第 42 卷 第 6A 期
2015 年 6 月

计 算 机 科 学
Computer Science

Vol. 42 No. 6A
June 2015

基于改进 LLE 的高维数据离散化方法

许统德

(广东农工商职业技术学院教务处 广州 510507)

摘 要 连续特征值离散化在数据挖掘、机器学习和模式识别等领域显得尤为重要。目前,现有的离散化方法主要处理低维数据,然而,现实世界中往往存在的是高维非线性数据。基于此,提出一种基于改进局部线性嵌入(LLE)的高维数据离散化方法——ILLE-HD3 方法。首先,通过考虑数据的类信息对 LLE 方法进行改进,使其有效降维,以便于数据在低维空间中离散化。其次,在降维的基础上,提出了基于差异-相似集合(DSS)的连续特征值离散化算法,该算法利用类与特征之间的关联程度来决定连续域中断点的选取位置,并通过 DSS 理论定义分类错误标准,以控制连续域划分过程中产生的信息损失。最后,使用决策树分类工具 C4.5 和 C5.0 进行性能分析,结果表明,提出的 ILLE-HD3 方法处理高维非线性数据时具有较好的效果,与现有的方法相比,得到了较高的分类精度。

关键词 高维数据,局部线性嵌入,离散化,类-特征相互关联,差异-相似集合

中图法分类号 TP18 文献标识码 A

High-dimensional Data Discretization Method Based on Improved LLE

XU Tong-de

(Office of Academic Affairs, Guangdong Agriculture Industry Business Polytechnic College, Guangzhou 510507, China)

Abstract Discretization algorithms for continuous features play a very important role in data mining, machine learning and pattern recognition. Existing methods mainly concentrate on discretizing low-dimensional data. However, there are high-dimensional nonlinear data in the real world. Based on this, this paper presented a high-dimensional data discretization method based on improved locally linear embedding (LLE), namely ILLE-HD3. First, LLE could be improved by considering class information of the data to effectively reduce dimensions of high-dimensional data. This facilitates the discretization method to be implemented in a low-dimensional space. Second, with the dimensionality reduction, we proposed a discretization algorithm for continuous features based on difference-similitude set (DSS). It uses class-feature interdependency to determine the selection of cut points in continuous value domain. Meanwhile, it defines a classification error criterion to control information loss generated by partition of continuous domain. Finally, by using the decision tree classification tools, C4.5 and C5.0, the proposed ILLE-HD3 algorithm achieves a better result on high-dimensional nonlinear data and higher classification accuracy than the existing algorithms.

Keywords High-dimensional data, Locally linear embedding (LLE), Discretization, Class-feature interdependency, Difference-similitude set (DSS)

1 引言

随着数据库系统中信息的大量增加,数据挖掘已经成为了研究热点。当应用机器学习从数据中提取知识时,涉及的数据通常包括数字(如 1,2,3),名词(如红,黄,蓝)和连续值(如温度,海拔等)。一些数据挖掘和机器学习方法[1,2]只能处理离散值表示的数据,即上述的数字和名词类数据。而另外一些机器学习工具既能处理离散数据,也能处理连续数据,但是在处理离散数据时效果会更好。因此,在应用机器学习之前,对连续特征值域进行合理划分(离散化)成为了数据挖掘和机器学习的一个重要方面。

目前,连续特征值离散化方法研究主要有如下几种形式:无监督与有监督、局部与整体、自上到下与自下到上。依据连续特征离散化时是否利用了决策类进行离散连续域,可将连续特征离散化算法分为两大类:有监督的和无监督的算法。传统的无监督方法有 Equal-W 和 Equal-F 算法[3],它们拥有实现简单和计算消耗低的特点,但离散后的结果在大多数情况下难以满足研究人员的要求。而近些年的连续特征离散化算法大部分都是有监督的方法,其中包括:基于统计学的 Chi2 相关算法[4]、基于熵最大化的信息理论离散化算法[5]和基于类-特征相互关联的连续域划分算法[6,7],它们都是目前非常有代表性的算法。其中,基于 Chi2 的相关算法能够使分类器具有相对较高的分类预测精度,因为它们在离散化过程中衡量了数据的不一致信息,避免了有效信息丢失。但这类方法在离散化的时间复杂度上要高于其它两类方法,因为在每次合并相邻区间后都要进行数据的不一致衡量。此外,最

本文受广东省省级教学管理 A 类课题(20120101005),广东省经济和信息化委员会项目(201210110600232)资助。

许统德(1980—),男,硕士,助理研究员,主要研究方向为数据挖掘与信息安全,E-mail:xtd-aib@163.com。

近也有一些新的方法被提出，如赵静娴等人[8]提出了高效的连续数据离散化算法，利用信息熵理论作为标准，快速进行连续域离散化；Jin 等人[9]研究了现有方法相互之间的内在关系，并将一些方法进行了统一化；史志才等人[10]提出了区间粒的概念，融合熵理论定义了区间粒的粒度，进而提出了基于粒计算的连续数值属性的离散化算法；汪凌[11]提出了一种基于改进粒子群的连续属性离散化算法，它结合了集群智能优化理论和粗糙集理论，将各属性离散化分割点初始化为粒子群体，通过粒子间的相互作用寻求最优离散化分割点。徐菲菲[12]提出了基于互信息的模糊粗糙分类特征基因快速选取方法。该方法考虑了粗糙集离散化会使得部分信息丢失，故采用模糊粗糙集，即结合模糊集和粗糙集两种理论的优点，将等价类的精确划分转变为模糊划分，确定对象对每个模糊等价类的隶属度，从而避免一定程度的信息丢失；Ruiz 等人[13]提出了 IDD 算法，它是一种有监督的方法，以区间距离作为离散化评判标准；Bondu 等人[14]提出一种半监督的离散化方法，其既不属于有监督也不属于无监督的新方法；Armengol 等人[15]提出了精炼的离散化方法，通过提出的离散化标准来精确地划分连续属性的值域；Salvador 等人[16]综述了目前存在的离散化方法，分析了各种技术的特点，统一了参数和各种指标的称谓。

