文章编号: 1001-0920 (2013) 06-0904-05

基于谱图和成对约束的主动半监督聚类算法

蒋伟进[1,2], 许宇晖[1], 王 欣[3]

(1. 湖南商学院 计算机与信息工程学院，长沙 410205；2. 武汉理工大学 计算机科学与技术学院，武汉 430070；3. 湖南工业大学 电气自动化学院，湖南 株洲 412008)

摘 要: 针对半监督聚类学习算法中缺乏主动学习的缺陷, 提出一种纠错式主动学习成对约束方法. 算法通过寻找一般聚类算法自身难以发现的成对约束信息, 同时避免这部分约束信息之间本身的关系, 将其引入谱聚类算法, 利用该监督信息调整谱聚类中点与点之间的距离矩阵对两点间距离进行排序, 采用双向寻找的方法, 使得学习器即使接收到没有标记的数据也能进行主动学习. 实验分析表明, 所提出算法能够获得较为满意的聚类效果.

关键词: 半监督聚类；主动式学习；成对约束；谱聚类

中图分类号: TP273 **文献标志码:** A

Active semi-supervised clustering algorithm based-on pair-wise constraints

JIANG Wei-jin[1,2], *XU Yu-hui*[1], *WANG Xin*[3]

(1. School of Computer and Information，Hu'nan University of Commerce，Changsha 410205，China；2. School of Computer Science and Technology，Wuhan University of Technology，Wuhan 430070，China；3. School of Electric Automatization，Hu'nan University of Technology，Zhuzhou 412008，China. Correspondent：JIANG Wei-jin，E-mail: nudtjwj@163.com)

Abstract: An active learning algorithm based on pair-wise constraints with error correction is proposed in this paper. The algorithm searches the pair-wise constraints information that the clustering algorithm cann't find, and tries its best to reduce the connections between these constraint informations, which is used in the spectral clustering. The suppervised information is used to adjust the distance matrix in the spectral clustering, and the distances are sorted. The learninger can study actively when the learinger receives the data without flags by using the two-way search method. Experiment analysis shows that better clustering result can be obtained by using the proposed method.

Key words: semi-supervised clustering；active learning；pairwise constraint；spectral clustering

0 引 言

信息技术的迅猛发展极大地帮助人们提高了数据收集、数据存储的能力, 在科学研究和社会生活的各个方面都积累了海量的数据, 对这些数据进行分析和发掘其中蕴含的有价值信息, 已经是各个领域的共同需求. 以往的机器学习算法一般只考虑有标记数据, 或者只考虑未标记数据, 但在现实环境中一般两者兼有, 因此, 如何更有效地利用这些数据成为一个亟待解决的问题. 近年发展起来的半监督学习是解决这一问题的有效方法, 它通过广泛利用无标号数据样本的先验知识来完成对样本数据的分类或聚类. 目前, 应用样本的先验信息解决聚类问题已成为智能信息处理的重要途径和研究热点[1-6].

半监督聚类一般通过两种先验信息引导聚类过程, 即标号点信息和成对约束信息. 由于标号点信息可以转化为成对约束信息, 通常用成对约束信息作为半监督聚类先验信息来监督聚类过程[1]. 作如下规定: 1) must-link, 若两样本为 must-link 约束, 则它们在聚类时必须被分配到同一类中 (即 Must-link 约束要求两个数据点必须在同一个聚类中); 2) cannot-link, 若两样本是 cannot-link 约束, 则它们在聚类时必须被分配到不同类中 (即 cannot-link 约束要求两个点不能在

收稿日期: 2012-02-22；**修回日期:** 2012-11-05.

基金项目: 国家自然科学基金项目(61074067, 21106036)；湖南省自然科学基金项目(10JJ5064, 11JJ6051)；教育部人文社科研究一般规划基金项目(11YJAZH039)；湖南省重点学科建设项目.

作者简介: 蒋伟进(1964–), 男, 教授, 博士, 从事机器学习、智能计算等研究；许宇晖(1969–), 女, 讲师, 从事计算机控制与应用的研究.

同一聚类中). 本文将所有 must-link 集合记为 ML, 所有 cannot-link 集合记为 CL, 成对约束信息包含 must-link 和 cannot-link 两种, 这些信息通过样本间的关联和约束来提高聚类效果.

