博士学位论文

时序、图像特征检测的理论、方法及应用研究

作 者 姓 名： 钟 清 流

学 科 专 业： 模式识别与智能系统

学院(系、所)：信息科学与工程学院

指 导 教 师： 蔡 自 兴 教 授

中 南 大 学

二零零八年九月

分类号 VDC_____ 密级 _____

博 士 学 位 论 文

时序、图像特征检测的
理论、方法及应用研究

Research on Theoreies,Methodologic and Application for Character Detection of Time Series and Image

作 者 姓 名： 钟 清 流

学 科 专 业： 模式识别与智能系统

学院(系、所)： 信息科学与工程学院

指 导 教 师： 蔡自兴 教授

论文答辩日期_____ 答辩委员会主席_____

中 南 大 学

2008 年 9 月

摘 要

待识对象的特征检测是各种智能系统(如机器人,医疗诊断仪器等)实现智能信息处理的基础.其应用的日益广泛性、任务的复杂性、工作环境的不确定性和特殊性、其自身资源的有限性及特征检测的时效性、决定了其研究任务的挑战性和艰巨性.然而,对特征检测可靠性和安全性日益增加的需求,又不断加大了研究解决这一问题的紧迫性. 因此,探索准确高效的特征检测理论、方法和技术已经成为研究的重要内容之一.

本文主要探讨如何以基于机器学习及相关途径相结合的方式,利用有限系统资源高效实现实时特征检测的相关理论与技术问题.本文着重针对时序和图像特征检测方面的问题,提出了一些相应的算法.并验证了其有效性. 本文的主要创新点包括

(1).在研究分析 OCSVM 及 PSO 模型特点和综合前人研究成果基础上,提出 OCSVM_CPSO 组合式异常检测模型,实现了系统的自适应调节,解决了检测系统的在线运行问题,从而为其现实应用扫除了障碍.并将其应用于解决机器人传感器故障检测的实际问题.取得较好效果.

(2).针对 SAX 模型容易丢失边界区信息问题,提出时序数据的 DLS 模型.它根据时序极值确定划分的上下边界,并根据最大熵确定最佳描述字符集,进而确定最佳划分间隔.从而能有效减少边界区的信息丢失;针对 EXT_SAX 模型缺陷,提出 VSB 模型,采用增加分量而非增加符号的途径来降低计算代价.且用实验证实它的有效性.

(3).提出时序矢量符号的 SFVS 模型和相应的确定时序数据最大压缩比的方法,此模型能够比 SAX 提供更全面的描述信息, 这有利于在时序特征检测的应用中实现更精确的分析.通过理论分析和实验证实了它的有效性.

(4).针对自调节谱聚类缺陷, 提出一种新的 ASC(自适应谱聚类)算法.它用全局平均N近邻距离作为比例参数 σ,利用本征矢差异来估计最佳聚类分组数 k, 这可在构造亲和力矩阵时减少计算代价,提高效率,并且更容易实现.在彩色图像检测与分割实际应用中的实验结果证实了其有效性.在此基础上提出改进聚类性能的相关半监督学习算法,且用实验证实了它的有效性.

关键词：特征检测,时序符号化,异常检测,相似查询,图像分割

ABSTRACT

Character_detection for identifying object is foundation on which intellective Systems such as robot and medical diagnose actualize intellective information disposal. It's increasingly universality in application,complexity in role, uncertainty and particularity in environmental, restricted self resource and real time demand of character detection , which have determined its research tasks are challenge and difficult . However , the demand ,for reliability and security of Character detection to be increasingly enhanced ,continuing enhance the pressure to resolve it.Thus,exploring true and resultful theory,method and technique about Character_detection has become one of important research contents.

In this paper, the relative theory and technique how to effectively implement real time Character_detection are discussed by restricted system resources and based on conjunction machine learning with others approach.For issue of Character detection in time series and image , some relevant algorithm have been put forward and its validity have been validated. Main contributions of the dissertation are shown as following,

By means of research and analysis the characteristic about the model of OCSVM and PSO and based on integration of father fruitm,an model of anomaly detection based on OCSVM_CPSO is put forward,which actualize that system adaptive adjustment and solve the problem about detecting system online run, and clean off obstacle for its real application. It is used as solving the real problem of fault detection of robot sensor,and a good result is abtained

The DLS model is put forword in order to overcome the bug that SAX easily lose information on boundary .DLS partition fluctuate boundary according to extremum of the time series ,and select optimization character set based on the most entropy ,more, neatly set optimization partition interval,thus can effectively reduce losing information on boundary .Another Model , for solving the bug of the EXT_SAX,the VSB model is put forword, which increase the component to reduce calculating cost . Its availability is proved by the experiments.

A vector symbolic model for Time Series Data Based on Statistic Feature,SFVS,and relavant mothod for estimating the most compress ratio about time series data, are put forward in order to surmount the bugs with which SAX

Algorithm can not describe time series information fully,which is helpful to implement more accurate analysis in application of feature detection of time series.Its validity have been proved by experiments

For the bug of Self-tuning spectral clustering, A new algorithm ASC(called as the adaptive spectral clustering), is put foward,which takes average distance of N-near-neighbour as scaling parameter σ, automatically estimates optimal clustering grouping k by means of information about enginvector difference .It can reduce calculating cost for constructing appetency matrix as well as get highter efficiency and implement easyer .Farther, the correlative semi supervised algorithm for improving its'performance are put forward on the base.The results of experiments in application about color image detection and segmentation shown that the algorithm is valid .

KEY WORDS: character detection, symbolize time series, abnormity drtection, similitude query, image segmentation

目 录

第一章. 绪 论

1.1 课题来源,研究背景及意义

1.1.1. 课题来源

本论文研究课题得到国家自然科学基金项目“未知环境中移动机器人导航的理论与方法研究”(60234030)和国家基础研究项目(A1420060159)资助. 本论文着重研究课题相关时序与图像特征(尤其是异常特征)检测方面相关的理论,技术及实际问题.

1.1.2. 研究意义

目标信息(如时序,图像)特征检测包括目标正常特征及异常(故障)特征检测两个范畴. 一般而言,正常特征描述研究对象的规则性状态或行为,异常特征描述研究对象的非规则性行为. 对研究对象这两种行为的特征检测,是各种智能系统(如机器人,医疗诊断仪器等)实现智能信息处理的基础. 其应用的日益广泛性. 任务的复杂性、不确定性, 和特殊性,其自身资源的有限性及特征检测的时效性. 决定了其研究任务的挑战性和艰巨性. 然而,对特征检测可靠性和安全性日益增加的需求,又不断加大了研究解决这一问题的紧迫性. 因此:探索准确高效地特征检测, 诊断理论,方法和技术,已经成为研究的重要内容之一.

目前的理论及方法大多只能解决某些特定领域问题,它们只是针对某类具体问题的局部模型或方法,对于特征检测研究而言,还有许多关键理论和技术问题有待解决和完善. 尚未形成统一的理论框架或体系. 因而研究探索能够解决综合复杂环境下的目标检测及诊断方法仍是一个十分艰巨的任务. 由于特征检测通常是在缺少先验知识情况下实时地进行. 是涉及多学科多技术综合的高难度问题之一. 融合多种理论,方法与技术是这一领域研究发展的趋势所在.

本文主要研究如何以基于机器学习及其它途径相结合的方式,利用有限计算资源高效实现实时特征 (尤其是异常特征) 检测的相关理论与技术问题. 以期取得实质性进展.

1.2 特征检测研究现状及进展

在特征检测尤其是异常(故障)特征检测方面,国内外已经积累了一些成功的经验,并建立了一些相应的理论,方法和模型,归纳起来,大致有以下三类:

1) 基于模型方法:

自主检测一直是人工智能的中心主题,尽管这些方法解决过很多不同的问题,包括静态环境,和/或定性输入问题. 一些基于分类的方法仅依赖当前传感器读数. 几何分类器把传

感器测量空间划分成不同空间和状态[1].这些方法通常不考虑时间信息,因而不适合类似于MYCIN[2]这样的基于规则的检测系统和一些广泛应用的专家系统[3].尽管这些系统可能包括传感器故障的假定,但它们不能用传感器数据来模拟这些不确定性.

Brown 等人从诊断研究的领域成果[4,5]中得到启发,在文献[6]中提出基于模型的推理,所定义的基于模型的的方法就是关于系统外在模型的的推理.使基于模型的的方法成为特征检测(如故障诊断)应用的重要方向之一.它以机器人典型的正常行为模型作为故障特征检测和诊断故障的准则.

文献[7-9]中给出了基于模型方法相关的实例.最典型的实例或许是故障树分析 -FTA.这是一种自顶向下的用图形表示设备状态为相关功能组件状态的分析模型.它能够定量或定性地分析发生在系统中可能导致顶级故障事件的各种可能故障事件组合.其主要缺点是容易遗漏一些重要的故障而造成误诊,并且由于它很费时,运算代价大,结果难以检查,使得很多情况下难以运用.

其它有代表性的模型方法还有 Visinsky 等人(1995)提出一种分层式状态检测和容错模型[10], Washington(2000 年),提出用于移动机器人状态估计和故障诊断的模型 MaKSI[11] .基于模型的诊断系统已经扩展到处理动态系统如 XDE[12]GMODS[14].为了解决连续动态系统问题, 通常采用定性模型(如 Livingstone system [11]) Livingstone 系统的状态估计是通过搜索整个模型的转换以寻找与传感器测量一致的状态.状态约束表达为一种逻辑表达式,它定义了代表整个系统状态的离散变量空间. Livingstone system 假定离散化的传感器测量中没有不确定性,并用它们作为硬约束.在真实信息不完全由当前传感器测量情况下,这一假定将导致空的假说空间,或不正确的识别.LivingstoneII[12]更新基于过去假说的保留记录,并得到成功的测试.由于这种方法需要建立模型,而系统故障模型的建立是非常困难的, 通常故障诊断的大部分时间都花在建模上,因而其应用范围受到限制.

Willeke 和 Dearden 等研究了建立混合模型遇到的一些特定问题,进而提出了相应的解决方法.由于基于模型的方法只能识别已建模的故障,因而关键是要确定哪些故障需要追踪.从而建立针对性的模型.然而这并未解决基于模型方法的根本问题.

近年来的一些新模型,方法和技术进展主要包括基于异常检测的模型.

采用异常检测来实现状态特征检测,尤其是故障状态检测是一种很有前景的高效算法.异常检测是识别新的,异常的或未知数据或信号的过程,它广泛用于信号处理,模式识别,机器人状态检测等.异常检测的基本思想是在为正常数据建模的基础上,通过测量目标数据与模型数据的差异,进而根据确定的阈值来识别异常.它通常用分类器作为检测器来检测输入数据中的异常成分(或未知成分),这一技术广泛用于状态检测[13-16],雷达目标检测[17] (Carpenter et. al., 1997), 医疗诊断[18] (Tarassenko, 1999), 手写体识别[19] (Tax and Duin, 1998), 欺诈或入侵侦测[20] (Manikopoulos and Papavassiliou, 2002),统计过程控制[21] (Guh et al, 1999).

基于模型方法的基本特点是：操作简单,容易实现,适合于大多数异常检测的任务.它以正常典型的行为作为标准,使用前须事先建立针对性的模型.根据待测目标偏离模型的程度来判断正常或异常.且只能检测已知的异常或故障.因而事先获取与检测相关的正确模型并能选择合适的划分标准或阈值成为检测成败的关键.然而,现实应用中的很多情况中,事先往往难以或根本无法得到标准的正常行为模型,此时,基于模型的方法便不再有效.

2）基于概率方法:

（1）统计性方法

统计性方法首先要选择描述主体行为的测度集,然后在采集到的正常事件集合中建立基于该测度集的检测模型,该模型有可能是用户的正常行为映像,也可能是正常工作流测度的概率分布,取决于异常检测系统的检测目标和实际的检测模型,某种度量算法被用来计算当前的主体行为与检测模型的背离程度,然后根据某种决策方法来决定是否异常.统计性模型的主要优点是能自适应地学习主体的行为,因此对异常行为比人更加敏感.另外统计性模型不要求训练数据全部是纯粹的正常行为,只要是真实环境的数据就行,这一点很重要,因为从数据源获取的数据一般是没有标记的.

采用统计性方法最有代表性的要数 SRI（Stanford Ressearch Institute）的 IDES（Intrusion－Detection Expert System）[22] 和 NIDES（Next-Generation Intrusion Detection Expert System ）[23],这两个状态检测系统包含了面向用户的异常检测模块,通过建立描述用户行为的各测度的概率分布函数作为其检测模型,并采用某种距离算法评价系统行为与其模型的差异并作出响应.

统计性模型的缺点是：(1)主观确定的故障阈值决定了误检率和漏检率的高低；(2)它对于依赖于事件之间关系的故障不敏感,因为这种模型忽略了事件之间的关系；(3)需要假设测度的概率分布. 目前一般采用正态分布或泊松分布. 而这有可能与实际不符合.

（2）预测方法

预测方法的检测对象是事件的时间序列,其目的是发现与构成故障的相关事件(指一条审计记录,或一个检测实体单元)集合在时间上的相关性,从而预测未来发生的事件. 如果实际发生的事件与预测结果有较大的差异,则表明有异常现象发生.Teng 和 Chen[24]提出基于时序的推导性归纳方法,产生时序性规则来建立系统的正常行为映像.这些规则在训练阶段会动态地调整,只有较高的预测准确性的规则被保留.

这种模型有以下的优点：(1)它能检测到传统的模型(例如统计性模型)不能检测的故障；(2)它对主体行为的变化有高度的适应性,因为低质量的时间序列模式会不断地被排除,而留下高质量的模式；(3)它能在故障完成前检测到并发出报警.

（3）其它方法：

贝叶斯网络也已用于故障诊断[25].其大多数应用于时间无关情况,但近年来动态贝叶

斯网络已经用于解决时间相关问题[26],其它概率方法如[27]采用 Dempster-Shafer 卡尔曼滤波器,这是动态系统状态追踪的流行方法.在过程的测量方程是线性,噪声是高斯情况下,卡尔曼滤波是最优选择.目前,卡尔曼已扩展了多种方法既可以解决单目标线性问题,也可以解决多目标非线性问题.卡尔曼滤波层残差的概率组合可以用来确定故障状态[28].

(4)基于粒子滤波方法

非线性动态系统的故障诊断通常同时涉及离散和连续混合状态.它无法用卡尔曼滤波器(KF)[61]解决,因为 KF 的前提是系统噪声为零均值且具有已知协方差的高斯过程(具有单峰的概率密度函数).基于蒙特卡罗(Monte Carlo,MC)和递推贝叶斯估计的粒子滤波器(PF)[62-65]成为这类问题的首选方案.

它可以应用于任何动态状态空间模型,在传统维纳滤波、卡尔曼滤波及扩展的卡尔曼滤波不能处理的非线性非高斯问题上,粒子滤波已经成为研究非线性非高斯动态系统最优估计问题的一个热点和有效方法.并已广泛应用到目标跟踪、信号处理和数字通信及故障诊断等多个领域.

作为一种基于递归贝叶斯估计的非线性滤波算法,在处理非高斯非线性时变系统的参数估计和状态滤波问题方面有独到的优势,但粒子滤波算法本身还不很成熟,仍有大量的问题亟待解决,主要体现在以下几个方面[66]:

重要性函数的选取直接影响粒子滤波性能的高低. 目前重要性函数选取仍是研究课题. 从粒子滤波的数学基础上看,粒子滤波的收敛性尚未解决,若能有效解决收敛性问题,则对粒子匮乏现象的抑制将有很大帮助. 同时作为滤波算法,如何评价粒子滤波的性能也是需要认真考虑的问题.

粒子滤波算法运算的实时性问题、状态初始概率选取问题,使粒子滤波距离工程应用尚有一定差距. 根据系统不同阶段体现的不同统计特性,将粒子滤波与其他非线性滤波方法相结合,避免非线性系统的近似线性化,减小非线性系统线性化后的高阶截断误差和非高斯噪声带来的影响,是实际问题中采用粒子滤波方法的又一研究方向. 针对这些问题,学术界已经进出或正在探索一些相应的解决方案.

例如:Rudolph van der Merwe 等人提出 UPF(Unscented Particle Filter)算法[67],它在 PF 算法的基础上利用 UKF 算法产生比普通 PF 更好的重要性函数.每一次采样后的粒子都由 UKF 算法进行更新,所得的均值和方差用于下次采样新的粒子, 因而是解决第一类问题的较好方案. Dearden and Clancy[68]利用“oracle”提供的一组可能使系统终止的候选状态,(假定的整个状态空间的当前分布),在重取样基础上,确保总有一些粒子粒子 oracle 粒子提供的状态. 解决了低概率故障问题.

对于第二类问题: Verma 等人(2003)提出变分辨率粒子滤波[VRPF](Variable Resolution Particle Filter)模型[69],VRPF 引入抽象粒子概念,可用一个抽象粒子表示并跟踪单个或一组相似状态,当占有某部分状态空间的邻域较低时,有限的粒子就足以代表

大部分的状态空间.当被编组状态的邻域增加,并且对这些状态重要性变得十分重要时,就须专门指定一些粒子代表单个状态. 变分辨率粒子滤波(VRPF)通过将多个相似低概率状态编组,仅为需要的状态特殊分组而减少了状态跟踪所需的粒子数.抽象粒子方法通常可以大幅减少状态转移数量和相应的状态估计方差.为提高计算效率,VRPF 动态地改变状态空间的分辨率：在信度强时,分辨率高；信度弱时,分辨率低,此时将多个相似状态抽象为一个抽象状态.VRPF 同样降低了粒子滤波器的维数,但仍停留在小故障空间的仿真研究阶段.

Gordon 等人提出了 Variable Resolution Unscented Particle Filter (VUPF)模型[70],通过提前一步考虑预期传感器的测量值以改进由于粒子数受限的那些状态的估计精度.采用 UKF 的预测一步的滤波器:在测量上的预测一步能大幅改进诊断性能.UKF 是一种递归式最小均方误差估计器,通常可以改进非线性模型的 EKF,UKF 用实际模型代替高斯分布的状态变量的近似分布.用最小组的确定性选择取样(sigma points samples)来确定高斯近似.各 sigma point 通过过程的测量模型独立传播,分析传播的 sigma point 集可以得到高斯近似的后验.

对第四类问题,Verma 等进出风险敏感粒子滤波(RSPF)模型[71] RSPF 结合了产生粒子的代价模型其灵感来自未追踪假说的的代价与风险相关的观察:不追踪一个罕见的,但有风险的状态可能付出高的代价.然而不追踪一个罕见的但良性的状态可能无碍大局.将代价模型合并到粒子滤波中,可以改进最关健状态的追踪性能.故障是低概率高代价事件.经典粒子滤波只产生正比于事件后验概率的的粒子,因而监测一个 CPF 的系统需要很多粒子,代价昂贵, RSPFs 按代价分解产生粒子,由于故障有高代价,即使它们发生概率低,也能产生能表示它们的粒子.故可用比 CPF 更少的粒子达到同样效果. 通过按更新粒子的建议分布合并潜在的损失减少了跟踪不稳定状态所需的粒子数. 通过引入风险权值来加大小概率故障状态的权重[72],然而这些算法并未提出有效确定风险权值的方法.

Frank Hutter 提出用于非线性混合系统粒子滤波的变体—高斯粒子滤波(GPF)[73].它的各粒子对应一个离散模型取样,并在各时间步用 unscented KF 更新的多元高斯来近似各连续变量.由于增强了粒子的代表能力,这有助于克服低故障概率引发的故障状态粒子数过少问题,并且由于该算法无需计算各后继模型的后验概率,因而大幅减少了计算代价.同时还提高了计算精度.尽管实验已证明了它的显著效果[76],但还值得进一步完善.

神经网络也可处理这些残差并用于识别相关的故障[29].另外,高斯混合模型,作为一个算法大家族,通过计算精确的后验概率,然后按要求减少混合,包括删去残差较高的滤波器[30],按权重对混合分量取样,并重复地将混合分量的最相似部分融合[31],以达到较好的估计结果.POMDP(partially observable Markov decision processes.)来表示离散状态模型. Haufbaur 在 [32]中用概率混合自动控制扩展连续动态的 HMMs 模型.

以上方法的弱点是:对动态系统.它们大都典型地局限于高斯后验模型.解决不确定

问题并用 POMDP 表示的另一途径是[33],其核心是从故障及其离散状态空间的运行中恢复信息.

3）机器学习方法:

基于机器学习的异常检测模型采用机器学习的方法来建立系统映像.它的最大特点是根据正常来分辨异常,与基于模型方法不同的是,由于它直接采用(可能含有少量异常的)工作流数据作为训练数据.而无需提供代表待检测故障模型的相关训练数据. 因而可以大幅减少需要处理的信息量.使学习器集中于某一特殊感兴趣信息,故减少了系统必须的学习代价.能够在普通数据流中察觉对象.在类概率密度分布或混合比率变化或新类出现时发现异常.机器人中用于训练机器人识别环境的正常特征(传感器已识别的),而将任何不能被识别的其它特征都视为可能的故障.这种方法的优点是检测速度快,而且误检率低.但该方法在用户动态行为变化以及单独异常检测方面还有待改善.复杂的相似度量和先验知识加入到检测中可能会提高系统的准确性. 但需要做进一步的工作.机器学习广泛应用于故障监测和医疗诊断. 一些有代表性的方法包括:

(1)基于数据挖掘的方法

数据挖掘能从审计记录或数据流中提取出感兴趣的知识,这些知识是隐含的、事先未知的、潜在的有用信息,提取的知识表示为概念、规则、规律、模式等形式,并可用这些知识去检测异常故障和已知的故障.Wenke Lee 和 Salvatore J. Stolfo 等在 1998 年和 1999 年提出通过对正常数据建立决策树的预测模型来作为检测模型[34-35],然后用该模型来检测实际发生的网络报文是否异常.他们在 2001 年提出将基于数据挖掘的检测应用到实时环境中的模型.着重解决三个关键问题：检测的准确性、效率和可用性.用于异常检测的数据挖掘典型方法还有分类和聚类如[36-40].

数据挖掘的优点是能自动、快速地产生异常检测模型,这在海量的历史数据中提取知识是非常重要的,通过人工建立的方法很难实现.数据挖掘方法的缺点在于：①误报率较高；② 由于在训练和评价时计算的复杂度较高,难以应用到实时环境中；③需要大量的训练数据,而且对数据的纯洁性要求较高.

(2)基于神经网络的方法

神经网络可以应用到各种异常检测模型中,有的基于神经网络的异常检测模型是分类器,它通过训练和学习,记忆了系统的正常行为或故障行为. 并能根据系统现状进行自我调节,有效地发现并阻止各种故障行为,这种神经网络与基于数据挖掘的决策树的作用是类似的,例如 Stephen Marsland 提出的基于神经网络的自主机器人在线异常检测模型[41].另一种神经网络可用来建立预测模型,例如通过在神经网络的输入端输入用户所用的命令序列可以预测下一个命令,如果不符合就可判定为异常,这种神经网络的例子是 Kevin[42]等设计的面向状态检测的神经网络模型.以及文献[13]提出的基于小波神经网络模型.

神经网络的优点是：① 它的实现不依赖对潜在数据的统计假设；②能较好地处理噪声

数据；③ 能自动调节影响输出的各测度的权重. 而这在传统的异常检测方法中通常是人为确定的.神经网络的缺点在于：① 神经网络的拓扑结构和各元素的权重只有在训练后才能确定；② 输入窗口的大小是该方法的一个主观因素,如果设得太低,该模型的检测能力就会下降,如果太高,就会碰到许多不相关的输入.

(3)基于 SVM(支持向量机)的方法

主要方式有 SVR 和 OCSVM 两类.其中,SVR 方式是利用 SVR 算法模拟工作流数据的动态特性,并通过比较输出结果与实际系统输出的残差来检测故障或异常.能够有效地保证了传感器故障诊断的准确性.例如, 肖健华,吴今培等人提出的模型[43]. 大多数检测算法都须要建立明确的统计模型,并将异常事件看作是统计模型的突变.相应的实现依赖于 Bayes 估计理论的最大邻域,然而现实中的大多数情况下很难得到一个健壮而可行的模型,因而无模型的解决方案更加容易实现.基于 OCSVM（一类支持向量机）的异常检测正是这种无模型解决方案.

2003 年,Scholkopf 等人基于 SVM 给出了估计在特征空间中估计大部分数据出现区域的方法[44],而由假设,大部分数据就是正常数据,因此,可以使用该算法来估计正常数据出现的区域,进而检测异常. Scholkopf 等人将 SVM 算法应用到非监督情况,提出了 OCSVM 算法,与传统的SVM不同,OCSVM试图用超平面将数据集中所有的点和原点分开,通过最大化分类边界压缩数据集在特征空间中所占据的区域来区分正常与异常点.从而奠定了用支持向量机实现异常检测的理论基础. 随后有很多的学者提出了众多的用支持向量机实现异常检测的有效方案[45-48].形成了一个研究热点.

Heller 等人(2003)提出了 OCSVM 的入侵检测模型.能够用较少的训练样本获得较其它方法更高的检测精度.还有一些基于 OCSVM 的较好异常检测算法和模型如文献[49-53].

Davy 等人(2006)在综合前人成果的基础上提出了基于OCSVM序列优化的无模型式的异常检测改进模型[54].通过引入系列优化算法,有效实现了无监督学习的在线异常检测,并在理论上和实践中验证了其有效性.

基于 OCSVM 异常检测的优点是它直接用工作流数据训练,能够发现各种已知或未知故障,计算代价小,精度高.并且是全局最优解,泛化性好.是比较理想且有开发前景的方法.其缺点是,它只能处理矢量数据,不能直接处理来自传感器的时序数据,时序数据须要经过相关的相空间转换才可使用.另外,它在高维情况下的表现还不尽人意.因而有待进一步探讨.

(4)基于免疫学原理的方法

主要包括:基于免疫学负向(阴性)选择[55]算法及克隆选择[56]算法:以及根据生物免疫系统原理建立免疫计算系统模型,包括:人工免疫网络模型和人工免疫系统模型两种形式.前者如 aiNET[57]免疫 PDP 网络[58]和免疫多值网络[59]等.各种免疫网络学说,如独特型网络、互联耦合免疫网络[60]等.

基于免疫学原理的异常检测模型受生物免疫系统的启发,试图为要保护的对象建立一个“免疫系统”,该方法的关键是如何有效地定义“自我”(self)和识别自我,并据此来排斥“异类”.基于免疫学负向(阴性)选择算法及克隆选择算法:其机理的一个重要应用是检测模式的变化.其中“自我”被认为是被监视系统的正常行为模式,所以任何所观察数据超过允许偏差的偏移,都被认为是异常模式[47].

免疫检测模型的优点:

是一种并行和分布自适应系统,可借鉴用于建立人工免疫网络(简称 AIN),是典型的交叉学科,其可借鉴的理论极为广泛和丰富,因而发展出的人工免疫系统形式也多种多样.算法实现方法简单,检测器通常以概率的方式进行检测而不需要有异常模式的先验知识. 识别能力强,精度高. 有极好的发展前景.

免疫检测模型的弱点:

现有各类免疫算法仍存在许多问题,有许多需要改进和研究的地方,比如:

AIS 领域还没有系统的数学基础.多数 AIS 用来自遗传算法、细胞自动机和 ANN 等思想表示一些混合和启发性算法.克隆选择算法缺乏处理动态问题的能力,阴性选择算法的处理对象过于简单,与实际需要有一定差距等等.因此迫切需要建立人工免疫系统自己的理论体系.总的来说:免疫算法效率低,计算代价大,不易实现实时分析.且知识表示方式有限;免疫模型尚未形成统一体系.免疫机制不够丰富.

总之,免疫模型在时序信息异常检测及故障诊断方面的应用能大大提高其智能度, 增强诊断效果.是一个值得注意的很有前景的新方向.

1.3 特征检测的特点和难点问题

特征检测问题具有以下显著特点: 检测数据的多样性,不完整性,随机性,问题的非线性,检测环境复杂性、不确定性、资源有限.这要求检测系统满足以下性能特点:准确性,鲁棒性、高效性、实时性、灵活性.

例如动态时序信息是一种随机变动信息,具有典型的不确定性,且时序数据又属高维数据,很难找到一种既能完整保持原始信息,同时又能简化运算并降低计算代价的理想降维方法和相应的描述模型.

而图像信息通常数据量大,含有大量干扰噪声,且由于受场景,亮度,视角等众多因素影响,因而如何有效预处理和寻找最佳描述模型本身就是一个难题.

同时,上述两类信息特征检测过程中,特征本身的选择, 确定最佳的正常与异常划分标准也非易事.需要经过大量的论证.

除此之外,还有由于时序与图像数据本身数据量通常过大带来的计算代价过高,运行缓慢,从而限制了它在需要实时控制或反应的系统(如自主机器人系统)的应用范围.

1.4 研究的关键问题

本文的研究目标是针对时序信息及图像信息特征检测(主要是异常特征)中的难点问题,提出相应的新理论及技术解决方案,并将这些理论和技术用于解决相关现实问题的实践,为机器人学,机器学习和模式识别等相关交叉学科的发展提供理论和技术支持.本文研究的主要内容概述如下:

(1) 研究特征检测系统的建模理论及相关技术问题,包括 OCSVM 及相容间重构理论和技术,目的是要提出一种简捷的无模型(无需事先建立正常行为模型而直接对待检测数据进行分析)解决方案;研究上述方案实现过程中相应检测系统参数确定问题.具体包括基于群智能理论和混沌理论及相关的全局最优搜索技术,目的是要达到自动确定检测系统的最佳模型参数以解决检测模型的可靠性及精度等问题,这为检测系统的实际应用奠定了基础.

(2) 着重研究时序数据符号化的理论及相关符号化模型在时序特征检测中的相关理论及应用问题,包括相关符号化理论及著名的 SAX 模型,其目的是要通过分析,比较,提出更简捷有效的时序符号化模型,以解决时序数据因维数过高导致的计算代价过大,运行缓慢,检测效率不高而难以实时化的问题.

(3) 研究基于谱图理论的谱聚类技术及在图像特征检测中的应用问题,目的是要将有竞争前景的全新的谱聚类技术引入图像分析领域,提出更有效简捷的图像特征检测算法,以充实图像特征检测的理论体系 .

(4) 考虑到混合式(兼有监督、半监督、无监督式)学习系统是机器学习的未来方向,文中还研究了半监督式学习问题,目的是拓展特征检测系统的应用范围,改善检测系统的性能.

1.5 论文的创新点

本论文的研究目标是针对时序信息及图像信息特征检测(主要是异常特征)中的难点问题,提出相应的新理论及技术解决方案,其主要创新如下:

(1).在研究分析 OCSVM 及 PSO 模型特点和综合前人研究成果基础上,提出 OCSVM_CPSO 组合式异常检测模型,克服了传统 OCSVM 不能处理时序数据的缺陷,实现了系统的自适应调节,解决了检测系统的在线运行问题,从而为其现实应用扫除了障碍.并将其用于解决机器人传感器故障检测的实际问题.取得较好效果.

(2).针对 SAX 模型容易丢失边界区信息问题,提出时序数据的 DLS 模型.它根据时序极值确定划分的上下边界,并根据最大熵确定最佳描述字符集,进而确定最佳划分间隔.从而

能有效减少边界区的信息丢失;针对 EXT_SAX 模型缺陷,提出 VSB 模型,采用增加分量而非增加符号的途径来降低计算代价.且用实验证实它的有效性.

(3).提出时序矢量符号的 SFVS 模型和相应的确定时序数据最大压缩比的方法,此模型能够比 SAX 提供更全面的描述信息, 这有利于在时序特征检测的应用中实现更精确的分析.通过理论分析和实验证实了它的有效性.

(4).针对自调节谱聚类缺陷, 提出一种新的 ASC(自适应谱聚类)算法.它用全局平均 N 近邻距离作为比例参数 σ,利用本征矢差异来估计最佳聚类分组数 k, 这可在构造亲和力矩阵时减少计算代价,提高效率,并且更容易实现.在彩色图像检测与分割实际应用中的实验结果证实了其有效性.在此基础上提出改进聚类性能的相关半监督学习算法,且用实验证实了它的有效性.

1.6 论文章节的安排

全文共分为 7 章,各章节内容安排如下:

第一章 对特征检测的国内外研究现状进行综述,总结该领域的研究难点问题以及研究趋势.概括论文的研究内容、关键问题以及论文的创新点.

第二章 讨论当前时序数据异常特征检测的基本技术,问题和改进方向,和特征检测系统的建模问题,提出基于 OCSVM_CPSO 组合式自适应故障特征检测模型.

第三章 提出几种不同的时序数据符号化方法,分别解决相应的特殊问题.

其一,时序数据的动态有界符号化(DLS) 方法.重点关注极值信息根据时序数据的极值确定划分的上下边界.并将其用于时序特征检测的实际中检验.

其二,VSB 矢量化符号方法.通过引入二个极值信息,将原来的 SAX 符号转化成具有三个分量的符号矢量.其符号值由各分量的加权和最终确定. 将其与 SAX 对比检验.

其三,提出的 SFVS 基于算法,首次将时序符号则看作是由某些分量构成的特征矢量.这是时序数据符号化算法探新的一种尝试.并将其与 SAX 对比检验.

第四章 提出一种新的用于彩图分割的自适应谱聚类算法,并将其与 STSC 算法比较.

第五章 提出改进学习器性能的半监督式学习算法.

第六章 总结全文要点,展望进一步的研究方向.

致 谢:

攻读博士期间完成的学术论文与科研工作

第二章. 时序特征检测系统模型

正如第一章所述,传统的特征检测大多须要根据样本数据建立起典型的正常行为的模型,检测实际上是用待测数据与正常模型比较.然而,在大多数情况下,利用样本数据建立正确的正常模型并非易事,尤其是一些复杂的动态时变系统,其建立模型本身就是非常困难的问题.然而,在许多情况下,可以借助一些数学寻优工具来解决这一困难.这为特征检测的建模提供了新的途径.

例如,本节所述的基于 OCSVM(一类支持向量机)的特征检测模型---就是一种无需事先建立正常行为模型的特征检测系统模型. 它不需要任何标记数据,能够从未标记的数据集中找出隐藏在其中特征异常,通过理论分析和实验证实它比有模型的解决方案更加容易实现.该模型是建立在时序数据表示,相空间重构及 OCSVM 等理论基础之上的.

另一方面,由于时序数据的动态随机性,以固定学习参数为特征的传统支持向量机检测模型无法胜任实时检测的任务.为此本节提出相应的 CPSO(混沌粒子群优化)自适应参数调节方法,与前面所述的 OCSVM 模型组合,构成 OCSVM_CPSO(基于混沌粒子群优化的一类支持向量机)特征检测模型.

2.1 时序特征检测的任务与挑战

2.1.1 时序特征检测的任务

时序(Time series)数据泛指按照时间先后顺序排列的各种观测记录的有序集合.它是工业,商业、经济以及科学观测等各个社会领域中广泛存在数量最大,应用最广的科学统计数据之一.分析研究时序数据的变动规律及特点,对各相关领域的科学发展和正确实践有深远的指导意义.

现实应用的情况千差万别,不但需要把握全局模型,而且也需要分析局部特征,例如分析不同时间段内的气象变化特征是否相似,或监测设备工作是否出现了异常.因此,时序数据的特征检测是实现智能控制和智能信息处理最基本的技术.

时序数据往往都蕴含着一些跟时间相关的现象甚至规律.因而其反映的大都是某个待观察过程在一定时期内的状态或表现.其研究的目的主要是以下两个方面:其一是学习并了解观察过程过去及现在的行为特征(如分析股票交易情况、发现异常交易,总结顾客消费规律)等; 其二是预测未来的可能状态或发展趋势,比如(如销售决策,传感器分布调整)等.这两个目的直接决定了时序数据特征检测中的主要任务:查找相似的行为模式;异常活动特征检测.

根据大多数文献的提法,时序数据分析研究的基本任务大致有如下几个:

1.相似查询:给定一个时序子串 Q,和一些相似/不相似测量 D(Q,C),在数据库中寻找最相似的时序[77,78].

2.聚类: 在给定相似/不相似测量 D(Q,C)[79].条件下在数据库中寻找时序数据的自然编组.

3.分类: 将给定的无标记时序 Q 正确地划分到它所属的预定义类中[80].

4.异常检测: 给定一个时序 Q,和一些正常行为模式,寻找 Q 中隐含的所有异常(新奇)的行为[81].其中时序数据特征检测的最核心内容就是寻找相似和发现异常.它是对这些与时间相关的数据进行分析并从中获取相关的信息的过程[82].

与课题相配合,本文将主要涉及时序分析的以下内容:

(1).时间序列相似测量:时间序列的相似度量研究是时间序列数据分析的基础问题.由于难以找到两条完全相同的时间序列,因此采用相似度来衡量时间序列之间的相似性.相似度和距离是相对的两个概念,是用来衡量两个(通常是两个,也有可能是多个)数据或实例之间关系的标准.在一定意义上,相似度和距离的表示能力是等价的,他们分别适用于一定的环境,也可以通过一定的方法互相转换.时间序列的相似性度量不一定支持距离三角不等式,必须能够应用到时间序列的数据分析中,例如时间序列的相似性查询、聚类和分类等等.

由于时间序列数据的复杂性,经常发生振幅平移和伸缩、线性漂移和时间轴伸缩等形变,因此相似度应该最大程度地支持时间序列的上述形变.相似度是整个数据挖掘过程中的许多其他工作(比如聚类问题等)的基础,相似度度量的好坏对这些问题的解决有着非常大的影响.

常见的相似度测量可以分为如下几类:基于形状的相似度[83]、基于特征的相似度[84]、基于模型的相似度[85]、基于数据压缩的相似度[86].而作为一种更有前景的相似度测量类型是本文后面章节要重点涉及的基于符号时序的相似度.

