符号数据的无监督学习:一种空间变换方法

王建新[1,3] 钱宇华[1,2,3]
(山西大学计算机与信息技术学院 太原 030006)[1] (山西大学智能信息处理研究所 太原 030006)[2]
(计算智能与中文信息处理教育部重点实验室 太原 030006)[3]

摘 要 近年来符号数据的无监督学习在模式识别、机器学习、数据挖掘和知识发现等诸多领域扮演着越来越重要的角色。然而现有的针对符号数据的聚类算法(经典的K-modes系列算法等),相比数值型数据的聚类算法,在性能方面仍然有很大的提升空间。其根本原因在于符号数据缺乏类似数值数据那样清晰的空间结构。为了能够有效地发掘符号数据内在的空间结构,采用了一种全新的数据表示方案:空间变换方法。该方法将符号数据映射到相应的由原来的属性组成的新的维度的欧氏空间中。在这一框架的基础上,为了找到符号数据更有代表性的模式,结合Carreira-Perpiñán提出的K-modes算法进行无监督学习。在9个常用的UCI符号数据集上进行了测试,与传统的符号数据聚类算法进行了实验比较,结果表明几乎在所有的数据集上提出的方法都是更加有效的。

关键词 符号数据,数据表示方案,空间变换
中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.1.021

Unsupervised Learning from Categorical Data: A Space Transformation Approach

WANG Jian-xin[1,3] QIAN Yu-hua[1,2,3]
(School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China)[1]
(Intelligence Information Processing Lab, Shanxi University, Taiyuan 030006, China)[2]
(Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Taiyuan 030006, China)[3]

Abstract The unsupervised learning method of categorical data plays a more and more important role in such areas as pattern recognition, machine learning, data mining and knowledge discovery in the recent years. Nevertheless, in view of many existing clustering algorithms for categorical data (the classical k-modes algorithm and so on), there is still a large room for improving their clustering performance in comparison with the performance of clustering algorithms for numeric data. This may arise from the fact that categorical data lack a clear space structure as that of numeric data. To effectively discover the space structure inherent in a set of categorical objects, we adopted a novel data representation scheme: a space transformation approach, which maps a set of categorical objects into a corresponding Euclidean space with the new dimensions constructed by each of the original features. Based on the new general framework for categorical clustering, we employed the Carreira-Perpiñán's K-modes algorithm for clustering to find more representative modes. The performance of the new proposed method was tested on the nine frequently-used categorical data sets downloaded from the UCI. Comparisons with the traditional clustering algorithms for categorical data illustrate the effectiveness of the new method on almost all data sets.

Keywords Categorical data, Data representation scheme, Space transformation

1 引言

聚类分析是数据挖掘中的一个重要的研究领域。由于聚类分析不对数据作任何统计假设,也常被称为一种无监督的学习方法,在文本挖掘、粒计算、信息检索、生物信息学、Web数据挖掘、客户分析和科学数据探索等很多方面有着非常广泛的应用[8-11],其研究也受到国内外学者的普遍关注。

聚类分析的作用是发现一组数据对象之间内在的组结构信息。而数据对象的组结构信息又与其类型紧密相关。由于数据对象类型的多样性,在实际数据的预处理过程中,通常的方法是将音频、视频、文本等非结构化的数据转换为数值型或符号型的结构化数据[12-15]。因此,目前存在的聚类算法主要是针对数值型数据或者符号型数据。

聚类分析的目标是将一组无标记的数据对象划分成有意义的类,使得在同一类中的对象之间具有较高的相似度,不在同一类的对象之间差异性较大。因此,对象之间的相似性或者差异性的度量是整个算法的核心步骤。

对于数值型数据的聚类,针对不同的数据分布,已经产生

到稿日期:2015-03-18 返修日期:2015-06-13 本文受国家优秀青年基金项目(61322211),教育部新世纪人才支持计划(NCET-12-1031),教育部博士点专项科研基金项目(20121401110013),山西省青年学术带头人(20120301)资助。
王建新(1990-),男,硕士,主要研究方向为大数据机器学习,E-mail:jianxinwang@email.sxu.edu.cn;钱宇华(1976-),男,教授,博士生导师,主要研究方向为计算智能、数据挖掘与知识发现,E-mail:jinchengqyh@sxu.edu.cn。

了许多成功的算法。其中最为经典的算法是 K-means 类型算法[16-18]。在 K-means 类型算法中，对象由一些数值属性表示，所有的对象都可以看作欧氏空间中的向量，进而通过欧氏距离或者余弦距离等来度量对象之间的相似性。K-means 类型算法由于简单、易实现，且具有线性的时间复杂度，因此在理论研究和实际应用中已经取得了丰硕的成果。所以，对符号型数据的聚类已经成为了一个热点课题。

