时间序列分类问题的算法比较

杨一鸣[1)] 潘 嵘[2)] 潘嘉林[2)] 杨 强[1),2)] 李 磊[1)]

[1)](中山大学软件研究所 广州 510275)

[2)](香港科技大学计算机科学与工程系 香港)

摘 要 时间序列分类是时间序列数据分析中的重要任务之一.不同于时间序列分析中常用的算法与问题,时间序列分类是要把整个时间序列当作输入,其目的是要赋予这个序列某个离散标记.它比一般分类问题困难,主要在于要分类的时间序列数据不等长,这使得一般的分类算法不能直接应用.即使是等长的时间序列,由于不同序列在相同位置的数值一般不可直接比较,一般的分类算法依然还是不适合直接应用.为了解决这些难点,通常有两种方法:第一,定义合适的距离度量(这里,最常用的距离度量是DTW距离),使得在此度量意义下相近的序列有相同的分类标签,这类方法属于领域无关的方法;第二,首先对时间序列建模(利用序列中前后数据的依赖关系建立模型),再用模型参数组成等长向量来表示每条序列,最后用一般的分类算法进行训练和分类,这类方法属于领域相关的方法.长期以来,研究者往往只倾向于使用其中一种算法,而这两类算法的比较却比较缺乏.文中深入分析了这两类方法,并且分别在不同的合成数据集和实际数据集上比较了两类方法.作者观测到了两类算法在不同因素影响下的性能表现,从而为今后发展新的算法提供了有力依据.

关键词 分类;时间序列;基于模型聚类;马尔可夫模型;统计学习

中图法分类号 TP311

A Comparative Study on Time Series Classification

YANG Yi-Ming[1)] PAN Rong[2)] PAN Jia-Lin[2)] YANG Qiang[1),2)] LI Lei[1)]

[1)](*Software Institute, Sun Yat-Sen University, Guangzhou* 510275)

[2)](*Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong*)

Abstract Time series classification or categorization is an important task in time-series analysis. Unlike traditional methods and problem formulations in time-series analysis, time series classification aims to take whole time sequences as input, and produce discrete labels that are assigned to each sequence. Compared to traditional classification problems, time series classification poses additional difficulties. A major difficulty is due to the fact that the time sequences are variable in length, making many traditional classification methods unable to apply directly. Even for sequences of uniform lengths, many methods can still not be applied directly because often the data located at different parts of the sequences are incomparable. Two methods have been tried separately in the past, including distance based methods such as DTW, and model based methods such as Markov models. Using either of these methods as preprocessing steps, a uniform length vector space can be built to enable the classification methods to be applied. In the past, there has been a lack of comparison between these two methods. This paper compares distance and model based methods on several data sets including synthetic and real data sets, to explicate the relative advantages and disadvantages of these methods. This paper presents several key observations on

收稿日期:2007-03-05;修改稿收到日期:2007-05-24.杨一鸣,男,1964年生,博士研究生,教授级高级工程师,研究方向为数据挖掘与数据分析. E-mail: yangym@gsta.com. 潘 嵘,男,1976年生,博士,研究方向为数据挖掘、基于案例推理. 潘嘉林,男,1980年生,博士研究生,研究方向为机器学习、数据挖掘. 杨 强,男,1961年生,副教授,博士生导师,研究领域为数据挖掘、智能规划、基于案例推理、机器学习. 李 磊,男,1951年生,教授,博士生导师,研究领域为数据库、数据挖掘、人工智能.

the relative merits of these two methods, and paves the way for further research in developing new methods for time series classification.

Keywords classification; time series; model based clustering; Markov model; statistical learning

1 引 言

时间序列数据分析在机器学习、数据挖掘以及数据仓库等领域中日益重要. 时间序列分类是时间序列数据分析中的重要任务之一. 本文研究的时间序列的分类问题定义如下:给定一个数据样本集合,每个数据样本包括:一个输入时间序列 $\boldsymbol{X}_i=\boldsymbol{x}_i^{(1)}$, $\boldsymbol{x}_i^{(2)},\cdots$及其离散的分类标签 C_i,其中,$\boldsymbol{x}_i^{(t)}\in R^n$是一个 n 维向量,$C_i\in\{1,2,\cdots,N_C\}$,目标是预测新的时间序列的类标签.

