Cancer data investigation using variable precision rough set with flexible classification

Sanjiban Sekhar Roy
SCSE, VIT University, Vellore
+919843121787
sanjibanroy09@gmail.com

Anupam Gupta
SCSE, VIT University, Vellore
+919597361737
anupamgupta587@gmail.com

ABSTRACT

The theory of Rough set is a new mathematical tool to deal with intelligent data mining proposed by Z. Pawlak. This paper implements the concept of tolerance relation on incomplete information system using variable precision rough set (VPRS) with flexible classification, which is an extension to the classical rough set theory. Here we have analyzed a cancer data set provided by national cancer institute from (1975-2008) and estimated the value of tolerance factor for each country people based on their age, using VPRS with flexible classification model.

Keywords: Rough Set, Tolerance relation, Incomplete information system, Variable precision rough set.

1. INTRODUCTION

Rough set theory[1,2] is based on the concept that every object in the world has some kind of information. It’s foundation based on some attributes and algorithm wise we have to generate some rules and classify objects into different classes and get information from them. Let X is an attribute, T = (U, A) and let \( B \subseteq A \) and \( X \subseteq U \), to get information of B using X by building the lower and upper approximations of X, represented by \( \bar{B}X \) and \( \tilde{B}X \) respectively, where

\[
\bar{B}X = \{x | [x]_B \subseteq X\}, \quad \tilde{B}X = \{x | [x]_B \cap X \neq \emptyset\}
\]

The accuracy of the approximation is given by,

\[
\alpha_x(x) = \frac{\text{card}(\bar{B}(x))}{\text{card}(B(x))}
\]

If \( \alpha_x(x) = 1 \), then X is a crisp set or if \( \alpha_x(x) < 1 \), then X is rough set.

In classical rough set theory the boundary region B of X is given by,

\[
BN_x(X) = \bar{B}X - \tilde{B}X
\]

consists of those objects that we cannot decisively classify in B. A set is called rough if its boundary region is non-empty, otherwise the set is called crisp.

Pawlak used to solve many decision tribulations but is not able to find solutions in the cases where data values are missing. The main idea of variable precision rough set (VPRS)[3] model is to obtain the class information of objects which can not be directly obtained by given values of attribute. VPRS model is an extension

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CCSEIT-12, October 26-28, 2012, Coimbatore [Tamilnadu, India]
Copyright © 2012 ACM 978-1-4503-1310-0/12/10...

Anvesha Sinha
SCSE, VIT University, Vellore
+918870651545
any_04@yahoo.in

Rohit Ramesh
SCSE, VIT University, Vellore
+919840695645
rrrohit992@yahoo.co.in

classical rough set theory[4]. The VPRS with flexible classification model is an slight modification in VPRS model, where some extra features are added have better performance and efficiency in results. This VPRS with flexible classification model is to applied to calculate the tolerance relation for different classes. Based on tolerance relations we calculate the tolerance factor for data set of objects.

1.1 Principle of VPRS (flexible classification)

Our main objective is to apply VPRS with flexible classification [5,6] on an incomplete information system. A data set which is having some uncertain or misclassified objects are to be identified using this model, where a set of attributes are selected then misclassified objects are to be identified among them an at last remaining objects have to be removed from those objects. Hence, we can find a unique decision rate out of them. It is the approximate value which can be used while classifying objects into different decision classes.

1.2 Tolerance Relations in Incomplete Information Systems (iis)

T(B) is symbol for the tolerance relation applied on attribute B. Assume a complete information system P is \( \{U,C\} \), where U is a nonempty finite set of objects or universe of discourse, C is a nonempty finite set of attributes, such that \( a : U \rightarrow V_a \) for any \( a \in C \).

Here the data set is having missing values and are denoted by *.

1.3 Analysis of Tolerance Relation

If \( P=\{U,C,U[\varnothing]\} \) is an incomplete information system then \( T(B) = \{(x,y) \in U \times U \mid x \sim y \} \). Here T(B) is called the tolerance relation applied on attribute B. To find out the tolerance factor for a given data set, first we make the decision classes and divide our data set objects into that decision classes where every class has some tolerance factor based on number of objects belong to that class. Tolerance factor is denoted by \( t(x) \), where \( \varnothing \) is one of the decision class. If a is the number of objects belongs to lower approximation class B and t is the number of objects

472
belong to upper approximation of class B, then tolerance factor of the class B in data set is given as $D_2(B) = d(u/t)$.

2. MISSING VALUE CALCULATIONS IN A DATA SET

If $P = (U, C \cup \{d\})$ is an incomplete information system $x \in U$, and $m(x) = \{c \mid c \in C \land c(x) = \#\}$
here $m(x)$ is the number of lost attributes. Cardinality of $m(x)$ is $|m(x)|$ and is given by[5],

\[
c(x) = 1 - \frac{m(x) + m(y)}{\beta}, x \neq y
\]

Let $T\beta$ is the tolerance class with respect to $B$ and value of $\beta$ ranges from $0 < \beta < 0.5$ which imply
$I_\beta(x) = \{ y \mid y \in T_{B(x)} \land c(x, y) > 1 - \beta \}$. Here, $B - \beta$ is upper approximation and $B + \beta$ is an lower approximation.

