成果与方法

基于 Weka 的地震数据挖掘系统的设计与实现*

董晓娜[1,2]，段会川[1]，张文娟[3]

(1. 山东师范大学信息科学与工程学院，山东 济南　250014；2. 山东省地震局，山东 济南　250014；3. 山东省地质科学实验研究院，山东 济南　250013)

摘要：基于开源的数据挖掘系统 Weka，使用 Java 语言及面向对象的思想，设计并实现了地震数据挖掘系统。根据地震数据资料的特点，将数据挖掘的核心技术(聚类分析、关联规则分析等)引入到该系统中，其中聚类分析选用 DBSCAN 作为核心算法，关联规则分析选用 Apriori 作为核心算法。用户使用该系统只需在交互界面选择相关参数，即可实现调用数据挖掘算法来分析地震数据，发现探索其隐含规律。

关键词：Weka；地震；数据挖掘；聚类分析；关联规则

中图分类号：TP131　　**文献标识码**：A

0　引言

所谓数据挖掘，就是从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、人们事先不知道的，但又是潜在的有用的信息和知识的过程[1]。目前，数据挖掘技术已在工业、农业、商业、金融、服务、医学、气象等很多领域得到很好的应用，因为其广泛的应用价值，数据挖掘学科领域汇聚了众多不同领域的研究者。地震预测是十分复杂的世界性科学难题，在长期的观察和研究实践中，已经积累了大量的十分宝贵的地震数据资料和经验知识，而且还有很多潜在的、没有被人们认识的知识和规律隐藏其中。从大量数据中提取规律、挖掘知识，正是数据挖掘技术的核心。因此，将数据挖掘的理论和技术引入地震研究领域是需求所驱[2]。

该文旨在设计并实现基于 Weka 的地震数据挖掘系统，为研究人员使用数据挖掘算法分析地震资料提供一个平台。研究人员只需通过该系统的交互界面，根据需要选择合适的算法，并输入相关核心参数，即可方便地调用数据挖掘的算法来分析数据，而不需要再详细研究算法本身或撰写代码。最后，对得到的挖掘结果进行分析，去伪存真，发现并探索其隐含规律，总结经验，进行地震预测研究。

1　Weka 简介

Weka 的全名是怀卡托智能分析环境(Waikato Environment for Knowledge Analysis)，是一款基于 Java 环境的、开源的机器学习(Machine Learning)及数据挖掘(Data Minining)软件[3]。Weka 作为一个公开的数据挖掘工作平台，得到了广泛的认可，被誉为数据挖掘和机器学习历史上的里程碑，是现今最完备的数据挖掘工具之一。

Weka 最大的优势在于其开源性，它的源代码可以很方便地在官方网站获得。因此开发人员可以根据实际需要，使用 Java 语言，通过实现其接口或是继承其方法，开发出适用于专业领域的数据挖掘应用系统。另外，还可以利用 Weka 的基本架构，改进原有的算法或开发新的算法，以满足实际应用。

2　地震数据挖掘系统的设计

2.1　基于 UML 的建模分析

该文使用统一建模语言(UML)作为需求分析、系统设计的表达语言，采用面向对象的可视化建模工具 Rational Rose 来进行可视化建模。

为了让系统使用方便、操作简易，该文从用户

* **收稿日期**：2008－09－04；**修订日期**：2009－07－25；**编辑**：陶卫卫

作者简介：董晓娜(1981—)，女，山东济南人，工程师，主要从事地震监测工作。

的角度来描述软件需求,分析软件所需的功能。在地震数据挖掘系统中共涉及 3 类用户,分别是:数据库管理员、数据分析人员和领域专家。其中数据库管理员负责管理数据库(包括数据的及时更新、数据库维护、升级、备份等);数据分析人员的操作有:打开数据源(包括文本文件、数据库或 URL),执行数据挖掘操作(包括数据预处理、聚类分析、关联规则分析等),查看日志(关于操作及最终结果的日志记录)等;领域专家则是利用领域知识、经验来解释和评价最终输出的结果,并给出所发现的知识(即隐含的规律)。使用 UML 的用例图对该系统进行建模(图 1)。

