工智能方法,等等.其中,分类算法作为数据挖掘中最重要的技术之一[11-13],它从数据集中提取描述分类特征指标,并利用指标把数据对象都归纳入某个已知类别中.它包含两个阶段:(1)构造分类模型;(2)利用分类模型将实际样本划分到相应类别.对于分类模型构造一般划分为训练和测试两个步骤.训练阶段主要是分析研究训练数据集的特点,为每个类别建立相应的数据描述.在测试阶段,利用类别描述对实际数据进行分类,测试其分类准确度.但是,对于构造分类模型的难点在于如何最优地划分分类区间,分类区间过大或者过小都将直接影响数据挖掘性能.

在以往工作的基础上,本文结合种群优化方法提出了一种新的 WSN 故障挖掘算法,首先利用小波变换降低故障数据的突发性,并基于关联系数给出故障数据分布区间的划分方法,和目标挖掘函数,同时采用菌群优化[14-16]实现对目标函数的求解.

2 故障数据挖掘方法

假设某时刻 t 存在 WSN 故障数据序列 $X=[x_1, x_2, \cdots, x_n]$,这类数据较正常数据而言通常具有高突发性,因此首先结合小波变换降低故障数据的突发性,然后将其进行对比,确定故障数据分类区间.根据式(1)将故障数据序列 X 进行小波变换,获得第 L 层的近似系数 a_{LN} 和小波系数 d_{LN}:

$$\begin{cases}\sqrt{2}d_{LN} = a_{L+1,2N} - a_{L+1,2N+1}\\ \sqrt{2}a_{LN} = a_{L+1,2N} + a_{L+1,2N+1}\end{cases} \tag{1}$$

其中,N 表示小波系数个数.对获得的近似系数 a_{jk} 和小波系数 d_{jk} 进行重构,得到消除突发性后的数据黄(t).同时结合式(2),利用黄(t)将故障数据 $X(t)$ 进行标准化处理,以此避免故障数据区间不均匀、过多冗余的缺点:

$$X = \frac{X - \dot{X}}{\sqrt{\frac{1}{n-1}\sum_t (X(t) - \dot{X}(t))^2}} \tag{2}$$

同时按照式(3)计算故障数据的关联系数 R,如果大于规定分区阈值 B,则按照上述方法继续缩小划分区间,直至大于规定阈值 B.

$$R = \frac{\sum_t ((x_1(t) - \dot{x}_1(t))((x_2(t) - \dot{x}_2(t))\cdots((x_i(t) - \dot{x}_i(t))}{\sqrt{\sum_t (x_1(t) - \dot{x}_1(t)) \sum_t ((x_2(t) - \dot{x}_2(t))\cdots \sum_t (x_i(t) - \dot{x}_i(t))}} \tag{3}$$

其中,$x_i(t)$表示 t 时刻故障数据 $X(t)$第 i 个样本属性值.

同时,这里定义样本方差 S^2、平均差 V 和熵 H 来优化故障数据挖掘性能,具体如下式所示.

$$S^2(x_i(t)) = \frac{1}{n-1}\sum_{i=1}^{n} (x_i(t) - \frac{1}{n}\sum_{i=1}^{n} x_i(t))^2 \tag{4}$$

$$V(x_i(t)) = \frac{1}{n}\sum_{i=1}^{n} (x_i(t) - \frac{1}{n}\sum_{i=1}^{n} x_i(t)) \tag{5}$$

其中,$\frac{1}{n}\sum_{i=1}^{n} x_i(t)$ 表示 t 时刻故障数据第 i 个样本属性的平均值;n 为样本数量.

