符号序列的概率向量聚类方法*

程铃钫[1]，陈黎飞[2†]
（1. 福建农林大学金山学院，福州 350002；2. 福建师范大学 数学与计算机科学学院，福州 350117）

摘　要：针对符号序列聚类中表示模型及序列间距离度量定义的困难问题，提出一种基于概率向量的表示模型及基于该模型的符号序列聚类算法。该模型引入符号序列的概率分布表示法，定义了一种基于概率分布差异的符号序列距离度量及该模型的目标函数，最后给出了一种符号序列 K-均值型聚类算法，并在来自不同领域的实际应用序列集上进行了实验验证。实验结果表明，与基于子序列表示模型的符号序列聚类算法相比，所提方法在 DNA 序列和语音序列等具有较多符号的实际数据上，在有效提高聚类精度的同时降低聚类时间 50% 以上。

关键词：数据聚类；符号序列；向量空间模型；概率向量；马尔可夫模型
中图分类号：TP391　　**文献标志码**：A　　**文章编号**：1001-3695(2018)06-1676-05
doi:10.3969/j.issn.1001-3695.2018.06.017

Clustering method for symbolic sequences using probability vectors

Cheng Lingfang[1], Chen Lifei[2†]
(1. *Jinshan College of Fujian Agriculture & Forestry University, Fuzhou* 350002, *China*; 2. *School of Mathematics & Computer Science, Fujian Normal University, Fuzhou* 350117, *China*)

Abstract: This paper proposed a representation model using probability vectors of symbolic sequences and a new clustering algorithm based on the model, to address the difficult problems in defining an efficient representation as well as a meaningful distance measure for symbolic sequences clustering. It proposed a probability-distribution-based representation method for symbolic sequences, on which first defined a new distance measure computed on the dissimilarity of the probability distributions, and also defined a clustering criterion for sequences clustering with the probability vector space model. Finally, it described a K-means-type algorithm for symbolic sequences clustering, and conducted a series of experiments on real-world sequence sets from various domains to evaluate its performance. The experimental results show that, on both gene sequences and speech sequences consisting of a relatively large number of symbols, the proposed method improves the clustering accuracy effectively with more than 50% decrease in the clustering time, compared with the existing algorithms using a subsequence-based representation model.

Key words: data clustering; symbolic sequence; vector space model; probability vector; Markov model

0　引言

符号序列这种具有内部结构的复杂数据类型广泛分布在数据挖掘的许多应用领域。例如，在计算生物学中，由碱基排列而成的基因就是一种符号序列，它是由 A、T、G、C 等代表不同碱基的符号组成的线性链。根据它们内在结构的相似性进行序列集的自动划分，即符号序列聚类[1,2]，具有重要的实际应用，譬如，经过基因序列聚类，人们可以预测基因功能和分析它们的演化关系等，实际上基因序列聚类已成为生物学家的重要研究内容之一[3,4]。

目前已提出了较多的聚类算法[2,3,5,6]，应用于符号序列聚类时，可以大致划分为基于模型的和基于数据的两种。前者着力于根据符号序列的特点构造适用的聚类模型，以 CLUSEQ[7] 等算法为代表，此算法利用了 Markov 链模型描述序列中的结构信息，并基于模型定义了序列和序列簇之间的相似性度量。由于简单、高效的缘故，基于数据的方案被广泛研究和应用，其出发点是将符号序列转换成普通聚类算法能够处理的简单数据类型[1,6]，通常是向量型数据（vector data），从而可以将 K-means（K-均值）等诸多成熟算法[8,9] 运用于符号序列聚类任务。

向量化符号序列的核心是提取隐藏在序列中的结构特征，进而将序列投影到由这些特征构成的向量空间中进行聚类。现有两类方法用于提取序列的结构特征[1]，即基于概率模型的方法和基于子序列的方法。前者使用隐 Markov 模型等为每条序列建立一个概率模型[10]，然后提取模型参数（如隐 Markov模型的状态转移矩阵等[11]）为描述序列结构的特征。基于子序列的方法其基本思想是利用符号序列的一系列短子序列描述结构特征，鉴于这样的特征提取过程易实现和所提取特征的实际有效性，现已成为一种主要的序号序列特征表示方法[8,9,12~14]。

尽管常用，基于子序列特征表示模型的序列聚类尚需解决若干难题。一方面，基于这种表示的向量空间模型通常具有较高的数据维数[8,9]，而高维数据聚类本身就是当前数据挖掘研究中面临的一个困难问题[15]；另一方面，多数面向向量的聚类算法基于特征独立假设衡量对象间的相似性（例如常用的欧

收稿日期：2017-01-16；**修回日期**：2017-03-10　　**基金项目**：国家自然科学基金资助项目(61672157)

作者简介：程铃钫(1983-)，女，山东滕州人，讲师，硕士，主要研究方向为机器学习、数据挖掘；陈黎飞(1972-)，男(通信作者)，福建长乐人，教授，博士，主要研究方向为统计机器学习、数据挖掘、模式识别(clfei@fjnu.edu.cn).

