Fuzzy Probabilistic Approximation Spaces and Their Information Measures

Qinghua Hu, Daren Yu, Zongxia Xie, and Jinfu Liu

***Abstract*—Rough set theory has proven to be an efficient tool for modeling and reasoning with uncertainty information. By introducing probability into fuzzy approximation space, a theory about fuzzy probabilistic approximation spaces is proposed in this paper, which combines three types of uncertainty: probability, fuzziness, and roughness into a rough set model. We introduce Shannon's entropy to measure information quantity implied in a Pawlak's approximation space, and then present a novel representation of Shannon's entropy with a relation matrix. Based on the modified formulas, some generalizations of the entropy are proposed to calculate the information in a fuzzy approximation space and a fuzzy probabilistic approximation space, respectively. As a result, uniform representations of approximation spaces and their information measures are formed with this work.**

***Index Terms*—Approximation space, fuzzy set, information measure, probability distribution, rough set.**

I. Introduction

ROUGH set methodology has been witnessed great success in modeling with imprecise and incomplete information. The basic idea of this method hinges on classifying objects of discourse into classes containing indiscernible objects with respect to some attributes. Then the indiscernible classes, also called knowledge granules, are used to approximate the unseen object sets. In this framework, an attribute set is viewed as a family of knowledge, which partitions the universe into some knowledge granules or elemental concepts. Any attribute or attribute set P can induce a partition Π_P of the universe. We say that knowledge P is finer than knowledge Q if Π_P is a refinement of Π_Q. An arbitrary subset X of the universe U can be approximated by two sets $\langle \underline{P}X, \overline{P}X \rangle$, called the lower approximation and upper approximation, respectively. If X can be precisely approximated by some knowledge granules of the partition, the set is called a definable set, where $\underline{P}X = \overline{P}X$; otherwise we say X is a rough set. The approximating power of an information system depends on the knowledge P. The finer the knowledge P is, the more accurately X can be approximated. This process is much similar to reasoning of human's mind. In real life, the objects are drawn together by indistinguishability, similarity, proximity and named with a concept. Then a concept system is formed and used to approximately describe unseen objects. Partition, granulation and approximation are the methods widely used in human's reasoning [10], [39]. Rough set methodology presents a novel paradigm to deal with uncertainty and has been applied to feature selection [1], [2], knowledge reduction [3], [36], [38], rule extraction [4]–[6], uncertainty reasoning [7], [8] and granulation computing [9], [33], [34], [42].

In Pawlak's rough set model, fuzziness and probability are not taken into consideration. Pawlak's model just works in nominal data domain, for crisp equivalence relations and equivalence classes are the foundation of the model [8], [37]. However, there are usually real-valued data and fuzzy information in real-world applications. To deal with fuzziness, some generalizations of Pawlak's model were proposed; the theories on rough and fuzzy sets were put together. Rough-fuzzy sets and fuzzy-rough sets were introduced in [11], [12], [35] and analyzed in detail [13]–[16]. The generalized methods were applied to hybrid data reduction [41], mining stock price [17], vocabulary for information retrieval [18] and fuzzy decision rule extraction [19].

Both the theory on classical rough sets and its fuzzy generalizations implicitly take an assumption that the objects in the universe are equally probable. Namely, the objects are uniformly distributed and the probability of each object is $1/n$, where n is the number of objects. In fact, this assumption just holds if the information about the probability of the objects is totally ignored. Sometimes, there is a probability distribution on the object or event set [23], [40]. A theory on probabilistic approximation space or a probabilistic rough set model is expected in this case. For example, there is an information system about the disease flu, which is described with three attributes: headache, muscle pain, and temperature. The values of the attributes headache, muscle pain and flu are yes and no, and those of the attribute temperature are high and normal. There are $2^4 = 16$ cases in all. If there are not any samples about the disease, but a probability distribution of the 16 cases, then the theory about probabilistic approximation spaces is desirable for reasoning with uncertainty of roughness and randomness. Probability distribution of the universe lays a foundation to employ statistical techniques into rough set model, which maybe lead to a tool to deal with inconsistency or noise in data.

In the rough set framework, attributes are called knowledge, which is used to form a concept system of the universe. Knowledge introduced by an attribute set implies in the partition of a referential universe. The more knowledge there is, the finer partition will be, and correspondingly we can get a more perfect approximation of a subset in the universe. Attributes induce an order or a structure of universe of discourse, which decreases uncertainty or chaos of the universe. Given a universe U, a probability distribution on U, and some nominal, real-value or fuzzy attributes, there comes forth an interesting problem: How do we measure the knowledge quantity introduced by an attribute set in the approximation space. In other words, it's interesting in con-

Manuscript received May 16, 2004; revised March 16, 2005 and June 8, 2005.

The authors are with Harbin Institute of Technology, Harbin, Heilonghiang Province 150001, China (e-mail: huqinghua@hcms.hit.edu.cn; yudaren@hcms.hit.edu.cn; xiezongxia@hcms.hit.edu.cn; liujinfu@hcms.hit.edu.cn).

Digital Object Identifier 10.1109/TFUZZ.2005.864086

structing a measure to compute the discernibility power induced by a family of attributes. This measure leads to the likelihood to compare the knowledge quantity formed by different attributes, and help us find the important attribute set and redundancy of an information system. Hartley captured the intuitive idea that the more possible results for an experiment, the less it can be predicted. Shannon [20] defined a measure of a random variable within the frame of communication theory. Forte and Kampe [21], [22] gave the axiomatic information measure, where the word "information" was associated both to measures of events and measures of partitions and suggested that the uncertainty measure is associated to a family of partitions of a given referential space. Zadeh [23] introduced a new uncertainty measure for fuzzy probabilistic space. Yager introduced some measures to calculate uncertainty implied in similarity relation [24]. In [25] a measure, suitable to operate on fuzzy equivalence relation domains, was introduced. Uncertainty measure on fuzzy partitions was analyzed in documents [26], [27], [37]. In this paper, Shannon's entropy is first introduced to compute the knowledge quantity of nominal attributes in Pawlak's approximation space, and then an extended information measure will be presented, which is suitable for the spaces where fuzzy attributes or fuzzy relations are defined on. Based on the extension, the solutions to measuring the information in fuzzy and fuzzy probabilistic hybrid approximation spaces are presented.

The rest of the paper is organized as follows. Some definitions in classical approximation spaces are reviewed in Section II. We introduce fuzzy probabilistic approximation spaces in Section III. Shannon's entropy is applied to calculating the information quantity in a classical approximation space in Section IV. Then we redefine the formulae of Shannon's entropy with a matrix representation and extend it to the fuzzy cases. The information measures for fuzzy approximation spaces and fuzzy probabilistic approximation spaces are presented in Section V. Finally, the conclusions and discusses are given in Section VI.

II. Preliminaries

In this section, we will review some basic definitions in rough set theory.

Definition 1: $\langle U, A\rangle$ is called an approximation space, where $U = \{x_1, x_2, \ldots, x_n\}$ is the universe; A is a family of attributes, also called knowledge in the universe. V is the value domain of A and f is an information function $f : U \times A \rightarrow V$. An approximation space is also called an information system.

Any subset B of knowledge A defines an equivalence (also called indiscernibility) relation $IND(B)$ on U

$$IND(B) = \{(x, y) \in U \times U| \qquad \forall a \in B, f_a(x) = f_a(y)\}. \tag{1}$$

$IND(B)$ will generate a partition of U. We denote the partition induced by attributes B as

$$\Pi_B = \frac{U}{B} = \{[x_i]_B : x_i \in U\} \tag{2}$$

where $[x_i]_B$ is the equivalence class containing x_i, the elements in $[x_i]_B$ are indiscernible or equivalent with respect to knowledge B. Equivalent classes, also named as elemental concepts, information granules, are used to characterize arbitrary subsets of U.

Definition 2: An arbitrary subset X of U is characterized by two unions of elemental concepts$\langle \underline{B}X, \overline{B}X\rangle$, called lower and upper approximations, respectively

$$\begin{cases} \underline{B}X = \cup\{[x_i]_B | [x_i]_B \subseteq X\} \\ \overline{B}X = \cup\{[x_i]_B | [x_i]_B \cap X \neq \phi\} \end{cases}. \tag{3}$$

The lower approximation is the greatest union of $[x_i]_B$ contained in X and the upper approximation is the least union of $[x_i]_B$ containing X. The lower approximation is also called positive region sometimes, denoted as $POS_B(X)$.

