Feature Selection Algorithm Using Fuzzy Rough Sets for Predicting Cervical Cancer Risks

Ms. J. Vandar Kuzhali (Corresponding author)
Senior Lecturer / Department of MCA & M.Sc.(S.E.), Erode Sengunthar Engineering College
Thudupathi, Erode(Dt), Tamil Nadu, INDIA – 638 057
Tel: 91-42-9423-2701 Fax: 91-42-9423-2705 E-mail: vandarkuzhali@yahoo.com

Dr. G. Rajendran
Professor & Head / Department of Mathemtics, Kongu Engineering College
Perundurai, Erode(Dt), Tamil Nadu, INDIA – 638 052
Tel: 91-42-9422-6645 Fax: 91-4294-220-087 E-mail: rajendranjv@gmail.com

Mr. V. Srinivasan
Senior Lecturer / Department of MCA, Velalar College of Engineering & Technology
Thindal, Erode(Dt), Tamil Nadu – 638 011
Tel: 91-424-227-0571 Fax: 91-424-243-1725 E-mail: newsrini@rediffmail.com

Mr. G. Siva Kumar
Senior Lecturer / Department of CSE, Erode Sengunthar Engineering College
Thudupathi, Erode(Dt), Tamil Nadu, INDIA – 638 057
Tel: 91-4294-232-701 Fax: 91-4294-232-705 E-mail: gsiva_g@yahoo.com

Abstract

Early detection or prediction is very important to reduce the fatalities of Cervical Cancer. Cancer cells affect the Cervix area initially, and then it will spread near by parts. A method using Fuzzy Rough sets is used to analyze the demographic dataset and identify the risk of Cervical Cancer. This method integrates Entropy, Information Gain (IG) and Fuzzy Rough sets for identifying the risk of Cervical Cancer earlier. Risk Factors are identified by IG. Rules are extracted by Fuzzy Rough sets. These rules can be used to identify the risk of Cervical Cancer efficiently than the decision trees. It is found that Human Papilloma Virus (HPV) and having Multiple Sexual Partners (MP) are the major risk factors increase the chances of affecting this cancer. If all the above factors are high the risk of affecting Cervical Cancer is high. Result of this paper will help to improve the clinical practice guidance for analyzing the risk of Cervical Cancer.

Keywords: Cervical cancer, Entropy, Information gain, Fuzzy rough sets, Demographic data, Feature selection

1. Introduction

Cervical Cancer is the commonest cancer in Indian Women and occupies the top rank among cancers in women. In most developing countries Cervical Cancer constitutes 34% of all women cancers. In India the incidence of this cancer in females is 100 000 / year. It is the 1/5 of the world cancer occurrence among women (Report of WHO consultation, 2009). HPV is the main factor that spreads and causes Cervical Cancer through sexual intercourse (Muñoz N et al, 2002 ; Kahn J. A et al, 2007). The co-factors are Multiple Sexual partners, Very younger age at first sexual act, Husband's extra family affairs, Low socio-Economic factors, Parity & Prolong use of Oral Contraceptive pills.

The American Cancer Society confirmed that the risk of occurring cervical cancer is low in women who have never experience in sexual intercourse. Waiting to have sex until the women is older can help to avoid HPV. It also helps to limit the number of sexual partners and to avoid having sex with someone who has had many other sexual partners. Women those with four full-term pregnancies (Parity) are having the high risk of developing cervical cancer (M. Klitsch, 2002) . There is a potential long term relationship between prolonged use of oral contraceptives and development of Cervical Cancer (Moreno V et al, 2002). Low socio-economic status (SES) is

recognized as a risk factor for many health problems, including cervical cancer, particularly in low-resource settings. Women with low SES often have limited income, restricted access to health care services, poor nutrition, and a low level of awareness about health issues and preventive behavior. All of these factors can make them more vulnerable to illness and preventable diseases such as cervical cancer (Ann L. Coker et al, 2006).

