摘 要

引力波作为广义相对论中的重要预言，其探测是当前科学前沿领域之一。目前世界上已经兴建了若干个激光干涉引力波天文台，天文台在运行期间全天候地采集大量数据。而作为探测引力波的最后一个环节，引力波数据处理的重要性不言而喻。

一方面，由于引力波信号过于微弱，各种噪声会掩盖真实的引力波信号。为了从充满噪声的数据中提取引力波信号，各个引力波探测器的数据均传送至一个数据处理中心节点集中处理，以提高信噪比，而这往往需要巨大的计算量进行数据分析，数据处理实时性难以保证。另一方面，在 2015 年左右，第二代激光干涉引力波天文台网络将在全球范围内开始运行，其对引力波数据的实时处理和探测器的快速诊断提出了新的要求。

综合这些因素，本文改进了现有的引力波数据处理系统。除了在系统结构上进行优化外，还对引力波数据处理的算法进行了改进，从而使其满足引力波数据实时处理的要求。

系统结构方面，本文提出将数据处理中心节点上的事件生成功能转移至探测器，从而为单探测器上的事件否决提供了可能。本文还介绍了为实时监测 Burst 类型引力波主频道事件所开发的工具，该工具为探测器实时诊断提供新的信息来源。

数据处理算法改进方面，利用辅助频道的信息以及模式识别的方法，否决探测器主频道上由噪声引起的事件，相比传统的否决算法，取得了更好的否决性能以及更快的否决速度，从而有力地支持了探测器实时诊断。此外，同样是利用辅助频道的信息，否决 Burst 类型引力波主频道事件，从而减少了中心节点上的计算负担，在不增加硬件计算能量的情况下，为数据处理提高精度和减少延迟提供了可能。最后，改进了 Burst 类型引力波信号能量恢复，在不降低实时性要求的情况下减少了引力波被误判为噪声的可能。

关键词：引力波　　实时处理　　事件否决　　能量恢复　　模式识别

Abstract

As one of the most important prediction in the General Relativity, gravitational wave's detection is the cutting-edge science currently. Now in the world, several laser interferometer gravitational-wave observatories have been built. During their science runs, large amounts of data are being collected all of the time. As the last mile of gravitational wave detection, the importance of gravitational wave data processing is self-evident.

On the one hand, due to the extreme weakness of gravitational wave, it is covered by various kinds of noises. To extract gravitational wave from the noisy data, all gravitational wave observatories transmit their data to a data processing central location in real-time for improving signal-to-noise ratio purpose, however this will consume large amount of computing power, namely it is hard to satisfy the real-time requirement. On the other hand, around 2015, the second generation of laser interferometer based gravitational wave observatory network will run in the world wide. This brings new challenge to the real-time gravitational wave data processing and also the real-time detector diagnostics.

Considering all these issues, in this thesis, modifications on the current gravitational wave data processing system are proposed. Besides improvements on system's structure, some enhancements are applied to the algorithms of gravitational wave data processing. All these make the system satisfy the requirement of real-time processing.

In terms of system structure, the event generation function is moved to single detector from the central location, which enables the single detector event veto. A tool is developed to monitor Burst type gravitational-wave main channel event, which provides a new source to the real-time detector diagnostics.

In terms of data processing algorithms, the information from all auxiliary channels is used to veto events originated by noises via pattern recognition method. Compared to a traditional veto method, a better veto performance and a faster veto speed are achieved, which significantly supports the real-time detector diagnostics.

Besides, the single detector event veto is achieved also by the information from all auxiliary channels. This can significantly reduce the computing burden on the central location, which enables high precision data processing and low latency without adding more computing hardware. Finally, the Burst type gravitational wave energy recovery is enhanced, which reduces the misclassification possibility on real gravitational wave without hampering the real-time performance.

Keywords: Gravitational Wave Real-time Processing Event Veto Energy Recovery Pattern Recognition

目 录

第 1 章 引言 1
1.1 引力波简介 1
1.2 引力波探测器简介 2
1.2.1 引力波探测器原理 2
1.2.2 引力波探测器发展情况 4
1.3 引力波数据分析简介 4
1.3.1 引力波数据分析的分工 4
1.3.2 受限的引力波数据分析 5
1.4 论文工作与安排 7
1.4.1 论文目标 7
1.4.2 论文主要贡献 8
1.4.3 论文结构安排 8
第 2 章 引力波数据实时处理系统 10
2.1 LIGO 数据实时处理 10
2.1.1 LIGO 现有的数据实时处理系统 10
2.1.2 实时处理的意义 15
2.1.3 Advanced LIGO 的挑战 17
2.2 改进后的数据实时处理系统 18
2.2.1 改进思路 18
2.2.2 改进后的系统 19
2.3 本章小结 20
第 3 章 引力波探测器表征中的信号否决 21
3.1 引力波探测器表征概述 21
3.1.1 噪声源研究 21
3.1.2 数据质量标记 23
3.1.3 噪声能谱分析 23
3.1.4 数据运行支持 23

3.1.5 数据标定 24
3.1.6 时钟同步 24
3.2 引力波主频道信号否决相关软件及算法 24
3.2.1 引力波主频道信号否决简介 24
3.2.2 引力波主频道信号否决依赖的软件 25
3.2.3 引力波主频道信号否决现有算法 25
3.3 引力波主频道信号否决算法设计 27
3.3.1 基于事件的否决存在的缺陷 27
3.3.2 模式提取 28
3.3.3 否决算法参数配置与流程 30
3.3.4 模式识别方法选择 31
3.4 引力波主频道信号否决性能比较 34
3.4.1 性能指标 34
3.4.2 与传统否决算法的比较 34
3.4.3 不同引力波数据上的比较 38
3.5 否决算法的在线实现 42
3.5.1 引力波信号特征提取 42
3.5.2 否决算法的在线运行 46
3.6 本章小结 47
第 4 章 Burst 类型事件实时监测与否决 48
4.1 Burst 类型事件实时监测 48
4.1.1 Burst 类型事件介绍 48
4.1.2 数据监测工具箱 DMT 简介 49
4.1.3 OmegaMon 的设计 50
4.1.4 OmegaMon 实时监测 51
4.2 Burst 类型主频道事件否决 52
4.2.1 已有的主频道事件否决方法 52
4.2.2 Burst 类型主频道事件否决 53
4.2.3 Burst 类型主频道事件否决对实时性的影响 55
4.3 本章小结 56

第 5 章 Burst 类型引力波信号能量恢复.....57
5.1 Omega Pipeline 中的信号能量恢复.....57
5.2 增强的 Omega Pipeline 能量恢复设计.....59
5.3 性能比较.....61
5.3.1 单探测器上的聚类融合.....62
5.3.2 多探测器上的聚类.....64
5.4 本章小结.....67
第 6 章 结论与展望.....69
6.1 结论.....69
6.2 未来工作.....70
6.3 展望.....70
参考文献.....72
致谢与声明.....76
个人简历、在学期间发表的学术论文与研究成果.....77

第1章 引言

1.1 引力波简介

1916 年，爱因斯坦在他的广义相对论中预言了引力波的存在：凡是有质量的物质，当其加速运动时都能产生引力辐射。作为自然界 4 种基本相互作用（电磁力，万有引力，强相互作用和弱相互作用）之一的引力的重要性与意义不言而喻。

从物理学的角度出发，首先引力波的探测能够进一步验证广义相对论的正确性，其次自从相对论和量子力学诞生以来，人们一直试图将两者融合到一个理论当中，即所谓的量子引力理论[1]，弦论和圈量子引力理论是其中的典型代表。引力波的探测在很大程度上影响着量子引力理论的进一步发展[2]。

从天文学的角度出发，尤其对于观测天文学来说，引力波是继电磁辐射、宇宙射线和中微子探测后又一个探索宇宙的重要手段。与电磁辐射相比，引力波与物质的相互作用十分微弱，在传播过程中几乎不会发生衰减或者散射，这就意味着引力波能够揭示一些宇宙隐蔽角落的信息，比如超新星爆发时的内部结构[3]。正由于引力波对于天文学有着如此重大的意义，于是从二十世纪六七十年代以来逐渐兴起了一个崭新的观测天文学分支——引力波天文学。尽管引力波仍未能通过实验探测到，但是理论上的引力波天文学业已存在并不断发展，且具有重大意义。比如在暗物质研究方面，由于现在天文学家的普遍共识认为宇宙中不发射任何电磁波的暗物质比例远大于已知的物质所占比例，这些暗物质与外界能产生相互作用的唯一途径即是引力相互作用，因此引力波天文学在暗物质探索方面作用不可替代。

鉴于引力波的重要性，研究和探测引力波的工作从未停止过。1974 年，小约瑟夫·泰勒和拉塞尔·赫尔斯发现脉冲双星 PSR 1913+16 的轨道衰减规律与理论上引力辐射所造成的动能衰减相符合，从而间接地证明了引力波的存在[4]，并因此获得了 1993 年的诺贝尔物理学奖。为了直接探测到引力波，人们相继研制出了两代引力波探测器：共振质量探测器和激光干涉引力波探测器。激光干涉引力波探测器是目前的绝对主流，目前世界上正在运行的大型引力波探测项

目均使用基于激光干涉原理的引力波探测器，主要有美国的激光干涉引力波天文台（Laser Interferometer Gravitational-wave Observatory，简称 LIGO）[5]，德国和英国合作的 GEO600[6]，日本的 TAMA300[7]以及法国和意大利合作的 VIRGO[8]。

1.2 引力波探测器简介

1.2.1 引力波探测器原理

引力波的一个重要性质是与物质交互，使得物质产生形变。它是一种四极辐射（Quadrupole Radiation），即在沿着引力波的传播方向上存在两组垂直的极化方向，组与组间的夹角为 45 度，具体如图 1.1 所示。图 1.1 显示的是引力波以垂直于纸面的方向穿过一圆环的中心，圆环形状随时间变化的情况。可见在时间轴上方的一组极化方向上，圆环的形变周期性地出现在 x 轴和 y 轴上；在时间轴下方的另一组极化方向上，圆环的形变则周期性地出现在$y = x$和$y = -x$方向上。圆环最终的形变是这两者的叠加。用h_+（h-plus）和$h_\times$（h-cross）分别代表两组极化方向上的形变强度（strain），即某特定方向上半径变化长度与半径的比值。h_+与$h_\times$的大小直接由引力波的幅度决定。幅度越大，形变强度越大。

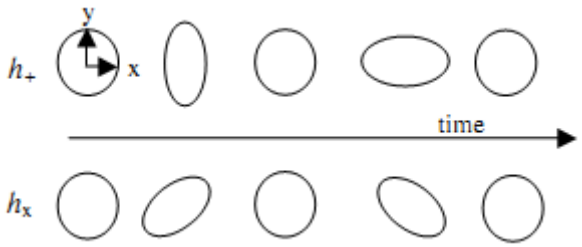

图 1.1 引力波四极辐射对物质的形变作用示意图[9]

目前世界上主流的引力波探测器均为激光干涉引力波探测器，如图 1.2 所示，其基本原理简单地讲就是一个大型的迈克尔逊干涉仪。

左下角的激光发生器发射出激光后，经过图 1.2 正中央的分光镜，分解成两束相互垂直的同相位和能量的激光，分别进入一条真空管道，到达各自管道尽头的镜面（即图 1.2 中左上角和右上角的 End Test Mass），反射回来，在分光镜处重新汇聚并产生相消干涉（不考虑噪声、引力波等因素的影响，此时两束激光应产生完全相消干涉），合成的光电流在图 1.2 右下角的光电转换装置处由

光信号转换成电信号。

可将探测器看成图 1.1 中的圆环，图 1.2 中央的分光镜对应于图 1.1 中圆环的中心，两个臂终点的 End Test Mass 则分别对应于圆环与 x 轴，y 轴的交点。于是探测器的臂长就恰好是圆环的半径。当引力波通过探测器时，两条真空管道的长度将不再相等，于是两束被反射回分光镜的激光束干涉所合成的光电流强度将会发生变化，这种变化转换成电信号，进而采样成为数字信号，这样就生成了一个数据流，它被称为引力波主频道（Gravitational-wave Channel，简称 GW Channel）。此外还有一系列的辅助频道供探测器硬件诊断使用。

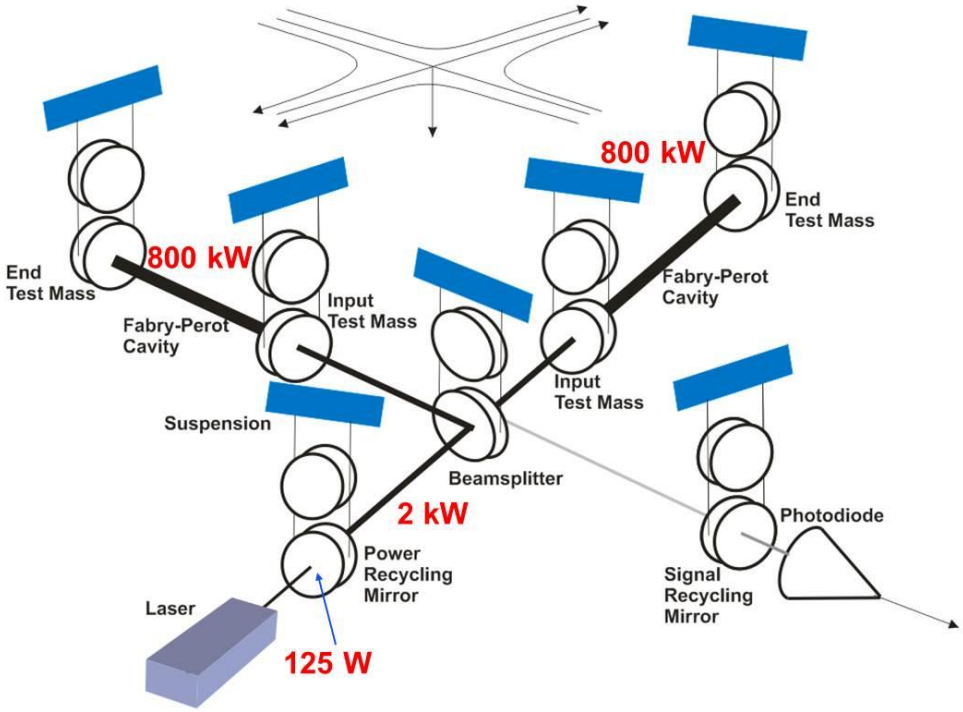

图 1.2 激光干涉引力波探测器基本原理示意图[10]

由于引力波很微弱，因此就需要增强引力波所产生形变而转化成的电信号的强度。显然，决定最终得到的电信号强度的因素主要有两个：激光的能量和分光镜处干涉的强度。激光能量的提高依赖于激光发生器的改进；而干涉的强度则依赖于两束激光的相位差的大小，而这又和真空管道的长度相关。长度越长，形变越明显，相位差越接近零，电信号越强。这就要求真空管道长度足够长，于是如图 1.2 中粗线标注的真空管道所示，激光在其中来回反射多次，变

相地成倍增加了真空管道的长度。

1.2.2 引力波探测器发展情况

美国是推动引力波探测和数据分析的绝对主力。第一代和第二代引力波探测器均是由美国率先提出和建造，而现在美国航空航天局联合欧洲航天局正在设计可以运行于太空中的第三代引力波探测器，激光干涉太空天线（Laser Interferometer Space Antenna，简称 LISA）[12]。而 LIGO 则已获得了美国国家科学基金的资助用于建设 Advanced LIGO[13]，计划于 2015 年正式运转。欧洲的 Virgo 也计划在 2015 年建成 Advanced Virgo[14]。日本则计划建造世界上第一个地下激光干涉引力波探测器——大型低温引力波望远镜（Large-scale Cryogenic Gravitational Wave Telescope，简称 LCGT）[15]，目前已经获得日本政府的资助。澳大利亚则计划在 LIGO 的帮助下建设南半球的第一个引力波天文台，LIGO Australia（Australian International Gravitational Observatory）[16]，计划于 2015 年下半年建成使用。

LIGO 拥有世界上精度最高的引力波探测器。目前 LIGO 共有两个天文台，一个位于美国华盛顿州，名为汉福天文台（Hanford Observatory，简称 H1），另一个位于美国的路易斯安娜州，名为利文斯顿天文台（Livingston Observatory，简称 L1）。自从 2002 年 8 月以来，LIGO 探测器已经进行了 6 次长时间连续的科学运行（Science Run）。最近的一次科学运行 S6（6^{th} Science Run）从 2009 年上半年持续到了 2010 年年底，进行了将近两年的数据收集。

1.3 引力波数据分析简介

1.3.1 引力波数据分析的分工

以 LIGO 为例，LIGO 将引力波数据分析分拆成四个组，分别是 Burst，CBC（Compact Binary Coalescence，致密星融合），Stochastic 和 CW（Continuous Wave）。它们分别对应四种类型的引力波[18]：短时脉冲信号（Burst Signals）[19]，线性调频信号（Chirp signals），随机性信号（Stochastic Signals）和周期性信号（Periodic Signals）。

短时脉冲信号的来源很多，核坍缩超新星爆炸，双星绕转融合，伽马射线爆发等相对论系统都是短时脉冲类型引力波的发生源。

线性调频信号的来源与短时脉冲信号有重合，最典型的代表是双中子星和双黑洞融合。

随机性信号典型的来源是宇宙大爆炸10^{-43}秒时放射出的宇宙残余引力波（Relic Gravitational Waves）[20]，由于引力波在传播过程中几乎不衰减或散射，因此宇宙残余引力波仍然在宇宙中传播着。

周期性信号的一个典型来源是快速自转的中子星，也就是通常所说的脉冲星（Pulsar）。

相对来说，由于短时脉冲信号和现行调频信号的引力波源十分丰富，且均为天体物理中研究的热点问题，因此 CBC 和 Burst 两个组是引力波数据分析中最重要和规模最大的两个组。

此外，LIGO 还设立了名为探测器表征（Detector Characterization，简称 DetChar）的工作组，也牵涉到很多数据分析的工作，但是其主要目的是给前面这四个组提供数据支持和探测器诊断。

1.3.2 受限的引力波数据分析

1.3.2.1 引力波事件率受限

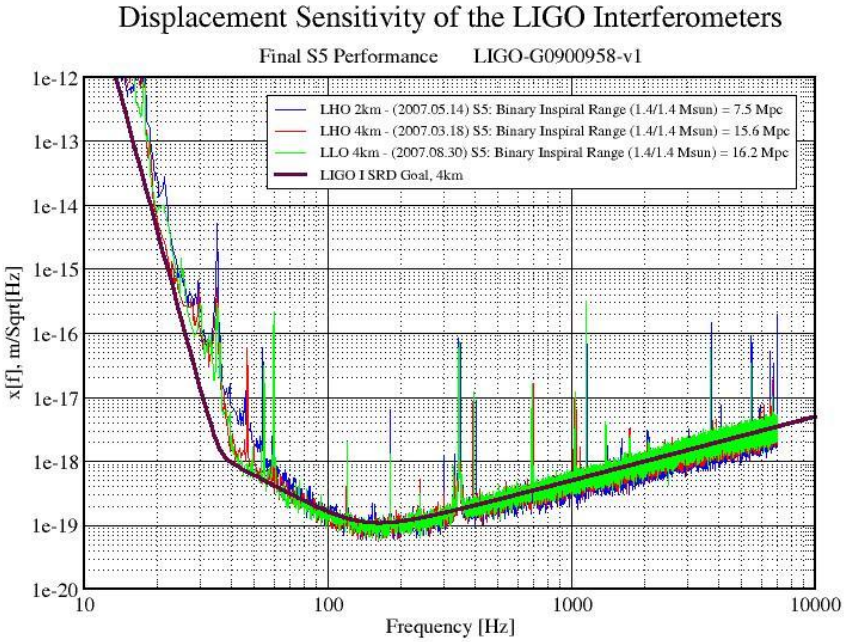

图 1.3 LIGO 第五次科学运行中的探测器灵敏度曲线[17]

如图 1.3 所示，显示的是 LIGO 第五次科学运行中随频率变化的位移灵敏度（Displacement Sensitivity）曲线。所谓位移灵敏度，在这里其实相当于单边功率谱密度函数的二分之一次幂，即$(S_n(f))^{1/2}$。曲线上的一个点表示的物理意义是：对于某个横轴上的特定频率，LIGO 探测器所能探测到的引力波的最低位移灵敏度。而即便是强度较高的引力波的位移灵敏度也在 10 的负 22 次量级，因此可以认为 S5 时期的探测器基本是无法探测到引力波的，即引力波事件率趋于零，事实上也未探测到。

接下来从数学上对引力波事件率进行推导[21]。

假设某类引力波信号的到达是一个泊松过程，设该泊松过程的参数为μ，那么在一段时间T内，有 n 个事件到达的概率为

$$P(\mathrm{n}|\mu)=\frac{\mu^n e^{-\mu}}{n!} \tag{1-1}$$

引入探测效率函数（detection efficiency function）f(X)[22]，自变量 X 为信号显著性。函数 f(X)表示的是最终被探测到且显著性高于 X 的引力波信号所占的百分比。那么有 m 个信号被探测到的概率为

$$P(\mathrm{m}|\mathrm{n},\mathrm{f(X)})=\frac{n!}{m!\,(n-m)!}f(X)^m(1-f(X))^{n-m} \tag{1-2}$$

那么联合概率为

$$P(\mathrm{m},\mathrm{n}|\mu,\mathrm{f(X)})=P(\mathrm{n}|\mu)\,P(\mathrm{m}|\mathrm{n},\mathrm{f(X)}) \tag{1-3}$$

那么观测到 0 个事件的概率为

$$P\big(\mathrm{m}=0\big|\mu,\mathrm{f(X)}\big)=\sum_{n=0}^{\infty}\frac{[1-f(X)]^n\mu^n e^{-\mu}}{n!} \tag{1-4}$$

将其看作指数函数的泰勒展开，则

$$P\big(\mathrm{m}=0\big|\mu,\mathrm{f(X)}\big)=e^{-\mu f(\mathrm{X})} \tag{1-5}$$

值得注意的是，这个概率等同于所有探测到的事件的显著性均小于 X 的概率。借助此概率，能推导出μ的上限。将时长T的实际数据上探测到的最显著事件的显著性用X_{max}表示，令$e^{-\mu f(\mathrm{X})}=1-p$，p 为需要的置信水平，则可推导出式 1-6 和式 1-7。

$$\mu_p = \frac{-\ln(1-p)}{f(X_{max})} \tag{1-6}$$

$$r_p = \frac{-\ln(1-p)}{Tf(X_{max})} \tag{1-7}$$

其中r_p就是置信度为p的情况下的引力波事件率上限，单位为天$^{-1}$（day^{-1}）。尽管由于不同的信号处理软件的$f(X_{max})$函数不同，得到的置信度 90%的事件率上限大多集中在 10^{-2} 至 10^{-1} 量级。如果考虑背景噪声的干扰，引力波事件率还会继续下降，可见目前引力波探测器确实很难探测到引力波，可能一年的时间才能探测到一两个甚至几十年的时间才能探测到一个。目前探测器基本都是在探测噪声。

1.3.2.2 精度受限

LIGO 中的数据分析往往对计算量有着较高的要求。但由于计算资源的限制，很多时候数据分析算法不得不牺牲计算精度，从而保证实时在线分析的需要，否则就很难实时在线处理数据。例如 LIGO 中名为 cWB（Coherent Wave Burst）的算法在 LIGO 第六次运行的 b 阶段中，由于分配的计算资源不足，竟然有多达 30%的数据段没有分析到[23]。

1.4 论文工作与安排

1.4.1 论文目标

本文主要针对前边提到的引力波数据分析受限问题，提出改进。研究目标包括以下四大方面内容。

1. 引力波数据实时处理系统的改进

如 1.2 中所述，LIGO 引力波探测器将于 2015 年正式开始运行 Advanced LIGO，2014 年即能取得部分工程数据。而去年结束的 LIGO 第六次科学运行中，引力波数据实时处理系统实现了延迟为十分钟量级的近似实时处理。针对 Advanced LIGO 中出现的新问题，本文提出了一个改进版本的引力波数据实时处理系统。

2. 引力波表征中主频道事件否决的改进

相对于引力波的强度，探测器中噪声背景极其强烈且复杂。出于对探测器诊断的需要，必须对引力波主频道中捕获到的噪声事件进行分析，确定其是由何种噪声源引起，从而改进探测器硬件，抑制乃至消除噪声源。而这就需要首先确定主频道中哪些事件最可能是噪声，并对其否决。

此外，LIGO 引力波表征中已有的主频道事件否决算法均无法支持实时在线分析，因此本文除了研究如何提高否决的性能外，还将探讨如何在实时运行方面进行改进。

3. Burst 类型引力波信号实时监测和筛选

目前缺乏一个能够实时显示捕获到的 Burst 类型引力波信号特性的工具。此外所捕获到的 Burst 类型引力波信号也需要判断其是否为噪声。本文将对这两点进行论述。

4. Burst 类型引力波信号能量恢复改进

目前 LIGO 中典型的 Burst 类型引力波信号捕获软件均使用显著性（例如信号幅值或置信度均可作为显著性指标）阈值来筛选捕获到的信号。若信号的显著性超过设定的阈值，那么这个信号就被认为是由引力波源产生；否则，该信号被认定为由噪声源引发。在信号捕获过程中，计算得到的信号显著性往往会低于信号的真实显著性，于是就可能发生由引力波源产生的引力波信号由于显著性亏损而被判定为噪声的情况。因此减少这种亏损，即恢复出更多的信号能量是一个值得研究的目标。

