计算机工程与应用
Computer Engineering and Applications
ISSN 1002-8331,CN 11-2127/TP

《计算机工程与应用》网络首发论文

题目：基于 E-t-SNE 的混合属性数据降维可视化方法
作者：魏世超，李歆，张宜弛，周晓锋，李帅
网络首发日期：2019-06-26
引用格式：魏世超，李歆，张宜弛，周晓锋，李帅．基于 E-t-SNE 的混合属性数据降维可视化方法[J/OL]．计算机工程与应用.
http://kns.cnki.net/kcms/detail/11.2127.TP.20190624.1731.006.html

网络首发时间：2019-06-26 15:42:33
网络首发地址：http://kns.cnki.net/kcms/detail/11.2127.TP.20190624.1731.006.html

基于 E-t-SNE 的混合属性数据降维可视化方法

魏世超[1,2,3,4]，李　歆[1,3,4]，张宜弛[1,3,4]，周晓锋[1,3,4]，李　帅[1,2,3,4]

WEI Shichao[1,2,3,4], LI Xin[1,3,4], ZHANG Yichi[1,3,4], ZHOU Xiaofeng[1,3,4], LI Shuai[1,2,3,4]

1.中国科学院 沈阳自动化研究所，沈阳 110016
2.中国科学院大学，北京 100049
3.中国科学院 网络化控制系统重点实验室，沈阳 110016
4.中国科学院 机器人与智能制造创新研究院，沈阳 110016

1.Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016,China
2.University of Chinese Academy of Sciences, Beijing 100049, China
3.Key Laboratory of Network Control System, Chinese Academy of Sciences, Shenyang 110016, China
4.Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang 110016, China

WEI Shichao, LI Xin, ZHANG Yichi, et al. Dimension reduction and visualization of mixed-type data based on E-t-SNE. Computer Engineering and Applications

Abstract: Aiming at the problem that the traditional t-SNE algorithm can only deal with single attribute data and can't handle mixed type data very well. An extended t-SNE dimensionality reduction visualization algorithm named E-t-SNE is proposed. The extension facilitates to handle mixed type data. Firstly, the concept of information entropy is introduced to construct the distance matrix of categorical data. Secondly, the distance matrix of mixed type data is constructed by combining the distance between categorical data and the Euclidean distance of numerical data. Finally, the combined matrix is used into t-SNE algorithm to reduce the dimension and display it in two-dimensional space. In addition, in order to verify the effectiveness of the algorithm, K-Nearest Neighbor (KNN) algorithm is used to evaluate. Experiments on UCI datasets show that this method not only has good visualization ability in dealing with mixed attribute data, but also can effectively reduce the dimension of different classes of data and improve the classification accuracy of subsequent classifiers.

Key words: t-SNE algorithm; mixed type data; dimension reduction; visualization

摘　要: 针对传统的 t 分布随机近邻嵌入(t-SNE)算法只能处理单一属型数据，不能很好的处理混合属性数据的问题，提出一种扩展的 t-SNE 降维可视化算法 E-t-SNE，用于处理混合属性数据。首先，该方法引入信息熵概念来构建分类属性数据的距离矩阵，其次采用分类属性数据距离与数值属性数据欧式距离相结合的方式构建混合属性数据距离矩阵，最后将新的距离矩阵输入 t-SNE 算法对数据进行降维并在二维空间可视化展示。此外，为验证算法有效性，采用 K 近邻(KNN)算法对混合数据降维后的效果进行评价。通过在 UCI 数据集上的实验表明，该方法在处理混合属性数据方面，不仅具有较好的可视化能力，而且能有效地对不同类别的数据进行降维分簇，提升后续分类器的分类准确率。

关键词: t-SNE 算法；混合属性数据；降维；可视化

文献标志码： A　**中图分类号：** TP391　doi：10.3778/j.issn.1002-8331.1903-0330

基金项目：沈阳市科技计划项目(No.Z18-5-102)。

作者简介：魏世超(1994－)，男，硕士研究生，研究领域为大数据、数据可视化，E-mail: weishichao@sia.cn；李歆(1981－)，男，硕士，副研究员,研究领域为企业信息化规划、制造执行系统平台技术；张宜弛(1986－)，男，硕士，副研究员，研究领域为大数据分析、分布式计算；周晓锋(1978－)，女，博士，副研究员，研究领域为数据挖掘、机器学习；李帅(1988－),男，博士研究生，助理研究员，研究领域为数据挖掘、机器学习、故障诊断。

1 引言

现实生活中的大量数据通常包含可能对决策者有用的隐藏模式，但这些数据通常维度较高。例如，在入侵检测、欺诈检测、医疗分析领域[1]的数据，通常包含数百维。模式识别、图像处理[2]领域的数据通常包含上千个特征。现实数据高维特性的存在带来了计算成本增加、维度灾难[3]等问题，不利于对数据的理解分析。解决数据高维特性问题的一种方法是降维，即寻求减少数据特征数量的技术。

众多研究者提出了多种降维技术。一种是基于特征选择的方法，根据一些标准选择原始特征的子集[4]。Danubianu[5]等人提出了一种应用相关性的筛选器来寻找相关特征的特征选择方法。另一种降维技术是基于特征变换的方法，通过指定的变换函数将高维数据映射到低维空间。与特征选择方法不同，生成的特征集不是原始特征的子集，而是根据原始特征新创建的。目前基于特征变换的降维方法主要有主成分分析(PCA)[6]，局部线性嵌入(LLE)[7]，等距映射(Isomap)[8]，局部切空间排列(LTSA)[9]，随机近邻嵌入(t-SNE)[10]，拉普拉斯特征映射(LE)[11]等。这些方法已经被应用于多个领域。Jamieson[12]等人将拉普拉斯特征映射和 t-SNE 应用于计算机提取的乳腺癌特征空间中，对医学图像映射进行分类和可视化检查。Garces 等人[13]运用 t-SNE 将不同风格的剪切画可视化。Liu 等人[14]采用 LLE 对肿瘤基因表达数据进行降维。

