文章编号：1000-8608(2012)01-0074-05

基于决策树方法的水库跨流域引水调度规则研究

习树峰[1]，彭　勇*[1]，梁国华[1]，王本德[1]，谢志高[2]，李学森[1,3]

（1.大连理工大学 水利工程学院，辽宁 大连　116024；
2.深圳市大鹏半岛水源工程管理处，广东 深圳　518008；
3.辽宁省汤河水库管理局，辽宁 辽阳　111000）

摘要： 目前跨流域引水调度属于常规引水规划调度方式，没有考虑实时信息.利用数据挖掘技术中的决策树方法将水库当前的气象预报信息和下垫面蓄水状态、水库多年的实际运行情况等资料与水库管理者的实际调度经验相结合，提出跨流域引水水库的实时调度规则.研究分三步，即首先选取旬初库水位、GFS预报与实际降雨量，旬前土壤含水状态，以及跨流域引水量等资料构成水库调度数据集；然后利用数据挖掘技术从中提取跨流域引水调度决策树；最后对调度决策树进行检验获取跨流域引水水库实时调度规则.实例计算结果表明，采用决策树跨流域引水调度规则进行水库引水调度，可提高水资源利用效率，增加水库综合效益.研究成果对跨流域引水水库实时调度的深入研究与应用有参考价值.

关键词： 跨流域引水；调度规则；数据挖掘；决策树；GFS预报降雨量
中图分类号： TV68；TV697.1　　**文献标志码：** A

0　引　言

随着人口的增长和经济的发展，各方面对水的需求量日益增加，现阶段，跨流域调水已经成为调节区域水资源分布不均匀、实现水资源合理配置的有效手段之一.由于跨流域调水系统存在多流域、多地区、多用途、多目标等特性[1]，多数采用优化方法计算多因素、多目标下的水库调度运行[2~5].目前对于跨流域调水水库调度的研究，大多集中于跨流域调水工程的规划设计阶段：通过定性分析，从理论上提出解决跨流域调水规划调度与实际调度之间矛盾的方法[6,7]；通过定量计算，从规划角度进行需水和供水预测，提出规划调水方案[8]；研究考虑旬径流预报信息的水库引水调度方式，得到跨流域引水水库引水预报调度图[9]等.以上研究为跨流域调水水库调度研究提供了一定的思路，但是其中也存在一些问题，如优化技术所依据的是历史资料，会导致方法研究与实际应用脱节；水文现象的随机性、不确定性和模糊性，使得规划调度方式难以适应调水工程的运行实际；预报信息、决策者的经验尚需结合等.

本文针对跨流域调水实时调度进行研究，将水库当前的气象预报信息和流域下垫面蓄水状态、水库多年的实际运行情况等资料与水库管理者的实际调度经验相结合，提出可以指导水库实时调度的规则.数据挖掘技术可以挖掘出各种水文信息与调度模式之间的内在联系，该方法在水库调度[10]和水文预报领域[11]都有一定的应用，所以本文利用数据挖掘技术分析美国全球预报系统(Global Forecasting System)发布的未来10 d的预报降雨数据(该数据经检验已经达到了可以利用的程度[12])、前期土壤含水量、当前库水位等水库调度数据，得到水库调度决策树，从而获得水库实时调度规则，并与常规调度规则下的水库引水调度结果进行分析比较，分析该方法的优越性和实用性.

收稿日期： 2010-05-01；　**修回日期：** 2011-10-11.
基金项目： 国家自然科学基金资助项目(50979011；51079015).
作者简介： 习树峰(1982-)，女，博士，E-mail：xishufeng820709@126.com；彭　勇*(1979-)，男，博士，E-mail：pyongcuidi@163.com.

1　基于决策树方法的水库跨流域引水调度

1.1　决策树方法

决策树方法能够从一组无规律的事例中利用信息论原理对大量样本的属性进行分析和归纳，推理出以决策树形式表示的分类规则，为决策者提供决策依据. 该方法主要采用自上而下的递归方式建立决策树. 树结构中的节点表示对一个属性的测试，测试算法使用信息增益或信息增益率作为启发信息，选择能够将样本分类的属性；树结构中的每一个分支代表一个测试的输出；树结构中每一个树叶或树叶节点代表一个类别或类别分布.

