文章编号: 1001-0920 (2011) 03-00397-05

基于排序融合的特征选择

杨 艺, 韩德强, 韩崇昭

(西安交通大学 电子与信息工程学院，西安 710049)

摘 要: 针对模式分类中的特征选择问题, 分别依据 ReliefF 算法、类间可分性及特征相关性等多个评价准则对待约简特征进行评价与排序, 基于排序融合方法实现对多个特征选择评价准则的综合利用. 基于多个数据集的实验结果表明, 该方法在有效降低特征维度的同时, 具有比单准则特征选择方法更高的分类性能.

关键词: 特征选择；ReliefF；排序融合；模式分类；相关性

中图分类号: TP391 **文献标识码:** A

Study on feature selection based on rank-level fusion

YANG Yi, *HAN De-qiang*, *HAN Chong-zhao*

(School of Electronic and Information Engineering，Xi'an Jiaotong University，Xi'an 710049，China. Correspondent: HAN De-qiang，E-mail：digital_king@263.net)

Abstract: For the feature selection problem in pattern classification, based on the criteria of ReliefF, discernability and correlation respectively, the evaluation and ranking of the features are implemented. All these evaluation criteria for feature selection are comprehensively used according to rank-level fusion. Experimental results based on several datasets show that the proposed approach can effectively reduce the feature dimensionality and, at the same time, outcomes better classification performance compared to the traditional single criterion approach.

Key words: feature selection；ReliefF；rank fusion；pattern classification；correlation

1 引 言

特征选择是从原始特征集中选取特征子集的过程. 适当的特征选择可有效去除不相关和冗余特征, 提升学习算法的效率[1-2]. 20 世纪 90 年代以来涌现出大规模机器学习问题, 使得已有的特征选择算法受到严峻的挑战, 迫切需要能适应大规模数据、综合性能 (如准确性及运行效率等) 较好的特征选择算法, 从而特征选择理论与方法引起了相关领域学者广泛的研究兴趣.

特征选择方法主要包括 Filter (过滤) 及 Wrapper (包裹) 2 种方式[3]. Wrapper 方式往往准确性更高, 但特征选择算法本身作为组成部分嵌入学习算法, 其实现较为复杂繁琐. Filter 方式简单易行, 使用 Filter 方式进行特征选择时应力求使评价准则最优. 国内外许多学者设计了多种评价准则来进行特征选择算法的研究, 取得了明显效果, 这些评价准则或方法包括: ReliefF[4], 特征相关性[5], 类间可分性[6], 不一致度[7], 信息增益[8]和基于粗集的智能评价方法[5]等.

单一的准则往往无法全面评价特征子集的好坏, 从而造成特征选择的普适性较差, 分类精度较低. 各种准则往往从不同侧面反映特征的好坏, 具有一定的互补性, 因此利用多准则进行特征评价, 并借助信息融合技术[9]实现有效的特征选择便成为一条可行的途径. 有部分学者依据这一思路提出了一些利用多准则进行特征选择的方法[10], 但其实现往往是“串行”方式, 即一种准则的使用是在另一种准则优选之后的基础上进行. 本文将提出一种基于“并行”方式的多准则排序融合的特征选择方法. 排序融合技术[11-13]是一种综合各种启发式条件影响因子的算法, 广泛应用于各种决策融合问题. 在本文的研究中, 基于多个准则同时对特征进行评价, 得到了针对特征的多个排序, 基于排序融合得到最终综合了多种启发式条件影响因子的单一排序结果, 并基于此完成有效的特征选择. 文中针对遥感数据集进行了仿真实验, 实验结果验证了所提出方法的合理性和有效性.

收稿日期: 2010-01-01；**修回日期:** 2010-03-17.

基金项目: 国家973计划项目(2007CB311006)；陕西省电子信息系统综合集成重点实验室项目(200910A).

作者简介: 杨艺(1980−), 女, 博士生, 从事信息融合、图像处理的研究；韩崇昭(1943−), 男, 教授, 博士生导师, 从事非线性系统控制、信息与图像融合等研究.

2 特征选择方法概要

特征选择可定义为已知一特征集, 从中选择一个子集使评价标准最优[3]. 以上定义可表述为: 给定一个学习算法 F, 一个数据集 S, 数据集 S 来自一个具有 n 个特征 $X_1, X_2, \cdots, X_n$, 具有类别标记 Y 以及符合分布 D 的例子空间, 则一个最优特征子集 $\boldsymbol{X}_{\rm opt}$ 是使得某个评价准则 $J = J(F, S)$ 最优的特征子集.

