Journal of University of Science and Technology Beijing
Volume 9, Number 3, June 2002, Page 237

Information

A new knowledge discovery method for scientific and technologic database

Dezheng Zhang[1], *Bingru Yang*[1], *and Lianying Sun*[2]

1) Information Engineering School, University of Science and Technology Beijing, Beijing 100083, China
2) China University of Mining and Technology, Beijing 100083, China
(Received 2001-11-28)

Abstract: A new algorithm for the knowledge discovery based on statistic induction logic is proposed, and the validity of the method is verified by examples. The method is suitable for a large range of knowledge discovery applications in the studying of causal relation, uncertainty knowledge acquisition and principal factors analyzing. The language field description of the state space makes the algorithm robust in the adaptation with easier understandable results, which are isomotopy with natural language in the topologic space.

Key words: knowledge discovery; statistic induction; fuzzy language field

[*This work was financially supported by the National Natural Science Foundation of China (No. 69835001).*]

1 Introduction

Knowledge discovery in database (KDD) is the process of extracting the novel and potentially useful knowledge from a very large database [1-3]. In the database of science and engineering, there are lots of hidden knowledge regarding the state and state changing of objects in their state spaces, such as trend and ratio of change *etc.*, which represent the state, the variation and the causal relationship of the objects and their attributes. They are the kernel or deep level knowledge, and useful for analyzing the mechanism of the development of objects in the real world. The knowledge discovery algorithm of association rule proposed by Agrawa [4] is in the logic form of $X \Rightarrow Y$, showing a general association between attributes in a database. The shortcomings of the algorithm are: (1) difficult to represent the knowledge of the states change; (2) too abstract to explain the causal relationship among the attributes. Therefore, for the knowledge discovery in science and technology database, a new method of KDD should be d eveloped to meet the needs of trend analysis, predication and forecasting.

2 Fuzzy language and fuzzy language value structure

The knowledge of the state and variation of state of objects in the state space are normally represented in the attributes and records, which are qualitative or quantitative representations of the complexity, abstraction and uncertainty of the objects' properties and behaviors. The theory of fuzzy set language field and language structure provide a fusing technique and a template for describing the complicated relation and structure, which are the knowledge people want to acquire [5].

Suppose X is the fuzzy language variable, which describes the state or state changing.

Definition 1 Given two real intervals L_1 and L_2, if either L_1 or L_2 is a subinterval of the other, and $L_1 \cap L_2 \neq \Phi$, then call L_1 and L_2 the overlapping interval pair.

Definition 2 Given a sequence of n real intervals, if every two adjacent intervals are overlapping interval pair, then call the sequence an overlapping interval sequence.

Obviously, all the corresponding base variable intervals of fuzzy language value X (in real domain) compose an overlapping interval sequence.

Definition 3 To set D consisting of n real intervals that may compose an overlapping interval sequence, the binary relation $\prec$ is defined as: to any two intervals $[x_1, x_2] \in D$ and $[y_1, y_2] \in D$, can get

$$[x_1, x_2] \prec [y_1, y_2] \Leftrightarrow (x_1 \leq y_2) \wedge (x_2 \leq y_2) \quad (1)$$

Theorem 1 The binary relation $\prec$ defined on D is a complete ordering relation.

Definition 4 In the corresponding base variable region of fuzzy language variables, the dots in the middle of every overlapping subinterval, ξ, and its adjacent region ε (ε is generally the tolerance error) are called the

Corresponding author: Dezheng Zhang **E-mail**:zhangdezheng@hotmail.com

standard samples (dots), the interval $(\xi-\varepsilon, \xi+\varepsilon)$ is called standard values; any other dots are called nonstandard samples (dots); they are called standard sample space and nonstandard sample space, respectively.

Definition 5 $l=\langle B, I, N, \prec_N\rangle$, if the following are satisfied:

(1) B is a set of all overlapping intervals of base variable region on R;

(2) $N\neq\phi$ is a finite set of fuzzy language value;

(3) $\prec_N$ is a complete ordering relation on N;

(4) $I: N\rightarrow B$ is a standard value mapping, and satisfies isotonicity.