We say Π_A is a refinement of Π_B if there is a partial order

$$\Pi_A \prec \Pi_B \Leftrightarrow \forall [x_i]_A \in \Pi_A, \exists [x_j]_B : [x_i]_A \subseteq [x_j]_B. \tag{4}$$

Theorem 1: $\forall X \subseteq U, \forall B \subseteq A : \underline{B}X \subseteq X \subseteq \overline{B}X$.

Theorem 2: $A \supseteq B \Rightarrow \Pi_A \prec \Pi_B, \overline{B}X \supseteq \overline{A}X \supseteq \underline{A}X \supseteq \underline{B}X$.

$\underline{B}X = \overline{B}X$, that is to say, X can be accurately characterized with knowledge B, and we say set X is definable, otherwise, X is indefinable and we say X is a rough set. $BN_B(X) = \overline{B}X - \underline{B}X$ is called boundary set. A set X is definable if it is a finite union of some elemental concepts, which let X precisely characterized with respect to knowledge B. Theorem 1 shows that the more knowledge we have, the finer partition we will get, accordingly, the more accurately subset X can be approximated and a less boundary we will get.

Definition 3: Given $\langle U, A\rangle$, $B \subseteq A$, $a \in B$, if $U/B = U/(B - a)$, we say knowledge a is *redundant* in B. Otherwise, we say knowledge a is *indispensable*. If each a in B is *indispensable*, we say B is *independent*. If a set $B \subseteq A$ is *independent* and $U/B = U/A$, we say B is a *reduct* of A.

A reduct of an information system has the same discernibility or representation power as that of the original system; however the reduct has a concise representation with respect to the original data.

There is often more than one reduct in an information system. The common elements of all reducts are called the *core* of the information system. The core is the attribute set which cannot be deleted from the system, or the discernibility of the system will decrease.

Definition 4: An information system $\langle U, A\rangle$ is called a decision table if the attribute set $A = C \cup D$, where C is the condition attribute set and D is the decision attribute set. We define the dependency $\gamma_C(D)$ between D and C as

$$k = \gamma_C(D) = \frac{|POS_C(D)|}{|U|} \tag{5}$$

where $|\bullet|$ denotes the cardinality of a set and $POS_C(D) = \cup \underline{C}X_i, X_i$ is ith equivalence class induced by D. Given $B \subseteq C$, we say $a \in B$ is *redundant* relative to D in B if $\gamma_{B-a}(D) = \gamma_B(D)$, otherwise a is *indispensable*. If $\forall a \in B$ is *indispensable* we say B is *independent* with respect to the decision D.

Dependency measures the capability of condition attributes B to characterize the decision D and can be used as a significance measure of condition attributes with respect to decision. $\gamma_B(D) = 1$ means that the decision can be approximated precisely by the knowledge granules induced with the attribute set B.

Definition 5: Given $\langle U, A\rangle$, we say B is the *D- relative reduct* if B satisfies

1) $\gamma_B(D) = \gamma_C(D)$;
2) B is independent relative to D.

The first term grantees the power of B to approximate D is the same as that of C; the second term means that there is no redundant attribute in C.

III. Fuzzy Probabilistic Approximation Spaces

Pawlak's approximation spaces work on the domain where crisp equivalence relations are defined. In this section, we will integrate three types of uncertainty: probability, fuzziness, and roughness together, and present the definition of fuzzy probabilistic approximation spaces.

Definition 6: Given a nonempty finite set U, $\widetilde{R}$ is a fuzzy binary relation over U, denoted by a matrix $M(\widetilde{R})$

$$M(\widetilde{R}) = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{pmatrix} \tag{6}$$

where $r_{ij} \in [0,1]$ is the relation value between x_i and x_j. We say $\widetilde{R}$ is a fuzzy equivalence relation if $\forall x, y, z \in U$, $\widetilde{R}$ satisfies

1) Reflexivity: $\widetilde{R}(x,x) = 1$;
2) Symmetry: $\widetilde{R}(x,y) = \widetilde{R}(y,x)$;
3) Transitivity: $\widetilde{R}(x,z) \geq \min_y\{\widetilde{R}(x,y), \widetilde{R}(y,z)\}$.

Some operations of relation matrices are defined as

1) $\widetilde{R}_1 = \widetilde{R}_2 \Leftrightarrow \widetilde{R}_1(x,y) = \widetilde{R}_2(x,y)$;
2) $\widetilde{R} = \widetilde{R}_1 \cup \widetilde{R}_2 \Leftrightarrow \widetilde{R}(x,y) = \max\{\widetilde{R}_1(x,y), \widetilde{R}_2(x,y)\}$;
3) $\widetilde{R} = \widetilde{R}_1 \cap \widetilde{R}_2 \Leftrightarrow \widetilde{R}(x,y) = \min\{\widetilde{R}_1(x,y), \widetilde{R}_2(x,y)\}$;
4) $\widetilde{R}_1 \subseteq \widetilde{R}_2 \Leftrightarrow \widetilde{R}_1(x,y) \leq \widetilde{R}_2(x,y)$.

A crisp equivalence relation induces a crisp partition of the universe and generates a family of crisp equivalence classes. Correspondingly, a fuzzy equivalence relation generates a fuzzy partition of the universe and a series of fuzzy equivalence classes, which are also called fuzzy knowledge granules [10], [39], [43].

Definition 7: The fuzzy partition of the universe generated by a fuzzy equivalence relation $\widetilde{R}$ is defined as

$$\frac{U}{\widetilde{R}} = \left\{[x_i]_{\widetilde{R}}\right\}_{i=1}^{n} \tag{7}$$

where $[x_i]_{\widetilde{R}} = \{(r_{i1}/x_1) + (r_{i2}/x_2) + \cdots + (r_{in}/x_n)\}$. $[x_i]_{\widetilde{R}}$ is the fuzzy equivalence class containing x_i. r_{ij} is the degree of x_i equivalent to x_j. Here, "+" means union of elements.

In this case, $[x_i]_{\widetilde{R}}$ is a fuzzy set and the family of $[x]_{\widetilde{R}}$ forms a fuzzy concept system of the universe. This system will be used to approximate the object subset of the universe.

Example 1: Assume $X = \{x_1, x_2, x_3\}$ is an object set, $\widetilde{R}_1$ is a fuzzy equivalence relation on X:

$$\widetilde{R}_1 = \begin{pmatrix} 1 & 0.9 & 0 \\ 0.9 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Then, the equivalence classes are

$$\begin{aligned} [x_1]_{\widetilde{R}_1} &= \left\{\frac{1}{x_1} + \frac{0.9}{x_2} + \frac{0}{x_3}\right\} \\ [x_2]_{\widetilde{R}_1} &= \left\{\frac{0.9}{x_1} + \frac{1}{x_2} + \frac{0}{x_3}\right\} \\ [x_3]_{\widetilde{R}_1} &= \left\{\frac{0}{x_1} + \frac{0}{x_2} + \frac{1}{x_3}\right\}. \end{aligned}$$

Theorem 3: Given an set U, $\widetilde{R}$ is a fuzzy equivalence relation on U. $\forall x, y \in U$, we have

1) $\widetilde{R}(x,y) = 0 \Leftrightarrow [x]_{\widetilde{R}} \cap [y]_{\widetilde{R}} = \phi$;
2) $[x]_{\widetilde{R}} = [y]_{\widetilde{R}} \Rightarrow \widetilde{R}(x,y) = 1$.

Theorem 4: Given an set U, $\widetilde{R}_1$ and $\widetilde{R}_2$ are two fuzzy equivalence relations on U, we have

$$\widetilde{R}_1 \subseteq \widetilde{R}_2 \Rightarrow \frac{U}{\widetilde{R}_1} \prec \frac{U}{\widetilde{R}_2}.$$

Definition 8: A three-tuple $\langle U, P, \widetilde{R}\rangle$ is a fuzzy probabilistic approximation space or a fuzzy probabilistic information system, where U is a nonempty and finite set of objects, called the universe, P is a probability distribution over U. $\widetilde{R}$ is a family of fuzzy equivalence relations defined on U.

Definition 9: Give a fuzzy probabilistic approximation space $\langle U, P, \widetilde{R}\rangle$. $\widetilde{X}$ is a fuzzy subset of U. The *lower approximation* and *upper approximation* is denoted by $\underline{\widetilde{R}}X$ and $\overline{\widetilde{R}}X$; then membership of x to X are defined as

$$\begin{cases} \mu_{\underline{\widetilde{R}}\widetilde{X}}(x) = \wedge\left\{\mu_{\widetilde{X}}(y) \vee \left(1 - \widetilde{R}(x,y)\right) : y \in U\right\}, x \in U \\ \mu_{\overline{\widetilde{R}}\widetilde{X}}(x) = \vee\left\{(\mu_{\widetilde{X}}(y) \wedge \widetilde{R}(x,y) : y \in U\right\}, x \in U \end{cases}. \tag{8}$$

where $\wedge$ and $\vee$ mean min and max operators, respectively, and $\mu_{\widetilde{X}}(y)$ means the membership of y to $\widetilde{X}$, seeing [28]. These definitions are the rational extension of some models. Let us derive the other model from these definitions.

