

一种改进的符号化时间序列聚类方法

李志刚, 牛 强

(中国矿业大学 计算机科学与技术学院, 徐州 江苏 221116)

摘 要: 符号化时间序列聚类是聚类研究中的热点之一, 其中关键问题是时间序列符号化相似度问题. 本文针对传统的基于欧式距离度量存在的缺陷, 以 LCS 度量为基础, 提出了 ELCS 相似性度量, 克服了 LCS 度量需要依赖线性函数选取的不足. 在两类数据集上进行的实验表明, 同其他常用度量的比较, 该度量有着更好的聚类效果.

关键词: 时间序列符号化; 聚类; 相似性度量; 最长公共子序列

中图分类号: TP311.13

文献标识码: A

文章编号: 1000-7180(2012)11-0074-04

An Improved Symbolization Time Series Cluster Method

LI Zhi-gang, NIU Qiang

(School of Computer Science & Technology, China University of Mining and Technology, Xuzhou 221116, China)

Abstract: The symbolization time series cluster is one of hot spots in the clustering research. The key to question is similarity in signifying of time series. This paper puts forward the ELCS similarity measure based on the defects existed in the traditional Euclidean distance metric, and on foundation of the LCS measure. It can overcome the shortcomings of the LCS measure which rely on the selection of linear function. The experiment on two types of data sets shows that compared with other common measure, this measure has a better clustering effect.

Key words: symbolization of time series; cluster; similarity measure; LCS

1 引言

符号化时间序列聚类是聚类研究中的热点之一, 其中关键问题是时间序列相似度问题.

Agrawal 等人^[1] 提出时间序列的相似性度量后, Faloutsos C. 等^[2] 最早提出基于欧氏距离的度量, 但其存在着对噪声十分敏感的问题, 从而使得聚类的效果很差. 之后文献^[3] 通过在欧氏距离上增加权重在一定程度上具有了消除变换偏移量、控制缩放幅度比例等优点, 但其中的参数需要用户干预.

Paterson^[4] 于 1994 年将最长公共子序列 (LCS) 度量应用于时间序列相似性度量中. Bollobas 等^[5] 在其基础上提出 LCSS, 使得经过振幅平移、时间轴伸缩和弯曲等变形的时间序列得到较好的相似性度量. 另有文献提出了基于斜率的^[6]、基于模型的^[7]、基于事件^[8] 和基于形态^[9] 的相似度测量.

本文主要研究了符号化后的时间序列相似度问题. 首先阐述了相关定义及经典的相似性度量, 并提出一种新的基于 LCSS 算法的相似性度量算法: 它相对于 LCSS 算法有效的避免了线性函数的选取, 提高了度量的准确性, 相对于 DTW 度量时间效率上有大幅提高. 然后做实验对所提出的聚类算法的效率进行验证.

2 LCS 和 LCSS 相似性度量模型

2.1 最长公共子序列度量

设有时间序列样本 $X, Y \in A$, 其向量形式表示: $X = \{x_1, x_2, \dots, x_n\}$, $Y = \{y_1, y_2, \dots, y_n\}$, 它们满足以下条件的最长公共子序列分别为 $X' = \{x_{i_1}, x_{i_2}, \dots, x_{i_l}\}$ 和 $Y' = \{y_{j_1}, y_{j_2}, \dots, y_{j_l}\}$ 其中 l 为公共子序列的长度, 那么时间序列 X 和 Y 之间的相似度定义为 $Sim(X, Y) = l/n$.

(1)任意 $1 \leq k \leq l$, 都满足 $i_k < i_{k+1}, j_k < j_{k+1}$;

(2)任意 $1 \leq k \leq l$, 都有 $x_{i_k} = x_{j_k}$.

2.2 LCSS

LCS 度量虽然能够避免时间序列的短期突变或者间断带来的相似性问题,但是对时间序列的振幅平移、时间轴伸缩和弯曲等变形无法得到较好的相似性度量结果. LCSS 度量即为针对以上问题的改进.

