A Hypothesis Analysis on the Proposed Methodology for Prediction of Polycystic Ovarian Syndrome

S. Rethinavalli[#1], Dr. M. Manimekalai[*2]

[#]*Assistant Professor, Department of Computer Applications*
Shrimati Indira Gandhi College, Trichy, Tamilnadu, India

[*]*Director and Head, Department of Computer Applications*
Shrimati Indira Gandhi College, Trichy, Tamilnadu, India

***Abstract*—The women's in the age of reproduction, the heterogeneous, complex and multifaceted disease called as PCOS (Polycystic ovary syndrome) affects about 5 to 10 % of the women. Metabolic syndrome (frequent metabolic traits) called as hypertension, hyperinsulinemia, abdominal obesity and dyslipidemia and these are characterized along with the resistance of insulin like chronic anovulation, hyperandrogenism and polycystic ovaries which ends with the serious diseases like coronary disease, endometrial hyperplasia and type 2 diabetes mellitus these are the long term consequences. For decision making purpose in PCOS phenotype (i.e to discover the main cause of PCOS with many number of attributes (tests)), an approach that used to store and manipulating the valued data for predicting the causes earlier and this process in information technology is called as data mining. In the previous research paper, a novel framework is proposed for the prediction of PCOS disease among women's. In this paper, a hypothesis test is made on the proposed methodologies to compare the effectiveness of the proposed techniques with existing techniques.**

***Keywords*— PCOS, Hypothesis, Chi-Square test.**

I. Introduction

The medicinal services industry gathers enormous measures of human services information which, sadly, are not "mined" to find concealed data for successful decision making. Data Mining and Knowledge Discovery have found various applications in business and investigative space. Important information can be found from use of data mining methods in social insurance framework. In this study, the potential utilization of classification based Data Mining methods [1], for example, Naïve Bayes, Artificial Neural Network, Decision Trees and Rule based to monstrous volume of medicinal services information are inspected. Utilizing medicinal profile, for example, age, sex, pulse, blood hormones and glucose, it can anticipate the probability of patients getting a Polycystic Ovarian Syndrome. It empowers huge information, e.g. relationships between medical factors, patterns identified with Polycystic Ovarian Syndrome to be set up.

Polycystic Ovarian Syndrome [2] is a typical female endocrine issue showed by hirsutism (extreme facial or body hair) and obesity, alopecia (male pattern baldness) unpredictable period, skin break out connected with enlarge ovaries, acanthuses, Nigerians' (brown skin patches), elevated cholesterol levels, weariness or absence of mental readiness, diminished sex drive, abundance male hormones and infertility. Different manifestations can incorporate rest apnea (breathing troubles amid dozing), thyroid issue, anxiety and depression. The reason for Polycystic Ovarian Syndrome is exceptional. Most analysts imagine that more than one component assumes a part in creating Polycystic Ovarian Syndrome. Qualities are thought to be one variable.

Data Mining based forecasts [4] gave the capacity to envision the onset of Polycystic Ovarian Syndrome and the discoveries can push the need to control the different component bringing about the infection. The investigation is critical in light of the fact that if these variables are left uncared, it may lead to excess body or facial hair, weight gain, infertility, sleep apnea, diabetes, skin problems, hormone imbalance and fatigue.

The health of populace, which is constructing principally in light of the aftereffect of therapeutic exploration, has a solid effect upon all human exercises. Among the most critical restorative perspectives are considered for the great translation of information and setting the analysis. Be that as it may, restorative basic leadership turn into a hard movement in light of the fact that the human specialists, who need to settle on choice, can scarcely prepare the enormous measures of information. So they require an apparatus that ought to have the capacity to help them to settle on a good choice. They could utilize some expert frameworks or Artificial Neural Network, which are a piece of Data Mining. PC technology has been progressed enormously and the interest has been expanded for the potential utilization of 'AI (Artificial Intelligence)' in Medicine and Biological Research [3]. The original dataset contains 32 attributes, so a novel algorithm called Neural Fuzzy Rough Set (NFRS) [4] feature selection method to reduce the number of attributes. It reduces the attributes size into 7. Then the Neural Fuzzy Rough Set is correlated with Artificial Neural Network [5] further to classify the reduced attribute set as with PCOS syndrome and without PCOS syndrome. The results obtained by using the above proposed algorithms are presented in the paper [6].

II. Chi-Square Test Of Independence

The chi-square (I) test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories. Remember, qualitative data is where you collect data on individuals that are categories or names. Then you would count how many of the individuals had particular qualities. If you were to do a hypothesis test, this is your alternative hypothesis and the null hypothesis is that they are independent. There is a hypothesis test for this and it is called the Chi-Square Test for Independence. Technically it should be called the Chi-Square Test for Dependence, but for historical reasons it is known as the test for independence.

