文章编号:1671-4598(2013)08-2181-04　　中图分类号:TP393.09　　文献标识码:A

改进贝叶斯垃圾邮件过滤技术的研究

计　宏
(西安科技大学 网络中心，西安　710054)

摘要：为提高贝叶斯垃圾邮件过滤器的精确率和召回率，提出一种改进加权贝叶斯模型（improved weighted bayes model，IWB），通过提高贝叶斯模型的准确性，改善垃圾邮件过滤性能；不同于朴素贝叶斯模型（naïve bayes model，NB）对邮件样本特征值所作的独立性和相同重要性的假设，通过给邮件样本的每一个特征值分配一个权值，减小贝叶斯模型与实际间的失配误差；根据贝叶斯公式建立基于最小二乘算法的目标函数，用于对IWB中权向量的优化；由于目标函数为非线性高维函数，提出一种新的粒子群优化算法，能够获得近似全局最优权向量，从而得到最优贝叶斯模型；通过仿真对NB、传统加权贝叶斯模型（weighted bayes model，WB）与IWB进行比较，仿真结果表明IWB能够显著地改善垃圾邮件过滤性能，提高邮件过滤的精确率和召回率。

关键词：垃圾邮件；贝叶斯；精确率；加权；粒子群

Research on Improved Bayesian Algorithm for Anti-Spam Filtering

Ji Hong
(Network Center Xi'an University of Science and Technology，Xian　710054，China)

Abstract：An improved weighted bayes model (IWB) is proposed in order to increase precision and recall. and anti-spam filtering performance is increased due to higher accuracy of bayes models. Differing from the assumption of naïve bayes (NB) model about that all attributes have the same independency and importance，each attribute of mail samples is assigned with a weight in order to decrease the mismatch error between bayes model and reality. Based on least squares method，the objective function is established with bayes formula to optimize the weight vector of IWB. Because the objective function is nonlinear and multivariable，a novel particle swarm optimization (PSO) method is proposed to obtain the approximately global optimal weight vector and obtain the optimal bayes model. Comparing IWB with NB and weighted bayes model (WB)，the simulation results show that IWB remarkably improves the anti-spam filtering performance and increases precision and recall.

Key words：spam；Bayesian；precision；weighted；PSO

0　引言

当今，电子邮件成为人们生活中越来越重要的一种通信手段。在使用电子邮件时，人们越来越关注垃圾邮件带来的问题。垃圾邮件不仅占用网络和系统资源，耗费用户精力，而且有可能携带不良信息甚至是病毒[1]。针对这一问题，垃圾邮件过滤技术越来越引起人们的关注。

贝叶斯方法是一个自适应统计智能技术。贝叶斯方法之所以备受瞩目是因为它具有高准确率与低假阳性率（false-positive rate），使其成为目前最有用的垃圾邮件过滤技术之一[2-3]。

但是，由于朴素贝叶斯模型（naïve bayes model，NB）在学习过程中假设全部特征值具有相同的重要性和独立性，这一假设与实际情况是不相符的，影响其垃圾邮件过滤性能。为了解决上述问题，学者提出很多改进朴素贝叶斯方法。两种最有代表性的方法是特征值权重法和贝叶斯网络分类器。贝叶斯网络分类器直接对NB的结构进行扩展，通过增加不同特征值间的依赖关系克服NB的缺点[4]，但是无法避免进行网络结构学习[5]。因此特征值权重法成为一个研究热点。

