ISSN1000-9825 Journal of Software 软件学报 2000,11(2):143～158

Mining Association Rules with Linguistic Cloud Models*

LI De-yi[1] DI Kai-chang[2] LI De-ren[2] SHI Xue-mei[3]

[1](Institute of China Electronic System Engineering Beijing 100036)

[2](School of Information Engineering Wuhan Technical University of Surveying and Mapping
Wuhan 430070)

[3](Department of Computer Science Hong Kong Polytechnic University Hong Kong)

E-mail: kcdi@public3.bta.net.cn

Abstract This paper presents linguistic cloud models for knowledge representation and uncertainty handling in KDD. Multi-dimensional cloud models are introduced as the extension of one-dimensional ones. The digital characteristics of linguistic clouds well integrate the fuzziness and randomness of linguistic terms in a unified way. Conceptual hierarchies based on the models can bridge the gap between quantitative knowledge and qualitative knowledge. In order to discover strong association rules, attribute values are generalized at higher concept levels, allowing overlapping between neighbor attribute values or linguistic terms. And this kind of soft partitioning can mimic human being's thinking, while making the discovered knowledge robust. Combining the cloud model based generalization method with Apriori algorithm for mining association rules from a spatial database shows the benefits in effectiveness, efficiency and flexibility.

Key words Linguistic cloud model, association rules, Apriori algorithm, virtual cloud, spatial data mining.

KDD (knowledge discovery in databases), or data mining, is the non-trivial extraction of implicit, previously unknown, and potentially useful knowledge from large number of small, very specific instances[1,2]. Mining association rules is an important issue in applications. A typical example of an association rule in retail industry is like "90% customers who by bread and butter also buy milk". Such rules are very useful for marketers to develop and implement customized marketing programs and strategies.

The problem of mining association rules was first introduced in Ref. [3]. Since then, various algorithms have been proposed and developed, such as AIS[3], SETM[4], Apriori[5], DHP[6], FUP[7], FUP2[8], etc. Among the algorithms reported, Apriori is an influential one. It makes multiple passes over the transaction data for discovering large itemsets. In the first pass, it counts the supports of individual items and determines the

* The research is supported by the National Natural Science Foundation of China (国家自然科学基金,No. 49631050) and by the National Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing (测绘遥感信息工程国家重点实验室基金,WKL(97)0302). **LI De-yi** was born in 1944. He is a professor and doctoral supervisor of the Institute of China Electronic System Engineering. His current research areas include computer engineering and artificial intelligence. **DI Kai-chang** was born in 1967. He is a Ph.D. student at School of Information Engineering, Wuhan Technical University of Surveying and Mapping (WTUSM). His current research interests include spatial data mining and knowledge discovery, integration of remote sensing and Geographic Information System (GIS). **LI De-ren** was born in 1939. He is a professor and doctoral supervisor of Wuhan Technical University of Surveying and Mapping. He is a member of both the Chinese Academy of Sciences and the Chinese Academy of Engineering. His current research areas include remote sensing, GIS, Global Positioning System (GPS) and their integration. **SHI Xue-mei** was born in 1967. He received a Ph.D. degree in Computer Science from Hong Kong Polytechnic University in 1998. His current research interests are data mining and financial analysis.

Manuscript received 1997-12-26, accepted 1998-08-05.

large 1-itemsets. In each subsequent pass, a number of candidate large items are generated based on the large itemsets found in the last pass, and then the actual supports for the candidate itemsets are counted by scanning the data. The process continues until no new large itemsets are found. Apriori outperforms AIS and SETM by a factor of three for small problems to more than an order of magnitude for large problems[5]. DHP algorithm uses hashing technique to further reduce the number of candidate itemsets[6].

Srikant and Agrawal introduced the problem of mining quantitative association rules in large relational tables containing both quantitative and categorical attributes[9,10]. The association rules mined from transaction data are referred to as Boolean association rules because a transaction consists of binary attributes (items). They mapped the quantitative association rules problem into Boolean association rules by representing each attribute with as many fields as the number of attribute values in stead of just one field. If the domain of values for a quantitative approach is large, one can first partition the values into intervals and then map each 〈attribute, interval〉 pair to a Boolean attribute. And then, any algorithm for finding Boolean association rules can be used to find quantitative association rules.

The problem of maintaining association rules, or incremental mining of association rules, was first studied in Refs. [7,11]. The FUP algorithm was proposed to update the association rules in a database when new transactions were added to the database. It stores the counts of all the large itemsets found in a previous mining operation to reduce the work of generating candidate large itemsets. FUP2 is a generalization of FUP. It can update the existing association rules when transactions are added and deleted from the database[8]. Both FUP and FUP2 are efficient.

The above algorithms all discover association rules at a single concept level. The problem of mining multiple level association rules was studied in Ref. [12]. It extended the existing single level association rule mining algorithms and explored the techniques for sharing data structures and intermediate results across levels. A similar work was reported in Ref. [13]. It discovered association rules at any level of the taxonomy by extension of the Apriori algorithm.

Parallel mining of association rules[14~16], meta rule guided mining of multidimensional association rules[17], mining of spatial association rules[18], have also been studied recently.

As a whole, researchers in KDD pay much attention to developing and implementing efficient data mining techniques, while studies on knowledge representation in KDD are seldom reported[19~22]. To make the knowledge mined from databases understandable and similar to the description of human experts, we need a representation that can integrate quantitative knowledge and qualitative knowledge and handle uncertainty with randomness and fuzziness. In Refs. [23,24], a linguistic cloud model was presented to meet these requirements. The mathematical description of a cloud integrates the fuzziness and randomness of linguistic terms in a unified way. Knowledge representation based on this model bridges the gap between qualitative knowledge and quantitative knowledge. Mapping between quantitatives and qualitatives knowledge becomes much easier and interchangeable[24].

The rest of this paper is organized as follows. In Section 1, we describe the linguistic cloud model and extend it to two- or multi-dimensional ones. In Section 2, we explore the application of the model to mine association rules at any higher level of conceptual hierarchies from a spatial database by means of Apriori algorithm. An experimental result is given to show the discovered association rules at higher levels mined from a Chinese geospatial and economic database. The conclusions are briefed in Section 3.

