Knowledge Discovery Applied to Agriculture Economy Planning

Bing-ru Yang and Shao-jun Huang
Information Engineering School
University of Science and Technology, Beijing, China, 100083
Email: bingru.yang@bj.col.com.cn

Abstract: The paper presents double-base cooperating mechanism by studying the knowledge discovery based on database (KDD) which changes the structure, running process and the mechanism of KDD. Then a new Knowledge discovery based on database is established as KDD*. Applied to agriculture economy planning, the KDD* provides scientific decision for instructing agricultural production.

Key Words: Knowledge Discovery, Agriculture Economy, Decision.

1. Introduction

In the agriculture research, management and its basic level department, a large amount of data, examples, knowledge and experiences have been accumulated. In the field of agricultural crop the data are not made full use. The accumulated data on seedling, soil, fertilization, water, harmful insect of all kinds of crops as well as weather and calamities are saved as archives. That is to say, the phenomenon of plentiful data and poor knowledge is more serious in agriculture than other. So the demands for knowledge discovery are more eager. If some new rules which are produced by dynamic changed factors can be found through finding interrelations of the factors from the plentiful data, examples, common experiences and knowledge, the economical and social benefits will be very great.

The agriculture is a large and complex system. The types of soil in the world are enormous. The kinds of crops are complex. The calamities of harmful insects appears frequently and their symptom changes constantly. The interrelations and its effects among fertilizer, water, density and weather haven’t been recognized. This is also the same with in the livestock, birds, fish and forestry. The relative database and knowledge base are characterized as large, multi-dimension, dynamic, incomplete and uncertain.

In recent years the market information didn’t flow smoothly in many places, especially the crop production planning isn’t instructed by the large dynamic market information. It causes blindness in the production planning and great fluctuation in the price which greatly affect agricultural market economy. How to collect the information in realization and find valuable and regular knowledge so as to effectively forecast and take measures in time will play an important role to the agricultural production.

Knowledge acquisition is always regarded as a bottleneck in the realization of intellectual system. Knowledge discovery partly solved the problem of knowledge acquisition. At present the development in knowledge discovery is mainly the traditional knowledge discovery based on database (KDD). Some intellectual methods, such as fuzzy logic, neural network, rough set and chaotic theory, are used in the KDD. But the KDD lacks means used by existing knowledge which helps to focus. The hypotheses and rules produced by KDD are directly evaluated. They are set into the knowledge base if passing the evaluation. Then the following defects are formed: first, many meaningless hypotheses are produced. It increases the burdens of evaluation and check on consistency and redundancy. It is close in the process of knowledge discovery; second, the KDD mines according to the need and interest of person, which lacks creative thought of computer itself to mine heuristically and directly. third, at present there are many experimental verification and original system but few practical system and tools.

In accordance with the above question, we first present the double base cooperating mechanism which is used to make basic knowledge base limit and drive KDD. This will lead to an open system of KDD: KDD* which is based on double base cooperating mechanism. KDD* breaks through the closeness of KDD. It makes database cooperate with knowledge base through interruptive and heuristic coordinator to find new knowledge.

2. The Introduction of KDD*

2.1 General Frame of KDD*
This figure shows the logic structure of the system and the relations between all the parts. From this figure we can see that the modules can be divided into the following parts:

**Pre-processing:** To process the original data by purifying the data, specific changing, etc. and create the DMDB which is used in the process of data mining and knowledge discovery.

**Focusing:** namely to chose data from data mining. The main method in focusing is clustering analysis and detecting analysis. The method to direct the focusing are: (i) the expert, through man-machine interaction, inputs the knowledge in which he is interested and direct the direction of the data mining. (ii) Data directional mining by using heuristic coordinator.

**Hypothesis rules:** It is the core process of KDD. It uncommonly abstracts the hidden, unknown and potential valuable information in database which has the character of large amount of data, incompleteness, uncertainty, structure and sparseness. In the system the abstracted information is causality relation rule. Thus the basic knowledge base will be further improved. The mining methods that are used are statistics induction reasoning and qualitative reasoning. The former method will be discussed in 2.2.

