基于数据挖掘分类法的农用地分等

王 璐[1,2]，田 剑[3]，刘建敏[3]

（1．华南农业大学信息学院，广州 510642； 2．中科院广州地球化学研究所，广州 510640；
3．合肥工业大学资源与环境工程学院，合肥 230009）

摘 要：应用决策树模型、BP 神经网络和 Logistic 回归模型等分类法，对龙川县农用地分等进行了实证研究，并对各方法的分等结果有效性进行了评价，同时利用混淆矩阵探讨了样本数量对 3 种模型分类精度的影响。结果表明，样本数量对模型影响有差异，其中对 BP 神经网络和决策树模型影响较大，在较多训练样本时，模型的精度较高。在较多样本支持下，BP 神经网络精度最高，但训练模型的时间较长，可解释性差；决策树模型既具有较高的精度又具有良好的可解释性；Logistic 回归模型表现较差。决策树模型最适合龙川县农用地分等工作。研究结果表明，数据挖掘分类法是有效而准确的土地评价方法，有助于提高土地评价的精度和准确性，对农用地分等方法的优化具有一定的借鉴意义。

关键词：分类，BP，神经网络， Logistic，决策树，农用地分等
doi：10.3969/j.issn.1002-6819.2009.08.047
中图分类号：TP183 **文献标识码**：A **文章编号**：1002-6819(2009)-8-0262-06

王 璐，田 剑，刘建敏．基于数据挖掘分类法的农用地分等[J]．农业工程学报，2009，25(8)：262－267.
Wang Lu, Tian Jian, Liu Jianmin. Farmland classification based on data mining classification method[J]. Transactions of the CSAE, 2009,25(8)：262－267.(in Chinese with English abstract)

0 引 言

数据挖掘是在大量的数据中发现潜在的[1]、有价值的模式和数据间关系（知识）的过程。利用数据挖掘进行分类、预测已经成为当前研究的热点[2]。从数据挖掘的角度来看，土地评价实质上属于分类预测问题。将数据挖掘应用到土地评价中，将极大地促进土地评价的自动化和智能化的水平；提高从土地评价数据库中提取相关属性的速度与准确性；并有助于建立合理的土地评价模型，探寻土地利用中各种影响因素间的潜在关系，对土地的质量变化做出科学评估和预测。

数据挖掘在土地科学领域中的应用尚处于起步阶段，其中以决策树模型、神经网络、数理统计和聚类分析应用较多[3]。决策树模型作为一种监督分析方法，具有良好的稳健性和鲁棒性，在土地覆盖分类领域中得到成功的应用[4-5]，在土壤质量划分方面有着显著的效果[6-8]。随着人工神经网络的成熟，它也在土地评价中有了初步的应用[9-10]。根据不同的评价目的，研究者逐步实现了对神经网络模型的优化研究[11]，通过从神经网络中抽取土地评价的模糊规则[12]，来增加神经网络模型在土地评价中的可理解性和评价结果的准确性。数理统计作为数据分析的基本手段，在土地评价中也有广泛的应用，如应用 Logistic 回归分析土地退化的态势[13]，探讨区域土地特征与土地利用之间的关系[14]，模拟分析县域土地利用格局的变化[15]等，从而为区域的土地利用规划提供科学依据。

目前对于数据挖掘分类模型在土地领域的应用已经有许多成功的案例，但对不同模型分类效果的研究还不够深入；如 Aalders 等[16]人虽然研究了农业普查中土地利用分类方法和各主要分类要素之间的关系，但缺乏对分类模型性能的探讨。

农用地分等是当前土地评价工作的主要内容之一，它是指在特定的目的下，根据农用土地潜在的、相对稳定的自然属性和经济属性，对农用土地的质量进行综合鉴定并划分出具有跨区域可比的性的土地质量等别。农用地分等的总体思路为：从光、温条件出发计算各有关作物的光温产量；按地块的条件评定各有关作物的理论产量；在标准耕作制度下计算土地总理论产量，评定土地潜力等级；根据土地利用水平进行修正，完成土地质量等级评价；按投入产出水平进行修正评定，实现土地经济评价；实现全国范围内的等级相互可比。