其中目标 2，3 和 4 均可看作目标 1 中改进系统的子模块具体实现。

1.4.2 论文主要贡献

本文的主要贡献包括以下几个方面。

1. 系统地介绍了 Burst 类型引力波数据处理的详细情况并提出了改进方案
2. 将模式识别引入到探测器表征与 Burst 类型引力波信号筛选中
3. 改进了 Burst 类型引力波信号能量恢复

1.4.3 论文结构安排

本文的结构如下：

第 1 章为引言，主要介绍了本文的研究背景与意义，简述了引力波和引力波探测的发展过程，介绍了引力波数据分析中客观存在的局限性，从而提出了

本文的研究目标与研究内容。

第 2 章介绍了引力波数据实时处理系统，并提出了改进版本的实时处理系统，为后文的各个子模块的介绍奠定了基础。

第 3 章为探测器表征中的主频道事件否决。先简单介绍了探测器表征，以及传统的主频道时间否决算法。然后引入模式识别方法进行否决，通过与传统算法的比较，验证了模式识别方法的优良性能。

第 4 章首先介绍了为 Burst 类型引力波信号实时监测开发的工具，然后按照第 3 章中引入模式识别方法的思路，建立了对 Burst 类型引力波信号的有效筛选。

第 5 章介绍了如何改进 Burst 类型引力波信号能量恢复，并与改进前进行了比较，分析了各自的优劣。

第 6 章为结论与展望。概括总结了本文的主要研究成果，并分析了未来可能的研究方向。

第2章 引力波数据实时处理系统

本章主要讨论的是引力波数据实时处理系统的总体结构。本章将首先介绍在 LIGO 中实际使用的数据实时处理系统，实时处理的意义以及 Advanced LIGO 时期数据实时处理可能遇到的问题，在最后提出了一个改进版的实时处理系统。

2.1 LIGO 数据实时处理

2.1.1 LIGO 现有的数据实时处理系统

尽管 LIGO 把引力波数据分析分成了四个大组，但是事实上每个组的引力波数据实时处理系统在结构上差别不是很大，只是具体的分析算法不同。因此，本章中将以 Burst 类型，即短时脉冲类型引力波数据处理系统作为代表性的例子加以介绍。

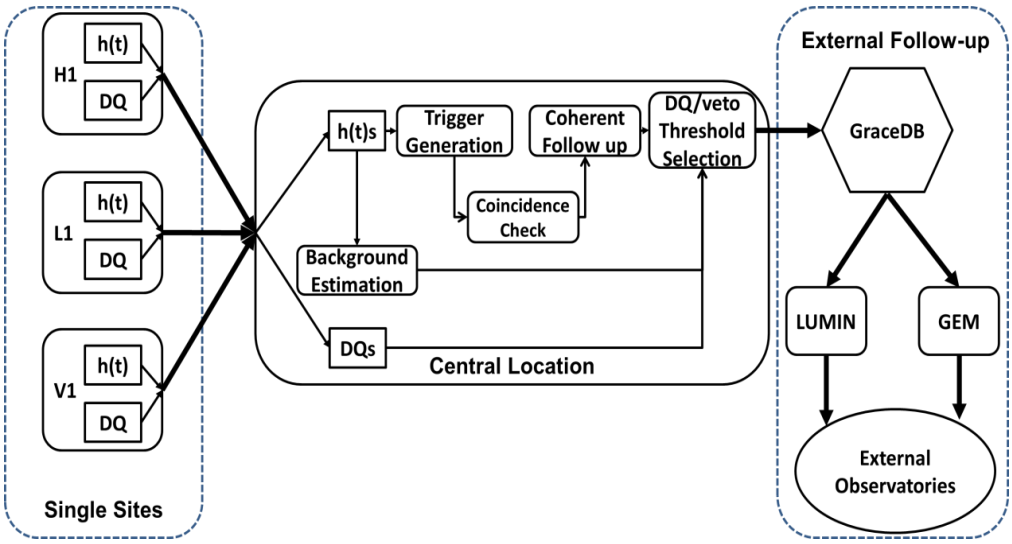

图 2.1 Burst 短时脉冲类型引力波数据实时处理现有系统示意图

如图 2.1 所示，Burst 类型引力波数据实时处理系统大致由三大部分组成：单天文台探测器（Single Sites）、引力波数据处理中心节点（Central Location）和传统电磁后续跟踪（External Follow-up）。接下来分别对这三部分进行详细介

绍。

2.1.1.1 单天文台探测器

目前LIGO与VIRGO合作，共享彼此的探测器数据，如图 2.1中所示的LIGO的H1和L1探测器以及VIRGO的V1探测器。之所以共享探测器数据，是因为探测器数量越多，越有可能找到引力波并确定其在天空中的方位。如图 2.2 所示，点 S 代表一个引力波源，由它传播至三个天文台的时间可以确立三个时间延迟（其中有两个独立），每个时间延迟可以确立一个引力波源可能存在其上的圆环，那么这三个圆环的交点即是引力波源的位置 S 或者是它的镜像S'。因此若想探测到引力波源的方位，至少三个探测器是必要的条件。且探测器数量增多的另一好处是可以提高信噪比（Signal-to-noise Ratio），提高引力波信号被探测到的几率。

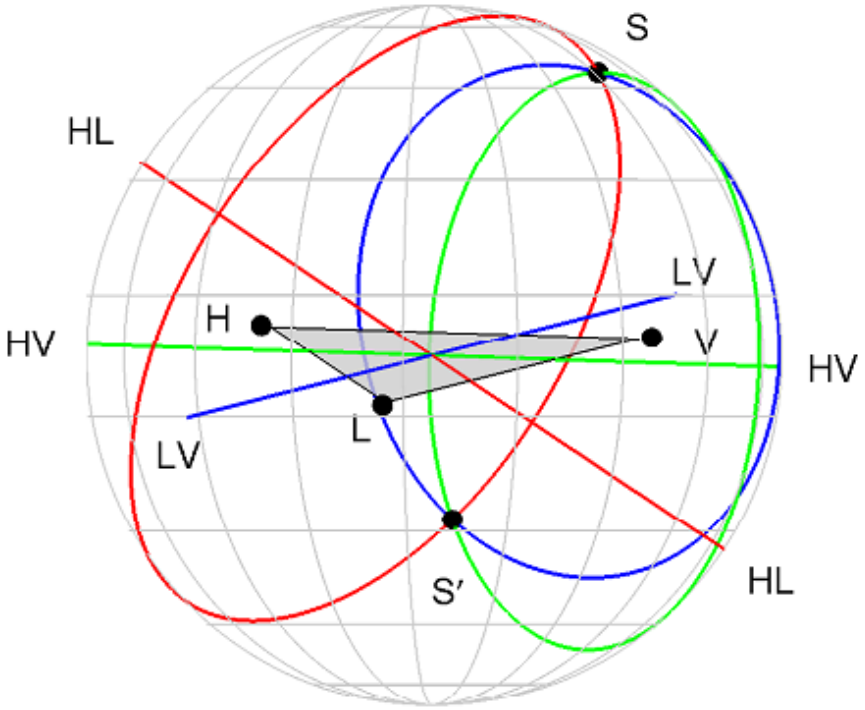

图 2.2 三个引力波探测器定位原理图[24]

在三个天文台上各自实时不断生成h(t)数据（即引力波主频道数据）和 DQ 数据（即数据质量数据），这些数据将会被传送到数据处理中心节点进行集中分析。

2.1.1.2 数据处理中心节点

三个天文台探测器的引力波主频道数据h(t)s在中心节点汇总后，开始进入正式的信号处理流程，以 Burst 类引力波数据分析实际使用的信号处理软件 Omega Pipeline（简称 Omega）[25]为例，它的信号处理流程大致分为三步：

1. 事件生成（Trigger Generation）

Omega 通过自定义的 Q 基底（Q Basis）模板[33]，进行匹配滤波，对每个探测器一定时间长度（实际中是每 64 秒）的h(t)数据进行多分辨率的时频分析，最终在时频平面上将一些显著性较强的时频块聚类成一个或多个事件，也称为 trigger。这些事件可能就代表着引力波信号。具体的事件生成过程如图 2.3 所示。子图(a)显示的是时域下的引力波数据，即信号和噪声的叠加，图中红色部分均为噪声，把蓝色的信号完全遮盖；经过多分辨率的时频转换后得到子图(b)所示的时频平面；取子图(b)的局部进行放大得到子图(c)；可发现子图(c)中是由众多时频块叠加而成，颜色越深，代表此处能量越高；将能量较低的时频块置为绿色，得到子图(d)。信号一般在时间和频率上是连续的，因此“相邻”（相邻有多种定义，字面上的相邻定义是最简单的一种）的能量较高的时频块聚类后就应该代表了一个信号，再经过一些筛选后，就得到了一个所谓的事件。

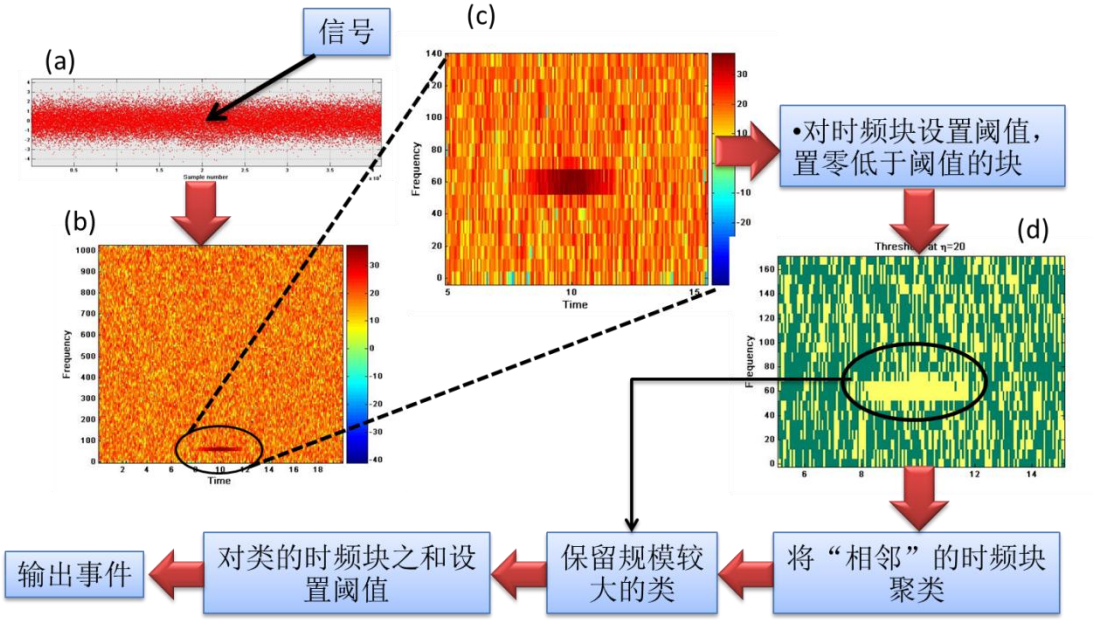

图 2.3 事件生成步骤示意图

2. 一致事件检验（Coincidence Check）

对于探测器 A 上的事件 1 和探测器 B 上的事件 2 而言，若事件 1 和事件 2 在时频平面上存在一定的重合（overlap），那么就可以认为事件 1 和事件 2 是一个一致事件，即可能由同一个引力波诱发产生；类似地，此定义可以推广到多个探测器。由于不同探测器上的噪声相互独立，因此两个探测器之间的一致事件就比单独的一个事件是引力波的概率要大，多个探测器上的一致事件更是如此。

3. 一致详细跟踪（Coherent Follow up）

为了进一步获取一致事件的信息，需要对其进行一致事件的详细跟踪，获取此一致事件的天空方位，距离，频率，能量等信息。该部分是信号处理流程计算量最大的部分，也是最不能满足实时在线分析要求的部分。

在以上的信号处理流程不断进行的同时，名为背景预测（Background Estimation）[34]的流程也在同步进行。背景预测的目的在于估计某个时间段内探测器背景噪声的强度。由于噪声的影响，探测器主频道的事件率很高，因此仍然有很多的噪声事件可以通过一致事件检验，最终引发误警（False Alarm）。为了否决这些由噪声产生的一致事件，就需要对这些一致事件所在时间段的噪声水平进行估计。背景预测利用人工时移（Time Shift，也被称为 Time Slide）的引力波主频道数据来估计噪声水平，如图 2.4 所示。

图 2.4 中显示的是一特定时间段$[T_1, T_2]$内，时移前后探测器 H1 和 L1 中一致事件的变化。图中每一个方块代表着一个事件，绿色虚线方框中囊括的方块表明此时刻有一个一致事件发生，用红色轮廓空心方块标记构成一致事件的 H1 和 L1 事件。在时移发生后，即 L1 探测器数据沿时间轴向右平移一段距离后，可以发现 H1 上的红色轮廓空心方块与 L1 上一个蓝色实心方块重新构成了一个一致事件。假设红色轮廓空心方块代表的事件由同一个引力波产生，那么经过时移后红色轮廓空心方块代表的事件和其他事件构成的一致事件就一定和该引力波无关（前提是时移的长度必须大于光在两探测器之间传播的时间），而可以认为是由噪声引发。也就是说对主频道数据进行人工时移后所产生的一致事件都可以认为是噪声，对这些噪声一致事件进行一致详细跟踪后计算出的一致事件能量就部分代表了该特定时间段$[T_1, T_2]$内噪声潜在的强度，这样经过时移后得到的一致事件被称为偶然重合事件（accidental coincidence）。进行多组人工时移则能得到该特定时间段内噪声的强度分布。显然人工时移的次数越多，得到的

噪声强度分布越准确。实际使用中，LIGO 分配了 100 个独占的 CPU 进行 100 组人工时移的背景预测，但是远远不能得到准确的噪声强度分布，这也属于第 1 章中谈到的数据分析精度受限的一种情况。

在未进行时移的主频道数据上得到的一致事件的能量如果落入该噪声强度分布（即比最强的噪声能量低），那么就有理由认为此一致事件由引力波产生的概率很小；或者可将偶然重合事件在时间上的到达看作泊松分布，于是根据偶然重合事件可计算出偶然重合事件到达率（也被称为 False Alarm Rate，误警率），若误警率过高，则在未时移情况下找到的一致事件也很值得怀疑；或者一致事件落入数据质量不可靠的时间段，此一致事件是引力波的可信度也就大大降低。这些就是图 2.1 数据处理中心节点部分里的 DQ/Veto Threshold Selection 所代表的含义。若有一致事件能够通过数据质量（对于实时在线分析而言，通常使用第一类和第二类数据质量）和噪声分布筛选，那它就成为了一个引力波候选事件（Gravitational-wave Candidate Event），可以进入后续跟踪环节。

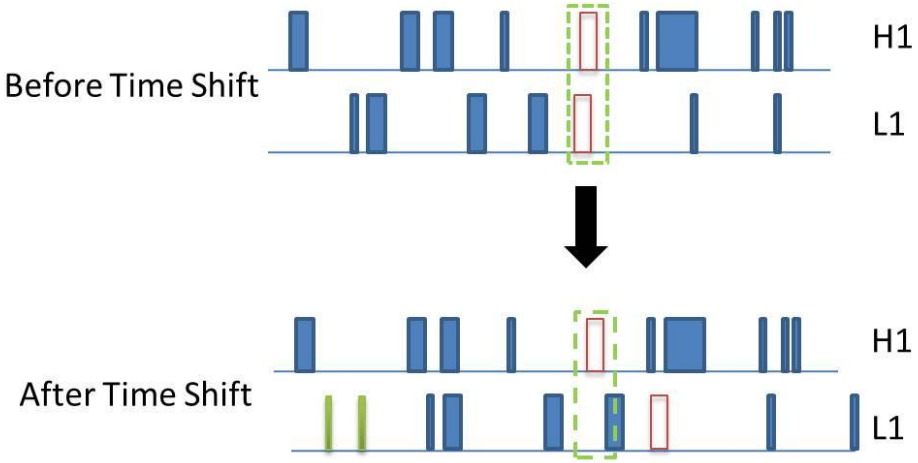

图 2.4 背景估计中时移示意图

2.1.1.3 电磁后续跟踪

后续跟踪的主要目的是用传统的电磁观测手段验证引力波候选事件，因为引力波源除了辐射引力波外，一般也会辐射大量电磁波，若在引力波候选事件所在的方位发现了传统电磁波现象，那么此引力波候选事件由引力波源产生的概率就大大增加。

数据处理中心节点捕获到的引力波候选事件将送入引力波候选事件数据库（Gravitational-wave Candidate Event DataBase，简称 GraceDB）[26]。

GraceDB 中的引力波候选事件会被 LUMIN（LoocUp Management & INterface）[27]和 GEM（Gravitational-to-ElectroMagnetic wave processor）[28]两个软件实时监视，它们对每个候选事件的各种属性进行分析后，根据候选事件所在的方位，距离和引力波源类型等信息挑选观测效果最佳的电磁观测方式来验证引力波候选事件的方位上是否有相应的引力波源存在。LUMIN 主要负责光学望远镜和射电望远镜的后续跟踪，而 GEM 则主要负责 x 射线天文卫星的后续跟踪。

2.1.2 实时处理的意义

在 LIGO 的第六次科学运行中，实现了延迟在十分钟量级的引力波数据处理。引力波数据实时处理的意义主要体现以下两方面。

2.1.2.1 快速的探测器诊断

引力波探测器的规模巨大，例如 LIGO 的 H1 和 L1 探测器的两条真空管道的长度均为 4 公里，而探测器又由许多精密的仪器构成，因此探测器常常出现各种故障。而探测器的故障将很快地反映在探测器收集的数据中。因此快速地从探测器数据中获知目前探测器的状态可以帮助科学家和工程师尽早地诊断故障并消除故障。如图 2.5，信号处理软件 Q-scan 的输出结果在时间上不连续，露出底部网格的部分即是探测器发生严重故障的时段。有输出结果的部分上分布的园点则代表着事件，星号代表的则是这 24 小时内能量最强的一个事件。天文学家能够根据这些事件的分布情况判断探测器的状态。

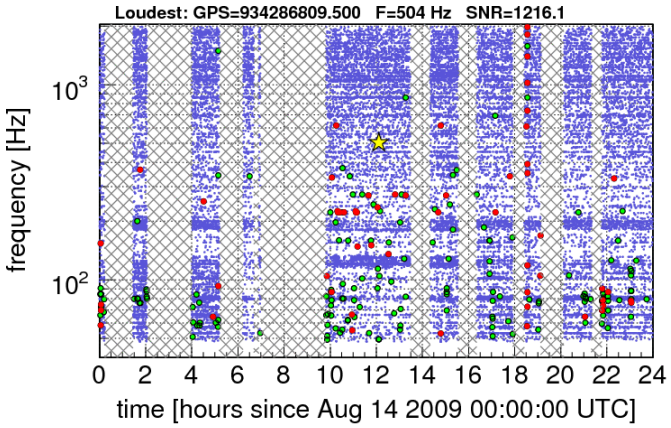

图 2.5 短时脉冲类型引力波信号处理软件 Q-scan[33]的输出结果

2.1.2.2 快速支持电磁后续跟踪验证

图 2.6 显示了四种典型的引力波源在各种观测波段上的持续时间长短。这四种典型的引力波源是：中子星融合（Neutron Star Binary，简称 NS/NS），核坍缩超新星爆炸（Core Collapse Supernova，简称 CCSN），长时间伽马射线爆发（Long-duration Gamma Ray Burst，简称 LGRB）和软伽马射线复现源（Soft Gamma Repeater，简称 SGR）。从图 2.6 可知，在伽马射线波段（γ），紫外线/X 射线波段（UV/X），光学波段（optical）和中微子辐射（ν）中均有持续时间不超过分钟乃至秒量级的四种典型引力波源。也就是说对于这些引力波源而言，若想在电磁后续跟踪验证中得以观测到，就要求前端的引力波数据处理提取出候选事件必须十分迅速，延迟不能超过若干分钟甚至是若干秒。

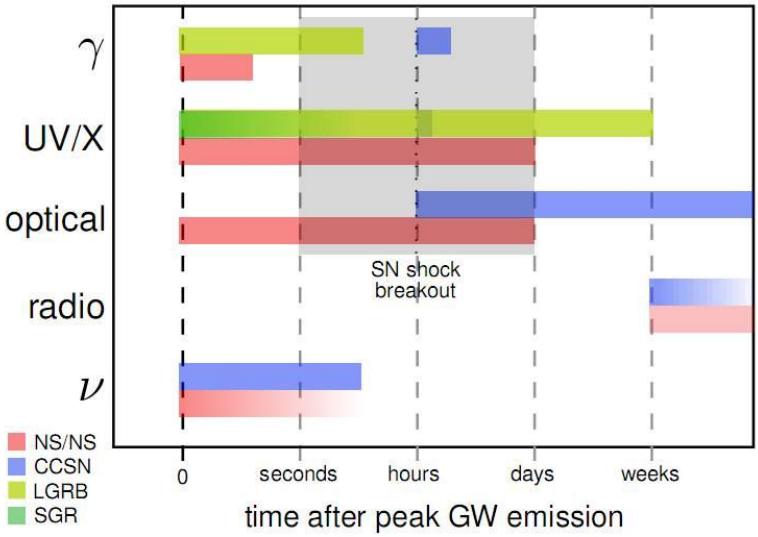

图 2.6 典型的引力波源在各个电磁观测波段上的持续时间示意图[29]

另一方面，快速的引力波数据处理意味着数据处理精度所受的限制可以得到缓解。例如信号处理流程中的第三步，一致事件详细跟踪中需要计算出一致事件所在的天空方位。若采用球坐标系，那么天空方位将由与 x 轴的夹角θ和与 z 轴的夹角φ决定。由于天空是连续的，θ与φ的组合是无穷的，由于计算量的限制，目前只能离散地选取一千余组θ与φ的组合，分别计算其是真正天空方位的概率，最终得到一张星空图（Skymap），如图 2.7，白色符号×所标记的是概率最高的θ与φ组合。若引力波数据处理能够更快速地进行，就能选取更多的θ与φ组合进行计算，从而提高方位精度。

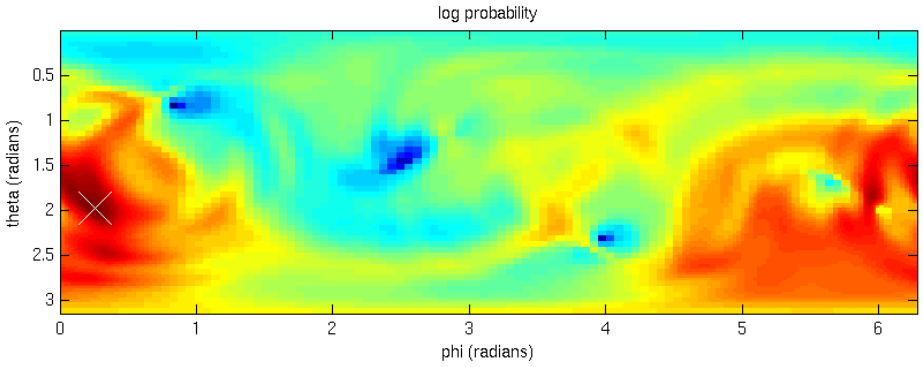

图 2.7 引力波候选事件所在天空方位概率分布图

2.1.3 Advanced LIGO 的挑战

首先，与第一代 LIGO 相比，Advanced LIGO 在探测灵敏度方面大约是前者的十倍。而由于探测器灵敏度与可探测的距离成正比，因此这就意味着 AdvLIGO 将能探测到比目前远十倍的引力波源。如图 2.8 所示，由于探测半径是原来的 10 倍，假设引力波源均匀分布的话，AdvLIGO 的引力波事件率将是目前的 1000 倍，那么对于某些引力波源而言，就可能一天就能探测到一个引力波事件。

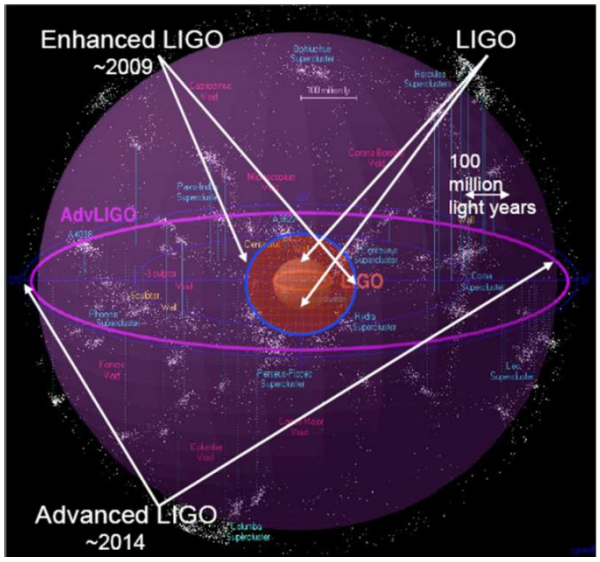

图 2.8 第二代激光干涉引力波探测器在灵敏度上的提高[10]

表 2.1 显示是 Advanced LIGO 中引力波主频道数据流的大小，尽管每个探测器上 10 兆每秒的数据流今天看来并不算大，但是除了引力波主频道外，还有

数以千计的辅助频道，若它们的采样频率与主频道相同的话，最终每个探测器的数据流可轻易超过百兆每秒。

表 2.1 估计的 Advanced LIGO 主频道数据流规模[11]

