基于多维行为分析的用户聚类方法研究

张林兵[1]，郭 强[1]，吴行斌[1]，梁耀洲[1]，刘建国[2*]

(1. 上海理工大学复杂系统科学研究中心 上海 杨浦区 200093；2. 上海财经大学合计与财务研究院 上海 杨浦区 200433)

【摘要】聚类分析是数据挖掘中一项重要的技术，通过对多维用户行为的聚类分析，可以从用户层面来帮助管理人员得到更为精确和有效的用户评价信息。该文首先从用户行为数据中提取多维用户行为特征，之后采用基于互信息的无监督特征选择（UFS-MI）模型对提取的特征进行排序、筛选并确定权重，得到每个用户行为的加权特征向量。根据用户行为之间的相似性构造网络，然后通过 Blondel 社团划分算法对用户行为网络进行聚类分析。在某公交线路的实证数据集上的实验结果表明，该方法的准确率为 92%，比传统聚类算法 K-means 的准确率有明显提升，研究结果可以为公交公司的管理层在进行统一管理和培训时提供参考。本文的工作拓展了网络科学在多维用户行为数据聚类分析的应用范围，丰富了多维驾驶行为数据聚类分析的思路，为决策者提供参考依据。

关 键 词 聚类分析； 特征筛选； 多维数据； 用户行为

中图分类号 N949 **文献标志码** A doi:10.12178/1001-0548.2018212

User Clustering Method Based on Multi-dimensional Behavior Analysis

ZHANG Lin-bing[1], GUO Qiang[1], WU Xing-bin[1], LIANG Yao-zhou[1], and LIU Jian-guo[2*]

(1. Research Center of Complex Systems Science, University of Shanghai for Science and Technology Yangpu Shanghai 200093;
2. Institute of Accounting and Finance, Shanghai University of Finance and Economics Yangpu Shanghai 200433)

Abstract Clustering analysis is an important technology in data mining. By clustering analysis of multi-dimensional user behavior, it can help managers get more accurate and effective user evaluation information from the user level. In this paper, multi-dimensional user behavior features are extracted from user behavior data, and then unsupervised feature selection based on mutual information (UFS-MI) is used to sort, filter and confirm the features of the extracted features, and the weighted feature vectors of each user's behavior are obtained. The network is constructed according to the similarity between user behaviors, and then the user behavior network is clustered and analyzed by Blondel community partition algorithm. The experimental results on an empirical data set of a bus line show that the accuracy of the method is 92%, which is significantly higher than the accuracy rate of the traditional clustering algorithm K-means. The results can provide a reference for the management and training of the public transport management. This paper expands the application scope of network science in multi-dimensional user behavior data clustering analysis, enriches the idea of multi-dimensional driving behavior data clustering analysis, and provides reference for managers.

Key words cluster analysis; feature selection; multi-dimensional data; user behavior

随着大数据技术的不断发展，人们收集到的用户行为数据维度越来越多，如何能够有效的对多维用户行为数据进行分析，是目前行为分析的难点之一[1-2]。聚类分析是数据挖掘领域中较为基础的数据处理手段，通过聚类算法对数据分类能够将一个数据集划分为若干个类内对象相似而类间对象相异的类簇[3]，从而在数据集中发现潜在的数据模式和内在联系[4]，为此国内外的众多专家学者们研究了各类聚类算法。其中传统聚类算法主要可以分为层次化聚类算法、划分式聚类算法和基于密度的聚类算法[5]。层次聚类算法又称为树聚类算法，它的优点是距离和规则的相似度容易定义、不需要预先制定聚类数、可以发现类的层次关系，缺陷[6] 在于没有全局待优化的目标函数；合并或分裂点的选择困

收稿日期：2018－07－23；修回日期：2018－12－13
基金项目：国家自然科学基金 (61773248, 71771152)
作者简介：张林兵 (1994－)，男，主要从事复杂网络方面的研究.
通信作者：刘建国，E-mail：liu.jianguo@sufe.edu.cn

