基于中药提取的数据挖掘系统设计与实现①

李耀芳 (天津城市建设学院 电子与信息工程系 天津 300384)

摘 要: 系统主要应用数据挖掘方法对中药提取数据进行分析和预测。首先对数据进行集成和离散化处理，得到适合数据挖掘的数据集，然后利用 k-means 和 DBSCAN 聚类算法对质检数据进行聚类，得到工艺参数质检区间；并对 Apriori 算法进行了改进，在算法中加入了用户兴趣度的概念，控制了候选集指数增长，得到工艺参数和固含量的关系；并利用三层 BP 神经网络算法训练网络模型，得出过程参数和结果质量参数的关系，发现数据中隐含的规律，为企业优化工艺以及提高其生产效率降低成本等提供科学的分析、决策辅助工具。

关键词: 挖掘分析；中药生产；聚类；关联规则；神经网络

Design and Implementation of Data Mining System Based on Traditional Chinese Medicine Extracted

LI Yao-Fang

(Tianjin Institute of Urban Construction, Tianjin 300083, China)

Abstract: Data mining methods are applied to the system to extract data for analysis of Chinese medicine and forecasting. First, the data integration and discrete treatment are made, to obtain appropriate data mining data sets. Then k-means and DBSCAN clustering algorithm are used to cluster the data quality control. Quality control process parameters interval are obtained. Apriori algorithm is improved by adding a user interest degree in the concept. The candidate set of exponential growth is controlled. The relationship between parameters and solids is got. By using three-layer BP neural network algorithm trained network model, the relationship between the process parameters and results quality parameters is obtained. The law implied in the data is found. It provides a scientific analysis and decision support for enterprises to optimize processes and improve their productivity and reduce the costs.

Keywords: mining analysis; Chinese medicine production; clustering; association rules; neural network

1 引言

论文研究背景来源于某中药制造企业的中药提取生产过程，该企业对工艺参数的设置、产品质量检验参数的确定等完全依靠操作人员的经验，导致了产品质量的波动性。此外，在提取过程中，温度、密度、压力、进液量等参数的变化对半成品的质量有影响，企业希望通过使用数据挖掘的算法进行科学的数据分析，得出与生产结果最密切的工艺参数，以及能够生产出合格半成品的工艺参数的数据区间。

为此，论文搭建数据挖掘系统平台，实现对工艺参数数据和质量检验数据的数据挖掘。系统采用 c#.net 程序设计语言设计，结合企业数据特点，使用 k-means 和 DBSCAN 算法实现对单个属性进行聚类[1]，得到中药提取工艺参数的质检区间；改进 Apriori 算法，缩减最大频繁集指数增长速度，发现工艺参数和固含量之间的关联规则；并利用神经网络的三层 BP 网络结构，以工艺参数作为输入，固含量作为输出，建立了过程和记过的函数模型，对输出结果固含量进行预测，得到固

① 基金项目:天津市软件专项基金(07FZRJGX03200, 08FZRJGX02100)

收稿时间:2009-12-15;修稿时间:2010-01-09

© 中国科学院软件研究所 http://www.c-s-a.org.cn

含量预测区间，进一步验证关联规则结果的准确度。

1　系统总体设计

1.1 系统体系结构

系统采用 c#.net 环境设计，使用的 C/S 架构，由四个主要模块组成：数据准备、数据整理、数据分析以及数据分析结果可视化模块。系统的主要功能模块如图 1 所示：

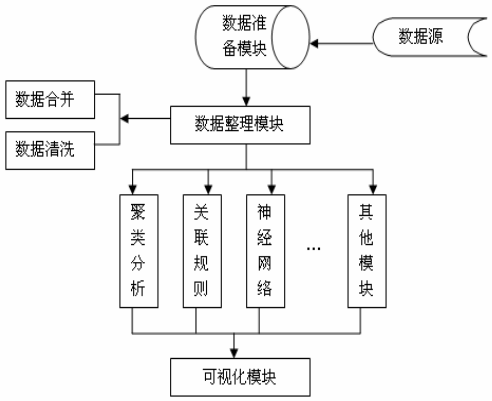

图 1　中药生产数据挖掘系统功能结构图

(1)数据准备与整理

系统包含了添加多个数据文件进行分析的功能，用户在进行数据分析时无需通过手动或第三方软件将多个文件合并成一个做数据准备，此论文涉及的中药生产数据按照日期分别存放在不同的文件中，在导入这些文件时，按记录合并成一个完整的数据集，并进行了数据清洗，以保证能够挖掘到更准确的信息。