在决策树方法中最常用的方法有 ID3[13] 算法和 C4.5[14] 算法. ID3 算法用信息增益来选择决策树中的节点属性；C4.5 算法采用基于信息增益率的方法选择测试属性，并在 ID3 算法的基础上加入了对数值属性的处理情况，也对属性值空缺情况进行了处理，用常用值、平均值或者采取概率分配的值来代替未知值. 本文选择 C4.5 算法建立水库跨流域引水预报调度模型.

1.2　跨流域引水决策树预报调度模型的建立

水库跨流域预报调度数据集合包括调度时段、旬初水库状态、GFS 预报降雨量，旬前土壤含水状态，跨流域引水量等多个属性. 每个属性有 p 个数据样本，其中确定跨流域引水量为决策属性，其他属性均为条件属性. 作为决策属性，跨流域引水量具有 m 个不同值，定义 m 个不同类别 $P_i(i=1,2,\cdots,m)$. 设 p_i 是类别 P_i 中的样本数. 对一个跨流域引水量进行分类时，其所需的期望信息为[13]

$$I(p_1,p_2,\cdots,p_m)=-\sum_{i=1}^{m}\frac{p_i}{p}\log_2\left(\frac{p_i}{p}\right) \quad (1)$$

设条件属性 A 具有 k 个不同值 $\{a_1,a_2,\cdots,a_k\}$. 属性 A 可将给定集合 S 划分为 k 个子集 $\{C_1,C_2,\cdots,C_k\}$，其中 C_i 包含 C 中的样本在 A 上具有值 a_j. 若将 A 作为测试属性，由包含集合 P 的节点生长出来的分支与这些子集相对应. 设 p_{ij} 是类别 P_i 在子集 C_i 中的样本数，则由 A 划分成子集的熵为[13]

$$E(A)=-\sum_{j=1}^{k}\left[\frac{p_{1j}+p_{2j}+\cdots+p_{mj}}{p}\times I(p_{1j},p_{2j},\cdots,p_{mj})\right] \quad (2)$$

其中 $(p_{1j}+p_{2j}+\cdots+p_{mj})/p$ 为第 j 个子集的权. 计算得的熵值越小，子集划分的纯度就越高.

根据上文给出的期望信息的计算公式，对于子集 C_j，其期望信息计算公式为[13]

$$I(p_{1j},p_{2j},\cdots,p_{mj})=-\sum_{i=1}^{m}\frac{p_{ij}}{|p_j|}\log_2\left(\frac{p_{ij}}{|p_j|}\right) \quad (3)$$

由期望信息和熵值可以得到对应的信息增益值. 对于在条件属性 A 上分支获得的信息增益可以由以下公式得到[13]：

$$Gain(A)=I(p_1,p_2,\cdots,p_m)-E(A)$$

则信息增益率的计算公式为[14]

$$GainRatio(A)=Gain(A)/SplitI(A) \quad (4)$$

其中 $SplitI(A)=-\sum_{j=1}^{k}\frac{p_{ij}}{|p_j|}\log_2\left(\frac{p_{ij}}{|p_j|}\right)$.

通过式(4)计算每个属性的信息增益率，选择具有最高信息增益率的属性作为分裂属性，创建一个节点，并以该属性为标记创建分支. 同时，再据此分支中的数据集计算信息增益率继续划分分支样本，直至给定节点的所有样本属于同一类或者没有剩余属性可以划分样本为止，最终形成按照水库跨流域引水各种信息进行分类的引水决策调度树. 对于形成的决策树，按照“if-and-then”的形式从树根到树叶获取调度规则. 例如数据集有属性 X、Y、Z，可以得到的调度规则为 if $X=$“x_1” and $Y=$“y_2” then $Z=$“z_2”.