(2).时间序列异常检测:所谓异常检测就是从海量日常时序数据中发现偏离正常模式的少量数据,例如机器人传感器故障检测或网络入侵检测等.异常检测最大的特点就是需要被识别的时间是较少发生的.很多模式识别方法都容易产生错误的演绎偏差,从而导致结果并不理想.有关异常检测实践中将面临的困难与问题可详见文献[80],

在异常检测中一个比较简单的方法是先对正常的时间序列建模,在通过检测那些与模型不符合的数据来发现异常.但这一方法有个前提是需要获得关于正常的时间序列的知识.变通的办法就是对所有拥有的数据进行建模,再用这个模型对数据进行筛选.文献[81]介绍了一种基于状态的方法,通过对时间序列进行时间点聚类得到一个近似的分割,然后基于这些分割和分割之间的关联建立一个包含异常状态的有穷状态自动机.再通过这个自动机来识别异常.此外,异常检测中还可以用反向选择等方法进行,详细内容参见文献[80].

2.1.2 时序特征检测面临的挑战

前面所述的时序数据相似度度量是时序数据分析最基本且重要的内容.许多上层的任务比如分类、聚类、相似查询、异常检测都直接依赖于数据的表示形式和相似度度量.虽然对这项工作已经有大量的研究和文献,提出了很多很实用的解决办法.但它们都与实际应用的要求差距很大,难以解决实际应用中的各种复杂问题,针对不同的问题找到比较合适的表示形式和相似度度量才能取得较为满意的结果.能否找到一个相对通用的解决办法,使之对大部分的问题都能有比较好的效果是大家普遍关心的问题.

大多数检测算法都须要建立明确的统计模型,并将异常事件看作是统计模型的突变.假设所有的异常行为都会偏离正常行为模式,相应的实现依赖于 Bayes 估计理论的最大邻域,然而现实中的大多数情况下很难得到一个健壮而可行的模型.如何克服这一难题,能否找到其它的突破方向,这是当前研究面临的严峻挑战.

时序特征检测研究未来发展的路在何方?首先,寻找更有效的表示方法来描述时序数据的特征将有助于更方便地建模和更精确地实现时序分析;其次,寻找表达能力更强的数据模型有助于提高分析的可靠性,实时性; 最后,利用数据本身统计特征直接进行时序分析的思路也许能给这一研究带来新的曙光.

2.2 时间序列的模式表示

时间序列数据分析的对象通常是海量的时间序列数据,其短期波动频繁、大量噪声干扰以及非稳态的特点使得直接采用原始时间序列进行相似性查询、时间序列分类和聚类、时序模式分析等工作不但效率低下,甚至会影响时间序列数据挖掘的准确性和可靠性.为此,许多研究者提出了时间序列的模式表示方法,从更高层次上对时间序列重新进行描述,在时间序列的模式表示上进行数据分析.

时间序列模式表示的基本思想是从时间序列中提取特征,将时间序列变换到特征空间,采用特征空间的特征模式来表示原始时间序列.时间序列的模式表示有两个优点:首先是压缩了时间序列数据,能够提高数据挖掘工作的效率;其次时间序列的模式表示在一定程度上保留了时间序列的主要特征,去除了一些次要细节,更能反应时间序列的变化情况,有利于时序数据分析的实现.

时间序列的模式表示主要可以有如下几种类型:奇异值表示[86],频域表示法,分段线性表示[87],基于模型的表示[88],基于空间的表示[89],基于符号的表示等[90].其中,基于符号的表示作为一种新的模式表示,以其良好的前景成为近年研究的热点[91].因而也是本文关注的重点.

时间序列的符号化表示就是通过一些离散化方法将时间序列的实数值或者一段时间

内的时间序列波形映射到有限的符号表上,将时间序列表示为有限符号的有序集合,即字符串.符号化表示的优点在于可以利用许多字符串研究领域的成果,缺点在于如何选择合适的离散化算法,解释符号的意义,以及定义符号之间的相似性度量.Agrawal 等人[92]将时间序列的波形符号化,引入多种符号算子;Park 等人[93]直接将时间序列的值采用等宽离散化方法和最大嫡方法符号化;其中,大多数符号表示方法都存在以下致命缺陷:

(1),时序数据是典型高维数据,时序数据降维是其高度期望的属性.但大多数符号时序无法降维;

(2),定义的符号时序无法进行距离计算,也就无法进行相似比较,因而缺乏实用价值.

2003 年,Eammon Keogh 博士在 2003 年提出的新型符号化表示方法-SAX(符号聚合近似)方法[90],首次有较地克服了上述致命缺陷.并在许多应用中性能都优于或至少不低于其它的方法[91],因而受到了广泛关注.本文也对此作了认真的研究,并取得了一些成果.在本文后面的第三章将详细论述有关时序数据符号化表示研究的新模型和相关算法.

2.3 特征检测系统基本模型

2.3.1 基本思路

一旦解决了时序数据的合适表示方式问题,下一个问题将是如何建立特征检测的模型.正如前述,这仍是一个十分艰巨的任务.但对于某些特殊的任务,例如异常检测.本文建议直接采用无模型的解决方案,因为它更加容易实现.基于 OCSVM 的异常检测正是这种无模型解决方案.

OCSVM 的检测是基于空间划分的无监督异常检测算法.与传统的异常检测方法不同的是.无监督的异常检测不需要任何标记数据,并能够从未标记的数据集中找出隐藏在其中异常.其基本思想是:在特征空间中,正常数据和故障数据（异常数据）占据不同的区域,如果能够估计出正常数据在特征空间中所占的区域,那么不在这个区域内的点,自然可以认为是异常点.2001 年,Scholkopf 等人基于 SVM 给出了在特征空间中估计大部分数据出现区域的方法[44],而由假设,大部分数据就是正常数据,因此,可以使用该算法来估计正常数据出现的区域,进而检测异常. Larry M 将 SVM 算法应用到非监督情况[94],提出了 OCSVM 算法,与传统的 SVM 不同, OCSVM 试图用超平面将数据集中所有的点和原点分开,通过最大化分类边界压缩数据集在特征空间中所占据的区域来区分正常与异常点.从而奠定了用支持向量机实现异常检测的理论基础. 随后有很多的学者提出了众多的用 OCSVM 实现异常检测的有效方案[95-96].形成了一个研究热点.

Davy 等人(2006)在综合前人成果的基础上提出了基于 OCSVM 序列优化的无模型式的异常检测改进模型[97].通过引入系列优化算法,有效实现了无监督学习的异常检测,并在理论上和实践中验证了其有效性.国内也有一些相关研究文献报道[52,98].OCSVM 的异常检测能

够发现各种已知或未知故障,计算代价小,精度高.并且是全局最优解,泛化性好.是比较理想且有开发前景的方法.但其缺点是只能处理静态矢量数据,不能直接处理动态时序数据.另外,它在高维情况下的表现还不尽人意.因而有待进一步改进.

本节利用相空间重技术克服以往的OCSVM无法处理非矢量且属于高维的时序数据的缺陷,提出基于OCSVM的时序特征检测模型.在此基础上实现OCSVM时序异常检测算法.从而解决了机器人传感器故障在线检测的实际问题.

2.3.2 OCSVM 模型

针对传统异常检测算法模型的弊端,文献[99]于2001年提出了基于SVM估计在特征空间中大部分数据出现区域的方法, 用 OCSVM,通过最大化分类边界压缩数据集在特征空间中所占据的区域,就可用一个超平面将正常数据与异常数据分开.后来, Tax 等人[100]进一步提出了用超球面作为划分面的模型,此模型假定正常的数据在球面以内,异常数据在球面以外;在求解这个最优化问题之后,对每一点实例如果判定函数为正就是正常点,反之,为异常点.

可以将上述方法表述为对一个正类样本集{ $\boldsymbol{x}_i \quad i=1,2,...l,\ x_i \in R^d$, }, 为了寻找一个以 a 为中心,以 R 为半径的包含全部样本点的最小超球体,求解如下优化问题

$$min \quad F(R,\boldsymbol{a},\boldsymbol{x}_i)=R^2+C\sum_{i=1}^{l}\boldsymbol{x}_i \tag{2-1}$$

$$\begin{aligned} S.t \quad & \left\|\Phi(\boldsymbol{x}_i)-a\right\|^2 \le R^2+\xi_i \quad i=1,2,...l \\ & \xi_i \ge 0 \quad i=1,2,...l \end{aligned} \tag{2-2}$$

这里,0<<x_i 为松驰项,非负的松驰项允许一些点处于超球面的另外一侧,这些点即异常点.而 C 作为惩罚调节项可以使其具备处理噪声数据的功能.相应于最优超球体解的判决函数可为:

$$f(\boldsymbol{x})=\mathrm{sgn}(R^2-\left\|\Phi(\boldsymbol{x})-a\right\|^2) \tag{2-3}$$

上面二次优化问题的拉格郎日函数为:

$$L(R,a,\xi)=R^2+C\sum_{i=1}^{l}\boldsymbol{x}_i+\sum_{i=1}^{l}a_i(\left\|\Phi(\boldsymbol{x}_i)-a\right\|^2-R^2-\xi_i)-\sum_{i=1}^{l}\beta_i\xi_i \tag{2-4}$$

其中, $a_i \ge 0, \beta_i \ge 0,$

令相关变量 R, a, ξ 的偏导数为零,可求得

$$\sum_{i=1}^{n} a_i = 1, \quad 0 \le a_i \le C, \quad a = \sum_{i=1}^{n} a_i$$

将此结果代入原式,并将高维空间优化中的内积运算采用满足 mercer 条件的核函数代替,即找一个核函数,使得 $K(\boldsymbol{x},\boldsymbol{y}) = <\Phi(\boldsymbol{x}),\Phi(\boldsymbol{y})>$

于是优化问题的对偶形式为:

$$\sum_{i=1}^{l} a_i K(\boldsymbol{x_i},\boldsymbol{x_j}) - \sum_{i=1}^{n}\sum_{i=1}^{l} a_i a_j K(\boldsymbol{x_i},\boldsymbol{x_j}) \qquad (2\text{-}5)$$

$$\sum_{i=1}^{l} a_i = 1$$

$$0 \le a_i \le C \qquad \text{i=1, 2}\cdots\text{. 1}$$

解这个对偶问题,可得 a_i,通常只有少量 $a_i \ge 0$ 相应的点为支持向量.超球体的中心可由式

$$a = \sum_{i=1}^{l} a_i \Phi(\mathbf{x}_i)$$

求出,而半径 R 可由任一支持向量代入下式求得.

$$R^2 - (K(\mathbf{x}_i,\mathbf{x}_j) - 2\sum_{j=1}^{l} a_j K(\mathbf{x}_i,\mathbf{x}_j) + a^2) = 0 \qquad (2\text{-}6)$$

判决函数的最终形式为

$$f(\mathbf{x}) = K(\mathbf{x}_i,\mathbf{x}) - 2\sum_{i=1}^{l} a_i K(\mathbf{x}_i,\mathbf{x}) + \sum_{i=1}^{l}\sum_{j=1}^{l} a_i a_j K(\mathbf{x}_i,\mathbf{x}_i) \qquad (2\text{-}7)$$

当然 $f(\boldsymbol{x})$ 也可写为:

$$f(\mathbf{x}) = \mathrm{sgn}[\sum_{i=1}^{n} a_i K(\mathbf{x}_i,\mathbf{x}) + b] \qquad (2\text{-}8)$$

$$b = -\frac{2}{N_{msv}} \sum_{k}\sum_{i} a_i K(\mathbf{x}_k,\mathbf{x}_i) \qquad (2\text{-}9)$$

式中, N_{msv} 为边界向量数.对于给定的 $\mathbf{x}$,当 $f(\mathbf{x}) > 0$ 时判定为正常点,而当 $f(\mathbf{x}) \le 0$ 则判定为异常点.由于传感器记录的数据属于时序数据,而支持向量机只能用于矢量数据. 因此,此模型无法直接运用.解决办法是用相空间重构技术行将时序数据转化成矢量数据.

2.3.3 相空间重构

（1） 相空间重构理论简述

混沌是一种低阶确定性的非线性动力系统所表现出来的非常复杂的行为,它对现代科学具有广泛而深远的影响,几乎覆盖了一切学科领域,尤其是在数学、物理学、化学、生物学、经济学等方面得到了广泛的应用.在对混沌时间序列的各种分析中,如混沌预测.动力学不变量的估计.混沌信号的诊断等,所要进行的第一步工作是要对混沌信号进行相空间重构.

1981 年 Takens 提出了相空间重构的延时坐标法,奠定了相空间重构技术的基础,这种方法用单一的标量时间序列来重构相空间,包括吸引子、动态特性和相空间的拓扑结构.现已成为最主要、最基本的相空间重构方法[101,102].

Packard 等[103]于 1999 年提出了由混沌时间序列重构相空间的两种方法:导数重构法和坐标延迟重构法.并采用导数重构法重构了 Rossler 吸引子,求出了 Lyapunov 指数.Grassberger 和 Procaccia 则提出了关联积分的概念及计算公式,采用时间延迟法,从一维时间序列中重构了 Henon 映射、Lorenz 方程、Logistic 方程和 Kaplan—York 映射等典型混沌系统,并求取了这些混沌系统的分数维.G—P 算法的提出,是混沌时间序列研究中的一个极其重要的突破,它使得对混沌时间序列的研究不仅仅局限在已知的混沌系统,如 Rossler 系统、Henon 映射和 Lorenz 系统等,而是任何实测混沌时间序列.这为混沌时间序列的研究进入实际应用奠定了新的基础.

从总体上讲,在时序存在噪声的情况下,系统的一些不变量如:分维数 d 和 Liapunov 指数等与 m (嵌入维)和 τ (延迟时间)的选取有关,所以相空间重构的中心任务便是合理的选取 m 和 τ .

相空间重构的基本方法有三种,它们分别是时间延迟法[104] ,导数法和基本分量坐标法[105].到目前为止这三种方法之间的关系还没有完全弄清楚.

目前的大多数混沌时序相空间重构都采用坐标延迟相空间重构法.它本质上是通过一维时间序列 $x=\{x(t_i)\}$, $i=1,2,...N$ 的不同时间延迟 0, t , $2t$,…, (m-1) t 来构造 m 维相空间矢量: $x(t_{ij})$ 其中, $j=0,1,2,...m-1$.

Takens 定理指出,对于理想的无限长和无噪声的一维时间序列,嵌入维数 m 和时间延迟 t 可以取任意值,但实际应用中的时间序列都是有限长度且存在噪声,因此必须细致地确定嵌入维数 m 与时间延迟 t ,才能保证重构的相空间的质量.

(2) 相空间重构技术实现:

为了能够从“一维”时间序列中“还原”动力系统相空间的几何结构,需要把一维时间序列嵌入到 m 维空间中:对已知观测的时间序列

$$x_i=x(t_i),\ t_i=t_0+i\Delta t,\ i=1,2,...N$$

在等距离采样,取时间延迟为 τ (必须为 Δt 整数倍)情况下,可用下式定义 m ($m>D$)维相空间.

$$\boldsymbol{X}(t_i)=\{x(t_i),...x(t_i+\tau),...x(t_i+(m-1)\tau)\} \qquad (2\text{-}10)$$

并可构造如下 m 维相空间分布矩阵:

$$\boldsymbol{X}=[\boldsymbol{X}(t_1),\boldsymbol{X}(t_2),...\boldsymbol{X}(t_i),...\boldsymbol{X}(t_{n'})]^T$$

其中,

$$\begin{aligned}
&\boldsymbol{X}(t_1)=(x(t_1),...x(t_1+\tau),...x(t_1+(m-1)\tau))\\
&\boldsymbol{X}(t_2)=(x(t_2),...x(t_2+\tau),...x(t_2+(m-1)\tau))\\
&\boldsymbol{X}(t_i)=(x(t_i),...x(t_i+\tau),...x(t_i+(m-1)\tau))\\
&\cdots\cdots\\
&\boldsymbol{X}(t_{n'})=(x(t_n),...x(t_n+\tau),...x(t_n+(m-1)\tau))
\end{aligned}\qquad (2\text{-}11)$$

式中, $t = KVt(k = 1,2,...)$. $X(t_i)$ $(i=1,2,...n')$ 为 m 维相空间的一个相点. $n'=n-(m-1)k$.

（3）τ 值的选取：

在相空间重构的过程中 τ 的取值非常重要,要保证相空间重构的正确性,所选用的延迟时间必须使重构相空间的各个分量保持相互独立.如果 τ 的值取的太少,时间序列的任意两个相邻延迟坐标点又非常接近,不能相互独立,将会导致数据的冗余.如果 τ 的值取的太大,轨线在相空间中会出现间断现象,不能反映整个系统的特性;这样一来有可能导致比较简单的几何图线在相空间中看起来非常复杂,系统的相图失真.于是围绕这一条件便先后出现了用自相关函数和互信息来确定延迟时间的方法[106].由于互信息法较自相关函数更适合本章研究的问题,因此,下面着重论及互信息法.

（4）互信息法

互信息方法是估计重构相空间时间延迟的一种有效方法，它在相空间重构中有很广泛的应用. Shaw 首先提出互信息第一次达到最小时的滞时作为相空间重构的延迟时间, Faster 给出了互信息计算的递归算法[196].

考虑两离散信息序列 $\{s_1,s_2,...s_n\}$ 和 $\{q_1,q_2,...q_n\}$ 构成的系统 S 和 Q.则由信息论,从两系统测量中所获得的平均信息量,即信息熵分别为

$$\begin{aligned}
H(S)&=-\sum_{i=1}^{n}P_s(S_i)\log_2 P_s(S_i)\\
H(Q)&=-\sum_{i=1}^{m}P_q(q_j)\log_2 P_q(q_j)
\end{aligned}\qquad (2\text{-}12)$$

其中 $P_s(S_i)$, $P_q(q_j)$ 分别为 S 和 Q 中事件的概率. 在给定 S 的情况下,我们能获取的系统 Q 的信息,即系统 S 和 Q 的互信息为：

$$I(Q,S)=H(Q)-H(Q\mid S)\qquad (2\text{-}13)$$

而

$$H(Q\mid s_i)=-\sum_j[P_{sq}(s_i,q_j)/P_s(s_i)]\log[P_{sq}(s_i,q_j)/P_s(s_i)]\qquad (2\text{-}14)$$

$$I(Q,S)=\sum_i\sum_j P_{sq}(s_i,q_j)\log_2[P_{sq}(s_i,q_j)/P_s(s_i)P_q(q_j)] \qquad (2\text{-}15)$$

这里 $P_{sq}(s_i,q_j)$ 为事件 s_i,q_j 的联合分布概率. 定义 $[s,q]=[x(t),x(t+\tau)]$ 即 s 代表时间序列 $x(t)$, q 为其延迟时间为 τ 的的时间序列 $x(t+\tau)$, , 则 $I(Q,S)$ 显然是与时间延迟 τ 有关的函数, 不妨记为 $I(\tau)$. $I(\tau)$ 的大小代表了在已知系统 S 即 $x(t)$ 的情况下, 系统 Q 也就是 $x(t+\tau)$ 的确定性的大小. $I(\tau)$ = 0, 表示 $x(t)$ 与 $x(t+\tau)$ 完全不可预测; 互信息法中的关键问题是联合分布概率 $P_{sq}(s_i,q_j)$ 的计算, Fraser 和 Swinney 采用的是等概率递推的方法, 其划分与计算很复杂, 杨志安等提出了等间距格子法, 其计算相对简单[106, 107]. 在最优时间延迟的估算中被广为采用的模型如[108], 国内的例子如[109-111].

对于一般离散变量序列 $X_0,X_1,...X_n$, 互信息为

$$I_0(X_0,X_1,...X_n)=\sum_j[H(X_j)-H(X_0,X_1,...X_n)]$$

如果向量是一个延迟时间重构, 则 $I_n(\tau)$ 第一次达到最小值的时滞可作为相空间重构的时间延迟. 同时如果 n 足够大则 $I_n(\tau)$ 应该是单调递减, 并且 I_n/n 给出系统熵的一个很好估计. 在实际应用中, FNN+互信息已成相空间重构中确定嵌入维数和时间延迟的普遍作法. 需要指出的是, 互信息法尽管是为确定相空间重构中的最优时延而提出的, 但它由于可以度量非线性相关性, 因此, 在判定两非线性序列的相关性方面得到了广泛的应用[112-114].

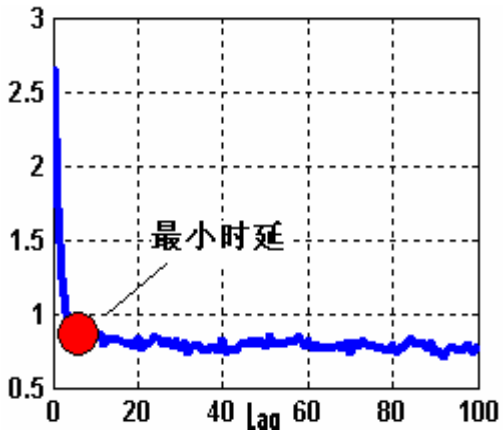


图 2-4 互信息法求时延(数据 1)

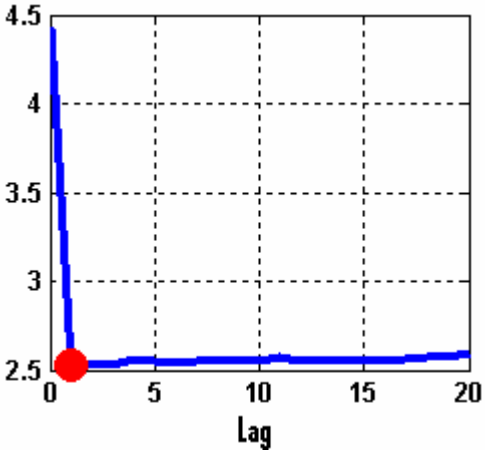


图 2-5 互信息法求时延(数据 2)

为实现时序数据的矢量化转换, 本章用互信息法确定两种典型时序延迟的结果如相关图表所示. 上面图 2-4, 2-5 显示的是用互信息法对不同时序数据集求时延的典型结果, 图中红圈中的点所对应的横坐标便是所求相应时序数据集的时延. 它正好对应于 $I_n(\tau)$ 第一次达到最小值的时滞. 显然, 对于时序数据 1, 数据 2 而言, 所求得的时延分别是 6 和 2.

（6） 嵌入维 m 的选取

嵌入维的确定过程中,关联维计算是关键所在,重构相空间维数的取值依赖于所分析和研究的关联维数,数据序列的关联维数不同,所选取的重构相空间维数亦不同.由单变量的时间序列重构相空间时,为了保证该相空间能包含原状态空间吸引子的特征,关联维应该取得足够大.按照 Takens 在 1980 年证明的嵌入维数大小的充分条件:

$$m \geq 2d+1 \qquad (2\text{-}16)$$

其中: m -重构相空间维数;d-原状态空间吸引子所处空间的关联维数.该定理表明,嵌入空间的维数至少是吸引子维数的两倍,这时重构的相空间和原系统的状态空间拓扑等价.

在实际的分形分析中,关联维数是所要求取的对象,并不知道其具体数值,因此,需首先估计出所求关联维的取值范围,从而得出重构相空间维数 m 的粗略估计值.在 m 粗略估计值的范围内对 m 取不同的值,然后分别求取系统的关联维数,当关联维数达到饱和时的嵌入维 m 取值,即为重构相空间的实际维数.

由于 Takens 只是嵌入维的充分条件.实际计算中,只要 $m>d$ 嵌入空间中点集的维数就等同于吸引子的维数.J.P.Eckman 等人已证明 m 可在 $d \leqslant m \leqslant 2d+1$ 中取值[115].

本文采用假近邻方法(False Nearest Neighbors,FNN)作为计算关联维数的具体方法.从几何的观点看,混沌时间序列是高维相空间混沌运动的轨迹在一维空间一时间上的投影,在这个投影的过程中,混沌运动的轨迹会扭曲.高维相空间中并不相邻的两点投影在一维时间-空间上时却会成为相邻的两点,即虚假邻点,这就是混沌时间序列呈现出无规律的原因所在.重建相空间,其实就是从混沌时间序列中恢复混沌运动的轨迹,不难想到,随着嵌入维数的增大,混沌运动的轨道将会逐渐打开、无扭曲、缠绕,虚假邻点也被剔除,从而混沌运动的轨迹得到恢复,这个思想就是 FNN 的出发点[116] .

d 维相空间中,每一个相点矢量 $X(k)=[x(k),x(k+T),...x(k+(d-1)T]$,都有一个某距离内的最近邻点 $x^{NN}(k)$ 其距离为

$$R_d(k)=\left\|X(k)-X^{NN}(k)\right\| \qquad (2\text{-}17)$$

$R_d(k)$ 应该是个很小的量,对于 N 个数据,其量值应大约在 $1/N^{1/d}$ 左右.当相空间的维数从 d 维增加到 $d+1$ 维时,这两个相点的距离就会发生变化,而成为 $R_{d+1}(k)$ 且

$$R_{d+1}^2(k)=R_d^2(k)+\left\|X(k+Td)-X^{NN}(k+Td)\right\|^2 \qquad (2\text{-}18)$$

若 $R_{d+1}(k)$ 比 $R_d(k)$ 大很多,可以认为这是由于高维混沌吸引子中两个不相邻的点在投影到低维轨道上时变成相邻的两点所造成的,因此这样的邻点是虚假的.令

$$S_n=\frac{\left\|X(k+Td)-X^{NN}(k+Td)\right\|^2}{R_d(k)} \qquad (2\text{-}19)$$

若, $S_n>R_T$ 则 X^{NN} (k)是 $X(k)$ 的假近邻,阈值 R_T 可在(10,50)内选择.若数据包含噪音.则当 $R_{d+1}(k)/R_A$ ³ 2时,可认为 $X^{NN}(k)$ 是 X(k)的虚假最近邻点,其中,

$$R_A=\frac{1}{N}\sum_{i=1}^{N}(X(k)-\bar{X})^2, \qquad (2\text{-}20)$$

$$\bar{x}=\frac{1}{N}\sum_{i=1}^{N}x(k) \tag{2-21}$$

对实测时间序列, 从嵌入维数的最小起始值开始, 计算虚假最近邻点的比例, 然后增加 d, 直到虚假最近邻点的比例小于 5%或者虚假最近邻点不再随着 d的增加而减少时, 可以认为混沌吸引子已完全打开, 此时的 d即为嵌入维数, 在相空间嵌入维数的确定方面, FNN 被认为或许是最有效的计算嵌入维的方法 [117]. FNN 的计算结果与 τ 有关, 给定不同的 τ 值, 会得出不同的最小嵌入维数. 考虑到这一点, Cao(曹氏法)[118]采用与 FNN 类似的思路, 对于时间序列 $x_1,x_2,...x_n$ 及相空间矢量

$$x_i(d)=(x_i,x_{i+\tau},...x_{i+(d-1)\tau}),i=1,2,...N-(d-1)\tau \tag{2-22}$$

定义

$$a(i,d)=\frac{\left\|x_i(d+1)-x_{n(i,d)}(d+1)\right\|}{\left\|x_i(d)-x_{n(i,d)}(d)\right\|}\quad i=1,2,...N-\tau \tag{2-23}$$

$a(i,d)$ 的均值为

$$E(d)=\frac{1}{N-d\tau}\sum_{i=1}^{N-d\tau}a(i,d) \tag{2-24}$$

则有 $E_1(d)=E(d+1)/E(d)$, 若 $E_1(d)$ 自某个 d_0 开始停止变化, 则 d_0+1即为所寻找的最小嵌入维数.

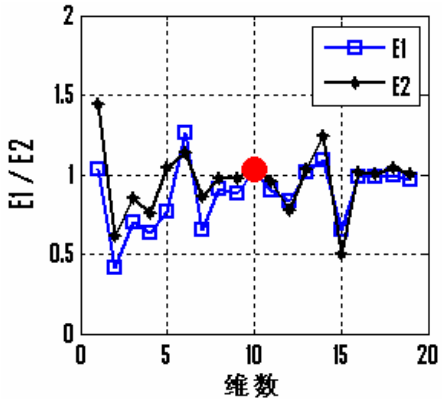


图 2-6 曹氏法计算嵌入维(数据 1)

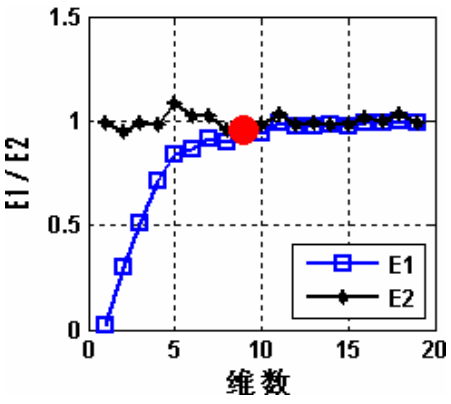
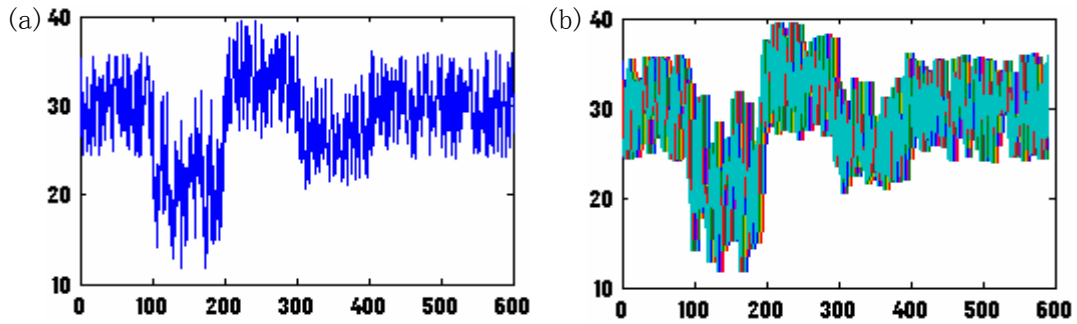


图 2-7 曹氏法计算嵌入维(数据 2)

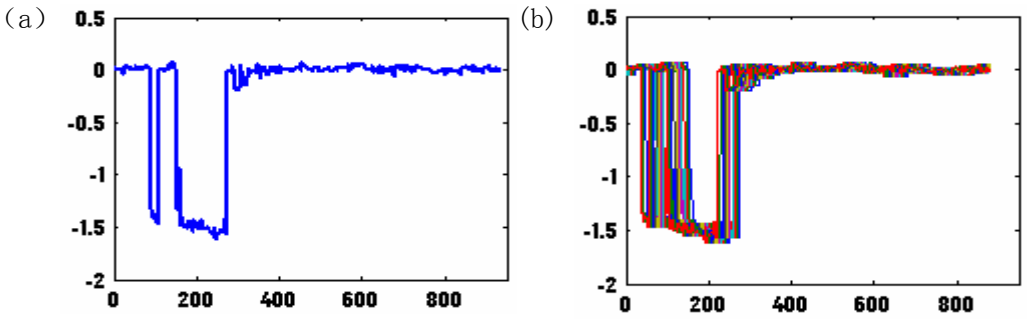
图 2-6, 2-7 是按照这个思路实现两个典型时序数据集的嵌入维的情况：其中, 图 2-7 反映的是按曹氏法计算所得的机器人速度传感器记录时序数据集的最佳嵌入维.



(a) 原始数据 (b) 重构数据

图 2-8 相空间重构(数据 3)

图 2-8 (a), (b)显示的是用相应的时延和嵌入维对不同时序数据集进行相空间重构的情况.



(a) 原始数据 (b)重构数据

图 2-9 (a), (b) 相空间重构(数据 4)

图 2-9(a), (b)是机器人速度传感器所记录的时序数据集片断的相空间重构情况.

2.3.4 特征检测在线实现算法

移动机器人传感器故障诊断的在线实现算法可简述如下:

1). 按前面所述方法对时序数据预处理,将其转化成矢量数据集. $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,...\boldsymbol{x}_n\}$;

2). 用固定平移窗长度 K 向后移动窗口,(每次)从 $\boldsymbol{X}$ 中选出 k 个数据用于训练检测模型,(其中新增一个最新的数据点,并移去一个最老的数据点.即 t 时刻的训练数据集为; $\boldsymbol{X}_t=\{\boldsymbol{x}_{t-k},\boldsymbol{x}_{t-k+1},...\boldsymbol{x}_{t-1}\}$, 这里, $t=k+1;t+1$ 时刻的训练集为 $\boldsymbol{X}_{t+1}=\{\boldsymbol{x}_{t-k+1},\boldsymbol{x}_{t-k+2},...\boldsymbol{x}_t\}$;

3). 用训练所得判决函数 $f(\mathbf{x})$ 来检验各点的异常,检验结果保存为 $I=I(t)=f(\mathbf{x})$;

4). 令 $t=t+1;(t=k+1,k+2,...n)$ 重复步骤 2~4.

2.4. 实验分析

本文采用本院实验室的移动机器人速度传感器的测量数据作故障诊断的仿真实验.

实验步骤:

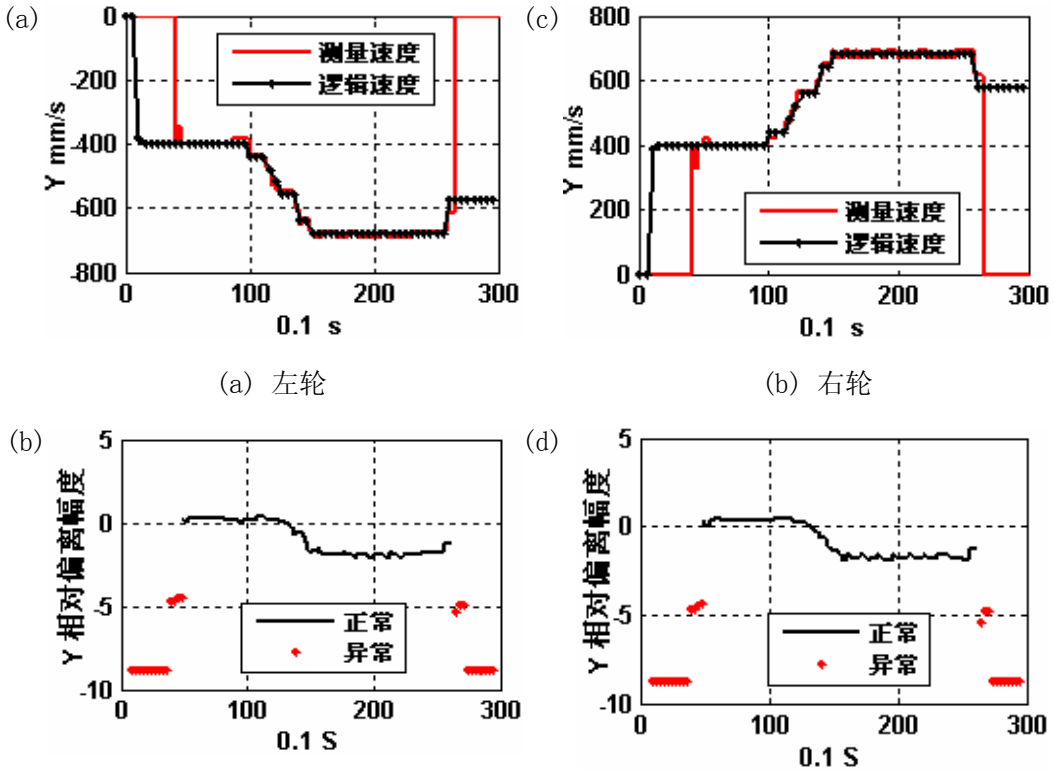
1. 分析时间序列的混沌特性

   对时序数据进行混沌特性的分析,当所分析数据满足:

   (1).频谱很宽; (2).最大 Lyapunov 指数大于 0.

   这两个条件时,可以认为它们具有混沌特性,于是可进行下面步骤.

2. 确定是否是混沌序列,如果是,则用前面所述互信息法求取最佳时延.
3. 对每一层分解的信号分别用曹氏方法求取最佳嵌入维数.
4. 根据求得最佳时延和嵌入维数,构造输入输出向量,并分别用 OCSVM 进行检测.



(a) 左轮 (b) 右轮

(c)左轮 (d) 右轮

图 2-10 t 时刻的异常检测结果(参数 1/C=0.73, s=1.5)

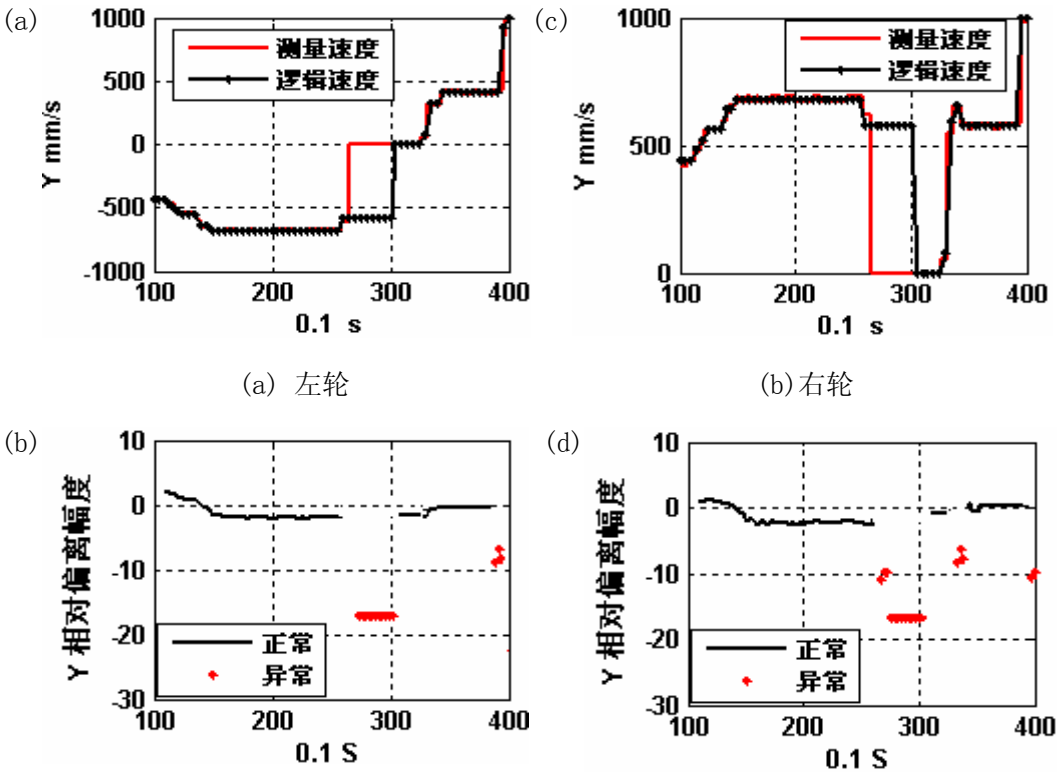
如图 2-10 所示,其中左图对应左轮(1 轮)速度传感器检测和诊断情况,右边是对应于右轮(2 轮)的检测及诊断结果.图 2-10 中的黑线代表该轮的逻辑速度,红线则是传感器测量速度,采样时间间隔为 0.1x300=30 s (秒).其中,从 0.5-4 秒及 26.5-30 秒 这两个区间内的测量速度被人为的设置为零,而逻辑速度为 400 mm/s(这是人为设置的传感器失效故障),而从 9-26.5 秒这个区间内测量速度与逻辑速度不相吻合的无规则曲线也是通过人为制造轮子打滑条件下形成的.目的是想检测本文所论故障诊断方法的灵敏度,测准率,误诊率.