**Double base cooperating mechanism:** to process the acquired rules by using interruptive coordinator and heuristic coordinator, and to exciting the data focusing for data mining by using relative strength. This will be discussed in 3.2.

**Evaluation:** this process is mainly used to evaluate the acquired rules in order to decide whether they will be stored into the derived knowledge base. The main methods are: (i) relative strength sets up a threshold value and be realized by computer; (ii) experts evaluate through man-machine interaction interface and also evaluates all kinds of figures and analysis materials provided by visual tools. Experts
evaluate mainly by using experiences and the relative strength of acquired rules. The rules are stored as new knowledge into the derivative knowledge base after passing evaluation.

2.2 The Reasoning Algorithm of Causality Statistics Induction

This algorithm uses incomplete induction approximate inference in statistics and credibility theory in uncertainty theory, by counting the examples in database and using the property with a large amount of examples as module, and gets a set of rules by credibility theory.

Possessed conditions: data focusing has been completed i.e. is ready to mine the two language variable A, B (e.g. the kinds of crop and its production). The mining process is as follows:

2.2.1 The Computer Decide the Relativity of the Corresponding Language Value through Statistic Analysis.

Divide A, B as A(A₁A₂...Aₘ)B(B₁B₂...Bₙ) according to their language values. If A and B are both single variable then we have A(A₁A₂A₃...Aₘ), B(B₁B₂B₃B₄B₅). Given A is the intersection of ml variables, m=5ᵐ⁻. Given B is the intersection of nl variables, n=5ⁿ⁻. Thus there are altogether m.n kinds of combination <Ai, Bj> i=1, 2,...m j=1,2,...,n. To the possibility factor Pk=Cnk/N k=1,2,..., m.n corresponding to each computation, P=0.5 is the highest possibility. If Pk>0.5, <Ai, Bj>is selected, otherwise it is eliminated and these two are considered to have no relativity.

2.2.2 Analyze A and B through Visual Tools

Experts can use visual tools, such as a distribution figure to decide the combination of the selected or eliminated areas. The areas here have one to one mapping relation with the language value mentioned above, i.e. the language value and the corresponding radius equals the corresponding area. The acquired area combination must be changed into corresponding language combination which is to be used in the later computation. Get the two highly relative properties e.g. Ai and Bj, and draw the corresponding values e.g. statistic value N, statistic value CnAiBj, appearing both in Ai and Bj, statistic value CnAi, appearing in Ai, and statistic value CnBj, appearing in Bj to decide which variable have causal relation.

2.2.3 Get Weight of the Premise in the Hypothesis Rule (Ai, Bj)

Given Ai is single premise, its weight is 1; given Ai is the interaction of many premises, i.e. rule R: Ai Bj is:

\[ R(P₁,p₁),(P₂,p₂),..., (Pₚ,pₚ),... \]

Then the corresponding ri in \( P₁,p₁, (Pᵢ,pᵢ) \) and ..... can be gotten from the following formula. The weight in its rule can be gotten according ri.

\[
r_i = \frac{m\sum_{j=1}^{m} q_j \cdot (p_j - (p_i - q)) - \left(\sum_{j=1}^{m} p_j - (p_i - q)\right)\sum_{j=1}^{m} q_j}{\sqrt{m\sum_{j=1}^{m} (p_j - (p_i - q))^2 - \left(\sum_{j=1}^{m} q_j\right)^2}}
\]

Causality statistics induction reasoning algorithm flow is shown as following:

3. Double Bases Cooperating Mechanism:

3.1 Basic Theory

The technological realization of double-base cooperating mechanism is to construct interruptive and heuristic coordinators. To realize them there are some requirements: The large (basic) knowledge base is divided into several correlative sub-knowledge bases according to each domain; Meanwhile, the real database is divided into
correlative sub-databases according to each domain. Thus the layers between knowledge nodes in mining knowledge base and data sub-class (structure) in mining database make a one to one mapping. The basic theory which is proposed by us is pan-homotopy conception and the following structure mapping theorem: (Details can be found in reference [1][2])

Theorem (Structure Mapping Theorem): Aiming at $X$, in the sub-database corresponding to sub-knowledge nodes, $<E_F>$ of knowledge nodes and $<F_D>$ of data sub-class (structure) are identical pan-homotopic type spaces.

