规范化相似度的符号序列层次聚类

张 豪 陈黎飞 郭躬德
(福建师范大学数学与计算机科学学院福建省网络安全与密码技术重点实验室 福州 350007)

摘 要 符号序列由有限个符号按一定顺序排列而成,广泛存在于数据挖掘的许多应用领域,如基因序列、蛋白质序列和语音序列等。作为序列挖掘的一种主要方法,序列聚类分析在识别序列数据内在结构等方面具有重要的应用价值;同时,由于符号序列间相似性度量较为困难,序列聚类也是当前的一项开放性难题。首先提出一种新的符号序列相似度度量,引入长度规范因子解决现有度量对序列长度敏感的问题,从而提高了符号序列相似度度量的有效性。在此基础上,提出一种新的聚类方法,根据样本相似度构建无回路连通图,通过图划分进行符号序列的层次聚类。在多个实际数据集上的实验结果表明,采用规范化度量的新方法可以有效提高符号序列的聚类精度。

关键词 符号序列,聚类,相似度,规范化因子
中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.5.023

Hierarchical Clustering of Categorical Sequences by Similarity Normalization

ZHANG Hao CHEN Li-fei GUO Gong-de
(Fujian Provincial Key Laboratory of Network Security and Cryptology, School of Mathematics and Computer Science, Fujian Normal University, Fuzhou 350007, China)

Abstract A categorical sequence is composed of finite symbols which are arranged in a certain order. Nowadays, categorical sequences, such as gene sequences, protein sequences, and speech sequences, etc., widely exist in many application domains of data mining. As a major method for sequence data mining, sequence clustering has a great value in identifying the intrinsic structural of sequence data, while it is also an open problem due to the difficulties in measuring the similarity between sequences. This paper proposed a new similarity measure for categorical sequences, and introduced a length-normalization factor to address the problem that the existing methods are sensitive to the sequences length, and to improve the effectiveness of measuring sequences similarity. Based on the new similarity measure, a new clustering method was proposed, where directed acyclic graphs are constructed according to the similarity between samples and a hierarchical clustering of categorical sequences is performed by graph partitioning. Experimental results on real-world datasets show that the new methods based on the normalized similarity measure are able to improve the clustering accuracy significantly.

Keywords Categorical sequence, Clustering, Similarity, Normalized variant

1 引言

符号序列是一组按一定顺序排列而成的符号串,包括生物序列、语音序列、行为序列等[1,2]。在科研和商务中,我们将遇到越来越多的序列数据,对序列的分析和处理也显得越来越重要。聚类是数据挖掘、模式识别等领域的重要研究内容之一[3],是根据数据集自身特性进行类别划分的一种无监督学习方法,其目的是使得同一簇类内样本是高相似的,而不同簇类的样本相似度低[1]。对符号序列的聚类分析是分析、处理序列的一种重要方法,在识别序列内在结构等方面起到重要作用。同其它类型数据的聚类一样,符号序列聚类也是以样本间的相似性度量为基础的。然而,不同于数值型向量数据[1],序列数据不能使用欧氏距离等常用的相似性度量方法,而且序列相似性度量还要考虑序列长度差异等问题[4]。定义合理的符号序列相似性度量是一个困难问题,当前的主要方法分为两类[1]:基于概率模型的方法和序列比对方法。前者的基本思想是把序列看作由变长符号串组成马尔科夫过程[5],由此,一类序列可以用一个马尔科夫统计模型来代表。这样,序列间的相似度转化为序列与代表一类序列的统计模型之间的相似度[1]。显然,对于序列中出现次数较少的重要符号串和较短的序列,这种基于统计的方法的有效性将大大降低;此外,该方法还存在如何初始化统计模型的棘手问题[1]。序列比对方法基于这样的思想[6]:两序列间相似或相同的符号子串越多,则相似性越大。相比于概率模型方法,以符号子串为度量单元的序列比对方法可解释性更强,但需要应对如何确定符号子串最小长度和如何处置序列间长度差异

到稿日期:2014-02-18 返修日期:2014-04-23 本文受国家自然科学基金(61175123),深圳市基础研究(重点)项目(JCYJ20120617120716224)资助。

张 豪(1987—),男,硕士生,主要研究方向为数据挖掘,E-mail:zhanghao_study@163.com;陈黎飞(1972—),男,博士,副教授,主要研究方向为数据挖掘、机器学习,E-mail:clfei@fjnu.edu.cn;郭躬德(1965—),男,博士,教授,主要研究方向为人工智能、数据挖掘、机器学习。

等难题。需要指出的是，序列间的长度差异是定义相似度度量时需要考虑的一个重要因素，然而，现有的序列比对方法，如N-gram[4]、SCS[7]等，都忽略了这个因素或未能有效处理序列长度差异给相似性度量带来的偏倚问题[4]。

