Interval Probability of Data Querying Based on Fuzzy Conditional Probability Relation*

Rolly Intan †

Abstract—**This paper discusses fuzzification of crisp domains into fuzzy classes providing fuzzy domains. Relationship between two fuzzy domains, X_i and X_j, is represented by a matrix, w_{ij}. If X_i and X_j have n and m elements of fuzzy data, respectively, then w_{ij} is $n \times m$ matrix. The primary goal of the paper is to generate and provide some formulas for predicting interval probability in the relation to data querying, i.e., given John is 30 years old and he has MS degree, what is his probability to getting high salary.**

Keywords: Fuzzy Conditional Probability Relation, Data Querying, Interval Probability, Mass Assignment, Point Semantic Unification.

1 Introduction

In this paper, we process a certain relational database by classifying every domain into several value of data or elements, i.e.[1], component *age* can be classified into *about_20*, *about_25*, $\cdots$. Assuming that every classified data is a fuzzy set, we must determine a membership function which represents degree of element belonging to the fuzzy set. Then, we construct a model of system data to describe interrelationship among all components of the system using conditional probabilistic theory. Relationship between two components, X_1 and X_2, of a system is expressed in a matrix w_{12}. If component X_1 has n elements, X_2 has m elements then matrix w_{12} is $n \times m$ matrix, where $a^{12}_{ij} \in w_{12}$ expresses *weight* or degree of dependency of $x_{2j} \in X_2$ from $x_{1i} \in X_1$, for $1 \le i \le n$, $1 \le j \le m$. Through this model, we generate some formulas to predict any value of data related to a given query of input data i.e., given John is 30 years old and he has MS degree, what is his probability of getting high salary? Given input of data querying can be precise as well as imprecise data (fuzzy data). First, before the data can be used to make prediction, we must find their probabilistic matching related to elements of components of system by using *Point Semantic Unification Process* as introduced in [6, 8, 11]. Here, *Point Value Semantic Unification* can be considered as a conditional probability between two fuzzy sets. Two different formulas are provided to calculate upper and lower bound probabilities of prediction. Hence, result of prediction works into a interval truth value $[a, b]$ where $a \le b$ as proposed in [7].

2 Basic Concept

2.1 Conditional Probability

$P(H \mid D)$ is defined as a conditional probability for H given D. Relation between conditional and unconditional probability satisfies the following equation [10].

$$P(H \mid D) = \frac{P(H \cap D)}{P(D)}, \tag{1}$$

where $P(H \cap D)$ is an unconditional probability of compound events 'H and D happen'. $P(D)$ is unconditional probability of event D.

2.2 Point Semantic Unification

Given f is a fuzzy set defined on the discrete space $X = \{x_1, x_2, ..., x_n\}$, namely
$f = \{\chi_1/x_1,\ \chi_2/x_2,\ ...,\ \chi_n/x_n\}$, where $\chi_i \in [0, 1]$ is a membership degree of x_i in fuzzy set f.
Suppose f is a normal fuzzy set whose elements are ordered such that $\chi_1 = 1,\ \chi_i \ge \chi_j \quad if \quad i < j$; The mass assignment corresponding to the fuzzy set f is [6, 8]

$$m_f = \{\{x_1, x_2, ..., x_i\} : \chi_i - \chi_{i+1}\}, \quad with\ \chi_{n+1} = 0. \tag{2}$$

Let $m_f = \{L_i : l_i\}$ and $m_g = \{N_j : n_j\}$ be mass assignments associate with the fuzzy sets f and g. Relation between m_f and m_g is represented by a matrix M. From the matrix,

$$M = \{m_{ij}\} = \left\{ \frac{card(L_i \cap N_j)}{card(N_j)} \right\} \cdot l_i \cdot n_j. \tag{3}$$

The probability $P(f \mid g)$ is given by [6]:

$$P(f \mid g) = \sum_{ij} m_{ij}. \tag{4}$$

For example, let $f = \{1/a, 0.7/b, 0.2/c\}$ and $g = \{0.2/a, 1/b, 0.7/c, 0.1/d\}$ are defined on $X = \{a, b, c, d, e\}$, as arbitrarily given by
$m_f = \{a : 0.3,\ \{a, b\} : 0.5,\ \{a, b, c\} : 0.2\}$,
$m_g = \{b : 0.3, \{b, c\} : 0.5, \{a, b, c\} : 0.1, \{a, b, c, d\} : 0.1\}$.
From the following matrix(i.e. $m_{13} = 0.01,\ m_{33} = 0.03$),

*Extended version of the paper presented in IMECS 2007[5]

†Department of Informatics Engineering, Petra Christian University, Surabaya, Indonesia. Tel/Fax: 62-31-2983420/ 62-31-8417658 Email: rintan@petra.ac.id

	0.3 {b}	0.5 {b,c}	0.1 {a,b,c}	0.1 {a,b,c,d}
0.3 {a}	0	0	0.01	0.00075
0.5 {a,b}	0.15	0.125	0.0333	0.025
0.2 {1,b,c}	0.06	0.1	0.02	0.015

the probability $P(f \mid g) = 0.53905$. It can be proved that Point Semantic Unification satisfies $P(f \mid g) + P(\bar{f} \mid g) = 1$. Thus, Point Semantic Unification is considered as a conditional probability.

2.3 Interval Probability

An interval probability $IP(E)$ can be interpreted as a scope of probability of event E, $P(E)$, i.e $IP(E) = [e_1, e_2]$ means $e_1 \leq P(E) \leq e_2$, where e_1 and e_2 are minimum and maximum probability of E respectively [7]. Given two probabilities $P(A) = a$ and $P(B) = b$ for event A and B, where $a, b \in [0, 1]$, interval probability of compound event 'A and B happened' is defined by

$$IP(A \cap B) = [\max(0, a + b - 1), \min(a, b)]. \tag{5}$$

Interval probability of compound event 'A or B happened' is defined by

$$IP(A \cup B) = [\max(a, b), \min(1, a + b)]. \tag{6}$$

3 Construction Model of System Data

A system data is defined as $S(Er, X)$. Here, Er is the number of data entries or number of records, and X is a set of domains or components in the system. If there are n components then $X = (X_1, ..., X_n)$. For example, given CAREER DATABASE in Table 2.1[1]. By assuming that CAREER is a system data, it has 10 entries and three components, *education*, *age*, and *salary*, therefore $Er = 10$, $X = \{X_1 : education, X_2 : age, X_3 : salary\}$. Now, we try to find relation among *education*, *age*, and *salary*.

