

基于概化闭包压缩规则的关联分类器构建算法研究^{*}

闵华清¹ 黄莉颖² 卢炎生¹ 蒋嘉之²

(华中科技大学 计算机科学与工程学院 武汉 430074)¹

(华南理工大学 计算机科学与工程学院 广州 510640)²

摘要 在分析关联规则挖掘领域中概化闭包(GC)项集压缩方法基础上,为克服训练数据集中的噪声干扰,改进 L_3 分类器,设计了一种基于概化闭包压缩规则的关联分类器(ACGCCR)构建算法模型。ACGCCR 改进了 GC 的容忍限度值设置方法,根据类分布状态自适应设置容忍限度值;并规定一种新的概化闭包类规则裁剪方法,避免概化过程中出现学习能力不足的问题。ACGCCR 分类规则在压缩存储表现、预测准确度、算法鲁棒性等方面性能表现良好。

关键词 数据挖掘,关联分类,概化闭包类规则,鲁棒性

Research on Algorithm Building of Associative Classifier Based on Generalized Closed Compact Rules

MIN Hua-Qing² HUANG Li-Ying¹ LU Yan-Sheng² JIANG Jia-Zhi²

(School of Computer Science & Engineering, South China University of Technology, Guangzhou 510640)¹

(School of Computer Science, Huazhong University of Science and Technology, Wuhan 430074)²

Abstract In order to overcome noise interference in training data set, an improved L_3 classification and an algorithmic model of Associative Classification Based on Generalized Closed Compact Rules (ACGCCR) are proposed. ACGCCR improves tolerance factor setting method of generalized closed rules mining, providing an adaptive tolerance factor setting method according to class distribution state. A new generalized closed class rules pruning technique is put forward to avoid insufficient learning ability problem. ACGCCR classification rules have better compact storage performance, stronger predictive accuracy and algorithmic robust capability.

Keywords Data mining, Associative classification, Generalized closed compact rules, Robust

1 概述

许多实际应用中,很多数据是高度相关的,这使得挖掘出来的规则数量非常大,一种有效的解决办法就是进行信息的无损压缩^[1~4]。已有研究提出了挖掘频繁闭包项集的方法,如 Pasquier 等. 提出的 Apriori-based 的频繁闭包项集挖掘算法 A-Close^[1], Zaki 和 Hsiao 提出的基于一种项集数据结构的 CHARM^[5] 算法以及 Jian Pei 和 Han Jiawei 提出的基于 FP 树和分区映射机制的 Closet^[6] 算法等。由于闭包集具有相同的支持度及覆盖相同的训练数据集,能大大压缩频繁项集的数量,特别是在稠密数据库中对长模式的挖掘。实验表明,对稠密数据的挖掘,通过闭包算法,压缩比能达到 50% 以上。这使得闭包频繁模式挖掘可以适应存储等方面的要求,从而使算法可行性提高。

2004 年 Elena Baralis 等结合闭包规则压缩表现的方法对分类器进行了改进,使得基于懒惰裁剪技术的分类器更加简洁完善,而预测能力并未受影响^[7]。文[7]详细说明了结合闭包压缩规则和懒惰裁剪技术的分类器 L_3 的生成技术。

基于闭包压缩规则的关联分类器 L_3 对噪声影响很敏感。因为闭包项集方法严重依赖于一个约束:频繁项集和它的超集具有严格相等的支持度。这意味着即使是对数据集的一个微小扰动,都能导致结果频繁项集数量的显著不同。这使得 L_3 分类器的适应性和鲁棒性存在问题。

