Decision Tree-Based Extension Strategy Generation and Knowledge Discovery

Kaijie Wang[1,a], Libo Xu[1,b*] and Jinghui Li[1,c]

[1]Ningbo Institute of Technology, Zhejiang University, Ningbo 315000, China

[a]805861498@qq.com, [b]xu_libo@163.com, [c]19317128@qq.com

*Author for correspondence.

Keywords: decision tree; extension transformation; extension strategy analysis; classification

Abstract. To overcome the drawbacks of both traditional extension strategy analysis methods and traditional decision tree classification, an extension strategy analysis method based on decision tree classification was proposed. First, original data were obtained through basic-element theory and optimized by extension data pre-processing. Second, optimized data were statically classified by dependent function, and an extension decision tree was constructed according to the results of information entropy and information gain calculations. Then, through extension strategy analysis, extension strategies were generated. Finally, the validity and feasibility of this method were proved by an experiment.

Introduction

An extension strategy analysis method starts from the extension of objects, solves incompatible problems with extension, transformation and evaluation[1-2].

Nowadays, with the improvement of the degree of recognition of extenics, extension strategy analysis methods have been applied in many fields. Reference [3] applied extension dependent prediction algorithm, which was more efficient and data compatible than nearest neighbor algorithm, to collaborative filtering recommendation. Reference [4, 5, 6] applied extension strategy generation to independent travel, job-seeking problems and tenement respectively. Reference [7] did a research on software architecture and its application for ESGS, which helped people to carry out more effective innovation activities. Reference [8] applied extension strategy generation to air pollution prevention and cure, which made it possible to solve problems qualitatively and quantitatively. However, the above methods can't ratiocinate and classify original data, and consequently, putting decision tree classification into traditional extension strategy analysis methods is necessary.

Decision tree is a popular tool for classification and its strength is the small amount of required data and there is no need for people who want to use it to be equipped with any other professional knowledge[9].

As an important method used for analyzing data intelligently, decision tree classification has been widely researched recently. Reference [10] proposed classification of MISR multi-angle imagery based on decision tree classifier. Reference [11] proposed extraction of Hongze Lake Wetland information based on the decision tree. Reference [12] applied decision tree to the risk assessment of exotic medica1-vector. Reference [13] proposed decision tree based online stability assessment scheme for power systems with renewable generations. Reference [14] proposed decision tree-based preventive and corrective control applications for dynamic security enhancement in power systems. Reference [15] proposed decision tree-based detection of denial of service and command injection attacks on robotic vehicles. However, the above methods can't ratiocinate data dynamically and adaptively. As a result, it is necessary for researchers to combine it with extenics.

In order to analyze data intellectually, this paper absorbed the advantages of both extenics and decision tree classification and proposed an extension strategy analysis method based on decision tree classification: use basic-element theory to obtain data, use basic methods of extension to pre-process data, use extenics and decision tree classification to construct an extension decision tree, and use extension transformation to generate the ultimate strategies. This method is equipped with high practicability and operability that it is able to analyze data intellectually. It fills the gaps in both

Copyright © 2018, the Authors. Published by Atlantis Press.
This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

traditional extension strategy analysis methods and traditional decision tree classification, which makes it possible to promote the development of intellectual extension strategy analysis.

Data Obtaining and Extension Data Pre-processing

Data Obtaining. First, towards the experimental case, build a basic-element model. Using matter-element[1], it can be described as Eq. 1 and be abstracted as Eq. 2.

$$M\begin{bmatrix} O_m, c_{m1}, v_{m1} \\ c_{m2}, v_{m2} \\ \dots, \dots \\ c_{m2}, v_{m2} \end{bmatrix} = (O_m, c_m, v_m) \tag{1}$$

$$M_i=(O_i, c_j, v_{ij}) \tag{2}$$

In Eq. 2, i=1,2,3, … ,n; j=1,2,3, … ,m. O_i is the object. c_j is the characteristic of the object. v_{ij} is the value of c_j towards O_i. i and j are the object number and the characteristic number respectively. n and m are the quantity of objects and characteristics respectively. (ps: the definitions and the values of i, j, n, m are the same in the following sections if there are no new instructions given). c_j is obtained through basic-element divergence. O_i and v_{ij} are obtained through investigations.

Then, build an original data table according to the basic-element model.

Extension Data Pre-processing. First, use basic methods of extension[1] to analyze and evaluate the original data table. Next, determine and erase redundant data in it. Basic methods of extension include conjugation analysis, dependency analysis, implication analysis and scalability analysis. Here, dependency analysis and implication analysis are mainly used. Dependency analysis is used to analyze the dependency relationships between the characteristics in the original data table. Implication analysis is used to analyze the internal logic relationships between the characteristics in this table. According to the dependency information obtained by the former analysis and the implication information obtained by the latter analysis, choose a part of data in the original data table as redundant data to erase.