并且对故障数据的熵 H 进行归一化处理,可得

$$H(x_i(t)) = -\frac{\sum_{i=1}^{n} p(x_i(t)) \log_2 p(x_i(t))}{\frac{1}{n}\log_2 n} \tag{6}$$

其中,$p(x_i(t))$为 t 时刻故障数据 $X(t)$第 i 个样本属性 $x_i(t)$的发生概率.因此,结合样本方差 S^2、平均差 V 和熵 H,这里建立目标函数 Z,以此获得故障数据 $X(t)$在所有组合下的最大偏差之和,如下式.

$$Z(x_i(t)) = \sum_{i=1}^{n} (\alpha S^2(x_i(t)) + \beta V(x_i(t)) + \gamma H(x_i(t))) \tag{7}$$

其中,$\alpha+\beta+\gamma=1$,并且$\alpha \geqslant 0, \beta \geqslant 0, \gamma \geqslant 0$.

对于上述目标函数 Z 的求解是一个非线性优化问题,常用的规划算法不能有效获得最优解,或者存在计算量太大的确定.本文曾经结合人工蜂群算法建立了上述目标函数 Z 的求解算法 FDMABC(Fault Data Mining algorithm based on Artificial Bee Colony),具体如图 1 所示.但是由于人工蜂群算法中涉及的个体种类较多,并且蜜源的适应度值难以确定.因此,本文结合一种新的种群智能方法——菌群优化算法来实现目标函数 Z 的求解.

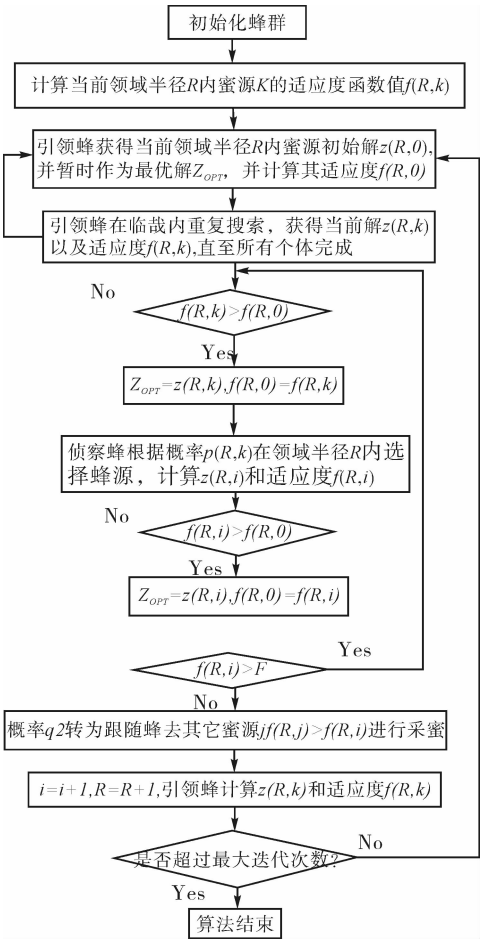

图 1　FDMABC 算法流程

Fig. 1　The algorithmic process of FDMABC

3　菌群优化算法

菌群优化(Bacteria Foraging Optimization)算法是基于杆菌在人体肠道内觅食和繁殖而提出的一种新型种群智能优化方法. 它主要包括趋化操作、复制操作和迁徙操作[17, 18]. 其中，趋化操作体现种群之间的协同性，复制操作根据健康度进行繁殖，迁徙操作则以一定概率将个体位置进行移动，以获得全局最优值. 对于上述建立的目标函数 Z，这里结合菌群优化算法建立一种新的求解方法 FDMBFO(Fault Data Mining algorithm based on Bacteria Foraging Optimization)，具体步骤如下.

(1) 在初始时刻 T 设置菌群规模 M，搜索邻域半径 r，最大迭代次数 MAX；

(2) 将故障数据 $X(t)=[x_1(t), x_2(t), \cdots, x_n(t)]$ 视作菌群个体，$X(t)$ 对应的目标函数 Z 视作个体适应度，随机确定种群中各个体位置 S，并对于个体 i 按照式(8)对其位置 S_i 进行更新.

$$S_i = S_i + d_i \tag{8}$$

其中，d_i 表示移动步长，并受到周围邻域半径 r 内其它 m_i 个个体影响.

$$d_i = \frac{1}{1 + e^{\frac{1}{m_i - 1}} \sum_{j=1, j\neq i}^{m_i} (S_i - S_j)} \tag{9}$$