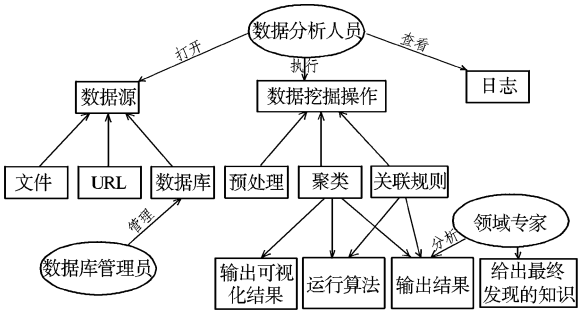


图 1　系统设计的用例图

对于该系统的核心算法,即聚类算法、关联规则算法的操作过程及其生命周期的描述是整个系统的重中之重。该系统的执行操作过程如下:首先打开数据源(可以选择文件或数据库),然后进行数据预处理(根据需要选择预处理的类型),再选择挖掘算法(聚类分析或是关联规则分析),然后点击 start 即可执行。如参数选择不当,可能导致挖掘失败,即未给出结果,或是给出的结果离期望相差较大,这时需要重新选择参数再次挖掘直到得到期望的结果为止。可见参数的选择对挖掘的最终结果起到了至关重要的作用。最后对输出结果给出解释评价并提取知识,即挖掘成功。使用 UML 的状态图对整个系统的执行操作过程及其生命周期进行建模(图 2)。

2.2　核心算法的选择

2.2.1　聚类

目前,聚类挖掘中较常用的聚类算法有 K - means 算法、K - medoids 算法及 DBSCAN 算法。前两者聚类结果趋于圆形,对于随机分布数据聚类效果较差。而 DBSCAN 算法是一个典型的基于密度的聚类方法,可以快速发现任意形状的类。鉴于地震数据在分布上的随机性,该文选择 DBSCAN 算法。

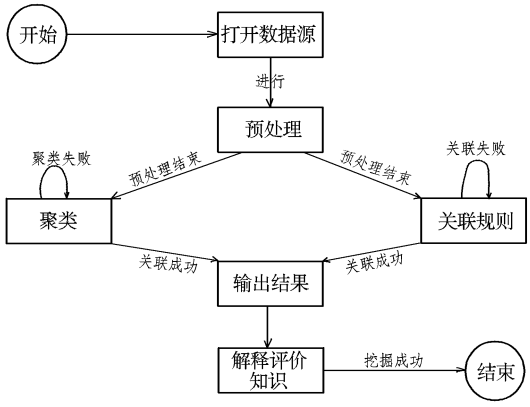


图 2　系统状态图

2.2.2　关联规则

目前,关联规则挖掘中较常用的算法有 Apriori 算法和 FP - Tree 算法。其中 Apriori 算法是挖掘产生关联规则所需频繁项集的基本算法,也是最著名的关联规则挖掘算法之一。该算法利用了一个层次顺序搜索的循环方法来完成频繁项集的挖掘工作。这一循环方法就是利用 k -项集来产生(k +1)-项集。鉴于地震数据的特点,选择 Apriori 算法。

3　地震数据挖掘系统的实现

3.1　开发环境

操作系统为 Windows XP,开发平台为 JCreator 4,JDK 版本为 J2SDK1.6.0。测试数据库为 Microsoft SQL server 2000 + SP3,数据库驱动为 Microsoft SQL server 2000 Driver for JDBC SP3。

3.2　主要开发过程

该文基于开源的数据挖掘系统 Weka,使用 Java 语言开发了地震数据挖掘系统。在数据的导入及最终挖掘结果的显示部分,该系统继承了 Weka 系统的源代码。在数据的预处理部分,该系统针对地震数据的特点(数据的海量、复杂、高维性、各属性之间的非线性关系、数据的缺值及干扰等),在 Weka 原有处理方法(如:数据类型转换、数据离散化等)的基础上增加了数据完整性及一致性的检验、噪声数据的处理等功能。在数据挖掘具体操作部分,根据地震数据资料的特点,聚类分析选用 DBSCAN 算法作为核心算法,关联规则分析选用 Apriori 算法作为核心算法。