农用地分等具体实施是在光温或气候生产潜力、标准耕作制度等国家参数的控制下，在省级国土资源部门的统一组织协调下，以县级行政区为基本区域通过逐级修订得到县域内部可比的农用地分等成果；在此基础上，通过检验、控制、调整、平衡、衔接等汇总技术，逐级汇总得到跨区域横向可比的地市级、省级乃至国家级农用地分等成果。

随着土地评价目的多样化，基于不同时空尺度差异，合理选用分类模型至关重要。本研究以广东省河源市龙川县农用地分等为例，探讨数据挖掘分类模型在土地评价中的实际应用效果。

收稿日期：2009-03-26 修订日期：2009-08-10
基金项目：国家自然科学基金（40671145，60573115）；国家星火计划（2006EA780057）；华南农业大学校长基金（5600-K05165）
作者简介：王 璐（1976－），女，河北博野人，讲师，博士生，主要研究方向：地理科学与地理信息系统应用。广州 华南农业大学信息学院，510642。Email: selinapple@163.com

1　材料与方法

1.1　数据准备

1.1.1　数据来源

研究数据来源于广东省河源市龙川县国土局、农业局、统计局等相关部门和龙川县农用地分等实地调查数据，具体包括：龙川县土地利用现状图、龙川县土壤普查资料、龙川县统计年鉴、龙川县农用地分等外业调查数据等。

在分等过程中，结合龙川县农业用地的实际情况，参考《农用地分等规程》和《广东省农用地分等定级与估价项目技术方案》中广东省二级区耕地评价因素因子的设置方案，选取地形、田面坡度、地下水位、有效土层厚度、土壤表层质地、剖面构型、表层有机质含量、pH 值、灌溉保证率、排水条件等 10 个评价因子。评价单元划分采用叠置法。在 MAPGIS 系统环境下，从龙川县 2004 年 1∶10 000 土地利用现状图中分离提取出耕地图斑，作为工作底图。将选取的评价因子要素图层与工作底图进行叠加，最终得到 30 281 个图斑为龙川县耕地评价单元。按照《农用地分等规程》计算得出龙川县农用地自然质量等别分为 4 等，处于 13 等级到 16 等级之间，主要集中在 14 等级和 15 等级；说明龙川市农用地生产潜力、土壤状况和立地条件等比较优越。其中，16 等级表示土壤质量最好，13 等级表示最差。所有评价因子的属性数据和土壤质量状况数据从龙川县土壤资源数据库中获取，其中农用地的样本总数为 30 281 个，其中 13 等地样本数为 4 854 个，14 等地样本数为 18 051 个，15 等地样本数为 7 016 个，16 等地样本数为 360 个。

为了证实数据挖掘分类的性能，本研究选取数据挖掘分类中常用的监督分类器决策树模型、BP 神经网络与数理统计中的 Logistic 回归模型 3 种模型，应用广东省河源市龙川县农用地分等数据，在大量数据情况下，对 3 种分类模型进行比较，探讨数据挖掘分类模型和常规统计分类模型在土地评价中表现出的性能差异，并与农用地分等的成果进行比较。

1.1.2　样本选取

训练数据的质量很大程度上影响着分类模型的精度，文献[17]中使用聚类的方法选取训练样本，取得了较好的效果，保证了选取样本的代表性。本研究运用聚类方法分别选取了 500、1 000、2 000、4 000、6 000 和 8 000 学习样本，测试样本采用全部的评价单元，即 30 281 个评价单元。