(3) 执行趋化操作：为了防止个体过分聚集造成局部最优，这里设置邻域半径 r 内个体之间的平均距离 d_r，如果 $d_i < d_r$，那么在邻域半径 r 内重新生成位置 S_i，跳转到步骤(2)重新执行更新操作.

$$d_r = \frac{1}{m_i} \sum_i \sum_j (S_i - S_j) \tag{10}$$

(4) 由式(11)计算邻域半径 r 内个体 i 的适应度 f_i，并将其最大适应度即记为 $f_{\max}$.

$$f_i = \frac{1}{1 + z_i} \tag{11}$$

(5) 根据上述趋化操作结果，定义健康度 $w(S_i)$ 为位置 S_i 处完成更新操作的个体适应度之和，如下式.

$$w(S_i) = \sum_i f_i(S_i) \tag{12}$$

其中，$f_i(S_i)$ 表示在位置 S_i 处完成更新操作的个体适应度.

(6) 执行复制操作：针对各位置的健康度 $w(S_i)$，这里从高到底进行排序，得到降序数组 $W=[w(S_1), w(S_2), \cdots, w(S_M)]$，同时根据式(13)淘汰最后 k 位对应位置的个体，并采用前 k 位对应位置的个体进行替代.

$$k = (1 - 0.5\text{rand}())M \tag{13}$$

其中，rand()产生(0, 1)之间的随机数.

(7) 执行迁徙操作：在邻域半径 r 内，按照式(15)给定概率 p 和移动步长 d_i 对个体 i 执行迁徙操作，将其移动到新的位置进行更新，以此获得全局最大适应度 $f_{\max}$.

$$p = \frac{1 + f_i{}^{\varphi} w(S_i)^{\varnothing}}{r + \sum_i f_i{}^{\varphi} w(S_i)^{\varnothing}} \tag{14}$$

其中，$0<\varnothing<1$；$0<\varphi<1$；

(8) 令 $T=T+1$，判断是否超出最大迭代次数 MAX，如果没有则跳转到步骤(2)进行下一次迭代操作，否则跳转到步骤(9)；

(9) 输出当前最大适应度 $f_{\max}$，即为目标函数 Z 最优值，算法结束.

对于参数 r 而言，它对算法收敛和多样性起着关键作用. 如果 r 较小将会使得细菌在极其有限的区域搜寻，容易陷入局部最优；反之如果 r 较大容易造成细菌个体跳过最优区域而转向局部最优值. 而

对于概率 p 起着对解空间以一定概率进行不精确搜索作用,当个体定位到的解的精度达到一定的阈值时,个体会进入开发状态;如果对该区域的搜索失败,个体会重新以较大的趋化概率迅速离开该区域,并开始对新的区域进行搜索.

4 实验分析

为了验证上述 WSN 故障数据挖掘算法 FDMBFO 的有效性,这里进行仿真实验. 首先在某企业提取了 3000 条 WSN 运行数据指标,包括丢包率(正常范围 10~15)、吞吐量(正常范围 240~300)和能耗,表 1 给出了部分 WSN 数据信息. 首先将前 2000 条数据作为先验信息,按照第 1 节给出的方法建立故障分类指标,总共建立 30 种故障类别;其次,再将后 1000 条数据作为样本,观察 FDMBFO 算法与 FDMABC 算法、以及人工神经网络算法 ANN(Artificial Neural Network)判断故障种类的准确率. 图 2 给出了对比结果. 从图 2 可以看出,本文提出的 FDMBFO 算法与实际样本信息更加吻合. 这里定义如下计算方差ε.

$$\varepsilon = (x_i(t) - \tilde{x}_i(t))/\tilde{x}_i(t) \tag{15}$$

其中,$x_i(t)$表示算法获得的样本数值;$\tilde{x}_i(t)$ 实际样本数值. 经过数据分析,FDMBFO 算法、FDMABC 算法和 ANN 算法的误差率分别为 11.72%、13.15%和 14.22%.

