计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200800011

一种基于符号关系图的快速符号数据聚类算法

张岩金[1]　白　亮[1,2]
1 山西大学计算机与信息技术学院　太原 030006
2 山西大学计算机智能与中文信息处理教育部重点实验室　太原 030006
(zhang17836204220@163.com)

摘　要　由于在实际应用中有大量的符号数据生成,符号数据聚类成为了聚类分析的一个重要研究领域。目前,已有许多符号数据聚类算法被提出,但将它们应用于大数据环境时,仍然存在计算成本高、运行速度慢等问题。文中提出了一种基于符号关系图的快速符号数据聚类算法。该算法使用符号关系图替代原始数据,缩小数据集的规模,有效地解决了这一问题。大量的实验分析显示新算法相比其他算法是有效的。

关键词:符号数据;相似性度量;关系图;数据挖掘;聚类
中图法分类号　TP391

Fast Symbolic Data Clustering Algorithm Based on Symbolic Relation Graph

ZHANG Yan-jin[1] and BAI Liang[1,2]
1 School of Computer and Information Technology,Shanxi University,Taiyuan 030006,China
2 Key Laboratory Computational Intelligence and Chinese Information Processing of Ministry of Education,Taiyuan 030006,China

Abstract　Since a large amount of symbolic data is generated in practical applications,clustering of symbolicl data becomes an important research area of cluster analysis. Currently,many symbolic data clustering algorithms are proposed. When they are applied in big data environment,there are still problems such as high computational cost and slow operation speed. This paper proposes a fast symbolic data clustering algorithm based on symbolic relation graphs. It effectively solves this problem by replacing the original data with a symbolic relation graph and reducing the size of the data set. A large number of experiments show that the new algorithm is more effective than other algorithms.

Keywords　Symbolic data,Similarity measure,Relation graph,Data mining,Clustering

1　引言

聚类分析是机器学习[1]和数据挖掘[2]的重要研究领域,其目的是根据对象之间的相似性对数据集进行划分,使得同一组中的对象比在不同组的对象更相似[3]。现有的数据一般分为3类:数值型数据、符号型数据和混合型数据。处理不同类型的数据需要不同的聚类方法。数值数据一般通过数据点间的数值差异性构造距离作为相似性度量(如欧氏距离、马氏距离等),从而使用聚类方法进行划分。与数值数据不同,符号数据的取值是有限且无序的,无法通过数据之间数值的差异得到合适的相似性度量,因此一些经典的数值聚类方法无法有效地处理符号数据。近年来,由于符号数据的大量存在,符号数据聚类[4]方法已经成为一个重要的研究领域,在零售业、电子商务、医疗诊断、生物信息学等领域都有大量的应用。目前,符号数据的聚类方法可以分为3类:基于相异测度的聚类算法、基于概率统计的聚类算法和基于信息熵的聚类算法。

基于相异测度的聚类算法是参照数值数据聚类的思想建立的算法,其引入了适合于符号数据的距离方法,包括对象间的距离以及对象与各个类间的距离,并以符号数据的距离方法为基础建立聚类模型。Huang[5]最初提出的K-Modes聚类算法,通过使用一种简单的相异匹配度量对符号数据进行聚类。Wang等针对K-Modes算法初始化k簇时误差率较高和k近邻对大样本数据分类不准确的问题进行了改进,并提出了k-modes-KNN算法[6]。Sudipto等提出的ROCK[7]算法提出了"链"的概念,用于测量一对数据点之间的相似度。Sharma等[8]提出了一种通用的相似度量(Generalized Similarity Metric,GSM),将5种流行的度量变成一个单一的参数化公式,然后运用ROCK算法展示其有效性。Ding等[9]提出

到稿日期:2020-04-30　返修日期:2020-08-05　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61773247,61876103);山西省基础研究计划(201901D211192)
This work was supported by the National Natural Science Foundation of China (61773247,61876103) and Technology Research Development Projects of Shanxi (201901D211192).
通信作者:白亮(bailiang@sxu.edu.cn)

了基于随机序列迭代和属性权重的符号数据聚类算法，通过对原始数据随机排序来消除数据输入序列对聚类质量的影响，并且基于属性熵提出了区分属性权重的计算方法。

基于概率统计的聚类算法的主要思想在于采取基于统计学的方式来处理符号数据。假设数据集中存在 k 个类满足相应个数的概率分布，对象以一定的概率属于某个类，进而根据其隶属概率值的大小来产生不同的类簇。这种聚类算法一般采用概率统计中的贝叶斯定理和极大似然估计，代表性的算法有 COBWEB[10] 算法和 Ryszard 等提出的 CLUSTER/2[11] 算法。2019 年，Mahamadou 等[12] 提出了 ECM 算法，定义了一种新的相异性度量，并引入了交替最小化方法获得分区，首次将证据聚类用于符号数据聚类。Ralambondrainy[13] 提出了一种将符号属性转化为二值属性，然后利用 K-Means 进行聚类的算法。