A. *Hypothesis Test for Chi-Square Test*

1. State the null and alternative hypotheses and the level of significance Ho: the two variables are independent (this means that the one variable is not affected by the other) HA : the two variables are dependent (this means that the one variable is affected by the other) Also, state your α level here.

2. State and check the assumptions for the hypothesis testa.

a. A random sample is taken.

b. Expected frequencies for each cell are greater than or equal to 5 (The expected frequencies, E, will be calculated later, and this assumption means $E \geq 5$)

3. Find the test statistic and p-value. Finding the test statistic involves several steps. First the data is collected and counted, and then it is organized into a table (in a table each entry is called a cell). These values are known as the observed frequencies, which the symbol for an observed frequency is O. Each table is made up of rows and columns. Then each row is totaled to give a row total and each column is totaled to give a column total.

The null hypothesis is that the variables are independent. Using the multiplication rule for independent events you can calculate the probability of being one value of the first variable, A, and one value of the second variable, B (the probability of a particular cell P(A and B)). Remember in a hypothesis test, you assume that H0 is true, the two variables are assumed to be independent.

P (A and B) = P (A) . P(B) if A and B are independent

= (Number of ways A can happen)/(total number of individuals) . (Number of ways B can happen)/(total number of individuals)

= (row total)/n *(column total)/n

Now you want to find out how many individuals you expect to be in a certain cell. To find the expected frequencies, you just need to multiply the probability of that cell times the total number of individuals. Do not round the expected frequencies.

Expected Frequency (Cell A and Cell B) = E (A and B)

= n ((row total)/n .(column total)/n)

= (row total. column total)/n

If the variables are independent the expected frequencies and the observed frequencies should be the same. The test statistic here will involve looking at the difference between the expected frequency and the observed frequency for each cell. Then you want to find the "total difference" of all of these differences. The larger the total, the smaller the chances that you could find that test statistic given that the assumption of independence is true. That means that the assumption of independence is not true. How do you find the test statistic? First find the differences between the observed and expected frequencies. Because some of these differences will be positive and some will be negative, you need to square these differences. These squares could be large just because the frequencies are large, you need to divide by the expected frequencies to scale them. Then finally add up all of these fractional values. This is the test statistic.

B. *Test Statistic for Distribution of Chi-Square*

The symbol for chi-square is χ2

$$\chi 2 = \sum \frac{(O - E)^2}{E}$$

where O is the observed frequency and E is the expected frequency. χ2 has different curves depending on the degrees of freedom. It is skewed to the right for small degrees of freedom and gets more symmetric as the degrees of freedom increases (see figure 6.1). Since the test statistic involves squaring the differences, the test statistics are all positive. A chi-squared test for independence is always right tailed.

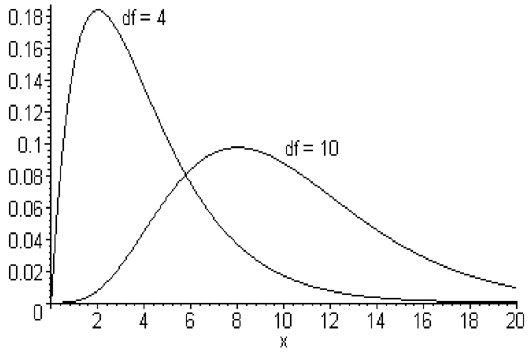

Figure 1: Chi-Square Distribution

p-value: Use χcdf (lower limit,1E99, *df*) Where the degrees of freedom is df = (# of rows −1)*(# of columns −1).

4. Conclusion : This is where you write reject H_o or fail to reject H_o . The rule is: if the p-value < α , then reject H_o . If the p-value ≥α , then fail to reject H_o

5. Interpretation: This is where you interpret in real world terms the conclusion to the test. The conclusion for a hypothesis test is that you either have enough evidence to show H_A is true, or you do not have enough evidence to show H_A is true.

III. Chi-Square Goodness Of Fit

In probability, you calculated probabilities using both experimental and theoretical methods. There are times when it is important to determine how well the experimental values match the theoretical values. An example of this is if you wish to verify if a die is fair. To determine if observed values fit the expected values, you want to see if the difference between observed values and expected values is large enough to say that the test statistic is unlikely to happen if you assume that the observed values fit the expected values. The test statistic in this case is also the chi-square. The process is the same as for the chi-square test for independence.