2　实例研究

大伙房水库是一座防洪、供水、灌溉、发电、养鱼等综合利用的大型水利枢纽工程，水库从邻近流域引水后，经过调节，向下游地区严重缺水的大、中城市供水，解决该地区工业与城市生活用水短缺问题. 本文以大伙房水库为实例，确立跨流域引水决策树预报调度规则.

2.1　引水决策的调度规则确立

在本文的研究中，按照目前各用水部门的需水定额，以满足各部门的用水保证率为约束条件，以跨流域引水量最小为目标函数，选择遗传算法对大伙房水库 1959～2007 年数据进行水库优化调度，得到跨流域引水水库兴利调度数据集，数据集

中包括调度时段,旬初水库水位状态、用水量,旬末水库水位状态,GFS 预报和实际降雨量,旬前土壤蓄水状态,弃水量和跨流域引水量等属性.选择数据集中的旬初水库水位状态、降雨量,旬前土壤蓄水状态,跨流域引水量这 4 个属性记录进行数据挖掘,确定跨流域引水量为决策属性,条件属性包括旬初水库水位状态、降雨量和旬前土壤蓄水状态.

旬初库水位,是影响调度决策的最主要因素,它是前一段时期水库蓄泄水量的累积效应,也是当前时段水库状态的直接反映;降雨量,是对下一时段径流过程的提前预测,反映未来水库状态的变化趋势(注:鉴于 GFS 预报降雨信息较短,数据挖掘模型和模拟检验中应采用实际旬降雨资料);旬前土壤蓄水状态,反映本时段初期的土壤含水量,主要影响径流的形成过程.旬初水库水位状态、降雨量,旬前土壤蓄水状态,跨流域引水量这 4 个属性均为连续属性.由于在实际应用过程中,准确的旬前土壤蓄水状态信息在实际中较难获得,本文决策树径流分级模型中采用上一旬后 n 天的径流量之和来表示旬初的土壤蓄水状态,试算和检验后确定 n 取值为 5.

选择大伙房水库 1959～2000 年的水库调度数据来建立数据挖掘模型,可得到跨流域引水决策树,结果如图 1 所示.使用 2001～2007 年实际降雨数据进行模拟校验,其调节结果与原调节结果相比合格率为 90.8%.

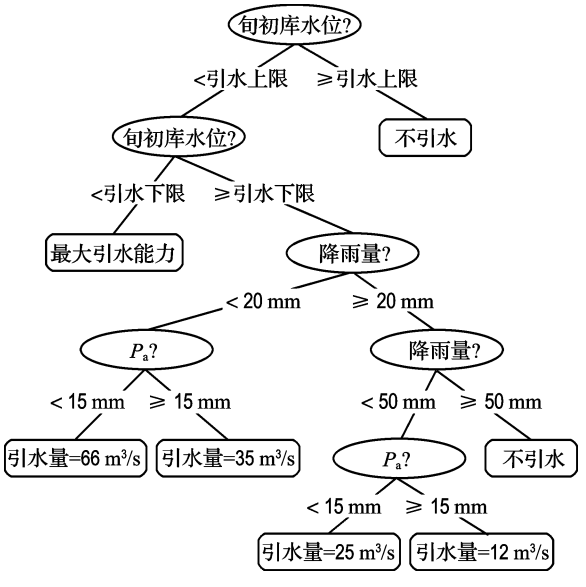

图 1 大伙房水库跨流域引水决策树

Fig. 1 Decision tree of water diversion in Dahuofang Reservoir

由跨流域引水分级决策树可得到跨流域引水分级等级:

(1)根据大伙房水库调度的特点,将旬前库水位分为三级,确定跨流域引水上限水位和下限水位,上下限水位值如表 1 所示.

表 1 大伙房水库跨流域引水上限水位值和下限水位值

Tab. 1 The inter-basin water diversion level constraint values of Dahuofang Reservoir

时间	引水上限/m	引水下限/m
7 月中旬	126.4	125.7
7 月下旬	126.4	125.7
8 月上旬	126.4	125.7
8 月中旬	126.4	125.7
8 月下旬至翌年 7 月上旬	126.6	125.6

(2)可以得到降雨量和 P_a 的分级标准,如表 2 所示.