Case 1: X is a crisp subset and R is a crisp equivalence relation on U

$$\begin{aligned} \mu_{\underline{\widetilde{R}}X}(x) = 1 &\Leftrightarrow \forall y \in U, \mu_X(y) \vee \left(1 - \widetilde{R}(x,y)\right) = 1 \\ &\Leftrightarrow \forall y \in U : y \notin X \rightarrow (x,y) \notin \widetilde{R} \\ &\Leftrightarrow \forall y \notin X \rightarrow y \notin [x]_{\widetilde{R}} \\ &\Leftrightarrow [x]_{\widetilde{R}} \subseteq X \\ \mu_{\overline{\widetilde{R}}X}(x) = 1 &\Leftrightarrow \exists y \in U : \mu_X(y) = 1, \widetilde{R}(x,y) = 1 \\ &\Leftrightarrow X \cap [x]_{\widetilde{R}} \neq \phi \end{aligned}.$$

These definitions are consistent with Pawlak's rough set model in this case.

Case 2: X is a fuzzy subset of U and R is a crisp equivalence relation on U

$$\begin{aligned}\mu_{\underline{R}X}(x) &= \wedge\left\{\mu_X(y) \vee (1 - R(x,y)) : y \in U\right\}\\ &= \wedge\left\{\mu_X(y) : R(x,y) = 1\right\}\\ &= \wedge\left\{\mu_X(y) : y \in [x]_R\right\}\\ \mu_{\overline{R}X}(x) &= \vee\left\{\mu_X(y) \wedge R(x,y) : y \in U\right\}\\ &= \vee\left\{\mu_X(y) : R(x,y) = 1\right\}\\ &= \vee\left\{\mu_X(y) : y \in [x]_R\right\}.\end{aligned}$$

Here, the rough sets are called rough fuzzy sets.

Case 3: X is a subset of U and $\widetilde{R}$ is a fuzzy equivalence relation on U:

$$\begin{aligned}\mu_{\underline{\widetilde{R}}X}(x) &= \min\left\{\mu_X(y) \vee \left(1 - \widetilde{R}(x,y)\right) : y \in U\right\}\\ &= \min_{y \notin X}\left\{1 - \widetilde{R}(x,y)\right\}\\ \mu_{\overline{\widetilde{R}}X}(x) &= \max\left\{\mu_X(y) \wedge \left(1 - \widetilde{R}(x,y)\right) : y \in U\right\}\\ &= \max_{y \in X} \widetilde{R}(x,y).\end{aligned}$$

From the previous analysis, we can conclude that the definitions of lower and upper approximations of fuzzy sets in fuzzy information systems are rational generalizations of the classical model.

The membership of an object $x \in U$, belonging to the fuzzy positive region is

$$\mu_{POS_{\widetilde{B}}(d)}(x) = \sup_{X \subseteq \frac{U}{d}} \mu_{\underline{\widetilde{B}}X}(x). \tag{9}$$

Definition 10: Given a fuzzy probabilistic information system $\langle U, P, \widetilde{A}\rangle$, B and D are two subsets of attribute set $\widetilde{A}$, the dependency degree of D to B is defined as

$$\gamma_B(D) = \sum_{x \in U} p(x)\mu_{POS_B(D)}(x). \tag{10}$$

The difference between fuzzy approximation spaces and fuzzy probabilistic approximation spaces is introducing probability distribution over U. This leads to a more general generalization of Pawlak's approximation space. The classic approximation space takes the uniform-distribution assumption. So $p(x_i) = 1/n$, $i = 1, 2, \ldots, n$. Then

$$\begin{aligned}\gamma_B(D) &= \sum_{x \in U} p(x)\mu_{POS_B(D)}(x)\\ &= \frac{1}{n}\sum_{x \in U} \mu_{POS_B(D)}(x)\\ &= \frac{\sum\limits_{x \in U} \mu_{POS_B(D)}(x)}{|U|}.\end{aligned}$$

This formula is the same as that in fuzzy approximation space [28], which shows that the fuzzy probabilistic approximation space will degrade to a fuzzy approximation space when the equality–probability assumption is satisfied.

Definition 11: Given $\langle U, P, \widetilde{A}\rangle$, $\widetilde{B} \subseteq \widetilde{A}$, $a \in \widetilde{B}$, if U/B *and* $U/(B - a)$ are two fuzzy partitions, we say knowledge a is *redundant* or *superfluous* in B if $U/B = U/(B-a)$. Otherwise, we say knowledge a is *indispensable*. If any a belonging to B is *indispensable*, we say B is *independent*. If attribute subset $\widetilde{B} \subsetneq \widetilde{A}$ is *independent* and $U/\widetilde{B} = U/\widetilde{A}$, we say $\widetilde{B}$ is a *reduct* of $\widetilde{A}$.

Definition 12: Given $\langle U, P, \widetilde{A}\rangle$, $\widetilde{A} = \widetilde{C} \cup \widetilde{d}$. $\widetilde{B}$ is a subset of $\widetilde{C}$. $\forall a \in \widetilde{B}$, a is redundant in $\widetilde{B}$ relative to d if $\gamma_{\widetilde{B-a}}(\widetilde{d}) = \gamma_{\widetilde{B}}(\widetilde{d})$, otherwise a is indispensable. $\widetilde{B}$ is independent if $\forall a \in \widetilde{B}$ is indispensable, otherwise $\widetilde{B}$ is dependent. $\widetilde{B} \subseteq \widetilde{C}$ is a reduct if $\widetilde{B}$ satisfies

1) $\gamma_{\widetilde{\mathrm{B}}}(\widetilde{d}) = \gamma_{\widetilde{\mathrm{C}}}(\widetilde{d})$;
2) $\forall a \in \widetilde{\mathrm{B}} : \gamma_{\widetilde{\mathrm{B}-a}}(\widetilde{d}) < \gamma_{\widetilde{\mathrm{B}}}(\widetilde{d})$.

Comparing the fuzzy probabilistic approximation space with fuzzy approximation space, we can find that the foundational difference is in computing the cardinality of fuzzy set, such as fuzzy equivalence classes, fuzzy lower approximations and fuzzy upper approximations. Accordingly it leads to difference in defining the function of dependency. Finding dependency of data is a foundational problem in machine learning and data mining. The difference in dependency will lead to great changes in reasoning with uncertainty. In classical fuzzy approximation space, we assume the objects are uniformly distributed and $p(x_i) = 1/|U|$. In the fuzzy probabilistic approximation space the probability of x_i is $p(x_i)$. When the probability $p(x_i) = 1/|U|$, the fuzzy probabilistic approximation space degrades to a fuzzy approximation space, and if the equivalence relation and the object subset to be approximated are both crisp, we get a Pawlak's approximation space.

IV. Shannon's Entropies on Pawlak's Approximation Space

Knowledge is thought as the discernibility power of the attributes in the framework of rough set methodology. An attribute set forms an equivalence relation; correspondingly generates a partition of the universe and a family of concepts. The quantity of knowledge measures the fineness degree of the partition. The finer the partition is, the more knowledge about the universe we have, and accordingly a finer approximation we will have. In this section, we will introduce Shannon's information measure to compute the knowledge quantity of a crisp attribute set or a crisp partition of U.

Given a universe U and two attribute sets A, B, we take the partitions Π_A and Π_B as two random variables in δ-algebra:

$$\begin{aligned}\Pi_A &= \{X_1, X_2, \cdots, X_n\};\\ \Pi_B &= \{Y_1, Y_2, \cdots, Y_m\}.\end{aligned}$$

The probability distributions of Π_A and Π_B are defined as

$$(X; P) = \begin{pmatrix} X_1 & X_2 & \cdots & X_n\\ p(X_1) & p(X_2) & \cdots & p(X_n)\end{pmatrix} \tag{11}$$

and

$$(Y;P) = \begin{pmatrix} Y_1 & Y_2 & \cdots & Y_m \\ p(Y_1) & p(Y_2) & \cdots & p(Y_m) \end{pmatrix} \tag{12}$$

where $p(X_i) = |X_i|/|U|$; $p(Y_j) = |Y_j|/|U|$. Correspondingly, the joint probability of X and Y is

$$(X \otimes Y;P) = \begin{pmatrix} X_1 \cap Y_1 & \cdots & X_i \cap Y_j & \cdots & X_n \cap Y_m \\ p(X_1 Y_1) & \cdots & p(X_i Y_j) & \cdots & p(X_n Y_m) \end{pmatrix} \tag{13}$$

where $p(X_i Y_j) = |X_i \cap Y_j|/|U|$.