Then l is called a fuzzy language field.

Definition 6 For the fuzzy language field $l=\langle B, I, N, \prec_N\rangle$, $F=\langle l, W, K\rangle$ is a fuzzy language value structure of l, if

(1) l satisfies definition 5;

(2) K is a natural number;

(3) $W: N\rightarrow[0, 1]^K$, it satisfies the following:

$\forall n_1, n_2\in N\ (n_1\leq_N n_2\rightarrow W(n_1)\prec_{dic} W(n_2))$,

$\forall n_1, n_2\in N(n_1\neq n_2\rightarrow W(n_1)\neq W(n_2))$,

in which, $\prec_{dic}$ is a lexicographic order in $[0, 1]^K$.

In F, the K-dimensional vector corresponding to the standard value in the subinterval of base variable region of every language value is called standard vector; otherwise it called nonstandard vector.

Definition 7 Given two fuzzy language fields l_1 and l_2, say that is an expansion of l_2, if there is a 1-1mapping $f: B_1\rightarrow B_2$, $g: N_1\rightarrow N_2$, satisfying

(1) f is monotonous;

(2) $\forall n_1\in N_1\ (f(I_1(n_1))=I_2(g(n_1)))$

in which $l_1=\langle B_1, I_1, N_1, \prec_{N_1}\rangle$, $l_2=\langle B_2, I_2, N_2, \prec_{N_2}\rangle$.

Definition 8 Given fuzzy language value structures $F_1=\langle l, W_1, K_1\rangle$ and $F_2=\langle l, W_2, K_2\rangle$ of $l=\langle B, I, N, \prec_N\rangle$, if there is a 1-1 mapping h: $[0, 1]^{K_1}\rightarrow[0, 1]^{K_2}$ that satisfies

(1) f is strictly monotonous in lexicography;

(2) $\forall n\in N, f(W_1(n))=W_2(n))$;

(3) $(\exists\varepsilon\in R)\ (\forall n,\ n'\in N)\ (\mathrm{dis}_1(W_1(n),\ W_1(n'))=\varepsilon\cdot \mathrm{dis}_2(W_2(n), W_2(n')))$;

in which dis_1: $[0, 1]^{K_1}\times[0, 1]^{K_1}\rightarrow[0, 1]$, dis_2: $[0,1]^{K_2}\times[0, 1]^{K_2}$; then, call F_1 and F_2 are $(\mathrm{dis}_1, \mathrm{dis}_2)$-isomorphic (the abbreviation is "dis-isomorphism").

Theorem 2 (expansion theorem) Given two fuzzy language fields l_1 and l_2, l_1 is an expansion of l_2, if l_1 and l_2 are the same-type (that is, $|N_1=N_2|$) language field.

Theorem 3 (dis-isomorphism theorem) Suppose that F is a fuzzy language value structure of F_{double}, then F and F_{double} (the double-extension of F) are dis-isomorphic under the weighted Hamming distance.

Since the same fuzzy language fields are not distinguished from each other in the expansion sense, the language fields can be described based on the language value of natural number such as "large", "small" and so on, and fuzzy language value structure can be built on different dimension spaces in the dis-isomorphism sense. The discrete type vector corresponding to each fuzzy language value can be chosen according to the application.