由于符号型数据的对象由非数值型的数据构成，在算法中如何学习符号数据对象之间的差异性是一个非常值得研究的难题。一般地，在算法中采用 0-1 距离或者它的扩展版本来度量符号数据对象之间的差异性。为了进一步发现符号数据内在的组结构，已经提出了许多算法和理论[19-21]。其中最有代表性的算法是由 Huang 在文献[3]提出的 K-modes 类型算法。本文为了不混淆概念，称其为 Huang's K-modes 类型算法。Huang's K-modes 类型算法采用了与 K-means 类型算法相同的范式，克服了 K-means 类型算法只对数值数据有效的限制，能够聚类较大的符号数据集。

尽管 Huang's K-modes 类型算法可以有效地聚类符号型数据，但是与数值型数据聚类算法相比，其在性能上仍然有很大的提升空间。原因在于：(1)差异性度量：0-1 距离并不能很好地揭示符号数据内在的组结构；(2)类中心更替：K-means 类型算法基于同一类内的所有对象各个属性下的均值来更新类中心，而 Huang's K-modes 类型算法基于属性频率的方式来更新类中心。从人类认知的角度可知，Huang's K-modes 类型算法得到的类中心并不是真正的模式。此外，不妨更为深远地思考一下，符号型数据和数值型数据能否内在地统一于一个聚类算法中。

基于上面的观点，为了有效地发掘符号数据内在的组结构，找到类中真正有代表性的模式，本文基于空间变换的方法，对符号数据进行了新的表示，然后用 Carreira-Perpiñán's K-modes 算法[2]进行无监督的学习(以下简称为 SBC_K-modes 算法)。与传统的符号数据聚类算法进行的实验比较表明新的方法更加有效。

2 SBC_K-modes 算法

这一节主要介绍 SBC_K-modes 算法的基本原理：符号数据的空间变换和 Carreira-Perpiñán's K-modes 算法。

2.1 符号数据的空间变换

为了解决符号数据分析问题，探索符号数据内在的空间结构，钱宇华等[1]提出了一种基于新的符号数据的表示方案，其具体的描述如下。

给定一个符号数据集 $U=\{x_1,x_2,\cdots,x_n\}$ 是 n 个对象的集合，$A=\{a_1,a_2,\cdots,a_m\}$ 是 m 个属性的集合，相应的权重 $W=\{w_1,w_2,\cdots,w_m\}$，$a_i(x_j)$ 是对象 x_j 在属性 a_i 下的符号属性值。传统的符号数据的表示如表 1 所列。区别于旧的数据表示方法，新的数据的表示中，首先需要计算每两个对象彼此之间相等的概率，形式化的定义如式(1)：

$$p(x_i=x_j)=\sum_{k=1}^{m}w_k\theta_k(x_i,x_j)/\sum_{k=1}^{m}w_i,\ (p(x_i=x_j)\text{记为 }p_{ij}) \tag{1}$$

其中，$\theta_k(x_i,x_j)=\begin{cases}1, & a_k(x_i)=a_k(x_j)\\ 0, & a_k(x_i)\neq a_k(x_j)\end{cases}$

表 1 传统的符号数据的表示

	a_1	a_2	…	a_m
x_1	$a_1(x_1)$	$a_2(x_1)$	…	$a_m(x_1)$
x_2	$a_1(x_2)$	$a_2(x_2)$	…	$a_m(x_2)$
⋮	⋮	⋮	⋱	⋮
x_n	$a_1(x_n)$	$a_2(x_n)$	…	$a_m(x_n)$

通过对符号数据对象的再表示，我们可以获得一个由这些对象之间的相似性的概率值构成的一个欧氏空间结构，使得原来的对象由一组新的维度$\{c_i=x_i,1\leqslant i\leqslant n\}$描述，具体如表 2 所列。

表 2 新的符号数据的表示

	$x_1(c_1)$	$x_2(c_2)$	…	$x_n(c_n)$
x_1	p_{11}	p_{12}	…	p_{1n}
x_2	p_{21}	p_{22}	…	p_{2n}
⋮	⋮	⋮	⋱	⋮
x_n	p_{n1}	p_{n2}	…	p_{nn}

性质 1[1] 用(U,C)表示由原来的符号数据集(U,A)得到的欧氏空间，x_C 表示(U,C)中的向量，x_A 表示(U,A)中的向量。易知具有如下的一些性质：

1) $\forall x_C$ 都在(U,C)的第一象限；

2)如果 $x_A=y_A$，那么 $x_C=y_C$；

3) $p_{ij}=1-\frac{1}{m}d(x_{iA},x_{jA})$；

4)$0\leqslant\langle x_C,y_C\rangle\leqslant 90°$。

定理 1[1] (U,A)表示一个符号数据集，$x_i,x_j,x_k\in U$，(U,C)表示通过(U,A)映射得到的欧氏空间。如果 $d(x_{iA},x_{jA})=d(x_{iA},x_{kA})$，那么 $D(x_{iC},x_{jC})=D(x_{iC},x_{kC})$不一定成立。

