Journal of System Simulation

Volume 31 | Issue 9 Article 4

12-12-2019

High-dimensional data hiding pattern mining based on topology data analysis

Bolong Liu
1. School of Electrical Engineering, Xinjiang University, Urumqi 830047, China; ;

Zhe Li
2. Network and Information Technology Center, Xinjiang University, Urumqi 830046, China;

High-dimensional data hiding pattern mining based on topology data analysis

Abstract

Abstract: Aiming at the limitation of traditional data analysis methods to find hidden patterns between high-dimensional complex data, *a method of high-dimensional data hiding pattern mining based on topological data analysis is proposed. By extracting the characteristics of complex high-dimensional data, the relationship between its shapes and samples is analyzed.* To get the dataset hidden mode, the topological data analysis is used to verify the gender recognition of high-dimensional dataset-voice.. At the same time, the relationship between the dataset data subgroups and related data subgroups is visually analyzed. The results show that the implicit relationship and pattern between data subgroups can be found by the proposed method, which cannot be found by traditional methods and it is more detailed and effective than traditional methods. The results also verify the power and effectiveness of the proposed method for high-dimensional data hiding mode mining.

Keywords

topological data analysis, hidden pattern mining, high dimensional data

Recommended Citation

Liu Bolong, Li Zhe. High-dimensional data hiding pattern mining based on topology data analysis[J]. Journal of System Simulation, 2019, 31(9): 1755-1762.

第 31 卷第 9 期 系统仿真学报© Vol. 31 No. 9
2019 年 9 月 Journal of System Simulation Sep., 2019

基于拓扑数据分析的高维数据隐藏模式挖掘

刘擘龙[1]，李喆[2]
（1. 新疆大学电气工程学院，新疆 乌鲁木齐 830047；2. 新疆大学网络与信息技术中心，新疆 乌鲁木齐 830046）

摘要：针对传统数据分析方法寻找高维复杂数据间隐藏模式存在局限性的问题，*提出基于拓扑数据分析的高维数据隐藏模式挖掘方法，通过提取复杂高维数据的特征，分析其形状和样本的相互关系来获得数据集隐藏模式*。利用拓扑数据分析对高维数据集-声音的性别识别进行实例验证，同时对数据集数据子组以及相关数据子组之间关系进行可视化分析，结果表明所提方法可发现传统方法无法发现的数据子组之间隐含的关系和模式，得到了比传统方法更精细有效的结果，验证了所提方法对高维数据隐藏模式挖掘的强大性和有效性。

关键词：拓扑数据分析；隐藏模式挖掘；高维数据
中图分类号：TP391.9 文献标识码：A 文章编号：1004-731X (2019) 09-1755-08
DOI: 10.16182/j.issn1004731x.joss.19-0401

High-dimensional data hiding pattern mining based on topology data analysis

Liu Bolong[1], *Li Zhe*[2]
(1. School of Electrical Engineering, Xinjiang University, Urumqi 830047, China;
2. Network and Information Technology Center, Xinjiang University, Urumqi 830046, China)

Abstract: Aiming at the limitation of traditional data analysis methods to find hidden patterns between high-dimensional complex data, *a method of high-dimensional data hiding pattern mining based on topological data analysis is proposed. By extracting the characteristics of complex high-dimensional data, the relationship between its shapes and samples is analyzed.* To get the dataset hidden mode, the topological data analysis is used to verify the gender recognition of high-dimensional dataset-voice.. At the same time, the relationship between the dataset data subgroups and related data subgroups is visually analyzed. The results show that the implicit relationship and pattern between data subgroups can be found by the proposed method, which cannot be found by traditional methods and it is more detailed and effective than traditional methods. The results also verify the power and effectiveness of the proposed method for high-dimensional data hiding mode mining.

Keywords: topological data analysis; hidden pattern mining; high dimensional data

引言

随着科学技术的发展，大规模数据不断产生，其结构的庞大性和复杂性使得从数据中提取有用的信息成为巨大挑战，因此如何对大规模数据进行有效的挖掘越来越具有挑战性[1]。目前使用的数据分析方法主要依赖于专家经验和假设，然而对于复杂庞大的数据集，利用专家经验做出有用假设变得非常困难，而通过数据挖掘技术可以从大量无序的数据中发现隐含的、有效的、有价值的、可理解的模式，进而发现有用的知识并得出相关的趋势和关

收稿日期：2019-07-31 修回日期：2019-08-02；
基金项目：国家自然科学基金(51767022, 51575469)；
作者简介：刘擘龙(1993-)，男，甘肃，硕士生，研究方向为深度学习、数据挖掘；李喆(通讯作者 1977-)，女，江苏，硕士，副教授，研究方向为大数据分析。

