

A Novel Hybrid Framework for Risk Severity of Polycystic Ovarian Syndrome

S. Rethinavalli* and M. Manimekalai**

ABSTRACT

Generally 5 to 15% of the ladies in the phase of reproduction face the sickness called Polycystic Ovarian Syndrome (PCOS) which is the complex, multifaceted and heterogeneous. In the grown-up age, the advancement of PCOS phenotype gets a noteworthy part the programming of utero fetal and it is imagined that it may be the reason for PCOS and it ought to be cleared up. The data mining methods are utilized to anticipate the danger seriousness of ladies with PCOS ailment. In this paper, a novel hybrid structure is proposed to discover the seriousness of ladies with PCOS. The proposed system incorporates the Neural Fuzzy Rough Set procedures in the pre-processing step, NFRS then combines with Artificial Neural Network to group the ladies with PCOS and without PCOS. At that point Artificial Neural Fuzzy Inference System is utilized to discover the seriousness of the ladies with PCOS infection.

Keywords: Neural Fuzzy Rough Set, Artificial Neural Network, Neuro Fuzzy System, PCOS.

1. 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]. 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.

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. 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].

* Assistant Professor, Department of Computer Applications.

** Director & Head, Department of Computer Application.

*,** Shrimati Indira Gandhi College, Trichy, Tamilnadu, India.

2. DEFINITION OF THE PROBLEM

By and large terms, an issue exists when there is a circumstance that presents uncertainty, perplexity, or trouble; or when an inquiry is offered for thought, talk, or arrangement. In the disease of PCOS (Polycystic Ovarian Syndrome), the issue is at whatever point it has a necessity or desire that is not being characterized or won't be met, whether because of doctrine, organization, policy or training and inadequate equipment. An acknowledgment is that an issue exists in the initial step while portraying it in significant terms. Easy as it might appear, it should ask ourselves whether it truly do have an issue. The issues with the PCOS ladies are no menstrual cycles or menstrual cycles (for ladies of reproductive age), skin break out, weight increase, overabundance hair development on the face and body, diminishing scalp hair, ovarian sores. Through this research work, these issues with the ladies can be recognized at before phase of the utilizing data mining procedures.

In proposing the conclusions, the upheld decision making and the valuable data is highlighted by method for the procedure of inspecting, cleaning, transforming and designing the information is called as Data Analysis [5]. In the range of science, sociology and the distinctive business, the different expansive systems underneath the format of names which has various methodologies and realities are known as Data Analysis [6]. The obscure data is revealed in a database which is recognized by the data mining. Prediction, Classification, Association, Clustering and these capacities are incorporated into the information mining [1]. From the different fields like Pattern Recognition [7], Information Retrieval [8], Neural Networks [9], Analysis of Spatial Data [10], Database Technology [11] and Statistics [12], and these procedures must be coordinated for the procedure of Data Mining. In addition, repetitious procedure with different strides is called as Data Mining.

3. INFORMATION ABOUT THE DATASET

In this research work, the dataset is taken from [13] and the title of the dataset is Polycystic Ovarian Syndrome Proliferative Phase Endometrial Cell Types. These samples are collected from obese/overweight women with PCOS. This dataset contains 31 attributes which is considered here as a sample id of the patient.

Table 1
Information about the PCOS dataset

<i>No</i>	<i>Name</i>	<i>Description</i>
1	Id_Ref	
2	Identifier	
3	Eenpcos103.Pco1 (Gsm1174425)	PCO WITH ENDOTHELIAL CELL
4	Eenpcos107.Pco7 (Gsm1174429)	
5	Eenpcos140.Uc271(Gsm1174436)	
6	Eeppcos105.Pco7_Epcam(Gsm1174427)	PCO WITH EPITHELIAL CELL
7	Eeppcos109.Pco8_Epcam (Gsm1174430)	
8	Eeppcos119.Pc11 (Gsm1174432)	
9	Eeppcos138.Uc271_Epcam (Gsm1174435)	
10	Emcpcos102.Pco7 (Gsm1174424)	PCO WITH MESENCHMAL CELL
11	Emcpcos106.Pco7 (Gsm1174428)	
12	Emcpcos120.Pc11 (Gsm1174433)	
13	Escpcos101.Pco1 (Gsm1174423)	PCO DISEASE STATE
14	Escpcos104.Pco7 (Gsm1174426)	PCO WITH STROMAL CELL
15	Escpcos118.Pc11 (Gsm1174431)	
16	Escpcos134.Uc271 (Gsm1174434)	
17	Eenctrl.Etb65 (Gsm1174409)	CONTROL WITH ENDOTHELIAL CELL
18	Eenctrl016.Uc182 (Gsm1174414)	
19	Eenctrl032.Uc208 (Gsm1174418)	

(contd...)