时间跨度	单主频道	三主频道	三重备份
一秒	10MB	30MB	57MB
一年	316TB	592TB	1.8PB

且第 1 章中已经介绍，到 2015 年 LIGO 和 Virgo 的第二代探测器将开始运行，澳大利亚的 LIGO Australia 届时也能运行，而日本的 LCGT 也有计划于 2015 年开始运行。那么届时将有至少 5 个引力波探测器同时收集数据，也就是说最终总的数据流规模可能超过 1GB 每秒，因此数据处理中心节点的数据 I/O 负担很重。在 LIGO 的第六次运行中，大部分辅助频道的信息并没有被传输到数据处理中心节点和利用。

其次，引力波信号分析算法的计算量与探测器个数紧密相关。例如之前提到的一致事件检验，三个探测器时，需要检验的组合个数为$C_3^2 + C_3^3 = 4$；5 个探测器时，则为$C_5^2 + C_5^3 + C_5^4 + C_5^5 = 25$；6 个探测器时（考虑加入德国和英国合作的 GEO600 探测器），则组合个数为$C_6^2 + C_6^3 + C_6^4 + C_6^5 + C_6^6 = 57$。即使摩尔定律成立，到 2015 年，处理速度最多增长 8 倍，勉强赶上计算复杂度的增加。

最后，理论上，Advanced LIGO 探测到引力波的可能性非常大，毕竟引力波事件率届时可以达到每天若干个的水平。一旦探测到引力波，引力波探测器的目标将不再是探测，而是研究引力波的具体属性，若两个引力波到达时间相距较近，则可能由于数据处理延迟而丧失后续电磁观测的机会，因此在 AdvLIGO 中更需要快速的引力波数据处理。

2.2 改进后的数据实时处理系统

2.2.1 改进思路

根据引力波数据实时处理中的实际问题，改进的方向应遵循以下几条原则：

1. 不影响数据处理结果的前提下，尽量减少向数据处理中心节点的数据传输

2. 支持快速便捷的探测器诊断

3. 尽量提高实时处理的速度和精度

2.2.2 改进后的系统

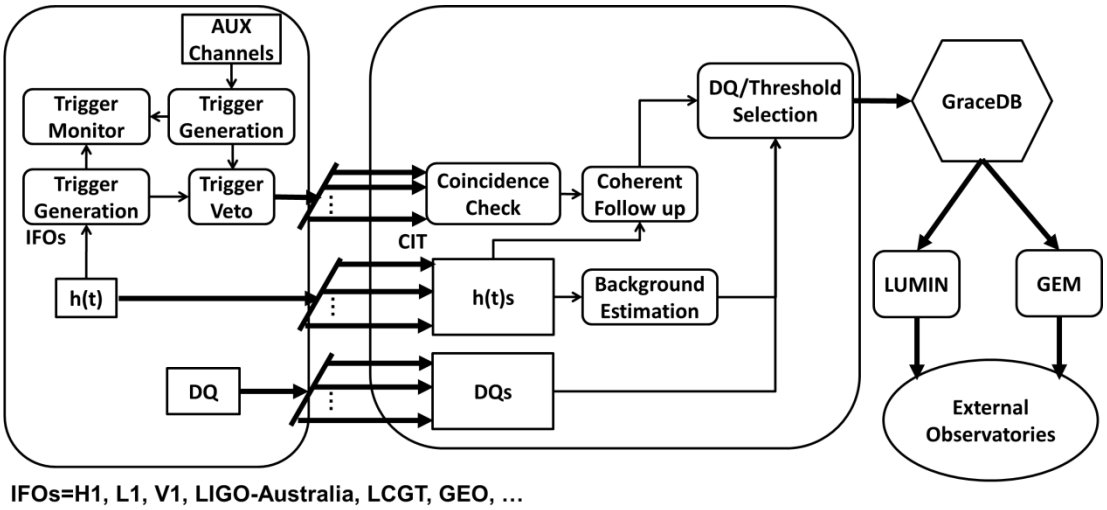

图 2.9 改进后的引力波数据实时处理系统

可见，与图 2.1 所代表的原系统相比，图 2.9 所示的新系统主要在单天文台探测器和中心节点部分进行了改进，接下来依次介绍。

1. 分散的事件生成

在新系统中，信号处理流程中的第一步，即事件生成被挪到了单天文台探测器上，也就是说离数据流更加靠近，延迟更小。而生成的事件将被事件监视器（Trigger Monitor）实时监视，从而为探测器诊断提供实时信息。尽管由于把事件生成挪到了单天文台探测器上而需要将生成的事件传输到中心节点，但由于用于存储事件的文件大小仅为h(t)数据大小的几百份之一，因此这部分多传输的数据可以忽略不计。

2. 基于辅助频道信息的事件否决

值得注意的是，对于辅助频道（AUX Channels）而言，也可提取出一系列的事件，而这些事件的信息将有助于判断从h(t)数据中提取的主频道事件是否是噪声。因此在单天文台探测器上生成事件后，系统将会依据辅助频道事件对主频道事件进行噪声判定并否决。只有未被否决的主频道事件才能被传输到数据处理中心节点进行一致事件检验。事实上，主频道事件大部分为噪声，因此若

否决能够去除大部分噪声事件，那么数据处理中心节点上找到的一致事件将大大减少，于是中心节点上的计算负担也就大大减轻了。且同时实现了不传输辅助频道数据至中心节点、利用辅助频道数据和降低误警率的三重目的。此部分以及事件监视器的内容将在第 4 章中详细讨论。

另一方面，主频道事件否决中判定为噪声的事件将有助于天文学家进行探测器硬件诊断和噪声源研究，因此主频道事件否决也需要实时在线运行。这部分内容将在第 3 章中详细介绍。

3. 改进的事件能量恢复

在信号处理流程的事件生成与一致事件详细跟踪中，都面临这样一个问题，即用阈值来分隔噪声与引力波信号的话，就可能有部分强度较弱的引力波信号被误判为噪声。为此，就需要增强这部分较弱信号的能量，途径则是在信号处理中恢复出更多的能量。这部分内容将在第 5 章中详细讨论。

2.3 本章小结

本章对引力波数据实时处理系统目前的状况进行了详细的描述，并介绍了未来升级后的引力波探测器给引力波数据实时处理的新要求。本章提出了一个改进版的引力波数据实时处理系统，将事件生成分散至各天文台探测器上，并通过基于辅助频道信息的否决机制减少数据传输以及中心节点上的计算量。具体的改进细节将在之后的第 3，第 4 和第 5 章中详细阐述。

第3章 引力波探测器表征中的信号否决

本章主要探讨探测器表征描述中的噪声信号否决机制，将从探测器表征概述、引力波信号否决算法设计、算法验证及算法在线实现等四大方面来介绍模式识别方法在信号否决中的应用。

3.1 引力波探测器表征概述

探测器表征（Detector Characterization，简称 DetChar）旨在理解和研究引力波探测器的各种详尽的物理性质，譬如探测器的响应函数，时间同步稳定性以及探测器噪声特性等。对物理性质的详尽准确的了解能够为后续的引力波数据处理提供可靠的校正和分析基础。它主要分为噪声源研究、数据质量标记、噪声能谱分析、实时数据监控、数据标定和时钟同步等若干部分。

3.1.1 噪声源研究

为了更好地对探测器性质进行研究，LIGO 在探测器内部以及周围环境中部署了上万个传感器。在探测器运行期间，这些传感器实时采集的数据将被实时监控和处理。每一个这样的传感器采集的数据流被称为一个频道（channel）。这些所谓的频道被统称为辅助频道（auxiliary channels），用于和引力波探测器主频道（Gravitational-wave Channel）进行区分。它包含两大类：设备频道（Instrumental Channels，简称 INST Channels）和环境监测频道（Physical Environmental Monitoring Channels，简称 PEM Channels）。典型的 INST Channels 包括激光锁频、镜面悬浮和定位等探测器模块中伺服系统上的传感器。典型的 PEM Channels 包括一系列监测周围环境的传感器，例如测量地面震动的地震检波器和加速度传感器；监视声学噪声的麦克风；监视影响线圈和电子设备磁场的磁力计；监测接近激光调制频率的无线电的接收机；监视交流电压波动的电压传感器[30]。

显而易见的是，由于普遍的设备故障以及环境扰动，引力波探测器主频道中收集的不仅仅是引力波，更多的情况下是各式各样的噪声。而 LIGO 部署的这些传感器则能捕捉到产生这些噪声的噪声源。譬如在 2001 年 2 月 28 日上午 11 时，美国华盛顿州发生了一次里氏 6.8 级的地震，同样位于华盛顿州的 LIGO

Hanford 天文台捕捉到了这次地震[31]，如图 3.1 所示。图 3.1 左侧显示的是引力波探测器主频道中由地震波诱发的突变信号，而图 3.1 右侧则显示的是探测器的地震检波器捕获到的信号，两者在时间上基本吻合。

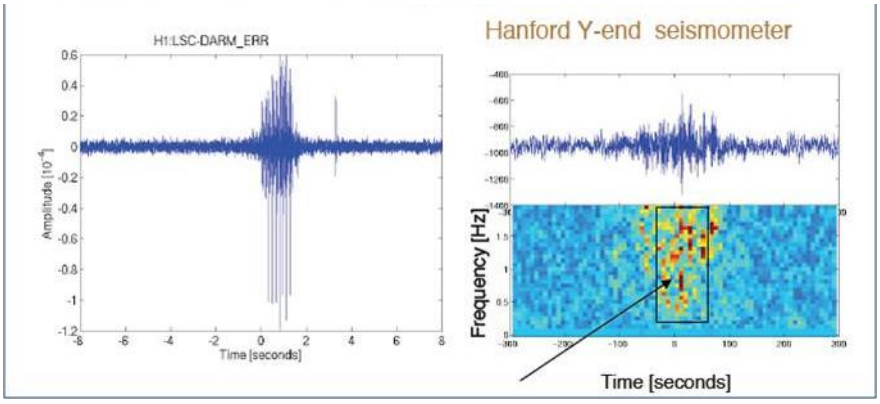

图 3.1 汉福引力波天文台探测到的地震波

由于引力波极其微弱，而噪声相对强烈，显然在引力波主频道的数据流中，引力波信号被淹没在噪声之中。而这严重限制了引力波探测器的探测能力。如图 3.2 所示，黑色曲线代表的是 LIGO 第五次科学运行（S5）中探测器的最大灵敏度曲线；橘红色曲线则代表第六次科学运行（S6）的灵敏度曲线；红色倒三角表示的是目前已知的所有脉冲星理论上产生的引力波辐射的强度。可见，无论对于 S5 还是 S6，都仅有一个脉冲星的强度满足探测器的灵敏度要求。也就意味着目前 LIGO 探测器几乎无法探测到脉冲星所产生的引力波。

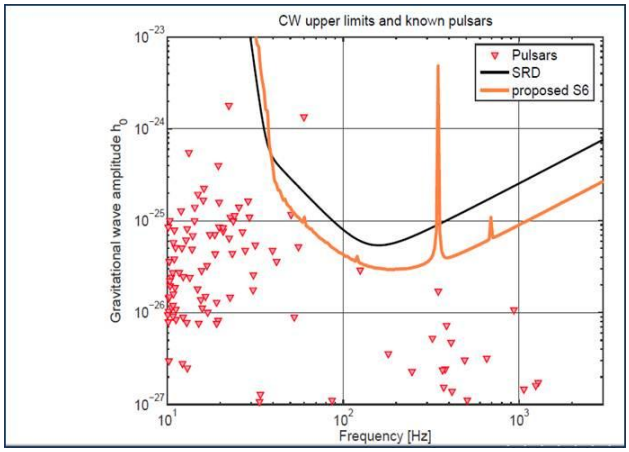

图 3.2 已知的脉冲星强度和引力波探测器灵敏度的比较

为了提高探测器灵敏度，消除噪声干扰是必要的。因此，对探测器噪声源的研究是探测器表征中最为重要的部分。

3.1.2 数据质量标记

在 LIGO 中，天文学家定义了一系列的数据质量标记[32]（Data Quality Flags，简称 DQ Flags），用其来描述探测器的运行状态和探测器主频道的数据质量。事实上，每一个数据质量标记都由一组时间区间组成。这些时间区间代表了影响数据质量的探测器故障和环境扰动所持续的时间范围。依照这些标记对数据质量的影响程度，LIGO 将其分为如下五大类。

- 第一类数据质量：当天文台探测器自身明显工作不正常的时间段，意味着这些时间段上的数据完全无法进行后续分析；
- 第二类数据质量：辅助频道与引力波探测器主频道间的耦合机制被完全理解的时间段，意味着这些时间段上的数据高度可疑；
- 第三类数据质量：辅助频道与引力波探测器主频道间的耦合机制被部分理解的时间段，意味着这些时间段上的数据质量不高；
- 第四类数据质量：辅助频道与引力波探测器主频道间的耦合显著性较低的时间段
- 第五类数据质量：较为特殊的一类，用于描述硬件仿真注射的时间段

总体而言，从第一类到第四类，所代表的数据质量逐渐好转。目前只有第一类数据质量和第二类数据质量实现了近似实时生成，且延迟在若干分钟量级。

数据质量标记与之前提到的噪声源研究有着密切关系，但不能将两者混为一谈。噪声源研究指导了第二、三、四类数据质量的标记，而数据质量标记则有效地缩小了噪声源研究中所需要关注的时间范围。

3.1.3 噪声能谱分析

噪声能谱分析主要研究引力波探测器主频道中的数据流的能谱特性，对随机背景引力波的探测有着尤其重要的意义。

3.1.4 数据运行支持

数据运行支持主要包括两方面。一方面对实时产生的数据流进行 24 小时不间断的人工监控，人工监控中就包含对数据质量进行标记的部分工作；另一方

面向引力波探测器注射人工制造的仿真引力波信号，目的是为了测试后续的引力波信号处理软件的可靠性，第五类数据质量即是为此目的而设立的。

3.1.5 数据标定

由于各种原因，引力波探测器生成的数据可能存在系统误差，因此需要对数据进行标定，尽量消除误差。

3.1.6 时钟同步

由于存在若干引力波探测器，而绝大部分引力波信号处理软件均需要对所有引力波探测器的同时段的数据流进行处理，因此如果各探测器间时钟同步不准确，将导致信号处理软件捕获到的事件的可信度大大降低。

3.2 引力波主频道信号否决相关软件及算法

3.2.1 引力波主频道信号否决简介

到目前为止，人们还未直接探测到引力波，尽管天文学家根据理论建模可以推测出部分类型引力波的信息，但实际上人们对于引力波的时域频域性质是不了解的。因此即使天文学家从引力波探测器主频道中捕获了一个信号，由于缺乏先验知识，是很难直接判断其为引力波的，只能先从判定其是否是噪声入手。并且由于引力波探测器的灵敏度限制，大多数情况下捕获的信号均是噪声，因此就需要对引力波主频道信号进行否决，从而降低误警率（false alarm rate）。

另一方面，数据质量标记往往将长度为数秒、数分钟的数据段标记为不可信，但这并不意味着剩余的数据段就是完全可靠的，仍然有大量的噪声充斥在这些数据段上。而随着 LIGO 天文探测器的每一次升级改造，数据质量标记的定义往往都需要修改甚至是重新设计。目前在 LIGO 探测器表征中，大概只有三分之一的噪声源是被物理学家所了解和掌握的，另外仍然有三分之二的噪声源处于未知状态，也就意味着数据质量标记是不充分的。基于以上三点原因，仅仅依靠数据质量标记来判断数据质量是远远不够的。在 LIGO 中，还需要对那些通过了数据质量标记筛选的数据段进行进一步分析。

因此在 LIGO 的探测器表征中引入了基于事件的否决机制（event-by-event veto），根据辅助频道中捕获到的事件的持续时间，把探测器主频道上对应时间

段的数据流标记为不可靠，那么探测器主频道中落入这些时间段的事件则被判定为有噪声引起。通常这样的时间段的长度在数百毫秒到 1 秒之间，被称为关联时间窗（coincidence time window）。换句话说一个探测器主频道上的事件的中心事件如果落入了以一个辅助频道事件中心时间为中点的巧合时间窗内，那么就认为这两个事件是关联的，均由某个噪声源诱发产生，因为被否决掉。

3.2.2 引力波主频道信号否决依赖的软件

只有从频道数据流中提取出事件，才能进行信号否决。KleineWelle[33]（简称 KW）就是一个这样的单天文台探测器频道事件提取软件。它基于二进小波变换[35]（Dyadic Wavelet Transform），将经过白化的频道时间序列投射到时移的多尺度基底上[33]，然后根据小波系数再对时频平面上若干局部化的信号能量进行阈值选择和聚类，最终聚类出的每一个类被称为一个事件。它具有良好的时频局部化特性，因此能有效地从任意探测器频道的数据流中捕获各种事件信号。

对于每个频道数据块，KleineWelle 的输出结果是一文本文件，并注册进数据库。每个文本文件均由一组 trigger 组成。所谓 trigger，即是代表一个事件的一组基本属性值。每个 trigger 有 8 个属性：类的开始时间（cluster start time），类的结束时间（cluster end time），由归一化能量加权的中心时间和中心频率，归一化和未归一化的类的能量，类所包含的基底的数量以及类的显著性（cluster significance，也被称为 KW significance 或 KW 显著性）。

3.2.3 引力波主频道信号否决现有算法

在 LIGO 的第六次科学运行中，主要有两种算法被用于引力波信号否决[36]：hveto[37]（hierarchical veto）和 UPV[38]（used percentage veto），且这两种算法都依赖于 KleineWelle 生成的 trigger。

3.2.3.1 Hveto

在 S6 的数据中，经过初步的数据质量筛选后，H1 和 L1 探测器上主频道的事件发生率集中在10^{-3}/秒的量级，且事件的持续时间均在若干秒的量级，因此我们可以合理地假设频道事件在时间上分布稀疏，对于探测器主频道和某个特定的辅助频道，给定长度为T的分析区间，设定关联时间窗的长度为T_{win}，则对于一个辅助频道事件而言，它与任意一个主频道事件存在 coincidence 的概率为：

$$P_{coin} = \frac{T_{win} N_{main}}{T} \tag{3-1}$$

其中N_{main}表示主频道上事件的数量。

于是对于辅助频道上数量为N_{aux}的事件而言，我们可以把其看成N_{aux}次 coincidence 试验，每次试验成功概率为P_{coin}的二项分布。由于事件发生率极低，因此概率P_{coin}甚小，可以看成稀有事件，根据泊松逼近定理[39]，可用泊松分布近似该二项分布。于是 coincidence 数量的泊松分布的概率分布函数如式 3-2 所示。值得注意的是，第二章中提及的误警率实际上与$\frac{\mu}{T}$相当。

$$\mathrm{PDF(x)}=\frac{\mu^{x}e^{-\mu}}{x!} \tag{3-2}$$

$$\mu=\frac{T_{win}N_{main}N_{aux}}{T} \tag{3-3}$$

Coincidence 发生的越多，表示主频道与辅助频道之间的关联越密切。在 hveto 中采用概率累积函数来描述这种关联显著性[40]：

$$S(x)=-\log_{10}(1-CDF(x-1)) \tag{3-4}$$

Hveto 的算法流程如下：

1. 选定一天时间范围内的引力波频道数据，对于辅助频道列表中的每一个辅助频道，使用不同的关联时间窗和 KW 显著性阈值，依照式 3-4，选择出与主频道 M 关联显著性最强的辅助频道 A；
2. 若 A 的关联显著性超过预先设定的阈值ε，则算法继续；否则，退出；
3. 否决掉所有与该辅助频道发生 coincidence 的主频道事件；
4. 将辅助频道 A 从辅助频道列表中删除后，返回步骤 1；

3.2.3.2 UPV

在 UPV 算法中，定义了名为使用率（used percentage）[38]的指标，用于描述一个辅助频道和主频道的关联性。

$$\text{Used Percentage}(\rho)=\frac{100\times N^{aux}_{coinc}(\rho)}{N^{aux}_{Total}(\rho)} \tag{3-5}$$

其中ρ代表设置在 KW 显著性上的阈值；$N^{aux}_{coinc}(\rho)$表示该辅助频道上与主频道有 coincidence 关系且 KW 显著性超过阈值ρ的事件的数量；$N^{aux}_{Total}(\rho)$则表示该

辅助频道上所有 KW 显著性超过阈值ρ的事件的数量。因此，很直观地，这个指标体现了一个辅助频道与主频道的相关程度。

UPV 的算法流程如下：

1. 选定一周时间范围内的引力波频道数据，设定关联时间窗为[-1s,+1s]，对于辅助频道列表中的每一个辅助频道，选定一个最低的 KW 显著性阈值，依照式 3-5，计算其使用率；
2. 若使用率低于 50%或$N_{coinc}^{aux}(\rho)$低于 10，则提升 KW 显著性阈值，依照式 3-5，计算其使用率，并停留在步骤 2；否则进入步骤 3；
3. 在 KW 显著性超过阈值的主频道事件中，否决掉所有与该辅助频道发生 coincidence 的主频道事件；

3.3 引力波主频道信号否决算法设计

3.3.1 基于事件的否决存在的缺陷

如前边所述，Hveto 和 UPV 都属于所谓的 event-by-event veto。虽然这两种算法能对主频道事件进行否决，但是它们也有一些缺陷。

首先 event-by-event veto 无法对主频道事件进行优先级排列。由 hveto 和 UPV 的算法流程可知，它们只能对主频道事件进行非黑即白的判断，要么被否决掉，要么被保留。在某些情况下，天文学家需要的是一个具有优先级的主频道事件排列。在引力波数据实时在线分析中，实时故障诊断是很重要的一个环节，天文学家可以根据否决算法中提供的优先级，对主频道事件所处的时间段进行依次排查，从而找出潜在的探测器故障或噪声源。

其次，无法对 hveto 和 UPV 设置惩罚因子。某些情况下，天文学家们对于否决算法的要求不同，有的希望一个严格的否决算法，对于主频道事件通过否决的要求较高，有的则恰好相反，这样就需要设置惩罚因子。而 hveto 和 UPV 由于没有优先级排列，无法设置惩罚因子。

再次，无法反映最新的物理特性。由于 hveto 和 UPV 的共同特点是以辅助频道为单位对辅助频道和主频道之间的相关性进行分析，若一个辅助频道被认为相关，则将该辅助频道所有事件作为基准对主频道进行否决，而 hveto 和 UPV 的分析是以天或周为时间单位，于是这里存在这一个前提假设，长度为天或周的时间内，该辅助频道和主频道间的物理相关性保持不变，否则用这种长时间

跨度的数据段进行一致分析是不合理的。但是这个前提假设是有问题的，因为天文台探测器所处的环境 极端复杂，天文台探测器的物理性质可能在任何时刻改变，辅助频道和主频道间的相关关系也可能随之改变。故此 hveto 和 UPV 的分析结果可能出现系统偏差。

最后，由于 hveto 和 UPV 分析所要求的时间单位过长，因此并不能直接用于实时在线分析。虽然通过对 hveto 和 UPV 的算法进行些许改造后能够用于在线分析，可又出现一个新问题，即对某一个主频道事件的否决结果前后不一致的情况会发生。比如现在 hveto 分析要求的时间单位为 86400 秒，为了满足实时在线分析要求，每隔十秒调用一次 hveto，即让 hveto 运行在[0,86400]，[10,86410]，[20,86420]等时间区间上，若在时间 86399 上存在一个主频道事件 E，则可能出现运行在[0,86400]上的 hveto 否决 E，运行在[10,86410]上的 hveto 不否决 E，运行在[20,86420]上的 hveto 又否决 E 的矛盾现象。

本文提出利用模式识别的方法来弥补上述四点缺陷。首先，引力波主频道信号否决问题从本质上可看做一个分类问题，因此理论上是可以转化成一个模式识别问题的。其次，它能有效地弥补 event-by-event veto 的不足。众多现有的模式识别方法支持概率输出，因此能够满足对主频道事件进行优先级排列的需求，也能对否决设置惩罚因子或权重。而由于转换成了模式识别问题，否决分析的最小单位不再是天或周这种时间跨度，也不是辅助频道，而是一个主频道引力波事件，于是就能避免第三和第四点缺陷。

3.3.2 模式提取

在本模式识别问题中，存在两类类别标签。第一类代表噪声不仅引发了主频道事件，而且诱发了某个或若干个辅助频道事件；第二类代表噪声仅仅诱发了某个或若干个辅助频道事件，而对主频道无影响。第一类类别被称为负样本（negative samples），第二类类别被称为正样本（positive samples）。接下来介绍如何提取正负样本的模式信息，建立训练集和测试集。

首先介绍负样本模式提取。第一章中曾经介绍过目前引力波探测器 90%置信度下的引力波事件率在 10^{-2}-10^{-1}/天，而在未经过数据质量筛选的 S6 实际数据中，探测器主频道事件率集中分布在 10^{4} 量级，可见绝大部分的探测器主频道事件均由噪声引起。因此实际上可以合理地把这些主频道事件全部看作由噪声引起，即使其中有若干个事件确实由引力波产生，对于整个负样本集来影响可以

