2009 International Joint Conference on Artificial Intelligence

Research on KDD process model and an improved algorithm

Zhen Peng
Department of Computer
North China Institute of Science and Technology
Beijing, China
e-mail: yx_dpzc@yahoo.com.cn

Bingru Yang
School of Information Engineering
Beijing University of Science and Technology
Beijing, China

Hongde Ren
Department of Computer
North China Institute of Science and Technology
Beijing, China

Abstract—**In order to improve the performance of KDD greatly, the paper researched KDD model process based on double bases cooperating mechanism and focused on one of its part that is heuristic coordinator, which simulated the creating intent of cognitive psychology feature. And an improved heuristic coordinator algorithm was proposed. The method used FCM representing knowledge and being effective inference to get non-association state of knowledge base for directional mining knowledge in massive database automatically. Furthermore, the method greatly reduced the searching space and the complexity of the algorithm and enhanced the self-cognition ability.**

Keywords-Knowledge Discovery in Database (KDD); Double Bases Cooperating Mechanism; Heuristic Coordinator; Fuzzy Cognitive Map

I. INTRODUCTION

With the fast growing amount of data stored in database, data warehouse or other data repositories, how to derive useful knowledge is an impending problem. KDD [1,2] offers us a powerful tool. Generally, KDD systems are in accordance to user demands to mine knowledge in massive database. However, the research in KDD has mostly been concentrated on good algorithms for various tasks. Relatively little research has been published about the theoretical framework or foundations of KDD. To overcome it and improve the performance of KDD greatly, we regard knowledge discovery as a cognitive system , and construct double bases (knowledge base and database) cooperating mechanism to improving the knowledge discovery process model.

There are two important parts in double bases cooperating mechanism. They are heuristic coordinator[2,3] and maintenance coordinator, in which heuristic coordinator could simulated the creating intent of cognitive psychology feature enhanced the self-cognition ability. And KDD systems based on heuristic coordinator can focus from the two perspectives of user demands and system, where the heuristic coordinator algorithm is able to be used to discover knowledge automatically in the knowledge base by searching the non-association state of knowledge nodes, then activate the corresponding data sub-class structure in the massive database, thus realizing the directional mining process, effective reducing searching space and the complexity of algorithm. So the heuristic coordinator conforms to the inner cognitive rule of KDD, improve the KDD model and have applied widely [4,5,6].

Based on the research of KDD model process, the paper proposed an improved heuristic coordinator algorithm based on FCM, which use FCM for knowledge representation and inference algorithm to achieving accessible matrix. Compared with the approach based on hyper graph, on the one hand, FCM-based algorithm has strong reasoning ability to effectively reduce searching space and complexity of algorithm, on the other hand, it fully reflects the inner cognitive law of KDD and the cognitive characteristics of cognitive psychology.

The rest of the paper is organized as follows: the section 2 introduces knowledge representation and knowledge reasoning based on FCM, the section 3 represents heuristic coordinator algorithm based on FCM and conclusion is in the section 4.

II. KNOWLEDGE REPRESENTATION AND INFERENCE MECHANISM BASED ON FCM

A. *Knowledge representation*

As a soft computing methodology, Fuzzy Cognitive Map (FCM)[7], in which knowledge stored in concepts and the relations between the concepts that are both fuzzy variable, are relative easy to use for representing structured knowledge, and the inference can be computed by numeric matrix operation.

In heuristic coordinator, the number of relational nodes are equal to the number of rules and each relational node corresponds to a probabilistic relationship rule "if C_i then C_j" ,in which C_i is the single concept node or co-node pointing to the relational node, while C_j is the single concept node or co-node pointed by the relational node. Each concept node C_j has a state value recorded as A_j. A_j amounts to $\delta(C_j)/N$, in which $\delta(C_j)$ means the record numbers that C_j is true in database and N is the total record numbers.

