OPEN ACCESS

International
Journal
Of Modern Engineering Research (IJMER)

Survival Rate of Patients of Ovarian Cancer: Rough Set Approach

Kamini Agrawal[1], Pragati Jain[2]
[1]*Department of Applied Mathematics, IET, Indore, India*
[2]*Department of Mathematics, SPIPS, Indore, India*

ABSTRACT: *Ovary cancer is becoming a record cause of death among women in the whole world. The survival of ovary cancer patients cannot be exactly predicted how long they will alive. In this paper, Rough Set Theory is used for feature extraction and rule generation. Experiments conducted on the database of ovary cancer patients who were operated for ovary cancer 10 years before. An experimental result shows that the type of surgery has a significant role in the survival of patients.*
Keywords: *Ovary Cancer, Rough Set, Johnson's Heuristic Algorithm, Decision Rule.*

I. INTRODUCTION

Cancer is one of the leading diseases in all over the world. Cancer diseases are developed by the abnormal cells that divide uncontrollably and destroy body tissues. Cancers are various types like Breast Cancer, Bone Cancer, Ovarian Cancer, Liver Cancer, Lung Cancer etc. Ovarian cancer ranks fourth among cancer deaths in women. This high death rate is due to the fact that almost 70% of women with epithelial ovarian cancer are not diagnosed until the disease gets advanced to Stage III (i.e. cancer has spread to the upper abdomen) or beyond [1]. The Multi Class Support Vector Machine was used in the previous researches for classifying the normal, early stage and late stage cases as an optimized attribute set. The optimized feature selection process was achieved by the hybrid Particle Genetic Swarm Optimization (PGSO) based rough set theory [2]. Different types of diagnosis are done for detecting cancer disease. After surgery, the survival of patient is unpredictable because some patients survive so many years and some for few years or months. For the prediction of survival rate of patients, Rough Set Theory is used as one of the tools.

The theory of Rough Sets (RS) proposed by Zdzislaw Pawalk in 1982 [3]. It is a mathematical tool for extracting knowledge from uncertain and incomplete data. It gives the optimal results for the analysis of data without loss of information. The advantage of this approach is that, it does not require the user to make any prior assumptions about the data. It can be used to find attributes efficiently; a critical step in the decision-making process, through the computing of reducts [4].

II. FUNDAMENTALS OF ROUGH SET THEORY (RST)

An information system (IS) is a pair *(U,A)*, where *U* is a nonempty, finite set of objects called the Universe and *A* is a nonempty, finite set of Attributes, such that $f_a : U \to V_a$ for any $a \in A$, where V_a is the set of values of *a*. If $R \subseteq A$, there is an associated Equivalence Relation, called as Indiscernibility Relation: $IND(R) = \{(x_i, x_j) \in U, b(x_i) = b(x_j), \forall b \in R\}$, where all identical objects of set are considered as elementary sets of the universe. The Partition determined by R, denoted as *U*/*IND*(*R*), or simply *U*/R. Let $X \subseteq U$, the R-lower approximation $R_*(X)$ and R-upper approximation $R^*(X)$ of set X are define as [5, 6]:

$$R_*(X) = \{x_i \in U : R(x_i) \subseteq X\}$$

$$R^*(X) = \{x_i \in U : R(x_i) \cap X \neq \phi\}$$

The difference between Upper and Lower Approximation is called as Boundary Region:

$$BN_R(X) = R^*(X) - R_*(X)$$

If the Boundary Region of *X* is the empty set, i.e., $BN_R(X) = \emptyset$, then the set *X* is *crisp* (*exact*) with respect to R; in the opposite case, i.e., if $BN_R(X) \neq \emptyset$, the set *X* is *rough* (*inexact*) with respect to R. Rough set can be also characterized numerically by the following coefficient:

$$\alpha_R(X) = \frac{card(R^*(X))}{card(R_*(X))}$$

which is called as *accuracy of approximation*. Obviously $0 \leq \alpha_R(X) \leq 1$. If $\alpha_R(X) = 1$, X is *crisp* with respect to R (X is *precise* with respect to R), and otherwise, if $\alpha_R(X) < 1$, X is *rough* with respect to R (X is *vague* with respect to R).

The reduct process for attributes reduces elementary set numbers, the goal of which is to improve the precision of decisions. After the attribute dependence process, the reduct attribute sets are generated to remove superfluous attributes. The complete set of attributes is called a reduct attribute set. There may be more than one reduct attribute set in an information system, but intersecting a number of reduct attribute sets yields a core attribute set.

$$RED(R) \subseteq A$$
$$CORE(R) = \bigcap RED(R)$$

2.1. Johnson's Heuristic Algorithm

The Johnson's Heuristic Algorithm is used to calculate reduct for a decision problem. It sequentially selects features by finding those that are most discernible for given decision feature [7]. It computes a discernibility matrix M, where

$$m_{ij} = \{\{f \in F_R : f(c_i) \neq f(c_j)\} \text{ for } f_d(c_i) \neq f_d(c_j), \text{and } \phi \text{ otherwise}\}$$