本项研究的目标是克服训练数据集中的噪声干扰,提供一种适应性强的、稳定的、具有较强预测能力的关联分类器。

2 ACGCCR 模型结构

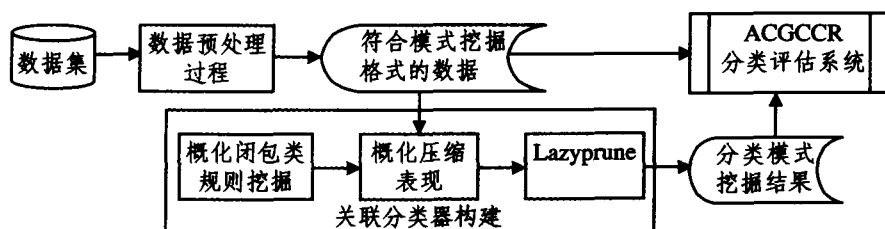


图 1 ACGCCR 模型框架

^{*}广东省工业攻关项目(2003C101007)和香港政府 2005 年粤港关键领域重大基金项目(GHS/054/04)。闵华清 教授,博士生,主要研究方向:人工智能、Web 数据库;黄莉颖 硕士,主要研究方向:数据库;卢炎生 教授,博士生导师,主要研究方向:数据库;蒋嘉之 硕士,主要研究方向:知识工程。

L_3^2 分类器结合了闭包挖掘思想和懒惰裁剪技术的特长,既提高了关联分类预测的准确度,又克服了过适应问题的瓶颈,是一种非常有前途的分类器。但是这项技术却极大依赖于项集和它的超集支持度严格相等的闭包运算条件限制,限制了 L_3^2 分类器的灵活适用性。为了克服闭包运算严格的条件限制,本项研究将在 Vikram Pudi 的概化闭包项集压缩思想^[8]的基础上,对 L_3^2 关联分类器进行改进,设计一种基于概化闭包压缩规则的关联分类器 (Associative Classification Based on Generalized Closed Compact Rules, 即 ACGCCR) 的算法模型,在基本不降低预测准确度的前提下,使得分类器更易于维护并且具有一定的容错性能。ACGCCR 模型框架如图 1 所示。

ACGCCR 在以下四方面提出了改进:

1) 概化闭包类规则挖掘模块将原来的闭包挖掘算法替换成概化闭包挖掘算法,放松了支持度严格相等的条件设置,允许项集支持度小范围的波动,使得框架对于数据库内容来说更加健壮,提供更为概化的信息。

2) 为了避免支持度波动影响过大,对 Vikram Pudi 提出的概化闭包挖掘算法中的容忍限度值设置进行改进,根据类分布具体的情况自适应设置适当的容忍限度值,避免对稀有类规则信息的干扰过大。

3) 为了使规则概化便于进行,在规则概化中提出概化闭包类规则裁剪方法,保持规则裁剪中规则信息大致不变。为了避免过分概化导致分类器学习能力不足,针对概化闭包思想在分类领域应用提出了对规则置信度变化的约束,对类规则执行概化闭包类规则裁剪。

4) 压缩表现模块也做相应改变。由于 Galois 闭包运算对概化项集是不可操作的,因此提出一种概化闭包压缩规则的规则压缩形式,压缩编码了一组概化闭包类规则。

3 ACGCCR 的定义与性质

针对概化思想在规则领域中的应用扩展,提出如下定义。

定义 1(ϵ -equal confidence) 在 $|\text{support}(X) - \text{support}(Y)| \leq \text{类分布 tol 值}$ 和 $|\text{support}(pX) - \text{support}(pY)| \leq \text{类分布 tol 值}$ 的条件下,这里 pX 、 pY 分别是指类规则项集 X 和 Y 的前提项集,认为项集 X 和 Y 形成的类规则具有 ϵ -相等的置信度值,记为 $\text{confidence}(X) \approx \text{confidence}(Y)$ 。

这里的 ϵ 值都选择自适应类分布 tol 值,是因为考虑到不平衡类分布的情况下项集支持度变化的类局部变化属性。

定义 2(概化闭包类规则项集) 对于类规则 $X \Rightarrow c$,若 X 为概化闭包项集, $X \cup c$ 也为概化闭包项集,则项集 $X \cup c$ 为概化闭包类规则项集。