难，好的局部合并选择不能保证高质量的全局聚类结果；算法的计算复杂度高，适合小型数据集的分类；对噪声、孤立点敏感，不适合非凸型分布数据集。K-means 算法是经典的划分式聚类算法，它的优点[7]是思想简单、易于实现，可用于大规模数据集的并行聚类挖掘，通常在对大型数据集聚类时，K-means 算法比层次聚类算法快得多，它的缺点是需要事先确定聚类个数k的大小，因为很多应用事先是无法确定的，如网络社团的划分；k个初始聚类中心是随机选择的，由于随机选择k个初始聚类中心，导致算法对异常数据敏感。DBSCAN 聚类算法是经典的基于密度的聚类算法，它的优点[8]是不需要事先确定簇的个数以及选择初始聚类中心，能够识别噪声数据点，且对数据点的输入顺序不敏感，缺点是需要事先确定 Eps 和 MinPts 这 2 个参数，而这 2 个参数的确定无规律可循且 DBSCAN 算法对这 2 个参数比较敏感，参数的轻微变化可能导致差别较大的聚类结果，DBSCAN 算法不能有效地处理数据分布比较均匀的数据集，也无法有效处理维数较大的数据集。上述的传统聚类方法在进行多维行为数据聚类分析时，存在很多问题，因而传统聚类算法不能直接应用到多维行为聚类分析。为了解决这个问题，本文尝试用网络科学[9-11]的方法对多维行为数据聚类分析。

与小世界性、无标度性[12-13]等基本统计特性相并列，网络簇结构 (network community structure，NCS) 是复杂网络最普遍和最重要的拓扑结构属性之一，具有同簇节点相互连接密集、异簇节点相互连接稀疏的特点，复杂网络聚类方法旨在揭示出复杂网络中真实存在的网络簇结构。复杂网络聚类算法主要分为启发式方法 (heuristic method，HM) 和基于优化的方法 (optimization based method，OBM)[14]。文献 [15] 提出的 GN 算法是经典的启发式方法，该方法的优点是思想简单而得到广泛应用，缺点是计算速度慢，不适合大规模的网络，同时又难以确定合适的终止条件。文献 [16] 提出的分级凝聚快速算法 (FN 算法) 是经典的基于优化的方法，与 GN 算法相比，时间复杂度大大降低，但准确性不如 GN 算法。文献 [17] 提出的 Blondel 算法是一种基于模块度最优化的启发式算法，与普通的基于模块度和模块度增益算法相比该算法的执行效率高且聚类效果非常明显，是目前国际上公认的执行速度最快且精度较高的非重叠社区发现算法[18]，因而本文选择用 Blondel算法进行聚类分析。

本文的主要贡献是：1) 将机器学习中的无监督特征选择方法与网络科学中的社团划分算法相结合，提出一种多维用户行为聚类分析方法。在某公交线路的实证数据集上的实验结果表明，该方法聚类准确率明显高于传统 K-means 算法；2) 本文提出的方法不仅为多维驾驶行为数据分析提供新的思路，还可以在不同的场景中广泛应用，例如金融市场的数据分析、互联网企业用户行为的数据挖掘等。

1 模型与方法

本文提出基于复杂网络多维用户行为聚类方法。首先对原始数据进行预处理，包括数据清洗和数据采样，之后从处理好的数据中提取多维用户行为特征，构建用户行为特征向量。然后用 UFS-MI 模型对多维用户行为特征向量降维并给特征确定权重，基于加权的用户行为特征向量计算不同用户之间的相似性构建网络。最后用 Blondel 算法对网络进行聚类分析。实验的流程图如图 1 所示。

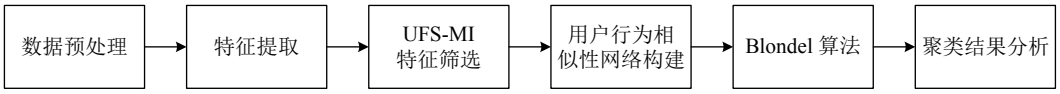

图 1 实验流程图

1.1 UFS-MI 特征选择模型

UFS-MI 是一种基于互信息的无监督特征选择模型，属于过滤型特征排序方法。UFS-MI 模型在进行特征选择时，首先计算出每个特征的相关度，再使用前向顺序搜索对特征进行重要性评价，最后输出一个有序特征序列。UFS-MI 模型的评价标准 UmRMR 综合考虑了特征的相关度和冗余度的信息度量[19-20]。

