XVIIth World Congress of the International Commission of Agricultural and Biosystems Engineering (CIGR)

Hosted by the Canadian Society for Bioengineering (CSBE/SCGAB)
Québec City, Canada June 13-17, 2010

A CASE-BASED FORECASTING SYSTEM FOR DANGSHANSU PEAR SCAB COMBINED WITH FUZZY CLUSTERING[1]

LI SHAOWEN, LI RUI, ZHANG PENG, FANG WENJUAN, LIU LI, WANG WEIWEI, GU LICHUAN

School of Information and Computer Science, Anhui Agricultural University, Hefei 230036, China; shwli@ahau.edu.cn

CSBE101677 – Presented at the 8th World Congress on Computers in Agriculture (WCCA)

ABSTRACT In accordance with the epidemic law of Dangshansu pear scab (Venturia nashicola Tanaka et Yamamoto) and based on the pear diseases, observations and weather data in the area of the ancient Yellow River valley, the case-based forecasting system for Dangshansu pear scab (PDFCBRS 1.0) was developed, combined with fuzzy clustering, and therefore had fuzzy reasoning capability and could be well emulated the expert reasoning with experience. The experimental results showed that the "CBR (Case-Based Reasoning) + Fuzzy" not only changed the forecasting method of pear scab from experience approach to computer processing but also significantly reduces the human effort required for forecasting Dangshansu pear scab.

Keywords: expert system, Dangshansu pea, pear scab, forecasting system, CBR, fuzzy reasoning

1 INTRODUCTION

The forecast of plant disease is the basis and premise of effective disease prevention. By far, there are mainly experience and mathematic model method in the fruiter disease forecast, but the effect is not satisfying. Thus, in practice, the expert experiential forecast according to the epidemic law is the dominant method. Case-Based Reasoning is one of the major reasoning paradigms in artificial intelligence. A CBR system solves a new problem by retrieving a similar one from a case base, which stores the experienced solutions to past problems (Zaluski et al., 2003). When it cannot find a solution that is similar enough to solve a new problem, a CBR system will adapt the solution of a relatively similar problem to the new one. The disease forecast based on CBR can not only overcome the weakness of mathematic model but also replace expert's experience with computer auto reasoning forecast. Therefore, aiming at the forecast and prevention of the key disease of Dangshansu pear scab in the area of ancient Yellow River valley, we developed the case-based forecasting system combined with fuzzy clustering (PDFCBRS 1.0) and attempted to offer a new approach for the forecast of the fruiter disease.

[1] Fund projects: National High Technology Research and Development Program of China (No. 2006AA10Z249); National Natural Science Foundation of China (No. 30971091, 30800663)

The remainder of the paper is organized as follows. Section 2 details the structure for PDFCBRS 1.0, and case representation, storage and retrieval; Section 3 presents a running example of the system; Section 4 discusses the results and limitations of the system.

2 SYSTEM DESIGN

2.1 System Structure

According to the principle of CBR, the system is made up of input module, output module, case retrieval, case match, case revisal, storage and maintenance of case-base. The whole process is as Fig1 shows. The function of input module is incepting known information pear scab data and weather data in the past period. Then the information is changed into being target case. In accordance with the epidemic law of Dangshansu pear scab and the practical forecasting demand (short forecast), we consider that it is more appropriate to forecast the next ten days of a month by the information of the previous two ten days of the month.