1 A Linguistic Cloud Model

1.1 Linguistic atoms and compatibility functions

Our point of departure in this paper is to represent linguistic terms in logical sentences or rules. We imagine a linguistic variable that is semantically associated with a list of all the linguistic terms within a universe of discourse. For example, "age" is a linguistic variable if its values are "young", "middle age", "old", "very old", and so forth, rather than the natural numbers. In the more general cases, a linguistic variable is a quintuple:

$$\{X, T(x), U, S, C_X(u)\}$$

X is the name of the variable. $T(x)$ is the term-set of X; that is, the collection of its linguistic values. U is a universe of discourse. S is a syntactic generator that generates the terms in $T(x)$. $C_X(u)$ is a compatibility function. It denotes the relationship between a term x in $T(x)$ and U. More precisely, the compatibility function maps the universe of discourse into the interval $[0,1]$ for each u in U.

The compatibility value has a relationship with both fuzzy logic and probability. Consider a set of linguistic terms T in a universe of discourse U, for example, the linguistic term "young" in the interval $[0,100]$. The compatibility of "28 years old" with "young" is 0.7. With the fuzzy logic point of view, the compatibility value "0.7" is an indication of the partial membership with which the element "age-value 28" belongs to the fuzzy concept "young". To understand the relationship with probability on the other hand, the correct interpretation of the compatibility value "0.7", given by one's conception, is that it is merely subjective indication. Human knowledge does not conform to such a fixed crisp membership degree "0.7" at "age-value 28". But there is always a tendency showing that the membership degree at "age-value 28" is a stable random number, under which a subjective probability distribution is obeyed. The degree of compatibility takes on random values itself. This type of randomness is adhered to the fuzziness.

Regarding syntactic generation, we shall usually assume that a linguistic variable is structured in the sense that it is associated with two rules. The first is the atom generator rule. It specifies the manner in which a linguistic atom, which cannot be sliced into any smaller parts, may be generated. The second, the semantic rule, specifies a procedure for computing composite linguistic terms, based on linguistic atoms.

In addition to linguistic atoms, a linguistic term may involve connectives (such as "and", "or", "either", and "neither"), the negation ("not") and the hedges (such as "very", "more or less", "completely", "quite", "fairly", "extremely" and "somewhat"). The linguistic connectives, hedges and negation may be treated as (some form of) soft operators that modify the meaning of their operands——linguistic atoms, in a soft computing fashion to become composite linguistic terms. That is the business of the semantic rule.

1.2 The concept of linguistic clouds

Cloud model is a model of the uncertain transition between a linguistic term of a qualitative concept and its numerical representation. In short, it is a model of the uncertain transition between qualitatives and quantitatives.

Let U be the set $U=\{u\}$, as the universe of discourse, and T a linguistic term associated with U. The membership degree of u in U to the linguistic term T, $C_T(u)$, is a random number with a stable tendency. $C_T(u)$ takes the values in $[0,1]$. A compatibility cloud is a mapping from the universe of discourse U to the unit interval $[0,1]$. That is,

$$C_T(u): U \to [0,1], \ \forall\, u \in U \quad u \to C_T(u).$$

Figure 1 shows appropriate compatibility clouds for the linguistic terms "about 20", "teenager" and "twenty-something" from the term-set of the linguistic variable "Age". The geometry of the compatibility cloud is a great aid in understanding the uncertainty of the transition between a linguistic term and its numerical representation. First of all, the mapping from all $u \in U$ to the interval [0,1] is a one-point to multi-point transition, producing a compatibility cloud, rather than a membership curve. Traditionally, the membership function of a fuzzy set is a one-point to one-point mapping from a space U to the unit interval [0,1]. After the mapping the uncertainty of an element belonging to the fuzzy concept becomes certain to that degree, a precise number. The uncertain characteristics of the original concept are not passed on to the next step of processing at all. Secondly, any particular drop of the cloud may be paid little attention to. However, the total shape of the cloud, which is visible, elastic, boundless and movable, is most important. That is why we use the terminology "cloud" to name it. Thirdly, the MEC (mathematical expected curve) of a compatibility cloud may be considered as its membership function from the fuzzy set theory point of view. Finally, the definition has effectively integrated the fuzziness and randomness of a linguistic term in a unified way.

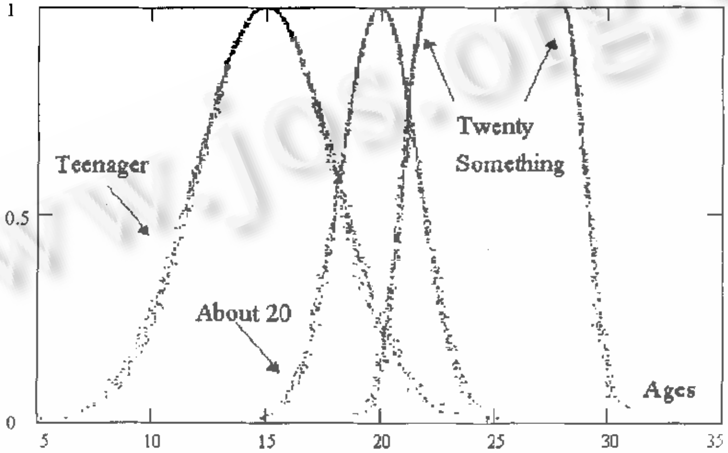


Fig. 1 Various linguistic terms for the linguistic variable "Age"

1.3 Digital characteristics of a cloud

The normal compatibility clouds are most useful in representing linguistic atoms because normal distributions have been supported by results in every branch of both social sciences and natural sciences. A normal compatibility cloud characterizes the qualitative meaning of a linguistic atom with three digital characteristics:

$$A(Ex,En,D)$$

where Ex, En, and D are the expected value, entropy and deviation of the cloud respectively. A given set $\{Ex, En, D\}$ uniquely defines a particular compatibility cloud A.

The expected value Ex of a compatibility cloud is the position at U corresponding to the center of gravity of the cloud. In other words, the element Ex in the universe of discourse is fully compatible with the linguistic atom A. It is very easy to determine Ex in practical applications.

The entropy En of a linguistic atom is a measure of the fuzziness of the concept within the universe of discourse. The entropy of a linguistic atom is defined by the bandwidth of the MEC (mathematical expected curve) of the normal compatibility cloud showing how many elements in the universe of discourse could be accepted to the linguistic atom. The MEC of the normal compatibility cloud to a linguistic atom A is:

$$MEC_A(u)=\exp\left[-\frac{(u-Ex)^2}{2En^2}\right].$$

The deviation D is a measure of randomness of the compatibility function. Looking at the normal compatibility cloud in detail we see that its thickness is uneven. Close to the waist, the degree of compatibility is most dispersed, while at the top and bottom the focusing is much better. Therefore, the maximum deviation D really comes from the randomness of the compatibility degree at the waist part of the cloud.

It should be noticed that the top, bottom, and waist of a cloud, however, do not need to be precisely defined at all. The three digital characteristics are good enough to represent a normal compatibility cloud.