假设集合$D=\{f_1,f_2,\cdots,f_n\}$表示完整的特征集合，$f_i(i=1,2,\cdots,n)$表示特征集合中的第i个特征，$P(f_t)$是特征为f_t的概率。特征f_t取值的初始不确定性可由如下信息熵度量：

$$H(f_t)=-\sum_{f_t}P(f_t)\log P(f_t) \tag{1}$$

在已知另一个特征$f_{t'}$的取值之后，f_t取值的不确定性由条件熵来度量：

$$H(f_t|f_{t'}) = -\sum_{f_{t'}} P(f_{t'}) \sum_{f_t} P(f_t|f_{t'}) \log P(f_t|f_{t'}) \quad (2)$$

两个特征f_t与$f_{t'}$之间的互信息定义为：

$$I(f_t; f_{t'}) = H(f_t) - H(f_t|f_{t'}) = I(f_{t'}; f_t) \quad (3)$$

特征选择过程从一个空集S开始，采用步进的方式，每次从特征全集中选择一个特征，令：

$$\text{score}(f_i) = \frac{1}{n} \sum_{t=1}^{n} I(f_i : f_t) \quad (4)$$

$$l_1 = \arg\max_{1 \leqslant i \leqslant n} \{\text{score}(f_i)\} \quad (5)$$

由式 (5) 可知，选择的第一个重要特征 g_1 为 f_{l_1}，因为在只选择一个特征的情况下，g_1 可以最大程度地降低其他未选特征的不确定性。

假设U为未被选择的特征集合，S_{m-1}为已经被选择的$m-1$个特征集合。在选择第m个特征 g_m 时，g_m 应该与U中的所有特征最大程度相关，同时与S_{m-1}中的$m-1$个特征最小程度的冗余。

一个特征f_i的相关度就是其与整个特征集合的平均互信息：

$$\text{Rel}(f_i) = \frac{1}{n}(H(f_i) + \sum_{1 \leqslant t \leqslant n, t \neq i} I(f_i; f_t)) \quad (6)$$

式中，$H(f_i)$表示特征f_i所包含的信息量，$H(f_i)$值越大，表明特征f_i能够提供给学习算法的信息越多；$\sum_{1 \leqslant t \leqslant n, t \neq i} I(f_i; f_t)$表示已知特征$f_i$的信息后，其他特征包含的信息量的减少量，其值越大，表明其他特征能够提供给学习算法除f_i信息以外的信息越少。所以选择具有最大相关度的特征f_i(即$\text{Rel}(f_i)$取最大值)，数据就可以最小程度地丢失信息。

一个特征 g_t 对特征f_i的相关度定义为：

$$\text{Rel}(g_t|f_i) = \frac{H(g_t|f_i)}{H(g_t)} \text{Rel}(g_t) \quad (7)$$

显然，条件相关度小于等于相关度 (当两个特征相互独立时相等)，将两特征之间的差别定义为冗余。

一个特征f_i对特征 g_t 的冗余度定义为：

$$\text{Red}(f_i; g_t) = \text{Rel}(g_t) - \text{Rel}(g_t|f_i) \quad (8)$$

所以在选择第m个重要特征时，综合考虑候选特征的相关度以及已选特征的冗余度，得到“无监督最小冗余-最大相关”特征重要性评价标准 (UmRMR)：

$$\text{UmRMR}(f_i) = \text{Rel}(f_i) - \max_{g_t \in S_{m-1}} \{\text{Red}(f_i; g_t)\} \quad (9)$$

$$l_m = \arg\max_{1 \leqslant i \leqslant n} \{\text{UmRMR}(f_i) | f_i \in U\} \quad (10)$$

由式 (10) 得，第m个特征选择为$g_m = f_{l_m}$，因为该特征最大程度地降低了其他特征的不确定性，同时带来最少的冗余信息，所以采用该方法逐个选取特征。

1.2　用户行为相似性度量

相似性度量，即综合评定两个事物之间相近程度的一种度量。两个事物越接近，它们的相似性度量也就越大，而两个事物越疏远，它们的相似性度量也就越小。假设每个特征具有不同的重要程度，可用权向量w表示，用来计算x，y两个用户行为之间的相关性。采用加权相关度对两个用户之间的行为特征进行计算。

加权相关度的计算公式为：

$$m(x; w) = \frac{\sum_i w_i x_i}{w_i} \quad (11)$$

$$\text{cov}(x, y, w) = \frac{\sum_i w_i (x_i - m(x; w))(y_i - m(y; w))}{\sum_i w_i} \quad (12)$$

$$\text{corr}(x, y; w) = \frac{\text{cov}(x, y; w)}{\sqrt{\text{cov}(x, x; w)\text{cov}(y, y; w)}} \quad (13)$$