几里德距离[6]),这个假设在子序列特征表示模型中很难成立,因为作为特征的子序列通常并非统计独立[8,10],例如,两个以相同的($n-1$)个符号为前缀的 n-元子序列。

本文提出符号序列的概率向量聚类法,在该方法中,每个序列被表示成一个规范化的概率向量,向量维数等于构成序列的符号数目,并基于符号概率分布的差异衡量序列间的相似性,从而避免了传统子序列特征表示模型的高维性和基于特征独立假设的距离定义带来的问题。基于概率向量表示模型,提出了一种简称为 PCS(probabilistic clustering of sequences)的符号序列聚类新算法,这是一种 K-means 型算法,但以规范化平均概率向量为序列簇的中心。最后在来自三个应用领域的实际序列集上进行了实验评估,其结果验证了所提方法的有效性。

1　相关工作

全文使用 $S=s_1s_2\cdots s_l\cdots s_L$ 表示长度为 L 的符号序列,其每个符号 $s_l\in X$。这里,$X=\{x_1,x_2,\cdots,x_a\}$ 是所有符号的集合,$|X|=a$ 为符号数目,X 中的任一符号简记为 $x\in X$。待聚类序列集 DB 包含 N 个序列对象,给定聚类数 $K(1<K<N)$,一个基于划分的硬聚类算法将 DB 划分为 K 个互不相交的序列子集 $\pi_1,\pi_2,\cdots,\pi_k,\cdots,\pi_K$,称其中的 π_k 为序列集的第 k 个簇,其包含的序列数用 $|\pi_k|$ 表示。

与向量型数据不同,衡量序列间的相似性或距离时,需要考虑隐含在各序列中的结构特征[1~3,7]。现有的序列相似性或距离度量可大致划分为两种类型[9,13,16],即基于序列对齐的(alignment based)和非对齐(non-alignment)的度量。前者包括编辑距离、符号对齐距离等[17],其基本思想是通过特定的对齐操作将一个序列转换成另一个序列,根据所涉对齐操作的代价衡量两个序列间的差异。由于其依赖于对齐算法,通常具有较高的时间复杂度。

本文算法基于非对齐型序列度量,因其具有较低的时间复杂度。文献中的该型度量可以细化为两种子类型。第一种类型基于概率模型:首先为一组序列(一个序列簇中的所有序列)建立一个概率模型,并视之为该组序列的生成模型(generative model),再根据序列的生成概率计算对象—簇间的相似性。常用的概率模型包括 Markov 模型[2,7,10]、隐 Markov 模型[11]等。该类型方法的长处包括能够有效捕捉全局序列结构信息等。

第二种非对齐型相似性度量的基础是序列集包含的短子序列,基本思想是将序列分解为一系列短子序列的集合,通过比较这些短子序列的共现程度衡量序列结构的相似性[1,8,9,16]。在序列聚类中,常用的子序列提取方法有针对序列对的和针对序列集的两种,前者包括 SCS[14]等,后者主要基于 n-阶 Markov 链假设,从所有序列中收集连续 n 个符号构成的子序列[12,18]。在此基础上,以不重复的子序列为特征,建立序列的向量空间模型,将序列转换成高维向量。与基于概率模型的方法相比,基于子序列的方法着重于序列的局部结构特征,因算法构造简单、易实现,现已成为一种主要的序号序列聚类方法[1,8,9,12~14,18]。