序列聚类除了要有合理的相似性度量方法外，还需选取合理的聚类方法，在序列聚类中层次聚类是最常用的聚类算法之一，其中又以凝聚型层次聚类法最受青睐[2]。凝聚型聚类法从只包含一个样本的 N 个簇开始，重复合并相似的簇，直到只有一个簇。有3种方法应用于凝聚聚类中簇间相似性度量：单链接、全链接和平均链接[8]。相比于全链接和平均链接，单链接方法具有不受簇合并顺序影响和能发现任意形状簇等优点，然而，由于单链接对噪声和孤立点敏感，导致相比于与孤立点间的距离，实际属于不同类别的簇间的单链接距离可能因为簇间噪声存在而非常小，对聚类树分割后将得到包含大多数样本的一个较大簇[2]。由于层次聚类方法的输出是一个称作系统树图的层次结构[8]，在实际应用中，通常需要用户指定一个阈值或所期望得到的聚类数目，据此分割聚类树以得到聚类结果。由于有效相似度度量的缺失及上述单链接方法固有的缺陷，应用于序列聚类时，现有的N-gram[4]、SCS[7]方法难以得到用户所期望的聚类结果。

本文提出一种新的序列相似性度量方法。该方法首先基于两序列间有统计意义的匹配符号数目计算初始相似度，继而采用长度规范因子来减小序列长度对相似度的影响。在该度量的基础上，提出了基于单链接的新凝聚型层次聚类方法。该方法自底向上地为符号序列构建无回路的连通图，通过引入样本间的链接度定义，计算链接样本间的链接度，然后根据用户设定的聚类数目对图的划分得到相应的聚类。新的聚类算法在分割连通图时，把簇间相似度转换为样本间链接度，该链接度综合了簇间距离和簇间链接样本与最近邻间的距离两个因素，避免了传统凝聚型算法仅根据簇间距离进行簇类划分所带来的问题。实验结果表明，本文提出的SCNS算法(Sequences Clustering bases a New Similarity measure)在不同领域的序列数据集上取得了良好的聚类效果。

本文第2节介绍相关工作；第3节对提出的序列比对算法进行描述；第4节对提出的聚类算法进行描述；第5节给出实验环境和实验结果分析；最后总结全文，并给出未来的研究方向。

2 相关工作

如前所述，相似度度量是一个聚类方法的基础构件，在符号序列聚类中相似度度量显得更为关键，这是因为组成序列的符号是非数值型的，且符号间存在复杂的联系，还具有序列长度不一等特点，这些因素都使得序列间的相似性度量变得困难。如何发现序列中关联性较强的一些符号并有效减小序列长度的影响是序列数据挖掘需要解决的首要问题之一。下面将介绍并分析现有若干代表性方法。

现有基于概率模型的代表性度量包括PST[5,9-11](概率后缀树)等，CLUSEQ[12]、DHCS[1]等算法都应用了此度量方法进行符号序列聚类。PST基于这样的假设：序列是由变长的符号串组成的马尔科夫链[5]，每个符号串内的相邻符号间有较高的统计意义上的关联性，而各个符号串间的关联性较低。具体方法是选取一些序列并根据这些序列的马尔科夫性质构建概率后缀树作为聚类的初始中心，通过计算各个概率后缀树模型产生一个序列的概率值大小来度量该序列与各个类间的相似度[1]。然而，对于较短的序列，很多符号可能出现很少的次数，这样统计本身的有效性将大大降低，而且符号间的关联性与概率上的关联性是否是正相关也是难以确定的；并且，这类方法存在初始中心选择问题[1,12]。因此，基于概率模型的方法用于序列间相似性度量时存在上述难以克服的困难。

序列比对算法的基本思想是寻找两序列匹配的符号串，匹配的符号串的个数越多，序列间的相似性越高。但如何确定匹配符号串所应满足的最短长度和如何降低两序列间长度影响是序列比对方法中的困难问题。定义序列比对算法中两序列间匹配的符号串的最小长度为 l，相似度函数为 $sim()$。如图1所示，若 $l=2$，两序列的相同字符串为“AA”、“FFF”；若 $l=3$，两序列的匹配字符串为“FFF”；若 $l\geqslant 4$，两序列的匹配字符串为空，所以如何确定符号串最小长度 l 是难点。从直观上理解，序列越长，序列间匹配符号串的个数很可能越多。然而对于图2中3个序列，不论 $S2$ 与 $S3$ 长度为多少，根据现有的序列比对算法都有 $sim(S1,S2)=sim(S2,S3)$，显然该度量方法并没有有效地处理序列长度对相似度的影响。

S1:AACFFF
S2:AADFFF

图1 有两对相同字符串的两序列

S1:AAAA
S2:AAAADDD……D
S3:AAAAEEE……E

图2 有相同字符串且长度不同的3个序列

常用的序列比对算法有N-gram[4]、ED[10,13](编辑距离)、SCS[7](基于符号串匹配的方法)。N-gram[4]是寻找两序列间最多的公共符号串的个数来度量两序列的相似度，然而此算法并没有给出如何选取 N 值使得相似性度量比较合理的方法。ED[10,13]通过计算两个序列间由一个转成另一个所需的最少编辑操作次数来计算相似度。然而，N-gram[4]和ED[10,13]只能找到两序列中顺序结构的各个公共字符串[7]，并且是找到全局意义上的匹配而不能找到局部上的匹配，很多情况下序列间局部匹配才更有意义[7]。SCS[7]找匹配的符号串的方法克服了上述缺点，构建以符号串作为“词”、序列作为文本的自然语言处理的方法，以此度量两序列的相似度，其找词的方法类似于生物序列寻找匹配片段的方法[7]。但是SCS[7]得到的列数很大的矩阵在处理起来比较困难，而且存在一个符号参与多次匹配，可解释性差。

