Chin. Ann. of Math.
22B: 4(2001),475-484.

THE 'HYBRID' TECHNIQUE FOR RISK ANALYSIS OF SOME DISEASES****

SHANG Hanji* LU Yuchu** XU Xuemei*** CHEN Qian***

Abstract

Based on the data obtained from a survey recently made in Shanghai, this paper presents the hybrid technique for risk analysis and evaluation of some diseases.

After determination of main risk factors of these diseases by analysis of variance, the authors introduce a new concept 'Illness Fuzzy Set' and use fuzzy comprehensive evaluation to evaluate the risk of suffering from a disease for residents. Optimal technique is used to determine the weights w_i in fuzzy comprehensive evaluation, and a new method 'Improved Information Distribution' is also introduced for the treatment of small sample problem.

It is shown that the results obtained by using the hybrid technique are better than by using single fuzzy technique or single statistical method.

Keywords Illness fuzzy set, Statistics, Fuzzy comprehensive evaluation, Information distribution, Optimization
2000 MR Subject Classification 03E72, 60P25
Chinese Library Classification O159 **Document Code** A
Article ID 0252-9599(2001)04-0475-10

§1. Introduction

Since the theory of fuzzy set was introduced by Lotfi A. Zadeh in 1965, fuzzy technique has developed rapidly and has been successfully applied into many fields.

In recent years, a new and promising way of using the fuzzy technique is combined with other deterministic and statistical methods. The so-called 'hybrid' technique has been applied to insurance business and achieved positive results (see [1, 2], etc.).

Based on the data obtained from a survey recently made in a community in Shanghai, this paper studies how some risk factors, such as age, family history and Body Mass Index (BMI), influence the prevalence rate of some diseases related to better living conditions. For the sake of brevity, we only put emphasis on the analysis of hypertension. The analysis methods for other diseases are quite similar.

Manuscript received March 10, 2000. Revised November 22, 2000.
*Department of Mathematics, Fudan University, Shanghai 200433, China.
E-mail: afac@fudan.edu.cn
**Department of Mathematics, Shanghai University, Shanghai 200436, China.
***Institute of Mathematics, Fudan University, Shanghai 200433, China.
****Project supported by the National Natural Science Foundation of China (No. 19831020).

§2. Determination of Main Risk Factors

There are many risk factors initiating hypertension included in the above-mentioned survey. To get the chief ones among these factors, we make analysis of variance on the influence of some risk factors, such as age, family history, BMI, smoking history and drinking history, using the survey data.

The steps of analysis are as follows:

(A) Divide every factor into levels. Age is divided with a level per ten years. The smoking history is divided into three levels, which are free from smoking, having smoked less than 54,750 cigarettes till the investigation and above 54,750 cigarettes. Family history is divided into three levels, which are 0 for having no family history, 1 for one of the parents or siblings suffering from hypertension, 2 for at least two of parents or siblings suffering from hypertension. We divide the BMI (from 16 to 36) into six levels, and the drinking history, into three levels, which are free from drinking, the total amount of consumed alcohol is less than 130 kg till the investigation and above 130 kg.

(B) Divide all the 18,749 adults under investigation into 5 groups according to their residential districts. In each group, we calculate the prevalence rate of hypertension for each level of certain risk factor. Then we make single-factor repeated-trials analysis of variance on every factor's influence on the prevalence rate of hypertension, that is, for every factor mentioned above, five experiments are made according to the five districts, which are restricted by the different levels divided before.

(C) Based on the observed values, we make analysis of variance using software S-PLUS. The result shows that age, family history and BMI make great impact on the prevalence rate of hypertension, the corresponding Pr(F) are 3.774758e-015, 1.110223e-016 and 1.890044e-012. In contrast, the influence of drinking history and smoking history is not so remarkable, for the corresponding Pr(F) are 0.04654151 and 0.6373128. So it is clear that age, family history and BMI may be regarded as three main risk factors influencing the prevalence rate of hypertension, while smoking history and drinking history may not.

Since the relationship between age and BMI might affect the result of risk evaluation, the whole group is divided into three subgroups: youth-group (age $15 \sim 34$), middle-age-group (age $35 \sim 54$) and old-people-group (older than 55). Thus we regard age and BMI as independent in every subgroup. The discussion below is aimed at each subgroup respectively.

