DOI:10.19651/j.cnki.emt.2105758

基于卡方差异性和 t-SNE 的定性数据分类研究

张 蕾
(陕西电子信息职业技术学院 西安 710500)

摘 要：针对定性数据环境下分类精度低且计算成本高的问题，提出了一种利用传统分类器和不同映射技术来提高类别可分性的分类变量识别方法。通过将初始特征(分类属性)映射到实数域空间，利用卡方距离(C-S)作为差异性的度量，增加特征空间的维数以提高类的可分性。运用 t-分布领域嵌入算法(t-SNE)将数据的维数降到 2 或 3 个特征，从而减少了学习方法的计算时间。通过在公共分类数据集上的实验证明，C-S 映射和 t-SNE 在保证识别精度的同时，大大减少了识别任务的计算量。同时，当只将 C-S 映射应用于数据集时，类别的可分性得到了增强，从而显著地提高了学习算法的性能。

关键词：卡方距离；t-SNE；数据分类；差异性
中图分类号：TP311.13 **文献标识码**：A **国家标准学科分类代码**：120.1010

Qualitative data classification based on chi-square dissimilarity and t-SNE

Zhang Lei
(Shaanxi Electrronic Information Institute, Xi′an 710500, China)

Abstract: To solve the problem of low classification accuracy and high computational cost in the qualitative data environment, a classification variable identification method was proposed to improve the classification separability by using traditional classifiers and different mapping techniques. By mapping the initial feature (classification attribute) to the real domain space and using the chi-square (C-S) as the measure of difference, the dimension of the feature space is increased to improve the class separability. The t-distributed domain embedding algorithm (t-SNE) is used to reduce the dimension of the data to two or three features, thus reducing the calculation time of the learning method. It is proved by experiments on the common classification data set that C-S mapping and t-SNE not only guarantee the recognition accuracy, but also greatly reduce the computation of recognition task. At the same time, when only C-S mapping is applied to the data set, the separability of categories is enhanced, thus significantly improving the performance of the learning algorithm.

Keywords: chi-square distance; t-SNE; data classification; differences

0 引 言

数据分类是通过其类别的属性或特征对具有某种共同属性或特征的数据进行区别归并[1]。正确的数据处理需要提前得到数据库类型，而用于数据分析的算法和方法主要集中在定量数据[2]。现有的数据分类大多采用决策树进行识别，该方法由于鲁棒性差而存在局限性，并且对于验证数据的性能也较低(泛化能力低)[3-4]。数据具有高度重叠的特殊性，为了实现自动标记并正确识别数据，文献[5]提出了基于模糊 C-均值的无监督方法，但其计算时间较长。为此，引入了相似系数、相似性度量、围绕中心点的划分(PAM)和聚类层次等方法对数据进行降维处理进而加快分类速度[6-7]。针对定性数据(调查、测试、投票等)的复杂性和重叠性，数据在未经处理或映射的条件下难以实现准确的分类识别。因此，定性数据的分类需要借助有监督的分类方法。文献[8]使用卡方距离(C-S)使 K 均值适应不同的空间，其目的是通过 C-S 将分类特征映射到另一个维数更高、类别更可分离的空间。同时，定性数据映射到更高维空间后面临计算复杂度高的困境。t-分布领域嵌入算法(t-SNE)作为一种非线性降维的机器学习算法[9]，在低维空间下，t-分布可替代高斯分布表达两点之间的相似度。文献[10]指出，t-SNE 可在更小的输入空间(二维或三维)中保留数据结构。因此，定性数据通过映射高维增加可分性后，再借助学习算法压缩计算时间并有效保留数据结构

收稿日期：2021-01-28

成为了研究的重点。

本文提出了一种 C-S 与 t-SNE 相结合的数据分类方法,利用 C-S 差异性将定性数据的整数域输入空间转换为实数域空间,增加特征空间的维数以提高类别的可分性,加快了传统分类器的处理性能。结合 t-SNE 来降低维数,从而减少学习算法的计算时间。最后从精度和计算时间两方面评估了所提出方法的性能。结果表明,C-S 映射和 t-SNE 在保证识别精度的前提下,大大减少了分类任务的计算量。同时,当只将 C-S 映射应用于数据集时,类别的可分性得到了增强,从而显著地提高了学习算法的性能。在不减少输入空间的情况下,使用 C-S 映射分类数据可以获得最佳的识别效果。

1　C-S

C-S 利用包含每个属性频率的列联表得到卡方统计量来衡量个体之间的差异性,C-S 比较两个或两个以上独立特征分类变量的响应计数为:

$$d_{ij}=\sqrt{\sum_{n=1}^{D}\frac{1}{\tilde{x}_n}(\tilde{x}_{in}-\tilde{x}_{jn})^2} \tag{1}$$