定理 1 暗示这样一个现象：当对象 y 和 z 在原始的属性集下不可区分时，它们在通过新的数据表示诱导出的欧氏空间中是可以区别开的，即新的数据表示可以提供更加细粒度的属性区分能力。

总的来说，符号数据可以通过空间变换的方法，使得所有原始属性下的属性值和相应的权重在不失信息的情况下应用到新的数据表示中。而且在得到的欧氏空间中，可以消除由于经典的符号数据对象之间没有清晰的结构带来的限制，从而能够利用许多现存的理论和方法来进一步挖掘符号数据集中隐含的知识。这也是本文的核心思想。

2.2 Carreira-Perpiñán's K-modes 算法

由于目前存在的较为经典的算法都没有返回精确的 K 个有意义的模式，Carreira-Perpiñán 等[2]提出了一种新的 K-modes 算法。这里简短地介绍一下 Carreira-Perpiñán's K-modes 算法，其最大化目标函数为

$$\max_{Z,C}\sum_{n=1}^{N}\sum_{k=1}^{K}Z_{nk}G\left(\left\|\frac{x_n-c_k}{\sigma}\right\|^2\right) \tag{2}$$

约束条件为：$z_{nk}\in\{0,1\}$，$\sum_{k=1}^{K}Z_{nk}=1,n=1,\cdots,N$，其中 N 表示对象个数，K 表示类个数，$z_{ij}=1$ 表示第 i 个对象属于第 j 类，Z 是样本数据的划分矩阵，C 则表示每次迭代中 K 个类中心的集合；$G(t)\propto e^{-t/2}$ 和参数 σ 可以参考文献[2,22]中的核密度估计。K-modes 算法中需要手动输入的参数有 2 个，分别是类个数 K 和核密度估计中的带宽值 σ。而关于参数 σ 和 K 对聚类性能的影响也将在 4.3 节参数选择部分给出。

这里之所以采用新的 K-modes 算法，是希望在空间变换后的符号数据的空间结构中找到更有意义的、更有代表性的

模式；另外，新的 K-mdoes 算法的时间复杂度（$O(KND)$）较低，适合处理转换后的空间结构矩阵。

3 SBC_K-modes 算法描述

基于符号数据的空间变换，结合 Carreira-Perpiñán's K-modes 算法，本文提出了一种新的基于空间变换的符号数据的无监督学习方法 SBC_K-modes 算法，其具体的描述如下。

算法 1 SBC_K-modes 算法

输入：原始符号数据集 $U=\{x_{1A},x_{2A},\cdots,x_{nA}\}$，属性集 $A=\{a_1,a_2,\cdots,a_m\}$ 以及相应的权重 $W=\{w_1,w_2,\cdots,w_m\}$

输出：K 个类的划分结果

(1)通过空间变换方法，生成新的符号数据的表示。

通过式(1)计算空间结构

$(U,C)=\{p_{ij},1\leqslant i,j\leqslant n\}$；产生相应的 n 个对象 $\{x_{iC}:1\leqslant n\}$，其中：$x_{iC}=(p_{i1},p_{i2},\cdots,p_{in})$。

(2)在空间结构矩阵中，随机地选择 K 个样本作为初始的类中心：$C^p=\{c_1^p,c_2^p,\cdots,c_K^p\}$，$p=0$（为了减少初始类中心选择带来的对算法性能上的影响，本文实验中的所有算法均采用相同的初始化类中心的方法；K 为真实的实验数据集的类个数，具体的设置参见 4.3 节参数选择部分）。

(3)对于固定的 C(即所有样本对象分配的过程)，Z 的优化是对每个样本点 $x_i(i=1,2,\cdots,n)$ 单独执行的。

For i=1 to n

x_i 被分配到类 l 中，当 $l=\arg\max_k G(\|(x_i-c_k)/\sigma\|^2)=\arg\min_k\|(x_n-c_k)\|$，即每个样本点 x_i 被分配到距其欧氏距离最近的类中心 c_k 所在的第 l 类中。

End For

(4)对于固定的 Z(即类中心更替的过程)，C 的优化是每一个类单独进行的，是对每一个类中心 c_k 单独的无约束的最大化问题，形式化的表示为：

$$L(c_k)=\sum_{n=1}^{N}Z_{nk}G(\|(x_i-c_k)/\sigma\|^2),$$

这样就产生了新的类中心的集合：

$C^{p+1}=\{c_1^{p+1},c_2^{p+1},\cdots,c_K^{p+1}\}$。

(5)如果 $C^{p+1}=C^p$，SBC_K-modes 算法停止；否则 $p=p+1$，转到步骤(3)。

4 实验结果与分析

在符号数据集上的实验证明了 SBC_K-modes 算法的收敛性、高效性和鲁棒性。表 3 列出了实验中用到的 9 个常用的符号数据集。

表 3 UCI 上的 9 个公用数据集的描述

	Data sets	Number of sample	Number of features	classes	Class distribution
1	Soybean-small	47	35	4	{10,10,10,17}
2	Vote	435	16	2	{168,267}
3	Breast cancer	699	9	2	{458,241}
4	Promoter	106	57	2	{53,53}
5	Fitting contact lenses	24	4	3	{4,5,15}
6	Hayes-Roth-Hayes-Roth	132	4	3	{51,51,30}
7	Balloon	20	4	2	{8,12}
8	Lymphography Domain	142	18	2	{81,61}
9	Space shuttle autolanding	15	6	2	{6,9}