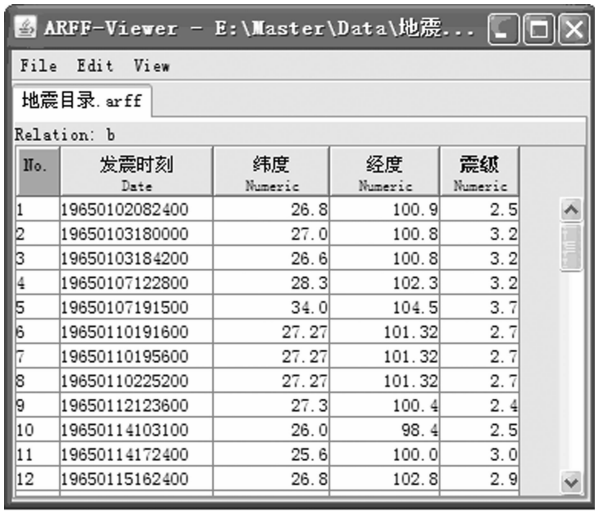
3.3　运行环境

该系统由 Java 语言开发,采用 C/S 模式的架构,具有可跨平台,不需安装,使用方便,操作简易等特点。用户可在任意操作系统使用该系统,使用之前只需安装与所使用操作系统匹配的 Java 运行环境(JRE)即可。如果用户选择数据库作为数据源,需根据数据库的类别下载安装相应 JDBC 驱动。

4　地震数据挖掘系统的功能及测试

该文以一组地震目录作为数据源,来测试该系统的功能及使用情况。

(1)打开数据源:根据要求可选择从文件或数据库中提取相关数据,即将目标数据导入到系统中(图 3)。


ARFF-Viewer - E:\Master\Data\地震...

File Edit View

地震目录.arff

Relation: b

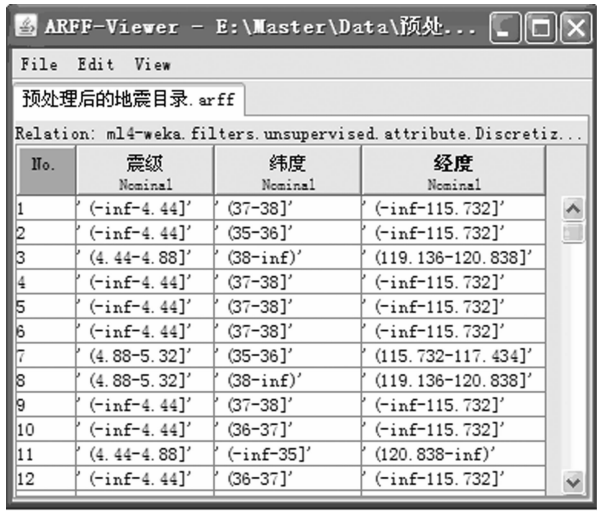
No.	发震时刻 Date	纬度 Numeric	经度 Numeric	震级 Numeric
1	19650102082400	26.8	100.9	2.5
2	19650103180000	27.0	100.8	3.2
3	19650103184200	26.6	100.8	3.2
4	19650107122800	28.3	102.3	3.2
5	19650107191500	34.0	104.5	3.7
6	19650110191600	27.27	101.32	2.7
7	19650110195600	27.27	101.32	2.7
8	19650110225200	27.27	101.32	2.7
9	19650112123600	27.3	100.4	2.4
10	19650114103100	26.0	98.4	2.5
11	19650114172400	25.6	100.0	3.0
12	19650115162400	26.8	102.8	2.9

图 3　目标数据导入到系统中

(2)数据变换:目的是对高维数据进行属性约简(降维),为便于数据挖掘,需要从初始特征中找出真正有用的特征,去除对挖掘结果无关的属性[4];这里主要关心地理位置上的特性,因此将“发震时刻”这个属性去掉,从而达到降维目的。

(3)数据预处理:先要对数据进行数据完整性及一致性的检验,对其中的噪声数据进行处理。然后根据需要,进行适当的数据类型转换,如做聚类分析,需要把连续型数据转换为离散型的数据(图 4)。

(4)聚类:使用 DBSCAN 算法来实现,结果将给出类的个数及每一条记录所属的类。用户可以根据需要来选择参数,参数选的不同,聚类结果就会不同。通过不断尝试,选择最佳参数以得出最佳结果。最终结果以数据形式或图形化形式给出,比较实用


ARFF-Viewer - E:\Master\Data\预处...