从特征选择的定义可见, 在给定学习算法、数据集以及特征集的前提下, 各种评价准则的定义和优化技术的应用将构成特征选择的重要内容.

如前所述, 特征选择方法主要分为Filter以及Wrapper两种方式. Filter特征选择方法是评估标准独立于分类器的学习算法, 直接由数据集求得, 而Wrapper则是直接用分类器准确率作为特征子集评估标准的特征选择算法. 由于采用学习算法的性能作为特征评估标准, Wrapper特征选择算法比Filter准确率高, 但其实现相对复杂, 算法效率较低. 本文将主要涉及基于Filter方式的特征选择方法.

2.1 ReliefF 算法

Relief系列算法是公认的效果较好的Filter式特征评估算法. Relief评估方法[14]最早由Kira提出, 1994年Kononenko[4]在Relief算法基础上进行扩展, 提出了ReliefF算法, 以解决多类问题以及回归问题. ReliefF系列算法的要点是根据特征对近距离样本的区分能力进行特征评估, 好的特征应使同类样本彼此接近, 不同类样本彼此远离. 为了减弱噪声的影响, ReliefF算法采用了 k-NN算法而不是最近邻算法(NN), 即

$$W[i] = W[i] - \sum_{j=1}^{k}[\mathrm{diff}(i, R_s, H_j)/(mk)] + \sum_{C\neq \mathrm{class}(R_s)} \frac{\dfrac{p(C)}{1-p(\mathrm{class}(R_s))}\displaystyle\sum_{j=1}^{k}\mathrm{diff}(i, R_s, M_j(C))}{mk}. \tag{1}$$

式中: $W[i]$ 为特征 i 的权值, R_s 为在训练集中随机选取的样本, H_j 为与 R_s 同类的 k 个最近邻样本中的第 $j(j\leqslant k)$ 个样本, M_j 为与 R_s 异类的 k 个最近邻样本中的第 j 个样本, $p(C)$ 为 C 类样本出现的概率.

ReliefF评估效率高, 对数据类型没有限制, 但ReliefF算法不能去除冗余特征, 并存在如下不足:

1) ReliefF算法随机选择样本, 样本个数 m 对各特征权值有影响.

2) 各类别样本数的差异影响特征的权值.

3) 随机选择样本会使小类别被选中进行权值计算的概率小, 有时甚至可能完全被忽略. 在无小类别参与的情形下, 所得到的特征权值既不准确又不合理.

4) 所计算的权值可能出现负值现象. 权值为负数表示同类近邻样本的距离比非同类近邻样本的距离还要大.

除ReliefF方法外, 还存在诸多其他类型的Filter式的特征评价方法, 如基于相关性、类别可分性的方法等, 分别简介如下.

2.2 相 关 性

求取每个特征与其余各个特征的总体相关性大小, 可以判别该特征的冗余程度. 总体相关性越小, 该特征冗余度越低.

设 P 为待处理样本数; N 为特征维数; x_n 为训练样本, 样本均值为

$$\mu = P^{-1}\sum_{n=1}^{P} x_n, \tag{2}$$

方差为

$$\Sigma = (P-1)^{-1}\sum_{n=1}^{P}(x_n-\mu)(x_n-\mu)^{\rm T}. \tag{3}$$

则由 $\Sigma = (v_{ij})_{i,j=1,\cdots,N}$ 可生成相关矩阵

$$\boldsymbol{R} = (R_{ij})_{i,j=1,\cdots,N},$$

其中 $R_{ij} = v_{ij}/\sqrt{v_{ii}v_{jj}}$.

