# Probabilistic Rough Classification in Information Systems with Fuzzy Decision Attributes

D.Latha(1) D.Rekha(2) K.Thangadurai(3) G.Ganesan(4)

(1)Department of Computer Science, Adikavi Nannaya University, Rajahmundry, Andhra Pradesh, India
(2)Department of Computer Applications, St.Peter's University, Chennai, Tamilnadu, India
(3)Department of Computer Science, Government College, Karur, Tamilnadu, India
(4)Department of Mathematics, Adikavi Nannaya University, Rajahmundry, Andhra Pradesh, India

***ABSTRACT:*** **Based on Pawlak's two way approximations on Rough Sets and using thresholds G.Ganesan et al in 2004 proposed a method of rough indexing an information system which has fuzzy decision attributes. The limitation of Pawlak's approximation is that it does not quantify the level of importance of the basic granules. Recently, Y.Y.Yao discussed Probabilistic Rough Set Model, which specified how basic granules could be quantified appropriately. In this paper, it is proposed to extend the work of G.Ganesan et al, taking into consideration the basic granule quantification mechanism of Probabilistic Set Model, thus generating more accurate rough indices for information systems with fuzzy decision attributes**

**Keywords: information system, rough set, probabilistic rough set, rough index**

*****

## 1. INTRODUCTION

The theory of Rough Sets introduced by Z. Pawlak [4,5] has applications in several areas, including fields of knowledge acquisition and discovery. According to this theory, using either of the two ways of performing union of Basic Categories, a given input or concept can be approximated. But this approximation does to take into consideration the degree of contribution of the basic categories. This limitation has been eliminated in Variable Precision Rough Sets model defined by W. Ziarko in 1993. This model has further been extended by Bing Zhou, YY Yao, Slezak and others to include probabilistic factors

The different hybridized rough and fuzzy models as proposed by Dubois, Prade, Nakamura, Biswas have varied real time applications. G.Ganesan et.al, in 2005 discussed the importance of defining thresholds in rough fuzzy computing and subsequently, in 2008, proposed a method of indexing an information system with fuzzy decision attributes using these thresholds. In [7], Slezak initiated a study on Bayesian Rough Set model and in [8], Yiyu Yao and Bing Zhou have discussed the Naïve Bayesian Rough Set Model. In this paper, we are extending the rough indexing to the information systems by incorporating the Probabilistic Naïve Bayesian Rough Set Model.

## 2. DECISION THEORETIC AND PROBABILISTIC ROUGH SETS

Rough Sets [4,5] theory defines two way approximations namely lower and upper approximations for a given input. For a given finite universe of discourse U and an equivalence relation E, the equivalence class of any x∈U is defined to be $[x]=\{y\in U / xEy\}$. The family of equivalence classes $U/E=\{[x]_E \mid x\in U\}$ is a partition of the universe U. For a given concept C, Pawlak defined the lower approximation as $\underline{apr}_E(C)=\{x\in U/[x]_E\subseteq C\}$ and upper approximation as $\overline{apr}_E(C)=\{x\in U/[x]_E\cap C\neq\Phi\}$.

For a given concept C, three disjoint regions, namely positive, negative and boundary regions are defined as follows:

Positive Region: $POS_E(C)=\{x\in U/[x]_E\subseteq C\}$

Boundary : $BND_E(C)=\{x\in U/[x]_E\cap C\neq\Phi\wedge[x]_E\not\subseteq C\}$

Negative region: $NEG_E(C)=\{x\in U/[x]_E\cap C=\Phi\}$

As Pawlak's model is restrictive, several researchers focused on generalizing this approach towards parameterized rough set model, probabilistic rough set model and generalized rough set model.