1.3　Blondel 算法

Blondel 算法常被用于社团划分问题[21]。在社交网络中，用户相当于每一个点，用户之间通过互相的关联关系构成了整个网络的结构，有的用户之间的连接较为紧密，有的用户之间的连接关系较为稀疏，在这样的网络中，连接较为紧密的部分可以被看成一个社团，其内部的节点之间有较为紧密的连接，而在两个社团间则相对连接较为稀疏，这便称为社团结构。为了评价社团划分的优劣，用模块度来衡量社团划分的好坏。模块度的计算公式如下：

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \quad (14)$$

式中，A_{ij}是实际网络的邻接矩阵；k_i和k_j分别为原网络中节点i和节点j的度；c_i与c_j分别表示节点i与节点j在网络中所属的社团，如果这两个节点属于同一社团，δ取值为 1，否则δ取值为 0。

Blondel 算法的思想是：首先将网络中的每个节点看成是一个独立的社团，慢慢将邻近的节点合并，如果合并之后整个网络的模块度提高，那么就合并，否则撤销；如此循环，直到网络的模块度无

法提高为止；接着再把每个社团当成一个节点，对每个社团进行如此的合并算法，直到整个网络的模块度无法提高为止。

2 实验及分析

本实验的数据来源于某公交线路 2017 年 9 月 1 日～2017 年 9 月 30 日，133 名司机、55 辆车、共 16 个字段的驾驶信息。

2.1 数据的预处理

为了让司机的驾驶行为特征具有可比性，设定一个具体场景，即从起点至终点的一趟行车记录作为每个司机的驾驶行为特征计算基准。对每趟行车路程设定阈值进行筛选，最终得到包含 103 名司机、51 辆车、879 趟完整行车记录的实验数据。

2.2 特征提取与筛选

结合原始数据和业务场景提取了车速平均值、车速中位数、车速标准差、加速度绝对值平均值、加速度标准差、电子刹车使用概率、油门踏板百分比平均值、油门踏板百分比标准差、脚刹使用概率、加速度绝对值大于 2 m/s^{2}的概率、行车过程中拉手刹的概率、空挡状态下的滑行概率共 12 个特征。

为了从初步提取的众多特征中筛选出需要的有效特征，用 UFS-MI 模型对特征进行重要性排序，然后选取具有代表性的特征。

2.3 司机驾驶行为相似性度量

选取排序靠前的 9 个特征，按照平均互信息值的大小确定权重，得到权向量，然后计算每两趟行车记录之间的皮尔森相关系数。设定阈值为 0.94，当两趟行车记录的行为相似性大于该阈值时，建立连边。最终构造成的网络包含 879 个节点，183 046 条连边。

2.4 聚类准确性度量

将构造成的网络用 Blondel 算法进行聚类，聚类结果将驾驶记录分为 3 类。第一类包含 55 个司机共 365 趟行车记录，第二类包含 64 个司机共 325 趟行车记录，第三类包含 21 个司机共 189 趟行车记录。

对于一个司机而言，如果司机驾驶行为是稳定的，那么他所有驾驶趟都会分到同一类中，但在司机驾驶行为发生变化的情况下，就会被分到不同的类中。因此本文定义了一个分类准确性指标：

$$p_c = \frac{1}{m}\sum_{i=1}^{m}\frac{\max\{n_i^{C_l}\}}{n_i} \qquad C_l = 1,2,3 \qquad (15)$$

式中，n_i为司机i行驶的总趟数；$n_i^{C_l}$为第C_l类中司机i的行驶趟数；$\max\{n_i^{C_l}\}$为司机i在C_l类中行驶最多的趟数；m为司机总数。对所有司机求平均，得到平均分类准确性。

根据平均分类准确性指标，计算得出 Blondel 算法分类准确率为 92%，而传统算法 K-means 算法在 k=3 时，分类准确率为 75%。

2.5 聚类结果分析

因为每一个类别中的司机驾驶行为是以趟的形式来度量的，所以根据 Blondel 算法聚类的结果实际上是不同趟的行车记录。由于公交司机驾驶的车辆会更换，同一司机在不同车辆上的驾驶行为可能不同，因此，需要先从趟的信息中，提取出司机的类别和驾驶车辆的类别，然后再进行用户行为分析。将 9 个驾驶行为特征转化为 3 个综合驾驶行为维度：驾驶不平稳性、刹车偏好性、车速偏好性。为了将特征对应到综合驾驶行为维度上，首先对特征进行 0～1 标准化处理，去除特征数据的单位限制。然后对无量纲的特征数值进行综合驾驶行为维度分析，得到如图 2 所示的驾驶行为偏好雷达图。最后，将 3 个综合驾驶行为维度的评分求平均，得到如图 3 所示的司机的综合评分直方图。

