DEVELOPMENT OF A DATA MINING METHOD FOR LAND CONTROL

WANG Shuliang
WANG Xinzhou
SHI Wenzhong

KEY WORDS land control; data mining; vector match method; rough set; GIS

ABSTRACT Land resources are facing crises of being misused especially for an intersection area between town and country, and land control has to be enforced. This paper presents a development of data mining method for land control. A vector match method for the prerequisite of data mining i.e., data cleaning is proposed which deals with both character and numeric data via vectorizing character string and matching number. A minimal decision algorithm of rough set is used to discover the knowledge hidden in the data warehouse. In order to monitor land use dynamically and accurately, it is suggested to set up a real-time land control system based on GPS, digital photogrammetry and online data mining. Finally, the means is applied in the intersection area between town and country of Wuhan city, and a set of knowledge about land control is discovered.

1 Introduction

Land is a fundamental resource for human being, and it is impossible for land to be reproduced and reverse-used (Sen, Wang, 1997). Currently, the balance of land is being destroyed because of land desertification, land erosion, saline-alkali soil rising, etc. (Jin, 1997). And land is also under the threat from urban development, building road and engineering construction. In an intersection area between town and country, land resources are facing much worse crises of being misused. Besides, many artificial phenomena are illegally happening to land.

An intersection area between town and country is a bridge between urban and country, and a front position of country urbanizing and modern agriculture. It owns both better infrastructures as a city and fresher natural environments like a country. Both society and economy develop actively. However, more and more land problems arise such as occupying land in disorder, misusing land, exchanging land lawlessly, wasting land etc. Firstly, city grows bigger and bigger while the intersection area between town and country becomes smaller and smaller, which leads to violent conflicts between construction and agriculture. Moreover, city expansion often occupies high quality plantation that is both peasants’ lives and citizens’ bases, whereas the efficiencies of land use are very low. The building density in an intersection area between town and country is far lower than that in the central district of a city. Secondly, there are two different land administrative systems for such an area (one for city, the other for country). And it is so weak in land planning and administrating that it causes the imbalance of land structure. Thirdly, exploiting and occupying land is so excessive that it destroys the original balance of the nature and lessens the function of an
intersection area between town and country to absorb city wastes. Hence, people have to control land use besides avoiding natural land crises in an intersection area between town and country. Wuhan is a good example on this.

In order to ensure the dynamic balance of plantation gross, a land management law has been revised to stress on land control plantation-protection. Land control means that the state makes and releases land planning ahead, carves up land subarea and works out land utilization limitation (permission, confined permission and impermissibility), so as to assure reasonable land utilization and keep economy, society and environment developing harmoniously (Land Management Laws of P. R. C., 1998). Therefore, land control is a synthetic decisive engineering that includes economic, lawful, technical and administrative measures. It is obvious that all measures are based on land information and its processing technology. More attention should be paid to technical measures, since they provide the essential information for monitoring land utilization and alteration carrying out control rules and deepening the system reform. Take the following example. Land planning ahead can give the methods for controlling land use, and lead to reasonable land use. Dynamic land monitoring may acquire the newest land information instead of manpower, and avoid some land problems on time. Visual reality on land crisis will make man cherish land consciously.

As we know, land control is a synthetic decision. It depends on all kinds of information from different sources. It is unable to avoid data errors or incomplete records when the information is collected. The same attribute of a land parcel may have more than one value, and either character or number represents the value. Before the information is employed for land control, data cleaning has to be done. Here, vector-match method is proposed to clean both character and numeric data via vectorizing character-string and matching number. After the land data being dealt with, a series of databases, such as position, history, laws, economy, population etc. can be set up, which is one of the important GIS contents. As the spatial information in land management is horizontal, the products are easier to be stored and utilized. After they are put into the computer, those data can be shared on the network in the scope of laws. With land databases cumulating a lot of knowledge perhaps is hidden in the databases. Unknown as the knowledge is, it is suitable to monitor, control and perfect land use. It is suggested to rebuild up a land data warehouse for data mining on the basis of these databases.

