Investigation on GIS Attribute Data Mining with Statistical Inductive Learning

Anmin Lu [1) 2)] Zongjian Lin [2)] Chengming Li [2)]

1) (School of Information Engineering, Wuhan Technical University of Surveying and Mapping, Wuhan 430079), E-mail: anminlu@263.net

2) (Chinese Academy of Surveying and Mapping, Beijing, 100039)

Abstract With the development of modern science and technology, huge amounts of data have been stored in spatial databases. This huge amount of data challenges traditional data analysis methods. Spatial data mining, i.e., discovery of interesting, implicit knowledge in spatial databases has attracted attentions in recent years. Spatial data mining is divided into there fields: graphics data mining, attribute data mining and their relations mining.

In this paper, a statistical inductive learning (SIL) approach is proposed to investigate GIS attribute data mining. This approach integrates statistical analysis with attribute oriented induction method. GIS attribute data mining is divided into three hierarchies, as follows:

1 From raw data to new data: With the help of GIS and statistical tools, we can calculate the minimum, maximum, sum, average, standard deviation of one column data in a table. We are also able to create thematic maps, such as bar charts, pie charts, dot charts, etc. All of these can be believed as new data from raw data.

2 From data to model : Building models from data is the creation of a mathematical model, that is, collecting data through investigation, studying the principle of the data, comprehension the main illogicality of the question, pointing out hypothesis, after abstracting and predigesting, building the mathematics relations of the problem. Then we use the methods and techniques of mathematics to solve practical problems. There are many methods of mathematics models, including network models, optimization models and random models, etc. The multiple linear regression model belonged to random models is one of the most useful models. The paper first use correlation analysis to study relations of two variables, then use the multiple linear regression model to build model from data.

3 From data to knowledge: Knowledge is cognition of the real world. Induction learning can obtain new concept and new rules. There are many methods about induction learning. Attribute-oriented induction (AOI) is one of the most useful methods to discover knowledge in databases. Data in databases often contains detailed information at primitive concept levels. AOI obtains general data from concept hierarchy generation, then changes the data into rules.

From raw data to new data can help us to know rough relations between two variables. From data to model can describe relations of dependent variables and independent variables with ration. From data to knowledge can obtain general rules in high levels.

Finally, an experiment on agricultural statistical data of China mainland shows that the statistical inductive learning approach is feasible and effective for GIS attribute data mining.

Keywords data mining and knowledge discovery, Attribute oriented induction, Multiple linear regression, Statistical analysis

1 Introduction

With wide applications of satellite and remote sensing technologies and automatic data collection tools, large amounts of spatial data and attribute data have been collected and stored in spatial databases. The extraction and comprehension of the knowledge implied by the huge amount of spatial data and attribute data, poses great challenges to currently available GIS technologies. Data mining, or knowledge discovery in databases, has been emerging as a new research field and a new technology for discovery of interesting, implicit, and previously unknown knowledge from large databases[1]. Data mining represents the confluence of several research fields, including artificial intelligence, database systems, statistics and data visualization. Spatial data mining, a branch of data mining, is divided into there fields: graphics data mining, attribute data mining and their relations mining. This paper deals with the attribute data mining. The attribute data mining includes three levels: from data to new data, from data to model and from data to knowledge. In this paper, the experiment data is from 1999 China statistical yearbook (table 4).

2 From raw data to new data

With the help of GIS and statistical tools, we can calculate the minimum, maximum, sum, average, standard deviation of one column data in a table. For example, based on table 4, we can calculate gross agricultural output of China in 1998 is 24082.29 (100 million yuan), total rural labor force number is 31682.8 (10000 persons), total cultivated area is 94997.6 (1000 hectares), average gross agricultural output per 10000 persons is 0.7601 (100 million yuan), average gross agricultural output per 1000 hectares is 0.2535 (100 million yuan), etc.

Otherwise, raw data can turn into many kinds of subject maps, such as bar charts, pie charts, dot charts, which is also believed as new data. Figure 1 depicts bar charts of rural labor force, cultivated area and gross agricultural output.

We can see the rough relations of rural labor force, cultivated area and gross agricultural output from figure 1. Such as cultivated area of Heilongjiang is the biggest one, the rural labor force of Henan is the most one, the gross agricultural output of Shandong is the highest one and the more the rural labor force, the more the gross agricultural output, etc.