上述序列相似性度量方法存在一个共同的缺点，即没有有效地处理两序列的长度对相似性度量结果带来的偏倚。虽然N-gram[4]采用除以比对的两序列中的最长序列的长度(LCSR)来消除偏倚的方法，但这种方法除了如图2所表明的问题外，还存在更一般性的长度偏倚问题：对于 S_1、S_2 和 S_3，它们的长度满足 $|S_1|>|S_2|>|S_3|$，因为对于两比对序列，序列越长，它们之间出现匹配的符号串的个数可能就越多，所以 $sim(L_1,L_2)>sim(L_1,L_3)$ 的可能性比较大。序列的长度问题是产生度量误差的重要因素，而现有序列比对算法并没有提出有效地处理序列长度影响的方法。

针对现有序列比对算法没有有效地消除长度对相似度影响的问题，本文提出了带有规范因子的序列比对算法，有效地消除了序列长度对相似度的影响。针对由于噪声和孤立点致

使分割基于单链接的凝聚聚类树难以得到所期望的聚类问题，本文提出新的聚类算法。该算法在用单链接度量簇间距离的基础上把凝聚聚类转化为图的构建问题和根据所引入的样本间链接度对图划分的问题，克服了上述缺点。

3 长度规范化的序列相似性度量

本节将详细阐述新的序列相似性度量方法；给出算法中匹配符号串长度取值的分析；提出规范因子所应满足的3个性质，提出满足这3个性质的规范因子并予以证明。

3.1 新的相似性度量

为了方便叙述相似性度量方法，引入下面两个定义。

定义1 记一个序列 $S=s_1\cdots s_m$，其中：

1) $s_i\cdots s_{i+l-1}$ 称为长度为 l 的子串，记为 $S_{i,l}$，$0\leqslant i\leqslant m-l+1$；

2)若 $i\leqslant j\leqslant i+l-1$，则有 $s_j\in S_{i,l}$；

3) $|S|$ 代表序列 S 的长度。

定义2 现有序列 $X=x_1\cdots x_m$ 与 $Y=y_1\cdots y_n$，其中：

1)对于 $\forall k$，若 $x_{i+k}=y_{j+k}$，则有 $X_{i,l}=Y_{j,l}$；用 Γ_l 表示两个相等的字符串的长度 l；$0\leqslant k\leqslant l-1$。

2)对于 $x_p=y_q$，若 $\exists X_{i,l}$，$\exists Y_{j,l}$，使得 $X_{i,l}=Y_{j,l}$ 且 $x_p\in X_{i,l}$，$y_q\in Y_{j,l}$，则称在 Γ_l 下，x_p 与 y_q 匹配。

定义3 序列 X 与序列 Y 的相似度为 $SIM(X,Y)$，$ESIM(X,Y)$ 表示序列 X 与序列 Y 的相似度量中的长度规范因子。

新的相似性度量方法通过寻找 Γ_l 下两比对序列中非重复的匹配符号的个数来得到两序列的初始相似度，然后初始相似度除以规范因子得到的值便是两序列的相似度。非重复匹配，是指序列中的任意一个符号最多只能与其比对序列的一个符号匹配，而不能与多个符号匹配。现举例说明新的度量方法寻找非重复匹配符号的过程。如图3所示，由短线连接起来的就是序列 S1 与 S2 之间在 Γ_l 取值为4的情况下所有非重复的匹配符号。例如两序列中的符号 K，两序列中都存在长度为4的符号串 KPQE，由定义2可知，两序列中符号 K 匹配，又因为两序列中的 K 都未与其他符号匹配，所以两序列中的符号 K 为非重复的匹配。同样，对于两序列中的符号 G，因为两序列中都存在长度为4的符号串 CDEG，所以由定义2可知，两序列中符号 G 匹配，又因为两序列中的 G 都未与其他符号匹配，所以两序列中符号 G 为非重复的匹配，而序列 S2 中字符串 CDEG 中的 CDE 都已经与其它符号匹配，由于规定不能重复匹配，S1 中符号串 CDEG 中的 CDE 不能与 S2 中符号串 CDEG 中的 CDE 匹配。图3所示两序列中非重复的匹配符号共有12个。可以看出新的序列相似性度量方法是局部的符号串匹配方法，而序列的局部匹配比全局匹配更有意义[7]。

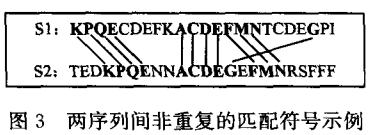

图3 两序列间非重复的匹配符号示例

3.2 计算方法

新的相似性度量方法的算法描述如下：

算法1 两序列相似性度量

输入：Γ_l，$X=x_1\cdots x_m$，$Y=y_1\cdots y_n$。

输出：序列 X 与序列 Y 的相似度 SIM(X,Y)。

步骤1 设 SIM(X,Y)=0，对于 X 中的任一个符号 x_i，设 $flag(x_i)=0$，同样对于 Y 中的任一个符号 y_j，设 $flag(y_j)=0$。

步骤2 对于序列 X 中的每个符号 x_i，在序列 Y 中寻找一个与之在 Γ_l 下匹配的符号 y_j，且满足 $flag(y_j)=0$；如果序列 Y 中存在与之匹配的符号，则令 $flag(y_j)=1$，SIM(X,Y)=SIM(X,Y)+1。