然而，现有的离散化方法主要处理低维数据，但现实世界中往往存在的是高维非线性数据。据了解，目前尚未出现基于 LLE 离散化方法的相关报道。基于以上研究，提出一种基于改进局部线性嵌入(LLE)的高维数据离散化方法——ILLE-HD3 方法，以解决高维连续数据的离散化问题。首先，考虑数据的类信息对 LLE 方法进行改进，使其有效降维，便于数据在低维空间中离散化。其次，在降维的基础上，提出了基于差异-相似集合(DSS)的连续特征值离散化算法，该算法利用类与特征之间的关联程度来决定连续域中断点的选取位置，并通过 DSS 理论定义分类错误标准，以控制连续域划分过程中产生的信息损失。最后，通过使用决策树分类工具 C4.5 和 C5.0 进行性能分析，结果表明，提出的 ILLE-HD3 方法对高维非线性数据具有较好的处理效果，与现有的方法相比，得到了较高的分类精度。

2 ILLE-HD3 方法

2.1 改进的 LLE 方法(ILLE)

2.1.1 局部线性嵌入(LLE)原理

局部线性嵌入(LLE)算法[17]主要用于高维数据降维，希望低维的样本局部依然可以保持高维空间中样本间的权值关系，并以样本邻域之间的重叠作为全局降维的连接信息，最终达到降维的目的。

给定 N 个采样点的数据集 $X=\{x_1,x_2,\cdots,x_N\}$，其中，$x_i\in R^D(i=1,2,\cdots,N)$，$D$ 是数据的维度。LLE 通过采样点 x_i 的 k 个近邻来构建采样点 x_i 的局部线性结构，其中，$x_i=\{x_{i1},x_{i2},\cdots,x_{iD}\}$ 是 $1\times D$ 维的向量。通过解决限制的最小二乘来获得最优的重构权值矩阵 W，如式(1)：

$$\begin{cases}\min\ \varepsilon(W)=\sum_{i=1}^{N}\parallel x_i-\sum_{j=1}^{k}w_{ij}x_j\parallel_2^2\\ \text{s.t.}\ \sum_{j=1}^{k}w_{ij}=1\end{cases}\tag{1}$$

其中，$\parallel\cdot\parallel_2$ 指 L2 范数。最优的重构权值矩阵 $W=\{w_i=(w_{i1},w_{i2},\cdots,w_{iN})\}^{\mathrm{T}}$ 可被看成稀疏矩阵，w_i 是第 i 个采样点的局部重构权值向量，$w_{ij}>0$。进而，LLE 将数据集 X 映射至低维空间 $R^d(d\ll D)$ 中，结果为 $Y=\{y_1,y_2,\cdots,y_N\}$。通过保留局部特性，求解式(2)：

$$\begin{cases}\min\ \Phi(Y)=\sum_{i=1}^{N}\parallel y_i-\sum_{j=1}^{k}w_{ij}y_j\parallel_2^2\\ \text{s.t.}\ YY^{\mathrm{T}}=I\end{cases}\tag{2}$$

其中，I 是 $N\times N$ 单位矩阵。结合约束条件，利用拉格朗日乘子法解决此问题。最终，LLE 获得 d 个具有代表性的特征向量，组成低维空间的数据集 Y。

2.1.2 LLE 方法

LLE 使用重构权值 $\{w_{ij}\}$ 来保证原始数据的固有几何拓扑结构，然而，它不能反映采样点 x_i 的 k 个最近邻的密度信息。因此，LLE 对分布不均匀的数据降维效果不佳。图 1 显示了 LLE 方法对不均匀数据从 3 维降至 2 维的局部嵌入，可明显看出，原始数据的局部片几何拓扑结构被严重损坏。

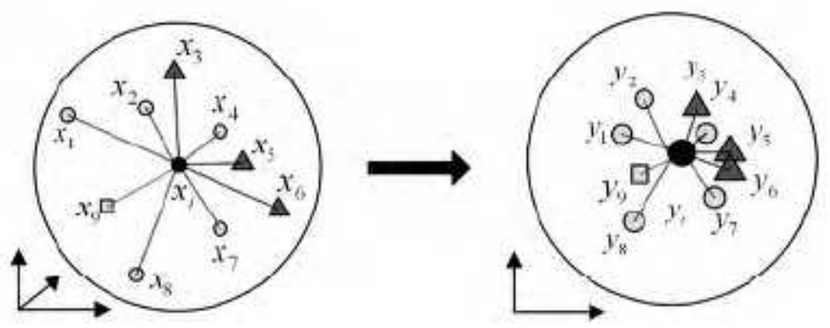


图 1 LLE 方法将不均匀数据从 3 维降至 2 维局部嵌入示意图

为了弥补此缺陷，考虑融合数据的类信息对局部片进行优化，使得降维后的低维空间中数据类间距离有较好的可分性，即相同类的数据点间距离越小越好，不同类的数据点间距离越大越好，这样可以避免在映射不均匀数据时，不同类的样本点聚集在一起。基于此，提出了改进的局部片优化 LLE 方法——ILLE 方法。具体地，$\forall y_i\in Y$，令 Y_{y_i} 为 y_i 的 k 个最近邻集合。注意，每个样本点仅属于一个类。根据类标签信息，将 Y_{y_i} 分割成两部分：$Y^1=\{y_1,y_2,\cdots,y_{k_1}\}$ 为与 y_i 相同类的 k_1 个最近邻点，$Y^2=\{y_{k_1+1},y_{k_1+2},\cdots,y_k\}$ 为与 y_i 不同类的 k_2 个最近邻点，这样，y_i 的局部片为 $\{y_i,Y^1,Y^2\}$。目标是寻找低维嵌入 Y，使得 y_i 与 Y^1 之间的距离最小化，y_i 与 Y^2 之间的距离最大化。图 2 显示了该思想，左侧为原始高维空间中样本点 x_i 的局部片，此局部片包含与 x_i 相同类的最近邻点 x_1、x_4、x_6 和 x_8，以及与 x_i 不同类的最近邻点 x_2、x_3、x_5 和 x_7。图 2 右侧为局部片降至低维空间的期望结果，即低维样本点 y_1、y_4、y_6、y_8 与 y_i 尽可能靠近，而 y_2、y_3、y_5、y_7 与 y_i 尽可能远离。