Table 2.1. CAREER DATABASE

Rec#	Education	Age	Sallary
#1	MS	35	400,000
#2	SHS	24	150,000
#3	PhD	44	470,000
#4	JHS	45	200,000
#5	ES	35	125,000
#6	SHS	37	250,000
#7	MS	39	420,000
#8	SHS	27	175,000
#9	MS	45	415,000
#10	SHS	56	275,000

First, all three domains are classified as follows.
$education = \{low\_edu, mid\_edu, hi\_edu\}$,
$age = \{about\_20, about\_25, ..., about\_60\}$,
$salary = \{low\_slr, mid\_slr, hi\_slr\}$,
where we assume that membership functions of *low_edu*, *mid_edu*, and *high_edu* are given by
$low\_edu = \{1/N, 0.8/ES, 0.5/JHS\}$,
$mid\_edu = \{0.2/ES, 0.5/JHS, 0.9/SHS, 0.2/BA\}$,
$hi\_edu = \{0.1/SHS, 0.8/BA, 1/MS, 1/PhD\}$.
In general formula, membership function of *age*,
$about\_n = \{0.2/(n-4), 0.4/(n-3), 0.6/(n-2), 0.8/(n-1), 1/n, 0.8/(n+1), 0.6/(n+2), 0.4/(n+3), 0.2/(n+4)\}$.
Membership functions of *salary* are given by trapezoidal or triangular membership functions, as follow.
$low\_slr = [1/0, 1/100000, 0/150000]$,
$mid\_slr = [0/100000, 1/150000, 1/250000, 0/300000]$,
$hi\_slr = [0/250000, 1/300000]$.
Through all membership functions above, we calculate and transform Table 2.1 into Table 2.2.

Table 2.2. CAREER FUZZY VALUE

Rec#	Education			Age				Salary		
	low_edu	mid_edu	hi_edu	about_20	about_25	...	about_60	low_slr	mid_slr	hi_slr
#1	0	0	1	0	0	...	0	0	0	1
#2	0	0.9	0.1	0.2	0.8	...	0	0	1	0
#3	0	0	1	0	0	...	0	0	0	1
#4	0.5	0.5	0	0	0	...	0	0	1	0
#5	0.8	0.2	0	0	0	...	0	0.5	0.5	0
#6	0	0.9	0.1	0	0	...	0	0	1	0
#7	0	0	1	0	0	...	0	0	0	1
#8	0	0.9	0.1	0	0.6	...	0	0	1	0
#9	0	0	1	0	0	...	0	0	0	1
#10	0	0.9	0.1	0	0	...	0.2	0	0.5	0.5
$\sum$	1.3	4.3	4.4	0.2	1.4	...	0.2	0.5	5	4.5

X_n is defined as compound attributes to express component of the system, where X_n is considered as a vector. If there are k elements of X_n then $Xn = (x_{n1}, ..., x_{nk})$, where x_{ni} is element i of compound attribute X_n and for further, x_{ni} is called attribute. For example, if system CAREER has three compound attributes, $X_1 : education$, $X_2 : age$ and $X_3 : salary$, then $x_{11} = low\_edu$, $x_{25} = about\_40$, $x_{31} = low\_slr$. e_j^{ni} is defined as membership's value of entry j for attribute x_{ni}. For example, as shown in Table 2.2., $e_4^{11} = 0.5, e_2^{12} = 0.9, e_2^{21} = 0.2$, etc. If compound attribute X_n has k attributes, it can be proved that $\forall j \sum_{1 \leq i \leq k} e_j^{ni} = 1$.
$N(x_{ni})$ is defined as sum of entries value for attribute x_{ni} as follows.

$$N(x_{ni}) = \sum_{1 \leq j \leq Er} e_j^{ni}. \tag{7}$$

If compound attribute X_n has k attributes, it can be proved that $Er = \sum_{1 \leq i \leq k} N(x_{ni})$.
For example, as shown in Table 2.2., $N(x_{11}) = N(low\_edu) = 1.3$.
$P(x_{ni})$ is defined as probability of attribute x_{ni} as follows.

$$P(x_{ni}) = \frac{N(x_{ni})}{Er}. \tag{8}$$

If compound attribute X_n has k attributes then, it can be proved that $\sum_{1 \leq i \leq k} P(x_{ni}) = 1$.

3.1 Relation Among Compound Attributes

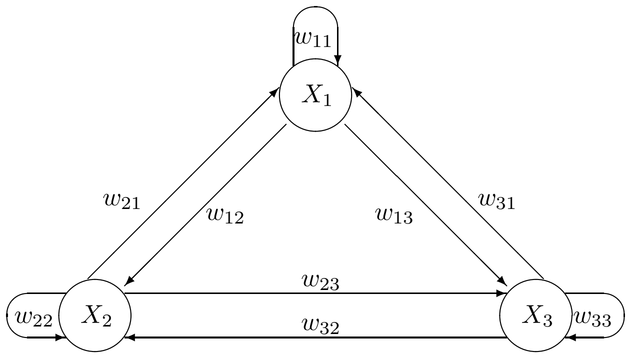


Figure 2.1: Relation Among Compound Attributes, X_1, X_2, and X_3.

