Mining Combined Detection of Tumor Marker Based on Cloud Model

Feng Guo Shaozi Li

Department of Cognitive Science, Xiamen University, Fujian 361005, P. R. China

Abstract

The cloud model is an effective tool in transforming between qualitative concepts and their quantitative expressions. In the field of tumor marker, detects combined markers can improve the performance of cancer detection. But the discovery of combined tumor markers depends on doctor's experience, and the markers are very difficult to find. This paper presents a multi-markers discovery method to mine combined detection tumor markers in detection histories. We mined top 10 combinations by our algorithm, and 8 of the results are fit for iatrical knowledge. Further this paper compares the results between cloud model and fuzzy set, and sums up their difference.

Keywords: Data mining, Association rule, Cloud model, Tumor marker, Combined detection.

1. Introduction

Data mining is the discovery of previously unknown potentially useful and hidden knowledge in databases. With the rapid increment of information, mining useful information from database has been a hotspot in many areas, such as artificial intelligence, database, and so on.

Mining association rules has been a hotspot study, since its algorithm was presented by R.Agrawa[1]. An example in supermarket about association rules is that "customers buy break → buy milk". If it satisfies its minimum support and minimum confidence, the rule is true. This example is about Boolean attributes, while in the real application many attributes are quantitative, such as age, income, etc. So the study of mining association rules must be extended to quantitative attributes. In traditional study, many researchers deal with quantitative attributes by fine-partitioning the values of the attribute and then combining adjacent partitions as necessary. In this way they can map the Quantitative Association Rules Problem into the Boolean Association Rules [2][3]. But this kind of methods is not good enough. Their boundaries are too sharp. First, values near the boundary will be excluded from partitions. Second, effects of the values in the cross partitions are over emphasized when using overlapping partitions.

To solve these problems, the fuzzy set is introduced in [4]. Fuzzy set can soften the sharp boundaries by using the membership functions. But this method has disadvantages，the completely certain membership function that has no any fuzziness at all has been the bottleneck of the applications of this theory.

In order to solve the sharp interval problem verily, in [5], cloud transform, which uses many concepts represented by cloud model to fit the real distribution of data, is introduced. In reference [6], cloud model is used to calculate support, confidence and relationship in the field of association rules mining. The method of mining the normal cloud association rules is provided and the mined rules are used for prediction in [7]. From these papers we can see that the cloud model not only broadens the form conditions of the normal distribution but also makes the normal membership function be the expectation of the random membership degree. It would be more applicable and universal in the representation of uncertain notions.

In the study of tumor marker, we find that the combined markers detection can increase the tumor's positively detectable rate by using the complementation of different markers [8]. In reference [8], three items, which have the most positive rate, are chose as the combined determination's markers. According to experience, alpha-fetoprotein (AFP), α-L-fucosidase (AFU), γ-glutamyl transpeptidase (γ-GT), tumor necrosis factor(TNF-α) and DR-70TM5 are used for the combined determination. The positive rate is above 98%. If the detection just uses AFP, the positive rate is only 64.0% [9]. So how to find the new combination is an important work. But nowadays, the finding of test-combination mostly depends on manual summarize but not computer aid.

This paper presents a new method based on cloud model to find the marker-combination for tumor determination. Cloud model can deals with fuzziness

and randomness well, so we use it to map the detection indexes into different partitions, and then we mine the marker-combination with the highest relativity. Our experiment data comes from the 2037 health examinations about tumor markers from 2002 to 2006. We use association rules based on cloud model and mine the highest relativity combinations. Most of the combinations are useful and others are waiting for the verification of future medical study. In experiment, we use rules-mining algorithm based on fuzzy set to compare with the algorithm based on cloud model. Results show that the latter can reflect human being's uncertainty of thinking.

2. Cloud model

2.1. Definition of cloud model

Definition1: Assuming X is a common set $X = \{x\}$, it is called discussion area. A fuzzy set $\overline{A}$ in X is defined as a membership function, $\mu_{\overline{A}}(x): X \to [0,1], x \in X$, which maps the factors in X to real number between 0 and 1, $\mu_{\overline{A}}(x)$ is called degree of x under fuzzy set $\overline{A}$, the membership degree of x to $\overline{A}$ for short. Membership function $\mu_{\overline{A}}(x)$ is the membership degree distribution of all factors in X belonged to fuzzy set $\overline{A}$. The fuzzy set is a factor set with different membership degree.

