文章编号：2095-2163(2019)03-0266-04 中图分类号：TP391 文献标志码：A

基于联合半监督学习的大数据聚类算法

谌裕勇
（广东工业大学 华立学院，广州 511325）

摘 要：为了提高对用户行为特征挖掘能力，需要对用户行为特征多维度文本数据进行优化聚类处理，提出一种基于联合半监督学习的大数据聚类算法。采用分段线性拟合方法进行用户行为特征大数据线性规划处理，提取用户行为特征大数据的互信息特征量，结合联合关联规则检测方法进行用户行为特征多维度文本数据的统计分析，构建大数据分布的关联属性样本集，采用联合半监督学习分类器进行数据分类，结合多传感量化跟踪识别方法进行聚类中心自动搜索，提高聚类收敛性。仿真结果表明，采用该方法进行用户行为特征多维度文本数据聚类处理的信息融合性能较好，数据聚类中心的自动搜索能力较强，提高了大数据分类检索能力。

关键词：联合半监督学习；大数据；用户行为特征；聚类

Big data clustering algorithm based on joint semi-supervised learning

CHEN Yuyong
（Huali College, Guangdong University of Technology, Guangzhou 511325, China）

【Abstract】 In order to improve the ability of user behavior feature mining, it is necessary to optimize the clustering of user behavior feature multi-dimensional text data. A big data clustering algorithm based on joint semi-supervised learning is proposed. The piecewise linear fitting method is used to deal with the user behavior feature big data, and the mutual information feature quantity of user behavior feature big data is extracted. Combined with the joint association rule detection method, the multi-dimensional text data of user behavior characteristics are analyzed, and the association attribute sample set distributed by big data is constructed, and the joint semi-supervised learning classifier is used to classify the data. The clustering center is automatically searched by multi-sensor quantization tracking and identification method to improve the clustering convergence. The simulation results show that this method has better information fusion performance and better automatic searching ability of data clustering center, which improves the ability of big data classification and retrieval.

【Key words】 joint semi-supervised learning; big data; household behavior characteristics; clustering

0 引 言

随着大数据信息技术的发展，在云环境中进行大数据的聚类处理，实现对数据的优化分类检索和识别，在社交网络中，需要对网络用户行为特征的文本大数据进行优化聚类处理，结合数据的聚类属性特征进行融合调度和分类识别，提高对用户行为特征的准确定位分析能力，研究基于大数据的用户行为特征多维度文本数据聚类方法，在提高社交网络的信息推荐能力和大数据信息处理能力方面具有重要意义[1]。对用户行为特征多维度文本信息聚类处理是建立在对数据的多维度特征提取和关联规则挖掘基础上，结合传感数据采集方法提取用户行为特征多维度文本信息的关联规则特征量，实现多维度文本数据分类识别[2]。本文提出一种基于联合半监督学习的大数据聚类算法。采用分段线性拟合方法进行用户行为特征大数据规划处理，提取用户行为特征大数据的互信息特征量，采用联合半监督学习分类器进行数据分类，最后进行仿真实验分析，展示了本文方法在提高用户行为特征多维度文本数据聚类能力方面的优越性能。

1 用户行为特征大数据采样及特征参量提取

1.1 用户行为特征多维度文本特征数据采样

在社交网络中，用户行为特征多维度文本信息结构复杂，系统耦合性强，通过对用户行为特征多维度文本数据分类，实现对用户行为特征的优化检测和分类识别，采用多维度文本信息融合方法进行社区网络用户行为特征检测和智能分析[3]。构建用户行为特征多维度文本特征数据分布结构模型如图 1 所示。

根据图 1，用户行为特征分布集合在 B 模型中的输出状态特征量为 $x_j = \{x_{1j}, x_{2j}, \ldots, x_{mj}\}^{\mathrm{T}}$，以 2 倍以上波特率进行采样，得用户行为特征多维度文本

作者简介：谌裕勇（1979-），男，硕士，讲师，主要研究方向：数据挖掘、数据分析、机器学习等。

收稿日期：2018-12-20

数据的状态特征分布为 $p(x_0)$，文本数据的关联规则联合特征挖掘结果为：

$$P_{ij}(k)=\frac{(l_j(k)-l_i(k))\eta_{ij}(k)}{\sum_{j\in N_i(k)}(l_j(k)-l_i(k))\eta_{ij}(k)},\tag{1}$$

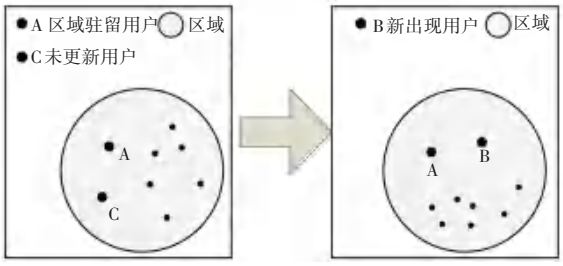

图 1　用户行为特征多维度文本特征数据分布结构模型

Fig. 1　Multi-dimensional text feature data distribution structure model for user behavior features