特征值权重法的核心是关于权向量的搜索方法。爬山法是一种局部搜索优化算法，其分别沿搜索空间的轴方向进行搜索，具有算法简单、搜索速度快的特点[6]。但是，在高维搜索空间、目标函数比较复杂的情况下，爬山法无法得到全局最优解，并且常常会出现搜索路径呈之字形、搜索步距减小，搜索时间增加。文献[7]采用最小二乘法确定目标函数中的权向量。为了减小计算目标函数偏导数的复杂度，文献对贝叶斯后验概率公式进行了简化及假设，在简化算法的同时降低了贝叶斯模型的准确性。文献[8]提出支持向量机法来对目标函数的权值进行寻优，从而减小过度拟合的危险性，改善分类器的性能。粒子群优化（particle swarm optimization，PSO）算法是一种基于种群搜索的自适应迭代优化技术[10]。但是，其存在收敛早熟和收敛速度慢的缺点，尤其是对于多变量目标函数。根据精确的加权贝叶斯模型提出二次型目标函数，利用改进PSO算法对加权贝叶斯模型中的权值进行全局寻优，得到更符合现实情况的加权贝叶斯模型，以此作为垃圾邮件分类器。

1　改进加权 Bayesian 模型

贝叶斯方法提供了一种概率的推理手段。在垃圾邮件过滤应用中，可以根据邮件样本和邮件类的先验概率估计邮件类后验概率。贝叶斯法则的公式表达形式可以写为[9]：

$$\varepsilon_j = P(c \mid x_j) = \frac{P(c) \cdot P(x_j \mid c)}{P(x_j)} \tag{1}$$

其中，$x_j = (x_{1j}, \cdots, x_{mj})$ 表示邮件样本 j 的 m 维特征向量，x_{ij} 表示邮件样本 j 中特征值 X_i 的值，$j=1, \cdots, n$，$i=1, \cdots, m$。c 为邮件类，分别为垃圾邮件类 c_s 和合法邮件类 c_h。$P(c)$、$P(x_j)$ 和 $P(x_j \mid c)$ 分别表示邮件类的先验概率、邮件样本的先验概率和在邮件类确定的情况下，邮件样

收稿日期:2013-03-16；　修回日期:2013-05-20。
基金项目:教育部博士点基金项目(20106121110003)。
作者简介:计　宏(1978-),女,工程师,博士生,主要从事校园网建设和维护,软件工程开发和研究工作。

本的先验概率；$P(c \mid x_j)$ 是关于 c 的后验概率。ε_j 可以显示邮件样本 j 为垃圾邮件的概率。如果 ε_j 接近于 1，则 j 确定是一个垃圾邮件；而如果 ε_j 接近于 0，则 j 更可能是一个合法邮件。因此，可以给 ε_j 设定一个阈值，用于调节辨识垃圾邮件的灵敏度。同时，给定阈值对于权衡邮件分类的假阴性（将垃圾邮件错误地划分为正常邮件）和假阳性（将正常邮件错误地划分为垃圾邮件）起着重要作用。

c_s 的先验概率可以很容易通过邮件样本集进行估计：

$$P(c_s)=\frac{\text{样本集合中属于 } c_s \text{ 的样本个数}}{\text{样本集合中总的样本个数}} \tag{2}$$

然而，在实践中应用贝叶斯方法的难度在于，一般情况下确定 $P(x_j \mid c)$ 的计算代价比较大。为了解决这个问题，提出朴素贝叶斯法则。与之对应的分类器基于一个简单的假设：邮件特征值之间相互条件独立，且具有相同重要性，即

$$p(x_j \mid c)=\prod_{i=1}^{m} p(x_{ij} \mid c) \tag{3}$$

进一步，由于式（1）中 $P(x_j)$ 为常值，通常式（1）简化为

$$\varepsilon_j = P(c \mid x_j) = P(c) \cdot P(x_j \mid c) \tag{4}$$

基于最大似然概率假设，即当邮件样本属于正确的邮件类时，其对应的后验概率最大；进一步，当 $P(c_s \mid x_j) > P(c_h \mid x_j)$ 时，邮件样本被认为是属于垃圾邮件类 c_s。

由于朴素贝叶斯法则中的条件独立假设在现实中通常是不成立的，因此学者提出特征值权重法，为每一个特征值乘上一个权重因子，其形式如下[6]

$$\varepsilon_j = P(c_s) \cdot \prod_{i=1}^{m} p^{\omega_i}(x_{ij} \mid c_s) \tag{5}$$