步骤3 SIM(X,Y)=SIM(X,Y)/ESIM(X,Y)。

序列 X 与序列 Y 分别有 $|X|$、$|Y|$ 个符号，序列 X 中每个符号在序列 Y 中寻找其匹配符号时，首先需要查找序列 Y 中是否有与之相同的符号，在最坏的情况下需要查找 $|Y|$ 个位置；如果找到相同的符号，在最坏的情况下需要判断这两个相同符号其后的 Γ_l-1 个符号是否相等，所以算法1的时间复杂度约为 $O(X_l\cdot|Y|\cdot\Gamma_l)$。在实现算法1时，采用以下索引和扩展技术来提高算法实际运行的效率：假设每个序列都由 m 个类别的符号构成。首先，把每个序列转化为 m 行的二维位置索引数组，数组的第 i 行存储该序列中第 i 个类别的符号的所有位置；然后，在寻找两序列匹配符号时，只需要在该符号对应的一维的位置索引数组上搜索即可；最后，如果一个符号找到了与之在 Γ_l 下匹配的符号，那么在两符号各自对应的序列的相应位置上，向左边和右边扩展，直到对应位置两符号不相等为止。扩展时，判断对应的两符号是否匹配。引用上述优化策略，算法1的时间复杂度由 $O(|X|\cdot|Y|\cdot\Gamma_l)$ 变为 $O(|X|\cdot|Y|\cdot\Gamma_l/m)$。

如何确定算法中两序列相等的符号串所应满足的最小长度 Γ_l，是影响序列相似度计算的一个重要因素，同时也是一个难点问题。下面对参数 Γ_l 的取值进行分析。根据算法1，两序列的相似度与两序列间匹配符号的个数有关，而匹配的前提是存在两个长度为 Γ_l 的相等子串。从统计学的角度看：Γ_l 越大，两序列间存在长度为 Γ_l 的相等子串的概率就越小，而此时在 Γ_l 下，两符号匹配的可信度越高，相应的匹配符号的个数就越少，且相比于取较小的 Γ_l 时，可能丢失很多匹配的符号；而 Γ_l 较小时，可能得到很多可信度低的匹配。因此选取合适的 Γ_l，才能得到合理的相似度。这里根据文献[6,7]中的方法来选取合适的 Γ_l 值：根据文献[14]的定理1，对于由 m 个类别的符号组成的序列 X 与序列 Y，设 X 与 Y 共同出现的概率即联合概率为 $P(X,Y)$，则出现概率为 $2P(X,Y)$ 的 X 与 Y 的最长公共字符串的期望长度 $K_{X,Y}$ 可用下式计算：

$$K_{X,Y}=\frac{\log(|X|^2+|Y|^2)+\log\lambda_{X,Y}(1-\lambda_{X,Y})+0.57}{-\log\lambda_{X,Y}}$$

$$\lambda_{X,Y}=\max\left(\sum_{i=1}^{m}(p_i^X)^2,\sum_{i=1}^{m}(p_i^Y)^2\right)$$

这里，p_i^X 和 p_i^Y 是第 i 个类别的符号在 X 和 Y 中出现的频率。

根据以上分析，在实际应用中用以下方法确定 Γ_l 的取值范围：对于一个样本集，求出每对序列 (X,Y) 的 $K_{X,Y}$，然后求出这些值的平均值 K_{AVR}，Γ_l 取区间 $[K_{AVR}-1,K_{AVR}+2]$ 内某个值(见5.3节的实验验证与分析)。

3.3 规范化因子

在 LCS(最长公共子序列)中，LCS 的长度随两序列的长度而增加。同样，算法1的步骤2结束后得到的相似度也会随着两序列长度的增加而增加。两个本来不属于同一类的序列，可能因为长度很大而有较高的相似度。规范化是减小序

列长度对相似度影响的一种方法[4]，较多采用的规范化方法是：$sim(X,Y)/\max(|X|,|Y|)$，即两序列规范化前的相似度除以这两序列中较长序列的长度[15]。这种方法存在的问题为：对于3个序列 X、Y、Z，假设 $|X|>|Y|>|Z|$，X 与 Y、X 与 Z 规范后的相似度分别为 $sim1=sim(X,Y)/|X|$、$sim2=sim(X,Z)/|X|$；由于 $|Y|>|Z|$，显然 $sim1>sim2$ 的可能性要比 $sim2>sim1$ 的可能性大，所以，用 $\max(|X|,|Y|)$ 作为规范因子并没有很好地减小长度效应的影响。下面将引入新的规范因子：$ESIM(S_1,S_2)$。

对于3个序列 X,Y,Z，且 $|X|\geqslant|Y|\geqslant|Z|$，由上面的分析可知，规范因子 $ESIM(S_1,S_2)$ 须满足以下性质：