1.2　分类基本理论

分类是数据挖掘中的一个重要问题，旨在生成一个分类函数或分类模型，该模型能把数据库中的数据项映射到给定类别中的某一个。分类和回归都可用于预测，预测的目的是从历史数据记录中自动推导出对给定数据的推广描述，从而对未知数据进行预测评价[18]。

数据分类可描述为：给定一训练数据的集合 T，T 中的元素记录有若干属性描述。在所有属性中有且仅有一个属性作为类别属性。属性集合用矢量 $X=(X_1, \cdots, X_n)$ 表示，其中 X_i（$1 \leqslant i \leqslant n$）对应各非类别属性，可以具有不同的值域，即对于任一属性 $X_i=\{x_1, \cdots, x_{mi}\}$，$mi$ 随属性的不同而变化。用 C 表示类别属性，$C=(C_1, \cdots, C_k)$，即数据集有 K 个不同的类别。那么，T 就隐含地确定了一个从矢量 X 到类别属性 C 的映射函数 $H: f(x) \to C$，分类的目的就是采用某种方法（模型）将该隐含函数 H 表示出来。各种数据挖掘获得知识的表示形式主要有 5 种：规则、决策树、知识基、网络权值和公式[16]。

分类过程主要包含以下两个步骤：

第一步，根据给定的训练集，找到合适的映射函数 $H: f(x) \to C$ 的表示模型。这一步通常称为模型训练阶段。

第二步，使用上一步训练完成的函数模型预测数据的类别，或利用该函数模型，对数据集中的每一类别进行描述，形成分类规则。

1.3　模型的建立

1.3.1　建立决策树

本研究建立决策树的目的，是对农用地的质量进行等级评价，即质量等级的划分，确定质量等级是决策树的分类属性。作为模型的输出属性，它是一个独立的数据变量，为离散型数据，分为 1 级、2 级、3 级和 4 级（分别对应农用地分等原始成果的 13 等、14 等、15 等和 16 等）。以农用地分等规程确定的 10 个评价指标作为模型的输入属性，并且在数据预处理中将这 10 个评价指标进行离散化处理。运用 Clementine8.1 数据挖掘软件，引入误差权重，采用 C5.0 算法生成土地评价的决策树。

1.3.2　建立 BP 神经网络模型

龙川县土地评价的指标因素共计 10 个，设计 BP 神经网络的输入层节点为 10 个，确定输出层节点为 4 个，即：1 等、2 等、3 等和 4 等（分别对应农用地分等原始成果的 13 等、14 等、15 等和 16 等）。对模型输出的变量用 0-1 的格式表示，1 等地输出变量为[0，0，0，1]，2 等地输出变量为[0，0，1，0]，3 等地输出变量为[0，1，0，0]，4 等地输出变量为[1，0，0，0]。本研究 BP 神经网络采用一个隐含层结构，根据 Kolmogoror 定理，确定了隐层的节点个数为 21 个，神经网络隐含层的神经元传递函数采用 S 型正切函数 tansig，输出层的神经元传递函数采用 S 型函数 logsig，需要函数的输出变量位于区间[0，1]中，经过规格化处理的训练样本和测试样本正好满足网络输出的要求。设定神经网络学习速率为 0.5，使用 Levenberg-Marquardt 学习算法建立评价模型。将学习样本输入到评价模型中，进行 BP 神经网络的学习，由此实现从输入（影响因素）到结果（质量等级）之间映射知识的获取，即分别获得网络单元之间的连接权值向量及各隐含层的阈值向量。

1.3.3　建立 Logistic 回归模型

在龙川县土地评价中，土地等级划分为 1、2、3 和 4 等级（分别对应农用地分等原始成果的 13 等、14 等、15 等和 16 等），系统将以变量值为 4 定义为参照类，建立因变量分别是 1、2 和 3 的 3 个二项分类 Logistic 回归模