Given three compound attributes, X_1,X_2 and X_3. Relation among them can be illustrated in Figure 2.1. w_{nm} is defined as a *weight matrix*, to express degree of dependency of X_m from X_n. For a k-compound attribute X_n and a j-compound attribute X_m, $w_{nm} = (a_{ih}^{nm})_{k\times j}$ and $w_{mn} = (a_{hi}^{mn})_{j\times k}$ present two different matrices as given by

$$w_{nm} = \begin{bmatrix} a_{11}^{nm} & a_{12}^{nm} & \cdots & a_{1j}^{nm} \\ a_{21}^{nm} & a_{22}^{nm} & \cdots & a_{2j}^{nm} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1}^{nm} & a_{k2}^{nm} & \cdots & a_{kj}^{nm} \end{bmatrix}$$

$$w_{mn} = \begin{bmatrix} a_{11}^{mn} & a_{12}^{mn} & \cdots & a_{1k}^{mn} \\ a_{21}^{mn} & a_{22}^{mn} & \cdots & a_{2k}^{mn} \\ \vdots & \vdots & \ddots & \vdots \\ a_{j1}^{mn} & a_{j2}^{mn} & \cdots & a_{jk}^{mn} \end{bmatrix}$$

Each element of matrix w_{nm}, entry $a_{ih}^{nm} \in [0,1]$ expresses numerical probabilistic value of relation from $x_{ni} \in X_n$ to $x_{mh} \in X_m$. a_{ih}^{nm} can also be interpreted as conditional probability as follows.

$$a_{ih}^{nm} = P(x_{ni} \mid x_{mh}) = \frac{P(x_{ni} \cap x_{mh})}{P(x_{mh})}. \tag{9}$$

If there are Er number of entries, then

$$a_{ih}^{nm} = \frac{\sum_{1\le j\le Er} \min(e_j^{ni}, e_j^{mh})}{\sum_{1\le j\le Er} e_j^{mh}}. \tag{10}$$

where $P(x_{ni} \cap x_{mh})$ expresses probability of entries which be inside x_{ni} and x_{mh}.

In [4], (10) is defined as a *fuzzy conditional probability relation.*

On the other hand, a_{hi}^{mn} expresses numerical probabilistic value of relation from $x_{mh} \in X_m$ to $x_{ni} \in X_n$. a_{hi}^{mn} can also be interpreted as a conditional probability as follows.

$$a_{hi}^{mn} = P(x_{mh} \mid x_{ni}) = \frac{P(x_{ni} \cap x_{mh})}{P(x_{ni})}. \tag{11}$$

If there are Er number of entries, then

$$a_{hi}^{mn} = \frac{\sum_{1\le j\le Er} \min(e_j^{ni}, e_j^{mh})}{\sum_{1\le j\le Er} e_j^{ni}}. \tag{12}$$

From equations (9-10) and (11-12), we conclude that a_{ih}^{nm} and a_{hi}^{mn} are in general different.

The above definition leads to the conclusion that every attribute can be used to determine itself perfectly.

$$\forall X_n, x_{ni} \in X_n, \quad \frac{P(x_{ni} \cap x_{ni})}{P(x_{ni})} = 1. \tag{13}$$

If compound attribute X_n has k attributes, then,

$$w_{nn} = \begin{bmatrix} 1 & a_{12}^{nn} & \cdots & a_{1k}^{nn} \\ a_{21}^{nn} & 1 & \cdots & a_{2k}^{nn} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1}^{nn} & a_{k2}^{nn} & \cdots & 1 \end{bmatrix} \tag{14}$$

3.2 Relation Among Attributes In System

Given three attributes, $x_{1u} \in X_1$, $x_{2v} \in X_2$ and $x_{3r} \in X_3$. Relation among these three attributes can be seen in Figure 2.2. as follows.

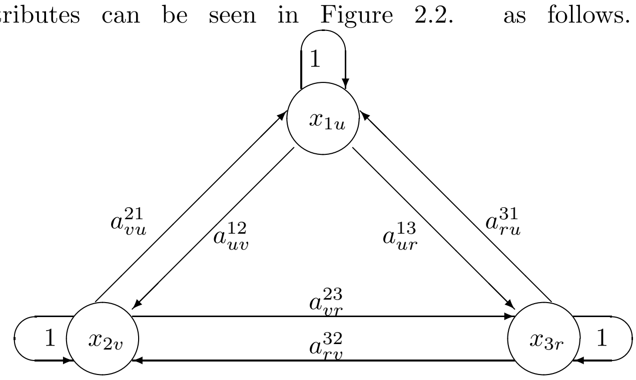
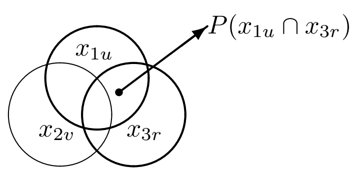


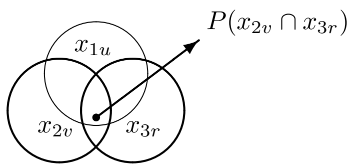
Figure 2.2: Relation Among Attributes, x_{1u}, x_{2v}, x_{3r}.

In order to understand the meaning of this connection, we use relation of sets.

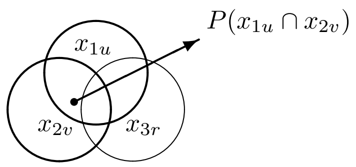


$$\begin{aligned} P(x_{1u} \cap x_{3r}) &= \frac{P(x_{1u} \cap x_{3r})}{P(x_{1u})} \cdot P(x_{1u}) = a_{ru}^{31} \cdot P(x_{1u}) \\ &= \frac{P(x_{1u} \cap x_{3r})}{P(x_{3r})} \cdot P(x_{3r}) = a_{ur}^{13} \cdot P(x_{3r}). \end{aligned}$$

Both $a^{31}_{ru} \cdot P(x_{1u})$ and $a^{13}_{ur} \cdot P(x_{3r})$, point to the same area or quantity, are intersection between x_{1u} and x_{3r}. In the same way, we can find two other relations, $a^{23}_{vr} \cdot P(x_{3r}) = a^{32}_{rv} \cdot P(x_{2v})$ and $a^{12}_{uv} \cdot P(x_{2v}) = a^{21}_{vu} \cdot P(x_{1u})$, which are proved as follows.



$$\begin{aligned} P(x_{2v} \cap x_{3r}) &= \frac{P(x_{2v} \cap x_{3r})}{P(x_{2v})} \cdot P(x_{2v}) = a^{32}_{rv} \cdot P(x_{2v}) \\ &= \frac{P(x_{2v} \cap x_{3r})}{P(x_{3r})} \cdot P(x_{3r}) = a^{23}_{vr} \cdot P(x_{3r}). \end{aligned}$$



$$\begin{aligned} P(x_{1u} \cap x_{2v}) &= \frac{P(x_{1u} \cap x_{2v})}{P(x_{1u})} \cdot P(x_{1u}) = a^{21}_{vu} \cdot P(x_{1u}) \\ &= \frac{P(x_{1u} \cap x_{2v})}{P(x_{2v})} \cdot P(x_{2v}) = a^{12}_{uv} \cdot P(x_{2v}. \end{aligned}$$