忽略。

2.2 节中介绍了 KW 的 trigger 有 8 个属性，首先对引力波主频道和所有辅助频道上的 trigger 按照第三个属性，即归一化能量加权的中心时间进行升序排列。定义第 i 个频道上的第 j 个 trigger 的 8 个属性为向量：

$$\boldsymbol{a}_{ij}=[a_{ij,1},a_{ij,2},...,a_{ij,8}] \tag{3-6}$$

其中$\mathrm{i}=1,\ldots,\mathrm{n};\mathrm{j}=0,\ldots,N_i$。$N_i$代表第 i 个辅助频道中的 trigger 数量。当$\mathrm{i}=0$时，trigger 属于引力波主频道事件（Gravitational Wave Channel Trigger），如图 3.3 中的红色实心点。再定义第 m 个主频道 trigger$\boldsymbol{a}_{0m}$与第 i 个辅助频道中的第 j 个 trigger$\boldsymbol{a}_{\mathrm{ij}}$之间的距离为两者的中心时间之差：

$$D_{mij}=a_{ij,3}-a_{0m,3} \tag{3-7}$$

对于每个辅助频道，我们各选择一个 trigger，使得它们与第 m 个主频道 trigger 的距离的绝对值在各自的辅助频道中最小。于是我们得到了这些 trigger 各自在其辅助频道的序号，即$\boldsymbol{s}_m=[s_{m1},s_{m2},\ldots,s_{mn}]$。通常，若辅助频道事件与主频道事件存在关联的话，它们在时间上应该是比较靠近的。因此设置一个关联时间窗$[-T_1,T_2],T_1>0;T_2>0$排除时间上相隔较远的辅助频道 trigger。即如果值$D_{mi,s_{mi}}$没有落入区间$[-T_1,T_2]$，则设置$s_{mi}=0$。从向量$\boldsymbol{s}_m$代表的辅助频道 trigger 中提取每个 trigger 的后 5 个属性值并附上距离属性：

$$\boldsymbol{v}_{mi}=\begin{cases}[a_{is_{mi},4},\ldots,a_{is_{mi},8},D_{mi,s_{mi}}], & s_{mi}>0\\ [0,0,0,0,0,0], & s_{mi}=0\end{cases} \tag{3-8}$$

其中，当$s_{mi}=0$，就意味着在第 i 个辅助频道上不存在与主频道时间上有关联的事件，因此用全零向量替代。将这些向量拼接起来，最后我们就获得了第 m 个主频道 trigger 的模式或称为特征向量$\boldsymbol{f}_m$。值得注意的是，若辅助频道数量较多时，特征向量$\boldsymbol{f}_m$的维度可上万，成为高维向量。

$$\boldsymbol{f}_m=[\boldsymbol{v}_{m1},\boldsymbol{v}_{m2},\ldots,\boldsymbol{v}_{mn}] \tag{3-9}$$

可能会出现所有的$\boldsymbol{v}_{mi}$全部是零向量的情况，这就意味着目前所有的辅助频道与主频道均没有时间上的关联性，提取出来的特征向量无效，不会用于后续的模式识别中。这种情况可能意味着有潜在的噪声源并没有被当前辅助频道代表的传感器覆盖探测到，可以为天文台硬件诊断提供些许参考。

正样本的模式提取与负样本大同小异。由于没有真实引力波的先验信息，

因此和负样本一样，也只能依赖辅助频道的信息。在同一个关联时间窗内，若辅助频道有事件发生，而主频道没有，那么这就意味着事件源（可能是引力波，也可能是噪声源）对辅助频道产生了影响，但却对主频道没有影响，这恰好是与负样本相对立的。因此，我们首先随机生成一些 GPS 时刻，并且排除那些落入以任意一个主频道事件中心时间为中心的关联时间窗的 GPS 时刻，即落入 $\bigcup_{j=1}^{N_0}[a_{0j3}-T_1, a_{0j3}+T_2]$的 GPS 时刻被排除，这是出于防止生成出与负样本相同的正样本的考虑。然后将剩余的 GPS 时刻伪装成主频道事件的中心时间，按照负样本的生成规则，生成正样本。

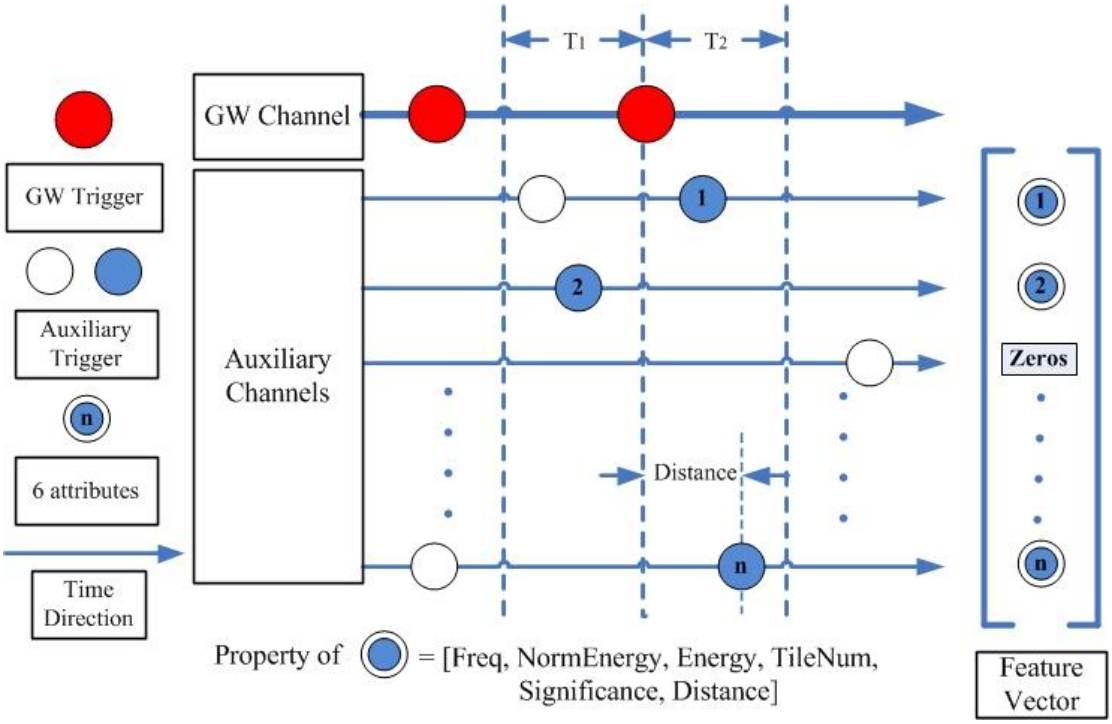

图 3.3 正负样本模式提取示意图

3.3.3 否决算法参数配置与流程

前边的模式提取中，牵扯到两个参数的选择：关联时间窗和正样本数量。

关联时间窗是一个很重要的参数，直接影响到特征向量的取值。首先要考虑的是选择对称窗还是非对称窗。由于事件源传递到主频道和引力波频道的时间不确定，因此非对称窗是不合适的。且在 hveto 和 UPV 算法中也都是选取了对称窗，故此选择对称窗$[-T_1, T_1]$。其次是T_1的取值。一般来说辅助频道事件与主频道事件时间上相隔过远的话，关联性就变的很弱，因此取值范围应在[0, 2s]

内。在 hveto 中，时间窗的取值是变化的，$T_1 = [0.05s\ 0.1s\ 0.2s\ 0.4s\ 0.5s]$，而在 UPV 中则是固定的，$T_1 = 1s$。出于对后边与 hveto 比较的考虑，$T_1$选择 hveto 中取值的中间值 0.25 秒。

负样本的数量取决于主频道事件的多寡，无法人工决定，但正样本的数量则需要考虑一下。一般而言，正样本与负样本数量大致平衡时，模式识别方法的准确性会比较稳定，因此正样本的数量选择与负样本数量一致。由于频道上时间的无差异性，按照均匀分布的方式随机生成正样本所需要的 GPS 时间。

此外，在负样本的生成中，还需要考虑对主频道事件和辅助频道事件进行预筛选，去除 KW 显著性极低的部分。这是因为 KW 显著性极低的事件几乎可以肯定就是噪声，故提前排除。采用 LIGO 第五次运行中常用的一组参数，主频道预先排除 KW 显著性低于 35 的事件，辅助频道则排除低于 10 的事件。

最后，生成样本时需要排除某些辅助频道。一般情况下引力波对于辅助频道的影响是可以忽略不计的，因此可以认为辅助频道上的事件均是由噪声引起。但少部分辅助频道受到引力波的影响较大，若也参与到样本生成，则可能导致真正的引力波信号被否决。因此，定义那些与引力波信号耦合关系不可忽略的辅助频道为不安全的频道（unsafe channels）[41]，必须在否决分析中予以排除。

基于模式识别的否决流程如下：

1. 读入时间长度为 T 的主频道和辅助频道事件
2. 根据数据质量和不安全频道的设置过滤部分事件
3. 根据 KW 显著性阈值再次过滤部分事件
4. 将长度 T 分为两部分：T_1和T_2，落入T_1的事件经模式提取后成为训练集；落入T_2的则成为测试集
5. 采用模式识别方法对训练集进行学习后，在测试集上测试性能

3.3.4 模式识别方法选择

模式识别方法大致可分为监督学习方法（supervised ）和非监督学习方法两大类，由于能够生成已知类别标签的样本集，故此选择监督学习方法。在监督学习方法中则需要选择能够支持概率输出或连续值输出的方法。监督学习方法的种类很多，不可能一一列举，且模式识别方法的选择在本章中并不是核心部分，且考虑到高维的样本，因此在这里选用两种最为常用且对高维分类支持较好的模式识别方法进行比较：支持向量机[42]（Support Vector Machine，简称 SVM）

与随机森林[43]（Random Forest）。

支持向量机本质上讲即是把数据按照某种映射关系转换到高维空间，并在高维空间中寻找一个最优分类面。给定由两类样本组成的数据集$\{\boldsymbol{x}_i, y_i\}$, i $= 1, \ldots,$ n, $y_i \in \{-1,1\}, \boldsymbol{x}_i \in \boldsymbol{R}^d$，那么支持向量机要解决的是一个二次优化问题[44]：

$$\min_{\boldsymbol{w},b,\xi} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{m}\xi_i, \qquad (3\text{-}10)$$

Subject to $$y_i(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_i) + b) \geq 1 - \xi_i,$$

$$\xi_i \geq 0, i = 1, \ldots, n,$$

其中数据$\boldsymbol{x}_i$经函数$\boldsymbol{\phi}(\boldsymbol{x}_i)$映射到一个高维空间；$C$是对训练错误设定的惩罚因子；$\xi_i$为松弛因子。由目标函数求得的$\boldsymbol{w}$和$b$可确定最优分类面为$\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}) + b$，对任意测试样本$\boldsymbol{x}$，可由$\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}) + b$的取值判断其的类别。

实现支持向量机的软件很多，但实际上大同小异，LIBSVM[45]是其中使用相当广泛的支持向量机软件包。本文中使用其最新的 C++3.1 版本。

随机森林本质上讲是一个包含多个决策树（例如 CART 分类回归树[46]）的分类器，输出的分类结果由这些决策树的结果共同决定，比如众数或者平均数。相比单个的决策树，随机森林具有更高的鲁棒性，不易受噪声影响。本文中使用名为 SPR（StatPatternRecognition）[47]的 C++软件包实现随机森林的功能。

由于 libsvm 与 SPR 均支持概率输出，取值范围在 0 与 1 之间。取值越接近 1，则表示主频道事件越不可能由噪声引起；越接近 0，则主频道事件越可能由由噪声引起。对该概率值进行阈值选择，那么就能灵活地对主频道事件进行严格或宽松的否决，且能得到模式识别中常用的接收者操作特性曲线（Receiver Operating Characteristic Curve，简称 ROC 曲线）[48]，并作为两者比较的基准。

支持向量机方法与随机森林方法均需设置一系列参数，在进行比较前，已经对这些参数进行了调优，以求用各自最佳的性能进行比较。

如图 3.4 所示，所有子图的横轴均为 false positive rate，即负样本中误判为正样本的百分比，也被称为误警率；纵轴的 true positive rate 则代表正样本中被正确判断为正样本的百分比。图 3.4 中，将两种方法在 969408000-969494400 和 969840000-969926400 两个 GPS 时间段（即一天的时间，86400 秒）的 H1 和 L1 探测器数据上测试了其在三种不同类别辅助频道上的性能（以其之前一天的数

据作为训练样本），即 INST 频道，PEM 频道以及两者的总和 BOTH 频道。总体上看，在 BOTH 与 PEM 辅助频道下，蓝色点划线与绿色实线代表的随机森林 roc 曲线往往爬升较快，且在低 false positive rate 的情况下均比 svm 的 true positive rate 明显要高，这就意味着随机森林能够以较低的误警率正确地判断正样本。

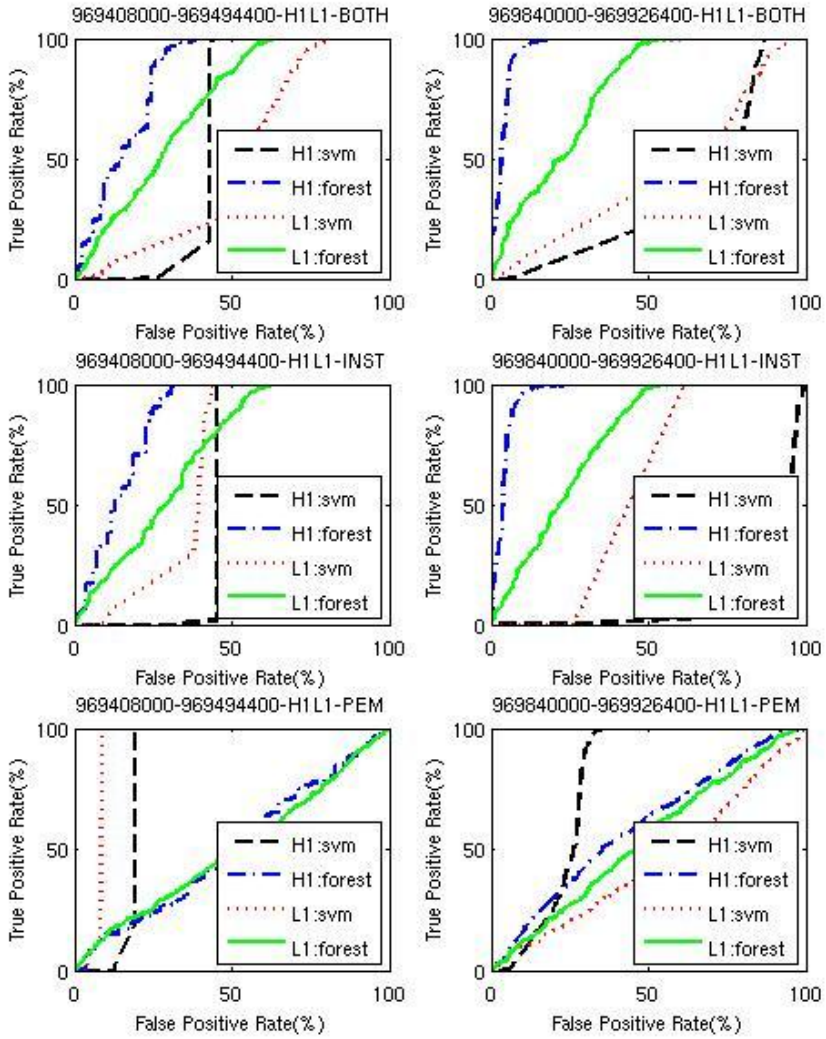

图 3.4 支持向量机与随机森林在 H1 和 L1 的不同辅助频道组合下的性能比较

而图 3.4 中的 PEM 频道情况下，红色点线和黑色虚线所代表的支持向量机 roc 曲线则在大部分情况下爬升较快，但在低误警率情况下，随机森林的性能并不比支持向量机差。PEM 下性能与 BOTH、INST 情况下的差异说明随机森林和

支持向量机的性能受数据类型的影响较大，而考虑到总体情况下随机森林的性能较好，且放弃一类辅助频道而进行信号否决不太合理，因此在后边的性能比较中，选用随机森林方法，且均以 BOTH 辅助频道为配置。

3.4 引力波主频道信号否决性能比较

3.4.1 性能指标

除了前边介绍的传统的 ROC 曲线外，LIGO 自己额外定义了一个性能指标，即死区时间-否决效率（dead time-efficiency curve）曲线。定义如下：

$$\text{Efficiency}(\rho) = \frac{100 \times N_{vetoed}^{GW}(\rho)}{N_{Total}^{GW}} \tag{3-11}$$

其中$N_{vetoed}^{GW}(\rho)$是负样本在概率阈值ρ下被否决掉的数量；N_{Total}^{GW}是所有负样本的数量，这个值体现了否决算法否决由噪声引起的主频道事件的能力。

$$\text{Dead Time}(\rho) = \frac{100 \times T_{vetoed}(\rho)}{T_{Total}} \tag{3-12}$$

其中$T_{vetoed}(\rho)$是在概率阈值ρ下被否决掉的时间，换言之是以所有被否决的主频道事件中心时间为中心的关联时间窗的并集；T_{Total}是否决前总的有效分析时间段的长度。

一个好的否决算法应该同时具有较高的否决效率和较低的死区时间，因此两者的比值常被用于评价否决的性能。比如对于数据质量标记而言，通常如果比值超过 5，就认为这个数据质量标记否决效果较好。

$$\text{ratio}(\rho) = \frac{\text{Efficiency}(\rho)}{\text{Dead Time}(\rho)} \tag{3-13}$$

3.4.2 与传统否决算法的比较

将基于随机森林的否决与 hveto 进行比较，为了保证比较的公平，两者在数据质量设置、预筛选中的主频道显著性阈值、关联时间窗大小、不安全辅助频道、最终的辅助频道设置和运行环境等方面完全保持一致，只有辅助频道显著性阈值设置由于 hveto 算法需要设置多个辅助频道阈值的的缘故而无法统一。

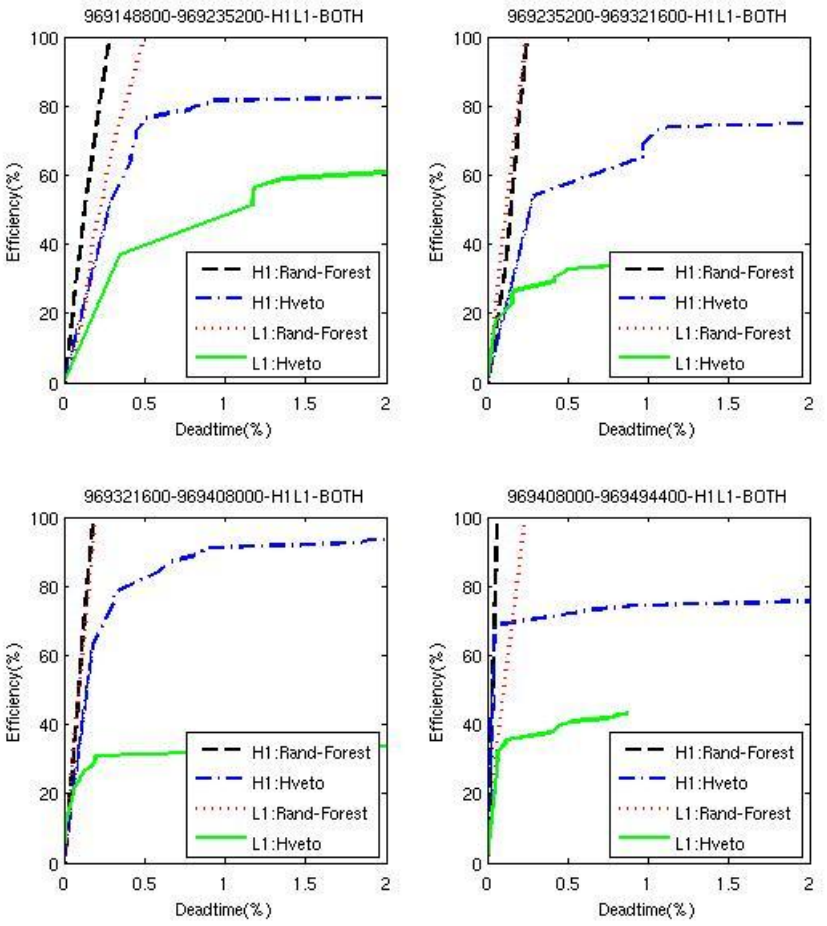

图 3.5 随机森林与 hveto 在死区时间-否决效率曲线之间的比较

图 3.5 中显示了随机森林与 hveto 在四个时间长度均为一天的数据集上测试的性能结果，由于 hveto 中没有正样本的概念，因此无法画出 roc 曲线，只能在死区时间-否决效率曲线与随机森林进行比较；也正由于 hveto 中没有正样本的概念，因此在随机森林计算 deadtime 的时候未考虑正样本被否决所造成的 deadtime，从而能够公平地与 hveto 的死区时间比较。每个数据集上，随机森林采用其前一天的数据作为训练集；hveto 由自身算法特点决定，不需要训练集。很明显地可以发现，无论是在 H1 还是在 L1 探测器的数据集上，黑色虚线和红色点线所代表的随机森林方法均比 hveto 的性能有着显著提高，几乎均在很低的死区时间下否决效率就接近了 100%。

此外 hveto 的曲线上值得注意的是，由于 hveto 是每次迭代过程中否决一个

辅助频道，因此 hveto 的曲线是一个很明显的折线图，上边的每一个明显的奇点代表着此处有一个辅助频道被判定为与主频道有显著关联，基本在经过头几个辅助频道之后，hveto 曲线的上升趋势就急速衰减。在右下角的子图中，甚至可以看到绿色实线的 hveto 曲线并没有像其他曲线一样向右延伸，这是因为 hveto 此时找到的关联最强烈的辅助频道已经小于 hveto 运行前设定的关联性阈值ε而退出了。这就说明在一段时间内，并不是所有的辅助频道和主频道关联都很强烈，应该只有若干辅助频道关联较强。

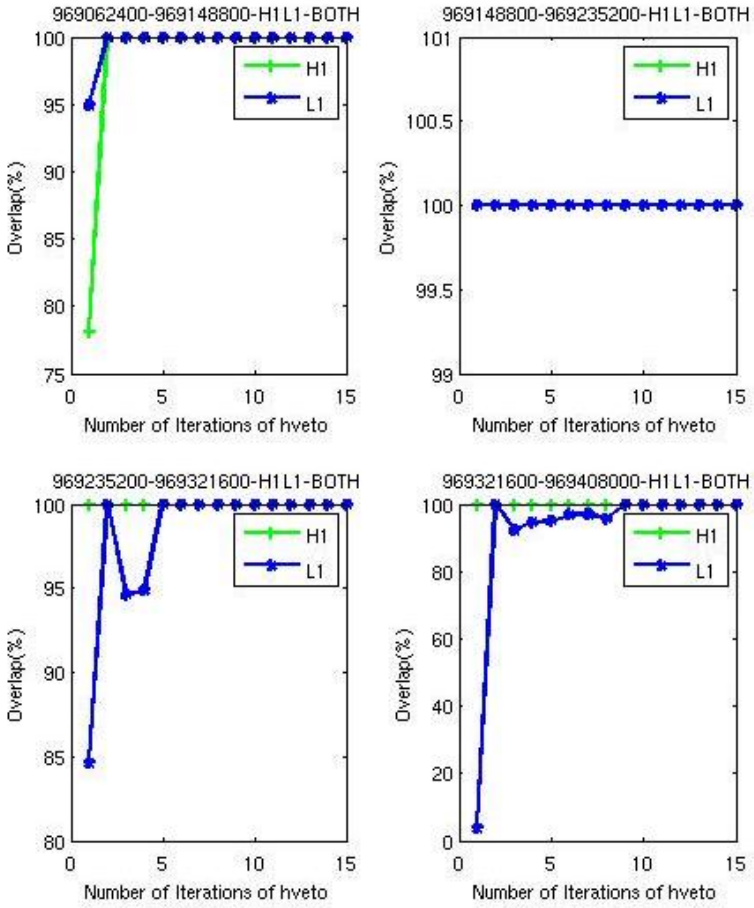

图 3.6 随机森林否决与 hveto 所否决的主频道事件组成的比较

图 3.5 初步表明了基于随机森林的否决在否决性能上较 hveto 否决优秀。然而我们需要进一步细粒度地考察随机森林否决的详细情况。为此，我们固定死区时间，比较基于随机森林的否决与 hveto 中被否决的样本的差异，如图 3.6 所示。

图 3.6 显示了在 4 个 GPS 时间段上随机森林与 hveto 在相同死区时间下所否决的主频道事件组成的差异。由于 hveto 每次迭代中只挑选一个关联性最强的辅助频道，因此一次迭代只能生成一个死区时间，于是以 hveto 若干次迭代所产生的死区时间作为基准进行比较。因此在图 3.6 中，横轴是 hveto 迭代的次数，纵轴为式 3-14 的取值。

$$\text{intersection}(\rho) = \frac{\bigcap(N_{hveto}(\rho), N_{rand}(\rho))}{N_{hveto}(\rho)} \tag{3-14}$$

其中ρ为死区时间；$N_{hveto}(\rho)$和$N_{rand}(\rho)$分别为 hveto 和随机森林在死区时间ρ下所否决的主频道事件的数量；符号$\bigcap()$则代表求交集运算，将$N_{hveto}(\rho)$和$N_{rand}(\rho)$中均被否决的主频道事件数量统计出来。若intersection(ρ)值较高，则表明 hveto 要否决的主频道事件，随机森林也大部分会否决。而图 3.6 中，可以明显地发现，在绝大多数情况下，intersection(ρ)接近 100%，也就相当于 hveto 否决的主频道事件集合是随机森林否决集合的一个子集。而 hveto 是 LIGO 中已经被证明幸而有效地否决方法，也就同时佐证了随机森林否决的有效性和随机森林相对 hveto 的优势。

最后讨论运行时间上的比较。随机森林否决中训练集和测试集的生成相关功能由 matlab 代码实现，随机森林学习和测试相关功能由 C++代码实现；hveto 则完全由 matlab 代码实现。为了公平起见，已将 hveto 中耗时的非否决功能的相关代码禁用（主要是生成各种中间过程图）。测试环境为 Linux CentOS5.3 系统，matlab2010b 版本，内存 8GB，随机森林与 hveto 各独占一个 AMD Athlon II 型 CPU 中的物理核。两者间执行时间比较如表 3.1 所示。