In 1994, Pawlak and Skowron [6] defined rough membership function by considering degrees of overlap between equivalence classes and a concept C to be approximated and is viewed as the conditional probability of an object belongs to C given that the object is in [x] (for simplicity, we denote $[x]_E$ with [x]) which is given as $\Pr\left(C/[x]\right)=\frac{|C\cap[x]|}{|[x]|}$

Using the definition quoted above, in [8], the positive, boundary and negative regions are defined as follows:

$$POS(C)=\left\{x\in U/\Pr\left(C/[x]\right)=1\right\}$$
$$BND(C)=\left\{x\in U/0<\Pr\left(C/[x]\right)<1\right\}$$
$$NEG(C)=\left\{x\in U/\Pr\left(C/[x]\right)=0\right\}$$

In 2009, Greco et.al [3] discussed the parameterized rough set model by generalizing the above said definitions. In this model, two thresholds namely $\alpha$ and $\beta$ are used to define the probabilistic regions and the positive, boundary and negative regions are modified as follows:

$$POS_{(\alpha,\beta)}(C)=\left\{x\in U/\Pr\left(C/[x]\right)\geq\alpha\right\}$$
$$BND_{(\alpha,\beta)}(C)=\left\{x\in U/\beta<\Pr\left(C/[x]\right)<\alpha\right\}$$
$$NEG_{(\alpha,\beta)}(C)=\left\{x\in U/\Pr\left(C/[x]\right)\leq\beta\right\}$$

These Probabilistic regions will lead three way decisions namely acceptance, deferment and rejection respectively for any object x in U. But, however, in several cases, it is easy to compute the probability of the existence of a category [x] for a given concept C using $\Pr\left([x]/C\right)=\frac{|[x]\cap C|}{|C|}$

However, by Baye's Theorem, the Positive, Boundary and Negative Regions are given by

$$POS^{B}_{(\alpha',\beta')}(C)=\left\{x\in U/\log\frac{\Pr([x]/C)}{\Pr([x]/C^c)}\geq\alpha'\right\}$$
$$BND^{B}_{(\alpha',\beta')}(C)=\left\{x\in U/\beta'<\log\frac{\Pr([x]/C)}{\Pr([x]/C^c)}<\alpha'\right\}$$
$$NEG^{B}_{(\alpha',\beta')}(C)=\left\{x\in U/\log\frac{\Pr([x]/C)}{\Pr([x]/C^c)}\leq\beta'\right\}$$

$$where\quad \alpha'=\log\frac{\Pr(C^c)}{\Pr(C)}+\log\frac{\alpha}{1-\alpha}$$

$$and\quad \beta'=\log\frac{\Pr(C^c)}{\Pr(C)}+\log\frac{\beta}{1-\beta}$$

## 3. ANALYSIS OF FUZZY SET USING A THRESHOLD

In this section we shall discuss the conventional approach of approximating fuzzy sets under rough fuzzy computing.

Consider a set D, called **R-domain** [1], satisfying the following properties:

a) $D\subset(0,1)$
b) If a fuzzy concept C is under computation, eliminate the values $\mu_C(x)$ and $\mu_C^c(x)$ $\forall x\in U$ from the domain D, if they exist.
c) After the computation using C, the values removed in (b) may be included in D provided C must not involve in further computation

Consider the universe of discourse $U=\{x_1,x_2,\ldots,x_n\}$. Let $\alpha,\alpha_1,\alpha_2,\beta$ be the thresholds assume one of the values from the domain D, where D is constructed using the fuzzy concepts A and B. For a given threshold $\alpha$ and a fuzzy set A, the Strong $\alpha$-Cut is given by $A[\alpha]=\{x\in U/\mu_A(x)>\alpha\}$. The union and intersection of fuzzy sets [10] are by the maximum and minimum of corresponding membership values respectively.

In 1972, Zadeh [9] introduced the concept of hedges. In fuzzy logic, in order to improve the efficiency of fuzziness, the concept of concentration and dilation were introduced by him.

For example, for the linguistic variable 'low' with the membership function $\alpha$, the hedges 'very' and 'very very' emphasis the efficiency of the variable with the corresponding membership values $\alpha^2$ and $\alpha^4$. They are called **concentration**, whereas the hedges 'slightly' and 'more slightly' dilutes the efficiency of the linguistic variables with the membership values with the corresponding membership values $\alpha^{1/2}$ and $\alpha^{1/4}$. They are called **dilation**.