Definition 13: Information quantity of attributes A is defined as

$$H(A) = -\sum_{i=1}^{n} p(X_i) \log p(X_i). \tag{14}$$

Definition 14: The joint entropy of A and B is defined as

$$H(AB) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(X_i Y_j) \log p(X_i Y_j). \tag{15}$$

Definition 15: The conditional entropy A to B $H(X|Y)$ is defined as

$$\begin{aligned} H(A|B) &= -\sum_{i=1}^{n}\sum_{j=1}^{m} p(X_i Y_j) \log p(X_i|Y_j) \\ &= -\sum_{j=1}^{m} p(Y_j) \sum_{i=1}^{n} p(X_i|Y_j) \log p(X_i|Y_j) \end{aligned} \tag{16}$$

where $p(X_i|Y_j) = |X_i \cap Y_j|/|Y_j|$.

Theorem 5: $H(A|B) = H(AB) - H(B)$.

Theorem 6: Given a universe U, A and B are two attribute sets on U, if $A \supseteq B$ then

1) $H(AB) = H(A)$;
2) $H(B|A) = 0$;
3) $H(A) \geq H(B)$, where AB means $A \cup B$.

Proof: The first two terms are straightforward. Here we just give the proof of the third term.

Taking that the probability distributions about knowledge A and B are

$$(X;P) = \begin{pmatrix} X_1 & X_2 & \cdots & X_n \\ p(X_1) & p(X_2) & \cdots & p(X_n) \end{pmatrix}$$

and

$$(Y;P) = \begin{pmatrix} Y_1 & Y_2 & \cdots & Y_m \\ p(Y_1) & p(Y_2) & \cdots & p(Y_m) \end{pmatrix}.$$

Without loss of generality, we assume that $Y_1 = X_1, \ldots, Y_{m-1} = X_{n-2}, Y_m = X_{n-1} \cup X_n$

$$H(A) = -\sum_{i=1}^{n-2} p(X_i) \log p(X_i) - \sum_{i=n-1}^{n} p(X_i) \log p(X_i)$$
$$H(B) = -\sum_{j=1}^{m-1} p(Y_j) \log p(Y_j) - p(Y_m) \log p(Y_m).$$

Here we have $p(Y_m) = p(X_{n-1}) + p(X_n)$

$$\begin{aligned} &H(A) - H(B) \\ &\quad = p(Y_m) \log p(Y_m) - \sum_{i=n-1}^{n} p(X_i) \log p(X_i) \\ &\quad = p(X_{n-1}) \log \frac{p(Y_m)}{p(X_{n-1})} + p(X_n) \log \frac{p(Y_m)}{p(X_n)} \end{aligned}$$

$(p(Y_m)/p(X_{n-1})) \geq 1$, $(p(Y_m)/p(X_n)) \geq 1$, so $H(A) - H(B) \geq 0$.

Theorem 7: Given $\langle U, A \rangle$, if $a \in A$ is *redundant*, then $H(A|A-a) = 0$, otherwise $H(A|A-a) > 0$.

Proof: The probability distributions of Π_A and Π_{A-a} are

$$(X;P) = \begin{pmatrix} X_1 & X_2 & \cdots & X_n \\ p(X_1) & p(X_2) & \cdots & p(X_n) \end{pmatrix}$$

and

$$(Y;P) = \begin{pmatrix} Y_1 & Y_2 & \cdots & Y_m \\ p(Y_1) & p(Y_2) & \cdots & p(Y_m) \end{pmatrix}.$$

Attribute a is *redundant*, U/A is the same as $U/(A-a)$. So

$$p(X_i|Y_j) = \frac{|X_i \cap Y_j|}{|Y_j|} = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}.$$

Then $H(X|Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(X_i Y_j) \log p(X_i|Y_j) = 0$.

Theorem 8: Given an approximation space $\langle U, A \rangle$, $B \subseteq A$ is a reduct if B satisfies

1) $H(A|B) = 0$;
2) $\forall a \in B : H(B|B-a) > 0$.

Theorem 9: Given a decision table $\langle U, A \rangle$, $A = C \cup d$, if $E \subseteq B \subseteq C$, then $H(d|E) \geq H(d|B)$.

Theorem 10: Given a decision table $\langle U, A \rangle$, $A = C \cup d$, where C is the condition attribute set and d is the decision, $B \subseteq C$, $a \in B$. a is redundant if $H(d|B-a) = H(d|B)$. B is independent if $\forall a \in B : H(d|B-a) > H(d|B)$. B is a reduct of the decision table if B satisfies

1) $H(d|B) = H(d|C)$;
2) $\forall a \in B : H(d|B-a) > H(d|B)$.

Example 2: Consider the decision Table I, where

$$U = \{x_1, x_2, \ldots, x_8\} \quad A = C \cup D = \{C1, C2, C3, C4, D\}.$$

TABLE I
HIRING DATA

U	C1	C2	C3	C4	D
x_1	MBA	Medium	Yes	Excellent	Accept
x_2	MBA	Low	Yes	Neutral	Reject
x_3	MCE	Low	Yes	Good	Reject
x_4	MSc	High	Yes	Neutral	Accept
x_5	MSc	Medium	Yes	Neutral	Reject
x_6	MSc	High	Yes	Excellent	Accept
x_7	MBA	High	No	Good	Accept
x_8	MCE	Low	No	Excellent	Reject

The partitions of U by $C1$, $C2$, ..., and D are

$$\Pi_{C1} = \frac{U}{C1} = \{\{x_1, x_2, x_7\}, \{x_3, x_8\}, \{x_4, x_5, x_6\}\}$$
$$\Pi_{C2} = \frac{U}{C2} = \{\{x_1, x_5\}, \{x_2, x_3, x_8\}, \{x_4, x_6, x_7\}\}$$
$$\ldots\ldots$$
$$\Pi_D = \frac{U}{D} = \{\{x_1, x_4, x_6, x_7\}, \{x_2, x_3, x_5, x_8\}\}.$$

First, we calculate the dependency between the condition attributes and decision attribute with Definition 5 and find that there are two reducts of the system: $C1, C2$ and $C2, C4$

$$\begin{aligned}
H(C1) &= -\sum_{i=1}^{3} p(X_i) \log p(X_i) \\
&= -\frac{3}{8}\log\frac{3}{8} - \frac{2}{8}\log\frac{2}{8} - \frac{3}{8}\log\frac{3}{8} \\
&= 1.5613 \\
H(C2) &= -\sum_{i=1}^{3} p(X_i) \log p(X_i) \\
&= -\frac{2}{8}\log\frac{2}{8} - \frac{3}{8}\log\frac{3}{8} - \frac{3}{8}\log\frac{3}{8} \\
&= 1.5613 \\
&\ldots\ldots \\
H(D) &= -\sum_{i=1}^{2} p(X_i) \log p(X_i) \\
&= -\frac{4}{8}\log\frac{4}{8} - \frac{4}{8}\log\frac{4}{8} \\
&= 1 \\
\Pi_{C1 \cup C2} &= \frac{U}{(C1 \cup C2)} \\
&= \{\{x_1\}, \{x_2\}, \{x_3, x_8\}, \{x_4, x_6\}, \{x_5\}, \{x_7\}\} \\
H(C1 \cup C2) &= -\sum_{i=1}^{5} p(X_i) \log p(X_i) \\
&= -\frac{1}{8}\log\frac{1}{8} - \frac{1}{8}\log\frac{1}{8} - \frac{2}{8}\log\frac{2}{8} - \frac{2}{8}\log\frac{2}{8} \\
&\quad -\frac{1}{8}\log\frac{1}{8} - \frac{1}{8}\log\frac{1}{8} \\
&= 2.5 \\
H(C2|C1) &= H(C1|C2) \\
&= 2.5 - 1.5613 \\
&= 0.9387 \\
H(D|C1C2) &= H(D|C2C4) = 0.
\end{aligned}$$

As we know, information entropy is greater than 0. The above computation shows the decision attribute can be totally precisely approximated if we have attribute set $C1, C2$. $C3, C4$ will make no refinement to the partition by $C1, C2$. Therefore no knowledge will be brought into the system by $C3$ or $C4$. $H(D|C1C2) = H(D|C1C2C3C4)$ and $H(D|C1)$ and $H(D|C2)$ are less than $H(D|C1C2)$. According to Theorem 7, we know $C1, C2$ is a reduct of the decision table. Analogously the set $C2, C4$ is also a reduct.