It is well known that land data almost change everyday, especially for an intersection area between town and country. At the same time, land control needs timely information. Unfortunately, it is difficult for only manpower to control land use duly, for the changes of land information are mostly intricate and non-linear and with acute conflicts of layer upon layer. Sometimes there are man-made errors in land statistic for some reasons. Moreover, not all ground control points could see each other because of artificial and natural objects on the ground, especially in the area with intensive high buildings. Fortunately, GPS-supported automatic aerial triangulation (Li et al., 1997) has been developed to extract land data of space, time and attribute without touching objects. It acquires aerial control points with GPS carried in a plane instead of ground control points. With the help of digital photogrammetry softwares including both analytical photogrammetry and digital image manipulation functions (e.g., DD-KIN, VirtuoZo and WuCAPSgps), it takes less time manpower and expense to get high grade of points. Moreover, it is easy to refresh land information. This makes it more reliable to extract land geometry and attribute information from the images.

In the network era, if we make full use of three-dimensional information and knowledge from online analytical mining, spatial land forecasting will truthfully and deeply show people land crises with the technologies of virtual realism, three-dimensional landscape etc. That will oblige them to work out land planning ahead, use land economically and protect plantation in advance.

In this paper, we focus on developing a data mining method for land control. In section 2, we propose vector-match method on data cleaning to integrate both character and numeric information from different sources. In section 3, a minimal decision
algorithm of rough set is used to discover the knowledge hidden in the data warehouse. In section 4, we apply the means in the intersection area between town and country of Wuhan city, and a set of reasonable knowledge about land control is found out. In section 5, a framework of real-time land control system is suggested. It is based on GPS, digital photogrammetry and online data mining so as to control land use dynamically.

2 Vector-Match method

Land information used for land control is from more than one source. They provide spatial, lawful, social and economical aspects of information etc. Obviously, it is the integrated information of heterogeneous data that land control utilizes. However, there are several problems hidden in original data, such as incomplete data, inaccurate data, ambiguous data, conflicting data and repetitive data.

If these kinds of data are allowed to enter data warehouse, the result of data mining will be incorrect with a very high chance. Therefore, cleaning, resetting and mining those data becomes necessary before they are put into GIS. That is to say, being a prerequisite of data mining, data cleaning should be carried out at first place. It is statistical that the data integration improved by 4% in a large corporation maybe leads to over one million dollars profits (Innovative Systems, Inc., 1997). The matching and amalgamation of heterogenous data is an important content of data cleaning. And it may be abstracted as the clustering analysis. Vector-match method is a new method for data cleaning, which is put forward to integrate multi-original data. This method gives a concept of similar degree to measure how two data from different sources on the same attribute are similar to each other. Each original value has a similar degree. Because character information is more difficult to deal with, we firstly do word analysis to find out the basic characters, of which a measure-space dimension is composed. All different original character data on the same object are vectorized according to the measure-space dimension. Each vector of original data has the same length that equals to the measure-space dimension. Then the derived inclination-cosine method is used to calculate the similar degree. As for numeric information, an index function is defined to match land parcel. Due to the confidence levels of different original data are different from each other, the creditable weight of data source is also given. Then a credibility degree is defined to decide the optimal which is computed according to its confidence levels and times. The confidence level focuses on its weight as regards times. Attention should be paid to its historical accumulation and inherit. The source information with the biggest credibility degree is selected in the end.

