一种基于决策粗糙集的自动聚类方法

于 洪 储双双
(重庆邮电大学计算机科学与技术研究所 重庆 400065)

摘 要 提出了一种基于决策粗糙集的面向知识的自动聚类方法。在面向知识的聚类算法中,获取初始聚类结果依赖人工阈值的设置。为此,首先根据物理学知识提出了一种差值排序方法来自动得到阈值。另外,讨论了决策粗糙集模型的损失函数,提出了一种聚类评估方法;通过对聚类结果的评估来实现自动聚类。实验结果表明新方法是有效的。

关键词 聚类,面向知识,决策粗糙集,自动
中图法分类号 TP18 **文献标识码** A

Novel Autonomous Clustering Method Based on Decision-theoretic Rough Set

YU Hong CHU Shuang-shuang
(Institute of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

Abstract This paper proposed an autonomous knowledge-oriented clustering method based on decision-theoretic rough set model. In order to obtain the initial clustering, the initial threshold values need to set in the knowledge-oriented clustering framework. Thus, a novel method, sort difference, was proposed to produce the initial threshold values autonomously in view of physics theory. Then, a cluster validity index based on the decision-theoretic rough set model was developed by considering various loss functions, which can estimate the quality of clustering. The results of experiments show that the new approach is valuable.

Keywords Clustering, Knowledge-oriented, Decision-theoretic rough set, Autonomous

1 引言

近几年来,聚类技术被广泛地应用于数据挖掘、模式识别、机器学习、信息检索、生物学以及市场营销等领域。

聚类算法一般分为以下几类:层次聚类、划分聚类、基于密度和网格的方法和其他聚类方法[7,9,10,13,14]。划分聚类需要预先给定聚类数目,然后遵循迭代最优策略把数据划分成预先设定数目的类别;层次聚类则需要选择算法终止点。这两种方法已取得了广泛的应用,但它们都是依靠局部数据特性来提炼聚类模式的,都存在产生扭曲数据本身特征的风险。通过结合层次聚类算法和粗糙集理论[8],Hirano 等[5]提出了面向知识(Knowledge-Oriented)聚类算法,其克服了传统聚类方法的一些弊端,能从局部和全局的数据特性上对数据进行聚类,达到了更好的聚类效果。此类算法的效率和性能很大程度上取决于初始阈值的选择,已经有很多改进的算法研究[1-3,10],例如,Bean 等利用物理上的质心原理来自动获取初始的分类阈值参数[2,3]。

聚类效果是一个聚类算法应用到实际中需要考虑的重要方面。一个好的聚类效果评估方法会对聚类中很多参数的选取有帮助,如聚类数目。有很多评估聚类有效性的方法[4],主要分为3种:外部有效性评估、内部有效性评估和相关性测试评估。Yao 提出的决策粗糙集模型 DTRS(Decision-theoretic rough set)[12]提供了一个更好的对分类的理解,其通过调整损失函数,能构造一个聚类效果评估指数。Lingras 等[6]提出了一个基于决策理粗糙集的评估方法。

本文针对自动聚类展开了研究。首先,基于物理学理论提出了一种差值排序方法来自动获得面向知识聚类算法框架中的初始阈值,从而避免了人为干扰因素。另外,讨论了决策粗糙集模型的损失函数,提出了一种聚类评估方法;通过对聚类结果的评估来实现自动聚类。实验表明新方法是有效的。

2 基本概念

本节将介绍面向知识聚类、决策粗糙集模型的有关基础概念等。

2.1 面向知识聚类

本文讨论的对象是一个信息系统。信息系统(Information System)又称知识表达系统,是一个四元组 $Info=(U,A,V,f)$。$U=\{x_1,\cdots,x_i,\cdots,x_n\}$是有限的对象集合,称为论域;$A=\{a_1,\cdots,a_m\}$是有限个非空集合;$V=\{V_1,V_2,\cdots,V_m\}$是属性的值域集,$V_i$ 是属性 a 的值域;f 是信息函数,$f:V_{ik}=f(x_i,a_k)\in V_k$。信息系统也可以简记为 $Info=(U,A)$。