其中,

$$\tilde{x}_{in}=\frac{x_{in}}{\sum_{n=1}^{D}x_{in}} \tag{2}$$

$$\tilde{x}_n=\frac{1}{D}\sum_{n=1}^{D}x_{in} \tag{3}$$

式中:D 为特征或维度的数量。C-S 可以提高类别的可分离性并允许进行分组。

2　t-SNE

由于数据映射到不同的空间而导致维数的增加。因此,使用 t-SNE 算法将维数降为 2 或 3 个属性。为了保持数据库的结构,本文在 t-SNE 的距离函数内实现 C-S 度量,在提高分类数据的可分性的同时减少学习算法的计算时间。

t-SNE 最小化了两个分布之间的差异:通过输入目标对 $X=(x_1,x_2,\cdots,x_N)\in R^{D_1}$ 来测量相似性的分布,结合嵌入 $Y=(y_1,y_2,\cdots,y_N)\in R^{D_2}$ 中的低维对应点对来度量相似性的分布,即 $D_1>D_2$。假设有 N 个输入目标组成的数据集 $X=(x_1,x_2,\cdots,x_N)$ 和计算两个目标之间的距离的函数 $d(x_i,x_j)$,t-SNE 定义联合概率 $p_{i,j}$ 来测量 x_i 和 x_j 之间的相似性。

$$p_{i,j}=p_{j,i}=\frac{p_{j|i}+p_{i|j}}{2N} \tag{4}$$

其中,

$$p_{j|i}=\frac{\exp\left(-\frac{d(x_i,x_j)^2}{2\sigma_i^2}\right)}{\sum_{k\neq i}\exp\left(-\frac{d(x_i,x_k)^2}{2\sigma_i^2}\right)} \tag{5}$$

$$p_{i|i}=0 \tag{6}$$

$$\sum_{i,j}p_{i,j}=1 \tag{7}$$

在式(4)中,高斯核的带宽 σ_i 设置为条件分布 p_i 的复杂度等于预定义的复杂度 μ。因此,在数据密度较高的数据空间区域内,σ_i 趋于变小,反之亦然。通过简单的二进制搜索或具有鲁棒的根搜索方法找到每个输入目标的 σ_i 最佳值。

t-SNE 的目标是找到 D_2 维映射 $Y=(y_1,\cdots,y_N)\in R^{D_2}$ 用于反映最优相似性 $p_{i,j}$。因此,以类似的方式测量两点 y_i 和 y_j 之间的相似性。

$$q_{j,i}=\frac{(1+\|y_i-y_j\|^2)^{-1}}{\sum_{k}\sum_{k\neq l}(1+\|y_k-y_l\|^2)^{-1}} \tag{8}$$

$$q_{i,i}=0 \tag{9}$$

归一化 t 检验的重尾允许不同输入目标 x_i 和 x_j 建模低维对应的 y_i 和 y_j。插入点 y_i 的位置通过最小化联合分布 P 和 Q 之间 Kullback-Leibler 的散度来确定。

$$C(\varepsilon)=KL(P\|Q)=\sum_{i\neq j}p_{i,j}\log\frac{p_{i,j}}{q_{i,j}} \tag{10}$$

由于 Kullback-Leibler 散度的非对称性,目标函数侧重于通过嵌入空间中邻近点的高值来模拟 $p_{i,j}$(相似目标)的高值。当沿梯度下降时,目标函数最小化为:

$$\frac{\partial C}{\partial y_i}=4\sum_{j\neq i}(p_{i,j}-q_{i,j})(y_i-y_j) \tag{11}$$

3　标准分类器

在监督学习阶段,本文测试了 4 种标准分类器:线性贝叶斯(LDC)、二次贝叶斯(QDC)、支持向量机(SVM)和最近邻节点(KNN)。目的是通过 C-S 映射来处理分类数据以此提高类别的可分性,并通过 t-SNE 来降低维数。

3.1　SVM

SVM 属于机器学习算法的一类,称为核方法。SVM 中常用的核函数包括:RBG、高斯、线性和多项式等[11-12]。本文选择 RBG 函数作为核函数,这是由于 RBG 函数对不同类型的数据具有灵活性。通过交叉验证设置 RBF 核的 Gamma 和 C 超参数。

3.2　贝叶斯分类器

根据贝叶斯规则,$E=(x_1,\cdots,x_D)$ 是类别 C 的概率为:

$$p(C\mid E)=\frac{p(C\mid E)p(C)}{p(E)} \tag{12}$$

当且仅在以下情况下,E 分类为 $C=+$ 类:

$$f_B(E)=\frac{p(C=+\mid E)}{p(C=-\mid E)}\geqslant 1 \tag{13}$$

式中:函数 $f_B(E)$ 为贝叶斯分类器。假设所有属性都独立于类变量,即:

$$p(E\mid C)=p(x_1,\cdots,x_D\mid C)=\prod_{i=1}^{D}p(x_i\mid C) \tag{14}$$

得到的分类器为：

$$f_N(E)=\frac{p(C=+|)}{p(C=-)}\prod_{i=1}^{D}\frac{p(x_i\mid C=+)}{p(x_i\mid C=-)} \tag{15}$$