目前, 国内对半监督聚类算法的研究已作了大量开创性工作. 王玲等[4]在谱聚类算法中引入空间一致性先验知识, 建立了一种数据样本间密度敏感的半监督谱聚类算法. 尹学松等[5]提出了一种基于成对约束的判别型半监督聚类分析方法来同时解决违反问题和高维数据问题, 充分运用监督信息来融合数据降维和聚类. 肖宇等[6]基于近邻传播算法, 使用已知的标签类数据或者成对点约束对数据形成的相似度矩阵进行调整, 进而达到提高近邻传播方法的聚类性能. 彭岩等[7]在典型相关分析算法基础上, 通过引入以成对约束形式给出的监督信息, 提出了一种半监督的典型相关分析算法, 算法除了考虑大量的无标号样本外, 还考虑成对约束信息. 金骏等[8]在鲁棒联机聚类算法的基础上, 引入以样本标记形式给出的监督信息, 提出半监督鲁棒联机聚类算法. 赵卫中等[9]基于摄动的模糊聚类算法对最优模糊等价矩阵的相关性质进行了分析. 郭崇慧等[10]提出了一种基于独立成分分析的时间序列谱聚类方法.

以上算法虽然利用成对约束来指导聚类, 但在求解时常会出现成对约束奇异等问题, 致使聚类效果不太理想. 为此, 本文提出一种新的结合谱图理论和成对约束的主动学习半监督聚类方法, 挖掘含有丰富聚类信息的监督信息, 并将其应用到谱聚类算法中. 调整点与点之间的距离矩阵, 使类内各点紧密分布, 类间距离尽量拉大, 形成主动半监督谱聚类算法, 提高聚类性能.

1　半监督聚类主动学习算法

1.1　学习方法

本文对成对约束信息进行量测时使用信息量的概念, 具体求解过程为: 设 C 是约束集合, CA 是聚类算法, P_{CA} 是算法 CA 在无任何约束下的数据集的划分. 约束集合 C 的信息量评价为

$$\mathrm{INF}_{\mathrm{CA}}(C)=\frac{1}{|C|}\sum_{c\in C}\mathrm{unsat}(c,P_{\mathrm{CA}}),$$

其中若根据 P_{CA} 得不到约束 c, 则 $\mathrm{unsat}(c,P_{\mathrm{CA}})=1$, 否则为 0. $\mathrm{INF}_{\mathrm{CA}}(C)$ 越大, 约束集涵盖的算法本身无法找到的约束越多, 便越具价值. 若 P_{CA} 得到的两点间的约束为不正确, 即相同类的两点由 CA 划分到了不相同簇, 或者不相同类的两点划分到了同一簇内, 则认为该两点间的约束是 CA 还无法找到的, 因此该约束具有学习价值. 此外, 若距离相近的两个点间具有 CL 关系, 而距离相对远的两个点之间具有 ML 关系, 则这类关系同样是算法难以找出的, 所以一般情况下其信息量价值均会较高.

运用逐步学习的方式, 将每次学习的 N 个新约束加到原先的成对约束里. 学习新成对约束时需要根据前次聚类的结果, 以便能发现上次划分中错误的成对关系并进行修正. CA 算法对两点之间的距离排序并进行双向寻找, CL 关系的学习从相距最近的两点开始观察该两点前次聚类的情况, 若一样则怀疑, 采用提问确认, 然后向距离较大的方向推进; ML 关系的学习从相距最远的两点开始. 学习的 ML 和 CL 约束关系各有一半.

CA 算法需要计算出两点间的欧氏距离作为输入参数, 即 $\mathrm{Dist}(i,j)=(\|x_i-x_j\|^2)^{1/2}$. 同时, 每次 CA 均需按照约束监督知识将距离矩阵进行修订, 即令 CL 关系距离为 ∞, ML 关系距离为 0. 若 $(x_i-x_j)\in\mathrm{CL}$, 则 $\mathrm{Dist}(i,j)=\infty$; 若 $(x_i-x_j)\in\mathrm{ML}$, 则 $\mathrm{Dist}(i,j)=0$.

算法 1　主动学习算法.

输入: 现有成对约束关系 ML 和 CL, 数据集合, Dist, 需学习的约束数目 N, 前次聚类情况 asgn (其中含有各点的簇标号).

输出: 新关系 new_ML 和 new_CL, 评估基准 evaluation.

Step 1: 按照成对约束关系的监督知识调整距离矩阵, 并对所有距离值由小到大排序, 得到 queue.