(Table 1 contd...)

No	Name	Description
20	Eenctrl036.Uc209 (Gsm1174421)	CONTROL WITH EPITHELIAL CELL
21	Eepctrl014.Uc182_Epcam (Gsm1174412)	
22	Eepctrl030.Uc208_Epcam (Gsm1174416)	
23	Eepctrl034.Uc209_Epcam (Gsm1174419)	CONTROL WITH MESENCHYMAL CELL
24	Emcctrl.Etb65 (Gsm1174408)	
25	Emcctrl015.Uc182 (Gsm1174413)	
26	Emsctrl031.Uc208 (Gsm1174417)	CONTROL WITH STROMAL CELL TYPE
27	Emcctrl035.Uc209 (Gsm1174420)	
28	Escctrl.Etb65 (Gsm1174410)	
29	Escctrl013.Uc182 (Gsm1174411)	
30	Escctrl029.Uc208 (Gsm1174415)	
31	Escctrl033.Uc209 (Gsm1174422)	

4. PROPOSED HYBRID FRAMEWORK

The proposed framework combines the core concepts of Rough set theory and Fuzzy system in the step of pre-processing. The proposed algorithm is named as Neural Fuzzy Rough Set Evaluation (NFRS) and this algorithm reduces the number of attributes from the original dataset. The resultant dataset obtained by this algorithm is called Reduct Dataset. In the next classification step, NFRS is again combined with ANN to produce the optimal dataset from reduct dataset. Then the optimal dataset is trained by Neuro-Fuzzy System to predict the risk severity of the PCOS patients. The proposed model for the risk prediction of the PCOS which contains NFRS and hybrid NFRS+ANN and Neuro-Fuzzy system is represented by the figure 1.

4.1. Feature Reduction by Proposed Neural Fuzzy Rough Set Evaluation Algorithm

This proposed Neural Fuzzy Rough Set Evaluation algorithm is presented in our previous published research paper [16]. The correlation between the decision feature E and a condition feature D_j is denoted by $RV_{j,e}$ which refers



Figure 1: Proposed Hybrid Framework for the risk severity of PCOS

RV that measures the above value. For the range of $[0, 1]$ its value is normalized by the symmetrical uncertainty to assure that they are comparable [14]. The knowledge of value of the conditional attribute D_j completely predicts the value of the decision feature E and it is indicated by the value of 1 and D_j and E values which are independent, then the attribute value D_j is irrelevant and it is indicated by the value zero [15]. Accordingly, the value of $RV_{j,e}$ is maximum, then the feature is strong relevant or essential is assumed. When the value of RV is low to the class such as $RV_{j,e} \leq 0.0001$ then we consider the feature is irrelevant or not essential and these are examined in this work.

Pseudo code: NFRS Algorithm:

Input: A Training Dataset represented by $\tilde{O} (d_1, d_2 \dots d_n, e)$

Algorithm:

Begin

When the forming of the set SN by the features, eliminate the features that have lower threshold value.

Arrange the value of $RV_{j,e}$ value in decreasing order in SN

Then initialize $SB = \max \{ RV_{j,e} \}$

To get the first element in SB the formula used for that is

$D_k = \text{getFirstElement}(SB)$

Then go to begin stage

for each feature D_k in SN

If $(\sigma D_k(SB) < \sigma(SB))$

$SN \rightarrow D_k; \text{new old} \{ \}$

$SB = SB \cup D_k$

$SB = \max\{I(SB_{\text{new}}), I(SB_{\text{old}})\}$

$D_k = \text{getNextElement}(SB);$

End until $(D_k == \text{null})$

Return SB ;

End;

Output:

Reduct Data set

4.2. Proposed Hybrid Classification algorithm by NFRS and Artificial Neural Network

This algorithm is published in our previous research paper [17]. In this research paper, a new Hybrid Classification algorithm by combining Neural Fuzzy Rough Set Evaluation and Artificial Neural Network.. The Neural Fuzzy Rough Set algorithm reduces the number of attributes based on the SU measure, In Neural Fuzzy Rough Set each attributes are compared pair wise to find the Similarity and the Attributes are compared to class attribute to find the amount of contribution it provides to the class value , based on these the attributes are removed. The selected attributes from the NFRS algorithm is fed into Artificial Neural Network for further reduction. Artificial Neural Network calculates the conditional probability for each attribute and the attribute which has highest conditional probability is selected. Both the Algorithms NFRS and ANN works on the Conditional Probability measure.