时间序列分类已经应用到很多不同的领域. 最早是 HMM 和 DTW 在语音处理和语音识别领域上的应用[1-3]. Berndt 和 Clifford[4]在 1994 年,首先把 DTW 引入到数据库领域中. 近十年来,时间序列分类被研究人员应用到更多的领域中,例如,生物信息学中的 RNA 数据的分类[5]、心电图 ECG 的模式匹配[6]、化学工程[7]以及手写识别[8]等等.

时间序列分类问题与一般分类问题的区别主要在两点:(1)时间序列的长度不相等,不能直接把每条序列看作是一个属性向量,作为一般分类算法的输入. 因而,一般的分类算法不能直接应用. (2)对于所有序列长度都相等的时间序列分类问题,由于不同序列的相同的属性值(不同序列在相同位置的数据值)不一定可比,如果直接套用一般的分类算法,如 SVM[9],k-近邻搜索[10]等,效果也不一定好. 以上两点即是时间序列分类问题的特点,又是它的难点.

为了解决这些时间序列分类问题特有的难点,学者们进行了大量的研究. 我们可以把这些研究工作分为两类:第一类是与领域无关的方法,这类方法的代表是动态时间变形(Dynamic Time Warping, DTW)算法[3,11]. DTW 是利用动态规划算法寻找两条序列的最优匹配,从而定义两条时间序列的距离度量. 我们将在本文中更详细地介绍 DTW. DTW 不需要利用领域知识,只是假设相近的序列存在低消耗(low cost)的平移匹配. 第二类是领域相关的方法. 这类方法主要是利用领域知识,对每条时间序列提取等长的特征向量,然后利用一般的分类算法进行训练和分类. 最常用的领域知识是序列中前后数据的依赖关系有一定规律,一般认为马尔可夫模型能较好地表示这一规律[8,12].

在以往的研究中,学者们通常在一个文献中只讨论一类方法. 但是,经过分析和比较,我们发现这两类算法各有优缺点,适用范围不同. 基于这一出发点,本文从三方面系统的比较了两类方法. (1)研究两类算法受训练样本大小的影响;(2)研究序列长度对算法的影响;(3)比较不同噪声情形下,两类算法的性能变化. 我们曾在文献[12]中介绍了一种基于模型的聚类方法,把变长的时间序列数据转变成固定维数的向量. 本文的重点在于系统地比较两类时间序列分类算法. 在本文后面部分,我们将讨论两类时间序列分类算法不同的适用范围.

2 问题描述

我们首先定义时间序列的分类问题. 给定一个序列集合 $\boldsymbol{S}=\{\boldsymbol{S}_1,\boldsymbol{S}_2,\cdots,\boldsymbol{S}_M\}$,其中每一个序列 S_i 由元素 $\boldsymbol{x}_i^{(t)}\in R^n(t=1,2,\cdots)$组成,而每一个元素 $\boldsymbol{x}_i^{(t)}$ 由一个属性值向量$(x_i^{t,1},x_i^{t,2},\cdots,x_i^{t,N})$组成. 例如,在我们 CRM 问题中,每一个 x_i^t可以是一个月的账单,包括不同种类的话费和其它费用. 每一个序列和一个类标签 $C_i(i=1,2,\cdots,l)$关联. 时间序列分类的目标是得到测试集合 S_{test}的分类结果,使得该结果达到某个量度标准. 例如,其中一个量度标准是测试结果的准确率. 在这里,我们使用一个更有效的量度标准 AUC,它是 ROC 曲线下的面积[13]. 我们将在本文后面的章节详细说明 AUC 度量.

3 领域无关的时间序列分类算法

过去十多年来,在时间序列分类问题上,出现了大量的研究工作. 在数据挖掘和数据库领域,基于距离的时间序列分类方法十分常用. 在文献[11,14-15]中,学者们提出了基于欧氏距离的 k-近邻搜索算法. 距离的计算是用查询序列与数据库里每条序列的欧氏距离. 容易看出,基于欧氏距离的算法主要有两个问题:(1)不能处理不等长序列;(2)时间的移位会对搜索结果造成很大的影响.

为了解决这些问题,需要对等长或不等长的序列进行移位匹配,这其实是一种基于动态规划的方

法. Berndt 和 Clifford[4] 在 1994 年，首先把动态时间变形(Dynamic Time Warping，DTW)引入到时间序列分类中. 此后，大量基于 DTW 的序列分类问题的研究表明，DTW 在很多时间序列分类的数据集①上有很好的性能. 下面，我们简单回顾基于 DTW 的距离计算.

