A Novel Construction of Heuristic Coordinator in KDD* Process Model

Zhen Peng, Lifeng Wu, Bingru Yang

Abstract

KDD* process has stronger theoretical foundations and better performance than KDD process. Heuristic coordinator as a key unit plays most important role in KDD* process model. In order to improve the heuristic coordinator algorithm, a kind of RBFCM model and its five important inference mechanisms are applied in the novel heuristic coordinator. The experiment in protein secondary structure prediction demonstrates its efficiency.

Keywords: KDD*, Heuristic Coordinator, RBFCM(Rule Based Fuzzy Cognitive Map), Protein Secondary Structure Prediction

1. Introduction

Knowledge Discovery in Database(KDD)[1,2] gives an effective way to solve the difficult that data in all works of life are rich but knowledge is poor. 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 the limitation of weak theoretical foundations and improve the performance of KDD greatly, we has proposed a new KDD process model-KDD* [3,4], which regards knowledge discovery as a cognitive system, incorporates double bases cooperating mechanism [5,6] constructing 1-1 mapping between databases and knowledge bases and two coordinators (heuristic coordinator and maintenance coordinator) into classical KDD process model, and improves some mining methods.

Heuristic coordinator [3,4], a most key unit in KDD*, simulates the “creating intent” of a cognitive process, is able to automatically find knowledge shortage, inspires the corresponding substructure in the database, starts the data mining and makes the self focus in order to directional mine in database, and determines the efficiency and the intelligence of KDD systems. But the proposed one directed hyper-graph based heuristic coordinator has some flaws such as only taking into account the rules with a single back part, only containing the existing co-nodes in the knowledge, limited reasoning mechanism and so on.

RBFCM[7,8] is a soft computing methodology and has stronger knowledge representation and inference ability than FCM[9,10]. Thus, it is suitable to uses RBFCM to represent knowledge and obtain accessible matrix by inference mechanism for discovering knowledge shortage and directional mining in the corresponding database. The paper uses one RBFCM model and the inference mechanisms to improve heuristic coordinator algorithm in KDD* process model.

At last, in order to validate the RBFCM based heuristic coordinator algorithm, we apply it in protein secondary structure prediction and get better prediction effect.

The rest of the paper is organized as follows: the section 2 introduces the KDD* process model and heuristic coordinator; knowledge representation based on RBFCM, the section 3 represents RBFCM based knowledge representation, inference rules. Section 4 evaluates the method in protein secondary structure prediction. Finally, we conclude this study in Section 5.

2. Heuristic coordinator in KDD* process model

In order to improve the performance of data mining greatly, we originally brought forward to a new research of inner cognitive mechanism based data mining, whose core idea is to regard the process of data mining as a process of cognizing, the knowledge discovery system as cognitive system, and form a new process model- KDD* shown in Figure 1.
The KDD* process model is based on incorporating double bases cooperating mechanism into classical KDD process model. Double bases cooperating mechanism constructs 1-1 mapping between databases and knowledge bases, demonstrates the structural corresponding theorem and resolves fundamentally the problem of “directional searching” and “directional mining”.

![KDD* Process Model Diagram](image)

**Figure 1.** KDD* process model

**Figure 2.** Knowledge single, compound node and their relations in Knowledgebase

Structural Corresponding Theorem: The reasoning category $C_r(N)$ of the universe of discourse $X$ and the accessibility category $C = \langle D, R_X \rangle$ of the complete data sub-class structure are equal. (The proof of this theorem can be found in Ref. [3]). This theorem establishes the one-to-one correspondence between knowledge single node and “data substructure” in database shown in Figure 2.