须要说明的是,本实验中逻辑速度只是作为参考对照用,而故障诊断实际上只用传感

器测量速度一种数据.即只用传感器测量数据本身来建模并检测其中的异常.

实验首先对输入数据进行了 0 均值和振幅为 1.的预处理.然后用相空间重构技术将时序数据重构为 m =2, t =11 的二维向量.其中 m , t 的最佳值是而是借助于前面所述的互信息法和曹氏法测定的.

在数据预处理基础上,用 OCSVM 进行学习训练并以此结果进行异常检测,其检测结果如图 2-10,2-11 各图的(c),(d)子图所示.其”相对偏离幅度”大于±3(允许误差范围)的点被视为所检测到的故障点.其偏离的距离反映了传感器测量速度与逻辑速度的偏离程度(需要说明的是,图中所示的偏差距离已与实际偏差数值不再严格对应,因为经过数据预处理和相空间重构后其幅度发生了很大变化);例如图中红色点为所检测到的传感器故障点 (即异常数据,为突出醒目这些点被特意放大),而黑线则代表正常状态.其中,图 2-10 是在一类向量机的调节参数设置为 1/C=0.73,而高斯宽度参数 s 设置为 1.5 时的结果.实验表明,本文所论的故障诊断方法有良好的表现.它不但能够发现传感器的编码器失效的重大故障(图 2-10 中 0.5-4 秒及 26.5-30 秒时间段),而且也能检测出由于轮子打滑造成的传感器测量速度不准确的小故障(图 2-10 中第 9-26.5 秒时间段-注:这与允许误差范围大小相关).



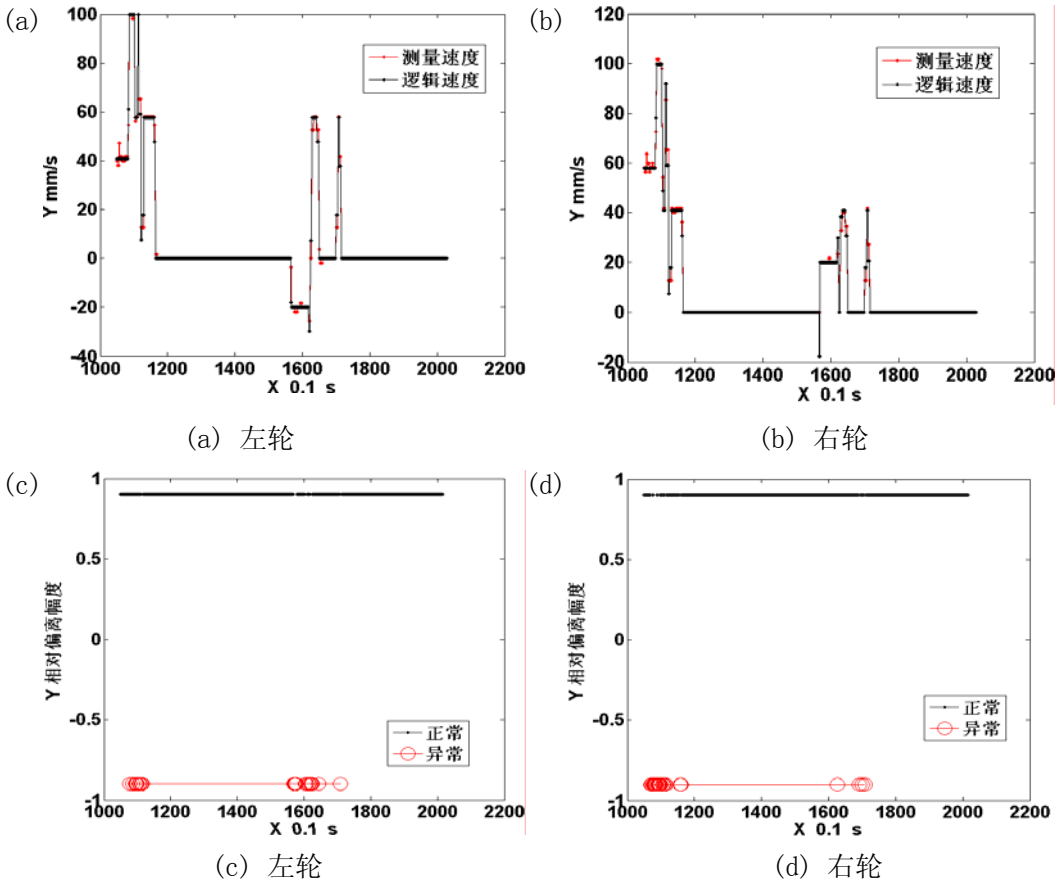
(a) 左轮 (b)右轮

(c) 左轮 (d) 右轮

图 2-11 t +10 秒时的异常检测结果(参数 1/C=0.73, s=1.5)

实验还表明,调节参数 nv=1/C 设置和高斯宽度参数 s 的设置将对故障诊断的灵敏度和

测准率产生较大影响. 例如. 调节参数 nv=1/C 和高斯宽度参数 s 分别设置为 0.45 和 1.5 时(注:未附图),除了较大的编码器失效故障仍可以被检测到以外,轮子打滑类的(位于 9-14 秒区间的)一些较小故障被漏检了. 另外,此算法也可方便地实现在线检测. 图 2-11 便是用 10 秒后用平移法所得更新时序数据在线检测的结果.



(a) 左轮 (b) 右轮

(c) 左轮 (d) 右轮

图 2-12 用于速度异常检测的实验系统检测结果

图 2-12 是用于在线异常检测的实验系统在某一时刻的检测结果,它主要用于检测异常或突变所发生时刻(位置)的情况. 图中的左、右边的子图分别对应待检测左、右机器人两轮的检测的实况,而同一边的上下子图对应同一轮子速度传感器数据的检测结果. 其中,上面的子图记录的是速度传感器的逻辑速度与实际测量速度. 下面的子图是用本文设计的检测系统进行检测的结果示例. 如图所示:在上子图中逻辑速度与测量速度差异较大的时点,相应的下子图中将检测出异常(图中的红色点所示). 例如左边图中,在 105s-120s 及 150s-170s 区间,测量速度和逻辑速度有较大差异(左上子图中),因而都检测出一些异常点(左下子图中).

机器人速度传感器数据的多次检测实验表明,本文的所述模型和相应算法在应用中表现良好,误差率一般小于 7%.因而能够较好地达到检测异常的预定目标.尤其是能够较准确地察觉突变点所发生的时刻(位置).因而对于智能机器的故障诊断有较好实用价值.

2.5 面临新问题及解决思路

因为 SVM 的学习性能与所选用的学习参数密切相关, 作为一种固定学习参数的模型,上面所述 OCSVM 检测系统只适用于特定的(与固定学习参数相适应的)时序数据.然而,实际应用中的数据变化差异巨大,用固定学习参数的 OCSVM 模型去处理变化巨大的动态工作流数据会产生截然不同的检测效果,有时甚至会出现错误.因而前面所述的 OCSVM 并不适用于变化过于复杂的动态时序数据,尤其是来自在复杂环境中移动机器人传感器的数据.解决 OCSVM 模型稳定性问题的基本方法有二:其一是根据不同的的工作流数据选择不同的学习参数,这显然是不方便也不适用的方案;其二是设计一个自适应学习算法,让系统根据实际情况自动确定学习参数,这才是较理想的方案.在这方面,已有一些学者作了有益的尝试,且取得一些成果. 最常用且可靠的参数选择方法是在判定参数范围基础上对整个参数空间进行网格式搜索,从中找出最佳值.然而其运算代价太大而不适合在线实时运用.其它算法如;基于实验设计的方法[120];代价估计法[121];基于在线高斯过程的搜索法[122];梯度下降法[123];混沌优化方法;最小最大化方法[124];在一定条件或范围内都能够解决一些特定问题.但都有计算代价大,费时长,难以用于在线实时情况.

本文受上述前人成果启发,综合他们的长处,提出自适应 CPSO(混沌粒子群优化)方法来实现动态环境下检测系统学习参数的自动选择.并以此为基础,构成基于 OCSVM(OCSVM)和 CPSO(混沌粒子群优化)组合式自适应特征检测模型.其基本思路是:采用 OCSVM 作为基本的检测模型学习器[130],用 CPSO 作为最优参数搜索器,正常情况下直接用 OCSVM 实现在线故障检测,当发现检测标准低于事先确定的阈值时,启动 CPSO 搜索器自动搜索新的最佳参数,一旦检测标准达到正常,则退出搜索过程,并启用新的参数来继续进行 OCSVM 在线故障检测.实验表明,此模型确实能够实现高精度的实时在线故障检测.是一种相当有效的解决方案.

2.6 粒子群优化理论及基本 PSO 算法

实现此自适应组合系统,涉及到 CPSO(混沌粒子群优化)算法.它是在基本 PSO 模块中嵌入混沌搜索子模块实现的.

2.6.1 基本 PSO(粒子群优化)算法:

基本 PSO[125]是一种模仿生物个体与群体相互作用特征,通过叠代寻找全局最优解的方法.它将每一个体称为”粒子”(粒子的位置即候选解 $\boldsymbol{x}_i$).在每一次叠代中,粒子通过跟踪两个“极值”来更新自己.一个是粒子本身所找到的最优解,即个体极值 $\boldsymbol{x}_p$.另一个是整个

种群目前找到的最优解,称之为全局极值 $\boldsymbol{x}_g$. 粒子在找到上述两个极值后,就根据下面两个公式来更新自己的速度与位置:

$$\boldsymbol{v}_{i+1} = w \times \boldsymbol{v}_i + r_1 \times rand(\ )(\boldsymbol{x}_{pi} - \boldsymbol{x}_i) + r_2 \times rand(\ )(\boldsymbol{x}_{gi} - \boldsymbol{x}_i) \qquad (2\text{-}25)$$

$$\boldsymbol{x}_{i+1} = \boldsymbol{x}_i + \boldsymbol{v}_{i+1} \qquad (2\text{-}26)$$

其中, $\boldsymbol{v}_i$ 和 $\boldsymbol{v}_{i+1}$ 分别是第 $i, i+1$ 代粒子群的当前速度和更新速度, $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_{i+1}$ 分别是第 $i, i+1$ 代粒子群的当前位置和更新位置, $\boldsymbol{x}_{pi}$ 与 $\boldsymbol{x}_{gi}$ 分别为第代个体最优与全局最优值. rand()是(0,1)之间的随机数 r_1 和 r_2 被称作学习因子,W 是加权系数,取值在 0.1 到 0.9 之间. 粒子通过不断学习更新,最终飞至解空间中最优解所在的位置,搜索过程结束. 最后输出的 x 就是全局最优解. 在更新过程中,粒子每一维的最大速率被限制为 $\boldsymbol{v}_{\max}$, 粒子每一维的坐标也被限制在允许范围之内.

该算法原理简单, 收敛快, 特别是在运行的早期, 但也存在着精度较低, 易发散等缺点. 若加速系数、最大速度等参数太大, 粒子群可能错过最优解, 算法不收敛; 而在收敛的情况下, 由于所有的粒子都向最优解的方向飞去, 所以粒子趋向同一化(失去了多样性), 使得后期收敛速度明显变慢, 同时算法收敛到一定精度时, 无法继续优化, 所能达到的精度也比 GA 低[126], 后期易于陷入局部极值点. 为了克服这一缺点. 需要作出相应的改进. 因此很多学者都致力于提高 PSO 算法的性能.

2.6.2 PSO 的改进算法

PSO 有全局版和局部版两种, 全局版收敛快,但有时会陷入局部最优. 局部版 PSO 通过保持多个吸引子来避免早熟.

假设每一个粒子在大小 l 的邻域内定义为一个集合

$$N_i = \{pbest_{i-l}, pbest_{i-l+1}, ...pbest_{i-1}, pbest_i, pbest_{i+1}, ...pbest_{i+l-1}, pbest_{i+1}\} \qquad (2\text{-}27)$$

从 N_i 中选出最好的,将其位置作为 $\boldsymbol{x}_l$ 代替公式中的 $\boldsymbol{x}_g$, 其他与全局版 PSO 相同. 实验表明,局部版比全局版收敛慢,但不容易陷入局部最优[126]. 在实际应用中, 可以先用全局 PSO 找到大致的结果,再用局部 PSO 进行搜索. 在此基础上,还衍生出许多改进算法.

由式(2-25)可以看出,公式的右边由三部分组成. 第一部分是粒子更新前的速度,而后两部分反映了粒子速度的更新. Shi 与 Eberharl 研究发现公式(2-25)等式的第一部分 $\mathbf{v}$; 由于具有随机性且其本身缺乏记忆能力,有扩大搜索空间,探索新的搜索区域的趋势. 因此,具有全局优化的能力. 在考虑实际优化问题时,往往希望先采用全局搜索,使搜索空间快速收敛于某一区域,然后采用局部精细搜索以获得高精度的解. 因此,在式(2-25)的第一项前乘以加权系数, w 较大算法具有较强的全局搜索能力, w 较小则算法倾向于局部搜索. 一般的做法是将 w 初始为 0.9 并使其随迭代次数的增加线性递减至 0.4,以达到上述期望的优化目的. 该方法加快了收敛速度, 提高了 PSO 算法的性能. 其 w 的调节公式可表为

$$w_i = (w_{ini} - w_{end})(G_k - g_i)/G_k + w_{end} \tag{2-28}$$

G_k 为最大进化代数, w_{ini} 为初始惯性权值, w_{end} 为迭代至最大代数时惯性权值. w 的引入使 PSO 算法性能有了很大提高，针对不同的搜索问题，可以调整全局和局部搜索能力，也使得 PSO 算法能成功的应用于很多实际问题.

2.6.3 混沌粒子群优化模型

上述改进算法为探索更有效的新算法提供了很好的启示.但考虑到上述算法中粒子初始化和进化过程的随机性,使 $\boldsymbol{x}_g$ 和 $\boldsymbol{x}_p$ 的更新带有一定的盲目性,影响了进化过程的收敛.为了使粒子群能够跳出进化过程中的陷入局部极小附近的停滞状态, 惰性粒子应由一个新的有活力的粒子取代. 因此,本节提出采用混沌搜索来完成“惰性”粒子的重新初始化. 由于混沌序列具有遍历性的特点[129],即混沌序列可以在一个特定区域内不重复地历经所有状态, 使其成为一种十分有效的搜索工具.

引入混沌序列的搜索算法可在迭代中产生局部最优解的许多邻域点, 以此帮助惰性粒子逃离局部极小点, 并快速搜寻到最优解.混沌搜索的主要思路是通过某特定格式迭代产生混沌序列; 然后通过载波的方式将混沌变量的值域“放大”到优化变量的取值范围空间.这种嵌入混沌序列的混合粒子群优化算法简称为 CPSO(Chaotic particle swarm optimization).CPSO 算法是混沌搜索算法与粒子群优化算法的综合应用.

CPSO 模型为解决无集中控制条件下分布式问题提供了框架.它利用群体中的个体交互搜索复杂空间的优化区[130].在这类动力系统中,智能来自于个体和群体间的混沌平衡.对一个给定的能量或代价函数,通过混沌遍历后[131],此混沌动力系统可能最终达到全局优化或解以较高概率达到其最好近似解.混沌平衡和它们在 PSO 中的特征对于更深刻理解,应用开发和设计新的计算模型非常重要.

1.CPSO（混沌的粒子群优化）算法.

CPSO 是以基本粒子群优化算法的运算流程作为主体流程,把混沌搜索机制引入其中,以此来增强全局搜索能力,摆脱局部极值点的吸引,同时又不降低收敛速度和搜索精度. 其基本的执行过程是先随机产生初始群体,然后开始随机搜索,通过基本的粒子群优化算法（(2-25),(2-26) 式）来产生一组新的个体. 当整个粒子群历史最优粒子位置 $\boldsymbol{x}_g$ 连续不变化或变化极小时,在 $\boldsymbol{x}_g$ 为中心的一定范围内进行混沌搜索,将混沌搜索得到的最优解作为新的 $\boldsymbol{x}_g$ 继续原粒子群算法的求解.

概括说:混沌优化算法CPSO的基本思想是把混沌变量从混沌空间映射到解空间,利用混沌变量具有遍历性、随机性和规律性的特点进行搜索.然后把混沌寻优的结果随机替换粒子群体中的一个粒子.它具有对初值不敏感、易跳出局部极小、收敛速度快、计算精度高、全

局渐近收敛的特点.

混沌搜索算法要点:

将 PSO 最优候选解 $\boldsymbol{x}_g=(\boldsymbol{x}_{g1},\boldsymbol{x}_{g2},...\boldsymbol{x}_{gn})$ 进行混沌优化. 将 $\boldsymbol{x}_{gi}(i=1,2,...n)$ 映射到 Logistic 方程的定义域[0,1],即

$$\boldsymbol{t}_{j,i}=\frac{\boldsymbol{x}_{j,i}-\boldsymbol{a}_i}{\boldsymbol{b}_i-\boldsymbol{a}_i} \tag{2-29}$$

产生本代序列,以此为初始值,进而利用下式

$$\boldsymbol{t}_{n+1}=\mu\boldsymbol{t}_n(1-\boldsymbol{t}_n)\quad n=0,1,2... \tag{2-30}$$

得到序列 $\boldsymbol{t}_{j,i}^{(0,l)}\ ,(l=0,1,...M_{j,l})$

生成家族成员,作逆映射可得下面序列,并用它们进行搜索寻优.

$$\boldsymbol{x}_{j,i}^{(0,l)}=a_i+(b_i-a_i)\boldsymbol{t}_{j,i}^{(0,l)} \tag{2-31}$$

$$i=0,1,2,...n\ ;\quad j=0,1,2,...k\ ;\quad l=0,1,2,...M_{ji}$$

本节直接采用粒子运动的前后速度差绝对值和当前适应度作为启动混沌搜索模块的依据.因而在CPSO运行初期相当于一个PSO模块,只有当检测到粒子速度前后变化小于某一阈值$\boldsymbol{v}_0$,并且群体当前适应度未达到阈值时,才会启动混沌搜索模块,即:

3. 自适应启动模块:

If $|\boldsymbol{v}_{i+1}-\boldsymbol{v}_i|\le\boldsymbol{v}_d$ *and* $f(\boldsymbol{x}_{gi})\le f(\theta)$ (最大化情况)

启动 嵌入了混沌搜索的 CPSO 模块 ;

else

运行 单纯的 PSO 模块

End

其中, $\boldsymbol{v}_d$ 为一事先预定的第 d 维最小速度阈值的正数. $f(q)$ 为适应度阈值. CPSO 算法主要步骤包括:计算家族成员的适应值 $f(\mathbf{x}_{ji})$;根据每代种群的适应值得到局部最优值 $\mathbf{x}_{pi}^{j}$ 和全局最优 $\mathbf{x}_g$;并用这个全局最优候选解随机替换当前群中的一个粒子; 在本例中,设定希望系统达到的检测最低准确率为阈值;因而适应度函数即为该阈值表示的常数.若达到最大叠代次数或适应值大于给定的阈值则结束.否则继续进行搜索.用此算法可尽量减少不必要的运算以提高效率.此组合检测系统的特点是:OCSVM检测模块的最佳学习参数是由CPSO模块在实际运行过程中通过对输入数据模式特征的学习后自动寻优并设置的.当输入数据模

式不变并且适应度值高于事先设定的阈值时,系统只用 OCSVM 模块快速实施在线检测任务.仅当输入数据模式有变导致适应度值低于阈值时才会启动 CPSO 模块重新进行最佳参数搜索.此时在线检测速度会略有下降. 见表 2-1.

3.基本 CPSO 算法流程:

(1).初始化种群： 给定群体规模 N， 随机产生每个粒子的初始位置和速度 $\boldsymbol{x}_i,\boldsymbol{v}_i$，
计算各粒子的适应值 $f(\boldsymbol{x}_i)$,有 $\boldsymbol{x}_{pi}=\boldsymbol{x}_i$,经比较得出 $\boldsymbol{x}_{gi}$.

(2).将 $\boldsymbol{x}_{gi}$ 的每个分量通过式 (2-29)的变换,映射为混沌变量 $\boldsymbol{t}_j$， 各分量 $\boldsymbol{t}_j\in(0,1)$；

(3).各粒子将通过式(2-25)、式(2-26)，计算速度 $\boldsymbol{v}_i$，并调整至新位置 $\boldsymbol{x}_i$，
进而计算各相应的适应值 $f(\boldsymbol{x}_i)$；

(4).混沌变量的各分量经式(2-30)作混沌运动,并变换为 $\boldsymbol{t}_{j+1}$；

(5)将 $\boldsymbol{t}_{j+1}$ 的每个分量通过 (2-31)式变换，映射为 $\{a_i,b_i\}_{i=1}^n$ 间的普通变量 $\boldsymbol{y}_i$，
并计算 $f(\boldsymbol{y}_i)$；

(6).比较 $f(\boldsymbol{x}_i)$、 $f(\boldsymbol{y}_i)$、 $f(\boldsymbol{x}_{pi}=\boldsymbol{x}_{pBest})$，确定下一叠代步的 $\boldsymbol{x}_{gi}=\boldsymbol{x}_{gBest}$。

(7).判断是否已满足终止条件，若是，则终止运行，输出当前最优解与
最优值；否则，返回(2)继续运行.

2.7 OCSVM-CPSO 组合式自适应故障检测模型

将前面的 OCSVM 模型与 CPSO 算法组合起来.其中 OCSVM 作为基本检测学习器[119],而 CPSO 作为最优参数搜索器,正常情况下直接用 OCSVM 实现在线故障检测,当发现检测适应度值低于事先定的阈值时,启动 CPSO 搜索器自动搜索新的最佳参数,一旦检测适应度值达到正常标准,则退出搜索过程,并启用新的参数来继续进行 OCSVM 在线故障检测.这样就可构成 OCSVM-CPSO 组合式自适应故障诊断模型.其框图如下:

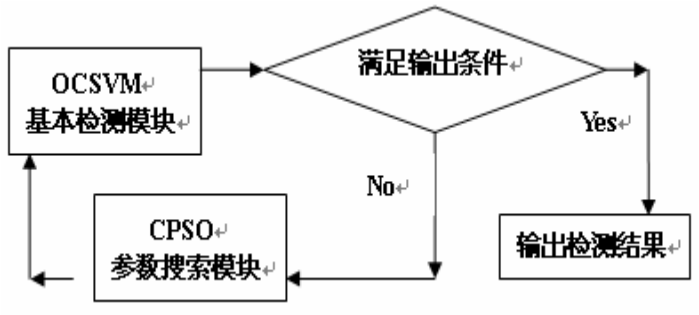


图 2-13 OCSVM-CPSO 组合式自适应故障诊断模型

2.8 实验分析与比较:

本节用本院实验室的移动机器人速度传感器测量数据作故障检测的仿真实验.采用的数据是来自机器人传感器的实际测量数据,不过传感器失效的故障是通过将某时段测量速度人为的设置为零来模拟的, 其它测量速度与逻辑速度不相吻合的无规则曲线也是通过人为制造轮子打滑条件下形成的.目的是想检测本文所论故障诊断系统的灵敏度,测准率,误诊率.须要说明的是,本实验中的逻辑速度只是作为参考对照用,而故障检测实际上只用传感器测量速度一种数据.即只用传感器测量数据本身来建模并检测其中的异常.实验采用OCSVM 与 OCSVM-CPSO 对照方式进行.

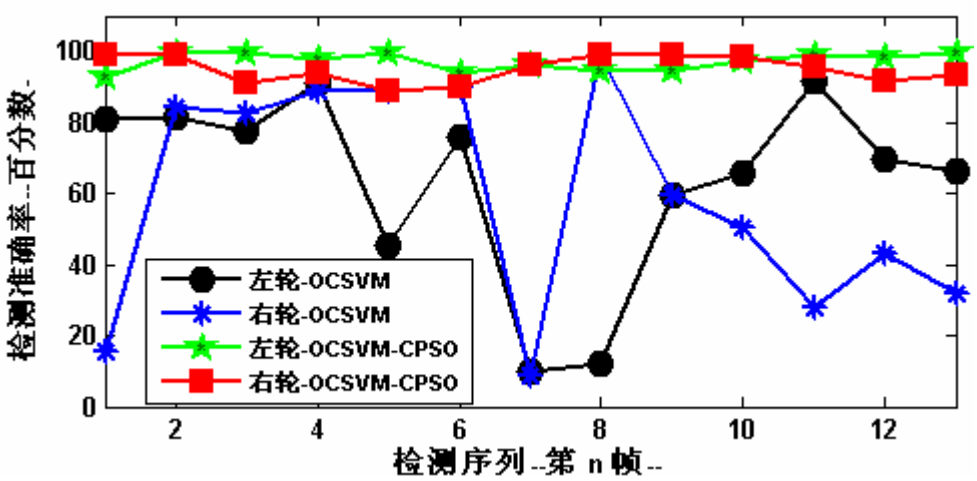


图 2-14 故障检测准确率对比

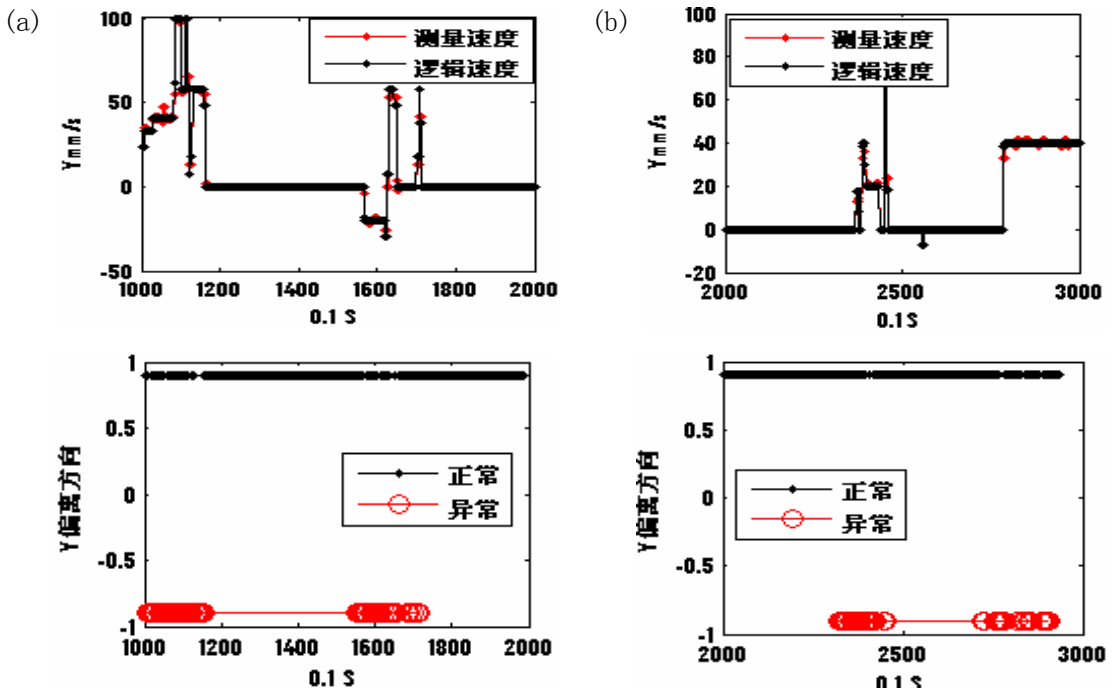
实验结果如图 2-14 所示,它是使用固定学习参数的 OCSVM 与组合自适应故障检测系统OCSVM-CPSO 的对比实验示例,从图中可看出,在运行过程中,随着输入数据的变化,OCSVM 的检测准确率会发生很大变化,例如本图中就有多次检测(对应于图中的兰色和黑色点)的准确率低于 40%,因而 OCSVM 检测系统是不太可靠的. 而从图中可以看出, 采用 OCSVM-CPSO 的系统在运行过程中波动远小于前者,且检测准确率明显高于前者(对应于图中的红色和绿色点).因为它所有点的检测准确率都在 88%以上,因而系统性能远高于前者.

下面的表 2-1 分别列出了 OCSVM 及 OCSVM_CPSO 两种检测系统的运行时间对照表,此表说明,提高检测准确率是以花费更多的计算时间为代价的.尽管 OCSVM-CPSO 组合式自适应检测系统较单纯的 OCSVM 运行稍慢一些,但实验表明这并不足以影响它的实时运行.因而它是行之有效的检测系统.

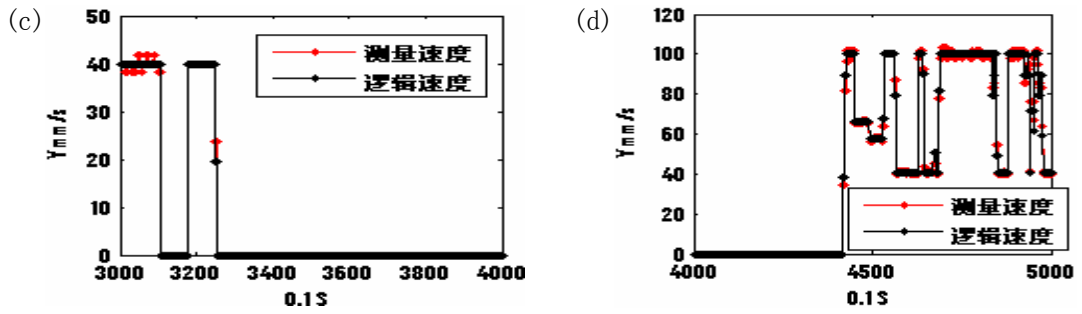
表 2-1 各次检测运行时间对照表. (秒/次)

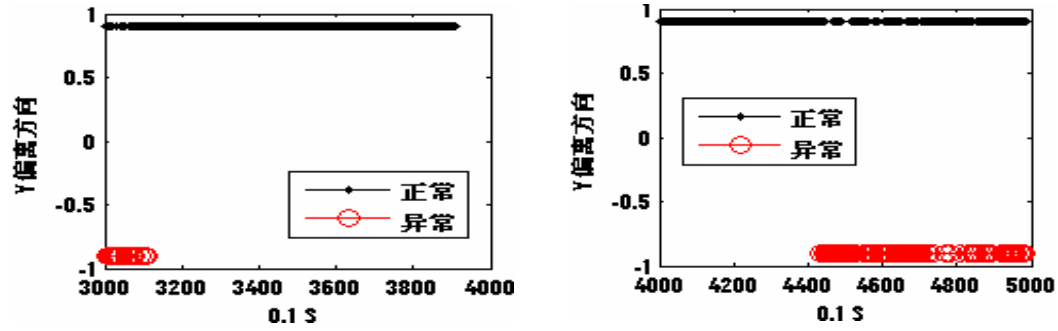
检测序号	1	2	3	4	5	6	7	8	9
样本点数	600	600	600	600	600	600	600	600	600
OCSVM 系统运行时间									
左轮时间	0.036	0.027	0.014	0.016	0.019	0.022	0.02	0.018	0.018
右轮时间	0.084	0.049	0.036	0.04	0.066	0.058	0.05	0.057	0.041
OCSVM-CPSO系统运行时间									
左轮时间	0.068	0.052	0.029	0.041	0.029	0.048	0.05	0.07	0.155
右轮时间	0.582	0.116	0.125	0.362	0.118	0.48	0.29	0.41	0.544

下面图 2-15 是采用机器人速度传感器数据,利用 OCSVM-CPSO 系统故障检测效果的一个实例. 其中图 2-15 (a)显示了实验中左轮运行到第 100-120 秒,及 160-175 秒时段时异常特征检测结果. (b)则是右轮在 100-120 秒时段发现异常的情况.



(a) 左轮 (b) 右轮





(c) 左轮 (d) 右轮

图 2-15 OCSVM-CPSO 组合故障检测的效果示例

图 2-15 中的(a),(b),(c),(d)各图分别对应左右两轮的检测情况,其中,上方的子图显示的是该轮测量/逻辑速度,下方的子图则显示检测结果.由图可见,当传感器测量速度与机器人逻辑速度相似时,检测结果为正常(图中黑线),否则为异常(图中红圈),对照此图.可见被检测到的异常点出现位置与实际情况基本对应,这也证实了本系统故障检测的准确率是相当高的.

2.9 小 结:

本章所述的基于 OCSVM 的特征检测模型,其主要优点是它无需事先建立正常行为模型,也不需要任何标记数据,能够从未标记的数据集中找出隐藏在其中特征异常,并通过理论分析和实验证实它比有模型的解决方案更加容易实现.该模型建立在时序数据表示,相空间重构及 OCSVM 等理论基础之上.它的实现是相关理论与技术综合运用的结果.

建立在 OCSVM 基础上的 OCSVM-CPSO 组合自适应故障检测模型,采用 OCSVM 作为基本检测模块,而用 CPSO 作为最优参数的搜索模块.当数据变化较大导致检测适应度值低于某确定的阈值时,启动 CPSO 搜索新的参数.而适应度值达或超过阈值时退出搜索过程,并用新参数继续后续检测过程.实验表明:此法的确能够有效地实现高准确率在线检测任务.是一种行之有效的组合解决方案. 为智能异常检测或故障诊断研究提供了一种新的思路.

第三章 高效时序特征检测的符号化模型

3.1.相关研究背景:

符号时间序列分析起源于上世纪90年代中期,它是由符号动力学理论[132]、混沌时序分析[133]和信息理论发展起来的一种新的信息分析方法,能够很好的抑制噪声. 大大缩短计算时间,便于实现可视化.它为时序数据特征检测提供了一种简单、快速且有效的处理方式.

符号时序的研究经历了一个逐步深入的过程：1983 年 Crutchfield 等人提出采用由STSA(Symbolic Time Series Analysis)估计的模式来描述动力系统统计复杂性的可能性[132],1991年Rechester等人基于符号动力学设计了一个符号运动方程并应用这一方程快速估计低维映射的不变性密度.1995 年 Kurths 等人首次将符号化方法应用于观测数据[133]. 1997年Tang等人将符号化方法应用于时间序列建模[135],Daw等人成功地将这一方法应用于内燃机实验数据模型拟合[136],R Brown 所做的工作表明符号序列统计量可用于时空系统的重构.同时Daw研究表明符号化方法可用来稳健地估计相关时间,检测隐藏在噪声或混沌信号中的周期信号[137]. 2000 年 Lehrman 等人研究了采用条件符号统计量辨识混沌过程中不同成分问关联性的可能性[138],蒋嵘,李德毅等提出了基于形态表示的时间序列模型[139],2000-2001年,Yi and Faloutsos,Keogh, E与 Geurt等人分别提出了实现时序数据符号化的分段线性近似算法（PAA）[140-141]; 李斌和倪世宏等分别于 2002,2005 年提出非同步多时间序列模式的提取算法[142 -144];2003 年,M.B. Kennel 提出了基于符号假近邻的模型（SFNN[145],它基于在符号空间相互接近的点在相空间也相互接近的假定.但这种分划在相空间维数很大,或在包含噪音时计算量太大. Venkatesh 等人2005年提出的基于小波变换的模型[146]克服了上述缺陷,并有助于平抑高维动态数据的噪音,增加计算效率; 此外还有2005年,任江涛等提出时序快速分段及符号化方法[147]; 2006年,王晓哗等.提出多维时序符号化方法[148],首次采用模糊自适应共振理论(Fuzzy ART)对多维时间序列数据进行聚类,实现了多维时间序列数据的符号化.

在上述各种时序表示方法中,大多数符号表示方法都存在以下致命缺陷:1,时序数据是典型高维数据,时序数据降维是其高度期望的属性.但大多数符号时序无法降维;2,定义的符号时序无法进行距离计算,也就无法进行相似比较,因而缺乏实用价值.而Lin, J. Keogh, E. 在2003年提出的新型的符号化表示方法- SAX（符号聚合近似）算法[149],首次有效地克服了上述致命缺陷,并在许多应用中性能都优于或至少不低于其它的方法,因而受到了广泛关注,成为倍受推宠的时序符号化算法之一[150].它的主要优点是,比其它符号算法更简便,高效;在符号化过程中实现了减维降噪,保证在符号空间计算出的两个符号序列的距离满足实际的两个时间序列的距离的下界要求,即不会出现漏报.SAX 符号化方法的优点使它在图像处理,生物信息,数据挖掘和机器学习等领域得到广泛应用.例如,2006年Li Wei ,Eamonn

Keogh,将其用于异常形状识别,据称能够使计算加快3~4个数量级[151],Xiaopeng Xi 等将其用于时序分类[152]；B. Lkhagva 等人将其用于金融时序的相似搜索[153],Keogh 等用它解决DNA,文本及视频时序的数据挖掘问题,并证明它在异常特征检测,兴趣点察觉,分类和聚类应用中的表现胜过其它方法[154]

3.1.1 SAX 时序符号化模型：

SAX 符号化方法基于 PAA 减维技术,PAA 能够把一个 n 维时序表示成 N 维时序,N<<n . 用下式表示.

$$\overline{C} = \overline{C_1},\ldots\overline{C_N}$$

其中第 i 个元素由下式计算：

$$\overline{C_i} = \frac{N}{n}\sum_{j=\frac{n}{N}(i-1)+1}^{\frac{n}{N}i} C_i \qquad (3\text{-}1)$$

采用 PAA 技术将连续时序数据转换为离散时序的原理如下图 3-1 所示.它把 n 维时序

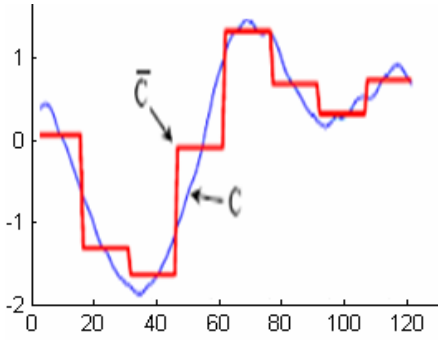


图 3-1,PAA 减维示意图

数据等分为 N 个时序子段,用各子段的平均值来近似表示各子段的原始时序,这个简单而直观的减维技术却令人惊雅的胜过付氏变换和小波变换[155].在对原始数据标准化的前提下,把转化成 PAA 时序,进而用 SAX[149]算法,对已标准化的 PAA 时序（服从高斯分布）按等概率区间分划原则转换成相应的离散 SAX 符号时序.