联关系，提供对数据洞察和理解能力。数据挖掘的主要任务包括分类、聚类、回归、关联、序列和偏差 6 种模式的识别[2]，使用数据挖掘方法对数据中的模型和关系进行挖掘时，较为关键的步骤就是降维和聚类。由于数据量增大，数据需要提取的特征越来越多，这就导致数据的维度越来越大，由于高维数据中包含大量的冗余信息以及高维空间中隐藏了数据间的相关性，所以通常对高维数据分析采用先降维后聚类的方法。线性降维方法主要有主成分分析(Principal Components Analysis，PCA)、奇异值分解(Singular Value Decomposition，SVD)、线性判别分析(Linear Discriminant Analysis，LDA)等[3]在数据挖掘中受到广泛的应用。但是这几个方法的局限是无法处理非线性特征的数据，处理非线性特征的数据的降维算法主要是流形学习，其中常见的算法有等距离映射(Isometric Feature Mapping，Isomap)、局部线性嵌入(Locally Linear Embedding，LLE)、t 分布随机近邻嵌入(t-distributed Stochastic Neighbor Embedding，t-SNE)等。聚类算法是一种重要的数据挖掘算法，应用比较广泛的主要有 K 均值聚类算法(K-means)、相似性传播聚类算法(Affinity Propagation，AP)均值漂移聚类算法(Mean-Shitf)和基于密度的含噪声空间聚类(Density-based Spatial Clustering of Applications with Noise，DBSCAN)算法[4]。以上算法在处理小规模数据集时有着比较好的性能和表现，但由于维数灾难，传统的聚类方法对高维数据进行聚类很难得到有意义的聚类结果，因此，对于高维数据聚类以及隐藏模式挖掘等问题的研究寻找新的解决方法是一种解决思路。通过阅读和研究大量文献，发现一种新的高维复杂数据挖掘方法-拓扑数据分析(Topological Data Analysis，TDA)[5]。

TDA 是一种解决高维数据全貌的方法，用来识别和挖掘数据的形状以及数据之间关系和模式。相比于主成分分析、聚类分析这些常用的方法，TDA 不仅可以有效地捕捉高维数据空间的拓扑信息，而且擅长发现一些用传统方法无法发现的小分类，最重要的是它不依赖于专家经验[6]。目前，利用 TDA 做数据挖掘和可视化分析正处在起步阶段，少数研究者将拓扑数据分析应用于医学和工业加工领域中。Lum 等[6]提出使用拓扑数据分析高维复杂数据集，通过提取其“数据形状”来研究复杂高维乳腺肿瘤的基因表达数据、找到了比标准方法更精确的子组和模式，为医学研究提供有利工具。Khasawneh 等[7]将机器学习与拓扑数据分析相结合，应用于加工过程中的颤振识别和检测，来获得可以检测到颤振的过程的描述，他们的结果显示该模型分类精度高达 97%。上述研究表明，使用拓扑数据分析方法在高维数据集隐藏模式的发现中具有巨大应用价值和潜力，本文将 TDA 应用于声音的性别识别数据集，计算和识别出高维数据的形状和数据样本之间的关系，并提出理解“数据形状”的重要性，来提取有意义的见解和重要相关性。

对数据集使用拓扑数据分析做数据挖掘和可视化，在拓扑数据分析得到的结果中使用可视化方法辅助数据挖掘过程[8]。本文验证了 TDA 可以处理高维数据并且能够挖掘数据中的隐藏模式，通过使用可视化方法辅助数据挖掘可以带来对于高维数据集新颖的洞察。

由于 TDA 强大的关系和模式挖掘以及良好的可视化性能，使得 TDA 从很大程度上改变了分析数据的方式，将拓扑分析与数据分析相结合，是一个很前沿和值得探索的技术，相信未来会有更多基于 TDA 的相关算法和应用被提出，并能够成功应用到数据挖掘领域。

1 拓扑数据分析

1.1 拓扑数据分析简介

拓扑学研究的是一些特殊的几何性质，这些在图形连续改变形状后还能继续保持不变的性质，称为“拓扑性质”，而在复杂的高维数据内部也存在着类似的拓扑性质，被形象地称之为“数据形状”。TDA 可以描述为在数据中找到结构的数据分析方法的集合，包括：聚类、流形估计、非线性降维、模式估计、岭估计和持久同调[9]。和通常研究的成

DOI: 10.16182/j.issn1004731x.joss.19-0401

对关系相比，这种相互关系的形状之中可能潜藏了巨大的研究价值，要理解数据的形状，就必须使用于拓扑学方法，TDA 所做的就是抽取这种形状并进行分析，在二维空间中，TDA 关心的是数据全貌的连接度和点(包含多个极度相似的样本)的数量；在立体的维度，TDA 关心的是数据全貌是空心球体的数量和它们之间的连接度；在更高的维度，通过代数工具的推算，也可以很严格的定义高维的数据全貌在拓扑空间上的形状，从而自然的量化这些比较恒定的形状，然而一切还得从如何勾勒数据的全貌开始。