2.2. The numerical characteristics of cloud

In real world, many fuzzy concepts' expectation curve is normal distribution or semi normal distribution.

And the normal cloud can be a most important, helpful tool to represent linguistic atom. A normal cloud is described by three numerical characteristics, which are Ex, En and He. Expectation Ex is the point of center that could represent this qualitative concept properly in the discussing area. Entropy En represents a measured granularity of a certain quality concept. Also, En represents fuzziness of a qualitative concept representing range of values that could be accepted in the discussing area. Hyper entropy He is the uncertain degree of entropy. He reflects the randomness of samples of a qualitative concept and reveals the relationship between fuzziness and randomness.

Figure 1 is a 1-dimension normal cloud generated by forward cloud generators.

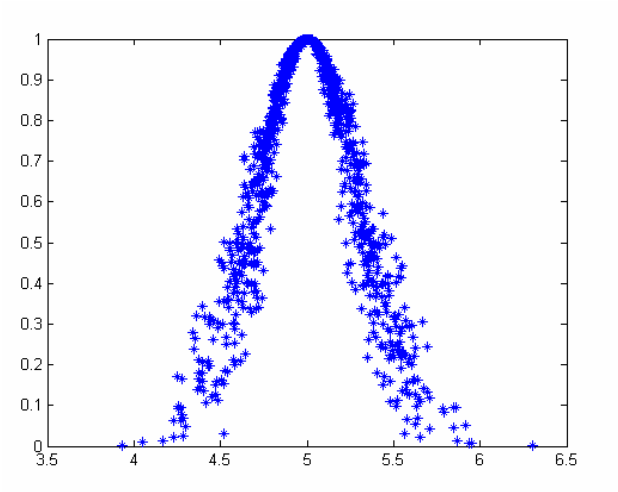

Fig. 1: a 1-dimension normal cloud model when Ex=5, En=0.3 and He=0.05

From figure 1 we can see the main characteristics of cloud listed as below:

1. The numerical values of same concept that we have described are agglomerate. That is, the nearer to the expectation of the cloud, the denser the points are, the farer to the expectation, the sparser the points are.
2. The cloud's expectation curve is a normal distribution, which is fit to express the daily fuzzy concepts.
3. To the same x in the discussion area, its membership degree will fluctuate randomly in certain scope, which reflects the uncertainness and randomness of concepts.

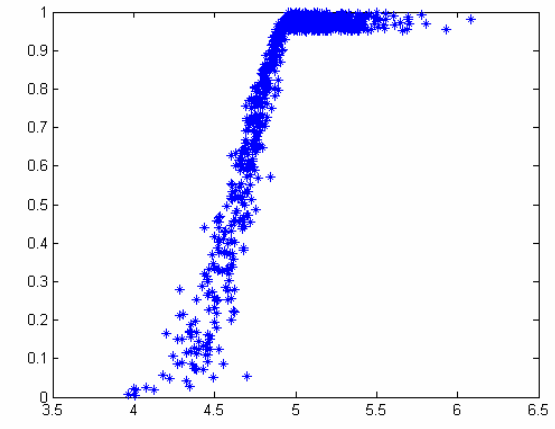

Fig. 2: the 1-dimensional semi cloud model when Ex=5, En=0.3, He=0.05.

In practical applications, there are many concepts that can not be described by normal cloud, for example, the superscalar in tumor markers. When an index of a

maker exceeds standard, the membership degree will keep high and won't decrease with the increment of index. So the semi cloud is needed. Figure 2 shows the 1-dimensional semi cloud.

2.3. Association rules mining based on cloud model.

Assume $T=\{t_1,t_2,\cdots t_n\}$ is a table in database, where t_j is the No.j record. $I=\{i_1,i_2,\cdots i_n\}$ is the attributes set of T, which is also called fields set. $t_j[i_k]$ is the No.j record's value in attribute i_k.

Fig 2: the 1-dimensional semi cloud model when Ex=5, En=0.3, He=0.05.