Step 2: 对该次学习过程中新加入的成对约束集合 $\mathrm{Pt\_con}=\Phi$ 进行初始化, 数组 $\mathrm{smalldis}=\Phi$, $\mathrm{bigdis}=\Phi$, 暂存新加的成对约束距离, 该次学习中已学习过的约束数目 $L=0$.

Step 3: 从距离近的两点开始处理.

Step 3.1: 在 queue 中由小到大找出第 1 个不为 0 的项 y, 设 m 和 n 为 y 的两个相关点;

Step 3.2: 若 $\mathrm{asgn}(m)\neq\mathrm{asgn}(n)$, 则考察 m 和 n 是否属于同一个类, 若 m 和 n 属于同类, 且 $(m,n)\notin\mathrm{CL}$, $m\notin\mathrm{Pt\_con}$, $n\notin\mathrm{Pt\_con}$, 则将 (m,n) 加入 new_ML, 在 Pt_con 中对 m 和 n 进行标记, 在 smalldis 中对 y 进行记录, 令 $L=L+1$;

Step 3.3: 在 queue 中由小到大找出下一个 y, 返回 Step 3.2 直至 $L=N$.

Step 4: 从距离远的两点开始处理.

Step 4.1: 在 queue 中由大到小找出第 1 个不为 ∞ 的项 x, 设 m 和 n 是 x 的相关两点;

Step 4.2: 若 $\mathrm{asgn}(m)=\mathrm{asgn}(n)$, 则考察 m 和 n 是否属于相同的一类, 若 m 和 n 非同类, 且 $(m,n)\notin\mathrm{CL}$, $m\notin\mathrm{Pt\_con}$, $n\notin\mathrm{Pt\_con}$, 则将 (m,n) 加入 new_CL, 在 Pt_con 中记录 m 和 n, 在 bigdis 中记录 x, 令 $L=L+1$;

Step 4.3: 在queue中由大到小找出下一个x, 返回Step 4.2直至$L > 0.5N$.

Step 5: 保存new_ML, new_CL, evaluation=average(bigdis)−average(smalldis).

在算法1中, 当现有聚类结果中距离较远的两个点的簇号相同时进行提问, 若两个点确实不属于同一类别且这种约束之前没有明确给出, 则作为新约束记录下来. 对距离较近的两个点有类似处理. 通常, 随着约束关系数的增加, 聚类情况会趋于更好. 但是, 在对UCI中的Iris开展实验时, 若CRI(一个聚类性能的评价指标)指标达到0.97 ∼ 0.99, 则会出现新增加的多个(如20个)约束均会关联到一个特定点的情形, 此时聚类效果会大幅下降, CRI下降至0.76左右, 表明此成对约束关系添加进来后反而影响了当前较好的聚类情况. 鉴于此, 规定算法1每次学习到的与约束关系对相关的点不重复出现, 即每个点最多被关联1次, 如Step 3.2和Step 4.2. 当然, 算法不限制先后两次学习到的约束对, 以便约束关系能够传递和调整.

用关系矩阵MR, CR分别代表ML约束关系和CL约束关系. ML为等价关系, 必须预先计算已有ML的等价闭包, 即MR ← tsr(MR). $r(\cdot)$, $s(\cdot)$和$t(\cdot)$分别代表自反、对称、传递闭包, $t(\cdot)$由Warshall方法完成[8]. CL关系根据上述MR和CR扩展, 描述为

$$\mathrm{CR} \leftarrow$$

$$\mathrm{MR} * \mathrm{CR} * \mathrm{MR} + \mathrm{MR} * \mathrm{CR} + \mathrm{CR} * \mathrm{MR} + \mathrm{CR}.$$

其中: $*$为矩阵乘法, $+$为逻辑或.

首先运用谱聚类结果学习新的约束关系(见算法2), 然后使用新的约束关系进行新的聚类. 同时, 在学习过程中还对学习停止条件和最大学习次数进行预设, 当学习停止条件出现时, 结束学习. 这时通常会获得较好的结果, 继续学习下去对聚类结果的提升没有太大效果. 另外, 如果学习次数达到了设定的最大次数, 则不会继续学习, 此时得到的聚类情况即为最终的学习结果.

使用算法1输出的evaluation作为终止继续学习的判断条件, 直到evaluation<0结束学习, 此时得到的远距离平均值低于近距离平均值, 所以结束学习.