表 3.1 中，显示了基于随机森林的否决与 hveto 在十组长度均为一天的数据集上的执行时间及两者间的时间比，其中随机森林将每组数据集之前一天的数据作为训练集。hveto($\varepsilon = 1$)代表 hveto 运行在关联性阈值$\varepsilon = 1$的情况下，当目前迭代中辅助频道与主频道关联性最高值低于 1 时，hveto 运行结束。$\varepsilon = 0.1$时，hveto 运行时间延长明显。由于与主频道关联强烈的辅助频道只占少数，因此，随着阈值降低，hveto 执行时间会显著增加，随机森林相对而言也就更加迅速了。并且实际上，由图 3.5 可知，此时对应于高否决效率，hveto 否决的死区时间将大大超过基于随机森林的否决，甚至 hveto 可能根本就达不到高的否决效率，而使得两者间在高否决效率下的比较没有意义。

表 3.1 随机森林与 hveto 执行时间的比较

序号	随机森林	hveto($\varepsilon = 1$)	时间比	hveto($\varepsilon = 0.1$)	时间比
1	99.9	354.1	3.5	674.1	6.7
2	156.6	606.3	3.9	774.8	4.9
3	160.5	509.4	3.2	858.1	5.3
4	98.0	381.8	3.9	834.0	8.5
5	58.9	343.3	5.8	454.5	7.7
6	73.9	511.2	6.9	1556.0	21.0
7	95.1	410.6	4.3	835.7	8.8
8	97.4	217.5	2.2	257.1	2.6
9	166.4	923.5	5.5	1468.5	8.8
10	205.1	998.3	4.9	2012.0	9.8

3.4.3 不同引力波数据上的比较

第一章中曾经介绍过 LIGO 的数次科学运行，在此将随机森林否决在第五和第六次科学运行上的结果进行比较，此外对不同探测器上的性能表现也同时进行比较。

为了尽量保证比较的公平性，除了否决参数、辅助频道列表和运行环境上保持一致外，还必须考虑到一个客观上无法消除的 S5 和 S6 之间的差异性，即 GPS 时间不一致。为了尽量减少这种差异性，特地选择只有年份不同的 S5 和 S6 的时间段，以此尽量避免季节、地球自转差异等造成的影响。例如图 3.7 左上角的子图中 S5 测试集的时间段[842856000, 843256000]等价于太平洋标准时间 2006 年 9 月 21 日 23:39:46 至 2006 年 9 月 25 日 14:46:26；S6 测试集的时间段[969086402, 969486402]则等价为 2010 年 9 月 21 日 23:39:46 至 2010 年 9 月 25 日 14:46:26，训练集为各自时间段之前的 100000 秒，即[842756000, 842856000]和[968986402, 969086402]。

对比图 3.7 中的四张子图，我们可以发现绿色实线代表的 S6H1 的 roc 曲线总要比黑色虚线代表的 S6L1 的要好；同样地，蓝色点划线代表的 S5H1 的 roc 曲线总要比红色点线代表的 S5L1 的要好。此外，绿色实线代表的 S6H1 的 roc 曲线总要比蓝色点划线代表的 S5H1 的要好；同样地，黑色虚线代表的 S6L1 的 roc 曲线总要比红色点线代表的 S5L1 的要好。

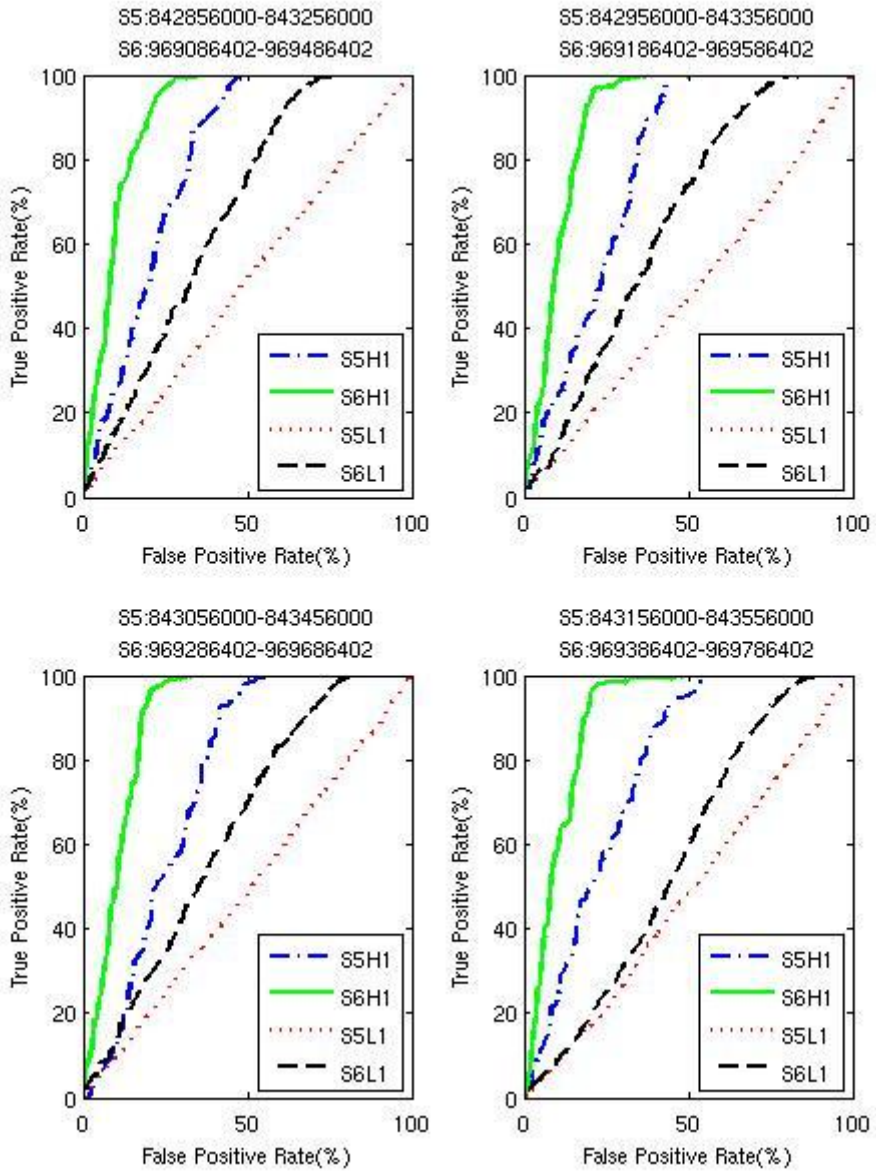

图 3.7 随机森林否决在不同探测器数据上的接收者响应曲线比较

这样明显的一致性透露出两个信息：对于同一个探测器而言，S6 的否决性能一致地比 S5 好；对于同一次科学运行而言，H1 的性能一直地比 L1 好。而这两个信息与实际恰好吻合。S6 相比 S5，灵敏度大幅提高；从第一章中提到的 LIGO 第五次运行探测器灵敏度曲线图，即图 1.3 可知，H1 的灵敏度是要比 L1 高些许的，尽管不太明显。由于灵敏度的提高，噪声的干扰降低，否决性能提升是符合逻辑的。这也从一个侧面佐证了随机森林的否决是正确有效的。

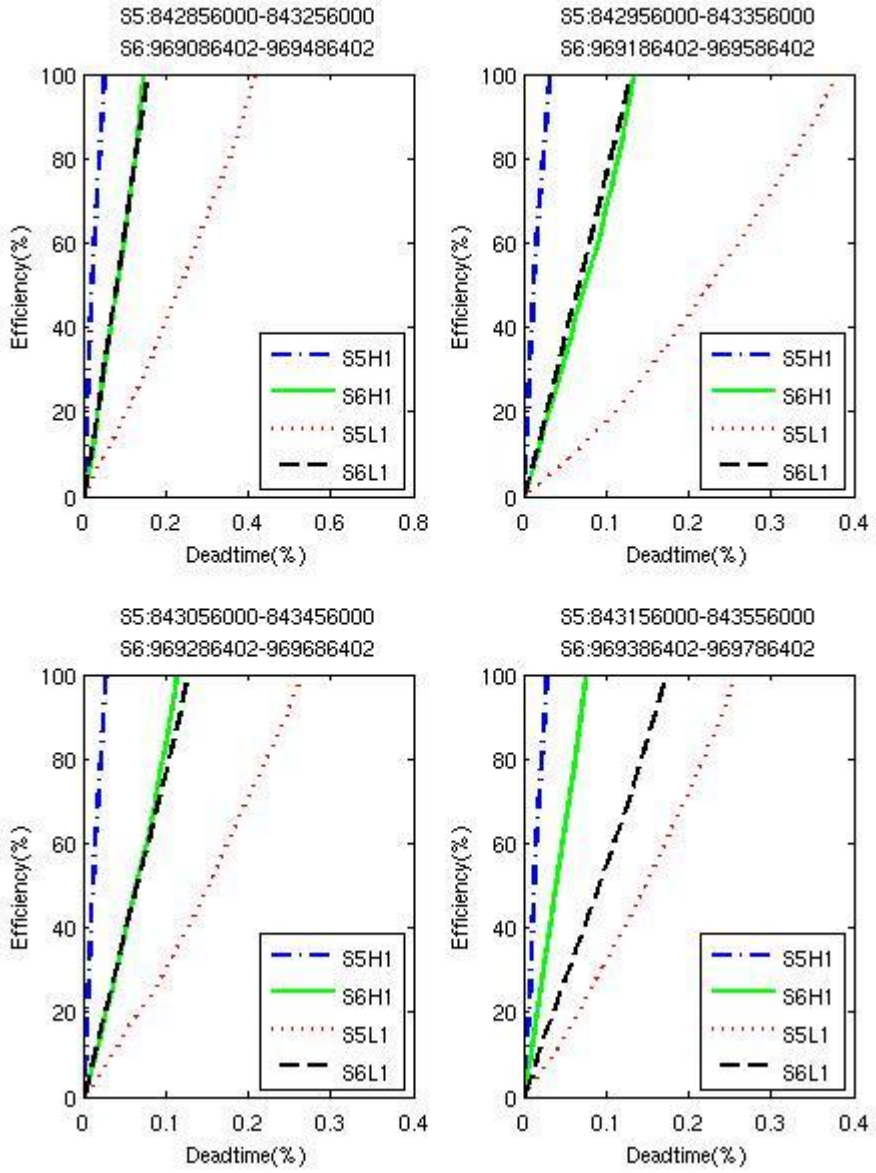

图 3.8 随机森林否决在不同探测器数据上的死区时间-否决效率曲线比较

图 3.8 和图 3.7 对应的随机森林否决是完全相同的，只是表现形式是死区时间-否决效率曲线，而不是 roc 曲线罢了。可以很明显地看出图 3.7 中的一致性在图 3.8 中不复存在，四张子图中爬升最快的不再是绿色实线 S6H1，而是蓝色点划线 S5H1；黑色虚线 S6L1 有时爬升地比绿色实线 S6H1 要快，表现的很随机。

解释这种不一致现象的关键在于理解死区时间和否决效率的定义。首先可明显地发现图 3.8 中的死区时间-否决效率曲线均近似成一条直线，即死区时间与否决效率成线性比。联想到否决效率是被否决的负样本的百分比，死区时间是被否决的时间百分比，那么只要所有负样本相互地不和其他样本靠近（即互相不落入对方的关联时间窗内，由于主频道样本的稀疏性，这个条件是满足的），一个负样本被否决，死区时间就同时增加一等份，自然而然地，死区时间-否决效率曲线趋于线性。其次，当否决效率达到 100%的时候，负样本已经全部被否决，那么死区时间-否决效率曲线的斜率就由曲线与Efficiency = 100%的截距决定：

$$\text{intercept} = \frac{T_{windows}}{\mathrm{T}_{livetime}} \tag{3-14}$$

其中$T_{windows}$代表以所有负样本中心事件为中点的关联时间窗的并集的大小；$\mathrm{T}_{livetime}$代表经过数据质量标记过滤后的总的有效分析时间，也被称为 live time。

如表 3.2 所示，尽管图 3.8 的四个子图中各自四组数据，共计 16 个组数据段，各自的总时间长度均为 400000 秒，但是经过数据质量标记过滤后，出现了明显差异。$\mathrm{T}_{livetime}$最高的可达3.66×10^4秒，最低为2.31×10^4秒。而$T_{windows}$的时间变化更为剧烈，最高可达 3984 秒，最低为 176 秒。正是由于这两者的共同作用，使得图 3.8 没有如同图 3.7 那般在 S6 和 S5 以及 H1 和 L1 的比较中表现一致。

通过以上分析，可以得到的结论是 LIGO 定义的死区时间-否决效率曲线并不适用于同一个否决算法在不同探测器或不同运行版本数据上比较的场景，只适合用于相同数据下不同否决算法间的比较，只有在性能比较使用相同数据的情况下，$\mathrm{T}_{livetime}$相等，进行的比较才有意义，因为此时$\mathrm{T}_{livetime}$是相同的，不会出现$\mathrm{T}_{livetime}$和$T_{windows}$相互纠缠的情况。而传统的 roc 曲线则不存在这种受限问题。

表 3.2 图 3.8 中四组数据的时间信息

子图	时间长度及斜率	S5H1	S6H1	S5L1	S6L1
	$T_{windows}$	356	666	3984	642
左上	$\mathrm{T}_{livetime} \times 10^4$	3.66	2.93	3.09	2.41
	intercept $\times 10^{-4}$	9.7	22.8	128.8	26.6
	$T_{windows}$	226	660	3146	653
右上	$\mathrm{T}_{livetime} \times 10^4$	3.53	3.11	2.63	2.78
	intercept $\times 10^{-4}$	6.4	21.2	119.7	23.4
	$T_{windows}$	182	598	2106	779
左下	$\mathrm{T}_{livetime} \times 10^4$	3.58	3.31	2.57	3.27
	intercept $\times 10^{-4}$	5.1	18.1	81.9	23.8
	$T_{windows}$	176	364	1847	1083
右下	$\mathrm{T}_{livetime} \times 10^4$	3.34	3.15	2.31	3.38
	intercept $\times 10^{-4}$	5.3	11.6	80.0	32.1

3.5 否决算法的在线实现

3.5.1 引力波信号特征提取

目前 LIGO 探测器总共有上万个传感器，即上万个辅助频道。前边的数据分析用到的辅助频道个数多在数百个，实际上大部分辅助频道并没有用于否决分析。一方面需要过滤掉前边介绍过的不安全的辅助频道，更重要的另一方面是因为辅助频道的多少直接影响到否决算法的执行时间，出于计算量上的考虑，只能挑选一些经验上认为最重要的辅助频道进行分析。

随着 Advanced LIGO 建造的开始，在 Advanced LIGO 中将会有更多的传感器频道加入。并且由于硬件的更换，以往经验将不再适用，因此就需要重新从这些辅助频道中挑选出重要性最高的。这里我们利用特征提取[49]中一个相当简单的方法，F-score[50]，来提取较为重要的一些辅助频道。给定训练集

$\boldsymbol{x}_k, k = 1, \ldots, l$，其第 i 个特征的 F-score 值为：

$$F(i) = \frac{(\bar{\mathbf{x}}_i^{(+)} - \bar{\mathbf{x}}_i)^2 + (\bar{\mathbf{x}}_i^{(-)} - \bar{\mathbf{x}}_i)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)} - \bar{\mathbf{x}}_i^{(+)})^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}(x_{k,i}^{(-)} - \bar{\mathbf{x}}_i^{(-)})^2} \quad (3\text{-}15)$$

其中n_+与n_-分别是训练集中正样本与负样本的数量；$\bar{\boldsymbol{x}}_i$，$\bar{\boldsymbol{x}}_i^{(+)}$和$\bar{\boldsymbol{x}}_i^{(-)}$分别为训练集全部，训练集正样本和训练集负样本的第i个特征的平均值；$x_{k,i}^{(+)}$和$x_{k,i}^{(-)}$分别是第 k 个正样本和负样本的第 i 个特征。F(i)的取值越高，就代表越容易根据第 i 个特征区分正负样本。

由于每个辅助频道在特征向量中占据 6 个特征，可以简单地将 6 个特征各自的 F-score 值相加得到该辅助频道的重要性。但是考虑到这 6 个特征物理上属于同一个辅助频道，因此看作一个整体求得每个辅助频道的 F-score 值可能更符合逻辑，于是将式 3-15 的 F-score 从标量形式推广到向量形式，第 i 个辅助频道的 F-score 值为：

$$F(i) = \frac{\left\|\bar{\mathbf{x}}_{i6}^{(+)} - \bar{\mathbf{x}}_{i6}\right\|^2 + \left\|\bar{\mathbf{x}}_{i6}^{(-)} - \bar{\mathbf{x}}_{i6}\right\|^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}\left\|\boldsymbol{x}_{k,i6}^{(+)} - \bar{\mathbf{x}}_{i6}^{(+)}\right\|^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}\left\|\boldsymbol{x}_{k,i6}^{(-)} - \bar{\mathbf{x}}_{i6}^{(-)}\right\|^2} \quad (3\text{-}16)$$

其中n_+与n_-分别是训练集中正样本与负样本的数量；$\bar{\boldsymbol{x}}_{i6}$，$\bar{\boldsymbol{x}}_{i6}^{(+)}$和$\bar{\boldsymbol{x}}_{i6}^{(-)}$分别为训练集全部，训练集正样本和训练集负样本的第i个辅助频道的 6 个特征的平均值所组成的列向量；$x_{k,i6}^{(+)}$和$x_{k,i6}^{(-)}$分别是第 k 个正样本和负样本的第 i 个辅助频道的 6 个特征组成的列向量；$\|\boldsymbol{x}\|$代表向量$\boldsymbol{x}$的二范数。

图 3.9 显示了分别在 H1 和 L1 探测器上，以 GPS 时间 969000000 至 969200000 为训练集，969200000 至 970000000 为测试集的基于随机森林的否决。将特征提取应用在训练集上，如图中右下角的曲线标注所示，分别提取重要性前 100%（即未进行特征提取），50%，25%和 10%的辅助频道，生成正负样本和测试。

可见绝大部分情况下，无论是 ROC 曲线，还是死区时间-否决效率曲线，去除越多的不太重要的辅助频道，性能表现越好。但是性能改善的幅度很小，这是因为这里使用的完整版本的辅助频道数量也才两百余个，已经是从众多的辅助频道中挑选出来的较重要辅助频道，因此性能改善小是符合逻辑的。

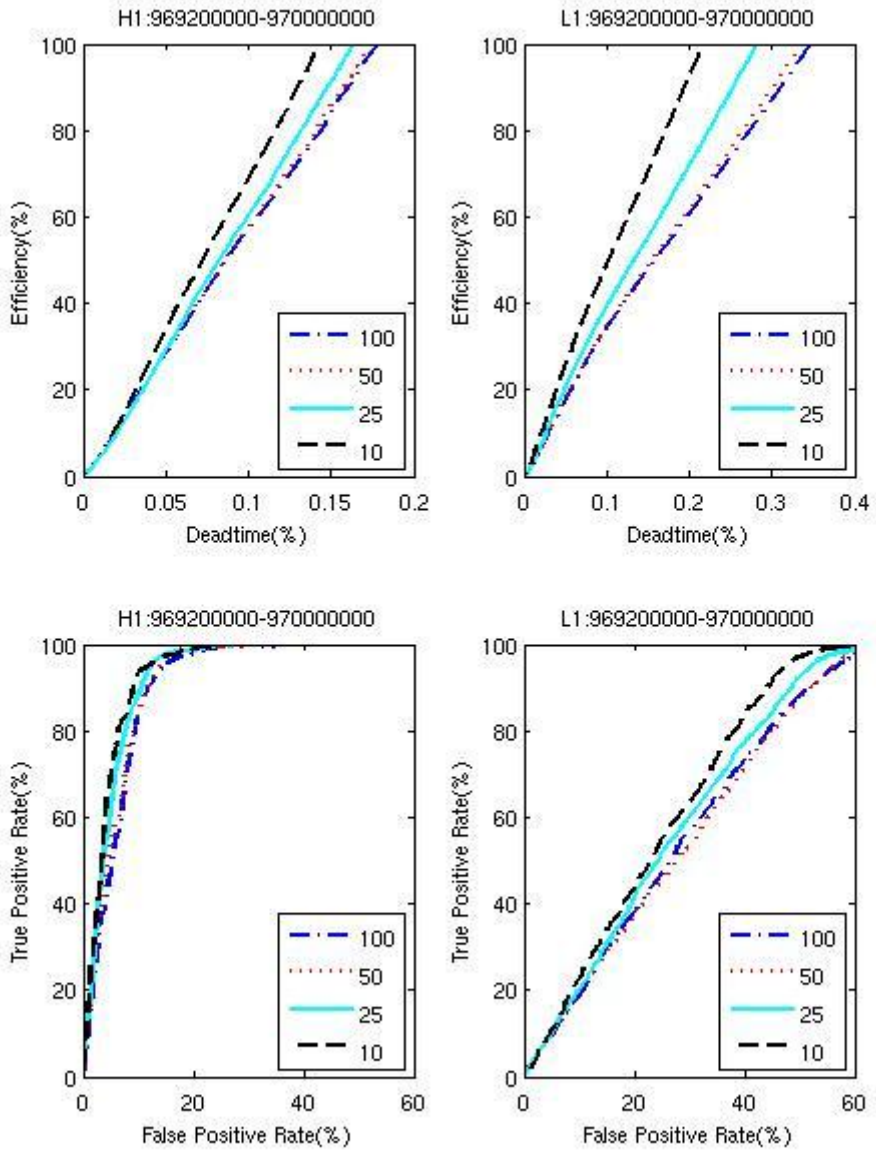

图 3.9 不同程度特征提取下否决性能的比较

此外在不同探测器的多组数据上对辅助频道重要性排序。假设有m组数据，N 个辅助频道，那么辅助频道重要性排序及变化程度可由式 3-17 至式 3-20 表示。

$$\mathrm{RankVar} = \frac{\sum_{i=1}^{n} \sqrt[2]{\sum_{j=1}^{m}(r_{ij} - \bar{r}_i)^2}}{n} \tag{3-17}$$

$$\mathrm{RankMean} = \frac{\sum_{i=1}^{n} \bar{r}_i}{n} \tag{3-18}$$

其中 n 为参与计算的辅助频道数量；r_{ij}代表第 i 个辅助频道在第 j 组数据中重要性排名；$\bar{r}_i$代表第 i 辅助频道在 m 组数据中重要性排名的均值。因此实际上，RankVar代表的是辅助频道排序变化的平均值。

$$\mathrm{RankVarPer} = \frac{\mathrm{RankVar}}{N} \tag{3-19}$$

$$\mathrm{RankMeanPer} = \frac{\mathrm{RankMin}}{N} \tag{3-20}$$

其中 N 为所有辅助频道的数量。RankVar与总的辅助频道数量的比值RankVarPer客观反映了辅助频道排序变化的激烈程度。RankMeanPer则反映了辅助频道在排序中所处位置的前后。

在 H1 和 L1 探测器各 23 组数据上计算，H1 的RankVarPer = 18.4%；L1 的RankVarPer = 18.84%。两者均接近百分之二十，也就是说 H1 和 L1 探测器上辅助频道重要性排序的正负变化幅度均接近 40%，这从一个侧面说明了引力波探测器中噪声分布存在很大的随机性。从另一侧面看，对 23 组数据中各组数据排序前 20%的辅助频道做交集后发现，H1 探测器上只有三个辅助频道一直属于前 20%的辅助频道；而 L1 探测器则完全没有辅助频道能一直停留在前 20%。

最后比较 INST 类辅助频道和 PEM 类辅助频道间的不同，如表 3.3 所示。可发现，INST 类辅助频道的整体重要性要高于 PEM 类辅助频道，但变化剧烈程度同样高于 PEM 类辅助频道。这可能是由于 INST 类辅助频道直接和探测器硬件设备相关，因此 INST 类辅助频道与探测器事件关联较大，导致 INST 类的整体重要性更高些。变化剧烈程度更高的原因可能是由于 INST 类辅助频道和 PEM 类辅助频道存在关联造成。因为环境的变化同样会影响到 INST 类辅助频道，再加上 INST 类辅助频道收到的来自设备的影响，导致其变化更剧烈些。

表 3.3 环境类和设备类辅助频道排序特征比较

频道类型	RankMean	RankMeanPer	RankVar	RankVarPer
PEM	148.8	0.61	42.5	0.17
INST	101.9	0.42	47.4	0.20

3.5.2 否决算法的在线运行

基于监督模式识别的否决方法的训练和测试可以拆分开来，分别独立进行，因此可以设计两个独立的进程：进程 1 和进程 2。进程 1 专门负责生成训练集和监督模式识别方法的训练；进程 2 则负责生成测试集和调用进程 1 训练好的识别器去进行否决。考虑到服务器均为多核的现实，可以让其运行在两个排他的 CPU 物理核心上，防止干扰。可以设计基于监督模式识别的否决算法的单探测器在线运行如下：