基于信息熵的聚类算法是将信息熵的概念引入到符号数据聚类算法中，代表算法有 Barbará 等提出的 COOLCAT[14] 算法。Gowda[15-17] 提出了基于位置、跨度和内容的相似度和差异度函数，这种距离测度被应用于符号属性数据的层次聚类。Sharma[8] 提出了 TEC 算法，当属性被描述为幂律时，提出的 TEC 算法优于基于 Shannon 熵的聚类算法。Nguyen 等[18] 提出了一种扩展 k-means 算法对符号数据进行聚类，该算法可以自动测量各属性对聚类的贡献。Jia[19] 等采用一种基于预聚类的初始中心选取方法，改善了传统 K-Modes 算法忽略属性间差异且受初始中心点影响大的缺点。

虽然已经有许多符号数据聚类算法被提出并取得了一定的研究成果，但其仍存在两个问题：1）在处理大规模数据时，会产生高昂的计算成本且运行速度慢；2）虽然有的方法降低了计算的时间复杂度，但受参数值的影响，聚类结果的稳定性差，算法的鲁棒性不高。针对以上问题，本文提出了一种基于符号关系图的快速符号聚类算法。该方法根据符号数据有限的特点构建符号关系图，缩小了数据集的规模并降低了计算成本。我们将此算法分别应用于谱聚类、K-Modes 和层次聚类中，提高了聚类结果的鲁棒性。所提方法的框架如图 1 所示。

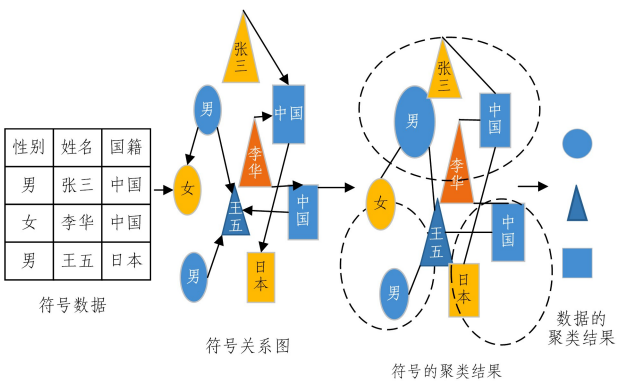

图 1 所提方法的框架图

Fig. 1 Framework of the proposed method

2 一种基于符号关系图的快速符号数据聚类算法

为了解决符号数据聚类在大规模数据集上的聚类问题，我们提出了基于符号关系图的快速符号聚类算法。该算法的核心步骤有：1）建立符号之间的关系图 G；2）分别使用谱聚类、k-means、层次聚类等具有代表性的图分割算法对符号关系图进行分割；3）根据不同符号对一个样本的划分找到其中概率最大的作为类划分。

2.1 符号关系图的构建

对于大规模数据，在进行符号数据聚类时由于其数据量大，会造成计算成本高、运行效率低等问题。因此，本文提出建立符号关系图的方法，通过此方法将数据集的规模缩小，从而降低时间复杂度。接下来，首先给出符号数据的相关定义。

定义 1 令 $X=\{x_1,x_2,\cdots,x_n\}$ 表示 n 个对象组成的一个数据集，其中，$x_i=\{x_{i1},x_{i2},\cdots,x_{im}\}$ 表示由 m 个属性 $A=\{a_1,a_2,\cdots,a_m\}$ 描述的第 i 个数据对象，a_j 为第 j 个属性，x_{ij} 表示对象 x_i 在第 j 个属性上的属性值。$D(a_j)=\{a_{j1},a_{j2},\cdots,a_{jn_j}\}$ 为 a_j 的值域，其中 a_{jh} 为属性 a_j 的第 h 个属性值，n_j 为 a_j 的属性值个数。在一个数据集 X 中，如果描述对象的每一个属性都是符号型变量，则称该数据集为符号型数据集，对象为符号型对象。

在每个属性上，符号数据不同对象属性值之间缺乏固有的几何性质，不能在空间中自然地定义对象之间的距离函数，因此符号数据之间的相似性计算变得非常困难。传统的符号数据聚类算法是直接在符号数据上进行聚类，为了提高符号聚类算法的运算效率，本文用符号聚类代替数据聚类。