式中：函数 $f_N(E)$ 为朴素贝叶斯分类器。LDC 和 QDC 的区别是协方差函数的假设。如果假设所有类别的协方差都相等，则为 LDC，这使得计算预测分布提供了数学便利性，但会丧失泛化能力。如果假设所有类别的协方差不同，则为 QDC，这使得可以更准确地分离非线性数据，但预测分布的计算更复杂。

3.3　KNN

KNN 方法的学习过程基于数据存储。方法描述如下：训练数据 $X=x_1,\cdots,x_N$，具有标签的 $Y=y_1,\cdots,y_N$（N 为数据样本数量）存储在内存中。对于新样本 $x_i\in R^D$，其中 D 是属性数量，在整个训练集中使用距离 d 找到 k 最近邻（k 可以是 1，3，5，7，…）。本文使用了 Manhattan 距离计算距离 d。Manhattan 距离定义为：

$$Manh(x,y)=|(x-y)^{\mathrm{T}}(x-y)| \tag{17}$$

此外，本文测试了 k 为 3、5 和 7 个近邻，但只给出了 $k=3$ 的最佳结果。

4　实验分析

4.1　数据集

本文测试了从 UCI 机器学习库下载的 7 个公共数据集。数据库及其主要特性，如表 1 所示。

表 1　从公共 UCI 机器学习库下载的分类数据集

数据集	样本	特征	类别	类分布
听力学（标准化）（A）	226	69	2	{124，76}
气球（B）	16	4	2	{12，8}
乳腺癌（诊断）（BC）	699	9	2	{458，241}
国际象棋（C）	3 196	36	2	{1 669，1 527}
淋巴影域（LD）	148	18	2	{81，61}
分子生物学（启动子基因序列）（MB）	106	57	2	{63，63}
投票记录（V）	435	16	2	{267，168}

首先，评估余弦（cosine）、杰卡德（Jaccard）、马氏（Mahalanobis）、切比雪夫（Chebychev）、闵氏（Minkowski）、布洛克（City block）、标准化欧氏（Seuclidean）、欧氏（Euclidean）和卡方（chi-square）等 9 种 t-SNE 距离来证明 C-S 度量与 t-SNE 算法相结合提高了数据库分类的可分性。然后，使用 4 种方法（LDC、QDC、SVM、KNN）对数据集进行分类，以此找出最准确的学习方法。

4.2　实验装置

为了对实验结果进行比较，在数据上测试了 4 种不同的实验设置：实验设置 1（单个分类器）、实验设置 2（C-S＋分类器）、实验设置 3（t-SNE＋分类器）和实验设置 4（C-S＋t-SNE＋分类器）。具体的实验设置，如表 2 所示。

在相同的条件下，计算每个设置中所有分类器的精度和计算时间。执行保留验证方案，每个实验重复 10 次，其中 70% 的数据用于训练，30% 的数据用于验证。在 Intel(R)Xeon(R)、CPU E5-2650 V2-2.60GHz、2 个八核处理器和 280 GB RAM 的电脑上使用 MATLAB 软件进行仿真。

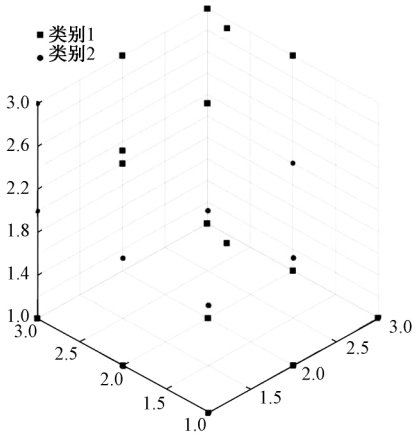
表 2　实验设置

实验设置	描述
（A）	数据集（A）＋分类器
（B）	数据集（B）＋分类器
（BC）	数据集（BC）＋分类器
（C）	数据集（C）＋分类器
（LD）	数据集（LD）＋分类器
（MB）	数据集（MB）＋分类器
（V）	数据集（V）＋分类器
（A）＋（C-S）	数据集（A）＋C-S＋分类器
（B）＋（C-S）	数据集（B）＋C-S＋分类器
（BC）＋（C-S）	数据集（BC）＋C-S＋分类器
（C）＋（C-S）	数据集（C）＋C-S＋分类器
（LD）＋（C-S）	数据集（LD）＋C-S＋分类器
（MB）＋（C-S）	数据集（MB）＋C-S＋分类器
（V）＋（C-S）	数据集（V）＋C-S＋分类器
（A）＋（C-S）＋（t-SNE）	数据集（A）＋C-S＋t-SNE＋分类器
（B）＋（C-S）＋（t-SNE）	数据集（B）＋C-S＋t-SNE＋分类器
（BC）＋（C-S）＋（t-SNE）	数据集（BC）＋C-S＋t-SNE＋分类器
（C）＋（C-S）＋（t-SNE）	数据集（C）＋C-S＋t-SNE＋分类器
（LD）＋（C-S）＋（t-SNE）	数据集（LD）＋C-S＋t-SNE＋分类器
（MB）＋（C-S）＋（t-SNE）	数据集（MB）＋C-S＋t-SNE＋分类器
（V）＋（C-S）＋（t-SNE）	数据集（V）＋C-S＋t-SNE＋分类器
（A）＋（t-SNE）	数据集（A）＋t-SNE＋分类器
（B）＋（t-SNE）	数据集（B）＋t-SNE＋分类器
（BC）＋（t-SNE）	数据集（BC）＋t-SNE＋分类器
（C）＋（t-SNE）	数据集（C）＋t-SNE＋分类器
（LD）＋（t-SNE）	数据集（LD）＋t-SNE＋分类器
（MB）＋（t-SNE）	数据集（MB）＋t-SNE＋分类器
（V）＋（t-SNE）	数据集（V）＋t-SNE＋分类器