§3. The Concept of Illness Fuzzy Set

To describe the situation of suffering from hypertension for a group, we first introduce an 'illness fuzzy set' as follows:

Definition 3.1. *Illness fuzzy set* $V' = (v_0', v_1', v_2')$ *is a kind of measure on the prevalence situation or risk for a certain disease of a certain group of residents. It is a fuzzy set, regarding evaluation set* $V = (v_0, v_1, v_2)$ *as its universal field, which fits* $\sum_{i=0}^{2} v_i' = 1$, *where* v_0, v_1, v_2 *stand for free from illness, mild illness and serious illness respectively, and* v_0', v_1', v_2' *stand for their membership grade accordingly.* V' *fully explains the hypertension risk of the certain group of residents.*

The method of calculating v_i' will be given later.

§4. Fuzzy Comprehensive Evaluation

Now let us consider the following problem: if we have known the levels of the three main risk factors for a certain resident or a certain group of residents, how can we evaluate its risk of suffering from hypertension (including mild and serious)?

To answer this question, we use the method of fuzzy comprehensive evaluation as follows.

4.1 Illness Fuzzy Set with Fixed Level of Certain Risk Factor

First consider a group of residents formed by fixing a certain level of a certain risk factor, for example, all old males with BMI ranging from 21 to 23, or all young females with no family history of hypertension, etc.

We hope to determine the illness fuzzy sets for all such groups (for every level of every appointed risk factor) and use these fuzzy sets to evaluate the risk of any person or any group.

The survey data are used to determine each membership grade of such group's illness fuzzy set. Under the law of great numbers, the frequency may approach the probability. So if the total number of residents of this group is large enough, the statistical frequency in every circumstance during the survey approximately replaces the membership grade directly. But, if the total number of residents or the number of residents suffered from hypertension is relatively small, then this method is unfavorable, for it can not reflect the true situation. We shall turn to information distribution method in this situation (see [3]).

4.2 Multi-Factor Fuzzy Comprehensive Evaluation

In medical science and medical insurance, it is often needed to evaluate the risk of hypertension of a group with fixed levels for all risk factors. To do this, we establish multi-factor fuzzy comprehensive evaluation. The method goes as follows.

First, fix the level of every risk factor and calculate the respective illness fuzzy set: $R_1 = (R_{10}, R_{11}, R_{12})$ for age, $R_2 = (R_{20}, R_{21}, R_{22})$ for BMI, $R_3 = (R_{30}, R_{31}, R_{32})$ for family history, with the method mentioned in 4.1, and give a note as below:

$$R = \begin{pmatrix} R_1 \\ R_2 \\ R_3 \end{pmatrix}.$$

Then, perform comprehensive evaluation.

Note the weights that each risk factor initiates the prevalence rate as w_1, w_2, and w_3. The selection of the weights will be discussed later.

Thus the result of comprehensive evaluation goes as follows:

$$A = (w_1, w_2, w_3) \circ R = W \circ R = (A_0, A_1, A_2). \tag{4.1}$$

Here the operation '$\circ$' is viewed as a generalized matrix multiplication, namely

$$A = W \circ R = \Big(\sum_{i=1}^{3} w_i \bullet R_{i0}, \sum_{i=1}^{3} w_i \bullet R_{i1}, \sum_{i=1}^{3} w_i \bullet R_{i2} \Big), \tag{4.2}$$

where $\bullet$ stands for ordinary multiplication, $\sum$, generalized addition, which means

$$x \oplus y = \min(1, x + y).$$

Fuzzy set A gained through this way is called the risk evaluation fuzzy set or risk vector of that group.

4.3 The Optimal Approximate Solution for the Inverse Problem of Fuzzy Comprehensive Evaluation

There are two ways of determining the weights w_i. One comes from experience or the opinions of several medical experts, which is quite simple but relatively arbitrary. Another way is to solve the inverse-problem of comprehensive evaluation. The latter method has been introduced in [4], where the fuzzy operator '$\circ$' is chosen as $M(\wedge, \vee)$.

We try to improve the solution method for the inverse problem by using fuzzy operator $M(\bullet, \oplus)$ and the optimization technique as follows.