The most common treatments are Hysterectomy, radiation therapy & chemotherapy. Surgery involves removing the uterus and nearby reproductive organs such as the fallopian tubes and ovaries. Lymph nodes near the tumor also may be removed during surgery to see if they contain cancer. After the treatment is finished, most women can lead normal lives. If their uterus was removed, however, they can no longer bear children. This often is not an issue for women in their fifties and sixties, but younger women in their twenties, thirties, and forties may find it hard to adjust to this reality.

The new treatment methods in the field of cancer are introduced everyday. But, the decision making is the complex practice should be done with extreme care and conscious. Sometimes making decisions with intuitive thinking may lead to wrong diagnosis and treatment. Methodological decision making is unfailing and will be the base for the decisions. Instead of identifying the stages and development of cancer, it is important to identify the risk possibility of Cervical Cancer for prevention.

This paper mainly concentrates on identifying the possible risk factors of cervical cancer. This was tested under a group of sample data sets. This is the main aim of this work. In the following sub-divisions the different processes involved in the system are given and the preliminary results were shown with sample data.

2. Review of Current Diagnosing Systems

The Pap smear has been the main test for number of years. But, the sensitivity and the specificity are not high. The next highly used technique is Colposcopy. It is the microscopic observation of the Cervix. In this test, Cervix is examined with low and high magnification. Acetic Acid and Lugol solution is applied in the Cervix for differentiating the normal cells and cancerous cells. The visual analysis of Colpscopic image is based on the color variations. This test is mainly used to assess the size, location and the distribution of the lesion. Colposcopy can determine not only the cancer, but also where the tumor is. But Colposcopy requires more experience for the correct analysis of the asymptomatic woman and recognizing the areas of biopsy. This method of test needs long-term experience and more training to get skill in Cancer cell pattern recognition. In these methods of tests clinicians may also add their intuitive decision making than analytical decision making. Additionally the methods for prevention or Early Detection are urgently needed.

In our method pure analytical decision making is done with Entropy and IG as the base, followed with Fuzzy Rough Sets, set of rules are framed. By applying this method one can identify the possible risk of Cervical Cancer from the Demographic factors. The main aim of this method is to prevent the Cervical Cancer.

3. Related Work

Decision trees are used to classify the objects. It is a structure that can be used to divide up a large collection of records into successively smaller sets of records by applying a sequence of simple decision rules. A decision tree model consists of a set of rules for dividing a large heterogeneous population into smaller, more homogeneous groups with respect to a particular target variable. These trees are like binary trees. They can be un-even in depth. It is useful to show the proportion of the data in each of the desired classes. Though they are good in many ways, the output must be categorical one. It is also limited to one output attribute. Decision trees are unstable. Trees created from numeric data sets are complex. (Quinlan J. R., 1986)

Here an algorithm named as Feature Selection Algorithm based on Fuzzy Rough Sets predicting Cervical Cancer risk is introduced in this paper. This will filter the irrelevant or noisy attributes from the data set. So, the prediction will be made easily. A major advantage of information theory is its nonparametric nature. Entropy does not require any assumptions about the distribution of variables. Also it does not assume a linear model. It can be applied on categorical time series data. After calculating Entropy and IG, a Fuzzy Rough set analysis is applied instead of Decision Trees for efficiency.

4. Methods

4.1 Feature Selection

Feature selection is a technique used to reduce the number of features before applying any algorithms to produce better results. Irrelevant features may have negative effects on a prediction task. Moreover, the computational complexity of a classification algorithm may suffer from the curse of dimensionality caused by several features. When a data set has too many irrelevant or noisy variables and only a few examples, over-fitting is likely to

occur. In addition, data are usually better characterized using fewer variables. Here for feature selection process Entropy and IG are used.

4.2 Entropy **:**

It is a measure of variability in a random variable. It is a measure of how pure or impure a variable is.

$$Entropy(S) = \sum - P(I) \log_2 P(I) \quad (1)$$

S- Collection of Samples.

c- Set of outcomes.

P(I) – Proportion of S to the class I.