V. INFORMATION MEASURES ON FUZZY PROBABILISTIC APPROXIMATION SPACES

Shannon's information entropy just works in the case where a crisp equivalence relation or a crisp partition is defined. It is suitable for Pawlak's approximation space. In this section, a novel formula to compute Shannon's entropy with a crisp relation matrix is presented, and then generalized to the fuzzy cases. Furthermore, we will propose another generalization applicable to the case where a probability distribution is defined on the universe and use the proposed entropies to measure the information in fuzzy probabilistic approximation spaces.

A. Shannon's Entropy for Crisp Equivalence Relations

Given a crisp approximation space $\langle U, A\rangle$, Arbitrary relation $R \subseteq U \times U \rightarrow \{0, 1\}$ can be denoted by a relation matrix $M(R)$

$$M(R) = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{pmatrix}$$

where r_{ij} is the relation value between elements x_i and x_j. If R satisfies $R(x, x) = 1$; $R(x, y) = R(y, x)$; and $R(x, y) = 1, R(y, z) = 1$ then $R(x, z) = 1$, we say R is an equivalence relation and $M(R)$ is an equivalence relation matrix.

Then the equivalence class contained x_i with respect to R is written as

$$[x_i]_R = \left\{\frac{r_{i1}}{x_1} + \frac{r_{i2}}{x_2} + \cdots + \frac{r_{in}}{x_n}\right\} \tag{17}$$

where $r_{ij} = 0$ or 1. "1" means that x_j is indiscernible with respect to the relation R and x_j belongs to the equivalence class; "0" means x_j doesn't belong to the class. The cardinality of $[x_i]_R$ is defined as

$$|[x_i]_R| = \sum_{j=1}^{n} r_{ij}. \tag{18}$$

Definition 16: Given an approximation space $\langle U, A\rangle$, an arbitrary equivalence relation R on U, denoted by a relation matrix $M(R)$, then we define the information measure for relation R as

$$H(R) = -\frac{1}{n}\sum_{i=1}^{n} \log \lambda_i \tag{19}$$

where $\lambda_i = |[x_i]_R/n$.

Example 3: There is an object set $\{x_1, x_2, x_3\}$, and a relation matrix of the set induced by a nominal attribute A

$$M(R_A) = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Then, the equivalence class of x_1 can be written as

$$[x_1]_R = \left\{ \frac{1}{x_1} + \frac{1}{x_2} + \frac{0}{x_3} \right\}.$$

And the information entropy of R is calculated by

$$H(R) = -\frac{1}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3}.$$

Intuitively, the object set is divided into two classes $\{\{x_1, x_2\}, \{x_3\}\}$. The information quantity of the relation is

$$H(A) = -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3}.$$

We can find that the information entropy with Definition 16 is equivalent to Shannon's entropy for crisp relations.

Theorem 11: Given an approximation space $\langle U, A\rangle$, $B \subseteq A$, R_B is the equivalence relation generated by attributes B. Then, we have $H(B) = H(R_B)$.

Theorem 12: Given an approximation space $\langle U, A\rangle$, $E, B \subseteq A$, R_E, R_B are two equivalence relations generated by attributes E and B. $[x_i]_E$ and $[x_i]_B$ are the equivalence classes induced by E and B. The joint entropy of E and B is

$$H(EB) = H(R_E R_B) = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{|[x_i]_E \cap [x_i]_B|}{n}$$

Here, EB means $E \cup B$ *and* $R_E R_B$ means $R_E \cap R_B$.

Theorem 13: Given an approximation space $\langle U, A\rangle$, $E, B \subseteq A$, R_E, R_B are two equivalence relation generated by attributes E and B. $[x_i]_E$ and $[x_i]_B$ are the equivalence classes induced by E and B. The conditional entropy E conditioned to B $H(E|B)$ is

$$H(E|B) = H(R_E|R_B) = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{|[x_i]_E \cap [x_i]_B|}{|[x_i]_B|}.$$

Proof: Please see the Appendix.

The previous work reforms Shannon's information measures into a relation matrix representation. The reformation will bring great advantages for generalizing them to the fuzzy cases.

B. Information Measure for Fuzzy Relations

As we know, fuzziness exists in many real-world applications. Dubois *et al.* presented the definitions of fuzzy approximation spaces [11], [12]. In this section, we will present a generalization of Shannon's entropy. The novel measure is with the same form as Shannon's one and can work in the case where fuzzy equivalence relations are defined.

Given a finite set U, $\widetilde{A}$ is a fuzzy attribute set in U, which generates a fuzzy equivalence relation $\widetilde{R}_{\mathrm{A}}$ on U. The fuzzy relation matrix $M(\widetilde{R}_{\mathrm{A}})$ is denoted by

$$M(\widetilde{R}_A) = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{pmatrix}$$

where $r_{ij} \in [0, 1]$ is the relation value of x_i and x_j.

The fuzzy partition generated by the fuzzy equivalence relation is

$$\frac{U}{\widetilde{R}} = \left\{ [x_i]_{\widetilde{R}} \right\}_{i=1}^{n} \tag{20}$$

where $[x_i]_{\widetilde{R}} = \{(r_{i1}/x_1) + (r_{i2}/x_2) + \cdots + (r_{in}/x_n)\}$.

Remark that r_{ij} takes a value in the range [0, 1] here. This is the key difference between the crisp set theory and the fuzzy one. As to a fuzzy partition induced by a fuzzy equivalence relation, the equivalence class is a fuzzy set. " $+$" means the operator of union in this case. The cardinality of the fuzzy set $[x_i]_{\widetilde{R}}$ can be calculated with

$$\left|[x_i]_{\widetilde{R}}\right| = \sum_j r_{ij} \tag{21}$$

which appears to be a natural generalization of the crisp set.

Definition 17: Information quantity of a fuzzy attribute set or a fuzzy equivalence relation is defined as

$$H(\widetilde{A}) = H(\widetilde{R}_A) = -\frac{1}{n}\sum_{i=1}^{n} \log \lambda_i \tag{22}$$

where $\lambda_i = |[x_i]_{\widetilde{R}}|/n$, called a fuzzy relative frequency, n is the number of objects in U.

This measure has the same form as the Shannon's one defined as Definition 16, but it has been generalized to the fuzzy case. The formula of information measure forms a map: $H : R \rightarrow \Re^+$, where R is a equivalence relation matrix, $\Re^+$ is the nonnegative real-number set. This map builds a foundation on which we can compare the discernibility power, partition power or approximating power of multiple fuzzy equivalence relations. The entropy value increases monotonously with the discernibility power of the fuzzy attributes.

Definition 18: Given $\langle U, \widetilde{A}\rangle$, $\widetilde{B}$, $\widetilde{E}$ are two subsets of $\widetilde{A}$. $[x_i]_{\widetilde{B}}$ and $[x_i]_{\widetilde{E}}$ are fuzzy equivalence classes containing x_i generated by $\widetilde{B}$, $\widetilde{E}$, respectively. The joint entropy of $\widetilde{B}$ and $\widetilde{E}$ is defined as

$$H(\widetilde{E}\widetilde{B}) = H(\widetilde{R}_E\widetilde{R}_B) = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{\left|[x_i]_{\widetilde{E}} \cap [x_i]_{\widetilde{B}}\right|}{n}. \tag{23}$$

Definition 19: Given $\langle U, \widetilde{A}\rangle$, $\widetilde{A}$ is the fuzzy attribute set. $\widetilde{B}, \widetilde{E}$ are two subsets of $\widetilde{A}$. The conditional entropy of $\widetilde{E}$ conditioned to $\widetilde{B}$ is defined as

$$H(\widetilde{E}|\widetilde{B}) = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{\left|[x_i]_{\widetilde{E}} \cap [x_i]_{\widetilde{B}}\right|}{\left|[x_i]_{\widetilde{B}}\right|}. \tag{24}$$

Theorem 14: $H(\widetilde{E}|\widetilde{B}) = H(\widetilde{B}\widetilde{E}) - H(\widetilde{B})$.