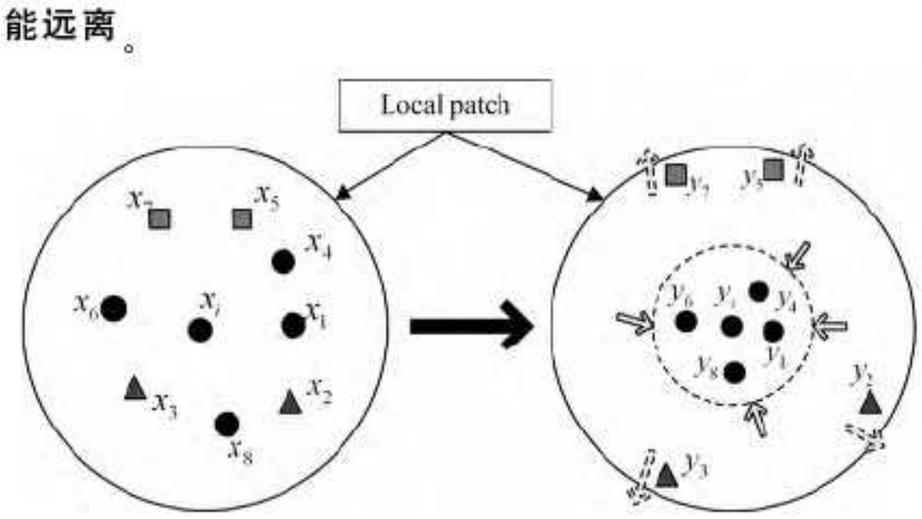


图 2 局部片优化，相同形状和颜色的样本点属于同一类

对于低维空间中 y_i 的局部片，希望 y_i 与其相同类近邻的距离尽可能小，则有：

$$\arg\min_{y_i}\parallel y_i-\sum_{j=1}^{k_1}w_{ij}^1y_j\parallel_2^2\tag{3}$$

其中，

$$w_{ij}^1=\begin{cases}\dfrac{w_{ij}}{\sum_{j=1}^{k_1}w_{ij}}, & j=1,2,\cdots,k_1\\ 0, & \text{其它}\end{cases}$$

同时，希望 y_i 与其不同类近邻的距离尽可能大，则有：

$$\arg\max_{y_i}\| y_i-\sum_{j=k_1+1}^{k}w_{ij}^2y_j\|_2^2 \tag{4}$$

其中，

$$w_{ij}^2=\begin{cases}\dfrac{w_{ij}}{\sum_{j=k_1+1}^{k}w_{ij}}, & j=k_1+1,k_1+2,\cdots,k\\ 0, & \text{其它}\end{cases}$$

基于以上分析，通过解决式(5)的最优化问题来求解最优嵌入：

$$\begin{cases}\arg\min_{y_i}\sum_{i=1}^{N}(\alpha\| y_i-\sum_{j=1}^{k_1}w_{ij}^1y_j\|_2^2-\beta\| y_i-\sum_{j=k_1+1}^{k}w_{ij}^2y_j\|_2^2)\\ \text{s.t. } YY^{\mathrm{T}}=I\end{cases} \tag{5}$$

其中，α 和 β 是衡量类内距离和类间距离的两个尺度参数，有 $\alpha+\beta=1,\alpha,\beta\geqslant 0$，因此，式(5)可以重写为：

$$\begin{cases}\arg\min_{y_i}\sum_{i=1}^{N}(2\alpha\| y_i-\sum_{j=1}^{k_1}w_{ij}^1y_j\|_2^2-\| y_i-\sum_{j=k_1+1}^{k}w_{ij}^2y_j\|_2^2)\\ \text{s.t. } YY^{\mathrm{T}}=I\end{cases} \tag{6}$$

为了解决此优化问题，首先将式(6)转化成式(7)，见定理1：

$$\begin{cases}\arg\min_{Y}[(2\alpha-1)Tr(Y^{\mathrm{T}}Y)+Tr(Y^{\mathrm{T}}((1-\alpha)W^2-\\ \quad\alpha W^1)^{\mathrm{T}}(2I-(W^1+W^2))Y)]\\ \text{s.t. } YY^{\mathrm{T}}=I\end{cases} \tag{7}$$

定理1 式(6)与式(7)等价

证明：

$$\begin{aligned}&\arg\min_{y_i}\sum_{i=1}^{N}(2\alpha\| y_i-\sum_{j=1}^{k_1}w_{ij}y_j\|_2^2-\| y_i-\sum_{j=k_1+1}^{k}w_{ij}y_j\|_2^2)\\ &=Tr(2\alpha Y^{\mathrm{T}}(I-W^1)^{\mathrm{T}}(I-W^1)Y-Y^{\mathrm{T}}(I-W^2)^{\mathrm{T}}(I-W^2)Y)\end{aligned}$$

由于 W^1 和 W^2 正交，则有 $(W^1)^{\mathrm{T}}W^2=0$ 和 $(W^2)^{\mathrm{T}}W^1=0$。根据矩阵迹的特性有：

$$\begin{aligned}&Tr(2\alpha Y^{\mathrm{T}}(I-W^1)^{\mathrm{T}}(I-W^1)Y-Y^{\mathrm{T}}(I-W^2)^{\mathrm{T}}(I-W^2)Y)\\ &=Tr(Y^{\mathrm{T}}(2\alpha(I-W^1)^{\mathrm{T}}(I-W^1)-(I-W^2)^{\mathrm{T}}(I-W^2))Y)\\ &=Tr(Y^{\mathrm{T}}((2\alpha-1)I+((1-\alpha)W^2-\alpha W^1)^{\mathrm{T}}+((1-\alpha)W^2-\alpha W^1)+\alpha(W^1)^{\mathrm{T}}W^1-(1-\alpha)(W^2)^{\mathrm{T}}W^2)Y)\\ &=(2\alpha-1)Tr(Y^{\mathrm{T}}Y)+Tr(Y^{\mathrm{T}}((1-\alpha)W^2-\alpha W^1)^{\mathrm{T}}Y)+Tr(Y^{\mathrm{T}}((1-\alpha)W^2-\alpha W^1)Y)+Tr(Y^{\mathrm{T}}(\alpha W^1-(1-\alpha)W^2)(W^1+W^2)Y)\\ &=(2\alpha-1)Tr(Y^{\mathrm{T}}Y)+Tr(Y^{\mathrm{T}}2((1-\alpha)W^2-\alpha W^1)^{\mathrm{T}}Y)+Tr(Y^{\mathrm{T}}(\alpha W^1-(1-\alpha)W^2)(W^1+W^2)Y)\\ &=(2\alpha-1)Tr(Y^{\mathrm{T}}Y)+Tr(Y^{\mathrm{T}}(W^2-2\alpha W^1)^{\mathrm{T}}(2I-(W^1+W^2))Y)\end{aligned}$$