Assume the discussion area of i_k can be partitioned to different concepts based on cloud models, we defined the concepts set as $F_{ik}=\{f_{ik}^1, f_{ik}^2,\cdots, f_{ik}^l\}$. f_{ik}^j is the No.j concept of attribute I_k. And each concept can be expressed by numerical characteristics, Ex_{ik}^j, En_{ik}^j and He_{ik}^j

Suppose $X=\{x_1,x_2,\cdots x_n\}$ and $Y=\{y_1,y_2,\cdots y_n\}$ are subsets of I respectively. In this paper we define $A=\{f_{x1}, f_{x2},\cdots f_{xn}\}$ and $B=\{f_{y1}, f_{y2},\cdots f_{yn}\}$, where f_{xi} and f_{yj} are concepts of the values in the attribute, x_i and y_j respectively. So rules can be defined as "if X is A then Y is B". When mining association rules, we need to calculate three measurements, including support, confidence and interest. These measurements are used to evaluate mined rules and discard bad rules. In this paper, the formula of association rule's support is:

$$S_{X\,is\,A\Rightarrow Y\,is\,B}=\frac{\sum_{i=1}^{n}[\prod_{j=1}^{p}\mu f_{xj}(t_i[x_j])*\prod_{k=1}^{q}\mu f_{yk}(t_i[y_k])]}{n} \quad (1)$$

Formula of association rule's confidence is shown as below:

$$C_{X\,is\,A\Rightarrow Y\,is\,B}=\frac{\sum_{i=1}^{n}[\prod_{j=1}^{p}\mu f_{xj}(t_i[x_j])*\prod_{k=1}^{q}\mu f_{yk}(t_i[y_k])]}{\sum_{i=1}^{n}\prod_{j=1}^{p}\mu f_{xj}(t_i[x_j])} \quad (2)$$

But confidence can't really reflect rule's validity. In some case, a rule's confidence exceeds the threshold, it will be consider as a good rule probably. But sometimes the rule's conclusion "Y is B" has a high probability and this probability even exceeds the rule's confidence. Then the rule's prerequisite and conclusion are negative correlation. So, we use interest instead of confidence to measure rules and the formula is:

$$I_{X\,is\,A\Rightarrow Y\,is\,B}=\frac{C_{X\,is\,A\Rightarrow Y\,is\,B}-S_{Y\,is\,B}}{Max(C_{X\,is\,A\Rightarrow Y\,is\,B}, S_{Y\,is\,B})} \quad (3)$$

Interest's interval is [-1, 1]. The 0 denotes prerequisite has no relationship with conclusion, 1 denotes positive correlation and -1 denotes negative correlation. According to the theory of cloud model, whenever calculating any x's membership degree in the discussion area, its corresponding X–condition cloud generator [10] will be used. That is to say, the mapping between x and its membership degree is not unique, which can make the calculating result more random. But there are still problems. Because membership degree calculating will generate different values, which will not only make the interest's absolute value in the formula (3) larger than 1 but also make CloudAprior algorithm (we will use it later) unavailable. So we constrain that once the formula has been calculated, the following value is the same as the calculated result.

3. Mining association rules of tumor marker based on cloud model

3.1. Concepts of tumor marker and corresponding numerical characteristics

Table 1 is the detection data of tumor markers and each record is a cancer examination result. All of the items of the examination are common tumor markers. Since people doesn't need all items, the item not detected is marked with NULL.

Name	Specimen date	Age	Sex	AFP	CEA	CA 199	…
Zhang san	2006.9.3	43	M	3.3	0.86	8.0	
Li si	2006.9.4	42	M	neg	neg	N	
Chen xiu	2004.3.21	70	F	186.50	4.95	108.30	
……							

Table1: Parts of detection data

In the above table, the "neg" means negative and "N" means NULL.

Each item has a standard. Take AFP for example,

its standard is between 0.0 and 20.0. For each item, we define four concepts: negative, normal, high, and very high. Four concepts' numerical characteristics of AFP are shown in table 2.

concepts	Ex	En	He
negative	NULL	NULL	NULL
normal	10	6	1
high	25	1.3	0.1
very high	100	23	3

Table2: AFP's four concepts and their numerical Characteristics

Where "normal" and "high" are described by normal cloud model and "very high" is described by semi-cloud model. "Negative" means can not find tumor marker in detection, so the marker's value is zero or smaller than zero. It's hard for us to use cloud model to map such values. So we simply map a marker to concept "negative" when its value is smaller than 0 and the membership degree fluctuates randomly around 1.