The corresponding diagonal matrix W of FCM is shown in Figure 1. The W_{ii} values on diagonal line are 0 forever, because the rule of $C_i \rightarrow C_i$ is always true and do not need to participate in the calculation. The W_{ij} of interconnection relationship/rule denotes $\{?, sup(C_i \rightarrow C_j), con(C_i \rightarrow C_j)\}$. "?" is the state value of the $C_i \rightarrow C_j$ rule indicating that the rule is knowledge rule or not. So the value of "?" is 1 or -1. $sup(C_i \rightarrow C_j)$ is equal to $\delta(C_i \wedge C_j)/N$ and means the support of the $C_i \rightarrow C_j$ rule. If $\delta(C_i \wedge C_j)/N$ is less than support threshold, the value of "?" is recorded as -1 indicating the rule is non-knowledge rule. $con(C_i \rightarrow C_j)$ represents the confident of the $C_i \rightarrow C_j$ rule and is equal to $\delta(C_i \wedge C_j)/ \delta(C_i)$. Similarly, if $con(C_i \rightarrow C_j)$ is less than confident threshold, the value of "?" is also recorded as -1 indicating the rule is non-knowledge rule. To sum up, each value of W_{ij} in matrix is 0 or the set of state value, support and confident.

978-0-7695-3615-6/09 $25.00 © 2009 IEEE
DOI 10.1109/JCAI.2009.15

$$\begin{array}{c|ccccccc}
 & C_1 & C_2 & C_3 & C_4 & C_5 & C_6 & C_7 \\
\hline
C_1 & 0 & 0 & \{1,s,c\} & 0 & 0 & 0 & 0 \\
C_2 & \{1,s,c\} & 0 & 0 & 0 & 0 & 0 & \{1,s,c\} \\
C_3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
C_4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
C_5:C_1\wedge C_3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
C_6:C_2\wedge C_4 & 0 & 0 & 0 & 0 & \{1,s,c\} & 0 & 0 \\
C_7:C_3\wedge C_4 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}$$

Figure 1. The FCM relational matrix

In FCM, each A_j and W_{ij} may be change with time and the changing of record numbers of database. Maintenance coordinator is responsible for the real time maintenance of information. Here is no longer described.

B. Inference mechanism

The Inference mechanism based on FCM is to calculate accessible knowledge using adjacency matrix W and state values of nodes of FCM automatically. It includes the following four conditions.

Inference 1. If the rule A→B and B→C is valid, then the rule A→C is also valid.

Inference 2. If the rule A→B is valid, in which back part B is a co-node, then the rule $A\rightarrow B_1$ must be valid to any $B_1\subset B$.

Because the rule of A→B is true, that is to say, $\delta(A\wedge B)/N$ is more than the support threshold and $\delta(A\wedge B)/\delta(A)$ is more than the confident threshold as well. And $\delta(A\wedge B_1)>\delta(A\wedge B)$ is sure to be true. So it is definitely valid that $\delta(A\wedge B_1)/N$ is more than the support threshold and $\delta(A\wedge B_1)/\delta(A)$ is more than the confident threshold too. Therefore, the rule of $A\rightarrow B_1$ is true.

Inference 3. If the rule A→B is valid, in which back part B is a co-node, then the rules of $B_1\rightarrow A\wedge(B-B_1)$ and $A\wedge B_1\rightarrow(B-B_1)$ can be determined to be true or false to any $B_1\subset B$.

The support of the rules of $B_1\rightarrow A\wedge (B-B_1)$ and $A\wedge B_1\rightarrow (B-B_1)$ is equal to $\delta(A\wedge B_1\wedge(B-B_1))/N$, also equal to $\delta(A\wedge B)/N$.

The confident of the rule $B_1\rightarrow A\wedge(B-B_1)$ is equal to $\delta(A\wedge B)/\delta(B_1)$, also equal to support×N/$\delta(B_1)$ and the confident of the rule $A\wedge B_1\rightarrow(B-B_1)$ is equal to $\delta(A\wedge B)/\delta(A\wedge B_1)$, also equal to support×N/$\delta(A\wedge B_1)$, where support, N, $\delta(B_1)$ and $\delta(A\wedge B_1)$ are known. Consequently, the confidents can be determined that the two rules are knowledge or not.

Inference 4. If the rule of A→B is valid, in which anterior part A is a co-node, then the rule $A_1\rightarrow B\wedge(A-A_1)$ can be determined to be true or false to any $A_1\subset A$. By the same token, the support $\delta(A_1\wedge B\wedge(A-A_1))/N$ and the confident $\delta(A\wedge B)/\delta(A_1)$ can be confirmed.