4.1 收敛性分析

本小节主要研究和检测了 SBC_K-modes 算法的收敛性。为了避免结果偶然性带来的影响，我们在 9 个数据集上分别实验了 100 次。每次实验中，初始的类中心都是随机产生的。

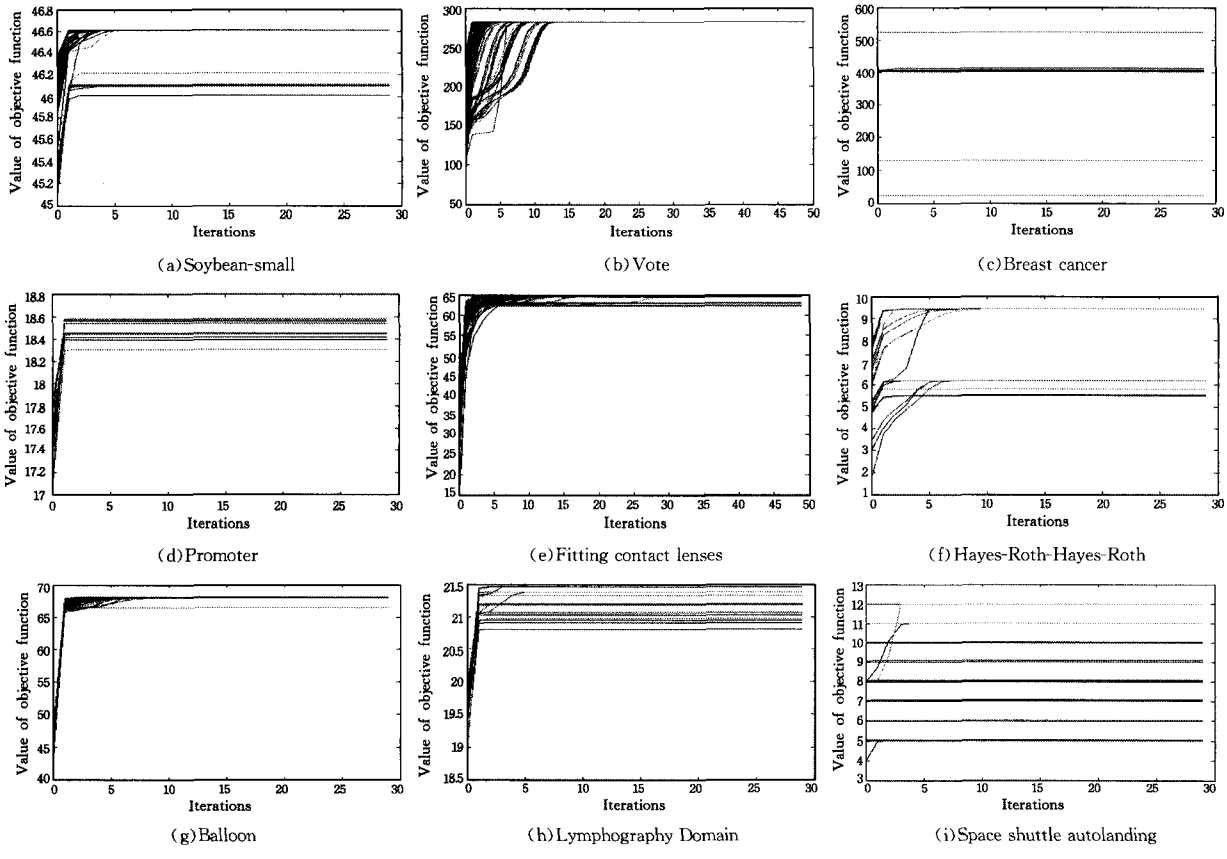

图 1 不同初始类中心条件下目标函数值随着迭代次数的变化

9 个数据集上，不同初始类中心条件下目标函数值随着迭代次数增加的变化趋势如图 1 所示。

从图 1 中的每幅子图可以看出，随着迭代次数的增加，每条曲线上目标函数的值增加得越来越快，在有限次的迭代后达到它的最大值。这一性质说明本文提出的 SBC_K-modes 算法的目标函数可以很快达到它的最大值。此外，每一条曲线在有限次的迭代之后目标函数的值不再变化。这说明对于每一个符号数据集，SBC_K-modes 算法可以在有限次的迭代后收敛。从这两个性质可以得出该算法是很快收敛的，可以有效地用在符号数据的聚类上。