给定两条时间序列 $T=t_1,t_2,\cdots,t_n$ 和 $R=r_1,r_2,\cdots,r_m$. DTW 的目标是寻找 T 和 R 之间的最优匹配 ϕ，使得匹配后的局部距离之和最小. 具体地，$\phi=(\phi_T,\phi_R)$，其中，$\phi_T=(\phi_T^1,\phi_T^2,\cdots,\phi_T^{K_\phi})$ $(1\le\phi_T^i\le n,\ 1\le i\le K_\phi)$ 并且 $\phi_R=(\phi_R^1,\phi_R^2,\cdots,\phi_R^{K_\phi})$ $(1\le\phi_R^i\le m,\ 1\le i\le K_\phi)$. 这样 T 和 R 的 DTW 距离定义如下：

$$DTW(T,R)=\min_{\phi}(D_\phi(T,R)),$$

其中，$D_\phi(T,R)=\frac{1}{M\phi}\sum_{k=1}^{K_\phi}d(t_{\phi_T^k}-r_{\phi_R^k})m_k$，$M\phi=\sum_{k=1}^{K_\phi}m_k$，$m_k$ 是第 k 个局部匹配的权值. 在一般的 DTW 距离的计算中至少有以下 3 个约束：

(1) 边界点约束：$\phi_T^1=\phi_R^1=1$，$\phi_T^{K_\phi}=n$ 和 $\phi_T^{K_\phi}=m$；

(2) 单调性约束：$\phi_T^k\le\phi_T^{k+1}$ 和 $\phi_R^k\le\phi_R^{k+1}$；

(3) 连续性约束：$\phi_T^{k+1}-\phi_T^k\le1$ 和 $\phi_R^{k+1}-\phi_R^k\le1$.

另外，在文献[11]中，Keogh 等人还介绍了其他不同的局部以及全局的约束. 这些约束在一定程度上降低了 DTW 的计算复杂度. 但是由于 DTW 是用动态规划计算的，在实际的应用中，k-近邻的计算开销还是很大. 有很多研究人员提出了不同的 DTW 的下界计算以进一步降低基于 DTW 的 k-近邻的计算复杂度[11,16-17]. DTW 的下界(Lower Bound，LB)计算的基本思想如下：在计算查询(query)序列与数据库中每条序列的 DTW 距离前，先计算它们的 DTW 距离的 LB，当 LB 小于当前最小 DTW 距离时，计算它们的 DTW 距离；否则，跳过 DTW 距离的计算. 可以看出，如果 LB 足够紧，可以有效减少 DTW 距离的计算次数；同时，如果 LB 的计算复杂度比 DTW 的计算复杂度要低，这种基于 DTW 的下界的 k-近邻搜索算法会降低计算开销.

以上我们回顾了一类领域无关的算法，在接下来两部分，我们将介绍一种基于模型的领域相关的时间序列分类算法.

4　基于分类的时间序列数据变换

4.1　概　述

时间序列分类问题与一般的机器学习、模式识别的分类问题有很大的不同. 一般的分类问题的输入 $x_i^{(a)}$ (input attributes)是等长的向量；然而，时间序列分类问题的输入是一些不等长的序列，这使得大多数标准的分类方法，如 k-近邻搜索[10]、决策树[18]、最大似然方法[19]、SVM[9]等方法不能直接使用. 另外，本文讨论的时间序列中的每个数据元素可以是一个多维向量，而不像某些股票分析数据集中的数据，只是一个实数. 这些问题使很多时间序列分析方法不适用.

我们的方法分两步. 第 1 步，将数据转化成等长向量. 一个关键问题是如何保留尽量多的时间和序列的信息. 在下一小节，我们将利用一个基于模型的聚类方法来进行数据的转化，由于建模利用了序列前后的依赖关系的特征，而这些特征是领域相关的知识，所以我们的这个方法属于领域相关的算法. 第 2 步，在经过转化的等长数据集上使用一般的分类算法作分类.

4.2　基于聚类模型的数据转换

算法框架如下：

算法首先根据已有的类标将训练集分为正例集和负例集，分别记为 DB_+ 和 DB_-. 然后再运用基于模型的聚类算法对正、负例集分别聚类. 特别地，我们将正、负例集分别聚成 p_+ 类和 p_- 类. 至于如何得到这些聚类将会在下一节有具体的介绍. 第 4 步，是将算法遍历所有的训练集. 一次循环就是一条数据根据最大似然法得到它在每一个聚类中的度量，这个度量产生一系列的概率值，一个概率值与一个聚类对应. 结果就是得到一条新的向量 ***vector***，维数为 p_++p_-，然后将这些向量组合成一个向量集合 ***Vectors***，这样就得到了经过转换的新的数据集. 算法 1 只讨论二维的情形，但事实上，我们可以很容易把它推广到多维的情形. 下一小节我们将针对算法的每个步骤的各个细节进行详细的介绍.