This theorem presents the mapping of layers between knowledge nodes in the sub-knowledge base and data sub-class in corresponding sub-database, shown in fig.3.

Fig.3 corresponding construct graph

On the basis of the research above, we can see that in the knowledge discovery system mathematical structure of database and knowledge base can be essentially come down to pan-homotopy category. Namely database is pan-homotopy category combined with data sub-type (structure ) set and “mining path”, which is called data mining category; and knowledge base is pan-homotopy category combined with knowledge nodes set and “reasoning arc”, which is called knowledge reasoning category. Moreover some results about the isomorphy and restricting mechanism of knowledge reasoning category $C_R,E,in <E,F>$ and data mining category $C_D,F,in <F,D>$ are got, and “directional searching” and “directional mining process” are solved.

3.2 The Technological Realization
3.2.1 Interruptive Coordinator
The main function of the interruptive coordinator is, when the rules (knowledge) have been created from the focusing of the data in the real base, to “interrupt” the process of the KDD and to search whether there is a repetition of the created rule in the corresponding position of the knowledge base. If so, cancel this created rule and return to the beginning of the KDD. There need some special technology and methods to process contradiction. If not, continue the process of the KDD i.e. evaluate and store the result.

Because the interruptive coordinator is introduced into KDD, the inconsistent and redundant knowledge can be canceled earlier. Only those who are possibly accepted as new knowledge are evaluated and the evaluation work is greatly reduced. At the same time redundancy is processed in real time. This avoids complication of problem accumulated in a long time. In practical expert system, the amount of rules which finally become new knowledge are rather small compared with the original knowledge (it is difficult to find new knowledge), and a great number of rules are repetitive and redundant, so the introduction of interruptive type coordinator into KDD enhance the efficiency.

The function of heuristic coordinator
Knowledge shortage is found. Data sub-class corresponding to real database uses heuristics and is
activated to produce “directional mining process”. To find the knowledge shortage in knowledge base especially in rule base, one of the methods is to compute the causality rule strength in each possible knowledge node in the whole causality network.

The causality rule strength consists of a group of three factors which can be expressed as

$$\pi (H,E) = \alpha \beta \gamma = CF(E) \cdot P(E)$$

$$\beta = CF(H,E)$$

$$\gamma = CF(H) \cdot P(H)$$

Among it CF(E) is the reliability of premise, P(E) is pre-probability, CF(H,E) is the reliability of rule, CF(H) is the reliability of conclusion, P(H) is pre-probability. CF(E) and CF(H,E) are known. It consists of the whole random and fuzzy uncertain information of the rule. According to the causality rule strength the priority of directional mining can be determined and those can not be mined will be excluded.

4. Properties of KDD*

Compared with KDD, KDD* is a new structure of knowledge discovery which blends KDD and double base cooperating mechanism. It has the following characters:

1) KDD* organically make new knowledge found by KDD* communicate and merge with the knowledge in knowledge base and become one organism.

2) In the process of knowledge discovery, KDD* processes those redundant, repetitive and incompatible information in real time. This effectively decreases the complication of problem caused by a accumulated process. At the same time the preconditions are given for the merge and fusion of new and old knowledge.

3) KDD* changes and optimizes the process, structure and running mechanism of knowledge discovery.

4) From cognition KDD* strengthens and provides intellectual degree of knowledge discovery and enhances the cognition of computer itself. This is the direction for a long team.

5) Double base cooperating mechanism, the core technology of KDD*, shows the mapping between sub-knowledge base and data sub-class under a certain principle of establishing base. It provides a valid technology to decrease search space and improve mining efficiency.

5. Knowledge Discovery in Agricultural Economy Planning

In agriculture system there are abundant data which form all kinds of database such as relation database, time-spatial database, object-oriented database and multimedia database. But the data in these database are not made full use and hold plenty of storing space. Therefore it is necessary to mine.