1.4 Two-Dimensional and multi-dimensional linguistic clouds

Following the above ideas, we extend the linguistic cloud model to two-dimensional one. Let U be the set $U=\{x,y\}$, as the universe of discourse, and T a linguistic term associated with U. The membership degree of u in U to the linguistic term T, $C_T(x,y)$, is a random number with a stabletendency. $C_T(x,y)$ takes the values in $[0,1]$. A two-dimensional compatibility cloud is a mapping from the two-dimensional universe of discourse U to the unit interval $[0,1]$. That is,

$$C_T(x,y):U\to[0,1],\ \forall\ (x,y)\in U(x,y)\to C_T(x,y).$$

The concept of 2-D clouds is pictured as three-dimensional graphics. Figure 2 is a surface plot of a 2-D cloud corresponding to the linguistic term "central". We can see that it is like a grave mound or a hill, it is smooth and slopes gently at the top and foot, but it is rugged and rough at the hill-side. This indicates that the degree of compatibility is more dispersed at the hill-side, while more focused at the top and foot. Therefore, the 2-D cloud is the natural extension of the 1-D cloud.

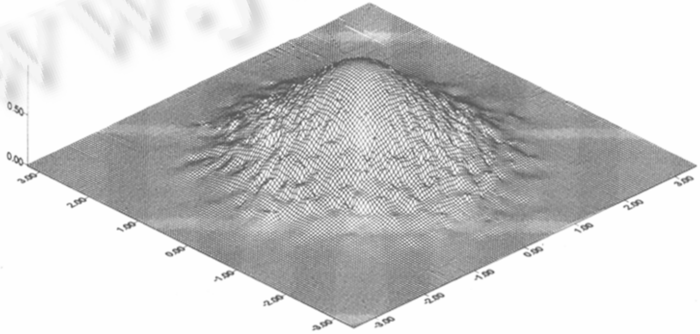


Fig. 2 A 2-D cloud corresponding to the linguistic term "central"

Suppose the two dimensions of a universe of discourse are independent, then the two-dimensional normal compatibility cloud for a linguistic term in the universe is characterized with six digital characteristics:

$$A(Ex,Enx,Dx,Ey,Eny,Dy)$$

where Ex and Ey are the expected values, Enx and Eny are entropies, and Dx and Dy are the deviations in the two dimensions x and y respectively. Similar to that in one-dimension case, (Ex,Ey) corresponds to the center of gravity of the 2-D cloud. Enx and Eny are the measures of the fuzziness of the concept. They are defined by the bandwidths of the MES (mathematical expected surface) of the 2-D cloud. The MES of the 2-D normal compatibility cloud to a linguistic atom is:

$$MES_A(x,y)=\exp\left[-\frac{1}{2}\left[\frac{(x-Ex)^2}{Enx^2}+\frac{(y-Ey)^2}{Eny^2}\right]\right],$$

where Dx and Dy are the measures of randomness of the two-dimensional compatibility function.

The projection of an MES to the $x-y$ plane is an ellipse (or a circle if Enx equals Eny). When the axes of the ellipse are not parallel to x and y axes respectively, we may add a digital characteristic θ to describe the cloud which is the angle between the corresponding axes. This cloud is referred to as rotated cloud, while the unrotated cloud is considered to be standard cloud. Figure 3 shows the shadows of the drops of a rotated cloud. If (x_i,y_i,μ_i) are the drops of a standard 2-D cloud, then (x'_i,y'_i,μ_i) are the drops of the rotated cloud, where (x'_i,y'_i) are computed as follows,

$$x'_i=(x_i-Ex)\cos\theta-(y_i-Ey)\sin\theta+Ex,$$
$$y'_i=(x_i-Ex)\sin\theta+(y_i-Ey)\cos\theta+Ey.$$

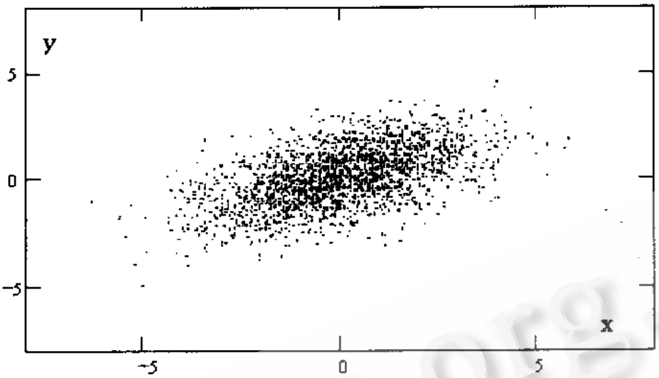


Fig. 3 Shadows of a rotated 2-D cloud

We can extend the linguistic cloud model to multi-dimensional one further. Suppose a universe of discourse has n dimensions that are independent from each other. The n-dimensional normal compatibility cloud for a linguistic term in the universe is characterized with $3n$ digital parameters:

$$A(E_1,En_1,D_1,E_2,En_2,D_2,\ldots,E_n,En_n,D_n),$$

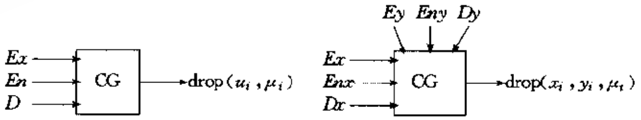
where E_1, E_2, ..., E_n are the expected values, En_1, En_2, ..., En_n are the entropies which represent the fuzziness of the concept, and D_1, D_2, ..., D_n are the deviations which are the randomness measures of the concept. The MEHS (mathematical expected hyper surface) of the multi-dimensional normal compatibility cloud to a linguistic atom is:

$$MEHS_A(x_1,x_2,\ldots,x_n)=\exp\left[-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-E_i)^2}{En_i^2}\right].$$

The two- and multi-dimensional cloud models are mainly used for two purposes in data mining: knowledge representation and soft computing. For example, "southeast" is represented by a 2-D cloud, and "warm color" is represented by a 3-D cloud because a color is composed of three components: red, green and blue. The soft "and" operation of "high education" and "high income" can be computed by a 2-D cloud model. The issue of cloud model based soft computing will be discussed in a further coming paper.

1.5 Cloud generators

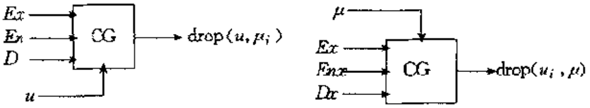
Given three digital characteristics Ex, En, and D to represent a linguistic atom, the generator could produce as many drops of the cloud as you like. All the drops obey the properties described above. Figure 4(a) shows a 1-D cloud generator. Correspondingly, given six digital characteristics Ex, Enx, Dx, Ey, Eny, and Dy, to represent a two-dimensional linguistic atom, the 2-D cloud generator could produce any number of drops of clouds. Cloud drops are three-dimensional points (x_i,y_i,μ_i), where $(x_i,y_i)s$ obey a two-dimensional normal distribution and $\mu_i's$ obey a one-dimensional distribution. Figure 4(b) shows a 2-D cloud generator.