To realize the goals, two coordinators, heuristic coordinator and maintenance coordinator, are constructed. The heuristic coordinator is constructed to simulate “creating intent”, thus the knowledge discovery system is equipped with the feature of discover knowledge shortage automatically. The implement techniques of the heuristic coordinator is mainly on searching the non-association state of knowledge nodes in the knowledge base, in order to discover “knowledge shortage”, then activate the corresponding data sub-class structure in the real data base, thus realizing the directional mining process. Therefore, compared with classical KDD process model, KDD* has the following characteristics:

1. KDD* integrates the newly discovered knowledge and the intrinsic knowledge of the basic knowledge base organically.
(2) In the discovery process, KDD* processes the redundant, repeated and contradicted hypothesis in real time and effectively reduces the complexity of the problem. Furthermore, it realizes the synchronous evolution of both the knowledge base and the database.
(3) In the accumulation of data in the database, the evolution of the knowledge base can be realized automatically by the double base cooperating mechanism rather than by intervention of field experts.
(4) KDD* changes and optimizes the process of knowledge discovery. 
(5) KDD* enhances the ability of self-cognition, overcomes the limitation of field experts and domain knowledge.
(6) The double base cooperating mechanism (the kernel technology of KDD*) reveals 1-1 mapping between database and knowledge base, thus directional mining and directional searching are realized reducing the complexity of KDD system greatly.
(7) Based on double bases cooperating mechanism and KDD*, we can propose new algorithms with better performance to mine association rules.

3. RBFCM based heuristic coordinator

RBFCM was proposed to replace the fuzzy values with fuzzy rules and fuzzy variables with a number of fuzzy member functions and has stronger knowledge representation and inference ability than FCM. We have introduced one kind of RBFCM\(^7\) based heuristic coordinator for more efficiency.

8.1. RBFCM

**Definition 1.** The topology of RBFCM is a 5-tuple \((C, R, E, W, S)\) (Figure 3), in which \(C=\{C_1, C_2,\ldots, C_n\}\) is a set of concept nodes. \(R=\{R_1, R_2,\ldots, R_m\}\) is a set of “and” relation nodes representing relationship rules among concept nodes. \(E=\{<C_i, R_j> or <R_i, C_j>| C_i, C_j \in C, R_i, R_j \in R\}\) is oriented arcs between concept node and relationship node. \(W=\{W_i|W_i\text{ is the fuzzy weight value of relation node } R_i \text{ linking concept nodes}\}\). \(S=\{S_i|S_i \text{ is the fuzzy state value of } C_i\}\).

As shown in Figure 3, each concept node represents one attribute of database. Those concept nodes correspond to a set of \(S\), in which \(S_j\) is the state value of the concept node \(C_j\) and equal to \(\sigma(C_j)/N\) in the intervals of \([0,1]\), where \(\sigma(C_j)\) means the record numbers that the value of \(C_j\) is true in database and \(N\) is the total record numbers.

In RBFCM, each relation node corresponds to a weight that represents there are a probabilistic relationship rule “if \(C_i\) then \(C_j\)” between \(C_i\) and \(C_j\), in which \(C_i\) is one or more concept nodes with the “and” relationship pointing to the relation node and \(C_j\) is one or more concept nodes with the “and” relationship pointed by the relational node.

Each RBFCM corresponds to a correlation matrix \(W\). The \(W_i\) values on diagonal line of \(W\) need not participate in the calculation, because the rule of \(C_i\rightarrow C_i\) is always true. The \(W_i\) of interconnection rule node \(R_k\) denotes \(\{\text{flag, sup}(C_i\rightarrow C_j)\}\). Each \(k\) corresponds a pair of \(<i,j>\). “flag” is the flag of the \(C_i\rightarrow C_j\) rule and it’s value is 1 or 0, which respectively indicates that the rule is knowledge rule or not. \(\text{sup}(C_i\rightarrow C_j)\) is the support of the rule \(C_i\rightarrow C_j\) indicating the fuzzy or probabilistic degree of the rule “if \(C_i\) then \(C_j\)”. 