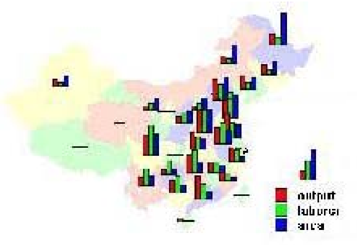
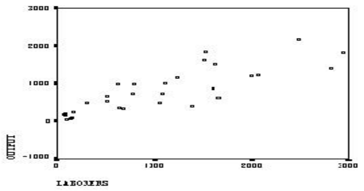


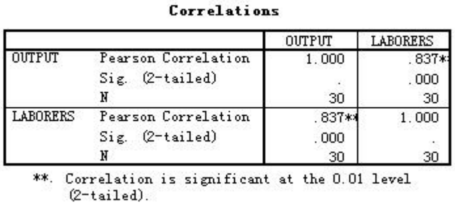
Figure 1 bar charts of rural labor force, cultivated area and gross agricultural output

3 From data to model

Building models from data is mathematics modeling, that is, collecting data through investigation, studying the principle of the data, comprehension the main illogicality of the question, pointing out hypothesis, after abstracting and predigesting, building the mathematics relations of the question. Then we use the methods and techniques of mathematics to solve practical problems. There are many methods of mathematics models, including network models, optimization models and random models, etc. The multiple linear regression model belonged to random model is one of the most useful models. The paper uses multiple linear regression models to build model from data. First, we use correlation analysis to analyze linear relation degree of two variables.

Figure 2 is the dot map of relation about rural labor force and gross agricultural output. We can see that there is a rule between rural labor force and gross agricultural output, that is, the more the rural labor force, the more the gross agricultural output. Table 1 represents the relation coefficient of rural labor force and gross agricultural output is 0.873. It shows that they have a high relativity.





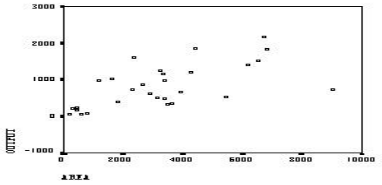
Correlations

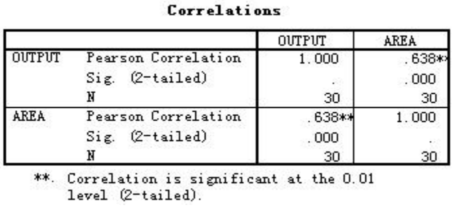
		OUTPUT	LABORERS
OUTPUT	Pearson Correlation	1.000	.837**
	Sig. (2-tailed)	.	.000
	N	30	30
LABORERS	Pearson Correlation	.837**	1.000
	Sig. (2-tailed)	.000	.
	N	30	30

**. Correlation is significant at the 0.01 level (2-tailed).

figure2 dot map of rural labor force and gross agricultural output table1 relation of rural labor force and gross agricultural output

Now, let us look at the relation of cultivated area and gross agricultural output, figure 3. There exists rough linear relation between cultivated area and gross agricultural output, that is, the more cultivated area, the more gross agricultural output. Table 2 represents the relation coefficient of cultivated area and gross agricultural output is 0.638, which shows the relativity of cultivated area and gross agricultural output is lower than those of rural labor force and gross agricultural output.





Correlations

		OUTPUT	AREA
OUTPUT	Pearson Correlation	1.000	.638**
	Sig. (2-tailed)	.	.000
	N	30	30
AREA	Pearson Correlation	.638**	1.000
	Sig. (2-tailed)	.000	.
	N	30	30

**. Correlation is significant at the 0.01 level (2-tailed).

figure3 dot map of cultivated area and gross agricultural output table 2 relation of cultivated area and gross agricultural output

Some elementary principles have been found through analyzing the relations among rural labor force, cultivated area and gross agricultural output with the method of correlation analysis. The relativity of rural labor force and gross agricultural output is higher than those of cultivated area and gross agricultural output. It represents gross agricultural output is main determined by rural labor force.

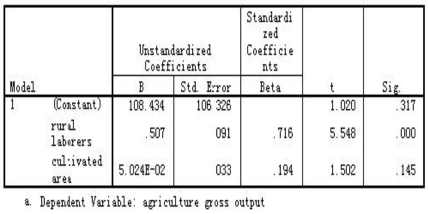
Now let us use multiple linear regression to analyze the relations of dependent variable and independent variable. The model helps us describe the relation of variables with ration.

Let X_1 refers to number rural labor force, X_2 refers to cultivated area, Y refers to gross agricultural output (table 4). From table 3, the regression equation is :

$Y=108.434+0.507X_1+0.050X_2$

Their standard regression coefficient of X_1 and X_2 are 0.716 and 0.194, respectively.

table3 regression equation coefficient and constant



Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	108.434	106 326		1.020	.317
	rural laborers	.507	091	.716	5.548	.000
	cul:ivated area	5.024E-02	033	.194	1.502	.145

a. Dependent Variable: agriculture gross output

The influence of rural labor force to gross agricultural output is bigger because the coefficient of X_1 is bigger. The influence of cultivated area to gross agricultural output is smaller because the coefficient of X_2 is smaller. The influence of rural labor force to gross agricultural output is much bigger than those of cultivated area to gross agricultural output because 0.716 is much bigger than 0.194. So, the gross agricultural output is main determined by the number of rural labor force. This result is as same as the result of relation coefficient analysis. Otherwise, we can build predictive model with regression model. For example, we can estimate agricultural loss in any special region if flood covers the region.