以往研究的不足之处在于研究都是在数值数据的背景下进行的。然而大多数真实世界的数据集同时包含分类属性和数值属性。例如，信用系统的数据包括年龄、年薪、储蓄金额等数值属性，以及教育背景、职业、婚姻状况等分类属性[15]。许多知识发现算法并不能处理混合类型数据。这些算法只分析数值或分类数据，并通过将一种类型的数据转换为另一种类型的数据来解决这一缺陷。对于只处理数值型数据的算法，独热编码(One-Hot Encoding)是一种普遍采用的方法，它将每个分类值转换为二进制向量。然而这种方法有几个显著的缺点。首先，转换后的数据增加了维度，计算成本也随之增加，其次将分类属性转换为二进制向量，未考虑属性间的相互关系，造成数据语义丢失，影响后续的分类聚类等算法的精度和性能。

本文主要针对 t-SNE 算法不能处理混合数据的缺点，提出一种 E-t-SNE 算法扩展其对混合属性数据的处理能力。该算法引入信息熵的概念，考虑不同分类属性值对距离计算过程中不同的贡献程度，然后采用与数值属性加权结合的方式，构造混合属性数据的距离矩阵。为了验证该算法处理混合属性数据的有效性，采用 k 近邻(kNN,k-Nearest Neighbor)[16]分类算法验证降维后数据的分类精度优于其他距离度量方案。此外，将混合属性数据投影到低维空间进行可视化，证明了该算法对混合数据集的可视化展示更符合人们对数据的直观理解。

2 相关定义

2.1 t 分布随机近邻嵌入算法

高维数据降维可视化处理是将维数大于 3 的数据转换为 2 维或者 3 维，降维后的数据能够在低维空间可视化展示并可以用于后续的机器学习算法。Hinton 及 Roweis[17]于 2002 年提出了一种随机近邻嵌入(SNE)降维方法。由于 SNE 方法的价值方程优化困难并且存在低维数据拥挤问题，因此 Maaten 及 Hinton 于 2008 年在 SNE 的基础上提出基于 t 分布的随机近邻嵌入方法(t-SNE)。t-SNE 的思想是在低维空间中采用重尾 t 分布构造一个概率分布，使其在高维空间中构造的概率分布相似。在概率分布中，相似的数据点被选中的概率较高，不相似的数据点被选中的概率较低。高维空间中，点 j 相对于点 i 的概率分布定义为

$$p_{j|i}=\frac{\exp(-\left\|x_i-x_j\right\|^2/2\sigma^2)}{\sum_{k\neq i}\exp(-\left\|x_i-x_k\right\|^2/2\sigma^2)} \tag{1}$$

其中 σ 是以 x_i 为中心的高斯函数的方差，由二进制搜索算法确定。设高维空间中的联合概率 p_{ij} 为对称条件概率，即

$$p_{ij}=\frac{p_{j|i}+p_{i|j}}{2n} \tag{2}$$

其中 n 是数据点的个数。对于高维数据 x_i 和 x_j 在低维空间中的映射 y_i 和 y_j，采用自由度为 1 的 t 分布表示低维联合概率分布，即

$$q_{ij}=\frac{(1+\left\|y_i-y_j\right\|^2)^{-1}}{\sum_{k\neq l}(1+\left\|y_k-y_l\right\|^2)^{-1}} \tag{3}$$

为保证低维空间映射点之间的联合概率分布 q_{ij} 较好的模拟高维空间数据点之间的联合概率分布 p_{ij}，采用梯度下降法最小化所有数据点的 KL 散度(Kullback-Leibler divergences)得到低维空间最佳模拟点。目标函数 C 和梯度下降法优化的定义如下：

$$C = KL(P \| Q) = \sum_{i,j} p_{ij} \mathrm{lb} \frac{p_{ij}}{q_{ij}} \tag{4}$$

$$\frac{\delta C}{\delta y_i} = 4\sum_j (p_{ij} - q_{ij})(y_i - y_j)(1 + \left\| y_i - y_j \right\|^2)^{-1} \tag{5}$$

为了加速优化过程，避免陷入较差的局部最小值，在梯度中加入一个动量项：

$$Y^{(t)} = Y^{(t-1)} + \eta \frac{\delta C}{\delta Y} + \alpha(t)(Y^{(t-1)} - Y^{(t-2)}) \tag{6}$$

2.2 独热编码

独热编码(One-Hot Encoding)是机器学习算法中处理分类属性数据的常用方法。由于很多现实世界中的数据集不总是连续的数值型数据，在应用回归，分类，聚类等机器学习算法时，数据集中包含的分类型数据无法直接进行距离和相似度计算。独热编码将具有 k 个不同值域的分类属性转换为一组 k 个二进制属性，每个二进制属性都对应于 k 个分类值中的一个。

然而这种对分类属性数据的处理方式缺点显著。首先转换后的数据大大增加了数据维度，计算成本增加。其次，将分类属性转换为二进制向量，未考虑属性间的相互关系，造成数据语义丢失，影响后续的分类聚类等算法的精度和性能。

2.3 k 近邻分类算法

k 近邻(kNN,k-Nearest Neighbor)分类算法是最常用的分类算法之一，被应用于数据挖掘、模式识别等多个领域。算法采用 k 个最近邻居的类标签来确定未知点的类标签，即 k 个最近邻文本中大多数属于某个类别，则样本也属于这个类别，故 k 值的选取至关重要。如果 k 值偏小，会提高噪声的干扰；如果 k 值偏大，并且测试样本中属于训练集中包含数据较少的类，则会增加噪声，降低分类准确性。

3 E-t-SNE 混合属性数据降维可视化方法

观察式(1)可以发现，t-SNE 算法将高维数据点间的欧式距离转换为表示相似性的条件概率[18]，常用的空间距离计算方法欧式距离并不适合处理分类型数据。本文采用一种新的距离计算方法，分别构建数值型数据的距离矩阵和分类型数据距离矩阵，然后将两个矩阵融合得到混合数据的距离矩阵，输入 t-SNE 算法进行降维并可视化。