型。在 Logistic 回归方程中，因变量是土地质量的等级，自变量是 10 个评价因子。

Logistic 回归方程求解参数采用最大似然估计法，因此通过似然函数值检验回归方程，由于似然函数值是个极小的小数，一般取其自然对数再乘以 2 检验，即-2 倍对数似然值（-2 L *L* 值）[19]。由表 1 的各项检验指标可以看出该模型整体和自变量因素的解释作用都是十分显著的。

表 1 Logistic 回归模型检验值

Table 1 Test value of Logistic regression model

模型	-2 倍对数似然值（-2 L *L* 值）	卡方检验（Chi-Square）	自由度（*Df*）	显著水平（*Sig*.）
只含常数项的模型	8607.867			
最终模型	1206.008	7401.859	30	0.000

Logistic 回归模型拟合信息表明，χ^2（Chi-Square）=7401.859，*P*（*Sig*.=0.000）＜0.001，本模型具有显著性意义，模型拟合度较好。模型的似然比检验表明，本模型的回归系数均有显著意义（*P*＜0.01）。此时获取知识形式是 3 个预测回归方程式，利用这些预测回归方程式可以预测其他区域的土地质量等级。

2 结果与分析

2.1 样本数量对分类精度的影响

在 6 组不同训练样本下的 3 种模型的总体分类精度变化见图 1，从图 1 中可得出：

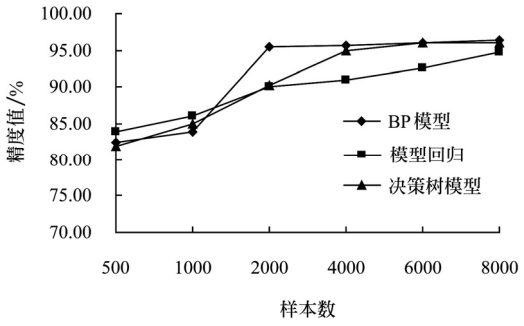

图 1 不同样本的 3 种模型的分类准确率曲线

Fig.1 Classification accuracy curve with three models by different samples

1）3 种模型随样本增多，精度曲线呈上升趋势。在 500 和 1 000 样本训练下，模型预测精度不高，且变化平缓。Logistic 回归模型精度高于决策树模型和 BP 模型精度。

2）在 2 000 个样本方面，3 种模型预测精度有明显上升，其中 BP 模型精度最高，达到 94.97%，比 1 000 个样本提高了 10.14%，体现了神经网络的优势，增加了 1 000 样本此时对神经网络影响大，对 logistic 回归模型和决策树模型影响较小。

3）在 4 000 个样本训练下，3 种模型总体精度较高，决策树模型精度变化最大，比 2 000 样本决策树模型精度高 4.26%，此时增加样本，对决策树模型影响最大，对神经网络影响最小。

4）在 6 000 和 8 000 个样本支持下，3 种模型精度有缓慢上升，提高程度不明显。只有 Logistic 回归模型变化较大，说明 4 000 个样本对评价模型有较好的训练效果，保证了模型的预测精度。

在 6 组样本空间进行训练下，对 3 种模型的总体分类精度进行了组内最大值和最小值差异计算，结果见表 2。从表 2 可知，训练样本大小对 Logistic 回归模型影响最小，对决策树模型和 BP 神经网络模型影响较大，但是对 3 种模型影响较大，均在 10%以上，可见 3 种分类模型对训练样本大小要求较高，因此进行分类挖掘时，需要严格地考虑训练样本大小和数据质量，在建立挖掘模型之前，要对数据进行预处理工作。