证毕。

这样，将通过 Lagrange 乘子法获得的 d 个具有代表性的特征向量作为最终嵌入的结果。下面，给出具体的 ILLE 算法。

ILLE 算法具体步骤如下：

输入：N 个样本点、维数为 D 的数据 X，以及样本点的近邻数 k；

输出：维数为 d 的低维嵌入结果 Y；

1) 使用最大似然评估 MLE 方法[18]对数据 X 的固有维度进行合理的评估，得到嵌入维度 d；

2) 寻找每个样本点 x_i 的 k 个近邻；

3) 根据式(1)计算局部重建权值 $\{w_{ij}\}$；

4) 映射数据 $X\to Y$ 至低维空间 R^d 中；

5) 通过解决最优化问题(6)来获得全局低维数据 Y。

注意，由于求解 LLE 目标函数中的 $Tr(Y^{\mathrm{T}}(I-W)^{\mathrm{T}}(I-W)Y)$ 和求解 ILLE 目标函数的 $(2\alpha-1)Tr(Y^{\mathrm{T}}Y)+Tr(Y^{\mathrm{T}}((1-\alpha)W^2-\alpha W^1)^{\mathrm{T}}(2I-(W^1+W^2))Y)$ 具有相同的时间复杂性，因此，ILLE 算法与 LLE 算法的时间复杂性相同。

2.2 基于 DSS 的类_特征关联离散化方法

本小节首先介绍一些相关知识，包括对 CAIM 算法和 DSS 理论进行简单介绍。

2.2.1 CAIM 算法简介

CAIM 算法是基于类_特征相互关联离散化算法的典型代表。其离散化的目标是将连续特征值划分成为有限个相邻区间。例如，一个决策表包含 M 个实例和 S 个决策类，那么总是存在着这样一个集合 $T\{d_1,d_2,\cdots,d_{n-1}\}$，可以将连续特征 C 划分为 n 个区间 $\{[d_0,d_1],(d_1,d_2],\cdots,(d_{n-1},d_n]\}$，其中，$d_0$ 是连续特征 C 的最小值，d_n 是连续特征 C 的最大值。因此，特征 A 中的各个值则会被划分至对应的区间中，且设置一个离散的值。如表1所列，CAIM 算法建立了决策类和连续特征 C 的离散区间的二维矩阵，并定义了离散标准 $caim$ 值的计算方法，如式(8)所示。

表1 特征 C 和离散断点集合 D 的二维矩阵

决策类	区间					类求和
	$[d_0,d_1]$	$\cdots$	$(d_{r-1},d_r]$	$\cdots$	$(d_{n-1},d_n]$	
D_1	q_{11}	$\cdots$	q_{1r}	$\cdots$	q_{1n}	M_1+
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	
D_i	q_{i1}	$\cdots$	q_{ir}	$\cdots$	q_{in}	M_i+
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	
D_s	q_{s1}	$\cdots$	q_{sr}	$\cdots$	q_{sn}	M_s+
区间数	M_{+1}	$\cdots$	M_{+i}	$\cdots$	M_{+n}	M

$$caim(D,T|C)=\frac{\sum_{r=1}^{n}\frac{\max_r^2}{M_{+r}}}{n} \tag{8}$$

其中，$q_{ir}(i=1,2,\cdots,S,r=1,2,\cdots,n)$ 表示实例中属于第 i 类且属于区间 $(d_{r-1},d_r]$ 的个数，max_r 是所有 q_{ir} 中的最大值，M_{i+} 是属于第 i 类的实例个数总和，M_{+r} 是属于区间 $(d_{r-1},d_r]$ 的实例个数总和，n 表示区间个数。$caim$ 值越大表明所选择的断点越合理。

CAIM 算法对各个连续特征按顺序进行离散化，在离散化过程中不需要人为输入参数。起初，设置 $globalcaim=0$，将各待定断点的 $caim$ 值计算出来，找到 $caim$ 的最大值，将其与 $globalcaim$ 值进行比较，如果 $caim>globalcaim$，则向离散断点集合添加此断点，且使得 $globalcaim=caim$，循环此过程；如果 $caim\leqslant globalcaim$，则离散化过程结束。按照顺序对下一个特征性进行离散化，直到将所有连续特征连续域划分完为止。

2.2.2 基于 DSS 的连续特征值离散化算法

DSS 是基于粗糙集理论而被提出的，最早应用于离散特征的约减[19]，在不断改进后，实现了对连续特征的先离散再

约减[20]。通过对类-特征相互关联离散化算法的研究，提出基于DSS理论的连续域特征值划分算法。首先，对具有代表性的CAIM算法进行研究，发现其存在不足：CAIM算法没有考虑分类错误率对离散数据的影响，即离散后的数据中存在着条件特征值相同、决策类不同的错误数据，产生信息丢失，这将对知识提取很不利，会影响后续的机器学习能力。因此，所提出的连续域特征值划分算法首先定义一个基于DSS理论的分类错误率标准，控制离散所导致的信息丢失。

信息系统通常可以表达为$S=\langle U,C,D,V,f\rangle$，其中，$U$是实例的集合，$C\bigcup D=B$是特征集合，子集$C$和$D$分别称为条件特征和决策类，$V$是特征值的集合，$f$是一个信息函数，它指定$U$中每一个对象的特征值。

定义1　i代表第i个实例，j代表第j个实例，DS_{ij}代表当i和j两个实例决策特征不相同时，条件特征也不相同的特征集合。

$$DS_{i,j}=\{\bigcup\{a\in C|f(a,x_i)\neq f(a,x_j)\wedge f(a,x_i)\neq *\wedge f(a,x_j)\neq *\},D(x_i)\neq D(x_j) \tag{9}$$