图 3 给出了这三种算法的平均判断误差率与分区阈值 B 之间的关系. 从图 3 可以看出,随着分区阈值 B 的增加,三种算法的平均判断误差率降低. 这与通常的理解是一致的,分区阈值 B 越大,在某区间内容纳的样本种类越多,出现误判的几率减小,从而误差率降低. 此外可以发现,当分区阈值 B 较小时,FDMBFO 算法的平均判断误差率低于 FDMABC 算法和 ANN 算法,而当分区阈值 B 较大($B \geqslant 0.5$)时,两种算法的平均判断误差率较为接近. 这说明 FDMABC 算法在处理较为细致的故障分类时存在一些不足.

同时,本文对 FDMBFO 算法与 FDMABC 算法的抗突发能力进行了比较. 这里将样本数据施加一系列强突发噪声,图 4 给出了两种算法的平均判断误差率与突发参数 H 之间的关系. 由于两种算法均采用小波变换来消除突发影响,所以其平均判断误差率均未受到明显影响,与无强突发噪声状态下的性能相近.

表 1 部分 WSN 样本数据

Tab. 1 A part of WSN data

样本	丢包率(%)	吞吐量	能耗	故障编号
8	14	213	偏高	5
52	11	256	偏高	24
98	21	226	正常	16
165	20	237	正常	19
218	16	245	偏高	23
327	10	228	正常	18
434	19	239	偏高	25
501	16	227	偏高	6
592	19	231	正常	11
686	17	225	偏高	17
762	13	201	正常	13
815	13	244	偏高	2
873	15	240	偏高	22
920	17	253	正常	12
981	18	242	偏高	7