In order to find knowledge from a database, it is necessary to process the database and establish corresponding basic knowledge base. Then All kinds of methods are used to mine the data in database. For example, Selenium (simplified as Se) is a necessary microelement for human and animal. It has many biological functions. Lack of Se is the main reason for many diseases, such as cataract, mastitis, cancer, large bone disease and so on. Rice is one of the main foods in the world. The content of Se is related to nutrition of Se in the human body. But most rice production areas are short of Se or have low content of Se. Therefore if we can find the dynamic changing rules under which rice sorbs Se, it will play an important role to instruct agriculture production and improve human health. Now there are some processed agricultural data which are shown in the following tables.

KDD* is applied to analyze the data in the table and finds that the accumulation of dry material isn’t at the same speed with that of Se in the rice. The peak of former is in the middle of growing period, the latter in the late period. This is a rule that will be stored in Knowledge base. According to the rule we should fertilize Se again before the period of filling starch in rice. On the other hand rice has certain ability to sorb Se. So fertilizing Se in those areas that lack Se or have low content of Se can greatly enhance the content of Se in rice and improve its nutrition quality. Doing so on one hand can instruct us to fertilize reasonably, on the other hand can instruct manufacturer of fertilizer to add different microelement in different stage so as to meet the demand of agriculture production. Other data of agricultural crop can be treated so.
<table>
<thead>
<tr>
<th>Bearing period</th>
<th>Growing time (d)</th>
<th>Accumulate speed ($\text{g pot}^{-1} \text{d}^{-1}$)</th>
<th>Stage accumulation ($\text{g pot}^{-1}$)</th>
<th>Stage comparative accumulation (%)</th>
<th>Accumulation ($\text{g pot}^{-1}$)</th>
<th>Comparative accumulation (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seedling period</td>
<td>30</td>
<td>0.755</td>
<td>22.65</td>
<td>6.70</td>
<td>22.65</td>
<td>6.70</td>
</tr>
<tr>
<td>Spike period</td>
<td>60</td>
<td>5.564</td>
<td>166.93</td>
<td>49.38</td>
<td>189.58</td>
<td>56.09</td>
</tr>
<tr>
<td>Filling starch period</td>
<td>80</td>
<td>3.818</td>
<td>76.35</td>
<td>22.59</td>
<td>265.93</td>
<td>76.01</td>
</tr>
<tr>
<td>Ripe period</td>
<td>100</td>
<td>3.605</td>
<td>72.09</td>
<td>21.33</td>
<td>338.02</td>
<td>100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bearing period</th>
<th>Growing time (d)</th>
<th>Accumulate speed ($\text{g pot}^{-1} \text{d}^{-1}$)</th>
<th>Stage accumulation ($\text{g pot}^{-1}$)</th>
<th>Stage comparative accumulation (%)</th>
<th>Accumulation ($\text{g pot}^{-1}$)</th>
<th>Comparative accumulation (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seedling period</td>
<td>30</td>
<td>24.30</td>
<td>729.00</td>
<td>5.82</td>
<td>729.00</td>
<td>5.82</td>
</tr>
<tr>
<td>Spike period</td>
<td>60</td>
<td>142.02</td>
<td>4260.68</td>
<td>34.03</td>
<td>4989.68</td>
<td>39.85</td>
</tr>
<tr>
<td>Filling starch period</td>
<td>80</td>
<td>151.61</td>
<td>3320.19</td>
<td>24.22</td>
<td>8021.78</td>
<td>64.07</td>
</tr>
<tr>
<td>Ripe period</td>
<td>100</td>
<td>224.95</td>
<td>4498.99</td>
<td>35.93</td>
<td>12520.77</td>
<td>100</td>
</tr>
</tbody>
</table>

6. Conclusion
Agriculture production is an important thing to a country and its people. Reasonable planning for agriculture production will take great effect on a country. The article provides a new method of scientific decision for agriculture economy planning. It decreases the loss caused by planless production and will be instructive to the development of agriculture.

On the basis of KDD, double base cooperating mechanism can be applied to mine knowledge automatically and directionally. It can also process repetitive, contradication and redundant rules. This will greatly improve the mining effeciency. The two kinds of coordinator can be independent system and install any existing KDD software to communicate with original knowledge base. It expands the function of original KDD greatly.

Reference