Then, combine the rest data and draw an optimized data table. Here, c_j transforms to the evaluation characteristic of the object and denotes as d_j.

The Construction of an Extension Decision Tree

Construct Dependent Function and Classify Statically. Towards evaluation characteristics d_j, its simple dependent function can be described as Eq. 3, its discrete dependent function can be described for example as Eq. 4 and its synthetic dependent function can be described as Eq. 5.

$$k_{ij}(v_{ij}) = \frac{v_{ij} \quad a_{j1}}{a_{j2}-a_{j3}} \tag{3}$$

$$k_{ij}(v_{ij}) = \begin{cases} 1, & v_{ij}=1 \\ 0.5, & v_{ij}=2 \\ 0, & v_{ij}=3 \\ -0.5, & v_{ij}=4 \\ -1, & v_{ij}=5 \end{cases} \tag{4}$$

$$K_{At} = \sum_{j=1}^{b} q_j k_{ij} \tag{5}$$

In Eq. 3, a_{j1} is the evaluation value of d_j and it is determined by people according to practical situations. a_{j2} is the maximum value of d_j. a_{j3} is the minimum value of d_j. In Eq. 5, $1 \leq b \leq m$. q_j is the

weight of d_j and it is determined by people according to practical situations. K_{Bi} can be gained in the same way and 4 situations can be listed as Eq. 6.

$$\begin{cases} K_{Ai} > 0, K_{Di} > 0; \\ K_{Ai} > 0, K_{Di} < 0; \\ K_{Ai} < 0, K_{Di} > 0; \\ K_{Ai} < 0, K_{Di} < 0; \end{cases} \tag{6}$$

According to these 4 situations, O_i can be statically classified into 4 categories: Class *A*, *B*, *C*, *D* for example. And a static classification table can be built.

Construct an Extension Decision Tree. First, according to the static classification table, calculate the information needed for constructing a decision tree:

$$\text{Info(D)} = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{7}$$

In Eq. 7, *i* is the category number of the static classification, for example, in this case, *i*=1, 2, 3, 4 on behalf of Class *A*, *B*, *C*, *D*. *n* is the quantity of the categories of the static classification, for example, in this case, *n*=4. p_i is the probability of Class *i* in this table.

Second, according to practical situations, the values of d_j can be classified artificially into 3 categories for example: [*a*,*b*)(low), [*b*,*c*)(middle), [*c*,*d*)(high). Here, *a*, *b*, *c*, *d* are constants and the information entropy of d_j can be calculated:

$$\text{Entropy}(d_j) = -\sum_{i=1}^{n} \left(\frac{a_{ij}}{a_j} \sum_{k=1}^{m} \frac{a_{ijk}}{a_{ij}} \log_2 \frac{a_{ijk}}{a_{ij}}\right) \tag{8}$$

In Eq. 8, *i* is the category number of the values of d_j, for example, in this case, *i*=1, 2, 3 on behalf of Class low, middle and high. *n* is the quantity of the categories of the values, for example, in this case, *n*=3. a_j is the quantity of the samples of d_j. a_{ij} is the quantity of the samples of d_j which belong to Class *i*. *k* is the category number of the static classification, for example, in this case, *k*=1, 2, 3, 4 on behalf of Class *A*, *B*, *C*, *D*. *m* is the quantity of the categories of the static classification, for example, in this case, *m*=4. a_{ijk} is the quantity of the samples of d_j which belong to both Class *i* and Class *k*.

Finally, calculate the information gain of d_j:

$$G(d_j)=\text{Info(D)}-\text{Entropy}(d_j) \tag{9}$$

Choose the d_j whose calculation result of $G(d_j)$ is the maximal as the root node and construct an extension decision tree, taking Fig. 1 for example.