1. 进程 1 每隔固定一段时间T_1，即在$T,\ T+T_1, T+2T_1$等时刻对$[\mathrm{T}+T_2-T_1-86400, \mathrm{T}+T_2-T_1]$，$[\mathrm{T}+T_2-86400, \mathrm{T}+T_2]$等区间频道数据进行处理，生成训练集并进行训练；
2. 进程 2 则在$T+T_2,\ T+T_1+T_2, T+2T_1+T_2$等时刻分别从$[\mathrm{T}+T_2-T_1, \mathrm{T}+T_2]$, $[\mathrm{T}+T_2, \mathrm{T}+T_2+T_1]$, $[\mathrm{T}+T_2+T_1, \mathrm{T}+T_2+2T_1]$等区间生成测试集并调用进程 1 在$T, T+T_1, T+2T_1$等时刻开始生成的分类器进行否决；

上述的设计中需要满足几个约束。约束一：在 86400 秒频道数据上生成训练集并训练所需的时间必须小于T_1和T_2；约束二：在时长为T_1的频道数据上生成测试集并测试所需的时间需小于T_1。

回顾表 3.1，随机森林否决在长度均为一天的测试集和训练集数据上执行时间均在 210 秒以下。随机森林否决消耗的时间主要由三块构成：测试集的生成，训练集的生成以及随机森林分类器的训练（分类器的测试基本不耗时），这三者时间消耗上基本持平，也就是说大致可以认为在一天的数据上三者均耗时 60 秒左右。

考虑约束一，由于当前 86400 秒训练集的生成可以使用之前的一个 86400 秒训练集上的大部分正负样本，因为两者相交的时间长度为$86400-T_1$，而T_1相对86400应很小，所以粗略地我们只需要T_1和T_2大于训练随机森林分类器的时间，即 60 秒，保守起见，T_1和T_2需不小于 90 秒。而由于测试几乎不需要时间，且在 86400 秒时间长度上生成测试集时间足够短，所以约束二自然满足。

当T_1和T_2均等于 90 秒时，可以估计出由基于随机森林的否决算法对主频道事件否决造成的延迟上下限。若想降低延迟上限，就必须对基于随机森林否决的代码进行优化和改进，实际上也是有改进空间存在的。

$$\text{Latency}_{max} = \text{T} + T_2 - (T + T_2 - T_1) + \frac{T_1}{86400} \times 60 \approx 90\text{s} \quad (3\text{-}21)$$

$$\text{Latency}_{max} = \frac{T_1}{86400} \times 60 \approx 0 \quad (3\text{-}22)$$

3.6 本章小结

本章主要对引力波探测器表征中最重要的噪声源研究领域中的噪声事件否决进行了详细的描述和分析。与 LIGO 以往使用的探测器主频道噪声事件否决机制不同，本章将噪声事件否决转换成了一个模式识别问题，以随机森林模式识别方法为例子与传统的噪声事件否决算法进行了比较。将不同探测器和不同科学运行数据上的比较结果与探测器间灵敏度差异的事实耦合，从而佐证了基于模式识别否决方法的有效性。同时也发现了 LIGO 自定义的性能指标存在的局限性。

此外，还引入了模式识别常用的特征提取，从而有助于挑选重要性较高的辅助频道；探讨了基于模式识别的否决方法在线运行的可能性，并给出了可能的延迟上下限。

第4章 Burst 类型事件实时监测与否决

本章主要讨论了 Burst 类型引力波数据分析中的事件实时监测与否决，将介绍主频道事件实时监测的意义，为其开发的监测软件以及主频道事件否决机制的结果。

4.1 Burst 类型事件实时监测

4.1.1 Burst 类型事件介绍

以 Burst 类型主频道事件为例，它们由第二章中介绍过的 Omega Pipeline 所生成。与第 3 章中介绍过的 KW trigger 有些类似，每个 Omega 主频道事件的属性有 5 个：中心时间，中心频率，持续时间，带宽和归一化能量。

在 LIGO 的第六次运行中，Omega 对每 64 秒的数据提取一次事件，并将事件的属性写入一文本文件中，该文本文件被称为 trigger file。由于不能排除一个事件恰好处于两个相邻的 64 秒数据的交界处，因此每两个相邻的 64 秒数据有 8 秒的重叠，如图 4.1 所示。

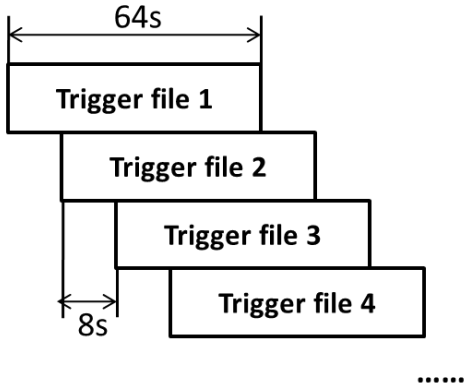

图 4.1 Omega 的结果输出文件在时间上的分布

第 2 章中使用了图 2.5 来介绍探测器诊断与引力波数据实时处理的相互作用，但需要注意的是图 2.5 中显示的是 Omega 主频道事件的 24 小时汇总图，显然这不能满足在线监测和诊断的需要，而 LIGO 中就缺乏一个对 Omega 生成的

事件进行实时监测的工具。此外，LIGO 的天文学家在实际应用中发现，在某些辅助频道上，Omega 生成的事件相比探测器表征专用的 KW 事件更能反映辅助频道的状况，因此开发这样一个实时监测的工具就显得更为必要了。

4.1.2 数据监测工具箱 DMT 简介

数据监测工具箱（Data Monitoring Tool，简称 DMT）[51]是由 LIGO 发起的一个 Unix/Linux 环境下的 C++软件项目，旨在开发支持连续的数据流监测的各种工具，例如 glitchMon[52]。

DMT 主要包含三大部分：为监测各种不同数据流而开发的 DMT Monitors，名称服务器（Name Server）和 DMT viewer。通常一个 monitor 负责监测的数据含有多种属性信息，DMT 将每种属性信息的监测结果定义为一个数据对象（Data Object），并以此作为基本元素定义了一系列的二次开发接口以方便开发各种 monitor。名称服务器的功能有些类似域名服务器，负责注册运行中的 monitor 的各种状态信息。DMT viewer 则是一个图形化的工具，用于显示监测结果，如图 4.2 所示。

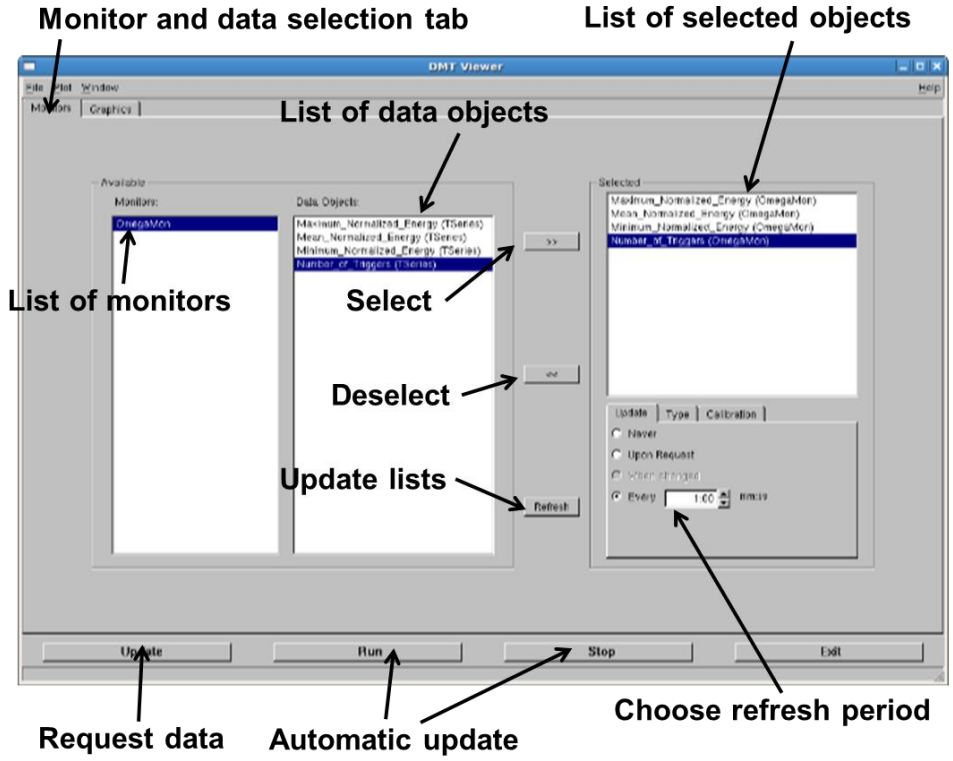

图 4.2 DMT viewer 的监视器选择界面

而图 4.3 显示的则是 DMT 这三大组成部分之间的关系。一般而言，各个 monitor 均运行在后台，因此也被称为 background monitor。首先，monitor 在开始执行监测任务的同时，将自己的名字，网络套接字和数据对象的各种固有属性注册到名称服务器。名称服务器接收到注册请求后开始实时更新该 monitor 的信息。当一个用户使用 DMT viewer 时，DMT viewer 向名称服务器请求目前所有运行中的 monitor 的信息并显示给用户。用户选择要查看的 monitor 及其名下的数据对象后，DMT viewer 将直接向该 monitor 发出请求，该 monitor 将数据对象实时传送至 DMT viewer 供显示。

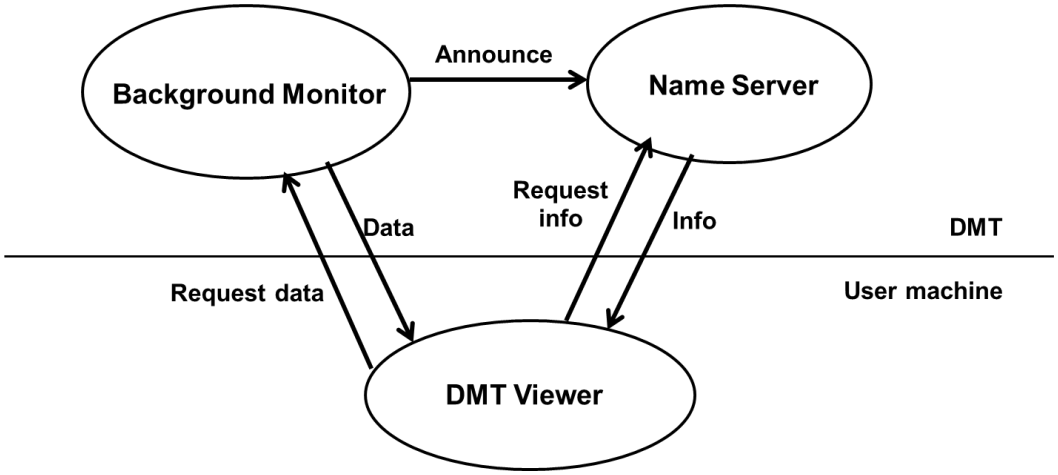

图 4.3 DMT 三大组成部分之间的关系

4.1.3 OmegaMon 的设计

监测 Omega 主频道事件的 monitor 被称为 OmegaMon。它继承了 DMT 中的 MonServer 类，主要有四个类函数[53]。OmegaMon 构造函数读入命令行参数，初始化 OmegeMon 的配置，例如 OmegaMon 和数据对象的名称。OmegaMon 析构函数负责在 OmegaMon 退出时做一些清理工作，例如清空输出缓冲区。OmegaMon 数据处理函数则是 OmegaMon 的核心函数，以预先定义的周期被循环调用，统计某一时间区间内 Omega 事件的属性。OmegaMon 中断函数主要处理 OmegaMon 运行时收到的各种中断信号，例如 Linux 中的 SIGINT 信号。

时间尺度上的监测尺度范围变化很大，从长期趋势到短期波动。与其他大部分的现有 monitor 只支持 24 小时和 1 小时时间尺度监测不同，OmegaMon 支持多个任意分辨率的数据监视，从秒尺度至年尺度。

从图 4.3 可知，DMT 可运行在分布式环境下。名称服务器，monitor 和 DMT

viewer 可分别处于不同的机器上，以方便用户的使用。但这也带来一个问题，即多个用户可能同时使用一个 monitor，而造成重名问题。在 OmegaMon 中，用户可以在 OmegaMon 这个 monitor 名称后边添加自定义的后缀，从而大大减少了重名的可能。

事件的时频散点图是 LIGO 中很重要的一类用于探测器诊断的图像。时频散点图描述了一段时间内，各种事件在时频平面上的分布情况。因此时频散点图中的各个离散点，即事件的中心时间一般均不是整数，而 DMT 在时间维度，即横轴只能支持离散值，因此 DMT 不能支持时频散点图。但是 OmegaMon 通过对时间维度进行坐标转换（例如将时间分辨率从秒转换到毫秒），从而间接支持了时频散点图的绘制。

最后，OmegaMon 保留了扩展接口，因此除了能够支持监测 Omega 事件外，还能监测其他分析软件生成的事件，例如第 3 章中介绍过的 KW 事件。

4.1.4 OmegaMon 实时监测

噪声源通常与各种频道上的事件属性存在关联，因此可以通过 OmegaMon 的监测发现噪声源的活动情况。例如，当地震发生在探测器周边或由于不稳定的供电导致激光能量波动。若过去一小时内平均事件率超过 500 个/分钟，即正常事件率的两倍，那么过去一小时内探测器的工作基本就可认定为不正常。

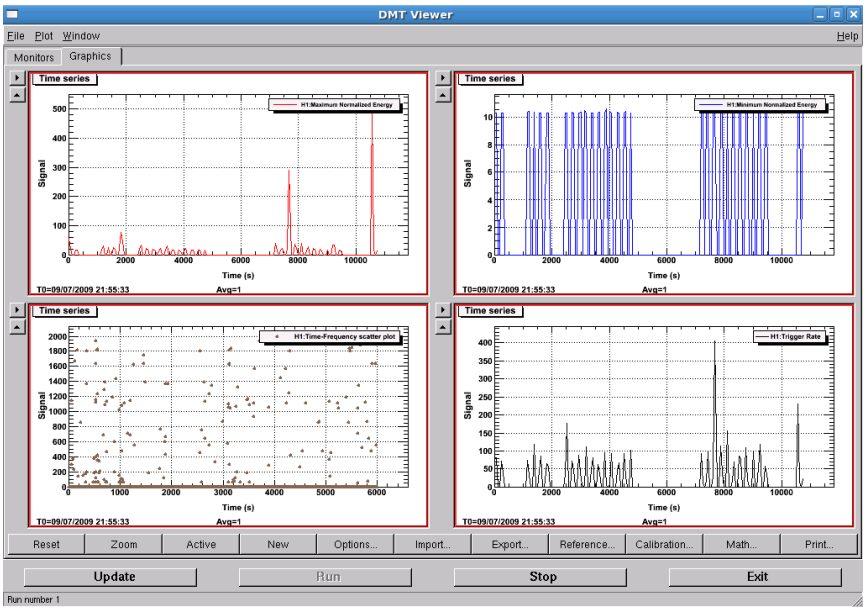

图 4.4 OmegaMon 实时监测

如图 4.4，显示的是 OmegaMon 对 LIGO 第六次科学运行探测器 H1 主频道事件的监测效果。左上角和右上角的子图中分别监测了每分钟内最高和最低的归一化能量。右下角显示的是每分钟的事件发生率。左下角的图则是前边提到过的时频散点图，显示了每分钟的事件在时间和频率上的分布情况。需要注意的是，时频散点图横轴的时间跨度是 6000 秒，而非 60 秒，这就是让 DMT 支持非整数的 GPS 时间而做的坐标转换，实际上横轴上的 1 秒相当于实际时间的 0.01 秒。

4.2 Burst 类型主频道事件否决

4.2.1 已有的主频道事件否决方法

第 2 章中提及的一致事件检验，数据质量和背景噪声估计是 LIGO 中所有类型引力波数据分析均会用到的三种主频道事件否决手段，但是仍然有大量的噪声事件能够通过这些否决，并被误判为一致事件乃至引力波候选事件。另一方面，对引力波候选事件的后续电磁跟踪验证的机会十分宝贵。比如美国的 Swift 伽马射线暴探测卫星仅仅允许 LIGO/Virgo 每个月提交一个电磁跟踪请求[54]，这是因为每探测一个新的天空方位，卫星都得在线调整姿态，从而指向该方位，探测是排他的，成本很高。因此就必须确保提交的引力波候选事件有很高的可信度，这就要求误警率必须处于一个极低的水平，LIGO 希望误警率达到每五十年一个的水平。若想降低误警率，就必须以否决掉占绝对多数的噪声事件为前提。

在 LIGO 的 CBC 类型引力波分析中，探测的引力波源均为双星融合类天文现象。这类天文现象在物理上建模已经较为成熟，因此引力波信号的各种属性是确定的，如图 4.5 所示，即为双星融合类引力波波形示意图。

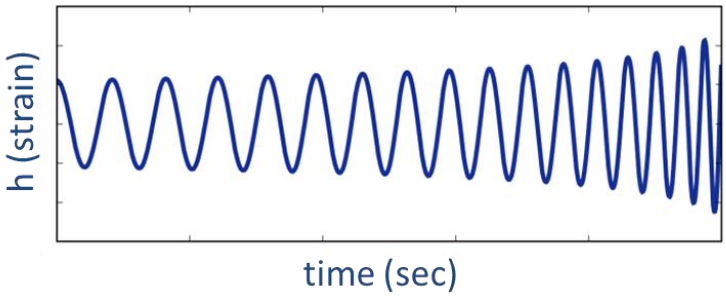

图 4.5 致密双星融合中典型的引力波信号

由于 CBC 类型引力波波形确定，因此可以向引力波数据中加入仿真 CBC 类型波形[55]，从而测试 CBC 类型引力波数据处理能否捕获到相应的 CBC 事件。由于这些 CBC 事件与真实 CBC 类型引力波事件相似，因此可以利用这些事件的信息去否决其余的噪声事件，为此 CBC 中发起了一个名为 MVSC（Multivariate Statistical Classifier）[56]的项目。

与 CBC 类型引力波分析不同的是，Burst 类型引力波是无确切模型的短时脉冲（Unmodeled Bursts），即无法通过仿真的手段创建引力波事件样本。因此到目前为止 Burst 组中并没有开发和 CBC 中 MVSC 类似的否决方法。

4.2.2 Burst 类型主频道事件否决

借鉴第 3 章中介绍过的探测器表征中的否决思想，用随机生成的 GPS 时间和 Omega 主频道事件的 GPS 中心时间从辅助频道事件中提取出引力波事件正样本和噪声负样本。模式识别方法则继续使用随机森林，且随机森林的配置参数没有改变。

采用的性能指标和第 3 章中一样，即传统的 ROC 曲线和死亡时间-否决效率曲线。对主频道和辅助频道事件均进行预筛选，将归一化能量低于 32 的主频道和辅助频道事件去除。

在采用何种辅助频道事件上，有两种选择。可依旧选择 KW 事件，也可以选用 Omega 事件。在此我们做个两者间的比较，如图 4.6 和图 4.7 所示。可见在 4 组不同数据上辅助频道采用 Omega 事件的 ROC 性能均比 KW 事件要好；而在死区时间-否决效率曲线比较中，辅助频道采用 Omega 事件在大部分情况下还是比 KW 事件要好。辅助频道采用 Omega 否决性能较好的原因可能是由于 Omega 事件属性维数比 KW 事件高一维，能够提供更多信息供否决使用。因此我们选择 Omega 的辅助频道事件作为否决主频道事件的信息来源。

接下来讨论如何确定否决阈值。前边画出的 ROC 曲线和死区时间-否决效率曲线均是在变动否决阈值的情况下生成。在实际应用中，必须设置否决阈值。在无任何先验知识的情况下，否决阈值是无法确定的。不过可以先对随机森林给出的 Omega 主频道事件的概率的分布进行统计，如图 4.8 所示。显然，在 6 个不同数据集上，正样本概率低于百分之十的主频道事件数量均占大多数，因此即使很保守地将否决阈值设置为百分之五，一半左右的 Omega 主频道事件将会被否决掉。

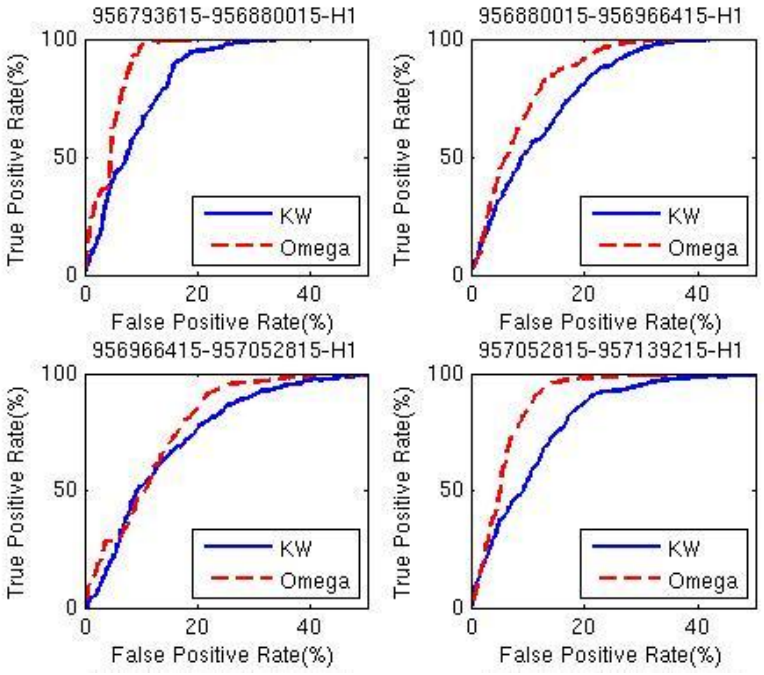

图 4.6 辅助频道为 KW 和 Omega 时的 Omega 事件否决接收者响应曲线性能比较

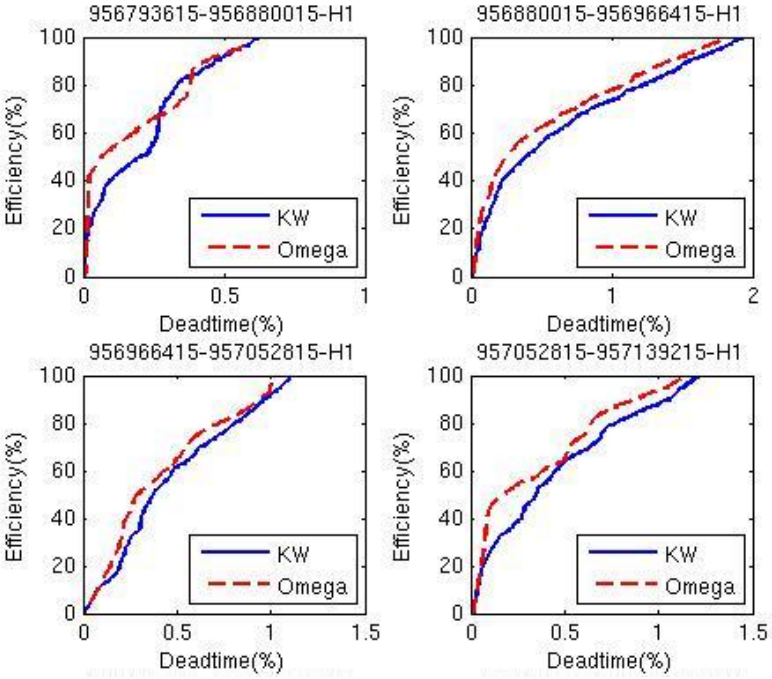

图 4.7 使用不同辅助频道时的 Omega 事件否决的死区时间-否决效率曲线

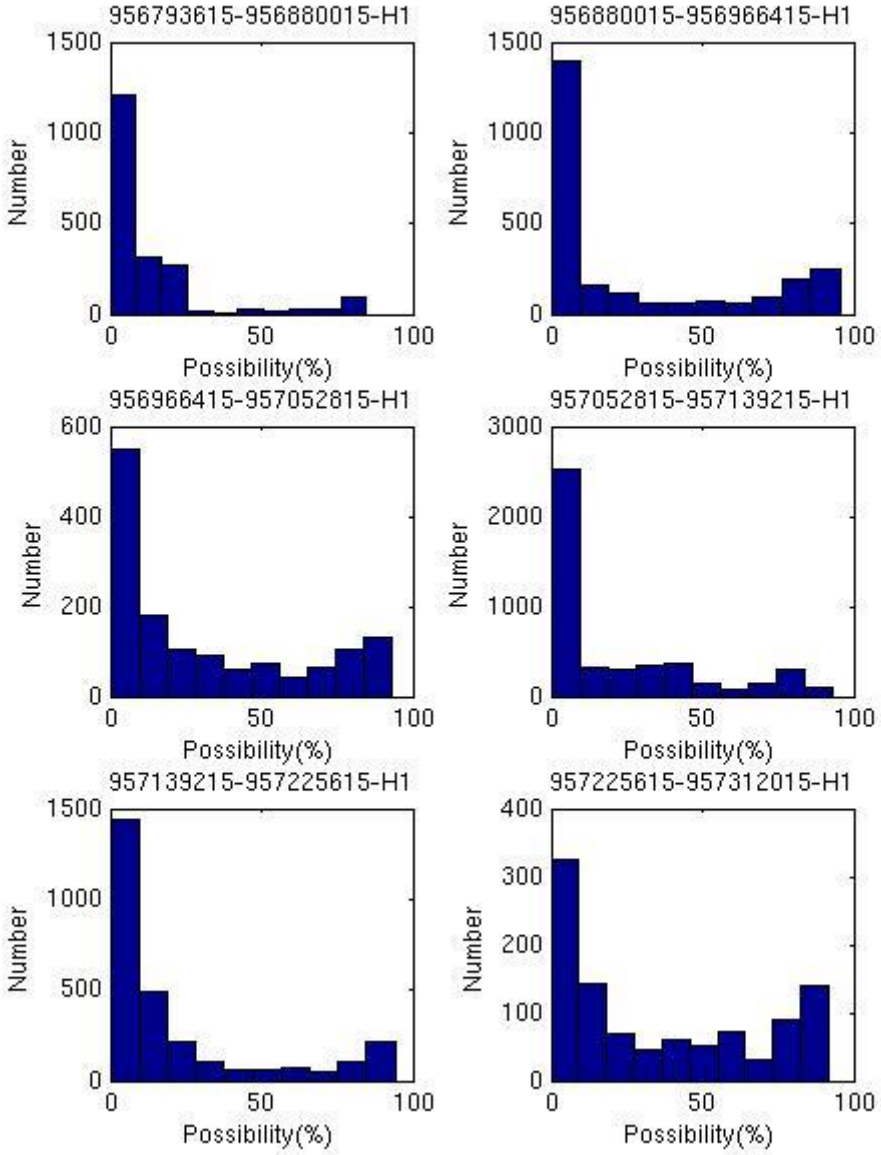

图 4.8 Omega 主频道事件的正样本概率柱状分布图

4.2.3 Burst 类型主频道事件否决对实时性的影响

根据式 3-3，很容易能推导出两个泊松过程发生一致事件的事件率如式 4-1 所示。即一致事件的事件率等于两个泊松过程各自事件率和关联时间窗三者的乘积。因此，若 H1 和 L1 探测器上均有一半的 Omega 主频道事件被否决，那么