从以上分析可知,常见的非对齐型度量多基于 n-阶 Markov链假设[10]:序列 S 中的符号 s_l 与其 n 元前缀子序列 $\delta_l=s_{l-n}\cdots s_{l-1}$ 相关,δ_l 具有固定的长度 n(其中 $i<0$ 的符号 s_i 用一个虚拟符号代替,表示序列的起点)。令 Y 为对应序列中观测符号的离散随机变量,Z 是与前缀子序列对应的随机变量,基于 Markov 链假设,序列 S 中符号 s_l 的概率通过以下条件概率估计:

$$p_S(Y=s_l|Z=\delta_l)=\frac{f_S(\delta_l s_l)+1}{f_S(\delta_l)+|X|} \tag{1}$$

式(1)是基于子序列频度的 Laplace 校正估计,$f_S(t)$ 表示子序列 t 在序列 S 中出现的次数。后文用 Δ 表示所有 n 元前缀子序列的集合。

显然,当 n 取值较大时,Δ 将包含为数众多的子序列(其数量可能达到 $|X|^n$)。这导致使用这些子序列构造的向量空间模型具有相当高的数据维度(第一种类型的非对齐序列相似性度量),也将导致基于概率模型的方法(使用第二种类型度量时)模型构造效率的大幅下降。本文构造的新模型以及序列相似性度量兼具此两类方法的特点,首先基于 n-阶 Markov 链假设模型化每条序列,然后将序列转换为一个概率向量,构造数据维度仅为 $|X|$ 的向量空间模型,因而可以在保证序列聚类精度的同时大幅提高聚类效率。

2　概率向量聚类算法

本章提出符号序列聚类新算法 PCS,首先给出符号序列的符号概率分布模型及在该模型中的序列距离度量。

2.1　符号概率分布模型

如前所述,符号序列间的相似性主要指序列结构的相似程度。对于由同一个 X 中的符号组成的序列 S 和 S',这种相似性自然体现在 $|X|$ 个符号分布的差异上,差异越小则序列的相似性就越高。因此,可以用离散随机变量 Y 的概率分布 $p_S(Y)$ 来刻画序列 S 蕴涵的结构特征,如下面的定义 1 所述。

定义 1　符号序列的概率分布表示。符号序列 S 的结构特征用其离散符号的概率分布 $p_S(Y)$ 来表示:

$$p_S(Y)=\sum_{\delta\in\Delta}p(\delta)p_S(Y|\delta) \tag{2}$$

其中:条件概率 $p_S(Y|\delta)$ 实为 $p_S(Y|Z=\delta)$,前缀子序列 δ 的概率 $p(\delta)$ 实为 $p(Z=\delta)$,为表达方便使用了式中的简化形式。此外,式(2)基于 n-阶 Markov 链假设,即假设序列中的每个符号与其 n 元前缀子序列 δ 相关,其中,条件概率 $p_S(Y|\delta)$ 根据式(1)估计。由于 Y 为离散随机变量,取值为 X 中某个符号 x,据式(2)可知 $\sum_{x\in X}p_S(x)=1$。

根据定义 1,序列 S 和 S'的结构相似性可以通过两个概率分布 $p_S(Y)$ 和 $p_{S'}(Y)$ 之间的差异来衡量。常用度量包括著名的巴氏系数(Bhattacharyya coefficient),其定义为

$$BC(p_S(Y),p_{S'}(Y))=\sum_{x\in X}\sqrt{p_S(x)p_{S'}(x)}$$

由于巴氏系数衡量的是两个分布之间的相关性,其数值越小表明两个分布之间差异越大,故序列 S 和 S'之间的距离(结构相异度)可以计算为

$$\mathrm{dis}(S,S')\sim 1-BC(p_S(Y),p_{S'}(Y))=$$
$$\frac{1}{2}\sum_{x\in X}\left(\sqrt{p_S(x)}-\sqrt{p_{S'}(x)}\right)^2 \tag{3}$$

在数值计算的角度看,式(3)实际上是一种平方欧几里德距离(squared Euclidean distance)函数,这为定义符号序列的新型向量空间表示模型提供了基础。

2.2 规范化概率向量表示模型

根据式(3),若将序列 S 和 S' 分别表示为式(4)定义的 $|X|$ 维向量 v_S 和 $v_{S'}$,则其距离的数值与 $\| v_S - v_{S'} \|^2$ 成正比,这里 $\| \cdot \|$ 表示向量的欧几里德范数。序列 S 对应的向量定义如下:

定义 2 符号序列的概率向量表示。符号序列 S 的概率向量表示为如下单位长度规范化向量 v_S:

$$v_S = (\sqrt{p_S(x_1)}, \sqrt{p_S(x_2)}, \cdots, \sqrt{p_S(x_{|X|})})^{\mathrm{T}} \tag{4}$$