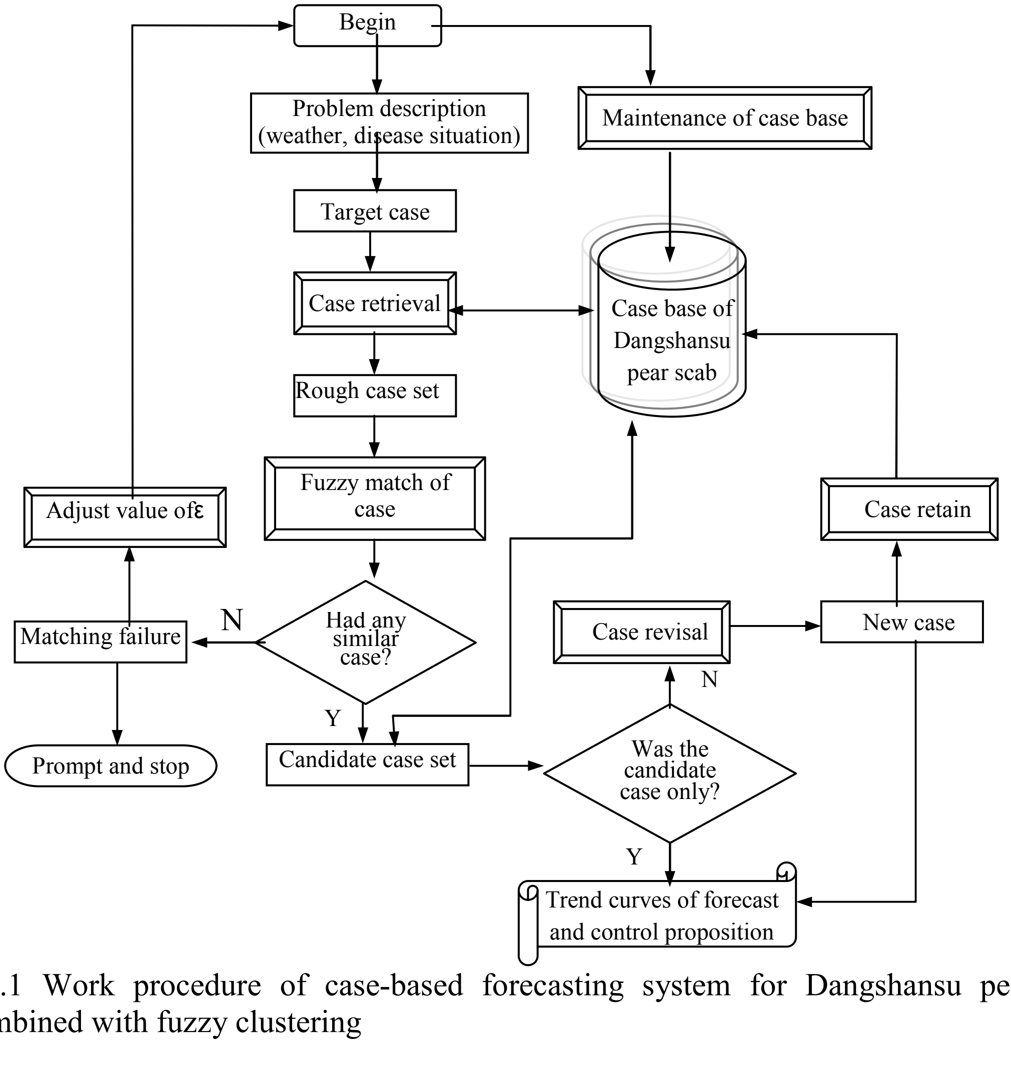


Fig.1 Work procedure of case-based forecasting system for Dangshansu pear scab combined with fuzzy clustering

2.2 Case representation, storage and retrieval

When we develop a case-based reasoning (CBR) system for various real-world applications, a significant challenge that faces us is how to create a high-quality case base and how to get a case from huge case base. Without a high-quality case base, it is impossible for any CBR to function well for solving problems (Yang et al., 2003). Through the research we found that the epidemic period of Dangshansu pear scab in the ancient area of Yellow River valley is between the middle ten days of March and the first ten days of October. And it is easy to come into being three disease fastigium periods, which is from the last ten days of April to the middle ten days of May, from July to August and in September. The degree of the disease is most correlative with rainfall. Besides, it also has something to do with air humidity and atmosphere temperature. Furthermore, the criterions of evaluating Duangshansu pear scab at present are mainly percentage of infected leaf (PIL) and percentage of infected fruit (PIF). In order to showing the inherent law, we defined a case as a group of features: date, rainfall, air humidity, atmosphere temperature, PIL, PIF and the suggestions of prevention. We also defined a year as an individual case unit. They are illustrated in table 1. All values except date and rainfall are the mean of PTD.

Table 1 Case representation of Dangshansu pear scab

Year	Month	PTD	RF / mm	RH / %	AT / □	PIL / %	PIF / %	CP
1999	…							
	4	LTD	30.0	78	16.4	18.9	18.9	
	5	FTD	6.3	79	16.4	20.3	20.2	
	5	MTD	17.7	76	16.8	19.2	19.7	
	5	LTD	3.5	79	21.2	11.1	16.9	
	6	FTD	3.2	54	25.7	8.2	10.1	
	6	MTD	4.3	53	25.5	4.3	7.9	
	6	LTD	0.0	65	26.7	3.2	5.6	
	7	FTD	61.1	73	30.2	5.1	8.8	
	7	MTD	156.1	86	25.7	10.3	10.5	
	7	LTD	55.8	86	27.7	7.9	9.9	
	…							

Note: PTD: a period of ten-days; FTD: the first ten-days of a month; MTD: the middle ten-days of a month; LTD: the last ten-day of a month; RF: rainfall; RH: relative humidity; AT: atmosphere temperature; PIL: percentage of infected leaf; PIF: percentage of infected fruit; CP: control proposition.

