CASE-BASED REASONING WITH ROUGH FEATURES

Hoang Xuan Huan

Faculty of Information Technology, College of Technologies, VNU

Abstract. Case-based reasoning (CBR) is an important analytical method in the knowledge engineering to support decision making. In many problems, a feature value may not be certainly determined but varies in an interval $[\underline{v}, \overline{v}]$ that its distribution is not known. In this paper, by proposing the notion of rough feature, we introduce an approach to solve these situations.

1. Introduction

Case-based reasoning (CBR) and rule-based reasoning (RBR) are two important methods in the knowledge engineering to support decision making (see [2, 7, 8, and 9]) in decision support systems (DSS). In rule-based reasoning the computer examines historical cases and generates rules, which are chained (forward or backward) to solve problems. Case-based reasoning, on the other hand, follows a different process, it finds those cases in memory that shoved problems similar to the current problem, and then adapts the previous solutions to fit the current problem by taking in to account any difference between the current and previous situations. As the main obstacle of RBR method of generating rules from experiments limited its capacity of application, then the CBR proved to be an extremely effective approach in complex case (see [9]).

In order to find similar cases, we have to assign appropriate features to each case. In many problems, a feature value may be uncertainly determined. For example, a number feature r may belong to an interval [a, b] and its distribution can not be known. Likes other areas (see [5]), this situation demands us to have another approach to extend the scope of application.

Basing on rough set theory [6] and interval algebra [1] Lingras [3, 4] proposed the concept of rough patterns, which are based on the notion of rough numbers (called rough values).A rough number consists of an upper and a lower bound which can be used to represent a range of values for variables such that daily temperature or daily financial indicators. By extending the Lingras's notion of rough values, in this paper, we propose an approach to use CBR in the case that the features are determined uncertainly.

Except the conclusion section, this paper is organized as follows. The basic notions and characters of DSS and CBR are briefly presented in section 2. A model of CBR with rough features and an example of its application to health care are presented in section 3.

2. Decision support systems and case-based reasoning

In this section, we first sketch of how is a DSS, and then give an overview of CBR (in detail can see [8, 9]).

2.1. What is a DSS?

Decision support systems assist decision making by combining data, analytical models and user-friendly software into a single powerful system that can support semistructured or unstructured decision making

In general, a typical DSS consist of following components:

♦ Database and management system. The DSS database is a collection of current or historical data from a number of applications or groups which is organized for easy access by a range of application

♦ Model base and management system. A model base is a collection of mathematical or analytical models that can be made accessible to the DSS user.

♦ Knowledge base. It is a collection of rules to solve relational problems in cases that model of rule-based reasoning is used.

♦ Dialog management system and users. The dialog management provides a graphical, easy-to-use, flexible user interface that supports the dialog between the user and DSS.

A conceptual model of a typical DSS can be described in figure 1.

