文章编号:1671-4598(2018)09-0215-04 DOI:10.16526/j.cnki.11—4762/tp.2018.09.046 中图分类号:TP3 文献标识码:A

地震前兆数据的大数据挖掘研究

李秀明[1,2], 尼 勇[2], 刘 磊[3]
(1. 青海民族大学 物理与电子信息工程学院, 西宁 810007; 2. 陕西师范大学 教育学院, 西安 710062;
3. 青海省地震局, 西宁 810000)

摘要: 为了能够合理使用地震前兆数据, 就要实现地震前兆数据的有效挖掘; 目前, 在地震千兆观测项及方法不断增加的过程中, 传统数据分析方法已经无法满足观测数据分析需求, 大数据挖掘技术的出现为地震前兆观测工作带来了较为积极的影响; 首先对现代大数据挖掘技术进行全面的分析, 之后分析地震预报过程中的大数据挖掘技术, 主要流程为分析地震预报方法、寻找地震地区的相关性、获得数据、实现数据预处理, 关联规则数据挖掘算法; 然后对传统地震前兆数据挖掘技术进行创新为基于时间序列数据流的增量式挖掘, 对地震前兆数据挖掘数据进行定义, 从时间序列中抽取模式, 确认重要点, 从而实现数据并行挖掘; 最后对设计的大数据挖掘算法进行验证, 表示文章所分析的地震前兆数据挖掘算法良好, 具有实际使用意义, 能够为相关方面提供参考。

关键词: 地震; 前兆数据; 大数据挖掘

Research on Big Data Mining of Earthquake Precursor Data

Li Xiuming[1,2], Nie Yong[2], Liu Lei[3]
(1. School of Physical and Electronic Information Engineering, Qinghai Nationalities University, Xining 810007, China;
2. School of Education, Shannxi Normal University, Xi'an 710062, China;
3. Qinghai Earthquake Agency, Xining 810000, China)

Abstract: In order to make reasonable use of earthquake precursor data, it is necessary to effectively mine earthquake precursor data. At present, in the process of increasing seismic gigabit observation items and methods, traditional data analysis methods can no longer meet the needs of observation data analysis. The emergence of big data mining technology has brought a positive impact on earthquake precursor observation work. Firstly, the modern big data mining technology is comprehensively analyzed, and then the big data mining technology in the earthquake prediction process is analyzed. The main processes are analyzing the earthquake prediction method, finding the correlation of the earthquake area, obtaining the data, realizing the data preprocessing, and correlating the rule data. Mining algorithm. Then the traditional earthquake precursor data mining technology is innovated into incremental mining based on time series data stream, which defines the data of earthquake precursor data mining, extracts patterns from time series, confirms important points, and realizes data parallel mining. Finally, the design of the big data mining algorithm is verified, which indicates that the seismic precursor data mining algorithm analyzed in this paper is good, has practical use significance, and can provide reference for relevant aspects.

Keywords: earthquake; precursor data; big data mining

0 引言

地震前兆观测指的是实现地震预报及其他地球物理科学研究的基础, 千兆预测数据质量及数量对此研究过程和结果具有直接的决定作用。所以, 前兆观测属于我国地震监测工作中的主要内容。通过我国多年的发展, 我国地震前兆观测系统已经创建成为覆盖全国单位、智能化及涉及多学科的网络观测系统。目前数字化前兆观测系统数据的采样率及精度有了进一步的提高, 也提高了数据量, 增加了前兆台站、前兆台网的数据及数据检查工作量。目前, 地震前兆数据预处理工作还是根据人工检查方式实现, 因为数据量较大, 人工检查方式效率较低。并且, 人工检查过程具有一定的直观性, 不同人员的判断各有不同。数据挖掘就是基于此种需求逐渐发展的学科, 其能够从随机、大量、模糊、有噪声及不完整数据中检测有用信息, 以此对人们提供决策根据。那么, 本文就将大数据挖掘应用到地震前兆数据分析中, 从而解决现代前兆台网数据人工检测效率较低的问题, 以此为目前观测数据大数据库量分析及使用工作的全新方法进行探索。

1 大数据挖掘技术的研究

数据挖能够使人们对信息数据进行分析、理解和使用的全新学科, 海量数据挖掘指的就是从不完全、大量、随机、模糊的实际收集信息中, 利用提炼隐藏在不轻易被人发现的有用信息及知识过程。此都是利用数据挖掘分析得到的知识及信息, 不仅能够被人们所理解, 还能够便于存储、使用及传播。大数据挖掘从出现之后, 此领域备受人