转换过程首先从确定划分点开始,例如,若欲将 PAA 时序转换成由 k 个字符组合表示的符号序列,则须用 k-1 个划分点将 N (0,1)高斯分布曲线下的区域分划成 k 个面积相等的区域(等概率面积),划分点可以由 SAX 算法计算得出.下表 3-1 是当 k=5 时所得到的划分点查找表.

表 3-1, k=5 时的划分点查找表

a	3	4	5
β*1*	-0.43	-0.67	-0.84
β2	0.43	0	-0.25
β3		0.67	0.25
β*4*			0.84

根据查找表, PAA 时序可按下式实施符号化转换：

$$\hat{C}_i = alpha_j \quad iif \quad \beta_{j-1} \le \overline{C}_i \le \beta_j \tag{3-2}$$

落入不同区间的 PAA 时序数据将被映射成不同的符号. 其中, 小于最低划分点值的 PAA 数据将被映射为 a, 高于最高划分点值的 PAA 数据将被映射为 c, 而大于最小划分点值但小于第二划分点值的 PAA 数据将被映射为 b, 如下图所示：

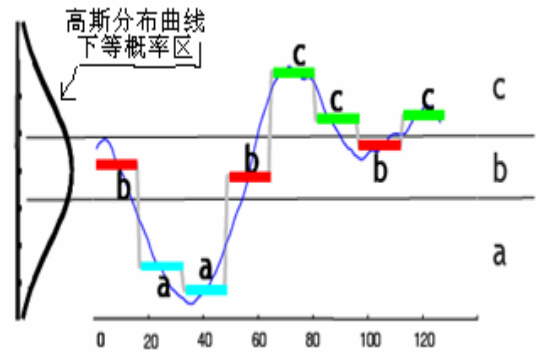


图 3-2 PAA 时序映射成 SAX 符号实例：

此例中, 原始时序长 n=128, 离散成 PAA 时序长 N=8, 字符集 k=3; 映射成 SAX 符号时序串为 b a a b c c b c , 于是, 原始维数为 128 的时序数据, 被压缩为 8 维的离散时序, 从而使其计算代价下降, 大大提高运算效率. 这正是符号化时序的魅力所在.

3.1.2 当前问题及本节目标：

由于 SAX 符号时序模型主要只适合于遵循高斯分布且在有限方差范围内有较高分布密度的时序数据. 并且 SAX 在作时序数据的符号化转换时, 只是利用了各相应时间子段数据的均值信息. 因而不可避免地也存在一些弱点. 例如在符号时序数据模式识别（异常检测或相似查询）中的精度不够理想. 在符号时序分析中所提供的描述信息不够充分. 难以表现时序数据更细微的特征等.

下图 3-3 是用 SAX 将某一时序转换为符号时数据与符号的对应映射图. 此图表明, 尽管位于上下边界区的数据都分别划分为相应的最大, 最小字符表示, 但由于这种划分过分粗糙, 实际上造成隐性的信息丢失, 形成上下边界的信息丢失区. 这将导致一些时序分析任务中的较大失误.

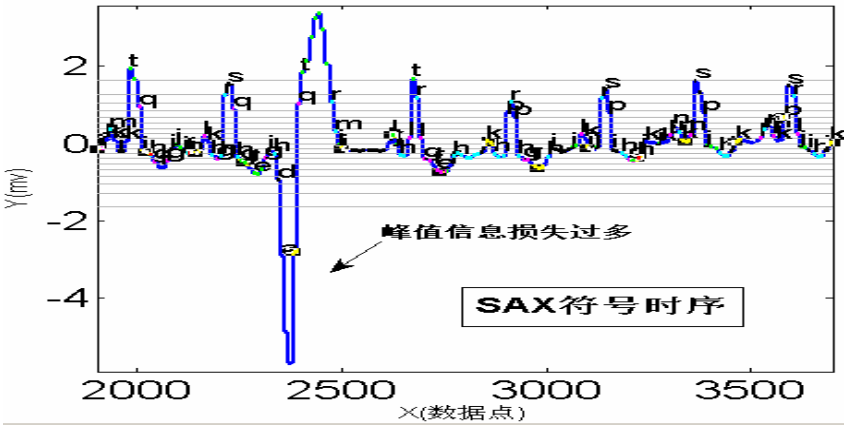


图 3-3　SAX 符号序列

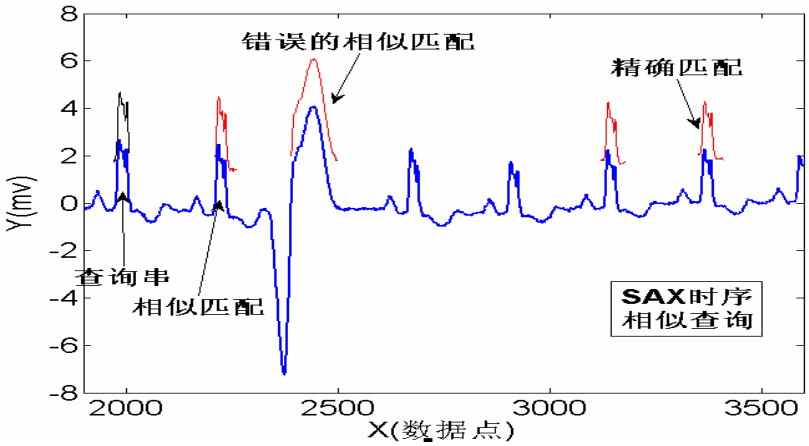


图 3-4.　SAX 符号用于相似查询

图 3-4 显示的是将这一 SAX 符号化结果用于相似查询分析的情况, 如果查询串恰好位于边界区, 则无论其幅度多大都会被转化为最大或最小字符, 尽管图 3-4 中第三个波峰幅度比第一个（查询串）高了几乎一倍, 但由于同属于边界区字符, 它们都会被转化为同样的字符(在计算时被当作相同的波幅对待), 因此, 在相似查询中出现错误匹配也就在所难免了.

由于很多时序分析任务中, 时序极值是重要的关键信息, 它们往往代表着某一趋势的重要转折点. 而 SAX 方法很难避免这种边界区信息丢失问题, 因而不可避免的限制了它在这些任务中的应用 . 为此, 本节提出 DLS (Dynamic and Limited Symbol--动态有界符号化）方

法:它根据时序数据的极值确定划分的上下边界,并根据最大熵确定最佳描述字符集,进而按照上下界范围及最佳字符集数动态地确定划分间隔.从而有效防止边界区的信息隐性丢失.实验表明,此方案能够适应于许多非高斯分布非线性时序的符号化问题,并对极值数据的提供更多描述信息.在一些与极值相关的时序相似搜索,分类,异常检测应用中能够得到比 SAX 更好结果.

3.2. DLS（动态有界符号化）方法

3.2.1 概述

本节要解决的问题是,将一段长度为 n 的任意时序数据 $X = x_1, x_2, ...x_n$ 转化成长度为 N 的,用字符集 $A = \{A_1, A_2, A_3...\} = \{a, b, c, ...\}$ 表示的符号数据 $\hat{X} = \hat{x}_1, \hat{x}_2, ...\hat{x}_N$. 这里, $N = n$.在转化过程中应当尽可能的减少极值信息丢失.为此,本节采用如下方案：首先计算待转化时序数据的最大压缩比 w(又叫最大分段长度,即每个符号所代表的时序数据点数),本节 3.2.3 节将详细说明其具体内容；据此再用分段集成近似（PAA[166]）方法计算各时段数据的近似表示,进而确定最佳字符集的分量数 A.与 SAX 方法不同的是,本节所用的符号化方法不按等概率原则来作为离散化的划分依据,而是根据时序数据的极值（最大值,最小值）及最佳字符集规模来动态决定划分间隔的基准,进而实现整个时序序列的符号化.

3.2.2 时序数据的降维

为了使符号化时序数据能够适用于一般情况,在降维前须将待转化连续时序数据转化为零均值,标准差为 1 的规范化数据.然后用 PAA 方法将标准化时序数据 $X = x_1, x_2, ...x_n$ 按适当的压缩比 w 降维,生成一个 N = n/w（$N = n$） 维空间向量 $\overline{X} = \overline{x_1}, \overline{x_2}, ...\overline{x_n}$.其中, $\overline{X}$ 的第 i 个元素可由下式计算：

$$\overline{X_i} = \frac{N}{n} \sum_{i=\frac{n}{N}(i-1)+1}^{\frac{n}{N}i} X_i \qquad (3\text{-}3)$$

上式即(3-1)式的更一般表述.

时序数据符号化目的是要在误差允许范围内通过时序数据的减维来降低计算代价.更具体地说是通过减维操作将维数较大的连续长时序数据压缩为维数较小的离散短符号序列,同时又保证经压缩后的符号时序保留原始时序的重要特征信息,以便在各种时序分析任务中能够以较小的计算代价达到与原始时序分析尽可能相同的效果.时序数据符号化的关键内容之一是寻找最佳压缩比.

3.2.3 最大压缩比的确定

压缩比相当于一个符号所能代表的时序数据点数,显然,压缩比越高,一个符号所能代表的时序数据点数越多,当然对所代表的那段时序描述也就越粗糙(信息丢失越多),计算代价越小.然而,压缩比越低,一个符号所能代表的时序数据点数越少,对所代表的那段时序描述也就越精细(信息丢失越少),同时也带来了计算代价变大的问题.在极端情况下,压缩比为 1,表示符号数据与原始时序等维,即没有任何压缩,当然也不会有任何信息损失,但这违背了时序符号化的初衷.而另一个极端是整个时序串压缩为一个符号表示,压缩比为原始时序的长度.此时信息几乎全部损失,同样没有意义.对时序数据符号化有意义的压缩比是指在时序分析误差增加不超过允许值前提下的压缩比.在此,本节提出最大压缩比的概念:

定义 1:在连续时序数据转化为离散符号数据过程中的最大压缩比,是指时序分析误差的增加不超过允许值的前提下,转化前连续时序数据的维数与转化后离散符号数据维数之比的各种可能比值中的最大值.

因而如何确定最大压缩比,是影响时序数据符号化的质量和成败的核心问题之一.不幸的是,在目前所报道的有关时序数据符号化的文献中,尚未发现相关的研究和探讨,大多数文献确定压缩比的方法是凭经验或实验地确定[156].

表 3-2,不同数据集的最大压缩比.

序号	数据名	数据集长	变化率	压缩比	数据源
1	chfdb_chf01_275	3751	0.02	20	ECG
2	chfdb_chf13_45590	3750	0.03	18	ECG
3	ECG106_test	1216	1.13	3	ECG
4	ltstdb_20221_43	3750	0.05	14	ECG
5	ltstdb_20321_240	3750	0.03	18	ECG
6	mitdb__100_180	5401	0.02	23	ECG
7	stdb_308_0	5400	0.03	20	ECG
8	xmitdb_x108	5400	0.01	26	ECG
9	synthetic_control	600	4.05	2	UCI
10	synthetic_data1	10000	0.0001	264	UCI
11	synthetic_data3	10000	0.00014	260	UCI
12	Coffee_TRAIN	287	0.38	5	UCR
13	OliveOil_TRAIN	571	0.015	26	UCR
14	JENKINS8	100	0.315	6	TSDL[12]
15	JENKINS9	100	0.31	6	TSDL[12]
备注	q ~5%	$2kq$ ~10	F	$W=\sqrt{2kq/F}$	

是否有规律可循？本文根据大量的仿真实验 ,发现压缩比与时序数据的变化频率呈相反变化关系,同时也与允许误差相关.而确切的关系目前难以分析地描述,在此,提出一个经验公式.

$$w=\sqrt{2k\theta/F} \tag{3-4}$$

$$F = \sum_{i=1}^{n-1} \left| x_{i+1,j} - x_{i,j} \right| / (n-1) \qquad (3\text{-}5)$$

这里, w —最大压缩比,F—时序数据变化的频率. q —允许误差, k---标准常数.n—数据点总数.表 3-2 是设定允许误差为 5%,标准常数 k=100 情况下测定的一些时序数据的最大压缩比.在不超过最大压缩比的情况下,符号化时序的结果基本满足时序分析应用的要求 .这在后面所述的一些实验中得到了验证.

3.2.4 时序数据的符号化

用 PAA 方法按选定的压缩比 w 降维后,就可根据字符集和时序数据的极值得到描述各个字符所代表数据区间的划分点.进而将已经降维的 PAA 数据离散化成符号数据,通常,若已经确定了最佳字符集规模为 k, 则一个有 k 个字符组成的字符集需要有 k-1 个划分点.因此划分点集可表为：

$$C = c_1, c_{2,\ldots} c_{k-1} \qquad (3\text{-}6\ )$$

而 $c_{0\,0}$ 与 c_k 分别定义为最小值,最大值(下界,上界)从第 c_{i-1} 个划分点到第 c_i 划分点的间隔按下式确定：

$$c_i - c_{i-1} = (c_k - c_0)/(k-1) \qquad (3\text{-}7)$$

按此方案,我们可以采用类似于 SAX 的方法来实现从 PAA 时序到符号化时序的转换.

定义: 若用 A_i 来表示字符集 A 中的第 i 个字符,则从 PAA 近似 $\bar{X}$ 到符号近似 $\hat{X}$ 的映射可由下式确定:

$$\hat{X}_i = A_i \quad iif \quad c_{i-1} \le \overline{X_i} < c_i \qquad (3\text{-}8)$$

将所有小于 c_1 的 PAA 时序数据映射为符号 $A_1 = A_{\min}$ 同理,也可将所有大于 c_1 而小于 c_2 区间的 PAA 时序数据映射为符号 A_2 ,而将所有大于 c_{n-1} 区间的 PAA 时序数据映射为符号 $A_n = A_{\max}$.若采用英语字符来表示,则可将一个 PAA 时序序列转化为类似于 $\hat{X} = \hat{x}_1, \hat{x}_2, \ldots \hat{x}_n = a, b, c, \ldots$ 这样的字符串.从而完成时序数据符号化的全部过程.

3.3. DLS 符号时序的距离计算：

由于上述从数据空间向符号空间的映射操作实际上存在着明显的对应关系,因而可以根据这些对应关系实现符号空间的距离计算及数据空间的欧式距离计算.回顾我们的符号化操作过程,第一步是采用 PAA 减维,将原始数据转化成了 PAA 平均值数据,两个原始时序 A,B 的欧氏距离(一维情况下)可表为：

$$D(A,B) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \tag{3-9}$$

相应的 PAA 时序 $\overline{A}$，$\overline{B}$ 的距离则为：

$$DR(\overline{A},\overline{B}) = \sqrt{\frac{n}{N}}\sqrt{\sum_{i=1}^{N}(\overline{a}_i - \overline{b}_i)^2} \tag{3-10}$$

可以证明，两个 PAA 时序 $\overline{A}$，$\overline{B}$ 的距离与原始时序的欧式距离之间存在下面关系[158].

$$D(A,B) \geq DR(\overline{A},\overline{B}) \tag{3-11}$$

当我们进一步将 PAA 数据转化成符号数据后：两个时序符号 A_i，B_i 间的距离可由(3-6)(3-7)(3-8)式得出并由下式计算:

$$dist(A_i,B_j) = \sqrt{(c_i - c_j)^2} = |c_i - c_j| \tag{3-12}$$

即 $dist(A_i,B_j)$ 可由相应符号的距离矩阵所构成查找表得出. 下面是某一特定时序数据在符号化过程中生成的距离矩阵查找表示例:

表 3-3, 某距离矩阵构成的查找表

	a	b	c	d	e
a	0	0	0.97	1.95	2.92
b	0	0	0	0.97	1.95
c	0.97	0	0	0	0.97
d	1.95	0.97	0	0	0
e	2.92	1.95	0.97	0	0

此表列出了某选定字符集中所代表任意两个字符 A_i,B_j 间的距离,其计算表达式为：

$$dist(A_i,B_i) = \begin{cases} 0 & if\ |A_i - B_i| \leq 1 \\ |c_i - c_j| & i,j = 1,2,...s\ otherwise \end{cases} \tag{3-13}$$

这里, c_i,c_j 为与该字符对应的划分点值. 例如, 当字符分别为 A_i = a，B_i = d 时, 由此表, 可查出 $dist(a,d) = dist(d,a) = 1.9455$. 由此可进一步计算它所代表相应的原始时序的最小距离为:

$$MinDist(A_i,B_i) = \sqrt{\frac{n}{N}}\sqrt{\sum_{i=1}^{N}(dist(A_i,B_j))^2} \tag{3-14}$$

同样可以证明 PAA 时序与符号时序距离间存在下面关系[158]：

$$DR(\overline{A},\overline{B}) \geq MinDist(A_i,B_i) \tag{3-15}$$

这与 SAX 方法的距离矩阵显然是等价的,因此,本算法实现了与 SAX 相同的维度压缩和下界距离[158].能够用与 SAX 相同的距离计算来进行符号时序分析比较.

SAX 符号时序与 DLS 时序法实验:

下图是分别用 SAX 方法与 DLS(动态符号)方法对数据进行处理时的效果比较:

作为对照,下面图是SAX及DLS两种符号化时序情况,为便于比较,后面的各图采用相同数据:

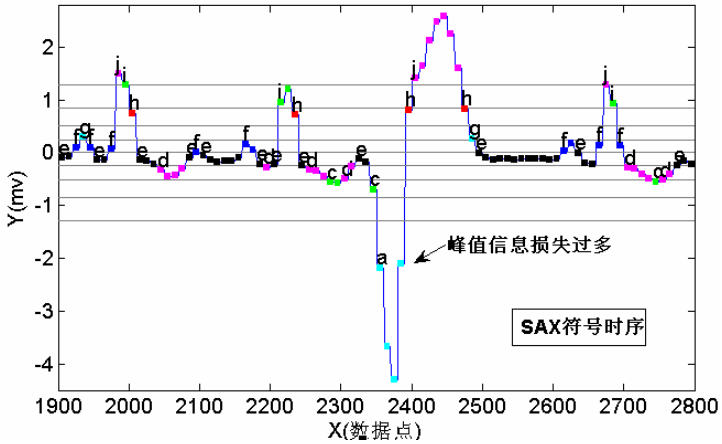


图 3-5.a. SAX 符号序列

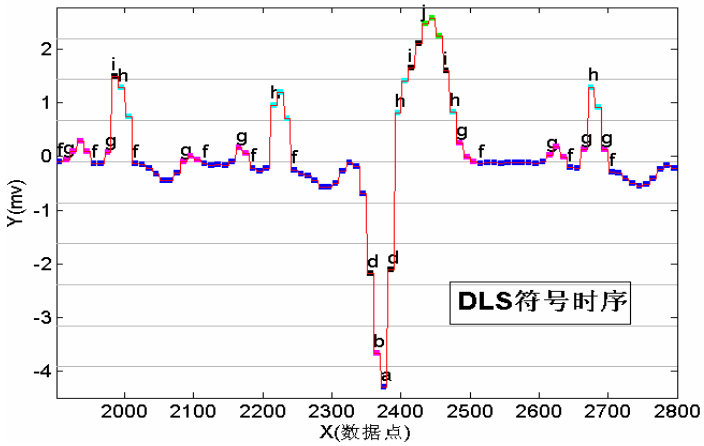


图 3-5. b. DLS 符号序列

需要说明的是,图3-5.a,图3-5.b 采用由6个字符{a,b,c,d,e,f}构成的字符集,目的是便于显示并比较不同算法的符号化结果.(图中的五条横向水印线将纵轴划分成6个区域,落入不同区域中的时序数据被转化为不同的符号,例如,从下往上数,落入第一条水印线以下的数据一概被转化为符号a,而落入第五条水印线以上的一概被转化为符号f,落入第一条水印线以上第二条以下的数据一概被转化为符号b,余此类推)

图 3-6.a,图 3-6.b 及图 3-7.a,图 3-7.b 则采用的是由 20 个字符构成的字符集,目的是保证实验精度.图中的横座标代表时序数据集中按时间排序的相应数据点,单位为 1.从图 5.5 可见,用 DLS 方法与 SAX 方法转换同一时序数据时,其结果是不同的.其应用效果可以从

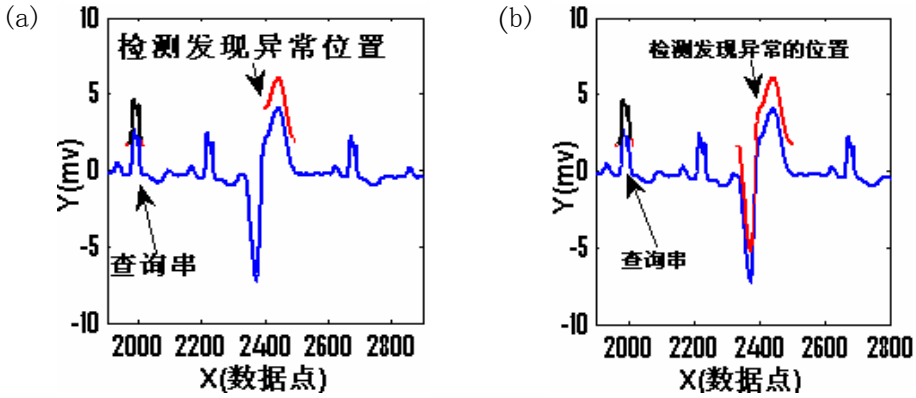
下面的实验分析中比较得出.

3.3.4, 实验分析

为了检验本节算法的有效性. 采用了本章表 3-5 所示的 UCI 和 EGC 共 15 个时序数据集, 分别用本节提出的 DLS 方法与 SAX 方法将各时序数据集转化成相应的符号时序数据集作对比实验. 其大致方法是: 分别将被查询序列子串 T_i 与待查询时序数据集 T 转化成相应的符号时序, 再用 Brute_Force[170]算法进行相应的时序分析实验, 对比实验的结果如下:

(1) 异常检测:

压缩比为 w=10; 字符集都取相同规模时, 选择一段包含有极值字符的正常时序子串 T_i 作为标准串, 用经改进的 Brute_Force 算法在整个时序数据集 T 中 检测异常, 其分析结果表明, 对大多数数据集而言, DLS 检测准确度都好于 SAX. 在有些情况下 DLS 显著好于 SAX.



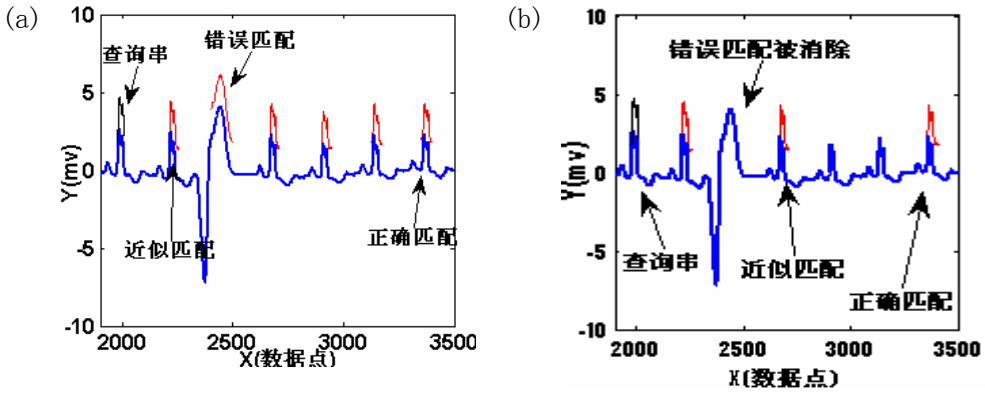
(a). SAX 符号用于异常检测　　(b). DLS 符号用于异常检测

图 3-6. 两种符号用于异常检测效果比较

图 3-6 a, b. 就是采用 ECG 数据源的 chfdb_chf01_275 时序数据集在异常检测时的一个对比实例, 此图说明, 在有些情况下, DLS 可能比 SAX 更准确. 在一般情况下至少不差于 SAX 符号时序. 这恰恰是因为异常检测对时序极值信息更敏感, 而本算法正好加强了极值信息描述的结果.

(2) 相似查询

通过对比相似查询实验, 可以从另一个角度来检验 DLS 符号时序与 SAX 各自的优点. 用 Brute_Force 算法, 在时序数据集 T 中(在一定误差范围内)查找能够与给定长度标准子串 T_i 相匹配的子序列 T_j. 相似查询实验结果表明将会出现两种典型结果:



(a).SAX 符号用于相似查询　　(b).DLS 符号用于相似查询

图 3-7.两种符号用于相似查询效果比较

情况一：

若标准子串 T_i 包含极值的字符串,则在实现符号化过程采用字符集规模相同的情况下,用 SAX 方法将导致相似查询中的误配率高,而用 DLS 符号时序在相似查询误配率较低.

将上图与前面的图 3-3 对比可以发现,由于 DLS 符号化方法防止了峰值信息过多丢失,因此在边界区(极值区)的表现好于 SAX.这在不同数据集的类似实验中都得到了证实.

表 3-4, 相似查询平均匹配数

符号化方法	相似查询平均匹配数	
	极值查询	非极值查询
SAX	5.81	3.21
DLS	2.53	4.02

表3-5, EXP_SAX与SAX运行时间比较(单位-秒)

	数据名	数据集长	算法运行时间		数据源
			EXP_SAX	SAX	
1	chfdb_chf01_275	3751	0.0305	0.009	ECG
2	chfdb_chf13_45590	3750	0.0337	0.0094	ECG
3	ECG106_test	1216	0.0307	0.01	ECG
4	JENKINS8	100	0.0025	0.0008	TSDL
5	stdb_308_0	5400	0.1527	0.0482	ECG
6	xmitdb_x108	5400	0.1761	0.0476	ECG
7	beefconvert	9098	0.2554	0.0811	UCR
8	Beeftrain	30	0.00028	9.00E-05	UCR
9	synthetic_control	600	0.0165	0.0049	UCI
10	synthetic_data	10000	0.2503	0.0894	UCI

情况二：

若标准子串 T_i 不包含极值字符时,在字符集相同(例如都用 8 个字符)情况下,SAX 方法

的匹配精度一般高于动态符号时序方法,而动态符号时序要达到与 SAX 方法同样的匹配精度则需要用更多的字符.这意味着将要付出更多的计算代价.表3-4是15个不同数据集在相似查询中 20 次实验的平均匹配结果.

表 3-4 中：在极值查询时,SAX 符号时序的平均匹配率远高于 DLS 符号时序,这说明此时它的匹配精度差于后者.因为这中间包含了如图 3-4 中所示的完全错误匹配和一些近似的匹配.而在非极值查询时,情况相反,此时DLS包含的近似匹配多于SAX,即其匹配精度差于SAX.因而两种符号化方法各有不同的应用范围.从已知的实验来看,DLS 似乎更适合于涉及极值(或转折点)的相关时序分析任务.

3.4. VSB 矢量化符号方法

3.4.1 算法描述

前面所述的 DLS 算法,仍有一些值得改进的地方.例如,由于注重边界极值(振幅较大的峰值)描述,而相对忽视了振幅较小的极值.然而有些时序分析任务需要关注除边界极值外的更多特征,即要求尽可能全面地时序信息描述.这需要提出新的符号时序模型.为此,本节提出 VSB 矢量化符号方法.通过引入二个极值信息,将原来的 SAX 符号转化成具有三个分量的符号矢量.其符号值由各分量的加权和最终确定.

由于VSB方案能够比SAX提供更多的描述信息,因而能够适应于许多非高斯分布非线性时序的符号化问题.在时序相似搜索、分类、异常检测应用中能够得到比 SAX 更好结果.

针对 SAX 符号时序算法容易丢失极值信息的缺陷,Lkhagva, B 等人于 2006 年曾经提出一种"改进方案",后面简称为 EXT—SAX[150].其基本思路是:分别引入各字符对应时序子段的极大极小值作为新的字符表示,从而将原来的每时序子段一个 SAX 字符改为三个字符,本节研究并验证过这种方案.发现这种方案效果有限.原因很简单,因为 SAX 在符号化过程中,若压缩比为 w,则通常是将 w 个时序数据点用一个代表其平均值的符号来表示.其中的极大值,极小值只是这 w 个点中的单个特殊点,只占其中的 1/w .若按"改进方案",将极大值,极小值单独作为字符,则原来的每时序子段将出现三个字符,即每个字符将代表 w/3 个时序数据点, 这意味着将极大值,极小值的代表性人为地扩大了 w/3 倍.这不但会造成人为的信息失真,而且计算代价也增加为 3 倍（有关 EXT—SAX 与 SAX 计算代价对比实验结果详见表 3-5）.因此本节通过具体地实验验证和理论分析后,认为文献[157]采用增加符号来描述极值信息的方案有待商榷.但它利用极值信息来改善 SAX 信息丢失问题的思路却给我们带来了灵感. 如果我们将各时序子段的极大极小值作为该字符对应的加权分量,则可在不增加字符数的基础上解决这个问题.其基本思路是:在利用 SAX 方法做时序数据的符号化转换时,除了用每个时序子段的均值作为代表该段时序的符号主分量外,还另外增加极大值,极

小值两个辅助加权分量. 这样, 每个时序子段转化成一个有三个分量的符号矢量. 其中, 各时序子段的均值作为刻划总体特征的主分量, 而极大值, 极小值则作为刻划细节特征的辅助加权分量. 二者从不同视角（维度）来描述该时序子段的特征, 因而能够在不增加符号数的前提下比 SAX 方法提供更为全面的描述信息. 相应的数学描述为：

$$x_{ik} = A_{j1} \quad iif \quad c_{j-1} \le \overline{x_i} < c_j \quad k = 1 \tag{3-16}$$

$$x_{ik} = A_{j2} \quad iif \quad c_{j-1} \le x_{i\min} < c_j \quad k = 2 \tag{3-17}$$

$$x_{ik} = A_{j3} \quad iif \quad c_{j-1} \le x_{i\max} < c_j \quad k = 3 \tag{3-18}$$

按各数据在相应的时序子段所占比例作为权重综合起来就得到该时序子段的符号矢量:

$$x_i = [((w-2)/w) \times x_{i1}, (1/w) \times x_{i2}, (1/w) \times x_{i3}]^T \tag{3-19}$$

于是, x_i 就是代表第 i 个时序子段的符号矢量, 其第一符号分量是主分量, 其权值取(w-2)/w, 而其余两个符号分量为辅助分量, 其权值各取 1/w . 由于只是增加了两个加权分量, 其它部分并未实质性改变, 因而可以进行与 SAX 方法[156]相同的距离计算:例如, 当两个符号时序矢量分别为 x_i, x_j 时, 它们代表的相应原始时序的最小距离可按(3-14)式计算其中:

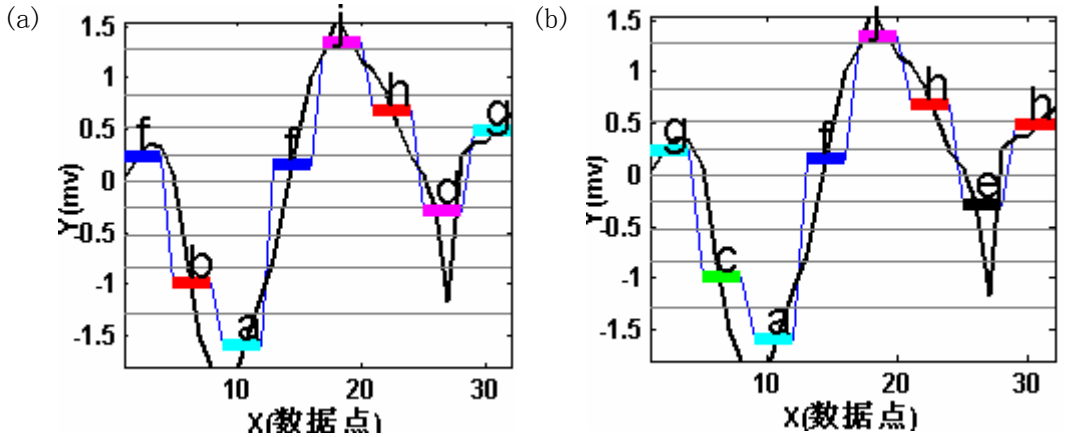
$$dist(x_i, x_j) = \sqrt{(((w-2)/w))^2 d_1^2 + ((1/w))^2 (d_2^2 + d_3^2)}$$

$$d_1^2 = (x_{i1} - x_{j1})^2; d_2^2 = (x_{i2} - x_{j2})^2; d_3^2 = (x_{i3} - x_{j3})^2 \tag{3-20}$$

这里, 根式中的后项

$$(1/w)^2 ? (d_2^2 \quad d_3^2)$$

即代表两个不同符号时序的极小值, 极大值间的距离, 当根式中的前项(即主分量)为零时, 若后项不为零, 则两个符号时序间的距离仍可不为零. 这意味着当我们比较两个时序符号串的相似性时, 不但要检验其相应的主分量(平均值)是否相似, 而且要检验其相应的辅助分量(极大值, 极小值)是否相似, 只有三个量同时相似时, 才能得出两个时序符号串相似的结论. 与 SAX 方法相比, VSB 方法显然提供了对时序数据更多的细节描述信息, 因而能够得到更精确的结果. 当然计算代价会有所增加.



(a). SAX 将 32 个数据点转化成 8 个符号　　(b). VSB 将 32 个数据点转化成 8 个符号

图 3-9 (a), (b) SAX 与 VSB 符号时序比较

上图是分别用 SAX 方法与 VSB 符号方法对数据进行预处理时的比较：图中可见，对同一时序数据的符号化结果是不同的. 其应用效果差异可以从下面的实验分析得出.

3.4.2 实验分析

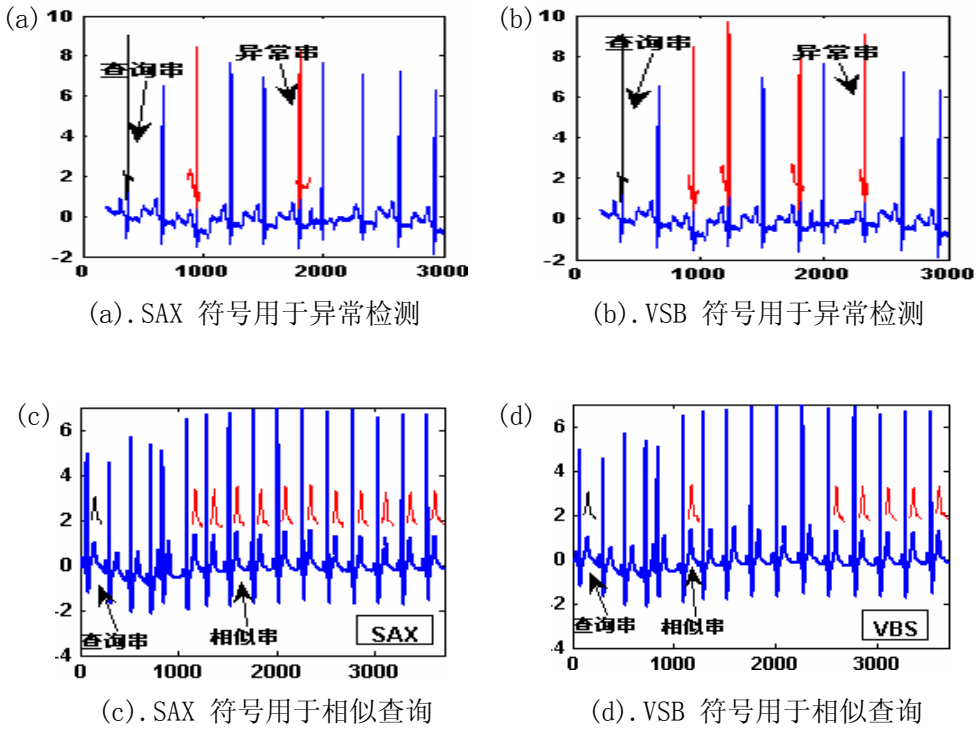
实验采用本节表 3-2 中的 UCI 和 EGC 等各种时序数据集，分别用本节提出的 VSB 方法与 SAX 方法将各时序数据集转化成相应的符号时序数据集作对比实验. 其大致方法是：分别将被查询序列子串 T_i 与待查询时序数据集 T 转化成相应的符号时序，再用 Brute_Force[157]算法进行相应的时序分析实验，对比实验的结果如下：

(1) 异常检测：

压缩比为 $w=10$；字符集都取相同规模时，选择一段包含有极值字符的正常时序子串 T_i 作为标准串，用经改进的 Brute_Force 算法在整个时序数据集 T 中检测异常，其分析结果表明，对表中大多数数据集而言，VSB 方法都好于或至少不差于 SAX 方法.

(2) 相似查询

通过对比相似查询实验，可以从另一个角度来检验 VSB 符号时序与 SAX 各自的特点. 用 Brute_Force 算法，在时序数据集 T 中(在一定误差范围内)查找能够与给定长度标准子串 T_i 相匹配的子序列 T_j. 对表 1 中 15 个不同数据集的相似查询实验结果表明：VSB 符号时序用于相似查询时的误配率低于 SAX 方法. 下面图 3-10 所示的就是一个典型实例：



(a).SAX 符号用于异常检测　　(b).VSB 符号用于异常检测

(c).SAX 符号用于相似查询　　(d).VSB 符号用于相似查询

图 3-10, 两种符号应用效果比较

图 3-10, (a), (c) 记录的是 SAX 符号应用效果片断, 其中, 最左边曲线是查询串, 右边则记录了其在应用中出错的真实情况. 图 3-10, (b), (d) 记录的是用 VSB 符号对同一数据集应用的情况, 图中显示 VBS 在异常检测时检出率较 SAX 要高, 而在相似查询中误配率较 SAX 要低. 下面表 2 是表 1 所列数据集在相似查询 20 次实验中的平均匹配结果.