与概化闭包项集裁剪不同,为了使规则概化便于进行,本文在规则概化中提出概化闭包类规则裁剪技术,在压缩冗余信息同时保持规则裁剪中规则信息大致不变,减少信息损失。

定义 3(概化闭包类规则裁剪) 假如类规则项集 $X \cup c$ 有一个 ϵ -相等支持度的直接超集 $Y \cup c$,项集 X 也有一个 ϵ -相等支持度的直接超集 Y ,那么我们裁剪去 $Y \cup c$,以避免生成任何是 $Y \cup c$ 的超级的候选集。这样相等于裁剪掉了 $Y \cup c$ 和 $Y \cup c$ 所有的超集。

由定义 3 可知,概化闭包类规则裁剪包含 ϵ -equal support 裁剪和 ϵ -equal confidence 裁剪两个部分。任何这些被裁剪掉的类规则项集的支持度和置信度,将 ϵ -相等于它的子集。保留下来的项集称之为类生成器(class generator)。

可以得到如下性质:

性质 1 $r_i: X \Rightarrow c$ 和 $r_j: Y \Rightarrow c$ 是任意两条规则。假如 $\gamma'(X) = \gamma'(Y)$, $\gamma'(X \cup c) = \gamma'(Y \cup c)$, 那么 ① $\text{support}(r_i) \approx \text{support}(r_j)$, ② $\text{confidence}(r_i) \approx \text{confidence}(r_j)$, ③ r_i 及 r_j 覆盖的训练数据集在概化意义上基本相同。

其中, γ' 代表采用自适应 tol 值的概化闭包操作。

4 ACGCCR 算法

4.1 自适应类分布 tol 值确定

产生概化闭包类生成器主要是基于概化闭包项集的思想,但是统一的 tol 值设置方法对稀有类的规则干扰很大,需要提供一种新的 tol 值设置方法。ACGCCR 模型中,在进行概化闭包生成之前,首先根据公式: 类分布 tol 值 = 平均容忍限度 $\times$ 该类实例数 / 平均类实例数,依据每个类在整个数据集中的频度来确定本类的 tol 取值。

4.2 产生全部概化闭包类生成器

主要是应用概化闭包挖掘算法挖掘出数据集中全部的概化闭包类生成器,但是针对分类领域问题对项集压缩中的 ϵ -equal support 裁剪进行了扩展,改成执行概化闭包类规则裁剪。

定义 4 对应于类规则 $X \Rightarrow c$ 的一个概化闭包类生成器是一个项集 G , G 满足 $\gamma'(G) = X \cup c$, $\gamma'(Gc) = X$, 并且 $\nexists G' \subset G$ 且 $\gamma'(G') = X \cup c$, $\gamma'(G'c) = X$ 。

其中 γ' 代表采用自适应 tol 值的概化闭包操作。

可以看出,规则 $(Gc) \Rightarrow c$ 就是一个概化的规则基。

4.3 组织生成概化闭包压缩规则

在 Elena Baralis 的压缩规则定义的基础上,针对概化思想在构建关联分类器中的应用,提出一种概化闭包压缩规则的特化规则通过相应的编码概括起来。每一概化压缩规则包括了一条概化的规则基和其特化规则的压缩编码。

定义 5(概化闭包压缩规则) $r_i: (Gc) \Rightarrow c$ 是 R 中的一条任意的概化的规则基,其中 G 为概化闭包类生成器。设 T 是任意一个项集,称为概化闭包压缩规则的尾部且 $T \subseteq (\gamma'(G) - G)$, 其中, γ' 代表概化闭包操作,则 $rc: \{Gc, T\} \Rightarrow c$ 是 R 中的一条概化压缩规则。

与闭包运算中的 Galois 闭包运算符 $\gamma = fg$ 不一样, γ' 包含了 ϵ 误差计算,它不能简单地用精确的数学运算计算出来,是 ϵ 范围内的概化闭包集合,可以在每次近似误差计算后判断是否属于同一概化闭包集合。

4.4 规则全局排序

为了进行 lazy 裁剪,构建关联分类器,需要对概化闭包挖掘阶段提供的候选规则进行全局排序,使得最可信的有效的规则被最先选择。

候选规则排序原则根据懒惰裁剪的需要为:

任意两个概化闭包压缩规则 r_i 和 r_j , $r_i > r_j$ (即 r_i 优先于 r_j), 仅当:

- 1) r_i 置信度高于 r_j , 或者
- 2) r_i 和 r_j 置信度相等, 但 r_i 的支持度高于 r_j , 或者
- 3) r_i 和 r_j 置信度和支持度都相等, r_i 包含的最长规则长于 r_j 包含的最长规则。概化压缩规则 r 包含的最长规则是指规则中包含了 G 和 T 中的所有项。

4) 两者置信度、支持度和长度都相等, 但 r_i 生成早于 r_j 。

即规则是按置信度降序、支持度降序、最长规则长度降序

的顺序来排列的。长规则优先于短规则的选择和文[7]中思想一致,是因为在 lazy 裁剪中,短规则保留在 level II 中,当特化规则没有可应用的时候,它们还可以被考虑,从而减弱过适应现象。

4.5 lazy 裁剪构造分类模型

由于概化闭包压缩只是在 ϵ 偏差范围内, ϵ 偏差的影响是非常有限的,因此覆盖大致相同的数据集。在懒惰裁剪中可以忽略 ϵ 值的影响。

假设 r_i 为同一概化闭包压缩规则中最长规则,根据以下 3 种情况构造分类器:

1) r_i 正确地分类了至少一个训练实例。这时, r_i 归入 level I, 同时从训练数据集中删除被 r_i 覆盖的所有实例。这样,对同属于一个概化闭包的其他规则来说就不再有匹配的训练实例,这些规则归入 level II 级别。

2) r_i 未匹配任一训练实例,同属于一个概化闭包中的其他规则也都不会有匹配的训练实例,因而全归入 level II 级别。

3) r_i 只导致分类错误, r_i 被裁剪掉,同一概化闭包中所有规则均被裁剪。

4.6 ACGCCR 关联分类器分类过程

首先考虑 level I 的规则。如果在 level I 中有规则匹配,则选取这条规则来分配类标识。

如果在 level I 中无规则匹配测试实例,则再考虑 level II 中的规则。这时只需要概化闭包类生成器 G 为测试实例属性集的子集,即 $G \subseteq d$,就能确定这条规则匹配测试实例,于是分配给测试实例相应的类标号。

5 ACGCCR 算法实验及结果分析

为了评估算法模型的效果,本研究进行了一系列的实验,目标主要在 3 个方面:

1) 评估概化闭包压缩的效果,包括规则数量的减少以及鲁棒性、预测准确度等。由于采用了 ϵ -equal support & ϵ -equal confidence 裁剪技术,提高了算法的数据噪声干扰下的稳定性。

2) 分类器的预测能力。由于根据类分布来调整 tolerance 值,使得对稀少类的预测能力不受太大影响。并且 ϵ -equal support & ϵ -equal confidence 裁剪技术使得算法即使在数据噪声干扰的情况下仍然取得较好的预测准确度。

3) 实际应用能力。通过实际应用数据检验算法的可行性和有效性。

实验数据来自 UCI^[9] 机器学习的 26 个标准数据集。本研究将对其中的一个弱相关数据集 HEART 和一个稠密数据集 ZOO 做代表分析。

5.1 分类规则压缩对比实验

下面的实验分别选择一个弱相关数据集 HEART(闭包规则压缩比为 72.39%^[4])和一个稠密数据集 ZOO(闭包规则比为 0.90%^[4])来做具体分析。

HEART 数据集的规则数对比如图 2(支持度为 1%,置信度为 0)。

图 2 可以看出,Pudi 提出的 g-closed 方法挖掘出的包含类属性的概化闭包项集数 < ACGCCR 方法挖掘出的概化闭包压缩类规则数 < 闭包类规则数。ACGCCR 方法挖掘出的概化闭包压缩类规则数比 g-closed 方法挖掘出来的包含类属性的概化闭包项集数要多。