收稿日期:2018-06-06; 修回日期:2018-07-16。

作者简介:李秀明(1978-),女,山西大同人,在读博士,讲师,主要从事计算机网络与远程教育方向的研究。

们的重视。在信息技术不断发展的过程中，通信水平也在不断的提高，大多数行业信息都实现了高度集中。所以，大数据挖掘技术被广泛应用到多领域中。

大数据挖掘属于全新的学科，其中具备了传统领域的思想，比如估计及假设检验、统计学抽样；模式识别、人工智能及机器学习搜索算法、学习理论和建模技术等。以上领域都包括进化计算、最优化、信息论、可视化及信号处理等技术，其被广泛应用到大数据挖掘中。另外，数据库系统还具有有效存储、查询处理、索引的支持，分布式技术能够帮助对海量数据进行处理，还能够在数据无法聚集的过程中一起处理。

大数据挖掘具有完整的方法对实际问题进行解决，根据此分类估计、预测分析、抽象聚类、相关性分组、建模描述可视化及复杂数据类型挖掘，能够实现大量信息的挖掘，此套完整方法在地震检测系统中使用，实现海量数据分析，能够使地震检测时效性及精准度得到有效的提高[1]。

2 地震预报中的数据挖掘

2.1 地震预报方法

地震预报复杂性及科学难度为世界公认，通过长时间的研究及探索，人们总结了地震学预报、千兆预报、地震活动大形势预报及综合预报等，本文所研究的为前兆预报方法。千兆预报是利用对大地形变场、地磁场、应力应变场、重力场、大地电场等地理物理场及物理量异常变化对未来大地震进行预报。对于中短期的地震预报探索分析，得到可靠地震前兆具有重要的意义。实现地震预测的主要内容就是确认地震前兆，地震前兆的重现性理为地震预测基础。

地震数据的主要特点为：1）具有较多的经验性知识。由于大部分预报知识和领域具有密切的关系，一般都是通过地震预报专家经验进行总结的；2）具有较大的数据量。地震前兆观测数据是通过传感器获得流数据，其中的采样频率每秒一次；3）具有较强的实践性。因为地震前兆预测监测要求具有实时性，从而方便对异常现象进行反应。并且要求地震数据具备时序性，由于地震数据和时间具有密切的联系，所以数据之间的时间约束关系较强。简单来说，地震数据和实践具有一定的关系，其是一种时间序列数据。4）具有大量干扰，具有较强的随机性及不确定因素[2]。

2.2 地震地区的相关性

地震和地质构造具有密切的联系，产生地震的原因和板块地震成因及内部地震成因、地震发生时间、地点及强度具有密切的联系。在地震预报科学中，通过长时间的观测研究及经验积累，专家表示的大范围地震活动高涨或者平静的时候，此地区地震活动具有同步涨落。此种距离的两个地区中某个指定震级以上显著地震相伴的现象就是地震相关现象，也就是地震地区相关性。比如华北北部三个地震活动区，图 1 为中山、东区及西区的范围，使用此三个相关地区中，做出图 2 的地震震级时间关系图。通过图中表示，在中区存在震级发生地震前后，和其相邻的东区和西区都有发生地震。