Pseudo code: Hybrid Classification Algorithm:**Input:** $T(K_1; K_2; \dots; K_{M:L})$ // a training data set δ // a predefined threshold*Begin**for j = 1 to M do begin**calculate $TU_{j,l}$ for K_j ;**if ($TU_{j,l} \geq \delta$)**append K_j to T'_{list} ;**end;**order T'_{list} in descending $TU_{j,l}$ value;* $K_p = \text{getFirstElement}(T'_{list});$ *do begin* $K_q = \text{getNextElement}(T'_{list}, K_p);$ *if ($K_q \neq \text{NULL}$)**do begin* $K'_q = K_q;$ *if ($TU_{p,q} > TU_{q,l}$)**remove K_q from T'_{list} ;* $K_q = \text{getNextElement}(T'_{list}, K'_q);$ *else $K_q = \text{getNextElement}(T'_{list}, K_q);$* *end until ($K_q == \text{NULL}$);* $K_p = \text{getNextElement}(T'_{list}, K_p);$ *end until ($K_p == \text{NULL}$);* $T_{best} = T'_{list};$ $T_{best} = \{X_1, X_2, \dots, X_N\}$ *for j=1 to N begin**for k=j+1 to N begin* $P[L_m/(X_j, X_k)] = P[(X_j, X_k)/L_m] * P(L_m)$ $P[L/(X_j, X_k)] = P[L_1/(X_j, X_k)] + P[L_2/(X_j, X_k)] + \dots + P[L_n/(X_j, X_k)]$ *If ($P[L/(X_j, X_k)] > \alpha$)*

{

*if ($P[L/X_j] > P[L/X_k]$)**Remove X_k from T_{best}* $T_{best} = \text{NFRS}(X)$

}

*Else**Remove X_j from T_{best} ;* $T_{best} = \text{NFRS}(X)$ *End;**} End;***Output:** T_{best} // a optimal dataset.

4.3. Risk Prediction by Neuro Fuzzy System (ANFIS)

The procedure of a fuzzy framework has three stages. These strides are Fuzzification, Rule Evaluation, and Defuzzification. In the fuzzification step, the information input values are changed into degrees of membership in the fuzzy sets. In the standard assessment step, each fuzzy guideline is allotted with a quality worth. The quality is controlled by the degrees of memberships of the crisp input values in the fuzzy arrangements of forerunner part of the fluffy guideline. The defuzzification stage transposes the fuzzy yields into input values. The fuzzy guideline calculation is produced for making the preparation dataset for the FNNM. The SQL inquiries are created utilizing the “RandAndOr” capacity for short-posting the unmistakable qualities contained in every field. These qualities symbolize the attributes of irregularity in the interruption information and typicality from the ordinary information. The tenet creation produced an intelligent grouping which contains the “and” and “or” legitimate administrators and effect the choice of irregularity or typicality are spoken to as far as weights relegated. Neuro Fuzzy System is developed utilizing the accompanying calculation.

Pseudo Code: Fuzzy Rule Algorithm

Input: *Optimal Data Set*

Start

S = selected attribute

K = subset of operation

A = next component from the accessible information

K = item[j]

For j = 1 to m-1

A = Data Field [j + 1]

K = K union A Select special thing of the field

End for

Store K

End

Initialize Increment to 1

Initialize Weight of Find Record to 0

Introduce Quct to 1

WHILE Increment < MJ

FOR every worth FuL

Record [FuL] = rand() mod Nfl

END FOR

FOR every worth JF

Qust = sql select proclamation where each

Field[JF] = Index[JF] + “ + RandAndOr();

END FOR

TotalFuzzyR = ExecuteQuery(Qust)

On the off chance that TotalFuzzyR is non zero THEN

Wht[Quct] = TotalR/TotalFuzzyR

Add 1 to Quct

ENDIF

Add 1 to Increment

ENDWHILE

Save Wht

Save Quct

Output: PCOS disease Severity: Healthy, Stage 1(Low), Stage 2(Medium) and Stage 3(High)

6. IMPLEMENTATION RESULT AND DISCUSSIONS

The above algorithms are implemented in MATLAB R 2016 a. The following results are obtained when the given dataset is utilized for implementation. From the table 2, it is found that the proposed NFRS method utilizes the genetic algorithm search method to get the reduct dataset. Using the proposed NFRS method the total numbers of attributes are reduced to 7 attributes, whereas the Information gain with Ranker search method gives 17 attributes from the total number of attributes. Since the genetic algorithm is the best optimization and therefore, it is considered with NFRS method.