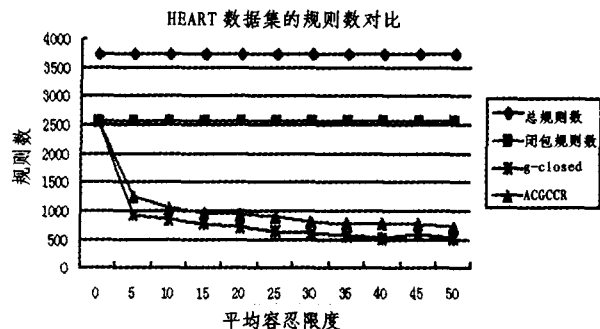


图 2 HEART 数据集规则数对比

ZOO 数据集的规则数对比如图 3(支持度为 1%,置信度为 70%)。

ZOO 数据集挖掘出来的总规则约有 220000 多条,而闭包规则数只为 2068 条,压缩比很大。为了在图中更好地表示 ACGCCR 规则和 g-closed 规则的数量,在上图中把总规则数省略了。

结果显示规则数之间关系和 HEART 数据集一致,即 g-closed 方法挖掘出的包含类属性的概化闭包规则数 < ACGCCR 方法挖掘出的概化闭包压缩类规则数 < 闭包类规则数。另外还可以发现,ZOO 数据集中 g-closed 规则数在 tol 值很小的时候下降很快,而 ACGCCR 相对下降要慢很多。这是由于 ZOO 数据集的非平衡类分布使少数类对 tol 值很敏感,即使 tol 值很小对它们的影响也很大,这很可能导致过压缩,这是 g-closed 在 tol 值很小的时候迅速下降的原因之一。这种过压缩很可能导致应用到分类器的预测准确度下降。而 ACGCCR 避免了过压缩现象。

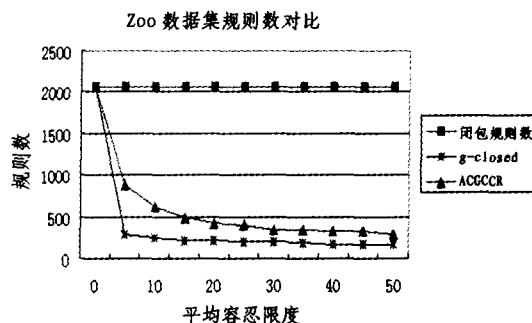


图 3 ZOO 数据集规则数对比

5.3 容错性分析实验

ACGCCR 算法模型采取概化闭包类规则裁剪,项集的支持度和置信度在 ϵ 变化范围内认为是相等的。 ϵ 值的设置使得分类器在一定程度上能够包容数据中的噪声干扰,提供相对稳定的分类模型。

使用 Weka 里提供的 Addnoise 方法给原数据集中添加干扰噪声,观察分类器的性能表 1。

表 1 表明 ZOO 数据集中,ACGCCR 算法预测准确度也比 C4.5 和 L_2 要高,在非平衡类分布情况下仍然性能良好。注意到,ZOO 数据集与 HEART 数据集取得最好预测效果的 tol 值是不一样的,反映出 ZOO 数据集与 HEART 数据集中数据相关程度不同,概化程度也不同,因此对不同的数据集我们需要调整 tol 取值的大小,以取得最佳的预测效果。

表1 ZOO数据集与C4.5分类器相同情况下的分类准确度对比

数据集	C4.5决策树 中叶子数量	C4.5决 策树大小	C4.5 准确度	L ₃ 准确度	ACGCCR 准确度
原数据集	13	21	92.08	90.10	96.04
添加1% 噪声后	13	21	92.08	90.10	96.04
添加2% 噪声后	13	21	92.09	90.10	94.06
添加5% 噪声后	13	21	87.13	87.15	88.12