3 State or state changing knowledge discovery based on statistic induction

The statistic induction is one of the important techniques of the inductive logic. The knowledge discovery based on the statistic induction (KDBSI) can be used to discover the abstract pattern, concept and *etc*. This is a process from concrete to the integration and from specific to abstract. Therefore, the statistic induction can be defined as a process of extracting the samples having attributes ϕ with probability p from a set s, that is if a is a random element of ϕ then the probability of $\phi(a)$ is p [6].

$$\phi(x)/R(x,s,\phi,h)\cdot h=p \tag{2}$$

where $\phi(x)$ is the set of characteristics, s is the sample set, h is the facts or evidences, x is the random samples with characteristics ϕ in s, and p is the probability. If there are samples with different n attributes, such as $\phi_1(x)$, $\phi_2(x)$, $\cdots$, $\phi_n(x)$, and put them in m observation groups to study the frequency, f_{ij} $(i=1, 2, \cdots, n; j=1, 2, \cdots m)$ of satisfying $\phi_1(x)$, $\phi_2(x)$, $\cdots$, $\phi_n(x)$ at the same time. If f_{ij} is larger than the given threshold, in the m observation groups, $\phi_i(x)$ is inductively associated with $\phi_j(x)$. This relationship is the knowledge people want to discover. The inductive association is a kind of hypothesis rule, the metric of relationship of attributes.

The verification of the hypothesis is based on the stability evaluation, which is the base or precondition of statistic induction knowledge discovery. Therefore, should verify the stability of the rules, which can be measured with the value of Q.

$$Q=R/r,\ R=\rho\sqrt{\frac{2[\delta^2]}{n-1}},\ r=\rho\sqrt{\frac{2v(1-v)}{g}}$$

Where ρ is a constant, v is the basic probability, g is the number of samples, $[\delta^2]$ is the mean of the squares of

frequency difference for each sample, n is the number of samples. If the samples from s_1 or s_2 are randomly chosen from a field u, the stability defined as: $Q=1$ is stable; $Q<1$ is sub-stable; $Q>1$ is unstable. If $Q=1$, the hypothesis rules are verified, and can be used as knowledge.

As to the driven mechanism, there are two different processes: hypothesis rule verification driven mechanism and discovery driven mechanism. The discovery driven mechanism means that the hidden knowledge is discovered automatically. The verification mechanism is completed through the verification of hypothesis rules in the database, and the hypothesis rule is determined by human experts according to their field knowledge. This process is helpful to discover the significant, useful and understandable knowledge. In addition, the verification mechanism can accelerate the speed of knowledge discovery, and is more efficient for the discovering of deep level knowledge.

In a transaction of T, $I=\{i_1, i_2, \cdots, i_m\}$ is the set of attributes, and the hypothesis rule is $\boldsymbol{X}\Rightarrow\boldsymbol{Y}$, $\boldsymbol{X}\subset I$, $\boldsymbol{Y}\subset I$ and $\boldsymbol{X}\cap\boldsymbol{Y}=\phi$. The attributes can be mapped to a standard vector with the language field and structure, for example, $\boldsymbol{X}=(x_1, x_2, \cdots, x_n)$, $\boldsymbol{Y}=(y_1, y_2, \cdots, y_n)$, have the ordering pairs $\langle x_i, y_j\rangle$, $(i=1, 2, \cdots, n; j=1, 2, \cdots, m)$, therefore, the association of state and state changing between $\boldsymbol{X}$ and $\boldsymbol{Y}$ can be presented by $H: x_i\rightarrow y_j$.

Definition 9 The support of the state (and state changing) association rule $H: x_i\rightarrow y_j$ is the probability of x_i and y_j in database occurring at same time, namely Support($x_i\rightarrow y_j$).

Definition 10 Intensity of the state (changing) association rules H_s describes the association degree between attributes, and is defined as: $H_s=\text{Support}(x_i\rightarrow y_j)/\text{Support}(x_i\rightarrow y_j)$.

Definition 11 The confidence of rules, CF, can be described by confidence factor in uncertain induction, namely

$$\text{CF}(H|e)=\text{MB}(H|e)-\text{MD}(H|e) \tag{3}$$

where MB($H|e$) and MD($H|e$) represent the increase of confidence of rule H induced by evidence e.