H1L1 一致事件的事件率将只有否决前的四分之一。也就是说在中心节点上，只需要运行否决前四分之一的一致事件详细跟踪就行了，而这还是在设置了一个很保守的否决阈值的基础上。因此对 Omega 主频道事件的否决将会极大地减少中心节点上的计算负担，而计算负担的减轻则允许中心节点运行精度更高的一致事件详细跟踪或降低处理中的延迟。

$$r_{coin}=\mathrm{T}_{\mathrm{win}}r_1r_2 \tag{4-1}$$

此外，对 Burst 类型主频道事件的否决也可以在线运行，如同第 3 章的 5.2 节中设计的在线运行。由于 Burst 类型主频道事件率较高，约为 KW 事件率的 2 倍，因此若想取得和 KW 在线否决相似的延迟，可将训练集的长度减半，由之前的一天时间缩短为半天。经过测试比较，半天时间的训练集与一天时间的训练集对于 Burst 类型主频道事件的否决性能影响不大。

4.3 本章小结

本章中主要介绍了如何对 Burst 类型主频道事件进行实时监测和否决。实时监测是为了支持探测器实时诊断，而主频道事件否决的目的则在于消除噪声事件，降低误警率。在降低误警率的同时，还必须注意降低未命中率（miss rate，也被称为第二类错误）。第 5 章中将讨论通过恢复引力波信号能量的方法来降低第二类错误。

第5章 Burst 类型引力波信号能量恢复

本章以 Burst 分析中使用的 Omega Pipeline 为例，主要介绍了如何更多地从引力波数据中恢复出引力波信号的能量，从而减少引力波信号误判为噪声的可能。

5.1 Omega Pipeline 中的信号能量恢复

第 2 章中曾简略介绍了 Omega 的基本原理。Omega 的 Q 变换所使用的 Q 基底是一组非正交的高斯包络的正弦基底（Gaussian Windowed Sinusoids）[57]。这样的基底可由中心时间τ，中心频率ϕ和品质因子 Q 刻画[33]，如式 5-1 所示的归一化形式。

$$\psi(t;\tau,\phi,\mathrm{Q}) = \left(\frac{8\pi\phi^2}{Q^2}\right)^{1/4} e^{-4\pi^2\phi^2(t-\tau)^2/Q^2} e^{-i2\pi\phi(t-\tau)} \tag{5-1}$$

信号探测理论中定义了一个名为模糊函数（Ambiguity Function）[59]的概念来描述使用了不匹配的基底后能剩余的能量的百分比，即相当于两个不同基底的内积，如式 5-2 所示。

$$\alpha(\delta\tau,\delta\phi,\delta\mathrm{Q}) = \int_{-\infty}^{\infty} \psi(t;\tau,\phi,\mathrm{Q})\psi^*(t;\tau+\delta\tau,\phi+\delta\phi,\mathrm{Q}+\delta\mathrm{Q})dt \tag{5-2}$$

根据式 5-2，则能计算出丢失掉的能量的百分比。

$$\mu(\delta\tau,\delta\phi,\delta\mathrm{Q}) = 1 - |\alpha(\delta\tau,\delta\phi,\delta\mathrm{Q})|^2 \tag{5-3}$$

对式 5-3 进行二阶展开，可得式 5-4[57]。

$$\mu(\delta\tau,\delta\phi,\delta\mathrm{Q}) \approx \frac{4\pi^2\phi^2}{\mathrm{Q}^2}\delta\tau^2 + \frac{2+Q^2}{4\phi^2}\delta\phi^2 + \frac{1}{2Q^2}\delta Q^2 - \frac{1}{\phi Q}\delta\phi\delta Q \tag{5-4}$$

实际上，式 5-4 可作为两个基底之间的距离度量，也就是说可作为一种相邻

远近的度量。第 2 章中曾经介绍过在事件生成中须对相邻的时频块进行聚类，对聚出的类的能量进行阈值挑选后即成为事件。在 Omega 中则是利用式 5-4 来判断时频块间的相邻与否，而聚类的方法则是采用基于密度的聚类（Density based clustering），如图 5.1 所示。首先选择一个基底，找到离它距离小于 D 的其他基底，即邻居。若该基底的邻居数量超过阈值 C，则以每个邻居为中心去寻找邻居的邻居，并依次类推，直至再也无法找到邻居数量超过阈值 C 的点。这些点就被聚为一类。

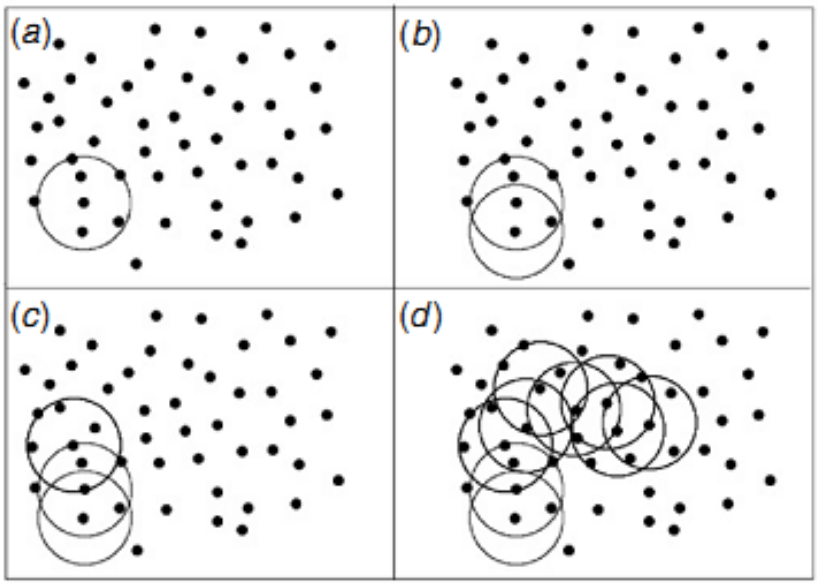

图 5.1 基于密度的聚类方法流程示意图[60]

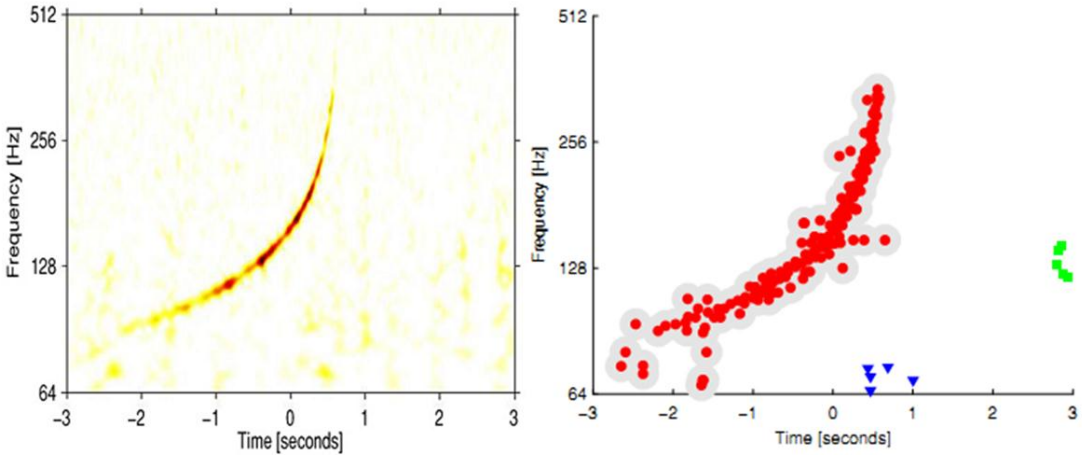

图 5.2 Omega 的聚类效果图[58]

Omega 中聚类的效果如图 5.2 所示，左边的子图显示的是两个 1.4 太阳质量的中子星双星系统融合的引力波仿真信号在时频平面上的表示。右边的子图则显示的是对含有该引力波仿真信号的数据进行 Q 变换且聚类后的效果。右边的子图中含有三个类，分别由红色圆点，蓝色倒三角和绿色方块表示。可见红色圆点代表的类基本与左边子图中的引力波仿真信号吻合，因此基于密度的聚类是有效的。

但是 Omega 中的基于密度的聚类也有其缺陷，它有两个主要参数，即最小邻居数量 C 和邻居最大半径 D，显然，聚类的效果和这两个参数的设置密切相关。可是 Burst 类型引力波信号的物理模型是未知的，因此预先设定的参数不可能对所有的 Burst 类型引力波有效，也就是说可能聚类会遗漏部分信号能量，特别是对于那些在时频平面上分布较广的引力波信号而言，遗漏会更严重。

另一个不足则是，目前在 Omega 中聚类方法的应用还局限于在单探测器数据上。直观上看，若能将聚类方法应用到多探测器数据上，譬如对一致事件的能量恢复，聚类方法的效果应能增加不少，因为多探测器相当于有多组基底，因此形象地讲，即基底密度变大，聚类中邻居的数量增多，聚类的效果应该会有所提高，能恢复出更多能量。

5.2 增强的 Omega Pipeline 能量恢复设计

根据目前 Omega 中聚类鲁棒性不强的特点，本章引入聚类融合（Clustering Fusion）[61]的方法，试图增强聚类鲁棒性。

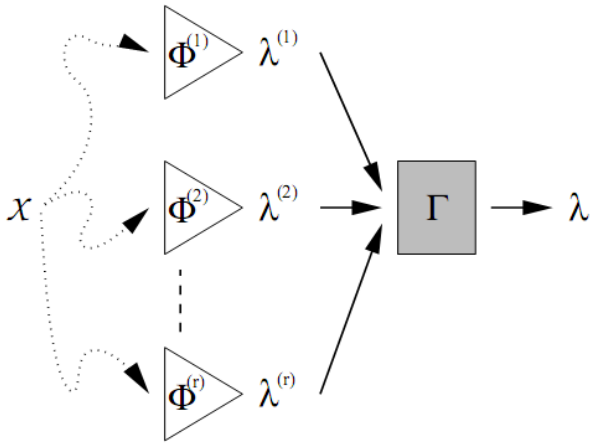

图 5.3 聚类融合原理示意图[61]

令样本集为$\chi = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$，将该样本集分割为 k 个类，即$\{C_i | \mathrm{i} = 1, \ldots, \mathrm{k}\}$，则样本集各个元素的类归属可用一个标签向量$\boldsymbol{\lambda} \in \boldsymbol{N^n}$表示。聚类融合的基本原理如图 5.3 所示。图中的每一个$\Phi^{(j)}, j = 1, \ldots, r$代表一种聚类函数。它们对样本集进行聚类分析后，会各自产生一个标签向量$\boldsymbol{\lambda}^{(j)}, j = 1, \ldots, r$。一致性函数（Consensus Function）Γ对这些标签向量进行一致分析，最终得到单一的标签向量$\boldsymbol{\lambda}$，并将其作为最终的聚类结果。

显然在聚类融合中，关键问题是如何设计一致性函数Γ。为了求得最优的一致性函数，先引入归一化互信息（Normalized Mutual Information，简称 NMI）估计函数ϕ[61]来度量两个不同聚类的结果，即两个标签向量之间的相似程度，如式 5-5 所示。

$$\phi\left(\lambda^{(a)}, \lambda^{(b)}\right) = \frac{\sum_{h=1}^{k^{(a)}} \sum_{l=1}^{k^{(b)}} n_{h,l} \log\left(\frac{n \cdot n_{h,l}}{n_h^{(a)} n_l^{(b)}}\right)}{\sqrt{\left(\sum_{h=1}^{k^{(a)}} n_h^{(a)} \log\frac{n_h^{(a)}}{n}\right)\left(\sum_{h=1}^{k^{(a)}} n_h^{(a)} \log\frac{n_h^{(a)}}{n}\right)}} \tag{5-5}$$

其中$k^{(a)}$和$k^{(b)}$分别为标签向量$\boldsymbol{\lambda}^{(a)}$和$\boldsymbol{\lambda}^{(b)}$所包含的类的数量；$n_h^{(a)}$为标签向量$\boldsymbol{\lambda}^{(a)}$对应的一个类$C_h$中的样本数量；$n_l^{(b)}$为标签向量$\boldsymbol{\lambda}^{(b)}$对应的一个类$C_l$中的样本数量；$n_{h,l}$则代表即在由$\boldsymbol{\lambda}^{(a)}$定义的第h个类，也在由$\boldsymbol{\lambda}^{(b)}$定义的第l个类中的样本数量。

将基于两个标签向量的归一化互信息估计函数进行推广，可得到一致性函数Γ生成的单一标签向量$\overline{\boldsymbol{\lambda}}$与$\boldsymbol{\lambda}^{(j)}, j = 1, \ldots, r$的相似程度度量，如式 5-6 所示。

$$\phi\left(\boldsymbol{\Lambda}, \overline{\boldsymbol{\lambda}}\right) = \frac{1}{r}\sum_{j=1}^{r} \phi\left(\overline{\boldsymbol{\lambda}}, \boldsymbol{\lambda}^{(j)}\right) \tag{5-5}$$

其中$\boldsymbol{\Lambda}$为标签向量的集合，即$\{\boldsymbol{\lambda}^{(j)} | j \in \{1, \ldots, r\}\}$。

优化目标则是找到使得$\phi\left(\boldsymbol{\Lambda}, \overline{\boldsymbol{\lambda}}\right)$最大的$\overline{\boldsymbol{\lambda}}$，其所对应的一致性函数即为最优。但这是一个组合最优化问题，因此求解很耗时。出于对引力波数据处理实时性的考虑，本章直接采用一种相当简单的一致性函数，即基于簇的相似度划分算法（Cluster-based Similarity Partitioning Algorithm，简称 CSPA）中的一个变种，即 Co-association 矩阵方法[62]。

Co-association 矩阵方法的基本思想是在样本集上运行 N 次聚类方法，聚类

方法的原理相同，只是每次运行参数不同而已。于是可以得到一个 Co-association 矩阵 A，矩阵 A 的元素A_{ij}表示在 N 次聚类中，第 i 个和第 j 个样本被聚到一类的次数。若A_{ij}与 N 的比值超过 0.5，即认为第 i 个和第 j 个样本应属于一类。

在 Omega 中，最小邻居数量 C 被设置为 3，邻居最大半径 D 被设置为 4。于是依照 Co-association 矩阵方法，令最小邻居数量 C 的取值集合为{2,3,4}，而邻居最大半径的取值集合则设置为{2:0.2:9.8}，因此运行了 120 次聚类，于是A_{ij}的阈值为 60。

5.3 性能比较

为了比较改进前后的聚类性能，需要某种 Burst 类型引力波信号作为比较基础。尽管 Burst 类型引力波信号没有确定的物理模型，但是天文学家们通过建模和数值仿真等手段推测出了一些 Burst 类型引力波信号的波形信息。本章中采用一个主流的超新星爆发模型仿真生成的引力波信号[63][64]，如图 5.4 所示。

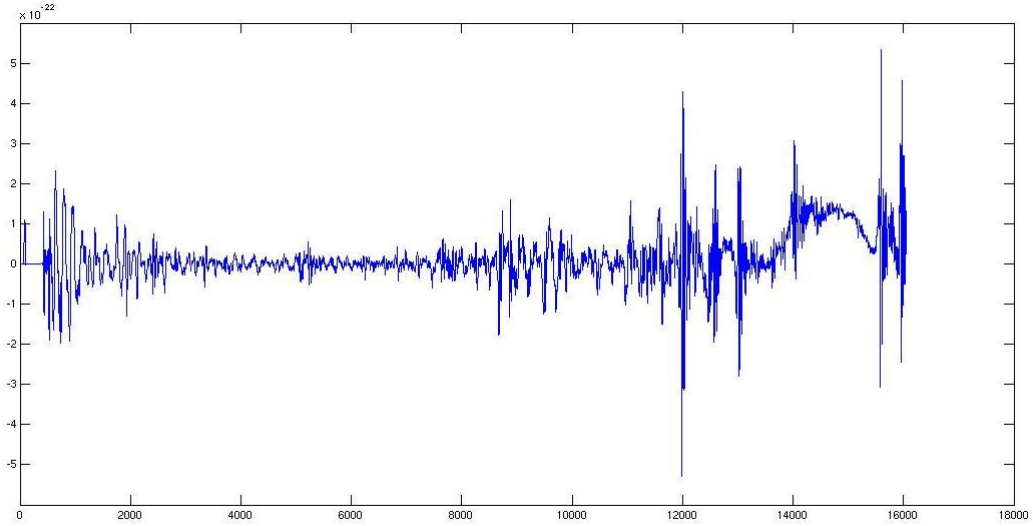

图 5.4 时域下超新星爆发引力波仿真波形

该仿真的引力波信号所代表的核坍缩超新星（也被称为 II 型超新星）拥有 15 个太阳质量，电子中微子光度（Electron Neutrino Luminosity，简称υ_e luminosity）为 32B（相当于3.2×10^{52}尔格/秒，1 尔格等于10^{-7}焦耳），离地球的距离为 10000 秒差距（1 秒差距约等于 3.26 光年）。该信号的持续时间为 0.978 秒，由于采样

频率为 16384Hz，因此图 5.4 所示的波形包含了 16049 个采样点。

用仿真引力波信号生成软件 GravEn（Gravitational-wave Engine）[65]将该信号人工地注射到 LIGO 第五次科学运行的 JW1 数据集（Joint Week 1 Data，即 LIGO 的 H1，L1 和 Virgo 的 V1 在第五次科学运行中联合采集的第一周数据）中。注入仿真信号到各个探测器时，GravEn 已经考虑到了仿真信号在天空的方位对仿真信号在探测器数据中幅值的影响。

由于第五次科学运行采集的引力波数据噪声相当强烈，因此为了测试聚类融合的效果，需要将信号幅值放大。这里选取的放大倍数组合为{1000, 100, 30}。幅值的放大倍数与超新星的距离成反比，因此若幅值放大 10 倍，超新星离地球的距离相当于是原来的十分之一，即 1000 秒差距。将幅值放大后的信号注入到 H1 探测器数据的 GPS 时刻 870052980.454770445 和 L1 探测器数据的 GPS 时刻 870052980.458621263。

5.3.1 单探测器上的聚类融合

图 5.5，图 5.6 和图 5.7 中，单探测器 H1 上的引力波信号分别被放大 1000,100 和 30 倍，从 Omega 使用现有聚类方法和聚类融合方法的表现来看，可知在三种放大情况下，聚类融合总能比现有聚类的表现要好，即能将更多的与信号匹配的基底聚为一类。且随着信号能量的降低，相对来说，聚类融合的性能越好。例如图 5.7 中，现有聚类将信号对应的基底聚成了若干个小类（不同的颜色代表不同的类，颜色越趋近于红色，类的能量越高；越趋近于蓝色，则类的能量越低），而聚类融合则相对完整地将信号能量恢复了出来，即红色的类。

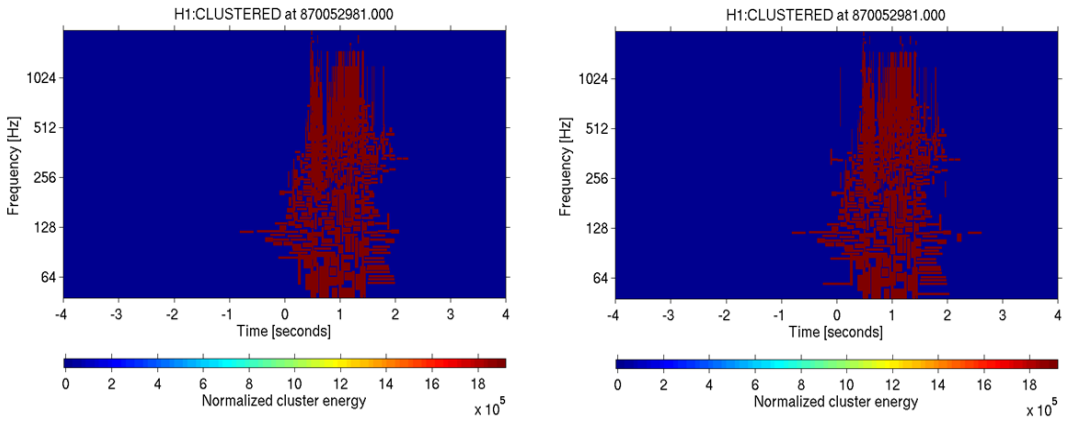

图 5.5 H1 数据上现有聚类与聚类融合的比较（放大倍数为 1000 倍）

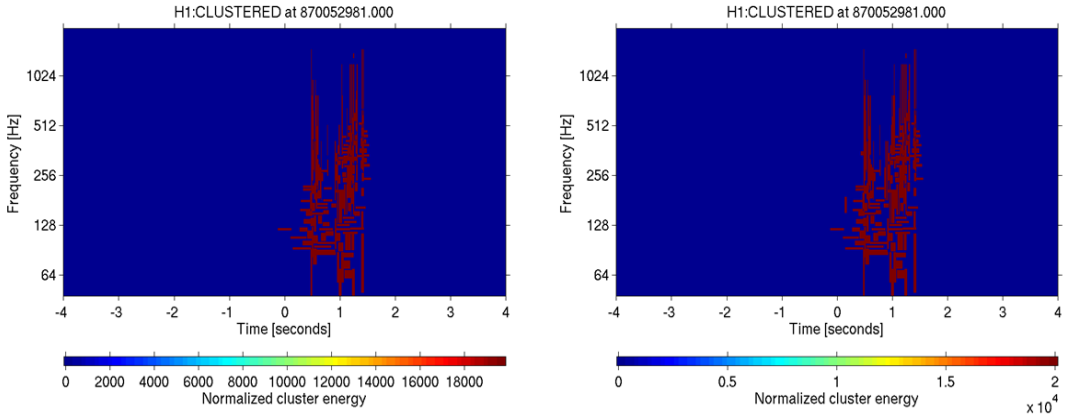

图 5.6 H1 数据上现有聚类与聚类融合的比较（放大倍数为 100 倍）

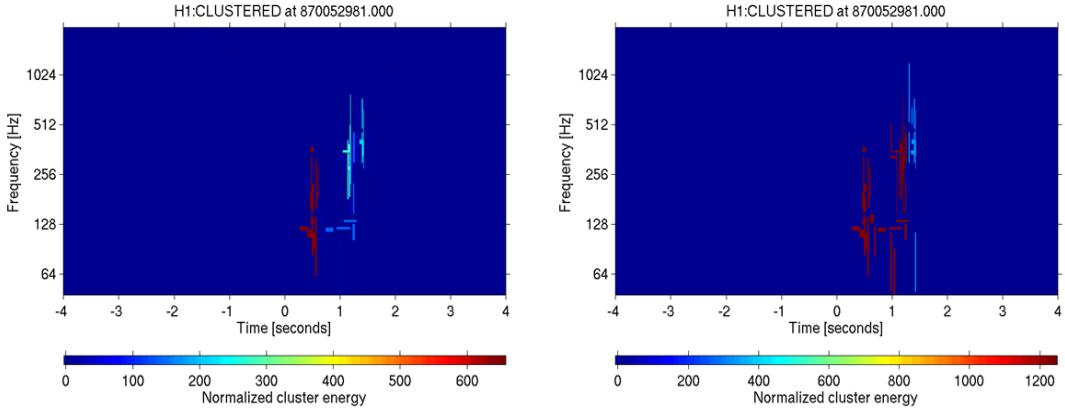

图 5.7 H1 数据上现有聚类与聚类融合的比较（放大倍数为 30 倍）

表 5.1 中，事件类的能量指的是通过聚类方法得到的所有类中能量最高的一个类的能量，即聚类恢复出的引力波信号的能量；事件类的大小则指的是事件类中含有的基底的数量；类的数量则指的是聚出的类的总数。从表 5.1 可以看出，几乎在所有的比较中，由聚类融合总能恢复出更多的信号能量，且聚出的类的总数相对较少，也就是说聚类融合将一些小类合并到了事件类当中。观察这些被合并的小类，可以发现它们的能量大多仅次于事件类的能量。且随着信号放大倍数降低，即信号能量降低，聚类融合多恢复出来的能量与现有聚类恢复出的能量的百分比比值越高，即聚类融合对于能量不高的信号的能量恢复能力比现有聚类方法要好，这对于引力波这种微弱的信号来说是个好消息。但是，必须承认的是，若信号能量继续下降，那么聚类融合的表现就与现有融合没有差别了。不过，这里使用的是 LIGO 第五次运行的数据，若使用 LIGO 第六