5.4 预测准确度分析实验

在UCI 26个数据集上执行了一系列实验,以验证ACGCCR算法的有效性。通过不断调整数据集支持度、置信度阈值以及容忍限度取值大小取得ACGCCR算法在不同数

据集上的最佳预测效果,如表2所示。为了比较算法的预测能力,在表中同时列出了C4.5^[10]、CBA^[11]、懒惰裁剪分类器L³^[12]、基于闭包压缩规则的分类器L₃²(支持度阈值取1%)^[7]的预测准确度。表2还列出了数据集属性个数、类标号个数和实例数等特征。

从表2中可以发现,ACGCCR算法在容忍限度大于0的情况下在26个数据集中的15个数据集取得比L₃²更好的预测准确度,平均预测准确度要高于L₃²0.34个百分点,更要高于C4.5(高3.01个百分点)、CBA(高1.71个百分点)、L³(高0.55个百分点)分类器。支持度阈值根据数据集中类分布的情况取值,取1%左右或者高一些,保证稀有类也有规则被发现。

实验结果证明了ACGCCR分类器有很好的规则压缩能力、预测准确度和抗干扰能力。

表2 ACGCCR在UCI 26个数据集上的预测准确度对比

数据集	属性数	类别数量	实例数	C4.5	CBA	L ³	L ₃ ²	ACGCCR	Sup(%)	conf	tol	Δ_{acc} ACGCCR-L ₃ ²
ANNEAL	38	6	898	94.8	97.9	96.10	96.32	96.14	1	0.95	3	-0.18
AUSTRAL	14	2	690	84.7	84.9	86.23	85.65	85.71	10	0.75	1	0.06
AUTO	25	7	205	80.1	78.3	77.56	80.97	81.1	2	0.8	5	0.13
BREAST	10	2	699	95	96.3	95.99	96.28	96.43	1	0.75	3	0.15
CLEVE	13	2	303	78.2	82.8	85.15	81.85	83.83	1	0.9	5	1.98
DIABETES	8	2	690	84.9	84.7	85.80	84.64	85.12	15	0.80	3	0.48
DIABETES	8	2	768	75.2	74.5	78.36	79.17	78.13	3	0.75	3	-6.96
GERMAN	20	2	1000	72.3	73.4	72.50	74.80	73.8	10	0.90	3	-1.0
GLASS	9	7	214	72.4	73.9	77.57	77.10	75.68	1	0.70	1	-1.42
HEART	13	2	270	80.8	81.9	82.96	84.44	85.97	1	0.5	3	1.53
HEPRTI	19	2	155	80.6	81.8	86.45	85.81	85.16	2	0.9	5	1.09
HYPO	25	2	3163	98.2	98.9	95.23	97.70	97.78	2	0.95	3	0.08
IONO	34	2	351	90	92.3	92.88	91.73	89.72	10	0.90	5	-2.01
IRIS	4	3	150	94	94.7	93.33	96	1	0.5	5	2.27	
LABOR	16	2	57	79.3	86.3	92.98	91.23	1	0.5	5	0.0	
LED7	7	10	3200	73.5	71.9	72.00	72.00	72.64	1	0.75	3	0.64
LYMPH	18	4	148	73.5	77.8	84.46	83.78	84.12	1	0.75	3	0.34
PIMA	8	2	768	85.5	72.9	77.99	78.39	85.97	1	0.5	6	7.58
SICK	29	2	2800	98.5	97	93.89	94.54	93.94	2	0.75	2	-0.6
SONAR	60	2	208	70.2	77.5	80.77	79.33	77.8	1	0.75	3	-0.53
TIC-TAC	9	2	958	98.4	99.6	99.48	98.96	99.17	3	0.9	5	0.21
VEHICLE	18	4	846	72.6	68.7	68.56	72.93	70.84	1	0.80	5	-2.09
AWVEFORM	21	3	5000	78.1	80	82.52	82.64	83.26	1	0.75	5	0.62
WINE	13	3	178	92.7	95	97.19	97.16	96.07	1	0.7	7	-1.09
ZOO	16	7	101	92.1	96.8	93.07	93.07	96.04	1	0.75	5	2.97
平均值				83.39	84.69	85.85	86.06	86.40				0.34