From the relations above, we find the following equation.

$$\frac{a^{12}_{uv} \cdot a^{23}_{vr}}{a^{32}_{rv}} = \frac{a^{21}_{vu} \cdot a^{13}_{ur}}{a^{31}_{ru}}. \quad (15)$$

Proof:

$$\begin{aligned} a^{12}_{uv} \cdot P(x_{2v}) &= a^{21}_{vu} \cdot P(x_{1u}) \\ a^{12}_{uv} \cdot (a^{23}_{vr} \cdot \frac{P(x_{3r})}{a^{32}_{rv}}) &= a^{21}_{vu} \cdot (a^{13}_{ur} \cdot \frac{P(x_{3r})}{a^{31}_{ru}}) \\ a^{12}_{uv} \cdot \frac{a^{23}_{vr}}{a^{32}_{rv}} &= a^{21}_{vu} \cdot \frac{a^{13}_{ur}}{a^{31}_{ru}}. \end{aligned}$$

Important characteristic of relation among attributes is *transitive relation*, i.e. given a^{12}_{uv}, a^{21}_{vu}, a^{23}_{vr}, a^{32}_{rv} and we would like to find interval value of a^{13}_{ur}, which satisfies the two following equations.
Lower bound of a^{13}_{ur},

$$a^{13}_{ur} \geq \max\{0, (a^{12}_{uv} + a^{32}_{rv} - 1)\} \cdot \frac{a^{23}_{vr}}{a^{32}_{rv}}. \quad (16)$$

Upper bound of a^{13}_{ur},

$$\begin{aligned} a^{13}_{ur} \leq\ & \min\{a^{32}_{rv}, a^{12}_{uv}\} \cdot \frac{a^{23}_{vr}}{a^{32}_{rv}} + \quad (17) \\ & \min\{(1 - a^{21}_{vu}) \cdot \frac{a^{12}_{uv}}{a^{21}_{vu}}, (1 - a^{23}_{vr}) \cdot \frac{a^{32}_{rv}}{a^{23}_{vr}}\} \cdot \frac{a^{23}_{vr}}{a^{32}_{rv}}. \end{aligned}$$

Proof

To find the upper bound of a^{13}_{ur}, first we take the maximum area inside x_{2v}, result of intersection between two intersection areas which are intersection between x_{1u} and x_{2v}, expressed in a^{12}_{uv} and intersection between x_{3r} and x_{2v}, expressed in a^{32}_{rv}. The maximum area that is result of overlapping between the two intersection areas can be expressed in min function applied to a^{12}_{uv} and a^{32}_{rv}. The next, we plus with maximum intersection between remain x_{1u} and x_{2v} which be outside of x_{2v}. Again, this area can be expressed in min function applied to $(1 - a^{21}_{vu})$ and $(1 - a^{23}_{vr})$. Value of these two area point to two different area, x_{1u} and x_{3r}. However, in order to be compared, they must point to the same area, in this case we use x_{2v} as base for their comparison. Therefore, we must convert them into x_{2v} by multiplying with $\frac{a^{12}_{uv}}{a^{21}_{vu}}$ and $\frac{a^{32}_{rv}}{a^{23}_{vr}}$, respectively. Finally, again we must convert all from x_{2v} into x_{3r} by multiplying with $\frac{a^{32}_{rv}}{a^{23}_{vr}}$.

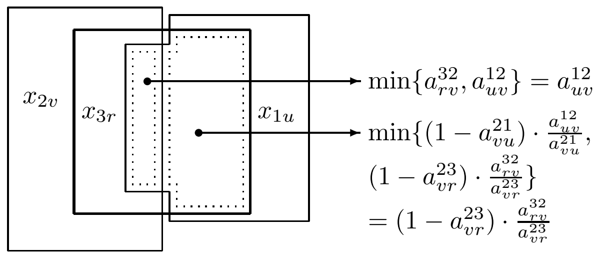


Figure 2.3: Maximum Area of Intersection between x_{1u} and x_{3r} inside x_{2v}.

To find the lower bound of a^{13}_{ur}, we take the minimum area inside x_{2v}, result of intersection between two intersection areas which are intersection between x_{1u} and x_{2v}, expressed in a^{12}_{uv} and intersection between x_{3r} and x_{2v}, expressed in a^{32}_{rv}. The minimum area which is result of as much as possible avoid overlapping between the two intersection areas can be expressed in *max* function applied to a^{12}_{uv} and a^{32}_{rv} as shown in (16). The next, we convert quantity of the maximum area from x_{2v} into x_{3r} by multiplying with $\frac{a^{32}_{rv}}{a^{23}_{vr}}$.

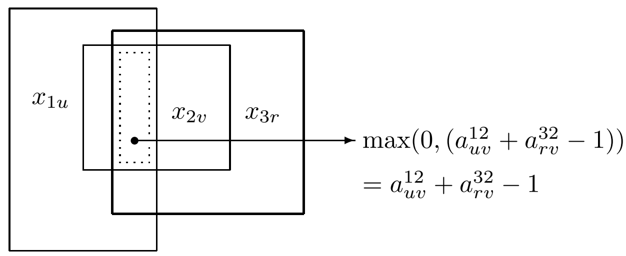


Figure 2.4: Minimum Area of Intersection between x_{1u} and x_{3r} inside x_{2v}.

4 Calculating Prediction

Constructed model of the system can be used to predict interval probability (find lower and upper bound) of any data querying. This section generates formulas to calculate interval probability of the data querying. First, user must give input related to the compound attributes. Q is defined as a set of input data given by user to do query for a certain data. If there are n compound attributes then $Q = \{q_1, ...q_n\}$ where q_i is data input related to compound attribute X_i. For example, suppose CAREER system has been constructed, given John is *old* man and has *MS* degree as input for *age* and *education* respectively, where $q_1 = old$ and $q_2 = MS$.