To evaluate the performance of the proposed algorithms, the following important parameters are considered: Classification Accuracy, Kappa Statistic, Mean Absolute Error, Root Mean Squared Error, Relative Absolute Error and Root Relative Squared Error. From the table 4, it is observed that the Hybrid Classification algorithm gives 90.12 % classification accuracy from the optimal dataset than the other methods like ANN and NFRS.

Table 2
List of Features Reduced by IG and NFRS methods

<i>Information Gain with Ranker Search</i>	<i>Proposed NFRS</i>
eENPCOS140.UC271	ID_REF
eENPCOS103.PCO1	eMCPCOS102.PCO7
eEPCtrl014.UC182_EpCAM	eSCPCOS134.UC271
eSCPCOS118.PC11	eENCtrl036.UC209
Ems0Ctrl031.UC208	eEPCtrl014.UC182_EpCAM
eSCCtrl033.UC209	eMCCtrl035.UC209
eMCCtrl.ETB65	eSCCtrl029.UC208
eEPPCOS109.PCO8_EpCAM	
eSCPCOS134.UC271	
eMCPCOS106.PCO7	
eENPCOS107.PCO7	
eENCtrl.ETB65	
eEPPCOS105.PCO7_EpCAM	
eSCCtrl029.UC208	
eENCtrl032.UC208	
eMCCtrl015.UC182	
eSCCtrl013.UC182	

Kappa Statistic value is also increased than the other methods. And the value of Mean Absolute Error, Root Mean Squared Error, Relative Absolute Error and Root Relative Squared Error is reduced than the ANN and NFRS. From the overall comparison, the proposed Hybrid Classification algorithm performs well in all considered aspects. And the optimal dataset gives higher prediction accuracy than the reduct dataset. The figure 2 represents classification accuracy of the proposed Hybrid Classification algorithm.

Using the fuzzy rules, the risk severity of the PCOS is predicted as Healthy, Stage 1(Low), Stage 2(Medium) and Stage 3(High). From the overall performance, Neuro Fuzzy System for optimal dataset gives more classification accuracy i.e. 93.64% than the NFRS and Hybrid Classification Algorithm.

Table 3
The optimal Dataset is produced by Proposed Hybrid Classification Algorithm

<i>Proposed NFRS</i>	<i>Proposed Hybrid Classification Algorithm</i>
ID_REF	eMCPCOS102.PCO7
eMCPCOS102.PCO7	eSCPCOS134.UC271
eSCPCOS134.UC271	eENCtrl036.UC209
eENCtrl036.UC209	eEPCtrl014.UC182_EpCAM
eEPCtrl014.UC182_EpCAM	eSCCtrl029.UC208
eMCCtrl035.UC209	
eSCCtrl029.UC208	

Table 4
Implementation result of the Proposed Hybrid Classification algorithms

	<i>Original Dataset</i>	<i>ANN</i>	<i>NFRS</i>	<i>Hybrid Classification Algorithm</i>
Correctly classified instance	78.54	84.32	86.85	90.12
Kappa statistic	0.41	0.65	0.69	0.72
Mean Absolute Error	0.36	0.31	0.29	0.24
Root Mean Squared Error	0.54	0.43	0.41	0.39
Relative Absolute Error	69.70	44.89	43.90	39.76
Root Relative Squared Error	87.64	79.24	76.74	75.20

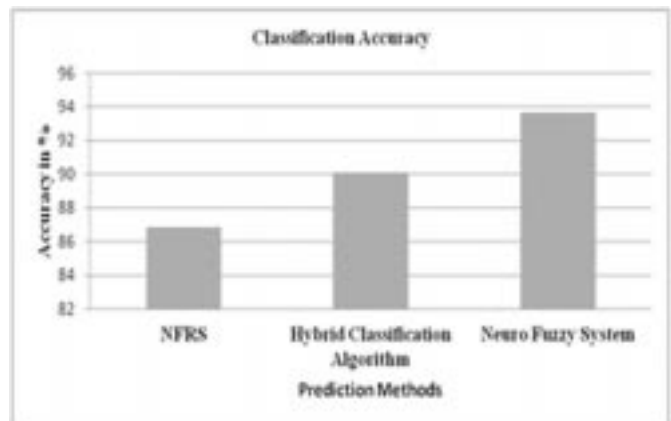
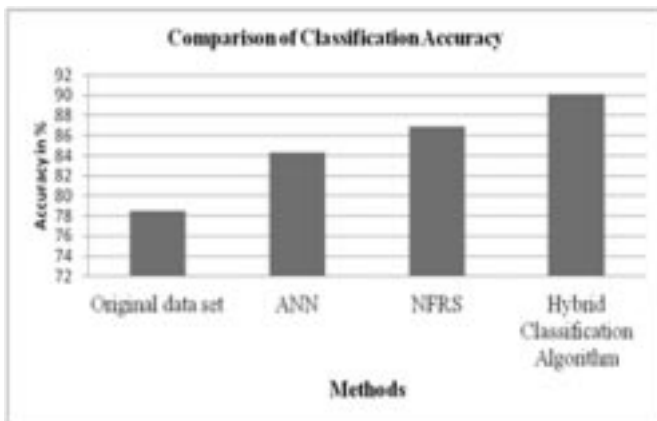


