CORRELATIVITY SETS BASED THEORETICAL FRAMEWORKS OF DATA MINING

XIAO-FENG WANG[1], JING LU[2], TIAN-RAN WANG[3]

[1]Shenyang Institute of Chemical Technology, 110142 Shenyang China
[2]Shenyang Institute of Automation, The Chinese Academy of Sciences, 110030 Shenyang China
[3]Shenyang Institute of Chemical Technology, 110142 Shenyang China
E-MAIL: Wangxf@mail.sy.ln.cn

Abstract:

The plausibility relation, one is generalization of fuzzy relation and probabilistic relation, is proposed in the paper. Data mining is a process of finding the plausibility relation from database and correlativity measure to be a particular plausibility relation based on correlativity sets. The critical calculation such as the accuracy of the rough sets, the confidence and the Bayesian form in data mining can be united which use the correlativity measure. The GPDM (General Process of Data Mining) represented the nature of data mining is proposed also. The data mining theoretical foundation and frameworks based on correlativity sets are given and discussed also in the paper.

1 Preface

At present, many of data mining methods have been discovered such as Apriori Algorithm (R. Agrawal, T. Imielinski, and A. Swami 1993), Decision Tree Method, Bayesian Method, Neural Network Method and so on. And various data mining theories are proposed and adopted, such as Fuzzy sets, Rough sets, Statistics, machine learning, neural network, decision tree, pattern recognition, high-performance computing and so on. Various of data can be mined, from traditional relational database to all kinds of text data, spatial data, image, video and Web information, biomedical information. Different kinds of data mining platforms have been developed, such as SAS Enterprise Miner of ASA company , Intelligent Miner of IBM company, DB Miner of Simon Fraser University in Canada, MS Miner of Institute of Computing Technology of Chinese Academy. But, with the deep research of data mining theory and application, we find that the fundamental research about data mining is not mature. It will be an important significance to advance the data mining if an unified theoretical framework can be constructed.

There are various answers for the question of what the theoretical foundation for data mining is, such as data inductive theory, data compress comment, pattern discovery, probability theory , induce database, microcosmic economics and so on. In 2001, Jiawei Han & Kamber M.of Simon Fraser University (Jiawei H., Micheline K.2000) proposed 3 requirements needed to meet for an ideal theoretical frame :

(1) able to model for a typical data mining tasks such as associations, classifications and clusters;
(2) having possibility characteristic;
(3) able to process various forms of data, considering the essence of iteration and interaction of data mining.

In 2001, the Granular Computing Model was given by Y.Y.Yao(Y.Y.Yao *2001*), depending on Granular Computing theory and the research of data mining modeling. In our opinion, due to the complexity of data mining objects, the diversity of users' requirements, it is impossible that an all-purpose and general data mining method could be applicable to all kinds of data mining tasks. Therefore, the current research result shows that it is feasible that some basic concepts, calculations and steps of different data mining methods are mixed together, and form a united abstract theoretical frame. For example, the calculation of rough degree in rough set and the calculation of confidence and support in Apriori Algorithm can be integrated as correlativity measure (Wang Xiaofeng, Wang Tianran 2002). logic return model in Simple Bayesian classification is consistent to the apperception function with sigmoid in neural network; the probability representation of confidence is consistent to Bayesian method, the integration on mathematics of these key conception, and basic calculation in different mining data methods is not happened coincidentally, which includes some inevitable relations. These relations hidden inside the diverse theories and methods which seem to be different is the foundation of a united theoretical framework.

Starting from the knowledge and knowledge representation, this paper explores the essence of data mining, proposes a new point, and correlativity set act as the foundation of data mining. We analyze the generality, the interactivity and mixing problem of the basic concepts, theory and calculations for the current data mining methods, and discuss the possibility of correlativity set as the foundation of data mining theoretical framework.