其中，ω_i 是对应每一个特征值的权重。但是，由于式（5）中 $\varepsilon_j \notin [0,1]$，使得无法得到最优权向量。为了提高加权贝叶斯模型的准确性，定义加权贝叶斯模型为

$$\varepsilon_j = \frac{P(c_s) \cdot \prod_{i=1}^{m} p^{\omega_i}(x_{ij} \mid c_s)}{P(x_j)} \tag{6}$$

当 ε_j 越接近 1 或者 0，分类结果越可信。在理想情况下可得

$$\varepsilon_{opt} = \begin{cases} 1, & x_j \in c_s \\ 0, & x_j \in c_h \end{cases} \tag{7}$$

实际应用中，ε_j 无法达到理想值，然而可以通过对邮件样本的训练使贝叶斯模型接近理想模型。采用最小二乘方法可以得到关于权值的目标函数

$$f(\omega) = \min \sum_{j=1}^{n} (\varepsilon_{opt} - \varepsilon_j)^2 \tag{8}$$

通过对式（8）求最小值，可以得到最优权向量。

2　改进粒子群优化算法

由式（6）可知，式（8）是一个 n 维非线性方程，为了获得（近似）全局最优权向量，同时克服标准 PSO 的缺点，提出改进 PSO 算法。

标准 PSO 算法首先在目标函数 $f(x)$ 的 n 维可行域内随机初始化粒子群中的每一个粒子

$$x_i(k) = (r_{\max} - r_{\min}) rand + r_{\min}, i = 1, \cdots, 5 \tag{9}$$

其中，$r_{\max}, r_{\min} \in \boldsymbol{R}^n$ 为目标函数可行域的上、下限向量；$rand$ 为服从 [0，1] 分布的随机数。第 i 个粒子的当前状态由一个 n 维位置向量 $x_i = (x_{i1} x_{i2} \cdots x_{in})$ 和一个 n 维速度向量 $v_i = (v_{i1} v_{i2} \cdots v_{id})$ 来表示。位置向量决定粒子在可行域中的当前位置，速度向量决定粒子进化飞行的方向和距离。每一个粒子根据目标函数值确定自己到目前为止发现的最好位置 $l_i = (l_{i1} l_{i2} \cdots l_{in})$，这被视为粒子自己的飞行经验。对于最小值优化问题，更新方程可以表示为

$$l_i(k+1) = \begin{cases} l_i(k), & f(x_i(k+1)) \geqslant f(l_i(k)) \\ x_i(k+1), & f(x_i(k+1)) < f(l_i(k)) \end{cases} \tag{10}$$

其中，k 为迭代次数。另外，通过比较每一个粒子的适应度，确定到目前为止整个粒子群中所有粒子发现的最好位置 $l_g = (l_{g1} l_{g2} \cdots l_{gn})$，这被视为粒子同伴的飞行经验，更新方程可以表示为

$$l_g(k+1) = \min(l_i(k+1) \mid f(l_i)), i = 1, 2, \cdots, s \tag{11}$$

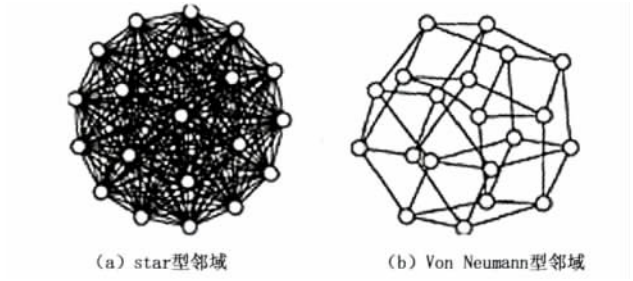
其中，s 为粒子群中粒子个数。粒子通过学习自己的经验和同伴的经验来决定下一步的运动。粒子 i 飞行轨迹由下面更新方程决定

$$v_{ij}(k+1) = \alpha v_{ij}(k) + c_1 r_1 [l_{ij} - x_{ij}(k)] + c_2 r_2 [l_{gj} - x_{ij}(k)] \tag{12}$$

$$x_{ij}(k+1) = x_{ij}(k) + v_{ij}(k+1), j = 1, 2, \cdots, n \tag{13}$$

其中，α 为惯性系数；c_1，c_2 为正的加速系数；r_1，r_2 是服从 [0，1] 分布的随机数；j 是更新的坐标。根据式（12）、（13），每一个粒子逐一更新速度、位置向量的每一维。随着进化过程，粒子飞越可行域空间，不断更新速度、位置向量，寻找自己 的最优位置。