定义 1(不可分辨关系,Indiscernibility Relation) 信息

到稿日期:2010-02-09 返修日期:2010-05-02 本文受重庆市科委项目(CSTC,2009BB2082)和重庆市教委项目(KJ080510)资助。
于 洪(1972—),女,博士,副教授,主要研究方向为数据挖掘、粗糙集理论、Web 智能等,E-mail:yuhong@cqupt.edu.cn;储双双(1985—),女,硕士,主要研究方向为智能信息处理、Web 智能。

系统 $Info$ 的不可分辨关系为二元关系 $R=\{(x,y)\in U\times U|\forall a\in A, f(x,a)=f(y,a)\}$，其中 x,y 为 U 中的对象。

不可分辨关系是一种等价关系，通过一个不可分辨关系，可以得到知识表达系统的一个划分。我们把划分后的等价类称为不可分辨类，通常用 $[X]_R$ 来表示包含对象 X 的不可分辨类。

定义 2（初始等价关系，Initial Equivalence Relation） 一个对象 $x_i\in U$ 的初始等价关系 R_i 为 $R_i=\{\{P_i\},\{U-P_i\}\}$，这里 $P_i=\{x_j|\mathrm{sim}(x_i,x_j)\geqslant Th_i\}$，$\mathrm{sim}(x_i,x_j)$ 表示对象 x_i,x_j 之间的相似度，Th_i 表示对象 x_i 的阈值。

选择一个合适的相似度公式是依赖于许多因素的，包括数据集的大小、应用以及数据特性等，它必须针对给定的聚类的问题选择合适的公式。本文的目的不是测量相似度，所以在后面的试验中我们就选用了使用得最广泛的欧式距离公式。

定义 3（对象之间的相似度，Similarity Between Objects）在知识表达系统中，$\forall x_i,x_j\in U$ 的相似度为

$$\mathrm{sim}(x_i,x_j)=1-\frac{\sqrt{\sum_{p=1}^{m}(f(x_i,a_p)-f(x_j,a_p))^2}}{\max\limits_{i,j}\sqrt{\sum_{p=1}^{m}(f(x_i,a_p)-f(x_j,a_p))^2}} \quad (1)$$

$0\leqslant\mathrm{sim}(x_i,x_j)\leqslant 1$。依据相似度和初始等价关系，可以得到对象间的不可区分度。

定义 4（对象间的不可区分度，Indiscernibility Degree of Objects） $\forall x_i,x_j\in U$ 的不可区分度为

$$\eta(x_i,x_j)=\frac{1}{|U|}\sum_{k=1}^{|U|}\gamma(x_i,x_j) \quad (2)$$

式中，$\gamma(x_i,x_j)=\begin{cases}1, & [x_i]_{R_k}=[x_j]_{R_k}\\ 0, & [x_i]_{R_k}\neq[x_j]_{R_k}\end{cases}$，$[x_i]_{R_k}$，$[x_j]_{R_k}$ 分别为 x_i,x_j 形成的等价类。

面向知识的聚类算法，其性能很大程度上依赖于相似度和不可区分度提供的信息。

2.2 决策粗糙集模型

根据贝叶斯决策过程构造概率近似，Yao[12] 提出了决策粗糙集模型。它继承了原有粗糙集的所有基本特点，同时增强了处理不确定信息的能力。

在决策粗糙集理论中，用状态集 $\Omega=\{X,\neg X\}$ 表示一个元素是否属于集合 X，动作集 $A_X=\{a_1,a_2,a_3\}$ 分别表示判定当前对象 x 属于 $POS(X)$，$NEG(X)$ 和 $BND(X)$ 的动作[15]。

a_1 表示判定当前对象 $x\in POS(X)$；

a_2 表示判定当前对象 $x\in NEG(X)$；

a_3 表示判定当前对象 $x\in BND(X)$。

用 $\lambda(a_i|x\in X)$ 表示当对象 $x\in X$ 时，执行动作 a_i 所引起的损耗；用 $\lambda(a_i|x\in\neg X)$ 表示当对象 $x\in\neg X$ 时，进行活动 a_i 所引起的损耗。