定义 2 给定 $G=(V,W)$ 是无向加权图，我们将符号之间相似关系重组到 G 中，其中 $V=\bigcup_{j=1}^{m}D(a_j)$ 表示所有属性的符号集合，$W_{a_{jl},a_{qh}}$ 表示符号值 a_{jl} 和 a_{qh} 之间的相似性权值。

根据定义 2 可知，原始数据中的符号属性是有限、无序的，为了找到每个符号对应的类并度量符号属性之间的差异性，本文引入了相似性的概念，作为符号关系图的权值。

定义 3 给定 $\boldsymbol{S}\in R^{d\times d}$ 是相似性矩阵，通过计算不同属性两两之间的相似性，可以得到相似性矩阵 $\boldsymbol{S}$ 的公式如下：

$$s_{a_{jl},a_{qh}}=\frac{|\{x_{ij}=a_{jl}\}\cap\{x_{iq}=a_{qh}\}|}{\sqrt{|\{x_{ij}=a_{jl}\}||\{x_{iq}=a_{qh}\}|}} \tag{1}$$

其中，$d=\sum_{j=1}^{m}n_j$；$q\in\{1,2,\cdots,m\}$；$l,h\in\{1,2,\cdots,d\}$；$\{x_{ij}=a_{jl}\}$ 表示符号属性值为 a_{jl} 的数据对象集合。

定义 4 基于高斯核函数，定义符号关系图的权值 $W\in R^{d\times d}$ 如下：

$$W_{a_{jl},a_{qh}}=\exp\left(-\frac{1-\cos\theta_{a_{jl},a_{qh}}}{\delta}\right) \tag{2}$$

其中，$\cos\theta_{a_{jl},a_{qh}}=\frac{s_{a_{jl}\cdot}\cdot s_{a_{qh}\cdot}}{\sqrt{|s_{a_{jh}}||s_{a_{qh}}|}}$，$s_{a_{jh}}$ 表示 $\boldsymbol{S}$ 中的符号 a_{jh} 所在的行，δ 为核参数。权值 W 越小，表示两个符号属性越相似；反之则不相似。

例 1 表 1 所列为 UCI 中 lymphography 的部分数据组成的关于淋巴系统的符号属性的信息。图 2 为例 1 形成符号关系图的过程表示。其中不同的几何图形表示不同的属性，如圆形代表 lymphatics；不同的颜色代表同一属性中的不同符号，如 lymphatics 中的蓝色代表 displaced。由图 2 可以明显看出数据规模的缩小。

表 1 关于淋巴系统的符号属性的信息

Table 1 Information about symbolic attribute of lymphatic system

Object	lymphatics	Changes-in-node	Changes-in-stru	Special-forms	class
x_1	displaced	lac. central	faint	no	Ma-lymph
x_2	displaced	lacunar	drop-like	chalices	metastases
x_3	displaced	lac. central	stripped	vesicles	Ma-lymph
x_4	deformed	lacunar	faint	no	metastases
x_5	arched	lac. margin	diluted	vesicles	metastases
x_6	deformed	lac. margin	grainy	vesicles	metastases
x_7	arched	lac. margin	diluted	vesicles	metastases

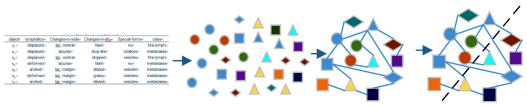

图 2 例 1 形成符号关系图的过程(电子版为彩色)

Fig. 2 Process of forming symbolic relation graph of example 1

表 1 中 $X=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7\}$,class 表示数据对象所属的类。首先将表 1 转化成矩阵的形式如下:

$$\begin{bmatrix}1&1&1&1&1\\1&2&2&2&2\\1&1&3&3&1\\2&2&1&1&2\\3&3&4&3&2\\2&3&5&3&2\\3&3&4&3&2\end{bmatrix}$$

通过计算可以得到类别数 $d=3+3+5+3+2=16$,根据定义 3 可以得到相似矩阵 $\boldsymbol{S}\in R^{16\times16}$,如 $s_{11}=\frac{|\{x_{11},x_{21},x_{31}\}\bigcap\{x_{11},x_{21},x_{31}\}|}{\sqrt{|\{x_{11},x_{21},x_{31}\}||\{x_{11},x_{21},x_{31}\}|}}=\frac{3}{\sqrt{3\times3}}=1$。