在标准 PSO 中，粒子间的通信方式如图 1（a），这使得 PSO 收敛速度很快，但是易于陷入收敛早熟，即搜索过程停滞于局部最优值。通过对各种标准测试函数的测试，Neumann 邻域通过构控制粒子间的信息流量来平衡局部和全局搜索能力，性能最好[11]，其拓扑结构如图 1（b）。



（a）star型邻域　　（b）Von Neumann型邻域

图 1　粒子群拓扑结构图

在 Neumann PSO 基础上提出动态重置 PSO（RPSO）。此时，将每一个 Neumann 邻域视为一个子群，定义子群最优值 l_l，并且在式中用 l_l 代替 l_g。将粒子群搜索过程分为两个阶段："全局搜索"阶段表示当前每一个子群都不存在收敛早熟的趋势，粒子群保持丰富的多样性。此时，采用 Neumann PSO 算法既有利于保持粒子群的多样性，增加搜索全局最优值能力，又能够保证较快的收敛速度。"动态重置"阶段表示当前至少有一个子群的 l_l 收敛到极值点，并认为子群发生收敛早熟。此时，对发生收敛早熟的子群进行重置操作，不改变其它子群的搜索状态。动态重置策略不仅消除了子群收敛早熟问题，同时保证其它子群充分利用当前信息，减小重复搜索的概率，提高搜索效率。

为了实现动态重置，定义子群重置条件，称为最大群半径

(Maximum Swarm Radius，MSR)：在第 k 次迭代，群半径 $\delta(k)$ 定义为子群中任意粒子与 l_l 之间的最大欧几里得距离

$$\delta(k) = \max_{i\in\{1,\cdots,s\}} \| x_i(k) - l_l(k) \| \tag{14}$$

其中，$\| \ \|$ 为欧几里得范数。$diam(D)$ 表示可行域直径，对于 n 维可行域：

$$diam(D) = \| [d_1, \cdots, d_n] \| \tag{15}$$

其中，$d_i, i = 1, \cdots n$ 为可行域的最大取值范围。那么，归一化群半径可以定义为

$$\delta_{norm} = \frac{\delta(k)}{diam(D)} < \sigma \tag{16}$$

当 $\delta_{norm} < \sigma$ 时，我们认为粒子群收敛于极值。通过实验发现 $\sigma = 1.1\times10^{-6} \sim \sigma = 1.1\times10^{-4}$ 能够满足重置机制。

由于目标函数为非线性高维函数，其解的分布及全局最优解是未知的，所以在应用 RPSO 对权向量寻优之前，首先利用标准测试函数验证 RPSO 的性能。因为标准测试函数具有解析表达形式和确定的全局最优值，可以对 PSO 算法进行定量的分析比较。Ackley function 是一个多峰（具有多个局部最优解）函数，其在搜索空间中具有全局最小值 $f(x_0) = 0$，$x_0 = [0, \cdots, 0]$。在 n 维搜索空间中，粒子位置向量 $x = [x_1, x_2, \cdots, x_n]$ 的取值范围是 $x_i \in [-32, 32]$，$i = 1, 2, \cdots, n$。Ackley function 表达式为：

$$f(x) = -20\exp\left(-0.2\sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2}\right) - \exp\left(\frac{1}{n}\sum_{i=1}^{n}\cos(2\pi x_i)\right) + 20 \tag{17}$$