Using the definitions of fuzzy sets mentioned above, the following properties were derived in [1].

a) $A[\alpha_1]\cup A[\alpha_2]=A[\alpha]$ where $\alpha=\min(\alpha_1,\alpha_2)$
b) $A[\alpha_1]\cap A[\alpha_2]=A[\alpha]$ where $\alpha=\max(\alpha_1,\alpha_2)$
c) $(A\cup B)[\alpha]=A[\alpha]\cup B[\alpha]$
d) $(A\cap B)[\alpha]=A[\alpha]\cap B[\alpha]$
e) $A^c[\alpha]=A[1-\alpha]^c$
f) $(A\cup B)^c[\alpha]=A^c[\alpha]\cap B^c[\alpha]$
g) $(A\cap B)^c[\alpha]=A^c[\alpha]\cup B^c[\alpha]$

Using the mathematical tool derived as above, in [1], rough set approach on fuzzy sets using a threshold is introduced as discussed below.

### *3.1 Rough Approximations on fuzzy sets using α*

Let $\Psi$ be any partition of U, say $\{B_1,B_2,\ldots,B_t\}$. For the given fuzzy concept, the lower and upper approximations with respect to $\alpha$ can be defined as ${}_{\alpha}C=\underline{(C[\alpha])}$ and ${}^{\alpha}C=\overline{(C[\alpha])}$ respectively.

#### *3.1.1 Propositions*

Here, by using the properties of rough sets, the following propositions [1] can be obtained.

a) $^{\alpha}(A\cup B) = {}^{\alpha}A \cup {}^{\alpha}B$
b) $_{\alpha}(A\cap B) = {}_{\alpha}A \cap {}_{\alpha}B$
c) $_{\alpha}(A\cup B) \supseteq {}_{\alpha}A \cup {}_{\alpha}B$
d) $^{\alpha}(A\cap B) \subseteq {}^{\alpha}A \cap {}^{\alpha}B$
e) $^{\alpha}(A^c) = ({}_{1-\alpha}A)^c$
f) $_{\alpha}(A^c) = ({}^{1-\alpha}A)^c$

Now, we shall hybridize the concepts dealt in the above two sections which gives the approach of dealing a fuzzy concepts under Naïve Bayesian Probabilistic Rough Sets.

## 4. NAÏVE BAYESIAN PROBABILISTIC ROUGH SETS MODEL FOR A FUZZY CONCEPT

Since, in the above both sections, the same threshold α has been used, for different purposes, to make the homogeneity, in this paper, we replace the threshold α to obtain a Strong Cut on fuzzy sets with δ.

Hence, for a given fuzzy concept F with the threshold δ, the probabilistic positive, boundary and negative regions are respectively defined on the approximation space U/E as

$$POS_{\delta}(F) = \left\{ x \in U / \Pr\left( F[\delta] / [x] \right) = 1 \right\}$$

$$BND_{\delta}(F) = \left\{ x \in U / 0 < \Pr\left( F[\delta] / [x] \right) < 1 \right\}$$

$$NEG_{\delta}(F) = \left\{ x \in U / \Pr\left( F[\delta] / [x] \right) = 0 \right\}$$

For given parameters α and β, the regions of the parameterized rough sets model are given by

$$POS_{(\alpha,\beta,\delta)}(F) = \left\{ x \in U / \Pr\left( F[\delta] / [x] \right) \geq \alpha \right\}$$

$$BND_{(\alpha,\beta,\delta)}(F) = \left\{ x \in U / \beta < \Pr\left( F[\delta] / [x] \right) < \alpha \right\}$$

$$NEG_{(\alpha,\beta,\delta)}(F) = \left\{ x \in U / \Pr\left( F[\delta] / [x] \right) \leq \beta \right\}$$

and the Regions of Naïve Bayesian Rough Sets Model are given by

$$POS^{B}_{(\alpha',\beta',\delta)}(F) = \left\{ x \in U / \log \frac{\Pr\left([x] / F[\delta]\right)}{\Pr\left([x] / (F[\delta])^{C}\right)} \geq \alpha' \right\}$$

$$BND^{B}_{(\alpha',\beta',\delta)}(F) = \left\{ x \in U / \beta' < \log \frac{\Pr\left([x] / F[\delta]\right)}{\Pr\left([x] / (F[\delta])^{C}\right)} < \alpha' \right\}$$

$$NEG^{B}_{(\alpha',\beta',\delta)}(F) = \left\{ x \in U / \log \frac{\Pr\left([x] / F[\delta]\right)}{\Pr\left([x] / (F[\delta])^{C}\right)} \leq \beta' \right\}$$

$$\text{where} \quad \alpha' = \log \frac{\Pr(C^c)}{\Pr(C)} + \log \frac{\alpha}{1-\alpha}$$

$$\text{and} \quad \beta' = \log \frac{\Pr(C^c)}{\Pr(C)} + \log \frac{\beta}{1-\beta}$$