表 2 样本数量对不同分模型的影响

Table 2 Influence of the number of samples on different models

分类模型	精度差异/%
决策树模型	13.66
BP 神经网络	13.58
Logistic 回归模型	11.48

2.2 模型混淆矩阵分析

按照《农用地分等规程》得出的龙川县农用地分等原始结果为：样本总数为 30 281 个，其中 13 等地样本数为 4 854 个，14 等地样本数为 18 051 个，15 等地样本数为 7 016 个，16 等地样本数为 360 个。以此为依据，对 3 种分类模型的农用地分等结果的比较分析采用混淆矩阵进行。混淆矩阵作为分类规则特征的表示，它包括了每一类被正确分类的个数和被错误分类的个数。混淆矩阵中的主对角线表示了每一类被正确分类的观测数据个数，非对角线上的元素则表示未被正确分类的观测数据个数。运用 4 000 个学习样本建立的决策树模型、BP 神经网络和 Logistic 回归模型等分类评价模型对龙川县的全部 30 281 个评价单元进行检验，评估 3 种模型的泛化能力，得到 3 种评价模型的混淆矩阵如表 3 所示。

表 3 3 种模型的混淆矩阵

Table 3 Confusion matrix of three models

			1 等	2 等	3 等	4 等
模型预测值	决策树模型	1 等	2 017	244	0	0
		2 等	168	10 910	301	0
		3 等	0	482	14 392	88
		4 等	0	0	96	1 423
	BP 神经网络	1 等	2 122	405	0	0
		2 等	233	11 029	448	0
		3 等	0	202	14 308	181
		4 等	0	0	33	1 330
	Logistic 回归模型	1 等	1 730	300	0	0
		2 等	615	10 372	480	
		3 等	0	964	14 237	71
		4 等	0	0	72	1 440

注：为计算方便，1 等、2 等、3 等、4 等分别对应农用地分等原始成果的 13 等、14 等、15 等和 16 等。

从混淆矩阵可以看出，3 个评价模型中没有出现“跳级”的误判，只有在相邻的等级出现了分类错误。从准确率来看，BP 神经网络模型的误判率比决策树模型的要略小一点，而 Logistic 回归分类模型的准确率最低。

表 3 的混淆矩阵中，矩阵对角线代表了各个等级预测的正确评价单元个数。由该表可知，决策树模型混淆矩阵中 2 等地和 3 等地之间的误判个数最多，其次是 1 等地与 2 等地之间的误判，而 3 等地与 4 等地之间误判最少，主要原因是与评价单元的分布个数有关。同时决策树模型引入了误差权重，提高了模型预测精度，使得模型的误判分布比较均匀，其评价结果与实际结果接近。

BP 神经网络预测结果中，BP 神经网络混淆矩阵对角线以下的误判比对角线以上的误判要多，2 等 3 等在 2 等地和 3 等地中尤为明显，表明误判质量差的评价单元个数要少于误判质量好的评价单元，说明了 BP 神经网络评价的结果整体要好于实际结果。且 BP 神经模型在预测 4 等地误差较大，是 3 个模型中准确率最低，可能原因是 4 等地学习样本较少，也体现了神经网络模型对学习样本的依赖。

Logistic 回归模型的混淆矩阵中对角线以上误判数要大于对角线以下的误判数，误判质量差的评价单元个数要多于误判质量好的评价单元，这与 BP 神经网络模型的结果相反，说明了 Logistic 回归模型评价的结果整体要劣于实际结果。并且在该模型对 4 等地的判断率最高。

由数据挖掘分类模型的分析结果可以得出，3 种评价模型在较多的学习样本支持下，保证了评价模型的学习精度。按模型评价标准来看：

1）从模型的拟合效果来看，建立的 BP 神经网络与决策树模型精度相差不大，小于 1%，相对来说，改进 BP 神经网络模型优于决策树模型，而 Logistic 回归模型分类模型精度最低。在龙川县数据方面，各种模型误差分布也不同，决策树模型误差分布比较均匀，BP 神经网络模型易于将优等地判定为劣等地，Logistic 回归模型易于将劣等地判定为优等地。