Based on the formation of above case representation, this system employs versatile Database Management System (DBMS), which can organize, store and manage the basic disease and weather data. The system retrieves case base using month and PTD as the indices and forms a temporary case-base (original case-base), and then matches them. It not only can meet the existing disease data from the department of plant protection and weather data from weather bureau, but also can avoid the tedious process of knowledge acquisition in the case design. So it is convenient to unify database and knowledge base and shorten the system development period (Xiong et al., 1995).

2.3 Case-based fuzzy reasoning

If the case-base is compared to the rule-base in the rule-based reasoning (RBR) expert system, then the mechanism of case retrieval, match and adaptation in CBR system resembles the reasoning mechanism of RBR system. Moreover case match and adaptation is quite easier to denote the uncertainty reasoning (Yang et al., 2003). Therefore, we applied the fuzzy ISODATA clustering to the uncertainty similar matching of case in the system and made similarity clustering evaluation on the features of problem description between target case and original case. Due to the different importance of each feature, we adopted average weighing (ω) of many experts. The weighing of each feature is following (Xiong et al., 1995). Rainfall (ω_1),0.45; relative humidity (ω_2),0.10; atmosphere temperature (ω_3), 0.05; percentage of infected leaf (ω_4),0.20; percentage of infected fruit (ω_5),0.20 (Li et al., 1994; Minor et al., 2000).

Supposing $X = \{x_1, x_2, \cdots, x_k, \cdots, x_n\}$ as target case set and original case set, each case x_k is represented as p features. That is $x_k = (x_{k1}, x_{k2}, \cdots x_{kh}, \cdots, x_{kp})$, $k = 1,2,\cdots,n; h = 1,2\cdots,p$. Then the algorithm steps are as following. □Decide the number of cluster center C= 4 and fuzzy exponential $m = 2$, then pre-process this feature data by this formula: $x'_{kh} = \dfrac{x_{kh} - \min\limits_k x_{kh}}{\max\limits_k x_{kh} - \min\limits_k x_{kh}}$. □Based on $\sum\limits_{i=1}^{c} \mu_{ik} = 1, \mu_{ik} \in [0,1], \forall k,$ initialize classification matrix $U^{(0)} = (\mu_{ik}^{(0)})_{c \times n}$, and appoint a small positive number $\varepsilon (\leq 10^{-1})$ as final criterion. □By $U^{(l)} = (\mu_{ik}^{(l)})_{c \times n}$ (1 from 0), calculate the cluster center according to $\upsilon_i^{(l)} = \sum\limits_{k=1}^{n} \left(\mu_{ik}^{(l)}\right)^m x'_k \bigg/ \sum\limits_{k=1}^{n} \left(\mu_{ik}^{(l)}\right)^m$ $(i = 1,2,\cdots,c)$. □Calculate the membership degree of the classification matrix $U^{(l+1)}$ with $\mu_{ik}^{(l+1)} = 1 \Bigg/ \left[\sum\limits_{j=1}^{c} \left(\dfrac{\left\| x'_k - \upsilon_i^{(l)} \right\|}{\left\| x'_k - \upsilon_j^{(l)} \right\|} \right)^{1/(m-1)} \right]$, $\left\| x'_k - \upsilon_j^{(l)} \right\| = \sqrt{\sum\limits_{h=1}^{p} \omega_h (x'_{kh} - \upsilon_{jh}^{(l)})^2}$ $(j = 1,\cdots,i,\cdots,c)$ □ ω is the feature weighing. □Calculate error $\delta = \max\limits_{i,k} \left| \mu_{ik}^{(l+1)} - \mu_{ik}^{(l)} \right|$. As $\delta \leq \varepsilon$, it is the best classification matrix $U^* = (\mu_{ik}^*)_{c \times n} = U^{(l+1)}$ and the iteration is over. Otherwise $l = l + 1$, return to the third step. □Based on the maximum principle to classify, if $\mu_{i_0 k}^* = \max\limits_i (\mu_{ik}^*)$, x_k is clustered into i_0. Every original case that is clustered into the target case will become candidate case, which is considered as the most similar to the target case (Schank, 1982).

2.4 Case revisal

Once a matching case is retrieved a CBR system should adapt the solution stored in the retrieved case to the needs of the current case. If the candidate case is only one after fuzzy clustering matching, then the system directly outputs the solution of candidate case as the forecasting result and doesn't retain it as a new case. Otherwise the system make weighted average based on the membership degree of all the candidate cases in the best fuzzy classification matrix U^*, then outputs the result as forecasting result of target case and retains the revised case as a new case into the case-base for the future (Watson, 1997; Portinale et al., 2000).