3.1 距离矩阵度量

关于混合属性数据的公式符号描述如下[19]：$X = \{x_1, x_2, \cdots, x_n\}$ 表示 N 个混合数据对象的数据集，对于每一个 $x_i (1 \le i \le n)$，混合型数据集 x_i 代表 M $(M = M_n + M_c)$ 个属性 $A_1^{(n)}, A_2^{(n)}, \cdots A_{M_n}^{(n)}, A_{M_n+1}^{(c)}, \cdots A_{M_n+M_c}^{(c)}$，其中 $A_1^{(n)}, A_2^{(n)}, \cdots A_{M_n}^{(n)}$ 表示 M_n 个数值型数据，$A_{M_n+1}^{(c)}, \cdots A_{M_n+M_c}^{(c)}$ 表示 M_c 个分类型数据。$x_{i,\mathrm{k}}^{(n)}$ 表示数值部分 $x_i^{(n)}$ 的第 k 个属性，$x_{i,\mathrm{k}}^{(c)}$ 表示分类部分 $x_i^{(c)}$ 的第 k 个属性。第 k 列分类属性的定义域表示为 $D(A_{\mathrm{k}}^{(c)}) = \{a_{k,1}, a_{k,2}, \cdots a_{k,t}\}$，$t$ 表示分类属性的个数。

3.1.1 数值型数据的距离矩阵

对于数值型数据，采用欧式距离计算 $\mathrm{x}_i^{(n)}$ 和 $\mathrm{x}_j^{(n)}$ 之间的距离 $d_{ij}^{(n)}(\mathrm{x}_i^{(n)}, \mathrm{x}_j^{(n)})$，公式如下：

$$d_{ij}^{(n)}(\mathrm{x}_i^{(n)}, \mathrm{x}_j^{(n)}) = \sqrt{\sum_{\mathrm{k}=1}^{\mathrm{N}} (x_{i,\mathrm{k}}^{(n)} - x_{j,\mathrm{k}}^{(n)})^2} \tag{7}$$

将计算得到的 $d_{ij}^{(n)}$ 构建成一个反应数值型数据点之间距离的矩阵 $D^{(n)}$，

$$D^{(n)} = \begin{bmatrix} 0 & \cdots & d_{1j}^{(n)} & \cdots & d_{1n}^{(n)} \\ \vdots & \ddots & \vdots & \cdots & \vdots \\ \vdots & \cdots & 0 & \cdots & \vdots \\ \vdots & \cdots & \vdots & \ddots & \vdots \\ d_{n1}^{(n)} & \cdots & d_{nj}^{(n)} & \cdots & 0 \end{bmatrix} \tag{8}$$

$D^{(n)}$ 是一个 $n \times n$(n=N)的矩阵，对角线上元素全是 0，表示数据点自身与自身的距离，$d_{ij}^{(n)}$ 表示 $x_i^{(n)}$ 和 $x_j^{(n)}$ 之间的距离，$d_{ij}^{(n)} = d_{ji}^{(n)}$。

3.1.2 分类型数据的距离矩阵

传统的定义分类属性数据之间距离的方式是建立在每个分类属性权重相同的情况下，数据相同距离为 0，不同距离为 1。本文表示为 $d_{ij}'^{(c)}(x_{i,k}^{(c)}, x_{j,k}^{(c)})$，公式如下：

$$d_{ij}'^{(c)}(x_{i,k}^{(c)}, x_{j,k}^{(c)}) = \begin{cases} 0, \ if \ x_{i,k}^{(c)} = x_{j,k}^{(c)} \\ 1, \ if \ x_{i,k}^{(c)} \ne x_{j,k}^{(c)} \end{cases} \tag{9}$$

这种定义分类属性数据的方法忽略了不同分布的不同属性值对距离计算时贡献度的差异。因此，为了考

虑贡献度的差异，为分类属性加入权重 w_k，公式(9)可修改为：

$$d_{ij}^{(c)}(\mathrm{x}_i^{(c)},\mathrm{x}_j^{(c)})=\sum_{k=1}^{M_c}w_k d_{ij}'^{(c)}(x_{i,k}^{(c)},x_{j,k}^{(c)}) \tag{10}$$

显然 w_k 是分类属性 $A_k^{(c)}$ 的权重，它表示对分类属性距离计算时的贡献程度。这里 $0\le w_k\le 1$，$\sum_{k=1}^{M_c}w_k=1$。

然后讨论每个分类属性权重 w_k 的计算方式，本文引入信息熵的概念计算权重[20]。由信息熵的性质可知，信息熵可以作为一个系统复杂程度的度量，如果系统越复杂，出现不同情况的种类越多，那么他的信息熵就越大，反之一个系统越简单，出现情况种类越少，信息熵越小。同理，如果数据集中分类型数据 $A_k^{(c)}$ 的定义域 $D(A_k^{(c)})=\{a_{k,1},a_{k,2},\cdots a_{k,t}\}$ 中 $a_{k,t}$ 种类越多，分布越不均匀，则 $A_k^{(c)}$ 的信息熵越大，即对距离的计算贡献程度越高，权重越大。因此 $A_k^{(c)}$ 的信息熵可表示为：

$$H_{A_k^{(c)}}=-\sum_{a_{k,t}\in DOM(A_k^{(c)})}p(a_{k,t})\mathrm{lb}(p(a_{k,t})) \tag{11}$$

这里 $p(a_{k,t})$ 是分类属性 $A_k^{(c)}$ 中 $a_{k,t}$ 的概率，

$$p(a_{k,t})=\frac{\sum_{i=1}^{N}d_{ij}'^{(c)}(x_{i,k}^{(c)},a_{k,t})}{N} \tag{12}$$

分子表示分类属性 $A_k^{(c)}$ 中分类值 $a_{k,t}$ 的个数，N 表示分类值的总个数。为了减少 $A_k^{(c)}$ 中过多分类值甚至唯一值对信息熵的影响，分类属性 $A_k^{(c)}$ 的信息熵计算公式可以优化为：

$$H'_{A_k^{(c)}}=-\frac{1}{r_k}\sum_{t=1}^{r_k}p(a_{k,t})\mathrm{lb}(p(a_{k,t})) \tag{13}$$