## 5. ROUGH INDICES

Let U be the universe of discourse and α be any value in (0,1). Let $X=\{W_1,W_2,\ldots,W_n\}$ be any partition defined on U. For any fuzzy set A define $A[\alpha]=\{x\in U/ \mu_A(x)>\alpha\}$ where α is chosen from R-domain satisfying the property that $dil^n(\alpha)$ and $con^n(\alpha)$ are the members of R-Domain for any positive integer n [dil represents dilation and con represents concentration]. The lower and upper approximations $A_{\alpha}$ and $A^{\alpha}$ are given by $A_{\alpha}=(A[\alpha])_{\sim}$ and $A^{\alpha}=(A[\alpha])^{\sim}$ respectively [1].

The following algorithm as in [2] illustrates the method of indexing the elements of U, by using the lower and upper approximations of the given fuzzy set A

Let M denote the largest number under consideration such that n+M is always positive and n-M is always negative for any integer n.

### *5.1: ALGORITHMS*

```
Algorithm rough index (x,A,α)
//Algorithm to obtain rough index of x an element of universe of discourse
//Algorithm returns the rough index
  1. Let x_index be an integer initialized to 0
  2. Pick the equivalence class K containing x.
        If μA(y)=0 for all y∈K
        begin
              x_index=-M
              goto 6
        end
     end
  3. If μA(x)=1
     begin
        If μA(y)=1 for all y∈K
        begin
              x_index=M
              goto 6
        end
     end
  4. compute Aα and A^α
  5. If x∈Aα
     begin
     x_index=M
     while (x∈Aα)
       begin
```

```
            α=dil(α)        //dilation of α
            x_index=x_index+1
            compute A_α
      end
    end
    else
            if x∉A^α
            begin
            x_index=-M
            while (x∉A^α)
              begin
                α=con(α)//concentration of α
               x_index=x_index-1
                compute A^α
               end
            end
            else
            Let β=α
            compute A^β
            while (x∉A_α and x∈A^β)
             begin
               α=con(α) //concentration of α
               β=dil(β)     // dilation of β
               compute A_α,A^β
               x_index=x_index+1
             end
            if x∈A_α then
             x_index= - x_index
    end
6.  return x_index
```

Consider the universe of discourse U={a,b,c,d,e,f,g,h} with the partition X={{a,e,f},{b,g},{c,h}, {d}}. Let α=0.5.

Consider the fuzzy set {(a,0.6),(b,0.4), (c,0.8), (d,0.24), (e,0.44), (f,0.56), (g,0.98), (h,0.77)}. By the above algorithm, `b' can be indexed by –1 and d can be indexed by –2-M. Similarly, other values of U can be indexed. These indices are called rough indices.

Here, it is to be noted that the elements of same equivalence classes have the same rough indices. Sometimes, the elements of different equivalence classes may have same rough index. Clearly, it depends upon the choice of α and the fuzzy set taken under consideration. But,, it is obvious that the elements of the same equivalence classes will have the same rough indices and therefore, instead of indexing the elements of U, one may follow the given algorithm for rough indexing the equivalence classes.