$P(X_i, q_i)$ is defined as probabilistic matching of compound attribute X_i toward given input data q_i. If there are k elements or attributes of compound attribute X_i, then,

$$P(X_i, q_i) = (p_{i1}, ..., p_{ik}), \tag{18}$$

where $p_{ij} = P(x_{ij} \mid q_i)$ expresses conditional probability for x_{ij} given q_i. In this case Point Semantic Unification Process [6, 8] can be used to calculate p_{ij}.

For example, given $q_i = old$ which is a *fuzzy set* defined as $q_i = [0/55, 1/60]$. $X_i = age$ has 9 attributes as defined in Section 2, as given by $X_i = \{about\_20, about\_25, \cdots, about\_60\}$. By applying *point semantic unification process* to membership function of *age* as defined in Section 2 and membership function of q_i, we calculate $P(X_i, q_i)$ as follows. First, we calculate *the mass assignment* for q_i. It is equivalent to the basic probability assignment of Dempster Shafer Theory as given by $m_{q_i} = \{56, 57, 58, 59, 60\} : 0.2, \{57, 58, 59, 60\} : 0.2\{58, 59, 60\} : 0.2, \{59, 60\} : 0.2, \{60\} : 0.2$. Next, i.e. *mass assignment* for $x_{i8} = 55$ as one attribute of X_i is given by $m_{x_{i8}} = \{51, ..., 59\} : 0.2, \{52, ..., 58\} : 0.2, \{53, ..., 57\} : 0.2, \{54, 55, 56\} : 0.2, \{55\} : 0.2$. Process to calculate *Point Value Semantic Unification* of relation between two fuzzy set, *old* and *about*_55 or $P(about\_55, old)$ is shown in the following table.

	0.2 {56,. . .,60}	0.2 {57,. . .,60}	0.2 {58,59,60}	0.2 {59,60}	0.2 {60}
0.2 {51,...,59}	0.032	0.03	0.026	0.02	0
0.2 {52,...,58}	0.024	0.02	0.013	0	0
0.2 {53,...,57}	0.016	0.01	0	0	0
0.2 {54,55,56}	0.008	0	0	0	0
0.2 {55}	0	0	0	0	0

From the table, we calculate $P(about\_55, old) = 0.199$. In the same way, we find $P(about\_60, old) = 0.799$, where $P(about\_20, old) = P(about\_25, old) = \cdots = P(about\_50, old) = 0$, because there is no intersection between their members. Finally, we found,
$P(X_i, q_i) = P(age, old) = (0, 0, 0, 0, 0, 0, 0, 0.199, 0.799)$.

$P(X_i, q_j)$ is defined as probability of attribute X_i influenced by given input data q_j. X_i and q_j have different type of data, therefore to find their probabilistic matching, first, we must find $P(X_j, q_j)$ and then apply max-multiply (*) operation between $P(X_j, q_j)$ and w_{ji} as follows. If X_i has k attributes and X_j has s attributes then,

$$\begin{aligned} P(X_i, q_j) &= P(X_j, q_j) * w_{ji} && (19)\\ &= (p_{j1}, ..., p_{js}) * \begin{bmatrix} a^{ji}_{11} & a^{ji}_{12} & \cdots & a^{ji}_{1k} \\ a^{ji}_{21} & a^{ji}_{22} & \cdots & a^{ji}_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a^{ji}_{s1} & a^{ji}_{s2} & \cdots & a^{ji}_{sk} \end{bmatrix} && (20)\\ &= (\max\{p_{j1} \cdot a^{ji}_{11}, \cdots, p_{js} \cdot a^{ji}_{s1}\}, \cdots, && (21)\\ &\quad \max\{p_{j1} \cdot a^{ji}_{1k}, \cdots, p_{js} \cdot a^{ji}_{sk}\}) \\ &= (P(x_{i1}, q_j), \cdots, P(x_{ik}, q_j)), && (22) \end{aligned}$$

where $P(x_{ir}, q_j) = \max\{p_{j1} \cdot a^{ji}_{1r}, \cdots, p_{js} \cdot a^{ji}_{sr}\}$. $P(x_{ir}, Q)$,which is defined as probability of attribute x_{ir} influenced by given set input data Q, is $\vee$ operation for all probabilities of relation between x_{ir} and all members of Q. $\vee$ operation will be explained latter. If there are n members of Q, $\{(q_1, ...q_n)\}$, then,

$$P(x_{ir}, Q) = \bigvee_{1 \le j \le n} P(x_{ir}, q_j). \tag{23}$$

$P(X_i, Q)$ is defined as probability of compound attribute X_i influenced by given set input data Q. If there are n members of Q and k attributes of X_i, then

$$\begin{aligned} P(X_i, Q) &= (P(x_{i1}, Q), ..., P(x_{ik}, Q)), && (24)\\ P(X_i, Q) &= (\bigvee_{1 \le j \le n} P(x_{i1}, q_j), \ldots, \bigvee_{1 \le j \le n} P(x_{ik}, q_j)). && (25) \end{aligned}$$

4.1 Calculating minimum probability truth of $P(x_{ir}, Q)$

Now, we generate formula for calculating minimum probability of attribute x_{ir} given $Q = \{q_1, ...q_n\}$, as input data. Related to (22), we defined minimum probability truth of $P(x_{ir}, Q)$ as follows.

$$P_{min}(x_{ir}, Q) = \bigvee_{1 \le j \le n}^{min} P(x_{ir}, q_j). \tag{26}$$

To simplify the problem, let's say that system just has three compound attributes, X_1, X_2, and X_3 and their relation shown in Figure 2.2. We calculate minimum probability truth of $x_{3r} \in X_3$ based on input $Q = \{q_1, q_2, q_3\}$.

$$P(x_{3r}, Q)_{min} = P(x_{3r}, q_1) \vee_{min} P(x_{3r}, q_2) \vee_{min} P(x_{3r}, q_3).$$

We separate formula above into two parts. The first, we call *direct predicted probability of* x_{3r} which is $P(x_{3r}, q_3) = P(x_{3r}|q_3) = p_{3r}$ and the second, we call *indirect predicted probability truth of* x_{3r} which is predicted from other attributes value, $P(x_{3r}, q_1) \vee_{min} P(x_{3r}, q_2)$. The next, we compare both of them by applying *max* function as follows.