结论与展望 本文针对数据挖掘领域中常见的过适应问题和算法鲁棒性问题,寻求解决问题的办法。主要贡献在于将关联规则挖掘领域中的概化闭包项集挖掘思想引入到关联分类领域,构建一种压缩的具有弹性容忍噪声性能的关联分类器ACGCCR。ACGCCR改进了概化闭包挖掘中的容忍限度值设置方法,提出根据类分布状态自适应设置容忍限度值的方法;在概化闭包挖掘中针对概化思想在规则领域中的应用提出概化闭包规则裁剪方法,避免概化过程中出现学习能力不足的问题。基于概化闭包压缩规则的关联分类器在分类规则压缩存储表现、预测准确度保持大致不变时,具有较好的容错性能,对小干扰数据不敏感,同时减少了过适应现象干

扰,算法鲁棒性也得到增强。目前ACGCCR分类器算法在一个公路交通征费系统中开始得到应用并取得很好的分类结果。

参考文献

- 1 Pasquier N, Bastide Y, Taouil R, et al. Discovering frequent closed itemsets for association rules. In: Proc. of Intl Conf. on Database Theory(ICDT), Jan. 1999. 101~109
- 2 Kryszkiewicz M. Concise representation of frequent patterns based on disjunction-free generators. In: ICDM, San Jose, California, USA: IEEE Computer Society, November-December, 2001. 305~312

(下转第226页)

分布信息,通过在散布矩阵的定义之中引入相应的隶属度函数以考虑每一样本对鉴别矢量所做的贡献,消除重叠(离群)样本对于分类产生的影响。

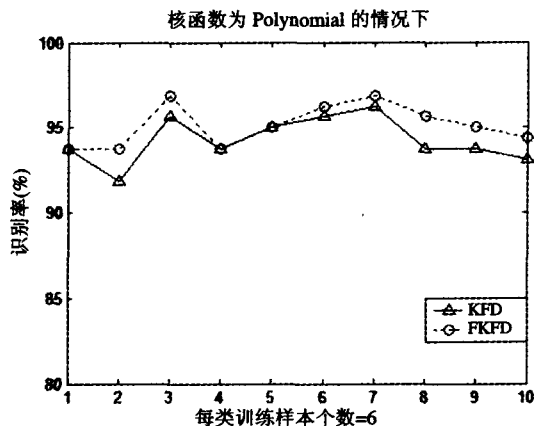


图2 KFD和FKFD特征提取方法识别率比较(采用 Polynomial 核函数)

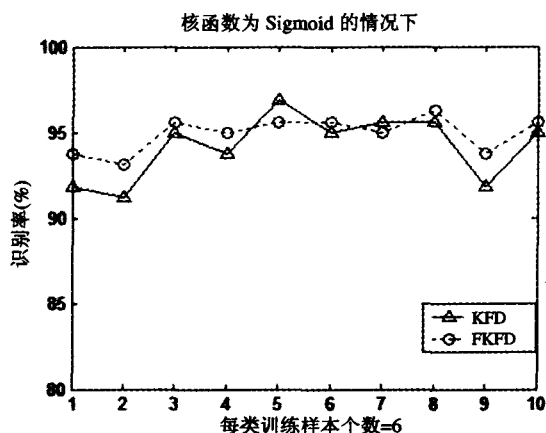


图3 KFD和FKFD特征提取方法识别率比较(采用 Sigmoid 核函数)

结束语 本文提出了一种新的基于模糊集理论的模糊核线性鉴别分析。随着模糊集理论的引入,样本的分布信息通过相应的隶属度函数被融入特征提取过程之中,有助于从重叠(离群)样本之中抽取得到有效的分类信息;并且基于 KP-CA+LDA的核线性鉴别分析方法给出了具体的算法。如何