这样，进行 3 种不同活动的估计损耗值为

$R(a_1|x)=\lambda_{11}P(X|x)+\lambda_{12}P(\neg X|x)$

$R(a_2|x)=\lambda_{21}P(X|x)+\lambda_{22}P(\neg X|x)$

$R(a_3|x)=\lambda_{31}P(X|x)+\lambda_{32}P(\neg X|x)$

式中，$P(X|x)$ 和 $P(\neg X|x)$ 分别表示 x 属于 X 和 x 属于 $\neg X$ 的概率。$\lambda_{i1}=\lambda(a_i|x\in X)$，$\lambda_{i2}=\lambda(a_i|x\in\neg X)$，$i=1,2,3$。

当 $\lambda_{11}\leqslant\lambda_{31}\leqslant\lambda_{21}$ 且 $\lambda_{22}\leqslant\lambda_{32}\leqslant\lambda_{12}$ 时，表示对象 $x\in X$ 时，将对象 x 划分到正区域 $POS(X)$ 的损耗小于等于将 x 划分到边界区域 $BND(X)$ 的损耗，并且两种损耗都严格小于将 x 划分到负区域 $NEG(X)$ 的损耗。反之，将一个不属于 X 的对象划分到 X 会得到相反的顺序。对于这种类型的损耗函数，最小风险决策规则 $RUL_{P,N,B}(RUL_P,RUL_N,RUL_B)$ 可以写为

RUL_P：如果 $P(X|x)\geqslant\beta$ 并且 $P(X|x)\geqslant\gamma$，那么 $x\in POS(X)$；

RUL_N：如果 $P(X|x)\leqslant\gamma$ 并且 $P(X|x)\leqslant\delta$，那么 $x\in NEG(X)$；

RUL_B：如果 $\delta\leqslant P(X|x)$ 并且 $P(X|x)\leqslant\beta$，那么 $x\in BND(X)$。

其中，

$$\beta=\frac{\lambda_{12}-\lambda_{32}}{(\lambda_{31}-\lambda_{11})+(\lambda_{12}-\lambda_{32})}$$

$$\gamma=\frac{\lambda_{12}-\lambda_{22}}{(\lambda_{21}-\lambda_{11})+(\lambda_{12}-\lambda_{22})}$$

$$\delta=\frac{\lambda_{32}-\lambda_{22}}{(\lambda_{21}-\lambda_{31})+(\lambda_{32}-\lambda_{22})}$$

由条件 $\lambda_{11}\leqslant\lambda_{31}\leqslant\lambda_{21}$ 且 $\lambda_{22}\leqslant\lambda_{32}\leqslant\lambda_{12}$，可知 $\beta\in(0,1)$，$\gamma\in(0,1)$，$\delta\in(0,1)$。此时，决策规则 $RUL_{P,N,B}$ 只依赖于参数 β，γ 和 δ。它们可以由用户给出的 λ_{ij} 值直接计算出来。

若 $\delta\leqslant\beta$，$\delta\leqslant\gamma\leqslant\beta$，根据决策规则 $RUL_{P,N,B}$，正区域、负区域和边界区域可以由 δ 和 β 来决定。若 $\beta<\delta$，就有 $\beta<\gamma<\delta$。根据 $RUL_{P,N,B}$ 可知边界区域为空，而此时正区域和负区域可由 γ 来决定。为了区分 3 个区域，令 $\delta<\beta$，即有 $\delta<\gamma<\beta$。此外，当判定 $x\in NEG(X)$ 和 $x\in BND(X)$ 的风险相同时，则判定 $x\in BND(X)$。

如果判定 $x\in POS(X)$ 和 $x\in BND(X)$ 的风险一样，则判定 $x\in POS(X)$。在这些假设前提下，决策规则 $RUL_{P,N,B}$ 可以简化为

RUL_P：如果 $P(X|x)\geqslant\beta$ 且 $\delta\leqslant\gamma\leqslant\beta$，那么 $x\in POS(X)$；

