International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 01,January-2015

A Novel Framework for Filtering the PCOS Attributes using Data Mining Techniques

Dr. K. Meena[1]
Former vice chancellor
Bharathidasan University
Trichy, India

Dr. M. Manimekalai[2], S. Rethinavalli[3]
Director and Head[2], Assistant Professor[3]
Department of Computer Applications
Shrimati Indira Gandhi College, Trichy, India

***Abstract*— From a large amount of data, the significant knowledge has been discovered by means of applying some techniques and this kind of techniques in knowledge management process are known as Data mining techniques. And for a specific domain, a structure of knowledge discovery is called as data mining and it is used to solving the problems. The classes of unknown data are detected by the technique called classification. Neural networks, rule based, decision trees, Bayesian are some of the existing methods used for the classification. Moreover, it is necessary to filter the irrelevant or unclosed attributes before applying any mining techniques. Embedded, Wrapper and filter techniques are different feature in selection techniques which is used for the filtering. The most common endocrinological issue which pretending the women are PCOS (Polycystic ovary syndrome). The higher prevalence for the patients with PCOS are obese than the general population about 50%. The long-term morbidity is resulted by means of insulin resistance due to the condition of the metabolic element. This present research focuses the attribute selection techniques like Information Gain Subset Evaluation (IGSE) and our proposed method Neural Fuzzy Rough Subsets Evaluation (NFRSE) for selecting the attributes from the large number of attributes, and search methods like BestFirst Search is used for neural fuzzy rough subset evaluation, and Ranker method is applied for the Information gain evaluation. The decision tree classification techniques like ID3 and J48 algorithm are used for the classification. In this paper, the above techniques are analyzed by the PCOS (Polycystic Ovary Syndrome) Dataset and generate the result and from the result it can be concluded which technique will be the suit for attribute selection in the decision making process.**

Keywords—ID3, J48, PCOS, NFRSE, IGSE.

I. INTRODUCTION

In the adult age, the development of PCOS phenotype brings a major role in the programming of utero fetal and it is conceived that it might be the cause for PCOS and it should be clarified. The phenotype of PCOS last expressions and the result of the menstrual disturbances and characteristic metabolic are due to the interaction of the genetic factors with the obesity [1] (environmental factors) which leads the women for developing of PCOS and it is genetically inclined. On ultrasonography the evidence in the ovary, that is the following changes may include such as showing of increased levels of serum androgen, hirsuitism acne, amenorrhoea or oligo, morphological change and an ovulation, these characteristics are demonstrated with the patients with PCOS. The symptoms of PCOS are as follows [2]: The unwanted or excess growth of hairs in the body or face. On the scalp, the hair might be thinning, around the waist, the weight may be increased or there will be problem of weight losing, the women with PCOS may experience infertility and the menstrual periods may be missing or it may be irregular. The treatment for PCOS are: using pills for birth control [3], alteration in the lifestyle [4], medicine for fertility [5], medicine for diabetes [6], preventing excess hair growth [7] and drilling of ovarian surgery [8].

In this digital world, data mining is the only reliable source available to solve the complexity of gathered data. The two categories of data mining tasks can be broadly classified such as descriptive and predictive [9]. Descriptive mining tasks characterize the general ascribes of the data in the database. Predictive mining tasks perform inference on the present data in order to make predictions. Data available for mining is raw data. Data comes in different source, so the format may be different. And it may consist of noisy data, irrelevant attributes, missing data etc [10]. Discretization – When the data mining algorithm cannot cope with continuous attributes, discretization needs to be implemented. This step consists of transforming a continuous attribute into an unconditional attribute, taking only a small number of detached values.

Discretization often improves the comprehensibility of the discovered knowledge [11]. Attribute Selection – not all attributes are relevant and so for selecting a subset of attributes relevant for mining, among all original attributes, attribute selection is required.

A Decision Tree Classifier consists of a decision tree generated on the basis of instances. The decision tree has two types of nodes: a) the root and the internal nodes, b) the leaf nodes. The root and the internal nodes are associated with attributes, leaf nodes are associated with classes. Basically, each non-leaf node has an outgoing branch for each possible value of the attribute associated with the node [12]. To determine the class for a new instance using a decision tree, starting with the root, consecutive internal nodes are inspected until a leaf node is reached. The root node and in the each internal node test is applied. The outcome of the test determines the branch traversed, and the next node visited. The class for the instance is the class of the final leaf node.

