上海大学出版社
2005年上海大学博士学位论文 96

地震预报中的数据挖掘方法研究

• 作者：吴绍春
• 专业：控制理论与控制工程
• 导师：吴耿锋

上海大学出版社
2005年上海大学博士学位论文 96

地震预报中的数据挖掘方法研究

● 作 者：吴绍春
● 专 业：控制理论与控制工程
● 导 师：吴耿锋

Shanghai University Doctoral Dissertation (2005)

Study on Novel Approaches of Data Ming for Earthquake Prediction

Candidate: Wu Shaochun
Major: Communication Control Theory and Control Engineering
Supervisor: Wu Gengfeng

Shanghai University Press
• **Shanghai** •

上海大学

本论文经答辩委员会全体委员审查，确认符合上海大学博士学位论文质量要求。

答辩委员会名单：

主任：	尤晋元	教授，上海交通大学计算机学院	200030
委员：	顾　宁	教授，复旦大学电子信息学院	200433
	王　炜	研究员，上海市地震局	200062
	乐嘉锦	教授，东华大学计算机学院	200051
	缪怀和	教授，上海大学计算机学院	200072
导师：	吴耿锋	教授，上海大学计算机学院	200072

评阅人名单：

顾　宁	教授，复旦大学	200433
郎志清	教授，华东理工大学	200237
王　娣	研究员，上海市地震局	200062

评议人名单：

周佩玲	教授，中国科技大学	230026
李国徽	教授，华中科技大学	430074
刘宗田	教授，上海大学	200072
郁松年	教授，上海大学	200072

答辩委员会对论文的评语

地震预报是一个国际公认的世界性难题。吴绍春同学的博士学位论文“地震预报中的数据挖掘方法研究”将数据挖掘技术引入到地震预报领域，研究地震数据处理与数据挖掘技术交叉结合的方法，探求提高地震预报准确性的新技术。选题新颖，有理论深度也有实用价值。

该论文的主要研究内容和成果包括：

（1）把地震地区相关性问题转化为时间序列的关联规则挖掘问题，提出并实现了一种基于主从模式设计的并行关联规则算法，通过相关实验和结果分析，得到了一些有价值的地震区域相关性知识。（2）从地震时、空、强三要素出发，提出并实现了基于地震相似度的地震时间序列相似性匹配算法，进行了不同粒度、不同时间差的地震序列相似性实验分析，取得了可信度较高的结果。（3）提出了一种基于广义约束规则的序贯模式挖掘算法，能有效地从地震震例数据中发现广义地震序列，为领域专家进行地震序列的相似性研究提供了支持。（4）基于时序分析技术，研究了地震前兆观测数据的处理方法，提出并实现了多变量时间序列相似性比较算法、动态划分时序模式挖掘算法等。（5）实现了一个基于集群系统的地震预报并行数据挖掘平台，为运行论文提出的各种挖掘算法提供了良好的平台。

论文结构合理，条理清晰，内容翔实，论述清楚，有创新

性。表明作者掌握了坚实宽广的基础理论和系统深入的专门知识，科学研究工作能力强。

吴绍春同学在答辩过程中，叙述清楚，回答问题正确。

答辩委员会表决结果

经答辩委员会表决，全票同意通过吴绍春同学的博士学位论文答辩，建议授予工学博士学位。

答辩委员会主席：尤晋元

2005 年 9 月 17 日

摘　要

地震预报是一个国际公认的世界性难题。我国地震预报事业经过30多年的发展，积累了丰富的宝贵经验和大量的数据资料，全国的地震台网更是每日都在记录着数以千兆计的海量地震前兆观测数据。本文将数据挖掘技术引入到地震预报领域中，研究现有地震数据处理与数据挖掘技术交叉结合的方法，充分应用现代高性能计算环境，从这些海量数据中挖掘出地震预报所需的规律性知识，以便辅助领域专家提高地震预报的准确性。