RUL_N：如果 $P(X|x)<\delta$，那么 $x\in NEG(X)$；

RUL_B：如果 $\delta\leqslant P(X|x)<\beta$，那么 $x\in BND(X)$。

3 基于决策粗糙集的聚类评估

聚类评估能为聚类数目的确定提供一些参考信息。已有很多有效的聚类评估参数[4]。Lingras 等[6] 提出了一种基于决策粗糙集模型[12] 的粗糙聚类评估方法，说明了基于决策理论中的损失函数能构造聚类评估方法的可行性。在文献[6]的基础上，本文提出一种新的面向知识聚类的聚类评估方法。

设论域 U 中的 n 个对象经过聚类后得到 $c_1,c_2,\cdots c_K$ 共 K 个互不相交的类，称其为论域 U 的一个聚类模式（Cluster Scheme），记为 $CS=\{c_1,c_2,\cdots c_K\}$，其中 $c_k\subseteq U,k=1,2,\cdots K$。

根据定义 4，可以得到对象和类之间的相似度。

定义 5（对象和类之间的相似度，Similarity Between Object and Cluster） 设 $x_j\in c_k$，则对象 x_i 和类 c_k 之间的相似度为

$$\mathrm{SimObjClu}(x_i,c_k)=\frac{\sum_{j=1}^{|c_k|}\eta(x_i,x_j)}{|c_k|} \quad (3)$$

式中，$|c_k|$ 表示类 c_k 中含有的对象的个数。

设 $\forall c_p,c_q\in CS$，类 c_p 和类 c_q 之间的不可区分度 ψ_{pq} 定义如下。

定义 6（类与类之间的不可区分度，Indiscernibility Degree of classes）

$$\psi(c_p,c_q)=\frac{1}{|c_p|\cdot|c_q|}\sum_{x_i\in c_p,x_j\in c_q}\mathrm{sim}(x_i,x_j) \tag{4}$$

式中，$|c_p|$，$|c_q|$分别为类 c_p 和 c_q 中对象的个数。

在一个聚类模式 $CS=\{c_1,c_2,\cdots c_K\}$ 中，用 $b_k(CS,x_i)$ 表示将 x_i 分配到类 c_k 中的动作，采取这个动作的代价 $Risk(c_k,x_i)=Risk(b_k(CS,x_i))$。如果假设条件概率 $p(c_k|x_i)$ 的值与对象 x_i 和类 c_k 之间的相似度相对应的话，即

$$p(c_k|x_i)=\frac{\mathrm{SimObjClu}(x_i,c_k)}{\sum\limits_{1\leqslant k\leqslant K}\mathrm{SimObjClu}(x_i,c_k)} \tag{5}$$

则 $p(\neg c_k|x_i)$ 表示将本来不属于 c_k 的对象 x_i 分到类 c_k 中的条件概率。

根据前面提到的决策粗糙集模型，当采取把对象分到 $POS(c_k)$ 的动作时，有

$$Risk(b_j(c_k,x_i)=\lambda_{11}p(c_k|x_i)+\lambda_{12}p(\neg c_k|x_i) \tag{6}$$

那么，一个类簇 c_k 的代价可定义为

$$Risk(c_k)=\sum_{i=1}^{|c_k|}Risk(c_k,x_i) \tag{7}$$

进而一个聚类模式的代价可定义为

$$Risk(CS)=\sum_{k=1}^{K}Risk(c_k) \tag{8}$$

很明显，$Risk(CS)$ 作为对整个聚类模式的一个评价，值越小说明聚类模式越好[6]。虽然找到合适的聚类数目已经被证明是 NP 难问题，但 $Risk(CS)$ 的值有利于我们进行聚类分析。我们的目标是最小化 $Risk(CS)$ 的值，为聚类模式获得较优的聚类数目。

为了实现聚类，这里考虑一种特殊类型的损耗函数。可以认为把一个实际上属于 X 的对象 x 划分到正区域 $POS(X)$ 的损耗为 0(也就是 x 被认为是正对象，因此没有损耗)。相反的情况下，把一个实际上属于 X 的对象 x 划分到负区域 $NEG(X)$ 的损耗即为最大损耗值 1。划分到边界区域的损耗是介于 0 和 1 之间的某个值。基于这种假设，有

$$\begin{aligned}&\lambda_{11}=0,\lambda_{12}=1\\&\lambda_{21}=1,\lambda_{22}=0\\&0\leqslant\lambda_{31}<1,0\leqslant\lambda_{32}<1\end{aligned} \tag{9}$$