分别计算每个特征对于其余 $N-1$ 特征的相关系数之和

$$\gamma_i = \sum_{j=1,\cdots,N,\ j\neq i} v_{ij}. \tag{4}$$

2.3 类间可分性

设第 c 类样本的个数为 M_c, 则第 c 类样本的第 j 维特征均值为

$$\bar{m}_c(j) = \frac{1}{M_c}\sum_{k=1}^{Mc} x_k^c(j), \tag{5}$$

其中 $x_k^c(j)$ 为第 c 类中第 k 个样本的第 j 维特征; 第 c 类样本第 j 维特征方差为

$$\overline{S}_c(j)^2 = \sum_{k=1}^{Mc}|x_k^c(j)-\bar{m}_c(j)|^2. \tag{6}$$

基于式(5)和(6), 类间可分性度量定义如下:

$$J(j) = \frac{\displaystyle\sum_{s,s\neq c}|\bar{m}_c(j)-\bar{m}_s(j)|^2}{\displaystyle\sum_{c=1,\cdots,M_c}\overline{S}_c(j)^2}. \tag{7}$$

可根据 J 的值判断基于不同特征的类间可分性大小. 式(7)中的分子为各类中心距离的平方和, 能反映各类之间的距离, 取值越大越好. $\bar{S}_c^2$ 为类内方差平方和, 代表某特征下同类样本的差异, 此差异越小越好. 满足这2个要求越好的特征子集越优良, 这样的

特征子集越可能带来更高分类正确率.

本文的研究目标是设法将多种评价准则或方法有机结合, 实现更有效的特征选择, 以服务于高维度模式分类问题的实际应用.

3 基于排序融合的特征选择

基于上述分析可知, 各种特征评价准则各有长短, 对于特征选择而言, 可将多种评价准则看作各种启发式条件影响因子. 为了能够综合利用各种不同条件影响因子, 进而从某种程度上抑制其各自的缺陷, 并做到优势互补, 本文提出了基于多准则的排序融合特征选择方法. 排序融合广泛应用于机器学习[12]以及社会选举, 常见的排序融合规则有 Borda 计数法、最高排序法、逻辑回归法等[13].

本文提出的基于多准则排序融合的特征选择方法流程如图 1 所示.

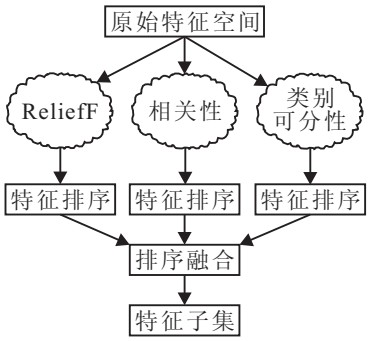

图 1　基于多准则排序融合的特征选择

1) 分别基于 ReliefF, 相关性及类间可分性对特征空间中所有特征进行评价, 生成相应的权值.

① 基于 ReliefF 的特征评价. 基于 ReliefF 产生对原始特征空间中各个特征权值的相反数进行升序排序 (取相反数的原因在于, 权值越大, 相应特征排序应越靠前, 排序值应越小), 依次分别记取排序后每个特征对应的序号, 记为

$$\Lambda_{\text{ReliefF}} = [s_R(1), s_R(2), \cdots, s_R(N)]. \tag{8}$$

其中: N 为原始特征空间特征总数, $s_R(i)$ 为第 i 个特征的权值在所有 N 个特征中的排序序号. 例如, 特征空间为 3 维, 基于 ReliefF 算法所得到的相应的特征权值大小分别为 $W[1]=0.6, W[2]=0.2, W[3]=0.7$. 由于 $-W[3]<-W[1]<-W[2]$, 其对应排序输出为 $\Lambda_{\text{ReliefF}}=[s_R(1), s_R(2), s_R(3)]$, 其中: $s(1)=2, s(2)=3, s(3)=1$.

② 基于相关性的特征评价. ReliefF 算法不能去除冗余性, 因而相关性准则是本文采取的多个准则之一. 依据式 (2) ∼ (4), 求取每个特征相对于其余特征的总体相关性 γ_i. 基于 γ_i 的值对所有特征进行升序排序, 其对应的排序输出为

$$\Lambda_{\text{Correlation}} = [s_C(1), s_C(2), \cdots, s_C(N)], \tag{9}$$

式中 $s_C(i)$ 为基于相关性准则获取的第 i 个特征的权值在所有 N 个特征中的排序序号.

③ 基于类可分性的特征评价. 分别基于原始特征空间中所有 N 个特征, 求取式 (7) 中类可分性度量, 得 $J=[-J(1), -J(2), \cdots, -J(N)]$. 取负号的原因在于, 可分性度量值越大, 相应特征排序应越靠前, 排序值应越小. 对 J 进行升序排序, 其对应的排序输出为

$$\Lambda_{\text{Discern}} = [s_D(1), s_D(2), \cdots, s_D(N)], \tag{10}$$

式中 $s_D(i)$ 为基于类间可分性准则求得的第 i 个特征的权值在所有 N 个特征权值中的排序序号.