In our experiments, we need to find the combined tumor markers that relate strongly to positive tumor markers. So, each conclusion of association rules that we mine has one tumor marker, and each of them has one concept. For example, a conclusion is "Y is B", where Y={AFP}，B={High}, or Y={AFP}，B={Very high}. Now our purpose is finding prerequisites, such as "X is A", and merging it with the conclusion above into rules. The rules' support and interest measurements should larger than minimum thresholds. The X consists of tumor markers, for example, X={CEA,PSA}, the B is corresponding concepts, such as B={high, very high}. That means "CEA is high and PSA is very high". We constrain all tumor markers in X and Y is unique.

3.2. CloudApriori: an algorithm to extract Rules

Our CloudApriori is a rules-extracting algorithm based on the classic Aprior. In order to express the CloudApriori clearly, we'd like to introduce some involved concepts first.

K itemset: when a rule's prerequisite contains k attributes and corresponding concepts, and the rule's conclusion contains one of tumor markers and the corresponding concept is high or very high, the rule is called K itemset.

Large K itemset: the k itemset whose support are larger than minsup. The minsup is a threshold which will be described later.

Candidate K itemset: the k itemset whose support is larger than the smallest threshold minsup probably.

R_k : set of all large k itemsets

C_k : set of all candidate k itemsets

In order to make our CloudApriori available, we must ensure the condition of Apriori algorithm. That is to say, the following theorem must be true.

Theorem 1: Any large itemset' non-empty subset must be large itemset.

Provement:

Assume R is a large itemset. According to the definition of large itemset, the following formula is true

$$S(R) > \min\sup \tag{4}$$

Suppose S is a non-empty subset of R, when S is extended to R, only prerequisite joints new conditions and conclusion does not change any more, so the

$\prod_{k=1}^{q} \mu f_{yk}(t_i[y_k])$ of R and S in formula (1) are the same and the value is between 0 and 1.

During the extension from S to R, the p in formula (1) is increasing and the number of attributes is increasing too. Because the membership degrees of different attributes multiply each other, and the membership degrees are between 0 and 1, R's $\prod_{j=1}^{p} \mu f_{xj}(t_i[x_j])$ is definite smaller than S's. That's the reason why R's support is smaller than S's.

At last the character of cloud model is taken into consideration. Since we need to use the random function when calculating the membership degree, so the degrees of the same value in a same attribute at different time are different, which will make R's support larger than S's and make the Theorem unavailable. In 2.3, we constrain that once the formula $\mu f_{xj}(t_i[x_j])$ has been calculated, the following value is the same as the calculated result.

According to the provement, we get the following result

$$S(S) > S(R) > \min\sup \tag{5}$$

So theorem 1 is true. That is to say, we can use Apriori's method to make joint operation between large itemsets.

Algorithm 1: CloudApriori

Input: cancer examination data table T, minimum support minsup, and minimum interest minint.

Output: association rule set R which satisfies minsup and minint.

Initialization:

For each tumor marker, generates a conclusion list according to the two concepts "high" and "very high". The conclusion is a "Y is B" described in 3.1. Y has only one attribute and B has only one corresponding concept.

Form a prerequisite set with each combination of T's attribute and concept, which are X and A described in 3.1. X has only one attribute and A has only one concept. The prerequisite set and conclusion list above form a rule set, which satisfies $X \cap Y = \Phi$. Then we calculate each rule's support, and discard those rules whose support doesn't satisfy the minsup. At last, put the satisfied rules into R_1.

Then we generate candidate itemsets and large itmesets in the way of recursion.

Call RecurApriori(T,minsup, R_1);

Procedure RecurApriori(T,minsup, R_k)

1) Empty large itemset R_{k+1}
2) To each large k itemset in R_k, make joint operation, then generate candidate itemsets C_k
3) Calculate support of all candidate k itemsets in C_k, and discard those k itemsets whose support are smaller than minimum support, then form the large k+1 itemsets. If the large k+1 itemset doesn't belong to R_{k+1}, add it to R_{k+1}.
4) If R_{k+1} is not empty, then call RecurApriori(T,minsup,minconf, R_{k+1}) recursively.

End

Sort the rules in $U_k R_k$ descending on interest, and put those rules that are larger than minint into R.