4.2 聚类性能分析

对于符号数据的聚类，目前已经提出了许多 Huang's K-modes 类型的算法。为了验证本文提出的 SBC_K-modes 算法的聚类性能，这里将其与其他 4 种 Huang's K-modes 类型的算法进行比较，分别是经典 K-modes 算法[3]、Chan's 算法[4]、Mkm_nof 算法[5]和 Mkm_ndm 算法[5]。

为了更好地说明算法的有效性，我们采用了两个广泛使用的聚类算法的性能评价指标——准确率(AC)和调整兰德指数(ARI)。AC[6]定义如下：

$$AC=\sum_{i=1}^{k}\frac{\max\{n_{ij}:j\leqslant k'\}}{n}$$

ARI[7]定义如下：

$$ARI=\frac{\binom{n}{2}\sum_{ij}\binom{n_{ij}}{2}-\sum_{i}\binom{b_i}{2}\sum_{j}\binom{d_j}{2}}{\frac{1}{2}\binom{n}{2}\left(\sum_{i}\binom{b_i}{2}\sum_{j}\binom{d_j}{2}\right)-\sum_{i}\binom{b_i}{2}\sum_{j}\binom{d_j}{2}}$$

具体的各个参数的意义和使用方法参考文献[1]。聚类结果与样本真实的类分布越相近，评价指标的值就越高。因为 Huang's K-modes 型的算法的性能与初始类中心有很大的关系，所以本文从统计的角度观察它们的性能。为了解决这个问题，本文在 9 个数据集上分别运行 5 个算法 100 次，每次的初始类中心都是随机产生的，然后计算 AC 和 ARI 100 次结果的平均值和方差。实验结果如表 4 和表 5 所列(在每一数据集上的最高的 AC 和 ARI 值由下划线标出，$\sigma=0.5$)。

表 4　9 个数据集上 5 个算法的 AC 值($\sigma=0.5$)

Data sets	SBC_K-modes	K-modes	Chan	Mkm_nof	Mkm_ndm
Soybean-small	0.9766±0.0044	0.9185±0.0087	0.8353±0.0096	0.7626±0.0051	0.9483±0.0078
Vote	0.8759±0.0000	0.8604±0.0001	0.8094±0.0088	0.8715±0.0027	0.8715±0.0027
Breast cancer	0.9914±0.0001	0.8608±0.0112	0.7717±0.0022	0.7697±0.0000	0.9464±0.0000
Promoter	0.8958±0.0045	0.6335±0.0057	0.7865±0.0028	0.7500±0.0026	0.7043±0.0121
Fitting contact lenses	0.7196±0.0037	0.6417±0.0013	0.6588±0.0016	0.6813±0.0026	0.6587±0.0020
Hayes-Roth-Hayes-Roth	0.4563±0.0019	0.4256±0.0004	0.4782±0.0020	0.4550±0.0019	0.4329±0.0026
Balloon	0.7880±0.0132	0.6910±0.0083	0.7045±0.0080	0.7310±0.0070	0.6710±0.0071
Lymphography Domain	0.7392±0.0065	0.6252±0.0030	0.5808±0.0007	0.5801±0.0000	0.6589±0.0046
Space shuttle autolanding	0.6853±0.0109	0.6240±0.0010	0.6293±0.0011	0.6140±0.0007	0.6367±0.0040
Average value	0.7920	0.6979	0.6949	0.6906	0.7254

表 5　9 个数据集上 5 个算法的 ARI 值($\sigma=0.5$)

Data sets	SBC_K-modes	K-modes	Chan	Mkm_nof	Mkm_ndm
Soybean-small	0.9699±0.0073	0.8247±0.0288	0.6956±0.0214	0.6330±0.0038	0.9111±0.0279
Vote	0.5639±0.0000	0.5187±0.0005	0.4123±0.0455	0.5599±0.0172	0.5599±0.0127
Breast cancer	0.9661±0.0014	0.5395±0.0792	0.2636±0.0126	0.2487±0.0000	0.7952±0.0000
Promoter	0.6411±0.0293	0.0859±0.0065	0.3334±0.0061	0.2545±0.0060	0.2084±0.0310
Fitting contact lenses	0.2057±0.0250	0.0169±0.0088	0.1009±0.0232	0.1232±0.0198	0.0621±0.0118
Hayes-Roth-Hayes-Roth	0.0238±0.0014	−0.0018±0.0001	0.0377±0.0010	0.0240±0.0007	0.0071±0.0011
Balloon	0.3647±0.0153	0.1356±0.0247	0.1536±0.0253	0.1987±0.0234	0.0981±0.0209
Lymphography Domain	0.2490±0.0214	0.0627±0.0053	0.0106±0.0011	0.0080±0.0000	0.1109±0.0106
Space shuttle autolanding	0.1215±0.0456	−0.0050±0.0017	0.0064±0.0020	−0.0155±0.0014	0.0262±0.0110
Average value	0.4562	0.2419	0.2238	0.2261	0.3089