由得到的相似矩阵 $\boldsymbol{S}$,根据定义 4 得到符号关系图 $G\in R^{16\times16}$,上述过程如图 2 所示。

2.2 基于符号关系图的聚类

将 2.1 节得到的符号关系图视作一个压缩的 $d\times d$ 数据矩阵,可以通过传统的聚类算法如 k-means、谱聚类和层次聚类算法等将符号划分为 k 类,即 $C=\{C_1,C_2,\cdots,C_k\}$,其中 $C_l\in V$ 为第 l 个符号集。每一个符号的类标签 CL 定义如下:

$$CL(a_{jh})=l,\text{if } a_{jh}\in C_l,1\leqslant j\leqslant m,1\leqslant h\leqslant n_j \tag{2}$$

由于符号关系图能有效地反映每对符号属性之间的相似性关系,因此相比其对应的传统聚类算法,得到的聚类结果与标准的聚类结果更一致。我们将上述得到的符号数据聚类结果进行重标记,重标记后的数据定义如下:

$$x_{ij}=CL(a_{jh}),\text{if } x_{ij}=a_{jh},1\leqslant i\leqslant n,1\leqslant j\leqslant m \tag{3}$$

进一步,根据不同符号属性对一个样本的划分找到其中概率最大的作为类划分,得到最终的数据聚类结果,其中每个对象的类标签 L 的计算方法如下:

$$L(x_i)=\arg\max_l|\{x_{ij}=l,1\leqslant j\leqslant m\}| \tag{4}$$

如图 3 所示,符号数据聚类结果的第一类中黄色出现的概率最大,将第一类都重标记为黄色,得到最终的类划分。

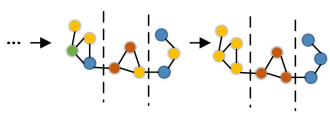

图 3 由样本的划分确定类的划分

Fig. 3 Division of the class determined by division of samples

2.3 算法的总流程和复杂度分析

综合上述过程,本文提出了基于符号关系图的快速符号数据聚类算法。在时间复杂度方面,由于计算符号之间的相似性所得到的相似性矩阵 $\boldsymbol{S}$ 为对称矩阵,因此得到相似性矩阵 $\boldsymbol{S}$ 的时间复杂度为 $O(d^2/2)$,得到符号关系图 G 的时间复杂度也为 $O(d^2/2)$,得到符号数据的聚类结果的时间复杂度为 $O(k)$,得到数据的聚类结果的时间复杂度为 $O(m)$。故该算法总的时间复杂度为 $O(d^2+k+m)$。所提算法的总流程如算法 1 所示。

算法 1 一种基于符号关系图的快速符号数据聚类算法

输入:数据集 X;聚类个数 k

输出:得到新的聚类标签 L

1. 统计数据集每一个属性存在的符号的类别数 n_j,并求和得到 d,即 $d=\sum_{j=1}^{m}n_j$;
2. 构建 $R^{d\times d}$ 的零矩阵;
3. 根据定义 3 求得各符号属性之间的相似性值 $s_{a_{jl},a_{qh}}$,得到相似图 $\mathbf{S}\in R^{d\times d}$;
4. 将相似图 $\mathbf{S}$ 作为新的符号数据,根据定义 4 计算得到符号关系图 $G\in R^{d\times d}$;
5. 使用传统聚类算法实现符号的聚类,得到类标签 CL;
6. 根据 CL 是 $R^{n\times d}$ 的属性标签矩阵,找到每行出现概率最大的类标签作为类划分,得到新的聚类标签 L。

3 实验分析

3.1 实验环境

本文中所有实验都是在 3.6 GHz CPU,8 GB 内存,Windows10 操作系统下完成的,提出的算法采用 Matlab2016b 实现。

3.2 实验数据

为了验证本文所提算法的有效性,我们从 UCI 机器学习库中选取了 13 个不同规模的数据集。数据集的详细描述如表 2 所列。

表 2 UCI 数据集描述

Table 2 Description of UCI data sets

Dataset	# Instance	# Features	# Classes
Soybean	47	35	4
Zoo	101	16	7
Lymphography	148	18	8
Dermatology	366	33	6
House-voting	435	16	2
Breast cancer	699	10	2
DNA	3190	60	3
Mushroom	8124	22	2
bc_pool_IS	2310	100	7
bc_pool_FCT	3780	100	7
bc_pool_MNIST	5000	100	10
bc_pool_ISOLET	7797	100	26
bc_pool_LR	20000	100	26

3.3 评价指标

本文使用标准互信息[20](Normalized Mutual Information,NMI)、调整兰德系数[21](Adjusted Rand Index,ARI)和精确度[22](precision,PE)来评价最终聚类质量。我们通过观察这些评价指标的值,可以很明显地看出划分与标准划分之间的差距,并比较不同算法完成聚类所用的时间。为了缩短在大规模数据集上聚类所耗费的时间,同时保证聚类结果的精度不降低,用 NMI,ARI 和 PE 来评价聚类结果的质量。