设 $\delta > 0$ 是个整型常数, $0 < \epsilon < 1$ 是实常数, $f \in L$ 其中 L 是线性函数集合. 给定实数时间序列 $X = \{x_1, x_2, \dots, x_n\}$ 和 $Y = \{y_1, y_2, \dots, y_m\}$, 它们满足以下条件的最长公共子序列分别为 $X' = \{x_{i_1}, x_{i_2}, \dots, x_{i_l}\}$ 和 $Y' = \{y_{j_1}, y_{j_2}, \dots, y_{j_l}\}$.

(1)任意 $1 \leq k \leq l$, 都满足 $i_k < i_{k+1}, j_k < j_{k+1}$;

(2)任意 $1 \leq k \leq l$, 都满足 $|i_k - j_k| \leq \delta$;

(3)任意 $1 \leq k \leq l$, 都有 $y_{j_k} / (1 + \epsilon) \leq f(x_{i_k}) \leq y_{j_k} (1 + \epsilon)$.

令 $S_{f, \epsilon, \delta}(X, Y) = l/n$, 那么时间序列之间的相似度定义为

$$Sim(X, Y) = \max_{f \in L} \{S_{f, \epsilon, \delta}(X, Y)\} \quad (1)$$

3 改进的 LCS 和 LCSS 相似性度量模型

虽然 LCSS 度量通过一些约束允许序列存在线性偏差,支持振幅平移、振幅伸缩和线性漂移等时间序列变形,并且允许子序列下标之间的距离存在一定偏差,在一定程度上也支持时间轴的伸缩和弯曲. 但存在如下问题:

(1)LCSS 给出的其实是一个解集,针对不同的时间序列数据集,线性函数 f 的选取也随之不同. 也就是说,只有通过一定的训练数据集预先求得对应的线性函数,才能进一步较为准确的度量序列的相似性. 而训练集和测试集总是存在差异,所以结果也就不太理想.

(2)LCSS 虽然也可以应用与两条不等长序列的比较,但当 $|i_k - j_k| \leq \delta$ 存在时,也限定了时间序列 $X = \{x_1, x_2, \dots, x_n\}$ 和 $Y = \{y_1, y_2, \dots, y_m\}$ 长度差 $|m - n| \leq \delta$. 否则,序列会出现漏检的候选系列(最早由 Keogh 等人^[10]在 DTW 算法中注意到这个问题). 由此,LCSS 算法对时间轴伸缩的支持是有限的.

针对 LCSS 度量存在的问题,本文提出 ELCS (Extended-LCS) 度量:

设 $\mu > 1$ 和 $\theta > 0$ 是实常数,给定实数时间序列 $X = \{x_1, x_2, \dots, x_n\}$ 和 $Y = \{y_1, y_2, \dots, y_m\}$, 对其进行

归一化,使得所有序列值位于 $[0, 1]$ 之间,得到 $X_{\text{norm}} = \{x'_1, x'_2, \dots, x'_m\}$ 和 $Y_{\text{norm}} = \{y'_1, y'_2, \dots, y'_m\}$, 他们满足以下条件的最长公共子序列分别为 $X'_{\text{norm}} = \{x'_{i_1}, x'_{i_2}, \dots, x'_{i_l}\}$ 和 $Y'_{\text{norm}} = \{y'_{j_1}, y'_{j_2}, \dots, y'_{j_l}\}$.

(1)任意 $1 \leq k \leq l$, 都满足 $i_k < i_{k+1}, j_k < j_{k+1}$;

(2)任意 $1 \leq k \leq l$, 都有 $\frac{1}{\mu} < \frac{i_k}{j_k} < \mu$ 并且 $\frac{1}{\mu} <$

$$\frac{i_k - m}{j_k - n} < \mu;$$

(3)任意 $1 \leq k \leq l$, 都有 $|x'_{i_k} - y'_{j_k}| < \theta$.

$$\text{令 } S_{\theta, \mu}(X_{\text{norm}}, Y_{\text{norm}}) = \frac{2l}{m+n}$$

那么时间序列之间的相似度定义为

$$Sim(X, Y) = \max \{S_{\theta, \mu}(X_{\text{norm}}, Y_{\text{norm}})\} \quad (2)$$

以上定义中,参数 μ 的引入使得度量的相似性矩阵的路径集中在一个菱形区域,不仅防止了序列的过匹配,同时降低时间复杂度. 而且路径区域的选取同各序列长度密切相关,不仅不会出现漏检序列,同时的大幅降低时间复杂度的情况下,很好的适应时间轴伸缩和变形的序列匹配.