对 PSO 算法有多种性能评价标准，分别从收敛效率角度定义 100 次仿真的收敛次数（Conv. Num.）；从收敛速度角度定义平均收敛目标函数计算次数（Mean Eva.）；从收敛质量角度定义 100 次仿真中的适应度最小和最大误差（min、max）。

3　仿真实验

3.1　参数设置

邮件训练样本来自于中国教育和科研计算机网紧急响应组(Data Sets of Chinese Emails，CCERT 2005－Jun)，该样本集包含合法邮件 9272 封，垃圾邮件 25088 封。从样本集中选择 1000 封合法邮件和 2700 封垃圾邮件作为邮件训练样本集，其中 100 封合法邮件和 270 封垃圾邮件用于 IWB 训练，其他用于测试。

RPSO 的粒子数由邮件特征值个数确定，其惯性系数和加速系数采用应用最广泛的标准 Clerc 设置：$\alpha = 0.729$，$c_1 = c_2 = 1.49455$；IWB 的权值取值范围为 [0，1]；搜索空间定义为 30 维。

3.2　RPSO 验证

独立进行 1 次仿真实验，目标函数计算次数作为仿真停止条件，定义为 20000；实验的目的是通过长时间搜索，以便得到（近似）全局最优值，并且直观地观察 RPSO 的收敛过程。由图 2 可以看到，粒子群在开始阶段缓慢收敛，当目标函数计算次数达到 1000 左右时，目标函数值迅速收敛于（近似）全局最优值，并且数量级达到 10^{-7}。

为了减小粒子群算法随机性对仿真结果的影响，独立进行 100 次仿真实验，当目标函数值小等于 5＊10e－5 时，认为目标函数达到最优值，仿真实验结果如表 1。

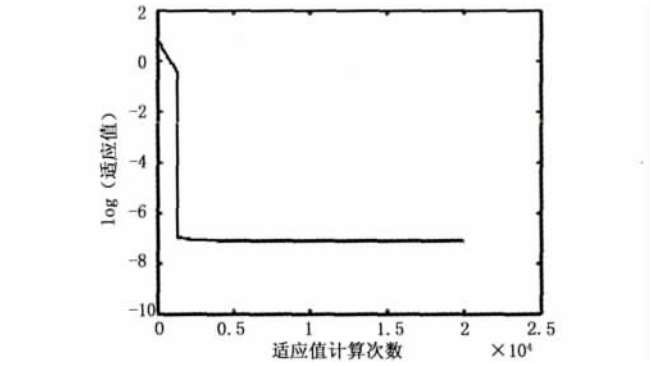


图 2　目标函数值收敛曲线

表 1　RPSO 仿真结果

Min	0
Max	2.3817e－006
Mean Eva	996.4800
Conv. Num	100

在 100 次仿真实验中，由 Conv. Num. 可知 RPSO 均收敛于（近似）全局最优值；由 Min 可知 RPSO 可以搜索到全局最优值，即 0；同时其搜索（近似）全局最优值的最差结果也达到了 10^{-6}；最后，由 Mean Eva. 可知，RPSO 大约需要计算 996 次目标函数即可搜索到（近似）全局最优值。也上两个仿真实验结果验证了 RPSO 具有优秀的全局搜索能力和搜索速度。

3.3　IWB 验证

在进行 IWB 验证之前给出检验其性能的标准测度。精确率和召回率是两个被广泛应用的标准测度，用来评价垃圾邮件过滤算法的有效性。其中，$n_{H\to S}, n_{S\to H}, n_{H\to H}, n_{S\to S}$ 分别表示假阳性错误数、假阴性错误数和正确分类合法邮件和垃圾邮件的数量。精确率（spam precision，SP）和召回率（spam recall，SR）分别定义为：