3.4 实验结果分析

我们将实验分为两部分。第一部分:将本文所提算法与 LCE[23], MCLA[24], COOLCAT[14], ROCK[7], Ng's K-Modes[25], San's K-Modes[26] 和 ACE[27] 算法进行比较分析。在比较过程中,本文设置聚类簇数 k 为数据集真实的簇个数,其他算法中所涉及的参数都与其原文保持一致,本文所涉及的高斯核参数 δ 设置为 0.0898。使用上述 7 种算法得到聚类结果后,利用 NMI,ARI 和 PE 这 3 种评价指标对上述聚类结果进行评价,并比较各算法的运行时间。实验结果如表 3—表 6 所列。第二部分:将所提算法分别应用在 K-means、谱聚类(SPC)和层次聚类(HC)上,将得到的聚类结果(用"after"表示)与 K-Modes、谱聚类(SPC)和层次聚类(HC)直接得到的聚类结果(用"before"表示)利用 NMI 和 ARI 指标进行评价,实验结果如表 7 和表 8 所列。

表 3 不同算法聚类结果的 *ARI* 值比较

Table 3 *ARI* comparison of clustering results of different algorithms

Datasets	LCE	MCLA	COOLCAT	ROCK	Ng's K-Modes	San's K-Modes	ACE	本文算法
Breast cancer	0.7178	0.5038	0.4513	0.6710	0.7240	0.6240	0.5040	**0.7335**
Soybean	0.7179	**0.8098**	0.1228	0.4326	0.6530	0.7228	0.6228	0.5543
Zoo	0.7404	0.2929	0.3834	0.4448	0.5834	0.6824	0.5734	**0.8534**
House-voting	0.4693	0.4840	0.1244	**0.9908**	0.3454	0.6204	0.7244	0.5487
Dermatology	0.0787	0.3137	0.0271	0.2546	0.2710	0.2710	0.2710	**0.4077**
DNA	0.1329	0.1323	0.3712	0.0474	0.3421	0.5326	0.0437	**0.5730**
Mushroom	0.1022	0.3637	0.0088	**0.6430**	0.0088	0.0088	0.0562	0.6241
Lymphography	0.0400	0.0630	0.1381	**0.4195**	0.1381	0.1381	0.0638	0.1002
bc_pool_IS	0.3491	0.5774	0.2739	0.1021	0.2739	0.5739	0.4323	**0.6051**
bc_pool_LR	0.4011	0.4070	0.1486	0.3908	0.1486	0.1486	0.4096	**0.4198**
bc_pool_FCT	0.1385	0.2209	0.0747	0.1023	0.0747	0.0747	**0.3722**	0.2369
bc_pool_MNIST	0.4852	0.5702	0.2124	0.5637	0.2246	0.2124	0.6072	**0.6534**
bc_pool_ISOLET	0.3575	0.4333	0.3321	0.3744	0.2453	0.2475	0.4539	**0.4627**

表 4 不同算法聚类结果的 *NMI* 值比较

Table 4 *NMI* comparison of clustering results of different algorithms

Datasets	LCE	MCLA	COOLCAT	ROCK	Ng's K-Modes	San's K-Modes	ACE	本文算法
Breast cancer	0.8265	0.5038	0.3802	0.8064	0.7130	0.5910	0.6590	**0.8356**
Soybean	0.6195	**0.8098**	0.7590	0.5284	0.3400	0.7590	0.7590	0.6854
Zoo	0.4590	0.2929	0.1488	0.5332	0.1488	0.1488	0.4880	**0.5524**
House-voting	0.5367	0.4840	0.1789	**0.9781**	0.1789	0.1789	0.1789	0.6638
Dermatology	0.0180	0.3137	0.0658	0.3815	0.1658	0.1658	0.1658	**0.4771**
DNA	0.1604	0.1323	**0.4756**	0.2546	0.0531	0.1296	0.1922	0.0385
Mushroom	0.1330	0.3137	0.1701	0.1021	0.1701	**0.6670**	0.4064	0.3119
Lymphography	0.0471	0.1630	0.1638	**0.6987**	0.1638	0.4380	0.1210	0.1740
bc_pool_IS	0.1992	0.4059	0.4660	0.4936	0.4660	0.2740	0.2807	**0.5136**
bc_pool_LR	0.1254	0.1439	0.2826	0.3381	0.2826	0.5037	0.1430	**0.6731**
bc_pool_FCT	0.0867	0.1209	0.5901	0.3406	0.6741	0.2826	0.4844	**0.6831**
bc_pool_MNIST	0.4355	0.4702	0.7410	0.3882	0.7901	0.5020	0.5326	**0.8954**
bc_pool_ISOLET	0.1394	0.2333	0.5648	0.4079	0.5648	0.2484	0.4406	**0.6032**