Johnsons _Reduct (F_p, F_d, C)
Input F_p: conditional features, f_d: decision feature, C: cases
Output R: Reduct $R \subseteq F_p$

1. $R \leftarrow \phi, F' \leftarrow F_p$.
2. $M \leftarrow$ Compute Discernibility Matrix (C, F', f_d)
3. Do
4. $f_h \leftarrow$ Select Highest Scoring Feature (M)
5. $R \leftarrow R \cup \{f_h\}$
6. For $(i = 0 \, to \, |C|, j = i \, to \, |C|)$
7. $m_{ij} \leftarrow \phi$ if $f_h \in m_{ij}$
8. $F' \leftarrow F' - \{f_h\}$
9. Until $m_{ij} = \phi \ \forall i, j$
10. Return R

For the standard Johnson's Algorithm, this is typically a count of the number of appearances an attribute makes within clauses; attributes that appear more frequently are considered to be more significant. The attribute with the highest heuristic value is added to the reduct candidate, and all clauses in the discernibility function containing this attribute are removed. As soon as all clauses are removed, the algorithm terminates and returns to reduct R.

2.2. Decision Rules

Let S= (U, P, Q) be a decision table. Every $x \in U$ determines a sequence $P_1(x).P_2(x),\ldots\ldots P_n(x)$, $Q_1(x), Q_2(x) \ldots\ldots\ldots, Q_n(x)$

where $\{P_1(x), P_2(x), \ldots\ldots\ldots\ldots P_n(x)\} = P$ and $\{Q_1(x), Q_2(x), \ldots\ldots\ldots\ldots Q_n(x)\} = Q$

The sequence is called as Decision Rule induced by x (in S) and denoted by $P_1(x), P_2(x), \ldots\ldots\ldots, P_n(x) \rightarrow Q_1(x), Q_2(x), \ldots\ldots\ldots, Q_n(x)$ or $P \rightarrow_x Q$.

The number $supp_x(P,Q) = |P(x) \cap Q(x)|$ called as Support of Decision Rule $\Phi \rightarrow_x \Psi$. The number,

$$\sigma_x(P,Q)=\frac{supp_x(P,Q)}{|U|}$$

is referred to as the Strength of the decision rule $P \rightarrow_x Q$, where $|X|$ denotes the cardinality of X. With every decision rule $P \rightarrow_x Q$, the certainty factor is denoted by $cer_x(P,Q)$ and defined as follows:

$$cer_x(P,Q)=\frac{supp_x(P,Q)}{|P(x)|}$$

If $cer_x(P,Q) = 1$, then $P \rightarrow_x Q$ is called as Certain Decision *R*ule; If $0 < cer_x(P,Q) < 1$ the decision rule is referred to as an uncertain decision rule.

Besides, a Coverage Factor of the decision rule, denoted as $cov_x(P,Q)$ and defined as

$$cov_x(P,Q)=\frac{supp_x(P,Q)}{|Q(x)|}$$

If $P \rightarrow_x Q$ is a decision rule then $P \rightarrow_x Q$ is called an inverse decision rule. The inverse decision rules are further used to give explanations *(*reasons*)* for a decision.

III. DESCRIPTION OF DATASET

The treatments for ovarian cancer are surgery, chemotherapy, biological therapy or radiotherapy, the type of treatment depends on the stage of the cancer. Sometimes both surgery and chemotherapy treatment are applied for advance stage of cancer. Advance stage means the cancer has spread away from the ovary. In surgery treatment surgeon removes as much of the cancer as possible. Chemotherapy applied before the surgery. It shrinks the cancer and make it easier to remove. In radiotherapy, X-rays are used to kill or damage cancer cell and reduce their activity [8].

The dataset of the Ovary Cancer disease is taken from Obel (1975) who studied on women, who were operated for ovary cancer 10 years before. The dataset consists of 5 attributes where 3 are conditional attribute, one is decision attribute and one is frequency attribute [9, 10, 11]. Table I shows the notation of ovary cancer attributes with their description.

Table I: Ovary Cancer Database Attributes Description

Notation	Description
A1	Stage of the cancer at the time of operation (e = early, a = advanced).
A2	Type of operation (r = radical, l = limited)
A3	X-ray treatment was received (yes, no).
D	Survival status after 10 years (yes, no).
F	Frequency

IV. EXPERIMENTAL RESULTS AND DISCUSSION

Table II represents the three factors affecting ovary cancer survival that concern fourteen cases. *U* is a universal set of objects. In table, columns $C=\{A_1, A_2, A_3\}$ represent conditional attributes and column *D* represents decision attribute. The elementary set of three attributes is $U/C=\{\{X_1, X_3\},\{X_2, X_4\},\{X_5, X_7\},\{X_6\},\{X_8, X_{10}\},\{X_9, X_{11}\},\{X_{12}\},\{X_{13}, X_{14}\}\}$ and equivalence class of decision attribute is $Y=\{Yes, No\}$.