From each of the three subgroups mentioned in §2, we select two representative groups. For the six groups, the method of calculating statistical frequency is used to get the risk vector

$$B_I = (B_{I1}, B_{I2}, B_{I3}) \ (1 \le I \le 6).$$

In practice, we assume the age and BMI contribute at least 10% to the hypertension risk, the family history is at least 40%. Therefore the W_I should satisfy

$$w_{I1} \ge 10\%, \quad w_{I2} \ge 10\%, \quad w_{I3} \ge 40\%, \quad \sum_{i=1}^{3} w_{Ii} = 1.$$

At last, six $W_I = (w_{I1}, w_{I2}, w_{I3})$ are obtained by approximately solving the equation (4.3) under this restriction.

$$W_I \circ R_I = B_I. \tag{4.3}$$

Before solving these equations, let us review the concept of approaching degree of fuzzy sets.

Definition 4.1. *Assume* $F(X) = \{$*all fuzzy sets defined on universal field* $X\}$*, if the mapping* $\sigma : F(X) \times F(X) \to [0, 1]$ *satisfies*:

(1) $\sigma(\underset{\sim}{A}, \underset{\sim}{A}) = 1$;

(2) $\sigma(\underset{\sim}{A}, \underset{\sim}{B}) = \sigma(\underset{\sim}{B}, \underset{\sim}{A}) \ge 0$;

(3) $\forall \underset{\sim}{A}, \underset{\sim}{B}, \underset{\sim}{C} \in F(X)$, $x \in X$, *if*

$$\sum \mid u_{\underset{\sim}{A}}(x) - u_{\underset{\sim}{C}}(x) \mid \ge \sum \mid u_{\underset{\sim}{B}}(x) - u_{\underset{\sim}{C}}(x) \mid,$$

then $\sigma(\underset{\sim}{A}, \underset{\sim}{C}) \le \sigma(\underset{\sim}{B}, \underset{\sim}{C})$*, where* $u(x)$ *stands for the membership function.*

We call $\sigma(\bullet, \bullet)$ the approaching degree on $F(X)$, $\sigma(\underset{\sim}{A}, \underset{\sim}{B})$ the approaching degree between $\underset{\sim}{A}$ and $\underset{\sim}{B}$.

Now let us turn back to the inverse problem of fuzzy comprehensive evaluation. In fact, it is almost impossible to get the exact solution (w_{I1}, w_{I2}, w_{I3}) of the equation (4.3). In practice, W_I is obtained by seeking a vector W_I which enables the Hamming approaching degree

$$\sigma_H(A_I, B_I) = 1 - \frac{1}{n} \sum_{i=1}^{n} \mid u_{A_I}(x_i) - u_{B_I}(x_i) \mid$$

between $A_I = W_I \circ R_I$ and B_I to reach its maximum using optimization method — linear search algorithm.

Under the optimization method, these six groups have six weight vectors respectively. In fact, these six $W_I s$ should have the same components, say $W_I = W(w_1, w_2, w_3)$, $\forall I$, so we

improve (4.3) to the problem (4.4).

$$\begin{aligned}&\mathrm{Max}\Big(\frac{1}{6}\sum_{I=1}^{6}\sigma_H(W\circ R_I, B_I)\Big)\\&\text{restriction:}w_1\geq 10\%, w_2\geq 10\%, w_3\geq 40\%, \sum_{i=1}^{3} w_i = 1.\end{aligned}\tag{4.4}$$

At last, we get $W = (0.1, 0.1, 0.8)$, the maximum is 0.9247, and the six $\sigma_H s$ all are larger than 0.8830.

§5. Information Distribution Method (IDM)

5.1 Introduction of IDM

It is well known that the membership function may be determined by using probabilistic distribution in some cases. As mentioned above, with the presumption that the total number of residents for a group is big enough, the membership grade for the illness fuzzy set of this group may be obtained by using statistical methods, even simply calculating the statistical frequency, so we may regard it as a standard. But the statistical method may fail in determining the illness fuzzy set if the group under consideration is a small one.

For instance, we consider two groups, one group G consists of all the 6495 records from old people under the above-mentioned survey, and the other G_1 consists of 700 records, drawn randomly from G. As the influence of BMI on serious illness of hypertension is concerned, a comparison of the results from G and from G_1 by calculating the occurring frequency of serious illness in each interval of BMI is shown below in Fig.1.