表 5 不同算法聚类结果的 *PE* 值比较

Table 5 *PE* comparison of clustering results of different algorithms

Datasets	LCE	MCLA	COOLCAT	ROCK	Ng's K-Modes	San's K-Modes	ACE	本文算法
Breast cancer	0.7548	0.8518	0.8188	0.8934	0.6501	0.6501	0.9472	**0.9573**
Soybean	0.6952	0.8324	**0.9063**	0.6259	0.3617	0.3617	0.7811	0.8241
Zoo	0.6196	0.6520	0.7065	0.6413	0.4059	0.4095	0.6853	**0.7108**
House-voting	0.7684	0.8871	0.7595	0.8076	0.6138	0.6138	0.8800	**0.9111**
Dermatology	0.6722	0.7804	0.7519	0.5490	0.3060	0.3060	0.7648	**0.8274**
DNA	0.4338	0.6880	0.6901	0.6273	0.3345	0.5180	0.4248	**0.6973**
Mushroom	0.4045	0.7390	0.6962	0.4396	0.5180	0.3864	0.4528	**0.7821**
Lymphography	0.5997	0.4234	0.4329	**0.7913**	0.3851	0.4608	0.6463	0.6581
bc_pool_IS	0.3854	0.7675	0.6541	0.6684	0.4291	0.4070	0.4167	**0.7759**
bc_pool_LR	0.2658	0.6397	0.6578	**0.9264**	0.0407	0.1300	0.2089	0.6200
bc_pool_FCT	0.2612	0.3370	0.3185	0.5747	0.1429	**0.3871**	0.1796	0.3365
bc_pool_MNIST	0.4759	0.6178	0.6338	0.5780	0.1000	0.3987	0.1467	**0.7675**
bc_pool_ISOLET	0.5128	0.3643	0.6584	0.5343	0.3850	0.3630	0.5975	**0.6708**

表 6 不同算法的聚类时间比较

Table 6 Comparison of clustering time of different algorithms

Datasets	LCE	MCLA	COOLCAT	ROCK	Ng's K-Modes	San's K-Modes	ACE	新算法
Breast cancer	17.164	4.005	0.656	131.002	0.625	0.721	905.942	**0.541**
Soybean	**0.225**	0.797	0.62	0.297	0.881	0.46	0.5	0.234
Zoo	5.563	0.375	0.32	1.187	0.26	0.31	3.094	**0.132**
House-voting	15.532	0.656	0.367	36.329	0.31	0.78	211.133	**0.209**
Dermatology	53.907	1.141	2.53	23.219	**0.402**	2.082	847.304	1.289
DNA	1109.4	**1.516**	43.139	3406	58.99	64.142	1587.9	8.371
Mushroom	10542	1.375	37.943	625.9	1.922	1.469	503.759	**1.053**
Lymphography	2.641	0.609	0.78	2.516	0.78	0.47	11.126	**0.23**
bc_pool_IS	24672	5.156	12.705	87253	**10.31**	56.431	555.9	**9.144**
bc_pool_LR	37962	359.73	2113.3	10763	19.21	79.372	7803	**12.4**
bc_pool_FCT	5362.5	129.9	39.397	9811	70.855	69.18	357.66	**19.5**
bc_pool_MNIST	1532	15.781	84.002	4536	33.75	43.27	307.04	**24.2**
bc_pool_ISOLET	2509.7	53.28	250.995	1052	18.658	25.609	964.8	**9.3**

表 7 3 种聚类算法应用本文方法前后聚类结果的 *ARI* 值比较

Table 7 Comparison of *ARI* values of three clustering algorithms before and after applying the proposed method