参数 θ 的引入,使得相似性度量算法在经过归一化后,进一步有柔性匹配的空间.

序列归一化的处理如公式(3)所示:

$$X_{\text{norm}} = \frac{X - \min_{i \in [1, m]} \{x_i\}}{\max_{i \in [1, m]} \{x_i\} - \min_{i \in [1, m]} \{x_i\}} \quad (3)$$

其中 $x_i \in X$, 这样避免了线性函数 f 选取的困难,同时又完整的保留了序列的数值趋势信息.

4 实验

相似性度量是其他数据挖掘处理过程的基础,测度的准确性直接影响到其他过程的处理效果. 反之,我们可以通过聚类结果的正确率来客观的衡量不同相似性测度方法的优劣.

4.1 实验环境与实验数据

在 Window XP Professional 操作系统下用 C# 语言实现了算法,在 CPU 为主频 2.20 GHz 的 E4500 和内存 1024M 的计算机上进行实验. 实验数据集使用 UCI 知识发现数据库档案(UCI KDD Archive)中的序列 I 控制图时间序列(Synthetic Control Chart Time Series)和 CBF 数据集.

序列 I 控制图时间序列数据集(以下简称为 SCC 数据集)的实验数据为 600 条,每条时间序列长度为 60,分为 6 类.

CBF 数据集中包含 Cylinder (C), Bell (B),

Funnel(F)三种事件,它是典型的合成数据集.数据为 30 条,每条时间序列长度为 256.

4.2 实验过程

对时间序列数据聚类是指通过相似性度量,将时间序列数据进行分组,使得同一组的时间序列彼此相似,不同组的时间序列相异.本文采用自底向上的层次聚类^[11],设定初始数据为 $C_1, C_2, \dots, C_n$,算法步骤如下:

(1)将每条时间序列作为一类 C_i .

(2)计算两两类之间的相似性,得到一个相似性矩阵.

(3)合并最为相似的两类,返回(2)循环执行,直到合并后的类总是等于设定的分类数目.

其中类簇间的距离使用 ELCS 相似性度量计算.

已知分类的标准结果 $C = C_1, C_2, \dots, C_k$ 和每种方法的聚类结果 $C' = C'_1, C'_2, \dots, C'_k$,则聚类精度由公式(4),(5)计算:

$$\text{Sim}(C_i, C'_j) = 2 \frac{|C_i \cap C'_j|}{|C_i| + |C'_j|} \quad (4)$$

$$\text{Sim}(C, C') = \frac{\sum_i \max_j \text{Sim}(C_i, C'_j)}{k} \quad (5)$$

$\text{Sim}(C', C)$ 的计算和 $\text{Sim}(C, C')$ 类似.因为 $\text{Sim}(C', C)$ 和 $\text{Sim}(C, C')$ 不对称,所以采用 $\frac{\text{Sim}(C', C) + \text{Sim}(C, C')}{2}$ 作为最终的评价标准.

对于每一组测试数据集分别采用以上几种方法进行聚类计算,并计算相应的聚类精度.通过所设计的实验进行验证.

4.3 实验结果与分析

4.3.1 参数的确定

ELCS 度量中涉及到的参数有 μ 和 θ 两个,其中参数 μ 对算法的性能影响十分显著,所以在 SCC 数据集上进行实验,随着参数 μ 的变化,聚类的准确率如图 1 所示.图 2 和图 3 分别显示了聚类结果的类内平均距离和类间平均距离随参数 μ 的变化.