其中，$*$表示可取该特征值域中的任何值。

定义2　差集是指当i和j两个实例的决策类不相同时，只有一个特征(比如特征a)值不同，而其余特征值均相同，那么，称所有的(i,j)所构成的集合是特征a的一个差集。

$$U_a=\{(i,j)|DS_{i,j}=a\} \tag{10}$$

在决策表中，如果不存在错误的分类，则称这个决策表是一致的，反之，则称这个决策表是不一致的。基于DSS理论，有如下定义：

定义3　$E_i\in U|IND(C)(i=1,2,\cdots,m)$代表$E_i$是$C$的一个等价类，其中$C$表示条件特征集合，$m$表示条件特征的个数。同样，$X_i\in U|IND(D)(i=1,2,\cdots,n)$表示$X_i$是$D$的等价类，其中$D$表示决策类，$n$表示决策类的个数，则特征$E_i$的一致率可由式(11)得到：

$$\mu_{\max}(E_i)=\max(\{|E_i\cap X_j|/|E_i|:X_j\in U|IND(D)\}) \tag{11}$$

定义4　分类错误率可由式(12)得到：

$$\mu(S)=1-\sum_{i=1}^{m}\frac{|E_i|}{|U|}\cdot\mu_{\max}(E_i) \tag{12}$$

其中，$0\leqslant\mu(S)\leqslant1$，表示被错误分类的实例个数与总实例个数的比值。

由于基于类-特征相互关联的CAIM算法未考虑分类错误率带来的影响，本节提出的连续域划分算法对其不足进行改进，即计算当前数据的分类错误率，如果数据存在分类错误，则将当前最大$caim$值的断点加入断点集合中；若加入此断点后，分类错误率不变，则删去此断点，加入次重要特征中最大$caim$值的断点，重复此过程，直到分类错误率不再降低为止。

ILLE-HD3算法具体步骤如下：

输入：M个样本的D维数据，m个连续特征和S个决策类；

输出：离散后的数据集，每个特征有k_i个区间；

第一阶段：使用ILLE算法将原始高维数据降维至d维空间；

第二阶段：在降维后的低维空间进行离散化；

1) 根据式(12)计算原始数据的分类错误率$\mu(S)_{original}$；

2) 对各个连续特征按顺序进行离散化：设置globalcaim=0，将各待定断点的caim值计算出来，找到caim的最大值，将其与globalcaim值进行比较，如果caim>globalcaim，则离散断点集合添加此断点，且使得globalcaim=caim，循环此过程；如果caim≤globalcaim，则离散化过程结束。

计算当前数据的分类错误率$\mu(S)_{current}$，与$\mu(S)_{original}$相比，如果$\mu(S)_{current}$有变化，则计算当前数据的分类错误率，如果数据存在分类错误，则将当前最大$caim$值的断点加入断点集合中；若加入此断点后，分类错误率不变，则删去此断点，加入次重要特征中最大$caim$值的断点，重复此过程，直到分类错误率不再降低为止。

3　实验与结果分析

本节实验的目的是评价ILLE-HD3算法的性能将其分别与类-特征相互关联的CAIM算法[6]、著名的基于Chi2的算法[4]和DDU算法[9]进行比较，对来自UCI机器学习数据库[21]的5组高维数据集进行了离散化处理。数据集的具体相关信息如表2所列。所有实验均在Windows XP操作系统、Xeon(TM) 3.4 GHz CPU以及3G SDRAM内存的PC机上运行。实验工具结合了VC++ 6.0、Matlab 7.x以及数据挖掘工具Waka来完成。

表2　数据集描述

数据集	条件特征	连续特征	类数	实例数
Heart	13	6	2	296
Ionosphere	34	32	2	351
Sonar	0	60	2	208
Isolet	0	617	26	6238
Multi-features	0	649	10	2000

采用ILLE-HD3、CAIM、基于Chi2的算法和DDU方法分别对表2中5组数据集进行离散化，利用10折交叉验证的方法对离散化后的结果进行训练集和测试集划分，此交叉验证方法与CAIM中用到的交叉验证方法相同。通过C4.5决策树和Bayesian分类器[1]对训练集构建分类树，然后对测试集进行分类，实验重复10次，获得平均识别率。

5种算法对数据离散后所产生的分类错误率如表3所列。由表3可以看出，ILLE-HD3算法产生的5组分类错误率是最小的。由于Ext Chi2方法考虑了不一致问题，因此该方法所产生的分类错误率也相对较小。

表3　离散后的分类错误率(%)

算法	Heart	Ionosphere	Sonar	Isolet	Multi-features
Ext Chi2	2.4	14.7	0.6	7.4	4.8
CAIM	5.7	38.1	14	10.7	12.5
ILLE-HD3	1.3	2.5	0	2..6	2.3
DDU	6.1	22.3	3.5	11.1	9.6

对于C4.5，建立分类决策树和Bayesian分类器，对二者识别精度和C4.5规则提取数进行统计，将10次测试结果取平均值，其结果如表4和表5所列，明显看出，ILLE-HD3达到了较好的分类预测结果。

下面对提出的方法在图像上的实验进行评价。首先在三维空间R^3中人工生成S-curve数据集，如图3所示，对改进的ILLE算法进行有效验证。图3(b)为S-curve数据2000个随机采样的散点图。将改进的ILLE算法与传统的LLE算法进行对比，分别选取样本点最近邻数$k=6、12、24$进行实验。从图4中可以清楚看到，ILLE算法保持了原始S-curve数据的几何拓扑结构，对不同的k得到了较好的低维嵌入结果。相反，LLE算法改变了原始S-curve数据的几何拓扑结构。