Fig.1 The results obtained by statistical method

It is found that the membership grade curve of serious illness of G_1 encircling the one of G fluctuates up and down, the information carried by which is incomplete mostly because there are few records falling into each interval of BMI, so it is unreasonable to determine the membership grade for the illness fuzzy set by calculating the occurring frequency. If we regard the statistic result from G as a standard, the total average relative error of G_1 is

19.56%. In this case, the information we obtained by incomplete data (small sample G_1) is the so-called fuzzy information. According to the theory of fuzzy information optimization, we can solve this kind of problems by using information distribution method (IDM).

5.2 Mathematics Models

The original concept of IDM comes from the case where significant sample data are scarce and it is hard to carry out statistics in knowledge engineering. In order to describe the IDM for the small sample in this paper, some basic definitions are needed. We provide the simple notion to IDM while focusing primarily on the aspects of IDM needed to solve small sample problem. For additional details on IDM, the reader is referred to [3].

Definition 5.1. *Let $W = \{w_1, w_2, \cdots, w_n\}$ be a given sample, and $U = \{u_1, u_2, \cdots, u_m\}$ be the discrete universal field of W. A mapping from $W \times U$ to $[0, 1]$*

$$\mu : W \times U \to [0, 1]$$

$(w, u) \to \mu(w, u)$ is called an information distribution of W on U if $\mu(w, u)$ has the following properties:

(i) *$\forall w \in W$, if $\exists u \in U$, such that $w = u$, then $\mu(w, u) = 1$, i.e., μ is reflexive.*

(ii) *for $w \in W$, $\forall u', u'' \in U$, if $\|u' - w\| \le \|u'' - w\|$, then $\mu(u', w) \ge \mu(u'', w)$, i.e., μ is decreasing when $\|u - w\|$ is increasing.*

(iii) $\sum_{j=1}^{m} \mu(w_i, u_j) = 1$, *$i = 1, 2, \cdots, n$, i.e., conservation of information.*

Definition 5.2. *$Q_i = \sum_{j=1}^{m} q_{ij}$ is called the information gross provided by W from controlling point u_i, and $Q = (Q_1, Q_2, \cdots, Q_n)$ is called the original information distribution matrix for W on U, in short, information distribution matrix.*

Definition 5.3. *The vector, which is made up of the quantities of information gains from each sample point w_j contributing to each controlling point u_i in U in some fashion μ, is called the information row vector of sample point w_j, denoted by $q_{ij} = \mu(u_i, w_j)$.*

Definition 5.4. *Let e_o denote the error obtained by former method and e_N denote the error by new method. Then $\rho = \frac{e_0 - e_N}{e_0} \times 100\%$ is called the reducible deviation by new method.*

5.3 The Applications of Information Distribution Method

Now let us apply IDM to solve the small sample G_1 problem above. At first, let us denote the universal field of BMI by $\{b | b \in [17, 30]\}$ and of risk of hypertension by $V = \{v_0, v_1, v_2\}$ =$\{$free from hypertension, mild, serious$\}$. Implementation of the IDM for group G_1 can be processed in the following steps.

(1) Determine controlling points: Having divided the universal field of BMI into 13 grades with equal space, we select mid point of each interval as controlling point, i.e.

$$B = \{b_1, b_2, \cdots, b_{13}\} = \{17.5, 18.5, \cdots, 29.5\}.$$

(2) Construct information distribution matrix Q: Q constructed from B and V is used to save up the information of G_1. The detail of construction is as follows.

Each record includes two segments b (the value of BMI) and v (v_0 or v_1 or v_2), and b could contribute its information to one or two adjacent controlling points subject to the condition that total amount of information is equal to one. Suppose that information distribution here

is conducted in linear form, i.e.

$$\mu(b, b_i) = \begin{cases} 1 - \frac{|b-b_i|}{\Delta}, & \mid b - b_i \mid \leq \Delta, \\ 0, & \mid b - b_i \mid > \Delta, \end{cases}$$

where $i = 1, 2, \cdots, 13$ and Δ denotes spacing.