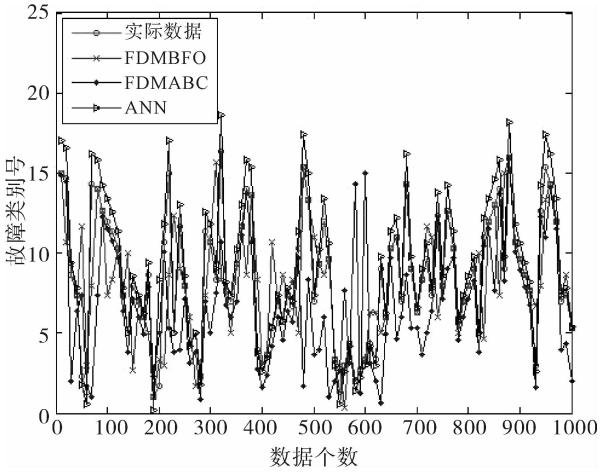

图 2 故障判断结果对比

Fig. 2 Compare of fault judgment results

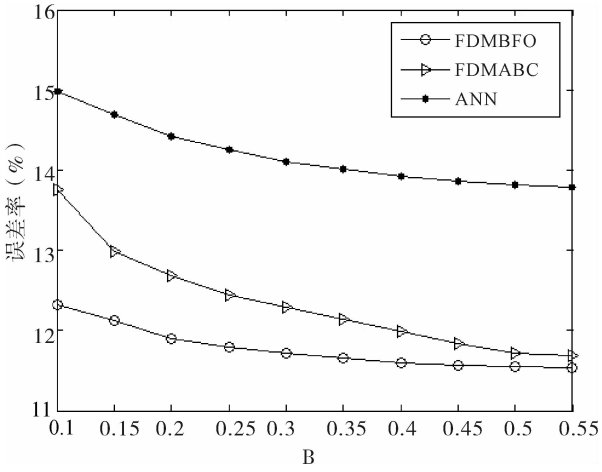

图 3 误差率与分区阈值 B 之间的关系

Fig. 3 The relation between error rate and partition threshold B

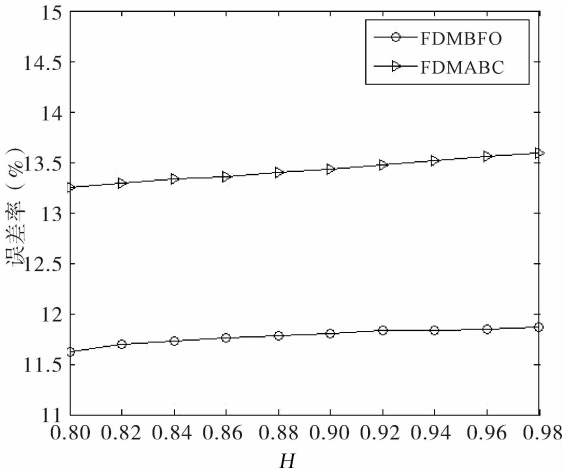

图 4　误差率与突发参数 H 之间的关系

Fig. 4　The relation between error rate and burst factor H

最后，为了进一步研究 FDMBFO 算法的影响因素，本文针对菌群优化算法关键参数进行分析. 图 5 给出了不同邻域半径 r 下 FDMBFO 算法的判断误差情况. 从图 5 可以看出，当在仿真开始阶段，邻域半径 r 越小对应的误差越小，而在后期出现突变，邻域半径 r 越小对应的误差越大. 从图 5 可以看出，当在仿真开始阶段，邻域半径 r 越小对应的误差越小，而在后期出现突变，邻域半径 r 越小对应的误差越大. 在开始阶段，种群整体性能较低，此时需要加大个体的更新速度才能尽快使得目标函数收敛，获得全局最优. 根据式(14)可以发现，邻域半径 r 越小迁徙操作概率越大，此时个体更新的可能性越大，对应的系统性能较能大幅度提高. 而后期随着种群性能的普遍提高，邻域半径 r 越大意味着邻域半径 r 内的个体越多，此时趋化的范围加大，可以有效避免陷入局部最优，因而邻域半径 r 越大对应的误差越小. 同时，图 6 给出了 FDMBFO 算法的平均判断误差率与迁徙概率 p 之间的关系. 从图 6 可以看出，随着迁徙概率 p 的增加，误差率呈现递减趋势，直至趋于稳定. 这说明迁徙概率 p 对改善误差率起到了一定作用，但是当迁徙概率 p 增加到一定程度时（如 $p\approx 0.8$），进一步提高迁徙概率 p 对误差率的影响较小. 此时应该考虑通过改善其它因素来降低误差率.