III. An improved algorithm

The first step realizing heuristic coordinator algorithm is to discover some knowledge by reasoning in order to get accessible matrix W for determining the non-association state of knowledge nodes according to the above inference mechanism. It includes two processes of constructing two-dimensional array W and one-dimensional array A and calculating accessible matrix W.

```
Function calculate_reach_matrix
//calling function construct_matrix for the first
  constructing process
Step1: construct_matrix(C,E,A,W);
//calling function calculate_matrix for the second
  calculating process
Step2: calculate_matrix (A,W,K);

Procedure construct_matrix(C,E,A,W)
Step1: In accordance with the order of all single nodes
         first and co-nodes back, all single nodes and
         existing co-nodes in FCM are given ID($C_i$)
         representing the subscription i of $C_i$. The total
         number of FCM is recorded as |C|. The |C| nodes
         form one |C|×|C| matrix W;
Step2:The state value A[i] of each node in one-
        dimensional array A is assigned as $\delta(C_i)$;
Step3:Initially, W[i][j] of each relationship in matrix is
        the value of 0;
Step4:e:=1;
Step5:Read the e^th rule $r_e:C_i\rightarrow C_j$;
Step6:W[i][j]:={1,sup,con}, sup and con are the support
        and confidence of the rule respectively;
Step7:if e<|E| (|E| is the relationship number of FCM)
Step8:   e:=e+1, go to the step 5.

Procedure calculate_matrix(A,W,K)
Step1:for r:=1 to |C|;
//r means the row number of the matrix W
Step2: for l:=1 to |C|;
//l means the line number of the matrix W
Step3:  if r!=l&&W[r][l]!=0 is true,
Step4:   if $C_r$ is a co-node,
 //the 4^th inference condition
Step5:    to any $C_{ri}\subset C_r$
Step6:     if $C_{ri}$ or $(C_r-C_{ri})\wedge C_l$ does not exist in W,
Step7:      A[ri]:= $\delta(C_{ri})$ /N or
            A[ID($(C_r-C_{ri})\wedge C_l$) ]:= $\delta((C_r-C_{ri})\wedge C_l)$ /N
Step8:      if $con := \frac{W[r][l].\text{sup}}{A[ri]} > \min\_con$ is true,
Step9:      W[ri][ ID($(C_r-C_{ri})\wedge C_l$)]:={1, W[r][l].sup,con}
Step10:     K:= K∪{$C_{ri}\rightarrow((C_r-C_{ri})\wedge C_l)$ }
  // Adding new knowledge in knowledge base K
Step11:  else,W[ri][ID($(C_r-C_{ri})\wedge C_l$)]:={-1,W[r][l].sup,con}
Step12:  if $C_l$ is a co-node,
// the 3^th inference condition
```

```
Step13: to each $C_{li} \subset C_l$
Step14: if $C_r \wedge C_{li}$ or $C_l$-$C_{li}$ does not exist in W,
Step15:   A[ID($C_r \wedge C_{li}$) ]:= δ($C_r \wedge C_{li}$) /N or
          A[[ID($C_l$-$C_{li}$) ]:= δ( $C_l$-$C_{li}$)/N
Step16:   if $con := \frac{W[r][l].\sup}{A[ID(C_r \wedge C_{li})]} > \min\_con$ is true,
Step17:     W[ID($C_r \wedge C_{li}$)][ID($C_l$-$C_{li}$)]:={1,W[r][l].sup,con}
Step18:     K:= K∪{ $C_r \wedge C_{li} \rightarrow C_l$-$C_{li}$}
Step19:   else,
          W[ID($C_r \wedge C_{li}$)][ID($C_l$-$C_{li}$)]:={-1,W[r][l].sup,con}
Step20: if $C_{li}$ or $C_r \wedge$($C_l$ -$C_{li}$ ) does not exist in W,
Step21:    A[li]:= $\sigma(C_{li})$ /N or
           A[ID($C_r \wedge$($C_l$ -$C_{li}$ ))]:= δ($C_r \wedge$($C_l$ -$C_{li}$)) /N
Step22:   if $con := \frac{W[r][l].\sup}{A[li]} > \min\_con$ is true,
Step23:    W[li][ ID($C_r \wedge$($C_l$ -$C_{li}$ ))]:={1,W[r][l].sup,con}
Step24:     K:= K∪{ $C_{li} \rightarrow$ ($C_r \wedge$($C_l$ -$C_{li}$ ))}
Step25:   else,
          W[li][ ID($C_r \wedge$($C_l$ -$C_{li}$ ))]:={-1,W[r][l].sup,con}
//the 2th inference condition
Step26: W[r][li]:={1,δ($C_r \wedge C_{li}$)/N, δ($C_r \wedge C_{li}$)/(A[r]×N)}
Step27: for k:=1 to |C|;
// the 1th inference condition
Step28:  if k!=r && W[l][k]!=0 is true,
Step29:   W[r][k]:={1, δ($C_r \wedge C_k$)/N, δ($C_r \wedge C_k$)/(A[r]×N)}
```