2) 基于 3 种不同准则, 可得到对所有特征的 3 种不同排序 Λ_{ReliefF}, $\Lambda_{\text{Correlation}}$ 和 Λ_{Discern}. 可采用如下规则进行排序融合 (记 $s_F(i)$ 为基于多种评价准则排序融合后获取的第 i 个特征的权值的排序值):

①“取前”规则排序融合. 针对每个特征在不同评价准则下获取的排序值中选取最靠前, 即最“小”排序值作为融合排序值.

$$\Lambda_{\text{Fusion}} = [s_F(1), s_F(2), \cdots, s_F(N)],\ \forall i = 1, \cdots, N;$$
$$s_F(i) = \min\{s_R(i), s_C(i), s_D(i)\}. \tag{11}$$

②“线性求和”规则排序融合. 将每个特征在不同评价准则下获取的排序值求和作为融合排序值.

$$\Lambda_{\text{Fusion}} = [s_F(1), s_F(2), \cdots, s_F(N)],\ \forall i = 1, \cdots, N;$$
$$s_F(i) = \alpha \cdot s_R(i) + \beta \cdot s_C(i) + \gamma \cdot s_D(i), \tag{12}$$

式中 α, β 和 γ 分别代表各个不同评价准则在融合中所占的权重. 通过对权重大小的调整与变化, 可实现不同评价准则对融合重要性的调整.

3) 假设特征选择后保留的特征维数为 L, 依据所得排序融合值矢量 Λ_{Fusion} 选取其中排序融合值最小的 L 个特征作为特征选择的最终结果. 具体实现步骤如下:

Step 1: 对排序融合值矢量 Λ_{Fusion} 中的各个元素按其取值升序排列, 得

$$\text{SCORE} = \text{sort}(\Lambda_{\text{Fusion}}) = [SC_1, SC_2, \cdots, SC_N], \tag{13}$$

其中 $SC_1 \leqslant SC_2 \leqslant \cdots \leqslant SC_N$.

Step 2: 选取排序融合值最小 (融合排序最高) 的 L 维特征 ($L<N$) 作为特征选择结果, 即

$$\boldsymbol{X}_L = \{X_i | S_F(i) \leqslant SC_L\}. \tag{14}$$

如遇排序平局, 则在平局特征中随机选取.

基于排序值进行融合进而实现特征选择, 简单易行, 可对多种因素综合考量. 综合考量多种启发式条件影响因子, 还可通过构造综合目标函数的方式实现, 但其构造过程相对复杂, 后续还要伴随优化求解的过程. 而基于排序融合的方式, 既可以综合各个启发式

条件因子的作用和影响, 又无需基于多种准则构造相对复杂的目标函数. 下面结合具体仿真实验来验证所提出方法的有效性与合理性.

4 实验与讨论

4.1 基于多光谱遥感数据集的实验

本实验采用UCI[15]的LandSat多光谱卫星遥感数据集(2 340×3 380像素), 针对的子区域包括82×100像素点. 地表共包括6个类别, 分别是红土、棉花作物、灰土、暗灰土、菜地及深灰土. 共有6 435个样本点, 其中训练样本4 435个, 其余2 000个样本作为测试样本. 每个样本对应用于一个3×3的矩形邻域, 包括36维特征, 即4个波谱段×9个像素点. 实验中将上述测试样本点与训练样本点合并, 从每类中随机抽取300个样本作为训练样本(合计1 800个), 每类中随机抽取300个样本作为测试样本(合计1 800个). 选取400个样本作为ReliefF特征评估时的训练样本. 特征选择维数为1~35维, 均进行计算实验. 分类器采用LS-SVM, 实验重复进行20次, 取其均值作为实验结果. 实验中的排序融合规则采用线性组合方式. 本例中, 式(12)的加权参数设置如下: $\alpha:\beta:\gamma=1:1:1$. 实验结果如图2和表1所示.

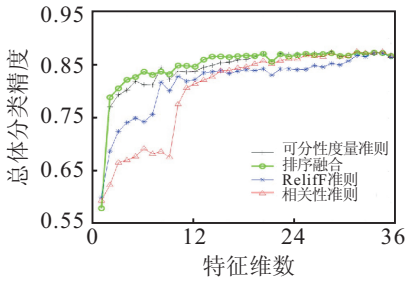

图 2 各种特征选择方法结果比较(UCI多光谱)

表 1 基于UCI多光谱数据集的实验结果

准则	平均分类精度/%
可分性	83.90
ReliefF	81.70
相关性	80.27
排序融合	84.76

表1中的实验结果是特征选取维度由1~35维分别进行实验后的分类精度均值.

