A REVIEW ON PROGNOSIS OF PCOS USING NFRS, HYBRID TECHNIQUE AND CHI-SQUARE TEST

Neetha Thomas
Research Scholar
Department of Computer Science
Kongunadu Arts and Science College
Coimbatore, Tamil Nadu, India

Dr. A.Kavitha, Ph.D
Associate Professor
Department of Computer Science
Kongunadu Arts and Science College
Coimbatore, Tamil Nadu, India

***Abstract*—Polycystic Ovarian Syndrome is an ambidextrous disease, which affects women in their reproductive age or during puberty, ie between adolescence age and menopause. It is a complex hormonal condition. Approximately 70 per cent of this kind of cases is undiagnosed. Multiple small cysts which are of pearl-sized are grown in an ovary, which affects 5–20% of women of child bearing age. This review incorporates study on foreseeing of disease with PCOS phenotypes, using neural fuzzy logic system, Chi-squrae test.**

***Keywords*—PCOS,NFRS,ANN,Chi-Square test**

I. INTRODUCTION

PCOS or polycystic ovarian morphology (PCOM) is an endocrinological issue which [4] is seen in women at the age of reproduction, The ovaries contain clusters of immature eggs, called cyst, it is referred as Polycystic or multiple cyst which are fluid sacks, that are harmless and develop in ovaries, which causes hormonal imbalance[1]. The excess production of androgen, a male hormone in female body that lead to hyperandrogenism which leads to ovulatory dysfunction and ends with the serious diseases like CVD, endometrial hyperplasia and type-2 diabetes mellitus (T2D). The risk of T2D in PCOS patients is higher than normal patients.

Polycystic Ovarian Syndrome [2] is an typical female endocrine issue with different symptoms shown, hirsutism (over growth of hair), obesity, Acne on face, Sleep Apnea, infertility, alopecia (male pattern baldness).mood disorders, unpredictable period, Anxiety, abundance male hormones, Preeclampsia (It is marked by high blood pressure in women),etc

This paper begins with pre-processing step to reduce the PCOS dataset, attribute filtering method NFRS produced the best result. Following with the classification techniques and hybrid algorithm which reduced the attributes further than NFRS, subsequently the technique is implemented for the prediction of PCOS; conclusively the chi-square test is performed to find the effect of methods used in the prediction of PCOS.

II. FEATURE SELECTION

Attribute Selection is an important step for the mining of data. A subset of attributes is relevant for mining, among the original attributes.

The main objective of feature selection is to reduce the inputs for processing, their by improve performance and to provide faster ,cost-effective models[3].feature engineering or feature extraction is the process in data mining to extract relevant information or features from large complex data sets.

Attribute selection is classified to two. Filter and Wrapper approaches [4]. The filter attribute selection, method is independent of the data mining algorithm, they look on to the native properties of the data. Whereas in Wrapper method, attribute selection method uses the result of the data mining algorithm and analyse its performance of attributes subset is. Also feature subsets are defined, and various subsets of features are generated and evaluated.

Figure-1 shows the processes of selecting features from complex dataset.Orginal feature set is fed for the generation of feature set, a candidate subset is selected as the best candidate in the current search, the decision is taken to continue with the available prototype, if so the attribute is selected, else the process will be iterated to generation of feature set, until the feature is selected.

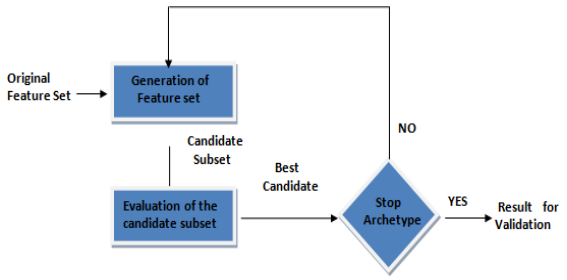

Figure- 1 The steps in feature selection

Dr.K.Meena et al, [3] proposed a new technique NFRS(Neural Fuzzy Logic Rough set) for the selection of attributes from the large number of attributes in PCOS dataset, when compared to Information Gain Subset Evaluation (IGSE),PCA (Principle Component Analysis), Correlation based Feature Selection (CFS) .greedy first search, ranker etc are the search algorithms that are used in the feature selection. The NFRS evaluation is performed based on the decision feature E and condition feature Dj, which is denoted by RVj,e,In the algorithm it is referred as *RV, such that* it Measures the value within range of [0, 1]. When the value of *RV* is low to the class,ie,

RVj,e≤ 0.0001 then the feature is irrelevant.
The training data set is represented as ϕ (d1, d2 . . . .dn,e).value o conditional feature is represented in SB,SN is the set of available features and the values re to be stored in descending order in RVj,e.After getting the first element compare (σ DK (SB) <σ (SB),then compare with the maximum current value in SB with the SB old value, this will continue until D_k is empty. Here, attributes are compared pair wise as well as compared with class attribute to find its contribution to class value, thus the irrelevant attributes are removed.