3. APPLYING VPRS MODEL WITH TOLERANCE RELATION ON AN DATA SET

Let $I = (U, A, F)$ is an incomplete information system.
For any $B \subseteq A$, the tolerance relation $TR(B)$ proposed by M.Kryskiewicz is defined as follows:
$TR(B) = \{(x, y) \in U \times U \mid a \in A, a(x) = a(y) \land a(x) = \# \}$
and the tolerance class of $x$ is defined as follows:
$TRB(x) = \{ y \in U \mid (x, y) \in TR(B) \}$

An improved tolerance relation definition as follows:
$TR = \{(x, y) \in U \times (U \setminus IRB(x, y)) \}$

Where $i$ is a determinate possibility threshold and $0 \leq i$. $RB(x)$ which represents the tolerance degree is the joint possibility of objects $x$ and $y$ having same values in attribute set $B$. Here the tolerance class of $x$ can be defined as:
$TR - B(x) = \{(y \in U \mid (x, y) \in TR(B) \}$

Here, $i$ must be determined by experience. It’s inconvenient or not objective.
Here, health is a decision attribute. Hence applying rough approximation we got the lower approximation as $30,50,50$; upper approximation is $35-50,18-27,27-35$ and the objects under boundary value would be $18-27,27-35$. Next we would apply VPRS with flexible classification to a cancer data set taken from National Cancer Institute, USA [4] which consists of all aged people ranging from 10, 20 and 30 years and the probability of being diagnosed from cancer and dying from cancer is given in table given below. Applying VPRS with variable precision relation we will find the tolerance factor for each count.

### Table 1: National Cancer Institute, USA

<table>
<thead>
<tr>
<th>Country People</th>
<th>Current Age</th>
<th>Diagnosing From Cancer</th>
<th>Dying From Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Races</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>0.01</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>70</td>
<td>0.02</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>0.02</td>
<td>*</td>
</tr>
<tr>
<td>White</td>
<td>0</td>
<td>0</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>0</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>0</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>0.01</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>70</td>
<td>0.03</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>0.03</td>
<td>*</td>
</tr>
<tr>
<td>Black</td>
<td>0</td>
<td>0</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>0</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>0</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>70</td>
<td>0.02</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>0.02</td>
<td>*</td>
</tr>
<tr>
<td>Asian/Pacific Islander</td>
<td>0</td>
<td>0</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>0</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>0</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>70</td>
<td>0.02</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>0.02</td>
<td>*</td>
</tr>
<tr>
<td>American Indian/Alaska native</td>
<td>0</td>
<td>0</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>0</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>0.02</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>70</td>
<td>0.06</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Hispanic</td>
<td>0</td>
<td>0</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>0</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.02</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>0.03</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>70</td>
<td>0.03</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>0.03</td>
<td>0.06</td>
</tr>
</tbody>
</table>
We are taking \( \beta \) value ranging from \( 0 < \beta < 0.5 \). In above data set, we have to find lower and upper approximation based on VPRS model. We made 2 classes based on death of patients from cancer. It would be based on tolerance relation \( \beta \) which is the calculation factor or range of approximation based on values of attributes. While calculating [7,8,9,10] the missing attribute values, we noticed that there are two attribute missing, for \( x = \text{age} + 20\text{yrs}, y = \text{age} + 30\text{yrs}; m(x) = 6; \ m(y) = 12; \ c = 12. \)

Hence, calculating the value of \( C(x,y) \) we have the result as 0.917. So, here unique decision rate is vary high, therefore, taking average value of other attributes would not effect our result to much extent.

Next, we have assumed that among two classes, class A and class B referred as:

Class A: objects who will die from cancer

Class B: objects who will be less effected from cancer

Subsequently, we have considered only class A and each time calculated tolerance value \( D_c \), lower approximation value namely \( A^L \) and upper approximation value \( A^U \) for class A while considering the ages of different races.

Calculation for all races:

\[
\begin{align*}
I_B(0) &= [0,10,20] & I_B(10) &= [0,10,20] \\
I_B(20) &= [0,10,20] \\
I_B(30) &= [30] & I_B(40) &= [40] & I_B(50) &= [50] \\
I_B(60) &= [60] & I_B(70) &= [70] & I_B(80) &= [80] \\
\end{align*}
\]

Lower approximation of ages \( A^L = \{30,40,50,60,70\} \), upper approximation of ages \( A^U = \{0,10,20,30,40,50,60,70\} \)

Hence, \( D_c(A) = 3/8 \)