在这个假设的基础上，如果能够估计 $P(X|x)$ 的值以及损耗值 λ_{31} 和 λ_{32}，就可以把任何对象划分到 3 个区域其中之一。本文只考虑了将对象划分到正域或负域，因此没有计算损耗值 λ_{31} 和 λ_{32}。

本文不考虑类和类之间的交叉现象，所以根据式(5)—式(9)可以得到聚类模式的代价值为

$$\begin{aligned}Risk(CS)&=\sum_{k=1}^{K}\lambda_{11}\sum_{x_i\in c_k}p(c_k|x_i)+\lambda_{12}\sum_{x_i\notin c_k}p(\neg c_k|x_i)\\&=\sum_{k=1}^{K}\lambda_{12}\sum_{x_i\notin c_k}p(\neg c_k|x_i)=\sum_{k=1}^{K}\sum_{x_i\notin c_k}p(\neg c_k|x_i)\\&=\sum_{k=1}^{K}\frac{\sum\limits_{x_i\notin c_k}\mathrm{SimObjClu}(x_i,c_k)}{\sum\limits_{1\leqslant k\leqslant K}\mathrm{SimObjClu}(x_i,c_k)}\end{aligned} \tag{10}$$

4 面向知识的自动聚类算法

本节将提出一个面向知识的自动聚类算法 AKOC-DTRS (Autonomous Knowledge-oriented Clustering based on DTRS)。

4.1 差值排序法

数据的初始分类在面向知识的聚类算法中是非常重要的一步。如果初始分类不准确，将会导致后来的聚类结果反映不了数据的真实结构。又因为初始分类是依赖于初始阈值 Th_i 的，所以如何选择这个 Th_i，是此类算法至关重要的步骤。

文献[2,3]用 COG 方法来选择 Th_i。COG 方法模拟的是物理上获取质心的原理。当公式

$$\left|\sum_{j=1}^{n}(\mathrm{sim}(x_i,x_j)-\omega\mathrm{sim}(x_i,x_k))\right| \tag{11}$$

的值取最小时，对象 x_i 的初始阈值 Th_i 取 $\mathrm{sim}(x_i,x_j)$，$i,j,k=\{1,2,\cdots,n\}$，这里的 ω 由用户设定。如果选择了最优的参数 ω，算法的迭代次数将很少，并且聚类结果很精确。相反，如果选择了一个不合适的参数值，算法的迭代次数将会增加，而且准确度降低，所以参数 ω 的选取将影响算法的效率和准确率。既然参数 ω 是由用户设定的，那么 COG 方法也不是完全的、无人为干扰因素的，并且这里计算阈值 Th_i 的时间复杂度为 $O(n^2)$，因此我们提出了一个新的自动选取阈值的方法——差值排序法，并且将计算阈值 Th_i 的时间复杂度减少到 $O(n\log n)$。

首先，将相似度矩阵 $S=\mathrm{sim}(x_i,x_j)$ 每行的元素进行从大到小排序，排序过后的矩阵为 $S'=\mathrm{sim}(x_i,x_j)$。然后我们按下式得到阈值

$$Th_i=\{\mathrm{sim}(x_i,x_{j+1})|\mathrm{sec}\ ond\ \max(\mathrm{sim}(x_i,x_j)-\mathrm{sim}(x_i,x_{j+1}))\} \tag{12}$$

$j=1,2,\cdots,n-1$。也就是将排序过后的相似度依次两两相减，如果差值最大，表示这两个相似度对应的对象在物理意义上隔得更远些。换句话说，类簇里面的对象之间的相似度之间跨度比较小。为了使算法收敛得更快，我们选择次大差值对应的减数作为阈值。实验证明，选取次大值能取得更好的效果。

4.2 AKOC-DTRS 算法步骤

结合前面介绍的聚类模式代价值估算的方法，可以将本文提出的基于决策粗糙集的面向知识的自动聚类算法 AKOC-DTRS (Autonomous Knowledge-oriented Clustering based on DTRS)描述如下：