(a) 1-D model (b) 2-D model

Fig. 4 Cloud generator

Cloud drops may be generated on conditions. Figure 5(a) shows the generator producing drops under a given numerical value u in the universe of discourse, U; while Fig. 5(b) shows the generator under the condition of a given membership degree μ. All the drops generated in Fig. 5(a) have the same value u in the universe of

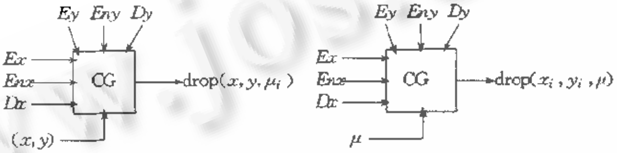


(a) On the condition of "*u*"　　(b) On the condition of "*μ*"

Fig. 5　Generator on conditions (1-D model)

discourse, and normal distributed membership degree μ_i; whereas all the drops generated in Fig. 5(b) have the same membership degree μ, and normal distributed numerical values u_i in the universe of discourse. A fuzzy rule in a spatial database, such as, "If elevation is low, then the road density is high", can be represented by two forward cloud generators under conditions of attribute values and membership degrees respectively.

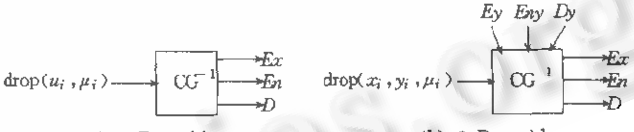
Similarly, 2-D cloud generator may produce cloud drops on conditions. All the drops generated in Fig. 6(a) have the same value (x, y), and normal distributed membership degrees μ_i; whereas all the drops generated in Fig. 6(b) have the same membership degree μ, and normal distributed values (x_i, y_i) along the membership ellipse related to μ.



(a) On the condition of "(x, y)"　　(b) On the condition of "*μ*"

Fig. 6　Generator on condition (2-D model)

It is natural to think about the generator mechanism in an inverse way. Given a limited set of drops, drop$_i(u_i, \mu_i)$, as samples of a compatibility cloud, the three digital characteristics Ex, En, and D could be produced to represent the corresponding linguistic atom (see Fig. 7(a)). The 2-D backward cloud generator is illustrated in Fig. 7(b). The two kinds of atom generators may be called forward and backward generators respectively. The atom generators implemented in both hardware and software have been a patented invention in China. The combination of the two kinds of generators can be used interchangeably to bridge the gap between quantitative knowledge and qualitative knowledge.



(a) 1-D model　　(b) 2-D model

Fig. 7　Backward cloud generator

When rules are discovered for databases, we can use the uncertainty reasoning mechanism in cloud theory for predictive data mining. When an input value enters the rule group, an uncertainty value is produced by combing the conditional forward and backward cloud generators and virtual cloud construction methods. A significant advantage over the conventional prediction in data mining is that there is an inherent uncertainty in the case of chaining rules.

2　Mining Association Rules with Linguistic Cloud Model

2.1　Problem description

Let $I=\{i_1, i_2, \ldots, i_m\}$ be a set of literals, called items. Let D be a database of transactions, where each transaction T is a set of items such that $T \subseteq I$. For a given itemset $X \subseteq I$ and a given transaction T, we say that T contains X if and only if $X \subseteq T$. An association rule is an implication of the form $X \Rightarrow Y$, where $X \subseteq I$, $Y \subseteq I$ and $X \cap Y = \varnothing$. The rule $X \Rightarrow Y$ holds in the databases D with confidence c if $c\%$ of transactions in D that contain X also contain Y. The rule $X \Rightarrow Y$ has support s in the database D if $s\%$ of transactions in D contain $X \cup Y$.

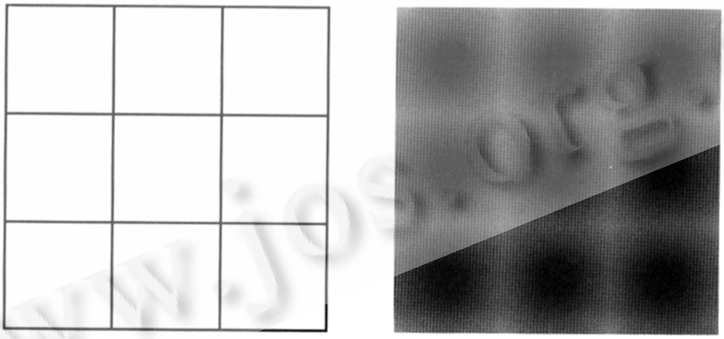
Given a database of transactions D, the problem of mining association rules is to generate all association rules that have support and confidence greater than user-specified minimum support and minimum confidence respectively.

2.2 Attribute generalization based on linguistic clouds and virtual clouds

2.2.1 Attribute Generalization

Generally, strong rules are likely to exist at high concept levels. Especially when attributes are numerical, mining at the primary level may not generate strong rules with large minimum support and minimum confidence thresholds or may generate many uninteresting rules with small thresholds. In this case, it is preferable to generalize the attributes first and then mine rules with the generalized data. This kind of attribute generalization problem is also called continuous data discretization.

The commonly used method for continuous data discretization is to map the continuous data to nonoverlapping intervals or areas, that is to say the attribute spaces are partitioned crisply. Overlaps between neighbor intervals are not allowed in this method, so we refer to this method as crisp partition method. Figure 8(a) shows a crisp partition of a two-dimensional attribute space. The space is partitioned into nine nonoverlapping rectangles, each rectangle corresponds to a unique label which can be considered to be a linguistic term or a generalized attribute value. But if an expert makes the partition, uncertainty is always adhered to it. In natural language, the linguistic terms associated with a linguistic variable often overlap and have ambiguous boundary. For example, "young" and "middle age" have some overlap, and it is difficult to partition them with a crisp age value. Different experts may make different partitions on one hand, and these partitions have stable tendencies on the other hand. Human beings can partition the attribute space flexibly and allow overlapping area between neighbor linguistic terms. Obviously, the partition mechanism of human being is not mimicked by the crisp partition method.