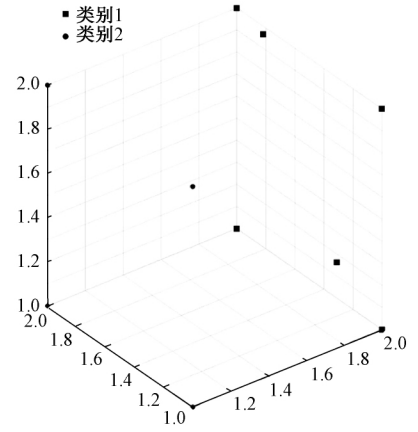
4.3　结果分析

C-S 因其数学性质而适用于分类数据，分类差异增加了数据的维度，将数据映射到实数域，并改进了类别的分离。使用 t-SNE 进行降维来避免计算复杂度。本文假设将分类属性映射到实数域，而不是整数域，并增加了可分离性。

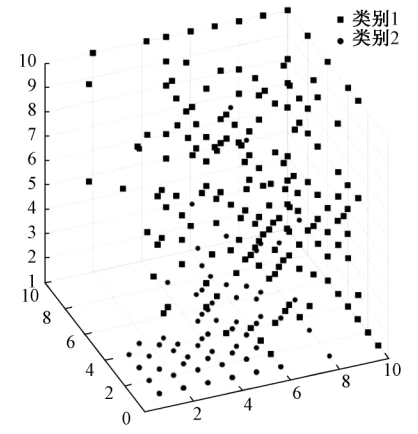
原始输入空间的 3 个属性,如图 1 所示。对应的 C-S+t-SNE 映射特征,如图 2 所示。其中,子图的每个维度都是随机特征。本文展示了 7 个数据库中的 3 个(投票记录、气球和乳腺癌)。可以看到图 1 中的原始输入空间高度重叠,特征只取整数值。相反,在图 2 中,利用 C-S 映射数据集,提高了数据的可分性,再使用 t-SNE 进行降维仍保留了 3 个维度。



(a) 投票记录(V)

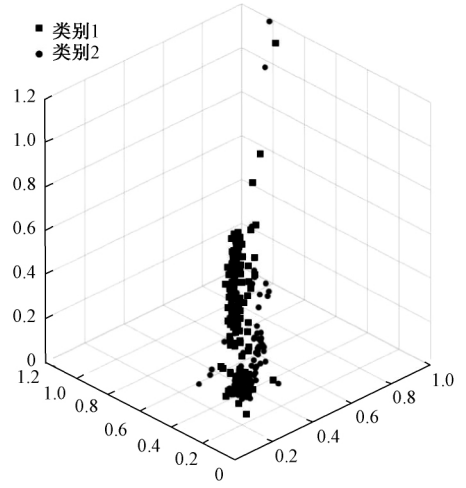


(b) 气球(B)

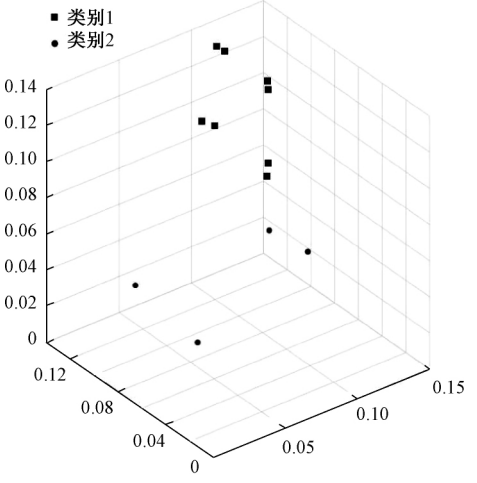


(c) 乳腺癌(BC)