4.2 基于高光谱数据集的实验

本文以采用OMIS高光谱成像仪获得的一幅我国延安地区的高光谱图像立方体为例. 对于植被分类问题, 只选用128个波段中的78个, 波长从455.7 nm到1 642.4 nm. 在延安枣园数据集[5]中选取如下类别的样本:

类1: 玉米地; 类2: 桃园地;

类3: 松树苗地; 类4: 葡萄园地.

每个样本都具有78维特征. 从每个类别随机选取500个样本作为训练样本(共计2 000个), 同时随机选取500个样本作为测试样本(共计2 000个), 选取300个样本作为ReliefF特征评估时的训练样本. 特征选择维数从1~60维选取, 均进行计算实验. 分类器采用LS-SVM, 实验重复进行20次, 取其均值作为实验结果. 实验中的排序融合规则采用线性组合方式. 本例中, 式(12)的加权参数设置如下: $\alpha:\beta:\gamma=5:4:1$. 实验结果如图3和表2所示.

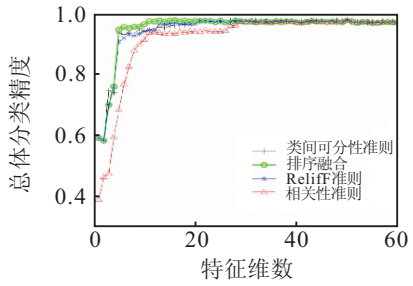

图 3 各种特征选择方法结果比较(枣园高光谱)

表 2 基于枣园高光谱数据集的实验结果

准则	平均分类精度/%
可分性	94.86
ReliefF	94.69
相关性	91.53
排序融合	96.36

表2中的实验结果是特征选取维数由1~60维分别进行实验后的分类精度均值.

从实验结果(图2, 图3, 表1, 表2)可以看出, 基于多准则排序融合的特征选择方法获取的特征子集, 在分类性能方面优于基于单准则的特征选择方法所获取的特征子集. 在保留特征维数较低时, 基于排序融合的特征的选择方法的优势更为明显.

5 结 论

本文提出了基于ReliefF和相关性以及类间可分性度量等多准则排序融合的特征选择方法, 使多个特征选择准则能够优势互补, 进而实现更为有效的特征选择. 基于高光谱及多光谱数据集的实验结果验证了所提出方法的有效性和合理性.

本文方法基于排序融合, 实现简单方便, 但存在参数(不同准则所获排序所占权重以及选择保留的特征维数等)选择的问题, 如何进行参数的优化选择, 如何选取融合规则以及如何设计更为合理有效的非参数的排序融合特征选择方法将是未来的研究方向.

差异和互补是有效融合的前提, 未来可以考虑引入差异性和互补性更大、构造更为合理的特征选择准则参与排序融合, 使排序融合能更有力地提升特征选择效果.

参考文献(References)

[1] Liu H, Yu L. Toward integrating feature selection algorithms for classification and clustering[J]. IEEE Trans

on Knowledge and Data Engineering, 2005, 17(3): 491-502.

[2] Guyon I, Elissee A. An introduction to variable and feature selection[J]. J of Machine Learning Research, 2003, 3(3): 1157-1182.

[3] Li Y, Lu B L. Feature selection based on loss-margin of nearest neighbor classification[J]. Pattern Recognition, 2009, 42(8): 1914-1921.

[4] Kononenko I. Estimation attributes: Analysis and extensions of RELIEF[C]. Proc of the 1994 European Conf on Machine Learning. Catania: Springer Verlag, 1994: 171-182.

[5] 孙亮. 高维遥感数据融合与分类的知识发现方法研究[D]. 西安: 西安交通大学电子信息工程学院, 2006.
(Sun L. Study on knowledge discovery method for fusion and classification of high-dimensional remote sensing data[D]. Xi'an: School of Electronic and Information Enginerring, Xi'an Jiaotong University, 2006.)

[6] 任双桥, 傅耀文, 黎湘, 等. 基于分类间隔的特征选择算法[J]. 软件学报, 2008, 19(3): 842-850.
(Ren S Q, Fu Y W, Li X, et al. Feature selection based on classes margin[J]. J of Software, 2008, 19(3): 842-850.)