1.2 数据形状刻画方法

形状是一个有点模糊的概念，缺少数学上精确和定量描述，在纯数学学科内拓扑和微分几何的研究成功的解决这个问题，它们提供明确的方法，在精确的意义上量化和描述数据的形状。此外，它们提供了离散化和压缩存在于几何对象中的信息的方法来提供有用的表示[10]。

TDA 使用拓扑技术分析数据集的方法，从通常具有挑战性的高维，不完整和高噪声的数据集中提取信息，提供了一种以特定度量不敏感的方式对所选择数据进行刻画通用框架，并提供降低维度和降低噪声的性能。TDA 提供强有力的方法来推断有关数据结构的稳健的定性信息，有时是定量信息，通过有根据的数学，统计和算法方法，以推断、分析和利用高维数据的复杂拓扑和几何结构，这些结构通常表示为欧几里德或更一般度量空间中的点云[11]。TDA 结合了代数拓扑和纯数学的其他工具，可以对“形状”进行数学上严谨的研究，TDA 的独特功能使其在数据分析和挖掘的研究领域具有广阔的潜力，可以广泛地探索和理解复杂的高维数据空间，主要方法包括持久同源性和 MAPPER 算法，MAPPER 帮助数据分析人员总结并可视化复杂的数据集，提供对数据直观的洞察[12]。

1.3 拓扑数据分析的优势和特点

拓扑数据分析的优势和特点可以总结为：

1）TDA 可以简明地解释数据中隐藏的形状。

2）捕捉高维数据空间的拓扑信息。

3）理解数据的形状，发现数据的结构。

4）数据降维、聚类、数据压缩。

5）分析总结数据及可视数据。

6）在同一个图表上同时分析千兆字节的数据和可视化数十万个参数乃至数百万行参数。

2 拓扑数据分析原理及算法

2.1 拓扑数据分析原理

TDA 的基本步骤如下：

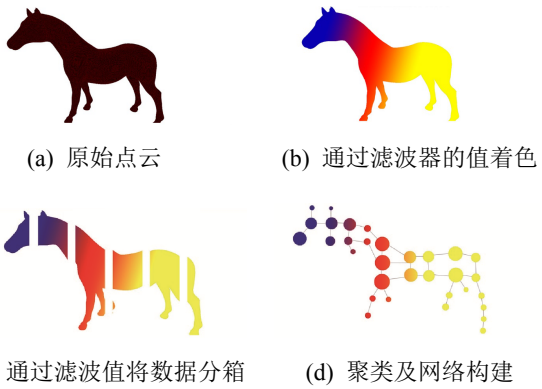
步骤 1：用一个滤波函数对每个数据点计算一个滤波值。

步骤 2：数据点按照其滤波值，从小到大被分到不同的滤波值区间里。相邻的滤波值区间设置有一定的重叠区域，重叠区域的点同时属于 2 个区间。

步骤 3：对每个区间里的数据做聚类分析。

步骤 4：各区间聚类的得到的小类放在一起，每一个小类用一个大小不同的圆表示。若 2 个类之间存在相同的原始数据点(这就是区间需要相互重叠的原因)，则在它们之间加上一条边。

步骤 5：对上述圆和边组成的图形施加一层力学布局，让其达到平衡，就得到最终的“数据图形”，TDA 工作原理的示意图如图 1 所示。



(a) 原始点云 (b) 通过滤波器的值着色

(c) 通过滤波值将数据分箱 (d) 聚类及网络构建

图 1 TDA 工作原理示意图

Fig. 1 Schematic diagram of TDA principle

图 1 中，A 图表示点云的 3D 对象(马)，如 B 图表示将过滤器值应用于点云对象后由过滤器函

数的值得到着色结果，C 图是通过着色值将数据集分成重叠组的结果，D 图中每个分割后的数据都为一个簇，通过簇与簇的相似程度来构建网络，通过以上示例可将 TDA 的工作流程图总结如图 2 所示。

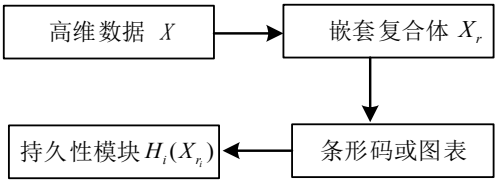


图 2 TDA 基本工作流程
Fig. 2 Schematic diagram of TDA principle

其中 X 是代表高维数据，X_r 代表的是嵌套的单纯复合体，该过程将高维数据转换为单纯复合物的过程，在该过程中考虑每种复合物的同源性给出了持久性 $H_i(X_{r_0}) \to H_i(X_{r_1}) \to H_i(X_{r_2}) \to ...$ 持久性模块通过可视化的方法转换为数据的条形码或者拓扑分析图。