Calculation for white race:

\[
\begin{align*}
I_B(0) &= [0,10,20] & I_B(10) &= [0,10,20] \\
I_B(20) &= [0,10,20] \\
I_B(30) &= [30] & I_B(40) &= [40] & I_B(50) &= [50] \\
I_B(60) &= [60] & I_B(70) &= [70] & I_B(80) &= [80] \\
\end{align*}
\]

Lower approximation of ages \( A^L = \{30,40,50,60,70\} \), upper approximation of ages \( A^U = \{0,10,20,30,40,50,60,70\} \)

Hence, \( D_c(A) = 5/8 \)

Calculation for white black race:

\[
\begin{align*}
I_B(0) &= [0,10,20] & I_B(10) &= [0,10,20] \\
I_B(20) &= [0,10,20] \\
I_B(30) &= [30] & I_B(40) &= [40] & I_B(50) &= [50,60] \\
I_B(50) &= [50,60] \\
I_B(70) &= [70] & I_B(80) &= [80] \\
\end{align*}
\]

Lower approximation of ages \( A^L = \{30,40,50,60,70\} \), upper approximation of ages \( A^U = \{0,10,20,30,40,50,60,70\} \)

Hence, \( D_c(A) = 3/8 \)

Calculation for asian/pacific islander race:

\[
\begin{align*}
I_B(0) &= [0,10,20] & I_B(10) &= [0,10,20] \\
I_B(20) &= [0,10,20] \\
I_B(30) &= [30] & I_B(40) &= [40] & I_B(50) &= [50] \\
I_B(60) &= [60] & I_B(70) &= [70] & I_B(80) &= [80] \\
\end{align*}
\]

Lower approximation of ages \( A^L = \{60\} \), upper approximation of ages \( A^U = \{0,10,20,30,40,50,60,70\} \)

Hence, \( D_c(A) = 8/9 \)

Calculation for american indian/alaska native:

\[
\begin{align*}
I_B(0) &= [0,10,20] & I_B(10) &= [0,10,20] \\
I_B(20) &= [0,10,20] \\
I_B(30) &= [30] & I_B(40) &= [40] & I_B(50) &= [50] \\
I_B(60) &= [60] & I_B(70) &= [70] & I_B(80) &= [80] \\
\end{align*}
\]

Lower approximation of ages \( A^L = \{0,10,20,30,40,50,60,70\} \), upper approximation of ages \( A^U = \{0,10,20,30,40,50,60,70\} \)

Hence, \( D_c(A) = 8/9 \)

Calculation for hispanic race:

\[
\begin{align*}
I_B(0) &= [0,10,20] & I_B(10) &= [0,10,20] \\
I_B(20) &= [0,10,20] \\
I_B(30) &= [30] & I_B(40) &= [40] & I_B(50) &= [50] \\
I_B(60) &= [60] & I_B(70) &= [70] & I_B(80) &= [80] \\
\end{align*}
\]

Lower approximation of ages \( A^L = \{40,50,60\} \), upper approximation of ages \( A^U = \{0,10,20,30,40,50,60,70\} \)

Hence, \( D_c(A) = 3/8 \)

Calculation for black people:

\[
\begin{align*}
I_B(0) &= [0,10,20] & I_B(10) &= [0,10,20] \\
I_B(20) &= [0,10,20] \\
I_B(30) &= [30] & I_B(40) &= [40] & I_B(50) &= [50,60] \\
I_B(50) &= [50,60] \\
I_B(70) &= [70] & I_B(80) &= [80] \\
\end{align*}
\]

Lower approximation of ages \( A^L = \{50,60,70\} \), upper approximation of ages \( A^U = \{0,10,20,30,40,50,60,70\} \)

Hence, \( D_c(A) = 3/8 \)

Calculation for native american

\[
\begin{align*}
I_B(0) &= [0,10,20] & I_B(10) &= [0,10,20] \\
I_B(20) &= [0,10,20] \\
I_B(30) &= [30] & I_B(40) &= [40] & I_B(50) &= [50,60] \\
I_B(50) &= [50,60] \\
I_B(70) &= [70] & I_B(80) &= [80] \\
\end{align*}
\]

Lower approximation of ages \( A^L = \{50,60,70\} \), upper approximation of ages \( A^U = \{0,10,20,30,40,50,60,70\} \)

Hence, \( D_c(A) = 3/8 \)

So finally tolerance factor of total data set \( D_c(A) = \text{all objects coming under lower approximation} / \text{objects coming under upper approximation} = 21/39 \)

4. CONCLUSION

In this paper, we have implemented the concept of VPRS with flexible classification model on an incomplete cancer data set. The tolerance value of American Indian/Alaska native country people was set up maximum. We calculated missing value of risk of being diagnosed from cancer for country people of different ages. We conclude that VPRS with flexible classification model is much useful mathematical tool for calculation of tolerance value along with lower and upper approximations from incomplete information system.
5. ACKNOWLEDGMENTS
We would akin to show our appreciation to the higher-ranking administration of VIT University (Vellore, India) for their kind shore up and encouragement towards our research work.

6. REFERENCES