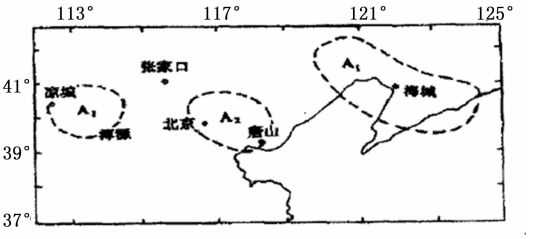

图 1 中山、东区及西区的范围

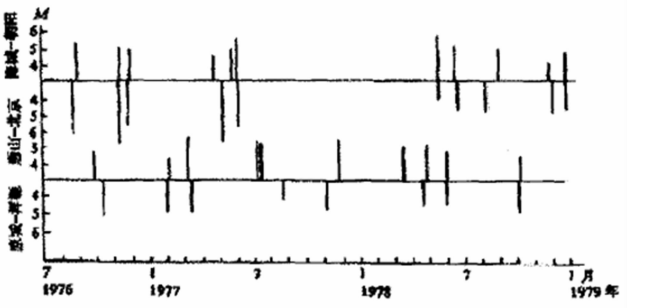

图 2 地震震级时间关系图

在长期观测积累中国地震目录中记录了全国的地震信息，大量地震信息中具有发生地震的规律，此数据挖掘技术也是挖掘隐藏在数据知识及规律产生的。通过地震数据，结合关联规则挖掘算法，能够找到其中的地震知识。具体的工作为：准备数据。选择太原台站数据，对数据进行预处理，之后将数据转换成为满足关联规则挖掘算法的地震事件序列；实现地震时间序列关联分析。对于通过数据整理的地震时间序列数据特征及数据量，对关联规则挖掘算法进行全面的研究，使用算法实现关联分析，找到不同地区地震的相关性；实验模拟及结果评价。选择最具代表性的地震数据实现模拟，对结果进行解释及分析[3]。

2.3 数据的获得

太原基准地震台在华北腹部，山西中部，其为最佳的测震地段。太原台站不仅台址的选择理想，而且测震设备及环境齐全。太原地震台被流体、形变、电磁三大学科观测手段覆盖，每项观测手段中都具有二十六个测项分量。观测仪器主要包括数字石英摆倾斜仪、数字体应变仪、数字伸缩仪、磁通门磁力仪、电离层斜测仪、数字化电阻率测量仪等，摒弃台站具有全省的流动地磁观测任务，一共有三十五个观测点。通过观测点得到数据[4]。

2.4 数据预处理

首先，对数据进行噪声处理。利用数据清洗技术能够对不同情况中的缺失问题进行适当处理，在数据清洗过程中使用聚类、桶分及回归技术实现异常点识别及平滑除燥。在地震数据中，噪声主要包括发生地震的时间有误、位置经纬度有误等。时间噪声私用手工处理，比如将 2015 年 12 月 10 日 16：60：00 中的时间替换成为 17：00。

其次，实现数据正规化。正规化能够使数据属性值从原本取值区间中到适当区间中映射，在实现数据挖掘之前实现正规化。一般正规化包括零均值、最小最大及小树尺度三种正规化。假设及为属性 A 的最小值及最大值，那么最小最大正规化的公式为：

$$\alpha' = \frac{\alpha - \min_A}{\max_A - \min_A}(new\_\max_A - new\_\min_A) + new\_\min_A$$

最后，对数据进行变换。为了满足关联分析及时间序列相似性的匹配算法需求，就要对地震目录数据根据地理区域实现划分，分别包括空间跨度、时间跨度及震级预处理，最后转换成为根据不同参数进行划分排列的地震事件序列。实现空间跨度预处理的主要方式就是划分地理位置，并且对其进行分片编号，使用区间标号代替实际经纬度数据值，从而满足空间属性与离散化需求，降低指定连续属性值数量[5]。

2.5　关联规则挖掘算法

关联规则指的是对事物及其他事物相互关联及依存关系的描述，关联规则挖掘属于工人的数据挖掘方法，能够寻找大量数据中项目集的相关联系。关联规则挖掘步骤为：

1）寻找所有频繁项集；

2）通过频繁项集得到强关联规则。

3　基于时间序列数据流的增量式挖掘

3.1　问题定义

因为要使用挖掘结果对模式库进行更新，假如每挖掘一次就更新模式库，不仅会增加服务器负担，并且会影响到挖掘效率。本文使用内存及外存两级式序列模式的存储结构，基于时间窗口找到最新的频繁模式，只需要将最近出现的频繁模式在内存中存储，在超过数量之后，就到模式库中发送更新，之后到内存中将此频繁模式去除，以此保证全部频繁模式状态的监控。图 3 为挖掘的步骤。

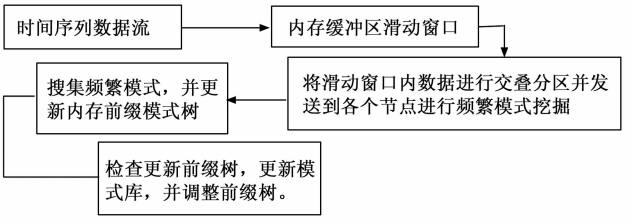

图 3　挖掘的步骤

1）基本窗口。假如 BW 属于基本窗口，其对应数据流子序列，长度为 $bw.size = |BW|$。一个基本窗口对一次模式的基本单元提取，也就是和一个时序模式集合对应。基本窗口 bw 为窗口 w 的分类，w 为基本窗口表示的方法。