(1) $ESIM(X,Y)\geqslant ESIM(X,Z)$，仅当 $|Y|=|Z|$ 时，$ESIM(X,Y)=ESIM(X,Z)$；

(2) $ESIM(X,Y)\geqslant ESIM(Y,Z)$，仅当 $|X|=|Y|=|Z|$ 时，$ESIM(X,Y)=ESIM(Y,Z)$；

(3) $ESIM(S_1,S_2)$ 与 $|S_1|$、$|S_2|$ 这两个因素有关。

设 $S_{\min}=\min(|S_1|,|S_2|)$，$S_{\max}=\max(|S_1|,|S_2|)$，$0<q<1$；构造 $ESIM(S_1,S_2)=S_{\min}\cdot(1-q^{S_{\max}/S_{\min}})$。下面的定理1和定理2表明，构造的 $ESIM(S_1,S_2)$ 函数满足性质(1)和性质(2)。但 q 取值过大或过小会使 $ESIM(S_1,S_2)$ 不满足性质(3)：q 较小时，如 $q=0.01$，$ESIM(S_1,S_2)\approx S_{\min}$，不满足性质(3)；$q$ 较大时，如 $q=0.99$，$ESIM(S_1,S_2)\approx S_{\max}$，不满足性质(3)。实验发现 q 在区间[0.8,0.95]内取值较好。

为了证明 $ESIM(S_1,S_2)$ 函数满足性质(1)，引入定理1。

定理1 $f(x)=1/x\cdot(1-q^x)$ 是关于 x 的减函数，其中 x 与 q 满足：$x\geqslant 1$，$0<q<1$。

证明：

(1)对 $f(x)$ 求导得 $f'(x)=-1/x^2\cdot(1-q^x)+1/x\cdot(-q^x)\cdot\ln q$。

(2)令 $g(q)=f'(x)$，对 $g(q)$ 求导得 $g'(q)=-q^{(x-1)}\cdot\ln q$。易知 $g'(q)>0$，所以 $g(q)$ 为增函数。

(3)把 q 的区间由(0,1)扩展为(0,1]，因为 $g(1)=0$，所以 $g(q)\leqslant 0$，即有 $f'(x)\leqslant 0$，仅当 $q=1$ 时，$f'(x)=0$。所以当 $x\geqslant 1$，$0<q<1$ 时，$f(x)=1/x\cdot(1-q^x)$ 是减函数。

为了证明 $ESIM(S_1,S_2)$ 函数满足性质(2)，引入定理2。

定理2 $h(t_1,t_2)=1-q_1^{t_1}-t_1\cdot(1-q_2^{t_2})\leqslant 0$，其中 t_1,t_2,q 满足：$t_1\geqslant 1$，$t_2\geqslant 1$，$0<q<1$。

证明：

(1) $h(t_1,t_2)=1-q_1^{t_1}-t_1\cdot(1-q_2^{t_2})\geqslant 1-q_1^{t_1}-t_1\cdot(1-q)$。

(2)令 $f(t_1)=1-q_1^{t_1}-t_1\cdot(1-q)$，对 $f(t_1)$ 求导得 $f'(t_1)=-q_1^{t_1}\cdot\ln q-1+q$，令 $g(q)=f'(t_1)$，对 $g(q)$ 求导得 $g'(q)=-t_1\cdot q_1^{(t-1)}\cdot\ln q-q_1^{(t-1)}+1$，易知 $g'(q)>0$，所以 $g(q)$ 为增函数。

(3)把 q 的区间由(0,1)扩展为(0,1]，因为 $g(1)=0$，所以 $g(q)\leqslant 0$，即有 $f'(t_1)\leqslant 0$，仅当 $q=1$ 时，$f'(t_1)=0$。所以当 $t_1\geqslant 1$，$0<q<1$ 时，$f(t_1)=1-q_1^{t_1}-t_1\cdot(1-q)$ 是减函数。因为 $f(1)=0$，所以 $f(t_1)\leqslant 0$，所以 $h(t_1,t_2)\leqslant 0$。

4 符号序列的层次聚类法

本节将详细阐述新的聚类方法，并给出该算法的时间复杂度分析。新聚类方法称为SCNS，其基本思想是用第3节定义的度量衡量符号序列间的相似性，构建无回路连通图，通过对图的划分得到聚类结果。图4和图5示出用该方法把10个样本聚为3簇的过程，图中黑点代表样本。图4表示把所有样本合并为一个簇后得到的无回路连通图，该连通图的点即是代表样本的黑点，连接两黑点的较粗的黑短线即是两样本之间的链接，图中的各个数字对应样本间"链接度"；把链接度从小到大排列起来得到(1.5,1.5,2,2,2,2,2,2,2,2)，选取第3个值'2'作为阈值，则把10个样本分为3簇，如图5所示。

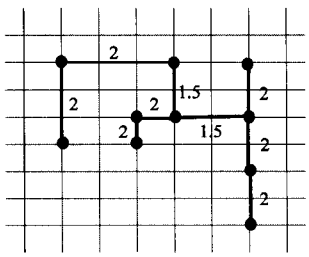

图4 1个例子：序列的无回路连通图

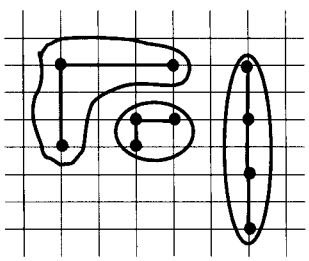

图5 3个子图：图4连通图的聚类划分结果

上述链接度用于衡量两链接的样本间相对相似度，其定义如下：

定义4 设 A、B 是链接的两样本，则这两点的链接度 $LR(A,B)$ 为：

$LR(A,B)=SIM(A,B)/\max\{SIM(A,\cdot)|\cdot\in P_+\}+SIM(A,B)/\max\{SIM(B,\cdot)|\cdot\in P_+\}$。其中 $SIM(A,B)$ 表示 A,B 两点间的相似度；$\max\{SIM(A,\cdot)|\cdot\in P_+\}$ 表示 A 与其链接的每个点间的相似度的最大值。