其中 r_k 为 $A_k^{(c)}$ 中分类值的数量。因此，数据集中分类属性 $A_k^{(c)}$ 对距离计算时贡献程度可用权重 w_k 表示：

$$w_k=\frac{H'_{A_k^{(c)}}}{\sum_{k=1}^{M_c}H'_{A_k^{(c)}}} \tag{14}$$

将式(14)带入式(10)中，得到分类属性数据的距离计算公式：

$$d_{ij}^{(c)}(\mathrm{x}_i^{(c)},\mathrm{x}_j^{(c)})=\sum_{k=1}^{M_c}\left(\frac{H'_{A_k^{(c)}}}{\sum_{k=1}^{M_c}H'_{A_k^{(c)}}}\cdot d_{ij}'^{(c)}(x_{i,k}^{(c)},x_{j,k}^{(c)})\right) \tag{15}$$

最后为适应算法的输入，将计算得到的 $d_{ij}^{(c)}$ 构建成一个反应分类型数据点之间距离的矩阵 $D^{(c)}$：

$$D^{(c)}=\begin{bmatrix}0 & \cdots & d_{1j}^{(c)} & \cdots & d_{1n}^{(c)}\\ \vdots & \ddots & \vdots & \cdots & \vdots\\ \vdots & \cdots & 0 & \cdots & \vdots\\ \vdots & \cdots & \vdots & \ddots & \vdots\\ d_{n1}^{(c)} & \cdots & d_{nj}^{(c)} & \cdots & 0\end{bmatrix} \tag{16}$$

$D^{(c)}$ 是一个 $n\times n$(n=N)的矩阵，对角线上元素全是0，表示数据点自身与自身的距离，$d_{ij}^{(c)}$ 表示 $x_i^{(c)}$ 和 $x_j^{(c)}$ 之间的距离，$d_{ij}^{(c)}=d_{ji}^{(c)}$。

3.1.3 混合型数据的距离矩阵

根据以上内容可以发现，对于混合型数据集的距离计算采用数值型和分类型分别计算的方法。数值型数据点直接采用欧式距离进行度量，分类型数据点引入信息熵的概念，将不同分类属性对距离计算的贡献程度量化成权重，用权重与分类值之间距离相乘的方法构造出分类属性数据点之间的距离。

最后将计算出来的数值型数据点之间的距离与分类型数据点之间的距离相加，得到混合属性数据点之间的距离 $d_{ij}(\mathrm{x}_i,\mathrm{x}_j)$，表示为：

$$\begin{aligned}d_{ij}(\mathrm{x}_i,\mathrm{x}_j)=&\frac{1}{M_c+1}\left(\sqrt{\sum_{\mathrm{k}=1}^{\mathrm{N}}(x_{i,k}^{(n)}-x_{j,k}^{(n)})^2}\right)+\\&\frac{M_c}{M_c+1}\sum_{k=1}^{M_c}\left(\frac{H'_{A_k^{(c)}}}{\sum_{k=1}^{M_c}H'_{A_k^{(c)}}}\cdot d_{ij}'^{(c)}(x_{i,k}^{(c)},x_{j,k}^{(c)})\right)\end{aligned} \tag{17}$$

由式(7)和(15)可知，$d_{ij}^{(n)}(\mathrm{x}_i^{(n)},\mathrm{x}_j^{(n)})$ 和 $d_{ij}^{(c)}(\mathrm{x}_i^{(c)},\mathrm{x}_j^{(c)})$ 取值范围是 0 到 1 之间，为保证两式相加结果范围在 0 到 1 之间，采用式(17)的加权方式。为适应 E-t-SNE 算法的输入，构造混合属性数据距离矩阵 D：

$$D = D^{(n)} + D^{(c)} \tag{18}$$

3.2 算法实现

E-t-SNE 算法流程如图 1 所示：

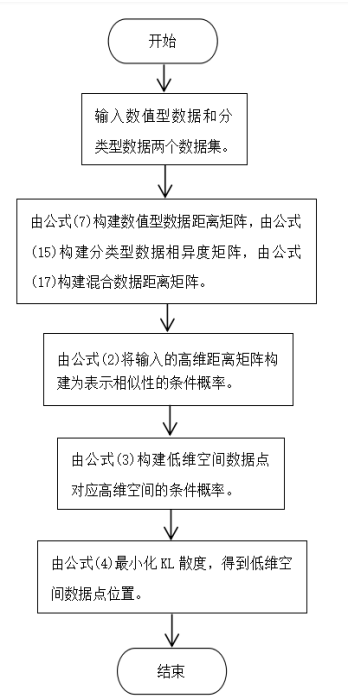


图 1　E-t-SNE 算法流程

E-t-SNE 算法描述如表 1 所示，其中迭代次数 太小容易导致优化过程不完全，太大增加算法执行时间，本文设置为 1000；较大的动量项可以使优化过程避免陷入较差的局部最优，根据经验，当迭代次数 $T < 250$ 时，动量项 $\alpha(T) = 0.5$，当 $T \geq 250$ 时， $\alpha(T) = 0.8$；学习率 η 初值为 100，每次迭代结束根据自适应学习率机制进行更新；维数 设置为 2[21]。

表 1　E-t-SNE 算法

输入：数值型数据集 $X^{(n)} = \{x_1, x_2, \cdots, x_n\}$，分类型数据集 $X^{(c)} = \{x_{n+1}, x_{n+2}, \cdots, x_{n+c}\}$，困惑度因子 $Perp$，迭代次数 T，学习率 η，动量 $\alpha(T)$，维数 m。

(1)分别根据式(7)和式(15)构建数值型数据距离矩阵 $D^{(n)}$ 和分类型数据距离矩阵 $D^{(c)}$；根据公式(17)构建混合数据距离矩阵 D；