**Figure 3.** One simple rule based fuzzy cognitive map
Definition 2. Given any node $C_j$ where $C_j \subseteq C_j$ and $C_j \subseteq C_k$, $C_j$ is the sub-node of $C_j$ and $C_k$ is the father node of $C_j$.

Theorem 1. The state value of any father node of any node is less than or equal to the state value of any sub-node of the node, and the state value of the node is between the smallest state value in its sub-nodes and the largest state value in its father nodes.

Proof. Given any node $C_j$, where $C_j \subseteq C_j$ and $C_j \subseteq C_j$, $\delta(C_j) = \delta(C_j)$ and $\delta(C_j) = \delta(C_k)$. And $\delta(C_j)/N$ is definitely less than and equal to $\delta(C_j)/N$. Therefore the state value of the $C_j$ node is between the smallest state value in its sub-nodes and the largest state value in its father nodes.

Theorem 2. The support of any rule of $C_i \rightarrow C_j$ is between the minimum state value in all sub-nodes of $C_i \cup C_j$ and the maximum state value in all father nodes of $C_i \cup C_j$.

Proof. The support of the rule $C_i \rightarrow C_j$ is equal to $\sigma(C_i \cup C_j)/N$. Assumed that $C_{i1} \subseteq C_i \cup C_i \subseteq C_{i2}$ is satisfied, $C_{i1}$ is the node whose state value is minimum in all sub-nodes and $C_{i2}$ is the node whose state value is maximum in all father nodes, the support of the rule of $C_i \rightarrow C_j$ is between the state value of $C_{i1}$ and the state value of $C_{i2}$ according to Theorem 1.

Theorem 3. If the support of the rule of $C_i \rightarrow C_j$ is $sup(C_i \rightarrow C_j)$, then the confidence is $sup(C_i \rightarrow C_j)/S_i$.

Proof. Each rule of $C_i \rightarrow C_j$ has a support $sup(C_i \rightarrow C_j)$ being $\sigma(C_i \cup C_j)/N$ and a confidence $conf(C_i \rightarrow C_j)$ being $\sigma(C_i \cup C_j) \sigma(C_i)$. The confidence can be denoted by $sup(C_i \rightarrow C_j)^* N(\sigma(C_i))$. And $\sigma(C_i)$ is $S_i * N$. So $sup(C_i \rightarrow C_j)/S_i$ is equal to $con(C_i \rightarrow C_j)$.

That is to say, the fuzzy degree of one rule can only be represented by its support. If the support of $sup(C_i \rightarrow C_j)$ and the confidence of $sup(C_i \rightarrow C_j)/S_i$ are both more than or equal to their thresholds, the rule $C_i \rightarrow C_j$ is knowledge rule and it’s flag is recorded as 1. Otherwise, the rule $C_i \rightarrow C_j$ is non-knowledge rule and it’s flag is recorded as 0.

8.2. Inference mechanism based on RBFCM

Each relation node in RBFCM implicates some knowledge inferences that are the base of heuristic coordinator algorithm. It aims to get accessible knowledge according to the existing information in RBFCM matrix for reducing the scope of shortage knowledge. The inference mechanisms mainly include Corollaries 1~Corollary 5, while Corollary 6 and Corollary 7 are uncertain inference.

Corollary 1. If the rule $A \rightarrow B$ is true, in which the back part $B$ is a co-node, then the rule $A \rightarrow B_1$ and $A \cup B_1 \rightarrow (B - B_1)$ must be true to any $B_1 \subseteq B$.