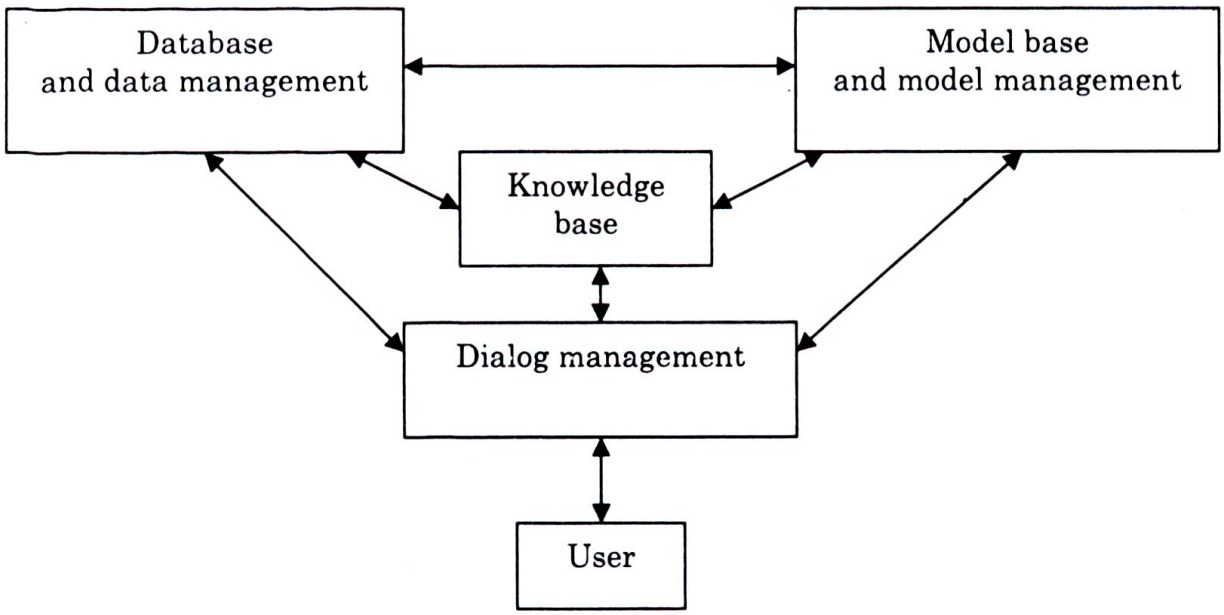

Figure 1. Conceptual model of DSS

2.2. Case-based reasoning

As mentioned above, analytical model may be case-based or rule-based reasoning. The basic idea of **case-based reasoning** is to adapt solutions that were used to solve old problems and use them for solving new problems. A case is a contextualized piece of

knowledge representing an experience. It contains the past lesson that is the content of the case and context in which the lesson can be used. A case can be account of an event, a story, or some record typically comprising

- The problem that describes the state of the word when the case occurred
- The solution that states the derived solution to that problem.

A collection of cases stored in database is called a case base. When we meet a new problem, case- base reasoning is following process. It

- finds those cases in the case base that solved problems are similar to the current problem, and
- adapt the previous solution or solutions to fit the current problem, taking into account any difference between the current and previous situations.

In order to facilitate of finding similar cases, each case (or problem) is characterized by assigning appropriate features, and the feature space is equipped a metric that defines the similarity of cases. A tuple of characteristic features for a case is called a case feature. In general, a case feature is an element of a relation r on a relational schema $R = \{A_1,...,A_n\}$, and feature space (also denoted by R) is the Cartesian product of value domains of attributes $A_1,...,A_n$.

Structure of the process of CBR is illustrated in figure 2.