As an illustration, consider a record with b=22.9 and $v = v_1$. When contributing its information to Q only controlling points $b_6 = 22.5$ and $b_7 = 23.5$ get the share. Therefore

$$q_{6,1} = 1 - \frac{\mid b - b_6 \mid}{\Delta} = 1 - \frac{\mid 22.9 - 22.5 \mid}{1} = 0.6,$$
$$q_{7,1} = 1 - \frac{\mid b - b_7 \mid}{\Delta} = 1 - \frac{\mid 22.9 - 23.5 \mid}{1} = 0.4.$$

After 700 original records of G_1 have been treated with this process and the information gains at each controlling point have been summed up, an information distribution matrix Q will turn out.

(3) Establish the fuzzy relationship matrix $\underset{\sim}{R}$. Each element in $\underset{\sim}{R}$ can be obtained by normalizing each element of information map Q, i.e.

$$r_{ij} = \frac{Q_{ij}}{\sum_{i=1}^{13}\sum_{j=0}^{2} Q_{ij}} \quad (i = 1, 2, \cdots, 13; \quad j = 0, 1, 2).$$

(4) Risk evaluation with single factor. Given the magnitude for BMI, we can calculate the membership grade for the illness fuzzy set by using fuzzy transformation $\underset{\sim}{V_0} = B_0 \circ \underset{\sim}{R}$, where '$\circ$' stands for fuzzy operator $M(\wedge, \vee)$. For example, given an old people with BMI= $29kg/m^2$, whose information row vector should be

$$B_0 = \{0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.5, 0.5\},$$

so his (her) risk vector is

$$\underset{\sim}{V_0} = B_0 \circ \underset{\sim}{R} = 0.058/v_0 + 0.032/v_1 + 0.030/v_2.$$

For defuzzification, it is normalized and the illness fuzzy set with BMI=29 (still denoted by $\underset{\sim}{V_0}$) is

$$\underset{\sim}{V_0} = 0.483/v_0 + 0.267/v_1 + 0.250/v_2.$$

Finally, we compare the result from large sample G statistics with the above result from small sample G_1 obtained by IDM (see Fig.2), and find that their curves are almost consistent. The total average relative error of G_1 by IDM is 12% and the reducible deviation by IDM compared with statistical method is 38.65%. It is shown that the IDM is a better method when dealing with small sample.

Fig.2 The comparison of results obtained by statistical method and IDM

§6. Improved Information Distribution Method (IIDM)

6.1 Selection of Information Distribution Functions and Controlling Points

In the IDM, the linear function and the mid-point of BMI interval are always chosen as the information distribution function and the controlling points. But in this paper, we try to make a more flexible choice and call the new method IIDM. Here we choose a non-linear distribution function:

$$\mu(b,b_i)=\begin{cases}1-\frac{|b-b_i|^\lambda}{\Delta^\lambda}, & |\,b-b_i\,|\le\Delta\\ 0, & |\,b-b_i\,|>\Delta\end{cases}\quad(\lambda>0)$$

and compare it with the original one. As to the controlling points, let θ be the shift from original controlling points, that is, we move the controlling points $|\ \theta\ |$ units from their original location. When $\theta<0$, move to the left, and $\theta>0$, the right.

For the sake of briefness, let BMI only be the integer from the interval $[17,30)kg/m^2$, $D_I(\text{BMI},\lambda,\theta)$ be a function of membership grade I of BMI by using IDM on G_I, and $S_I(\text{BMI})$ be the occurring frequency of membership grade I in the BMI-th stage by using statistical method on G, $I=0,1,2$, where 0 stands for free from disease, 1 for mild, and 2 for serious. The function

$$f(\lambda,\theta)=\sum_{I=0}^{2}\sum_{\text{BMI}=17}^{29}|\,D_I(\text{BMI},\lambda,\theta)-S_I(\text{BMI})\,|^2$$

(where λ, θ are above-mentioned) is called the objective function in this paper.