表 3-6, 相似查询平均匹配数

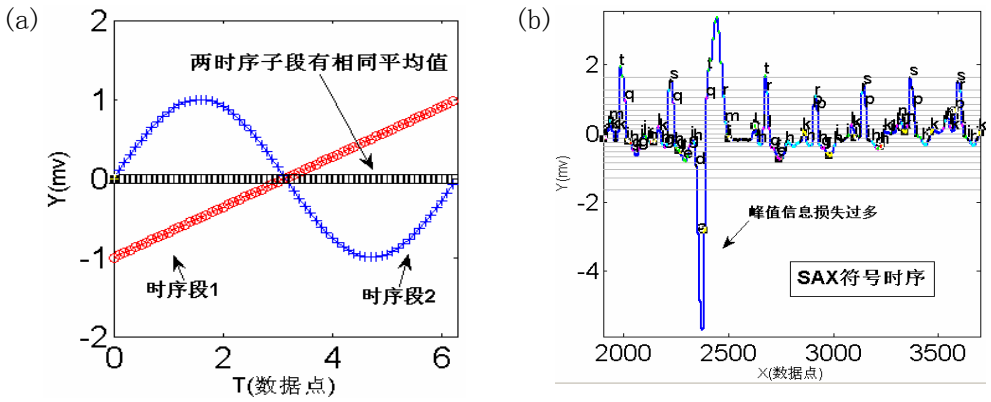
符号化方法	相似查询平均匹配数	
	极值查询	非极值查询
SAX	5.81	3.21
VSB	2.97	2.16

表中：在极值查询时, SAX 符号时序的平均匹配率高于 VSB 符号时序, 因为这中间包含了如图 3-4, 图 3-10 中所示的完全错误匹配和一些近似的匹配. 这说明它的匹配精度差于后者. 因此 VSB 确实是一种比 SAX 精度更高的符号化方法.

3.5 基于统计特征的符号化时序方法

进一步的研究表明,尽管上述二种(DLS,VSB)方法从不同的角度对 SAX 方法有较大的改进.但时序数据在符号化过程中还有一些信息未能充分利用.更好的改进方案是充分利用反映时序数据本身主要统计特征的信息来实现符号化.这就是本节提出的基于统计特征的符号化时序方法.

让我们从一个具体实例开始：下面图 3-11.a 中时序子段转化成一个符号时,尽管时序段 1 及时序段 2 数据均方差完全不相同,但由于它们的平均值相同.SAX 算法会把它们转化成相同的符号.而在做时序相似分析时,这两个在图中看来完全不同的时序段将会被判定为相同.显然,这是一个明显的缺陷.这种算法的直接后果是符号化过程中丢失了重要的数据方差（均方差）信息,包括一些其它重要信息.因而不适合于那些需要更精确分析的场合.



(a).　时间子序列　　(b). SAX 符号序列

图 3-11 SAX 符号化的缺陷

SAX 算法只能近似描述时序数据的大致特征的缺陷限制了它的应用范围.在时序数据分析中,数据的均方差信息对分析结果至关重要,在很多情况下甚至会对时序分析的结果产生决定性影响.例如,对前面所说的图 3-11.a,只需进一步比较二者的均方差,立即就会推翻前面的错误结论.基于这个认识.本节提出的 SFVS 基于（统计特征矢量符号化）算法作为时序数据符号化算法探新的一种尝试,其基本思路是:利用描述时序数据统计特征的均值与均方差分别作为描述其”平均值”特征及”发散程度”特征的分量,而时序符号则看作是由这些分量构成的特征矢量.这样,在实施时序数据符号化过程中,每个时序子段将转化成一个有二个分量的符号矢量.其中,各时序子段的均值作为刻划其“平均值”特征的分量,而相应时序子段的均方差则作为刻划发散程度特征的另一个分量.一个时序符号的总体特征则是该二分量的矢量和.由于二者从不同视角（维度）来描述该时序子段的特征,因而能够在不增加符号数的前提下比 SAX 方法提供更为全面的描述信息.这有利于在时序模式识别的应用中实现更精确的分析.

3.5.1. SFVS（统计特征矢量符号化）算法

(1)　算法描述

考虑到时序数据大多可以看作是一种随机系列,时序分析实质上是对所观测到的时序系列进行统计分析,而时序数据的最基本统计特征可以由其均值与方差来确定.因此当人们将时序数据转化为符号表示时,自然应当充分利用这些统计特征信息.本节提出的 SFVS 算法的基本要点是:利用描述时序数据统计特征的均值与方差(或均方差—注:本节采用方差)分别作为描述其平均值及发散程度的分量,而时序符号则是由这些分量构成的矢量.这样,在实施时序数据符号化过程中,每个时序子段将转化成一个有二个分量的符号矢量.其中,各时序子段的均值作为刻划其平均值特征的分量,而相应时序子段的方差则作为刻划数据对平均值 $\bar{X}$ 的平均偏离程度(显然也包括了极值的影响)特征的分量.而一个时序符号的总体特征则是该二分量的矢量和.由于二者从不同视角(维度)来描述该时序子段的特征,因而能够在不增加符号数的前提下比 SAX 方法提供更为全面的描述信息.相应的数学描述为:

$$\overline{x_i}=\frac{1}{w}\sum_{k=1}^{w}x_{ik}\quad,\quad s_i=\frac{1}{w}\sum_{k=1}^{w}(x_{ik}-\overline{x_i})^2$$

$$x_{ik}=A_{i1}\qquad iif\quad c_{j-1}\le\overline{x_i}<c_j\quad k=1 \tag{3-21}$$

$$x_{ik}=A_{i2}\qquad iif\quad c_{j-1}\le s_i<c_j\quad k=2 \tag{3-22}$$

其中,c_{i-1},c_i 分别代表第 i 个字符划分区的上限与下限.

于是可按(3-21)(3-22)两式将落入不同的划分区间的 $\overline{x_i}$,s_i 值转化为相应的符号分量.这里,$\overline{x_i}$ 为第 i 个时序子段的均值,s_i 为相应时序子段方差 s:

$$x_i=x_{i1}\cdot i+x_{i2}\cdot j=A_{i1}\cdot i+A_{i2}\cdot j \tag{3-23}$$

式中,x_i 就是代表第 i 个时序子段的符号矢量.可以按下面的(3-24)式进行类似的距离计算(证明见下节): 当两个时序符号矢量分别为 x_i,x_j 时,它们间的矢量距离可由下式表示:

$$dist(x_i,x_j)=\sqrt{\sum_{d=1}^{2}(x_{id}-x_{jd})^2}=\sqrt{(A_{i1}-A_{j1})^2+(A_{i2}-A_{j2})^2} \tag{3-24}$$

其中, 根式中的前一项代表两个不同符号时序的平均值间的差异,而根式中的第二项则代表了两个不同时序方差间的差异.如果我们把二个不同符号时序间的距离定义为两个符号时序间的(统计)特征距离.则按(3-23)式,仅当两个符号时序的特征距离为零(即它们全部的对应分量相等)时,两符号时序才能相等.同样地.仅当两符号时序的特征距离小于阈值

时,它们才会相似.这只有当其对应分量之差异分别小于阈值时才有可能.与 SAX 方法相比,SFVS 算法显然提供了对时序数据更多的细节描述信息,因而能够得到更精确的结果.

与 SAX 符号不同的是,这里的 $dist(\vec{x}_i,\vec{x}_j)$ 是由相应符号距离矩阵构成的二个不同查找表得出,但查找表的构成方法却与 SAX 相同.除了维数不同之外,一些相关计算与 SAX 相同.

在计算代价方面, SFVS 符号化算法虽然比 SAX 算法上多了一个分量,但是计算代价增加并不大,这可以在后面的实验中得到证实(见表 3-7,表 3-9).

(2) SFVS 算法实现要点:

1, 用 PAA 方式按(3-3) 式对时序数据降维;

2, 计算各时序子段的数值平均及方差两个数值分量;

3, 按(3-21),(3-22)式分别将两数值分量转化为符号分量;

4, 按(3-23) 式,二者的合成可构成符号矢量;

5, 重复以上步骤直到完成全部数据的符号化

表 3-7 符号化过程时间复杂度比较实验

数据集名	数据集长	总运行时间(秒)		预处理部分时间(秒)	
序号	详见附录	SFVS	SAX	SFVS	SAX
1	3751	0.0034	0.0032	1.90E-04	3.90E-05
3	1216	0.00131	0.0013	1.20E-04	3.10E-05
5	5400	0.0036	0.0034	2.10E-04	4.10E-05
10	10000	0.0049	0.0046	2.20E-04	4.50E-05

3.5.2 SFVS 与 SAX 的比较.

在计算时间代价方面,与SAX方法类似,SFVS方法的计算代价主要包括符号化过程的代价和符号距离计算的代价.符号化过程的代价与符号化过程的所有环节相关,由于符号化过程包括数据预处理,数据划分((3-21),(3-22)式),数据符号转换等众多环节,很难得出一个精确的分析结论,但可以通过对表3-7中典型时序数据的符号化过程20次实验平均结果对比来得出大致结果.表3-7的第5,6两列记录的是SFVS及SAX两种方法在符号化过程中的数据预处理环节的时间代价比较,显然,此环节中的SFVS方法的计算代价较SAX方法大约几倍,然而,由于预处理环节的代价在整个符号化过程中所占比例微不足道,因此,两种方法在符号化过程中的时间复杂度相差不大,具体差异请参见表3-7中的第3,4列.

下面来比较两种方法应用最多的符号距离计算代价. SFVS与SAX的符号距离计算可分别用(3-24)式和(3-24)式根式中除去第二项构成的式子表示,两种符号距离计算式之比可表为:

$$MinDist(\hat{X}_i,\hat{X}_j)/MinDist(A_{i1},A_{j1})$$
$$=\sqrt{\frac{n}{N}}\sqrt{\sum_{i,j=1}^{N}((A_{i1}-A_{j1})^2+(A_{i2}-A_{j2})^2)}\Big/\sqrt{\frac{n}{N}}\sqrt{\sum_{i,j=1}^{N}(A_{i1}-A_{j1})^2}$$

显然,除了作为分子的根式内部比分母根式多出一项外,其余的计算形式是相同的,这这意味着SFVS与SAX符号距离计算的时间复杂度是同一数量级的,只有渐进常数的差异.考虑到分子根式中第二项与第一项的计算代价相同,因此,两种符号距离计算的时间复杂度之比大致为：$1<T_{SFVS}(n)/T_{SAX}(n)\le\sqrt{2}=1.414$.即,与SAX相比,SFVS符号距离计算增加的计算代价不大于50%.后面的实验对此作了具体的比较(详见表3-9).

3.5.3 算法相关的符号距离计算

与DLS类似,同样可以根据SFVS各分量对应关系实现符号空间的距离计算及数据空间的欧式距离计算.回顾我们的符号化操作过程,第一步是采用 PAA 减维,进而求方差,将原始数据转化成了预处理数据,考虑到 SFVS 是二维符号,两个原始时序 A,B 的欧氏距离可相应地表为:

$$D(\boldsymbol{A},\boldsymbol{B})=\sqrt{\sum_{i=1}^{n}\sum\nolimits_d(\boldsymbol{A}_{id}-\boldsymbol{B}_{id})^2} \qquad (3\text{-}25)$$

式中的 d 代表数据的维数,相应的预处理时序A',B'的距离则为:

$$DR(\boldsymbol{A}',\boldsymbol{B}')=\sqrt{\frac{n}{N}}\sqrt{\sum_{i=1}^{N}\sum\nolimits_d(\boldsymbol{A}'_{id}-\boldsymbol{B}'_{id})^2} \qquad (3\text{-}26)$$

可以证明, 两个预处理时序A',B'距离与原始时序的欧式距离之间同样存在下面关系[158].

$$D(\boldsymbol{A},\boldsymbol{B})\ge DR(\boldsymbol{A}',\boldsymbol{B}') \qquad (3\text{-}27)$$

与前面类似,当我们进一步将预处理数据转化成符号数据后:两个时序符号$\hat{A}$,$\hat{B}$间的距离可由下式计算:

$$dist(\hat{A}_i,\hat{B}_j)=\sqrt{\sum\nolimits_d(\hat{A}_{id}-\hat{B}_{jd})^2} \qquad (3\text{-}28)$$

若用实际的分量表述上式则有:

$$\begin{aligned}dist(\hat{x}_i,\hat{x}_j)&=\sqrt{\sum_{d=1}^{2}(\hat{x}_{id}-\hat{x}_{jd})^2}\\&=\sqrt{(A_{i1}-A_{j1})^2+(A_{i2}-A_{j2})^2}\end{aligned} \qquad (3\text{-}29)$$

由此可进一步计算它所代表相应的原始时序的最小距离为:

$$MinDist(\hat{X}_i,\hat{X}_j)=\sqrt{\frac{n}{N}}\sqrt{\sum_{i,j=1}^{N}(dist(\hat{x}_i,\hat{x}_j))^2} \qquad (3\text{-}30)$$

同样可以证明预处理时序与符号时序距离间存在下面关系 [158]:

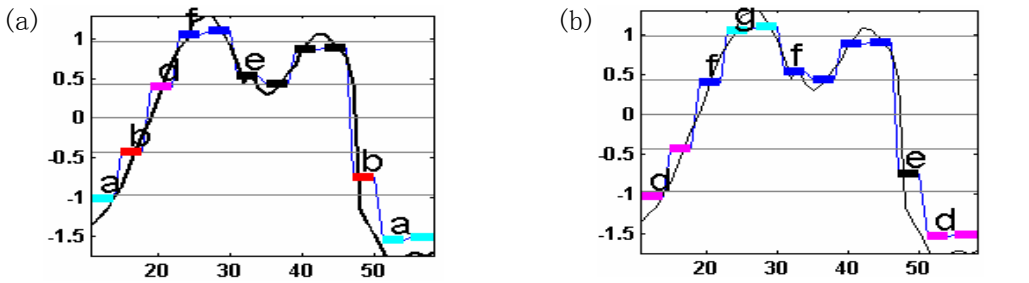
$$DR(\boldsymbol{A}',\boldsymbol{B}') \geq MinDist(\boldsymbol{X}_i,\boldsymbol{X}_j) \tag{3-31}$$

综合考虑(3-29)及(3-33)两式可得：

$$D(\boldsymbol{A},\boldsymbol{B}) \geq DR(\boldsymbol{A}',\boldsymbol{B}') \geq MinDist(\boldsymbol{X}_i,\boldsymbol{X}_j) \tag{3-32}$$

即按 SFVS 符号算得的符号距离同样是其时序数据相应欧式距离的下限.

3.5.4 基于统计特征的符号化方法示例



(a) SAX 符号化算法的前十位字符　　(b) SFVS 符号化算法的前十位字符

图 3-12 同一数据的两种算法的符号化结果对比

表 3-8 同一数据的 SAX 符号与其它符化算法前十位字符

序号	1	2	3	4	5	6	7	8	9	10
SAX	a	b	d	f	f	e	e	e	e	b
SFVS	d	d	f	g	g	f	f	f	f	e

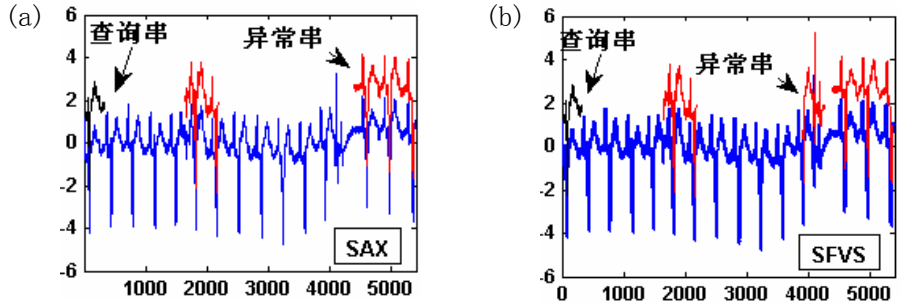
上面的图表反映了不同符号化方法的差异情况,其应用效果请见下面实验分析.

3.5.5　实验分析

为了检验本节算法的有效性.下面仍采用表 3-5 中所示各种不同来源及不同长度的时序数据集,分别用本节提出的 SFVS 算法与 SAX 方法将各时序数据集转化成相应的符号时序数据集作对比实验.其方法是:分别将被查询序列子串 T_i 与待查询时序数据集 T 转化成相应的符号时序,再用改进的 Brute_Force 算法进行相应的时序分析实验,对比实验的结果如下:

(1) 异常检测:

字符集都取相同规模时,随机选择一段时序子串 T_i 作为标准串,并设定相应的阈值,用经改进的 Brute_Force 算法在整个时序数据集 T 中搜索比较,通过寻找超出阈值的时序子串来检测异常,图 3-13 a, b 分别记录了 SAX 及 SFVS 符号时序用于同一数据集作异常检测的真实情况片断,图中较粗的黑色曲线片断是用作检测的标准串,而较细的红色曲线片断是检测出来的异常串,从图中可以直观地看到,用 SAX 符号检测出的异常串为 3 个,而用 SFVS 符号时,检测出的异常串为 4 个,显然 SFVS 的漏检率要低于 SAX. 为了全面比较两种符号时序的优劣,本节还用附录表 2 中的数据集作了对比实验,其具体结果详见后面表 1.

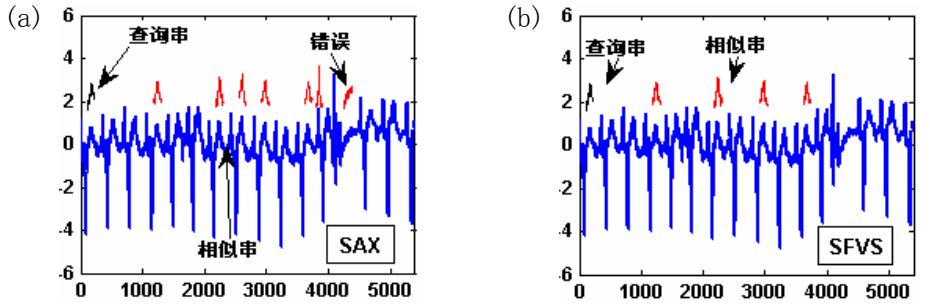


(a)SAX 符号用于异常检测　　(b)SFVS 符号用于异常检测

图 3-13 用于异常检测的两种符号时序效果比较

(2) 相似查询

通过对比相似查询实验,可以从另一个角度来检验 SFVS 符号时序与 SAX 各自的特点. 用 Brute_Force 算法,在时序数据集 T 中(在一定误差范围—阈值内)查找能够与给定长度标准子串 Ti 相匹配的子序列 Tj . 下面图 3-14 所示的就是一个典型实例:



(a), SAX 符号用于相似查询　　(b),SFVS 符号用于相似查询

图 3-14 用于相似查询的两种符号时序效果比较

图 3-14,a 记录的是用 SAX 时序作相似查询的结果片断,其中,最左边曲线是查询串,右边则记录了其在相似查询中出错的真实情况. 图 3-14,b 记录的是用 SFVS 对同一数据集作相似查询时的情况,图中显示它排除了类似 SAX 那样的错误.

下面表3-9记录的是表3-5所列数据集在作异常检测及相似查询20次实验中的平均结果.为简便和便于比较,表3-9中第3,4,5,6列记录了在实验中用选定查询串测出的异常检测漏检率与相似查询误检率而第7,8列记录了两种不同算法的运行时间.

表3-9, SFVS与SAX应用比较

数据名	数据集	异常检测漏检率 %		相似查询误检率%		算法运行时间 s	
详见附录	长度	SFVS	SAX	SFVS	SAX	SFVS	SAX
1	3751	18.7	36.7	5.9	23.5	0.0103	0.009
2	3750	16.7	38.1	20	44.2	0.0106	0.0094
3	1216	22.2	41.7	5	35	0.011	0.01
4	100	28.6	47.4	25	35.1	0.0008	0.00075
5	5400	10	37.5	5.6	16.7	0.054	0.0482
6	5400	14	31	23.3	43.7	0.0573	0.0476
7	9098	7.6	42	20	41.6	0.0931	0.0811
8	30	16	48.3	33	33	0.00013	0.00009
9	600	14.2	27.1	16.7	33.3	0.0055	0.00493
10	10000	7.14	57.12	21.1	58.9	0.0952	0.0894

表3-9中的每一行代表在相同条件下用两种算法对同一数据集分别进行异常检测和相似查询的比较结果,而不同行代表不同数据集的实验比较结果.实验表明,无论是用于异常检测还是相似查询.SFVS算法都要好于或至少不差于SAX算法.从表3-9中第3,4列的异常检测对比结果可见:对所列的10个数据集而言,SFVS检测出异常的漏检率要低于SAX.其原因是,当用SAX符号来检测异常时,实际上是检测两个符号所代表的时序子段平均值之间的距离,当此距离小于阈值时,被检测时序子段将被判定为正常;然而,采用SFVS符号检测时,即使两个符号的均值分量间距离小于阈值,两个符号的矢量距离仍可能大于阈值(例如当两个符号的方差分量相差很大时)而被检测为异常.表中第5,6列的相似查询对比结果则说明,SFVS的误检率要低于或至少不高于SAX,其原因是SFVS的相似匹配需要同时满足两个分量分别匹配的条件,而SAX只需满足一个均值匹配条件即可,因此. SAX更容易产生较多的错误或不精确匹配.另外,表3-9中第7,8列记录的两种不同算法在实验中平均运行时间的数据说明: SFSV增加的计算代价远不及SAX的 50%, 因此,综合算法的精度和代价这二个因素考虑,应当可以说,对比实验确实证实了SFVS是一种比SAX更为出色的符号化方法.

3.5.6 小结

针对 SAX 符号时序方法在边界区信息丢失较多所带来的缺陷,本章首先提出了 DLS 符号时序方法,它根据数据集的极值来动态确定最佳字符集及时序数据的划分间隔,通过估算最大压缩比来指导最好的降维比例.并采用 PAA 作降维预处理,从而实现了与 SAX 同样的符号化时序转换和相同的距离计算方式.但与 SAX 不同的是,DLS 符号时序防止了位于划分边

界区的峰值信息的丢失,因而在一些对峰值信息敏感的时序分析中比 SAX 有更好的表现.当然,为此付出的代价是,在相同字符集情况下,它在非边界区内的信息描述不及 SAX 那么精细,为达到同样精度则需增加它的字符集,就将增加计算代价.因此,下一步的工作是要探索既能少丢失边界区峰值信息,又不影响非边界区描述精度的可能方案.

为改进SAX符号时序方法极值信息丢失较多的缺陷,本章提出了VSB解决方案,VSB方法是基于SAX的改进方案,它通过引入两个极值分量将原来的SAX符号转化为具有三个分量的符号矢量,其符号值由各分量的加权和最终确定.由于 VSB 方案能够比 SAX 提供更多的描述信息,因而它在时序分析中能够获得比 SAX 更好的结果.是一种行之有效的方案.

针对 SAX 符号时序方法对数据描述不充分,难以实现时序特征检测精确分析的缺陷,本节提出了 SFVS 解决方案,SFVS 方法可以看作是 SAX 的改进方案,它通过引入两个统计特征分量将原来的 SAX 标量符号转化为矢量符号.由于 SFVS 方案能够比 SAX 提供更充分的对时序数据的描述信息,因而它在时序分析应用中能够获得比 SAX 更出色的表现.实验证明它确实是一种值得推荐的符号化新方案.

第四章 图像特征检测的自适应谱聚类

近年来,基于谱图理论的谱分析方法的兴起为图像特征检测开辟了一条新途径,其代表性方法谱聚类成为最有竞争力的分析工具之一,受到学术界的高度重视,并广泛应用于特征检测的各个方面.本节针对 STSC(Self-Tuning Spectral Clustering-自调节谱聚类[159]算法的缺陷,提出一种新的自适应谱聚类算法.它用全局平均 N 近邻距离作为比例参数 σ,利用本征矢差异来估计最佳聚类分组数 k,达到了比前者更好的效果,并且更容易实现.在彩色图像分割实际应用的实验结果表明.该算法适应性强,计算代价小,精度较高,性能好于或至少不差于以往的类似算法.

4.1, 相关工作及问题的提出;

图像特征检测和目标识别是模式识别的重要内容.从广义上讲,图像的特征包括基于文本的特征（如关键字、注释等）和视觉特征（如色彩、纹理、形状、对象表面等）两类.视觉特征又可分为通用的视觉特征和领域相关的视觉特征.前者用于描述所有图像共有的特征,与图像的具体类型或内容无关,主要包括色彩、纹理和形状; 后者则建立在对所描述图像内容的某些先验知识（或假设）的基础上,与具体的应用紧密有关,例如人的面部特征或指纹特征等. 图像特征检测与识别是图像分析（如图像分割,边沿检测等）的基础.核心问题,也是它的经典难题之一.

图像分割的主要技术有四类,基于阈值的技术;基于边界的技术;基于区域的技术;混合技术;阈值技术是基于有相似属性的邻接像素,尽管能够得到较好结果,但计算效率不高[160];边界技术以边界像素突变为基础,主要用于候选边的搜索,且通常计算复杂;区域生长技术利用同一区域相邻像素的相似属性,其算法性能与选定的相似标准相关[161];混合方法通常组合几种方法因而能得到更好的结果[162].近年来彩色图像分割由于能够克服灰色图的缺点而展现出广阔前景[163].区域生长直接用于彩色空间,在分割过程中考虑彩色分布并按空间区域重新划分,因而更适合于彩色图像[164].由谱聚类实现的图像分割,它将借标记块将图像分割成同质区块,用主成份算法将彩图投影到二维子空间.从而使图像分析技术进入了新阶段.

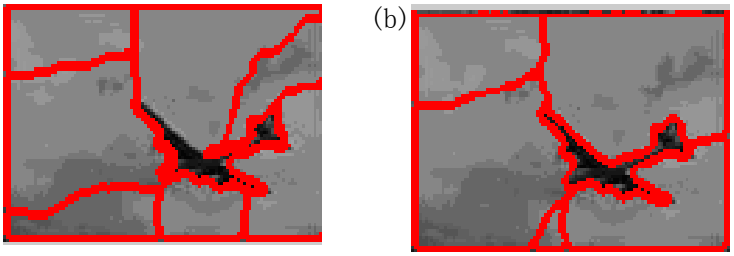
基于谱图理论的谱分析方法作为一种模式识别新方法,以其优异的性能成为其它分类方法强大的竞争对手.谱分析方法[165,166]利用基于数据点间距离矩阵的本征矢分解来实现分类.它并不需要估计明确的数据分布模型,而是依据数据点对相似矩阵的谱分析.通过计算、分析矩阵本征谱,最终利用某种分类算法将数据点分类成组.它容易实现,组合了处理复

杂多维数据集的能力.有望成为最具竞争力的模式识别新技术之一[166].

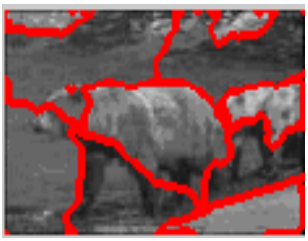
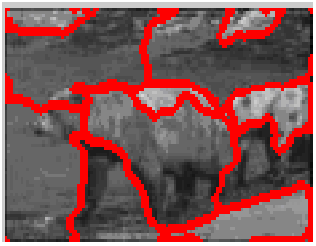
在利用谱分析结果实现分类的各种算法中,最简单且有效的莫过于谱聚类方法[167,168].然而,基于划分的标准谱聚类算法通常需要事先估计聚类参数(如聚类分组数κ,比例参数σ). 聚类结果的优劣通常与聚类参数估计正确性密切相关.对复杂数据集而言,准确估计其聚类参数本身就是一个难题.因而在一般情况下难以保证聚类效果. 由于聚类参数与被处理数据的分布密切相关,现实应用的数据分布通常千差万别.因此在应用中就必须针对不同的数据集设定不同的聚类参数.这难免限制了它的实际应用.如果能够建立聚类参数与数据分布关系的数学模型,或找到一种能够根据数据分布自动估计聚类参数的方法,从而实现对任意数据都能自动估计最佳聚类参数的”自适应谱聚类”, 就能使谱聚类方法如虎添翼.为了解决这个问题,近年来不少学者在这方面做了不懈努力,并且已经取得了一些成果[169,170].

Jordan 和 Y.Weiss 于 2001 年提出了利用数据分布通过分析亲和力矩阵本征值来确定聚类分组数的方法[169],并指明 L 矩阵的第一个最大本征值是一个取值为 1 的重复本征值,其数量等于聚类数,这意味着可以通过计算等于 1 的本征值的数量来估计聚类数.然而,这只适用于各聚类有明显分隔而其间无噪音情况,一旦有噪音,其本征值将与 1 偏离,此法就不再有用;为此,M. Polito 和 P. Perona 于 2002 年在文献[170]中提出了另一种选择,即寻找降幅最大的本征值,但它缺乏足够的理论依据.因为 L 的本征值是与各聚类相应的子矩阵本征值的组合,它们都与单个聚类结构相关.2005 年,Lihi Zelnik-Manor 提出 STSC(Self-Tuning Spectral Clustering-自调节谱聚类)算法模型[159]. 它能根据数据分布状况来自动估计相应数据点的比例参数,利用对 L 矩阵排序后的本征矢结构,用梯度下降法寻找代价相应最小的本征矢组合来自动确定最佳聚类分组数 k;因而实现真正意义上的自动聚类.实验表明,此模型能够处理一些复杂问题,且能达到较高精度,在解决一些现实应用问题中表现出色.然而,该算法的不足之处是它通过反复重组旋转矢量来搜索代价最小的分组数 K 并借此确定最佳分组数 K 的做法,带来了过高计算代价而导致运行缓慢;另一个缺陷是,为了使该算法能够正确估计最佳聚类分组数,仍需要针对不同的数据设定不同的阈值参数,否则就可能出错,这使它的”自调节”性大打折扣.

(a)　(b)



(a)　STSC-图像分割　　(b)　STSC-图像分割

(c)  　(d) 

(c) STSC-图像分割　　(d) STSC-图像分割

图 4-1 自调节谱聚类的图像分割

图 4-1(a),(b),(c),(d)是将该文自己提供算法的源代码用于图像数据集进行二值图像分割的情况;其中,图 4-1(a)是 STSC 算法随机运行时的分割出错结果,而图 4-1(b)为经调整参数后得到的正确结果; 图 4-1(c), 图 4-1(d) 则是它用于复杂图像分割的情况.它说明,对于复杂的图像,STSC 算法的分割结果很难令人满意. 实际上,图 4-1 (c),(d)是经过多次参数调节尝试才找到的相对而言”较好”结果.因此该算法仍需进一步改进.

在许多现实应用中(例如动态目标识别与追踪),实时性和自适应性往往是必须考虑的重要因素之一.理想情况下最好是无需调节任何参数就能快捷地处理各种情况并保证足够小的出错概率.为此,基于上述算法[159,170]的启发,本节提出一种新的自适应谱聚类(Adaptative Spectral Clustering) 算法,以下简称为 ASC 算法.它采用全局平均 N 近邻距离的比例参数 σ_g 代替局部 N 近邻距离的比例参数 σ_l;这可在构造亲和力矩阵时减少计算代价,而采用自适应 κ 选择算法来自动估计最佳聚类分组数 κ,从而避免了 STSC 算法计算代价过高的缺陷.实验表明: 与 STSC 相比,ASC 实现更容易,计算代价更低,出错概率更小.能够满足实际应用的大多数要求.可实现一般意义上的全自动聚类.该算法建立在详谱图理论基础之上.

4.2, 谱图理论

谱聚类是建立在谱图理论基础之上的分类方法,谱图理论[171-172]本质上是一种利用求解数据相关矩阵本征矢来实现数据分类,并通过谱分析获取数据的低维空间表示,从而使分类更容易实现的方法.

谱图理论将数据集 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,...\boldsymbol{x}_n\}$ 用 $\boldsymbol{G}=(\boldsymbol{V},\boldsymbol{E})$ 表示成由顶点集 V 和边集 E 构成的图.其中,图中各顶点 $\boldsymbol{V}$ 代表数据集中的相应数据点, $\boldsymbol{V}=\{\boldsymbol{x}_i\}$;图中联接顶点间边 $\boldsymbol{E}$ 的连接权 $\boldsymbol{W}$ 代表相应数据点间的相似性. 两顶点 $\boldsymbol{v}_i$ 和 $\boldsymbol{v}_j$ 带有非负连接权 $\mathbf{w}_{ij}\geq 0$; 图的加权邻接矩阵是 $\boldsymbol{W}=(w_{ij})$,$i,j=1,2,...n$,表示顶点 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 之间的边的连接权,是一个 $n\times n$ 矩阵.而 $w_{i1}=0$ 表示两个顶点 $\boldsymbol{v}_i$ 和 $\boldsymbol{v}_j$ 无连接.

$$w_{i,j}=e^{-\frac{\|X(i)-X(j)\|_2^2}{\sigma^2}} \tag{4-1}$$

$$\begin{bmatrix} w_{11} & \mathrm{K} & w_{1n} \\ \mathrm{M} & & \mathrm{M} \\ w_{n1} & \mathrm{K} & w_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \mathrm{M} \\ x_n \end{bmatrix} = \lambda \begin{bmatrix} x_1 \\ \mathrm{M} \\ x_n \end{bmatrix} \tag{4-2}$$

由于图 G 是无向的. 故有 $w_{ij}=w_{ji}$. 顶点 $v_i \in V$ 的度数定义为:

$$\boldsymbol{D}(i,i)=\sum_j w_{ij} \tag{4-3}$$

表示与 x_i 相关的边连接权之总和. 是一个 $n\times n$ 对角矩阵. 事实上, 此式只对与 v_i 相邻的顶点求和, 而对其它顶点 $\mathbf{v}_j$ 的权 $w_{ij}=0$; 而度数矩阵 $\boldsymbol{D}$ 定义为具有度为 $d_1,d_2,...d_n$ 的对角矩阵, 于是分类问题就转化成了相似图的分割问题.

为实现分类, 可将在一组内只有一个顶点的连接边集看作是边割集, 将分割目标表示为待划分边割集的函数. 分类过程就是寻找最小切割分组的过程.

$$cut(\boldsymbol{A},\boldsymbol{B})=\sum_{i\in A,j\in B} w_{ij} \tag{4-4}$$

目前采用的分类标准主要有两种:其一是最小切割标准

$$\min cut(\boldsymbol{A},\boldsymbol{B})=\sum_{i\in A,j\in B} w_{ij} \tag{4-5}$$

它按组间最小连接权分组. 简单明了, 但它只考虑了分组外部连接而未考虑内部密度. 且运行效率不高, 更好的方法是采用规范化最小切割标准[186]

$$\min\ \ Ncut(\boldsymbol{A},\boldsymbol{B})=\frac{cut(\boldsymbol{A},\boldsymbol{B})}{vol(\boldsymbol{A})}+\frac{cut(\boldsymbol{A},\boldsymbol{B})}{vol(\boldsymbol{B})}, \tag{4-6}$$

式中. $vol(\boldsymbol{A})$ 代表源于组 A 的边之总权数. 它考虑了与各组密度相关的组间连通性. 能获得更平衡的分割. 最小规范化切割与最大规范化关连方法等价.