A. *Hypothesis Test of Goodness of Fit Test*

1. State the null and alternative hypotheses and the level of significance. Ho : The data are consistent with a specific distribution. HA : The data are not consistent with a specific distribution. Also, state your α level here.
2. State and check the assumptions for the hypothesis test
a. A random sample is taken.
b. Expected frequencies for each cell are greater than or equal to 5 (The expected frequencies, E, will be calculated later, and this assumption means $E \geq 5$).
3. Find the test statistic and p-value
Finding the test statistic involves several steps. First the data is collected and counted, and then it is organized into a table (in a table each entry is called a cell). These values are known as the observed frequencies, which the symbol for an observed frequency is O. The table is made up of k entries. The total number of observed frequencies is n. The expected frequencies are calculated by multiplying the probability of each entry, p, times n.
Expected Frequency (entry i) = E = n *p.

B. *Test Statistic*

$$\chi 2 = \sum \frac{(O-E)^2}{E}$$

where *O* is the observed frequency and *E* is the expected frequency Again, the test statistic involves squaring the differences, so the test statistics are all positive. Thus a chi-squared test for goodness of fit is always right tailed. p-value: Use χcdf (lower limit,1E99, *df*), Where the degrees of freedom is $df = k - 1$.
4. Conclusion: This is where you write reject H_o or fail to reject H_o . The rule is: if the p-value < α , then reject H_o . If the p-value ≥α , then fail to reject H_o .
5. Interpretation: This is where you interpret in real world terms the conclusion to the test. The conclusion for a hypothesis test is that you either have enough evidence to show H_A is true, or you do not have enough evidence to show H_A is true.

IV. Result Of Chi-Square On The Proposed Methodology

The chi-square test is implemented on our proposed methods to know whether the methods will give effect when it is implemented for the prediction of polycystic ovarian syndrome. The sample output for the chi-square test will be as follows:

A. *Chi-Square Test on Types of Menstrual Cycle*

In this result, the method 1 represents the proposed method (Neural Fuzzy Rough Set) [5] with obese and without obese patients records. Method 1 is NFRS with obese and Method 2 gives NFRS without obese.

To find the expected frequencies, we assume independence of the rows and columns. To get the expected frequency corresponding to the 50 at top left, we look at row total (311) and column total (80), multiply them, and then divide by the overall total (538). So the expected frequency is:

$$\frac{80 * 311}{538} = 46.245$$

So to complete the expected table, draw up another table similar to that above and having the same row and column totals. For each entry in this table, we simply calculate (row total*column total)/538.

The number of degrees of freedom is calculated for an m-by-n table as (m-1)(n-1), so in this case, (2-1)(9-1) = 1*8=8. To calculate the χ^2, we then have a further table 3.

From the above table 3, at the significance level of 5%, since $\chi^2 \geq$ critical value (i.e 8.32 ≥ 6.44) it can conclude that the proposed method with obese and without obese rejects the null hypothesis (H_0).

B. *Chi-Square Test on Clinical Hyperandrogenism*

To find the expected frequencies, we assume independence of the rows and columns. To get the expected frequency corresponding to the 19 at top left, we look at row total (96) and column total (44), multiply them, and then divide by the overall total (185). So the expected frequency is:

$$\frac{96 * 44}{185} = 22.83$$

So to complete the expected table, draw up another table similar to that above and having the same row and column totals. For each entry in this table, we simply calculate (row total*column total)/185. The completed table is table 5. The number of degrees of freedom is calculated for an m-by-n table as (m-1)(n-1), so in this case, (2-1)(4-1) = 1*3=3. To calculate the χ^2, we then have a further table 6: From the above table 6.6, From the above table 6.3, at the significance level of 5%, since $\chi^2 \geq$ critical value (i.e 3.70 ≥ 1.41) we can conclude that the proposed method with obese and without obese rejects the null hypothesis (H_0)

V. Conclusions

From the results obtained, it is concluded the proposed methodology NFRS for the prediction of PCOS which reduces the dataset size in the pre-processing step. The proposed NFRS (Neural Fuzzy Rough Set) is compared with Information gain by considering the type of menstrual cycle when the patient with obese and without obese, it is found that the NFRS remove the null hypothesis i.e the value of chi-square is greater than critical value. There is a positive, significant and correlation between the types of menstrual and the PCOS disease.