The attributes are selected by NFRS Evaluation using Best First Search method and compared with Information Gain Subset Evaluation using Ranker Method; found that NFRS is the best method for feature selection.

Input: A training set is represented by ϕ (d1, d2 . . .dn,e)
Output: The conditional feature D_k is represented by SB
Begin
Step 1: The features SN =Eliminate the features that have lower threshold value.
Step 2: RVj,e = Arrange the value of value in decreasing order in SN
Step 3: SB = max { RVj,e }
Step 4: To get the first element in SB,ie, Dk = getFirstElement (SB).
Step 5: Then go to begin stage
Step 6: If (σ DK (SB) <σ (SB)
Step 7: SN →Dk ; new old { }
SB = SB ∪ Dk
Step 8: SB = max{I (SBnew),I (SBold)}
Step 9: Dk = getNextElement(SB) ;
Step 10: E**nd until** (Dk == null)
Step 11: Return SB;
End;

Thus among the 26 attributes of PCOS data set, with NFRS using best first search attributes are reduced to 5 with the accuracy 80.85. Table 1 shows the accuracy percentage of corresponding filtering methods.

Table 1- list of Attribute filtering method with Accuracy percentage

Attribute filtering method	**No: of filtered Attributes**	**% of Accuracy**
CFS	3	80.25
IG	9	72.65
NFRS	5	80.85
PCA	11	69.65

The root mean squared error of NFRSE- ID3 gives less error rate .and it is concluded that NFRS Evaluation given better result.

III. CLASSIFICATION

The original the dataset contains 26 attributes and 303 instances, different classifiers including SVM, Naive Bayesian, and classification tree compared along with ANN to acquire the best accuracy.

Table-2 shows the Classification accuracy of algorithms used.SVM, Naive Bayesian, classification tree produced accuracy percentage 76.58%, 82.85%,75.23%. The classifier accuracy was 83.70% for ANN.
Table 2 Percentage of Accuracy of Classification Algorithm

Classification Algorithm	**Correctly classified instance from 303 instances.**	**Percentage of Accuracy**
SVM	232	76.58
ANN	254	83.83
Classification Tree	228	75.23
Naive Bayes	251	82.85

IV. FEATURE SELECTION AND CLASSIFICATION

K. Meena [5] et al, proposed a new algorithm known as Hybrid algorithm, which combines NFRS with ANN (Artificial Neural network).Those attributes selected using NFRS Evaluation are transferred to Artificial neural network for further reduction. Attribute which has highest conditional probability is calculated in artificial neural network. Both NFRS and ANN calculate Conditional Probability measure.

The Input is provide with training set δ, with a predefined threshold value T (K1; K2; ; KM;L) ,the expected output is an optimal subset .Experimentation is done using Orange Data mining tool, Table 3 shows Reduced Attributes using NFRS and Hybrid Technique. Originally the dataset contains 26 attributes and 303 instances, from that by using the attribute filtering method NFRS with ANN the attributes are reduced to 2.

Table 3 Reduced Attributes using NFRS and Hybrid Technique.

Attribute filtering method	**Total Attributes**	**Number of filtered Attributes**
NFRS	26	5
Hybrid Feature: NFRS +ANN	26	2

Feature selection and classification algorithms combine together to form a Hybrid algorithm (NFRS+ANN), which is best for predicting the PCOS disease from PCOS dataset.

V. IMPLEMENTATION OF HYBRID ALGORITHM IN PCOS DATASET

Dr.meena et al [6] proposed a work in PCOS dataset to predict PCOS disease; they have implemented the algorithm using MATHLAB for patients with different age groups, menstrual cycle, with/without obese, presence/absence of clinical hyperandrogenism.

These reports are compared with the feature selection techniques like Principal Component Analysis (PCA), Information Gain (IG), Neural Fuzzy Rough Set Evaluation; etc among them NFRS shows the best result less number of features than the other techniques.

The test is preceded with feature selection from 100 patients Table 4 shows the result of NFRS and NFRS+ANN: patients with obese

Table 4 Results of feature selection by NFRS and classification accuracy NFRS+ANN: patients with obese.

Different age group	**NFRS**	**NFRS+ANN (Accuracy %)**
Less than 20	20	85
21-25	22	80
26-30	30	81
31-35	21	88
36-40	18	75
More than 41	33	83

When comparing: patients with /without obese using NFRS and Hybrid Algorithm, NFRS + ANN gives the best accuracy result of classification.