*4.3 Information Gain***:**

The information gain is based on the decrease in entropy after a dataset is split on an attribute. First the attribute that creates the most homogeneous branches are identified.

$$Gain(S, A) = Entropy(S) - \sum ((|S_y| / |S|) * Entropy(S_y)) \quad \text{(Ross Quinlan, 1993)} \quad (2)$$

4.4 Fuzzy Rough Set:

A Rough set is a formal approximation of a Crisp set, in terms of a pair of sets which give lower and upper approximation of the original set. The Lower and Upper approximation sets are crisp sets.

This is the mathematical tool to process the uncertain knowledge. A knowledge representation system is defined as K = (U, A) where, $X \subseteq U$. U is a non-empty finite set of objects. A is the finite set of primitive attributes. R is an equivalence relation defined on U. U/R indicates the partition of R on U. An ordered pair (U, R) is called the approximation space and any subset is called a concept. Each concept X can be defined as Lower and Upper approximation. (Z. Pawlak, 1982; Z. Pawlak, 1991)

The target set X can be approximated using only the information contained with in P by constructing the P-lower approximations of X.

$$\underline{P}X = \{x / [x]_P \subseteq X\} \quad (3)$$

It is the union of all the equivalence classes in $[x]_P$ which are contained by the target set.

5. Proposed Method & Experiment

A sample fuzzified data set is given to show the proposed method. This method is used to extract decision rules to find the risk of Cervical Cancer. Each patient record contains the set of attributes and one decision attribute specifies the risk of Cervical Cancer.

5.1 Membership Degree

The membership function of a fuzzy set represents the degree of truth as an extension of valuation. For any set X, a membership function on X is any function from X to the real unit interval [0,1]. It is represented as μ_A. $\tilde{A}$ is the fuzzy set. $\mu_A(x)$ is the membership degree of x in the fuzzy set. $\mu_A(x)$ computes the grade of membership of the element x to the fuzzy set $\tilde{A}$. If x is a member of fuzzy set, then the value is 1, other wise 0. (. L. Zadeh, 1965 ; Goguen J. A. ,1967) This can be defined as

$$\mu_A(x) = \begin{cases} 1, & if \quad x \in A \\ 0, & if \quad x \notin A \end{cases}$$

This is shown in figure I.

5.2 Fuzzification

Initially the data set is represented using a function called membership function for mapping the elements according to the degree of membership. Here the quantitative value is transferred to fuzzy sets. Two linguistic

terms used here Y(Yes) & N(No). These two membership values are produced for each attribute according to the membership functions. The sample fuzzified result is shown in Table I. This is done for the calculation efficiency.

5.3 Feature Selection Algorithm to Predict Cervical Cancer

1) Find Entropy from the fuzzified dataset.

Entropy (S) =Σ – P(I)$\log_2$ P(I).

2) Calculate IG. Gain(S,A) = Entropy(S) - Σ((|S_v|/|S|)* Entropy(S_v))

3) If Gain (attribute(i)) > Threshold value Select attribute(i) for further processing. Other wise discard it.

4) Find the Equivalence Class. Equiv(i) = Collection of attributes with similar membership values.

5) Generate Discerning Matrix from the Equivalence class. D=a or b or c or avb or avc or bvc or avbvc if Equiv(i) ≠ Equiv(j). D=X(Null) otherwise.

6) Find the Reduct set.

7) Extract Rules for Disease analysis.

Entropy is a mathematical method of study and used here for analyzing the risk of Cervical Cancer. It is used to find the initial result of the total clinical sample data set that is used for further analysis in IG. Entropy is the base used to find the initial result of the total dataset. This Entropy is applied for positive examples and Negative examples in an attribute set. For example 'High' Risk of HPV comes under the group of positive and the 'Low' Risk of HPV is in Negative.

Based on the Entropy results of each attribute or factor, IG is calculated. Each IG is compared with the threshold value. IG values which produce a higher gain than the threshold value are taken as major risk factors of the Cervical Cancer. Entropy and IG values range from 0 to 1. If all factors of the sample data set belong to same class i.e. YES, the value of Entropy is 0.