2）从模型的可解释性来看，决策树模型能提取相关的评价规则，可以同时处理数值型和非数值型数据，清楚地显示某个预测变量的相对重要程度，可用来构建明确的统计规则模型，取代传统土地评价中使用的不明确的思维模型；BP 神经网络评价是“黑箱”操作，得到的学习知识是各个神经元权值，可以获取神经网络评价规则，需要进一步的研究，如：文献[12]研究了从神经网络中抽取模糊规则的方法，则增加了知识挖掘的难度，限制了 BP 神经网络的实际运用范围，这样比决策树模型得到有效的规则要复杂；而在 Logistic 回归模型中，结合数理统计，简单直接，通过学习得到 3 个预测回归方程，仅需要对这些预测方程进行分析即可。

3）从模型计算的复杂度来看，在相同的硬件条件下，在龙川县农地分等研究中，决策树模型和 Logistic 回归模型建立较快，而 BP 神经网络模型训练的时间比决策树模型和 Logistic 训练模型的时间要长。

综合考虑分类模型的评价标准，可知决策树模型比 BP 神经网络和 Logistic 回归模型更适合对龙川县农地质量进行评价，该结果与 Aalders 等学者[16]研究结果一致。实践证明，运用决策树模型进行的龙川县农地分等研究，满足了实际工作的要求。

3　结论与讨论

3.1　结论

本研究利用广东省龙川县农用地分等数据，采用决策树模型、BP 神经网络和 Logistic 回归模型进行了农用地分等研究，具体结论如下：

1）样本数量直接影响到模型的精度。其中，样本数量对决策树模型和 BP 神经网络模型的影响要比 Logistic 回归模型要大；在样本数量增加时，BP 神经网络模型比决策树模型和 Logistic 回归模型精度提高要显著；在一定学习样本支持下，决策树模型和 BP 神经网络能得到较好的精度。

2）在模型预测准确率方面，BP 神经网络和决策树模型精度较高，Logistic 回归模型准确率较低；在模型可解释性方面，以决策树模型最好， Logistic 回归模型次之，BP 神经网络最差；在模型训练时间上，决策树模型和 Logistic 回归模型建立较短，BP 神经网络模型建立时间较长。

3）通过对 3 种模型的比较研究表明，决策树模型最适合龙川县土地评价，使用决策树评价模型可以为农用地分等工作提供一种的新的解决方案，减少了评价中人为因素的影响，保证评价结果科学性。

3.2　讨论

1）数据挖掘分类法相对传统土地评价方法而言，具有更高的分类精度。

龙川县农用地分等成果是按照《农用地分等规程》的方法实现的，其方法简单易行、移植性强、易于推广，但在权重确定中过分依赖经验知识，不能对知识的不完整性做出调整，不准确的知识往往带来较大的偏差；而本研究所采用的数据挖掘分类模型采用规则形式表达，与常规统计方法相比较，具有易理解性和较高的精准性，可面向大量数据的分类；该方法克服了传统评价方法过于依赖与经验知识的缺陷，从而为土地评价研究提供一种新的思路与方法，并对进一步完善《农用地分等规程》提供支持。

2）应用数据挖掘方法进行农用地分等及土地评价研究，将弱化人为因素的影响，使评价结果更接近实际。

《农用地分等规程》中主要采用累加求和的计算方法，在计算农用地自然生产潜力的基础上，利用土地利用系数和土地经济系数逐级修正得到不同层次的农用地等别，这种方法在实际操作中相关指标和参数可调整的余地比较大，即人为影响较大；而应用数据挖掘方法则对数据样本的有效性要求比较高，在操作过程中的实际调整的空间比较小，与农用地外业调查和基础数据处理的精度直接相关；因此，在具体应用中可以结合不同地区数据获取的实际情况，因地制宜，选择恰当的方法进行农用地分等实践，以期逐步完善农用地分等的方法体

系。

3）土地评价目的和数据样本的可靠性与数量直接影响到分等成果的有效性和准确性。

在农用地分等以及土地评价实践中，模型的分类效果不但与分类特征、训练样本和分类模型本身有关，还与土地评价的目的有关，这主要取决于数据挖掘过程中对问题的分析与理解程度。此外，原始数据的准确性、样本数量和可靠性也影响到分类结果的有效性和准确性。