0-7803-7865-2/03/$17.00 ©2003 IEEE

2 Essence of Data Mining

2.1 Knowledge Representation in Data Mining

The set is one of the basic concepts in mathematics, the first step that can be applied to abstract mathematical notation from specific object. The relation is another basic concept created in set. If a set represents a concept or an object, then a relation is the relationship (including function and operation) among concepts or objects. Knowledge is one kind of relationship about the states, activity rules and relationships between one object and all other object around it. In Data Mining area, knowledge can be classified into 4 categories: association rules, classification rules, sequential rules and clusters. Association rules describe the inherent relationship among data (or objects); clusters can discriminate categories automatically; classification can calculate classified function from the given categories; and a sequential rule is the time sequence of association or classification, it is one kind of special association rules or classification rules. Those four kinds of rules can be classified into two categories: one is the "discrimination" of the knowledge about objects, for example, depending on the feature of the data, clusters and classification rules can "discriminate" objects to form different sets about data; The other one describes the "inherent relationship" among objects, such as association rules. The two kinds of knowledge, "discrimination" of objects and description of relationship, all can be represented by "relation" in mathematics. The "discrimination" of objects can create relations of featured attribute sets and concept sets, but the relationship among objects is a "relation" of concept sets. So it is appropriate to represent data mining using sets and relations. comparing the other theories, they are closer to data and knowledge.

For Data Mining, objects exist in database in the form of data. Mining the data characteristic of objects or the relationship (knowledge) among objects is just discovering the "relation" from database. Since the data that represents the objects is usually incomplete, imperfect and inaccurate, the mined knowledge is incomplete and inaccurate, a result of conjecture. To easily describe the conjecture result, we introduce plausibility.

2.2 Plausibility Relation

Let X and Y be two finite sets, and R is the binary relation defined on $X \times Y$. If for any $x R y$, there exists a η, where $\eta \in [0,1]$, $x \in X$ and $y \in Y$, then R is called a binary plausibility relation on $X \times Y$ and η is the plausibility degree of R.

Obviously, plausibility is the extension of Fuzzy relation and Probabilistic Relation. When η is a membership value, R is a Fuzzy relation. And if η is probabilistic value, then R is a Probabilistic Relation. In general cases, η is correlativity measure, it can be applied to representing mined knowledge along with R. "Plausibility" means that it seems correct, plausibility implies that R seems be hold, but the degree of the holding depends on η. When $\eta = 1$, R is an ordinary relation; when $\eta = 0$, relation is false (there is no relation). Generally, when $0 < \eta < 1$, R seems hold, when η approaches to1, the confidence is bigger, otherwise, it is smaller.

Note: Professor Daphne Koller of Stanford University in USA proposed a Probabilistic Relational Models (PRM) in IJCAI99 Conference in 1999 (N. Friedman, L. Getoor, D. Koller and A. Pfeffer. 1999), and Bayes network methods is applied to studying this kind of Probabilistic Relational Models from the data directly, it has two ways of learning: parameter learning and structure learning.

Daphne Koller considered the result of Data Mining as one kind of probabilistic relation. But the probabilistic relation cannot represent all the cases of the relation, the indefinite of relation is not random event completely, it can be caused by fuzzy data, indistinct boundary, and incomplete data, so an accurate relation cannot be defined. For fuzzy data, the influences are discussed in the definition of fuzzy relation in Fuzzy mathematics, and the result has already been accepted. In addition, considering on the shortcomings of rough set theory, professor He Huanchan, a Chinese scholar, proposed the concept of "relational flexibility" . He points out that the relationship among fuzzy propositions can be continuously variable (He Huachan, Wang Hua, Liu Yonghuai, Wang Yongjun, Du Yongwen 2001). We wonder if the result of Data Mining is also one kind of "relational gentleness". The plausibility provides us a new idea.

The relation described by plausibility is indefinite, inaccurate or nondeterministic. It can be a probabilistic relation, a fuzzy relation, or a partial relation formed by incomplete data. There is no specific request in the definition of plausibility relation. So the plausibility relation can represent various mining results, such as association rules, classification rules, cluster rules and so on. Considering on the plausibility relation and the discussion of knowledge, we can get the following data mining formula:

Data Mining = Sets + Plausibility + Mining Algorithm

And the relationship among them is shown in Figure 1.