Datasets	K-Modes		SPC		HC	
	before	after	before	after	before	after
Breast cancer	0.5925	**0.7830**	**0.8389**	0.7335	0.6515	**0.9370**
Soybean	**0.8192**	0.5543	0.4436	**0.5543**	**0.8298**	0.5319
Zoo	**0.7716**	0.4874	0.4134	**0.5534**	**0.6436**	0.5842
House-voting	0.4318	0.4450	0.5487	**0.6068**	0.6161	**0.8138**
Dermatology	**0.5797**	0.3540	0.3264	**0.4077**	0.3224	**0.3579**
DNA	0.0530	**0.1186**	**0.6357**	0.0573	0.5191	**0.5191**
Mushroom	0.0470	**0.7750**	**0.6349**	0.2411	0.5224	**0.5643**
Lymphography	**0.2009**	0.0941	**0.4324**	0.1002	0.4257	**0.4257**
bc_pool_IS	0.6154	**0.6415**	0.5960	**0.6051**	**0.4835**	0.4489
bc_pool_LR	0.3270	**0.3263**	0.2943	**0.4198**	0.0747	**0.0920**
bc_pool_FCT	0.1849	**0.1976**	**0.2841**	0.2369	**0.1444**	0.1429
bc_pool_MNIST	0.5354	**0.6992**	0.4559	**0.6534**	0.1018	**0.1980**
bc_pool_ISOLET	0.7090	**0.8014**	0.4627	**0.5089**	0.1507	**0.3424**

表 8 3 种聚类算法应用新方法前后聚类结果的 *NMI* 值比较

Table 8 Comparison of *NMI* values of three clustering algorithms before and after applying the proposed method

Datasets	K-Modes		SPC		HC	
	before	after	before	after	before	after
Breast cancer	0.6833	**0.8742**	0.7335	**0.8356**	0.5200	**0.7630**
Soybean	**0.6553**	0.3054	**0.4468**	**0.3054**	**0.6738**	0.1228
Zoo	**0.6707**	0.2694	0.4158	**0.5024**	**0.4381**	0.3834
House-voting	0.4905	**0.5632**	0.6138	**0.6638**	0.0037	**0.3922**
Dermatology	**0.4149**	0.2063	0.1477	**0.3197**	0.0045	**0.2170**
DNA	0.0272	**0.0341**	**0.5188**	0.0385	**0.2140**	0.1600
Mushroom	0.0604	**0.5540**	**0.6355**	0.3119	**0.7370**	0.0088
Lymphography	0.0505	**0.1228**	0.0174	**0.4257**	0.0270	**0.1381**
bc_pool_IS	0.4607	**0.5104**	0.4136	**0.5286**	0.2520	**0.2624**
bc_pool_LR	0.1205	**0.1823**	0.1573	**0.1759**	**0.1900**	0.0056
bc_pool_FCT	0.1024	**0.1042**	**0.2839**	0.1306	0.4510	**0.8930**
bc_pool_MNIST	**0.3972**	0.3098	0.4522	**0.4895**	0.1710	**0.4520**
bc_pool_ISOLET	0.4709	**0.5690**	0.2603	**0.4200**	0.3440	**0.3503**

表 3—表 5 列出了不同算法在 13 个真实数据集上的聚类效果,结合表 6 所列的运行时间可以发现,在大多数数据集上,如 Breast cancer,Zoo 和 bc_pool_IS 数据集,本文所提算法的聚类效果更好且其聚类运行时间相比其他符号数据聚类算法大幅缩短。本文算法虽然在一些数据集上的 *ARI*,*NMI* 和 *PE* 值不是最优,如 Mushroom 数据集上的 *ARI* 值不是最优,但是与该数据集上的最优值接近。

表 7 和表 8 列出了将符号关系图应用于这 3 种聚类算法对聚类结果的影响。在多数数据集上,将构建的符号关系图应用于传统算法可以产生更好的聚类结果,如在 Breast cancer 上,聚类效果明显变好。虽然在某些情况下聚类精度达不到最优,但证明了本文所提算法的优势。综上所述,本文提出的基于符号关系图的快速符号数据聚类算法与其他算法相比聚类结果更好,运行时间也大幅缩短。

结束语 本文针对传统符号数据聚类算法计算成本高昂的问题,提出了一种基于符号关系图的快速符号数据聚类算法。该算法通过构建符号关系图来代替原数据,获得符号之间的相似性关系,并通过图分割方法得到符号数据的聚类结果,提高了符号数据聚类的鲁棒性。在不同数据集上的实验结果表明,该算法在 NMI,ARI 和 PE 评价指标上的效果都有所提高,且运行速度快,证明了该算法的有效性。

参 考 文 献

[1] ZHOU Z H. Machine learning and its applications[M]. Beijing: Tsinghua University Press, 2009: 15-20.

[2] ZHONG X, MA S P, ZHANG B, et al. A survey of data mining [J]. Pattern Recognition and Artificial Intelligence, 2001, 3(1): 50-57.

[3] JAIN A K, MURTY M N, FLYNN P J. Data clustering: a review[J]. Acm Computing Surveys, 1999, 31(3): 264-323.