对于给定的符号序列集和目标聚类数 k，在聚类的初始阶段，我们将每个样本作为一个簇，且约定簇间的相似度为隶属于不同簇的所有样本对的相似度中的最大值，称对应的这两个样本为近邻样本。接着，合并具有最大簇间相似度的两个簇，同时用一条边链接这两个簇中的近邻样本；重复这个合并过程直到所有的样本被合并到一个簇中，此时，得到一个无回路的连通图。在这个连通图上，依据定义4计算每条边的链接度，调用一种排序算法将链接度从小到大排列，取出第 k 个链接度作为聚类划分阈值。聚类划分方法如下：检查该无回路连通图的每条边，若边的链接度小于阈值，就断开这条边，这样就可以把一个该连通图划分为 k 个无回路的连通子图。这里，每个子图中的样本(对应于一个符号序列样本)属于同一簇，这样就得到了 k 个目标聚类。由上述聚类过程可以看出，构建无回路连通图的方法类似于分层聚类中单链接方法，但避免了由于噪声和孤立点存在而影响聚类的划分。

SCNS聚类方法的形式描述如下：

算法2 SCNS聚类算法

输入：样本集S，聚类数k。

输出：k个簇的集合。

步骤1 根据算法1求出S中的每两个样本间的相似度。

步骤 2　用 C 表示各个簇的集合。初始化 C={所有样本}。

步骤 3　选取 C 中的属于不同簇的两个样本 p_1, p_2 使得 $SIM(p_1,p_2)=\max\{SIM(p_3,p_4)\mid p_3\in C, p_4\in C\}$，$p_1$ 链接到 p_2，把 p_1, p_2 合并为一个簇。

步骤 4　若 C 仅包含一个簇，转到步骤 5；否则，转到步骤 3。

步骤 5　计算所有已链接的两点的链接度；

步骤 6　对所有的链接度按从小到大的顺序排序，然后选取第 k 个值作为阈值。

步骤 7　对于所有的链接，如果链接度小于阈值，则断开该链接；最后，有 k 个链接断开，得到 k 个无回路的连通图，每个连通图内的样本为同一类别，不同连通图内的样本为不同类别，这样就得到 k 个聚类，输出这 k 个聚类。

对于 N 个样本，需要计算 $N(N-1)$对样本的间相似度，故算法 2 中步骤 1 的时间复杂度为 $O(N^2\cdot|X|\cdot|Y|\cdot\Gamma_l/m)$；在步骤 3、4 中，使用快速排序法对样本间相似度按从小到大顺序排序，时间复杂度是 $O(N^2\log N)$，从左向右依次判断排序后的每个相似度对应的两样本是否属于不同的簇，若属于不同的簇，则链接这两个样本，每次判断的时间复杂度 $O(1)$，最多需要 $N(N-1)$次这样的判断，因此步骤 3、4 的时间复杂度为 $O(N^2)$；步骤 6、7 的时间复杂度为 $O(N\log N)$。综上，SCNS 的时间复杂度为 $O(N^2\log N)$。

5　实验与结果分析

本节在实际应用数据集上检验所提出的 SNCS 算法。首先介绍数据集，接着介绍实验对比算法和性能评价指标，最后根据实验结果分析算法的有效性和算法效率。

5.1　实验数据集

为了验证算法的有效性，选取了基因、语音和蛋白质这 3 种不同领域的序列样本集。实验使用了 4 个数据集，分别用 DS1、DS2、DS3、DS4 表示。DS1 来自于 PBIL(http://pbil.univ-lyon1.fr/) 的同源脊椎动物基因数据库 HOVERGEN[16]；DS2 来自于 PBIL 的同源微生物基因数据库 HOMOLENS[16]。由于 PBIL 在不断更新，因此 DS1 和 DS2 样本数比文献[16]有所增加。DS3 是 5 个单独的法语元音('a','e','i','o','u')，它经常以序列的形式应用于语音识别[1]。DS4 是传统的生物序列比对方法很难区分的"(α/β)8-barrel"蛋白质序列[6]。各数据集的详细参数如表 1 所列，其中 *NS* 代表数据集中序列个数；*AEL* 代表该数据集中序列的平均长度；*NC* 代表数据集中序列类别的数目；*NCN* 代表数据集中各类别包含的样本数。

表 1　实验数据集的相关信息

DB	Description of the dataset	NS	AEL	NC	NCN
DS1	homologous vertebrate genes	285	1307	6	23:50:26:35:92:59
DS2	homologous Ensembl genes	251	1074	6	29:28:55:70:21:48
DS3	speech sequences	50	1898	5	10:10:10:10:10
DS4	(α/β)8-barrel protein sequences	33	871	5	6:7:6:8:6