算法 1.　时间序列转换算法.

输入：训练集 DB，正、负例分别所聚的类数 p_+，p_-

输出：转换后的训练集向量

步骤：

1. 将训练集分为正例集 DB_+ 和负例集 DB_-；
2. 正例的类＝基于模型的聚类(DB_+,p_+)；
   负例的类＝基于模型的聚类(DB_-,p_-)；
3. $Model$＝(正例的类，负例的类)；***Vectors***＝{}；
4. For $seq_i\in TrainDB$
5. $\boldsymbol{vector}_i=maxlikelihood(seq_i,Model)$；

① Keogh E, Xi X, Wei L, Ratanamahatana C A. The UCR time series classification/clustering homepage: www.cs.ucr.edu/~eamonn/time_series_data/

6. $Vectors = Vectors \cup vector_i$;

7. end For

8. Return *Vectors*.

4.3 基于模型的聚类

聚类算法的第 2 步根据输入的正、负例集合建立一些类. 但是一条时间序列的组成可能是非常复杂的,可能同时包括类别数值和连续数值. 因而,我们首先的任务是运用基于模型的聚类算法将这些数据统一转换为离散的状态. 先将所有的数值特征离散化,这一步可以通过一些标准的监督或非监督的离散化算法实现. 在这里,我们就是通过这种途径对数据集中的所有时间序列做了离散化处理(在实验中,我们是用 k-means[20] 聚类算法来获得状态). 通过这一步就得到了 Q 个状态:$S_i(i=1,2,\cdots,Q)$,其中 Q 是一个参数,这可以通过调整得到恰当的值.

从这些状态信息中可以看出,每个时间序列转换为一条状态转移链. 我们运用基于 EM 的聚类算法分别得到正、负例集(DB_+,DB_-)的聚类个数,即 p_+,p_- 的值. 假设马尔可夫模型包含 Q 个状态,而在每个聚类中又有 k 条马尔可夫链,又假设先验状态分布 $Pr(s_i)=a_i$,转移矩阵为

$$v_p=\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix},\quad T_p=\begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ t_{21} & t_{22} & \cdots & t_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ t_{n1} & t_{n2} & \cdots & t_{nn} \end{pmatrix},$$

其中,t_{ij} 是表示从类 s_i 到类 s_j 的转移概率. 那么就有

$$a_i=\frac{\sum_{t=1}^{t=k} I(c_t=i)}{k},$$

并且

$$t_{ij}=\frac{\sum_{t=1}^{t=k}\sum_{s=1}^{s=s_t-1} I(c_{ts}=i,c_{t(s+1)}=j)}{\sum_{t=1}^{t=k}\sum_{s=1}^{s=s_t-1} I(c_{ts}=i)},$$

其中,如果 x 为真,$I(x)=1$,否则 $I(x)=0$. 通过输入数据,我们可以得到如下形式的包含 K 个类的混合马尔可夫模型:

$$p_k(v|\theta)=\sum_{k=1}^{K} p(c_k|\theta)p_k(v|c_k,\theta),$$

其中,$p(c_k|\theta)$ 是指第 k 类的边沿概率并且 $\sum_{k=1}^{K} p(c_k|\theta)=1$,$p_k(v|c_k,\theta)$ 是指统计模型描述一条时间序列属于类 k 的概率,θ 表示模型的参数. 这里的 $v=v_1v_2\cdots v_L$ 是一个随机长度的特征向量序列,我们假设每个模型均是一阶马尔可夫模型,可表示成

$$p_k(v|c_k,\theta)=p(v_1|\theta_k^I)\prod_{i=2}^{L} p(v_i|v_{i-1},\theta_k^T),$$

其中,θ_k^I 和 θ_k^T 均是概率分布的参数,前一个概率分布是通过计算类 k 中所有时间序列的初始特征向量得到的,后一个则是通过类 k 中一个特征向量到另一个特征向量的转换得到. 这个模型刻画了时间序列中不同状态之间转换的特征,包括他们的原始状态的特征和两个连续状态之间的转换关系.