Now, we shall modify this algorithm for three way approach on rough sets as follows:

```
Algorithm Three_Way_ rough index (x,A,δ)
//Algorithm returns the Three_Way_rough index of x
1.  Let x_index be an integer initialized to 0
2.  Pick the equivalence class K containing x.
            If μ_A(y)=0 for all y∈K
            begin
                  x_index=-M
                  goto 6
            end
    end
3.  If μ_A(x)=1
    begin
             If μ_A(y)=1 for all y∈K
            begin
                  x_index=M
                  goto 6
            end
    end
4.  compute POS_δ(A), BND_δ(A) and NEG_δ(A)
5.  If x∈ POS_δ(A)
    begin
    x_index=M
    while (x∈ POS_δ(A))
      begin
            α=dil(δ)        //dilation of δ
            x_index=x_index+1
            compute POS_δ(A)
      end
    end
    else
            if x∈ NEG_δ(A)
            begin
            x_index=-M
            while (x∈ NEG_δ(A))
              begin
                δ=con(δ)//concentration of δ
               x_index=x_index-1
                compute NEG_δ(A)
               end
            end
            else
            Let γ=δ
            compute NEG_γ(A)
            while (x∉(POS_δ(A)∪NEG_γ(A)))
             begin
               δ=con(δ) // concentration of δ
               γ=dil(γ)     // dilation of γ
               compute POS_δ(A)∪NEG_γ(A)
               x_index=x_index+1
             end
            if x∈ POS_δ(A) then
             x_index= - x_index
    end
6.  return x_index
```

This algorithm can be illustrated in the same manner as mentioned in the previous example. Now, we parameterize the algorithm using parameters α and β.

```
Algorithm Naïve Bayesian_ rough index (x,A,α,β,δ)
//Algorithm returns Naïve Bayesian_rough index of x
1.  Let x_index be an integer initialized to 0
```

2. Pick the equivalence class K containing x.

```
        If μA(y)=0 for all y∈K
        begin
                x_index=-M
                goto 6
        end
   end
3. If μA(x)=1
   begin
        If μA(y)=1 for all y∈K
        begin
                x_index=M
                goto 6
        end
   end
4. compute POS^B(α',β',δ)(A), BND^B(α',β',δ)(A) and
   NEG^B(α',β',δ)(A)
5. If x∈ POS^B(α',β',δ)(A)
   begin
   x_index=M
   while (x∈ POS^B(α',β',δ)(A))
     begin
        α=dil(δ)        //dilation of δ
        x_index=x_index+1
        compute POS^B(α',β',δ)(A)
     end
   end
else
        if x∈ NEG^B(α',β',δ)(A)
        begin
        x_index=-M
        while (x∈ NEG^B(α',β',δ)(A))
          begin
            δ=con(δ)//concentration of δ
           x_index=x_index-1
            compute NEG^B(α',β',δ)(A)
          end
        end
        else
        Let γ=δ
        compute NEGγ(A)
        while
   (x∉(POS^B(α',β',δ)(A)∪NEG^B(α',β',γ)(A)))
          begin
            δ=con(δ) // concentration of δ
            γ=dil(γ)     // dilation of γ
          compute
            POS^B(α',β',δ)(A)∪NEG^B(α',β',γ)(A)
            x_index=x_index+1
          end

        if x∈ POS^B(α',β',δ) (A) then
          x_index= - x_index
   end
6. return x_index
```

## 6. NAÏVE BAYESIAN INDEXING IN INFORMATION SYSTEM WITH FUZZY DECISION ATTRIBUTE

According to the perspective of Z.Pawlak, any information system is given by T=(U, A, C, D), where U is the universe of discourse, A is a set of primitive attributes, C and D are the subsets of A called condition and decision features respectively [C and D may not exist in a few of the information systems].

Consider an information system with conditional attributes $C=\{a_1,a_2,\dots,a_n\}$ and decision attributes $\{d_1,d_2,\dots,d_s\}$ with the records $U=\{x_1,x_2,\dots,x_m\}$. For any index key 'a' in C, the indiscernibility relation is given by $x_i \approx_{a_k} x_j$ (read as $x_i$ is related to $x_j$ with respect to $a_k$) if and only if $a_k(x_i)=a_k(x_j)$. Clearly, this indiscernibility relation partitions the universe of discourse U. However, the procedure of selecting the appropriate minimal attributes [reducts] for effectiveness is not discussed in this paper.