By considering the resultant major Risk Factors, Equivalence classes were created. Positive region is built by using Lower Approximations. Decision tables are generated for extracting the rules. Instead of making intuitive decisions, this analytical decision making will assist the clinicians efficiently. Intuitive decision making may lead to over treatment for an asymptomatic woman. Using this method, women with the possibility of high risk of Cervical Cancer can be identified, screened and advised for biopsy.

The sample data set after IG calculation is shown in Table II. This is named as Feature Selection. Here, Entropy for the total data set is calculated. Entropy(S) = 0.97104. IG for each attribute is calculated. It is shown in table III. Threshold value is fixed. The factors that are exceeding the threshold value are considered as the Major Risk Factors. It is shown in figure II. These are taken for creating Equivalence classes. It is shown in Table IV.

From the above table discerning matrix is built from the equivalence classes. HPV is assigned as the factor a, MP is assigned as b and Low SES is c. A 5 X 5 matrix is built. Discern = $(D)_{5X5.}$ Here if Equiv-i ≠ Equiv-j then they can be considered in the discerning matrix. The result is shown in Table V. Reduct set of Table V is shown in Table VI. From Table VI rules can be extracted in the IF-THEN format.

From the table VI, the following rules are extracted.

(i) IF HPV Risk = HIGH AND MP>1 THEN Risk of Cervical Cancer=HIGH

(ii) IF HPV Risk = LOW AND MP>1 THEN Risk of Cervical Cancer = HIGH

(iii) IF HPV Risk = HIGH AND MP>1 AND Low SES = YES THEN

Risk of Cervical Cancer = HIGH

(iv) IF HPV Risk = HIGH AND MP=1 THEN Risk of Cervical Cancer = LOW

(v) IF Low SES = Yes THEN Risk of Cervical Cancer = LOW

While using decision trees, they can produce only binary results. More over, they induce the sequential results. Some times class overlap problem may occur. Decision trees are having complex production rules. Also, a decision tree can be sub-optimal. A sample decision tree is given in Figure III.

6. Conclusion

In this paper an algorithm is suggested for predicting the risk of Cervical Cancer. Our result shows that the factors HPV, Multiple partners and Low SES are the major factors that will drive to Cervical Cancer. Extracting rules from fuzzy rough sets are producing better results than the decision trees. Studies have reported that the women with a Lower SES are having the risk of affecting Cervical Cancer. Diagnosing Cervical Cancer with the help of symptoms may some times lead to wrong decisions and over treatment. But, this method of detecting the risk of Cervical Cancer will surely be an aid for Clinicians with High Sensitivity and Specificity.

References

Ann L. Coker, Xianglin L. Du, Shenying Fang and Katherine S. Eggleston. (2006). Socio Economic Status and cervical cancer survival among older women: Findings from the SEER–Medicare linked data cohorts, *Gynecologic Oncology,* Volume 102, Issue 2.

A Report of a WHO consultation. (2009). *about Cervical Cancer in Developing Countries.*

Goguen J. A. (1967). "L-Fuzzy Sets", *Journal of Mathematical Analysis and Applications* 18.

Kahn J. A, Lan D, Kahn R S. (2007). Socio Demographic Factors Associated with High Risk of HPV Infection. *Obstet Gynecol.* 110(1).

L. Zadeh. (1965). *Fuzzy Sets Information and Control,* Vol.3(8).

M. Klitsch. (2002). Long-term pill use, high parity raise cervical cancer risk among women with Human Papilloma Virus infection - Digests - Brief Article, *Perspectives on Sexual and Reproductive Health 6.*

Moreno V, Bosch FX, Muñoz N, et al. (2002). Effect of oral contraceptives on risk of cervical cancer in women with human papillomavirus infection: *the IARC multi -centric case-control study.* Lancet 359(9312).

Muñoz N, Franceschi S, Bosetti C, et al. (2002). Role of parity and human papillomavirus in cervical cancer: *the IARC Multi-centric case-control study.* Lancet 359 (9312).