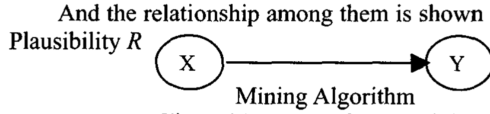

Figure1.Essence of Data Mining

Here, X and Y denote data sets respectively, they are mining objects; R is plausibility, it represents the mining result; Mining algorithm is a specific mining process composed by series of key computation steps. Thus, data mining can be abstracted to a simply fundamental model. From this model, it's possible to construct the theoretic framework of data mining by further studying and searching the key steps and consistencies of various mining methods.

2.3 Correlativity Measure —A computing method of plausibility

It is easy to understand the definition of plausibility, but the key point is how to compute the plausibility degree. Especially compute it in data mining. The authors proposed the concept of correlativity measure (Wang Xiaofeng, Wang Tianran 2002), and proved that the confidence of association rules, rough degree (one kind of fuzzy degree) in Rough sets are all correlativity measures, they are consistent in the form. In order to get the method of computing the correlativity measure, we will further discuss the relationship among plausibility relation, plausibility degree, and correlativity degree.

For discussion, the definition of correlativity measure is given first. (see Wang Xiaofeng, Wang Tianran 2002 for details).

Definition of Correlativity Measure

Suppose U is a finite set (Not empty), $P(U)$ is a power set, and R is a binary relation defined in $P(U)$. If $A \in P(U)$, for set $X \subseteq U$, there is a relation R defined between A and X, then $F_A(X)$: $P(U) \to [0, 1]$ is called the correlativity measure of set A and set X on relation R. Here, $F_A(X)$ is the mapping from $P(U)$ to [0, 1] reduced by relation R, and satisfies:

(1) $F_A(\Phi)=0$;
(2) $F_A(A)=1$;
(3) $0 \leqslant F_A(X) \leqslant 1$;
(4) $F_A(X_i \cup X_j) \leqslant F_A(X_i) + F_A(X_j)$

Where, X_i、$X_j \subseteq U$, there is relation R on A, X_i and X_j.

Note 1: R is the binary relation defined on power set or sets. For example, there is a common attribute relation between A and X, A includes X (containing relation) and so on. Through the relation, a relationship (or association) can be established between two sets. And the correlativity measure is the quantity of this association degree.

Note 2: The correlativity measure describes the numerical characteristic of the given relation and the correlativity set, but the real $F_A(X)$ is dependent on the specific definition of relations. If the relation is defined by some operations on the set, then it can be used as a $F_A(X)$ as long as it satisfies the definition of measures; If it is defined by language or characteristic declaration, then it must be described accurately by using operations on set. That is, transforming the relation into operation forms among sets. For example, let $U=\{t1, t2, t3, t4, t5\}$, $P(U)$ is the power set on U, R is the containing relation on $P(U)$. Then R can be represented accurately by: $R=\{<A, B>| A、B \in P(U), A \supseteq B\}$, the correlativity set of A *is* $[A]_R = \{B \mid A \supseteq B, B \in P(U)\}$, then $F_A(X) = |[X]_R \cap [A]_R| / |[A]_R|$ satisfies the definition of measures, and it is one kind of correlativity measures.