In the implementation of this algorithm, the bottle-neck is slow database access speed. So we use the following dynamic programming when calculating support and confidence in formula (1) and (2). By dynamic programming, all k itemsets can be jointed to k+1 itemsets that scaning database just once. When reading a new record, calculating all k itemsets' $\prod_{j=1}^{p} \mu f_{xj}(t_i[x_j])$ in formula (1) and (2), then adding it into $\sum_{i=1}^{n}[\prod_{j=1}^{p} \mu f_{xj}(t_i[x_j])]$. In this way, after scanning database one time, all k itemsets' support and confidence can be calculated by division operation.

If the largest number of attributes in prerequisite is K, the database is needed to be scanned for K times. After support and confidence is calculated, the calculation of interest needn't access database any more, so the calculating speed is very fast.

4. Experiments and results

4.1. Dataset of experiments

There are 11 tumor markers in our experiments. And there are about 2037 cancer examination records. The sparsity degree is 73.8%. We set K, the maximum number of attributes in prerequisite of rules, as 5. Rules mined by CloudApriori are sorted descending on interest and the top 10 rules are obtained.

4.2. Mining results of association rules under different settings.

When minsup is set as 0.0001, minint is set as 0.8, the 10 highest interest rules are shown in table 3.

Rules	Support	Interest
AFP is high and CA199 is very → CEA is very high	0.00109	0.99896
NSE is high ,CYFRA is very high, and SCC is very high → PSA is very high	0.00026	0.99883
NSE is high, CYFRA is very high and SCC is high → CA724 is very high	0.00109	0.99879
NSE is high and SCC is very high → PSA is very high	0.00102	0.99771
NSE is high, CYFRA is very high and SCC is very high → FPSA is very high	0.00106	0.99413
NSE is high → PSA is very high	0.00105	0.99329
CA153 is high, NSE is high and CYFRA is very high → CA724 is very high	0.00104	0.99198
CA153 is high and CYFRA is very high → CA724 is very high	0.00112	0.98994
CA153 is high, NSE is normal and CYFRA is negative → CA724 is very high	0.00012	0.98909
CA724 is high NSE is normal and CYFRA is very high → SCC is very high	0.00100	0.98811

Table 3: Top 10 rules in interest score when minsup is 0.0001, minint is 0.8

When minsup is set as 0.001, minint is set as 0.8, the 10 highest interest rules are shown in table 4.

Rules	Support	Interest
CA724 is high → SCC is very high	0.01392	0.98344
CA724 is high and NSE is normal → SCC is very high	0.01351	0.96113
CEA is very high → AFP is very high	0.08283	0.95987
CEA is high → AFP is very high	0.02407	0.93165
CEA is very high → AFP is high	0.09451	0.93003
CA125 is normal → CYFRA is very high	0.09287	0.91098
CA153 is normal → PSA is very high	0.05959	0.90057
CEA is high and CA125 is normal → AFP is high	0.09106	0.88135
CA125 is normal and CA153 is normal → CEA is very high	0.05906	0.88049
CA125 is normal and CA153 is normal → CA724 is very high	0.09046	0.87772

Table 4:Top 10 rules in interest score when minsup is 0.0001, minint is 0.8

Comparing table 3 with table 4, we can find that when the support is increased the top 10 rules in table 3 don't exist in table 4. Moreover, there are much more prerequisite attributes in table 3, most rules have three, while in table 4, each rule has no more than 2 attributes.

The reason is that in formula (1), when the number of attributes in x is increased, the value of $\prod_{j=1}^{p} \mu f_{xj}(t_i[x_j])$ is extremely turning small. So the number of prerequisite attributes in table 4 is small than table 3.

Meanwhile high support dues to the high frequency of the marker appears in examination records such as "CA125 is normal", "CA153 is normal". In the cancer examination of tumor markers, each marker has indicative function to different organic symptom. That is to say, if some index is high, the other indexes aren't necessary high too. Generally, for diagnosis doctors may ask their patients to do more marker detection, such as CA125、CA153 and so on. Mostly the CA125 and CA153 are normal, so the frequencies of "CA125 is normal" and "CA153 is normal" are high too. That's the reason why their supports are high, and they appear in the mined rules some times. The "NSE is normal" is the same instance. After comparing results under different minimum supports, we find that setting minsup as 0.0001 is preferable. Eight rules in table 3 are fit for iatrical knowledge. For example, in digestive neoplasm detection, AFP、CEA and CA199 are frequently used, while in lung cancer detection, CA724、NSE、SCC and CYFRA are usually used. So the mined combinations are significant to the tumor markers combined detection.