表 2 大伙房水库降雨量和 P_a 分级标准

Tab. 2 Grading standards of rainfall and P_a in Dahuofang Reservoir

降雨量分级/mm			P_a 分级/10^6 m³	
Ⅰ	Ⅱ	Ⅲ	1 级	2 级
[0,20)	[20,50)	≥50	[0,15)	≥15

(3)在不同的旬前水库水位状态、降雨量和旬前土壤蓄水状态条件下,将跨流域引水流量离散为 0、12、25、35、66 和 70 m³/s,这 6 个数值分别代表 1～6 级.

根据分级结果可提取出跨流域引水调度规则.具体的描述跨流域引水决策调度规则如下:

①当水库当前水位高于引水限制水位上限时,水库跨流域引水量为 1 级;当水库当前水位低于引水限制水位下限时,引水量为 6 级.

②若当前水库水位处于跨流域引水限制水位上、下限之间,降雨量为Ⅰ级,旬前土壤含水量为 1 级时,引水量为 5 级;降雨量为Ⅰ级,旬前土壤含水量为 2 级时,水库引水量为 4 级.

③当前库水位处于引水限制水位上、下限之

间，降雨量为Ⅱ级，旬前土壤含水量为 1 级时，引水量为 3 级；降雨量为Ⅱ级，旬前土壤含水量为 2 级时，水库引水量为 2 级.

④当前库水位处于引水限制水位上、下限之间，降雨量为Ⅲ级时，引水量为 1 级.

2.2　结果对比

文献[12]分析了 GFS 旬降雨预报数据的误差分布，认为 GFS 分级降雨预报数据可以利用到水库实际调度当中，所以本文用 GFS 分级预报降雨量代替实际分级降雨量，采用决策树法跨流域引水调度规则进行水库跨流域引水调度.

依据决策树法得到的跨流域引水预报调度规则和水库常规引水调度规则，分别采用大伙房水库 2001～2007 年资料进行长系列引水调度，调度比较结果如表 3 所示. 由于采用新调度规则进行水库供水调度的风险分析关键是要论证其是否降低了水库的供水保证率，表中列出供水保证率项.

从表 3 调度结果可以看出，水库供水保证率仍保持原设计值 85.7%，即在供水风险不增加的前提下多年平均引水量可减少 0.9×10^8 m^3，多年平均弃水量减少 0.88×10^8 m^3，多年平均缺水量略有增加，调节水库末水位抬高了 0.24 m. 以上数据表明，利用 GFS 降雨预报信息的决策树调度规则进行水库调度的经济效益明显优于常规调度，既不增加供水风险，又减少了引水量和弃水量，提高了水资源利用率.

表 3　调度结果对比

Tab. 3　Operation result comparison

调度规则类别	水库供水保证率/%	多年平均引水量/10^6 m^3	多年平均弃水量/10^6 m^3	多年平均缺水量/10^6 m^3	水库末水位/m
决策树预报调度规则	85.7	1 714	292	14	120.92
常规调度规则	85.7	1 804	380	10	120.68

3　结　语

本文研究的基于决策树方法的跨流域引水预报调度规则，集中反映了历史水文数据和调度数据的规律性；增加了 GFS 预报降雨量和土壤含水量两个决策指标，考虑了未来水库的入流情况，使得调度决策更加符合实际；弥补了常规调度图中仅以引水限制线确定是否引水的局限性.

在提取水库引水调度规则的过程中，决策树从根节点到叶子节点的路径非常直观，全面的和复杂的调度决策规则可以通过一系列简单和局部的决策近似取得.

采用决策树跨流域引水调度规则，可在不增加供水风险的前提下，提高水资源利用效率，增加水库综合效益. 研究的方法结论有理论意义和实用价值.

另外，在应用决策树方法中，由于方法的局限性可能会出现“过拟合”现象，而且 GFS 预报数值和前期土壤含水量值本身存在一定的误差，所以今后应从模型和数值不确定性两个方面分析其风险.