From the definition of the correlativity measure we know that, given a binary relation R, a mapping can be defined from $P(U)$ to [0, 1], and when the mapping satisfies some basic properties such as additive, it becomes the correlativity of relation R. Obviously, the value of this correlativity measure is a numerical result on [0, 1]. Comparing the definition of plausibility, it is easy to find that both of them are unified. The definition of plausibility only requires when any two elements of the set have the relation R, and there is a corresponding value (plausibility) between [0, 1]. The specific compute method and constraint condition are not given. But the correlativity measure has specific constraint conditions for the computing rules of $F_A(X)$. Plausibility describes the holding case of relation from the view of elements, whereas correlativity measure describes the confidence of relation from the view of sets, then how to integrate them? It should be noticed that in data mining, sometimes the same problem can be expressed in two different ways. For example, one concept may contain multiple attributes, and those attributes make up a set. The concept may belong to a concept set, and it is an element of the set. Thus, the concepts can be discussed from the view of elements or from the view of sets. That is, one concept can be an element belonging to a concept set, and it can also be a set containing attributes. For the same reason, an attribute is a set of attribute values, so it possesses the dual meaning. If there is a relation R between two concepts, then there are two ways to measure the holding of R: one is based on elements, the other one is based on sets. plausibility degree is a measure based on elements, and correlativity measure is based on sets. Both of them measure a relation, so the value should be the same. Their relation is shown in Figure2.

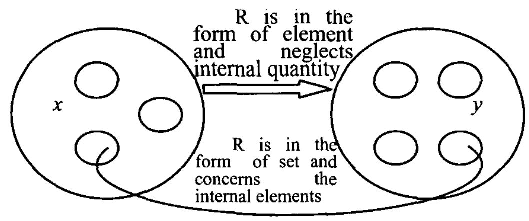

Figure2: Two expression of the same relation

The mining knowledge in database is a conceptual relation that uncovers the 'high level' from 'low level' attribute sets, a process from detailed data to abstract concept. Getting the plausibility of concepts from the correlativity measures of attribute sets is just the reflection of ideas of mining 'high level' conceptual ideas from the 'low level' data.

3 Correlativity Set - Foundations of Data Mining

As stated above, we know that the result of data mining can be represented as a plausibility relation, and the plausibility can be 'mined' from correlativity measures. Is this 'mining' common in data mining fields? Can it satisfy the current data mining methods? What is foundation of data mining theoretical framework.

3.1 Correlativity Measure— Foundations of Typical Data Mining Methods

In reference (Wang Xiaofeng,Yin Danna, Shihchuan Cheng 1999),we proved that the confidence of association rules is a correlativity measure, and it is based on the intersection operation of the set. The main ideas are shown below.

Let X and Y be two related item sets, the confidence of association rule X⇒Y in transaction database is the ration of transaction number including X item and Y item and the one including X item. It should be noticed that X and Y have dual meanings: from the notation view, they are item names; in addition, they are transaction sets comprising of items respectively. So

$$\text{confidence}(X \Rightarrow Y) = |X \cap Y| / |X|$$

But, from the definition of correlativity measure, we have

$$|X \cap Y| / |X| = F_A(X)$$

Where, $X \cap Y$ represents item X and item Y are in the same transaction, it is the intersection of two transaction sets comprising of item X and item Y, and X represents the transaction set comprising of item X. So

$$|X \cap Y| / |X| = (|X \cap Y| / |D|) / (|X| / |D|)$$

$$= P(X \cap Y) / P(X) = P(X|Y) * P(Y) / P(X)$$

$$= P(Y \mid X)$$

Here,

|D| is the total number of transaction in transaction database;

P(X) = |X| / |D| represents the possibility that item X occurs in transaction database;

$P(X \cap Y) = |X \cap Y| / |D|$ represents the possibility that item X and item Y occur in transaction database at the same time;

P(Y)= |Y| /|D| represents the possibility that item Y occurs in transaction database.

Therefore, $F_A(X) = \text{confidence}(X \Rightarrow Y) = |X \cap Y| / |X|$

$= P(Y \mid X) = P(X|Y) * P(Y) / P(X)$,

That is the confidence of rule X⇒Y is equal to conditional probability P(Y | X) and correlativity measure. We can also conclude that correlativity measure summarize the confidence and conditional probability.

Considering the relationship between Simple Bayesian Formula and conditional probability, since the Bayesian formula satisfies the definition of correlativity measure, therefore, the correlativity measure can be used to represent the Bayesian formula, the confidence and the conditional probability.