2.2 MAPPER 算法

高维数据无法直接可视化，虽然科学家已经发明了许多方法来从数据集中提取低维结构，例如主成分分析、局部线性嵌入和多维缩放等方法[13]。然而这些方法在处理高维数据时并没有良好的表现，因为在同一数据集中可以找到许多不同的拓扑特征，因此 TDA 对高维空间可视化的研究至关重要，Singh 等[14]，提出了一种称为 MAPPER 的通用方法，不仅可以展示集群、变量之间的关系而且可以获得高维数据结构的更高水平理解，以下是 MAPPER 算法的原理、实现步骤和算法。

2.2.1 MAPPER 算法原理

MAPPER 算法是通过一个映射 $f: X \to G$ 将数据集的拓扑结构总结到一张图上，是一种从数据构建图形的一种方式，它揭示了高维数据空间的拓扑特征，能够估计数据底层空间的重要连接方式在可视化拓扑图中进行数据挖掘和探索，适用于任何简单或复杂数据而且非常灵活。MAPPER 可以对数据进行可视化或聚类，是一种生成数据可视化表示的无监督方法，通常可以揭示其他方法无法获得的数据的新见解和新发现，数据分析人员就可以使用它来探索复杂高维数据集的结构和拓扑性质。

2.2.2 MAPPER 算法实现步骤

在实际的应用中主要使用的是分布式 MAPPER 算法，为了保证分布式 MAPPER 的输出与顺序 MAPPER 相同(顺序 MAPPER 算法的细节将在算法 1 中给出)需要在覆盖上进行一些处理，以获得最终的 MAPPER 输出。在分布式 MAPPER 中，考虑 N 链覆盖 $A_1, A_2, ..., A_N$ 间隔的一个[a, b]以及它们的覆盖 $u_1, u_2, ...u_N$。覆盖预处理的细节在算法 2 中描述。在完成覆盖的预处理并获得集合 $\{A_i, U_i\}_{i=1}^N$ 之后，然后将每对 (A_i, U_i) 映射到特定处理器 P_i，该处理器 P_i 执行一些计算以产生子图 G_i，最后将子图合并为一个图 G。

算法 1：顺序式 MAPPER

输入：数据集 X，且数据点之间具有度量概念;
标量函数： $f: X \to R^n$
一个有限 $f(X)$ 的覆盖 $u = \{U_1, U_2, ...U_k\}$

输出：代表 $N_1(f^*(u))$ 的图

step 1：对于每个集合 $X_i := f^{-1}(U_i)$，它的簇 $X_{i,j} \subset X_i$ 是通过使用聚类算法计算的。

step 2：每个聚类都被视为 MAPPER 图中的顶点。此外，只要 $X_{ij} \cap X_{kl} \neq \phi$，就在两个节点 X_{ij} 和 X_{kl} 之间插入一条边。

算法 2：覆盖预处理

输入：点云数据 X;
标量函数： $f: X \to$ [a, b];
N 个处理器(P)集合：

输出：对 $\{A_i, U_i\}_{i=1}^N$ 的集合，其中 $\{A_i\}_{i=1}^N$ 是[a, b] N 链覆盖，U_i 是 A_i 的覆盖。

step 1：构建一个[a, b]的 N 链覆盖，[a, b]由 N 个开区间 $A_1, A_2, ..., A_N$ 所覆盖，当 $|i-j|=1$ 和空集时，$A_{i,j} := A_i \cap A_j \neq \phi$。

step 2：对于每个开放的集合 A_i 构造一个开放的覆盖 U_i，$\{U_i\}_{i=1}^N$ 覆盖满足以下条件：

(1) $A_{i,i+1}$ 是覆盖 U_i 和 U_{i+1} 的开集，即 $U_i \cap U_{i+1} = \{A_{i,i+1}\}$；

(2) 如果$U_i \in u_i$ 和$U_{i+1} \in u_{i+1}$ 使得$U_i \cap U_{i+1} \neq \phi$；那么对每个 i=1, 2, …, N-1；有 $U_i \cap U_{i+1} = A_{i,i+1}$。