5.2　对比算法与评价指标

为了测试新的序列比对算法的性能，选用 N-gram[4] 和 SCS[7] 这两个主流的序列相似性度量方法作为对比算法；选取基于单链接的凝聚型聚类算法作为比对算法，以此衡量新的聚类算法的性能。为了进一步测试 SCNS 的性能，另选用 CLUSEQ[12] 这个基于 PST 的序列聚类算法作为比对算法。由于 N-gram[4] 中 N 值的选取对聚类效果有影响，实验中仅给出经过对 N 值大范围测试后 N-gram[4] 在上述聚类算法下最好的聚类精度。实验中所有算法的聚类数目都选用实际类别数目。

常用的聚类有效性评价方法有外部评价法、内部评价法和相对评价法[17]。由于实验数据的实际类别已知，这里采用常用的聚类有效性评价外部指标——F-measure[18] 来评价聚类结果的有效性。对于实际类别 i 和聚类 j，评价指标 $F(i,j)$ 定义为：

$$F(i,j)=\frac{2*prcision(i,j)*recall(i,j)}{prcision(i,j)+recall(i,j)}$$

其中，$i=1,2,\cdots,e, j=1,2,\cdots k$，$precision(i,j)=n_{ij}/n_j$，$recall(i,j)=n_{ij}/n_i$，$e$ 是实际类别数目，k 是聚类数目，n_{ij} 是聚类 j 中属于实际类别 i 的样本的数目，n_i 是实际类别 i 中样本的数目，n_j 是聚类 j 中样本的数目。聚类结果的总体 F-measure 定义如下：

$$F=\sum_i\frac{n_i}{N}\max(F(i,j))$$

其中，N 是样本集样本数目。

5.3　实验结果及分析

实验中的各个算法分别在每个数据集上进行了 5 次测试，取 5 次聚类精度的平均值，实验结果如表 2 所列。SCNS 在序列的相似度量中，通过理论估计[14] 匹配字符串所应满足的长度值，然后在具体的数据上通过实验来选取最合适的长度值。表 3 给出在各个数据集上 Γ_l 的取值与 K_{AVR} 的值，可以看出，Γ_l 与 K_{AVR} 比较接近，因此 Γ_l 的选取并不十分困难。

表 2　基于不同聚类方法的各个序列相似度算法的聚类精度 F-measure

DB	新聚类方法			凝聚聚类(单链接)			CLUSEQ
	SCNS	N-gram	SCS	SCNS	N-gram	SCS	
DS1	**1.0000**	0.7630	0.7608	**1.0000**	0.8433	0.7608	0.4656
DS2	**0.8302**	0.7065	0.3351	0.7675	**0.8590**	0.3351	0.3601
DS3	**1.0000**	0.8481	0.8421	**1.0000**	0.7241	0.8421	0.6095
DS4	**1.0000**	0.8248	0.8330	**0.8717**	**0.8717**	0.8330	0.4927

表 3　各个数据集上 Γ_l 的取值与 K_{AVR} 的值

DB	Γ_l	K_{AVR}
DS1	12	10.97
DS2	12	10.71
DS3	7	7.28
DS4	5	5.05

从表 2 可以看出，在新的聚类方法和基于单链接的凝聚聚类上，相比于 N-gram[4] 与 SCS[7] 等其它算法，新的序列相似性度量方法表现良好，这验证了使用规范因子可以有效地处理序列长度的影响从而提高序列相似性度量的有效性。相比于基于单链接的凝聚聚类，新的序列相似性度量使用基于图划分的聚类算法能有效地提高聚类精度。相比于表 2 中的 3 个序列比对算法在各个聚类算法上的精度，CLUSEQ[12] 的聚类精度是很低的，说明在序列聚类中使用序列比对方法比使用统计模型方法更合理。

CLUSEQ[12] 采取的随机选取聚类中心的方法会影响聚类精度，除此之外，序列的重要片段并不一定具有较高的统计频率，所以以 PST 模型为基础的 CLUSEQ[12] 并不可靠；序列

(下转第 141 页)

机应用,2011,31(10):2869-2871
[10] 黎亮,谭世海,师伟. 基于聚类的多传感器数据融合方法研究[J]. 计算机工程,2013,39(5):61-64,68
[11] 董赞强. 基于网络编码的数据通信技术研究[D]. 南京:南京邮电大学,2013
[12] Li S Y R, Yeung R W, Cai N. Linear network coding[J]. IEEE Transactions on Information Theory, 2003, 49(2): 371-381
[13] Bhardwaj M, Garnett, Chandrakasan A P. Upper bounds on the lifetime of sensor networks[C]// IEEE International Conference on Communications, 2001(ICC 2001). IEEE, 2001, 3: 785-790
[14] Rashmi R R, Soumya K G. Adaptive data aggregation and energy efficiency using network coding in a clustered wireless sensor network: An analytical approach[C] // Computer Communications. Volume 40, March 2014: 65-75

(上接第 118 页)

间可能有大量在非顺序的匹配符号[7],而 N-gram[4] 只能找到顺序匹配符号,从而会大幅降低聚类精度;SCS[7] 采用类似生物序列比对中常用的方法来找到匹配的子序列,然后构建符号子串和序列间的矩阵,并没有考虑序列长度的影响,且存在一个子串和多个不同子串同时匹配的情况,可解释性较差;更重要的是,SCS[7] 与 N-gram[4] 并没有严格考虑序列长度对序列间相似度的影响,导致序列间相似度量存在偏倚而影响聚类精度。