根据用户行为特征多维度文本信息传输码元特征量，进行信息重构，采用模糊数据聚类分析技术[4]，得到用户行为特征多维大数据传输的比特序列分布为：

$$x(t)=\sum_{i=0}^{p}a(\theta_i)s_i(t)+n(t),\tag{2}$$

求得用户行为特征多维度文本数据的语义概念集，对用户行为特征多维度文本数据进行粗糙集调度和频繁性挖掘[5]，根据数据聚集树分层特征得到用户行为特征多维度文本数据分类状态特征量为 $z(t)$，数据聚类中心的粗糙概念分布子集 $S_i(i=1,2,\cdots,L)$ 满足半监督学习的收敛性条件为：

$$p(y\mid\alpha,\theta)=\sum_{k=1}^{K}\alpha_k p_k(y\mid\mu_k,\sum{}_k),\tag{3}$$

根据上述分析，采用一种网格聚类方法进行用户行为特征多维度文本数据分类处理，结合小扰动抑制方法避免聚类中心扰动，提高聚类的收敛性。

1.2　用户行为特征大数据线性规划处理

采用分段线性拟合方法进行用户行为特征大数据线性规划处理，提取用户行为特征大数据的互信息特征量，描述为：

$$P_i(t)=\sum_{n=1}^{N}\frac{A}{r}e^{-jkr}R_{in}\frac{1}{r}e^{-ikr},\tag{4}$$

对于用户行为特征多维度文本数据的标量时间序列为 $x(t),t=0,1,\cdots,n-1$，给定用户行为特征多维度文本数据信息流的一向量组 $x_1,x_2,\cdots,x_n\in C^m$（$m$ 维复数空间），结合线性规划方法，得到用户行为特征多维度文本数据集分布的有限集合为：

$$\Sigma=diag\{\max\{|\rho_1^+|,|\rho_1^-|\},\cdots,\max\{|\rho_n^+|,|\rho_n^-|\}\}=diag\{\rho_1,\cdots\rho_n\},\tag{5}$$

$$\Sigma_1=diag\{\rho_1^+\rho_1^-,\cdots,\rho_n^+\rho_n^-\},\tag{6}$$

对融合数据进行分段样本组合设计，得到用户行为特征多维度文本数据的关联规则集特征提取的时间间隔为 $O(d)$ 和 $O(N^{\frac{1}{d}})$，数据聚类空间的嵌入维数 $m\to1$ 时，$sn\xi\to\tanh\xi$，由此得到用户行为特征多维度文本数据准确聚类的边值收敛条件满足：

$$\Delta E=-\eta\left[\left(\frac{\partial E}{\partial\omega}\right)^2+\left(\frac{\partial E}{\partial b}\right)^2\right],\tag{7}$$

如果 $C_o(x^*)=0$，则：

$$Y(P,Q,\beta)=Y[red(P,Q,\beta),Q,\beta],\tag{8}$$

设计 3 种核函数分别表示用户行为特征多维度文本数据聚类的线性核函数、随机分布特征核函数和均匀分布核函数[6]，表达式分别为：

$$K(x_i,x_j)=<x_i,x_j>;\tag{9}$$

$$K(x_i,x_j)=(<x_i,x_j>+1)^d;\tag{10}$$

$$K(x_i,x_j)=\exp(\|x_i-x_j\|^2/2\sigma^2).\tag{11}$$

根据上述三个核函数进行用户行为特征多维度文本数据准确聚类的线性规划设计，结合半监督学习算法，提高数据聚类过程中的收敛控制能力[7]。

2　大数据聚类优化

在上述采用分段线性拟合方法进行用户行为特征大数据线性规划处理的基础上，进行大数据聚类算法的优化设计，本文提出一种基于联合半监督学习的大数据聚类算法。提取用户行为特征大数据的互信息特征量[8]，得到用户行为特征大数据聚类的几何邻域 (t,f) 在非线性空间的特征分布值为：

$$f(x)=\begin{cases}f(x),x\in Levf;\\ a,\quad x\in Levf.\end{cases}\tag{12}$$

结合联合关联规则检测方法进行用户行为特征多维度文本数据的统计分析，在聚类空间矩阵 $\boldsymbol{G}=[\boldsymbol{E}_{k\times k}\mid\boldsymbol{A}]$ 中，求得数据聚类的基向量 $(x_1,x_2,\cdots,x_n)$，构建用户行为特征多维度文本数据聚类的联合扰动特征方程组为：

$$\begin{cases}a(H_{ac})=1-\dfrac{H_{ac}}{\max(H_{ac})+l};\\ \max(H_{ac})=\log_2 k.\end{cases}\tag{13}$$