Table 5 results of feature selection by NFRS and classification accuracy NFRS+ANN: patients different menstrual cycle- with /without obese

TYPE OF MENSTRUAL CYCLE	**NFRS**		**NFRS+ANN (Accuracy %)**	
	With Obese	Without Obese	With Obese	Without Obese
Abnormal	50	30	80	73
Withdrawal	45	25	73	75
Bleed	30	24	75	78
Delayed	51	40	80	79
Early	32	25	79	80
Variable	35	0	78	80
Normal	22	18	85	85
Oligo-Ovulation	18	26	89	86
Normal Ovulation	28	19	86	89

When comparing: patients different menstrual cycle-with /without obese, using NFRS and Hybrid Algorithm, NFRS + ANN gives the best accuracy result of classification, the results are shown in Table5.

Table 6- Results of feature selection by NFRS: clinical hyperandrogenism - with /without obese.

CLINICAL HYPERANDROGENISM	**NFRS**	
	With Obese	Without Obese
Hirustism	19	25
Acne &Oily skin	23	26
Hirustism, Acne &Oily skin	26	20
Absence of Hyperandrogenism	28	18

Table-6 shows results of clinical hyperandrogenism are stated with the most common symptom hirustisum, acne, oily skin etc.Result of absence of androgen hormone is mentioned.

VI. CHI-SQUARE TEST

S.Rethinavalli et al [7], proposed a hypothesis analysis using chi-square test. As per the above conclusions, on PCOS Menstrual cycle and Clinical Hyperandrogenism, chi-square test is implemented for a vivid understanding about the effect when it is implemented in PCOS dataset.

The chi-square (I) test is used to regulate difference between the expected and observed frequencies in one or more categories. This is hypothesis test; the null hypothesis is stated to be independent. This hypothesis test is called the Chi-Square Test for Independence.

The Summary of hypothesis test for Chi-square test is -

1. State the Null and Alternative hypotheses and the level of significance H_o,H_A, also, state your α level here
2. State and check the assumptions, The expected frequencies, E, will be calculated for each cell
3. Test statistic and p-value should found, as well as find the degrees of freedom, *df* = (# of rows −1)*(# of columns −1).
4. This is the step to either reject or accept 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. Portray the conclusion that; either enough evidence to show H_A is true/No evidence to show H_A is true.

The symbol for chi-square is x2.

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

Where O is the observed frequency and E is the expected Frequency. Chi-square is used in proposed methodology for the test on types of menstrual cycle-obese and without obese and also for the test on clinical hyperandrogenism- obese and without obese. By using the result in Table 5 and Table 6 using the NFRS, chi-square value is calculated.

Expected frequencies 'E' is calculated using the independence of rows and columns.ie

Expected frequencies = Column total *Row total / Overall Total

Example**: - 80*311 = 46.245**
538

Here [7], 80 is the column total of menstrual type Abnormal - obese and without obese

311 is the row total of obese menstrual types.

538 is the overall total of both obese and without obese types.

Observed frequency 'O' is calculated from Table 5 and Table 6 determined by NFRS technique.

Thus the chi-square value for menstrual type withdrawal is calculated as

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

Example:-

$$X^2 = \sum \frac{(50-46.25)^2}{46.25} = 0.30$$

Where the degrees of freedom is *df* is calculated as (2-1) (9-1) = 1*8=8.

The significance level / α level is redefined with 5%, According to the rule if the $\chi 2 < \alpha$, then reject *Ho* . $\chi 2 \geq \alpha$, then fail to reject H_o. *Here,* $\chi 2 \geq$ critical value ie, (8.32 ≥ 6.44).Similarly chi-squre test on clinical Hyperandogenism [7] is also calculated. $\chi 2 \geq$ critical value (i.e 3.70 ≥1.41)

In both cases the proposed method with obese and without obese rejects for (menstrual cycle and clinical hyperandrogenism) the null hypothesis (H_0)

VII. CONCLUSION

Data mining is the exploration of large datasets of information from hidden and unknown patterns .medical mining plays a key role to extract the information from the patient's medical history for the deeper understanding of the medical data for the prediction of disease, it will reduce medical costs as well as providing information to patients regarding effective treatments and best practices.

This review undergone through different methodologies like filtering of attributes, classification, also verified the effect of method implementated.

VIII. REFERENCES

[1]Ranjitha Sitheswaran, S. Malarkhodi **"**An effective automated system in follicle identification for Polycystic Ovary Syndrome using ultrasound images", Date Added to IEEE Xplore: 08 September 2014.

[2]Neetha Thomas,,Dr.A.Kavitha," A literature inspection on polycystic ovarian morphology in women using data mining methodologies", International Journal of Advanced Research in Computer Science, Volume 9, No. 1, January-February 2018.

[3] 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

[4] Sunita Beniwal , Jitender Arora," Classification and Feature Selection Techniques in Data Mining", International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 6, August – 2012

.[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

[7] S. Rethinavalli1,Dr. M. Manimekalai," A Hypothesis Analysis on the Proposed Methodology for Prediction of Polycystic Ovarian Syndrome",IJCSET,| November 2016 | Vol 6, Issue 11, 396-400