Theorem 15:

1) $H(\widetilde{R}_A) \geq 0$, " $=$" holds if and only if $r_{ij} = 1, \forall i, \forall j$.
2) $H(\widetilde{R}_A\widetilde{R}_B) \geq \max\{H(\widetilde{R}_A), H(\widetilde{R}_B)\}$.
3) $\widetilde{R}_A \subseteq \widetilde{R}_B \Leftrightarrow H(\widetilde{R}_A\widetilde{R}_B) = H(\widetilde{R}_A)$.
4) $\widetilde{R}_A \subseteq \widetilde{R}_B \Leftrightarrow H(\widetilde{R}_B|\widetilde{R}_A) = 0$.

C. Information Quantity on Fuzzy Probabilistic Approximation Space

Shannon's entropy and the proposed measure work on the assumption that all the objects are equality-probable. In practice, probability of elements in the universe are different. In this section, we will give a generalization where a probability distribution is defined on U.

Given a fuzzy probabilistic approximation space $\langle U, P, \widetilde{A}\rangle$, $\widetilde{A}$ is the fuzzy attribute set, and generates a family of fuzzy equivalence relations on U; P is the probability distribution over U and $p(x_i)$ is the probability of object x_i. A fuzzy equivalence relation $\widetilde{R}_B \subseteq U \times U$ generated by the attribute subset $\widetilde{B}$ is denoted by a relation matrix:

$$M(\widetilde{R}_B) = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{pmatrix}$$

where $r_{ij} \in [0, 1]$. $U/\widetilde{R} = \{[x_i]_{\widetilde{R}}\}_{i=1}^{n}$.

Definition 20: The expected cardinality $\bar{\lambda}_i$ of a fuzzy equivalence class $[x_i]_{\widetilde{R}}$ is defined as

$$\bar{\lambda}_i = \sum_{j=1}^{\mathrm{n}} p(x_j) \cdot r_{ij} \tag{25}$$

Definition 21: The information quantity of a fuzzy attribute set $\widetilde{\mathrm{B}}$ or fuzzy equivalence relation $\widetilde{\mathrm{R}}_B$ is defined as

$$H(\widetilde{B}, P) = -\sum_{i=1}^{n} p(x_i) \log \bar{\lambda}_i. \tag{26}$$

This measure is identical with Yager's entropy [24] in the form, but different in goal. The information measure we give is to compute the discernibility power of a fuzzy attribute set or a fuzzy equivalence relation where a probability distribute is defined on U. while Yager's entropy is to measure the semantics of a fuzzy similarity relation.

Here. we will present a smooth generalization of the definitions of joint entropy and conditional entropy in Shannon's information theory. And the novel generalizations overcome this problem.

Definition 22: Given $\langle U, P, \widetilde{A}\rangle$, $\widetilde{B}$, $\widetilde{E}$ are two subsets of $\widetilde{A}$. The fuzzy equivalence relations induced by $\widetilde{B}$, $\widetilde{E}$ are denoted by $\widetilde{R}$ and $\widetilde{S}$.The joint entropy of $\widetilde{B}$ *and* $\widetilde{E}$ is defined as

$$H(\widetilde{E}\widetilde{B}, P) = H(\widetilde{R}\widetilde{\mathrm{S}}, P) = -\sum_{i=1}^{n} p(x_i) \log \hbar_i \tag{27}$$

where $\hbar_i = \sum_{j=1}^{n} p(x_j)(r_{ij} \wedge s_{ij})$.

Definition 23: The conditional entropy of $\widetilde{E}$ to $\widetilde{B}$ is defined as

$$H(\widetilde{E}|\widetilde{B}, P) = -\sum_{i=1}^{n} p(x_i) \log \frac{\hbar_i}{\bar{\lambda}_i} \tag{28}$$

where $\bar{\lambda}_i = \sum_{j=1}^{\mathrm{n}} p(x_j) \cdot r_{ij}$ and $\hbar_i = \sum_{j=1}^{n} p(x_j)(r_{ij} \wedge s_{ij})$.

Theorem 16: $H(\widetilde{E}|\widetilde{B}, P) = H(\widetilde{B}\widetilde{E}, P) - H(\widetilde{B}, P)$

The forms of the proposed information measures are identical with that of Shannon's ones, however they can be used to measure the information generated by a fuzzy attribute set, a fuzzy equivalence relation or a fuzzy partition.

The previous work presents an information measure for fuzzy equivalence relations when a probability distribution is defined. Here, we will apply it to the fuzzy probabilistic approximation space.

Theorem 17: Given a fuzzy probabilistic approximation space $\langle U, P, \widetilde{A}\rangle$, $\widetilde{A}$ is a fuzzy attribute set; P is the probability distribution on U. $\widetilde{B}$, $\widetilde{E}$ are two subsets of $\widetilde{A}$. The fuzzy equivalence relations induced by $\widetilde{B}$, $\widetilde{E}$ are denoted by $\widetilde{R}$ and $\widetilde{S}$. Then, we have

1) $\forall \widetilde{B} \subseteq \widetilde{A} : H(\widetilde{B}, P) \geq 0$;
2) $H(\widetilde{E}\widetilde{B}, P) \geq \max\{H(\widetilde{E}, P) H(\widetilde{B}, P)\}$;
3) $\widetilde{B} \supseteq \widetilde{E}$ or $\widetilde{R}_B \subseteq \widetilde{R}_E$: $\mathrm{H}(\widetilde{B}\widetilde{E}, P) = H(\widetilde{B}, P)$;
4) $\widetilde{B} \supseteq \widetilde{E}$ or $\widetilde{R}_B \subseteq \widetilde{R}_E$: $H(\widetilde{E}|\widetilde{B}, P) = 0$.

Theorem 18: Given a fuzzy information system $\langle U, \widetilde{A}, P\rangle$, $\widetilde{B} \subseteq \widetilde{A}$, $a \in \widetilde{B}$, $H(\widetilde{B}, P) = H(\widetilde{B} - a, P)$ if a is redundant; $H(\widetilde{B}, P) > H(\widetilde{B} - a, P)$ if $\widetilde{B}$ is independent. $\widetilde{B}$ is a reduct if $\widetilde{B}$ satisfies

1) $H(\widetilde{B}) = H(\widetilde{A})$;
2) $\forall a \in \widetilde{B} : H(\widetilde{B}) > H(\widetilde{B} - a)$.

Theorem 19: Given a fuzzy information system $\langle U, \widetilde{A}, P\rangle \widetilde{A} = \widetilde{C} \cup \widetilde{d}$. $\widetilde{B}$ is a subset of $\widetilde{C}$. $\forall a \in \widetilde{B}$, $H(\widetilde{B} - a|d) = H(\widetilde{B}|d)$ if a is redundant in $\widetilde{B}$ relative to d; $H(\widetilde{B} - a|d) > H(\widetilde{B}|d)$ if $\widetilde{B}$ is independent. $\widetilde{B}$ is a reduct of $\widetilde{C}$ relative to $\widetilde{d}$ if $\widetilde{B}$ satisfies

1) $H(\widetilde{B}|\widetilde{d}) = H(\widetilde{C}|\widetilde{d})$;
2) $\forall a \in \widetilde{B} : H(\widetilde{B} - a|\widetilde{d}) > H(\widetilde{B}|\widetilde{d})$.

Example 4: Given a set $X = \{x_1, x_2, x_3\}$. The probability distribution is $p(x_1) = p(x_2) = 2/5$ and $p(x_3) = 1/5$ Some fuzzy equivalence relations on X are shown as follows:

$$M(\widetilde{R}_1) = \begin{pmatrix} 1 & 0.9 & 0 \\ 0.9 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \; M(\widetilde{R}_2) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0.8 \\ 0 & 0.8 & 1 \end{pmatrix}$$

$$M(\widetilde{R}_3) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0.9 \\ 0 & 0.9 & 1 \end{pmatrix}, \; M(R_d) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

where $M(\widetilde{R}_1)$, $M(\widetilde{R}_2)$ and $M(\widetilde{R}_3)$ are fuzzy equivalence matrices induced by fuzzy condition attributes $a1$, $a2$, and $a3$, $M(R_d)$ is the relation matrix induced by decision d.