Quinlan, J. R. (1986). *Induction of Decision Trees.* Mach. Learn. 1.

Ross Quinlan J. (1993). *Machine Learning,* 1st ed. Morgan Kaufmann Publishers Inc.

www.cancer.org American Cancer Society.

Z. Pawlak. (1982). "Rough Sets", *International Journal of Computer and Information Sciences,* Vol.11.

Z. Pawlak. (1991). *Rough Sets - Theoretical Aspect of Reasoning about Data,* Kluwer Academic Pubilishers.

Table 1. Fuzzified Data Set

Patient No.	Risk of HPV	Multiple Partners (MP)	Young Age at First Sexual Act (AFSA)	Husband's Extra Family Affairs (EFA)	Low Socio Economic Status (Low SES)	Parity (Minimum of 4 full term pregnancies)	Oral Contraceptive Pills	Risk of Cervical Cancer
1	Y	Y	Y	Y	Y	Y	Y	High
2	N	Y	Y	N	Y	N	N	High
3	N	N	Y	Y	Y	N	N	High
4	Y	Y	N	N	Y	Y	Y	High
5	Y	N	N	N	Y	Y	N	Low
6	N	Y	Y	Y	Y	N	N	High
7	N	N	Y	N	N	Y	Y	Low
8	N	Y	Y	N	Y	N	N	High
9	Y	N	N	N	Y	Y	N	Low
10	N	N	N	N	N	N	N	Low
11	N	N	Y	N	N	Y	Y	Low
12	Y	N	N	N	Y	Y	N	Low
13	N	Y	Y	Y	Y	N	N	High
14	N	Y	Y	N	Y	N	N	High
15	N	Y	Y	N	Y	N	N	High
16	N	N	N	N	N	N	N	Low
17	Y	N	N	N	Y	Y	N	Low
18	Y	N	N	N	Y	Y	N	Low
19	N	Y	Y	Y	Y	N	N	High
20	N	N	Y	Y	Y	N	N	High
21	N	N	Y	N	N	Y	Y	Low
22	Y	Y	Y	Y	Y	Y	Y	High
23	Y	N	N	N	Y	Y	N	Low
24	Y	Y	N	N	Y	Y	Y	High
25	Y	Y	N	N	Y	Y	Y	High
26	N	Y	Y	N	Y	N	N	High
27	N	Y	Y	N	Y	N	N	High
28	N	N	Y	Y	Y	N	N	High
29	N	N	N	N	N	N	N	Low
30	Y	Y	Y	Y	Y	Y	Y	High

Table 2. Feature Selection

Patient No.	Risk of HPV	Multiple Partners (MP)	Low Socio Economic Status (Low SES)	Risk of Cervical Cancer
1	Y	Y	Y	High
2	N	Y	Y	High
3	N	N	Y	High
4	Y	Y	Y	High
5	Y	N	Y	Low
6	N	Y	Y	High
7	N	N	N	Low
8	N	Y	Y	High
9	Y	N	Y	Low
10	N	N	N	Low
11	N	N	N	Low
12	Y	N	Y	Low
13	N	Y	Y	High
14	N	Y	Y	High
15	N	Y	Y	High
16	N	N	N	Low
17	Y	N	Y	Low
18	Y	N	Y	Low
19	N	Y	Y	High
20	N	N	Y	High
21	N	N	N	Low
22	Y	Y	Y	High
23	Y	N	Y	Low
24	Y	Y	Y	High
25	Y	Y	Y	High
26	N	Y	Y	High
27	N	Y	Y	High
28	N	N	Y	High
29	N	N	N	Low
30	Y	Y	Y	High

ISSN 1913-1844 E-ISSN 1913-1852

Table 3. Gain Results for Risk Factors

Risk Factors	Gain
HPV	0.420026
MP	0.61006
AFSA	0.256424
EFA	0.28128
Low SES	0.32197
Parity	0.12453
Oral Pills	0.00721