(2)分别根据式(1)和(2)计算 $p_{j|i}$ 和 p_{ij}；

(3)初始解 $Y^{(0)} = \{y_1, y_2, \cdots, y_n\}$ 采样与正态分 $N(0, 10^{-4}I)$

(4)for t =1 to T

(a)根据式(3)计算 q_{ij}；

(b)根据式(5)计算梯度；

(c)根据式(6)计算 $Y^{(t)}$；

(5)得到低维数据表示 $Y^{(T)} = \{y_1, y_2, \cdots, y_n\}$。

输出：混合数据集在低维空间的数据表示。

4　实验分析

4.1　实验数据集和评价方法

为了验证 E-t-SNE 算法的有效性，本文选用 UCI 机器学习知识库中的四个真实混合数据集：Credit Approval、Australian Credit Approval、Heart、Adult。数据集 Credit Approval 包含 2 个类，6 个数值属性和 8 个分类属性，共 653 条数据；数据集 Australian Credit Approval 包含 2 个类，6 个数值属性和 9 个分类属性，共 690 条数据；数据集 Heart 包含 2 个类，7 个数值属性和 6 个分类属性，共 270 条数据；数据集 Adult 包含 2 个类，3 个数值属性和 3 个分类属性，共 1100 条数据。表 2 列出了四个数据集数据量，数值属性个数，分类属性个数以及类的个数。

表 2　混合型数据集的详细信息

Dataset	Data Points	Categorical Attributes	Numeric Attributes	Class
Credit Approval	653	8	6	2
Australian Credit Approval	690	9	6	2
Heart	270	6	7	2
Adult	1100	3	3	2

为了评价 E-t-SNE 算法的性能，本文使用了四个真实混合数据集。由于不知道数据在高维空间中的结构，所以无法通过对比高维和低维空间投影映射的方法来直接的评估性能。E-t-SNE 算法本身是一种混合数据降维可视化算法，按照理论，降维以后的数据应该保留了原始高维空间中数据集的分类特征。因此，可以对混合数据集降维之后的数据进行评价，验证此降维可视化算法对后续数据分析的影响。本文采用 k 近邻(kNN)分类算法，将降维以后的数据中 80%用做训练集，20%做测试集，通过对测试数据分类准确度的比较，对本文 E-t-SNE 算法和独热编码方式、余弦距离度量方法进行了评价。

4.2　实验结果分析

为验证 E-t-SNE 算法的有效性，本文采用多种距离计算方法构建混合数据集的距离矩阵，并与本文算法对比。其中独热编码方式是将分类属性转换为 k 个二进制属性构建距离矩阵；余弦距离方法是将混合属性数据转换为向量，通过计算向量间夹角的余弦值构建距离矩阵。原始的 t-SNE 算法不能直接处理混合数据集，但是本文将混合数据集不做处理直接输入 t-SNE 算法，作为对比降维以后分类准确率的基准，以凸显本文算法的有效性。

文献[10]指出，t-SNE 算法的性能对困惑度 Perp 变化较为敏感。根据 Perp 的定义，若 Perp 太小，则低维

嵌入点孤立，看不到低维空间中的聚簇效果；若 perp 太大，则所有低维嵌入点聚集成一个簇，无法辨析数据的真实结构。因此，本文以代价函数 $KL(P\|Q)$ 的值在 0.05-0.35 之间为依据，经过多次实验，对四个混合数据集 Credit Approval、Australian Credit Approval、Heart、Adult 的困惑度 Perp 分别设定为 50，50，20，100。

文献[16]指出，KNN 分类算法中 k 的选择至关重要，为避免 k 值选取不当对实验结果准确新造成的干扰，本文选取多个 k 值作为参数，每个 k 值分别进行 5 次实验取平均值。

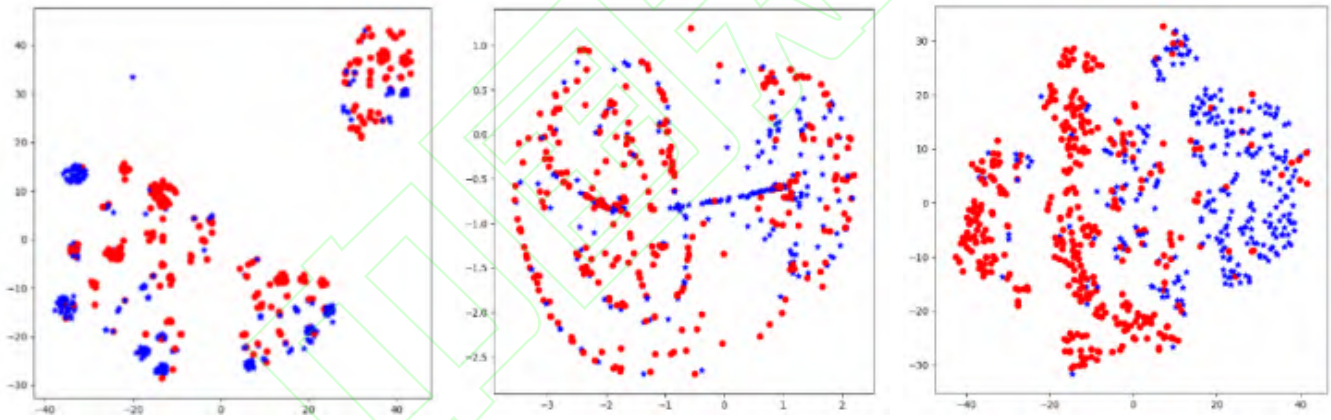
表 3　Credit Approval 数据集实验结果

k 值	t-SNE	余弦距离	One-Hot 编码	E-t-SNE
1	0.6610	0.6987	0.8518	0.8214
5	0.7236	0.7701	0.8151	0.8653
11	0.6915	0.7404	0.7959	0.8931
15	0.6519	0.7786	0.7938	0.8654
平均准确率	0.6820	0.7469	0.8142	**0.8613**

从表 3 可以看出，k 取不同值时，算法 t-SNE、余弦距离、One-Hot 编码、E-t-SNE 在 Credit Approval 数据集上的分类准确率分别是 0.6820、0.7469、0.8142、0.8613。通过比较可以发现，E-t-SNE 比其他三种算法在准确率上分别提高了 17.93%、11.44%、4.71%。因此 E-t-SNE 算法性能更好。