Under the condition

$$\begin{cases}g_1(\lambda,\theta)=\lambda-3\le0,\\ g_2(\lambda,\theta)=-\lambda\le0,\\ g_3(\lambda,\theta)=\theta-1\le0,\\ g_4(\lambda,\theta)=-\theta-1\le0,\end{cases}$$

we minimize the object function $f(\lambda,\theta)$ in the model by solving the sub-problem of quadratic programming based on quadratic approximation to Lagrangian function

$$L(\lambda,\theta,l)=f(\lambda,\theta)+\sum_{i=1}^{4}l_ig_i(\lambda,\theta),$$

and applying the linear search algorithm. The final result is as follows:

$$\min\{f(\lambda,\theta) \mid g_i(\lambda,\theta) \le 0, i = 1,2,3,4\} = 0.0175,$$

where $\widehat{\lambda} = 1.67, \widehat{\theta} = 0.18$. Then we apply the results, with appropriate distribution function

$$\mu(b,b_i) = \begin{cases} 1 - \frac{|b-b_i|^{1.67}}{\Delta^{1.67}}, & \mid b-b_i \mid \le \Delta, \\ 0, & \mid b-b_i \mid > \Delta, \end{cases}$$

and controlling points moving $\mid \widehat{\theta} \mid = 0.18$ units to the right from their original locations, to the 700 original data, and make comparisons with the result from large samples statistics. The result is quite good (see Fig.3). The total average relative error of G_1 by improved IDM is 9.72% and the reducible deviation ρ by improved IDM compared with the original IDM is 19%.

Fig.3 The comparison of results obtained by various methods

6.2 Selection of Fuzzy Operators

The comparison are made between different fuzzy operators '$\circ$' used in the fuzzy transformation $\underset{\sim}{V_0} = B_0 \circ \underset{\sim}{R}$. '$\circ$' is set to be $M(\wedge,\vee)$, $M(\bullet,\vee)$,$M(\wedge,\oplus)$ and $M(\bullet,\oplus)$ respectively. It is found that the best one is $M(\bullet,\oplus)$ in Table 1.

Fuzzy Operator $\circ$	$M(\wedge,\vee)$	$M(\bullet,\vee)$	$M(\wedge,\oplus)$	$M(\bullet,\oplus)$
Total Average Relative Error of 700	12%	10.02%	8.36%	8.01%
ρ by IDM Compared with Statistic	38.65%	48.77%	57.26%	59.05%

Table 1 Comparison made between different fuzzy operators $\circ$

§7. Conclusions

7.1 Discussion on the Inverse Problem of Comprehensive Evaluation

The method of finding the optimal approximate solution for the inverse problem of fuzzy comprehensive evaluation introduced in this paper is an improved method based on the conventional methods, it gives a new idea to solve this kind of inverse problem. However,

there are many aspects worthy to study to perfect this method, for instance, how to select the representative groups and which approaching degree of fuzzy sets is most suitable, etc.

7.2 Advantages and Problems Concerning IIDM

The IIDM given in this paper shows a way to choose optimal information distribution function and optimal controlling points, so it is better than IDM in some cases. This method may not only be applied to the risk evaluation for some diseases, but also be applied to some other small sample problems provided that there is a reference large sample. In addition, there are many problems worthy to be discussed, such as how to choose optimal step of BMI, etc.

7.3 Advantages of the Hybrid Technique

Among various techniques applied in the area of risk analysis and evaluation, the hybrid technique advanced in this paper is a new and promising one. It is found that the result obtained by Hybrid Method is better than that by single Fuzzy Method or by single Statistical Method in some cases. Indeed, the idea integrating Fuzzy Method with Statistical Method and Optimization Method has some remarkable advantages and shows itself a vital force.

Acknowledgement. Dr. Huang Chongfu from Beijing Normal University gave us great help for this paper, we hereby express our heartfelt thanks to him.

References

[1] Cummins, J. D. & Derrig, R. A., Fuzzy financial pricing of property-liability insurance [J], *North American Actuarial Journal,* **1**(1997), 21–40.

[2] Chen Jengwo & He Zesheng, Using analytic Hierarchy process and fuzzy set theory to rate and rank the disability [J], *Journal of Fuzzy Sets and Systems,* **88**(1997), 1–22.

[3] Huang Chongfu & Wang Jiading, Technology of fuzzy information optimization processing and applications [M], Beijing University of Aeronautics and Astronautics Press, 1995.

[4] Zhang Shiwei & Lu Yuchu, Fuzzy mathematics and its applications [M], Tong Ji University Press, 1991.