In addition, we proved that the accurate degree in rough sets is also one kind of correlativity measure. So, the important tools in classification mining —rough set theory is integrated with Bayesian Network and with Bayesian formula as the core, as well as is integrated with association rule mining and with confidence as the critical component. (Note that support is the special example of confidence). Those methods have the same correlativity measures.

Thus, correlativity measure provides a common theoretic framework for those typical data mining methods. This satisfies the first basic demand of ideal theoretic framework proposed by Jiawei Han and Kamber M.

In addition, by studying the correlativity sets, we also developed some efficient data mining methods, such as the dual space searching method based on correlativity set function, top-down mining frequent items using the property of reduction on rough set, and correlativity set method exclusively used on easing knowledge base (Wang Xiaofeng,Yin Danna, Shihchuan Cheng 1999), All of these methods include the mining of classification rules and association rules. All kinds of generalized rough sets, confidence, possibility computing and fuzzy sets can all be put into the framework of correlativity set. This reflects the universality of correlativity from another aspect.

3.2 General Process of Data Mining(GPDM)

Correlativity measure provides an abstract, formal computing formula for various data mining methods. From this formula and the knowledge representation of correlativity set, we have the following general data mining process.

Let universe U be the given finite set, each of its element is an attribute, and the various combinations of attributes make up different concepts. Suppose there are two concept sets C、D, $c_i \in C$, $d_j \in D$ are two specific concepts, then $c_i \subseteq U$, $d_j \subseteq U$ are all sets of attributes. Assume that relation $c_i \rightarrow d_j$ holds first, then verify if it is true. If the correlativity measure of $\boldsymbol{c_i \rightarrow d_j}$ is greater than the given confidence, then $c_i \rightarrow d_j$ holds, otherwise it fails. This algorithm is shown below.

GPDM (General Process of Data Mining):

(1) According to some rules, such as support of concept(s), generate the candidate concept(s) from the given universe, such as C or D. If there is no candidate concept(s), then end the data mining process.

(2) Select candidate concept attributes (or concept) from

concept(s), such as $c_i \in C$, $d_j \in D$.

(3) Try to generate a implication (or causality) such as $c_i \rightarrow d_j$.

(4) If the plausibility (got from correlativity measure) of concept attributes (or concept) satisfies the condition defined in advance, then the implication (or causality) $c_i \rightarrow d_j$ together with the plausibility constitute one rule (knowledge).

(5) Are there candidate concept attributes (or concept) that are not processed yet? If yes, go to (2), otherwise go to (1) to calculate the next group of concept(s).

Note that various candidate concept(s) processing method and supporting algorithm constitute various data mining algorithms.

For example, in decision tree methods, concept attributes are selected according to the size of information entropy. In correlativity set data mining method, the combination of conditional attribute values acts as candidate concept sets; The correlativity intensity between two kinds of attributes is computed depending on having the same object (intersection) relation, and determines whether the implication is hold.

In typical association rules mining method——Aprior Algorithm, item sets (concept attributes) are selected depending on the size of support degree. And depending on the confidence of rule X⇒Y in transaction database D determines whether the rule is hold.

In rough sets theory, reductions are determined depending on knowledge dependency value (one kind of measure). In Bayesian network method, candidate parameters are determined depending on the function of probability density distribution, and possibility value is used to judge whether the argument or structure are hold.

All of those methods satisfy GPDM (General Process of Data Mining), so GPDM summarize all kinds of mining methods including classification rules, association rules and so on, it is the basic framework of data mining.

4 Framework Architecture of GPDM System Oriented to Relational Database

For Data Mining, a field in relational database is an attribute, the set of some attributes represents a concept, a generalized tuple (record) is a specific instant or object of this kind of concept. Discovering knowledge in database using correlativity set method is to analyze the correlativity of attribute sets (concepts) and determining the plausibility relation. The process of GPDM has been discussed, now the framework architecture of GPDM system will be given.