为了更加直观的反应降维以后的数据分布情况，证明 E-t-SNE 算法在可视化方面的优势，本文结合低维空间可视化视图作为一种最直观的评价方法。

图 2 展示了 Credit Approval 数据集在低维空间中的映射视图(类标签只用于染色，不同的颜色表示不同的类)。其中(a)采用独热编码方式计算混合数据集数据点之间的距离降维后得到的视图；(b)采用余弦距离方式计算混合数据集数据点之间的距离降维后得到的视图；图(c)采用 E-t-SNE 算法处理混合数据集降维后得到的视图。通过观察图 2 可以发现，(a)和(b)中两个类的数据点混淆在一起，没有很好的分离，在低维空间中不能明显的发现数据集的类别情况；而(c)则可以直观的发现数据集中存在两个类。



(a)独热编码　　(b) 余弦距离　　(c) E-t-SNE

图 2　Credit Approval 数据集二维空间可视化

从表 4 可以看出，k 取不同值时，算法 t-SNE、余弦距离、One-Hot 编码、E-t-SNE 在 Australian Credit Approval 数据集上的分类准确率分别是 0.7028、0.7175、0.8006、0.8442。通过比较可以发现，E-t-SNE 比其他三种算法在准确率上分别提高了 14.14%、12.67%、4.36%。因此 E-t-SNE 算法性能更好。

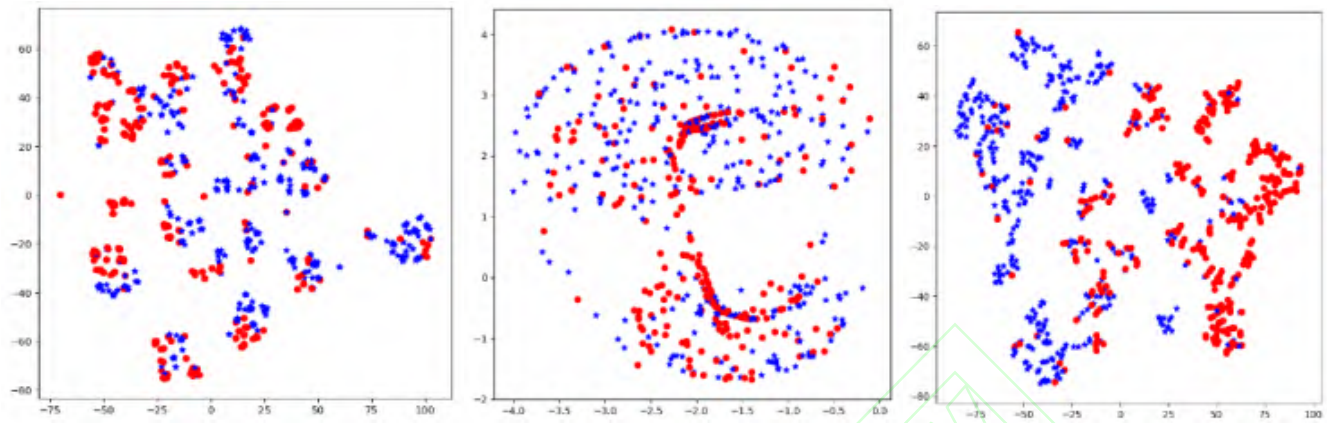
图 3 展示了 Australian Credit Approval 数据集在低维空间中的映射视图。其中(a)采用独热编码方式计算混合数据集数据点之间的距离降维后得到的视图；(b)采用余弦距离方式计算混合数据集数据点之间的距离降维后得到的视图；图(c)采用 E-t-SNE 算法处理混合数据集降维后得到的视图。通过观察图 2 可以发现，(a)中两个类的数据点混淆在一起；(b)中数据点分布松散无明显规律；(c)中两个类别区分明显。

表 4　Australian Credit Approval 数据集实验结果

k 值	t-SNE	余弦距离	One-Hot 编码	E-t-SNE
1	0.7014	0.7175	0.7883	0.8101
5	0.7086	0.7265	0.8202	0.8638
11	0.7231	0.6938	0.7927	0.8462
15	0.6782	0.7324	0.8014	0.8565
平均准确率	0.7028	0.7175	0.8006	**0.8442**

从表 5 可以看出，k 取不同值时，算法 t-SNE、余弦距离、One-Hot 编码、E-t-SNE 在 Heart 数据集上的

分类准确率分别是 0.6786、0.6896、0.7560、0.7958。通过比较可以发现，E-t-SNE 比其他两种算法在准确率上分别提高了 11.72%、10.62%、7.74%。因此 E-t-SNE 算法性能更好。



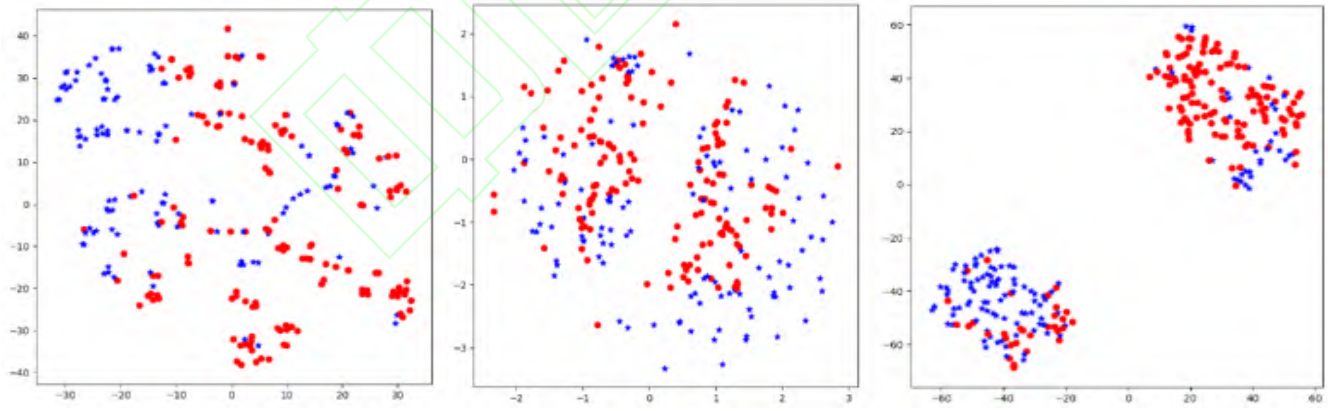
(a)独热编码　　(b) 余弦距离　　(c) E-t-SNE

图 3　Australian Credit Approval 数据集二维空间可视化

表 5　Heart 数据集实验结果

k 值	t-SNE	余弦距离	One-Hot 编码	E-t-SNE
1	0.6370	0.6456	0.6954	0.7814
5	0.7037	0.7283	0.7963	0.8241
11	0.6962	0.7025	0.7812	0.7968
15	0.6777	0.6822	0.7512	0.7806
平均准确率	0.6786	0.6896	0.7560	**0.7958**