定义 2 给出了符号序列的一种新型向量空间表示模型:将每条序列投影到一个 $|X|$ 维的空间,空间的每个维度与 X 中的一个符号 x 相对应,因而称之为符号空间;序列 S 投影在新空间每个维度 x 上的值取为 $\sqrt{p_S(x)}$,这样,$\| v_S \| = 1$,S 在该空间中映射为一个规范化的概率向量。因此,称这种映射关系为符号序列的规范化概率向量表示。

向量 v_S 的每个元素根据式(2)计算,涉及条件概率 $p_S(x|\delta)$ 和每个 n-元前缀子序列 δ 的概率 $p(\delta)$。由于 δ 作用于所有序列(不仅仅是序列 S 本身),$p(\delta)$ 应从整个序列集估计。这里采用频度估计法,即

$$p(\delta) = \frac{1}{\gamma} \sum_{S \in DB} f_S(\delta) \tag{5}$$

其中:γ 表示序列集包含的 n-元前缀子序列的总数。

给定序列集 DB,其概率向量空间表示模型的构造通过以下两个步骤实现。在第一个步骤中,给定 Markov 模型的阶数 n,扫描 DB 中的每条序列,收集其中的 n-元前缀子序列,记其总数为 γ,不重复的子序列组成的集合为 Δ,根据式(5)计算所有子序列 $\delta \in \Delta$ 的概率。此步骤的算法时间复杂度与基于子序列的向量空间模型构造算法一样[9,12],都为 $O(nM)$,其中 M 表示序列总长度。第二个步骤再次扫描 DB 中的每条序列 S,依式(1)计算每个符号的条件概率,再根据式(2)计算 S 上各符号的概率。使用这些概率值,根据定义 2 构造出每条序列的向量 v_S。因此,第二个步骤的算法时间复杂度也是 $O(nM)$。上述模型构造过程是聚类算法的一个预处理步骤。

2.3 聚类算法

本节提出 K-means 型算法 PCS,用于聚类规范化概率向量空间表示模型下的符号序列。在此模型中,每条序列已经被表示为一个数值型向量,因而,传统的 K-means 算法[6]是可以直接运用的。但是,传统的 K-means 算法基于向量与聚类中心(均值向量)的欧几里德距离,而非式(3)所示的序列间结构相异度。显然,对于一个序列聚类,其中心也是一条可以用定义 2 表示的序列,是一个规范化的概率向量,如定义 3 所示。

定义 3 序列聚类中心。序列聚类 π_k 的中心是一个单位长度规范化的向量 c_k:

$$c_k = (\sqrt{c_k(x_1)}, \sqrt{c_k(x_2)}, \cdots, \sqrt{c_k(x_{|X|})})^{\mathrm{T}}$$
$$\text{s.t.} \quad \sum_{x \in X} c_k(x) = 1 \tag{6}$$

基于定义 3 和序列间距离度量式(3),算法 PCS 的目标设定为最小化以下聚类优化目标函数:

$$J_0(\Pi, C) = \sum_{k=1}^{K} \sum_{S \in \pi_k} \| v_S - c_k \|^2$$

其中:$\Pi = \{\pi_k | k = 1,2,\cdots,K\}$ 和 $C = \{c_k | k = 1,2,\cdots,K\}$ 为待优化参数。应用拉格朗日乘子法,引入式(6)定义的约束条件,并代入式(4)和(6),聚类优化目标变为

$$\min J(\Pi, C) = \sum_{k=1}^{K} \sum_{S \in \pi_k} \sum_{x \in X} (\sqrt{p_S(x)} - \sqrt{c_k(x)})^2 + \sum_{k=1}^{K} \lambda_k (1 - \sum_{x \in X} c_k(x))$$

其中:$\lambda_k (k = 1,2,\cdots,K)$ 为拉格朗日乘子。

给定序列集 DB 的概率向量空间表示模型和聚类数 K,算法 PCS 基于 K-means 算法结构[5,6]求取 $J(\Pi, C)$ 的局部优解。算法描述如下:

算法 1 符号序列概率向量聚类算法 PCS

输入:符号序列概率向量 $v_S, S \in DB$;聚类数 K。

输出:聚类集合 $\Pi = \{\pi_1, \pi_2, \cdots, \pi_K\}$。

begin

1. 随机选择的 K 个概率向量组成初始聚类中心集合 $\hat{C}$。

2. repeat

2.1 令 $C = \hat{C}$ 固定,求解最小化 $J(\Pi, \hat{C})$ 的聚类划分 $\hat{\Pi}$;