随着数据获取技术的提高和数据获取手段的多样化，大量与空间位置相关的数据被收集，人们迫切需要强有力的数据分析工具来从这些数据中获取信息或知识，空间数据挖掘的出现便满足了这种需求，使得数据挖掘技术和 3S 得到了有机的结合，并为农用地分等研究提供了更为有效的方法与手段，这也是今后农用地分等及土地评价方法研究的重点方向。

［参 考 文 献］

[1] 罗泽旺. 数据挖掘在水资源分析评价中的应用研究[D]. 南京：河海大学，2006.
Luo Zewang. The Research of Application in the Water Resources Analysis and Assessment Based on Data Mining[D]. Nanjing: Hohai University，2006. (in Chinese with English abstract)

[2] 范明，范宏建. 数据挖掘导论[M]. 北京：人民邮电出版社，2006：89－122.
Fan Ming, Fan Hongjian. Introduction to data mining[M]. Beijing: Posts & Telecom Press，2006：89－122. (in Chinese)

[3] 鲍艳，胡振琪，柏玉，等. 主成分聚类分析在土地利用生态安全评价中的应用[J]. 农业工程学报，2006，22(8)：87－90.
Bao Yan, Hu Zhenqi, Baiyu, et al. Application of principal component analysis and cluster analysis to evaluating ecological safety of land use[J]. Transactions of the Chinese Society of Agricultural Engineering，2006，22(8)：87－90. (in Chinese with English abstract)

[4] Debeir O, Van D S I, Latinne P, et al. Textural and contextual land-cover classification using single and multiple classifier systems[J]. Photogrammetric Engineering and Remote Sensing，2002，68(6)：597－605.

[5] Pal M, Mather P M. An assessment of the effectiveness of decision tree methods for land cover classification[J]. Remote Sensing of Environment，2003，86(4)：554－565.

[6] 孙微微，胡月明，刘才兴，等. 基于决策树的土壤质量等级研究[J]. 华南农业大学学报，2005，26(3)：108－110.
Sun Weiwei, Hu Yueming, Liu Caixing, et al. Soil quality grade evaluation based on decision tree[J]. Journal of South China Agricultural University，2005，26(3)：108－110. (in Chinese with English abstract)

[7] 周斌，王繁. 基于决策树模型的土壤性质空间推断[J]. 土壤通报，2004，35(4)：385－390.
Zhou Bin, Wang Fan. Spatial prediction of soil properties based on decision tree modeling[J]. Chinese Journal of Soil Science，2004，35(4)：385－390. (in Chinese with English abstract)

[8] Elisabeth N B, Brent L H, Karin V. Knowledge discovery from models of soil properties developed through data mining[J]. Ecological Modelling，2006，191(3/4)：431－446.

[9] 赵霈生，陈百明. 在土地评价中应用人工神经网络专家系统的理论与实践[J]. 中国土地科学，1998，12(2)：28－34.

[10] 唐南奇，潭明军. 基于人工神经网络农用地分等研究 I 分等模型与精度检测[J]. 福建农林大学学报(自然科学版)，2004，33(2)：241－244.
Tang Nanqi, Tan Mingjun. Grade of farming land by artificial neural network Ⅰ. The model for grade of farming land and accuracy measurement[J]. Journal of Fujian Agricultural and Forestry University(Natural Science Edition)，2004，33(2)：241－244. (in Chinese with English abstract)

[11] 刘耀林，焦利民. 基于计算智能的土地适宜性评价模型[J]. 武汉大学学报信息科学版，2005，30(4)：283－287.
Liu Yaolin, Jiao Limin. Model of land suitability evaluation based on computational intelligence[J]. Editorial Board of Geomatics and Information Science of Wuhan University，2005，30(4)：283－287. (in Chinese with English abstract)