$$SP = \frac{n_{S\to S}}{n_{H\to S} + n_{S\to S}}$$

$$SR = \frac{n_{S\to S}}{n_{S\to H} + n_{S\to S}} \tag{18}$$

由上式可知，精确率反映了对合法邮件的识别能力，即精确率越高，假阳性错误比率越低；召回率反映了对垃圾邮件的识别能力，即召回率越高，假阴性错误比率越低。为了对不同过滤算法进行性能的对比，给出 F_1 测度，其是上述两个互补测度（精确率与召回率）的调和平均测度，并定义为

$$F_1 = \frac{2 \cdot SP \cdot SR}{SP + SR} \tag{19}$$

表 2　IWB 仿真结果

测度	贝叶斯模型		
	NB	*WB*	*IWB*
SP（%）	86.40	86.86	93.94
SR（%）	92.85	93.41	96.40
F1（%）	89.51	90.02	95.15

仿真实验对 *NB*、*WB* 和 *IWB* 的垃圾邮件辨识能力进行比较验证，结果如表 2。从 F_1 测度来看加权贝叶斯模型的性能比朴素贝叶斯模型提高了 0.51%，而精确率和召回率分别提高了 0.46%和 0.56%。我们可以看到加权贝叶斯模型的性

能只是略好于朴素贝叶斯模型，而这一结论也与文献[7]给出的实验结果相符合。另一方面，提出的改进贝叶斯模型由于没有对贝叶斯模型进行简化处理，同时利用粒子群优化算法对权向量进行寻优，得到近似全局最优的贝叶斯模型，从 F_1 测度可以看到性能比朴素贝叶斯模型提高了 5.64%，而比加权贝叶斯模型提高了 5.13%；精确率和召回率分别提高了 7.54%、7.08%和 3.55%、2.99%。仿真实验表明，采用一个精确模型和一个合适的权向量优化策略对分类准确性有显著的影响。提出的基于精确贝叶斯模型的粒子群优化策略具有明显优势。

4 结论

提出一种改进加权贝叶斯模型，用于垃圾邮件的辨识。为了提高贝叶斯模型的准确性，文中采用精确贝叶斯公式定义加权贝叶斯模型；利用最小二乘算法定义目标函数，用于优化模型中的权向量；针对高维非线性目标函数，文中提出 RPSO 算法对权向量进行寻优。仿真结果表明，基于 RPSO 的 IWB 通过提高模型的准确性，改善了对垃圾邮件的辨识能力，具有良好的实用价值。

参考文献：

[1] 郑 炜，沈 文，张英鹏. 基于改进朴素贝叶斯算法的垃圾邮件过滤器的研究 [J]. 西北工业大学学报，2010，4 (28)：622－627.

[2] 朱志勇，徐长梅，刘志兵，等. 基于贝叶斯网络的客户流失分析研究 [J]. 计算机工程与科学，2013，35 (3)：155－158.

[3] 陈春雷，张新家，张 荔. 电子邮件集成管理系统设计与实现 [J]. 计算机测量与工程，2009，1 (1)：54－55.

[4] J. C，A. B D，W. L. An algorithm for Bayesian belief network construction from data：Proceedings of AI & STAT'97 [Z]. Florida：19978－－90.

[5] Jiang L，Zhang H，Cai Z. A Novel Bayes Model Hidden Naive Bayes [J]. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING，10 (2009)，21，1361－1371.

[6] Provost F J，Domingos P. Tree induction for probability based ranking [J]. Machine Learning，3 (2003)，52，199－215.

[7] Orhan U，Adem K，Comert O. Least Squares Approach to Locally Weighted Naive Bayes Method [J]. Journal of New Results in Science，(2012) 71－80.

[8] N. C，J. S. An introduction to Support Vector Machines and other kernel－based methods [M]. Cambridge University Press，2000.

[9] 李 雯. 基于贝叶斯技术的邮件过滤研究 [D]. 山东师范大学，2008.

[10] Mhamdi B，Grayaa K，Aguili T. Hybrid of particle swarm optimization，simulated annealing and tabu search for the reconstruction of two－dimensional targets from laboratory－controlled data [J]. Progress In Electromagnetics Research B，28 (2011) 1－18.

