Expanded Researching on Knowledge Discovery

Bing-ru Yang[1], De-zheng Zhang, Jing Tang, Ying zhou

(Information Engineering School, University of Science and Technology, Beijing, China, 100083)

ABSTRACT: The development of database technology, every-increasing needs of enterprise management for the decision-making support and the comprehensive application of AI are driving the theories and technology of data mining and knowledge discovery growing rapidly. In recent years, data mining and knowledge discovery has been paying more attention to and all kinds of algorithms and tools are blooming. Their characteristics are overlapping of subjects, fusion of several techniques, generalizing of data mining and integrating of knowledge discovery.

This paper discussed the development trend of knowledge discovery based on the summary and analysis of the theory of knowledge discovery in database and the actuality of technological method, and introduced the mechanism of total data mining process, researching in expansible structure and algorithm of KDD, and serially developing of software, according to the theory of double base cooperation proposed by author.

KEYWORDS: Data mining, Knowledge discovery, Double-bases cooperation.

1. The Actuality of Data Mining and Knowledge Discovery

Because of the application of Database and Data Warehouse, a great many of data and information have been accumulated in the field of enterprise management, finance activities, commerce running, science and technology researching and etc. The extending of internet result in soaring of the total of information relevant to the internet. Rich information and lots of data which are included in Web database, e-mail, web pages and so on can be used in all kinds of fields and meet different needs, but they can't be extracted by the traditional statistic analysis methods. Consequently, The Data Mining and the Knowledge Discovery on Database(KDD) come in to using. Data mining technology can be used to overcome the shortage of the traditional statistic methods, and be adapt to the analyses of large scale data. Data mining is suitable to extract useful information from data. The knowledge discovery is the kernel of data mining, and guide the way of data mining. The concept of knowledge discovery based on database can be defined as a uncommon process of extracting effective, novel, potential and understandable patterns. A lot of methods and tools are suitable for the discovering of knowledge, such as classification of rules, prediction models, data summarizing, cluster, association rules, sequence models, dependence relationship or model, abnormality, trend and etc.

The researchers in different fields are studying and developing the data mining from different viewpoint by way of corresponding theories and analysis methods. The methods used by data mining and knowledge discovery involve machine learning, statistics, analysis of database, recognition of models, machine discovery, artificial intelligence, knowledge acquirement, artificial neural network, data visualizing, uncertainty reasoning, intelligence data analysis, genetic algorithms, fuzzy logic, rough set and so on.

The objective of data mining and knowledge discovery involve structure, semi-structured and unstructured data sources which include relationship database, object oriented database, spatial relationship database, reasoning database, multimedia database, text database, graphic, audio and video data sources.

With the widely using of data mining and knowledge discovery theory, the development of utility tools are paid more attention to. A lots of algorithm and tools are developed, and a variety of commercializing utility data mining tool come forth rapidly to meet kinds of application needs.

In recent years, the studying of data mining and knowledge discovery show the
characteristics of subjects overlapping, techniques fusion, data mining generalizing and knowledge discovery unifying. Then, a new knowledge discovery frame is needed to achieve the uniform of knowledge discovery, maturity and cognizing independence.

2. New Knowledge Discovery Frame-- KDD*

Nowadays, the development of data mining and knowledge discovery has made great progress, and at the same time, many challenge points are proposed. Several of them are: 1) To break through the closed system which are the knowledge discovery based on the database with the simple frame, to fusion the two knowledge discovery mechanism which can cooperate with knowledge base, and to obtain the new knowledge discovery according to the drive and restriction of basic knowledge base; 2) Now the situation is that more attention is paid to the research of algorithms, but we think it is time to research the general frame of knowledge discovery from different knowledge level and multiple abstract classes in macro background; 3) The knowledge discovery in database should be fused with the knowledge evolution in knowledge base, and integrated in the whole process of general knowledge discovery; 4) Further studying is necessary for the intelligibility of discovered knowledge and the utility of priori knowledge in the knowledge discovery; 5) The development of practical software and tools of knowledge discovery are urgent.

According to the background and the results of cognition logic development, the author proposed the double bases cooperation mechanism based on the knowledge discovery and overlapping of cognition science and artificial intelligence, and developed the expanding structure of double bases cooperation of knowledge discovery---KDD* through continuously studying.