从表 4 可以看出，我们提出的 SBC_K-modes 算法比另外 4 种算法在统计学意义上的性能好得多。对于 9 个符号数据集，本算法在 8 个数据集上的结果是最优的，而在另外一个 Hayes-Roth-Hayes-Roth 数据集上的结果也与最优的结果相近。值得注意的是，SBC_K-modes 算法在绝大部分数据集上都显著地提高了 AC 和 ARI 的值。比如说，对于 AC 值，SBC_K-modes算法在数据集 Promoter上的平均 AC 值提高了 89.58%−78.65%=10.93%，在数据集 Lymphography 上的平均 AC 值提高了 73.92%−65.89%=8.03%；对于 ARI 值，SBC_K-modes 算法在数据集 Promoter 上的平均 ARI 值提高了 64.11%−33.34%=30.77%，在数据集 Balloon 上的平均 ARI 值提高了 36.47%−19.87%=16.60%。

从上述的实验分析可以看出，SBC_K-modes 算法是收敛的，而且相对于其他 4 种算法具有更优的聚类性能。

4.3 参数选择

本小节主要研究 SBC_K-modes 算法中参数对聚类性能指标 AC 和 ARI 的影响。

本文提出的 SBC_K-modes 算法有两个参数：聚类个数 K 和带宽 σ。对于 K 的取值，将其设置为与真实类个数相同，具体的设置参考表 3 中对于各个数据集的描述。对于带宽 σ，本文仅选取部分值($\sigma=0.5,1,1.5,2.5$)进行实验。图 2 和图 3 给出了在 9 个数据集上性能指标 AC 和 ARI 与带宽 σ 的关系。