采用最小规范化切割时须要考虑方程(4-1)~(4-4)有

$$\begin{aligned} &(\boldsymbol{D}-\boldsymbol{W})\boldsymbol{X}=\lambda \boldsymbol{D}\boldsymbol{X} \\ &\boldsymbol{X}_A(i)=\begin{cases} 1 & if \quad i\in A \\ 0 & if \quad i\notin A \end{cases} \end{aligned} \tag{4-7}$$

与之相关的属于组 A 的边之总权数为

$$vol(\boldsymbol{A})=\sum_{i\in A}\boldsymbol{D}_i, A\subseteq V$$

图中的割集表示为:

$$cut(A,\overline{A}) = \sum_{i\in A, j\in \overline{A}} w_{i,j} \tag{4-8}$$

分割矩阵 $\boldsymbol{X}$ 为

$$\boldsymbol{X} = [\boldsymbol{X}_1, ...\boldsymbol{X}_k] \tag{4-9}$$

拉普拉斯矩阵 $\boldsymbol{L}$:

$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W} \tag{4-10}$$

注意到非规范化图拉普拉斯并不依赖于邻接矩阵 $\boldsymbol{W}$ 的对角元素. 在所有非对角位置与 W 相应的各矩阵导致相同的非规范化图拉普拉斯 $\boldsymbol{L}$, 非规范化图拉普拉斯和它的本征值和本征矢可以描述许多图形属性. 例如用于谱聚类:令 $\boldsymbol{G}$ 是一个具有非负连接权的无向图则在图中, $\boldsymbol{L}$ 的 0 本征值乘数 k 等于其连接分量数 $A_1, A_2, ...A_n$. 这些分量的指示矢量 $1_{A1}, 1_{A2}, ...1_{Ak}$ 跨越 0 本征值的本征空间. 为计算要求通常将拉普拉斯矩阵规范化为下式

$$\boldsymbol{L} = \boldsymbol{D}^{-0.5} \times (\boldsymbol{D} - \boldsymbol{W}) \times \boldsymbol{D}^{-0.5} \tag{4-11}$$

4.3. 谱图理论的本征问题:

根据谱图理论[171, 172]若第 i 个顶点度数表示为 $\boldsymbol{D}_i = \sum_j w_{ij}$, 属于组 A 的边之总权数为

$$vol(\boldsymbol{A}) = \sum_{i\in B_1} \boldsymbol{D}_i \ , \ vol(\delta \boldsymbol{A}) = cut(\boldsymbol{A}) = \sum_{i,j\in E, i\in A, j\in \overline{A}} W_{ij} \tag{4-12}$$

这里, $\delta \boldsymbol{A}$ 为 $\boldsymbol{A}$ 与 $\overline{\boldsymbol{A}} = \boldsymbol{V} - \boldsymbol{A}$ 间连接边界的边集, 而 $vol(\delta \boldsymbol{A})$ 为连接边界的总权数. 若定义平

衡割集

$$b_w(g) = \min_{A\in V, vol(\boldsymbol{A}) = vol(\overline{A})} cut(\boldsymbol{A}) \tag{4-13}$$

定义矢量 $\boldsymbol{f}$ 的子集 $\boldsymbol{f}_A$ 为

$$\boldsymbol{f}_{A_i} \in \boldsymbol{R}^N = \begin{cases} 1, [i] \in A \\ -1, [i] \in \overline{A} \end{cases} \tag{4-14}$$

令 $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$ 并考虑到仅有源于不同集间有连接边的顶点才有计算意义:

$$\begin{aligned} \boldsymbol{f}^T\boldsymbol{L}\boldsymbol{f} &= \boldsymbol{f}^T\boldsymbol{D}\boldsymbol{f} - \boldsymbol{f}^T\boldsymbol{W}\boldsymbol{f} = \sum_{i=1}^{n}\boldsymbol{D}_i\boldsymbol{f}_i^2 - \sum_{i=1}^{n}\boldsymbol{D}_i\boldsymbol{f}_i\boldsymbol{f}_j \\ &= \frac{1}{2}(\sum_{i=1}^{n}\boldsymbol{D}_i\boldsymbol{f}_i^2 - 2\sum_{i,j=1}^{n}\boldsymbol{f}_i\boldsymbol{f}_j w_{ij} + \sum_{i=1}^{n}\boldsymbol{D}_j\boldsymbol{f}_j^2) \\ &= \frac{1}{2}\sum w_{ij}(\boldsymbol{f}_i - \boldsymbol{f}_j)^2 = 4vol(\delta\boldsymbol{A}_1) \end{aligned} \tag{4-15}$$

因每一边都是按对角线求和

$$\boldsymbol{f}^T\boldsymbol{D}\boldsymbol{f} = \sum \boldsymbol{A}_{ij} = vol(\boldsymbol{v})\,; \tag{4-16}$$

$$\boldsymbol{f}_{A_1}^T\boldsymbol{D}\mathbf{1} = \sum_{i\in A_1}\boldsymbol{D}_i - \sum_{i\in\overline{A_1}}\boldsymbol{D}_i = vol(\boldsymbol{A}_1) - vol(\overline{\boldsymbol{A}_1}) \tag{4-17}$$

上式仅当　　$vol(\boldsymbol{A}_1) = vol(\boldsymbol{A}_1')$　时才会为 0 ， 经代数变换有

$$b_w(g) = \min_{f\in(1,-1),f\boldsymbol{D}1=0} \boldsymbol{f}^t\boldsymbol{L}\boldsymbol{f} \tag{4-18}$$

$$b_w(g) = \frac{vol(\boldsymbol{v})}{4}\min_{f\in(1,-1),f\boldsymbol{D}1=0}\frac{\boldsymbol{f}^t\boldsymbol{L}\boldsymbol{f}}{\boldsymbol{f}^t\boldsymbol{D}\boldsymbol{f}} \tag{4-19}$$

将此式推广到 n 维空间 R^n 后

$$b_w(g) = \min_{f\in\boldsymbol{R}^n,f\boldsymbol{D}1=0}\frac{\boldsymbol{f}^t\boldsymbol{L}\boldsymbol{f}}{\boldsymbol{f}^t\boldsymbol{D}\boldsymbol{f}} \tag{4-20}$$

综合可得:

$$\begin{aligned} &\min_x Ncut(\boldsymbol{x}) = \min_f \frac{\boldsymbol{f}^T(\boldsymbol{D}-\boldsymbol{W})\boldsymbol{f}}{\boldsymbol{f}^T\boldsymbol{D}\boldsymbol{f}} \\ &Subject\quad to:\quad \boldsymbol{f}^T\boldsymbol{D}\mathbf{1} = 0 \end{aligned} \tag{4-21}$$

或

$$\begin{aligned} &\min_x Ncut(\boldsymbol{x}) = \min_y \frac{\boldsymbol{f}^T(\boldsymbol{D}-\boldsymbol{W})\boldsymbol{f}}{\boldsymbol{f}^T\boldsymbol{D}\boldsymbol{f}} \\ &Subject\quad to:\quad \boldsymbol{f}^T\boldsymbol{D}\boldsymbol{f} = 1 \end{aligned} \tag{4-22}$$

由此得本征方程　　$(\boldsymbol{D}-\boldsymbol{W})\boldsymbol{f} = \lambda\boldsymbol{D}\boldsymbol{f}$ (4-23)

注意到　　$(\boldsymbol{D}-\boldsymbol{W})\mathbf{1} = 0$ (4-24)

因此, 第一个本征矢为 $\boldsymbol{f}_0 = 1$, 本征值 l =0 ;它不能提供任何信息. 第二个最小本征矢是实数值, 对应于问题的解. 用此本征矢结合划分阈值, 即可用相应的分类算法实现分类. 当然, 也可用多个本征矢结合相关的分类算法进行分类.

上述规范化拉普拉斯矩阵有下列属性:

1. 是一个$n \times n$对称矩阵 ;

2. 本征值是非负实数.

3. 本征矢为实数且互相正交.

4. 本征值和本征矢提供了图连通性信息.

且矩阵$\boldsymbol{L}$满足下列属性:

1, 对每一个矢量 $\boldsymbol{f} \in \boldsymbol{R}^n$ 有:

$$\boldsymbol{f}^T \boldsymbol{L} \boldsymbol{f} = \sum_{i,j=1}^{n} w_{ij} (\boldsymbol{f}_i - \boldsymbol{f}_j)^2 \qquad (4\text{-}25)$$

2. λ是$\boldsymbol{L}_{rw}$ 的本征矢为$\boldsymbol{u}$的一个本征值. 当且仅当λ是$\boldsymbol{L}_{sym}$的本征矢为

$\boldsymbol{W} = \boldsymbol{D}^{1/2} \boldsymbol{u}$ 的本征值 (4-26)

3. λ是$\boldsymbol{L}_{rw}$的本征矢为$\boldsymbol{u}$ 的一个本征值. 当且仅当λ和$\boldsymbol{u}$解广义本征问题

$$\boldsymbol{L}\boldsymbol{u} = \lambda \boldsymbol{D}\boldsymbol{u} \qquad (4\text{-}27)$$

4. 0 是$\boldsymbol{L}_{rw}$的一个本征矢恒为 1 的本征值,0 也是$\boldsymbol{L}_{sym}$的本征矢为 1 的本征值.

5. $\boldsymbol{L}_{sym}$ 和 $\boldsymbol{L}_{rw}$ 是正半定且有 n 个非负实值本征值

$$0 \le \lambda_1 \le \lambda_2 \le, \ldots \lambda_n \qquad (4\text{-}26)$$

因此,通过解此矩阵本征值问题, 就可利用矩阵本征值和相应的本征矢提供的数据分布结构信息来进行”谱”分析,例如实现谱聚类.这里所说的”谱”,是指由(按本征值大小排序对应的)一组本征矢构成的序列.

$$\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots \boldsymbol{u}_n \qquad (4\text{-}27)$$

4.4, 自适应谱聚类算法:

(1)基本算法描述

将基于谱图理论的谱聚类算法 [174]应用于图像分析（如图像分割）问题,是图像特征检测或识别技术研究的新热点.简单地说,基于谱分析的算法主要包括以下三个基本步骤:

1,预处理: 根据谱图理论构造一个表示数据集的矩阵.即将数据集表示成用

G = (V,E)由顶点集 V 和边集 E 构成的有向图.

2,分解: 计算该矩阵本征值和本征矢.并利用一个或多个本征矢将各点映射到低维空间.

3.分类. 以新的表示和相应的分类算法为基础,将各点分配到相应的类别中.

若上面第 3 步采用 k-means 聚类算法作为划分器,则上面的算法就是谱聚类[175];

谱聚类的实现有两种基本的方法: 即二分法谱聚类与多本征矢谱聚类.

二分法的缺点是低效不稳定,在此,我们不予考虑;而多个本征矢谱聚类用多个本征矢

构造一个减维空间,它能够用于近似最佳规范化割集[176].增加数据分布的不均衡,使相似点间的关连被放大,而不相似点间关连被削弱. 能够将数据转换到由 k 个正交基构成的新”嵌入空间”,同时采用多个本征矢有利于防止由于信息损失造成的不稳定[177].

本征矢谱聚类算法是上述基本算法用于多本征矢谱聚类的具体实例.它主要包括

数据预处理:　将输入数据标准化.

构造相似矩阵:计算相似点对间的距离,并构造相应的相似矩阵.

本征分解:　找 A' 的本征值和本征矢,用与 k 个最大本征值相应的本征矢构造嵌入空间

分组:　用 k-means 算法将空间划分以形成 k 个聚类.

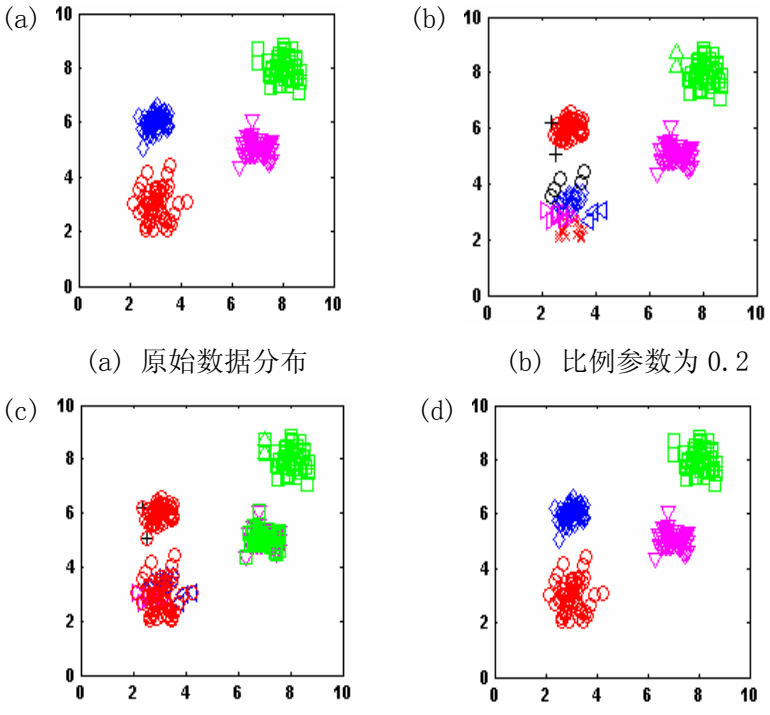
本征间距:　两个连续本征值间的差值

$$\Delta_k = \left|\lambda_k - \lambda_{k-1}\right| \tag{4-28}$$

通过解此矩阵本征值问题,就可利用本征值和相应的本征矢矩阵提供的数据分布结构信息来进行”谱”分析.

(2)　比例参数 σ 的估计:

一般聚类算法中需要用户设定的参数主要有比例参数 σ 和聚类分组数 κ;前一个参数的主要作用是根据数据分布状况(密度/间距)调节测量尺度,以确保测量对象规范化并落入测量范围内.不合适的比例参数将导致检测错误.另一个参数的作用主要是用最佳的划分数确保误差最小.



(a) 原始数据分布　(b) 比例参数为 0.2

(c) 比例参数为 0.6　(d) 比例参数为 0.01

图 4-2 不同比例参数取值的谱聚类效果比较

图 4-2 是对同一数据采用不同比例参数的聚类结果,其中,图 4-2(a)是数据的原始真实分布,图 4-2(b)~(c)是聚类结果误差较大的情况,尤其是图 4-2(c),居然把 4 个聚类识别为 3 类.图 4-2(d)则是当比例参数设置合适时所得的正确聚类结果,它和原始分布几乎一致.

上面结果清楚地说明,比例参数选择对聚类结果正确与否至关重要.目前最常用的方法是通过实验或经验地确定.典型方案是网格式搜索所有可能的σ值,进而通过结果比较而选择最佳σ值.这显然存在很多缺陷.自然也难以达到理想的结果.

为解决σ值的有效选择问题,许多学者提出了一些基于模型的方法.文献[171]提出了自动选择比例参数σ的方法,其基本思路是通过对反复运行不同σ值所得结果的比较,从中选出能够产生最佳聚类结果的σ值,这无疑是一个成功的尝试,但由此带来的计算时间剧增却不利于它的现实应用.此外此算法σ值的检测范围仍需人工设定,且只适用于数据分布具有全局统计特征的情况.当数据分布具有不同局部统计特征时,对不同的区域就应采用不同的比例参数σ值,才能得到较理想的结果.为此 L. Zelnik-Manor and P. Perona 在文献[178]中提出了采用局部 N 近邻距离作为σ参数的方法,其基本要点是:按数据分布的不同逐点计算确定相应的σ参数,因而一个数据集可能由于分布的局部统计特征差异会有多个反映此差异的不同比例参数的σ值.文献[179]的实验也表明此法在解决一些数据分布复杂的问题时确实能够得到较好的聚类结果.尽管能够改进聚类精度,然而逐点计算确定σ参数的做法同样极大的增加了计算代价.对那些实时性要求很强的应用就带来不利.

实际上.在大多数现实问题中,对同一个数据集而言,直接用全局N近邻平均距离作为σ参数足以满足精度要求,这既能省去了文献[171]的由于反复比较σ值带来的计算代价,又能节省文献[178]那样由于逐点计算σ值带来的巨额计算代价;出于这种考虑,本节采用数据集全局 N 近邻平均距离作为σ参数的近似估计.

$$s_i = d(\mathbf{x_i}, \mathbf{x}_N) \tag{4-29}$$

这里, X_N 为,X_i 的第 N 个邻域 (在一般情况下,N 的取值为 5-9.)

$$s = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{N}\sum_{i=1}^{N} s_i \tag{4-30}$$

在一般情况下,N 的取值为 5-9 为佳.

(3) 聚类分组数 k 的选择

仔细分析本征问题(4-23)的解可知.按L本征值排序的本征矢矩阵X中的每一行,代表特征空间中的一个点;而每一列对应于特征空间的一个特征(座标)分量;通常估计最佳聚类分组数,实质上是选择能够实现对数据准确划分的最佳特征数.由于本征矢矩阵X的各列是按其相应本征值的大小排序生成,因此最佳划分问题最终成为选择多少列本征矢问题.

确定最佳聚类数的一种可能途径是分析亲和力矩阵的本征值,文献[177]中给出的分析表明, L 的第一个最大本征值将是一个大小为 1 的重复本征值,其重复数等于分组数 k,这意味着可以通过计算本征值等于 1 的次数来估计 k. 然而,文献[178]指出,这个结论只在无噪音数据集的情况成立. 一旦含有噪音,不同分组间就难以明确划分,相应的本征值也不再为 1,于是又需另外选择估计标准. 一种可能的途径是寻找幅度下降最大的本征值,然而这种方法目前尚缺少理论依据. L 的本征值是各聚类对应子矩阵本征值的组合,这意味着本征值与各聚类结构相关,尤其是,第 k 个本征值与第 $k+1$ 个本征值间的的间隔大小并不确定,而与数据集的分布模式密切相关,因而无法假定可以简单地用此法一定能够取得满意的结果. 于是,文献[166,178]提出利用本征矢结构确定最佳聚类数的方法:在按聚类对 L 排序后,在理想情况下(当 L 是严格地块对角时-有块 $L(c)$,C=1,…C)),它的本征值和本征矢是适当填补了零的块本征值和本征矢的组合. 只要各块的本征值是不同的,各本征矢将仅为单一块时才有非零值.

$$\boldsymbol{X}=\begin{bmatrix} x^{(1)} & 0 & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & 0 & x^{(C)} \end{bmatrix}_{n\times C}$$

其中, $\boldsymbol{x}^{(C)}$ 是相应于聚类 C 的子阵 $\boldsymbol{L}^{(c)}$ 的本征矢,由于本征值 1 是多个等于分组数 K 的重复本征值,所以能够用选择 $\boldsymbol{X}$ 的跨越同一子空间的正交矢其它列的方法来选择本征矢,也就是,可以用满足 $R\in R^{C\times C}$ 的任何正交矩阵 $\boldsymbol{X}=\boldsymbol{X}\boldsymbol{R}$ 来代替 $\boldsymbol{X}$. 这意味着即使本征解提供的是旋转矢量集,仍可保证存在一个旋转 $\boldsymbol{R}$. 以致于在矩阵 $\boldsymbol{X}\boldsymbol{R}$ 各行有单一非零项. 由于 $\boldsymbol{L}$ 的本征矢是各单块子矩阵本征值的组合,而与本征值间大小无关,文献[178]正是基于此思路在所有可能的旋转中用梯度下降法寻找最小代价函数的方案来选择最佳聚类数 k 的.

相应的实验也证实了这个方案的可行性和可靠性. 然而,该算法的不足之处是它通过反复重组旋转矢量来搜索代价最小的分组数 k 并借此确定最佳分组数 k 的做法,带来了过高计算代价而导致运行缓慢;另一个缺陷是,为了使该算法能够正确估计最佳聚类分组数,仍需要针对不同的数据设定不同的阈值参数,否则就可能出错,这使它的”自调节”性大打折扣.

为此,本文利用相邻本征矢平均差异的方式来确定近似最佳分组数的方案. 即通过计算 $\boldsymbol{X}$ 矩阵中前 k 个相邻列间的平均差异与第一列的比值,直到此差异比小于给定的阈值时,此时相应的 k 值便是近似最佳选择. 若(4-23)式的解按本征值排序生成的本征矢矩阵为 $\boldsymbol{X}$,则相邻列矢量间的平均差异的绝对值可表为:

$$H_{i-1,j}=\left|\overline{\boldsymbol{X}_{i-1}-\boldsymbol{X}_{i}}\right| \quad i=1,2,\ldots k,\ldots n \tag{4-31}$$

它与第一列(最大本征值对应的本征矢)平均值之比为:

$$J_i = H_{i-1,j} \Big/ \left|\overline{\boldsymbol{X}_1}\right| \quad i = 1,2,...k,...n \tag{4-32}$$

其中 J_i 是随着的 i 增大而减小的变量,随着 $\boldsymbol{X}$ 矩阵中的相邻列(分量)差异减小,相应列对聚类划分的作用也减小,当它减少到一定程度时,继续增加列矢量数对聚类划分已经不再有益.此时的 i 便是最佳聚类数.以下是实现上述思路的算法伪代码:

(4)自适应 k 选择算法:

1.初始化: 输入 $\boldsymbol{L}$ 的按本征值排序规则生成的相应本征矢矩阵 $\boldsymbol{X}$,并设下列初值.

```
i=1;H(i)=c1;J(i)=c2; q=c3;
%% c1=c2 为大于 1 的任意初始常数,c3=0.1~0.2.
```

2.计算相邻各列差异平均比值:

```
while J(i)> q  % q—阈值
    i=i+1;
    H(i)=abs(mean(X(:,i-1)-X(:,i)));
    J(i)=H(i)/abs(mean(X(:,1)));
End
```

3.输出选定的 k 值.

```
k = i;
```

4.5　用于彩图分割的 ASC 算法:

给定一组数据点

$$\boldsymbol{S} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, ...\boldsymbol{s}_n\} \in \boldsymbol{R}^l \tag{4-33}$$

为了将这些原始数据点聚类成 k 个子集须要进行以下步骤:

1, 按(4-29)(4-30)式计算平均比例参数 σ_I, σ_s ;其中,I ,S 分别代表像素点亮度,距离信息.由此构造亲和力矩阵为:

$$\boldsymbol{W}(i,j) = e^{-\left(\left\|\boldsymbol{I}(i)-\boldsymbol{I}(j)\right\|_2^2 \Big/ \boldsymbol{\sigma}_I^2 + \left\|S(i)-S(j)\right\|_2^2 \Big/ \boldsymbol{\sigma}_s^2\right)} \tag{4-34}$$

if $i \neq j$; $w_{ii} = 0$;

2, 定义对角矩阵

$$\boldsymbol{D}_{ii} = \sum_{j=1}^{n} w_{ij}$$

并由此构造标准拉氏矩阵

$$\boldsymbol{L} = \boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$$

3, 解 $\boldsymbol{L}$ 的本征方程 ,求本征值及本征矢;

4, 排列 $\boldsymbol{L}$ 的本征矢,以形成新矩阵 $\boldsymbol{X}$ 的的各列矢量:

$$\boldsymbol{x}_1, \boldsymbol{x}_2, ...\boldsymbol{x}_k, ...\boldsymbol{x}_n$$

5, 运行自适应 k 选择算法,确定最佳聚类数 k;

6, 以此为依据(自动)选择 X 矩阵中的前 k 列本征矢;形成新的 X 矩阵;

7, 将 $\boldsymbol{X}$ 的各行按下式重新规范化为单位长矢量 $\boldsymbol{Y}$,

$$\boldsymbol{Y}_{ij} = \boldsymbol{x}_{ij} \Big/ (\sum\nolimits_j \boldsymbol{x}_{ij}{}^2)^2 \qquad Y \in R_X^{n\times k}$$

8, 将 $\boldsymbol{Y}$ 的各行视为 $\boldsymbol{R}^k$ 空间的点,用 K-means 算法聚类成 k 个聚类.;

9, 标记原始数据集.当且仅当 $\boldsymbol{Y}$ 的第 i 行标记为聚类 j 时,把原始数据点 s_i 标记为聚类 j.

4.6. 实验分析:

为了验证算法的有效性,本节采用 Berkeley 彩色图像数据集[180],相应的图像文件名详见下表 4-1,并分别用 ASC 与 STSC 算法对相同图像数据进行图像分割对比实验. 需要说明的是,STSC 算法直接采用了原始论文所提供的源代码进行二值图像分割,其结果详见图 4-1,而 ASC 算法则用于进行彩色图像分割,其结果详见图 4-3.下面表 4-1 所列是各相应算法在处理不同图像分割任务时 50 次运行的平均结果.实验表明:与 STSC 算法相比,ASC 算法时间代价少于 STSC. 说明 ASC 确实较 STSC 更高效.

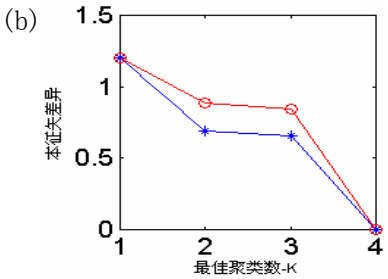
表 4-1: ASC 与 STSC 算法比较

	算法平均运行时间 (秒)	
图像名	STSC	ASC
3096.jpg	27.20	**11.25**
302003.jpg	23.38	**8.92**
100075.jpg	25.91	**11.52**
113044.jpg	28.90	**7.07**
161062.jpg	22.44	**13.10**

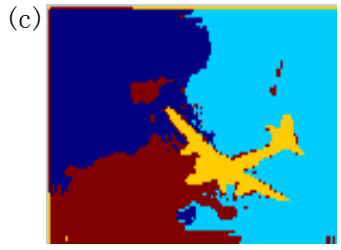
前面的图 4-1(a),(b),(c),(d)和下面的图 4-3,图 4-4,图 4-5 的(a),(b),(c),(d)子图分别记录了 STSC 及 ASC 算法用于图像分割实验的情况.其中, 图 4-3,图 4-4,图 4-5 的

子图(a)是原始图像(照片);子图(b)反映了自适应 k 选择算法运行情况,图中的红色曲线总体向下的趋势显示了本征矢差异比随着聚类分组数 k 的增加而下降的特点.而兰色曲线则显示了随着所选本征矢列数的增加相邻本征矢差异逐渐减小的事实,当本征矢差异比减小到相应的阈值而导致程序终止时,此时的 k 便是最佳分组数;子图(c)反映的是相应图像的分割聚类分布情况;子图(d)则是按分割聚类进行色彩平均填充后的最终结果.

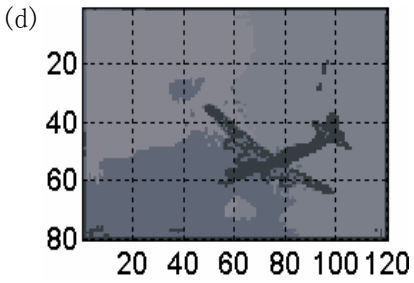
(a) 原始图像



(b) K 值选择



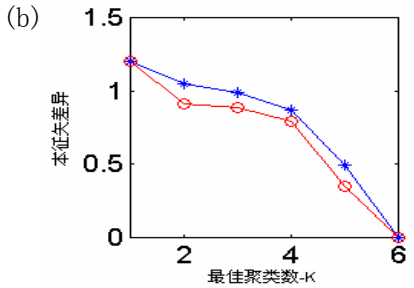
(c) 分割聚类



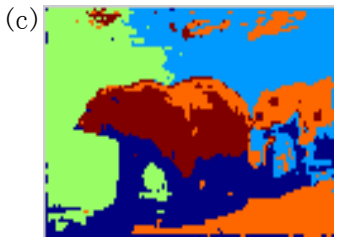
(d) 图像分割

图 4-3 ASC 彩色图像分割

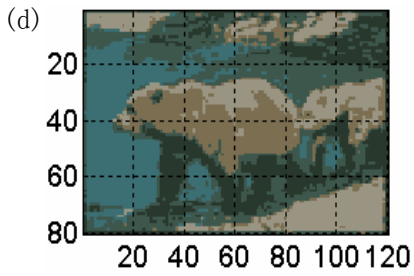
(a)原始图像



(b) K 值选择

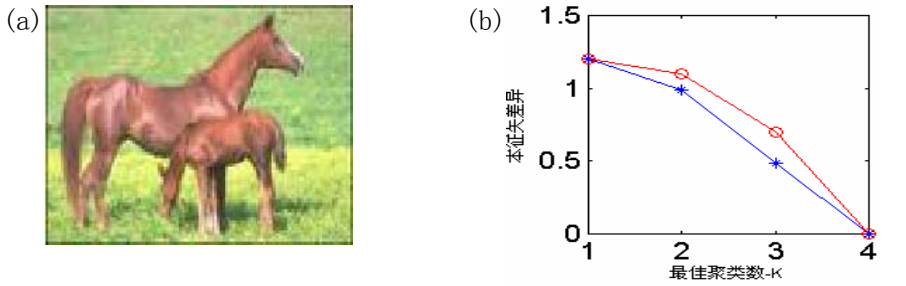


(c) 分割聚类

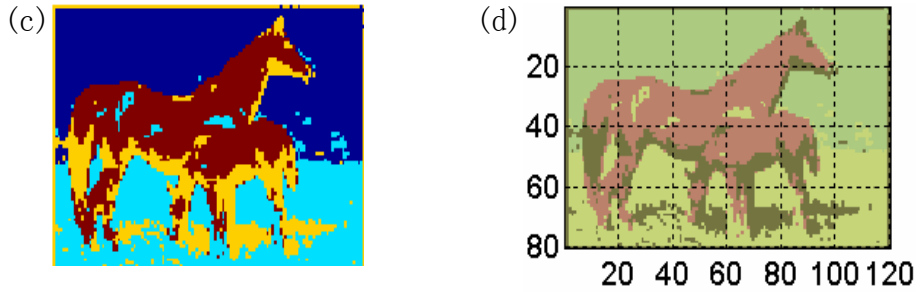


(d) 图像分割

图 4-4 ASC 彩色图像分割



(a) 原始图像　　(b) K 值选择



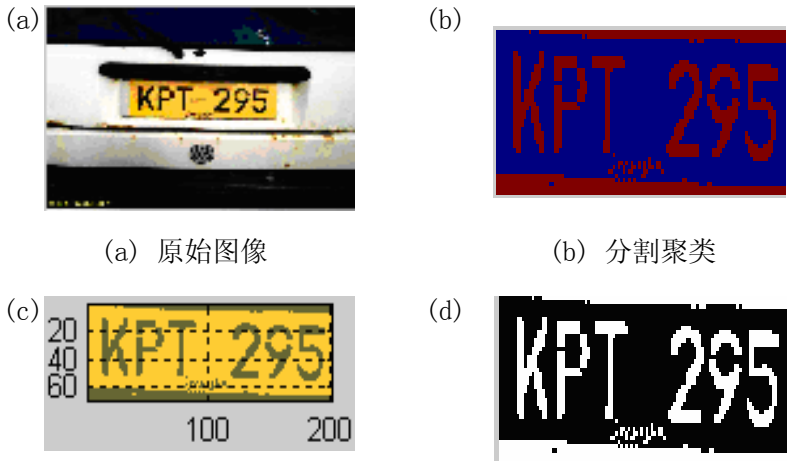
(c) 分割聚类　　(d) 图像分割

图 4-5　ASC 彩色图像分割

比较图 4-3(d)，6-4(d)，6-5(d)所示 ASC 彩图分割与前面 6-1 的 STSC 图像分割结果，显然，ASC 彩图分割的结果能够更准确地反映被检测对象的真实情况.

应用真实数据的典型实验：

下面图 4-6，图 4-7 是将 ASC 算法应用于二个典型真实图像数据的结果. 其中，图 4-6 (a)是一张待识别的车牌照片，(b) (c)是用 ASC 算法得到的分割聚类和彩色图像分割，(d)是用分割结果识别得出的目标检测结果.



(a) 原始图像　　(b) 分割聚类

(c) ASC 图像分割　　(d) 目标检测

图 4-6　ASC 彩色图像分割

(a) 原始图像　　(b) 分割聚类　　(c) ASC 图像分割

图 4-7　ASC 彩色图像分割

图 4-7 则是直接一张双人照,从上面图像分割的结果来看,ASC 算法可以得到满意的结果,确实能够胜任处理复杂的实际问题的任务.

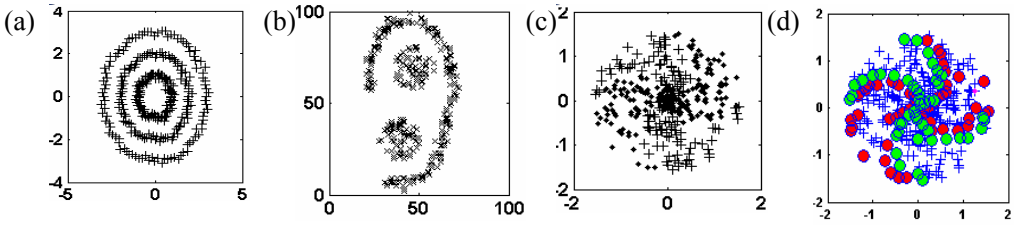
4.7,小结.

根据数据分布自动估计聚类参数及复杂背景的图像分割是一个难题, 而 ASC 算法为解决这一难题提供了又一种新的选择方案.尽管此方案能够较好解决彩色图像分割的大量应用问题.较以往的算法有了一些进步,然而,谱分析理论和应用研究远未完结.更高效实用的算法有待进一步探索.

第五章. 改善特征检测性能的半监督式学习

5.1. 问题的提出

前面所述的谱分析方法[181,165]利用基于数据点间距离矩阵的本征分解来实现分类.对于有明显稠密区且无噪声的数据集,这显然是一种简单有效的分类方法.然而,现实应用的很多情况下,待处理数据可能根本就没有明显稠密区或者含有大量噪声,此时,谱分析方法效果就难尽人意.



(a),(b),(c)-原始数据 (d)加入了标记的数据

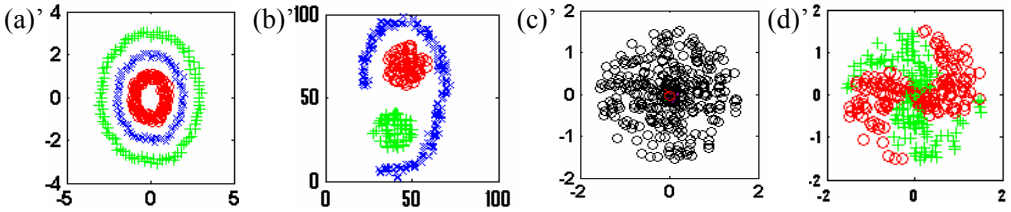


图 5-1 (a)',(b),(c)',(d)'--对数据(a),(b),(c)(d)的谱聚类

图 5-1(a)'(b)'(c)'(d)'是将谱聚类用于(a),(b),(c)所示的原始数据集(图 5-1(c)是 UCI 数据集four_spiral.)所得的结果.其中, 由于图 5-1 (a),(b)数据集中各数据块有明显的稠密区且无噪声,因而能够得到较好的聚类结果,见图 5-1(a)'(b)'.然而图 5-1(c)的数据分布无明显稠密区或者说数据块边界区有较大的噪声,用谱聚类分类的结果就会出现错误(将两种不同数据聚为一类) ,见图 5-1(c)'.显然,单纯的谱聚类无法有效处理无明显稠密区或含有大量噪声数据的问题.然而,这一问题在图像特征检测中却普遍存在,例如, 在复杂图像分析中,当被检测的对象本身与背景界线模糊,或由于色彩过于丰富导致数据分块边界间特征差异过小等.此时,为了能够让聚类方法继续有效,解决的可能途径有二：其一是为待处理数据集增加相应的特征分量,利用已知先验知识的基本特征作为部分数据的标记,用相应半监督式学习算法学习其它数据的未知标记,进而将标记完整的特征分量加入到原始数据集中,就可用谱聚类实现正确分类了,例如上图 5-1 (d)是原始数据中加入标记数据的示意图,图 5-1 (d)'就是用加入标记数据进行半监督式学习前提下进行谱聚类的结果,图中显示的谱聚类结果与原始数据几乎一

致(图中用色彩标记的聚类与原始数据分布一致)；其二是,直接将传统的无监督式谱聚类改进为半监督式聚类,这同样需要利用已知的先验知识(已知的部分数据标记)来改造聚类;总之,为了拓展图像特征检测的聚类模型的应用范围,使其能够适应于更加复杂的实际情况,本文下面提出相应的半监督式学习算法.其中,基于支持向量机的半监督式学习模型对应于上面所述的途径一；而交互式半监督式聚类对应于上面所述的途径二.

5.2, 半监督式学习：

半监督式学习（SSL）是介于监督式与无监督式学习之间的一种学习方式.它利用少量有标记数据和大量无标记数据进行反复监督式学习.半监督式学习的数据集通常由少量的标记数据 $x_l=(x_1,...x_l)$, $y_l=(y_1,...y_l)$.与大量无标记数据 $x_u=(x_{l+1},...x_{l+u})$ 组成,无标记数据的标记是未知待求的.

5.2.1 半监督式学习的基本假定：

半监督式学习是建立在以下假定之上的.

（1）平滑假定：

两个相互接近的点 x_1,x_2 其相应的标记也应接近.此假定保证了从有限训练集向无限测试集的推广；

（2）聚类假定：包括

a, 在同一聚类中的各点可能属于同一个类,这可看作平滑假定的特例;

b, 低密度分隔,决策边界应该位于低密度区.

（3）流形假定:

(高维)数据基本上位于低维流形之上.流形学习理论有助于解决高维数据在统计算法中的维数灾难问题.

5.2.2 半监督学习的主要研究方向：

在过去几年,半监督式学习已经成为机器学习领域新的研究热点之一. 目前的半监督式学习主要有以下研究方向:

1,基于产生式模型:主要包括,半监督式学习的分类方法,半监督式期望最大分类,有约束的半监督式聚类,转换式支持向量机 TSVM;

2,基于低密度分隔:主要包括转换式支持向量机 TSVM 高斯过程,熵调节,数据相关调节等方法;

3,基于图形的方法：主要有标记传播和二次规范方法;

4,谱图理论方法：谱转换的图形核,维数约简的谱方法;

5.基于矩阵的方法.

在众多半监督式学习方法中：本节首先着重研究产生式和基于低密度分割的综合方法.具体地说,就是基于支持向量机的渐近式自训方法.

自训是一种最通用的半监督学习方式.用于自训的分类器首先用少量有标记数据训练自己,然后用训练模型分类无标记数据.训练过程中不断将最可能为某标记的无标记数据点连同它们的标记移送到训练集.进行反复训练.、训练过程始终以其自身的预测来教自己.自训技术已被应用于自然语言处理. Yarowsky (1995)[182]将它用于词意解疑,Riloff et al. (2003)[183]用它识别主题词.Maeireizo et al. (2004)[184]用自训分类器将对话分类为情绪性和非情绪性两类.Rosenberg et al. (2005)[185135]将自训用于图像目标检测系统.Haffari, G. & Sarkar, A.[186]还有 Culp, M. & Michailidis, G.(2007)通过分析几个基于规则的算法.提出了对 Yarowsky 算法的改进方案[187].并取得较好结果.