The KDD* system represents the fusion of KDD and double base cooperating mechanism. This is not a simple addition, but fusing double base cooperating mechanism into the technology i.e. build the “path” of inherent correlation between database and fundamental knowledge base in order to restrict and drive the mining process of KDD, change the inherent running mechanism of KDD, and form an open, optimum enlargement relative to KDD. The mechanism of double bases cooperation improve the theory and function of KDD, shown in two aspects: 1) Under the attribute-based constructive principle, it will find “Knowledge Shortage” and produce “original idea intend” through searching the irrelevant status of “Knowledge Nodes” in knowledge base to elicit and activate the corresponding “Data Class” and produce “directional mining process”. 2) The structure is convenient to the maintenance and management of knowledge base during the process of KDD through the cooperation of double bases.

Overall structure figure of KDD* system (see fig.1)

Relative to KDD, KDD* is a new knowledge discovery structure of the fusion of KDD and double base cooperating mechanism. It has the following features:

1) KDD* organically fuses the newly discovered knowledge of KDD* and the inherent knowledge of basic knowledge base and make them become an organic whole; i.e. the previous knowledge of users and “the knowledge discovered before are coupled into the process of discovery.”

2) In the process of knowledge discovery, KDD* processes the redundant, repetitive, and incompatible information in real time. It can effectively reduce the complexity of problems caused by the process of accumulation. And at the same time provides precondition for the fusion of the old and the new knowledge; achieves “the synchronization evolution of knowledge and database.”

3) KDD* changes and optimizes the process and running mechanism of knowledge discovery, realizes “multi-headstream” focusing and reduces the quantity of evaluation.

4) From the perspective of cognition science, KDD* strengthens and provides the intellectual degree of knowledge discovery, enhances the cognition of self-determination of computer (this is the main theme of research in the long run), effectively overcomes the limit of field experts, uses “the focusing of domain knowledge assisted initial discovery”.

5) The kernel knowledge of KDD* the research on double base cooperating mechanism shows, under certain principle of base construction, the corresponding relation between knowledge sub-base and data sub-structure, provides effective technical
method to reduce searching space, enhance mining efficiency for the realization of “limited searching”.

3. Theory fundation of double bases cooperating mechanism:

The key of KDD* is double base mechanism, while the technological realization of double-base cooperating mechanism is to construct interruptive and heuristic coordinator. The requirement of realizing interruptive and heuristic coordinator is that the large (basic) knowledge base should be divided into several correlative sub-knowledge bases according to each domain; Meanwhile, extracts in real database the correlative sub-databases according to each domain. Thus the layers between knowledge nodes in knowledge base and data sub-class (structure) make a one to one mapping. The basis theory which is proposed by us is pan-homotopy conception and the following definitions and structure mapping theorem: (Details can be found in reference [4][5])

3.1 Knowledge node, data sub-class and their relation

3.1.1 Knowledge node

Definition 3.1: In the sub-knowledge bases corresponding to domain X, knowledge expressed in the following ways is called uncertainty rule type knowledge:
1) \( P(X) \Rightarrow Q(X) \)
2) \( P(X) \Rightarrow \bigwedge_{j=1}^{n} Q_j(X) \)
3) \( \bigwedge_{j=1}^{n} P_j(X) \Rightarrow \bigwedge_{j=1}^{n} Q_j(X) \)
4) \( \bigwedge_{j=1}^{n} P_j(X) \Rightarrow \bigwedge_{j=1}^{n} Q_j(X) \)

Where, \( P(X) \), \( P(X) \), \( Q(X) \), \( Q(X) \) are all in the form of “Attribute Word” (or “State Word”) + “Degree Word”. For example, the following knowledge “If water line in the cooling water expanding box of diesel engine drops unconventionally and the exhaust temperature is low, then the supercharger cracks” belongs to type 3.

**Definition 3.2**: In definition 1, \( P(X) \) and \( P(X) \) are called knowledge initial nodes, \( Q \), \( X \), and \( Q \), \( Q \), \( Q \), \( Q \) are called knowledge terminative nodes, both of them are called prime knowledge nodes; \( j \), \( j \) are called knowledge initial nodes, \( Q \), \( X \), and \( Q \), \( Q \), \( Q \), \( Q \) are called prime knowledge nodes.

**Definition 3.3**: Aiming at domain \( X \), in the sub-databases corresponding to correlative sub-knowledge base, \( S=<U, N, I, W> \) is called data sub-class structure which corresponds to each prime knowledge nodes.