参考文献：

[1] 邵东国. 跨流域调水工程规划调度决策理论与应用[M]. 武汉：武汉大学出版社，2001

[2] 刘建林，马　斌，解建仓，等. 跨流域多水源多目标多工程联合调水仿真模型——南水北调东线工程[J]. 水土保持学报，2003，**17**(1)：75-79

[3] 卢华友，沈佩君，邵东国，等. 跨流域调水工程实时优化调度模型研究[J]. 武汉水利电力大学学报，1997(5)：11-15

[4] 张建云，陈洁云. 南水北调东线工程优化调度研究[J]. 水科学进展，1995，**6**(3)：198-204

[5] 雷声隆，覃强荣，郭元裕，等. 自优化模拟及其在南水北调东线工程中的应用[J]. 水利学报，1989，**18**(5)：1213

[6] 李　光. 跨流域调水工程供水调度运行初探[J]. 水科学与工程技术，2008(14)：18-20

[7] 董延军，蒋云钟，韩亦方，等. 南水北调中线供水调度特性浅析[J]. 中国水利，2007(4)：22-24

[8] 叶新霞. 区域水资源合理配置及跨流域调水水资源系统问题研究[D]. 南京：河海大学，2005

[9] 梁国华，王国利，王本德，等. 大伙房跨流域引水工程

预报调度方式研究[J]. 水力发电学报, 2009, **28**(3):32-36

[10] 许惠君,李彩林,刘晓安. 数据挖掘技术在水库调度中的研究与应用[J]. 计算机与数字工程, 2006, **34**(9):61-63

[11] 张 弛,王本德,李 伟. 数据挖掘技术在水文预报中的应用及水文预报发展趋势研究[J]. 水文, 2007, **27**(2):74-77

[12] 梁国华,习树峰,王国利,等. GFS 预报在大伙房水库分级利用方式中的应用[J]. 水电能源科学, 2009, **27**(2):4-6

[13] KAMBER M, WINSTONE L, GON Wang, *et al*. Generalization and decision tree induction: efficient classification in data mining [C] // **Proceedings of RIDE'97**. Birmingham:[s n], 1997:111-120

[14] QUINLAN J R. **C4. 5: Programs for Machine Learning** [M]. San Mateo: Morgan Kaufmann Publishers, 1993

Research on reservoir operation rules of inter-basin water transfer based on decision tree method

XI Shu-feng[1], PENG Yong[*1], LIANG Guo-hua[1], WANG Ben-de[1], XIE Zhi-gao[2], LI Xue-sen[1,3]

(1. School of Hydraulic Engineering, Dalian University of Technology, Dalian 116024, China;
2. Dapeng Half Island Administrative Office of Water Source Engineering, Shenzhen 518008, China;
3. Tanghe Reservoir Management Bureau of Liaoning Province, Liaoyang 111000, China)

Abstract: The inter-basin water transfer operation belongs to the conventional water transfer planning operation mode, and the real-time information is not considered in the operation. To solve this problem, the decision tree method in data mining is used combining the current reservoir forecast information, underlaying surface water storage condition, perennial reservoir running situation and other data with the reservoir managers' actual operation experiences, and then, the inter-basin water transfer real-time operation rules can be realized. Research process has three steps. Firstly, initial reservoir water level, actual rainfall, GFS forecasting rainfall, soil moisture, diversion water quantity, etc. are selected to compose the reservoir operation data set. Secondly, the inter-basin water transfer operation decision tree is extracted by using data mining technology. Finally, the operation decision tree is tested and the inter-basin water transfer real-time operation rules are obtained. The actual example results show that using decision tree inter-basin water transfer operation rules in the reservoir operating can increase the water resource efficiency and the reservoir comprehensive benefits. This research result has some reference value for the further study and application of the inter-basin water transfer real-time operation.

Key words: inter-basin water transfer; scheduling rules; data mining; decision tree; GFS forecasting rainfall