Table 1: The Observed Frequency On The Types Of Menstrual Cycle For The Proposed Methods

	Types of Menstrual Cycle									
Methods	Abnormal	Withdrawal	Bleed	Delayed	Early	Variable	Normal	Oligo Ovulation	Normal Ovulation	Total
Method-1	50	45	30	51	32	35	22	18	28	311
Method-2	30	25	24	40	25	20	18	26	19	227
	80	70	54	91	57	55	40	44	47	538

Table 2: The Expected Frequency On The Types Of Menstrual Cycle For The Proposed Methods

	Types of Menstrual Cycle									
Methods	Abnormal	Withdrawal	Bleed	DelayEd	Early	Variable	Normal	Oligo Ovulation	Normal Ovulation	Total
Method-1	46.5	40.46	31.22	52.60	32.95	31.7	23.12	25.43	27.17	311
Method-2	33.75	29.54	22.78	38.40	24.05	23.21	16.88	18.57	19.83	227
	80	70	54	91	57	55	40	44	47	538

Table 3: The Chi-Square Result Of The Types Of Menstrual Cycle On The Proposed Methods

Observed (O)	Expected (E)	|O-E|	$\|O-E\|^2$	$\|O-E\|^2/E$
50	46.25	3.75	14.10	0.30
45	40.46	4.54	20.57	0.51
30	31.22	1.22	1.48	0.05
51	52.60	1.60	2.57	0.05
32	32.95	0.95	0.90	0.03
35	31.79	3.21	10.28	0.32
22	23.12	1.12	1.26	0.05
18	25.43	7.43	55.28	2.17
28	27.17	0.83	0.69	0.03
30	33.75	3.75	14.10	0.42
25	29.54	4.54	20.57	0.70
24	22.78	1.22	1.48	0.06
40	38.40	1.60	2.57	0.07
25	24.05	0.95	0.90	0.04
20	23.21	3.21	10.28	0.44
18	16.88	1.12	1.26	0.07
26	18.57	7.43	55.28	2.98
19	19.83	0.83	0.69	0.03
	Total			8.32

Table 4: The Observed Frequency On The Clinical Hyperandergenism For The Proposed Methods

Clinical Hyperandrogenism					
Methods	Hirustism	Acne & Oily Skin	Hirustism, Acne, Oily Skin	Absence of Hyperandrogenism	Total
Method-1	19	23	26	28	96
Method-2	25	26	20	18	89
	44	49	46	46	185

Table 5: The Expected Frequency On The Clinical Hyperandergenism For The Given Methods

Clinical Hyperandrogenism					
Methods	Hirustism	Acne & Oily Skin	Hirustism, Acne, Oily Skin	Absence of Hyperandrogenism	Total
Method-1	22.83	25.43	23.87	23.87	96
Method-2	21.17	23.57	22.13	22.13	89
	44	49	46	46	185

Table 6: The Chi-Square Result On The Clinical Hyperandergenism Of The Proposed Methods

Observed (O)	Expected (E)	|O-E|	$\|O-E\|^2$	$\|O-E\|^2/E$
19	22.83	3.83	14.69	0.64
23	25.43	2.43	5.89	0.23
26	23.87	2.13	4.54	0.19
28	23.87	4.13	17.05	0.71
25	21.17	3.83	14.69	0.69
26	23.57	2.43	5.89	0.25
20	22.13	2.13	4.54	0.20
18	22.13	4.13	17.05	0.77
	Total			3.70

REFERENCES

[1] Illhoi Yoo, Patricia Alafaireet, Miroslav Marinov, Keila Pena-Hermandez, Rajitha Gopidi, Jia-Fu Chang, Lei Hua, "Data Mining in Healthcare and Biomedicine: A Survey of the Literature", Journal of Medical Systems-Springer, Volume 36, Issue 4, pp. 2431-2448, August 2012.

[2] S.S. Lim, R.J. Norman, M.J. Davies and L.J. Moran, "The effect of obesity on Polycystic Ovary Syndrome: A Systematic Review and Meta-Analysis", Obesity Review-Wiley Online Library, Volume 14, Issue 2, pp. 95-109, February 2013.

[3] Casey C. Bennett and Kris Hauser, "Artificial Intelligence Framework for Simulating Clinical Decision-Making: A Markov Decision Process Approach", Artificial Intelligence in Medicine-Elsevier, Volume 57, Issue 1, pp.9-19, January 2013

[4] Dr. K. Meena, Dr. M. Manimekalai and S. Rethinavalli, “A Novel Framework for Filtering the PCOS Attributes using Data Mining Techniques”, International Journal of Engineering Research & Technology (IJERT), Vol. 4 Issue 01,January-2015, pp.no 702-706.

[5] Dr. K. Meena, Dr. M. Manimekalai and S. Rethinavalli, “Correlation of Artificial Neural Network Classification and NFRS Attribute Filtering Algorithm for PCOS Data”, International Journal of Research in Engineering and Technology, Volume 4, Issue 3, pp. 519-524, March 2015.

[6] Dr. K. Meena, Dr. M. Manimekalai and S. Rethinavalli, “Implementing Neural Fuzzy Rough Set and Artificial Neural Network for Predicting PCOS”, International Journal on Recent and Innovation Trends in Computing and Communication, Volume: 3 Issue: 12, pp.6722-6727, December 2015