[7] Dash M, Liu H. Consistency-based search in feature selection[J]. Artificial Intelligence, 2003, 151(1/2): 155-176.

[8] Battiti R. Using mutual information for selecting features in supervised neural net learning[J]. IEEE Trans on Neural Network, 1994, 5(3): 537-550.

[9] 韩崇昭, 朱洪艳, 段战胜. 多源信息融合[M]. 北京: 清华大学出版社, 2006: 1-13.
(Han C Z, Zhu H Y, Duan Z S. Multi-source information fusion[M]. Beijing: Tsinghua University Press, 2006: 1-13.)

[10] 李勇明, 张素娟, 曾孝平, 等. 轮询式多准则特征选择算法的研究[J]. 系统仿真学报, 2009, 21(6): 2010-2013.
(Li Y M, Zhang S J, Zeng X P, et al. Research of poll mode and multi-criteria feature selection algorithm based on chain-like agent genetic algorithm[J]. J of System Simulation, 2009, 21(6): 2010-2013.)

[11] 刘明, 袁保宗, 苗振江. 一种双目标排序层分类器融合方法[J]. 自动化学报, 2007, 33(12): 1276-1282.
(Liu M, Yuan B Z, Miao Z J. A double-objective rank level classifier fusion method[J]. Acta Automatica Sinica, 2007, 33(12): 1276-1282.)

[12] Renda M E, Straccia U. Web metasearch: Rank vs score based rank aggregation methods[C]. Proc of the 2003 ACM Symposium on Applied Computing. Melbourne: ACM Press, 2003: 841-846.

[13] Ding C, He X F, Husbands P, et al. Rank aggregation methods for the web[C]. Proc of the 10th Int World Wide Web Conf. Hong Kong: ACM Press, 2001: 613-622.

[14] Kira K, Rendell L A. The feature selection problem: Traditional methods and a new algorithm[C]. Proc of the 10th National Conf on Artificial Intelligence. San Jose, 1992: 129-134.

[15] Blake C L, Merz C L. UCI repository of machine learning databases[DB/OL]. [1998-08-10]. http://www.ics.uci.edu/˜mlearn/MLRepository.html.

(上接第396页)

[4] Liu J, Feng Z, Petrovic D. Information-directed routing in ad hoc sensor networks[J]. IEEE J on Selected Areas in Communications, 2005, 23(4): 851-861.

[5] Long Z, Ruixin N, Varshney P K. Posterior crlb based sensor selection for target tracking in sensor networks[C]. IEEE Int Conf in Acoustics, Speech and Signal Proc, ICASSP. Honolulu: IEEE Press, 2007, II: 1041-1044.

[6] Long Z, Ruixin N, Varshney P K. A sensor selection approach for target tracking in sensor networks with quantized measurements[C]. IEEE Int Conf in Acoustics, Speech and Signal Processing, ICASSP. Las Vegas NV: IEEE Press, 2008: 2521-2524.

[7] 王睿, 梁彦, 潘泉, 等. 无线传感器网络信息感知中的自组织算法[J]. 自动化学报, 2006, 32(5): 829-833.
(Wang R, Liang Y, Pan Q, et al. A self-organization algorithm in wireless sensor networks[J]. Acta Automatica Sinica, 2006, 32(5): 829-833.)

[8] 王睿, 潘泉, 程咏梅, 等. 基于动态分簇的传感器网络协同算法[J]. 计算机工程与应用, 2005, 41(11): 9-11.
(Wang R, Pan Q, Cheng Y M, et al. A cluster based CSIP algorithm in sensor networks[J]. Computer Engineering and Applications, 2005, 41(11): 9-11.)

[9] 王权, 王睿, 梁彦. 无线传感器网络的信息收益函数研究[J]. 系统仿真学报, 2007, 19(24): 5812-5817.
(Wang Q, Wang R, Liang Y. Performance analysis on information utility function in wireless sensor networks[J]. J of System Simulation, 2007, 19(24): 5812-5817.)

[10] Bhardwaj M, Chandrakasan A P. Bounding the life-time of sensor networks via optimal role assign-ments[C]. INFOCOM 2002, 21st Annual Joint Conf of the IEEE Computer and Communications Societies. New York: IEEE Press, 2002, 3: 1587-1596.