2.1 Description of the method

Suppose that attribute A of the same land parcel L has n (n ≥ 1) values \( \{ V_1, V_2, \ldots, V_n \} \), which come from different data sources \( \{ O_1, O_2, \ldots, O_m \} \). The credibility degree of \( V_i \) is defined as \( R(L, A, V_i) \) in Eq. (1).

\[
R(L, A, V_i) = [ W(O_k) \cdot R_0(L, A, V_k) ] \cdot \frac{1}{n} \sum_{j=1}^{n} S(V_i, V_j)
\]

(1)

where, \( L \) is a land parcel; \( A \) is an attribute of \( L \). \( R_0(L, A, V_j) \in [0, 1] \) \((j = 1, 2, \ldots, n)\) is the original value of \( R(L, A, V_j) \), which mainly aims at data cleaning for many times when the original sources are from different times. It can be given as 1 for the first times. \( W(O_k) \in (0, 1) \) \((k = 1, 2, \ldots, m)\) is the creditable weight of data source \( O_k \), which can be given by the user. \( W(O_k) \cdot R_0(L, A, V_k) = \max \{ W(O_j) \cdot R_0(L, A, V_j) | V_j = V_i, j = 1, 2, \ldots, n \} \). \( S(V_i, V_j) \in [0, 1] \) is the similarity degree between \( V_i \) and \( V_j \). Its value is got via Eqs. (2) and (3), where Eq. (2) is for character string values and Eq. (3) is for numeric values.

\[
S(V_i, V_j) = \frac{\sum_{m=1}^{n} | V_{im} \cdot V_{jm} |}{\sum_{m=1}^{n} V_{im}^2 \cdot \sum_{m=1}^{n} V_{jm}^2}
\]

(2)

\[
S(V_i, V_j) = \frac{1}{2} \exp \left( - \sqrt{V_i - \frac{1}{n} \sum_{r=1}^{n} V_r} \right)
\]

(3)

In Eq. (2), \( V_i, V_j \) are the character strings. \( V_i = (v_{i1}, v_{i2}, v_{i3}, \ldots, v_{ip}) \) and \( V_j = (v_{j1}, v_{j2}, v_{j3}, \ldots, v_{jp}) \). The vector dimension points to a word, whose value is
the times of the word appearing in land \( L' \). Especially, if the value of \( V_i \) is empty then \( M(V_i) = 0 \), \( R(L, A, V_i) = 0 \) and if \( V_i = V_j \), then \( R(L, A, V_i) = R(L, A, V_j) = 1 \). These satisfy the function definition. In the end, \( V_i \) with the biggest \( R(L, A, V_i) \) value of \( V_1, V_2, \ldots, V_i, \ldots, V_n \) is selected to be the value of attribute \( A' \) of land \( L' \), while those empty values or repetitive values \( (V_1, V_2, \ldots, V_i, \ldots, V_n; V_i \neq V_j) \) are eliminated.

### 2.2 Operation of algorithms

Here, we give an example of how to apply vector-match method. There is a land parcel (Table 1) in Jiang' an district intersection area between town and country of Wuhan city. By this vector-match method we carry out data cleaning on the land use change between 1988 and 1995. Due to the attribute ‘owner name’ belongs to character string, the measure-space dimension should be firstly extracted from its sources via word analysis, eg. A, B, C, D, E, F, G. Then \( V_1 \) and \( V_2 \), “Wuhan A B C D E Corporation” and “Wuhan A B C F G Corporation” are changed into \( V_1 = “A B C D E” \), \( V_2 = “A B C F G” \). \( V_2 = (1, 1, 1, 1, 1, 0, 0) \) and \( V_2 = (1, 1, 1, 1, 0, 0, 1) \). Hence, \( S(\langle V_1, V_2 \rangle) = 0.730 \) according to Eq. (2).

### Table 1 Data cleaning on a land parcel in Jiang’ an district

<table>
<thead>
<tr>
<th>Data origin</th>
<th>Owner name</th>
<th>Land area/\text{m}^2</th>
<th>Reliable weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Photogrammetry</td>
<td>Wuhan A B C D E Corporation</td>
<td>330.24</td>
<td>0.4</td>
</tr>
<tr>
<td>Land survey</td>
<td>Wuhan A B C D F G Corporation</td>
<td>332.00</td>
<td>0.5</td>
</tr>
<tr>
<td>Land statistic</td>
<td>Wuhan A B C F G Corporation</td>
<td>284.67</td>
<td>0.2</td>
</tr>
<tr>
<td>RESULTS</td>
<td>Wuhan A B C D F G Corporation</td>
<td>332.00</td>
<td></td>
</tr>
</tbody>
</table>

At the same time, land area is a numeric value. As to \( V_1 \) and \( V_2 \), \( S(\langle V_1, V_2 \rangle) = 0.995 \) according to Eq. (3). It is the same with other attributes. Finally, we get the results as in Table 1. Obviously, there are man-made errors in land statistics. Perhaps this owes to that the owner wants to pay less land tax, while the information of land survey is the most reliable.