We have got same results through analyzing the relation of rural labor force, cultivated area and gross agricultural output with from raw data to new data and from data to model.

4 From data to knowledge

Knowledge is cognition of the real world. Induction learning can obtain new concepts and new rules. There are many methods about induction learning. Attribute-oriented induction (AOI) is one of the most useful methods to discover knowledge in databases[4]. Data in databases often contains detailed information at primitive concept levels. It is often desirable to summarize a large set of data and present it at a high concept level. An attribute-oriented concept tree ascension technique is applied in generalization, which substantially reduces the computational complexity of database learning processes. AOI obtains general data from concept hierarchy generation, then changes the data into rules. This method is useful in many fields, such as data classification. AOI demands background knowledge, which can obtain by data analysis automatically or given by experts in their fields. We direct give background knowledge.

rural labor force: 0----599⊏few, 600----1499⊏middle, 1500----3000⊏many,

{few, middle, many}⊏ANY（rural labor force）

cultivated area: 0----1999⊏small, 2000----3999⊏middle, 4000----10000⊏big,

{small, middle, big}⊏ANY（cultivated area）

gross agricultural output: 0----699⊏low, 700----1299⊏middle, 1300----2000⊏high,

{low, middle, high}⊏ANY（gross agricultural output）

{beijing, tianjin, hebei, sanxi, neimenggu}⊏north,

{liaoning, jilin, helongjiang}⊏northeast,

{shanghai, jiangsu, zhejiang, anhui, fujian, jiangxi, shandong}⊏east,

{henan, hubei, hunan, guangdong, guangxi, hainan}⊏south,

{sichuan, guizhou, yunnan, xizang}⊏southwest,

{shanxi, gansu, qinghai, ningxia, xinjiang}⊏northwest,

{ north, northeast, east, south, southwest, northwest}⊏ANY（region）

table4 information of gross agricultural output

province city	rural laborers（10000 persons）	cultivated area（1000 hectares）	gross agricultural output （ 100 million yuan）	province city	rural laborers（10000 persons）	cultivated area（1000 hectares）	gross agricultural output （ 100 million yuan）
Beijing	67.7	399.5	176.58	Henan	2940.3	6805.8	1822.99
Tianjin	79.4	426.1	156.17	Hubei	1232.9	3358.0	1147.51
Hebei	1635.5	6517.3	1505.94	Hunan	2062.9	3249.7	1232.75
Sanxi	639.9	3645.1	359.15	Guang dong	1508.2	2317.3	1614.64
Neimenggu	512.4	5491.4	534.39	Guangxi	1604.1	2614.2	865.91
Liaoning	633.0	3389.7	969.79	Hainan	170.2	429.2	242.54
Jilin	517.0	3953.2	666.47	Sichuan	2811.9	6189.6	1394.14
Helong jiang	760.3	8995.3	736.34	Guizhou	1388.4	1840.0	402.29
Shanghai	76.3	290.0	206.78	Yunnan	1661.8	2870.6	614.50
Jiangsu	1531.5	4448.3	1849.19	Xizang	89.3	222.1	42.34
Zhejiang	1102.7	1617.8	1003.71	Shanxi	1047.4	3393.4	479.36
Anhui	1992.9	4291.1	1202.27	Gansu	683.8	3482.5	335.79
Fujian	776.8	1204.0	973.39	Qinghai	138.2	589.9	60.78
Jiangxi	1073.7	2308.4	734.87	Ningxia	146.6	807.2	78.76
Shandong	2487.0	6696.0	2174.54	Xinjiang	310.7	3128.3	498.41

From table 4, we can not get any obvious rules. So we generalize table 4 according to the background. For example, beijing is replaced with north, rural laborers of beijing is replaced with few, etc (table 5). A new field, which used to count the number of generalization, is added in table 5.

table5　information of gross agricultural output after generalization

No.	Region	laborers	area	output	Count	No.	Region	laborers	area	output	count
1	North	few	small	low	1	16	South	many	big	high	1
2	North	few	small	low	1	17	South	middle	middle	middle	1
3	North	many	big	high	1	18	South	many	middle	middle	1
4	North	middle	middle	low	1	19	South	many	middle	high	1
5	North	few	big	low	1	20	South	many	middle	middle	1
6	Northeast	middle	middle	middle	1	21	South	few	small	low	1
7	Northeast	few	middle	low	1	22	Southeast	many	big	high	1
8	Northeast	middle	big	middle	1	23	Southeast	middle	small	low	1
9	East	few	small	low	1	24	Southeast	many	middle	middle	1
10	East	many	big	high	1	25	Southeast	few	small	low	1
11	East	middle	small	middle	1	26	Northwest	middle	middle	low	1
12	East	many	big	middle	1	27	Northwest	middle	middle	low	1
13	East	middle	small	middle	1	28	Northwest	few	small	low	1
14	East	middle	middle	middle	1	29	Northwest	few	small	low	1
15	East	many	big	high	1	30	Northwest	few	middle	low	1