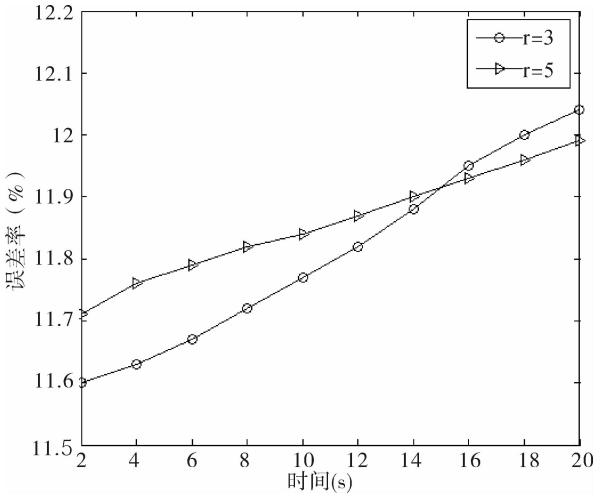

图 5　不同邻域半径 r 下误差率比较

Fig. 5　Compare of error rate in the different neighborhood radius r

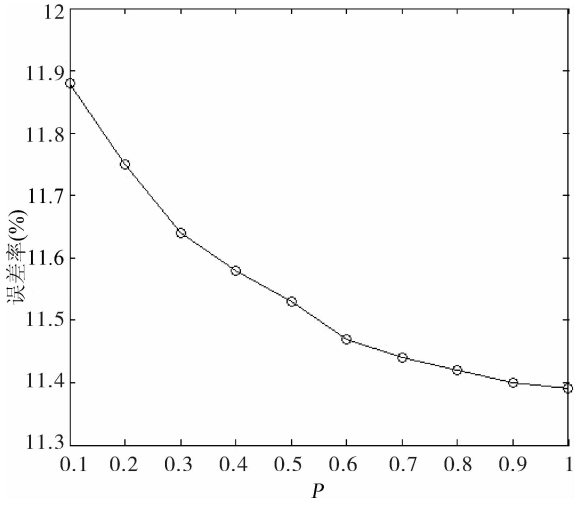

图 6　误差率与迁徙概率 p 之间的关系

Fig. 6　The relation between error rate and migration probability p

5　结束语

针对 WSN 故障数据挖掘性能问题，本文在以往的研究基础上，结合菌群优化算法提出了一种新的故障挖掘算法FDMBFO. 该算法首先基于小波变换降低故障数据的突发性，并通过关联系数 R 划分故障数据分布区间，以此建立故障数据的目标挖掘函数 Z. 最后利用实际样本数据，通过仿真实验对比研究了 FDMBFO 算法与 FDMABC 算法、ANN 算法之间的性能状况，结果发现 FDMBFO 算法较 FDMABC 算法有所改善. 在今后研究中，可以考虑结合粗糙集和云模型来完善 WSN 故障数据挖掘算法.

参考文献：

[1]　Yick J, Mukherjee B, Ghosald. Wireless sensor network survey [J]. IEEE Comput Networks, 2008, 52(12): 2292.

[2]　蔚赵春，周水庚，关佶红. 无线传感器网络中数据

存储与访问研究进展[J]. 电子学报，2008，36(10)：2001.

[3] 刘亮，秦小麟，刘亚丽，等. 顽健的无线传感器网络K近邻查询处理算法[J]. 通信学报，2010，31(11)：171.

[4] 肖伟，徐明，吕品，等. 无线传感器网络事件簇的数据聚集容错机制[J]. 通信学报，2010，31(6)：112.

[5] 李光，王亚东. 一种改进的基于奇异值分解的隐私保持分类挖掘方法[J]. 电子学报，2012，40(4)：739.

[6] 陈玉明，吴克寿，李向军. 一种基于信息熵的异常数据挖掘算法[J]. 控制与决策，2013，28(6)：867.

[7] 吴芝明，钱程，伍少梅. 关联规则挖掘的PredictiveApriori算法的研究及改进[J]. 四川大学学报：自然科学版，2012，49(1)：97.

[8] He D，Li R Y，Zhu J D. Plastic bearing fault diagnosis based on a two-step data mining approach[J]. IEEE Trans Industr Electr，2013，60(8)：3429.

[9] Wei X P，Kusiak A，Sadat H R. Prediction of influent flow rate：data-mining approach[J]. J Energ Eng-Asce，2013，139(2)：118.

[10] Velasquez J D. Web mining and privacy concerns：Some important legal issues to be consider before applying any data and information extraction technique in web-based environments[J]. Expert Syst Appl，2013，40(13)：5228.

[11] 霍纬纲，邵秀丽. 一种基于多目标进化算法的模糊关联分类方法[J]. 计算机研究与发展，2011，48(4)：567.

[12] 何波. 基于频繁模式树的分布式关联规则挖掘算法[J]. 控制与决策，2012，27(4)：618.

[13] 熊伟清. 基于二元蚁群优化算法的分类规则挖掘[J]. 模式识别与人工智能，2008，21(4)：500.

[14] 章国勇，伍永刚，谭宇翔. 一种具有量子行为的细菌觅食优化算法[J]. 电子与信息学报，2013，35(3)：614.

[15] 王雪松，程玉虎，郝名林. 基于细菌觅食行为的分布估计算法在预测控制中的应用[J]. 电子学报. 2010，38(2)：334.

[16] Chatzis S P，Koukas S. Numerical optimization using synergetic swarms of foraging bacterial populations[J]. Expert Syst Appl，2011，38(12)：15332.

[17] Mishra S. A hybrid least square-fuzzy bacteria foraging strategy for harmonic estimation[J]. IEEE Trans Evol Comput，2005，9(1)：61.

[18] Saber A Y. Economic dispatch using particle swarm optimization with bacterial foraging effect[J]. Elec Power Energ Syst，2012，34(1)：38.