2.2 令 $\Pi = \hat{\Pi}$ 固定,求解最小化 $J(\hat{\Pi}, C)$ 的聚类中心集合 $\hat{C}$。

until $C = \hat{C}$

end

算法 1 的步骤 2.1 通过将每条序列 S 重新划分到与其最相似的序列聚类 π_k 来实现,S 与 π_k 的相似度根据其概率向量 v_S 和 π_k 之间的序列距离式(3)计算。此步骤的划分规则为

$$k = \mathrm{argmin}_{k' = 1,2,\cdots,K} \| v_S - c_{k'} \|^2$$

步骤 2.2 中,聚类划分 Π 固定,最小化 $J(\Pi, \hat{C})$ 的中心集合 $\hat{C} = \{\hat{c}_k | k = 1,2,\cdots,K\}$ 对所有的 $x \in X$ 和 $k = 1,2,\cdots,K$ 有

$$\frac{\partial J}{\partial \hat{c}_k(x)} = 0, \frac{\partial J}{\partial \lambda_k} = 0$$

经过推导得到

$$\hat{c}_k(x) = \frac{[b_k(x)]^2}{\sum_{x' \in X} [b_k(x)]^2}$$

其中

$$b_k(x) = \frac{1}{|\pi_k|} \sum_{S \in \pi_k} \sqrt{p_S(x)}$$

由此可知,序列聚类中心是一个规范化的平均概率向量。

PCS 算法的时间复杂度为 $O(NK|X|)$,这里的符号数目 $|X|$ 实际上也是序号序列概率向量空间的数据维数。作为对比,基于子序列的向量空间模型其维数通常达到 $|X|^n$;从这个意义上说,PCS 可以看做是传统方法 $n = 1$ 的情形。但是,PCS 使用的序号序列概率向量表示实际上利用了 n-阶 Markov 链模型,因而兼具 Markov 模型化方法可以捕捉序列全局结构的优点,同时有效提高了聚类效率。

3 实验

实验包括 PCS 算法聚类结果有效性评估和聚类效率验证两个方面,并与若干相关工作相比较。

3.1 实验数据与实验设置

实验在六个序列集上进行,它们来自三个实际应用领域,详细信息如表 1 所示。如表 1 所示,第一组序列集是银行客户行为序列[19],两个数据集简称为 B1 和 B2,分别包含 2 000 个信用卡客户长度为 10 和 12 的交易行为序列,每条序列由三个符号组成,每个符号代表客户在一个月内的行为类型。目的是根据客户连续 10 个月或 12 个月的行为预测其是否破产(因此分为两类)。

第二个领域是计算生物学,包括 D1 和 D2 两个取自 PBIL 微生物同源基因家族库(http://pbil.univ-lyon1.fr)的基因序列集[12]。D1 中的六类基因序列在 PBIL 库分别名为

HBG065748、HBG000013、HBG010471、HBG000080、HBG060165 和 HBG000026,符号集为{A,C,G,N,T}。D2 中的序列分属 HBG050644、HBG423057、HBG093787、HBG099893、HBG415481 和 HBG364776 基因序列类,符号集为{A,C,G,N,T,Y}。目的是根据结构相似性将序列归类到各基因家族。从表 1 可知,基因序列集中的序列数较少但长度较长。

表 1　实验使用的实际应用序列集

数据集	符号数(|Σ|)	类数(K)	序列数(N)	平均长度	描述
B1	3	2	2 000	10	银行客户
B2	3	2	2 000	12	行为序列
G1	5	6	251	1 075	基因序列
G2	6	6	310	1 536	
S1	20	5	50	560	语音序列
S2	19	5	50	1 088	

表 1 中的 S1 和 S2 来自语音识别领域,两个序列集分别命名为 locmelovoy 和 locfmtrvoy[20]。其中的每条序列是五个元音('a','e','i','o','u')之一的法语语音序列,由音频信号分箱取样而成,因此具有较多的符号数目,如表 1 所示,分别达到 20 和 19。

实验使用了基于其他两种符号序列表示模型的 K-means 算法为对比算法。第一种对比模型是子序列特征表示模型,即由 n-元子序列为特征的向量空间模型,使用新近出版的文献[9]建议的规范化子序列频度(normalized term frequency,NTF;这里使用了文档挖掘领域的术语,将子序列看做组成文档的词,称为 term)向量表示法,以下简称基于该表示模型的 K-means 算法为 NTF-KM。第二种是加权子序列频度表示模型,使用了文档挖掘中常用的 IDF(inverse document frequency)加权法[21],在此基础上,再应用文献[9]的方法对向量作规范化处理。基于第二种对比型的算法简称 IDF-KM。