$$P(x_{3r}, Q)_{min} = max\{P(x_{3r}, q_1) \vee_{min} P(x_{3r}, q_2), p_{3r}\}. \tag{27}$$

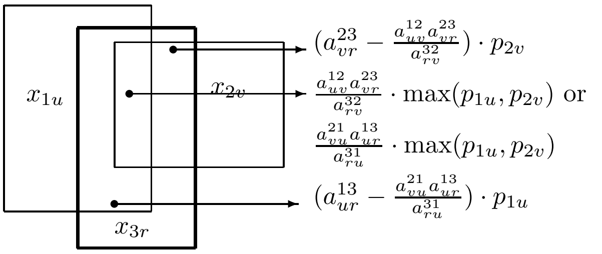
The problem now, is how to calculate $P(x_{3r}, q_1) \vee_{min} P(x_{3r}, q_2) = \delta_{min}$.i.e. X_1 has s attributes, X_2 has t attributes. Let's say that,

$$P(x_{3r}, q_1) = \max\{p_{11} \cdot a^{13}_{1r}, \cdots, p_{1s} \cdot a^{13}_{sr}\} = p_{1u} \cdot a^{13}_{ur},$$

$$P(x_{3r}, q_2) = \max\{p_{21} \cdot a^{23}_{1r}, \cdots, p_{2t} \cdot a^{23}_{tr}\} = p_{2v} \cdot a^{23}_{vr}.$$

We solve this problem by imaging interrelationship among x_{1u}, x_{2v} and x_{3r} as shown in Fig. 2.2, in the following three conditions.

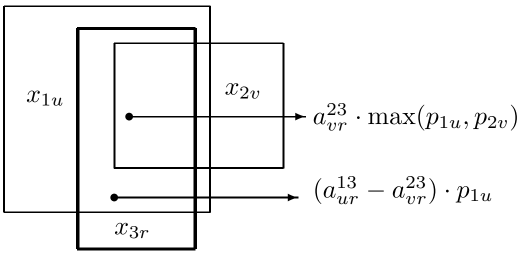
1. If $|(x_{1u} \cap x_{2v})| \leq |(x_{1u} \cap x_{3r})|$ and $|(x_{1u} \cap x_{2v})| \leq |(x_{2v} \cap x_{3r})|$, then $(x_{1u} \cap x_{2v})$ will be put in x_{3r}.



$$\delta_{min} = (a^{23}_{vr} - \frac{a^{12}_{uv}a^{23}_{vr}}{a^{32}_{rv}}) \cdot p_{2v} + (a^{13}_{ur} - \frac{a^{21}_{vu}a^{13}_{ur}}{a^{31}_{ru}}) \cdot p_{1u} + \frac{a^{12}_{uv}a^{23}_{vr}}{a^{32}_{rv}} \cdot \max(p_{1u}, p_{2v}).\ or$$

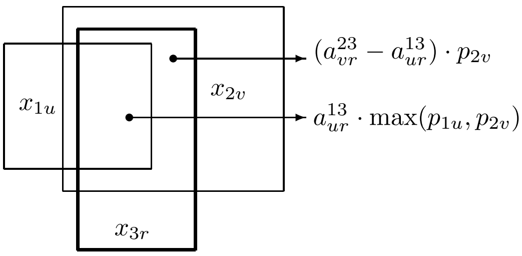
$$\delta_{min} = (a^{23}_{vr} - \frac{a^{12}_{uv}a^{23}_{vr}}{a^{32}_{rv}}) \cdot p_{2v} + (a^{13}_{ur} - \frac{a^{21}_{vu}a^{13}_{ur}}{a^{31}_{ru}}) \cdot p_{1u} + \frac{a^{21}_{vu}a^{13}_{ur}}{a^{31}_{ru}} \cdot \max(p_{1u}, p_{2v}).$$

2. If $|(x_{3r} \cap x_{2v})| \leq |(x_{1u} \cap x_{3r})|$ and $|(x_{3r} \cap x_{2v})| \leq |(x_{2v} \cap x_{1u})|$, then $(x_{3r} \cap x_{2v})$ will be put in $(x_{1u} \cap x_{3r})$.



$$\delta_{min} = (a^{13}_{ur} - a^{23}_{vr}) \cdot p_{1u} + a^{23}_{vr} \cdot \max(p_{1u}, p_{2v}).$$

3. If $|(x_{1u} \cap x_{3r})| \leq |(x_{1u} \cap x_{2v})|$ and $|(x_{1u} \cap x_{3r})| \leq |(x_{2v} \cap x_{3r})|$, then $(x_{1u} \cap x_{3r})$ will be put in $(x_{2v} \cap x_{3r})$.



$$\delta_{min} = (a^{23}_{vr} - a^{13}_{ur}) \cdot p_{2v} + a^{13}_{ur} \cdot \max(p_{1u}, p_{2v}).$$

From the above conditions, we generate a formula that satisfy all conditions as follows.

$$\delta_{min} = (a^{23}_{vr} - \min(\frac{a^{12}_{uv}a^{23}_{vr}}{a^{32}_{rv}}, a^{23}_{vr}, a^{13}_{ur})) \cdot p_{2v} + (a^{13}_{ur} - \min(\frac{a^{12}_{uv}a^{23}_{vr}}{a^{32}_{rv}}, a^{23}_{vr}, a^{13}_{ur})) \cdot p_{1u} + \min(\frac{a^{12}_{uv}a^{23}_{vr}}{a^{32}_{rv}}, a^{23}_{vr}, a^{13}_{ur}) \cdot \max(p_{1u}, p_{2v}). \quad (28)$$

Finally, we find that $P_{min}(x_{3r}, Q) = \max\{\delta_{min}, P(x_{3r} \mid q_3)\}$.