(2)挖掘分析功能

①聚类分析功能

系统设计的聚类分析[2]主要针对所有字段均为连续型的整型、浮点型等数字类型数据，详见 2.1.1。

②关联规则分析功能

包含了对属性、置信度等的自动设置，详见 2.2。

③神经网络分析功能

以关联规则的结果作为原始数据，训练神经网络预测模型，详见 2.3。

2　数据挖掘分析功能

包括聚类分析、关联规则和神经网络三个功能模块。

2.1 聚类分析模块

根据企业对工艺参数质检数据的需求，聚类分析模块实现了单属性聚类。本节详细阐述了聚类分析模块的设计思路、设计流程和实现方法。如图 2 所示聚类分析流程图：

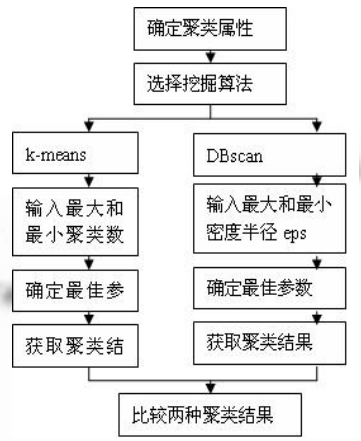

图 2　聚类分析流程图

2.1.1 聚类算法的参数选择

很多聚类算法需要输入聚类个数或其他如密度、半径等参数当作初始值，而这些参数对于决定聚类结果的优劣至关重要，不同的参数能够得到完全不同的聚类结果。系统设计由程序完成对参数的判断选择，采用了 goodness Index 值判断法[3]，具体描述如下：

①给出一个参数范围，得到不同的聚类结果。

②通过推导公式计算每个聚类的 goodness Index。

③寻找最小的 goodness Index，由此得出相应的参数，这个参数即为所求。

2.1.2 根据企业数据特点确定聚类算法

分析企业需要聚类的数据特点，得知所有数据属性均为连续属性，而企业的需求是尽量将取值相近的数据分为一组，所以最好的分组方法是根据数据间的曼哈顿距离进行分组，将取值相近的数据分为一组。

系统提供两种聚类算法供用户选择：k 均值和 DBSCAN，用户可以选择某种算法，也可以对应用两种算法的聚类结果，通过比较数据同构度和异构度来得到最好的聚类结果。

2.2 关联规则模块

针对企业的数据特点和需求设计了数据挖掘系统，图 3 显示了系统关联规则模块分析的流程。

2.2.1 改进的 Apriori 算法

本节对 Apriori 算法[4]进行了一些改进，在支持数

© 中国科学院软件研究所 http://www.c-s-a.org.cn

的基础上增加了左右侧属性的限制，基于用户感兴趣的属性降低了候选项集指数增长，提高了算法速度。

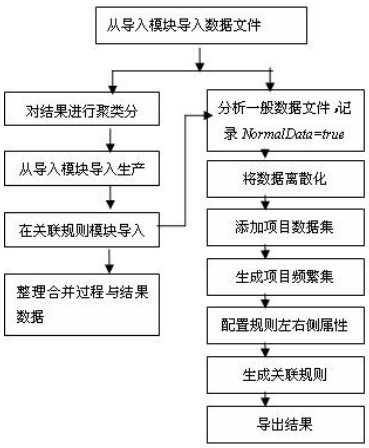

图 3　关联规则模块运行流程图

Apriori 算法描述如下：

①C1={candidate1-itemsets};
②L1={c∈C1|c.count≥minsupport};
③For(k=2,Lk-1≠Φ,k++)
④Ck=sc_candidate(Lk-1);
⑤for all transactions　t∈D
⑥Ct=count_support(Ck,t);
⑦for all candidates c∈Ct
⑧c.count=c.count+1;
⑨next
⑩Lk={c∈Ck|c.count≥minsupport};
⑪next
⑫resultset=resultset∪Lk

2.2.2 根据改进 Apriori 算法获取最大频繁集

在限制最小支持度的基础上，系统增加了限制左右侧属性的条件，假设系统设定关联规则左侧允许出现的属性集合为 SL，右侧属性为 SR，对于产生的每一个候选 k 项集 CKi，只选择系统设定的项集 CKi (SL∪SR)的项集，这样在一定程度上缩小了候选集的数量。