为对比不同类型算法的性能,实验还使用了层次聚类算法 bisection K-means (简称 BKM)[12]为对比算法。鉴于子序列特征表示模型的高维性,另选择了一种代表性的高维数据子空间聚类算法 EWKM[15]作性能对比。在实验中,BKM 和 EWKM 都使用了上述第一种对比模型;EWKM 算法的参数 β 设为建议值 0.5[15]。

对比算法及本文 PCS 算法都需要给定数据集的聚类数目 K。鉴于估计序列集最优聚类数的实际困难[7,9],在实验中,各算法聚类每个数据集时均设定参数 K 为如表 1 所示的实际类数。另外,这些算法的起点都是随机选择的一组初始聚类中心。为公平比较各算法的性能,在实验中,为每个序列集随机选择了 100 组初始聚类中心,各算法每次按顺序从中提取一组初始中心开始聚类,算法在同一个数据集上的 100 次运行结果作为性能对比的基础。除时间效率外,聚类性能的对比分析将依据聚类结果的质量,评价指标为聚类精度(clustering accuracy),其定义为聚类结果中类划分与实际类别标号相匹配的序列数目占总体的比例,精度越高意味着聚类结果具有越好的质量。

3.2　聚类结果

表 2 以"平均指标值 ± 1 个标准差"的格式列出了各算法在六个实际序列集各运行 100 次取得的平均聚类精度,每个序列集上最高的平均精度以加粗方式标注。本组实验使用个 5-阶 Markov 模型。如表 2 所示,除 B2 外,本文 PCS 算法都取得了最高的平均聚类精度;在序列长度较长的基因序列集 D1 和 D2 上,较之于对比算法,平均精度提高幅度达到 15% 以上;在符号数目较多的语音序列集 S1 和 S2 上,也有 5% 以上明显的精度提升。

表 2　各种聚类算法聚类精度对比(n = 5)

数据集	NTF-KM	IDF-KM	EWKM	BKM	PCS
B1	0.58 ± 0.01	0.57 ± 0.02	0.55 ± 0.02	0.50 ± 0.01	**0.58 ± 0.03**
B2	**0.58 ± 0.00**	0.57 ± 0.00	0.54 ± 0.03	0.53 ± 0.02	0.57 ± 0.03
D1	0.65 ± 0.09	0.62 ± 0.09	0.65 ± 0.09	–	**0.80 ± 0.10**
D2	0.53 ± 0.07	0.56 ± 0.07	0.50 ± 0.07	0.43 ± 0.09	**0.68 ± 0.11**
S1	0.73 ± 0.13	0.70 ± 0.14	0.72 ± 0.13	–	**0.79 ± 0.16**
S2	0.69 ± 0.14	0.69 ± 0.14	0.69 ± 0.14	–	**0.74 ± 0.11**

表 2 显示 B1 和 B2 上,基于子序列特征表示模型的算法可以取得与 PCS 相当的聚类结果质量。注意到这两个序列集的符号数目较少且序列较短(表 1),因此,所生成模型的维数较低,这与其他四个序列集有较大反差。层次聚类法 BKM 的结果类似,该算法在 D1、S1 和 S2 上的 100 次聚类中多数失败。表 2 结果也表明聚类子序列特征表示模型下的序列数据时,通过特征加权(包括 IDF-KM 和软子空间聚类算法 EWKM)对提高聚类结果质量作用有限。

给定序列集,所用 Markov 模型的阶数与子序列特征表示模型的数据维数密切相关,因而也是影响聚类算法性能的一个因素。图 1 ~ 3 分别显示两种算法的平均聚类精度随模型阶数的变化情况。图中仅对比 PCS 和 NTF-KM 算法,因为它们代表两种不同的符号序列向量空间表示模型:PCS 算法的符号概率向量模型以及对比算法的子序列特征表示模型。如图 2 所示,NTF-KM 算法的平均精度随阶数增加呈下降趋势,在符号数目和每序列符号数都较大的语音序列集上下降幅度更大。这是随阶数增加表示模型维数剧增的缘故。