随着 μ 选取的增大,分类的正确率由低到高,之后降低.在 $\mu = 2.2$ 时,对应的聚类正确率最高,类内平均距离最小,类间平均距离最大.这是因为 SCC 数据集中时间序列的长度都为 60,即 ELCS 度量中序列长度 $m = n$,当 μ 太大时,不能很好的限定待测序列的位置对应信息,容易出现线性漂移,得到无意义的相似序列段;当 μ 太小时,相似性矩阵的搜索范围十分有限,出现漏检.所以随着 μ 值的减小,分类的正确率急剧下降.

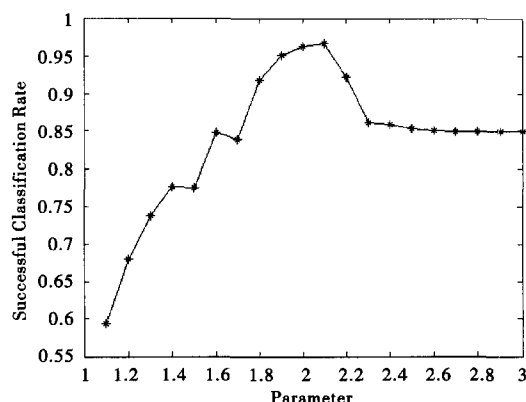


图 1 聚类的准确率

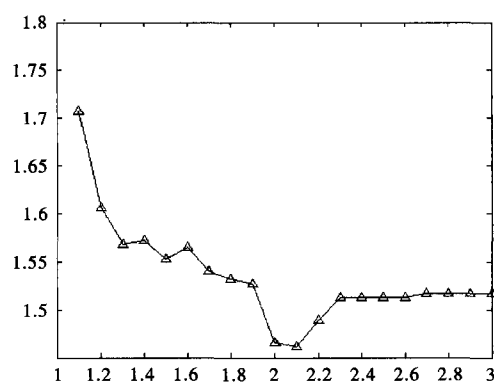


图 2 类内平均距离

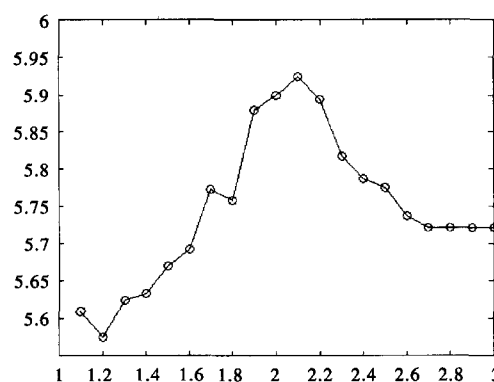


图 3 类间平均距离

4.3.2 基于三种测度的聚类比较

为比较 DTW、LCSS 和 ELCS 三种距离度量的耗时情况,分别将其作为层次聚类中的相似性度量标准.对 LCSS 和 ELCS 算法选取适当的参数,使得最终分类的准确率最高时,可以看到 DTW 算法的耗时明显高出其他度量,ELCS 度量因 $\frac{1}{\mu} < \frac{i_k}{j_k} < \mu$ 并且 $\frac{1}{\mu} < \frac{i_k - m}{j_k - n} < \mu$ 较 LCSS 度量的 $|i_k - j_k| \leq \delta$ 更为复杂,所以耗时有明显增加.

从 DTW、LCSS 和 ELCS 度量的聚类正确率比

较. 可以看到, 各种度量对于 SCC 数据集都有较好的效果, 因为 SCC 数据集本身数据的区分度较高, 数据的噪声较少. CBF 数据集因为是随机生成的数据集, 每条时间序列都有大量的毛刺, 无疑增加了聚类的难度. 但无论对哪一个数据集, ELCS 算法都表现出了较好的度量效果, 即聚类的正确率是最高的.

上面提到的数据集差异, 在图 4、图 5 中也可以看到, 相对于 SCC 数据集, CBF 数据集的聚类结果的类内平均距离更大, 而类间平均距离更小. 而基于 LCS 的 LCSS 和 ELCS 算法不存在 DTW 算法一点对多点的问题, 局部的噪声可以很好的忽略.