2）滑动窗口。假如 SLW 属于滑动窗口，其和连续基本窗口序列对应，表示为 SLW＝$bw1$，$bw2$，bwk。

3）TPSS。对滑动窗口中的时序模式来说，在其中的每个基本窗口都具有 TPSS[6]。

3.2　基于重要点分段

从时间序列中实现模式抽取的方法为：先分割原始时间序列，将其中的子序列转换成为某个高级的数据，之后实现模式发现。本文使用基于重要点分段，此方法的计算时间比较小，能够避免时间序列受到噪声的影响，掌握序列整体变化的特点，其方法较为简单并且有效。重要点指的是序列变化过程中视觉具有主要影响的观测点，也就是序列汇总某部分局部极大、极小的点，图 4 为序列的重要点。

图 4　序列的重要点

假设 $S = < X_1 = (v_1, t_1), \ldots, X_n = (v_n, t_n) >$ 属于时间序列，其中的 v_1 指的是时间 t_1 中的观测值，本文假设△＝1，并且 t_1＝0。

虽然时间序列和全文序列中的基本元素具有区别，时间序列为连续取值实数构成，全文序列为有限字符构成。实现序列的分段，每个字段表示变化模式。之后在模式中实现相似性定义，从而实现符号化，最终将时间序列转化成为符号序列。

对给定常量 $R>1$ 及时间序列 $< X_1 = (v_1, t_1), \ldots, X_n = (v_n, t_n) >$ 进行定义，假如数据点 X_m（$1 \leqslant m \leqslant n$）为主要极小点，那么其要满足以下需求：

在 $1<m<n$ 的时候，下标 i 及 j 为 $1 \leqslant i < m < j \leqslant n$。

给定常量直观的含义就是：X_m 为序列 $X_i, \ldots, x_j$ 的最小值，在此段中的两个端点值比 XmR 大，R 属于可控制选取参数，R 值越大，那么被选中的相对重要点就会越少，时间序列线段的描述就会越粗。所以，利用 R 的选择，能够在不同精细度程度中实现数据挖掘。

因为是在数据流中实现划分，为了能够对流连续性进行保证，只要寻找小于 n 的重要点，最后重要点后面数据在后续数据中分析。

算法 Selesc _ Improtant _ （S，R，S^C）

输入：时间序列 S，选择参数 R；

输出：通过重要点构成序列 S^C；

步骤：（1）i＝find _ first _ importmt _ point（S^C）；

（2）if $I<N$ and $vi \geqslant v_1$ then i＝find _ minimum（i）；

（3）while $i<N$ do

（4）{i＝find _ maximum（i）}

（5）if $i<N$ then i＝find _ minmum（i）

以此表示，此算法只要扫描一次序列，在分段过程中进行简单对比计算，不需要复杂的最小二乘法计算。并且此算法支持序列在线选择。

以算法所获得的重要点集，能够实现序列的逐段线性

化，以此得出通过线段所表示的趋势变化特征模式构成的序列集合 S^C。图 5 为序列分段。

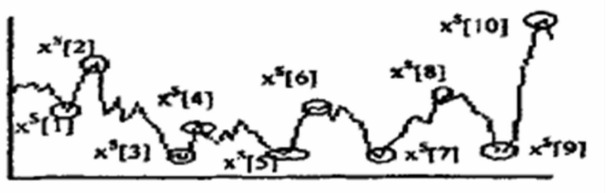

图 5 序列分段

3.3 数据并行挖掘

在数据挖掘过程中不仅要考虑速度，还要考虑最近数据挖掘中挖掘，其结果的精准性比静态数据挖掘结果要差，所以要尽量提高挖掘结果。充分考虑此点问题，在数据挖掘过程使用并行挖掘方法，其思想为：

假如具有 $N+1$ 个处理器，其中 0 属于主处理器。假如 TPSS 具有 K 个基本窗口。算法在全新基本窗口中产生之后删除传统窗口，之后触发模式分析。在第一次执行算法的时候，通过挖掘模式创建前缀树，并且将其在频发模式中使用，创建互联之后到处理机 1 中保存。之后每次触发算法，在处理器 2，...N 中对此数据进行处理，然后创建互联后继树，之后和处理机 1 中的后继树进行合并，并行实现模式挖掘，对频繁模式前缀树进行更新，以频繁模式前缀树实现模式库的封信[7]。图 6 为并行挖掘模型。