Considering for not missing any rule, the procedure of calculate_matrix use firstly the forth inference mechanism and lastly the first inference mechanism. Finally, the rules with 0 weight value in matrix W are non-association states comprising knowledge rule and non-knowledge rule. Knowledge rule is these rules that conform to support and confidence threshold, do not yet appear in knowledge base and are still unable to be got by knowledge inference mechanism.

The second step of heuristic coordinator algorithm is to activate the corresponding data sub-class structure of the massive database and to realize the directional mining process to get the shortage knowledge by calling the procedure Heuristic_Coordinator.

```
Procedure Heuristic_Coordinator(A,W,K)
Step1:for i:=0 to |C|
Step2: for j:=0 to |C|
Step3:  if i!=j && W[i][j]==0 is true, then
Step4:   Directional mining to the data tables
           corresponding to i and j;
Step5:   if the support and confidence of the rule $C_i \rightarrow C_j$
           satisfy with thresholds,
Step6:     W[i][j]:={1,support,confidence};
Step7:     K= K∪{$C_i \rightarrow C_j$};
Step8:     calculate_matrix(A, W，K);
Step9:   else,
Step10:    W[i][j]:={-1, support, confidence}
```

IV. Conclusion

The paper proposed one new heuristic coordinator algorithm that is based on knowledge presentation method and inference mechanism of FCM. Compared with hyper graph based algorithm, the heuristic coordinator algorithm is able to effectively reason out more rules, especially the knowledge whose length is greater than 2, and more accord with cognitive characteristics. At the same time, KDD systems with the heuristic coordinator can achieve double focus from user demands and system self that simulate the "creating intent" of cognitive psychology feature. So the new heuristic coordinator algorithm based on FCM reduced the searching space and the complexity of the algorithm and enhanced the self-cognition ability and intelligence degree, which will be definite to extremely promote the mainstream development of KDD.

Acknowledgment

We would like to thank the Science and Research Project of North China Institute of Science and Technology (No.2008-B-13) and National Nature Science Foundation (No.60675030) support.

References

[1] P. S. Gregory, "Knowledge Discovery in Database: 10 Years After," SIGKDD Explorations, vol.1, 2000, pp: 59-61.

[2] H. Mannila, "Theoretical frameworks for data mining," SIGKDD Explorations Newsletter, vol.1, 2000, pp:30-32.

[3] B. Yang, T. Zhang, W. Song, J. Gao, "Coordinatiors based on cognitive psychology features and the cooresponding KDD process model," Journal of university of science and technology of china, vol.37, 2007, pp:212-216.

[4] B. Yang, W. Song, Z. Xu, "New structure of Expert System based on Knowledge Discovery Innovation Technology," Science in China(Series E:Information Sciences) , vol.37, 2007, pp:738-747.

[5] B. Yang, J. Wang, H. Sun, "A Study on Double Bases Cooperating Mechanism in KDD (Ⅱ) ," Engineering Science , vol.4, 2002, pp:34-44.

[6] B. Yang, W. Song, Z. Xu, "Knowledge Discovery theory and application based on inner cognitive mechnism," Progree in Natuarl Science, vol.16, 2006, pp:107-115.

[7] J. Aguilar. "A Survey about Fuzzy Cognitive Maps Papers (Invited Paper) ," International Journal of Computational Cognition, vol.3, 2005, pp:27-33.

[8] J. P. Carvalho, A. B. Tomé J, "Rule Based Fuzzy Cognitive Maps and Fuzzy Cognitive Maps –A Comparative Study," Proceeding s of the 18th International Conference of the North American Fuzzy Information Processing Society, New York, USA, 1999, pp.115-119.

[9] J. P. Carvalho, A. B. Tomé J, "Rule Based Fuzzy Cognitive Maps-Fuzzy Causal Relations," In: Mohammadian, M. (Ed.), Computational Intelligence for Modelling, Control and Automation-Evolutionary Computation and Fuzzy Logic for Intelligent Control, Knowledge Acquisition and Information Retrieval, IOS Press. pp. 276-281.