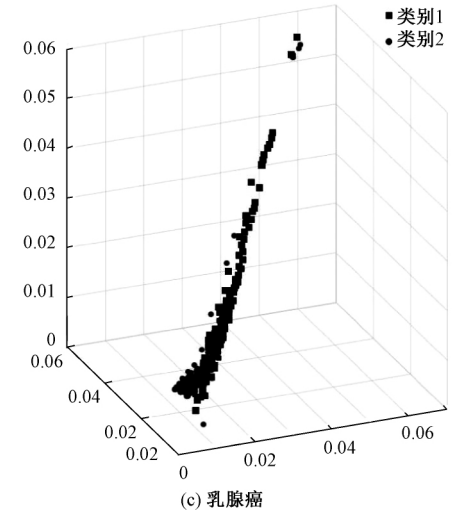
图 1　原始输入空间的属性



(a) 投票记录



(b) 气球



(c) 乳腺癌

图 2　对应的 C-S+t-SNE 映射特征

在数据库中使用 t-SNE 算法时,LDC、QDC、SVM 和 KNN 的精度和标准差,如表 3 所示。目的是评价 t-SNE 方法中常用的距离,并证明 C-S 最适合分类属性。由表 3 可以看到,在大多数情况下,C-S 优于其他距离方法,具有

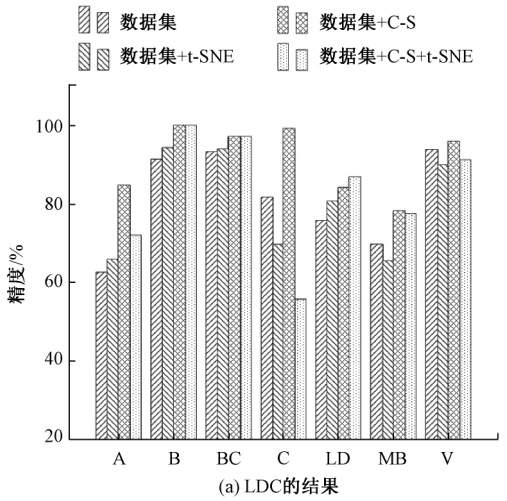
统计显著性差异。同时，t-SNE 在不丢失相关信息或数据　　结构的情况下，降低了映射数据的维数。