步骤 1　根据式(1)计算每个样本间的相似度，得到相似度矩阵 $S=\{\mathrm{sim}(x_i,x_j)\}$。

步骤 2　根据式(12)对每个对象 x_i 计算阈值 Th_i，根据每个对象的阈值将所有样本进行初始分类，得到 n 个初始分类结果，再对所有的分类结果取交集，得到初始的聚类模式。

步骤 3　根据式(2)计算每个样本间的不可区分度，得到不可区分度矩阵 $\mu=\{\eta(x_i,x_j)\}$。

步骤 4　根据式(10)计算初始聚类模式的聚类代价值。

步骤 5　根据式(4)计算类和类间的不可区分度，合并不可区分度最大的两个类，得到新的聚类模式。

步骤 6　根据式(10)计算新的聚类模式的代价值。

步骤 7　重复步骤 5、步骤 6，直到获得一个稳定的聚类结果。

5 实验结果

为了验证本文算法的有效性，我们对算法进行了实验仿真。实验环境：CPU：Pentium(R) 2.60GHz；内存：1.96GB；硬盘：250G。实验平台：Microsoft Visual C++6.0。

实验 1　采用了一个简单的人工数据集进行实验。数据集包含 100 个样本，每个样本两个属性，如图 1 所示。图 2 给出了初始聚类结果，图 3 是稳定后的聚类结果，图 4 则是代价变化曲线图。可以看到，代价值随着聚类个数的减少在降低，当迭代 7 次也就是当类别数到达 5 后，变化趋势有所平缓。

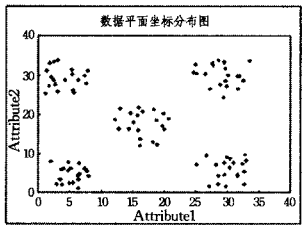

图 1　数据平面坐标分布图

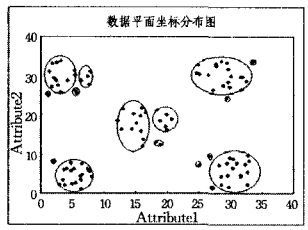

图 2　初始聚类结果

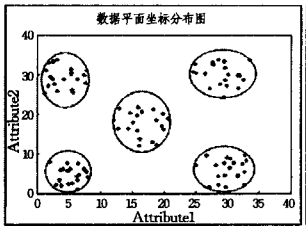

图 3　稳定后的聚类结果

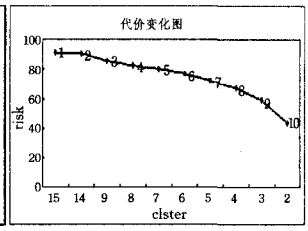

图 4　代价变化曲线图

实验 2　为了验证算法的有效性，我们还选用了 UCI[11] 机器学习数据库中的 3 个数据集：Iirs，Wine 以及 Letter 里随机选取的部分数据。Letter 数据集总共有 26 个类，这里分别选取了一些类别进行实验。Letter1 中选了 A，C 两个类；Letter2 选了 A，C，D 3 个类；Letter3 选了 A，I，L，M，W，Y 6 个类；Letter4 选了 A，I，L，M，W，Y，O 7 个类；Letter5 则选用了 A，I，L，M，W，Y，O，E 8 个类。其中，运行时间单位是秒。实验相关结果如表 1 所列。

表 1　算法准确率和运行时间

数据集	对象个数	属性个数	类别数	准确率	时间(s)
Iris	150	4	3	70%	0.41
Wine	178	13	3	68%	0.5
Letter 1	680	16	2	98%	31
Letter 2	580	16	3	99%	30
Letter 3	101	16	6	87%	0.25
Letter 4	116	16	7	85%	0.30
Letter 5	131	16	8	79%	0.35

AKOC-DTRS 算法具有以下特点：不需要设定任何的阈值，避免了人为干扰；算法收敛速度较快，运行时间相对较短；在不考虑类与类有交叉的情况下准确率很高。在下一步的工作中，我们还需要考虑孤立点对聚类结果的影响、考虑类类间的重叠交叉现象以及研究代价值对聚类结果选取的潜在意义等。