Figure 2,3: Comparison of the Performance and Classification Accuracy of Proposed Hybrid Classification algorithm, ANN and NFRS

6. CONCLUSIONS

From the results obtained, it is concluded that the proposed Neural Fuzzy Rough Set Evaluation methods produces the reduct dataset which reduces the original dataset of 31 attributes to 7 attributes. The proposed Hybrid Classification Algorithm gives only 5 attributes which is named as optimal dataset from the reduct dataset. Using Neuro Fuzzy System, the prediction of risk severity level of Polycystic Ovarian Syndrome and it would be better diagnostic tool for predicting the risk severity. The proposed model gives the prediction accuracy upto 93.64 % which is higher than the other existing methods. In future, the real time dataset would be considered to evaluate the proposed system to achieve our results. The proposed NFRS and Hybrid Classification Algorithm could be suggested for Feature Reduction algorithm in all other medical fields.

REFERENCES

- [1] Illhoi Yoo, Patricia Alafaireet, Miroslav Marinov, Keila Pena-Hernandez, Rajitha Gopidi, Jia-Fu Chang, Lei Hua, "Data Mining in Healthcare and Biomedicine: A Survey of the Literature", *Journal of Medical Systems-Springer*, **36(4)**, 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*, **14(2)**, 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*, **57(1)**, 9-19, January 2013.
- [4] Bruno Fernandes Chimieski, Rubem Dutra Ribeiro Fagundes, "Association and Classification Data Mining Algorithms Comparison Over Medical Datasets", *Journal of Health Informatics*, 44-51, 2013.
- [5] Andrew Kusiak, "Data Mining and Decision Making", *Proceedings of the SPIE Conference on Data Mining and Knowledge Discovery: Theory, Tools, and Technology IV*, **4730**, SPIE, Orlando, FL, 155-165, April 2002.
- [6] Mohammed J. Zaki, Wagner Meira, "Data Mining and Analysis-Fundamental Concepts and Algorithms", 1-25.
- [7] Noriyasu Homma, "Pattern Recognition in Medical Image Diagnosis", 319-336.
- [8] Ammar Yassir and Smitha Nayak, "Issues in Data Mining and Information Retrieval", *International Journal of Computer Science & Communication Networks*, **2(1)**, 93-98.
- [9] Dr. K. Usha Rani, "Analysis Of Heart Diseases Dataset Using Neural Network Approach", *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, **1(5)**, September 2011.
- [10] Diansheng Guo, Jeremy Mennis, "Spatial data mining and geographic knowledge discovery—An introduction", *Computers, Environment and Urban Systems*, **33**, 403–408, 2009.
- [11] Amir Netz, Surajit Chaudhuri, Jeff Bernhardt, Usama Fayyad, "Integration of Data Mining and Relational Databases", *Proceedings of the 26th International Conference on Very Large Databases, Cairo, Egypt*, 719-722, 2000.
- [12] David J. Hand, "Statistics and Data Mining: Intersecting Disciplines", *SIGKDD Explorations, ACM SIGKDD*, **1(1)**, pp. 16-19, 1999.
- [13] Dataset Source Link- <ftp://ftp.ncbi.nlm.nih.gov/geo/datasets/GDS4nnn/GDS4987/>.
- [14] M.L. Raymer, W.F. Punch, E.D. Goodman, L.A. Kuhn and A.K. Jain, "Dimensionality Reduction Using Genetic Algorithms", *IEEE Transactions On Evolutionary Computation*, **4(2)**, 2000
- [15] Yongheng Zhao and Yanxia Zhang "Comparison of decision tree methods for finding active objects", 2-8.
- [16] 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)*, **4(1)**, 702-706, January-2015.
- [17] Dr. K. Meena, Dr. M. Manimekalai, S. Rethinavalli, "Correlation of Artificial Neural Network Classification and NFRS Attribute Filtering Algorithm for PCOS data", *IJRET – International Journal of Research in Engineering and Technology*, **4(3)**, 519-524, 2015.