After data cleaning, different databases from different sources in different times have become clean, complete and re-engineered. They can rebuild up a data warehouse of land control for data mining. The data warehouse is an extreme large database which stores huge amounts of data, however, it differs from general databases it is the integrated data set with the exact topic compositive, dynamic but stable. Its data includes integrated data, detailed data and summarized data, historical data interpretable data etc. These make good preparation for data mining.

### 3 Data mining based on rough set

Once the data warehouse has been set up successfully, it is the focus for data mining to discover useful information in the warehouse. Data mining is to discover knowledge, which is concealed, potential effective useful and understood. There are many kinds of data mining methods such as rough set, cloud theory (Di, 1999). Here a minimal decision algorithm based on rough set and data auditing is studied to discover land control knowledge.

#### 3.1 Principles of rough set

A database can be taken as a special case of a knowledge representation system. Let \( S = (U, C, D, V, f) \) be a knowledge representation system, where \( U \) is a nonempty set of objects (i.e. \( U = \{u_1, u_2, \ldots, u_n\} \), \( C \) is a nonempty set of conditional attributes and \( D \) is a nonempty set of decision attributes, \( A = C \cup D \) and \( C \cap D = \emptyset \) is the set of all attributes. Let \( V = \bigcup\{V_a|a \in A\} \), where \( V_a \) is a finite attribute domain and the elements of \( V_a \) are called values of attribute \( a \). \( f \) is an information function that \( f(u, a) \in V_a \), for every \( a \in A \) and \( u \in U \). Every object that belongs to \( U \) is associated with a set of values corresponding to the conditional attributes \( C \) and decision attributes \( D \). Table 2 shows an example of a knowledge representation system.

In Table 2, DJ is the districts of Wuhan City, and JA, JH, QK, HY, HS, WE, CD, JX are respectively Jiang’ an, Jianghan, Qiaokou, Hanasyang, Hongshan.
Table 2 Knowledge representation system

<table>
<thead>
<tr>
<th>$U$</th>
<th>DI</th>
<th>TO</th>
<th>PL</th>
<th>VE</th>
<th>FO</th>
<th>GR</th>
<th>BI</th>
<th>TR</th>
<th>WA</th>
<th>VI</th>
</tr>
</thead>
<tbody>
<tr>
<td>$U_1$</td>
<td>JA</td>
<td>$-1$</td>
<td>$-1$</td>
<td>$0$</td>
<td>$-1$</td>
<td>$0$</td>
<td>$3$</td>
<td>$1$</td>
<td>$-2$</td>
<td>$-1$</td>
</tr>
<tr>
<td>$U_2$</td>
<td>JH</td>
<td>$-2$</td>
<td>$-2$</td>
<td>$0$</td>
<td>$-1$</td>
<td>$0$</td>
<td>$1$</td>
<td>$1$</td>
<td>$-1$</td>
<td>$0$</td>
</tr>
<tr>
<td>$U_3$</td>
<td>QK</td>
<td>$-1$</td>
<td>$-1$</td>
<td>$0$</td>
<td>$0$</td>
<td>$0$</td>
<td>$1$</td>
<td>$0$</td>
<td>$1$</td>
<td>$-1$</td>
</tr>
<tr>
<td>$U_4$</td>
<td>HY</td>
<td>$-1$</td>
<td>$-1$</td>
<td>$0$</td>
<td>$-1$</td>
<td>$0$</td>
<td>$1$</td>
<td>$0$</td>
<td>$1$</td>
<td>$-1$</td>
</tr>
<tr>
<td>$U_5$</td>
<td>HS</td>
<td>$-1$</td>
<td>$-3$</td>
<td>$-1$</td>
<td>$-1$</td>
<td>$0$</td>
<td>$3$</td>
<td>$1$</td>
<td>$-2$</td>
<td>$-1$</td>
</tr>
<tr>
<td>$U_6$</td>
<td>WE</td>
<td>$-1$</td>
<td>$-3$</td>
<td>$-1$</td>
<td>$-1$</td>
<td>$0$</td>
<td>$3$</td>
<td>$1$</td>
<td>$2$</td>
<td>$-1$</td>
</tr>
<tr>
<td>$U_7$</td>
<td>CD</td>
<td>$-2$</td>
<td>$-3$</td>
<td>$-1$</td>
<td>$-1$</td>
<td>$0$</td>
<td>$3$</td>
<td>$0$</td>
<td>$2$</td>
<td>$-1$</td>
</tr>
<tr>
<td>$U_8$</td>
<td>JK</td>
<td>$-1$</td>
<td>$-2$</td>
<td>$-1$</td>
<td>$-1$</td>
<td>$2$</td>
<td>$1$</td>
<td>$1$</td>
<td>$-1$</td>
<td>$-1$</td>
</tr>
</tbody>
</table>