1.2 主动学习谱聚类算法

结合谱聚类算法和上述成对约束学习算法建立一种主动学习谱聚类算法. 为了简便, 采用K-means算法完成谱空间的聚类. 首先, 基于给定的样本数据, 对点与点间距离形成的距离矩阵进行距离测度函数的学习; 在此基础上得到相似矩阵, 对其特征值、特征向量进行分解, 进而构成新空间下对原先数据的新的描述, 实现在新描述中同一个聚类里的样本均能够更紧密地聚积于一起; 最后, 在新的空间中建立聚类核函数, 应用梯度下降法求解聚类核半监督分类器中的非凸优化问题, 完成分类. 具体流程见算法2.

算法2 半监督谱聚类算法.

给定待处理的数据集$S = \{S_1, S_2, \cdots, S_n\}$, 将其分为$k$类, 具体步骤如下.

Step 1: 计算邻接矩阵$A \in R^{n\times n}$, 其中

$$A_{ij} = \exp(-||S_i - S_j||^2/2\sigma^2),\ i \neq j,\ A_{ii} = 0.$$

Step 1.1: 若有一对点(i, j)属于must-link集, 则$A_{ij} = A_{ji} = 1$;

Step 1.2: 若有一对点(i, j)属于cannot-link集, 则$A_{ij} = A_{ji} = 0$.

Step 2: 构造矩阵$L = (A + d_{\max}I - D)/d_{\max}$. 其中: D为对角矩阵, 对角元素为$d_{ij} = \sum_{j=1}^{n} A_{jk}$; $d_{\max}$为对角线上元素的最大值.

Step 3: 计算L的k个最大特征值所对应的特征向量$x_1, x_2, \cdots, x_k$, 构造$X = [x_1, x_2, \cdots, x_k] \in R^{nk}$.

Step 4: 对X中的每一行进行单位化处理, 得到矩阵Y, 即

$$Y_{ij} = X_{ij} \Big/ \Big(\sum_j x_{ij}^2\Big)^{1/2}.$$

Step 5: 将Y的每一行看作R^k空间中的一点, 运用K-means算法完成聚类.

Step 6: 如果将Y中第i行划进第j类, 则将原数据点s_i同样分配到第j类.

2 实验分析

2.1 实验环境和数据

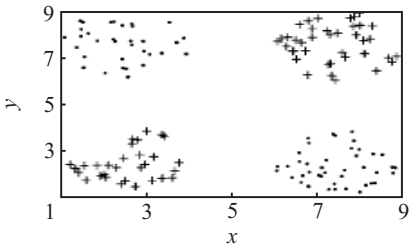
选取真实世界数据集合和人工自造数据集合构成实验数据, 数据特征如表1所示. 真实数据集来自UCI基准数据iris, glass和heart. 自造数据为rings, triples和balls, 如图1所示. 图1(a)为4个球型数据团, 以对角线各为一类; 图1(b)为2个环型数据团, 内、外圈各自一类; 图1(c)包含3个S型数据团, 共3类.

表1 数据集的特征

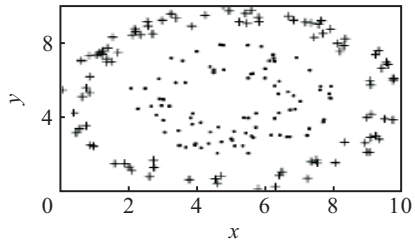
数据集	样本数	维数	类数
iris	161	4	5
glass	237	9	8
heart	289	13	6
rings	198	8	3
balls	165	7	7
triples	193	2	4

2.2 实验方法与评价准则

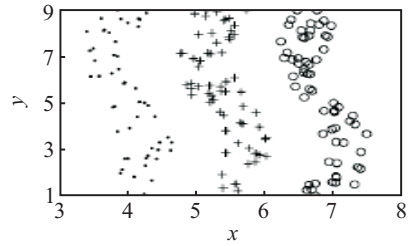
为了对聚类算法进行准确评价, 实验过程中采用2种不同的评价指标: 成对F-评测指标和CRI指



(a) balls



(b) rings



(c) triples

图 1　人工数据集

标. 成对 F- 评测指标针对以往的信息重获评测指标提出的评测标准, 由于成对 F 指标能够对部分数据集的聚类情况进行评测, 可以对五重交叉验证中实验数据集的聚类情况作出不同的评价. 成对 F 指标包含准确率 (precision) 和召回率 (recall) 两个方面, 描述为

$$F = 2 \times \text{precision} \times \text{recall}/(\text{precision} + \text{recall}),$$

$$\text{precision} = \frac{\text{pairs correctly predicted in same cluster}}{\text{total pairs predicted in same cluster}},$$

$$\text{recall} = \frac{\text{pairs correctly predicted in same cluster}}{\text{total pairs in same cluster}}.$$