通过上述对用户行为特征多维度文本数据准确聚类的边值收敛条件分析，保证了整个数据聚类数学模型的稳定收敛性[9]。采用半监督学习方法，构建用户行为特征多维度文本数据聚类的边界解向量函数为：

$$w_{ji}(k+1)=w_{ji}(k)-\alpha\frac{\partial F}{\partial w_{ji}},\quad(14)$$

$$z_{kj}(k+1)=z_{kj}(k)-\alpha\frac{\partial F}{\partial z_{kj}},\quad(15)$$

结合联合关联规则检测方法进行用户行为特征多维度文本数据的统计分析[10],得到统计特征方程描述为:

$$\dot{x}(t)=Ax(t)+Bx(t-d_1(t)-d_2(t)),\quad(16)$$

其中,$x(t)=\phi(t)$,$t\in[-h,0]$,为了实现数据优化聚类,在有限维空间中输入新的训练向量:

$$x(t)=(x_0(t),x_1(t),\cdots,x_{k-1}(t))^{\mathrm{T}},\quad(17)$$

对于 $x(t)$,采用联合半监督学习进行迭代过程的收敛性约束控制,得到数据聚类中心的空间聚类为:

$$d_j=\sum_{i=0}^{k-1}(x_i(t)-\omega_{ij}(t))^2,\ j=0,1,\cdots,N-1.\quad(18)$$

其中,$\omega_j=(\omega_{0j},\omega_{1j},\cdots,\omega_{k-1,j})^{\mathrm{T}}$,当聚类中心满足半监督学习的收敛性条件时,用户行为特征多维度文本数据的检测统计量满足聚类收敛性条件。综上分析,得到本文设计的大数据聚类算法的实现流程如图 2 所示。

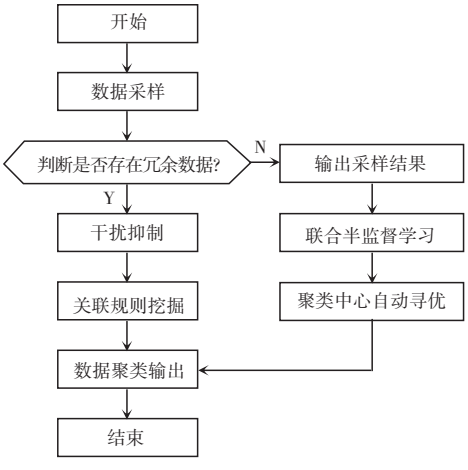

图 2 大数据聚类算法的实现流程

Fig. 2 Realization flow of big data clustering algorithm

3 仿真实验分析

为了测试本文方法在实现用户行为特征多维度文本数据的聚类中的性能,进行仿真实验,实验建立在 Deep Web 数据库基础上,结合 Matlab 进行数据聚类算法设计,大数据样本的属性设置为 6,数据聚类的初始置信度为 95%,临界值 $Q_c=1.24$,判断阈值为 0.13,特征空间分布的嵌入维数设定为 $m=4$,测试样本集的数据长度为 2 000,仿真时长为 120 s,根据上述仿真环境和参数设定,进行用户行为特征多维度文本大数据聚类分析,得到原始数据分布如图 3 所示。

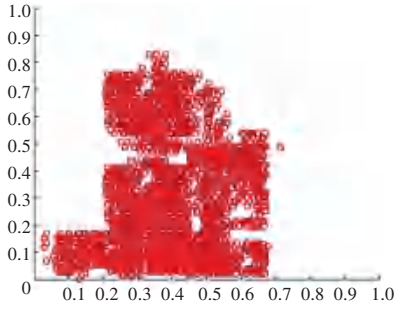

图 3 原始数据分布

Fig. 3 Raw data distribution

以图 3 的数据为研究对象,进行数据聚类处理,采用联合半监督学习分类器进行数据分类,得到聚类输出结果如图 4 所示。

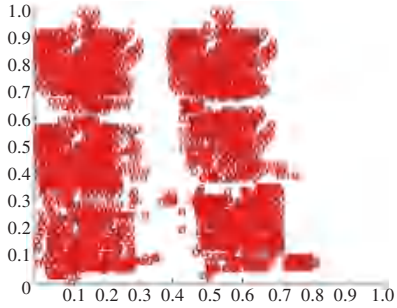

图 4 大数据聚类输出

Fig. 4 Big data cluster output

分析图 4 得知,采用本文方法能有效实现大数据聚类处理,数据分类的准确性较高,误分率较小,测试不同方法进行大数据聚类的性能,得到对比结果如图 5 所示,分析图 5 得知,本文方法进行大数据聚类的误分率较低,性能优于传统方法。

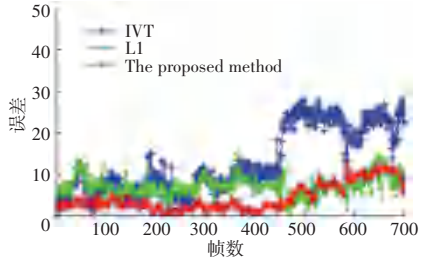

图 5 数据聚类的性能对比

Fig. 5 Performance comparison of data clustering

4 结束语

结合传感数据采集方法提取用户行为特征多维度文本信息的关联规则特征量,实现多维度文本数据分类识别,本文提出一种基于联合半监督学习的

(下转第 272 页)