图 4 展示了 Heart 数据集在低维空间中的映射视图。其中(a)采用独热编码方式计算混合数据集数据点之间的距离降维后得到的视图；(b)采用余弦距离方式计算混合数据集数据点之间的距离降维后得到的视图；图(c)采用 E-t-SNE 算法处理混合数据集降维后得到的视图。通过观察图 4 可以发现，(a)中两个类的数据点分散，很难发现数据集中的隐藏模式。(b)中可以大体发现有两个类别，但类内数据点没有很好的区分；(c)中两个类之间区分明显。



(a)独热编码　　(b) 余弦距离　　(c) E-t-SNE

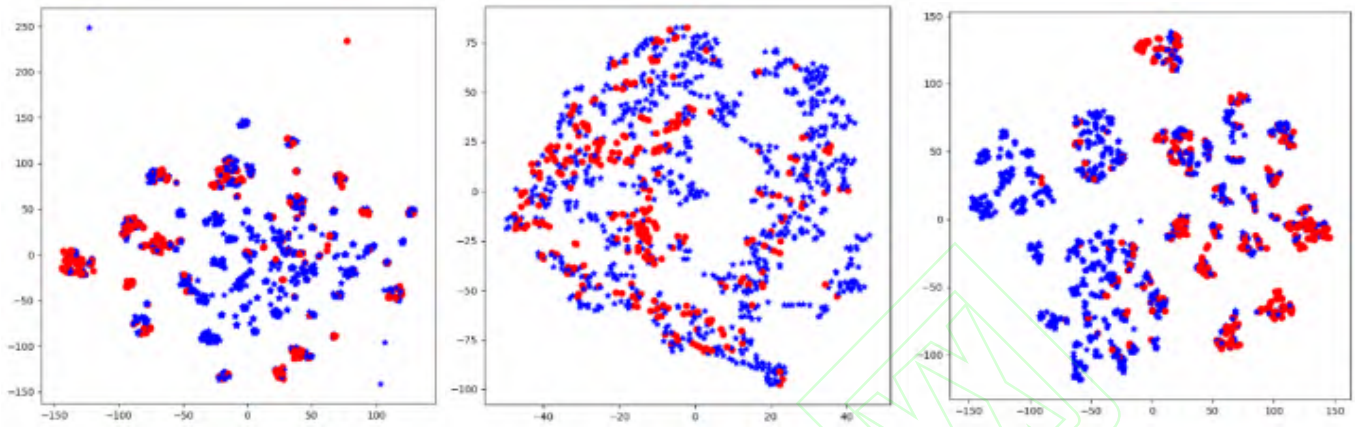
图 4　Heart 数据集二维空间可视化视图

从表 6 可以看出，k 取不同值时，算法 t-SNE、余弦距离、One-Hot 编码、E-t-SNE 在 Adult 数据集上的分类准确率分别是 0.7370、0.7408、0.7534、0.7910。通过比较可以发现，E-t-SNE 比其他两种算法在准确率上分别提高了 5.40%、4.93%、3.76%。因此 E-t-SNE 算法性能更好。

表 6　Adult 数据集实验结果

k 值	t-SNE	余弦距离	One-Hot 编码	E-t-SNE
1	0.7101	0.7242	0.7106	0.7569
5	0.7241	0.7469	0.7536	0.7672
11	0.7593	0.7502	0.7680	0.8159
15	0.7545	0.7421	0.7817	0.8239
平均准确率	0.7370	0.7408	0.7534	**0.7910**

图 5 展示了 Adult 数据集在低维空间中的映射视图。其中(a)采用独热编码方式计算混合数据集数据点之间的距离降维后得到的视图；(b)采用余弦距离方式计算混合数据集数据点之间的距离降维后得到的视图；图(c)采用 E-t-SNE 算法处理混合数据集降维后得到的视图。通过观察图 5 可以发现，(a)(b)两个类的数据点混淆在一起，不能很好的区分类与类之间的关系。(c)中两个类之间虽有混淆重叠，但类间区分明显。



(a)独热编码　　(b)余弦距离　　(c) E-t-SNE

图 5　Adult 数据集二维空间可视化

图 6 汇总了 E-t-SNE 算法和对比算法在四个 UCI 数据集上的分类准确度。由图 6 可以看出，将选取的四个混合属性数据集用不同的距离度量方法降维到二维平面后，用 KNN 做分类处理，E-t-SNE 算法具有较高的分类准确性。