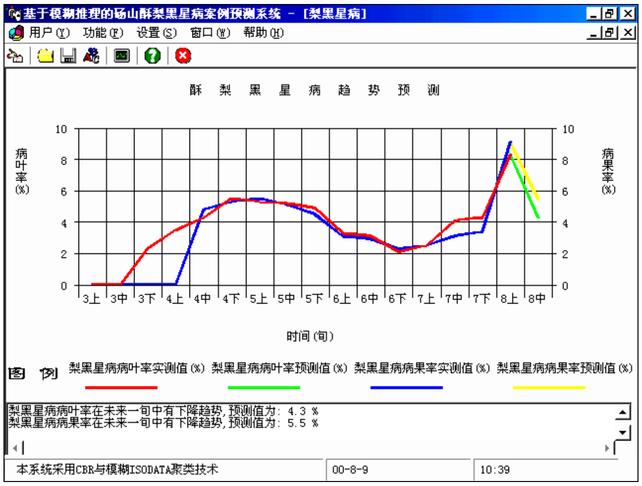
Supposing target case x_t and the best fuzzy classification matrix $U^* = (\mu_{ik}^*)_{c\times n}$, then the forecasting output of target case is $x_{t\cdot} = \sum_{k=1}^{n'} \mu_{\cdot k}^{*\,m} \cdot x_{k\cdot} \Big/ \sum_{k=1}^{n'} \mu_{\cdot k}^{*\,m}$. And $x_{t\cdot}$ is the PIL and PIF of ten days that will be predicted in target case. $x_{k\cdot}$ is the corresponding PIL and PIF in candidate case. $\mu_{\cdot k}^*$ is the membership degree of candidate case in U^*. n' is the number of candidate case. m is the fuzzy degree.

3 THE SYSTEM REALIZATION AND RUNNING

Based on the ideas proposed above, the system named PDFCBRS 1.0 was developed by Visual Basic 6.0 for forecasting Dangshansu pear scab, including main controlling, reasoning part, case base of Dangshansu pear scab, management subsystem and interface of user.

According to the infected degree of Dangshansu pear scab from 1975 to 2009, we divided them into four clusters: heavy disease year (PIL, PIF>=20%), middle disease year (20%> PIL>=10%, 20%>PIF>=10%), light disease year (10%>PIL>3%, 10%>PIF>3%), little disease year (PIL, PIF<=3%). There is mainly the dynamic data having the feature of above cluster and relevant factor in the source case-base. The system deletes the quite similar and overlapped data among different years by the maintenance program in order to improve the system efficiency and insure the case is typical and effective. As long as not influencing the typicality of case-base, the system should try best to retain newer year data, especially the data latest to the forecasted year, which is more significant to direct the forecast. For example, the state of disease in 1978 is similar to that in 1979, both of which belong to little disease year, so the system will delete the dynamic data of 1978 and retain 1979.

As the system is running, at the beginning the system retrieves the source case-base according to the information of the target case and defines the retrieved cases similar to the target case as original cases. Then calculate the similarity between original cases and target case by fuzzy clustering method. And defines the original case clustered in target cases as candidate cases. Then it gets new forecasting information after modification by synthetically considering each candidate cases. In accordance with the new forecasting information, the system will re-adopt information from the case-base further, and then output the result with forecasting curves and prevention suggestion. Finally in accordance with the practice, the system will decide whether define it as new case information and retain it in the case-base. When cases are matched, if there is no one similar case, you should adjust clustering precision (ε). If there is still none after adjusting many times, the system will prompt and stop and it shows that there is no essential case in the case-base and new cases should be appended.



梨黑星病病叶率在未来一旬中有下降趋势，预测值为：4.3 %
梨黑星病病果率在未来一旬中有下降趋势，预测值为：5.5 %

本系统采用CBR与模糊ISODATA聚类技术 | 00-8-9 | 10:39

Fig. 2 Display of forecasting results

An example as following was given to show the whole process above. Supposing today was August 9th in 2009, what was the development trend about Dangshansu pear scab in the next ten days? The following are the steps: □Check whether case-base has retained the essential previous typical dynamic year data (source case). Time span of every year is possible disease development period (from MTD of March to FTD of October). □Input the dynamic data of 2009 that is PTD data from March to FTD of August in the form of table 1, which has been known as target case. □Decide the forecasting PTD (MTD of August in 2009) and clustering precision(ε). The system will be automatic to recognize the data of the first two PTD (LTD of July and FTD of August in 2009) as described information. Then fuzzy clustering matches this information to the corresponding data of the two PTD in previous year. □Click "run". The system will run according to the process showed in Fig1. □Display the result. The Fig 2 shows the changing curves on PIL and PIF of pear scab since MTD of March in 2009.The segment from FTD of August to MTD of August in Fig 2 is the forecasting trend. The textbox under the figure also shows the forecasting description and value. If the matched candidate case is only one, the advice of preservation will be shown in the textbox too.