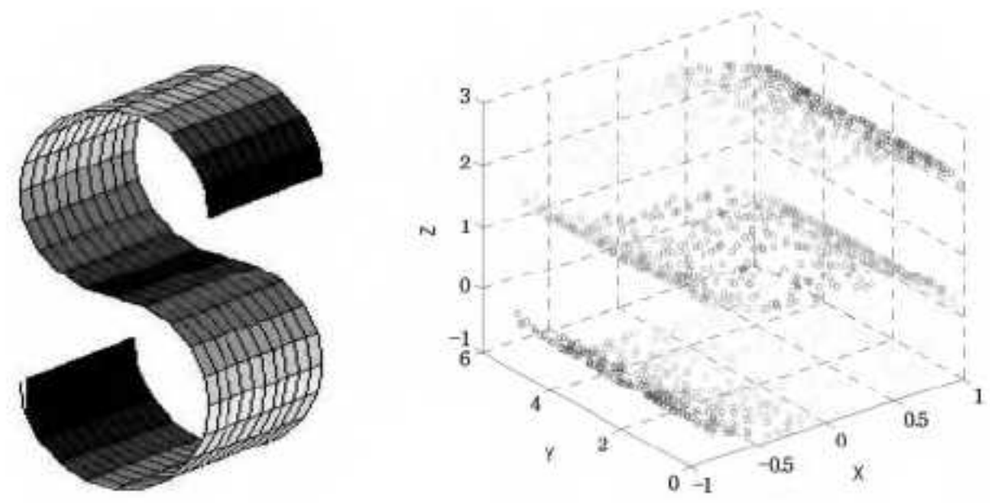
另外一组实验数据是 Frey 人脸数据集[15]，包含 1965 副人脸图像，均为一个人的不同表情，每幅图像处在 560 维空间中。所有图像依据人脸表情分为 5 类：558 副自然表情、630 副高兴表情、627 副不高兴表情、77 副吐舌表情、73 副撅嘴表情。图 5 呈现了由提出的 ILLE-HD3 方法获得的 Frey 人脸数据集二维分类可视化结果（ILLE 选取的最近邻点数为 10），结果表明：在二维空间中本方法仍然能够将高兴、不高兴与自然表情的图像分开，并且可以从细节上分开。

表 4　C4.5 实验结果

数据集	识别精度(%)				规则数 Rules			
	Ext Chi2	CAIM	ILLE-HD3	DDU	Ext Chi2	CAIM	ILLE-HD3	DDU
Heart	76.7	75.6	78.3	77.9	23	25	25	28
Ionosphere	94.9	94.6	97.3	96.5	29	27	30	33
Sonar	94.6	93.5	95.3	94.6	12	12	15	18
Isolet	92.2	91.6	95.3	93.8	152	140	166	174
Multi-features	92.8	92.2	94.5	93.4	86	79	89	94

表 5　Bayesian 分类器实验结果

数据集	分类识别精度(%)			
	Ext Chi2	CAIM	ILLE-HD3	DDU
Heart	77.8	77.4	79.7	78.5
Ionosphere	95.4	95.8	97.9	97.1
Sonar	95.2	93.4	95.9	95.4
Isolet	93.5	93.1	96.3	95.1
Multi-features	93.5	93.1	94.6	93.2



(a) S-curve 数据　(b) S-curve 散点数据，样本数 $N=2000$

图 3

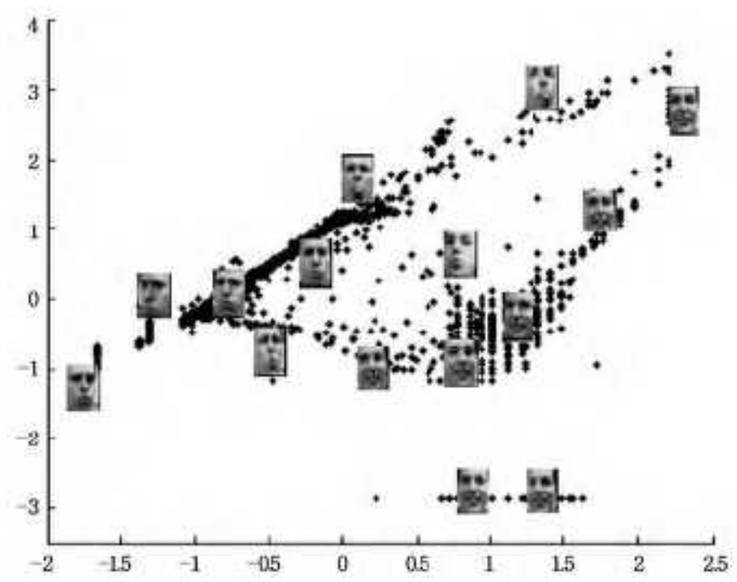
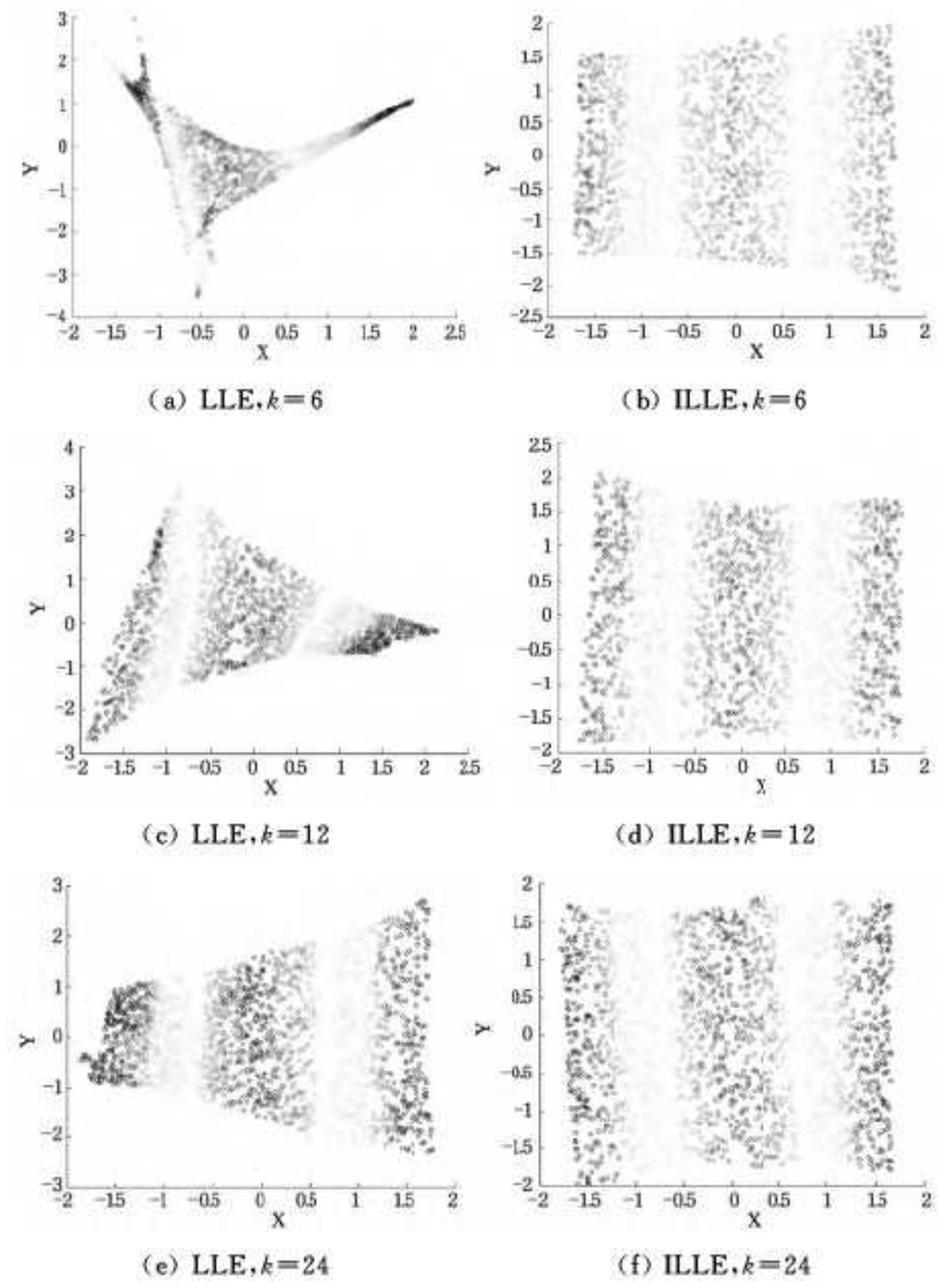