For example, consider the decision table with C={a,b,c,d} and D={E}.

| | a | b | c | d | E |
|---|---|---|---|---|---|
| $x_1$ | 1 | 0 | 2 | 1 | 1 |
| $x_2$ | 1 | 0 | 2 | 0 | 1 |
| $x_3$ | 1 | 2 | 0 | 0 | 2 |
| $x_4$ | 1 | 2 | 2 | 1 | 0 |
| $x_5$ | 2 | 1 | 0 | 0 | 2 |
| $x_6$ | 2 | 1 | 1 | 0 | 2 |
| $x_7$ | 2 | 1 | 2 | 1 | 1 |

Let us consider the index key as 'c'. As $x_1,x_2,x_4,x_7$ have the values 2; $x_3,x_5$ have the values 0 and $x_6$ has the value 1. Hence, the partition on U with respect to c can be defined as $\{\{x_1,x_2,x_4,x_7\},\{x_3,x_5\},\{x_6\}\}$.

However, in real time systems we can find several information systems with fuzzy decision attributes and hence the scope of the algorithms discussed above would be applicable for such information systems. Here, the Naïve Bayesian rough indexing of the data can be derived from the fuzzy decision attribute as discussed in the previous section.

For example, consider knowledge representation of the information system with C={a,b,c,d} and D={E} where E is of fuzzy natured.

| | a | b | c | d | $\mu_E(x_i)$ |
|---|---|---|---|---|---|
| $x_1$ | 1 | 0 | 2 | 1 | 0.45 |
| $x_2$ | 1 | 0 | 2 | 0 | 0.7 |
| $x_3$ | 1 | 2 | 0 | 0 | 0.65 |
| $x_4$ | 1 | 2 | 2 | 1 | 0.1 |
| $x_5$ | 2 | 1 | 0 | 0 | 0.91 |
| $x_6$ | 2 | 1 | 1 | 0 | 0.6 |
| $x_7$ | 2 | 1 | 2 | 1 | 0.35 |

On considering 'c' as the index key, the partition obtained is $\{\{x_1,x_2,x_4,x_7\}, \{x_3,x_5\},\{x_6\}\}$. Let $\delta$=0.5. Here, $E[\delta] = \{x_2,x_3,x_5,x_6\}$. For a given $\alpha$ and $\beta$, the Naïve Bayesian indexing algorithm would be implemented further.

**7. CONCLUSION**

In this paper, by using the concept of Naïve Bayesian rough sets the approach of indexing the records of the information system is dealt. These rough indices are useful to analyze and index a database when the fuzzy information about the entire key values is obtained.

**REFERENCES**

[1]. G.Ganesan, C.Raghavendra Rao, Rough Set: Analysis of Fuzzy Sets using thresholds, UGC Sponsored National Conference on Recent Tends in Computational Mathematics, March 2004, Gandhigram Rural Institute, Tamilnadu, by Narosa Publishers, pp:81-87, 2005

[2]. G.Ganesan, D.Latha, C.Raghavendra Rao, Rough Index in Information System with Fuzziness in Decision Attributes, International journal of Fuzzy Mathematics, Vol: 16, No: 4, 2008

[3]. Greco, S., Matarazzo, B. and Slowinski, R. Parameterized rough set model using rough membership and Bayesian con_rmation measures, International Journal of Approximate Reasoning, 49, 285-300, 2009

[4]. Pawlak, Z. Rough sets, International Journal of Computer and Information Sciences, 11, 341-356, 1982.

[5]. Pawlak, Z. Rough Sets, Theoretical Aspects of Reasoning about Data, Dordrecht: Kluwer Academic Publishers, 1991.

[6]. Pawlak, Z. and Skowron, A. Rough membership functions, in: Yager, R.R., Fedrizzi, M. and Kacprzyk, J., Eds., Advances in the Dempster-Shafer Theory of Evidence,John Wiley and Sons, New York, 251-271, 1994.

[7]. Slezak, D. and Ziarko, W. The investigation of the Bayesian rough set model, International Journal of Approximate Reasoning, 40, 81-91, 2005.

[8]. Yiyu Yao and Bing Zhou, Naive Bayesian Rough Sets, Proceedings of RSKT 2010, LNAI 6401, pp. 719-726, 2010.

[9]. Kohavi R, Useful feature subsets and Rough set reducts, Proceedings, Third International Workshop on Rough Set and Soft Computing, pp:310-317, 1994

[10]. Zadeh L.A. Fuzzy Sets, Journal of Information and Control, 8, pp. 338-353, 1965