File Edit View

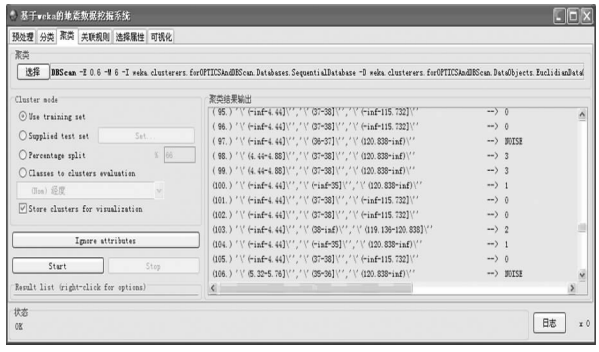
预处理后的地震目录.arff

Relation: ml4-weka.filters.unsupervised.attribute.Discretiz...

No.	震级 Nominal	纬度 Nominal	经度 Nominal
1	'(-inf-4.44]'	'(37-38]'	'(-inf-115.732]'
2	'(-inf-4.44]'	'(35-36]'	'(-inf-115.732]'
3	'(4.44-4.88]'	'(38-inf)'	'(119.136-120.838]'
4	'(-inf-4.44]'	'(37-38]'	'(-inf-115.732]'
5	'(-inf-4.44]'	'(37-38]'	'(-inf-115.732]'
6	'(-inf-4.44]'	'(37-38]'	'(-inf-115.732]'
7	'(4.88-5.32]'	'(35-36]'	'(115.732-117.434]'
8	'(4.88-5.32]'	'(38-inf)'	'(119.136-120.838]'
9	'(-inf-4.44]'	'(37-38]'	'(-inf-115.732]'
10	'(-inf-4.44]'	'(36-37]'	'(-inf-115.732]'
11	'(4.44-4.88]'	'(-inf-35]'	'(120.838-inf)'
12	'(-inf-4.44]'	'(36-37]'	'(-inf-115.732]'

图 4　预处理后的数据

方便。该文使用一组地震目录进行聚类(图 5)。


基于weka的地震数据挖掘系统

预处理 分类 聚类 关联规则 选择属性 可视化

聚类

选择 DBScan -E 0.6 -M 6 -I weka.clusterers.forOPTICSAndDBScan.Databases.SequentialDatabase -D weka.clusterers.forOPTICSAndDBScan.DataObjects.EuclideanData

Cluster mode

Use training set

Supplied test set　Set...

Percentage split　%　66

Classes to clusters evaluation

(Num) 经度

Store clusters for visualization

Ignore attributes

Start　Stop

Result list (right-click for options)

聚类结果输出

(95.) '(-inf-4.44]','(37-38]','(-inf-115.732]'　--> 0

(96.) '(-inf-4.44]','(37-38]','(-inf-115.732]'　--> 0

(97.) '(-inf-4.44]','(36-37]','(120.838-inf)'　--> NOISE

(98.) '(4.44-4.88]','(37-38]','(120.838-inf)'　--> 3

(99.) '(4.44-4.88]','(37-38]','(120.838-inf)'　--> 3

(100.) '(-inf-4.44]','(-inf-35]','(120.838-inf)'　--> 1

(101.) '(-inf-4.44]','(37-38]','(-inf-115.732]'　--> 0

(102.) '(-inf-4.44]','(37-38]','(-inf-115.732]'　--> 0

(103.) '(-inf-4.44]','(38-inf)','(119.136-120.838]'　--> 2

(104.) '(-inf-4.44]','(-inf-35]','(120.838-inf)'　--> 1

(105.) '(-inf-4.44]','(37-38]','(-inf-115.732]'　--> 0

(106.) '(5.32-5.76]','(35-36]','(120.838-inf)'　--> NOISE

状态

OK

日志　x 0

图 5　聚类结果

在地震预测中聚类的应用领域也非常广泛,如地震异常与正常数据的聚类、有震样本与无震样本的聚类、地震序列类型的划分、地震知识的获取[4,5]等。吴绍春[6]用 DBSCAN 算法对地震目录数据进行聚类,得出的结果与地震预测专家根据地质构造和地震活动分布的情况划分的地震带基本吻合。使用该系统,用户可以很方便地调用聚类算法对地震数据进行分析。

(5)关联规则:使用 Apriori 算法来实现,结果给出的是所挖掘出关联规则的最优规则[7](Best rules found)。用户可以根据需要选择不同的参数,如设置 numRules = 10,则返回找到的关联规则中最优的 10 条。该文使用一组地震目录进行关联规则分析(图 6)。

关联规则算法在地震预测中也有很多应用,如邢殿勇等[8]和吴绍春等[9]使用关联规则算法寻找地震相关地区,发现了一些以往人们没有发现的地震相关地区,同时也验证了专家的一些结果。关联