[12] 胡月明，薛月菊，李波，等. 从神经网络中抽取土地评价模糊规则[J]. 农业工程学报，2005，21(12)：93－97.
Hu Yueming, Xue Yueju, Li Bo, et al. Extracting fuzzy rules from neural networks for land evaluation[J]. Transactions of the Chinese Society of Agricultural Engineering，2005，21(12)：93－97. (in Chinese with English abstract)

[13] 王静，何挺，郭旭东，等. 基于逻辑回归模型的环北京地区土地退化态势分析[J]. 地理科学进展，2005，24(5)：23－32.
Wang Jing, He Ting, Guo Xudong, et al. Study on land degradation trend by applying logistic multivariate regression model in northwest region of Beijing[J]. Progress in Geography，2005，24(5)：23－32. (in Chinese with English abstract)

[14] Gobin A, Campling P, Feyen J. Logistic modeling to identify and monitor local land management systems[J]. Agricultural Systems，2001，67(1)：1－20.

[15] 张永民，周成虎，郑纯辉，等. 沽源县土地利用格局的多尺度模拟与分析[J]. 资源科学，2006，28(2)：88－96.
Zhang Yongmin, Zhou Chenghu, Zheng Chunhui, et al. Spatial land use patterns in Guyuan County: simulation and analysis at multi-scale levels[J]. Resources Science，2006，28(2)：88－96. (in Chinese with English abstract)

[16] Aalders I H, Aitkenhead M J. Agricultural census data and land use modelling[J]. Computer Environment and Urban Systems，2006，30(6)：799－814.

[17] 田剑，胡月明，王长委，等. 聚类支持下决策树模型在耕地评价中的应用[J]. 农业工程学报，2007，23(12)：58－62.
Tian Jian, Hu Yueming, Wang Changwei, et al. Application of evaluation in farmland with decision tree model based on clustering[J]. Transactions of the Chinese Society of Agricultural Engineering，2007，23(12)：58－62. (in Chinese with English abstract)

[18] 陈文伟，黄金才. 数据挖掘技术[M]. 北京：北京工业大学出版社，2002：8－14.

[19] 李雪平，唐辉明，周顺平. 区域滑坡因子敏感性的 Logistic 回归分析[J]. 地球科学与环境学报，2005，27(4)：14－18.
Li Xueping, Tang Huiming, Zhou Shunping. Logistic regression analysis on sensitivity of regional landslide factors [J]. Journal of Earth Science and Enivronmental，2005，27(4)：14－18．(in Chinese with English abstract)

Farmland classification based on data mining classification method

Wang Lu[1,2], Tian Jian[3], Liu Jianmin[3]

(1. *College of Informatics*, *South China Agricultural University*, *Guangzhou* 510642, *China*;
2. *Guangzhou Institute of Geochemistry*, *Chinese Academy of Sciences*, *Guangzhou* 510640, *China*;
3. *School of Resources and Environment Engineering*, *Hefei University of Technology*, *Hefei* 230009, *China*)

Abstract: Decision tree, BP neural network, and logistic model were used to explored farmland classification of Longchuan Country. The effectiveness of results was analyzed. Confusion matrix was adapted to probe into accuracy of the classification. The results showed that the influences of the number of samples were different to three models. With more training samples, BP neural network and decision tree had heavier influence and higher accuracy in comparison with logistic model. Besides of three models, BP neural network had the highest accuracy and needed a longer time to train model with poor interpretation; decision tree had higher accurate and good interpretation; Logistic model performed worst, Therefore, decision tree might be the best model for farmland classification in Longchuan Country. So data mining classification is an effective and exact method for farmland evaluation, which will enhance the precision and accuracy of land evaluation, and is of significance for the optimization of farmland classification method.

Key words: classification, BP, neural network, Logistic, decision tree, farmland classification