5.3. 典型的半监督式学习模型

(1)用于自训的 SVM 数学模型

SVM 求解可归结为二次规划问题.对于给定的训练数据集,

$$(x_i,\ y_i,\ i=1,2...n),\quad x \in R^d;\ y \in (+1,-1); \tag{5-1}$$

考虑一般的非线性情况：并以 ξ_i 表示松驰变量,则问题的数学表达式可为:

$$Min\quad \Phi(w\ ,\xi_i)=\frac{1}{2}(w*w)+C\sum_{i=1}^{l}\xi_i$$

$$\text{i=1, 2}\cdots\text{l} \tag{5-2}$$

$$\text{s.t}\qquad y_i(w\cdot x_i+b)\geq 1-\xi_i\ ;$$

$$\xi_i\geq 0;\quad i=1,2..l$$

SVM 模型为解决各种模式识别、函数模拟及预测问题提供了有力的模式分析工具.然而,SVM 只能用于有标记数据的模式分析问题,而现实生活中的大量数据都是无标记的,这使它的应用范围受到了限制.目前,解决这一问题的有效途径之一是采用半监督式学习[186]

(2) 半监督式学习模型

半监督式学习是利用少量有标记数据训练初始分类器.进而用大量无标记数据来进一

步改进初始分类器的性能最终达到精确学习的一种综合学习方法.尽管它是一个新的研究领域,目前已经取得一些初步成果[188].半监督式学习最早的模型或许是产生式方法,具有产生式特点的一种半监督式学习途径是将数据转换成由产生式模型决定的特征表示,然后用判决式分类器来处理这些新的特征表示. 由于判决式方法直接利用类条件概率$P(\mathbf{y}|\mathbf{x})$,在参数估计循环中有偏离$P(\mathbf{x})$的危险,而直推式支持向量机通过引导决策边界远离稠密区的方法构建决策边界与$P(\mathbf{x})$间的联系,因而成为一种克服这种危险的较好选择.尽管精确的TSVM解是NP难问题,但一些近似的方法已经提出并有积极的效果[189,190].典型的直推式支持向量机 (T.Joachims-TSVM)算法原理[191]可以表述为,给定一组独立同分布的有标记数据集

$$\begin{gathered}(\boldsymbol{x}_1,\boldsymbol{y}_1),(\boldsymbol{x}_2,\boldsymbol{y}_2),...(\boldsymbol{x}_n,\boldsymbol{y}_n)\\ \boldsymbol{x}_i\in R^d,\boldsymbol{y}_i\in(+1,-1)\end{gathered} \tag{5-3}$$

及无标记数据集 $\boldsymbol{x}_1^{'},\boldsymbol{x}_2^{'},...\boldsymbol{x}_k^{'}$,求其相应的标记 $\boldsymbol{y}_1^{'},\boldsymbol{y}_2^{'},...\boldsymbol{y}_k^{'}$ (5-4)

在一般情况下,该 TSVM 可用如下数学公式描述:

$$\begin{aligned}&Min(\boldsymbol{y}_1^{'},\boldsymbol{y}_2^{'},...\boldsymbol{y}_k^{'},w,b,\boldsymbol{\xi}_1,...\boldsymbol{\xi}_n,\boldsymbol{\xi}_1^{*},...\boldsymbol{\xi}_k^{*})\\ &=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}\boldsymbol{\xi}_i+D\sum_{i=1}^{n}\boldsymbol{\xi}_i^{*}\end{aligned} \tag{5-5}$$

$$\begin{aligned}&S.t\quad \boldsymbol{y}_i[w.\boldsymbol{x}_i+b]\geq 1-\boldsymbol{\xi}_i\quad i=1,2,...n\\ &\boldsymbol{y}_i^{*}[w.\boldsymbol{x}_i^{*}+b]\geq 1-\boldsymbol{\xi}_i^{*}\quad j=1,2,...k\\ &\boldsymbol{\xi}_i\geq 0,\quad \boldsymbol{\xi}_i^{*}\geq 0\end{aligned} \tag{5-6}$$

式中,C,D 为参数,其中 D 为影响因子.这个算法比单一使用有标记样本训练的分类器有较大的性能改进,而其主要缺陷是须要事先估计训练正负样本分布比率和正样本数.不正确的估计将导致较差的结果.对这个问题.陈毅松等人提出了一种改进算法 PTSVM [192].该算法通过成对标注和标记重置的办法改进了 TSVM 的性能,但只适合于无标记样本较少的情况:无标记样本较多时,这种频繁的标记与标记重置将导致算法的复杂性迅速增加.并且远超过一般的 TSVM 算法.现实应用的大多数情况是无标记样本远多于标记样本,因而需要开发适应于此类情况的相应算法. 受文献[142,143]的共同启发,针对 TSVM 学习需要事先精确估计正负样本分布比率的难题及 PTSVM 算法须要经过多次重复计算和训练,因而导致计算代价太大的问题,本节提出一种直接基于支持向量机的渐近式半监督式学习算法.

5.4 基于支持向量机的渐近式半监督式学习算法

对比式(5-2)与(5-6)中所表示SVM与TSVM的目标函数,不难发现,二者只有一项不同,即TSVM的目标函数多了一项,影响因子项,这一项的关键作用,实际上是利用新标记样本去影响新决策面的动态调整方向.以便使目标函数朝着最小化方向变化.本节下面将要说明的是同样的效果也可以直接用SVM通过渐近式算法实现. 本节将此算法称为”基于SVM的渐近式半监督式算法—GSSVM”,这个算法成功的关键是根据误标记最少原则来确定适当的取样阈值.它通过每次选取等量正负样本数的办法避开了TSVM算法中需要事先准确估计样本分布比率的难题,同时又以批量标记新样本的方法克服了成对标记样本引起的大量重复计算问题.其基本思路是:用基本SVM进行半监督式学习时, 用渐近式方法把每次新标记的样本分布都控制在上一次训练集附近的一个不太大(合适的)邻域内,以确保在尽可能少的误标记前提下选择尽可能多的待标记样本.这一方面保证了新的训练集沿着正确的方向逐步扩大范围,最终达到正确包含整个无标记样本集的目标.另方面也可以有效地地减少重复计算的次数,从而使效率大大提高. 减少每次的误标记非常重要,因为错误的标记会产生积累效应,从而引导后面的学习走向错误的方向.而且为纠正它将使学习代价增大.为了确保每次的误标记尽可能少及减少重新标记样本数,需要根据原始标记数据集与无标记数据集空间分布差异的特点来确定相应的取样范围 ,以控制某一特定范围以内被选择的样本数量的方法来控制误标记数.即按下式

$$\boldsymbol{y}_i^* = \left| f(\boldsymbol{x}_i^*) \right| \geq \theta \tag{5-7}$$

这里$\mathbf{y}_i^*$为新选定的样本标记,并且

$$\boldsymbol{y}_+^* = f(\boldsymbol{x}_i^*) \geq \theta \tag{5-8}$$

$$\boldsymbol{y}_-^* = f(\boldsymbol{x}_i^*) \leq -\theta \tag{5-9}$$

式中$\mathbf{y}_+^*$, $\mathbf{y}_-^*$分别为正,负标记.选取欲标记的样本, (5-7)式中,$\theta \geq 0$,是一个用于调节选择范围的阈值,它通常须要根据有标记样本和无标记样本分布差异的实际情况确定.为了说明θ值是如何通过对取样样本数量的影响进而导致对决策面调整方向的影响.下面不妨对(5-6) 式作一个分析: 当用上次训练集学习一个决策面并以此为依据来选择下次新增的标记样本以扩大训练集时,须满足:

$$\boldsymbol{y}_i^*[w.\boldsymbol{x}_i^* + b] \geq 1 - \boldsymbol{\xi}_i^* \quad i = 1,2,...k$$

$$\xi_i^* \geq 0 \qquad (5\text{-}10)$$

为得到最优分割结果,在一般情况下, 对于所有新增训练样本 $f(x_i^*)$ 应满足

$$\left|f(\boldsymbol{x}_i^*)\right| \geq 1-\xi_i^* \qquad (5\text{-}11)$$

根据 $f(\mathbf{x}_i^*)$ 的符号,即可初步预测相应的类标记符号 $\mathbf{y}_i^*$. 当取样规则选择使 θ 满足

$$\left|f(\boldsymbol{x}_i^*)\right| \geq \theta \geq 1-\xi_i^* \qquad (5\text{-}12)$$

上式左边取等号则有:

$$\left|f(\boldsymbol{x}_i^*)\right| = \theta \geq 1-\xi_i^* \qquad (5\text{-}13)$$

据 $q, x_i^* \ ^3\ 0$ 和(5-16)式可知, q, x_i^* 二者变化趋势是相反的.

$$\xi_i^* \geq 1-\theta \qquad (5\text{-}14)$$

上式说明,当 q 取值变大时, 相当于 ξ_i 取值小的情况, 由(5-2)(5-7)式可知,这将导致目标函数值也减小.而 q 取值变小时情况则相反.故调节 q 值实际上相当于调节 ξ_i ,或者说, 调节了所取样本分布与决策超平面间的距离.也就调节了实际取样的数量.当 q 值变大时, 每次所取的新样本分布在距上次训练的决策面距离更远的区域内.或者说所取新样本分布在与上一次训练集更接近的区域内(实际取样数量也就减少了).而新的决策面的的调整方向主要受上次训练结果和新加入的部分样本共同影响.故调节 q 事实上也将影响新决策面的调整方向.由下面（5-15）式

$$w.\boldsymbol{x}_i^* + b \geq \theta \geq 1-\xi_i^* \qquad (5\text{-}15)$$

可得: $$\boldsymbol{x}_i^* \geq \frac{\theta-b}{w} \geq \frac{1-\xi_i^*-b}{w} \qquad (5\text{-}16)$$

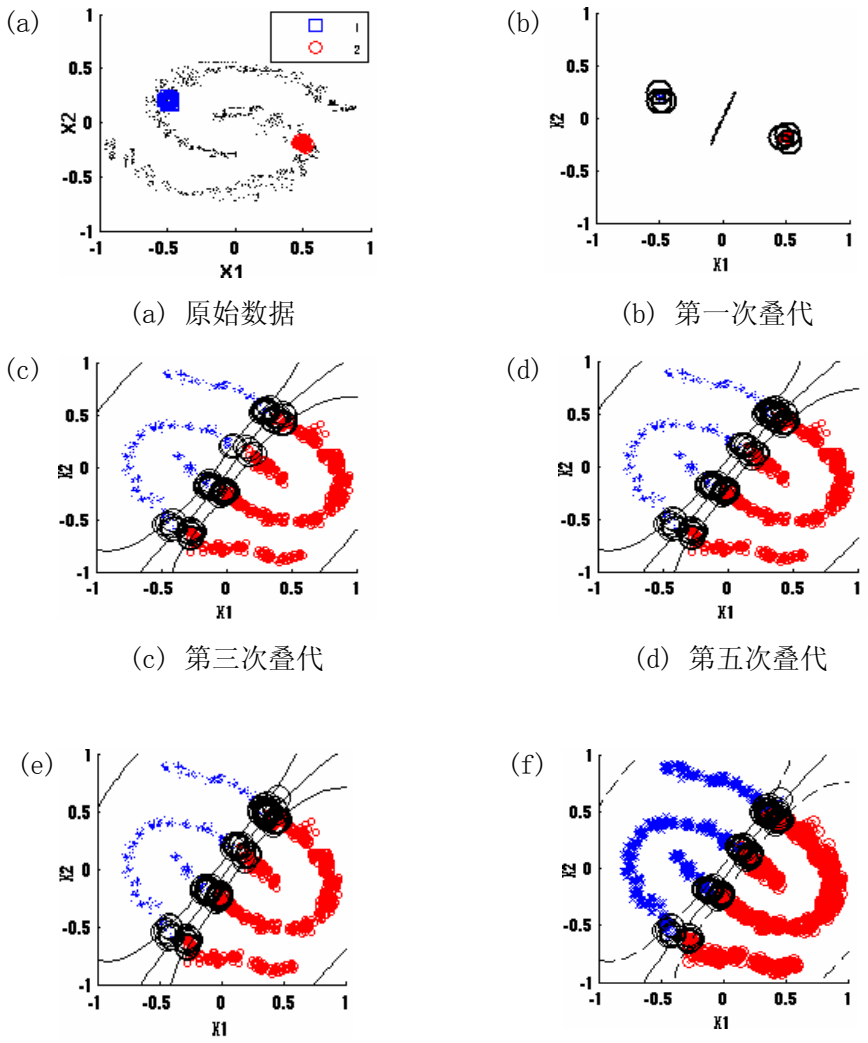
上式说明,当按(5-7)式来选择新样本时,实际上只选取了分布特征满足

$$\boldsymbol{x}_i^* \geq \frac{1-\xi_i^*-b}{w} \qquad (\theta = 1-\xi_i^*) \qquad (5\text{-}17)$$

区域的数据点. 这在一定程度上减少了误标记的可能性. 确保了训练能够逐渐地逼近真实目标.另外,当核函数为 RBF 类型时,选择不同的参数 σ 对训练结果也有较大的影响.通常, σ 取值小,则容易得到正确的学习结果,但同时学习时间也变长,而 σ 取值变大,虽能节省学习时间,但学习误差也相应变大甚至得到错误的结果.通常,这个参数可以用交叉确认

方法来确定.

图 5-2, 5-3 的(a)-(e)对应子图显示了不同θ, σ取值时的学习结果比较. 其中, 图 5-2 是随机选择学习参数所产生的结果. 图 5-2 是按上面(5-8), (5-9)式选择学习参数所得的结果. 显然, 图 5-2 的结果表现了较大的学习误差. 较理想的学习结果可见下面图 5-3

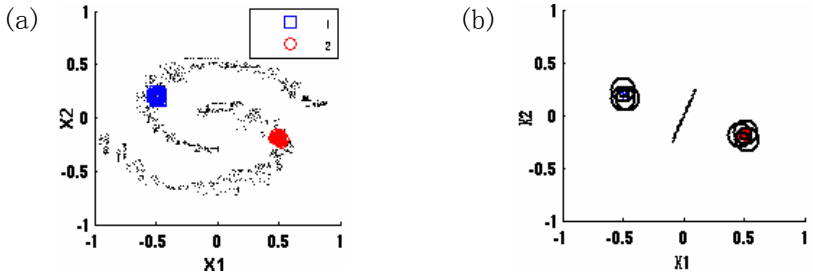


(a) 原始数据 (b) 第一次叠代

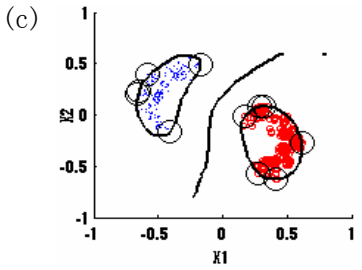
(c) 第三次叠代 (d) 第五次叠代

(e) 第七次叠代 (f)最终结果

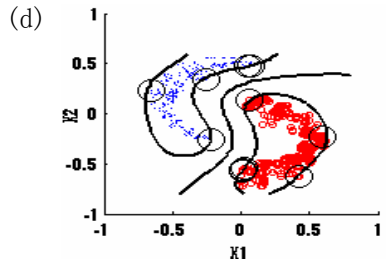
图 5-2 学习参数θ= 0.3; σ= 0.5 时的半监督式学习效果



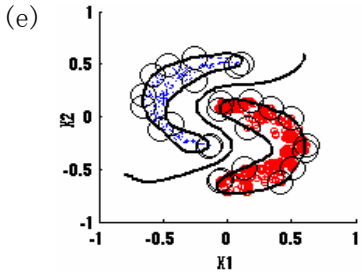
(a) 原始数据 (b) 第一次叠代



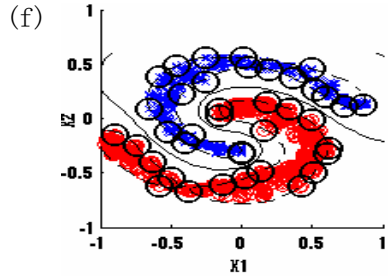
(c) 第三次叠代



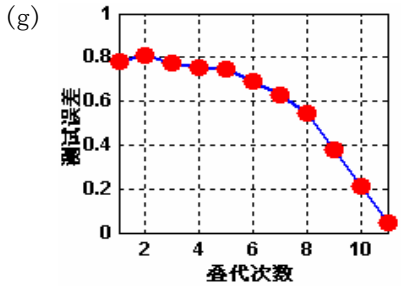
(d) 第五次叠代



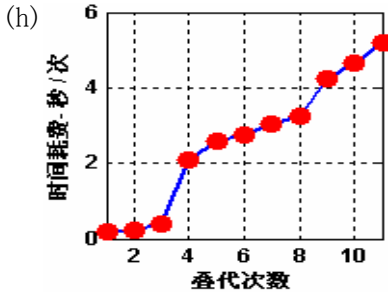
(e) 第七次叠代



(f)最终结果



(g) 训练误差



(h)时间耗费

图 5-3 学习参数θ= 0.43;σ= 0.2 时的半监督式学习效果

显然,从上面各对应子图的比较中,可以一目了然的发现,与图 5-2 相比,图 5-3 的各对应子图所得结果远远好于前者.尤其是图 5-3 的(f)子图,学习结果与原始数据几乎相同,这说明按(5-8),(5-9)式的规则选择学习参数能够确保学习的正确性.

进一步地,考虑到非支持向量对决策面的调整没有影响,为了提高学习速度,本节采取每次学习循环都删除上次训练集中的非支持向量的措施,以减少不必要的计算量.实验表明,这一改进确实能显著提高学习效率.下面图 5-4 就明显反映了这种效果.

综上所述. 新算法可以归纳为以下步骤:

步骤 1: 选择合适的核函数类型及相关参数,原始标记样本训练初始分类器.

步骤 2. 以初始分类器预测各无标记样本的判决函数值, 根据样本分布特点确定适当的 q 值,从预选集中选取最大可能的等量正负样本合并到训练样本集中.若预选集

中最后只剩下单一标记(全正或全负)的样本,则将其全部加入到训练集中.

步骤 3: 用新构成的训练样本集重新训练分类器.并删除训练集中的全部非支持向量.

步骤 4: 若无标记样本集中样本未被移空,则一直重复步骤 2--步骤 3.否则结束此轮训练.若结果满足预定要求,则直接输出结果,否则进入步骤 5.

步骤 5: 将 q 递增到一个较大的值,重复前面步骤 2-4,结束时比较两轮的结果,以优的那一轮为依据.沿着(或逆着)此方向调 (大/小) q 值,再重复 2-4,比较并选择满意的结果,输出结果.

这个算法的计算复杂性可以这样来估计:由于普通 SVM 的时间复杂度可用下式表示.Ts≈O(cL2),这里,L 为训练样本数,而 c 为算法相关常数. 由于这里讨论的算法本质上同为二次规划问题,故可以用相似的公式来估计;本算法可为 Tss=O(kcL2),k 为对 L 的选择性批量取样次数).对 TSVM 算法而言,它的一次训练相当于 n1*n2 次 SVM[145],故为 O(n1n2c*L2).其中,n1,n2 分别为 TSVM 的内,外循环次数 基于 TSVM 改进型的 PTSVM 成对标记法则可表为 Tp=O(n1n2c*L2/2).与本算法相比较有: Tp / Tss =O(n1.n2c*L2/2) / O(2kc*L2) 因 2k<<n1.n2 故本算法计算复杂性低于 PTSVM.再考虑到本算法每次循环都删除了训练集中占比例超过 90%以上的非支持向量,这更使计算代价显著低于 PTSVM[193].

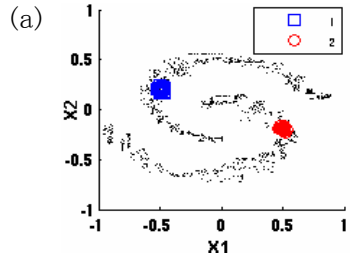
5.5 实验分析:

本节给出相关的实验,它证实了本算法的可行性和有效性,为了说明它对不同分布的适应情况,实验首先采用一个具有复杂分布的人工合成数据集, 其特点是:原始有标记数据样本 40 个,无标记数据样本 1080 个,其样本分布如 图 1 所示.核函数采用 RBF 类,C 取值为 10.实验的一些其它相关内容详见表 5-1.

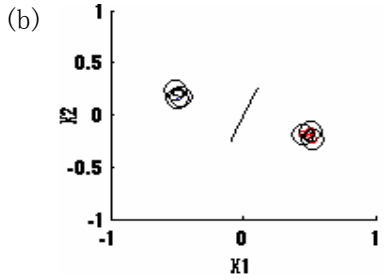
表 5-1 ϴ, σ 不同取值对时间(T)和测试误差(TE)的影响

编号	ϴ	σ	T(s)	TE
1	0.2	0.2	5.8	17%
2	0.3	0.2	10	9.7%
3	0.4	0.2	8	26%
4	0.5	0.2	8.3	<0.89%
5	0.6	0.2	8.8	< 0.5%
6	0.7	0.2	12	< 0.3%
7	0.8	0.2	19	< 0.1%
8	**0.43**	**0.2**	**4.7**	**<0.65%**
9	0.53	0.2	6.2	<0.55%
10	0.53	0.3	11.9	<0.65%

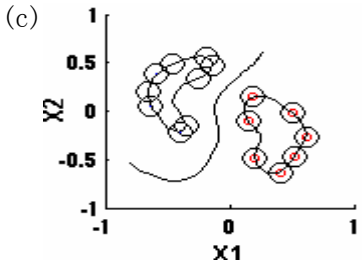
实验表明:如果只是要求学习器得到误差允许范围内的结果(即允许次优解),则对q,σ值的选择要求并不高,只需在合适的取值区间任取一值即可,例如本例中,取值范围是0.42<q<0.7; 0.1.8<σ<0.25.从这个意义上说,如果参数选择合适(这一般不难),上面算法只需前面 4 个步骤经一轮学习即可实现.当然,若要得到最优结果,则需进入步 5,经过多轮学习比较后才能实现.



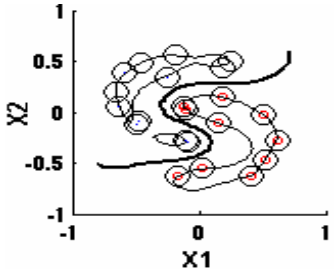
(a) 原始数据



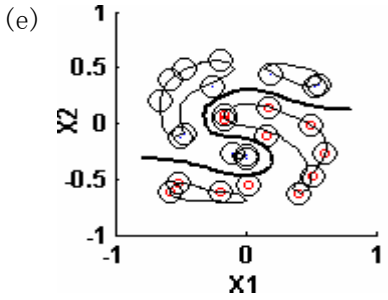
(b) 第一次叠代



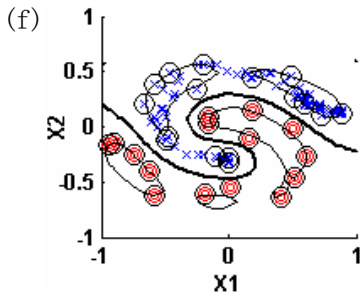
(c) 第三次叠代



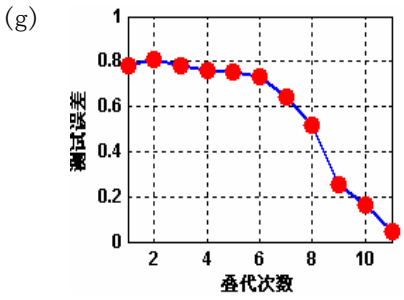
(d) 第五次叠代



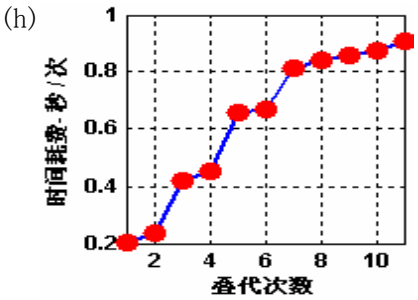
(e) 第七次叠代



(f)最终结果



(g) 训练误差



(h)时间耗费

图 5-4 基于 SVM 的渐进式半监督式学习算法

图5-4是实验的详细情况.其中,粗大醒目的样本点为有标记点,细小的为无标记点. 图5-4 (a)—(c) 展示的是此算法在验证实验过程中随着新标记样本增加而引起的决策面动态调整的情况.图5-4中决策面的变化情况也正好说明了这一点.图5-4(g)所示的的误差曲线显示了最终的学习精度(在第11次叠代时) 误差接近于0 的的过程.

与前面的方法相比,图5-4表明,基于SVM的渐进式半监督式学习算法的优势在于,它既能达到与图5-3同样的学习精度,同时还大大减少了计算代价(学习中所用的训练数据远少于普通算法,这只需比较两图各相应子图的各次叠代图中的数据点数量便可得出结论),从而大幅提高了运算效率.例如,对同一数据的学习,图 5-3 (g)显示,达到误差接近于零的叠代需要11次,且运行时间为5秒; 而图5-4 (g)显示,当学习误差接近于零时,叠代次数虽然相同,但运行时间只有0.9秒,二者相差5倍以上.

表5-2 真实数据集的测试结果

数据集名	训练集	测试集	GSSVM误差%	CSVM误差 %
tulip	25	75	8.5	6.1
riply	250	1000	14.5	13.2
bupa	20	356	9.9	7.1
pima	80	700	28	24.5

UCI 数据集仿真: 为了说明它在处理真实数据时的效果,本节进一步用几个UCI 低维数据集仿真实验,其结果也证实了本算法的有效性.表5-2是20次平均的结果:与标准的监督式学习支持向量机 (CSVM算法)相比,尽管误差稍大些,但作为半监督式学习 (GSSVM算法)能够达到这种效果已经是很不错了.

5.6 交互式半监督聚类

如前述,标准的无监督式学习的谱聚类在许多复杂的情况下,由于缺少先验知识的指导不稳定甚至不可靠.通过引入先验知识,半监督式学习能较好克服这些问题.

针对谱聚类算法难以处理无明显稠密区数据及对噪音敏感的缺陷,本节提出交互式半监督聚类算法,这是一种基于文献[197]半监督式聚类的改进算法.不同之处在于,它采用交互式方法输入有标记数据点,采用半监督式学习方法来改进聚类,并利用标记数据点的种类数自动确定聚类数(这免去了由于自动确定最佳聚类数的计算代价),通过增加平均相似系数来调节抗噪能力.将其用于标准数据及图像分割的实验表明,与标准谱聚类相比,此法能够获得更好的应用效果.

本节提出的交互式半监督式聚类算法 ISC,主要特点是充分利用少量有标记数据的先验信息来指导学习过程.为了保证先验信息的最佳性,本文采用了交互式输入有标记数据的方式来获取数据.

1) 交互式输入：

交互式是指根据原始数据分布的实际情况,用鼠标直接从屏幕上灵活地选择标记数据点的位置,数量和类型,通过屏幕直接获取部分有标记数据点的方法.这能确保先验信息的最佳和有效性.在某些领域（如医疗图像分割或检测）的应用中尤其有效.

2) 半监督式学习

半监督式学习的主要内容是预测无标记点的标记.设有 N 维输入空间数据集 $\boldsymbol{X}=\{\boldsymbol{x}_1,...\boldsymbol{x}_l,\boldsymbol{x}_{l+1},...\boldsymbol{x}_n\}\subset\boldsymbol{R}^N$,其中,标记数据集 $L=\{1,...c\}$, x_l（$1\le l\le c$）, $y_l\in L$ 为有标记数据点. x_u（$l+1\le u\le n$）为无标记数据点.

若用 $\boldsymbol{F}$ 表示 $n\times c$ 维非负矩阵. $\boldsymbol{F}=[\boldsymbol{F}_1^T,...\boldsymbol{F}_n^T]^T\in\boldsymbol{F}$ 与标记为 $y_i=\arg\max_{j\le c}F_{ij}$ 的有标记数据点 x_i 的数据集 $\boldsymbol{X}$ 相对应.可以把 F 理解为矢量函数 $\boldsymbol{F}$： $\boldsymbol{X}\to\boldsymbol{R}^c$,为每一个点 x_i 指定一个矢量 $\boldsymbol{F}_i$.若 x_i 被标记为 $y_i=j$ 则定义一个 $Y_{ij}=1$ 的 $n\times c$ 维矩阵 $\boldsymbol{Y}\in\boldsymbol{F}$;否则 $\mathbf{Y}_{ij}=0$.

实现半监督式学习的基本步骤如下[195]：

(1) 计算亲和力矩阵:

$$w_{ij}^{'}\equiv w(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp(-\left\|\boldsymbol{x}_i-\boldsymbol{x}_j\right\|^2\Big/2\sigma^2),\forall i\ne j \qquad (5\text{-}18)$$

与前面相似,定义 $w_{ii}^{'}=0$

(2) 为特征空间规范化,构造矩阵：

$$\boldsymbol{D}(i,i)=\sum_j w_{ij}^{'} \qquad (5\text{-}19)$$

$$\boldsymbol{S}=\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2} \qquad (5\text{-}20)$$

(3) 多次重复计算下面函数,直到收敛.

$$\boldsymbol{F}(t+1)=\alpha\boldsymbol{S}\boldsymbol{F}(t)+(1-\alpha)\boldsymbol{Y} \qquad (5\text{-}21)$$

式中,α 为参数（$0<\alpha<1$）,重复上面步骤到收敛.令 F^* 代表序列$\{F(t)\}$的极限,对各点 x_i 的估计可用下式计算： $y_i=\arg\max_{j\le c}F_{ij}^*$ 根据[195], $\boldsymbol{F}^*$ 能够直接用下式计算而无需叠代.

$$\boldsymbol{F}^*=\lim_{t\to\infty}\boldsymbol{F}(t)=(1-\alpha)(\boldsymbol{I}-\alpha\boldsymbol{S})^{-1}\boldsymbol{Y}, \qquad (5\text{-}22)$$

(4) 在此基础上实现聚类的其它步骤.

3) 本文改进方法

上面算法只适合于理想情况.在数据含有噪音时,实际效果将大打折扣.甚至得出错误的

结果.为克服此缺陷本文提出一个改进方案.考虑到噪音或局外数据在整个数据集中通常只占极少比例.可以采用均值法加以抑制.由（5-18）式,考虑 x_i 的 N 个最近邻点, $A_i^{'}$ 为点 x_i 的联结权,

$$A_i^{'} = \sum_{x_j \in N_i} w_{ij}^{'} \qquad (5\text{-}23)$$

x_i 与其 N 个最近邻的平均相似度

$$\overline{w_i} = A_i^{'} / N \qquad (5\text{-}24)$$

显然 $0 < \overline{w_i} \le 1$

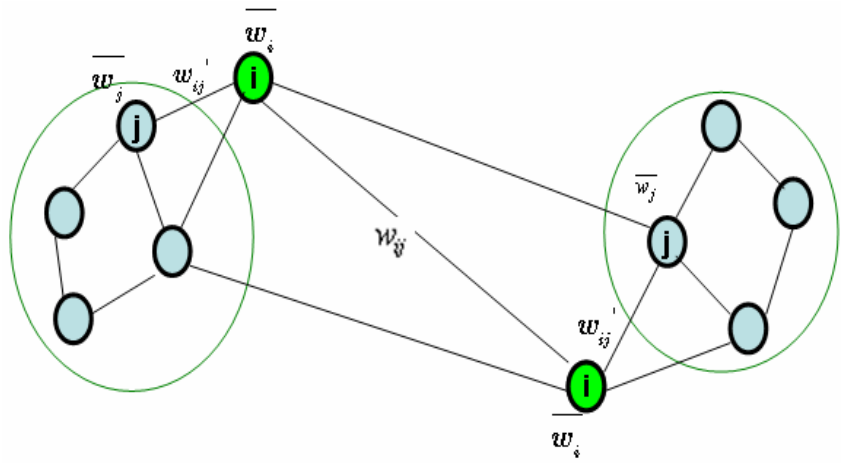


图5-5 . 抑制噪音或局外数据的方法

如上图所示,如果设定两数据点的相似测量项为下式：

$$w_{ij} = \overline{w_i} \cdot \overline{w_j} \cdot w_{ij}^{'} \qquad (5\text{-}25)$$

则从图中可看出：仅当 $\overline{w_i}, \overline{w_j}, w_{ij}^{'}$ 之积大于某阈值时, x_i, x_j 才可能相似且位于同一聚类区中,只要三者中任何一项有较低值,则 x_i, x_j 必然位于不同聚类或者其中之一是噪音或局外值.于是借此可滤除噪音或局外数据.从而改进聚类性能.

将（5-25）式中的 w_{ij} 代替(5-19)式中的 $w_{ij}^{'}$,并解(5-19)~(5-22)即为改进的半监督式聚类.

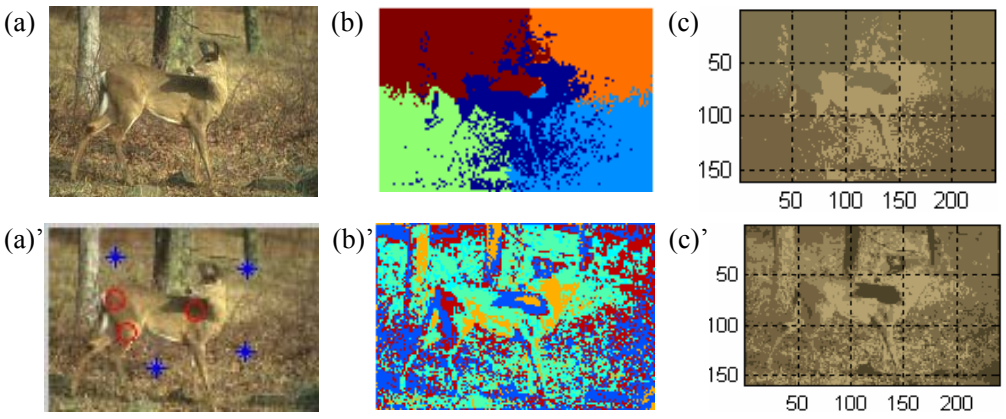
概括上述各步,实现交互式半监督聚类的基本过程可简述如下：

在对待处理数据实施聚类时,首先为其加入带有先验知识的数据分量(例如直接从屏幕上交互式提取部分有标记数据点),再通过半监督式学习获取其它相应未标记点的标记,并将这些标记作为特征分量加入到原数据集来实施谱聚类.从而达到改进性能的效果.

5.7. 实验分析:

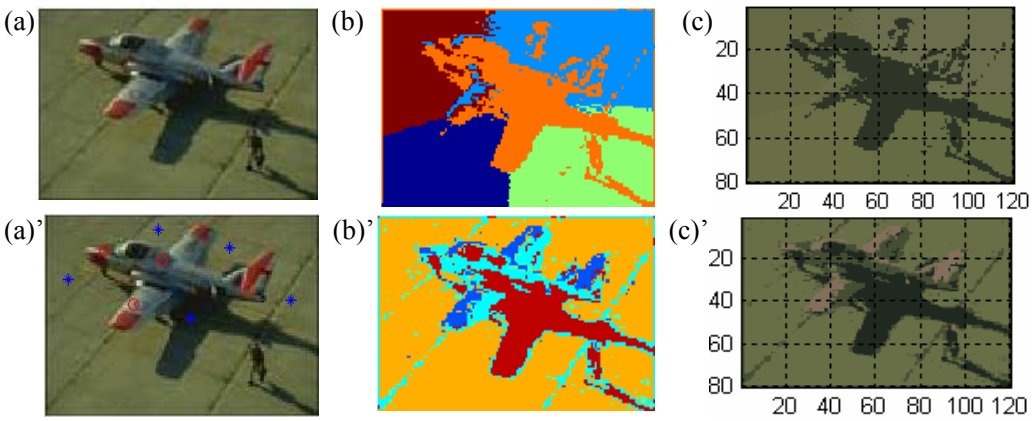
为了验证本算法的有效性,下面采用来自Berkeley图像数据库[180]的两种数据集作实验,将其与普通的无监督式谱聚类方法的同类运用效果进行对比, 结果如下:

图5-6, 图5-7是交互式半监督聚类与普通谱聚类的应用于图像分割(数据名236017.jpg,37073.jpg)的效果比较.其中,图5-6.(a), 图5-7.(a)是原始图像; 图5-6.(a)', 图5-7.(a)' 中的红圈和兰点表示交互式半监督聚类算法人为交互地加入标记点; 图5-6.(b),(b)图5-7.(b),(b)'是原始图对应的分割聚类; 图5-6.(c),(c)', 图5-7.(c),(c)' 则分别是两种方法对应的图像分割结果.比较二者可见, 交互式半监督聚类的图像分割效果明显好于单纯的无监督式谱聚类.



(a) ,(a)' 原始图像 (b), (b)'分割聚类 (c) (c)' 图像分割

图 5-6, 交互式半监督学习用于图像分析 1



(a) ,(a)' 原始图像 (b), (b)'分割聚类 (c) (c)' 图像分割

图 5-7 , 交互式半监督学习用于图像分析 2

5.8 小结:

本节针对多维数据如图像特征检测的聚类模型存在的缺陷,提出相应的改进方案.其中,基于支持向量机的渐近式半监督式学习算法,采用的特定取样规则和核参数可以确保减少误标记数量并控制决策面的动态调节进程,通过删除非支持向量来提高训练速度.实验表明,这种算法能够适应不同的样本分布情况,并取得较好的效果,是一种值得注意的新尝试.当然,此算法还需进一步完善, 例如核函数类型及核参数值对决策面动态调节进程的影响方式和规律,以及如何将本学习算法与主动式学习结合起来以解决更复杂问题的方案等还有待进一步研究.

另外,交互式半监督式算法,通过屏幕交互提供特定先验知识,并用改进相似系数抑制噪声,用相应半监督式学习算法学习其它数据的未知标记,进而将标记完整的特征分量加入到原始数据集中的方法,经实验证明:的确可改进聚类性能,取得更好的应用效果.

第六章 结论与展望

6.1 本论文工作小结

本文主要探讨了如何以基于机器学习及相关途径相结合的方式,利用有限系统资源高效实现实时特征检测的相关理论与技术问题. 本文着重针对时序和图像特征(尤其是异常特征)检测方面的问题,提出了一些相应的算法.并验证了其有效性.主要研究成果包括:

(1).在研究分析 OCSVM 及 PSO 模型特点和综合前人研究成果基础上,提出 OCSVM_CPSO 组合式异常检测模型, 克服了传统 OCSVM 不能处理时序数据的缺陷,实现了系统的自适应调节,解决了检测系统的在线运行问题,从而为其现实应用扫除了障碍.并将其用于解决机器人传感器故障检测的实际问题.取得较好效果.

(2).针对 SAX 模型容易丢失边界区信息问题,提出时序数据的 DLS 模型.它根据时序极值确定划分的上下边界,并根据最大熵确定最佳描述字符集,进而确定最佳划分间隔.从而能有效减少边界区的信息丢失;针对EXT_SAX模型缺陷,提出VSB模型,采用增加分量而非增加符号的途径来降低计算代价.且用实验证实它的有效性.

(3).提出时序矢量符号的 SFVS 模型和相应的确定时序数据最大压缩比的方法,此模型能够比 SAX 提供更全面的描述信息, 这有利于在时序特征检测的应用中实现更精确的分析.通过理论分析和实验证实了它的有效性.

(4).针对自调节谱聚类缺陷, 提出一种新的 ASC(自适应谱聚类)算法.它用全局平均 N 近邻距离作为比例参数 σ,利用本征矢差异来估计最佳聚类分组数 k, 这可在构造亲和力矩阵时减少计算代价,提高效率,并且更容易实现.在彩色图像检测与分割实际应用中的实验结果证实了其有效性.在此基础上提出改进聚类性能的相关半监督学习算法,且用实验证实了它的有效性.

6.2 进一步研究工作

本文以时序和图像特征检测为中心内容,着重研究了时序数据的分类,相似查询和异常检测,以及作为彩色图像检测基础的图像分割.这两类待识对象的特征检测都具有计算代价大,噪音,干扰多, 难以实现实时性动态检测的问题.尽管已经出现了许多相应的解决方案,然而,现实应用的需求日新月异,有待探索更加高效的解决方案.

因此作为本文的后续工作,将从以下方面展开:

(1)探索更好的特征检测建模方式和有效的实施技术.使之能更有效地解决分类,相似查询,

异常检测等问题.