算法 3：分布式 MAPPER

输入：点云数据 X：标量函数：$f: X \to [a, b]$；

N 个处理器(P)集合：成对的集合$\{A_i, U_i\}_{i=1}^{N}$从覆盖预处理算法中获得。

输出：分布式 MAPPER 图

step 1：for ($i \leftarrow 1$ to $i = N$)do。

step 2：$P_i \leftarrow (A_i, u_i)$ //映射每个 A_i 和它的覆盖 u_i 到处理器 P_i。

step 3：确定点的集合 $X_i \subset X$ 通过 f 它被映射到 A_i 并且在覆盖$(f | x_i) * (u_i)$(i=1, 2, …, N)上同时运行顺序 MAPPER 构造，获得 N 个图 $G_1, G_2, ..., G_N$，若 N=1，则返回图 G_1。

step 4：设 $C_{j_1}^i, C_{j_2}^i, ..., C_{j_i}^i$ 是从 $f^{-1}(A_{i,i+1})$ 获得的簇。通过选择覆盖 u_i 和 u_{i+1}，这些簇由 G_i 和 G_{i+1}(每个 v_k^i 对应簇 C_k^i)中的顶点 $v_{j_1}^i, v_{j_2}^i, ...v_{j_i}^i$ 表示。

step 5：以如下方式合并图 $G_1, G_2, ...G_N$，通过构建 $A_{i,i+1}$，u_i 和 u_{i+1}，每一个 $f*(u_i)$ 和 $f*(u_{i+1})$ 公用一个在 $f*(A_{i,i+1})$ 中的簇 $C_{j_k}^i$，因此 $C_{j_k}^i$ 是由在图 G_i 和 G_{i+1} 中的一个向量表示，合并是通过考虑不相交的联合图 $G_1 \cup G_2 \cup ... \cup G_N$ 然后取这个图的商来确定 G_i 和 G_{i+1} $(1 \leqslant i \leqslant N-1)$ 中的相应顶点。

3 实验及结果分析

3.1 数据集描述

使用声音的性别识别数据集进行试验验证，基于声音和语音的声学特性，该数据集被创建用来识别男性或女性的语音。实验所用数据集包括 3 168 个记录的语音样本，这些样本来自男性和女性发言者。使用 Seewave 和 TuneR 软件包在 R 中通过声学分析对语音样本进行预处理，分析频率范围为 0~280 Hz (人类声音范围)测量每种声音的以下声学特性并包含在 CSV 中：基本属性如表 1 所示，共有 3 168×21=66 528 个数据。

表 1 语音样本预处理后的声学特性

Tab. 1 Acoustic characteristics of speech samples after preprocessing

声学特性	说明
meanfreq	平均频率/kHz
sd	频率的标准偏差
median	中位频率/kHz
Q25	第一个分位数/kHz
Q75	第三个分位数/kHz
IQR	间隔范围/kHz
skew	偏斜
spent	谱熵
sfm	光谱平坦度
mode	模式频率
centroid	频率质心
peakf	峰值频率(能量最高的频率)
meanfun	声学信号测得的基频平均值
minfun	声学信号上测量的最小基频
maxfun	声学信号上测量的最大基频
meandom	在声学信号上测量的主导频率的平均值
mindom	声学信号上测量的最小主频率
maxdom	在声学信号上测量的主导频率的最大值
dfrange	在声学信号上测量的主导频率范围
modindx	调制指数
label	男性或女性

3.2 实验结果比较分析

在参数 nr_cubes=15 和 nr_cubes=10 下利用拓扑数据分析对数据集拓扑结构进行聚类和可视化分析，其中参数 nr_cubes 表示生成簇的多少，越大生成的簇越多，越小生成的簇越少。实验结果如图 3~4 所示。其中红色圈表示女性，黄色圈表示男性，蓝色圈表示男性和女性都有。

比较 TDA 的降维效果与传统的降维算法：主成分分析(PCA)，局部线性嵌入(LLE)，多维缩放(MDS)，随机树嵌入(Random Tree embedding)，等距离映射(ISOMAPE)，t 分布随机近邻嵌入(t-SNE)的降维效果。得到结果如图 5~7 所示。

比较 TDA 的聚类效果与比较常用的聚类算法：相似性传播聚类算法(Affinity Propagation)，K 均值聚类算法 (K-means)，均值漂移聚类算法(Mean-shitf)、基于密度的含噪声空间聚类(DBSCAN)的效果，得到实验结果如图 8~9 所示。

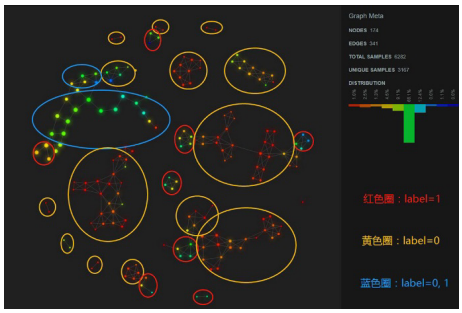


图 3 拓扑数据分析可视化结果
(参数：nr_cubes=15，overlap_perc=0.9)
Fig. 3 Topology data analysis visualization results
(Parameters: nr_cubes=15, overlap_perc=0.9)