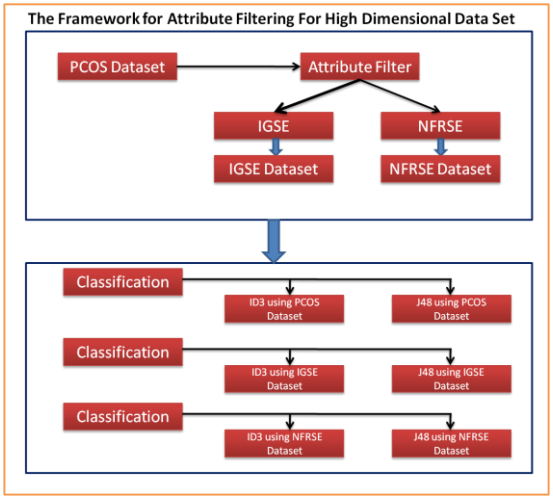

Figure 1: The Framework for Attribute Filtering For High Dimensional Data Set

II. FEATURE SELECTION

Many irrelevant attributes could be presented in data to be mined. Therefore they need to be removed. And Also many mining algorithms don't perform well with large amounts of features or attributes. Therefore feature selection techniques need to be applied before any kind of mining algorithm is applied [13]. The main objective of feature selection is to avoid over fitting and improve model performance and to provide faster and more cost-effective models.

The selection of optimal features appends an extra layer of complexity in the modeling as instead of just finding optimal parameters for full set of features, first optimal feature subset is to be found and the model parameters are to be optimized. Attribute selection methods can be broadly divided into filter and wrapper approaches. In the filter approach, the attribute selection method is independent of the data mining algorithm to be applied to the selected attributes and assess the relevance of features by looking only at the intrinsic properties of the data. In most cases a feature relevance score is calculated, and low scoring features are removed [14]. Moreover, in subset of features left after feature removal is presented as input to the classification algorithm. The important benefit of filter techniques are easily scale to high dimensional datasets that are computationally simple and fast, and as the filter approach is independent of the mining algorithm so feature selection needs to be performed only once, and then different classifiers can be evaluated.

III. INFORMATION GAIN SUBSET EVALUATION TECHNIQUE

The information gain can be measured by using the following algorithm steps:

To estimate the gain generated by a split over attributes is computed by the following algorithm:

Step 1: Let R be the sample:

Step 2: K_i is Class J; j = 1,2,…,n

$$J(r1,r2,..,rn) = - \Sigma\, qj \log2\,(qj)$$

Step 3:Rj is the no. of samples in class j

The binary algorithm is Qi = Ri /R, log2

Step 4: u be the distinct values for the attribute B

Step 5: A(E) = Entropy is

$$\Sigma\{(R1_k+R2_k+..+Smk)/R\}*J(r1k,..rnk)$$
$$k=1$$

Step 6: Where Rjk is samples in Class j and subset k of Attribute B.

$$J(R1k,R2k,..Rnk) = - \Sigma\, qjk \log2\,(qjk)$$
$$Gain(B)=J(r1,r2,..,rn) - A(E)$$

Step 7: From among the tests with at smallest amount average gain, then the gain ratio is then choosen

Step 8: Q(B) is the Gain Ratio

$$\sum_{i}^{t} \frac{Si}{S} \log \frac{Si}{S}$$

Gain Ratio (B) = Gain (B)/ Q(B)

IV. ID3 DECISION TREE TECHNIQUE

By using the predictive machine-learning known as decision tree to find the target value of a new sample by using the different attribute gives the dependent variable from the available data [15]. The terminal nodes are associated with the final classification result value of the dependent variable, the branches between the nodes are shown by the possible values of these attributes and an observed sample, and these attributes referred by the internal nodes of a decision trees. It depends on the values of all other attributes and the consideration of the needed variable value. And it is the attribute with the purpose of predicted and it looks on it. The values of the dependent variables are predicted by the independent variable's attribute.