Where, \( U \neq \Phi, U, \{u_1, u_2, \ldots, \} \), \( u_i \) is the set of data and produced by following I), it is the class (called data sub-class) which describes the “Attribute Word” and “State Word” of prime knowledge nodes;

\( N \neq \Phi \) is the finite set of language values, which describes the “Degree Word” of prime knowledge nodes;

\( I, N \rightarrow U \) is the mapping that divides the classes of data set according to language values. Generally it is divided into several overlapping interval when data continuously distributes (namely \( \forall_{i,j}(u_i \cap u_j \neq \Phi) \) )

\( W, N \rightarrow [0,1]^k \) (in positive integer) satisfies the following conditions:

\( \forall_{n_1, n_2} \epsilon N, n_1 \leq n_2 \rightarrow W(n_1) \leq W(n_2) \)

\( \forall_{n_1, n_2} \epsilon N, n_1 \neq n_2 \rightarrow W(n_1) \neq W(n_2) \)

Where, \( \leq_N \) is the full order on \( N \), \( \leq_{dc} \) is the dictionary order \( [0,1]^k \), \( W(n) \) (\( n \epsilon N \)) is the standard vector of language value (i.e. the sample choose from the vector which corresponds with language value and the language value corresponding the middle point of interval and its adjacency).

**Definition 3.4**: In the structure of data sub-class \( S=<U, N, I, W, > \), three element group \(<u_i, n_i, r_i>\) is called the layer of S which satisfies the following conditions:

\( u_i \epsilon U, u_i(i=1, 2, 3, \ldots, v) \) is the sample’s data set of originally divided No. i interval;

\( n_i \epsilon N, n_i(i=1, 2, 3, \ldots, v) \) is language value that distinguished by which interval sample data set drops.

The determination of \( r_i(i=1, 2, 3, \ldots, v) \):

i) When the data of \( u \) sample drops in uncrossed interval, \( r_i \) is standard vector, \( r_i \epsilon W(n) \);

ii) When the data of \( u \) sample drops in crossed interval we can get \( r_i \) by the interpolating formula:

\[ r_o^i = A_i(I - \frac{|u_i - u|}{l_i}) + A_{adjacent} |\frac{u_i - u}{l_i}| \]

where, \( u_o \) is the standard sample data of interval i, \( l_i \) is the length of interval i, \( A_i \) is the standard vector of interval i, \( A_{adjacent} \) is the standard vector of the interval that be adjacent to that one \( u \) drops.

Then according to the measurement of \( r_o^i \) and \( r_i^*, r_{i+1}^* \), or measurement of \( r_i^* \) and \( r_i, r_{i+1}, r_{i+1} \) or \( r_{i+1} \) or \( r_{i+1} \) is decided to choose, and this datum is reserved in layer i or moved to layer i+1 or layer i-1.

3.1.3 Mapping from knowledge node to data subclass

It is obvious that all the data sub-class and data sub-class structure make a one to one mapping.

**Theorem 3.2**: Aiming at X, in the sub-database corresponding to sub-knowledge base, all the data sub-class are F(definite set), if D\( \subseteq \rho \), \( F \)

(D includes \( \Phi \) and \( F \)), then \( <F, D> \) make a topological space. (Proof omitted)

**Definition 3.5**: Supposed X and Y are arbitrary topological space, we call the following continuous mapping \( F : X \times [0,1]^k \rightarrow Y \) as the path homotopy of the mapping from X to Y.(the extension of usual homotopy)
**Definition 3.6:** Suppose f and g are continuous mapping from topological space X to Y, if there is a pan-homotopy \( F(x,t) = f_t(x) \), making that for any \( x \in X \) there are \( f(x) = F(x, (0, ..., 0)) \), \( g(x) = F(x, (1, ..., 1)) \), then we call g and f are pan-homotopy and F is the pan-homotopy that relates mapping f and g.

**Definition 3.7:** We call the continuous mapping f which is from topological space X to Y as equivalence of the pan-homotopy, if there is continuous mapping g that is from topological space Y to X, making synthetical mapping \( g \circ f \) and \( f \circ g \) are respectively mapping that are from X and Y to itself and pan-homotopy to identical mapping \( I_X \) and \( I_Y \) in corresponding space, and expressed as \( g \circ f \sim I_X \), \( f \circ g \sim I_Y \). Mapping g is also equivalence of pan-homotopy and it is called the adverse equivalence of equivalence f.

**Definition 3.8:** Suppose two topological spaces, if there is at least one pan-homotopy equivalence mapping from one space to another space, then the two spaces are called the spaces of identical pan-homotopy type.