Table 4. Equivalence Class

	HPV	MP	Low SES	Risk
Equiv – 1	Y	Y	Y	Y
Equiv – 2	N	Y	Y	Y
Equiv – 3	N	N	Y	Y
Equiv – 4	Y	N	Y	N
Equiv - 5	N	N	N	N

Table 5. Discerning Matrix

	E1	E2	E3	E4	E5	R
E1	-	a	aVb	b	aVbVc	R1
E2	a	-	b	avb	bVc	R2
E3	aVb	b	-	a	c	R3
E4	b	aVb	a	-	aVc	R4
E5	aVbVc	bVc	c	aVc	-	R5

Table 6. Reduct Set

Patient No.	HPV(a)	MP(b)	Low SES(c)	Risk of Cervical Cancer
E1	Y	Y	X	Y
E2	N	Y	X	Y
E3	Y	Y	Y	Y
E4	Y	N	X	N
E5	X	X	N	N

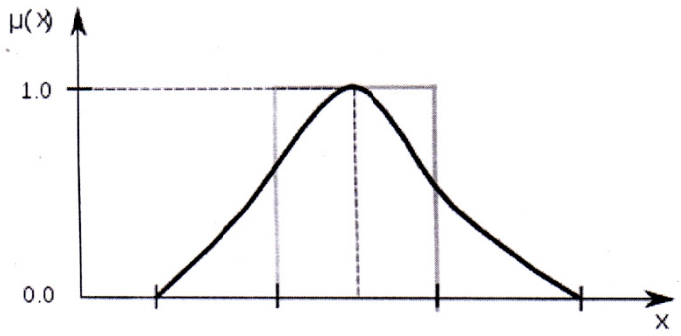

Figure 1. Membership Function

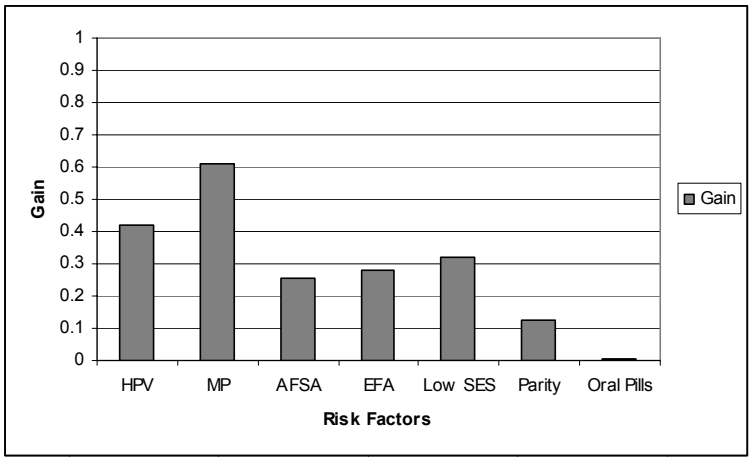

Figure 2. Results of Gain Calculation

 ISSN 1913-1844 E-ISSN 1913-1852

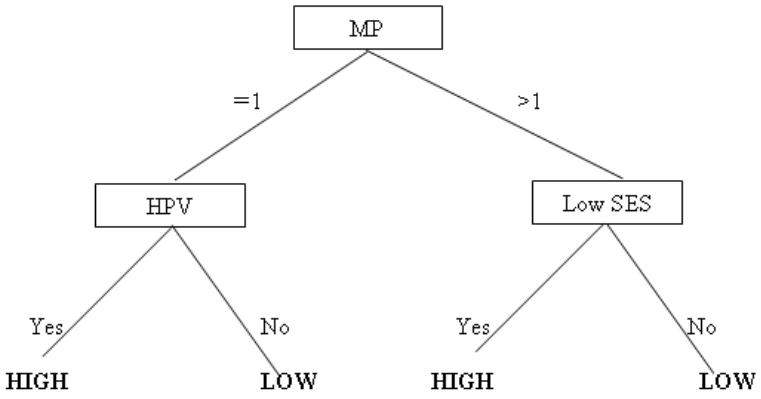

Figure 3. Decision Tree