5 利用一般的分类算法对时间序列数据进行的分类

根据上述算法,我们得到了 $K=p_-+p_+$ 个聚类,在本文中,我们取 $p_+=1$. 对任意给定的时间序列,都可以计算出这条序列在 K 个聚类中的概率分布,然后用一个向量来表示 $V=(p_1,p_2,\cdots,p_K)$. 通过这个方法可以将所有的训练集和测试集转换为长度为 K 的向量. 这个概率度量将会用来对所有的测试集进行排序,从而得到一条排好序的列表,在实验中将会利用原有的类标和这个排好序的列表计算出 AUC 的值.

在实际中可以尝试任意一个标准的分类算法,例如,k-近邻搜索、最大似然法等等,对时间序列数据进行分类. 具体的,我们首先将测试数据序列转换为状态向量形式:$S_1,S_2,\cdots,S_K$. 进一步可以生成长度为 K 的向量 V_i,其中 K 表示聚类的个数. 这个向量 V_i 的每个元素就是这条序列属于不同聚类模型的概率. 这样我们就可以把训练和测试数据转换为相同的向量. 根据这个转换后的向量,我们可以用一般的分类算法进行分类.

6 实验与分析

在这一部分,我们将讨论在合成数据集和真实数据集上的实验,以比较两类算法的性能和不同特点. 在实验中,我们比较了 3 个算法,基于 DTW 距离的 k-近邻搜索算法属于领域无关的算法,另外两个是基于模型聚类的最大似然估计(MC-MLE)和 k-近邻搜索算法(MC-k-NN),它们属于领域相关的算法. 在时间序列分类问题中,有很多因素会影响分类算法的性能. 训练样本的大小、序列中的噪声和序列长度是 3 个很重要的方面. 为了能够系统地进行实验比较,我们使用仿真的隐马尔可夫模型产生数据,以控制这 3 个因素. 另外,我们还在 8 个公开数

据集以及一个检查电信客户逃离的数据集上比较了3个时间序列分类算法.

我们合成的是一个两类的时间序列分类的数据集.每一类由一个随机产生的隐马尔可夫模型产生1000条序列.隐马尔可夫模型隐层状态数量为5,隐层节点到观测节点由高斯分布产生.我们把每一类的500条序列用作训练样本,其余的作为测试数据,即训练和测试数据各包含1000条序列.

在第1组实验中,我们变换训练数据的数量,从50～1000,测试3个算法在测试数据上的错误率(如图1所示).我们可以发现,两个算法MC-k-NN与MC-MLE在这个数据集上比DTW算法的错误率小.另外,MC-k-NN与MC-MLE算法受训练样本影响不大,在训练样本很少的情况下能够得到理想的效果.相比之下,DTW算法随着训练样本减少,误差上升较快.我们相信,基于模型的算法由于利用了领域知识,因此需要的训练样本较少.

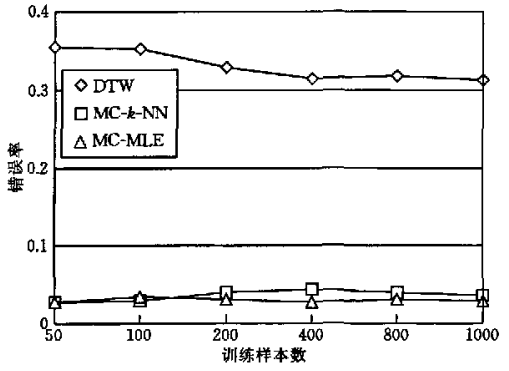

图1　训练样本大小对算法的影响

在第2组实验中,我们变换每条时间序列的长度,由50～400,训练样本与测试样本大小均为1000.从图2可以看到,MC-k-NN与MC-MLE算法的错误率在不同的序列长度上都比DTW算法小.随着序列长度的增加,3个算法的错误率总体上都往下降,并且,MC-k-NN与MC-MLE算法与DTW算法的错误率的差距增大.