即在上面的算法④和⑤之间加入下面的控制：

Cs=belong（SL∪SR）//选择只属于 SL∪SR 的项集

将⑥改为 Ct=count_support(Cs,t)，以下几行均由 Cs 代替 Ck。

2.2.3 关联规则系统分析界面

界面图示如图 4 所示，图中给出了主要功能的设计：合并数据、数据清洗、生成项目频繁集、关联规则发现。系统可分析普通数据文件，也可分析多条过程数据对应一条结果数据的数据文件、可配置规则左右侧属性、可配置划分区间数量、置信度。

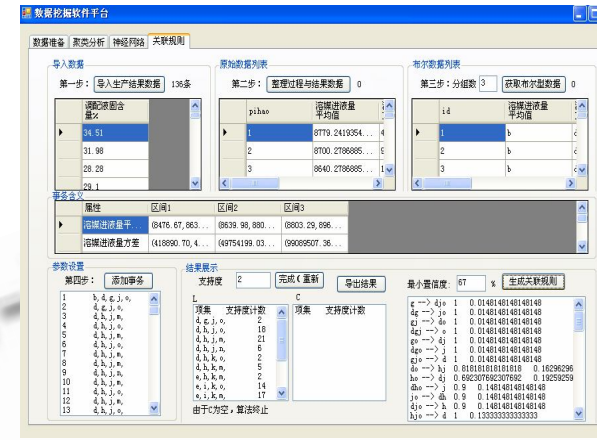

图 4　关联规则发现模块界面

2.3 神经网络模块

神经网络模型[5]采用的基本算法是 BP 神经网络算法，并采用三层网络结构。

下图为神经网络模块流程图：

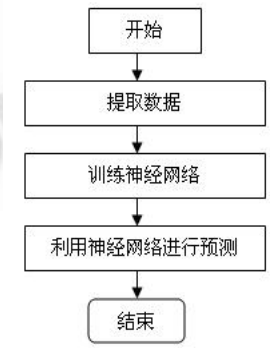

图 5　神经网络模块流程图

2.3.1 确定输入和输出节点

为了确定工艺参数和固含量的关系，我们选择数据中除固含量外的各个属性作为输入节点，包括：

提取进液量均值、提取进液量方差、提取温度均值、提取温度方差四个属性，系统通过这几个生产过程的信息对固含量进行预测；

选择固含量作为输出节点，如图 6 所示，这样可以通过训练好的网络对中药提取生产的半成品进行固含

© 中国科学院软件研究所 http://www.c-s-a.org.cn

量的预测。中间层单元数由公式 $m=\sqrt{n+l}+\alpha$ 确定[4]，其中 m 为中间节点个数，n 为输入结点个数，l 为输出节点个数，α 为 1-10 之间的常数，据此可取 m=10 作为中间节点个数。

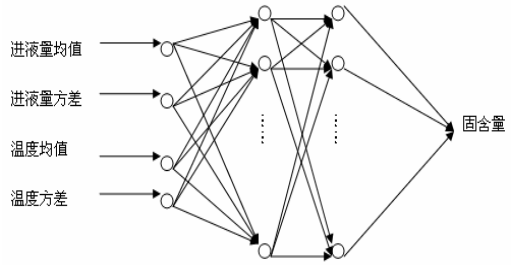

图 6　三层 BP 神经网络应用示意图

2.3.2 训练模型

该模块的重点是训练神经网络，程序从全部导入数据中提取三分之二作为训练数据，计算中间变量已得到很好的训练模型。BP 算法训练网络包括单样本调整误差和全部样本综合调整误差两种情况，分析中药提取的数据样本，发现数据之间的差异性不强，这样采取单样本调整误差对就可以达到训练的效果，因此本文采取了单样本调整误差[5]的方式。

预测模块处理需要和上面的模型训练相结合，以上系统对 2/3 的数据进行了模型训练，建立好训练模型，接下来对剩余的 1/3 数据进行预测，进一步判断训练模型的准确性。

3　系统运行结果

将采集的提取数据导入到系统中，分别对数据进行聚类、关联规则和神经网络挖掘，得到系统运行结果。

3.1 聚类分析结果

对质检数据的七个质检参数进行聚类，得到水煮液相对密度、检测温度、固含量、调配液的检测温度、相对密度、固含量以及喷雾干燥的水分这七个属性的聚类数据结果和图形结果，同时对原来的质检指标进行了优化，如表 1 所示。