4.2 Calculating maximum probability truth of $P(x_{ir}, Q)$

Next, we generate formula for calculating maximum probability of attribute x_{ir} given $Q = (q_1, ...q_n)$, as input data. Related to (22), we defined maximum probability truth of $P(x_{ir}, Q)$ as follows.

$$P_{max}(x_{ir}, Q) = \bigvee_{1 \leq j \leq n}^{max} P(x_{ir}, q_j). \quad (29)$$

To simplify the problem, let's say that system just has three compound attributes, X_1, X_2, and X_3 and their relation shown in Figure 2.2. We calculate maximum probability truth of $x_{3r} \in X_3$ based on input $Q = \{q_1, q_2, q_3\}$.

$$P(x_{3r}, Q)_{max} = P(x_{3r}, q_1) \vee_{max} P(x_{3r}, q_2) \vee_{max} P(x_{3r}, q_3).$$

We separate formula above into two parts. The first, we call *direct predicted probability of* x_{3r} which is $P(x_{3r}, q_3) = P(x_{3r}|q_3) = p_{3r}$ and the second, we call *indirect predicted probability truth of* x_{3r} which is predicted from other attributes value,$P(x_{3r}, q_1) \vee_{max} P(x_{3r}, q_2)$. The next, we compare both of them by applying *min* function as follows.

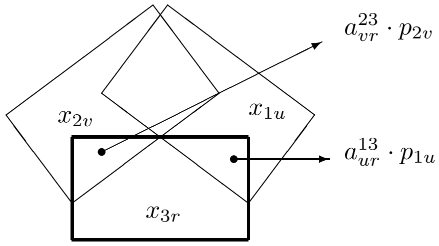
$$P(x_{3r}, Q)_{max} = \min\{1, (P(x_{3r}, q_1) \vee_{max} P(x_{3r}, q_2)) + p_{3r})\}. \quad (30)$$

The problem now, is how to calculate $P(x_{3r}, q_1) \vee_{max} P(x_{3r}, q_2) = \delta_{max}$.i.e. X_1 has s attributes, X_2 has t attributes. Let's say that,

$$P(x_{3r}, q_1) = \max\{p_{11} \cdot a^{13}_{1r}, \cdots, p_{1s} \cdot a^{13}_{sr}\} = p_{1u}.a^{13}_{ur}$$
$$P(x_{3r}, q_2) = \max\{p_{21} \cdot a^{23}_{1r}, \cdots, p_{2t} \cdot a^{23}_{tr}\} = p_{2v}.a^{23}_{vr}$$

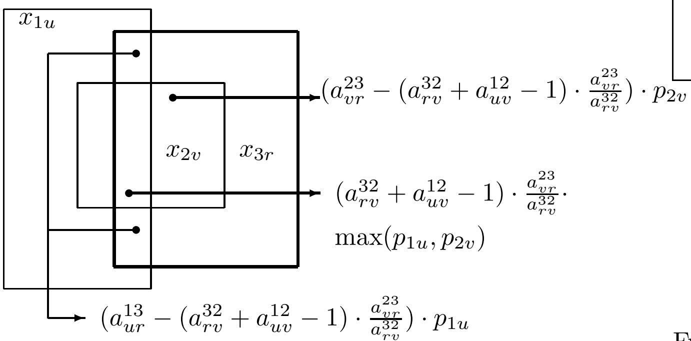
We solve this problem by imaging interrelationship among x_{1u}, x_{2v} and x_{3r} as shown in Fig. 2.2, in the following four conditions.

1. If $(a^{32}_{rv} + a^{12}_{uv} \leq 1)$ and $(a^{31}_{ru} + a^{21}_{vu} \leq 1)$ and $(a^{13}_{ur} + a^{23}_{vr} \leq 1)$, then



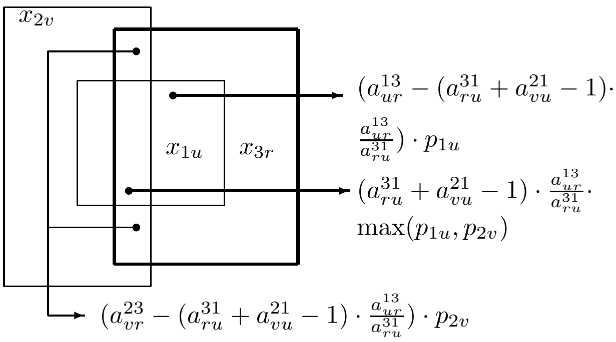
$$\delta_{max} = a^{23}_{vr} \cdot p_{2v} + a^{13}_{ur} \cdot p_{1u}.$$

2. If $(a^{32}_{rv} + a^{12}_{uv} > 1)$ and $((a^{32}_{rv} + a^{12}_{uv} - 1) \cdot \frac{a^{23}_{vr}}{a^{32}_{rv}}$ ¿ $(a^{31}_{ru} + a^{21}_{vu} - 1) \cdot \frac{a^{13}_{ur}}{a^{31}_{ru}})$ and $((a^{32}_{rv} + a^{12}_{uv} - 1) \cdot \frac{a^{23}_{vr}}{a^{32}_{rv}} > (a^{13}_{ur} + a^{23}_{vr} - 1))$, then



$\delta_{max} = (a^{23}_{vr} - (a^{32}_{rv} + a^{12}_{uv} - 1) \cdot \frac{a^{23}_{vr}}{a^{32}_{rv}}) \cdot p_{2v} +$
$(a^{13}_{ur} - (a^{32}_{rv} + a^{12}_{uv} - 1) \cdot \frac{a^{23}_{vr}}{a^{32}_{rv}}) \cdot p_{1u} +$
$(a^{32}_{rv} + a^{12}_{uv} - 1) \cdot \frac{a^{23}_{vr}}{a^{32}_{rv}} \cdot \max(p_{1u}, p_{2v})$.