(a) Crisp partition (b) Soft partition

Fig. 8 Attribute space partition

With the linguistic cloud model, each attribute (or several relevant attributes) is considered as a linguistic variable. Several linguistic clouds are given for each linguistic variable to represent linguistic terms of the linguistic variable. The clouds may be interactively specified by the user or automatically generated by analyzing the histogram of the attribute values. Our method allows overlapping area between neighbor clouds. Figure 8 (b) shows the same attribute space partitioned by nine 2-D linguistic clouds. At the center of each cloud, the corresponding attribute value is fully compatible with the linguistic term, which is pictured by black points. While at the middle of two neighbor clouds, the image is gray, it indicates the overlapping area and ambiguous boundary between the two clouds. So the cloud based generalization method can be referred to as a soft partition method.

Suppose clouds $A_1(Ex_1, En_1, D_1)$, $A_2(Ex_2, En_2, D_2)$, ..., $A_N(Ex_N, En_N, D_N)$ are given for a numerical attribute. Input the attribute value u to the cloud generators $CG_1, CG_2, \ldots, CG_N$, and we get the outputs μ_1, $\mu_2, \ldots, \mu_N$, which are the compatibility values of u with $A_1, A_2, \ldots, A_N$ respectively. The maximum compatibility value is retrieved, and u is assigned to A_i if μ_i is the maximum. If two compatibility values, say μ_i and μ_j, are both equal to the maximum, u is assigned to A_i or A_j randomly. This mechanism is illustrated in Fig. 9. In a similar way, two-dimensional attribute space can be partitioned by 2-D clouds softly. For example, x and y positions are generalized to two-dimensional linguistic terms, "southwest", "southeast", "northeast", "central", etc.

We know from the properties of the cloud generator that μ_1, $\mu_2, \ldots, \mu_N$ are random numbers with stable tendency, rather than fixed values. Therefore, the same attribute values within the overlapping area of two neighbor clouds may be assigned to different clouds at different occasions. This is just like what human beings do, thus the soft partition method can mimic human being's thinking better than the crisp partition method.

Not only numerical attributes but also the nominal and categorical attributes, can be generalized with the cloud model, and the attributes can be generalized to multiple concept levels. This is the work of virtual linguistic atoms and virtual clouds.

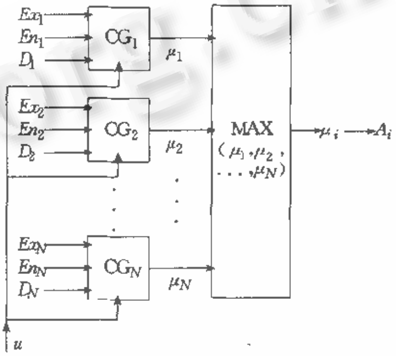


Fig. 9 The mechanism for assigning an attribute value to a linguistic atom

2.2.2 Virtual clouds

In our previous papers, we have introduced the concept of virtual linguistic atom and presented the virtual cloud construction methods[24,25]. There are two kinds of linguistic clouds: floating clouds and synthesized clouds.

Suppose we have two neighbor linguistic atoms, $A_1(Ex_1, En_1, D_1)$ and $A_2(Ex_2, En_2, D_2)$, over the same universe of discourse U. A virtual floating linguistic atom, $A(Ex, En, D)$, may be located at any point u between the two clouds in U using the following definition:

$$Ex = u,$$
$$En = \frac{En_1(Ex_2 - Ex) + En_2(Ex - Ex_1)}{Ex_2 - Ex_1},$$
$$D = \frac{D_1(Ex_2 - Ex) + D_2(Ex - Ex_1)}{Ex_2 - Ex_1}.$$

From a geometrical point of view, this definition satisfies the property that when the virtual cloud is floating towards A_1, it will be more and more affected by A_1, while less and less affected by A_2, till totally overlapped at the position of A_1, and vice versa. In other words, the virtual cloud constructed at u between Ex_1 and Ex_2 in the universe of discourse is a balance from both sides of A_1 and A_2 on a distance based weighting.

Floating clouds are very useful when the user specified clouds for a linguistic variable are not enough to cover the universe of discourse. There may be some blank areas between some clouds. As a result, the maximum compatibility values of the attribute within these areas must be equal to zero or very small. These attribute values can not be assigned to any cloud. Generating floating clouds in the blank areas can easily solve the above problem. In fact, the only indispensable work of the user is to specify key clouds at the key positions. Other clouds can be automatically generated by the floating cloud construction method.

The mechanism of 1-D floating cloud construction can be extended to 2-D one naturally. Suppose we have three neighbor linguistic atoms, $A_i(Ex_i, Enx_i, Dx_i, Ey_i, Eny_i, Dy_i)$, $i = 1, 2, 3$, over the same 2-D universe of discourse U. $P_1(Ex_1,Ey_1)$, $P_2(Ex_2,Ey_2)$ and $P_3(Ex_3,Ey_3)$ form a triangle $\Delta P_1P_2P_3$. A 2-D virtual floating linguistic atom, $A(Ex,Enx,Dx,Ey,Eny,Dy)$ may be located at any point $p(x,y)$ in the triangle. The floating cloud is constructed based on the following idea: points (Ex_1,Ey_1,z_1), (Ex_2,Ey_2,z_2), (Ex_3,Ey_3,z_3) form a plane, the z value at $p(x,y)$ is determined by the intersection point of the plane and the line that passing through p and perpendicular to the $x-y$ plane, where z represents Enx, Eny, Dx and Dy. The plane equation is:

$$\begin{vmatrix} x-Ex_1 & y-Ey_1 & z-z_1 \\ Ex_2-Ex_1 & Ey_2-Ey_1 & z_2-z_1 \\ Ex_3-Ex_1 & Ey_3-Ey_1 & z_3-z_1 \end{vmatrix}=0.$$

We can easily get the explicit representation by expanding the equation and replacing z by Enx, Eny, Dx, Dy respectively. We omit the expanding process here.

The 2-D virtual cloud is a balance for the three clouds. The closer to one cloud, the more influence by that cloud. Given the key clouds over a universe of discourse, a TIN (triangulated irregular network) can be automatically constructed to control the generation of floating clouds. When a point $p(x,y)$ is given to generate a floating cloud, the triangle which contains p is retrieved from the TIN first, and then the floating cloud is constructed by the three clouds corresponding to the vertices of the triangle.

A synthesized cloud is used to synthesize linguistic terms into a generalized one. For example, the concept of "teenager" may be considered as the parent node of the concepts "about 14 years old" and "about 18 years old" in a concept hierarchy.