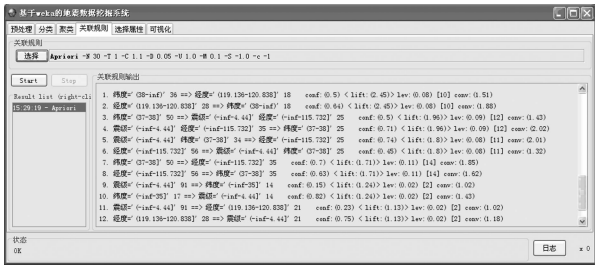


图 6　关联规则输出结果

规则算法还可用于地震预测知识发现、异常和干扰发现等方面。使用该系统,用户可以很方便地调用关联规则算法对地震数据进行分析。

(6)解释评价:数据挖掘得出的结果是潜在的、未知的、不确定的。因而也不排除有可能得到模糊的、错误的结果。对于挖掘出的结果,需要专家利用领域知识、经验做出评价,摒弃其中无用的错误的信息,找到真正有用的信息(发现知识)。然后将发现的知识以用户能理解的方式表示,再根据实际情况对知识发现过程中的具体处理阶段不断进行优化,直到满足要求。

5　结论

该文基于开源的数据挖掘系统 Weka 设计并实现了地震数据挖掘系统,根据地震数据资料的特点,将数据挖掘的核心技术(聚类分析、关联规则分析等)引入到该系统中。该系统由 Java 语言开发,采用 C/S 模式的架构,可跨平台、不需安装、使用方便、操作简易。随着研究深入,可能随时需要对系统进行升级,因此采用面向对象的思想,这样无论是改进原有的算法还是增加新的算法,都不需对源代码做大的修改,具有良好的可扩展性。随着数据挖掘技术在地震研究领域应用的深入,将会有更多的算法被引入,有助于研究人员进一步开展地震预测研究。

参考文献:

[1] (加)HAN JW,KAMBER M. 数据挖掘概念与技术[M]. 范明,孟小峰,译. 北京:机械工业出版社,2005:3-6.

[2] 王炜,林命遇,马钦忠,等. 数据挖掘及其在地震预报中的应用前景[M]. 国际地震动态,2005,(12):1-2.

[3] (新西兰)W ITTEN LH, FRANK E. 数据挖掘实用机器学习技术(2 版)[M]. 董琳,邱泉,于晓峰,等译. 北京:机械工业出版社, 2006: 241-243.

[4] 吴淑芳,吴耿锋,王炜,等. 一种新的模糊规则提取方法[J]. 计算机工程,2005,31(6):157-160.

[5] 王炜,吴耿锋. 模糊联想记忆神经网络模型在地震预报中的应用[J]. 地震学报,1997,19(3):254-260.

[6] 吴绍春. 地震预报中的数据挖掘方法研究[D]. 上海大学博士学位论文,2005:16-21.

[7] 许军. 基于公安信息的数据挖掘应用研究[D]. 南京工业大学工学硕士学位论文, 2006:33-39.

[8] 邢殿勇,吴绍春,王炜,等. 并行关联规则算法在地震地区相关性预报中的应用[J]. 计算机应用研究, 2005, 22(10):175-177.

[9] 吴绍春,吴耿锋,王炜,等. 寻找地震相关地区的时间序列相似性匹配算法[J]. 2006,17(2):185-192.

Design and Implementation of Seismic Data Mining System Based on Weka

DONG Xiao-na[1,2], DUAN Hui-chuan[1], ZHANG Wen-juan[3]

(1. Information Science and Engineering College of Shandong Normal University, Shandong Jinan 250014, China; 2. Shandong Earthquake Administration Bureau, Shandong Jinan 250014, China; 3. Shandong Institute and Laboratory of Geological Sciences, Shandong Jinan 250013, China)

Abstract: Based on open source data mining system of the Weka, by using Java language and object-oriented ideas, seismic data mining system is designed. According to characteristics of seismic datas, the core technology of datas mining (cluster analysis, association rules analysis, etc.) are led into the system. Among them, the DBSCAN algorithm is regarded as a core selection in cluster analysis, while Apriori is selected as the core of association rule analysis. The system user can simply select the relevant parameters in intersection boundary, and data mining algorithms which are used to analyze the seismic datas and find the laws can be gained.

Key words: Weka; seismology; data mining; cluster analysis; association rules