Let $S=\{U, A, H, V, f\}$ be a specific information system. Where

U: a finite set of objects, $U=\{x_1, x_2, \ldots, x_n\}$.

A: a finite set of attributes,, $A=C\cup D$, C is a set of conditional attributes, D is a set of decision attributes.

H: a finite set of concepts, $H=\{A_1, A_2, \ldots, A_m\}$, $A_i \subseteq A$ is a part of attribute sets.

V: $V=\bigcup_{p\in A} V_p$, V_P is a domain of attributes P.

f: $U\times A \rightarrow V$ is a function from object attributes to domain. If $p \in A$, $x \in U$, then $f(x,p) \in V_P$

A specific relational database is this kind of information system, where field is attribute; row is object, and the value of field p in the x^{th} row is the attribute value of object x corresponding to p. The combination of several fields represents a concept.

According to the GPDM process introduced above, the architecture of GPDM system is shown in Figure 3. Here data mining is divided into 5 levels: concept level, attribute level, attribute value level, attribute value-object level and object value.

Here, the concept set to be mined that makes up the concept space. Through candidate rules, concepts (attribute sets), which may have some relation, and can be put together to form a specific mining object by concept space.

Concept is a set of attributes;concept is represented by attributes, and all the sets of concept attributes make up the attribute level. Each attribute is a set composed of several attribute values, and an attribute value represents an attribute, and all attribute values make up the attribute value level together.

In general sense, all the sets of tuples (records) constitute the object level. Objects that have the identical attribute values are put together to constitute a set and represent attribute values. All these attribute value-object sets make up a set class: attribute value –object level. The concept level is an abstract level that is suitable to the thought of human beings and it is on the top of the architecture. Object level is the data of the database, suitable to machine working. Through the intermediate level, the relation hold on `concepts can be transformed into the plausibility among tuple objects. On the attribute value-object level, the correlativity measure of sets determines whether the plausibility relation is true. If it holds, then it will be passed to concept level as the result of data mining.

It should noticed that the concept is the reflection of objects in human's mind, the concept space is oriented to people, whereas attributes, attribute values and tuples are data stored in database. The transform process between concepts and attributes is a human-machine interactive process. The relation between concepts or between concepts and attributes can be determined in advance. According to some rules (program), attributes can also be organized to constitute concepts using computer. The former is the 'concept tree' constructed by people, and the latter is the result induced or deduced by computer. Membership Cloud method proposed by academician Li deyi (Li Deyi, Meng Haijun, Shi Xuemei 1995) is a transformation method between levels, Bayes network (Jiawei Han, Micheline Kamber 2000, and Y.Y. Yao *2001*), is a specific hierarchical structure induce method. According to the structure showed in Figure 3, the correlativity set algorithm for KDD (Wang Xiaofeng, Yin Danna, Shihchuan Cheng 1998, and Wang Xiaofeng, Tang Zhong 2000) searches for the implication of concepts, instead of searching for the containing relation of correlative object sets. The correlativity intensity is used to determine whether the containing relation between condition object sets and decision object sets holds.

5 Conclusions

The plausibility relation is the extension and generalization of fuzzy relation and probability relation, and data mining discovers the plausibility relation implied by data from database. Correlativity measure is a specific plausibility degree. By using correlativity measures, the critical calculate in data mining such as rough degree of rough sets, confidence of association rules (Note: supports is a special case of confidence) and Bayesian formula can be integrated. Base on this, a General Process of Data Mining (GPDM) is created to represent the essence of data mining. Based on the correlativity set, We consider that a theoretical foundation and framework of data mining may be constructed to describe the essence of data mining. This framework satisfies the basic demand proposed by Jiawei Han and Kamber M, containing current typical mining methods, and becomes the important theory foundation for developing data mining language and modeling. The research proposed in this paper is elementary, and there are some problems that need study further such as the property of plausibility relation, the formalization of data mining, and the transformation between data mining which have different formats or different structures (especially for semi-structure on non structure).