表 3　在 7 个 UCI 公共数据集上 t-SNE 算法几个距离的分类精度　　%

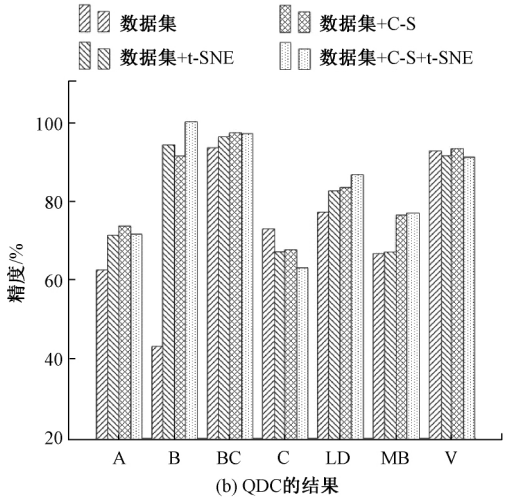
数据集（A）	余弦距离	杰卡德距离	马氏距离	切比雪夫距离	闵氏距离	布洛克距离	标准化欧氏距离	欧氏距离	卡方距离
LDC	62.5±0.0	70.3±0.1	71.1±0.1	60.0±0.4	69.2±0.0	62.8±0.0	61.2±0.0	66.4±0.0	73.4±0.1
QDC	72.8±0.1	83.1±0.0	70.7±0.0	58.5±0.1	73.6±0.3	73.9±0.0	55.6±0.0	69.3±0.0	84.6±0.0
KNN	82.8±0.0	84.8±0.1	77.2±0.0	79.3±0.1	84.4±0.1	85.1±0.0	63.9±0.1	80.8±0.0	88.9±0.0
SVM	62.3±0.0	70.8±0.1	71.1±0.1	62.3±0.0	62.6±0.0	60.3±0.0	62.3±0.0	64.3±0.0	76.7±0.1
平均值	70.1	77.2	72.5	65.0	72.5	70.5	60.7	70.2	80.9
数据集（B）	余弦距离	杰卡德距离	马氏距离	切比雪夫距离	闵氏距离	布洛克距离	标准化欧氏距离	欧氏距离	卡方距离
LDC	74.3±0.2	82.9±0.1	75.7±0.2	78.6±0.2	72.9±0.1	92.9±0.1	54.3±0.1	92.9±0.1	97.1±0.1
QDC	71.4±0.2	74.3±0.1	71.4±0.1	75.7±0.2	84.3±0.1	84.3±0.1	75.7±0.2	81.4±0.2	91.4±0.0
KNN	75.7±0.1	85.7±0.1	88.6±0.1	78.6±0.2	85.7±0.1	94.3±0.1	67.1±0.1	95.7±0.0	100±0.0
SVM	72.9±0.1	84.3±0.1	85.7±0.2	78.6±0.2	70.0±0.1	94.3±0.1	58.6±0.1	90.0±0.1	97.1±0.1
平均值	73.6	81.8	80.4	77.9	78.2	91.5	63.9	90.0	96.4
数据集（BC）	余弦距离	杰卡德距离	马氏距离	切比雪夫距离	闵氏距离	布洛克距离	标准化欧氏距离	欧氏距离	卡方距离
LDC	88.3±0.0	94.3±0.0	78.3±0.0	96.3±0.0	95.6±0.0	96.6±0.0	96.5±0.0	96.4±0.0	96.9±0.0
QDC	90.6±0.0	93.4±0.0	89.2±0.0	96.6±0.0	96.6±0.0	97.3±0.0	96.5±0.0	96.4±0.0	97.3±0.0
KNN	90.1±0.0	95.3±0.1	91.8±0.0	96.7±0.0	96.7±0.0	97.5±0.0	96.7±0.0	97.1±0.0	97.4±0.0
SVM	88.1±0.0	94.4±0.0	79.1±0.0	95.5±0.0	95.5±0.0	96.6±0.0	96.5±0.0	96.5±0.0	97.2±0.0
平均值	89.3	94.4	84.6	96.1	96.1	95.7	96.5	96.6	97.2
数据集（C）	余弦距离	杰卡德距离	马氏距离	切比雪夫距离	闵氏距离	布洛克距离	标准化欧氏距离	欧氏距离	卡方距离
LDC	60.8±0.0	59.7±0.0	57.8±0.0	50.3±0.0	60.9±0.0	55.3±0.0	62.4±0.0	60.8±0.0	68.2±0.0
QDC	65.4±0.0	60.1±0.0	58.9±0.0	53.9±0.0	62.1±0.0	63.1±0.0	64.1±0.0	65.2±0.0	65.5±0.0
KNN	88.5±0.0	70.8±0.0	84.3±0.0	53.0±0.0	89.4±0.0	89.5±0.0	85.9±0.0	89.1±0.0	89.7±0.0
SVM	62.6±0.0	60.7±0.0	58.6±0.0	52.2±0.0	61.5±0.0	60.8±0.0	62.5±0.0	61.1±0.0	68.7±0.0
平均值	69.3	62.8	64.9	52.4	68.5	67.2	68.7	69.1	73.8
数据集（LD）	余弦距离	杰卡德距离	马氏距离	切比雪夫距离	闵氏距离	布洛克距离	标准化欧氏距离	欧氏距离	卡方距离
LDC	76.6±0.0	71.6±0.1	68.9±0.1	64.3±0.1	65.9±0.1	76.1±0.0	76.6±0.0	72.0±0.1	81.6±0.1
QDC	77.3±0.1	76.1±0.0	64.1±0.1	67.5±0.1	67.0±0.1	77.5±0.0	78.6±0.0	73.6±0.1	81.1±0.1
KNN	79.1±0.1	76.4±0.1	79.1±0.1	72.5±0.1	74.8±0.1	80.7±0.0	83.4±0.0	78.6±0.1	84.0±0.0
SVM	75.0±0.0	71.1±0.1	68.1±0.1	66.1±0.1	68.6±0.1	75.9±0.5	78.9±0.0	70.7±0.0	81.8±0.1
平均值	77.0	73.8	70.0	67.6	69.1	77.6	79.4	73.7	82.9
数据集（MB）	余弦距离	杰卡德距离	马氏距离	切比雪夫距离	闵氏距离	布洛克距离	标准化欧氏距离	欧氏距离	卡方距离
LDC	47.8±0.1	56.2±0.1	60.0±0.1	43.1±0.1	60.3±0.0	72.5±0.1	62.5±0.1	55.0±0.1	76.2±0.1
QDC	57.2±0.1	71.2±0.1	54.1±0.1	57.5±0.1	68.7±0.0	74.7±0.1	65.3±0.1	58.4±0.1	78.7±0.1
KNN	62.5±0.1	70.9±0.1	65.6±0.1	50.9±0.1	66.2±0.0	75.6±0.1	68.1±0.1	70.6±0.1	80.3±0.1
SVM	52.2±0.1	55.9±0.1	61.6±0.1	44.4±0.1	56.6±0.1	70.3±0.1	63.7±0.1	54.1±0.1	76.6±0.1
平均值	54.9	63.6	60.3	49.0	63.0	73.3	64.9	59.7	78.0
数据集（V）	余弦距离	杰卡德距离	马氏距离	切比雪夫距离	闵氏距离	布洛克距离	标准化欧氏距离	欧氏距离	卡方距离
LDC	90.5±0.0	88.9±0.0	90.0±0.0	73.7±0.0	90.2±0.0	91.4±0.0	80.8±0.0	90.1±0.0	91.5±0.0
QDC	90.5±0.0	90.9±0.0	90.0±0.0	74.5±0.0	92.1±0.0	91.7±0.0	81.4±0.0	90.6±0.0	91.4±0.0
KNN	92.6±0.0	91.7±0.0	91.4±0.0	76.6±0.0	92.3±0.0	93.3±0.0	82.7±0.0	92.3±0.0	93.8±0.0
SVM	91.4±0.0	90.9±0.0	89.8±0.0	75.2±0.0	91.9±0.0	92.3±0.0	81.7±0.0	90.8±0.0	92.6±0.0
平均值	91.2	90.6	90.4	75.0	91.6	92.2	81.6	90.9	93.1