Suppose P(H) is the prior probability of rule H, P($H|e$) describes probability of H occurring under the condition e, then

$$\text{MB}(H|e)=\begin{cases}0, & \text{P}(H)\geq\text{P}(H|e)\\ \dfrac{\text{P}(H|e)-\text{P}(H)}{\text{P}(H)}, & \text{P}(H)<\text{P}(H|e)\\ 1, & \text{P}(H)=0\end{cases} \tag{4}$$

$$\text{MD}(H|e)=\begin{cases}0, & \text{P}(H)\leq\text{P}(H|e)\\ \dfrac{\text{P}(H|e)-\text{P}(H)}{\text{P}(H)}, & \text{P}(H)>\text{P}(H|e)\\ 1, & \text{P}(H)=0\end{cases} \tag{5}$$

The threshold of support and confidence is S_c and C_c separately, which are compared with the calculating results. If the support and confidence of hypothesis rule surpass the given threshold, then this rule is proven and accepted, otherwise cancel it.

4 Algorithm and test

Based on the theory of language field and statistic induction a new algorithm of KDD can be summarized as: (1) Mapping the real world state into the standard state space; (2) discovering the hypothesis rules through the statistic inductive logic; (3) the verification of hypothesis rules.

4.1 Mapping between the real state and language vector

The first step of the algorithm is mapping the attributes of a real database into a standard vector by structure transformation of language field and language value to create a database (D') used for knowledge discovery. For a single language field, $P=\{\langle t_k, s_k\rangle\}$ is used to describe sample value t_k and s_k in the state space and state changing space. The mapping can be implemented by the following equation, and can get a state or a state changing value (a_t).

$$a_t=\boldsymbol{A}_i\cdot\left(1-\frac{|t_i-t_{i0}|}{t_i}\right)+\boldsymbol{A}_{\text{neighbor}}\cdot\frac{|t_i-t_{i0}|}{l_i} \tag{6}$$

Where t_i is the input data of the i^{th} interval, t_{i0} is the midpoint of the i^{th} interval, l is the length of the interval, $\boldsymbol{A}_i$ is the state (or state changing) standard vector, and $\boldsymbol{A}_{\text{neighbor}}$ is a adjacent standard vector of state (or state changing) in the left or the right determined by the position of t, then can get a_t.

Determine the type of the state (or state changing) vector a_t, such as $\boldsymbol{A}_i$ (k=1,2,3,4,5) calculating the metric, d_H, between a_t and each standard vector $\boldsymbol{A}_i$, selecting the minimum as the state (or state changing) type of a_t.

$$d_H(a_{tw}, \boldsymbol{A}_i)=\sum_{j=1}^{5}|\mu a_{tw}^{(j)}-\mu\boldsymbol{A}_i^{(j)}| \tag{7}$$

Where, $\mu a_t^{(j)}$ and $\mu\boldsymbol{A}_i^{(j)}$ are the elements of the corresponding vectors, respectively.

4.2 Knowledge discovery based on statistic induction

The form of discovered knowledge with the statistic induction is a hypothesis rule, and can be formally de-

scribed as: supposing $I=\{i_1, i_2, \cdots, i_m\}$ attributes sets, the hypothesis rules are $X \Rightarrow Y$, $X \in I$, $Y \in I$ and $X \cap Y = \phi$. The attribute value is mapped into standard vectors, such as $X=(x_1, x_2, \cdots, x_n)$, $Y=(y_1, y_2, \cdots, y_n)$, according to the dis-isomorphic structure of language field and language value, in which the ordering pair is: $\langle x_i, y_j \rangle$, $(i=1, 2, \cdots, n, j=1, 2, \cdots, m)$. The statistic association of the state (or state changing) between X and Y can be described as: $H: x_i \rightarrow y_j$. Based on the theory of statistic induction, an algorithm of the knowledge discovery that has been proposed, and can be described as:

Step 1 Determine the rule template according to the domain knowledge and the interesting degree, such as $X \Rightarrow Y$;

Step 2 Combine the elements of the standard vector corresponding to X and Y, calculate the statistical parameters after scanning the database, and compute support and intensity of a hypothesis rule. For an ordering pair $\langle x_i, y_j \rangle$, $(i=1, 2, \cdots, n, j=1, 2, \cdots, m)$, the support is $\text{Support}=c_k/T$.