Proof. The rule $A \rightarrow B$ is true. So it’s support of $\sigma(A \cup B)/N$ is more than or equal to the support threshold of $sup_{th}$ and it’s confidence of $\sigma(A \cup B)/\sigma(A)$ is more than or equal to the confidence threshold of $con_{th}$. To any $B_1 \subseteq B$, the support of the rule $A \rightarrow B_1$ of $\sigma(A \cup B_1)/N$ is certainly more than or equal to $\sigma(A \cup B)/N$ and the confidence of the rule $A \rightarrow B_1$ of $\sigma(A \cup B_1)/\sigma(A)$ is certainly more than or equal to $\sigma(A \cup B)/\sigma(A)$. So the rule of $A \rightarrow B_1$ is true. By the same token, to the rule of $A \cup B_1 \rightarrow (B - B_1)$, the support $\sigma(A \cup B_1 \cup (B - B_1))/N = \sigma(A \cup B)/N$ of $\sigma(A \cup B)/\sigma(A)$ is $sup_{th}$ and the confidence $\sigma(A \cup B)/\sigma(A \cup B_1) = \sigma(A \cup B)/\sigma(A) = con_{th}$. So the rule of $A \cup B_1 \rightarrow (B - B_1)$ is definitely true.

Corollary 2. If the rule of $A \rightarrow B$ is true, in which the back part $B$ is a co-node, then the rule of $B_1 \rightarrow A \cup (B - B_1)$ can be determined to be true or false to any existing $B_1 \subseteq B$.

Proof. The support of the rule $B_1 \rightarrow A \cup (B - B_1)$ is $\sigma(B_1 \cup A \cup (B - B_1))/N$ that is equal to $\sigma(A \cup B)/N$ and is greater than $sup_{th}$. The confidence of $\sigma(A \cup B)/\sigma(B_1)$ is equal to $sup(A \rightarrow B)/S(\text{id}_{B_1})$. For the existing $B_1 \subseteq B$, $\text{id}_{B_1}$ is the NO. of $B_1$. So the confidence value can be got and the rule can be determined to be true or false according to the state value of the node $B_1$.

Corollary 3. If the rule of $A \rightarrow B$ is valid, in which the anterior part $A$ is a co-node, then the rule of $A_1 \rightarrow B \cup (A - A_1)$ can be determined to be true or false to any existing $A_1 \subseteq A$.

Proof. By the same token, Corollary 3 can be proved.

Corollary 4. If the state value of $A$ is more than or equal to the support threshold and is also more than or equal to the product of the state value of $A_1$ and the confidence threshold to any existing $A_1 \subseteq A$, then the rule of $A \rightarrow (A - A_1)$ is definitely true, or else the rule of $A \rightarrow (A - A_1)$ is definitely false.

Proof. The co-node $A$ can be as $A_1 \cup (A - A_1)$. The support of $A_1 \rightarrow (A - A_1)$ is $\sigma(A_1 \cup (A - A_1))/N$ being equal to $\sigma(A)/N$ and the confidence of $A_1 \rightarrow (A - A_1)$ is $\sigma(A_1 \cup (A - A_1))/\sigma(A_1)$ being equal to $\sigma(A)/\sigma(A_1)$. So if the state value of $A$ of $\sigma(A)/N$ is more than or equal to the support threshold of $sup_{th}$ and is also more
than or equal to \((\text{sup th})\times\text{con th}\), then the rule of \(A_1 \rightarrow (A - A_1)\) is definitely true. Otherwise, the rule of \(A_1 \rightarrow (A - A_1)\) is definitely false.

**Corollary 5.** If the rule \(A \rightarrow B\) is not true, then the rule \(A \rightarrow C\) is certainly false to any \(B \subseteq C\).

**Proof.** The \(A \rightarrow B\) being false means the support \(\text{sup th}\) or the confidence \(\text{con th}\). Because of \(\text{sup th}\) or \(\text{con th}\) is true. Therefore, the rule \(A \rightarrow C\) is certainly false.

**Corollary 6.** If the rule of \(A \rightarrow B\) is valid, in which the anterior part \(A\) is a co-node, the rule of \(A_1 \rightarrow B\) is not necessarily valid to any \(A_1 \subseteq A\).