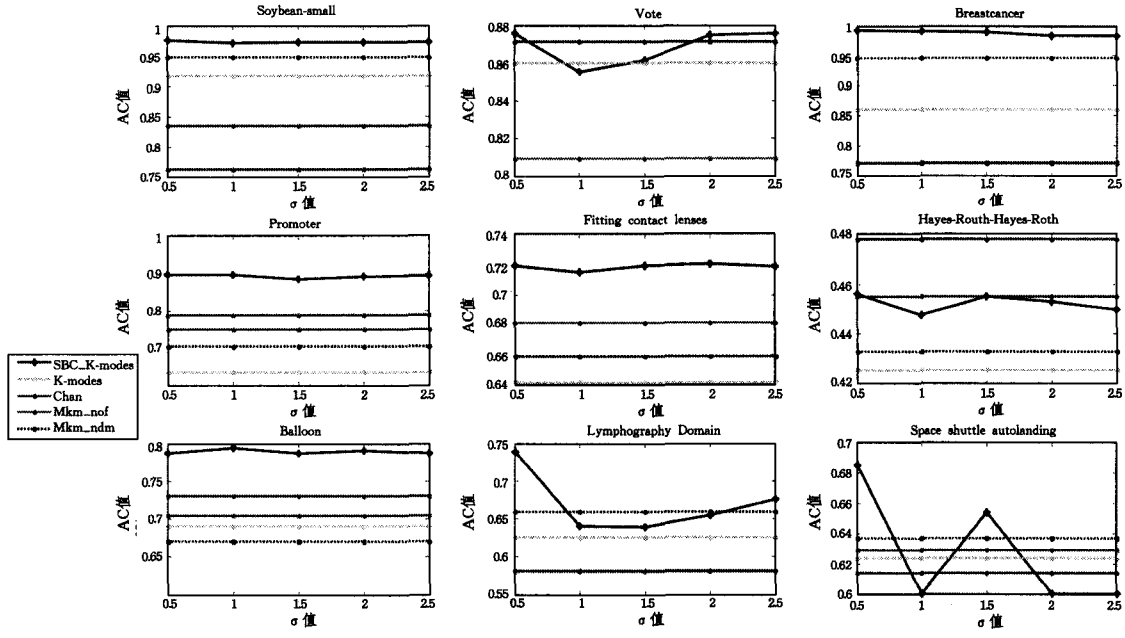

图 2　9 个数据集上不同带宽下的 AC 值

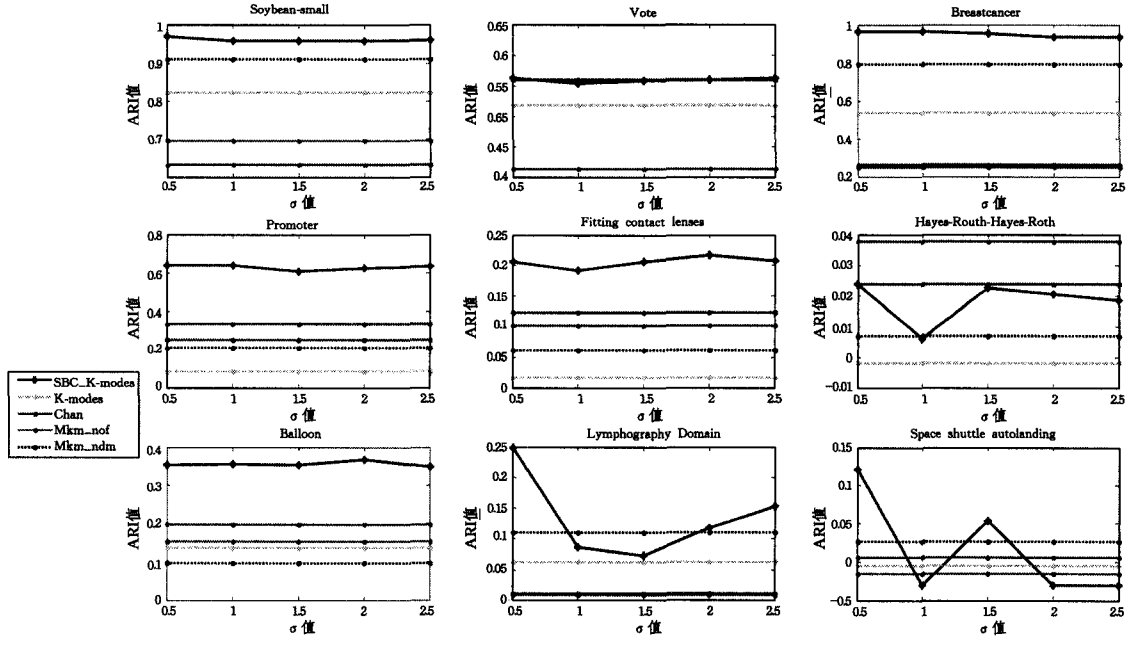

图 3　9 个数据集上不同带宽下的 ARI 值

所提算法在数据集 Soybean-small、Promoter、Breastcancer、Fitting contact lenses 和 Balloon 上，AC 和 ARI 的结果都是最优的，更为值得注意的是，在数据集 Promoter 和 Balloon 上有显著地提高；对于数据集 Vote，AC 和 ARI 的结果与最优的结果相近；对于数据集 Lymphography Domain 和 Space shuttle autolanding，部分带宽 σ 的取值使得 AC 和 ARI 的结果是最优的；而对于数据集 Hayes-Roth-Hayes-Roth，带宽 σ 的取值使得性能指标并不理想。可以看出，SBC_ K-modes 算法有较高的鲁棒性。

结束语　本文提出了一种新的基于空间变换的符号数据的聚类方法 SBC_K-modes 算法。该算法将所有符号对象映射到新的维度的欧氏空间中，有效地发掘了符号数据内在的空间结构。其不仅提供了处理符号型和数值型混合数据语义上统一的框架，而且在通过空间变换得到的空间结构中，获得了符号数据对象之间更加细粒度的差异，并通过 Carreira-Perpiñán's K-modes 算法找到了类中真正的模式。通过在常用符号数据集上的实验验证了算法的收敛性和高效性。值得注意的是，SBC_K-modes 中对带宽的选取和对混合型数据的处理都是非常有趣的研究课题。

参 考 文 献

[1] Qian Yu-hua, Li Fei-jiang, Liang Ji-ye, et al. Space structure and clustering of categorical data[J]. IEEE Transactions on Neural Networks and Learning Systems, 2015, 99: 1-13

[2] Carreira-Perpiñán M A, Wang Wei-ran. The K-modes algorithm for clustering[J]. arXiv preprint arXiv: 1304. 6478, 2013

[3] Huang Zhe-xue. A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining[M]// Research Issues on Data Mining & Knowledge Discovery. 1998: 1-8

（下转第 121 页）

参 考 文 献

[1] Park J, Sandhu R, Cheng Y. Acon: Activity-centric access control for social computing[C]// 2011 Sixth International Conference on Proc. of Availability, Reliability and Security (ARES). IEEE, 2011: 242-247

[2] Mahmood S. Online Social Networks: Privacy Threats and Defenses[M]// Security and Privacy Preserving in Social Networks. Springer Vienna, 2013: 47-71

[3] Hu Hong-xin, Gail-Joon Ahn, Jan Jorgensen. Multiparty Access Control for Online Social Networks: Model and Mechanisms[J]. Proc. of IEEE Transactions on Knowledge and Data Engineering, 2013, 25(7): 1614- 1627

[4] Thomas K, Grier C, Nicol D M. unfriendly: Multi-party privacy risks in social networks[C]// Proc. of Privacy Enhancing Technologies. Springer Berlin Heidelberg, 2010: 236-252

[5] Squicciarini A C, Shehab M, Wede J. Privacy policies for shared content in social network sites[J]. The VLDB Journal-The International Journal on Very Large Data Bases, 2010, 19(6): 777-796

[6] Amrutha P, Sathiyaraj R. Privacy Management of Multi User Environment in Online Social Networks (OSNs)[J]. GJCST-E: Network, Web & Security, 2013, 13(10): 01-07