**Theorem 3.3 (Structure Mapping Theorem):** For domain X, in the sub-database corresponding to sub-knowledge base, topological spaces <\( E, F \)> of knowledge nodes and <\( F, D \)> of data sub-class structure are identical pan-homotopic type spaces.

This theorem presents the mapping of layers between knowledge nodes in the sub-knowledge base and data sub-class in corresponding sub-database, shown in fig.2.

---

**3.2 The mechanism of double bases cooperation**

The technical realization of double-bases cooperating mechanism is to construct R type coordinator and S type coordinator. The main function of R type coordinator is as follows: when the great number of data in real database are focused to produce new knowledge, R type coordinator will cause the process of KDD to produce interrupt and search the repetition of reasoning category \( C_R(F) \) in <\( F, D \)> are got, and the problem of “directional searching” and “directional mining process” are resolved.

On the basis of the research above, we can see that in the knowledge discovery system mathematical structure of database and knowledge base essentially come down to pan-homotopy category. Namely database is pan-homotopy category combined with data subtype (structure) set and “mining path”, which is called data mining category; and knowledge base is pan-homotopy category combined with knowledge nodes set and “reasoning arc”, which is called knowledge reasoning category. Additionally more results about the isomorphy and restricting mechanism of knowledge reasoning category \( C_D(E) \) in <\( E, F \)>and data reasoning category \( C_R(F) \) in <\( F, D \)> are got, and the problem of “directional searching” and “directional mining process” are resolved.
this piece of knowledge in the knowledge base. If repetition is found, new knowledge is canceled and KDD is restarted at the beginning point. If there is no repetition, KDD will go on to evaluate and put this piece of knowledge into base. The main function of S type coordinator is as follows: under the attribute-based constructive principle, it will find “Knowledge Shortage” and produce “original idea intend” through searching the incorrelate status of “Knowledge Nodes” in knowledge base to elicit and activate the corresponding “Data Class” and produce “directional mining process”.

The necessary condition of realizing above function of R and S type coordinator is to construct the mapping relation between knowledge nodes in knowledge base and data sub-type structure in the true database and produce directional searching for repetition knowledge (reduce searching space) and elicit directional mining process. Typified on uncertain rule type knowledge of language variables, discussion is done in real database and large knowledge base.

Because the interruptive coordinator is introduced into KDD, the contradictory and redundant knowledge can be canceled earlier. Only those which are possibly accepted as new knowledge are evaluated and the evaluation work is greatly reduced. The efficiency of KDD is enhanced.

The KDD* is an open structure, which fuse the knowledge discovery and double bases cooperation mechanism, and the basic knowledge is used to drive and limit the process of knowledge discovery. The theory have changed the inertial mechanism of knowledge discovery, and become an extended mechanism with optimal and open structure and function compared to the ordinary knowledge discovery.

4. The realization algorithm of KDD*
4. 1 The realization algorithm
The modules can be divided into the following steps:

Step 1: Pre-processing
To process the original data by purifying the data, specific changing, etc. and create the mining database DMDB in order to use it in the process of data mining and knowledge discovery.

Step 2: Focusing
Namely to chose data from DMDB. The main methods in focusing are clustering analysis and detecting analysis. The methods to direct the focusing are: (i) the expert, through man-machine interaction, inputs the knowledge in which he is interested and direct the direction of the data mining; (ii) directive data mining by using heuristic coordinator.

Step 3: Acquire hypothesis rules
This is the kernel process of KDD, which uncommonly extracts in real database (has a large number of data, and incomplete, uncertain, structural, sparse characteristics) the previously unknown information which had the value of potential application and hid in the data. Cause-and-effect correlation rules are extracted in the system. It further enriches the basic knowledge base. The mining method used is statistical induction and cause-and-effect correlation qualitative reasoning.

Step 3.1 Determining the relationship of language values with statistic method
Divide A, B as A(A1 , A2 , .... , Am),B(B1 , B2 , .... , Bn) according to their language values. If A and B are both single variable then we have A(A1, A2, A3, A4, A5), B(B1, B2, B3, B4, B5). Given A is the intersection of m1 variables, then m=m1. Given B is the intersection of n1 variables, then n=n1. Thus there are altogether m×n kinds of combination< Ai , Bj > i=1, 2,...,m ; j=1,2,...,n. To calculate the possibility factor Pk=Cnk/N ( k=1,2,..., m×n) corresponding to each combination, P=0.5 is the highest possibility. If Pk>0.5, < Ai , Bj >is selected, otherwise it is eliminated and they are considered to have no relativity.

