时间序列聚类方法

聚类

10.1.1 什么是聚类分析？

聚类分析简称聚类，是把一个数据对象划分成子集的过程。每个子集是一个簇(cluster)，使得簇中的对象彼此相似，但与其他簇中的对象不相似。由聚类分析产生的簇的集合称做一个聚类。

显然，在相同的数据集上，不同的聚类方法可能产生不同的聚类。聚类可能导致数据内事先未知的群组的发现。

聚类分析应用广泛，包括商务智能、图像模式识别、Web搜索、生物学和安全等。

聚类分析可作为一种独立的工具，用来洞察数据的分布，观察每个簇的特征，将进一步分析集中在特定的簇集合上。聚类分析亦可作为其他算法（如特征化、属性子集选择和分类）的预处理步骤，

之后这些算法将在检测到的簇和选择的属性或特征上进行操作。

聚类又称作数据分割，因其根据数据的相似性把大型数据集合划分成组。聚类还可以用于离群点检测，其中离群点可能比普通情况更值得注意。离群点检测的应用包括信用卡欺诈检测和电子商务中的

犯罪活动监控。

在机器学习领域，分类被称作监督学习。分类预先给定了类标号信息，即学习算法是监督的，因其被告知每个训练元组的类隶属关系。聚类被称为无监督学习，因其没有提供类标号信息。聚类通过

观察学习，而非通过示例学习。

10.1.3 基本聚类方法概述

1）划分方法(partitioning method)

给定一个n个对象的聚合，划分方法构建数据的k个分区，其中每个分区表示一个簇，并且k<=n。

大部分划分方法基于距离。给定要构建的分区数k,划分方法首先创建一个初始划分。然后，它采用一种迭代的重定位技术，通过把对象从一个组移动到另一个组来改进划分。

为达到全局最优，基于划分的聚类可能需要穷举所有可能的划分，计算量极大。实际上，大多数应用都采用了流行的启发式方法，如k-均值和k-中心点算法，渐进地提高聚类质量，逼近局部最优解。

2）层次方法(hierarchical method)

层次方法创建给定数据对象集的层次分解。根据层次分解如何形成，层次方法可以分为凝聚的或分裂的方法。凝聚的方法，自底向上，开始将每个对象作为单独的一个组，然后逐次合并相近的对象和组，

直到所有的组合并成一个组（层次的最顶层），或满足某个终止条件。分类的方法，自顶向下，开始将所有的对象置于一个簇中。在每次相继迭代中，一个簇被划分成更小的簇，直到最终每个对象在单独的

一个簇中，或者满足某个终止条件。

层次聚类方法可以是基于距离或者基于密度和连通性的。层次聚类方法的一些扩展也考虑了子空间聚类。

3）基于密度的方法(density-based method)

4）基于网格的方法(grid-based method)

时间序列的聚类方法

1.传统的聚类方法针对静态数据，即特征不随时间变化的数据。时间序列的特征包含随时间变化的值，并非静态数据，对于时间序列的聚类来说，有以下两种思路：

1.修改现有的聚类方法以适用于时间序列

2.将时间序列转化为静态数据，再使用现有的聚类方法

第一种思路直接作用于原始数据，称为基于原始数据的方法，难点在于要找到适用于时间序列的距离/相似性度量。

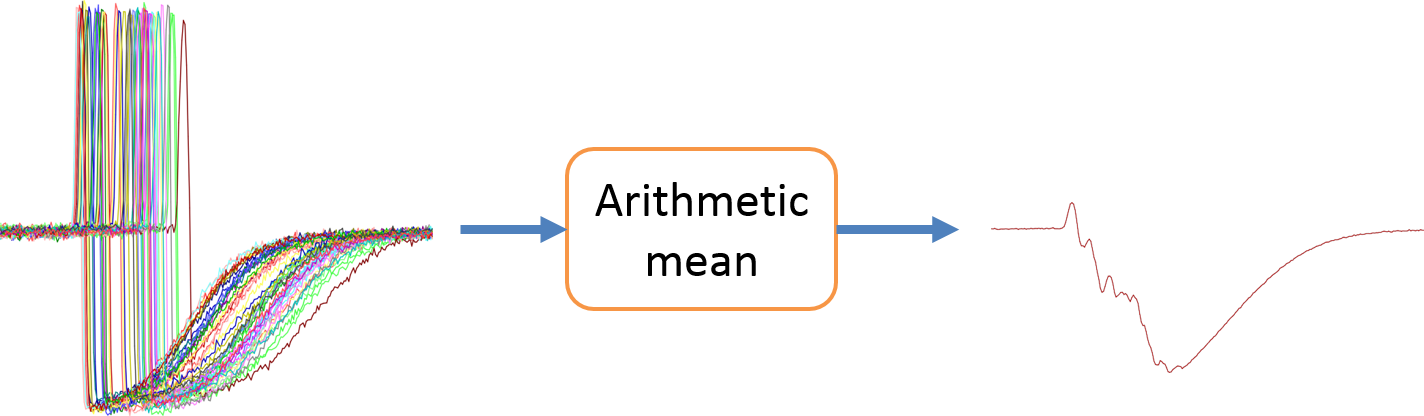
1. DTW/DBA+KMedoids算法（https://blog.csdn.net/qq\_37960007/article/details/108000094）