图 5　ILLE-HD3 方法的 Frey 人脸分类 2D 可视化结果

结束语　本文提出一种基于改进局部线性嵌入(LLE)的高维数据离散化 ILLE-HD3 方法，解决了高维数据的离散化问题。在有效降维的基础上，提出了基于类_特征相互关联离散化算法，考虑了分类错误率对离散化的影响，弥补了此类算法产生不一致所导致的缺陷。提出的 ILLE-HD3 方法使得数据集的分类错误率改善较大，实验也取得了较好的分类预测结果，充分显示出所提算法的合理性。



(a) LLE，$k=6$　(b) ILLE，$k=6$

(c) LLE，$k=12$　(d) ILLE，$k=12$

(e) LLE，$k=24$　(f) ILLE，$k=24$

图 4　两种降维方法在 S-curve 数据集上的降维效果对比

参 考 文 献

[1] Wu X D. Top 10 algorithms in data mining [J]. Knowledge Information System, 2008, 14(1): 1-37

[2] Vadera S. CSNL: a cost-sensitive non-linear decision tree algorithm [J]. ACM Transactions on Knowledge Discovery from Data, 2010, 4(2): 1-25

[3] Dougherty J, Kohavi R, Sahami M. Supervised and unsupervised discretization of continuous feature [C] // Proceedings of the 12th International Conference of Machine Learning. San Francisco: Morgan Kaufmann, 1995: 194-202

[4] Su C T, Hsu J H. An extended Chi2 algorithm for discretization of real value attributes [J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(3): 437-441

[5] Fayyad U, Irani K. Multi-interval discretization of continuous-valued attributes for classification learning [C] // Proceedings of the 13th International Joint Conference on Artificial Intelligence. San Mateo, CA: Morgan Kaufmann, 1993: 1022-1027

[6] Cios K J, Kurgan L. CAIM discretization algorithm [J]. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(2): 145-153

[7] 杨萍，杨天社，杜小宁，等. 一种基于类别属性关联程度最大化离散算法[J]. 控制与决策，2011，26(4)：592-596

[8] 赵静娴，倪春鹏，詹原瑞，等. 一种高效的连续属性离散化算法[J]. 系统工程与电子技术，2009，31(1)：195-199

(下转第 157 页)