Acknowledgements

This paper obtained national 863 plan (2001AA413410)and sponsored by LiaoNing Province natural science funds(9910200205).Wang xiaofeng , Professor, research direction: Data Mining, Intelligence Information Process;

References

[1] R. Agrawal, T. Imielinski, and A. Swami (1993) Mining association rules between sets of items in large databases. Proceedings of the ACM SIGMOD Conference on Management of data, pp. 207-216, .

[2] N. Friedman, L. Getoor, D. Koller and A. Pfeffer. (1999), Learning Probabilistic Relational Models,T *Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI)*, Stockholm, Sweden, pages 1300-1307.

[3] Jiawei Han, Micheline Kamber. (2000). Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers

[4] Y.Y. Yao *(2001)*, On modeling data mining with granular computing, *Proceedings of the 25th Annual International Computer Software and Applications Conference (COMPSAC 2001),* Chicago, USA, IEEE Computer Society, Los Alamitos, California, pp. 638-643.

[5] Shi Zhongzhi (2002), Knowledge Discovery, Tsinghua University Publishers.

[6] Wang Xiaofeng,Yin Danna , Shihchuan Cheng (1999), Mutuality Sets, Journal of Shenyang Institute Chemical Technology, 67-76.

[7] Wang Xiaofeng, Wang Tianran (2002), Correlativity Measure and Incremental Computation of Support and Confidence , Journal of software.

[8] Wang Xiaofeng, Wang Tianran (2002), Methods of mining frequents based dual space searching , Computer Science

[9] Wang Xiaofeng,Yin Danna , Shihchuan Cheng (1998), Correlativity set and its application in simplified knowledge base, Journal of Tsinghua University

[10] Wang Xiaofeng, Tang Zhong (2000), Correlativity set and its application in knowledge discovery from database, Journal of Nanjing University, Computer special (36)11, 52-57

[11] He Huachan, Wang Hua, Liu Yonghuai, Wang Yongjun, Du Yongwen (2001), Universal Logics Principle, Science Publisher

[12] Li Deyi, Meng Haijun, Shi Xuemei (1995), Membership cloud and membership cloud generator, Computer research and development, 32 (6)

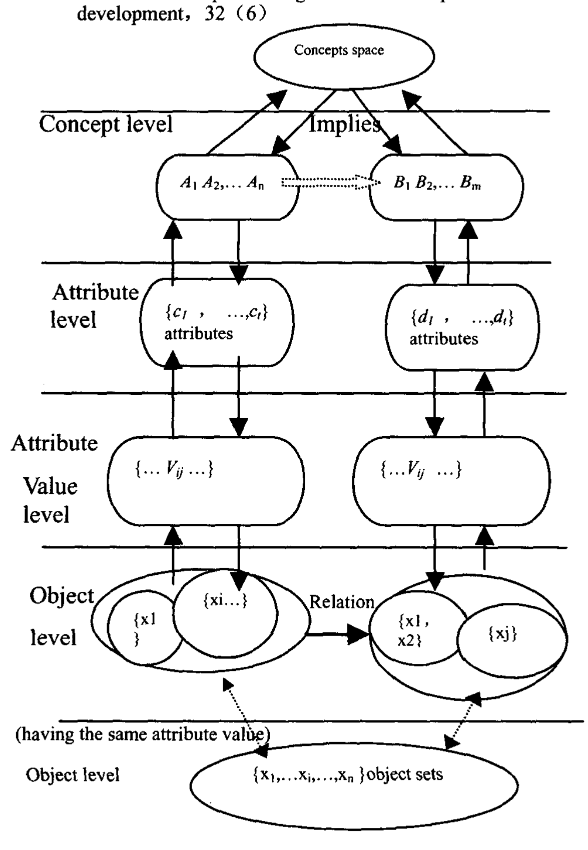

Figure 3: The Structure Model of GPDM System