East & West Lake, Caidian Jiangxia. TO, PL, VE, FO, GR, BI, TR, WA, VI are respectively the change areas of total land, plantation vegetable land, forest land, grass land, transportation land, urban-rural building & industrial-mineral land, virgin land. $-3, -2, -1, 0, 1, 2$ respectively denote big down ($-8 \sim -12 \text{ km}^2$), down ($-2 \sim -8 \text{ km}^2$), small down ($0 \sim -2 \text{ km}^2$), unchanged ($0 \text{ km}^2$), small up ($0 \sim +2 \text{ km}^2$), up ($+2 \sim +8 \text{ km}^2$), big up ($+8 \sim +12 \text{ km}^2$), $U = \{ u_1, u_2, \ldots , u_8 \}$. Each object $u_i$ ($i = 1, 2, \ldots , 8$) is described by a set of condition attributes $C = \{ \text{TO}, \text{PL}, \text{VE}, \text{FO}, \text{GR}, \text{BI}, \text{TR}, \text{WA}, \text{VI} \}$, with attribute values $V_{\text{TO}} = \{-1, -2\}$, $V_{\text{PL}} = \{-1, -2, -3\}$, $V_{\text{VE}} = \{(0, -1)\}$, $V_{\text{FO}} = \{(0, -1)\}$, $V_{\text{GR}} = \{(0, -1)\}$, $V_{\text{BI}} = \{1, 2, 3\}$, $V_{\text{TR}} = \{(0, 1)\}$, $V_{\text{WA}} = \{-1, -2\}$, and $V_{\text{VI}} = \{0, -1\}$. And $V_{\text{DI}} = \{(\text{TA}, \text{JH}, \text{QK}, \text{HY}, \text{HS}, \text{WE}, \text{CD}, \text{JK})\}$ of decision attribute $D$ represents the set of concept descriptions to be learned from the attribute values of $C$.

Suppose $B$ is a nonempty subset of $A$. $u_i, u_j$ are members of $U$, and $R$ is an equivalence relation over $U$ ($R = U \times U$). Define a binary relation, called an indiscernibility relation as $\text{IND}(B) = \{(u_i, u_j) \in R | u_i, u_j \in U, \forall a \in B, f(u_i, a) = f(u_j, a)\}$. It is believed that $u_i$ and $u_j$ are indiscernible by a set of condition attributes $B$ in a knowledge representation system if and only if $\forall a \in B, f(u_i, a) = f(u_j, a)$.