表 1　降压避风片提取生产检测指标优化前后对比

	水煮液			调配液			喷雾干燥
	相对密度（g/ml）	检测温度（℃）	固含量（%）	相对密度（g/ml）	检测温度（℃）	固含量（%）	水分（%）
原来	1.054-1.066	53-60	12.21-17.19	1.121-1.456	55-58	24.1-29.96	<=6.52
优化	1.058-1.095	55-58	12.21-17.19	1.11-1.22	50-61	24.1-29.5	<=6.42

3.2 关联规则结果

针对过程和结果集成的数据集，使用本系统对数据进行关联规则挖掘分析，得到提取和浓缩过程属性即提取进液量、提取温度、浓缩液密度、浓缩液温度对固含量的关系，取得过程与结果之间定量的分析结果，如表 2 所示：

表 2　关联规则运行结果

规则	置信度	支持度
提取过程对固含量关系		
进液量方差(418890.70,37420371.95),循环温度均值(62.15,62.60), 循环温度方差(766.06,27129.81)--> 调配液固含量(25.11,29.96)	0.846	0.096
浓缩过程对固含量关系		
温度方差(16093.41,31750.07)，密度均值1.02，密度方差(0.46,0.91) --> 调配液固含量(23.05,29.10)	0.676	0.251

3.3 神经网络结果

神经网络模块采用导出的关联规则模块整合数据，属性包括提取进液量均值、提取进液量方差、提取温度均值、提取温度方差和调配液固含量。并以这四个过程参数为输入节点，固含量为输出节点，利用 BP 算法训练网络，对固含量进行取值区间的预测。

经过训练的模型的预测精度为 82.1%，预测的固含量范围基本准确。利用关联规则的结果，任意选取得出的规则左侧属性各个取值区间的数值当做输入，预测输出属性的取值区间，得到了固含量的预测区间，并且和关联规则右侧属性的区间相同。

© 中国科学院软件研究所 http://www.c-s-a.org.cn

4 系统设计难点问题的解决方案

该系统设计的第一个难点在于系统的运行要适用于不同数据，而且属性、数据量、数据格式等不固定，系统要根据不同的数据类型进行不同的预处理，例如对数字类型数据进行均值填充空白数据，而对其他非数值型数据则没有这一项分析；对于某个属性字段为矩阵的数据，需要将数字矩阵按转换为数值。

第二个设计难点在于神经网络的输入节点和输出节点的设计，一般神经网络都是具体问题具体分析，没有一个固定的模式适用于全部问题，而本系统的设计适用范围广，并不是说一个神经网络结构适用与很多实际问题，而是用户可以自行设置网络结构，根据不同的实际问题自己配置输入层和输出层以及中间节点，从而达到预测结果的目的。

第三个设计难点在于关联规则模块的数据离散化系统由程序自动完成数据的离散化，只需用户给出划分区间的数目即可，在离散化时，需要记录不同属性不同区间的符号以及属性名称、属性标记，使系统增加了难度，但是却方便了用户操作。

5 结语

论文针对中药提取生产面临的实际情况和当前数据挖掘软件的特点，研究并提出了一个数据挖掘软件平台，详细论述了聚类分析、神经网络、关联规则的数据分析结构设计与难点解决，并最终实现了各种连续型数据的挖掘分析工作。该系统在实现分析中药提取质量分析数据以及过程数据的同时，能够对各种其他领域的不同数据进行数据挖掘功能分析，具有灵活部署、自由配置和可扩展的优点，较好地解决了中药提取生产数据的质量检验标准优化以及工艺参数优化的问题。

参考文献

1 JIAWEI HAN. 数据挖掘概念与技术. 北京:机械工业出版社, 2006.

2 孙吉贵,刘杰,赵连宇.聚类算法研究.软件学报，2008,19(1):48－61.

3 Gyenesei A. Fuzzy Partitioning of Quantitative Attribute Domains by a Cluster Goodness Index. Turku Centre for Computer Science TUCS Technical Report NO 368, October 2000

4 Tang ZH, 麦克雷南.邝祝芳,焦贤龙,高升译.数据挖掘原理与应用.北京:清华大学出版社, 2007.

5 韩力群.人工神经网络理论、设计及应用. 北京:化学工业出版社, 2007.

© 中国科学院软件研究所 http://www.c-s-a.org.cn