SCNS 通过分析序列长度对相似度的影响,提出了规范因子的 3 要素,进而构建了规范因子函数。通过定理 1 和定理 2 证明可以看出这个函数满足规范因子所应具备的 3 个性质,实验表明,该规范因子有效地处理了序列长度对相似性度量带来的影响。

下面测试新方法的可伸缩性。为检验序列长度对相似性度量计算时间的影响,对含有 AGCT 4 个符号的合成的随机序列集进行测试,测试结果如图 2 所示。测试方法如下:选用 N-gram 与 SCS 作为对比算法;SCNS 中参数的 Γ_l 取为 12,N-gram[4] 中参数的 N 也取为 12;生成 6 个随机序列集 1—6,每个序列集包含 10 个样本,其中序列集 i 中每个随机序列的长度为 $1000 * i$;记录上述各个算法在每个数据集上运行 5 次的度量两样本间相似度花费的平均时间。因为这 3 个算法时间复杂度都为 $O(\alpha \cdot L_1 * L_2)$形式(其中 α 具体取值与具体算法有关),所以时间呈二次函数型增长。但从图 2 可以看出,SCNS 时间开销最低,且在实际应用中序列长度较短时,时间开销与序列长度呈线性关系。

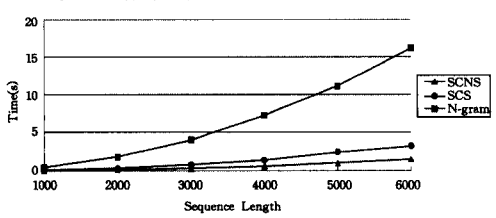

图 6 各算法相似性度量平均时间比较

结束语 现有的序列相似性度量算法没有有效地处理序列长度对相似度的影响,针对此问题,本文提出了新的相似性度量算法和新的聚类方法。在相似性度量算法中,首次提出了规范因子 3 个性质,并引入了满足此 3 要素的规范因子,新的相似性度量算法有效地减小了序列相似性度量中长度的影响。针对基于单链接的凝聚聚类算法的缺点,提出了基于图划分思想的聚类方法,该方法有效地克服了基于单链接的凝聚聚类易受噪声和孤立点影响的缺点,大大提高了聚类精度。实验表明,与现有序列聚类算法相比,SCNS 在多个领域的符号序列数据集上都有很好的聚类质量。下一步工作将研究 SCNS 参数选取策略和寻找提高其运行速度的优化算法。

参 考 文 献

[1] Xiong T, Wang S, Jiang Q, et al. A new Markov model for clustering categorical sequences[C] // Proceedings of the International Conference on Data Mining (ICDM). 2011: 854-863
[2] Dong Guo-zhu, Pei Jian. Sequence Data Mining[M]. New York: Springer-Verlag New York Inc., 2007: 1-65
[3] 孙吉贵,刘杰,赵连宇. 聚类算法研究[J]. 软件学报,2008,19(1):48-61
[4] Kondrak G. N-gram similarity and distance [C]// String Processing and Information Retrieval. 2005: 115-126
[5] Ron D, Singer Y, Tishby N. The power of amnesia: Learning probabilistic automata with variable memory length[J]. Machine learning, 1996, 25(2/3): 117-149
[6] Kelil A, Wang S, Brzezinski R, et al. CLUSS: Clustering of protein sequences based on a new similarity measure[J]. BMC bioinformatics, 2007, 8(1): 286
[7] Kelil A, Wang S. SCS: A new similarity measure for categorical sequences[C]// International Conference on Data Mining. 2008: 343-352
[8] Alpaydin E. 机器学习导论[M]. 范明,等译. 北京:机械工业出版社,2009: 95-96
[9] Grossi R, Vitter J. Compressed suffix arrays and suffix trees with applications to text indexing and string matching[C] // Proc. of ACM STOC. 2000: 397-406
[10] Gusfield D. Algorithms on strings, trees, and sequences [J]. ACM SIGACT News, 1997, 28(4): 41-60
[11] Ukkonen E. On-line construction of suffix trees[J]. Algorithmica, 1995, 14(3): 249-260
[12] Yang J, Wang W. CLUSEQ: Efficient and effective sequence clustering[C] // International Conference on Data Engineering. IEEE, 2003: 101-112
[13] Hirschberg D S. Pattern Matching Algorithms [M]. Oxford Univ. Press, London, 1997: 123-142
[14] Karlin S, Ghandour G. Comparative statistics for DNA and protein sequences: single sequence analysis[J]. Proceedings of the National Academy of Sciences, 1985, 82(17): 5800-5804
[15] Melamed I D. Bitext maps and alignment via pattern recognition [J]. Computational Linguistics, 1999, 25(1): 107-130
[16] Wei D, Jiang Q, Wei Y, et al. A novel hierarchical clustering algorithm for gene sequences[J]. BMC bioinformatics, 2012, 13(1): 174
[17] Halkidi M, Batistakis Y, Vazrgiannis M. On clustering validation techniques[J]. Intelligent In formation Systems, 2001, 17(2/3): 107-145
[18] Larsen B, Aone C. Fast and effective text mining using linear-time document clustering[C] // Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 1999: 16-22