Suppose we have two neighbor linguistic atoms, $A_1(Ex_1,En_1,D_1)$ and $A_2(Ex_2,En_2,D_2)$, over the same universe of discourse U. A virtual linguistic atom, $A(Ex,En,D)$, may be created by synthesizing the two atoms using the following definition:

$$Ex=\frac{Ex_1En'_1+Ex_2En'_2}{En'_1+En'_2},$$
$$En=En'_1+En'_2,$$
$$D=\frac{D_1En'_1+D_2En'_2}{En'_1+En'_2},$$

where En'_1 and En'_2 are calculated as follows.

Suppose $MEC_{A_1}(u)$ and $MEC_{A_2}(u)$ are the mathematical expected curves of A_1 and A_2 respectively. Let

$$MEC_{A_1}(u)=\begin{cases} MEC_{A_1}(u), & \text{when } MEC_{A_1}(u)\geqslant MEC_{A_2}(u) \\ 0, & \text{otherwise} \end{cases},$$
$$MEC_{A_2}(u)=\begin{cases} MEC_{A_2}(u), & \text{when } MEC_{A_2}(u)\geqslant MEC_{A_1}(u) \\ 0, & \text{otherwise} \end{cases},$$

then,

$$En'_1=\frac{1}{\sqrt{2}}\int_U MEC'_{A_1}(u)\mathrm{d}u,$$
$$En'_2=\frac{1}{\sqrt{2}}\int_U MEC'_{A_2}(u)\mathrm{d}u.$$

From a geometrical point of view, En_1 and En_2 are the areas covered by MEC_{A_1} and MEC_{A_2} reduced by a factor of $1/\sqrt{2\pi}$. En'_1 and En'_2 are the areas covered by MEC'_{A_1} and MEC'_{A_2} reduced by a factor of $1/\sqrt{2\pi}$. MEC'_{A_1} and MEC'_{A_2} are the unoverlapped parts of MEC_{A_1} and MEC_{A_2} which are created by cutting the curves at the intersection point of the curves. So we call En'_1 and En'_2 entropy cuts. The synthesized cloud is the entropy cut weighted average of the two clouds.

The 2-D synthesized clouds are constructed in a similar way. Suppose the two dimensions of a universe of discourse are independent, then we can calculate the digital characteristics in two dimensions separately. 2-D synthesized clouds also obey the properties described above from the geometrical point of view.

If we use the mechanism of synthesized cloud construction recursively from low concept levels to high concept levels, we can get concept hierarchies for linguistic variables. The higher the concept levels, the less the number of linguistic atoms. Till the top level, there is only one linguistic atom "any".

In brief, the linguistic cloud model and the mechanism of virtual cloud construction provide a new solution to data preprocessing for mining association rules. Combining the existing algorithms for mining association rules with the concept hierarchies, we can mine association rules at any concept level or at multiple concept levels.

2.3 Experiment and discussion

In order to verify the feasibility and effectiveness of the linguistic cloud model in mining association rules, we conducted experiment using a spatial database about the geospatial and economic information in China (see Table 1). What we are interested in is the relation between the geospatial information and economic situation. The task relevant data are extracted from the large database by spatial query. It has six attributes, x, y, elevation, road density, distance to the sea, and average income per year. The data size is about 600K bytes. The locations x and y are planimetric rectangular coordinates, which are transformed from geographical coordinates (longitudes and latitudes). The attributes are all numerical, among which road density is represented by the average total road length per square kilometer. The spatial database is managed by Geographic Information System, Arc/Info. The detail of spatial query of the spatial database is omitted here because of the space limitation. After query, the task relevant data are organized as a relational database or a relation table (see Table 1).

Table 1 The experimental database on Chinese geospatial and economic data for mining association rules

city name	x	y	elevation	road density (m/km^2)	distance to sea (km)	average income per year
Shanghai	1 091 703.29	2 420 920.22	27.82	950.06	19.33	16 000
Guangzhou	349 979.58	1 450 967.29	200.41	916.18	60.06	18 000
Wenzhou	1 038 667.52	1 992 814.97	51.97	848.49	24.51	17 000
Beijing	549 675.77	3 333 817.10	29.82	1 049.66	149.24	12 000
Changchun	1 224 839.09	3 828 302.91	187.72	899.34	458.45	8 000
Huhehot	102 736.67	3 447 929.21	2 300.30	485.09	581.90	7 500
Lanzhou	−562 917.30	2 877 368.66	3 503.85	400.90	1 439.41	5 600
Xining	−838 688.23	2 972 462.09	4 297.27	402.57	1 619.54	4 800
Urumqi	−1 818 150.51	3 904 377.65	3 603.35	349.92	2 590.17	4 900
Kuerle	−1 979 809.33	3 676 153.43	3 501.82	178.71	2 700.18	3 800
Zhengzhou	292 923.52	2 706 200.50	69.83	911.39	538.71	8 200
Laohekou	131 264.70	2 430 429.57	85.35	398.55	762.52	4 200
Lhasa	−1 865 697.22	2 363 864.17	6 202.52	280.81	2 200.32	4 800
Naqu	−1 780 113.14	2 496 994.96	5 897.91	168.97	2 098.59	3 700
Changsha	292 923.52	1 897 906.39	208.59	792.15	720.79	8 100
Xianning	435 563.67	2 183 186.66	175.14	581.12	698.01	5 500
…	…	…	…	…	…	…

Since the attributes are all numerical, there are lots of distinct values for each attribute. It is impossible to discover strong association rules at the primary level directly. Therefore, we define linguistic atoms for attribute generalization. We consider x and y attributes as one linguistic variable "location", and specify eight 2-D linguistic atoms for it, such as "southwest", "northeast", "north by east", "southeast", "northwest", "north", "south" and "central", most of them are rotated clouds. Other attributes are considered to be 1-D linguistic variables. Three linguistic atoms are defined for each attribute, such as "low", "middle" and "high" for "elevation", "road density", and "average income per year"; "close", "middle" and "far" for "distance to the sea". The linguistic atoms "low" and "close" are half-down clouds, "high" and "far" are half-up clouds as described in Refs. [24,25]. Neighbor linguistic atoms for each linguistic variable have some overlaps.

The attribute generalization process is straightforward and simple: the database is scanned only once, and the consuming time is linear with the size of the database. After attribute generalization, many different tuples at the primary level become identical at the higher concept level and are merged to one tuple, thus the database shrinks prominently. A new attribute "count" is added into the generalized database to record the numbers of primary tuples merged to it. This attribute is not treated as a "real" attribute when mining rules. It is just used for counting the supports of the itemsets. The generalized and reduced database makes the succeeding process easy and fast. Any existing algorithm for mining association rules, such as Apriori, can be applied to the generalized database.