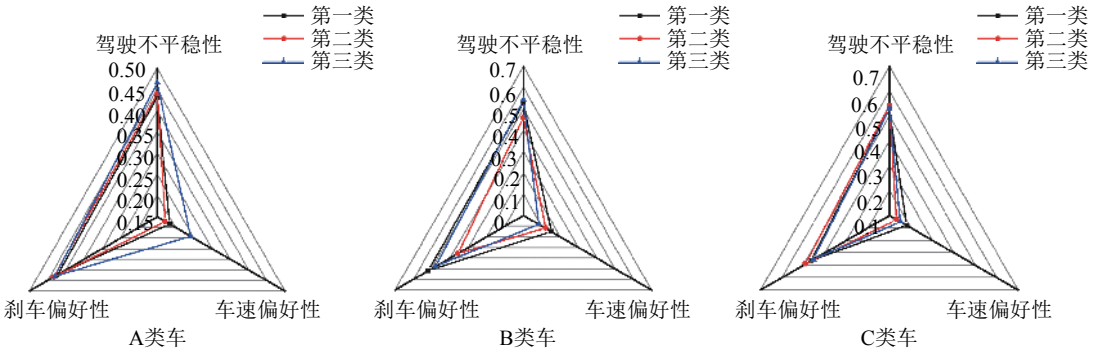

图 2 不同车型下各类司机的驾驶行为图

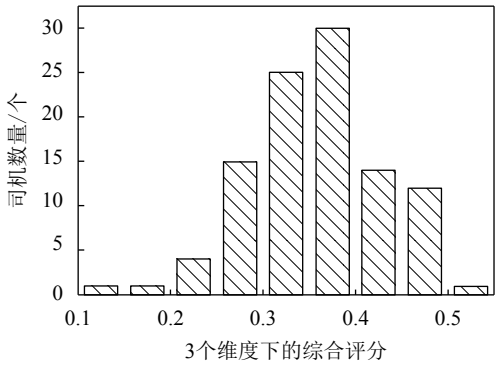

图 3　司机驾驶行为综合评分直方图

由图 2 可以看出，在驾驶 A 类车时，3 类司机在刹车偏好性维度上具有显著的区别。驾驶 B 类车时，第二类司机的驾驶行为在 3 个维度上都要优于其他两类司机。而对于 C 类车，3 类司机在刹车偏好性维度上也有明显区别。由图 3 可以看出，司机在 3 个维度下的综合评分接近正态分布，综合评分较低的司机较少，这表明驾驶行为优秀的司机占极少数而综合评分较高的司机相对较多，表明大部分司机驾驶行为需要改善。

3　结 束 语

对多维用户行为进行聚类分析，可以帮助管理人员得到更为精确和有效的用户评价信息，为管理层决策参考提供依据。本文从多维用户行为数据中提取用户行为特征，采用 UFS-MI 模型对提取的用户行为特征进行排序并筛选，然后按照平均互信息的值给特征确定权重，得到用户行为的加权特征向量。通过计算用户行为之间的皮尔森相关系数，设定阈值并构建网络，再结合复杂网络理论，采用 Blondel 社团划分算法对用户行为网络进行聚类分析。在某公交线路的实证数据集上的实验结果表明，该方法的准确率为 92%，比传统聚类算法 K-means 的准确率有明显提升。

本文提供的方法还有众多的应用场景，例如根据股票价格波动的相似性构建股票关联网络，对股票进行聚类分析。根据个股进行相关股的推荐，为投资者提供参考。通过对互联网企业用户簇集进行数据挖掘，有助于企业及时掌握和研究用户的总体变化，为不同类型的用户提供更有针对性的个性化服务，从而增加企业市场份额和利润。此外，本文根据 UFS-MI 模型进行特征筛选，没有结合具体的业务，未来的工作可以结合具体业务对特征进行筛选，从而提高聚类的效果。

参 考 文 献

[1] CHEN D, TANG J, LI J, et al. Discovering the staring people from social networks[C]//Proceedings of the 18th International Conference on World Wide Web. [S.l.]: ACM, 2009: 1219-1220.