在探讨现阶段数据挖掘算法模型及其实现基础上，本文首先对地震预报的传统方法（地震震例数据和前兆观测数据分析）进行探讨。同时，围绕地震地区相关性分析、地震序列分析和地震前兆的规律性认识等关键问题进行分析研究，实现并行关联规则算法、基于地震相似度的时间序列相似性匹配算法以及序贯模式挖掘算法。然后，基于时序分析技术，提出一系列地震前兆观测数据处理模型和并行实现算法。最后，结合实际应用实现一个地震预报并行数据挖掘平台，为地震预报数据挖掘的海量数据处理提供强大技术支持。

本文的主要创新性工作包括：

1. 基于关联分析技术研究地震相关地区的搜索方法，提出并实现了一种基于主从模式设计的并行关联规则算法 FPM-LP（Fast Parallel Mining of Local Pruning）。本文把地震地区相关性问题转化为时间序列的关联规则挖掘问题，通过相关的实

验和结果分析，挖掘出许多有价值的地震区域相关性知识。

2. 基于时间序列相似性匹配技术对地震地区相关性进行分析，实现了基于相似度的地震时间序列相似性匹配算法 WSM3S（Whole Sequence Matching Based-on Seismo Similiarity Support）。本文从地震三要素时、空、强的三维角度，给出了地震相似度定义和时间序列相似性匹配模型及算法。通过分析近二十年来我国地震活动频繁区域的历史数据，应用该算法进行多种不同粒度、不同时间差的序列相似性实验分析，取得了可信度较高的结果。

3. 基于序贯模式挖掘技术进行地震序列分析的研究，提出并实现一种基于广义约束规则的序贯模式挖掘算法 SPBGC（Sequential Pattern Mining Based on General Constrains）。本文将地震序列的相关领域知识定义为一组广义约束规则，应用该算法从地震震例数据中挖掘广义地震序列，为领域专家进行地震序列的相似性研究提供强有力的支持。

4. 基于时序分析技术重点研究地震前兆观测数据的处理方法，提出一系列实用地震前兆观测数据处理并行实现算法。首先，提出基于动态规划的时间扭曲方法进行子序列搜索的相似性度量，能有效地进行考虑噪声、幅度、偏移等问题的不精确时间序列匹配；其次，实现基于奇异值分解的多维时间序列相似性比较算法 SSVD（Similarity Based Singular Value Decomposition）；最后，应用自底向上的时间序列划分方法，提出基于频度的动态划分时序模式挖掘算法 SSMBF（Sub-Sequence Matching Based-on Frequency-Partition）。通过实验证明，该算法能有效地从海量历史数据中挖掘出单变量和多变量组合的时间序列频繁模式，从而辅助领域专家及时发现地震前兆观测数据中的异常信息。

5. 依托本单位自强2000高性能计算环境，实现一个基于Cluster机群系统的地震预报并行数据挖掘平台PDMPEP (Parallel Data Ming Platform for Earthquake Prediction)。以Cluster机群系统作为地震预报数据挖掘平台的并行支撑环境，基于面向对象技术封装平台的数据挖掘算法库；同时，提出并行数据挖掘管理中间件的概念，运用软件分层思想，将并行挖掘算法的运行控制及相关的数据操纵功能从平台中分离出来，极大提高了系统的鲁棒性和可扩展性。

注：本文背景课题“地震预报中的数据挖掘方法研究”得到上海市自然科学基金(项目编号03ZR14038)与国家地震科学联合基金(项目编号104090，与上海市地震局合作)的共同支持。

关键词 地震预报，数据挖掘，关联规则，时间序列，序贯模式，地震序列，地震地区相关性，地震前兆观测数据，并行数据挖掘平台

Abstract

Earthquake prediction is a worldwide challenging problem. With the development of earthquake prediction in the past 30 years, a large amount of prior knowledge and billions of data have been accumulated in our country. The gigantic auspice data under earthquake conditions is recorded by the sensor network of seismological observatory everyday. In this paper, we introduce the advanced data mining techniques into the earthquake prediction field, and several novel approaches between data mining and seismological data analysis are studied. Meanwhile, just by using the techniques of high performance computing and parallel data mining, seismological domain knowledge hidden in the gigantic data can be efficiently discovered to support earthquake prediction, therefore the accuracy of the earthquake prediction can be improved effectively.