[4] EL-SONBATY Y, ISMAIL M A. Fuzzy clustering for symbolic data[J]. IEEE Transactions on Fuzzy Systems, 1998, 6(2): 195-204.

[5] HUANG Z. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values[J]. Data Mining and Knowledge Discovery, 1998, 2(3): 283-304.

[6] WANG Z H, LIU S T, LUO Q. KNN Classification Algorithm based on improved K-modes clustering[J]. Computer Engineering and Design, 2019(8): 2228-2234.

[7] SUDIPTO G, RAJEEV R, KYUSEOK S. Rock: A robust clustering algorithm for categorical attributes[J]. Information Systems, 2005(5): 345-366.

[8] SHARMA S, SINGH M. Generalized similarity measure for categorical data clustering[C]// 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE Press, 2016: 21-24.

[9] DING X, TAN J, WANG M. A categorical data clustering algo-

rithm and its efficient parallel implementation[C]// 2016 5th International Conference on Computer Science and Network Technology(ICCSNT). IEEE Press, 2017: 224-228.

[10] FISHE R, DOUGLAS H. Knowledge acquisitionvia incremental conceptual clustering[J]. Machine Learning, 1987, 2(2): 139-172.

[11] MICHALSKI R S, STEPP R E. Automated Construction of Classifications Conceptual Clustering Versus Numerical Taxonomy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1983, 5(4): 396-410.

[12] MAHAMADOU A J D, ANTOINE V, CHRISTIE G J, et al. Evidential clustering for categorical data[C]// 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). IEEE Press, 2019: 1-6.

[13] RALAMBONDRAINY H. A conceptual version of theK-means algorithm[J]. Pattern Recognition Letters, 1995, 16(11): 1147-1157.

[14] BARBARÁ D, LI Y, JULIA C. COOLCAT: an entropy-based algorithm for categorical clustering[C]// International Conference on Information and Knowledge Management. 2002: 582-589.

[15] GOWDA K C, RAVI T V. Divisive clustering of symbolic objects using the concepts of both similarity and dissimilarity[J]. Pattern Recognition, 1995, 28(8): 1277-1282.

[16] GOWDA K C, DIDAY E. Symbolic clustering using a new dissimilarity measure[M]. Elsevier Science Inc. 1991.

[17] DINESH M S, GOWDA K C, NAGABHUSHAN P. Unsupervised classification for remotely sensed data using fuzzy set theory[C]// Geoscience and Remote Sensing(IGARSS '97). IEEE Press, 1997.

[18] NGUYEN T H T, HUYNH V N. A k-Means-Like Algorithm for Clustering Categorical Data Using an Information Theoretic —Based Dissimilarity Measure[C]// International Symposium on Foundations of Information & Knowledge Systems. Springer-Verlag New York, 2016.

[19] JIA B, LIANG Y, SU H. An improvedK-Modesclustering algorithm[J]. Software Guide, 2019, 18(6): 60-64.

[20] MCDAID A F, GREENE D, HURLEY N. Normalized Mutual Information to evaluate overlapping community finding algorithms[J]. arXiv: 1110. 2515.

[21] WARRENS M J. On the Equivalence of Cohen's Kappa and the Hubert-Arabie Adjusted Rand Index[J]. Journal of Classification, 2008, 25(2): 177-183.

[22] YANG Y M. An Evaluation of Statistical Approaches to Text Categorization[J]. Proc. Amia. Annu. Fall. Symp., 1999, 1(1/2): 358-362.

[23] IAMON N, BOONGOEN T, GARRETT S, et al. A Link-Based Cluster Ensemble Approach for Categorical Data Clustering[J]. IEEE Transactions on Knowledge andData Engineering, 2012, 24(3): 413-425.

[24] STREHLA, GHOSH J. Cluster Ensembles-A Knowledge Reuse Framework for Combining Multiple Partitions[J]. Journal of Machine Learning Research, 2003, 3(3): 583-617.

[25] MICHAEL K, LI J J, HUANG Z X, et al. On the impact of dissimilarity measure in k-modes clustering algorithm[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(3): 503-507.

[26] SAN O, HUYNH V, NAKAMORI Y. An alternative extension of the k-means algorithm for clustering categorical data[J]. Pattern Recognition, 2004, 14(2): 241-247.

[27] CHEN K, LIU L. "Best K": critical clustering structures in categorical datasets[J]. Knowledge and Information Systems, 2009, 20(1): 1-33.

ZHANG Yan-jin, born in 1995, postgraduate. Her main research interests include categorical data clustering.

BAI Liang, born in 1982, Ph.D, professor, is a member of China Computer Federation. His main research interests include cluster analysis and so on.