[7] Subhani S, Rajasekhar M. A photo privacy for tagged images using rule-based access control in social networks[J]. International Journal of Research Sciences and Advanced Engineering, 2012, 2(5): 45-49

[8] Yeung C A, Kagal L, Gibbins N, et al. Providing Access Control to Online Photo Albums Based on Tags and Linked Data[C]// Proc. of AAAI Spring Symposium: Social Semantic Web: Where Web 2. 0 Meets Web 3. 0. 2009: 9-14

[9] Zhong Yong, Zhang Hong, Liu Feng-yu, et al. A Digital Rights Management Mechanism and Implementation Based on Logic Framework[J]. Journal of Computer Research and Development, 2010, 47(2): 223-230(in Chinese)
钟勇, 张宏, 刘凤玉, 等. 一种基于逻辑框架的数字版权管理机制和实现[J]. 计算机研究与发展, 2010, 47(2): 223-230

[10] Bertino E, Catania B, Gori R, et al. Active-U-Datalog: integrating active rules in a logical update language [C]// Proc. of International Seminar on Logic Databases and the Meaning of Change, LNCS 1472. Berlin: Springer, 1998: 107-133

[11] Montesi D, Bertino E, Martelli M. Transactions and updates in deductive databases[J]. IEEE Trans. Knowl. Data Eng., 1997, 9(5): 784-797

[12] Carminati B, Ferrari E, Perego A. Rule-based access control for social networks[C]// Proc. of on the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops. Spring-Verlag Lecture Notes in Computer Science, LNCS 4278, 2006: 1734-1744

[13] Jajodia S, Samarati P, Sapino M L, et al. Flexible support for multiple access control policies[J]. ACM Transaction on Database System, 2001, 26(2): 214-260

(上接第 93 页)

[4] Chan E Y, Ching W K, Ng M K, et al. An optimizationalgorithm for clustering using weighted dissimilaity measure[J]. Pattern Recoginzation, 2004, 37(5): 943-952

[5] Bai Liang, Liang Ji-ye, Dang Chuang-yin, et al. The impact of cluster representatives on the convergence of the K-modes type clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(6): 1509-1522

[6] Yang Yi-ming. An evaluation of statistical approaches to text categorization[J]. Information Retrieval, 1999, 1(1/2): 69-90

[7] Information G M. Uncertainty and the utility of categories[C]// Proc. of the Seventh Annual Conf. on Cognitive Science Society. Lawrence Erlbaum, 1985: 283-287

[8] Barbará D, Li Yi, Couto J. COOLCAT: an entropy-based algorithm for categorical clustering[C]// Proceedings of the Eleventh International Conference on Information and Knowledge Management. ACM, 2002: 582-589

[9] Aggarwal C C, Procopiuc C, Yu P S. Finding localized associations in market basket data[J]. IEEE Transactions on Knowledge and Data Engineering, 2002, 14(1): 51-62

[10] Cao Fu-yuan, Liang Ji-ye, Bai Liang, et al. A framework for clustering categorical time-evolving data[J]. IEEE Transactions on Fuzzy Systems, 2010, 18(5): 872-882

[11] Wrigley N. Categorical data analysis for geographers and environmental scientists[M]. Blackburn Press, 2012

[12] Chmielewski M R, Grzymala-Busse J W. Global discretization of continuous attributes as preprocessing for machine learning[J]. International Journal of Approximate Reasoning, 1996, 15(4): 319-331

[13] Dash M, Liu Huan. Consistency-based search in feature selection [J]. Artificial Intelligence, 2003, 151(1): 155-176

[14] Guyon I, Elisseeff A. An introduction to variable and feature selection[J]. The Journal of Machine Learning Research, 2003, 3: 1157-1182

[15] Zhou Zhi-hua. Three perspectives of data mining[J]. Artificial Intelligence, 2003, 143(1): 139-146

[16] Huang Zhe-xue. 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

[17] Lee M, Pedrycz W. The fuzzy C-means algorithm with fuzzy P-mode prototypes for clustering objects having mixed features [J]. Fuzzy Sets and Systems, 2009, 160(24): 3590-3600

[18] Yu Jian. General C-means clustering model[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(8): 1197-1211

[19] Alamuri M, Surampudi B R, Negi A. A survey of distance/similarity measures for categorical data[C]// 2014 International Joint Conference on Neural Networks (IJCNN). IEEE, 2014: 1907-1914

[20] Andritsos P, Tsaparas P, Miller R J, et al. LIMBO: Scalable clustering of categorical data[M]// Advances in Database Technology-EDBT 2004. Springer Berlin Heidelberg, 2004: 123-146

[21] Chan E Y, Ching W K, Ng M K, et al. An optimization algorithm for clustering using weighted dissimilarity measures[J]. Pattern Recognition, 2004, 37(5): 943-952

[22] Comaniciu D, Meer P. Mean shift: A robust approach toward feature space analysis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(5): 603-619