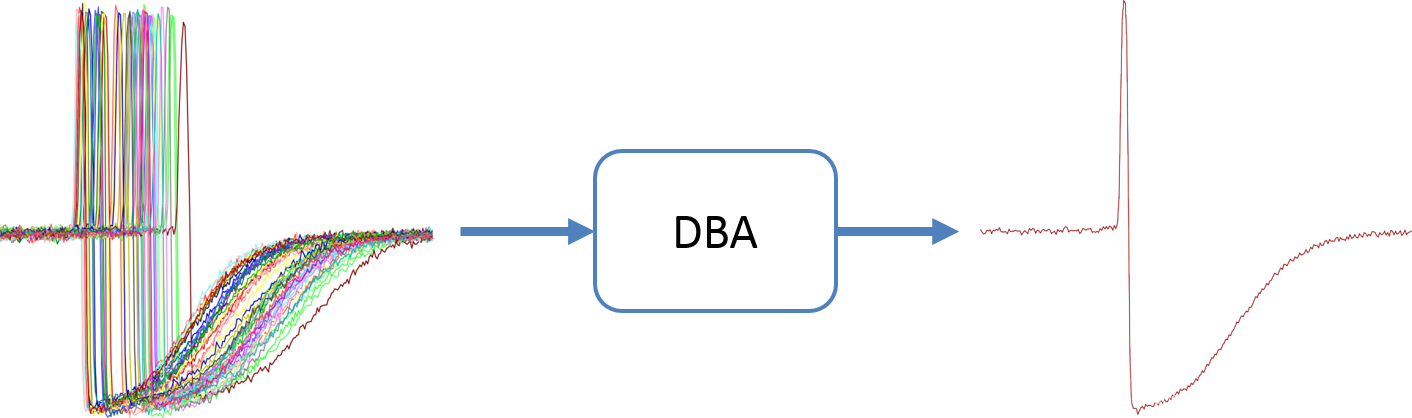
DTW(Dynamic Time Warping):

DBA(DTW Barycenter Averaging)

DBA 是在DTW的基础上做重心平均，为什么这么做，因为往往我们需要比较的时间序列不是一条而是一个集合，所以通过把时间序列压缩成一条来进行比较.

DBA代表动态时间扭曲重心平均。DBA是一种与动态时间扭曲一致的平均方法。下面给出一个时间序列集合的传统算术平均值与DBA之间的区别的例子。





第二种思路首先将原始的时间序列数据转换为低维的特征向量或若干模型参数，称为基于特征的或基于模型的方法。

例如金融交易数据普遍存在信噪比低、非线性、非平稳、非正态的特点，因此直接利用原始数据进行聚类往往效果不好。所以第二种思路中的两种方法是研究的重点。

这两种方法都是通过一定的方式将原始时间序列转化为静态的特征或模型参数，然后通过传统的静态数据的聚类方法进行聚类。

基于模型的方法认为相似的时间序列应该产生相似的模型，所以通过比较模型之间的相似性可以达到时间序列聚类的目的。困难在与模型选择和参数确定两个问题。常用的模型：

·ARMA模型

·HMM模型

·马尔科夫链

基于特征提取的聚类主要应用于高维时间序列，特别是高频金融时间序列，经过特征提取后可有效降维。针对时间序列的数学特性。

对时间序列进行特征提取，用提取的特征项对时间序列进行时间序列的重新描述，然后对重新描述的时间序列聚类。

2.时间序列聚类的步骤

·数据简化

·两个时间序列相似性的度量

·时间序列聚类的通用算法

·评价聚类结果的标准

2.1数据简化

·DFT离散傅里叶变换

·DWT离散小波变换

·SVD奇异值分解

·PLA分段线性估计

·PAA分段聚合近似

·SAX符号化聚合近似

2.2相似性/距离度量

·欧式距离、Minkovski距离

·Pearson相关系数

·DTW(Dynamic Time Warping)

·DBA(DTW Barycenter Averaging)

2.3时间序列聚类的通用算法

2.4评价聚类结果的标准

·有ground truth：将聚类结果与ground truth比较

Rand Index

Mutual Information

Cluster Similarity Measure(CSM)

Cluster purity

...

·无ground truth: 用于比较不同聚类方法得到的结果

SSE （均方误差）

Silhouette Coefficient (轮廓系数）

...