From table 5, we can see that region and rural laborers are much important to gross agricultural output, whereas cultivated area has a little influence to gross agricultural output. This result is as same as the result of part 3. Unite the same row in table 5. The result is in table 6.

table6　information of gross agricultural output after uniting same row

No.	Region	laborers	area	output	Count	No.	Region	laborers	area	output	count
1	North	few	small	low	2	13	South	many	big	high	1
2	North	many	big	high	1	14	South	middle	middle	middle	1
3	North	middle	middle	low	1	15	South	many	middle	middle	2
4	North	few	big	low	1	16	South	many	middle	high	1
5	Northeast	middle	middle	middle	1	17	South	few	small	low	1
6	Northeast	few	middle	low	1	18	Southwest	many	big	high	1
7	Northeast	middle	big	middle	1	19	Southwest	middle	small	low	1
8	East	few	small	low	1	20	Southwest	many	middle	middle	1
9	East	many	big	high	2	21	Southwest	few	small	low	1
10	East	middle	small	middle	2	22	Northwest	middle	middle	low	2
11	East	many	big	middle	1	23	Northwest	few	small	low	2
12	east	middle	middle	middle	1	24	Northwest	few	middle	low	1

Now remove redundant attribute value in table 6. An attribute is redundant when the decision-making result does not change if the attribute value is removed. For example, from row 22 to row 24, the gross agricultural output is always “low” when rural laborers and cultivated area are all removed. So we replace rural laborers and cultivated area with “----”. Finally unite the same row again (table 7).

Every row in table 7 is a rule, where the count number is its support degree. No.3 and No.4, No.5 and No.7, No.10 and No.11 are disaccord rules, whereas the others are accordant rules. Rule 1 represents “In every province or city in China mainland, if rural laborers is few, then gross agricultural output is low ”. The rule can be represented as:

rural laborers∈few→gross agricultural output ∈ low (support is 7).

Rule 6 represents “In east of China, if rural laborers is middle, then gross agricultural output is middle”. This rule can be represented as:

（province and city∈east）∧（rural laborers∈middle）→gross agricultural output∈middle（support is 3）.

table7　information of gross agricultural output after removing redundant attribute value

No.	region	laborers	Area	output	count	No.	Region	laborers	area	output	count
1	---	few	---	low	7	8	South	many	big	high	1
2	north	many	---	high	1	9	South	middle	---	middle	1
3	north	middle	---	low	1	10	South	many	middle	middle	2
4	north	middle	---	middle	2	11	South	many	middle	high	1
5	east	many	Big	high	2	12	Southwest	many	big	high	1
6	east	middle	---	middle	3	13	Southwest	middle	---	low	1
7	east	many	Big	middle	1	14	Southwest	many	middle	middle	1
						15	Northwest	---	---	low	5

Rule 8 represents “In south of China, if rural laborers is many, cultivated area is big, then gross agricultural output is high”. Rule 15 represents “In the northwest of China, gross agricultural output is low", etc.

From raw data to new data obtains one result "the more the rural laborers, the more the gross agricultural output". From data to model obtains one result "gross agricultural output is main determined by rural laborers". From data to knowledge obtains one result "In south of China, if rural laborers is many, cultivated area is big, then gross agricultural output is high". All these results are main the same, whereas there is a little different with them.

5 conclusion

In this paper, a statistical inductive learning (SIL) approach is proposed to investigate GIS attribute data mining. This approach integrates statistical analysis with attribute oriented induction method. From raw data to new data can help us to know rough relations of two variables. From data to model can describe relation of dependent variables and independent variables with ration. From data to knowledge can obtain general rules in high levels. An example on agricultural statistical data of China mainland shows that the statistical inductive learning approach is effective for GIS attribute data mining.

References

[1] Krzysztof Koperski, Jiawei Han, Junas Adhikary. Mining knowledge in geographical data. In Comm. Acm (to appear), 1997

[2] Liu Hong, Lu Chunheng, Zhai Ligong. China statistical yearbook. Beijing: China statistics press, 1999

[3] Zhang Raoting, Fang Kaitai. Multiple statistics analysis. Beijing: science press, 1982

[4] Han J, Cai Y, Cercone N, Knowledge Discovery in database: An attribute oriented approach. In: Proceedings of the 18th VLDB Conference: Vancouver, British Columbia, Canada, 1992, 547-559