First, let's not take the decision into account, and analyze the approximation space without the decision d.

$$H(\widetilde{R}_1,P)=0.54 \quad H(\widetilde{R}_2,P)=0.71 \quad H(\widetilde{R}_2,P)=0.70.$$

Looking at $R1$ and $R3$, we find although the relation matrices of $R1$ and $R3$ are similar, the information quantities are different. The difference comes from the probability distribution of the objects. The probabilities of x_1 and x_2 are greater than that of x_3. x_1 and x_2 are discernible as to R_3, so the total discernibility power of relation R_3 is greater than that of R_1, and $H(\widetilde{R}_1,P) < H(\widetilde{R}_3,P)$

$$M(\widetilde{R}_1\widetilde{R}_2\widetilde{R}_3)=\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}$$

and

$$M(\widetilde{R}_1\widetilde{R}_2)=M(\widetilde{R}_1\widetilde{R}_3)=\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}.$$

We have $U/\widetilde{R}_1\widetilde{R}_2=U/\widetilde{R}_1\widetilde{R}_3$.

$$M(\widetilde{R}_1)\neq M(\widetilde{R}_1\widetilde{R}_2\widetilde{R}_3)$$
$$M(\widetilde{R}_2)\neq M(\widetilde{R}_1\widetilde{R}_2\widetilde{R}_3)$$
$$M(\widetilde{R}_3)\neq M(\widetilde{R}_1\widetilde{R}_2\widetilde{R}_3).$$

We can conclude $\{\widetilde{R}_1,\widetilde{R}_2\}$ and $\{\widetilde{R}_1,\widetilde{R}_3\}$ are independent and have the same discernibility power as $\{\widetilde{R}_1,\widetilde{R}_2,\widetilde{R}_3\}$, respectively. So $a1,a2$ and $a1,a3$ are two reducts

$$H(a_1a_2a_3,P)=1.055, H(a_1a_2,P)=H(a_1a_3,P)=1.055$$
$$H(a_{1_1},P)<H(a_2,P)<H(a_3,P)<H(a_1a_2a_3,P)$$

From Theorem 19, $a1,a2$ and $a1,a3$ are reducts of the approximation space.

In the same way, we can find $a1,a2$ and $a1,a3$ are relative reducts of the space.

VI. Conclusion and Discussion

The contribution of this paper is two-fold. On one side, we generalize fuzzy approximation spaces to fuzzy probabilistic approximation spaces by introducing a probability distribution on U. On the other side, we reform Shannon's information measures into relation matrix representations and extend them to the fuzzy probabilistic approximation spaces. The proposed definitions of fuzzy probabilistic approximation spaces integrate three types of uncertainty: fuzziness, probability and roughness into one framework. The analysis shows that the fuzzy probabilistic approximation space will degrade to fuzzy approximation space if the uniform distribution assumption holds. Furthermore, an approximation space is Pawlak's one if equivalence relations and the subsets to be approximated are crisp. Therefore the fuzzy probabilistic approximation spaces unify the representations of the approximation spaces. Accordingly, the information measures for fuzzy probabilistic approximation spaces give uniform formulas to calculate the information quantity of the spaces.

The probability characterizes the uncertainty of randomness of event sets and is an efficient tool to deal with inconsistency and noise in data. Introducing probability into an approximation space presents a gate for statistical techniques applying to rough set methodology, which maybe lead to a tool for randomness, incompleteness, inconsistence and vagueness in real-world applications.

Appendix

Theorem 13: Given a set U with n elements and two crisp equivalence relation matrices R, S, the K,L equivalence classes generated by R and S are denoted by X_k and Y_l, respectively. The equivalence classes contained x_i are denoted by $[x_i]_R$ and $[x_i]_S$, then we have

$$\frac{1}{n}\sum_{i=1}^{n}\log\frac{|[x_i]_R\cap[x_i]_S|}{|[x_i]_S|} = \sum_{l=1}^{L}P(Y_l)\sum_{k=1}^{K}P(X_k|Y_l)\log P(X_k|Y_l)$$

where

$$|U|=n \quad P(Y_l)=\frac{|Y_l|}{|U|}=\frac{|Y_l|}{n} \quad P(X_k|Y_l)=\frac{|X_k\cap Y_l|}{|Y_l|}.$$

Proof:

$$\frac{1}{n}\sum_{i=1}^{n}\log\frac{|[x_i]_R\cap[x_i]_S|}{|[x_i]_S|} = \frac{1}{n}\left(\sum_{i=1}^{n}\log|[x_i]_R\cap[x_i]_S| - \sum_{i=1}^{n}\log|[x_i]_S|\right)$$

$$\sum_{l=1}^{L}P(Y_l)\sum_{k=1}^{K}P(X_k|Y_l)\log P(X_k|Y_l) = \sum_{l=1}^{L}\frac{|Y_l|}{n}\sum_{k=1}^{K}\frac{|X_k\cap Y_l|}{|Y_l|}\log\frac{|X_k\cap Y_l|}{|Y_l|}$$
$$=\frac{1}{n}\sum_{l=1}^{L}\sum_{k=1}^{K}|X_k\cap Y_l|\log\frac{|X_k\cap Y_l|}{|Y_l|}$$
$$=\frac{1}{n}\left(\sum_{l=1}^{L}\sum_{k=1}^{K}|X_k\cap Y_l|\log|X_k\cap Y_l\ | - \sum_{l=1}^{L}\sum_{k=1}^{K}|X_k\cap Y_l|\log|Y_l|\right).$$

$$\sum_{k=1}^{K}|X_k\cap Y_l| = |Y_l|, \text{ then}$$

$$\sum_{l=1}^{L}\sum_{k=1}^{K}|X_k\cap Y_l|\log|Y_l|) = \sum_{l=1}^{L}|Y_l|\log|Y_l|.$$

Because there are the following properties between: $[x_i]_S|_{i=1}^n$ and $U = \{Y_1, Y_2, \ldots, Y_L\}$

- $\cup_{l=1}^L Y_l = U$, that is $|\cup_{l=1}^L Y_l| = n$;
- $\forall Y_a, Y_b \in \{Y_1, Y_2, \ldots, Y_L\}$, $Y_a \cap Y_b = \Phi$;
- $\forall Y_l \in \{Y_1, Y_2, \ldots, Y_L\}$, $|Y_l| = t$, such that

$$Y_l = [x_{l1}]_S = [x_{l2}]_S = \cdots = [x_{lt}]_S.$$

Then, we have

$$\sum_{i=1}^n \log|[x_i]_S| = \sum_{l=1}^L |Y_l| \log|Y_l|. \tag{29}$$

Now, we just require proving

$$\sum_{i=1}^n \log|[x_i]_R \cap [x_i]_S| = \sum_{l=1}^L \sum_{k=1}^K |X_k \cap Y_l| \log|X_k \cap Y_l|.$$

Just the same as before, we have

- $\bigcup_{l=1}^L \bigcup_{k=1}^K (X_k \cap Y_l) = U \sum_{l=1}^L \sum_{k=1}^K |X_k \cap Y_l| = \sum_{l=1}^L |Y_l| = n$;
- $\forall X_a, \quad X_b \quad \in \quad \{X_1, X_2, \ldots, X_K\}; \quad \forall Y_c$, $Y_d \in \{Y_1, Y_2, \ldots, Y_L\}$, $(X_a \cap Y_c) \cap (X_b \cap Y_d) = \Phi$;
- Assuming that

$$|X_k \cap Y_l| = n_{kl} \quad X_k \cap Y_l = \{x_{n_1}, x_{n_2}, \ldots, x_{n_{\mathrm{kl}}}\}.$$

Then

$$\begin{aligned}|X_k \cap Y_l| \log|X_k \cap Y_l| &= \log|[x_{n_1}]_R \cap [x_{n_1}]_S| \\ &\quad + \log|[x_{n_2}]_R \cap [x_{n_2}]_S| + \cdots + \log\left|[x_{n_{kl}}]_R \cap [x_{n_{kl}}]_S\right|\end{aligned}$$

Now, we have

$$\sum_{l=1}^L \sum_{k=1}^K |X_k \cap Y_l| \log|X_k \cap Y_l| = \sum_{i=1}^n \log|[x_i]_R \cap [x_i]_S|. \tag{30}$$

Combine (29) with (30), we can reach the conclusion.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their helpful and insightful comments and suggestions.

REFERENCES

[1] W. Swiniarski, Roman, and L. Hargis, "Rough sets as a front end of neural-networks texture classifiers," *Neurocomput.*, vol. 36, no. 1–4, pp. 85–102, 2001.

[2] W. Swiniarski and A. Skowron, "Rough set methods in feature selection and recognition," *Pattern Recog. Lett.*, vol. 24, no. 6, pp. 833–849, 2003.

[3] Q. Hu, D. Yu, and Z. Xie, "Reduction algorithms for hybrid data based on fuzzy rough set approaches," in *Proc. 2004 Int. Conf. Machine Learning and Cybernetics*, pp. 1469–1474.

[4] S. Tsumoto, "Automated extraction of hierarchical decision rules from clinical databases using rough set model," *Expert Syst. Appl.*, vol. 24, no. 2, pp. 189–197, 2003.