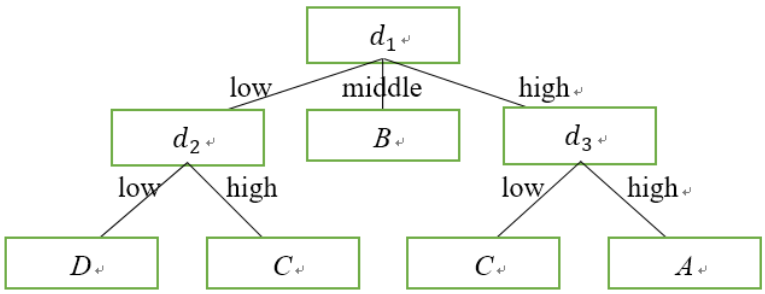

Fig. 1 The extension decision tree

Extension Strategy Analysis and Generation

The extension decision tree can be analyzed to get:

1) rule knowledge:

if(d_1 =low)and(d_2=low)then(category=D) if(d_1=high)and(d_3=low)then(category=C)
if(d_1=low)and(d_2=high)then(category=C) if(d_1=high)and(d_3=high)then(category=A)
if(d_1=middle)then(category=B)

2) condition transformation:
$T_1(d_1=\text{low})=(d_1=\text{middle})$ $T_5(d_1=\text{high})=(d_1=\text{low})$ $T_9(d_3=\text{low})=(d_3=\text{high})$
$T_2(d_1=\text{low})=(d_1=\text{high})$ $T_6(d_1=\text{high})=(d_1=\text{middle})$ $T_{10}(d_3=\text{high})=(d_3=\text{low})$
$T_3(d_1=\text{middle})=(d_1=\text{low})$ $T_7(d_2=\text{low})=(d_2=\text{high})$
$T_4(d_1=\text{middle})=(d_1=\text{high})$ $T_8(d_2=\text{high})=(d_2=\text{low})$

3) result transformation:
$T(A)=B$ $T(A)=D$ $T(B)=C$ $T(C)=A$ $T(C)=D$ $T(D)=B$
$T(A)=C$ $T(B)=A$ $T(B)=D$ $T(C)=B$ $T(D)=A$ $T(D)=C$

4) category transformation knowledge (including reversed category transformation knowledge, which isn't listed here):
$(d_1=\text{low})\text{and}(T_7(d_2=\text{low})=(d_2=\text{high}))\rightarrow T(D)=C$
$(d_1=\text{high})\text{and}(T_9(d_3=\text{low})=(d_3=\text{high}))\rightarrow T(C)=A$
$(d_3=\text{low})\text{and}(T_4(d_1=\text{middle})=(d_1=\text{high}))\rightarrow T(B)=C$
$(d_2=\text{high})\text{and}(T_2(d_1=\text{low})=(d_1=\text{high}))\rightarrow T(C)=A$
$(d_3=\text{high})\text{and}(T_4(d_1=\text{middle})=(d_1=\text{high}))\rightarrow T(B)=A$
$(d_2=\text{low})\text{and}(T_1(d_1=\text{low})=(d_1=\text{middle}))\rightarrow T(D)=B$
$(d_2=\text{high})\text{and}(T_3(d_1=\text{middle})=(d_1=\text{low}))\rightarrow T(B)=C$

According to the experimental case, choose some suitable strategies in the category transformation knowledge as the ultimate strategies.

An Experiment and Its Analysis

In this experiment, elderly people are the objects and denote as O_i. Here, 10 of them are selected, so i=1, 2, 3, … ,10. c_j is the characteristic of elderly people. Here, 7 characteristics are chosen according to some investigations and research and they are retirement circumstances (c_1), gender (c_2), living places (c_3), monthly pensions (c_4), pre-retirement work (c_5), age (c_6) and hobbies (c_7). The aim of this experiment is to improve the life quality of elderly people through the analysis method introduced above.

First, according to basic-element theory, obtain relevant data and build an original data table of elderly people:

Table 1 The original data table of elderly people

	c_1	c_2	c_3	c_4/¥	c_5	c_6	c_7
O_1	1	1	3	3500	1	62	3
O_2	3	1	2	4500	4	70	1
O_3	2	2	1	3300	3	65	2
O_4	2	1	5	5000	2	72	3
O_5	1	2	3	3200	1	80	1
O_6	2	2	4	6000	2	75	1
O_7	3	1	1	5500	4	64	2
O_8	3	2	3	4800	4	79	2
O_9	1	2	1	3600	1	82	3
O_{10}	2	1	2	4200	3	85	2

Second, pre-process the data in the original data table of elderly people using extension data pre-processing. According to dependency analysis, there are dependency relationships between retirement circumstances (c_1), pre-retirement work (c_5) and monthly pensions (c_4). So, c_1 and c_5

should be erased and c_4 should be reserved as it is more representative than the other two. According to implication analysis, there are internal logic relationships between gender (c_2), age (c_6) and elderly people (O_i): O_i and-implicate c_2 and c_6. What's more, c_2 and c_6 cannot change artificially in general. So, c_2 and c_6 should be erased. Combine the rest data and build an optimized data table of elderly people. Here, d_1, d_2 and d_3 refer to living places, monthly pensions and hobbies respectively.