表 2 中 4 种不同实验设置测试所达到的精度,如图 3 所示。

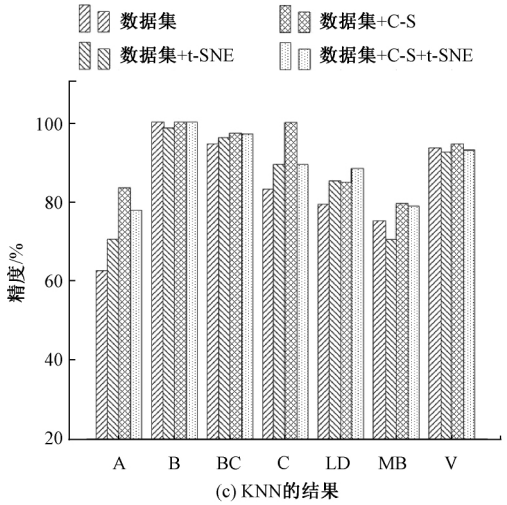
实验设置 1(单个分类器):对分类数据库中的标准分类器进行评估,而不需要对数据进行任何处理或映射。可以观察到,分类结果并不是每个数据集的最佳选择,因此,分类数据必须在识别任务之前进行处理或映射。



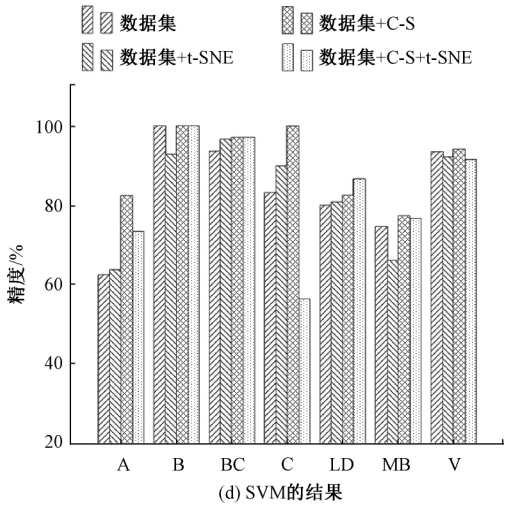
(a) LDC的结果



(b) QDC的结果



(c) KNN的结果



(d) SVM的结果

图 3　在 7 个 UCI 数据库中的公共数据集中测试精度结果

实验设置 2(C-S+分类器):在 C-S 映射的数据集上测试分类器可获得更好的可分性,更高的维数导致计算时间的增加,但 C-S 映射为所有数据集生成最佳的分类结果。C-S 映射将分类数据转换为定量数据,学习方法在这种情况下表现得更佳。这是由于 C-S 映射的主要功能是增加数据的维数来减少分类特征的重叠。分类属性是整数:$X \in Z^D$。当 X 与 C-S 差异映射时,特征域也进行了变换,即当 $K > D$ 时,$X \in Z^D \rightarrow X^* \in R^K$。因此,C-S 映射实现了从分类数据到定量数据的转换。

实验设置 3(t-SNE+分类器):利用 t-SNE 算法将属性维数减少到 3 个。降维减少了计算时间,同时保留了数据结构。精度结果和直接利用单个分类器(实验设置 1)相当,但计算时间得到明显的改善,该设置可适用于在线识别系统。

实验设置 4(C-S+t-SNE+分类器):将 C-S+t-SNE 直接应用于分类数据集上。虽然训练学习算法所需的计算时间较低,但精度受到影响。

从图 3 中看到,在精确度方面最好的设置是实验设置 2(C-S+分类器),当分类特征(整数值)映射到具有更高维度的实数空间(定量数据)上时,可以获得更好的可分性。同时,KNN 作为分类器的分类效果最佳。标准分类器的计算时间,如表 4 所示。

表 4　标准分类器的计算时间

实验设置	计算时间/s	实验设置	计算时间/s
(A)+C-S	33.8	(A)+C-S+t-SNE	24.6
(B)+C-S	1.4	(B)+C-S+t-SNE	0.2
(BC)+C-S	397.0	(BC)+C-S+t-SNE	86.0
(C)+C-S	30.9	(C)+C-S+t-SNE	21.5
(LD)+C-S	25.9	(LD)+C-S+t-SNE	18.3
(MB)+C-S	155.4	(MB)+C-S+t-SNE	52.6
(V)+C-S	2 530.5	(V)+C-S+t-SNE	523.5