时间序列分类方法

1.1 分类的定义：

分类和数值预测是预测问题的两种主要类型。

分类是一种数据分析形式，提取刻画数据类的模型。通过构建称为分类器的模型，用于预测数据的类标号。若要对一个连续值函数或有序值进行数值预测，而非确定类标号，则构建的模型为预测器。回归分析是数值预测最常用的统计学方法，时常可将回归分析与预测器作为同义词使用。

1.2 分类的一般方法

数据分类包括两个阶段：

1. 学习阶段（构建分类模型）

建议描述预先定义的数据类或概念集的分类器。分类算法通过分析或从训练集中“学习”来构造分类器。训练集由数据库元组和相关联的类标号组成。

2）分类阶段（使用模型预测给定数据的类标号）

2.2 决策树归纳

决策树归纳是从有类标号的训练元组中学习决策树。决策树的结构类于树状的流程图，其中每个内部结点表示在一个属性上的测试，每个分枝代表该测试的一个输出，而每个树叶节点存放一个类标号。树的最顶层结点是根节点。

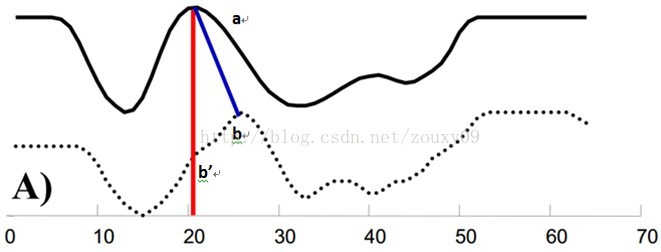
2.3 SVM(Support Vector Machine, SVM)支持向量机

SVM是一类按监督学习方式对数据进行二元分类的广义线性分类器，其决策边界是对学习样本求解的最大边距超平面。支持向量(Support Vector)指离分隔超平面最近的那些点，寻找最大间隔指寻找最大化支持向量到分隔超平面的距离，在此条件下求出分隔平面。

2.4 DTF+K-NN

2.4.1 DTW(Dynamic Time Warping，动态时间归整)

时间序列是数据的一种常见表示形式。对于时间序列处理来说，一个普遍的任务就是比较两个序列的相似性。



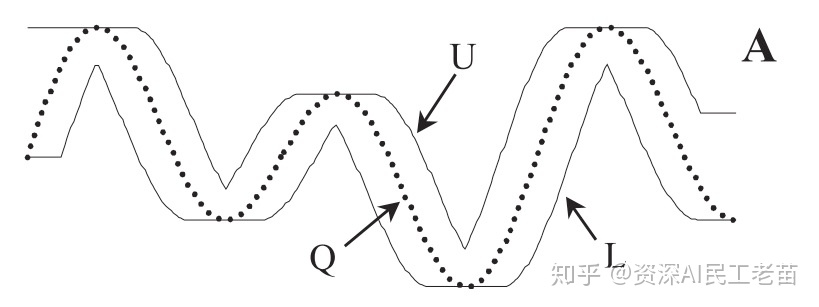
大部分情况下，两个序列整体上具有非常相似的形状，但是这些形状在X轴上并不是对齐的。所以在对其进行相似度比较时，需将其中一个（或多个）序列在时间轴下进行扭曲，以达到更好的对齐。如上图所示，a应对应扭曲后的b点，而非与具有相同x值得b’对齐。

DTW算法的时间复杂度很高，匹配长度分别为m和n的两个时间序列时，其时间复杂度为。在搜索数据很大时，逐个用DTW算法比较每一条是否匹配非常耗时，可使用一种计算较快的近似算法，LB Keogh算法(Lower Bounding Keogh algorithm)。通过LB处理掉大部分不可能是最优匹配序列的序列，对于剩余的序列再使用DTW逐个比较。

LB\_keogh距离定义：

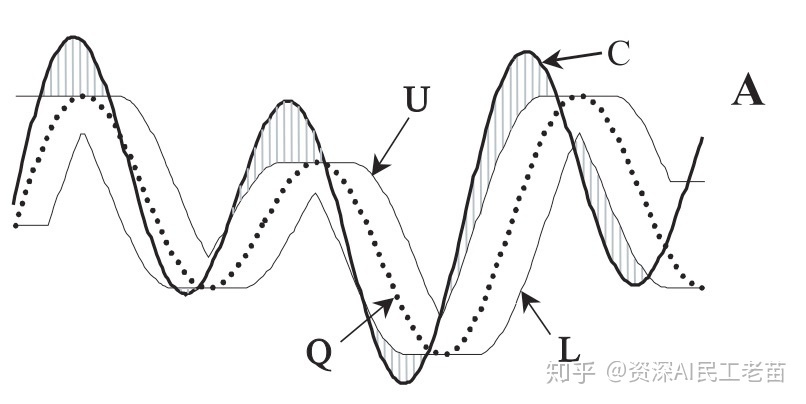
第一部分为Q的{U, L}包络曲线，给Q序列的每个时间步长定义上下界。定义如下

其中r是一段滑行窗距离，可自行定义。示意图如下



U为上包络线，及在当前时间步为Q，滑动窗口为r的序列中的最大值。L为下包络线，同理。

如图



阴影部分为对LB的贡献。