**Proof.** Because of \(\text{sup th}\) or \(\text{con th}\) is true. So the rule \(A \rightarrow C\) can not be reasoned out by the support and confidence of \(A \rightarrow B\) and \(B \rightarrow C\). They may be less or more than the thresholds. So the rule \(A \rightarrow C\) is not necessarily true.

4. Experiment of heuristic coordinator in protein secondary structure prediction

We use CMP (Compound Pyramid Model)\(^{[12,13]}\) for protein secondary structure prediction. The compound pyramid model is composed of 4 layers that are comprehensive analysis layer, kernel judgment layer, assistant judgment layer and optimization layer. Comprehensive analysis layer combines improved SVM method and improved homogenous analysis method. In the kernel judgment layer, the SAC (Structural Association Classifier) module is in the core method of classification, which takes on the classification of data that is hard to judge in compound analysis layer after the classification via comprehensive analysis layer. The AAC (Attribute Association Classifier) module is located in the assistant judgment layer of CPM. Through the association analysis of the physicochemical properties of the amino acid, refinement rule is created to predict the data of lower layers that is not identified. The KDD* model and the novel heuristic coordinator algorithm RBFCM_HC based on RBFCM and Inference mechanism are applied in kernel judgment layer, assistant judgment layer.

We select RS126\(^{[15]}\) and CB513\(^{[16]}\) datasets in our experiment. At the same time, we use Q3 as evaluation standard denoting a percent of accurately predicting amino acid in total amino acid. Each layer’s results of CPM are shown in Table 1 and Table 2.

<table>
<thead>
<tr>
<th>Module</th>
<th>Accuracy</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comprehensive Analysis Layer</td>
<td>16937/18646</td>
<td>90.83%</td>
</tr>
<tr>
<td>Kernel judgment layer</td>
<td>3891/6053</td>
<td>64.28%</td>
</tr>
<tr>
<td>Assistant judgment layer</td>
<td>16/107</td>
<td>14.95%</td>
</tr>
<tr>
<td>Total</td>
<td>20844/24806</td>
<td>84.03%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Module</th>
<th>Accuracy</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comprehensive Analysis Layer</td>
<td>105106/113638</td>
<td>92.49%</td>
</tr>
<tr>
<td>Kernel judgment layer</td>
<td>19970/32194</td>
<td>62.03%</td>
</tr>
<tr>
<td>Assistant judgment layer</td>
<td>88/401</td>
<td>21.95%</td>
</tr>
<tr>
<td>Total</td>
<td>125153/146233</td>
<td>85.59%</td>
</tr>
</tbody>
</table>

Reviewing on the typical research literatures, As shown Figure 4 and Figure 5, SVM method used by HU\(^{[16]}\) got 78.8% on RS126, nerve network used by Xie\(^{[17]}\) got 79.65% and 69.11% on RS126 and CB513, layer nerve network used by Chen\(^{[18]}\) got 74.38% on RS126, cell automatic machine method used by Chopra\(^{[19]}\) got 58.21% and 56.51% on RS126 and CB513, context analysis used by Liu\(^{[20]}\) got 69.8% and
69.6% on RS125 and CB513, double layer SVM method used by Guo\cite{21} got 70.8% and 73.5% on RS125 and CB513.

Now we can see that the CPM achieves the excellent predict accuracy compared to other methods and servers. And CPM method is a universal prediction system, which can be used in different types of protein secondary structure prediction.

5. Conclusions

The paper first introduces the KDD* process model and its theoretical foundation of double bases cooperating mechanism, emphasizes on the effect of heuristic coordinator in KDD* system. Then a novel heuristic coordinator algorithm based on RBFCM and the inference is presented. At last, the experiment in protein secondary structure prediction demonstrates the heuristic coordinator in KDD* process model has better efficiency.

Acknowledgements

We would like to thank Knowledge Engineering Institute of Beijing University of Science and Technology and National Natural Science Foundation of China (No.60875029, No.61175048) support.

References