On the basis of discussing the existing data mining algorithms, the paper mainly focuses on the domain knowledge of seismology and the traditional methods for earthquake prediction. Then, by using the relativity analysis on earthquake zones, the earthquake sequences and the rules of earthquake auspice data, it carries out several parallel data mining algorithms such as association rules based parallel

mining algorithm, the seismological similarity and similarity-matching algorithm realization, and the sequential pattern mining algorithm etc. Furthermore, the earthquake auspice data processing method and a series of parallel implement algorithms are proposed based on the technique of time series analysis. Finally, the parallel seismological data mining platform is implemented, which integrates all of the algorithms proposed in this paper.

The main contribution of the dissertation is shown as follows:

1. By analyzing and discovering the earthquake catalogue data on the relativity of earthquake zones, a Master/Slave mode based parallel mining algorithm FPM-LP (Fast Parallel Mining of Local Pruning) is put forward by using association rules, just as well as the relative preprocessing algorithm is presented. The experimental results demonstrate that the algorithm is satisfactory to find relative earthquake zones.

2. On the basis of analyzing the relative earthquake zones on the technique of time series similarity matching, the seismological similarity and similarity-matching model on the relative earthquake zones and its algorithm WSM3S (Whole Sequence Matching Based-on Seismo Similarity Support) are proposed according to the three earthquake essential factors, which are named time, space, intensity separately. Just by analyzing the historical earthquake data of seismological zones, the useful relative earthquake can be easily found according to the sequence matching experiment with different

kind of granularity.

3. Just taking seismological knowledge as general constraints to restrict the sequential pattern, a sequential pattern mining algorithm based on general constrains (SPMGC) is presented in this paper. The remarkable result of this algorithm is to discover the earthquake sequence and can help the domain experts to study the similarity of earthquake sequence.

4. Based on the time series analyzing method, it put emphasis research on the processing method on earthquake auspice data, and then several data processing algorithms are proposed for those data. Firstly, a sub-sequence similarity measurement by using dynamic time warping is proposed, which can be applied to process imprecise matching among time series with considering those problem, such as white noise, amplitude, excursion. Secondly, a time series similarity measurement using singular value decomposition is put forward to compare different multivariate time series. Thirdly, with the bottom-up segment approach to segment time series, the algorithm SSMBF (Sub-Sequence Matching Based-on Frequency-Partition) is implemented in this paper. Experimental results show that: the algorithm can be used to discover singular and multivariate time series frequent pattern efficiency from tremendous history data, so as to help the domain experts to find abnormal information rapidly and accurately in the large amount of earthquake auspice data.

5. Depending on the ZIQANG 2000 high performance

computing system in Shanghai University, a Cluster based data mining platform PDMPEP(Parallel Data Ming Platform for Earthquake Prediction) is implemented, which integrates several data mining toolbox engine, therefore both sequential and parallel programs can be executed under the parallel environment. Meanwhile, data mining toolboxes are developed by using object-oriented techniques, and a middleware for parallel data mining management is also proposed. Furthermore, the flow control of parallel data mining algorithm is separated with the data management on the key idea of software modular, which can help to improve the robustness and extensibility of PDMPEP.

Key words Earthquake Prediction, Dada Mining, Association Rules, Time Series, Sequential Pattern, Earthquake Sequence, Relativity of Earthquake Zones, Earthquake Auspice Data, and Parallel Data Ming Platform.

目　录