3. If $(a^{31}_{ru} + a^{21}_{vu} > 1)$ and $((a^{31}_{ru} + a^{21}_{vu} - 1) \cdot \frac{a^{13}_{ur}}{a^{31}_{ru}} > (a^{32}_{rv} + a^{12}_{uv} - 1) \cdot \frac{a^{23}_{vr}}{a^{32}_{rv}})$ and $((a^{31}_{ru} + a^{21}_{vu} - 1) \cdot \frac{a^{13}_{ur}}{a^{31}_{ru}} > (a^{13}_{ur} + a^{23}_{vr} - 1))$, then

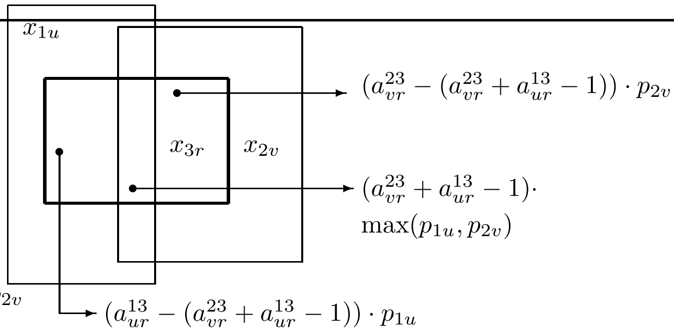


$$\delta_{max} = (a^{13}_{ur} - (a^{31}_{ru} + a^{21}_{vu} - 1) \cdot \frac{a^{13}_{ur}}{a^{31}_{ru}}) \cdot p_{1u} +$$

$$(a^{23}_{vr} - (a^{31}_{ru} + a^{21}_{vu} - 1) \cdot \frac{a^{13}_{ur}}{a^{31}_{ru}}) \cdot p_{2v} +$$

$$(a^{31}_{ru} + a^{21}_{vu} - 1) \cdot \frac{a^{13}_{ur}}{a^{31}_{ru}} \cdot \max(p_{1u}, p_{2v}).$$

4. If $(a^{23}_{vr} + a^{13}_{ur} > 1)$ and $((a^{23}_{vr} + a^{13}_{ur} - 1) > (a^{31}_{ru} + a^{21}_{vu} - 1) \cdot \frac{a^{13}_{ur}}{a^{31}_{ru}})$ and $((a^{13}_{ur} + a^{23}_{vr} - 1) > (a^{32}_{rv} + a^{12}_{uv} - 1) \cdot \frac{a^{23}_{vr}}{a^{32}_{rv}})$, then



$$\delta_{max} = (a^{23}_{vr} - (a^{23}_{vr} + a^{13}_{ur} - 1)) \cdot p_{2v} +$$
$$(a^{13}_{ur} - (a^{23}_{vr} + a^{13}_{ur} - 1)) \cdot p_{1u} +$$
$$(a^{23}_{vr} + a^{13}_{ur} - 1) \cdot \max(p_{1u}, p_{2v}).$$

From the above conditions, we generate a formula that satisfy all condition as follows.

$$\delta_{max} = (a^{23}_{vr} - \max(0, (a^{32}_{rv} + a^{12}_{uv} - 1) \cdot \frac{a^{23}_{vr}}{a^{32}_{rv}}, (a^{31}_{ru} + a^{21}_{vu} - 1) \cdot \frac{a^{13}_{ur}}{a^{31}_{ru}}, (a^{23}_{vr} + a^{13}_{ur} - 1))) \cdot p_{2v} + (a^{13}_{ur} - \max(0, (a^{32}_{rv} + a^{12}_{uv} - 1) \cdot \frac{a^{23}_{vr}}{a^{32}_{rv}}, (a^{31}_{ru} + a^{21}_{vu} - 1) \cdot \frac{a^{13}_{ur}}{a^{31}_{ru}}, (a^{23}_{vr} + a^{13}_{ur} - 1))) \cdot p_{1u} + \max(0, (a^{32}_{rv} + a^{12}_{uv} - 1) \cdot \frac{a^{23}_{vr}}{a^{32}_{rv}}, (a^{31}_{ru} + a^{21}_{vu} - 1) \cdot \frac{a^{13}_{ur}}{a^{31}_{ru}}, (a^{23}_{vr} + a^{13}_{ur} - 1)) \cdot \max(p_{1u}, p_{2v}). \quad (31)$$

Finally, we find that $P_{max}(x_{3r}, Q) = \max\{1, \delta_{max} + P(x_{3r} \mid q_3)\}$

5 Conclusions

This paper proposed a method based on conditional probability relation to approximately calculate interval probability of dependency of data for data querying. Theoretically the formulation is quite interesting. However, it seems to be too complicated to calculate interaction of three or more components. Practically the formulas should be simplified, even though the accuracy of prediction may be decreased.

References

[1] Intan, R., Mukaidono, M., ‘Application of Conditional Probability in Constructing Fuzzy Functional Dependency(FFD)’, *Proceedings of AFSS’00*, (2000), pp.271-276.

[2] Intan, R., Mukaidono, M., ‘A proposal of Fuzzy Functional Dependency based on Conditional Probability’, *Proceeding of FSS’00 (Fuzzy Systems Symposium)*, (2000), pp. 199-202.

[3] Intan, R., Mukaidono, M., 'Fuzzy Functional Dependency and Its Application to Approximate Querying', *Proceedings of IDEAS'00*, (2000), pp.47-54.

[4] Intan, R., Mukaidono, M., 'Conditional Probability Relations in Fuzzy Relational Database ', *Proceedings of RSCTC'00, LNAI 2005, Springer & Verlag*, (2000), pp.251-260.

[5] Intan, R., 'Predicting Interval Probability in Data Querying', *Proceedings of IMECS 2007*, (2007), pp.690-695.

[6] Baldwin J.F., 'Knowledge from Data using Fril and Fuzzy Methods',*Fuzzy Logic*,John Wiley & Sons Ltd, pp. 33-75, 1996.

[7] Yukari Yamauchi, Masao Mukaidono , 'Interval and Paired Probabilities for Treating Uncertain Events',*The Institute of Electronics, Information and Communication Engineers*, Vol. E82-D, pp. 955–961, May 1999.

[8] Baldwin J.F., Martin T.P., and Pilsworth B.W. *FRIL-Fuzzy and Evidential Reasoning in AI*, Research Studies Press and Willey, 1995.

[9] Shafer G. *A Mathematical Theory of Evidence*, Princeton Univ. Press, 1976.

[10] Richard Jeffrey, 'Probabilistic Thinking', Priceton University, 1995.

[11] Baldwin J.F., Martin T.P., 'A Fuzzy Data Browser in Fril', *Fuzzy Logic*, John Wiley & Sons Ltd, pp. 101-123, 1996.