Rand 指标是一种常用的聚类结果评测方法, 用于判断聚类中每一成对约束点是否出于同一类. 定义如下: 如果某数据集包含 n 个样本, 则其中有 $n(n-1)/2$ 个成对约束. Wagstaff 等[4]按照算法所得到的正确约束对数评测聚类算法的性能, 即 Rand, 有

$$\text{RI} = \frac{\text{correctdecisions}}{\text{totaldecisions}} = \frac{\text{Cd}}{(n \times (n-1))/2},$$

其中 Cd 为由算法所获得的正确决策对数. 为了方便将 Rand 用于半监督聚类算法的评测, Klein 等对 Rand 指标进行改进, 建立了 CRI 指标, 有

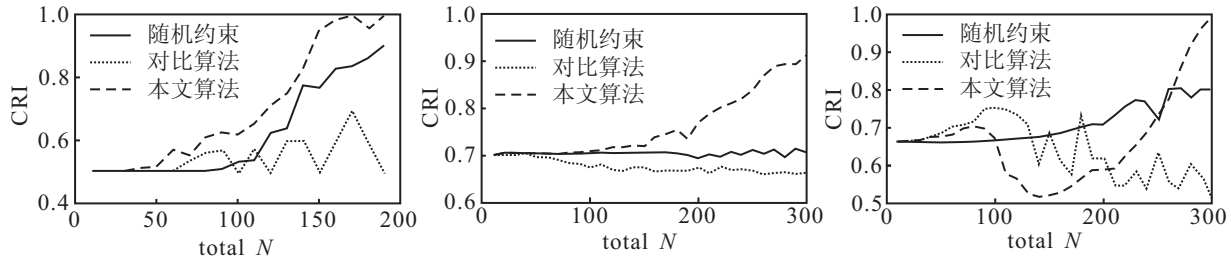
$$\text{CRI} = \frac{\text{correctdecisions}}{\text{totaldecisions}} = \frac{\text{Cd}}{(n \times (n-1))/2}.$$

其中: Cn 为成对约束数, correct free decisions 为划分正确的数据对数目减去约束对中划分正确的数据对数目.

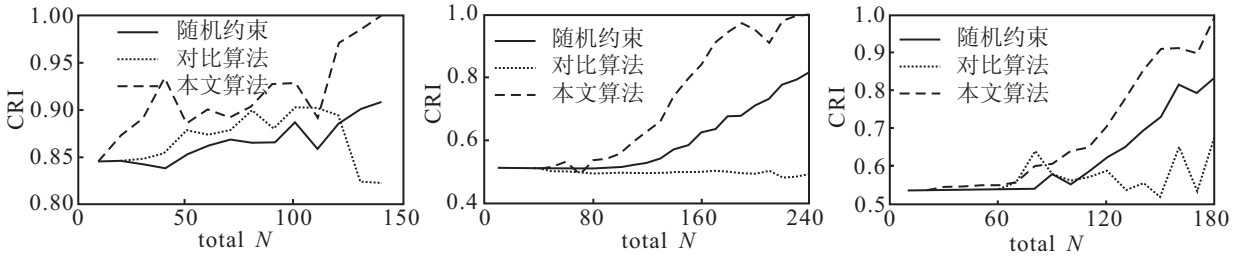
上述两个评测指标各自侧重点不同, F 指标基于同类数据点对的聚类情况评价, 偏重于对同类数据点约束对的聚类结果进行评价; CRI 指标基于数据点是否来自同一类开展判断. 在实验过程中, 对部分数据集采用 F 指标评价, 对全部数据集采用 CRI 指标评价, 用两个评价指标对聚类算法进行评价能够更充分地评价聚类情况.