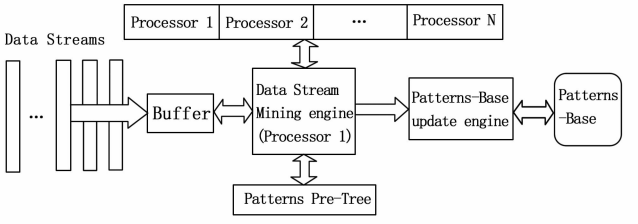

图 6 并行挖掘模型

算法步骤为：

1）在 SIRST 中寻找最早的基本窗口线段序列；

2）对序列中每类线段序列 Cidi 出现次数进行计算；

3）实现 SIRST 中字数的遍历，并且将叶子节点中 Cidi 相应的 tagj 进行修改成为 tagj－Num。

在每次模式分析之后，处理器 0 就会对频繁模式进行收集，对前缀树中的频发模式进行更新，之后对前缀树进行检查，并且将满足规测需求的频繁模式到模式库中发送实现模式库更新，之后对模式前缀树进行调整。以此表示，每次分析序列能够使算法效率得到提高，并且还能够对分析时序列完整性进行保障，是模式库更新频率得到降低。

4 算法验证

将以上所分析的算法通过 Weka 数据挖掘工具实现验证，安装 Weka 工具。Weka 自身是以 Java 所编写的，本文使用可扩展及开放的集成开发工具 Eclipse。算法验证的过程为：

将 Weka 打开，在预处理面板中加载本文所选择的地震前兆数据。之后切换到 Cluster 面板中，单机 Choose 按钮，就能够到下拉菜单中实现 DFCM 算法的导入。单击文本输入框，在所弹出的对象编辑器中使 epsilon 参数设置为 0.5。之后单击 Ignore，在弹出的 Select 窗口中选择 time 属性，关闭此窗口。在 Result 中右击结果列表中的新添加条目，在弹出的菜单中选择簇分配可视化菜单，Weka 就会弹出可视化窗口，可视化界面的结果利用坐标选择不同测项，以此显示所有的结果。

根据以上步骤，通过多次的参数选择，将 DFCM 两个关键参数进行反复的设置对比，以台站测试需求，每年使监测仪器调整为相对零值，并且每年的数据分布形态相同，实验聚类数量为 6，阈值为 0.5。图 7 为聚类结果。

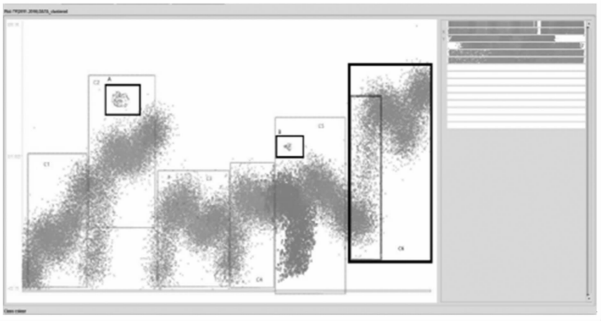

图 7 聚类结果

通过图 7 可以看出来，中间的黄色部分数据为聚类质心群，以颜色对数据程度进行区分，黄色为 0.8～1.0，浅黄色为 0.6～0.8，此种颜色的数据为此簇类。灰色为 0.4～0.6，浅蓝色为 0.2～0.4，其隶属簇权值较大，标记蓝色数据为 0.0～0.2，此属于此类以外数值，以此为相对孤立值。

对测项进行总结，2012 年～2017 年的数据异常结果详见表 1。

表 1 前兆数据的检测结果

测项代码	总天数	异常值		实际异常	准确率
		ε=0.5	ε=0.3		
2231	2191	667	503	593	88.9
2232	2191	635	489	568	89.4
2241/2242	2191	431	372	489	87.1
2243/2244	2191	456	383	512	89.0
232A/232B/2329	2191	615	568	211	34.3
4313	2191	×	×		
9140	2191	217	139	201	92.6

本文一共选择了十三个测项中的十二测项数据实现分析，为 2791 天的数据记录，以实验过程选择聚类结果较好的参数阈值 0.3 和 0.5 实现结果罗列，得到前兆数据处理结

（下转第 241 页）