[5] N. Zhong, J. Dong, and S. Ohsuga, "Rule discovery by soft induction techniques," *Neurocomput.*, vol. 36, no. 1–4, pp. 171–204.

[6] T. P. Hong, L. Tseng, and S. Wang, "Learning rules from incomplete training examples by rough sets," *Expert Syst. Appl.*, vol. 22, no. 4, pp. 285–293, 2002.

[7] L. Polkowski and A. Skowron, "Rough mereology: A new paradigm for approximate reasoning. Intern," *J. Approx. Reason.*, vol. 15, no. 4, pp. 333–365, 1996.

[8] Z. Pawlak, "Rough sets, decision algorithms and Bayes' theorem," *Eur. J. Oper. Res.*, vol. 136, no. 1, pp. 181–189, 2002.

[9] Z. Pawlak, "Granularity of knowledge, indiscernibility and rough sets," in *Proc. 1998 IEEE Int. Conf. Fuzzy Systems*, 1998, pp. 106–110.

[10] L. Zadeh, "Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic," *Fuzzy Sets Syst.*, vol. 19, pp. 111–127, 1997.

[11] D. Dubois and H. Prade, "Rough fuzzy sets and fuzzy rough sets," *Int. J. Gen. Syst.*, vol. 17, no. 2–3, pp. 191–209, 1990.

[12] D. Dubois and H. Prade, "Putting fuzzy sets and rough sets together," in *Intelligent Decision Support*, R. Slowiniski, Ed. Dordrecht, The Netherlands: Kluwer, 1992, pp. 203–232.

[13] N. Morsi Nehad and M. M. Yakout, "Axiomatics for fuzzy rough sets," *Fuzzy Sets Syst.*, vol. 100, no. 1–3, pp. 327–342, November 16, 1998.

[14] R. Anna Maria and E. E. Kerre, "A comparative study of fuzzy rough sets," *Fuzzy Sets Syst.*, vol. 126, no. 2, pp. 137–155, 2002.

[15] W. Wu and W. Zhang, "Constructive and axiomatic approaches of fuzzy approximation operators," *Inform. Sci.*, vol. 159, no. 3–4, pp. 233–254, 2004.

[16] J. Mi and W. Zhang, "An axiomatic characterization of a fuzzy generalization of rough sets," *Inform. Sci.*, vol. 160, no. 1–4, pp. 235–249, 2004.

[17] Y.-F. Wang, "Mining stock price using fuzzy rough set system," *Expert Syst. Appl.*, vol. 24, no. 1, pp. 13–23, 2003.

[18] S. Padmini and R. Miguel *et al.*, "Vocabulary mining for information retrieval: rough sets and fuzzy sets," *Inform. Process. Manage.*, vol. 37, no. 1, pp. 15–38.

[19] Q. Shen and A. Chouchoulas, "A rough-fuzzy approach for generating classification rules," *Pattern Recog.*, vol. 35, no. 11, pp. 2425–2438, 2002.

[20] C. Shannon and W. Weaver, *The Mathematical Theory of Communication*. Champaign, IL: Univ. Illinois Press, 1964.

[21] B. Forte, "Measure of information: The general axiomatic theory," *RIRO*, vol. R2, no. 3, pp. 63–90, 1969.

[22] J. Kampe de Feriet and B. Forte, "Information etc Probabilite CRAS Paris," in ser. A, vol. 265, 1967, pp. 110–114, 143-146.

[23] L. Zadeh, "Probability measures of fuzzy events," *J. Math. Anal. Appl.*, vol. 23, pp. 421–427, 1968.

[24] R. Yager, "Entropy measures under similarity relations," *Int. J. Gen. Syst.*, vol. 20, pp. 341–358, 1992.

[25] E. Hernandez and J. Recasens, "A reformulation of entropy in the presence of indistinguishability operators," *Fuzzy Sets Syst.*, vol. 128, pp. 185–196, 2002.

[26] R. Mesiar and J. Rybarik, "Entropy of fuzzy partitions: a general model," *Fuzzy Sets Syst.*, vol. 99, pp. 73–79, 1998.

[27] C. Bertoluzza, V. Doldi, and G. Naval, "Uncertainty measure on fuzzy partitions," *Fuzzy Sets Syst.*, vol. 142, pp. 105–116, 2004.

[28] R. Jensen and Q. Shen, "Fuzzy-rough attribute reduction with application to web categorization," *Fuzzy Sets Syst.*, vol. 141, pp. 469–485, 2004.

[29] J. F. Peters, Z. Pawlak, and A. Skowron, "A rough set approach to measuring information granules," in *Computer Software and Applications Conf.*, 2002, pp. 1135–1139.

[30] R. Yager, "On the entropy of fuzzy measures," *IEEE Trans. Fuzzy Syst.*, vol. 8, no. 4, pp. 453–461, Aug. 2000.

[31] S. Greco, B. Matarazzo, and R. Słowiński, "Rough sets methodology for sorting problems in presence of multiple attributes and criteria," *Eur. J. Oper. Res.*, vol. 138, no. 2, pp. 247–259, 2002.

[32] ——, “Fuzzy extension of the rough set approach to multicriteria and multiattribute sorting,” in *Preferences and Decisions Under Incomplete Knowledge*. Heidelberg, Germany: Physica-Verlag, 2000, pp. 131–151.

[33] Y. Yao, “Information granulation and rough set approximation,” *Int. J. Intell. Syst.*, vol. 16, no. 1, pp. 87–104, 2001.

[34] T. Y. Lin, “From rough sets and neighborhood systems to information granulation and computing in words,” *Proc. Eur. Congr. Intelligent Techniques and Soft Computing*, pp. 1602–1606, Sep. 8–12, 1997.

[35] W. Pedrycz, “Shadowed sets: bridging fuzzy and rough set,” in *Rough Fuzzy Hybridization: A New Trend in Decision Making*, S. K. Pal and A. Skowron, Eds. Berlin, Germany: Springer-Verlag, 1999.

[36] G. Wang, H. Yu, and D. Yang, “Decision table reduction based on conditional information entropy,” *Chinese J. Comp.*, vol. 25, no. 7, pp. 1–9, 2002.

[37] Q. Hu and D. Yu, “Entropies of fuzzy indiscernibility relation and its operations,” *Int. J. Uncertainty, Fuzziness Knowledge-Based Syst.*, vol. 12, no. 5, pp. 575–589, 2004.

[38] D. Li, B. Zhang, and Y. Leung, “On knowledge reduction in inconsistent decision information systems,” *Int. J. Uncertainty, Fuzziness Knowledge-Based Syst.*, vol. 12, no. 5, pp. 651–672, 2004.

[39] L. Zadeh, “Fuzzy logic equals computing with words,” *IEEE Trans. Fuzzy Syst.*, vol. 4, no. 2, pp. 103–111, Apr. 1996.

[40] J. Casasnovas and J. Torrens, “An axiomatic approach to scalar cardinalities of fuzzy sets,” *Fuzzy Sets Syst.*, vol. 133, no. 2, pp. 193–209, 2003.

[41] Q. Hu, D. Yu, and Z. Xie, “Information-preserving hybrid data reduction based on fuzzy-rough techniques,” *Pattern Recognit. Lett.*, vol. 27, no. 5, pp. 414–423, 2006.

[42] L. Zadeh, “A new direction in AI—Toward a computational theory of perceptions,” *AI Mag.*, vol. 22, no. 1, pp. 73–84, 2001.

[43] Y. Zhang, “Constructive granular systems with universal approximation and fast knowledge discovery,” *IEEE Trans. Fuzzy Syst.*, vol. 13, no. 1, pp. 48–57, Feb., 2005.

Qinghua Hu received the M.S. degree in power engineering from Harbin Institute of Technology, Harbin, China, in 2002. He is current;y working toward the Ph.D. degree at Harbin Institute of Technology.

His research interests are focused on data mining and knowledge discovery in historical record database of power plants with fuzzy and rough techniques. He has authored or coauthored more than 20 journal and conference papers in the areas of machine learning, data mining, and rough set theory.

Daren Yu was born in Datong, China, in 1966. He received the M.Sc. and D.Sc. degrees from Harbin Institute of Technology, Harbin, China, in 1988 and 1996, respectively.

Since 1988, he has been with the School of Energy Science and Engineering, Harbin Institute of Technology. His main research interests are in modeling, simulation, and control of power systems. He has published more than one hundred conference and journal papers on power control and fault diagnosis.

Zongxia Xie , photograph and biography not available at the time of publication.

Jinfu Liu , photograph and biography not available at the time of publication.