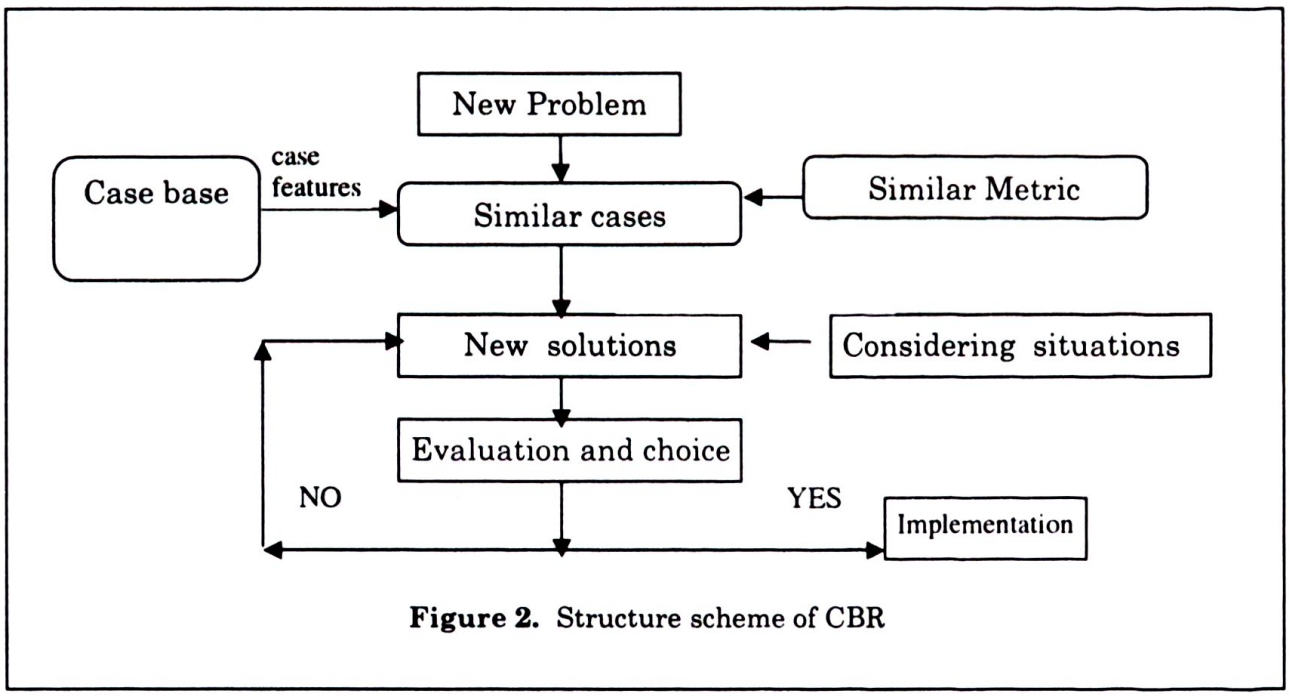

Figure 2. Structure scheme of CBR

The procedure CBR of finding solution for a new case can be specified in figure 3.

Procedure CBR

Begin

feature extraction;// extracting the case feature for the new case;

find similar case features;//find one or k most similar case features

recall similar cases;// recall cases corresponding to found case features.

building new solution;// in comparing the solutions of old cases and new situations.

end.

Figure 3. Specification of CBR

3. CBR with rough feature

In this section we first present definition of rough features, then a model for case-based reasoning in cases that have rough feature.

3.1. Rough feature

In the feature space, a domain of attributes may be a categorical, ordered categorical or number set.

Definition 3.1.1. Let $DOM(A_j)$ be the domain of attribute A_j. We define the following notions.

i) Categorical attribute. A_j is called a categorical attribute if $DOM(A_j)$ is unordered, e.g., for any values $a,b \in DOM(A_j)$, either $a = b$ or $a \neq b$.

ii) Real attribute. A_j is called a real attribute if $DOM(A_j)$ is a set of real numbers.

ii) Ordered attribute. In case that A_j is not a real attribute then it is called an ordered attribute if $DOM(A_j)$ is a finite and totally ordered set.

For example, a categorical attribute may be virus kind (which infects patients) as: virus A, B,C for a liver patient; an ordered attribute may be body state as: no pain, little pain, pain or very pain; a real attribute may be body temperature.

. In many problems, a feature value may not be certainly determined but varies in an interval $[\underline{v}, \bar{v}]$ that its distribution is not known. For example, in a determined period, body temperature of a patient varies in interval [38°C, 39°C] and its distribution isn't known. In this case, we can not take average value to replace this interval because they are quite different. This situation demands us to have rough approach to extend the scope of application.

Definition 3.1.2(rough feature). A rough feature (in feature space with respect to relational schema $R= \{ A_1,...,A_n\}$) is a tuple $x = (x_1, x_2,...,x_n)$ such that

i) if A_j is a categorical attribute then

$$x_j \in DOM(A_j), \tag{1}$$

ii) if A_j is a ordered or real attribute then

$$x_j = (\underline{x}_j, \overline{x}_j), \tag{2}$$

where $\underline{x}_j, \overline{x}_j \in DOM(A_j)$ and $\underline{x}_j \leq \overline{x}_j$ (with respect to the ordering relation in A_j); $\underline{x}_j$ is the lower bound and $\overline{x}_j$ is the upper bound of x_j. A value x_j taken by (2) is called *rough value* or *rough number* (corresponding to ordered or real attribute).