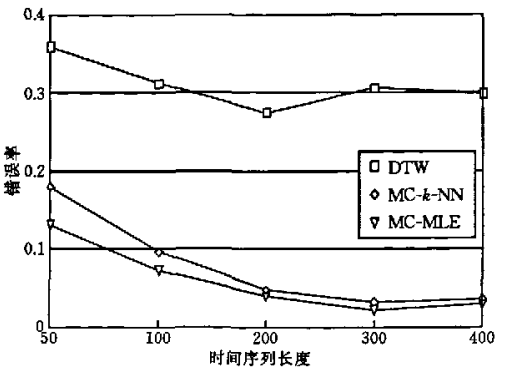

图2　序列长度对算法的影响

在下一组实验中,我们改变隐马尔可夫模型中高斯的标准方差,由0.125～2,这相当于在数据中加入不同的噪声.在这组实验中,训练样本与测试样本大小还是1000.实验结果如图3.我们可以看出虽然MC-k-NN与MC-MLE算法的错误率比DTW算法小,但随着噪声增加,MC-k-NN与MC-MLE算法的误差越来越接近DTW的误差.这说明,DTW算法受噪声的影响相对较少.

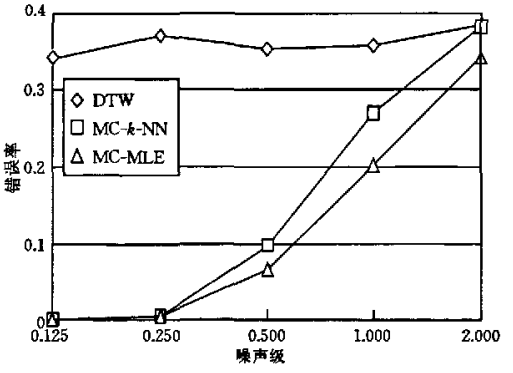

图3　噪声的影响

我们的第2个实验数据来源于前面注①中的8个时间序列分类数据集.这些数据集的基本属性在表1中第2列～第5列显示.表1中的最后3列是DTW,MC-k-NN和MC-MLE的分类错误率.在这

表1　前面注文①中的8个时间序列分类数据集的属性及3个算法的错误率(最后3列)

数据集	分类数	训练集大小	训练集大小	时间序列长度	错误率		
					DTW	MC-k-NN	MC-MLE
Synthetic Control	6	300	300	60	0.007	0.08	0.0867
Face (all)	14	560	1690	131	0.192	0.4183	0.4219
OSU Leaf	6	200	242	427	0.409	0.4545	0.4711
Swedish Leaf	15	500	625	128	0.21	0.2848	0.2912
50Words	50	450	455	270	0.310	0.622	0.6923
Trace	4	100	100	275	0.0	0.0	0.0
Adiac	37	390	391	176	0.396	0.3939	0.4118
Fish	7	175	175	463	0.167	0.3029	0.32

些数据集上，领域无关的算法(DTW)的错误率相对较低.经过观察，这些数据比较规则，同类的数据中容易出现形状"相似"的序列，因而在这类数据集上，我们认为用 DTW 这类领域无关的分类算法较好.

我们的第 3 个实验数据来源于电信客户关系管理(CRM)小组进行客户分类所用的数据.在我们的 CRM 的例子中，一个序列是一个客户的月账单的连续序列 $S_i=x_i^{(1)},x_i^{(2)},\cdots$，其中 $x_i^{(t)}$ 表示用户在某个月使用的电信服务的情况.例如有长途服务、区域服务、直呼服务等等.每一个 $x_i^{(t)}$ 中的元素是一个实值向量$\langle e_1,e_2,\cdots,e_k\rangle$，其中每一个 e_i 是用户在相应月份支付相应服务的钱.如果我们使用离散整数记录每一种服务，那么从一个服务日志中得到的序列可以用表 2 表示.一个序列由电信客户的电话呼叫记录组成.其中每一个电话呼叫记录由呼叫的持续时间、服务种类、费用、时间、日期和其它很多因素组成.

表 2 服务日志中抽取的有分类标签的序列数据

索引号	服务队列	类别
1	17 2 17 16 16 16	+
2	16 27 27 16 19 18	−
3	20 20 23 23 24 17 17 16 16 16	+
4	16 16 17 17 17 16 16	+
5	16 1 2 6 26	+
6	15 16 27 21 27 19 16	−

我们用数量庞大的电信客户月帐单数据来作为实验的训练集和测试集.下面，我们介绍实验数据是如何构建的，同时给出聚类模型和排序模型的实验结果.

我们在电信的数据库中随机的抽取 20000 个客户的数据.这里包括 96%的非逃离客户(负例)和 4%的逃离客户(正例).每个客户都对应一系列月帐单记录，这些记录由客户的月通话记录组成，它们的时间长度在 4～24 个月之间不等.最后我们会将数据十等分法(10−fold)来做交叉验证.