[11] Kennedy J，Mendes R. Population structure and particle swarm performance [C]. Honolulu，HI：2002，1671－1676.

(上接第 2157 页)

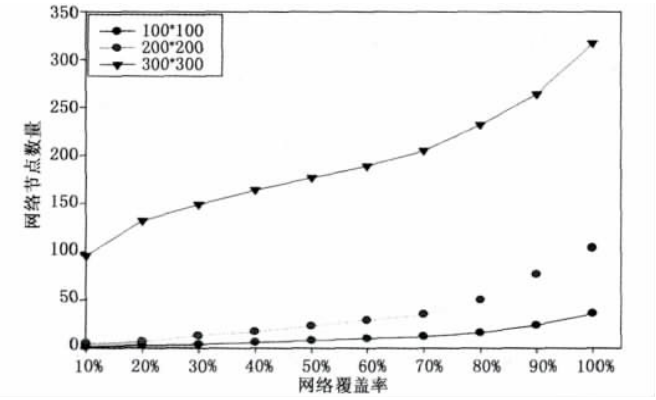


图 5 不同网络规模下的节点覆盖率变化曲线

随着网络覆盖与连通率的增大，需要节点的数量大幅度增加；当网络的覆盖与连通率增大时，对边界影响逐渐变小，最后趋于平衡状态。

图 6 反应的是在有无考虑边界影响条件下，实现不同的覆盖与连通率需要部署节点的数量图。与考虑边界影响下相比，所部署节点数量稍微增加，随着节点数量增加，节点之间的密度也会随之变大，对边界影响有所降低。

4 结束语

针对无线传感器网络中随机部署节点的网络覆盖和连通问题，给出了传感器节点与目标节点关联关系模型。在考虑边界影响的情况下，提出了一个局部覆盖优化算法模型。它能够简化网络覆盖与连通的计算复杂度，提高了算法执行效率，实现了在满足一定网络覆盖和连通率要求下，更加精确地求解出所需部署节点的数量值。最后，通过模拟实验结果验证了理论求解的正确性和算法的有效性。以后的主要工作是立足于传感器网络拓扑结构、路由控制策略以及对不规则区域的有效覆盖和整个网络节能问题的研究。

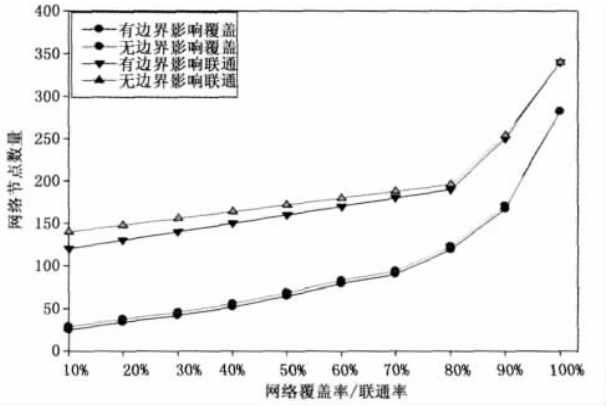


图 6 有无考虑边界影响网络覆盖率/连通率变化曲线

参考文献：

[1] 杨俊刚，史浩山，段爱媛. 无线传感器网络节点软件设计及实现 [J]. 计算机测量与控制，2009，17 (11)：2306－2308.

[2] 孙泽宇，赵国增. 基于节点调度策略的能量有效覆盖算法 [J]. 光通信研究，2012，38 (6)：52－55.

[3] 黄 晓，程宏兵，杨 庚. 无线传感器网络覆盖连通性研究 [J]. 通信学报，2009，30 (2)：129－135.

[4] 宁菲菲，王国军，邢萧飞. 无线传感器网络中一种基于节点序列的覆盖算法 [J]. 中南大学学报，2011，42 (7)：2028－2013.

[5] 孙泽宇，邢萧飞. WSN 中一种规则区域最优覆盖与连通算法研究 [J]. 计算机科学，2011，38 (5)：79－82.