次运行所采集的数据，聚类融合的表现理应有所改善，应能支持更低的信号能量。

表 5.1 现有聚类和聚类融合在不同信号能量和探测器上的比较

放大倍数	聚类方法	事件类的能量	事件类的大小	类的数量
1000	H1 现有聚类	1.9106e6	795	61
	H1 聚类融合	1.9114e6	828	34
	L1 现有聚类	2.2338e5	442	73
	L1 聚类融合	2.2415e5	469	51
100	H1 现有聚类	1.9758e4	212	49
	H1 聚类融合	2.0026e4	225	40
	L1 现有聚类	1.0147e3	26	67
	L1 聚类融合	1.9358e3	55	56
30	H1 现有聚类	6.5421e2	14	66
	H1 聚类融合	1.2441e3	38	51
	L1 现有聚类	4.1166e1	1	62
	L1 聚类融合	4.1166e1	1	62

5.3.2 多探测器上的聚类

直到目前，Omega 中的现有聚类方法只应用在了单探测器数据上。直观地看，若能将聚类方法应用在多探测器数据上，那么聚类的效果将有明显提高。而进行多探测器聚类的话，就必须知道引力波信号在天空中的方位，从而消除引力波信号达到各个探测器的时延，使得各个探测器上的基底能够对齐且集中在一起聚类。但是引力波信号的真实天空方位是未知的，不过在第 2 章中曾经提及的一致事件详细跟踪中能够计算出一致事件的天空方位，因此聚类融合可以使用该天空方位，从而恢复出多探测器上一致事件的能量。这将有助于其超过对一致事件设置的能量阈值，从而可能减少未命中率。

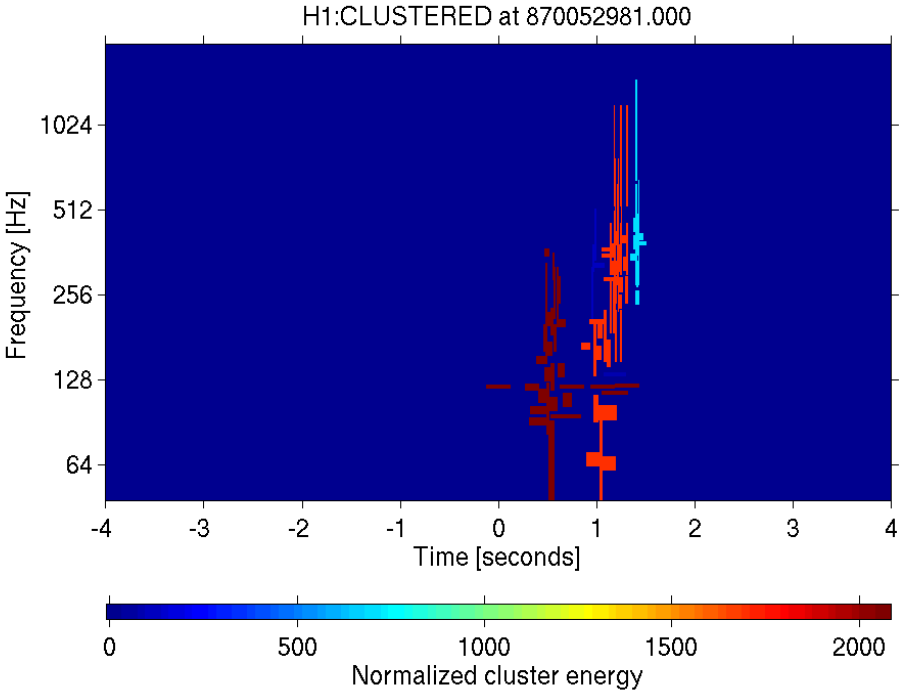

图 5.8 H1 上现有聚类的效果（放大倍数为 50 倍）

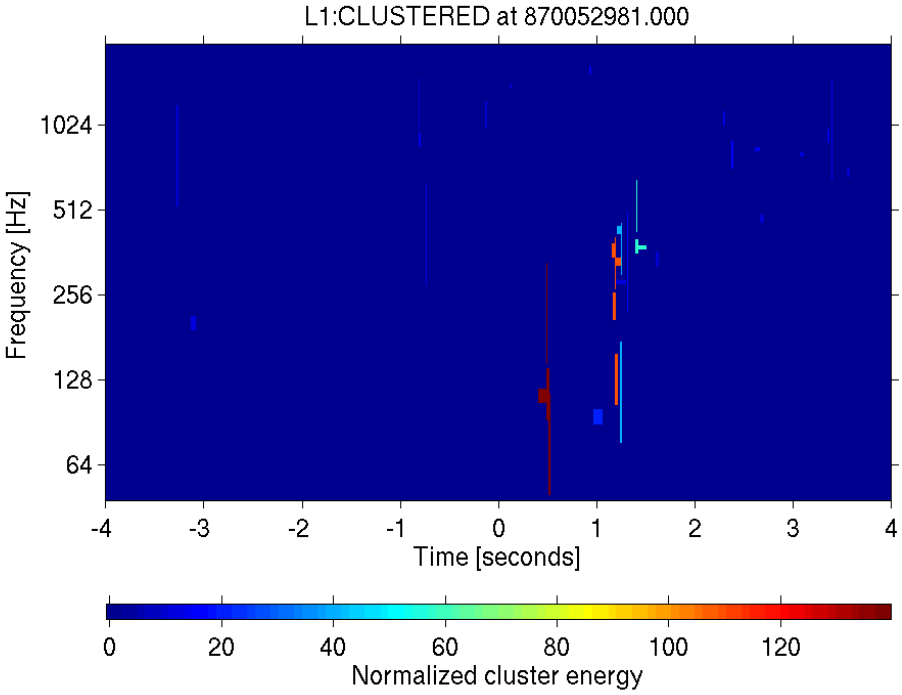

图 5.9 L1 上现有聚类的效果，以 H1 为基准进行时移（放大倍数为 50 倍）

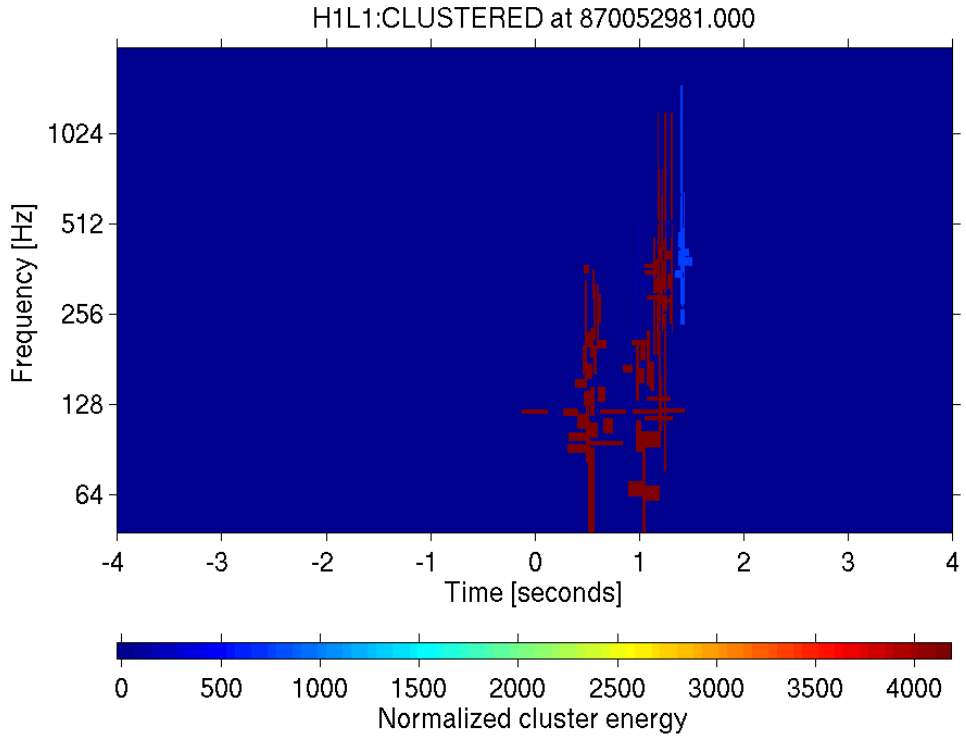

图 5.10 H1L1 上的多探测器聚类（放大倍数为 50 倍）

图 5.8，图 5.9 和图 5.10 中的聚类均采用的是 Omega 中已有的聚类方法，而不是聚类融合算法。尽管如此，可见图 5.10 中的多探测器聚类方法比图 5.8 和图 5.9 的单探测器聚类性能要好的多。图 5.8 和图 5.9 中基底分裂成了若干个不同颜色的小类，而图 5.10 则将大部分小类合并到了一起，大大提高了事件类的能量。

表 5.2 中 H1+L1 聚类表示的仅仅是将 H1 聚类和 L1 聚类的结果相加，以方便和 H1L1 聚类(即多探测器聚类)相互比较。在放大倍数为 50 和 40 的情况下，H1L1 多探测器聚类的效果比 H1、L1 单探测器聚类的效果提高很多，与 H1+L1 聚类相比也提高地相当多，这就意味着这里的 H1L1 多探测器聚类起到了一加一大于二的效果。而随着信号放大倍数的降低，这种效果消失了。这是可以理解的，毕竟随着信噪比下降，能够超过时频块阈值的基底变少，即可供聚类的基底减少，那么基底的邻居也在减少，即使使用多探测器聚类，它的效果下降也是理所应当的。因此，多探测器聚类和前边的聚类融合类似，随着信噪比降低，性能退化严重。

表 5.2 H1L1 多探测器聚类与单探测器聚类的比较

放大倍数	聚类方法	事件类的能量	事件类的大小	类的数量
50	H1 聚类	2.0733e3	43	55
	L1 聚类	1.3896e2	5	66
	H1+L1 聚类	2.2123e3	48	121
	H1L1 聚类	4.1654e3	90	112
40	H1 聚类	1.2334e3	21	62
	L1 聚类	9.1079e1	4	66
	H1+L1 聚类	1.3245e3	25	128
	H1L1 聚类	1.4047e3	29	117
30	H1 聚类	6.5421e2	14	66
	L1 聚类	3.0052e1	1	64
	H1+L1 聚类	6.8426e2	15	130
	H1L1 聚类	6.8426e2	15	122

5.4 本章小结

本章通过引入聚类融合的方法提高了单探测器上 Burst 类型引力波信号能量恢复效果。另外提出了多探测器聚类的思路，从而能够大大提高一致事件能量恢复的程度。

但是这两种方法还需要进一步研究和改进，毕竟它们在低信噪比的情况下，性能表现退化严重。但可以理解的是，本章测试用的数据均为 LIGO 第五次科学运行采集的数据，噪声水平很高。若采用第六次科学运行的数据，相信在低信噪比情况下会有所改善。更重要的是，未来的 Advanced LIGO 时代，噪声水平将下降明显，因此这两种方法还是很有希望的。

此外，这两种方法相比 Omega 中现有的聚类而言，肯定耗时将增多一些，但是考虑到聚类消耗的时间仅仅是 Omega 中时间消耗的极小的一部分，且随着

2015 年到来，计算处理能力将继续指数级提高，对于这两种方法而言，并不会对 Burst 类型引力波数据处理的实时性造成多少影响。

最后，在本章中的性能测试仅仅使用了核坍缩超新星引力波仿真信号，对于大量的无确定物理模型的 Burst 类型引力波而言，一种测试信号是不够的，还需要对其他的 Burst 类型引力波仿真信号进行测试。

第6章 结论与展望

6.1 结论

本文所做的工作及结果主要有以下几点：

1. 详细介绍并改进了引力波数据实时处理系统

本文详细介绍了在 LIGO 第六次科学运行中使用的 Burst 类型引力波数据实时处理系统，并指出了在下一代引力波探测器网络中，引力波数据实时处理可能遇到的问题，并针对这些问题，提出了改进的方案。根据该方案，事件生成将分散到各个天文台探测器上；各个探测器的辅助频道数据可在本地被完全利用，而不用传输到数据处理中心节点。数据处理中心节点的计算负担也将由于主频道事件的否决而大大降低。更为重要的是，LIGO 中四大类型引力波的数据实时处理相当类似，因此对 Burst 类型引力波数据实时处理系统的改进对其他类型引力波数据的实时处理有借鉴意义。

2. 将模式识别方法引入到引力波探测器表征中的噪声信号否决

与以往关于探测器表征中的噪声信号否决不同，本文将其转化为一个模式识别问题，除了在否决性能上大幅提升外，在否决所消耗时间上也明显较已有否决算法少。一方面，在与已有算法的具体比较中，否决的事件集合几乎完全包含已有算法所否决的事件，即证明了基于模式识别的否决是已有否决算法的有效补充；另一方面，在不同探测器的引力波数据上的否决比较结果与引力波探测器的硬件特征相符合，印证了基于模式识别的否决的有效性。由于基于模式识别的否决耗时较少且仍然有改进空间，因此完全可以适用于引力波数据的在线分析，从而为探测器实时诊断提供新的信息来源。此外，还针对 Burst 类型引力波主频道事件的监测开发了一个数据监测器，同样可用于探测器实时诊断。

3. 利用辅助频道事件否决短时脉冲引力波事件

Burst 类型引力波信号与其他三类引力波信号相比最大的不同即是其没有确定的物理模型，因此很难像其他类型引力波信号分析那样，利用仿真的引力波信号的信息来否决噪声事件。但是通过利用辅助频道事件的信息，仍然可以对 Burst 类型主频道事件进行有效的否决。且通过在单探测器上的实时否决，可有

效减少由噪声产生的虚假事件，从而减轻数据处理中心节点上的计算负担。计算负担的减轻意味着数据处理中心节点能够腾出更多的计算能力，对一致事件处理的精度因为计算能力的空闲而得以增加，例如一致事件天空方位的精度就能增加。且计算能力的解放能降低引力波数据实时处理中的时延。这种否决方法也可推广到其他类型引力波数据处理中。

4. 改进了短时脉冲引力波信号搜索中的能量恢复

同样是由于短时脉冲引力波信号的物理模型不明，因此在对其进行能量恢复时，传统的能量恢复手段不够鲁棒，且未能同时利用多探测器的信息，从而可能导致引力波信号被误判为噪声而被否决。为此，本文提出了用聚类融合以及多探测器同时聚类的方法，从而解决了以上问题。

6.2 未来工作

1. 优化基于模式识别的事件否决

在本文中，只测试了支持向量机与随机森林这两种较为常用的模式识别方法的性能，其他的模式识别方法，诸如神经网络，压缩传感并未尝试。另外，无论是探测器表征中的噪声信号否决还是 Burst 类型引力波主频道事件的否决，现有的否决机制中含有大量的 matlab 代码，若将其改写为 C/C++程序，势必能大大提高否决的速度。速度的提高能降低事件否决在实时处理中的时延，从而为探测器实时诊断以及引力波数据实时处理提供更好的支持。

2. 短时脉冲引力波信号搜索中的能量恢复

在本文中，目前的测试结果表明聚类融合方法和多探测器聚类方法均有不错的信号能量恢复能力，但是还可以对多探测器的聚类融合方法进行尝试。此外，低信噪比的情况下，如何提高聚类融合与多探测器聚类方法的表现是一个很有挑战性的问题，毕竟在低信噪比情况下表现不佳是正常的。而另一个值得关注的问题是聚类融合方法与多探测器聚类可能会将部分噪声吸收进来，这就需要对背景估计进行测试，看背景噪声的水平是否由于聚类融合而增长，若增长了，那么就需要和信号能量的增长相比较，看哪个增长地更多。

6.3 展望

引力波探测是一个多学科交叉的科学前沿，而引力波数据分析是其中重要

的一部分。国内在引力波数据分析方面几乎是一片空白，主要原因在于中国没有自己的引力波探测器，也就没有自己的引力波数据。在我国面向 2050 年科技发展路线图中，空间科学发展的战略目标中明确提及了引力波的直接探测。希望本文能够为有志从事这方面研究的同学提供些许帮助，也希望中国自己的引力波探测器早日建成。

参考文献

[1] Dewitt B, Quantum theory of gravity. I. The canonical theory. Physics Review, 1967, 160:1113~1148

[2] Amelino-Camelia G, Gravity-wave interferometers as quantum-gravity detectors. Nature, 1999, 398:216~218

[3] Schutz B, Gravitational Wave Astronomy. Classical and Quantum Gravity, 1999, 16:A131~A156

[4] Taylor J and Weisberg J, Futher experimental tests of relativistic gravity using the binary pulsar PSR 1913 + 16. The Astrophysics Journal, 1989, 345: 434~450

[5] Abramovici A, el al., LIGO: the laser interferometer gravitational-wave observatory. Science, 1992, 256:325~333

[6] The German-British Gravitational Wave Detector. http://www.geo600.org

[7] TAMA300. http://tamago.mtk.nao.ac.jp

[8] VIRGO. http://www.virgo.infn.it

[9] Hough J and Rowan S, Laser interferometry for the detection of gravitational waves. Journal of Optics A: Pure and Applied Optics, 2005, 7:S257

[10] Smith J, Roadmap to the enhanced and advanced LIGO detectors. LIGO technical document, LIGO-G080418-00-D, 2008

[11] Anderson S, Advanced LIGO data and computing. LIGO technical document, LIGO-G0900008-v1, 2009

[12] LISA – Laser Interferometer Space Antenna. http://lisa.nasa.gov

[13] Advanced LIGO. http://www.ligo.caltech.edu/advLIGO

[14] Advanced Virgo. https://wwwcascina.virgo.infn.it/advirgo/

[15] LCGT. http://tamago.mtk.nao.ac.jp/spacetime/lcgt_e.html

[16] LIGO Australia. http://www.aigo.org.au/

[17] Zweizig J, Displacement sensitivity of the LIGO interferometers, S5 performance. https://dcc.ligo.org/cgi-bin/DocDB/ShowDocument?docid=6564

[18] Barish B and Weiss R, LIGO and the detection of gravitational waves. Physics Today, 1999, 52:44~50

[19] Cutler C and Thorne K, An overview of gravitational-wave sources. General Relativity and Gravitation, 2002, 2001:72-111

[20] Zhao W and Zhang Y, Relic gravitational waves and their detection. Physical Review D, 2006, 74:043503

[21] Brady P, Creighton J and Wiseman A, Upper limits on gravitational-wave signals based on loudest events. Classical and Quantum Gravity, 2004, 21: S1775~S1781

[22] Abbott B, el al., First upper limits from LIGO on gravitational wave bursts. Physical Review D, 2004, 69:102001

[23] Rollins J, Burst online analyses and position reconstruction. LIGO technical document, LIGO-G0900861-v2, 2009

[24] Chatterji S, el al., Coherent network analysis technique for discriminating gravitational-wave bursts from instrumental noise. Physical Review D, 2006, 74: 082005

[25] Omega Pipeline. https://trac.ligo.caltech.edu/omega/

[26] GraceDB. https://archie.phys.uwm.edu/gracedb

[27] LUMIN. https://ldas-jobs.ligo.caltech.edu/~lumin/s6/

[28] Hughey B, el al., GEM processor status and MDC results. LIGO technical document, LIGO-G1000578-v3, 2010

[29] Rollins J, Low latency transient searches. LIGO technical document, LIGO-G1000405, 2010

[30] Christensen N, el al., LIGO S6 detector characterization studies. Classical and Quantum Gravity, 2010, 27:194010

[31] 6.8 Magnitude Washington Earthquake Rattles Hanford Observatory. http://www.ligo.caltech.edu/LIGO_web/news/0228quake.html

[32] Slutsky J, el al., Methods for reducing false alarms in searches for compact binary coalescences in LIGO data. Classical and Quantum Gravity, 2010, 27: 165023

[33] Chatterji S, Blackburn L, Martin G and Katsavounidis E, Multiresolution techniques for the detection of gravitational-wave bursts. Classical and Quantum Gravity, 2004, 21:S1809~S1818

[34] The LIGO Scientific Collaboration, Tuning matched filter searches for compact binary coalescence. LIGO technical document, LIGO-T070109-01-Z

[35] Mallat S. A wavelet tour of signal processing. Academic Press, 1999

[36] Bizouard M, el al., UPV and hveto results for Virgo. LIGO technical document, LIGO-G1000282, 2010

[37] Saulson P, Garofoli J and Smith J, S6A glitch classification by hveto. LIGO

technical document, LIGO-G0901087, 2009

[38] Isogai T and the LIGO Scientific Collaboration and the Virgo Collaboration, Used percentage veto for LIGO and Virgo binary inspiral searches. Journal of Physics: Conference Series, 2010, 243:012005

[39] 葛余博. 概率论与数理统计. 北京：清华大学出版社，2005. 47~49

[40] Smith J and Leroy N for LIGO DetChar and DQV groups, hveto application for detector characterization. LIGO technical document, LIGO-G0900878-v3

[41] Hild S, et al., A statistical veto method employing an amplitude consistency check. Classical and Quantum Gravity, 2007, 24:3783~3798

[42] Cortes C and Vapnik V, Support-vector networks. Machine Learning, 1995, 20: 273~297

[43] Breiman L, Random forests, Machine Learning, 2001, 45:5~32

[44] Burges C, A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 1998, 2:121~167

[45] Chang C and Lin C, LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[46] Breiman L, el al., Classification and Regression Trees. Boca Raton: Chapman & Hall, 1984.

[47] Narsky I, StatPatternRecognition and Machine Learning in HEP, http://www.hep.caltech.edu/~narsky/spr.html

[48] Bradley A, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognition, 1997, 30:1145~1159

[49] Guyon I and Elisseeff A, An introduction to variable and feature selection. Journal of Machine Learning Research, 2003, 3:1157~1182

[50] Chen Y and Lin C, Combining SVMs with Various Feature Selection Strategies. Feature Extraction, 2006, 207:315~324

[51] Data Monitor Tool. http://www.ligo.caltech.edu/~jzweizig/dmt/DMTProject/

[52] glitchMon. http://www.ligo.caltech.edu/ jzweizig/dmt/Monitors/glitchMon/

[53] hMon. http://emvogil-3.mit.edu/~shourov/fan/dmt/update_20080822.html

[54] Hughey B, el al., GWM status and swift follow-up of G19377. LIGO technical document, LIGO-G1000921-v1, 2010

[55] Brown D, Testing the LIGO inspiral analysis with hardware injections. Classical and Quantum Gravity, 2004, 21:S797-S800

[56] Hodge K, MVSC to your ears?. LIGO technical document, LIGO-G0901106

[57] Chatterji S. The search for gravitational wave bursts in data from the second

LIGO science run:[PhD Thesis]. Massachusetts: MIT Physics, 2005

[58] Khan R and Chatterji S, Enhancing the capabilities of LIGO time-frequency plane searches through clustering. Classical and Quantum Gravity, 2009, 26: 155009 (14pp)

[59] Owen B and Sathyaprakash B, Matched filtering of gravitational waves from inspiraling compact binaries: Computational cost and template placement. Physics Review, 1999, D60:022002

[60] Ester M, el al., A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, 1996, 1996:226-231

[61] Strehl A and Ghosh J, Cluster ensembles – A knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research, 2002, 3: 583-617

[62] Fred A, Finding consistent clusters in data partitions. Proceedings of the 2nd International Workshop on Multiple Classifier Systems, 2001, 2096:309-318

[63] Murphy J, et al., A model for gravitational wave emission from neutrino-driven core-collapse supernovae. The Astrophysical Journal, 2009, 707:1173

[64] Supernovae Data. http://www.stellarcollapse.org/gwcatalog/murphyetal2009

[65] Stuver A and Finn L, GravEn: software for the simulation of gravitational wave detector network response. Classical and Quantum Gravity, 2006, 23:S799

致　谢

衷心感谢导师曹军威研究员对本人的精心指导。感谢曹老师为本人提供了一个研究当今科学前沿问题的机会。他的言传身教将使我终生受益。

在美国加州理工物理系进行三个月的合作研究期间，承蒙 Antony Searle 博士后的热心指导与帮助，不胜感激。

感谢清华计算机系的王小鸽副教授和都志辉副教授，以及清华天体物理中心的周建峰副教授，两年多来的 LIGO 组会不能没有你们！

本课题承蒙国家自然科学基金资助，特此致谢。

声　明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签　名：＿＿＿＿＿＿＿＿　　日　期：＿＿＿＿＿＿＿＿

个人简历、在学期间发表的学术论文与研究成果

个人简历

1987 年 07 月 24 日出生于江西景德镇市。

2004 年 9 月考入清华大学水利水电工程系水利水电工程专业，2005 年 9 月转入清华大学自动化系控制科学与工程专业，2008 年 7 月本科毕业并获得工学学士学位。

2008 年 9 月免试进入清华大学自动化系攻读控制科学与工程硕士至今。

发表论文

[1] Cao J W and Li J W. Real-time gravitational-wave burst search for multi-messenger astronomy. International Journal of Modern Physics D, 2011（SCI 检索，影响因子 1.046）

[2] Li J W and Cao J W. Development of a DMT monitor for statistical tracking of gravitational-wave burst triggers generated from the omega pipeline. Proceedings of 9th Asia-Pacific International Conference on Gravitation and Astrophysics, Wuhan ,China, 2010, 92-101

攻读学位期间参加的研究项目

[1] 美国加州理工访问学生项目，“Omega Pipeline 中的增强能量恢复”（2010.2~2010.5）

[2] LIGO 科学组织（LIGO Scientific Collaboration）正式会员（2009.10 至今）