我们实验的设计是用来测试本文提出数据转换算法区分正负例的能力，如前所述，由于数据中正负样本的比例不平衡，我们用 AUC 这个度量标准.AUC 是用 ROC 曲线下的面积来表示.根据文献[13]，可以用公式

$$AUC=\frac{\sum_{i=1}^{n_0}(r_i-i)}{n_0n_1}$$

来计算 AUC 的值，其中 n_0 和 n_1 分别是正、负例子的数目，而 r_i 是第 i 个正例的位置.

在实验中，我们还是比较 DTW，MC-k-NN 和 MC-MLE 3 个算法.我们的实验是通过变换在特征空间上所聚的类数进行的.当所聚的类数越大，即得到的状态越多，从一个状态转换到另一状态的分布就越趋于均匀分布，并且聚类转换的实现难度也越高.结果已经显示在图 4 中.从图中可以看出，不论是在多少状态数上，MC-MLE 算法所得到的结果有明显的优势，DTW 的在 AUC 上的性能较差.

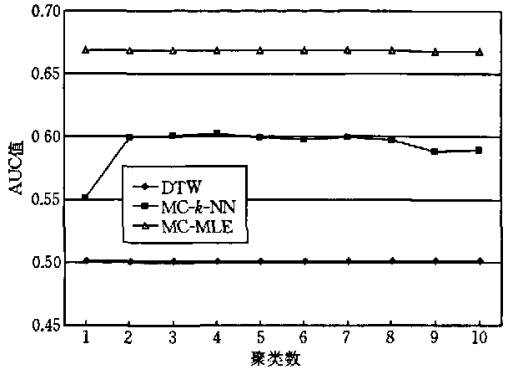

图 4 3 个算法在不同状态数上所计算的结果

这是因为我们先用聚类转换算法对数据进行转换，然后再结合基于 MLE 的算法可以更好地利用 AUC 这个度量的理论原理.通过对变换不同聚类数进行实验可以进一步验证这一点.事实上我们可以看出在聚类的类数从小到大的变化过程中，MLE 算法始终表现出很强的鲁棒性.所以，我们整个算法过程的一个优点就是可以用合成的数据训练出一个对客户进行分类排序的鲁棒性很强的分类器.

7 结 论

时间序列分类是时间序列数据分析中的重要任务之一.它比一般分类问题困难的原因主要在于要分类的时间序列数据不等长，这使得一般的分类算法不能直接应用.即使是等长的时间序列，由于不同序列在相同位置的数值一般不可直接比较，一般的分类算法依然还是不适合直接应用.本文分析了两类方法，并且分别在不同的合成数据集和实际数据集上比较了领域无关和领域相关的两类方法.我们发现在训练数据较少时，使用领域相关的算法比较合适；另一方面，领域无关的算法受噪声的影响相对较少.在电信客户逃离的数据集上，我们发现利用了领域知识的基于模型聚类的方法有较好的 AUC 性能.在一些形状比较规则的数据集上，领域无关的算法在分类效果上有一定的优势.

在今后的工作中，我们将会把具体的分类算法考虑进去来做进一步的预处理，考虑怎样的属性组合可以更好地提高我们算法分类的性能. 我们也计划把算法运用到其他的多维序列数据集中.

参 考 文 献

[1] Itakura F. Minimum prediction residual principle applied to speech recognition. IEEE Transactions on Acoustics Speech Signal Process(ASSP), 1975, 23(1): 52-72

[2] Kruskall J B, Liberman M. The symmetric time warping algorithm, From continuous to discrete//Time Warps, String Edits and Macromolecules. Addison, 1983

[3] Myers C, Rabiner L, Roseneberg A. Performance tradeoffs in dynamic time warping algorithms for isolated word recognition. IEEE Transactions on Acoustics Speech Signal Process(ASSP), 1980, 28(6): 623-635

[4] Berndt D, Clifford J. Using dynamic time warping to find patterns in time series//Proceedings of the AAAI-94 Workshop on Knowledge Discovery in Databases. Seattle, WA, USA, 1994: 229-248

[5] Aach J, Church G. Aligning gene expression time series with time warping algorithms. Bioinformatics, 2001, 17: 495-508

[6] Caiani E G, Porta A, Baselli G, Turiel M, Muzzupappa S, Pieruzzi F, Crema C, Malliani A, Cerutti S. Warped-average template technique to track on a cycle-by-cycle basis the cardiac filling phases on left ventricular volume. IEEE Computers in Cardiology, 1998, 25: 73-76