It is obvious that a precise case feature presented in section 2 also is a rough feature by taking

$$\underline{x}_j = \overline{x}_j = x_j \tag{3}$$

for every value x_j of characteristic feature in attribute A_j.

Now, we introduce a model for case-base reasoning with rough feature which is denoted by RCBR.

3.2. Model for RCBR

We first define a metric on rough feature space, then RCBR is performed in the same CBR framework mentioned in section 2.

In order to determine distance between rough features x and y in rough feature space with respect to relational schema $R = \{A_1,...,A_n\}$ we have to define distance between two values in domain of attributes.

Distance between rough numbers. Let $x_j = (\underline{x}_j, \overline{x}_j)$ and $y_j = (\underline{y}_j, \overline{y}_j)$ be rough numbers in domain of a real attribute A_j. The distance $d(x_j,y_j)$ between them is defined as:

$$d(x_j,y_j) = |\underline{x}_j - \underline{y}_j| + |\overline{x}_j - \overline{y}_j| \tag{4}$$

Distance between rough values. Let A_j be a ordered attribute and $DOM(A_j) = \{a_j^1,...,a_j^k\}$ where $a_j^1 < a_j^2 < ... < a_j^k$. We take a monotone function $f_j : DOM(A_j) \rightarrow [0,1]$ such that $f_j(a_j^1) = 0; f_j(a_j^k) = 1$ (this function can be chosen by an expert). The distance between two rough values $x_j = (\underline{x}_j, \overline{x}_j)$ and $y_j = (\underline{y}_j, \overline{y}_j)$ in $DOM(A_j)$ is defined as:

$$d(x_j,y_j) = |f_j(\underline{x}_j) - f_j(\underline{y}_j)| + |f_j(\overline{x}_j) - f_j(\overline{y}_j)|. \tag{5}$$

The function f_j is called scale function of attribute A_j.

Distance between categorical values.

Let A_j be a categorical attribute and $x_j, y_j \in DOM(A_j)$. The distance $d(x_j,y_j)$ is defined as:

$$d(x_j, y_j) = \begin{cases} 0 & if: x_j = y_j \\ 1 & if: x_j \neq y_j \end{cases}. \tag{6}$$

Now, we can give the definition for metric on rough feature space.

Distance between rough features. The distance between rough features $x = (x_1,...,x_n)$ and $y = (y_1,...,y_n)$ can be calculated as:

$$d(x,y) = \sum_{j=1}^{n} \rho_j d(x_j, y_j) \qquad (7)$$

where, for all j, $d(x_j,y_j)$ are calculated by equations (4-6) and coefficients ρ_j are positive weight chosen by experts.

The procedure RCBR can be specified as those of CBR in figure 3.

In order to illustrate this model, we introduce an example of building DSS for health care, (presented feature space is assumed).

An example.

Suppose that we design a DSS to help doctors to finds treatment schemes for a new patient. In order to build a convenient treatment scheme for a new patient, doctors first search in clinical records of old patients to find similar cases, then, by referring to treatments for found cases build a convenient one for new patient.

In this DSS, a case is a treatment scheme which is written in a clinical record collected in data base. Case feature can be disease, symptoms and bio-chemical indexes of patient in treating periods. Case feature for liver clinical scheme may consist of infected virus kind, sex, weight, enzyme indexes such that :AST; ALT;GGT; cholesterol; glucose, liver size, sensation. The relational schema of feature space is following

Infected virus	sex	weight	AST	ALT	GGT	cholesterol	glucose	liver size	Sensation

Ranging from lower to upper bound, DOM(liver size) may be {normal, level 1, level 2,..., level k} and DOM(sensation) may be {no pain, little pain, pain, very pain}. In this relational schema, infected virus and sex attributes are categorical, weight; indexes of {AST; ALT, GGT, cholesterol, glucose} are real and liver size; sensation are ordered attributes. Case features for three patients are showed in table 1. In this table, we use the following abbreviations: M= male; F = female; L = level ; lp= little pain; vp = very pain and a\b = (a,b) for a rough value or rough number.