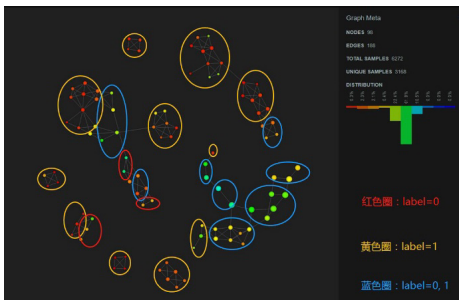


图 4 拓扑数据分析可视化结果
(参数：nr_cubes=10，overlap_perc=0.9)
Fig. 4 Topology data analysis visualization results
(Parameters: nr_cubes=10, overlap_perc=0.9)

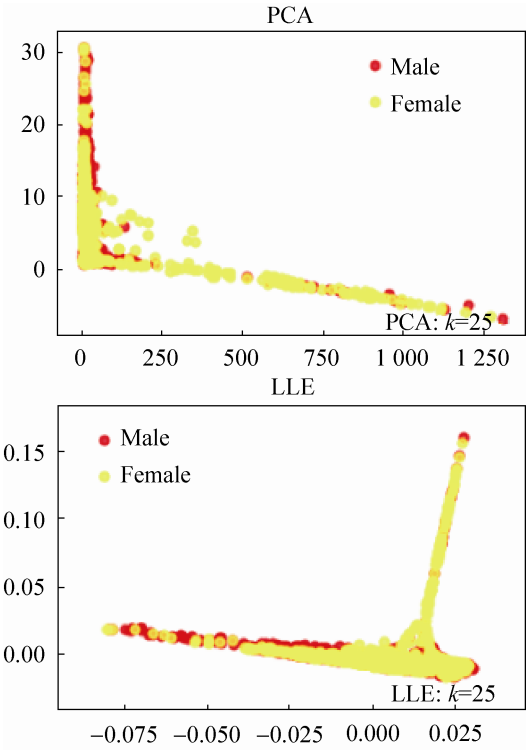


图 5 PCA，LLE 降维算法的结果
Fig. 5 Result of the PCA and LLE dimensionality reduction algorithm

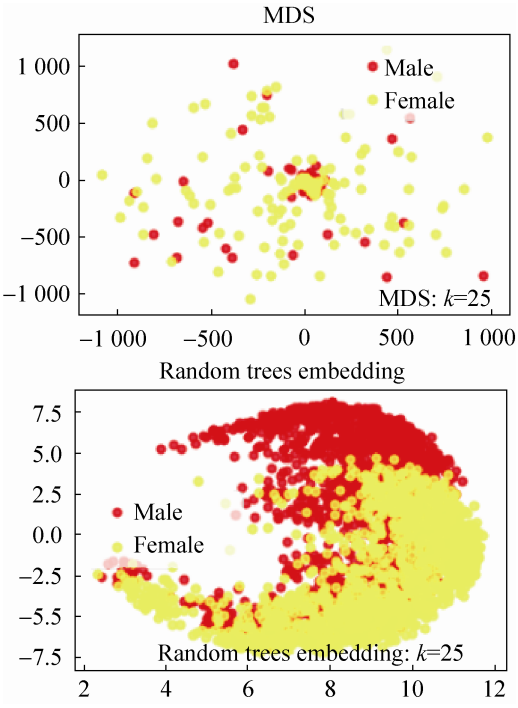


图 6 MDS，Random trees embeding 降维算法的结果
Fig. 6 Result of the MDS and Random trees embeding dimensionality reduction algorithm

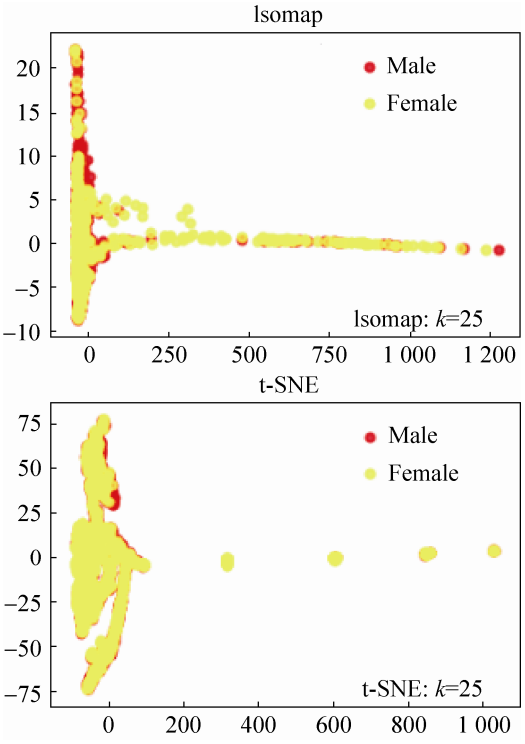


图 7 ISOMAPE，t-SNE 降维算法的结果
Fig. 7 Result of the ISOMAPE and t-SNE dimensionality reduction algorithm

https://dc-china-simulation.researchcommons.org/journal/vol31/iss9/4
DOI: 10.16182/j.issn1004731x.joss.19-0401