The indiscernibility relation partitions $U$ into equivalence classes. Equivalence classes of the relation $R$ are called elementary sets in an approximation space $\text{APR} = (U, R)$. For any object $u_i \in U$, the equivalence classes of the relation $R$ containing $u_i$ are denoted as $[u_i]_R$. If $X \subseteq U$, then the lower approximation of the upper approximation are respectively $\text{APR}(X) = \{ u_i \in (U \mid [u_i]_R \subseteq X) \}$, $\text{APR}(X) = \{ u_i \in (U \mid [u_i]_R \cap X \neq \emptyset) \}$. $X$ is rough with respect to $\text{IND}(B)$ if $\text{APR} \neq \text{APR}$. And a subset (eg. $X \subseteq U$) defined with the lower approximation and upper approximation is called Rough Set (Pawlak, 1982). Obviously, we are sure that the data is in the set if it is in the lower approximation or we are sure that the data is not in the set if it is not in the upper approximation (Ahlqvist et al., 2000).

### 3.2 Elimination of superfluous attribute

During a data collection, all features that are believed to be useful and relevant are collected into the land databases. In a database system, we describe each object by the attribute values of $C$. Very often turns out that some of the attributes in $C$ may be redundant in the sense that they do not provide any additional information about the objects. Thus it is necessary to eliminate those superfluous attributes to improve learning efficiency and accuracy. Some objects may have the same attribute values because of elimination. After they are combined together, such objects are taken as a unified object in the following calculation.

Suppose $B \subseteq C$, then $\text{POS}(C) = \{ \text{POS}(C) \mid X \in \text{IND}(D) \}$. An attribute $a \in C$ is superfluous in $C$ with respect to $D$ if $\text{POS}(C) = \text{POS}(C \setminus \{a\})$. Otherwise, $a$ is indispensable in $C$ with respect to $D$.

If an attribute is superfluous in the information system, it should be removed from the information system without changing the dependency relationship of the original system. Knowledge is often represented in the form of rules that indicate the degree of association between conditional attribute $C$ and decision attribute $D$. And these rules support the decision of land control. A rule is a combination of values of some condition attributes. Traditionally, the rule is denoted as an implication:

$$ (C_1 = V_{i1}) \land (C_2 = V_{i2}) \land \ldots \land (C_m = V_{im} \implies (D = V_d)) \quad (4) $$

where $C_1, C_2, \ldots , C_m$ are the condition attributes and $D$ is a decision attribute.

The process by which the maximum number of condition attribute values of a rule are removed without decreasing the classification accuracy of the rule is called value reduction, and the resulting rule
is called a minimal decision algorithm rule. Thus a minimal decision algorithm rule is optimal in the sense that no condition could be removed without decreasing the classification accuracy of the rule. The calculation of a minimal decision algorithm rule is of importance with respect to data mining since they represent the most general patterns existing in the data.

4 Case study

Wuhan succeeds in controlling land use besides avoiding natural land crises, and land changes typically between 1988 and 1995 in its intersection area between town and country. So we take it for example. Table 2 shows its knowledge representation system. The original data have been not only cleaned by vector-match method, but also transformed with the method of histogram equilibria. Each object is described by a set of condition attributes C= \{TO, PL, VE, FO, GR, BI, TR, WA, VI\}, with attribute values \(V_{TO}=\{-1, -2\}\), \(V_{PL}=\{-1, -2, -3\}\), \(V_{VE}=\{0, -1\}\), \(V_{FO}=\{0, -1\}\), \(V_{GR}=\{0, -1\}\), \(V_{BI}=\{1, 2, 3\}\), \(V_{TR}=\{0, 1\}\), \(V_{WA}=\{-1, -2\}\), and \(V_{VI}=\{0, -1\}\). The set of values \(V_{DI}=\{JA, JH, QK, HY, HS, WE, CD, JX\}\) of decision attribute \(D\) represents the set of concept descriptions which are to be learned from the attribute values of \(C\).