进一步对样本的分布情况进行研究,有效地利用样本的分布信息,特别是重叠(离群)样本的信息是我们今后要进一步关注的问题。

参考文献

- 1 Vapnik V. The Nature of Statistical Learning Theory [M]. Springer, N Y, 1995
- 2 Schölkopf B, Smola A, Müller K R. Nonlinear component analysis as a kernel eigenvalue problem [J]. Neural Computation, 1998, 10(5): 1299~1319
- 3 Bach F R, Kernel M I. Independent Component Analysis [J]. Journal of Machine Learning Research, 2002, 3: 1~48
- 4 Mika S, Rätsch G, Weston J, et al. Fisher discriminant analysis with kernels [A]. In: IEEE International Workshop on Neural Networks for Signal Processing IX, Madison (USA), August, 1999, 41~48
- 5 Mika S, Rätsch G, Schölkopf B, et al. Invariant feature extraction and classification in kernel spaces [A]. Advances in Neural Information Processing Systems 12. Cambridge: MIT Press, 1999
- 6 Baudat G, Anouar F. Generalized discriminant analysis using a kernel approach [J]. Neural Computation, 2000, 12(10): 2385~2404
- 7 Müller K R, Mika S, Rätsch G, et al. An introduction to kernel-based learning algorithms [J]. IEEE Transactions on Neural Networks, 2001, 12(2): 181~201
- 8 Yang J, Jin Z, Yang J Y, et al. Essence of kernel Fisher discriminant: KPCA plus LDA [J]. Pattern Recognition, 2004, 37: 2097~2100
- 9 Yang J, Frangi A F, Yang J Y, et al. KPCA plus LDA: A Complete Kernel Fisher Discriminant Framework for Feature extraction and Recognition [J]. IEEE Trans Pattern Analysis and Machine Intelligence, 2005, 27(2): 230~244
- 10 Zadeh L A. Fuzzy sets [J]. Info Control, 1965, 8: 338~353
- 11 Keller J M, Gray M R, Givens J R. A fuzzy k-nearest neighbor algorithm [J]. IEEE Trans Syst Man Cybernet, 1985, 15(4): 580~585
- 12 Kwak K C, Pedrycz W. Face recognition using a fuzzy fisherface classifier [J]. Pattern Recognition, 2005, 38: 1717~1732
- 13 ORL face database. <http://www.uk.research.att.com/facedatabase.html>

(上接第 185 页)

- 3 Bykowski A, Rigotti C. A condensed representation to find frequent patterns. In: PODS, ACM, Santa Barbara, USA, pages May 2001. 267~273
- 4 Kryszkiewicz M, Gajek M. Concise representation of frequent patterns based on generalized disjunction-free generators. In: PAKDD, Springer, Taipei, Taiwan, pages May 2002. 159~171
- 5 Zaki M J, Hsiao C J. CHARM: An Efficient Algorithm for Closed Itemset Mining. In: Proc of 2002 SIAM Intl Conf Data Mining (SDM '02), Apr 2002. 457~473
- 6 Per J, Han H, Mao R. Closet: An efficient algorithm for mining frequent closed itemsets. SIGMOD Intl Workshop on Data Mining and Knowledge Discovery, May 2000
- 7 Baralis E, Chiusano S, Garza P. On support thresholds in associa-

- tive classification. In: Proc of the 2004 ACM Symposium on Applied Computing, Nicosia, Cyprus, 2004. 225~232
- 8 Pudi V, Haritsa J. Generalized Closed Itemsets for Association Rule Mining. In: Proc of Intl Conf on Data Engineering (ICDE'03), Mar. 2003. 1057~1073
- 9 Merz C J, Murphy P. UCI repository of machine learning database. <http://www.cs.uci.edu/~mlearn/MLRepository.html>
- 10 Quinlan J. C4. 5: program for classification learning. Morgan Kaufmann. 1992
- 11 Liu B, Hsu W, Ma Y. Integrating classification and association rule mining. KDD'98, New York, NY, Aug. 1998
- 12 Yin X, Han J. CPAR: Classification based on predictive association rules. In: SIAM International Conference on Data Mining (SDM'03), San Francisco, CA, May 2003. 957~965