4.3. Compared with mining method based on fuzzy set

In the following, we'll mine combined markers by mining method based on Fuzzy set, which will be compared with the method based on cloud model. Formulas of calculating support and interest are same as cloud model. The difference between the two methods is that membership degree is calculated certainly in Fuzzy set, but it's calculated with random normal distribution function in cloud model.

In this paper, minsup is set as 0.0001 and minint is set as 0.8 when mining combined markers based on Fuzzy set. Comparing these two methods, based on cloud model and based on Fuzzy set, we find that their 10 highest interest rules are almost the same and only two of them are different. For example rule "NSE is high, CYFRA is very high → PSA is very high" only exists in the former. Reason is that a rule's certain membership function of Fuzzy set make the rule's support far lower than minsup, which makes the rule be discarded. In the way of cloud model, randomness can increase a rule's support and makes the rule satisfy the minsup, and then add it to large itemsets R_k .

The obvious difference of these two methods happens when interest is set as 0.8. Since the membership degree function of Fuzzy set is certain, every time of running the results are same. But in the way of cloud model, the randomness of the membership degree function helps those rules that cann't be accepted before exist in the large itemsets R_k . Table 5 lists some of these rules.

Rules	Support	Interest
CYFRA is high → PSA is very high	0.00964	0.80411
CA153 is high and CA724 is very high → CEA is high	0.00931	0.80411
CA724 is very high and NSE is high → FPSA is high	0.00696	0.87314

Table5: Rules only given in the way of cloud model.

5. Conclusions

This paper takes the advantage that cloud model can deal with fuzzy boundary in transferring between quantitative values and qualitative concepts. We import semi-cloud model and also present the formula

of calculating support and interest according to association rules of tumor markers. We also present the CloudApriori algorithm which finds the optimal combined tumor markers in the way of mining association rules. This paper compares our algorithm with the method based on Fuzzy set and obtains the following results. That is, because of randomness, the method based on cloud model will get more new rules which are near to the boundary. Combined markers given by our experiments are mostly fit for iatrical knowledge while some unknown combinations are waiting for verification of future medical study. In application, since the quality concept is defined manual, we'll study auto-definition instead in the future.

Acknowledgement

This project is supported by the Science & Technology Project of Fujian Province (Project Number: 2006H0037)

References

[1] R. Agrawal, Imielinske, T. Swami, A. Mining association rules between sets of items in large databases. *Proceedings of the ACM SIGMOD International Conference on the Management of Data. Washington, DC: ACM Press*, pp.207-216, 1993.

[2] R. Srikant, R. Agrawal, Mining quantitative association rules in large relational tables. P*roceedings of the 1996 ACM SIGMOD International Conference on Management of Data. Montreal, Canada*, ACM Press, pp.1-12, 1996.

[3] Z.H .Zhang, Y.C. Lu, B. Zhang, An algorithm for mining quantitative association rules. *Journal of Software*, 11:801-805, 1998.

[4] C. Kuok, A. Fu, M. Wong, Mining fuzzy association rules in databases. *SIGMOD Record*, 27-32, 1998.

[5] D.Y. Li, C.Y. Liu, L.Y. Liu. Study on the Universality of the Normal Cloud Model. *China Engineering Science*, 3:18-24, 2004.

[6] Y Du, Z.L. Song, D.Y. Li, Mining Association Rules Based on Cloud Model. *Journal of PLA University of Science and Technology*, 3:29-34, 2003.

[7] J.J. Lu, Z.P. Qian, Z.L. Song, Application Normal Cloud Association Rules on Predition. *Journal of Computer Research & Development*, 37:1317-1320, 2000.

[8] Y.Y. Li, J. Fan, T.G. Wu, X.L. Gan. The Significance of Combined Tumor Marker Detection in Clinic Diagnose. *Chinese Journal of Clinical Oncology and Rehabilitation,* 7:83-84, 2000.

[9] J.X. Ma, Q.L. Gong, M.Q. Lin, Y. Gong. *Combined five tumor markers in detecting primary hepatic carcinoma. Chinese Journal of Surgery* ,38:14-16, 2000.

[10] D.Y. Li, H.J. Meng, X.M. Shi, Cloud model and cloud model generator. *Computer Research and Development*, 32:15-20, 1995.