结束语　目前，大部分聚类方法都需要人为设置阈值。因此，本文从面向知识的聚类框架出发，基于决策粗糙集理论研究了自动聚类方法。为了确定初始聚类，提出了差值排序法来自动获取初始分类阈值。此外，本文采用基于决策粗糙集理论模型的聚类评估方法，考虑了一种特定的损失函数，估计聚类模式的代价并以此来指导选择算法的终止点。实验结果表明，本文算法是有效的。为了更好地指导聚类数目的选择，如何改进评估方法并形式化，将是我们下一步的工作。

参 考 文 献

[1] Asharaf S, Murty M N. An adaptive rough fuzzy single pass algorithm for clustering large data sets[J]. Pattern Recognition, 2003(36): 3015-3018

[2] Bean C, Kambhampati C. Knowledge-Based Clustering: A Semi-Autonomous Algorithm Using Local and Global Data Properties [J]. IEEE International Joint Conference on Neural Networks, 2004, 11(3): 95-100

[3] Bean C, Kambhampati C. Autonomous clustering Using Rough Set Theory[J]. International Journal of Automation and Computin, 2008, 5(1): 90-102

[4] Halkidi M, Batistakis Y, Vazirgianni M. Clustering Validity Checking Methods: Part II[J]. ACM SIGMOD Conf. Record, 2002, 31(3): 19-27

[5] Hirano S, Tsumoto S A. Knowledge-oriented Clustering Technique Based on Rough Sets[C]// Proceedings of 25th IEEE International Conference on Computer and Software Applications. Chicago, USA, 2001: 632-637

[6] Lingras P, Chen M, Miao D Q. Rough Cluster Quality Index Based on Decision Theory[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(7): 1014-1026

[7] 殷钢，苗夺谦，段其国. 一种新的粗糙 Leader 聚类算法[J]. 计算机科学，2009，36(5)：203-205

[8] Pawlak Z. Rough sets[J]. International Journal of Computer and Information Sciences, 1982(11): 341-356

[9] 郑吉，苗夺谦，王睿智，等. 一种基于粗糙集理论的谱聚类算法[J]. 计算机科学，2009，36(5)：193-196

[10] Serban G, Campan A. Hierarchical Adaptive Clustering[J]. Informatica, 2008, 19(1): 101-112

[11] UCIrvine Machine Learning Repository[EB/OL]. http://archive.ics.uci.edu/ml/

[12] Yao Y Y. Decision-theoretic Rough Set Models[J]. Lecture Notes in Computer Science, 2007, 4481: 1-12

[13] Yu H, Luo H. A Novel Possibilistic Fuzzy Leader Clustering Algorithm[C]// Sakai H, et al., eds. RSFDGrC 2009. LNAI, 5908: 423-430

[14] 孙吉贵，刘杰，赵连宇. 聚类算法研究[J]. 软件学报，2008，19(1)：48-61

[15] 赵文清，朱永利，高伟. 一个基于决策粗糙集理论的信息过滤模型[J]. 计算机工程与应用，2007，43(7)：185-187

(上接第 220 页)

[7] Bao C L, Huang Y F, et al. Steganalysis of compressed speech [C]// Multiconference on Computational Engineering in Systems Applications. Beijing, China, Oct. 2006, 1: 5-10

[8] ETSI GSM 06.10 version 8.0.1-1999, Digital Cellular Telecommunications System(Phase 2+). Full Rate Speech, Transcoding [S]. 1999

[9] ITU-T Recommendation G.723.1: Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s [S]. 1996

[10] ITU-T Recommendation. G.729 Coding of Speech at 8kbit/s Using Conjugate Structure Algebraic-Code-Exited Linear-Prediction(CS-ACELP) Speech Codec [S]. 1996

[11] Harmsen J, Pearlman W. Steganalysis of Additive-noise Modelable Information Hiding [C]// Proc. SPIE, Security Watermarking Multimedia Contents. 2003, 5020: 131-142

[12] Ker A D. Steganalysis of LSB matching in grayscale images [J]. IEEE Signal Processing Letters, 2005, 12(6): 441-444

[13] Chang C C, Lin C J. LIBSVM--A Library for Support Vector Machines[EB/OL]. http://www.csie.ntu.edu.tw/~cjlin/libsvm/, 2009-04