In order to validate the forecasting effect, we took the latest three years (2007~2009) as the target year and forecasted the disease of each PTD, then made correlation statistical analysis between forecasting value and observations as Table 2 displays. The result shows that the correlation coefficient between forecasting value and observations is great significant, which shows the fitting degree between dynamic curve of forecasting value and observations is quite well. So the system is feasible to predict pear scab.

Table 2 Correlation analysis of forecasting data and observations

Year	CCPIL	CCPIF
2007	0.810**	0.793**

2008	0.641^{**}	0.601^{**}
2009	0.867^{**}	0.873^{**}

Note: CCPIL: correlation coefficient of percentage of infected leaf; CCPIF: correlation coefficient of percentage of infected fruit; **: great significantly different (*P*<0.01).

4 DISCUSSION

The process of CBR closely resembles human reasoning. The monograph 'Dynamic Memory' by Schank in 1982 is widely held to be the origins of CBR, and it had been subsequently further validated in many intelligent system architecture. So it becomes a bright foreground method in artificial intelligence and expert system. And it also has a great potential in the agriculture application, especially in insect and disease forecast, which can replace the expert experience with auto forecast by computer reasoning. What's more, the system based on CBR has the ability of self-learning, thus as it running gradually, the forecasting result will be more and more accurate.

The epidemic of plant disease is a dynamic system that is influenced by many infectors and the law is complex, changeful and experiential. It is unpractical to design a cause and effect model or elicit them into rule that is employed by traditional expert system. However, in practice, it is more practical and more convenient to give concrete examples. Even though a statistic forecasting model is established, the forecasting accuracy is always awful. Why? Firstly, the factors that can be taken into account are limited. Secondly, much useful information has been deserted as error in the statistic modeling. Thirdly, the model is fixed after it is established. So it is not easy to suit the frequent change. Therefore it has inherent limitations. Before the system PDFCBRS 1.0 was designed, we had already tried to solve the forecasting pear scab problems by statistic model and traditional expert system. However, both were aborted because of bad suitability and difficulty in knowledge elicitation. In fact, not all the development and epidemic law of plant disease can be all represented by case, but CBR is effective on many occasions.

Aiming at the problem traits, the system introduced fuzzy clustering into uncertain match of case. Though it didn't follow the traditional similar measurement method, such as nearest neighbor, induction and template retrieval etc, the system attempt to match cases by fuzzy ISODATA clustering, which can cluster the cases that are similar to the target case and act as similar matching. Moreover, the fuzzy ISODATA clustering is based on the idea of fuzzy soft division, which closely resembles human thinking habit (experience reasoning). Therefore, the method has better performance than others, which has been validated by system PDFCBRS 1.0 working.

REFERENCES

Zaluski M, Japkowicz N, Matwin S. 2003. Case authoring from text historical experiences. In: Proceedings of the sixteenth Canadian conference on artificial intelligence (AI'2003). Dalhousie University, Halifax, Nova Scotia, Canada

Yang C, Orchard R, Farley B, Zaluski M. 2003. Authoring cases from free-text maintenance data. In: Proceedings of IAPR international conference on machine learning and data mining (MLDM2003). Leipzig, Germany

Xiong J T, Xiong F L, Tu R S. 1995. Case-based reasoning in forecasting insect pests.

Proc. of PACES’95. Beijing: Publishing House of Electronics Industry

Li X G, Li H X, Chen S Q. 1994. Fuzzy clustering analysis and application. Guiyang: Science and Technology Press of Guizhou

Minor M, Hanft A. 2000. The life cycle of test cases in a CBR system. In: Proceedings of advances in case-based reasoning: 5th European workshop, EWCBR 2000, Trento, Italy

Schank R C. 1982. Dynamic memory. Cambridge: Cambridge University Press

Watson L. 1997. Applying case based reasoning: Techniques for enterprise systems. Los Altos CA: Morgan Kaufmann Publishers

Portinale L, Torasso P. 2000. Automated case base management in a multi-model reasoning system. In: Proceedings of advances in case-based reasoning: 5th European workshop, EWCBR. Trento, Italy