Table 2 The optimized data table of elderly people

	d_1	d_2/ ¥	d_3
O_1	3	3500	3
O_2	2	4500	1
O_3	1	3300	2
O_4	5	5000	3
O_5	3	3200	1
O_6	4	6000	1
O_7	1	5500	2
O_8	3	4800	2
O_9	1	3600	3
O_{10}	2	4200	2

Then, according to practical situations, determine that the evaluation values of d_1 and d_2 are 2 and 4000 respectively and construct dependent function:

$$k_{i1}(v_{i1}) = \frac{v_{i1}-2}{5-1} \tag{10}$$

$$k_{i2}(v_{i2}) = \frac{v_{i2}-4000}{6000-3200} \tag{11}$$

$$k_{i3}(v_{i3}) = \begin{cases} -1, & v_{i3} = 1 \\ 0, & v_{i3} = 2 \\ 1, & v_{i3} = 3 \end{cases} \tag{12}$$

Determine that the weights of d_1 and d_3 are 0.6 and 0.4 respectively and construct synthetic dependent function of spiritual life as Eq. 13 and of material life as Eq. 14.

$$K_{Si}=0.6k_{i1}+0.4k_{i3} \tag{13}$$

$$K_{Mi}=k_{i2} \tag{14}$$

Determine classification rules:

1) While $K_{Si} \geq 0$ and $K_{Mi} \geq 0$, the elderly people O_i has high quality of life (abbreviation: HQ);
2) While $K_{Si} > 0$ and $K_{Mi} < 0$, the elderly people O_i has low quality of material life (abbreviation: LM);
3) While $K_{Si} < 0$ and $K_{Mi} > 0$, the elderly people O_i has low quality of spiritual life (abbreviation: LS);
4) While $K_{Si} < 0$ and $K_{Mi} < 0$, the elderly people O_i has low quality of life (abbreviation: LQ);

And build a static classification table of elderly people's life quality:

Table 3 The static classification table of elderly people's life quality

	d_1	d_2/¥	d_3	K_S	K_M	categories
O_1	3	3500	3	0.55	-0.17857	LM
O_2	2	4500	1	0	0.178571	HQ
O_3	1	3300	2	-0.15	-0.25	LQ
O_4	5	5000	3	0.85	0.357143	HQ
O_5	3	3200	1	0.15	-0.28571	LM
O_6	4	6000	1	-0.1	0.714286	LS
O_7	1	5500	2	-0.15	0.535714	LS
O_8	3	4800	2	0.55	0.285714	HQ
O_9	1	3600	3	0.25	-0.14286	LM
O_{10}	2	4200	2	0	0.071429	HQ

Determine that d_1={1 , 2} refers to low, d_1={3 , 4 , 5} refers to high ; d_2=[3000,4000) refers to low, d_2= [4000,5000) refers to middle, d_2= [5000,6000] refers to high ; d_3=1, 2, 3 refers to low, middle, high respectively. Calculate information, information entropy and information gain to get that:

Info(D)=1.846 Entropy(d_1)=1.722 Entropy(d_2)=0.6 Entropy(d_3)=1.351

G(d_1)= Info(D)-Entropy(d_1)=0.124 G(d_2)= Info(D)-Entropy(d_2)=1.246

G(d_3)= Info(D)-Entropy(d_3)=0.495

So, choose d_2 as the root node and build an extension decision tree of elderly people's life quality:

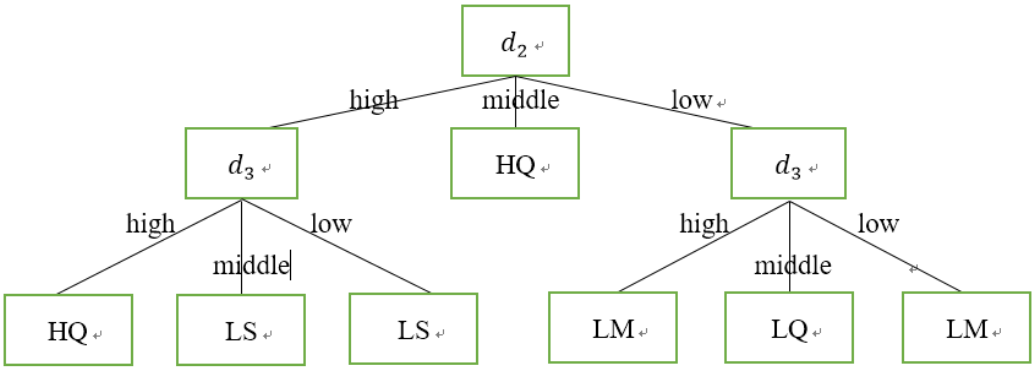

Fig. 2 The extension decision tree of elderly people's life quality

According to extension strategy analysis, the life quality of elderly people has nothing to do with living places (d_1) and is relevant to monthly pensions (d_2) and hobbies (d_3). To improve the life quality of the elderly people who have low quality of spiritual life (LS), finding and developing their hobbies are needed. And to improve the life quality of the elderly people who have low quality of material life or low quality of life (LM or LQ), increasing their monthly pensions is needed.