Table II: Database of Ovary Cancer

U	***A1***	***A2***	***A3***	***D***	***F***
X1	e	r	No	No	10
X2	e	r	Yes	No	17
X3	e	r	No	Yes	41
X4	e	r	Yes	Yes	64
X5	e	l	Yes	No	3
X6	e	l	No	Yes	13

X7	e	l	Yes	Yes	9
X8	a	r	No	No	38
X9	a	r	Yes	No	64
X10	a	r	No	Yes	6
X11	a	r	Yes	Yes	11
X12	a	l	No	No	3
X13	a	l	Yes	No	13
X14	a	l	Yes	Yes	5

Thus, the Lower approximation is $C_*(yes)=\{X_6\}$

Upper approximation of set is

$C^*(yes)=\{X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}, X_{13}, X_{14}\}.$

Boundary region of set is $BNB(yes)=\{X_1, X_2, X_3, X_4, X_5\ X_7, X_8, X_9, X_{10}, X_{11}, X_{12}, X_{13}, X_{14}\}.$

For finding the Reduct, the Johanson Heruristic Algorithm is used. The set of attributes thus obtained is $\{A_1, A_2, A_3\}$. The decisions rules are generated from this reduct are as follows:

1. $((A_1, e)\wedge(A_2, l)\wedge(A_3, no))\rightarrow(D, yes)$
2. $((A_1, a)\wedge(A_2, l)\wedge(A_3, no))\rightarrow(D, no)$
3. $(A_2, r)\rightarrow(D, yes)$
4. $(A_2, r)\rightarrow(D, no)$
5. $(A_3, yes)\rightarrow(D, yes)$
6. $(A_3, yes)\rightarrow(D, no)$

The certainty and coverage factors of the decision rules are given is Table III.

Table III: Certainty and Coverage factors

Rules	Support	Strength	Certainty	Coverage
R1	13	0.029	1	0.09
R2	3	0.007	1	0.021
R3	122	0.269	0.48	0.82
R4	129	0.285	0.51	0.87
R5	89	0.196	0.48	0.59
R6	97	0.214	0.52	0.65

From the certainty and coverage factors of all decision rules, It is observed that the patient could survive after 10 years of surgery if the type of surgery is limited and without X-ray treatment in early or advance stage of ovarian cancer. The patient would be difficult to survive after 10 years, if the surgery is radical type or receives X-ray treatment.

V. CONCLUSION

In this paper, rough set theory approach is used to predict the survival rate of patients after the 10 years treatment of ovary cancer. Johnson Heuristic Algorithm is used for feature selection of ovary cancer survival. Decision rules are generated on the basis of the features. Certainty and coverage factors are calculated on the basis of decision rules. Both are helpful to predict the survival of patients after 10 years of surgery. It is concluded that patients survive after 10 years, if the surgery is not radical type.

REFERNCES

[1]. T. Z. Tan, C. Quek and G.S. Ng, Ovarian Cancer Diagnosis by Hippocampus and Neocortex – Inspired Learning Memory Structures, Neural Networks, 18, 2005, 818–825.

[2]. P. Yasodha and N. R. Anathanarayanan, Analysis Big Data to Build Knowledge Based System For Early Detection of Ovarian Cancer, Indian Journal of Science and Technology,8, 2015, 1-7.

[3]. Z. Pawlak, Rough Sets, International Journal of Computer and Information Sciences, 11, 1982, 341–356.

[4]. H. Bai, Y. Ge, J. Wang and Y. L. Liao, Using Rough Set Theory to Identify Villages Affected by Birth Defects: The Example of Heshun, Shanxi, China, International Journal of Geographical Information Science, 24, 2010, 559–576.
[5]. Q. Shen and R. Jensen, Rough Sets, Their Extensions and Application, International Journal of Automation and Computing, 4, 2007, 100-106.
[6]. B. Walczak, D.L. Massart, Rough Set Theory, Chemometrics and Intelligent Laboratory Systems, 47, 1999, 1–16.
[7]. A. H. El-Baz, Hybrid Intelligent System Based Rough Set and Ensemble Classifier for Breast Cancer Diagnosis, Neural Computing and Applications, 26, 2015, 437–446.
[8]. http://www.cancerresearchuk.org/about-cancer/type/ovarian-cancer/treatment/which-treatment-for-ovarian-cancer.
[9]. E. B. Obel, A Comparative Study of Patients with Cancer of the Ovary Who Have Survived More or Less Than 10 Years. Acta Obstetricia et Gynecologica Scandinavica, 55, 1975, 429-439.
[10]. E. B. Andersen, The Statistical Analysis of Categorical Data. 2nd edition, (Berlin, Springer-Verlag, 1991), 192-193.
[11]. https://vincentarelbundock.github.io/Rdatasets/datasets.html