（2）探索能够适应不同时序数据的符号化新技术.使之能够处理动态时序数据的实时性分析和检测问题.

(3)探索能够用监督式,半监督式和无监督式多种学习方式实施训练的特征检测算法,以适应复杂环境应用的现实要求.

参 考 文 献

[1] B.Dasarthy, Nearest neighbor pattern classification techniques, IEEE Computer Society Press. Los Alamitos, CA, USA, 1991:276-284

[2] E.Shortliffe, MYCIN: Computer-Based Medical Consultations, Elsevier, New York,Publisher: Elsevier May 1976: 264

[3] Paul Harmon and Rex Maus and William Morrissey, Expert Systems: Tools and Applications, John Wiley and Sons, New York. US. 1988.289

[4] J.de Kleer and B. C. Williams, Diagnosing multiple faults, Readings in Nonmonotonic Reasoning, Artificial Intelligence, 1987,32: 97–130,

[5] M. Ge, R. Du, G. Zhang and Y. Xu, “Fault diagnosis using Support vector machines with an application in sheet metal stamping operations”, Mechanical systems and signal processing, 2004,18:143-159.

[6].R.J.Patton,J.Chen.Robust Model-based Fault Diagnosis for Dynamic Systems, Kluwer Academic Publishers,London,1999, 23

[7] A.Drake,Observation of a Markov process through a noisy channel,Ph.D. thesis. MIT, 1962. 5

[8] D.Dvorak and B.J. Kuipers, Model-based monitoring of dynamic systems, IJCAI,1989, 1324-1330.

[9] W.Hamscher, Modeling Digital Circuits for Troubleshooting, AI, 1991, 51(1-3): 223-271

[10] A.Doucet and N.de Freitas and N.Gordon,Sequential Monte Carlo Methods in Practice,Springer Verlag,New York. Springer Verlag, 2001,70: 197-217.

[11] Monahan, G.E."A survey of partially observable Markov decision processes: theory, models and algorithms", Management Science, 1982, 28:.1-16

[12] J.Kurien and P.Nayak,Back to the Future with Consistency based Trajectory Tracking, AAAI 2000, 370-377

[13] 刘丹红,张世英. 基于小波神经网络的非线性误差校正模型及其预测. 控制与决策.2006,21(10):1114-1118.

[14] 冯力, 孙杰, 等. 基于 Windows Native API 序列的异常检测模型.西安交通大学学报.2006,41(4):407-412.

[15] 谭小彬,奚宏生,王卫平,殷保群.基于支持向量机的异常检测.中国科学技术大学学报.2003,10. 33(5): 599-604.

[16] 杨奎河 1,2 单甘霖 赵玲玲.基于支持向量机的旋转机械故障诊断研究.微计

算机信息(测控自动化)2006,22(12-1): 184-185.

[17].G.A.Carpenter,M.A.Rubinand,W.W.Streilein,“ARTMAP-FD:familiarity discrimi -nation applied to radar target recognition”, Proc. International Conference on Neural Networks, 1997, 3:1459-1464,

[18] L. Tarassenko, A. Nairac, N. Townsend and P. Cowley, “Novelty detection in jet engines”,IEE Colloquium on Condition Monitoring, Imagery, External Structures and Health, 1999,4: 41-45,

[19].D. M. J. Tax and R.P.W. Duin, “Outlier detection using classifier instability”, In Advances in Pattern Recognition, the Joint IAPR International Workshops, 1998, 593-601,

[20].C. Manikopoulos and S. Papavassiliou, “Network intrusion and fault detection: a statistical anomaly approach”, accepted for publication in IEEE Communications Magazine, October 2002.

[21].R.S. Guh, F Zorriassatine and J.D.T. Tannock, “On-line control chart pattern detection and discrimination -a neural network approach”, Artificial Intelligence in Engineering, 1999,13:413-425,

[22] Lunt T F．Tamaru A．Gilham F．A real—time intrusion detection expert system (IDES)-Final technical report；Computer Science Laboratory.SRI 2001.

[23] Anderson D．Anderson D．et a1．Detecting unusual program behavior using the statistical component of the next—generation intrusion detection expert system(NIDES)；Computer Science Laboratory．SRI 2002

[24] chen K.An inductive engine for the acquisition of temporal knowledge：[PH.D Thesis].Department of computer science. University of I1linoist Urbana champaign 2007

[25] P.Szolovits and S.G.Pauker,Categorical and probabilistic reasoning in medical diagnosis revisited,Artificial Intelligence. 1993,59:167-180 ,

[26] U.Lerner and R. Parr and D. Koller and G. Biswas, Bayesian Fault Detection and Diagnosis in Dynamic Systems, Proceedings of the 17th National Conference on Artificial Intelligence, 2000 (AAAI-00), 658-664

[27] J.Gordon and E.H.Shortliffe,A Method for Managing Evidential Reasoning in a Hierarchical Hypothesis Space,Artificial Intelligence, 1985 26 (3): 323-357

[28] R.Kalman, A new approach to linear filtering and prediction problems, Journal of Basic Engineering, 1960, 82: 34-45

[29].李尔国,俞金寿．一种基于输入训练神经网络的非线性 PCA 故障诊断方法

[J]．控制与决策,2003,18(2)：229—232

[30] R.Washington, On-Board Real-Time State and Fault Identification for Rovers, Proceedings of the 2000 IEEE International Conference on Robotics and Automation. 2000.129-132

[31] U. Lerner and R. Parr and D. Koller and G. Biswas, Bayesian Fault Detection and Diagnosis in Dynamic Systems, Proceedings of the 17th National Conference on Artificial Intelligence, 2000

[32] M. Hofbaur and B. Williams, Mode estimation of probabilistic hybrid systems, Hybrid Systems: Computation and Control, 2002 LNCS 2289 : 253–266

[33] J.L.Feranández, Supervision, detection, diagnosis and exception recovery in autonomous mobile robots, University of Vigo, March 2000

[34] Lee W ．Stolfo S J,Data mining approaches for intrusion detection． In：Proc．of the 7 USENIX Security Symposium ． San Antonio, TX．Jan．1998

[35] Lee W,Stolfo S J.Mok K W.Mining audit data to build intrusion detection models.In：Proc of the 4 Int1.Conf.on Knowledge Discovery and Data Mining. New York,NY．AAAI Press．1998

[36] 郑建国,焦李成.偏差检测挖掘方法研究[J] .计算机工程, 2001,27(8) : 33- 35.

[37] 史东辉, 张春阳, 蔡庆生. 离群数据的挖掘方法研究[J] ． 小型微型计算机系

统, 2001 , 22 (10) : 234 -236.

[38] 史东辉, 蔡庆生, 倪志伟, 等. 基于规则的分类数据离群挖掘方法[J] ． 计算机研究与发展, 2000 , 37(9) : 109-110.

[39] 姜灵敏. 基于相似系数和检测孤立点的聚类算法[J] ． 计算机工程, 2003, 29(11) : 183-185.

[40] 陆声链, 林士敏. 基于距离的孤立点检测及其应用[J] ． 计算机与数字工程, 2003, 12(22) : 31 -33.

[41] Stephen Marsland, Ulrich Nehmzow and Jonathan Shapiro, On-line Novelty Detection for Autonomous Mobile Robots, J. Robotics and Autonomous Systems, 2005, 51:191-206,

[42] Fox K L．Henning R R,et a1．A neural network approach towards intrusion detection．In：proc．of the 1 3‘ national computer security conference ;1990.

[43] 肖健华,吴今培,孙德山. 基于 SVR 的异常数据检测. 计算机工程与应用.2003,26:23-28

[44].Scholkopf,B,Platt,J.Schawe-Taylor,J.Smola,A.J.and WilliamsonR.C. Estimating

the Support of a High-Dimensional Distribution. Technical Report 2001. 99-87

[45]周小平, 晏蒲柳, 吴 静. 基于支持向量机的网络故障在线诊断方法研究. 武汉大学学报(工学版).2006,39(3):102-106.

[46] 钱权,耿焕同,王煦法. 基于 SVM 的入侵检测系统. 计算机工程.2006,32(9):136-139.

[47]周红刚, 杨春德. 基于免疫算法与支持向量机的异常检测方法. 计算机应用.2006,26(9):2145-2147.

[48] 魏藜, 宫学庆,钱卫宁,等. 高维空间中的离群点发现[J].软件学报,2002,13(2):280- 290.

[49].杨敏,张焕国,等. 基于支持向量数据描述的异常检测方法.计算机工程.2005,31(3).: 39-42.

[50]. K. Heller, K. Svore, A. Keromytis, and S. Stolfo."One Class Support Vector Machines for Detecting Anomalous Windows Registry Accesses". In Proceedings of the ICDM Workshop on Data Mining for Computer Security (DMSEC), Melbourne, Florida, November 19, 2003.

[51].张家凡,黄之初,王小明. 基于支持向量异常检测算法的新故障检测. 武汉理工大学学报. 2006, 28(12):109-112.

[52]. 王婉湘,一种基于一类支持向量机的时序异常检测算法. 微型机与应用.2005,1:23-27

[53] T. Hastie, S. Rosset, R. Tishbirani, and J. Zhu. The entire regularization path for support vector machines. J. Machine Learning Research, 2004,5:1391-1415,

[54] M. Davy, F. Desobry, A. Gretton and C. Doncarli, "An Online Support Vector Machine for Abnormal Events Detection", Signal Processing, to appear, 2006, 86(8):72-73

[55] Forrest S,Perelson A,Allen L,et a1．Self-Nonself Discrimination in a Computer [A].Proc of the IEEE Symp on Research in Security and Privacy [C].1994,202-212.

[56] lshida Y,Adaehi N．Active Noise Control by an Immune Algorithm：Adaptation in Immune System as an Evolution[A]．Proc ICEC 96[c]．1996,150-153．

[57] Mori K,Tsukiyama M,Fukuda T．Immune Algorithm and Its Application to Factory Load Dispatching Planning[A]．Proc of Japan USA Symp on Flexible Automation[C]．1994,1343-1346．

[58] Mori K,Tsukiyama M,Fukuda T．Application of an Imune Algorithm to Multi-Optimization Problems[J]．Eleotrical Engineering,1998,122(2)：30-37．

[59] Tang Z,Yamaguchi T,Tashima K．Multiple-Valued Immune Network Model and Its Simualtions[A].Proc 27th lnt'1 Symp on Multiple Valued Logic[C]. 1997, 233-238．

[60] De Castro L N,yon Zuben F J．Artificial Immune Systems：Part I.Basic Theory and Applications[R]．Technical Repot-RT DCA 01／99,1999．

[61] A Doucet, N de Freitas, and N J Gordon, editors. Sequential Monte Carlo Methods in Practice. Springer-Verlag, 2001,6 .429-444

[62] S J Godsill, A Doucet, and M West. Maximum a posteriori sequence estimation using Monte Carlo particle filters. Ann. Inst. Stat. Math. March 2001. 53(1): 82–96

[63] H Shiqiang.JIN G Zhong—liang Overview of particle filter algorithm Control and Decision. Apr.2005,20 :4-6

[64] Rudolph van der Merwe, Arnaud Doucet, Nando de Freitas, Eric Wan The Unscented Particle Filter (2000) Advances in Neural Information Processing Systems 13.

[65] R. Dearden and D. Clancy, Particle Filters for Real-Time Fault Detection in Planetary Rovers, Twelfth International 2004.

[66] Geoffrey J. Gordon Bayesian Methods for Identifying Faults on Robots for Planetary Exploration ISBA 2004 - International Society for Bayesian Analysis.

[67] Rudolph van der Merwe, Arnaud Doucet, Nando de Freitas, Eric Wan The Unscented Particle Filter (2000) Advances in Neural Information Processing Systems 13

[68] R. Dearden and D. Clancy, Particle Filters for Real-Time Fault Detection in Planetary Rovers, Twelfth International Workshop on Principles of Diagnosis, DX-2002.

[69] V. Verma, S. Thrun, and R. Simmons. Variable resolution particle filter. In Georg Gottlob and Toby Walsh, editors, IJCAI-03, Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, Acapulco, Mexico, August 9-15, 2003, 976

[70] Geoffrey J. Gordon Bayesian Methods for Identifying Faults on Robots for Planetary Exploration ISBA 2004 - International Society for Bayesian Analysis. 2004. 89-93

[71] S. Thrun and J. Langford and V. Verma, Risk Sensitive Particle Filters, Neural Information Processing Systems (NIPS), Dec 2001

[72] Verma V, Gordon G, Simmons R, et al. Real-time fault diagnosis [robot fault diagnosis]. IEEE Robotics & Automation Magazine, 2004,11(2):56-66

[73] Frank Hutter and Richard Dearden:The Gaussian Particle Filter for Diagnosis of Non-Linear Systems. In Proceedings of the 14th International Conference on Principles of Diagnosis(DX'03), Washington, DC, USA, June 2003, 65--70,

[74] Frank Hutter and Richard Dearden: Efficient On-line Fault Diagnosis for Non-Linear Systems. In Seventh International Symposium on Artificial Intelligence and Robotics in Space (i-SAIRAS-03), May, 2003.

[76] S. Thrun and J. Langford and V. Verma, Risk Sensitive Particle Filters, Neural Information Processing Systems (NIPS), Dec 2001

[77] Verma V, Gordon G, Simmons R, et al. Real-time fault diagnosis [robot fault diagnosis]. IEEE Robotics & Automation Magazine, 2004,11(2):56-66

[78] E. Keogh, K. Chakrabarti, and M. Pazzani. Locally Adaptive Dimensionality Reduction for Indexing Large Time Series Databases. In proceedings of ACM SIGMOD Conference on Management of Data. Santa Barbara. May 21-24, 2001.123-130

[79] K. Kalpakis, D. Gada, and V. Puttagunta. Distance Measures for Effective Clustering of ARIMA Time-Series. In proceedings of the 2001 IEEE In'l Conference on Data Mining. San Jose, CA. Nov 29-Dec 2, 2001. 273-280

[80] P. Geurts. Pattern Extraction for Time Series Classification. In proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery. Freiburg, Germany. 2001. 37-45

[81] E. Keogh, S. Lonardi, and B. Chiu. Finding Surprising Patterns in a Time Series Database in Linear Time and Space. In proceedings of the 8th ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining. Edmonton, Alberta, Canada. Jul 23-26, 2002. 550-556

[82] Mörchen, F. Time-series Knowledge Mining, Phd thesis, Dept. of Mathematics and Computer Science, University of Marburg, Germany, 2006

[83] K.P.Chan and W. C. FU. Eficient Time Series Matching by Wavelets[C]. Proceed -ings of the International Conferenceon Data Engineering.Washington: IEEE Computer Society: 1999.126-133.

[84] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids: Cambridge University Press. 1998. 15:446-454,

[85].Eamonn Keogh1 ,Stefano Lonardi1 Compression based data mining of sequential data, Data Mining and Knowledge Discovery, Springer Netherlands, 2007. 14(1):99-129

[86] B. Ripley. Pattern recognition and neural networks [M]. London: Cambridge University Press. 1996. ISBN 0-521-46086-7 (hardback), 403.

[87] Xiao, Hui; Feng, Xiao-Fei and Hu, Yun-Fu. A new segmented time warping distance for data mining in time series database[C]. In proceedings of 2004 International Conference on Machine Learning and Cybernetics. Shanghai,China. 2004.4:1277-1281.

[88] A. Panuccio, M. Bicego, and V. Murino. A Hidden Markov Model-based approach to sequential data clustering. In T. Caelli, A. Amin, R. P. W. Duin, M. S. Kamel, and D. de Ridder, eds, Proceedings Joint IAPR International Workshops Structural, Syntactic, and Statistical Pattern Recognition, Springer, 2002: 734-742.

[89] P. Indyk, N. Koudas, and S. Muthukrishnan. Identifying representative trends in massive time-series data sets using sketches. In A. E. Abbadi, M. L. Brodie, S. Chakravarthy, U. Dayal, N. Kamel, G. Schlageter, and K.Y. Whang, eds, Proceedings of the 26th International Conference on Very Large Data Bases (VLDB'00), Morgan Kaufmann, 2000. 363-372.

[90] Lin, J, Keogh, E, Lonardi, S. & Chiu, B. “A Symbolic Representation of Time Series, with Implications for Streaming Algorithms”, In proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. 2003.1-10

[91].Lkhagva, B. Suzuki, Y. Kawagoe, K.“Extended SAX: Extension of Symbolic Aggregate Approximation for. Financial Time Series Data Representation.” DEWS2006 4A-i8.

[92] R. Agrawal, G. Psaila, E. L. Wimmers and M. Zait. Querying shapes of histories[C]. In Porc.of the 21st Int'l Conferenceo nVeryLargeDatabases. San Francisco: Morgan Kaufmann Publishers Inc. 1995:502-514.

[93] Sanghyun Park, Wesley W Chu, Jeehee Yoon, Chihcheng Hsu. Efficient searches for similar subsequences of different lengths in sequence databases[C]. Proceedings of the 16th International Conference on Data Engineering. Washington: IEEE Computer Society.2000:23-32.

[94]. Larry M. Manevitz. One-Class SVMs for Document Classification. Journal of

Machine Learning Research 2 (2001) 139-154.

[95] Hyun Joon Shina, Dong-Hwan Eomb, Sung-Shick Kim. One-class support vector machines-an application in machine fault detection and classification Computers & Industrial Engineering 2005 ,48:395–408

[96] Ingo Steinwart Don Hush Clint Scovel A Classification Framework for Anomaly Detection Journal of Machine Learning Research 2005,6:211–232

[97] M. Davy, F. Desobry, A. Gretton and C. Doncarli, "An Online Support Vector Machine for Abnormal Events Detection", Signal Processing, August 2006, 86, Issue 8: 2009-2025.

[98] 邓乃扬,田英杰. 数据挖掘中的新方法—支持向量机.北京：科学出版社,2004 .

[99] Scholkopf,B. Estimating the Support of a High-Dimensional Distribution. Neural Computation 2001,13：1443–1471

[100] Tax, D. und Duin, R. Data domain description by support vectors. In: Verleysen, M. (Hrsg.), Proc. ESANN. S. Brussels. D. Facto Press. 1999. 251–256.

[101] 杨绍清, 贾传荧. 两种实用的相空间重构方法 . 物 理 学 报, 2002,51(11); 2452-2456

[102] 吕金虎,陆君安,陈士华.混沌时间序列分析及其应用【M】武汉大学出版社,2002

[103] Packard N H, Crutchifield J P , Farmer J D , et al . Geometry f rom a time series [J] . Phys Rev Lett ,1980 ,45 (6) :712 -716.

[104] Agnon Yehuda,Golan Amos, Shearer Matthew.Nonparametric,nonlinear, short-term forecasting： theory and evidence for nonlin—earities in the commodity markets rJ]．Economics Lett．,1999,65:293-299

[105] Broomherd D S. Extracting qualitatire dynamics f rom experimental data [J] . Phys D ,1987 ,20 (11) 67-69

[106] YING Cheng-lai,David Lerner.Effective scaling regime for computing the correlation dimension f rom chaotic time series [J] .Phys D,1998,115 (5) :1-18.

[107] 李玉霞, 吴百海, 邢志鹏.单变量时间序列相空间重构及应用研究 . 组合机床与自动化加工技术 2004,2:51-55

[108] 杨志安,王光瑞,陈式刚．用等间距分格子法计算互信息函数确定延迟时间[J]．计算物理,1995,12(4)：442-448

[109] Jeong J,Gore J C.Peterson B S．Mutual informa tion analysis of the EEG in patients with Alzheimer’s disease[J]．Clinical Neurophysiology,2001,112：872~835

[110] 赵鸿,柴路,王浩.等．互信息在时间序列分析中的应用[J].1996,14(1):48-52

[111] 徐健学,杨红军,等.皮层脑电时间序列的相空间重构及非线性特征量的提取[J]．物理学报,2002,51(2)：205~213

[112] 陈阳.新的独立性度量及其在混沌信号分析中的应用[J].东南大学学报,2003,11(33)：121-125.

[113]王泽,朱贻盛,李音．独立量在混沌信号分析中的应用[J]．电子学报,2002,30(10)：1505~1507

[114]樊重俊,王浣尘．度量两个序列非线性相关性的一种方法[J]．信息与控制,1998,27(3)：183-189

[115]H.S. Kim1; a, R. Eykholt b; J.D. Salas. Nonlinear dynamics, delay times, and embedding windows. Source, Physica D archive March 1999,127(1-2): 48 - 60

[116]Kennel.Mathew B,Brown R,Abarbanel H D I．Determi ning embedding dimension for phase··space reconstruction using a geomet—·rical construction[J]．Phy Rev A,1992,45：3403~3411

[117]Abarbanel H D I,Brown R,Sidorowich J J,et a1．The analysis of observed chaotic data in physical systems[J]．Reviews of Modern Physics.1993,65(4)：1331-1392

[118]Cao Liangyue．Practical method for determining the minimum em—．bedding dimension of a scalar time series[J]．Physica D,1997,110:43-50

[119] 钟清流,蔡自兴.基于一类支持向量机的传感器故障诊断 计算机工程与应用 2006 42 (19). 1-4.

[120] Carl Staelin, Parameter selection for support vector machines. HP Laboratories Israel1 HPL-2002-354 (R.1) November 10th , 2003.341

[121] T. Hastie, S. Rosset, R. Tishbirani, and J. Zhu. The entire regularization path for support vector machines. J. Machine Learning Research, 2004.5:1391-1415,

[122] Frohlich, H. Zell, A. Efficient parameter selection for support vector machines in classification and regression via model-based global optimization. Neural Networks, 2005. IJCNN '05. Proceedings. 2005 IEEE International Joint Conference on. 31 July-4 Aug. 2005,3.1431- 1436

[123] O. Chapelle, V. Vapnik, O. Bousqet, and S. Mukherjee. Choosing Multiple Parameters for Support Vector Machines. Machine Learning, 2002.46(1): 131-159.

[124]V. Cherkassky and Y. Ma. Practical selection of svm parameters and noise estimation for svm regression. Neural Networks, 2004.17(1):113-126,

[125] Eberhart R C , Shi Y. Particle swarm optimization :developments , applications and resources [A] .Proceedings of the IEEE Congress on Evolutionary Computation [C] . Piscataway , NJ : IEEE Service Center,2001. 81-86

[126]徐海, 刘石, 马勇, 等. 基于改进粒子群游优化的模糊逻辑系统自学习算法[J] . 计算机工程与应用,2000 , (7) : 62-63

[127]van den Bergh F , Engelbrecht A P. Using cooperative particle swarm optimization to train product unit neural networks [R] . IEEE International Joint Conference on Neural Networks , Washington D C , USA , 2001.79

[128]曾建潮,崔志华, 一种保证全局收敛的 PSO 算法[J]. 计算机研究与发展,2004,41(8):1333-1338

[129]Gwo Ching Liao, Ta Peng Tsao. Applicat ion Embed ded Chaos Search Immune Genet icA lgo rithm fo r Sho rt2 term U nit Comm itment [J]. E lectric P ow er S y stem s R esearch, 2004, 71 (2) : 135-144

[130]A. Abraham, H. Guo, H. Liu, Swarm intelligence: foundations, perspectives and applications, in: N. Nedjah, L. Mourelle (Eds.), Swarm Intelligent Systems, Studies in Computational Intelligence, Springer-Verlag, Germany, 2006.236

[131] J.-P. Eckmann, D. Ruelle, Ergodic theory of chaos and strange attractors,Rev. Mod. Phys. 1985,57:617.

[132] CRUTCHFIELD J P,PACKARD N H·Symbolic dynamics of reversibility Using Symbolization[c].Proceeding of Fifth Experi．noisy chaos[J]·Physica D, 1983,7：201-223

[133] KURTHS J,VOSSA,SAPARIN P,etal·Quantitative analysis of variations in a spark igniti0n engine[J].SAE Paper,1996,29 heart rate variability[J] Chaos, 1995,5(1)：88-94．

[134] RECHESTER A,WHITE R B Symbolic kinetic equation for DAVISL J , FELDKAMP L,HOARD J,en1. Controlling cyclic chaotic attractor[J] Physics Review LettersA. 1991,156：419-424·

[135] TANG x z,TRACY E R,BROWN R·Symbol statistics and spa chanicaI Engineering Con and Exposition(ASME).Califotio-temporal systems [J] Physical Review D· 1997,102：253-261．

[136] DAW Cs,KENNEL M B,FINNEY C E A． Application of Sym. Instit Ⅱ te,KY:Lexington,1998．boIic Dynamics‘o Modeling and Control of an Internal Co mbustion [233 TANG x z．TRACY E R．Data ComDression and InIom ation Engine[C]-Snowbird：SIAMDS97,1997. 243

[137] DAW C s,FINNEY C E A,KENNEL M B．Measuring Time Irreversibility Using Symbolization[c].Proceeding of Fifth ExperimentaI Chaos Conference Orlando,USA：FIorida,1999.219

[138] Lehrman M,Rechester A B,White R B.Symbolic analysis of chaotic signals and turbulent fluctuations. C S,Finney C E A,Kennel M B.Symbolic approach for measuring temporal "irreversibility".Physical Review E,2000,62(2): 1912-1921

[139] 蒋嵘,李德毅,程辉. 基于形态表示的时间序列相似性搜索[J].计算机研究与发展,2000,37 (5) :601～608

[140] Keogh, E.Chakrabarti,K.Pazzani, M. & Mehrotra “Dimensionality reduction for fast similarity search in large time series databases”, Journal of Knowledge and Information Systems. 2000,91-93

[141] Keogh E, pazzani M. Relevance Feedback Retrieval of Time Series Data. Proceedings of the 22th Annual International.Semimarkov Models for Endpoint Detection in Plasma Etching IEEE Transactions on Semiconductor Engineering, 2001 .325

[142] 李斌,谭立湘,解光军,等．非同步多时间序列中频繁模式的发现算法[J]．软件学报,2002,13(3)：410—416．

[143] 金宁德 李伟波. 非线性时间序列的符号化分析方法研究. 动力学与控制学报.2004,2(3):54-58.

[144] 倪世宏,王刚,史忠科. 一种非同步时间序列特征提取算法. 计算机应用研究 2005, 5:87-89.

[145] M.B. Kennel, M. Buhl, Estimating good discrete partitions form observed data: symbolic false nearest neighbors, Phys.Rev. E 2003, 91 (8) ,84-102.

[146] V. Rajagopalan, A. Ray, Wavelet-based space partitioning for symbolic time series analysis, Proceedings of IEEE Conference on Decision and Control (CDC) and European Control Conference (ECC), Seville, Spain, 2005, 5245–5250.

[147] 任江涛,何武,等. 一种时间序列快速分段及符号化方法. 计算机科学 2005 32(9):166-169.

[148] 王晓哗,徐晓曩,等.多维时间序列数据符号化表示方法的研究.计算机工程.2006,32(12):52-54.

[149] Lin, J,Keogh, E., Lonardi, S. & Chiu, B. “A Symbolic Representation of Time Series, with Implications for Streaming Algorithms”, In proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. 2003,273

[150].Lkhagva, B., Suzuki, Y., Kawagoe, K., “Extended SAX: Extension of Symbolic Aggregate Approximation for. Financial Time Series Data Representation.” DEWS2006 ,4A-i8.

[151] Dragomir Yankov, Eamonn Keogh, Li Wei, Xiaopeng Xi, Wendy Hodges. Fast Best-Match Shape Searching in Rotation Invariant Metric Spaces. SIAM International Conference on Data Mining (SDM'07), to appear, 2007.351

[152] Xiaopeng Xi, Eamonn Keogh, Christian Shelton, Li Wei & Chotirat Ann Ratanamahatana (2006). Fast Time Series Classification Using Numerosity Reduction. ICML. 2006.371-374

[153] B. Lkhagva; Yu Suzuki; K. Kawagoe New Time Series Data Representation ESAX for Financial Applications Data Engineering Workshops, 2006. Proceedings. 22nd International Conference 2006 , 115 -117

[154] Keogh, Eamonn1; Lonardi, Stefano2; Ratanamahatana, Chotirat3; Wei, Li4; Lee, Sang-Hee5; Handley, John6 Compression-based data mining of sequential data Data Mining and Knowledge Discovery, Volume 14, Number 1, February 2007 , (31):99-129

[155] E. Keogh and S. Kasetty. On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration. In proceedings of the 8th ACM SIGKDD Int'l Conference on KnowledgeDiscovery and Data Mining. Edmonton, Alberta, Canada. July 23-26, 2002. 102-111

[156] J. Lin and E. Keogh. Group SAX: Extending the Notion of Contrast Sets to Time Series and Multimedia Data. In proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases. Berlin, Germany. Sept 18-22, 2006. 284-296

[157].E. Keogh,J. Lin and A. Fu. HOT SAX: Efficiently Finding the Most Unusual Time Series Subsequence. In Proc. of the 5th IEEE International Conference on Data Mining (ICDM 2005),Houston,Texas,Nov 27-30, 2005.226 - 233.

[158].Jessica Lin, Eamonn Keogh Li Wei and Stefano Lonardi Experiencing SAX: a Novel Symbolic Representation of Time Series. DMKD Journal. ctober 2007, 15(2): 107 - 144

[159] Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. In Lawrence K. Saul, Yair Weiss, and L´eon Bottou, editors, Advances in Neural Information Processing Systems 17. MIT Press, Cambridge, MA, 2005. 97

[160] Hui-Fuang Ng. Automatic thresholding for defect detection [J]. Pattern

Recognition Letters,2006,10(15):1644-1649.

[161] Marina Mueller, Karl Segl, Hermann Kaufmann. Edge and region based segmentation technique for the extraction of large, man-made objects in high resolution satellite imagery [J]. Pattern Recognition, 2004,37(8): 1619-1628.

[162] Chen Guang Zhao, Tian Ge Zhuang. A hybrid boundary detection algorithm based on watershed and snake [J]. Pattern Recognition Letters,2005,26 (9): 1256–1265.

[163] Lin Kai-yan, Wu Jun-hui, Xu Li-hong. A survey on color image segmentation techniques [J]. Journal of Image and Graphics,2005,10(1):11-18

[164] Tremeau A, Borel N A. Region growing and merging algorithm to color segmentation [J]. Pattern Recognition,1997,30(7):1191-1203.

[165] F.R. Bach, and M.I. Jordan. Learning Spectral Clustering. Advances in Neural Information Processing Systems (NIPS), 2004,16.79

[166] Arik Azran, Zoubin Ghahramani Spectral Methods for Automatic Multiscale Data Clustering,Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - 2006 ,1 :190 - 197

[167] F. R. Bach and M. I. Jordan. Blind one-microphone speech separation: A spectral learning approach. Advances in Neural Information Processing Systems (NIPS) 17, 2004.

[168] Chang, Hong Yeung, Dit-Yan.Robust path-based spectral clustering with application to image segmentation Proceedings 10th IEEE International Conference on Computer Vision, 2005, ICCV 2005. Beijing, China, 15-21 October 2005. 1: 278-285.

[169] A.Ng.M.Jordan and Y.Weiss “On spectral clustering: Analysis and an algorithm” In Advances in Neural Information Processing Systems 14, 2001.

[170] M. Polito and P. Perona “Grouping and dimensionality reduction by locally linear embedding” Advances in Neural Information Processing Systems 14, 2002

[171] Jacob Lurie Review of Spectral Graph Theory: by Fan R. K. Chung ACM SIGACT News archive June 1999, 30 ,Issue 2: 14 - 16

[172] Chung, Fan R. K. (1997). Spectral Graph Theory. Providence, RI: American Mathematical Society. ISSN 0160-7642; no 92.

[173] Jianbo Shi, Jitendra Malik, Normalized Cuts and Image Segmentation (1997) , IEEE Transactions on Pattern Analysis and Machine Intelligence.79-80.

[174] F. R. Bach, and M. I. Jordan. Learning spectral clustering, with application to speech separation. Journal of Machine Learning Research, 2006, 7:1963-2001,

[175]Wang Chongjun; Li Wu jun; Ding Lin; Tian Juan; Chen Shifu,Image segmentation using spectral clustering.Tools with Artificial Intelligence, 2005. ICTAI 05. 17th IEEE International Conference .2005,14-16

[176] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. In NIPS 15, 2003.

[177] H. Zha, C. Ding, M. Gu, X. He, and H. Simon. Spectral relaxation for K-means clustering. In NIPS 14, 2002.58-59.

[178] Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. In Lawrence K. Saul, Yair Weiss, and L´eon Bottou, editors, Advances in Neural Information Processing Systems 17. MIT Press, Cambridge, MA, 2005. 97

[179] C. J. Alpert and A. B. Kahng. Multiway partitioning via geometric embeddings, orderings and dynamic programming.IEEE Transactions on Computer-aided Design of Integrated Circuits and Systems, 1995.14(11):1342–58.

[180]http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/segbench/BSDS300/html/dataset/images.html

[181] Ulrike von Luxburg. A Tutorial on Spectral Clustering.Statistics and Computing 2007 ,17(4):45-48

[182]Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics 1995:189–196.

[183] Riloff, E., Wiebe, J., &Wilson, T. (2003). Learning subjective nouns using extraction pattern bootstrapping. Proceedings of the Seventh Conference on Natural Language Learning (CoNLL-2003).271

[184] Maeireizo, B., Litman, D., & Hwa, R. (2004). Co-training for predicting emotions with spoken dialogue data. The Companion Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL). 2004.78

[185] Rosenberg, C., Hebert, M., & Schneiderman, H. (2005). Semi-supervised selftraining of object detection models. Seventh IEEE Workshop on Applications of Computer Vision. 2005,310

[186] Haffari, G., & Sarkar, A. (2007). Analysis of semi-supervised learning with the Yarowsky algorithm. 23rd Conference on Uncertainty in Artificial

Intelligence(UAI). 2007. 217

[187] Culp, M., & Michailidis, G. (2007). An iterative algorithm for extending learners to a semisupervised setting. The 2007 Joint Statistical Meetings (JSM).1-30

[188] hapelle, O., & Zien, A. (2005). Semi-supervised classification by low density eparation. Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics (AISTAT 2005).173

[189] Demirez, A., & Bennett, K. Optimization approaches to semisupervised learning. In M. Ferris, O. Mangasarian and J. Pang (Eds.), Applications and algorithms of complementarity. Boston: Kluwer Academic Publishers. 2000. 76-80

[190] Chapelle, O., & Zien, A. (2005). Semi-supervised classification by low density separation. Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics 2005,217.

[191] Joachims, T. (1999). Transductive inference for text classification using support vector machines. Proc. 16th International Conf. on Machine Learning. Morgan Kaufmann, San Francisco, CA. 1999. 200–209.

[193] 陈毅松,汪国平,董士海 基于支持向量机的渐进直推式分类学习算法,软件学报, 2003,14(3):451-460.

[194] Altun.Y.,McAllester, D.& Belkin,M. (2005). Maximum margin semi-supervised learning for structured variables. Advances in Neural Information Processing. 2005,148: 145-152

[195] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Sch¨olkopf, “Learning with local and global consistency,” in Advances in Neural Information Processing Systems, NIPS2004, Vancouver, Canada, Dec 2004, vol. 16, MIT Press.

[196] A. M. Fraser and H. L. Swinney, Independent coordinates for strange attractors from mutual information, Phys. Rev. A, 33, 1134 (1986).

致 谢

转眼间,在中南大学里度过了五年的求学历程！能够在这里学有所成,首先要归功于我敬爱的导师蔡自兴教授.

在科研的道路上,恩师用他特有的人格品质言传身教.他严谨治学的精神、自强不息的拼搏意志、持之以恒的追求真理、未雨绸缪的超前意识、无一不时刻激励和鞭策着我,成为我战胜困难、攻克科研难题的精神力量.他的高尚的人格深深地感染着我,并将成为我终生的楷模.在此,谨向我的导师致以最衷心的感谢！

在漫长的求学生涯中,我要衷心感谢我的妻子！如果没有她的理解,鼓励,全力支持和帮助,几乎没有完成学业的可能.

在论文的写作过程与项目的研究中,得到了多位同学的启发,直接帮助,他们或对我的研究提出了中肯的意见,或在研究过程中给予了重要的支持！他们的工作对于本论文的最后完成不可或缺！为此,特向曾经为我提供支持和帮助的王璐博士、于金霞博士、龚涛博士、崔益安、高平安、刘丽珏、陈爱斌、陈白帆、文志强、屈太国、孔志周、潘薇等表示感谢.

在长期的学习过程中,得到过魏世勇、肖晓明、孙国荣、石跃祥等老师的帮助和指导 在此表示最诚挚的谢意！同时也衷心感谢各位专家在百忙中对本论文的评阅与指导！

特别感谢国家自然科学基金项目“未知环境中移动机器人导航的理论与方法研究”(项目号：60234030)、国家基础研究项目(项目号：A1420060159)的资助！

攻读博士期间完成的学术论文与科研工作

[1]. 钟清流.蔡自兴.基于统计特征的时序符号化算法.计算机学报 2008,31(10): 1857-1864. (EI)

[2]. 钟清流.蔡自兴. 时序数据的动态有界符号化方法. 控制与决策 2008, 23(10): 1109-1113 (EI)

[3]. 钟清流.蔡自兴.时序数据的矢量化符号方法小型微型计算机系统 (已录用)

[4]. 钟清流.蔡自兴.用于彩图分割的自适应谱聚类算法.计算机应用研究(已录用)

[5]. 钟清流.蔡自兴.基于 OCSVM-CPSO 的自适应故障诊断.计算机工程与应用 2007 43 (8): 18-20

[6]. 钟清流.蔡自兴.基于一类支持向量机的传感器故障诊断.计算机工程与应用.2006.42(19):1-4.

[7]. 钟清流.蔡自兴. 基于支持向量机的渐近式半监督式学习算法.计算机工程与应用. 2006.42(25):19-23

[8]. 董本清,钟清流. SVM 在金属塑性成形摩擦系数预测中的应用研究.科学技术与工程. 2006,22(6);3571-3574

[9].任文进;钟清流,基于混沌粒子群的支持向量机参数优化.科学技术与工程 2007.7(18) 23-27:4597-4600

科研工作

[1] 国家自然科学基金重点项目"未知环境中移动机器人导航控制的理论与方法研究" (项目号：60234030)

[2] 国家基础研究项目(项目号：A1420060159)