2.3　实验结果与分析

实验运行 10 次取平均, 每次迭代学习 10 个新的约束, 最多迭代 30 次. 因为本文算法可能会提早出现 CRI = 1 的情形, 所以从 10 次中选择迭代次数最长的



(a) balls　(b) glass　(c) heart



(d) iris　(e) rings　(f) triples

图 2　各数据集的对比实验结果

一次作为结束控制点. 6个数据集的测试情况如图2所示. 随机生成约束方法随着约束对的增多, CRI会逐渐增大; 对比策略在部分约束时会出现好的聚类, 但随着约束对的增加效果反而会变坏; 本文算法能够较快地获得好的聚类结果, 这是因为本文谱空间内的聚类算法使用了K-means算法. 在约束对增加的过程中, CRI指标可能会出现某些减小, 但随后一般均会出现上升趋势. 引起减小的成因是, 新学习的约束与现有约束未涉及的数据点有关联, 根据约束对的传递关系, 获得许多新的约束联系, 这些关系会引起距离矩阵作相应修改, 进而导致谱聚类情况出现较大改变. 但随着成对约束关系的逐渐增多, 约束点间的约束关系会越来越稳定, 聚类效果更趋完善.

算法设置了停止学习条件evaluation <0, 为了验证其有效性, 在上述实验过程中记录了evaluation <0时的迭代次数和CRI值, 重复实验10次, 聚类结果见表2. 由表2可知, NEW算法在6个数据集上均有较为理想的结果; glass集在297个约束数内无法达到停止条件, 直到最大学习次数时才会停止; 其他数据集不到30次即会出现停止条件, CRI值也比较理想.

表 2 对6个数据集的实验结果

数据集	平均迭代次数	最大/最小迭代次数	平均CRI	最大/最小CRI
iris	5.5	10/5	0.978 3	1.000 0/0.939 1
glass	30	30/30	0.901 7	0.942 3/0.862 9
heart	27.1	29/26	0.987 2	0.988 7/0.921 2
balls	14.8	19/12	0.953 7	1.000 0/0.890 5
rings	18.4	26/17	0.961 3	0.989 7/0.930 6
triples	17.1	18/15	0.969 9	1.000 0/0.931 8

3 结 论

本文提出一种主动学习成对约束算法, 基于被调整距离矩阵所形成的谱特征矩阵进行目标函数聚类. 实验结果表明, 该算法聚类性能优于基于随机选取监督信息的半监督谱聚类, 但约束集合对聚类效果影响较大. 下一步的研究方向是探索至少需要多少标记样本才能实现有效的半监督学习、半监督学习在何种情况下奏效、如何避免半监督学习可能造成学习器泛化能力显著下降以及如何将半监督学习用于解决更多实际问题.

参考文献(References)

[1] Dan Klein. From instance level constraints to space-level constraints: Making the most of prior knowledge in data clustering[C]. Proc of the 19th Int Conf on Machine Learning. New York: IEEE Press, 2002: 307-314.

[2] Basu S, Banerjee A. Active semi-supervised for pairwise constrained clustering[C]. Proc of the 4th SIAM Int Conf on Data Mining. San Francisco: IEEE Computer Societ, 2004: 333-344.

[3] Wagstaff K, Cardie C. Clustering with instance-level constraints[C]. Proc of the 17th Int'l Conf on Machine Learning. Stanford: Morgan Kaufmann Publishers, 2000: 1103-1110.

[4] Wang Ling, Bo Lie-feng, Jiao Li-cheng. Density-sensitive semi-supervised spectral clustering[J]. J of Software, 2007, 18(10): 2412-2422.

[5] Yin Xue-song, Hu En-liang, Chen Song-can. Discrim inative semi-supervised clustering an alysiswith pairwise constraints[J]. J of Software, 2008, 19(11): 2791-2802.

[6] Xiao Yu, Yu Jian. Semi-supervised clustering based on affinity propagation algorithm[J]. J of Software, 2008, 19(11): 2803-2813.

[7] Peng Yan, Zhang Dao-qiang. Semi-supervised canonical analys is algorithm[J]. J of Software, 2008, 19(11): 2822-2832.

[8] Jin Jun, Zhang Dao-qiang. Semi-supervised robust on-line clustering algorithm[J]. J of Computer Research and Development, 2008, 45(3): 496-502.

[9] Zhao W Z, He Q, Shi Z Z. Analysis for the properties of the optimal fuzzy equivalent matrix obtained by FCMBP algorithm[J]. Systems Engineering－Theory & Practice, 2010, 30(7): 1238-1245.

[10] Guo Chong-hui, Su Mu-ya. Spectral clustering method based on independent component analysis for time series[J]. Systems Engineering - Theory & Practice, 2010, 31(10): 1921-1931.