Now we apply a minimal decision algorithm of rough set to find out knowledge that land control is interested in. Thanks to rough set, TO, VE, FO, GR, TR, VI are superfluous attributes in Table 2. After removing them from Table 2, Table 3 is obtained. As we can see, Table 3 is simpler but has the same discernibility as Table 2. However, Table 3 is not yet the optimal knowledge representation system. \(U_3\) and \(U_4\) have the same condition attributes, and the same with \(U_5\) and \(U_6\) and \(U_7\). Such similar objects should be united respectively, and all repeated attributes should be deleted except that only one attribute is reserved. After simplifying Table 3 further, Table 4 is obtained, which depicts a decision matrix. From Table 4, we can get several minimal decision rules for the decision attribute "DI".

<table>
<thead>
<tr>
<th>(U)</th>
<th>DI</th>
<th>PL</th>
<th>BI</th>
<th>WA</th>
</tr>
</thead>
<tbody>
<tr>
<td>(U_1)</td>
<td>JA</td>
<td>-1</td>
<td>3</td>
<td>-2</td>
</tr>
<tr>
<td>(U_2)</td>
<td>JH</td>
<td>-2</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>(U_3)</td>
<td>QK</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>(U_4)</td>
<td>HY</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>(U_5)</td>
<td>HS</td>
<td>-3</td>
<td>3</td>
<td>-2</td>
</tr>
<tr>
<td>(U_6)</td>
<td>WE</td>
<td>-3</td>
<td>3</td>
<td>-2</td>
</tr>
<tr>
<td>(U_7)</td>
<td>CD</td>
<td>-3</td>
<td>3</td>
<td>-2</td>
</tr>
<tr>
<td>(U_8)</td>
<td>JX</td>
<td>-2</td>
<td>2</td>
<td>-1</td>
</tr>
</tbody>
</table>

\[(PL=-1) \land (BI=3) \land (WA=-2) \rightarrow (DI=JA)\];
\[(PL=-2) \land (BI=1) \land (WA=-1) \rightarrow (DI=JH)\];
\[(PL=-1) \land (BI=-1) \land (WA=-1) \rightarrow (DI=QK, HY)\];
\[(PL=-3) \land (BI=3) \land (WA=-2) \rightarrow (DI=HS, WE, CD)\];
\[(PL=-2) \land (BI=2) \land (WA=-1) \rightarrow (DI=JX)\].

The knowledge shows that plantation decreased most in Hongshan district, East & West Lake district, Caidian district, Urban-rural building & industrial-mineral land increased most in Jiang’an district, Hongshan district. East & West Lake district, Caidian district, Water area decreased most in Hongshan district, East & West Lake district, Caidian district. Urban-rural building & industrial-mineral land occupied lots of land in the intersection area between town and country of Wuhan city, especially for plantation and water area. Obviously, Wuhan city has to strengthen land control and plantation protection in its intersection area between town and country. To keep the suitable use of land resources, it is suggested that the strictest measures should be implemented in Jiang’an district, Hongshan district, East & West Lake district.
In the light of modern city development, city and country should fuse with each other in function and configuration. The first industry, the second and the third should develop harmoniously, so as to form a modern, dispersed and high-effective urban and rural amalgamation body. At present, there are two development modes on an intersection area between town and country. One is the service district of city, the other is the dispersed group. In Wuhan city, small towns around it have not become a certain size, but the main city district owns a stronger political, economical and cultural attraction. And it develops a characteristic land use mode, city and country fusing with each other (Fig. 1), which can be seen from Wuhan city images of satellite remote sensing (Fig. 2). So Wuhan city should develop and improve 9 satellite cities such as Houhu, Qingshan, Baisha and so on in order to set up a compound ecosystem.

![Fig. 1 Fusion mode](image1.png)

![Fig. 2 Satellite images of Wuhan city](image2.png)

5 Dynamic land control

Obviously, all above are based on whether land information is true or not. It is well known that land data almost change every day, especially for an intersection area between town and country. Because land control needs timely information, data mining always discovers knowledge in a data warehouse from all spatiotemporal databases. In order to daily control land use in an intersection area between town and country, Wuhan city is advised to develop a real-time land control system.