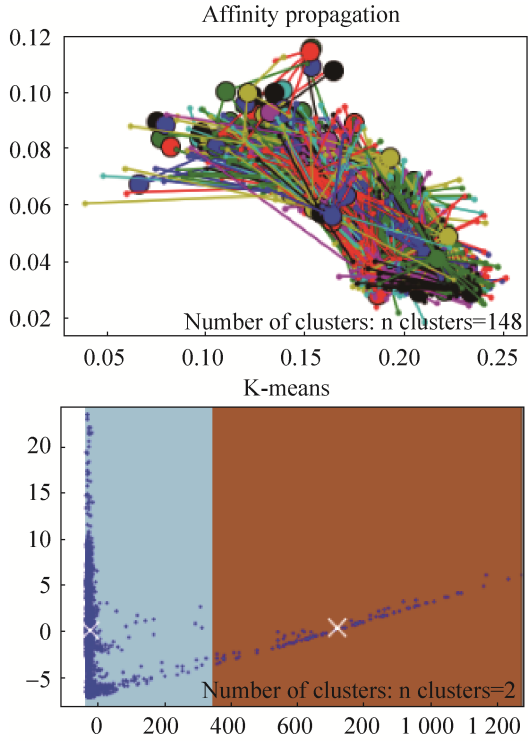


图 8　Affinity Propagation，K-means 聚类算法的结果
Fig. 8　Result of the Affinity Propagation and K-means clustering algorithm

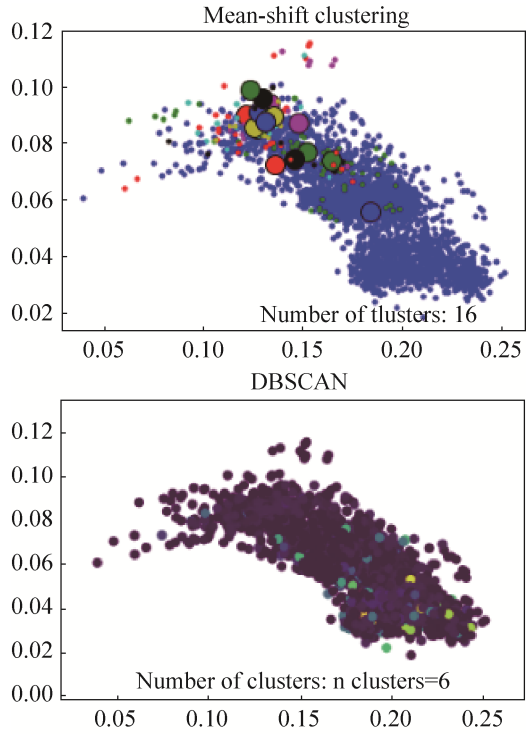


图 9　Mean-shitf，DBSCAN 聚类算法的结果
Fig. 9　Result of the Mean-shitf and DBSCAN clustering algorithm

TDA 研究的是与坐标无关的数据拓扑形状，完全不受坐标的限制，这意味着拓扑形状的构建依赖于距离函数的定义，或者说相似度概念的定义。所以在 TDA 生成的结果图中：点越大该集合所含的样本点越多，有边连接的部分表明两个集合相似度比较高，不同的颜色代表原始高维空间与低维映射空间之间的不同映射关系。

TDA 使用有限的点和边来表示大量的数据，并且保留了数据重要的特征，因此可以容忍数据小范围的变形与失真，有助于发现高维数据的形状和隐藏模式。图 3 和图 4 清楚地向我们展现了数据中隐藏的形状，以及从不同尺度观察数据的结果。由图 3 可知，男性和女性样本被有效区分，在同性中存在不同的簇，说明同性之间的声音也有不同的类别，而在异性间存在共同的簇，说明即使异性之间的声音也会有相似特征和成分。由图 4 可以明显看到簇的减少，以及更加明显的聚类和数据形状，虽然整个数据的形状有点扭曲变形，但是 Y 形耀斑这样的基本特征仍然存在，说明 TDA 的确对小误差的容忍度很大，数据簇更加明确。

通过调节 nr_cubes 这样的参数，可以以不同的尺度来观察高维数据的特征和数据样本之间的关系，从可视化图中分析数据的模式和依赖关系。拓扑数据分析(TDA)像使用放大镜和显微镜一样来在不同尺度上观察和探索高维数据，从而更好的发现高维数据样本之间的关系和模式，有助于发现使用传统方法无法发现的数据结构、模式、关系和分类。