[7] Gollmer K, Posten C. Detection of distorted pattern using dynamic time warping algorithm and application for supervision of bioprocesses//Proceedings of the IFAC Workshop on On-Line Fault Detection and Supervision in Chemical Process Industries, 1995

[8] Abou-Moustafa K T, Cheriet M, Suen C Y. A generative-discriminative hybrid for sequential data classification//Proceedings of the IEEE International Conference on Acoustics and Signal Processing. Montreal, 2004: 805-808

[9] Vapnik Vladimir. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1999

[10] Dasarathy Belur V. Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques. Los Alamitos: IEEE Computer Society Press, 1990

[11] Keogh E, Kasetty S. On the need for time series data mining benchmarks: A survey and empirical demonstration//Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Edmonton, Alberta, Canada, 2002: 102-111

[12] Yang Yi-Ming, Yang Qiang, Lu Wei, Pan Jia-Lin, Pan Rong, Lu Chen-Hui, Li Lei, Qin Zhen-Xing. Preprocessing time series data for classification with application to CRM//Proceedings of the 18th Australian Joint Conference on Artificial Intelligence (AI 2005). Sydney, Australia, 2005: 133-142

[13] Ling C X, Huang J, Zhang H. AUC: A statistically consistent and more discriminating measure than accuracy//Proceedings of the 18th International Joint Conference on Artificial Intelligence. Acapulco, Mexico, 2003: 329-341

[14] Faloutsos Christos, Ranganathan M, Manolopoulos Yannis. Fast subsequence matching in time-series databases//Proceedings of the 1994 ACM SIGMOD International Conference on Management of Data. Minneapolis, Minnesota, United States, 1994: 419-429

[15] Chan Franky Kin-Pong, Fu Ada Wai-Chee, Yu Clement. Haar wavelets for efficient similarity search of time-series: With and without time warping. IEEE Transactions on Knowledge and Data Engineering, 2003, 15(3): 686-705

[16] Yi Byoung-Kee, Jagadish H V, Faloutsos Christos. Efficient retrieval of similar time sequences under time warping//Proceedings of the 14th International Conference on Data Engineering. Orlando, Florida, 1998: 201-208

[17] Kim Sang-Wook, Park Sanghyun, Chu Wesley W. An index-based approach for similarity search supporting time warping in large sequence databases//Proceedings of the 17th International Conference on Data Engineering. Heidelberg, Germany, 2001: 607-614

[18] Quinlan J R. C4.5: Programs for Machine Learning. San Francisco: Morgan Kaufmann Publishers, 1993

[19] Kay Steven M. Fundamentals of Statistical Signal Processing: Estimation Theory. Upper Saddle River, NJ, USA, Prentice Hall, 1993

[20] MacQueen J B. Some methods for classification and analysis of multivariate observations//Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, 1967, 1: 281-297

YANG Yi-Ming, born in 1964, Ph. D. candidate, professor. His research interests include data mining and data analysis.

PAN Rong, born in 1976, Ph. D., postdoctoral fellow. His research interests include data mining and case-based reasoning.

PAN Jia-Lin, born in 1980, Ph. D. candidate. His research interests include machine learning and data mining.

YANG Qiang, born in 1961, associate professor, Ph. D. supervisor. His research interests include data mining, AI

planning, case-based reasoning, and machine learning.

LI Lei, born in 1951, professor, Ph. D. supervisor. His research interests include data base, data mining, and artificial intelligence.

Background

Time-series learning has long been an important topic in machine learning and data mining research due to its wide-ranging impact in applications such as stock-market analysis, speech recognition, hand-written character and word recognition, and sensor-network-based activity recognition to name a few. There are several different aspects of the time-series learning problem, among which the whole time-sequence classification problem, which determines how to classify an entire sequence into one of several discrete labels, is an important sub-problem with many industrial applications. Unlike traditional methods and problem formulations in time-series analysis, time series classification aims to take whole time sequences as input, and produce discrete labels that are assigned to each sequence. Compared to traditional classification problems, time series classification poses additional difficulties. A major difficulty is due to the fact that the time sequences are variable in length, making many traditional classification methods unable to apply directly. Even for sequences of uniform lengths, many methods can still not be applied directly because often the data located at different parts of the sequences are incomparable. This paper focuses on the supervised whole-time-sequence classification problem and offers a novel solution for solving it.