为了证明所提出方法的有效性，本文与已有文献的几种分类方法进行了比较：稀疏加权朴素贝叶斯分类器(SWNBC)[13]、耦合属性相似性方法(C4.5)[14]、基于布尔核的分类器(BK)[15]和具有广义概率的朴素分类器(NPC)[16]。使用 7 个数据库中的 5 个数据集与其他方法方法相比，C-S 可以获得更好的分类精度，如表 5 所示。

表 5　C-S 与其他分类法的识别精度　　%

数据集	SWNBC[13]	C4.5[14]	BK[15]	NPC[16]	C-S
(C)	87.59±1.23	97.48±1.85	97.22±1.94	88.67±1.72	100.0±0.00
(V)	90.08±3.71	93.28±3.18	92.36±3.23	94.23±3.62	94.53±1.60
(BC)	72.50±7.71	71.33±6.33	66.45±6.92	73.81±7.11	97.35±1.30
(LD)	83.60±9.82	73.12±8.63	73.82±8.47	87.76±9.60	88.30±4.80
(B)	100.0±0.00	100.0±0.00	100.0±0.00	100.0±0.00	100.0±0.00

5　结　　论

本文实现了一种定性数据的分类识别方法。将范畴属性映射到个具有 C-S 相异性的高维空间。通过将分类数据集的特征域从整数转换为实数，从而缓解了重叠问题，分类数据的映射提高了识别的精度。其次，在 t-SNE 方法中引入了基于 C-S 距离的替代距离方法，提高了数据的可分性。在 UCI 机器学习库下载的公共数据集上测试标准分类器 LDC、QDC、KNN 和 SVM 时，将 C-S＋tSNE 应用于分类数据，在提高了数据的可分性的同时减少了分类的计算时间。此外，通过对比 SWNBC、C4.5、BK 和 NPC 等其他分类方法，从而验证了 C-S 分类识别精度的优越性。本文方法仅对定性数据的可分部分进行了变量识别，在未来的研究中，将该方法改进后推广到定性数据的不可分性的快速分类。

参考文献

[1] 黄裕. DSM-Forest 算法对计算机多类数据学习分类性能的影响[J]. 信息技术，2019，43(5)：148-150，154.

[2] 徐玲玲，迟冬祥. 面向不平衡数据集的机器学习分类策略[J]. 计算机工程与应用，2020，56(24)：12-27.

[3] 刘帅，刘长良，甄成刚. 基于数据分类重建的风电机组故障预警方法[J]. 仪器仪表学报，2019，40(8)：1-11.

[4] 易明雨，肖赤心，潘晖，等. 用于大数据分类的快速隐层优化分布式极限学习机[J]. 计算机工程与应用，2019，55(16)：165-169，203.

[5] 佐磊，胡小敏，何怡刚，等. 小样本数据处理的加速寿命预测方法[J]. 电子测量与仪器学报，2020，34(11)：26-32.

[6] 肖连杰，郜梦蕊，苏新宁. 一种基于模糊 C-均值聚类的欠采样集成不平衡数据分类算法[J]. 数据分析与知识发现，2019，3(4)：90-96.

[7] 史荧中，王士同，邓赵红，等. 基于核心向量机的多任务概念漂移数据快速分类[J]. 智能系统学报，2018，13(6)：935-945.

[8] 刘欢，胡德敏. 类不平衡数据的卡方聚类算法研究[J]. 软件，2019，40(4)：7-10.

[9] 姚舜禹，王雪，邹德财，等. 基于 t-SNE 算法的 ABPSK 信号个体识别[J]. 时间频率学报，2019，42(4)：336-344.

[10] 夏雨薇，石美红，贺飞跃，等. 基于降维融合特征和集成学习的织物疵点分类[J]. 国外电子测量技术，2019，38(7)：86-91.

[11] 魏世超，李歆，张宜弛，等. 基于 E-t-SNE 的混合属性数据降维可视化方法[J]. 计算机工程与应用，2020，56(6)：66-72.

[12] LUIS C P，MARTIN C，ALFONSO R D，et al. A novel formulation of orthogonal polynomial kernel functions for SVM classifiers：The Gegenbauer family[J]. Pattern Recognition，2018，84(1)：211-225.

[13] 赵博文，王灵矫，郭华. 基于泊松分布的加权朴素贝叶斯文本分类算法[J]. 计算机工程，2020，46(4)：91-96.

[14] 李春生，焦海涛，刘澎，等. 基于 C4.5 决策树分类算法的改进与应用[J]. 计算机技术与发展，2020，30(5)：185-189.

[15] MYRIAM B，HENRI P，GILLES R，et al. Oddness/evenness-based classifiers for Boolean or numerical data[J]. International Journal of Approximate Reasoning，2017，82(8)：81-100.

[16] 段尧清，林平，李施展. 基于多类型分类器装袋技术的数据分类模型研究[J]. 情报科学，2019，37(4)：59-65.

作者简介

张蕾，讲师，硕士，主要研究方向为数据分析。

E-mail：ehh6473@163.com