V. J48 DECISION TREE TECHNIQUE

In Weka data mining tool, J48 is an implementation of the C4.5 algorithm by developing java code, it creates a decision tree based set of label indexed input data. [16] Ross Quinlan was developed this algorithm. C4.5 algorithm can be utilized for classifying the data, generating decision trees. Therefore, C4.5 algorithm is frequently concerned as a statistical data classifier.

VI. PROPOSED TECHNIQUE:NEURAL FUZZY ROUGH SET EVALUATION

The correlation between the decision feature E and a condition feature Dj is denoted by RVj,e which refers *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 [17]. 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 [18]. Accordingly, the value of *RVj*,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 *RVj*,e≤ 0.0001 then we consider the feature is irrelevant or not essential and these are examined in this paper.

Input: *A training set is represented by ϕ (d1, d2 . . .dn,e)*
Output: *A reductant accuracy of the conditional feature D is represented by SB*
Begin
Step 1: When the forming of the set SN by the features, eliminate the features that have lower threshold value.
Step 2: Arrange the value of RVj,e value in decreasing order in SN
Step 3: Then initialize SB = max { RVj,e }
Step 4: To get the first element in SB the formula used for that is D_k = getFirstElement (SB).
Step 5: Then go to begin stage
Step 6: for each feature D_K in SN
Step 7: If (σD_K (SB) <σ (SB)
Step 8: SN →D_k ; new old { }
SB = SB ∪ D_k
Step 9: SB = max{I (SBnew),I (SBold)}
Step 10: D_k = getNextElement(SB) ;
*Step 11: **End until** (D_k == null)*
Step 12: Return SB;
End;

VII. IMPLEMENTATION RESULT

The attributes that are selected by the Neural Fuzzy Rough Subset Evaluation using Best First Search method and Information Gain Subset Evaluation using Ranker Method are as follows: For the experimental results, the PCOS patients dataset is used here [19]:

S.No	Attributes
1	ID_REF
2	IDENTIFIER
3	eENPCOS103.PCO1
4	eENPCOS107.PCO7
5	eENPCOS140.UC271
6	eEPPCOS105.PCO7_EpCAM
7	eEPPCOS109.PCO8_EpCAM
8	eEPPCOS119.PC11
9	eEPPCOS138.UC271_EpCAM
10	eMCPCOS102.PCO7
11	eMCPCOS106.PCO7
12	eMCPCOS120.PC11
13	eSCPCOS101.PCO1
14	eSCPCOS104.PCO7
15	eSCPCOS118.PC11
16	eSCPCOS134.UC271
17	eENCtrl.ETB65
18	eENCtrl016.UC182
19	eENCtrl032.UC208
20	eENCtrl036.UC209
21	eEPCtrl014.UC182_EpCAM
22	eEPCtrl030.UC208_EpCAM
23	eEPCtrl034.UC209_EpCAM
24	eMCCtrl.ETB65
25	eMCCtrl015.UC182
26	eMSCtrl031.UC208
27	eMCCtrl035.UC209
28	eSCCtrl.ETB65
29	eSCCtrl013.UC182
30	eSCCtrl029.UC208
31	eSCCtrl033.UC209

Table 1: Attribute Table-Given Dataset

A. *Information Gain Subset Evaluation Technique using Ranker Search Method:*

S.No	Raking	Attributes
1	8.0596	5 eENPCOS140.UC271
2	8.0487	3 eENPCOS103.PCO1
3	7.9952	21 eEPCtrl014.UC182_EpCAM
4	7.9801	15 eSCPCOS118.PC11
5	7.9644	26 Ems0Ctrl031.UC208
6	7.9644	31 eSCCtrl033.UC209
7	7.9549	24 eMCCtrl.ETB65
8	7.9549	7 eEPPCOS109.PCO8_EpCAM
9	7.9482	16 eSCPCOS134.UC271
10	7.9482	11 eMCPCOS106.PCO7
11	7.9482	4 eENPCOS107.PCO7
12	7.9316	17 eENCtrl.ETB65
13	7.9316	6 eEPPCOS105.PCO7_EpCAM
14	7.9316	30 eSCCtrl029.UC208
15	7.9316	19 eENCtrl032.UC208
16	7.9146	25 eMCCtrl015.UC182
17	7.9146	29 eSCCtrl013.UC182

Table 2: Result dataset from IGSE

IGSE using Ranker serach method, Selected Attributes Are: 5,3,21,15,26,31,24,7,16,11,4,17,6,30,19,25,29 -17 Attributes.