[2] HUANG X, WU Q. Micro-blog commercial word extraction based on improved TF-IDF algorithm[C]// TENCON 2013-2013 IEEE Region 10 Conference (31194). Xi'an, China: IEEE, 2013: 1-5.

[3] JAIN A K, DUBES R C. Algorithms for clustering data[J]. Technometrics, 1988, 32(2): 227-229.

[4] DU K L, SWAMY M N S. NEURAL networks and statistical learning[J]. Science Business Media, 2014: 727-745.

[5] 孙吉贵, 刘杰, 赵连宇. 聚类算法研究[J]. 软件学报, 2008, 19(1): 48-61.
SUN Ji-gui, LIU Jie, ZHAO Lian-yu. Clustering algorithms research[J]. Journal of Software, 2008, 19(1): 48-61.

[6] 金阳, 左万利. 一种基于动态近邻选择模型的聚类算法[J]. 计算机学报, 2007, 30(5): 756-762.
JIN Yang, ZUO Wan-li. A clustering algorithm using dynamic nearest neighbors selection model[J]. Chinese Journal of Computers, 2007, 30(5): 756-762.

[7] HUANG Z. A fast clustering algorithm to cluster very large categorical data sets in data mining[J]. Research Issues on Data Mining & Knowledge Discovery, 1997: 1-8.

[8] HOLDEN N P, FREITAS A A. A hybrid PSO/ACO algorithm for classification[C]//Conference Companion on Genetic and Evolutionary Computation. [S.l.]: ACM, 2007: 2745-2750.

[9] LIU J G, HOU L, PAN X, et al. Stability of similarity measurements for bipartite networks[J]. Scientific Reports, 2016, 6: 18653.

[10] LIU J G, LIN J H, GUO Q, et al. Locating influential nodes via dynamics-sensitive centrality[J]. Scientific Reports, 2016, 6: 21380.

[11] 周卿, 郭强, 刘建国. 基于交互频率的动态网络上的社会知识传播研究[J]. 上海理工大学学报, 2017, 39(1): 25-29.
ZHOU Qing, GUO Qiang, LIU Jian-guo. Social knowledge diffusion on dynamical networks in terms of interaction frequency[J]. Journal of University of Shanghai for Science and Technology, 2017, 39(1): 25-29.

[12] WATTS D J, STROGATZ S H. Collective dynamics of 'small-world'networks[J]. Nature, 1998, 393(6684): 440.

[13] BARABÁSI A L, ALBERT R. Emergence of scaling in random networks[J]. Science, 1999, 286(5439): 509-512.

[14] 杨博, 刘大有, 金弟, 等. 复杂网络聚类方法[J]. 软件学报, 2009, 20(1): 54-66.
YANG Bo, LIU Da-you, JIN Di, et al. Complex network clustering algorithms[J]. Journal of Software, 2009, 20(1): 54-66.

[15] GIRVAN M, NEWMAN M E J. Community structure in social and biological networks[J]. Proceedings of the National Academy of Sciences, 2002, 99(12): 7821-7826.

[16] NEWMAN M E J. Fast algorithm for detecting community structure in networks[J]. Physical Review E,

2004, 69(6): 066133.

[17] BLONDEL V D, GUILLAUME J L, LAMBIOTTE R, et al. Fast unfolding of communities in large networks[J]. Journal of Statistical Mechanics, 2008(10): 155-168.

[18] ROBERT B M, CHRISTOPHER J F, JASON J J, et al. A 61-million-person experiment in social influence and political mobilization[J]. Nature, 2012, 489(7415): 295.

[19] 徐峻岭, 周毓明, 陈林, 等. 基于互信息的无监督特征选择[J]. 计算机研究与发展, 2012, 49(2): 372-382.
XU Jun-ling, ZHOU Yu-ming, CHEN Lin, et al. An unsupervised feature selection approach based on multual information[J]. Journal of Computer Research and Development, 2012, 49(2): 372-382.

[20] WANG Y, WANG J, LIAO H, et al. An efficient semi-supervised representatives feature selection algorithm based on information theory[J]. Pattern Recognition, 2017, 61: 511-523.

[21] YANG K, GUO Q, LI S N, et al. Evolution properties of the community members for dynamic networks[J]. Physics Letters A, 2017, 381(11): 970-975.

编 辑 叶 芳