We developed the Apriori algorithm on PC. The implementation was slightly different from the standard Apriori[5] in that it can deal with relational table directly so that relational table need not be transformed to binary transactions[9]. The generalized database is mined with the minimum support 6% and minimum confidence 75%. Eight large 4-itemsets are mined at the generalized database. Eight association rules are generated with the "average income per year" as the seccedent and the conjunction of other attributes as the antecedent. The rules can be described in the form of production rules, such as:

Rule 1. If location is "southeast", road density is "high", and distance to the sea is "close", then average income is"high".

Rule 2. If location is "north by east", road density is "high", and distance to the sea is "close", then average income is "high".

Rule 3. If location is "northeast", road density is "high", and distance to the sea is "middle", then average income is "middle".

Rule 4. If location is "north", road density is "middle", and distance to the sea is "middle", then average income is "middle".

Rule 5. If location is "northwest", road density is "low", and distance to the sea is "far", then average income is "low".

Rule 6. If location is "central", road density is "high", and distance to the sea is "middle", then average income is "middle".

Rule 7. If location is "southwest", road density is "low", distance to the sea is "far", then average income is low.

Rule 8. If location is "south", road density is "high", distance to the sea is "middle", then average income is "middle".

These rules are visualized as ellipses with different colors gradually changing to gray from the centers of the ellipses (see Fig. 10). The numbers marked on the ellipses are the rule numbers.

As intermediate results, large 2-itemsets and 3-itemsets are discovered during the mining process. The association rules generated from them are also interesting and useful. For example, the following rules demonstrate the relation between the road density and elevation and location. The rules are visualized in Fig. 11.

Rule 1. If elevation is "low", then road density is "high".

Rule 2. If elevation is "high", then road density is "low".

Rule 3. If elevation is "middle" and location is "northwest", then road density is "low".

Rule 4. If elevation is "middle" and location is "north", then road density is "middle".

In order to discover multi-level association rules, we use the virtual cloud construction method to generalize the attribute "location" further. Two virtual clouds are generalized by synthesizing "northwest" and "southwest" to "the west", "south" and "central" to "south and central China". As a result, the eight rules in Figure 10 are reduced to six rules, in which rules 1, 2, 3, 4 are the same, rules 5, 7 and rules 6, 8 are generalized to the following new rule 5 and rule 6 respectively. The six rules are visualized in Fig. 12.

Rule 5. If location is “the west”, road density is “low”, and distance to the sea is “far”, then average income is “low”.

Rule 6. If location is “south and central China”, road density is “high”, and distance to the sea is “middle”, then average income is “middle”.

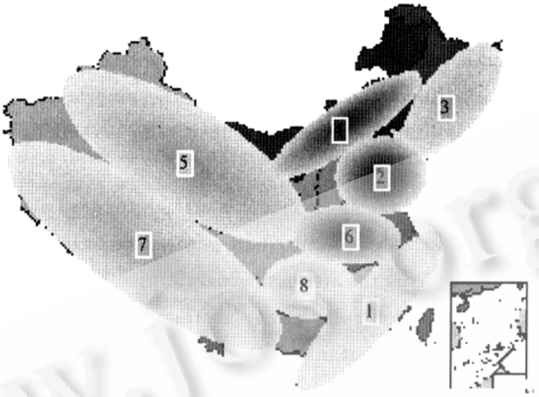


Fig. 10 Discovered association rules for “average income”

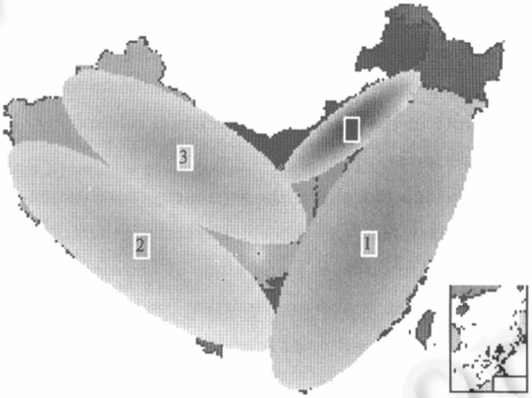


Fig. 11 Discovered association rules for “road density”

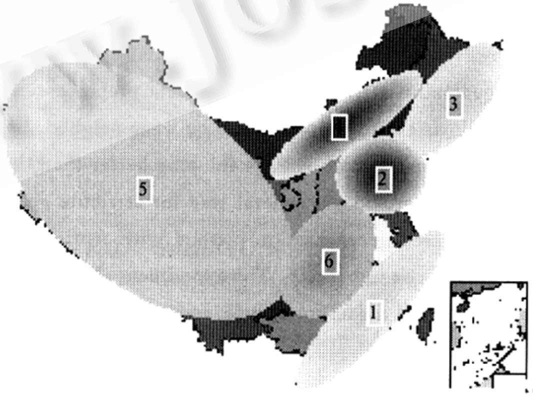


Fig. 12 Discovered association rules for “average income” at a higher concept level

It should be noted that these rules are uncertain in that the supports and confidences are slightly different at different occasions. That is to say, because of the properties of the linguistic clouds and the overlaps of neighbor clouds, the attribute generalization results in different generalized relation tables in which the counts of the tuples are slightly different at different occasions. If we use the conventional generalization method, we simply specify several thresholds for the attributes. The generalized relation table is identical at any occasion, and so are the rules. This crisp partition method ignores the uncertainty in the generalization process. Therefore, the cloud based generalization method is superior to the conventional method in holding uncertainties. The above results show the effectiveness of the linguistic cloud model for the preprocessing in mining association rules. That is, the cloud model based generalization can mimic human being's thinking, while making the discovered knowledge robust.

The attribute generalization based on the cloud model is straightforward and efficient, Apriori is efficient, and the combination of them is a two-step process: attribute generalization first and rule mining next. Therefore, the whole process is efficient. Of course, the attribute generalization method can be combined with any existing efficient algorithm for mining association rules at any abstract level or at multiple concept levels. Mining multiple level association rules can be carried out in top-down or bottom-up manner. The concept hierarchies are automatically generated based on the mechanisms of synthesized cloud construction, the user should specify enough linguistic atoms at a relatively low level.

The extensive study of the linguistic cloud model and the combination or integration with the other association rule mining algorithms, such as incremental mining algorithms, parallel mining algorithms, etc., are the future directions of our work. For example, we can specify the minimum support and minimum confidence as linguistic atoms to hold the uncertainty in the process of association rule mining, however we specified the two thresholds as fixed values in the experiment.

In the mean time, the linguistic cloud model and the attribute generalization method provide a general way for data preprocessing in KDD. For example, the concept hierarchy generation method can be combined with attribute oriented induction method[19,26] to discover many kinds of knowledge. The integration of the linguistic cloud model with the other KDD algorithms is also an important issue for further study.