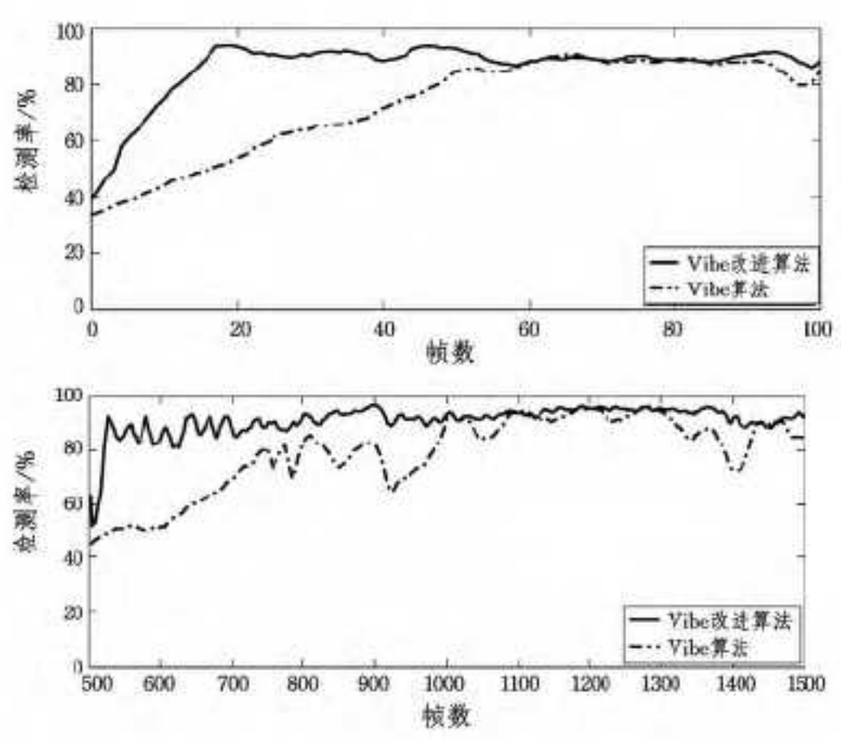


图 6 Pedestrian 与 Highway 视频序列的检测率曲线

Pedestrian 视频序列的行人运动行为简单，从图 6 可以发现，由于鬼影的原因，开始的 60 帧内，检测率不高，而改进的算法由于快速抑制了鬼影，能够快速提高检测率，同时由于本文算法能够自适应调整背景更新率，检测率比较稳定。对于 Highway 视频序列，由于道路的交通情况比较复杂，车辆运动变化较大，传统的 Vibe 算法检测结果不佳，同时检测率也不稳定。改进的 Vibe 算法由于可以根据车辆的运动情况自适应调整更新率，因此检测效果较好，同时比较稳定，鲁棒性较好。实验说明本文提出的算法能够快速地提高目标检测的准确度，同时目标检测的检测率也较为稳定。

结束语 Vibe 算法是一种快速实用的目标检测算法，但是由于背景初始化的问题，容易出现鬼影；同时由于背景更新率恒定，对于前景运动变化较大的情况，检测的准确率有所下降。针对这些问题，本文引入了基于 Otsu 阈值的鬼影抑制以及基于质心运动速度的背景自适应更新方法。对于出现的鬼影，利用 Otsu 算法进行宏观的判别抑制，同时根据当前的检测前景的运动情况，自适应调整背景的更新速率，以适应前景变化较大的情况。为了验证改进算法的有效性，采用了两组不同的视频序列进行了验证，并计算了其每一帧的检测率。实验结果表明，本算法能够很好地抑制鬼影，同时保持较高、较稳定的检测率。

参 考 文 献

[1] 张磊，傅志中，周岳平. 基于 HSV 颜色空间和 Vibe 算法的运动目标检测[J]. 计算机工程与应用，2014(4)：181-185

[2] Wren C R, Azarbayejani A, Darrell T, et al. Pfinder: Real-time tracking of the human body [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19(7): 780-785

[3] Barron J L, Fleet D J, Beauchemin S S. Performance of optical flow techniques [J]. International journal of computer vision, 1994, 12(1): 43-77

[4] 李刚，邱尚斌，林凌，等. 基于背景差法和帧间差法的运动目标检测方法[J]. 仪器仪表学报，2006，27(8)：961-964

[5] Stauffer C, Grimson W E L. Adaptive Background Mixture Models for Real-Time Tracking [C] // Proc. Computer Vision and Pattern Recognition 1999 (CVPR'99). June 1999

[6] Barnich O, Van Droogenbroeck M. iBe: A powerful random technique to estimate the background in video sequences [C] // Proc. Int. Conf. Acoust. Speech Signal Process. Apr. 2009: 945-948

[7] 陈亮 陈晓竹，范振涛. 基于 Vibe 的鬼影抑制算法[J]. 中国计量学院学报，2013，24(4)：425-429

[8] Van Droogenbroeck M, Paquot O. Background subtraction: Experiments and improvements for ViBe [C] // 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, 2012: 32-37

[9] Barnich O, Van Droogenbroeck M. ViBe: A universal background subtraction algorithm for video sequences [J]. IEEE Transactions on Image Processing, 2011, 20(6): 1709-1724

[10] Otsu N. A threshold select ion method from gray-level histograms [J]. IEEE Transactions on System Man and Cybernetic, 1979, 9(1): 62-66

[11] 闵华清，吕居美，罗荣华，等. 基于 GMM 和 MRF 的自适应阴影检测[J]. 华南理工大学学报：自然科学版，2011，39(7)：115-120

[12] 张德才，周春光，周强，等. 基于轮廓的孔洞填充算法[J]. 吉林大学学报：理学版，2011，49(1)：82-86

(上接第 150 页)

[9] Jin R, Breitbart Y, Muoh C. Data discretization unification [C] // The Seventh IEEE International Conference on Data Mining (ICDM Best Paper). 2007: 183-192

[10] 史志才，夏永祥，周金祖. 基于粒计算的离散化算法及其应用[J]. 计算机科学，2013，40(6A)：133-135

[11] 汪凌. 一种基于改进粒子群的连续属性离散化算法[J]. 计算机工程与应用，2013，49(21)：29-32

[12] 徐菲菲，魏莱，杜海洲，等. 一种基于互信息的模糊粗糙分类特征基因快速选取方法[J]. 计算机科学，2013，40(7)：216-221

[13] Ruiz F J, Anguio C, Agell N. IDD: a supervised interval distance-based method for discretization [J]. IEEE Transactions on Knowledge and Data Engineering, 2008, 20(9): 1230-1238

[14] Bondu A, Boulle M, Lemaire V, et al. A non-parametric semi-supervised discretization method [C] // The Eighth IEEE International Conference on Data Mining (ICDM). 2008: 53-62

[15] Armengol E, Garcia-cerdana A. Refining discretizations of continuous-valued attributes [C] // Modeling of Decisions of Artificial Intelligence Conference, LNAI. Springer, Heidelberg, 2012: 258-269

[16] Salvador G, Julian L, Antonio S J, et al. A survey of discretization techniques: taxonomy and empirical analysis in supervised learning [J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(4): 734-750

[17] Roweis S, Saul L. Nonlinear dimensionality reduction by locally linear embedding [J]. Science, 2000, 290(5500): 2323-2326

[18] Levina E, Bickel P J. Maximum likelihood estimation of intrinsic dimension [C] // Advances in Neural Information Processing Systems. 2005

[19] Wu M, Xia D L, Yan P L. A new knowledge reduction method based on difference-similitude set theory [C] // Proceedings of the Third International Conference on Machine Learning and Cybernetics. 2004: 1413-1418

[20] Wu M, Xia D L, Yan P L. Discretization algorithm based on difference-similitude set theory [C] // Proceedings of the Fourth International Conference on Machine Learning and Cybernetics. 2005: 1752-1755

[21] Blake C L, Merz C J. UCI repository of machine learning databases [OL]. http//: www. ics. uci. edu/~mlearn/MLRepository. html