The experiment results show that the extension strategy analysis method based on decision tree classification has high operability and practicability, which fully combines both extenics and decision tree classification and provides the most value of them and also promotes their development.

Conclusions

The extension strategy analysis method based on decision tree classification is a wonderful combination of both extenics and decision tree classification because it brings their superiority into

full play. As a result, it is a scientific strategy analysis method which is not only practical, but also reliable.

This paper starts from the introduction of data obtaining and extension data pre-processing, fully expounds the construction of an extension decision tree and the details of extension strategy analysis and generation. At last, this paper use an experimental case to prove the practicability and reliability of this method.

In the future, a large amount of research towards the extension strategy analysis method based on decision tree classification is still needed. For example, the optimization of the construction of an extension decision tree and the pruning of it, the processing of fuzzy and incomplete data, and the optimization of extension strategy generation. These will be the main part of the further research.

Acknowledgements

This work was financially supported by National Science Foundation Project of Zhejiang of China (Grant: LY16G010010, LY18F020001) and Ningbo Innovative Team: The intelligent big data engineering application for life and health (Grant: 2016C11024).

References

[1] Xu, L., Tian, Y., Florentin S., & Rajan A. (2015). An extension collaborative innovation model in the context of big data. International Journal of Information Technology & Decision Making, 14 (1), 1-23.

[2] Xu, L., Pan, X., Yuan, P., et al. (2017). Knowledge innovation by intelligent emergence: concept, framework and its pathway. CAAI Transactions on Intelligent Systems, 12(1), 47-54.

[3] Xu, L. B., Li, X. S., & Guo, Y. (2018). Gauss-core extension dependent prediction algorithm for collaborative filtering recommendation. Cluster Computing.

[4] Fang, Z., Li, W., & Li, C. (2009). Research and realization of extension strategy generating system for independent travel. Journal of Guangdong University of Technology, 26(2), 83-89.

[5] Chen, Y., & Li, W. (2012). Research on the extension strategy generating system for job-seeking problems. Journal of Guangdong University of Technology, 29(1), 88-93.

[6] Li, C., & Li, W. (2011). Research on a tenement extension strategy generation system. CAAI Transactions on Intelligent Systems, 6(3), 272-278.

[7] Fan, R., Yan, S., Peng, Z., Liao, Y., Chen, Y., Luo, X., Lin, H., & Tan, Z. (2011). A research on software architecture and its application for ESGS. Journal of Guangdong University of Technology, 6(3), 272-278.

[8] Ye, G., Li, W., & Zhang, X. (2007). Research and realization of extension strategy generating system for air pollution prevention and cure. Journal of Guangdong University of Technology, 24(4), 42-48.

[9] Guo, Q., & Zou, G. (2017). Prediction methods for extension architecture programming based on decision tree classification. CAAI Transactions on Intelligent Systems, 12(1), 117-123.

[10] Yang, X., & Wang, X. (2016). Classification of MISR multi-angle imagery based on decision tree classifier. Journal of Geo-information Science, 18(3), 416-422.

[11] Zhang, L., & Ruan, R. (2015). Extraction of Hongze Lake Wetland information based on the decision tree. Geomatics & Spatial Information Technology, 38(2), 87-91.

[12] Qiu, J., Sun, Z., Wang, J., Zheng, J., & Yang, D. (2016). The application of decision tree on the risk assessment of exotic medica1-vector. Chin J Hyg Insect & Equip, 22(2), 137-144.

[13] Tong, W., Bi, T., Wang, H., & Liu, J. (2015). Decision tree based online stability assessment scheme for power systems with renewable generations. CSEE Journal of Power and Energy Systems, 1(2), 53-61.

[14] Istemihan, G., Ruisheng, D., Vijay, V., Sharma, K., & Sujit, M. (2010). Decision tree-based preventive and corrective control applications for dynamic security enhancement in power systems. IEEE Transactions on Power Systems, 25(3), 1611-1619.

[15]Tuan, P. V., George, L., Diane, G., & Anatolij, B. (2016). Decision tree-based detection of denial of service and command injection attacks on robotic vehicles. IEEE International Workshop on Information Forensics & Security, 1-6.