3 Conclusion

Mining of association rules is an important task in KDD. The linguistic cloud model is introduced to enhance the existing algorithms. The two-dimensional and even multi-dimensional linguistic cloud models are presented as the extension of the previous 1-D model. The mathematical representation of 2-D linguistic atoms with six digital parameters for uncertainty is given to explore the essential uncertainty with both fuzziness and randomness. The model is used to generalize attribute values for data preprocessing of mining association rules. Cloud model based generalization allows the overlapping area between neighbor linguistic terms, and it is a soft partition of attribute spaces that can mimic human being's thinking better than the crisp partition method. The mechanism of virtual cloud construction provides a general way for attribute generalization at multiple concept levels. Combining the cloud model based generalization method with Apriori algorithm can mine association rules at any concept level or at multiple concept levels. The experiment shows the benefits of effectiveness, efficiency and flexibility of the linguistic cloud model for mining association rules.

References

1 Fayyad U M, Piatetsky-Shapiro G, Smyth P *et al.* Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996

2 Fayyad U M, Piatetsky-Shapiro G, Smyth P. Knowledge discovery and data mining: towards a unifying framework. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD 96). Portland, Oregon, 1996

3 Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases. In: Proceedings of the ACM SIGMOD Conference on Management of Data. Washington, DC, 1993

4 Houtsma M, Swami A. Set-Oriented Mining of Association Rules. Research Report RJ 9567, IBM Almaden Research Center, San Jose, California, 1993

5 Agrawal R, Srikant R. Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th International Conference on Very Large Data Bases. 1994. 478～499

6 Park J S, Chen M S, Yu P S. An effective hash-based algorithm for mining association rules. In: Proceedings of the 1995 ACM-SIGMOD International Conference on Management of Data. San Jose, CA, 1995. 175～186

7 Cheung D W, Han J, Ng V *et al.* Maintenance of discovered association rules in large databases: an incremental updating technique. In: Proceedings of the 1996 International Conference on Data Engineering. New Orleans, Louisiana, 1996

8 Cheung D W, Lee S D, Kao B. A general incremental technique for maintaining discovered association rules. In: Proceedings of the 5th International Conference on Database Systems for Advanced Applications. Melbourne, Australia, 1997

9 Srikant R, Agrawal R. Mining quantitative association rules in large relational tables. In: Proceedings of the ACM SIGMOD Conference on Management of Data. Montreal, Canada, 1996

10 Agrawal R, Ghosh S, Imielinski T *et al.* An interval classifier for database mining applications. In: Proceedings of the 18th International Conference on Very Large Data Bases. 1992. 560～573

11 Cheung D W, Ng V, Tam B W. Maintenance of discovered knowledge: a case in multi-level association rules. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96). Portland, Oregon, 1996

12 Han J, Fu Y. Discovery of multiple-level association rules from large databases. In: Proceedings of the 21st International Conference on Very Large Data Bases. 1995. 420～431

13 Srikant R, Agrawal R. Mining generalized association rules. In: Proceedings of the 21st VLDB Conference. Zürich, Switzerland, 1995

14 Park J S, Chen M S, Yu P S. Efficient parallel data mining for association rules. In: Proceedings of the 4th International Conference on Information and Knowledge Management. 1995. 31～36

15 Cheung D W, Han J, Ng V *et al.* A fast distributed algorithm for mining association rules. In: Proceedings of the 4th International Conference on Parallel and Distributed Information System (PDIS-96). Miami Beach, 1996

16 Cheung D W, Ng V, Fu A W *et al.* Efficient mining of association rules in distributed databases. IEEE Transactions on Knowledge and Data Engineering, 1996,8(6):911～922

17 Kamber M, Han J, Chiang J Y. Metarule-guided mining of multi-dimensional association rules using data cubes. In: Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining. Newport Beach, California, 1997

18 Koperski K, Han J. Discovery of spatial association rules in geographic information data-bases. In: Proceedings of the 4th International Symp. on Large Spatial Databases (SSD'95). Portland, Maine, 1995. 47～66

19 Han J. Data mining techniques. In: Proceedings of the ACM-SIGMOD'96 Conference on Tutorial. June 1996

20 Chen M S, Han J, Yu P S. Data mining: an overview from database perspective. IEEE Transactions on Knowledge and Data Engineering, 1997

21 Di Kai-chang, Li De-ren. KDD and its application and extension in GIS. In: Proceedings of the 2nd Annul Meeting of Chinese Association of GIS. Beijing, 1996

22 Di Kai-chang, Li De-yi, Li De-ren. Knowledge representation and discovery in spatial databases based on cloud theory. Proceedings of the International Archives of Photogrammetry and Remote Sensing, 1998,32(3/1):544～551

23 Li De-yi, Shi Xue-mei, Meng Hai-jun. Membership clouds and cloud generators. The Research and Development of Computers, 1995,42(8):32～41

24 Li De-yi, Han J, Chan E *et al.* Knowledge representation and discovery based on linguistic atoms. In: Proceedings of the 1st Pacific-Asia Conference on KDD & DM. Singapore, 1997

25 Li De-yi, Cheung D W, Shi Xue-mei *et al.* Uncertainty reasoning based on cloud models in controllers. Computers and Mathematics with Applications, Elsevier Science, 1998,35(3):99～123

26 Han J, Cai Y, Cercone N. Data-driven discovery of quantitative rules in relational databases. IEEE Transactions on Knowledge and Data Engineering, 1993,5:29～40

用语言云模型发掘关联规则

李德毅[1]　邸凯昌[2]　李德仁[2]　史雪梅[3]

[1](中国电子系统工程研究所　北京　100036)
[2](武汉测绘科技大学信息工程学院　武汉　430070)
[3](香港理工大学计算机系　香港)

摘要　该文提出用语言云模型用于KDD中知识表达和不确定性处理,引入了多维云模型作为一维模型的扩展.语言云的数字特征量将语言值的模糊性和随机性用统一的方式巧妙地综合到一起,基于云模型的概念层次结构可以跨越定量和定性知识之间的鸿沟.为了发现强关联规则,属性值要在较高的概念层上泛化,同时允许相邻属性值或语言项间有重叠.这种软划分可以模仿人类的思想,使发现的知识具有稳健性.将基于云模型的泛化方法与Apriori算法结合起来,从空间数据库中发掘关联规则.试验显示了其有效性、高效性和灵活性.

关键词　语言云模型,关联规则,Apriori算法,虚拟云,空间数据发掘.

中图法分类号　TP311