Where c_k is a statistical count of the k-th vector. T is the support of the rule template in the database. While rule intensity is: $H_s=\text{Support}(x_i \rightarrow y_j)/\text{Support}(X \wedge Y)$.

Step 3 Calculate the confidence CF of hypothesis rule with its support larger than the threshold,

$$\text{CF}(x_i|y_j)=[C_{(x_i,y_j)}/T-(T-C_{(x_i,y_j)})/T]\cdot\max[0, C_{(x_i,y_j)}/T].$$

Step 4 Select the hypothesis rules from the output as the discovered knowledge, according to the present threshold of support (S_c) and confidence (C_s), or by the verification of hypothesis rules.

4.3 Algorithm verification

In order to prove the correctness of the algorithm proposed in this paper, a software was developed to implement the algorithm. The software coded in VC++ and run to discover the knowledge from a database of physical factors of atmosphere. The database contains twenty attributes of weather of the investigated area, such as temper, humidity, pressure and *etc.*, the number of records in the database is 15 000. After running the software, sixteen rules are acquired. Select one of them, such as the rule 'if the humidity is heavy and pressure is higher then it rains with higher probability ', has support and confidence 0.45, 0.61 respectively, which are over the given thresholds. The support equaling 0.25 means that there are more than 25% records satisfying the condition 'the humidity is heavy and pressure is higher' and conclusion 'it rains' at same time in the database. The confidence, 0.61, means that in the records having the attributes 'humidity is heavy' and 'pressure is high', there are 61% of them with the attribute 'it rains' in the database. The thresholds of support and confidence affects the number of discovered rulers (**table 1 and 2**). In practice, the thresholds of support and confidence should be given according to the studied problems and the background knowledge. In order to analyze the effective of the algorithm, the program has run based on CPU of Intel PIII 766 with 256 M primary memory, and the results of run time *vs.* number of records are shown in **table 3**.

Table 1 Support *vs.* number of discovered rules (confidence=0.5)

Support	0.35	0.36	0.37	0.38	0.39	0.40	0.42	0.44	0.46
Number of ruler	144	112	92	72	63	45	30	19	15

Table 2 Confidence *vs.* number of discovered rules (support=0.4)

Confidence	0.4	0.5	0.6	0.7
Number of ruler	35	33	13	8

Table 3 Number of rules *vs.* run time

Number of records	1 000	5 000	10 000	15 000
Run time/s	2	4	5	6

5 Conclusion

KDD deals with all kinds of databases, which are consisted of qualitative and quantitative attributes. Rakesh Agrawal's algorithm for mining the association rule can't be adapted for describing the knowledge of the state and the state variation. The method of this paper can acquire the knowledge of causality and the relationship of state and its variation in the attributes. The method can be used not only in discovering knowledge from the science and engineering database, but also suitable for the general transaction database.

References

[1] M.S.Chen, J.W.Han, and P.S.Yu, Data mining : an overview from a database perspective [J], *IEEE*, 8(1996), No.6, p.125.

[2] B.R.Yang, ESKD-a new structure of expert system based on knowledge discovery [J], *Journal of University of Science and Technology Beijing*, 7(2000), No.1, p.63.

[3] B.R.Yang and D.Z.Zhang, Expanded research on KDD system (in Chinese) [J] , *Journal of University of Science and Technology Beijing*, 22(2000), No.1, p.84.

[4] R.Agrawal, Database mining: a performance perspective [J], *IEEE Trans on Knowledge and Data Engineering*, 5(1993), No.6, p.914.

[5] B.R.Yang, FIA and CASE based on fuzzy language field [J], *Fuzzy Sets and Systems*, 95(1998), p.83.

[6] I.Hacking, *Logic of Statistical Inference* [M], University of British Columbia, 1965.