As to land use, we can integrate GIS, GPS, charge-coupled device camera, remote sensing RS and online analytical mining (OLAM) into a real-time land control system (Fig. 3). It can be set up in a car. When the car goes in the monitoring area, the real land data can be acquired, then they are all put into basic land databases. In Fig. 3, digital photogrammetry system includes both analytical photogrammetry and digital image manipulation functions that can be used to extract land information such as feature and topographic essentials. It employs land survey information, land images (from satellite, aircraft and unmeasured camera) and so on. When the system is applied in land control it is firstly necessary to put the basic parameters into the system, and calculate the orientation parameters including interior orientation, relative orientation and absolute orientation. Secondly, do aerial triangulation, array stereo-images in the direction of epipolar line, and match images. Thirdly, on the basis of digital image, make landscape map on land dynamic use, get buildings information on the ground with computer auxiliary survey. After all operations are finished, many products on land control can be automatically obtained, e.g., imagine map, cadastral map, section map. All information should be put and stored into GIS. As the spatial information in land management is horizontal, the products are easier to be stored and utilized. The database of other related land data such as images, laws, economy, population etc. are also respectively built up. Then these databases are put into GIS, and with them a land data warehouse is rebuilt up.
after data cleaning with vector-match method. In the network, OLAM can help to use information and knowledge fully. With the user interfaces, we can let users construct data warehouses, select the desired sets of data, perform constraint-based interactive online analysis processing and mining, visualize and explore the results. Because OLAM takes advantage of widely available, comprehensive information processing infrastructure, it makes good use of existing and real-time infrastructure rather than constructs everything from scratch. Furthermore, OLAM provides an exploratory data analysis environment. It becomes possible to find out different subsets of data and at different levels of abstraction by drilling, pivoting, filtering, slicing, and dicing a multidimensional database and the intermediate data mining results. And it facilitates online, interactive selection of data mining functions and interestingness thresholds. Moreover, all land workers can share with the information and discover their own interested knowledge via graphical user interface API. This will speed the decision and avoid land problems those are important for land control. In the end, land control can be achieved before land is misused.

6 Conclusion

This paper presents a development of data mining method for land control. We proposed a vector-match method for the prerequisite of data mining that is data cleaning. And rough set is adopted to discover land control knowledge. Then a case study is carried out for applying the method. When the method is applied in the intersection area between town and country of Wuhan city, a set of knowledge was mined out. As its plantation is reduced seriously, Wuhan city should strengthen land control and plantation protection in such area and develop a mode that city and country fuse with each other. That indicates the method is feasible.

Vector-match method is able to do word analysis and numeric matching together, which benefits the amalgamation of heterogenous source data. It gives a similar degree concept to measure how two data from different sources on the same attribute are similar to each other. As for difficult character in formation, it firstly does word analysis to find out the basic characters of which a measure-space dimension is composed. All different original character data on the same object are vectorized according to the dimension. The length of each original data vector equals to the measure-space dimension. Then the derived inclination-cosine method is used to calculate the similar degree. As for numeric information, an index function is defined to match land parcel. The credibility degree of each source is calculated according to its confidence levels and times.

The confidence level focuses on its weight, and times pay attention to its historical accumulation and inherit. Finally, the source with the biggest credibility degree is selected. The minimal decision algorithms of rough set can eliminate the superfluous condition attributes and improve the speed to make a decision. The knowledge represents in the form of decision rules which matches the demands of land control.
At the same time, land crises become worse and worse. And information for land control changes every day. Therefore, it is necessary to develop a real-time land control system with GIS, GPS, charge-coupled device camera, remote sensing OLAM. Performing those functions interactively and viewing the results with data/knowledge visualization tools will greatly enhance the power and flexibility of exploratory data mining. With mining constraint, the land control knowledge can be discovered with OLAM. The knowledge database should also be set up in order to make full use of the knowledge when data mining. With the help of GIS, the knowledge can be represented as imagine map, tridimensional map, virtual forecasting map, cadastral map, land section map etc. according to the demand of land control.

References