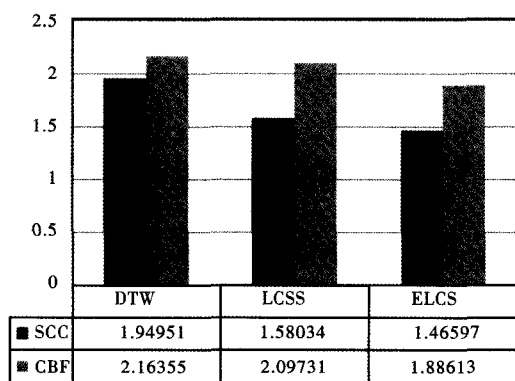


图 4 三种距离度量的聚类类内平均距离比较

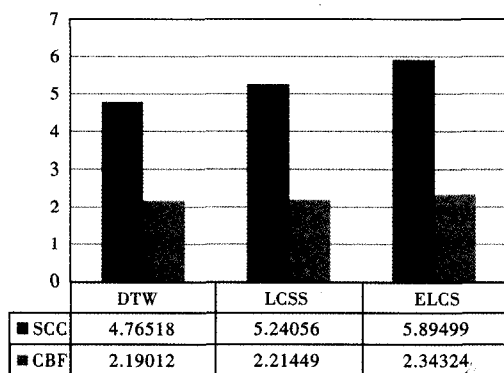


图 5 三种距离度量的聚类类间平均距离比较

5 结束语

本文针对传统聚类算法基于欧式距离度量相似性存在的缺陷, 以 LCS 度量为基础, 通过引入参数, 规范相似性矩阵的搜索路径, 提高了相似性度量的准确程度. 通过在两类数据集上的实验, 基于 ELCS 度量的聚类算法能够取得较高的聚类正确性, 在一定程度上, 这是以牺牲部分时间效率为代价的. 综合表明, 该聚类算法能有效地应用于各种时间序列的聚类应用.

参考文献:

- [1] Agrawal R, Faloutsos C, Swami A. Efficient similarity search in sequence database[C]// Proceedings of 4th International Conference on Foundations of Data Organization and Algorithms. Berlin: Springer, 1993: 69-84.
- [2] Faloutsos C, Ranganathan, Manolopoulos Y. Fast subsequence matching in time-series databases [J]. ACM SIGMOD Record, 1994, 23(2):417-429.
- [3] Chung F L, Fu T C, Ng V, et al. An evolutionary approach to pattern-based time series segmentation[J]. IEEE Transactions on Evolutionary Computation, 2004, 8(5):471-489.
- [4] Paterson M, Dancik V. Longest common subsequences [J]. Mathematical Foundations of Computer Science, 1994(841):127-142.
- [5] Bollobas B, Das G, Gunopulos D, et al. Time-series similarity problems and well-separated geometric sets [C]// Proceedings of the Thirteenth Annual Symposium on Computational Geometry. Nice: ACM Press, 1997:454-456.
- [6] 张雪丽, 牛强. 基于角点弯曲度的时间序列相似性搜索算法[J]. 计算机工程, 2011, 37(15):37-39.
- [7] Hawkins D M. Fitting multiple change-point models to data[J]. Computational Statistic & Data Analysis, 2008, 37 (3):323-341..
- [8] 吴学雁, 黄道平. 基于事件的时间序列相似性度量方法[J]. 计算机应用, 2010 (7):1944-1946.
- [9] 董晓莉, 顾成奎, 王正欧. 基于形态的时间序列相似性度量研究[J]. 电子与信息学报, 2007(5):1228-1230.
- [10] Keogh E J, Pazzani M J. Derivative dynamic time warping[DB/OL]. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.23.3383&rep=rep1&type=pdf>, 2001.
- [11] HAN Jiawei, KAMBER M. 数据挖掘概念与技术[M]. 2 版. 范明, 孟小峰, 译. 北京: 机械工业出版社, 2007.

作者简介:



李志刚 男, (1988—), 硕士研究生. 研究方向为时间序列分析与挖掘.

牛强 男, (1974—), 博士, 副教授. 研究方向为智能优化算法和数据挖掘.