Step 3.2 Visual tools used to analysis A and B
Experts can use visual tools, such as a distribution figure to decide the combination of the selected or eliminated intervals. The intervals here have one to one mapping relation with the language value mentioned above, i.e. the language value and the corresponding radius equals the corresponding interval. The acquired interval combination must be changed into corresponding language value combination which is to be used in the later calculation. To get two kinds of properties which are highly relative. e.g. Ai and Bj, and extract the corresponding values such as general statistic value N, statistic value Cnk(Ai,Bj) which is
the times of together appearing of $A_i$ and $B_j$, statistic value $C_n(A_i)$ expresses the times of appearing of $A_i$, and statistic value $C_n(B_j)$ expresses the times of appearing of $B_j$, to decide which variable have causal relation.

**Step 3.3 Getting the weight of a single premise in rule $A_iB_j$**

Given $A_i$ is single premise, its weight is 1; given $A_i$ is the interaction of many premises, i.e. rule $R$: $A_iB_j$ is:

$$R,(P_1 , p_1),(P_2,p_2)....,(P_n , pn),(Q , q)$$

Then the mutuality number $r_i$ in $(P_i , p_i),(P_i , p_i)$ and $(Q , q)$ can be gotten from the following formula. The weight in its rule can be gotten according $r_i$.

$$r_i = \frac{\sum_{j=1}^{m} q_j (p_j - (p_i - q)) - (\sum_{j=1}^{m} (p_j - (p_i - q))) \sum_{j=1}^{m} q_j}{\left[ \sum_{j=1}^{m} (p_j - (p_i - q))^2 - (\sum_{j=1}^{m} (p_j - (p_i - q)))^2 \right] \left[ \sum_{j=1}^{m} q_j^2 - (\sum_{j=1}^{m} q_j)^2 \right]}$$

**Step 4: Double base cooperating mechanism**

To process the acquired hypothesis rules by using interruptive coordinator and heuristic coordinator, and to inspire the data focusing for data mining by using relative intensity.

**Step 5: Evaluation:**

This process is mainly used to evaluate the acquired hypothesis rules in order to decide whether they will be stored into the knowledge base. The main methods are: (i) according to relative intensity, we can set up a threshold value and be achieved by computer; (ii) experts evaluate through man-machine interaction interface and also evaluates through all kinds of figures and analysis materials provided by visual tools. To store the rules that evaluated and accepted into the derivative knowledge base as new knowledge.

4. 2 The prototype of the software

Today, many of studying in prototype and application system pay more attention to material technology of mining, but we emphasis on the mechanism of total date mining process, external structure and running mechanism of KDD, and serial development of software system. The major characteristics are Combination between macropscopical and microcosmic, between abstract and materiality, between theory and application. We have taken up research in mechanism and algorithm, developed KDD stars software called independent KDD stars software, which includes and expands all function of prior developed software. Through the long term of research and practice, we have built up the mechanism of double-bases cooperation and achieved the prototype of technology, developed the software and used it in practice.

The theory and technology of knowledge discovery have wide application fields. They are showing a promising future in area of decision-making support, estimation of commercial activities, analyses and prediction, and scientific study. The following is an example of the software application based on the data of surveying in a community. The attributes of the real database include work status, marriage condition, age of first marriage, age of their children, years of education, income per year, sense and so on.

- Starting with data mining according to the interesting of expert.

Choosing the attribute interested by expert to start with data mining, the rules of the software running are shown in the following figure(Fig 3), in which the condition is ‘education level’ and ‘year income of 1991’ as the result.
• Using heuristic coordinator to drive the process of knowledge discovery (KDD*)

According to “Knowledge Shortage” in knowledge base, heuristic coordinator can realize the knowledge discovery automatically, which can be chosen by experts of field or evaluation methods. The result of KDD running based on the same database shows in the figure 4.

5 Conclusion

The theory proposed in the paper is the abstract of the theory and technology of knowledge discovery and data mining development. The mechanism discover the natural kernel of knowledge discovery or perceiving. Two coordinators can become system independently, namely it can form “interface” to be loaded in any KDD system to communicate with knowledge base. Then the mining efficiency is improved greatly.

This system supports the ordinary database such as Oracle, Access, FoxPro etc. And it adopts data dictionaries. According to different field database, it is only necessary to change the data dictionary of different field. Therefore, it is general. In addition, it can be used on the internet.

Reference