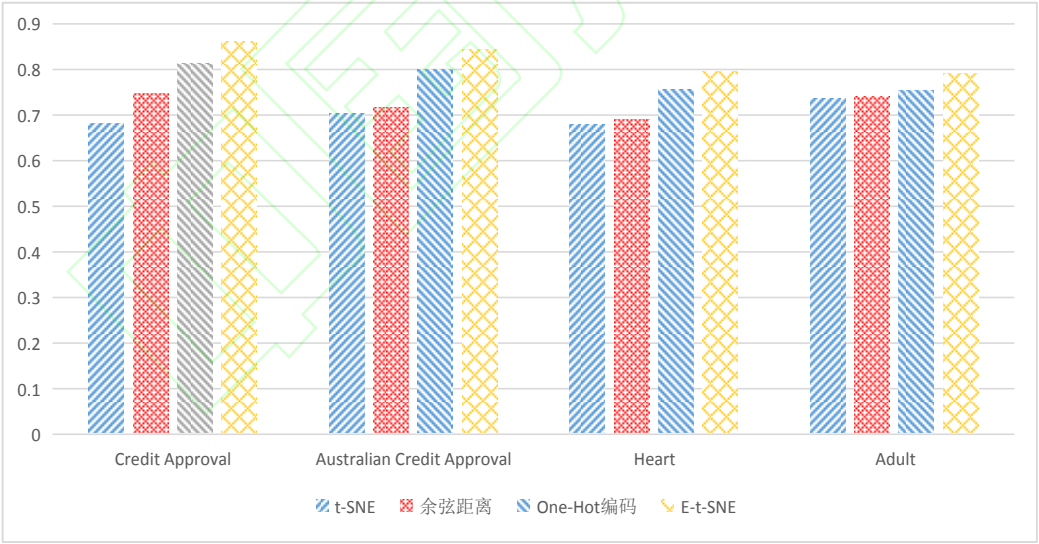


图 6　E-t-SNE 算法与其他算法对比

以上实验证明了 E-t-SNE 算法在对混合数据降维方面有更大的优势。由于引入了信息熵概念，考虑了不同的分类型属性对距离计算时的不同贡献度，使得降维以后的数据更多的保留了数据在高维空间的结构分布，使其在后续的分类算法中获得了更高的准确度。

5　结束语

本文总结了现有的降维可视化算法的原理及特点，指出它们不能有效的处理混合属性数据的缺点。在 t-SNE 算法的基础上提出了一种扩展的 E-t-SNE 算法，使其能处理混合型数据集。该方法引入信息熵的概念度量混合属性数据点之间的距离，相比于传统的将分类型数据转换成 0/1 的独热编码方式和将分类型数据转换成向量的余弦距离计算方法，该方法既能处理混合型数据，又不增加数据维度，并且将不同类标签的数据映射到低维空间，有利于更好地理解和发现高维空间数据的结构。通过实验发现，经过 E-t-SNE 降维后的混合类型数据集，在后续的数据分析算法中表现出比独热编码方

式更高的准确率，证明了 E-t-SNE 算法的有效性。

E-t-SNE 算法能有效的对混合类型数据集进行降维可视化操作。但并没有考虑算法的复杂度和执行时间问题。当数据量太大时，复杂的矩阵运算会占用大量内存和消耗大量的时间。在后续的研究工作中，将考虑算法的复杂度问题，优化求解过程，缩短求解时间。

参考文献：

[1] Mohammed R A, Wong K W, Shiratuddin M F, et al. Machine learning techniques for highly imbalanced credit card fraud detection: a comparative study[C]//Pacific Rim International Conference on Artificial Intelligence. Cham: Springer, 2018: 237-246

[2] 胡彬, 邵叶秦. 基于核学习和距离相似度量的行人再识别[J]. 信息与控制, 2017, 46(5): 525-529.

[3] Cherchi E, Guevara C A. A Monte Carlo experiment to analyze the curse of dimensionality in estimating random coefficients models with a full variance-covariance matrix[J].Transportation Research Part B: Mechodological, 2012, 46(2): 321-332.

[4] Junchin A, Andri M. Supervised, Unsupervised, and Semi Supervised Feature Selection: Review on Gene Selection[J]. Transactions on Computational Biology and Bioinformatics, 2016, 13(5) : 971-989.

[5] M.Danubianu, S.G. Pentiuc, Data dimensionality reduction for data mining: a combined filter-wrapper framework[J]. International Journal of Computers Communications & Control, 2014,9(3): 576-580.

[6] Turk M . Eigenfaces for recognition[J]. Journal of Cognitive Neuroscience, 1991, 3(1): 71-86.

[7] S.T. Roweis, L.K Saul, Nonlinear dimensionality reduction by Locally Linear Embedding[J]. Science, 2000, 290(5500): 2323-2326.

[8] Tenenbaum J B, Silva V, Langford J C. A global geometric framework for nonlinear dimensionality reduction[J]. Science, 2000, 290(5500): 2319-2323.

[9] Zhang P , Qiao H , Zhang B . An improved local tangent space alignment method for manifold learning[J]. Pattern Recognition Letters, 2011, 32(2): 181-189.

[10] Van Der Maaten L, Hinton G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9(11): 2579-2605.

[11] Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation[J]. Neural Computation, 2003, 15(6): 1373-1396.

[12] Jamieson A R , Giger M L , Drukker K , et al. Exploring nonlinear feature space dimension reduction and data representation in breast CADx with Laplacian eigenmaps and t-SNE[J]. Medical Physics, 2010, 37(1): 339-351.

[13] Garces E , Agarwala A , Gutierrez D , et al. A similarity measure for illustration style[J]. ACM Transactions on Graphics, 2014, 33(4): 1-9.

[14] Liu W , Wang C , Wang B , et al. Application of improved locally linear embedding algorithm in dimensionality reduction of cancer gene expression data[J]. Journal of Biomedical Engineering, 2014, 31(1): 85-90.

[15] Hsu C C, Huang W H. Integrated Dimensionality Reduction Technique for Mixed-Type Data Involving Categorical Values[J]. Applied Soft Computing, 2016, 43: 199-209.

[16] Yang Y, Liu X. A re-examination of text categorization methods[C]//International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 1999: 42-49.

[17] Hinton G E, Roweis S T. Stochastic neighbor embedding[C]// Proceedings of Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2002: 833-840.

[18] 董迎朝, 王彬, 马洒洒, 等. 基于 t-SNE 的脑网络状态观测矩阵降维方法研究[J]. 计算机工程与应用, 2018(1): 42-47.

[19] 姜智涵, 朱军, 周晓锋等. 一种基于信息熵的混合属性数据谱聚类算法[J]. 计算机应用研究, 2018, 36(8).

[20] Ding Shifei, Du Mingjing, Sun Tongfeng, et al. An entropy-based density peaks clustering algorithm for mixed type data employing fuzzy neighborhood[J]. Knowledge-Based Systems, 2017 (133): 294-313.

[21] 徐森, 花小朋, 徐静,等. 一种基于 T-分布随机近邻嵌入的聚类集成方法[J]. 电子与信息学报, 2018, 40(6): 50-56.