在 PCS 算法中,表示模型的数据维数等于符号数,与 Markov 模型阶数无关;但序列的概率向量基于 Markov 链模型估计,因而其聚类结果质量也存在关联。如图 1 ~ 3 所示,PCS 的平均聚类精度对阶数变化的敏感程度依次为银行客户行为序列、语音序列、基因序列。对基因序列集 D1 和 D2,PCS 的聚类结果精度变化在 n = 5 和 n = 3 处形成拐点,这与基因序列聚类相关研究[2,12,14]中的结论基本吻合。采用这些设置时,PCS 的聚类精度较之于 NTF-KM 有大幅提高,如图 2 所示,n = 3 时,D2 的聚类精度提高了 20% 以上。

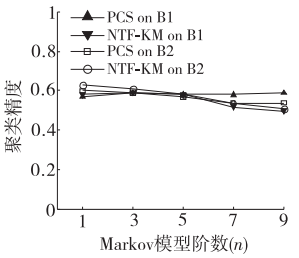

图1　客户行为序列集上两种算法的平均聚类精度随模型阶数变化情况

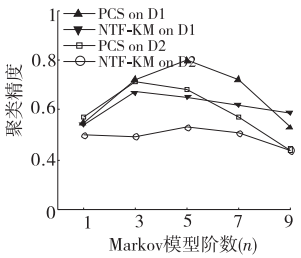

图2　语音序列集上两种算法的平均聚类精度随模型阶数变化情况

3.3　聚类效率

本组实验的目的是评估符号序列不同表示模型下的 K-means聚类效率。给定序列集,影响效率的主要因素是表示模型的数据维数,在子序列特征表示模型中,这与使用的 Markov 模型的阶数有关。表 3 列出了六个实际序列集上 n-元子序列的数目(等于表示模型的数据维数)。如表中所示,随 n 递增,n-元子序列数目大幅增加,例如,在基因序列集 D2 上,

$n=9$ 的子序列数近九倍于 7-元子序列数。而 PCS 算法使用的特征数等于 1-元子序列数，这不是维度约简的结果，而是本文提出的序列概率向量表示方法的结果，在有效压缩空间规模的同时还提高了聚类结果的质量(3.2 节)。

表 3 符号序列不同表示模型的特征数对比

数据集	子序列数($n=1$)	子序列数($n=3$)	子序列数($n=5$)	子序列数($n=7$)	子序列数($n=9$)	PCS 特征数
B1	3	27	243	1 286	1 751	3
B2	3	27	243	1 693	3 400	3
D1	5	94	1 118	16 372	114 919	5
D2	6	72	1 039	16 335	146 106	6
S1	20	2 941	16 627	22 935	25 313	20
S2	19	3 302	26 127	39 324	45 403	19

图 4 是 $n=5$ 时三种 K-means 算法的聚类时间对比图，数据取自各算法 100 次运行时间的平均值。层次聚类法 BKM 未参加本组实验，因为与 K-means 相比该类算法时间复杂度较高。IDF-KM 算法的运行时间同 NTF-KM，在图中统一用 KM 表示。如图所示，EWKM 算法需要额外的特征赋权步骤，聚类所需时间多于 KM 算法；PCS 算法由于较低的空间表示模型维数，更重要地，在这个新空间中 PCS 基于 Markov 链模型构造的序列概率向量可以反映序列中的全局结构特征，减少了聚类过程的迭代次数，进一步提高了聚类效率。对于实验中的六个实际序列集，与基于子序列特征模型的算法相比，PCS 所需聚类时间减少了 50% 以上。

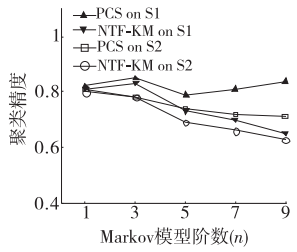

图3 基因序列集上两种算法的平均聚类精度随模型阶数变化情况

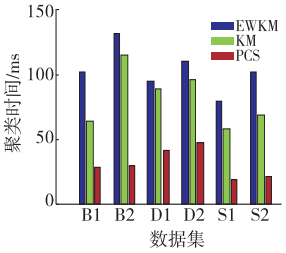

图4 三种K-means型算法的平均聚类时间对比(n=5)

4 结束语

本文首先提出了一种符号序列的新向量空间表示模型。在这个新模型中，每条序列映射为一个用 Markov 链模型估计的规范化概率向量。向量的长度等于组成序列的符号数，因此，与传统基于子序列特征的表示模型相比，大大压缩了数据空间的维数。其次，基于该表示模型提出了一种称为 PCS 的符号序列聚类新算法，与经典 K-means 算法不同，在 PCS 算法中，每个序列簇的中心用一个规范化的平均概率向量表示。实验在六个实际序列集上进行，其结果表明与基于子序列特征表示模型的现有算法相比，新方法可以有效提高聚类精度和聚类效率。

后续研究将着重于以下几个方面：将提出的单序列概率向量表示法推广到多维序列，研究多维序列聚类新算法；将有关方法运用到符号序列分类领域，研究有监督的概率向量模型构造方法和基于中心(原型)的符号序列分类器等。

参考文献：

[1] Dong Guozhu, Pei Jian. Sequence data mining [M]. Berlin: Springer, 2007.