TDA 具有直接在高维空间中对数据进行操作的能力使其具有良好的降维能力，由图 5~7 可以看出传统降维算法面对高维数据已经无法实现有效降维，而 TDA 降维效果和可视化效果非常理想。由图 8~9 可以看出传统聚类算法的聚类效果也没有 TDA 好。这表明 TDA 不仅具有高维数据隐藏模式挖掘能力而且具有良好降维和聚类能力。TDA 的这种精细数据分析能力，使其数据科学家进行数据探索和挖掘的优秀方法和框架。

4 结论

本文提出了基于拓扑数据分析的高维数据隐藏模式挖掘方法，通过对 Gender Recognition by Voice 高维数据集的聚类和可视化分析研究表明，TDA 方法可以将相似度高的集合聚为一簇，而相似度较低的集合被有效分割，通过可视化所展现的拓扑结构有效解释数据中隐藏的拓扑形状，从而得到传统方法无法发现的数据隐藏关系。

通过实验分析得出以下结论：

(1) 本文提出使用拓扑数据分析来挖掘高维数据，实现了高维数据空间隐藏模式的挖掘和发现，提高了数据挖掘隐藏模式发现的能力，增强了对高维数据的理解。

(2) 通过可视化呈现加深对数据的理解和分析，辅助数据挖掘得到更好的效果，与传统数据挖掘方法相比，本文所提方法能够有效过滤噪声、降维和聚类，同时直观呈现模式挖掘的结果。

参考文献:

[1] 海沫. 大数据聚类算法综述[J]. 计算机科学, 2016, 43(增 1): 380-383.
Hai Mo. Overview of Big Data Clustering Algorithms[J]. Computer Science, 2016, 43(S1): 380-383.

[2] 王光宏, 蒋平. 数据挖掘综述[J]. 同济大学学报(自然科学版), 2004(2): 246-252.
Wang Guanghong, Jiang Ping. A Survey of Data Mining[J]. Journal of Tongji University (Natural Science), 2004(2): 246-252.

[3] 孙喜利. 高维数据的降维及聚类方法研究[D]. 兰州: 兰州大学, 2016.
Sun Xili. Research on Dimensionality Reduction and Clustering Method of High Dimensional Data[D]. Lanzhou: Lanzhou University, 2016.

[4] Treshansky A, Mcgraw R M. Overview of clustering algorithms[J]. Proceedings of SPIE - The International Society for Optical Engineering, 2001, 4367: 41-51.

[5] Medina P S, Doerge R W. Statistical Methods in Topological Data Analysis for Complex, High-Dimensional Data[J]. arXiv: 1607.05150 [stat] (S2331-8422), 2016. https://arxiv.org/abs/1607.05150.

[6] Lum P Y, Singh G, Lehman A, et al. Extracting insights from the shape of complex data using topology[J]. Scientific Reports (S2045-2322), 2013, 3(1): 1236

[7] Khasawneh F A, Munch E, Perea J A. Chatter Classification in Turning Using Machine Learning and Topological Data Analysis[J]. IFAC-Papers On Line (S2405-8963), 2018, 51(14): 195-200.

[8] 马昱欣, 曹震东, 陈为. 可视化驱动的交互式数据挖掘方法综述[J]. 计算机辅助设计与图形学学报, 2016, 28(1): 1-8.
Ma Yuxin, Cao Zhendong, Chen Wei. A Review of Visualized Driven Interactive Data Mining Methods[J]. Journal of Computer-Aided Design & Computer Graphics, 2016, 28(1): 1-8.

[9] Carlsson G. Topology and data[J]. Bulletin of the American Mathematical Society (S0273-0979), 2009, 46(2): 255-308.

[10] Liu S, Maljovec D, Wang B, et al. Visualizing High-Dimensional Data: Advances in the Past Decade[J]. IEEE Transactions on Visualization and Computer Graphics (S1077-2626), 2017, 23(3): 1249-1268.

[11] Chazal F, Michel B. An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists[J]. arXiv: 1710.04019 [cs, math, stat] (S2331-8422), 2017. https://arxiv.org/abs/1710.04019.

[12] Wasserman L. Topological Data Analysis[J]. arXiv:1609.08227 [stat] (S2331-8422), 2016. https://doi.org/10.1038/s41566-018-0148-x.

[13] Dey T K, Memoli f, Wang Y. Mutiscale MAPPER: A Framework for Topological Summarization of Data and Maps[J]. arXiv: 1504.03763 [cs, math] (S2331-8422), 2015. https://arxiv.org/abs/1504.03763.

[14] Singh G, Mémoli F, Carlsson G E. Topological methods for the analysis of high dimensional data sets and 3d object recognition[C]// SPBG. 2007: 91-100. https://www.ayasdi.com/wp-content/uploads/2015/02/Topological_Methods_for_the_Analysis_of_High_Dimensional_Data_Sets_and_3D_Object_Recognition.pdf.