Table 1

patient	Infected virus	sex	weight	AST	ALT	GGT	cholesterol	glucose	Liver size	sensation
HHT	A	M	54\56	30\35	28\34	40\45	4,2\4,8	5\5,2	L1\l2	lp\pain
NLA	B	F	48\50	36\40	34\37	20\25	4,6\5,2	6\7,2	L1\L3	lp\lp
NTT	A	M	57\60	38\50	35\38	24\28	4,4\5,8	6\6,5	L2\L3	Lp\vp

Scale function of liver size attribute can be defined as: $f_{ls}(\text{level j}) = \frac{j-1}{k-1}$,and scale function of sensation attribute may be defined as $f_s(a_j) = \frac{j-1}{3}$ where a_j is the value at order j in the ordering of DOM(sensation). Coefficients ρ_j in equation (6) will be chosen by doctors.

4. Conclusion

In applications of CBR methods, we meet an obstacle when feature value can not be certainly determined. In many cases, other processing approachs are not convenient. For these problems, we propose rough feature notion and develope the RCBR methods to solve them. By sketching out an application example, we hope that this framework can be effectively applied to other problems.

REFERENCES

1. Allen J.F., Maintaining knowledge about Temporal Interval, *Communication of ACM.26* (1983), 832-834.
2. Burn-Thornton K.E., Making the most of data in order to Provide accurate clinical decision support systems for the use in the determination of heart disease, *Proceeding of third international conference on data mining ,Bologna 2002*(Data mining 2002), 839-848.
3. Lingras P., Rough neural networks, *proceeding of sixth international conference on Information processing and management of uncertainty in knowledge-based systems*, Grenada Spain (1996), 1445-1450.
4. Lingras P., Application of rough pattern, *Rough set in data mining and knowledge discovery 2, series soft computing, Physical Velag (Springer)*, vol2 (1998),369-384.
5. Ohsuga S., to what extent can computers? Toward second phase information technology, *proceeding of second International Conference of RSCTC, Banff, Canada*, October 16-19 (2000), 8-29.
6. Pawlak Z., *Rough Sets Theoretical aspects of reasoning about data*, Kluwer Academic Publishers, 1991, 229p.
7. Riha A.,Svatek V, Nemec P , Zvarova J., Medical guideline as prior knowledge in electronic health care record mining, *Proceeding of third international conference on data mining*, Bologna 2002(Data mining 2002), 809-818.
8. Turban E., *Decision support and expert systems Management support systems*, Prentice hall, 1995, 887p.
9. Watson I., *Applying Case-based reasoning: Techniques for Enterprise Systems*, Mogan Kaufmann Publishers, Inc San Francisco, California, 1997, 289p.

TẠP CHÍ KHOA HỌC ĐHQGHN, KHTN & CN, T.XXI, Số 1, 2005

LẬP LUẬN DỰA TRÊN TÌNH HUỐNG VỚI ĐẶC TRƯNG THÔ

Hoàng Xuân Huấn

Khoa Công nghệ Thông tin, Trường Đại học Công nghệ, VNU

Lập luận dựa trên tình huống là một cách tiếp cận quan trọng để xử lý hỗ trợ ra quyết định trong các DSS nhờ tham khảo kinh nghiệm đã có. Trong nhiều trường hợp, các giá trị đặc trưng không xác định đúng được mà chỉ biết khoảng [a,b] mà mỗi giá trị thay đổi trên đó. Khi đó các tiếp cận thống kê thường kém hiệu quả. Trong bài này chúng tôi giới thiệu khái niệm đặc trưng thô và phát triển một mô hình lập luận dựa trên tình huống với kiểu đặc trưng này để khắc phục khó khăn đã nêu.