[2] Xiong Tengke, Wang Shengrui, Jiang Qingshan, *et al*. A novel variable-order Markov model for clustering categorical sequences [J]. IEEE Trans on Knowledge and Data Engineering, 2014, 26(10): 2339-2353.

[3] 唐东明，朱清新，杨凡，等. 一种有效的蛋白质序列聚类分析方法[J]. 软件学报，2011，22(8)：1827-1837.

[4] 陶华，唐旭清. 蛋白质序列的聚类结构分析[J]. 生物信息学，2012，10(4)：269-273，279.

[5] Xu Rui, Wunsch D. Survey of clustering algorithms [J]. IEEE Trans on Neural Networks, 2005, 16(3): 645-678.

[6] Aggarwal C C. Data mining: the textbook [M]. Berlin: Springer, 2015.

[7] Yang Jiong, Wang Wei. CLUSEQ: efficient and effective sequence clustering [C]//Proc of the 19th International Conference on Data Engineering. Washington DC: IEEE Computer Society, 2003: 101-112.

[8] Yuan Liang, Hong Zhiling, Chen Lifei, *et al*. Clustering categorical sequences with variable-length tuples representation [C]//Proc of the 9th International Conference on Knowledge Science, Engineering and Management. Berlin: Springer, 2016: 15-27.

[9] Guo Gongde, Chen Lifei, Ye Yanfang, *et al*. Cluster validation method for determining the number of clusters in categorical sequences [J]. IEEE Trans on Neural Networks and Learning Systems, 2017, 28(12): 2936-2948.

[10] Fink G A. Markov models for pattern recognition: from theory to applications [M]. Berlin: Springer-Verlag, 2008.

[11] 郭彦明，陈黎飞，郭躬德. 基于隐马尔科夫模型的 DNA 序列分类方法[J]. 计算机系统应用，2014，23(7)：24-30.

[12] Wei Dan, Jiang Qingshan, Wei Yanjie, *et al*. A novel hierarchical clustering algorithm for gene sequences [J]. BMC Bioinformatics, 2012, 13(1): DOI: 10.186/1471-2105-13-174.

[13] 郑宏珍，初佃辉，战德臣，等. 基于数据挖掘的符号序列聚类相似度量模型[J]. 计算机工程，2009，35(1)：178-179，194.

[14] Kelil A, Wang Shengrui. SCS: a new similarity measure for categorical sequences [C]//Proc of the 8th IEEE International Conference on Data Mining. Piscataway, NJ: IEEE Press, 2008: 343-352.

[15] Jing Liping, Ng M K, Huang J Z. An entropy weighting K-means algorithm for subspace clustering of high-dimensional sparse data [J]. IEEE Trans on Knowledge and Data Engineering, 2007, 19(8): 1026-1041.

[16] Vinga S, Almeida J. Alignment-free sequence comparison: a review [J]. Bioinformatics, 2003, 19(4): 513-523.

[17] Herranz J, Nin J, Sole M. Optimal symbol alignment distance: a new distance for sequences of symbols [J]. IEEE Trans on Knowledge and Data Engineering, 2011, 23(10): 1541-1554.

[18] Kondrak G. N-gram similarity and distance [C]//Proc of the 12th International Conference on String Processing and Information Retrieval. Berlin: Springer-Verlag, 2005: 115-126.

[19] Xiong Tengke, Wang Shengrui, Mayeers A, *et al*. Personal bankruptcy prediction using sequence mining [C]//Proc of KDD Workshop on Data Mining for Business Applications. Quebec: University de Sherbrooke, 2008.

[20] Loiselle S, Rouat J, Pressnitzer D, *et al*. Exploration of rank order coding with spiking neural networks for speech recognition [C]//Proc of the IEEE International Joint Conference on Neural Networks. Piscataway, NJ: IEEE Press, 2005: 2076-2080.

[21] Gupta V, Lehal G. A survey of text mining techniques and applications [J]. Journal of Emerging Technologies in Web Intelligence, 2009, 1(1): 60-76.