B. *Proposed Technique: Neural Fuzzy Rough Set Using Genetic Search Algorithm:*

S.No	Attributes
1	ID_REF
2	eMCPCOS102.PCO7
3	eSCPCOS134.UC271
4	eENCtrl036.UC209
5	eEPCtrl014.UC182_EpCAM
6	eMCCtrl035.UC209
7	eSCCtrl029.UC208

Table 3: Result dataset from NFRSE

NFRSE using Genetic search algorithm, Selected Attributes are: 1, 10, 16, 20,21,27,30: 7 Attributes.

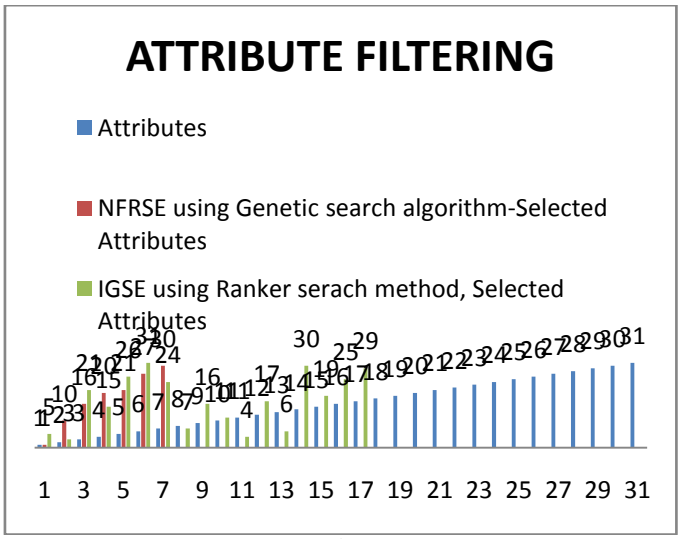

Figure 2: Graphical Representation of NFRSE and IGSE Attribute Result

C. ID3 Classification Result for Given Dataset:

Correctly Classified Instances	10	3.3445%
Incorrectly Classified Instances	289	96.6555%
Kappa statistic	0	
Mean absolute error	0.007	
Root mean squared error	0.0592	
Relative absolute error	99.9285%	
Root relative squared error	100.0634%	
Coverage of cases (0.95 level)	65.2174%	
Mean rel. region size (0.95 level)	63.3803%	
Total Number of Instances	299	

D. J48 Classification Result for Given Dataset:

Correctly Classified Instances	297	99.3311%
Incorrectly Classified Instances	2	0.6689%
Kappa statistic	0.9933	
Mean absolute error	0	
Root mean squared error	0.0049	
Relative absolute error	0.6717%	
Root relative squared error	8.1921%	
Coverage of cases (0.95 level)	100%	
Mean rel. region size (0.95 level)	0.3592%	
Total Number of Instances	299	

E. ID3 Classification Result for Result Dataset from IGSE Technique:

Correctly Classified Instances	297	99.3311%
Incorrectly Classified Instances	2	0.6689%
Kappa statistic	0.9933	
Mean absolute error	0	
Root mean squared error	0.0049	
Relative absolute error	0.6717%	
Root relative squared error	8.1971%	
Coverage of cases (0.95 level)	100%	
Mean rel. region size (0.95 level)	0.3592%	
Total Number of Instances	299	

F. J48 Classification Result for Result Dataset from IGSE

Correctly Classified Instances	297	99.3311%
Incorrectly Classified Instances	2	0.6689%
Kappa statistic	0.9933	
Mean absolute error	0	
Root mean squared error	0.0049	
Relative absolute error	0.6717%	
Root relative squared error	8.1971%	
Coverage of cases (0.95 level)	100%	
Mean rel. region size (0.95 level)	0.3592%	
Total Number of Instances	299	

G. ID3 Classification Result for Result Dataset from NFRSE

Correctly Classified Instances	281	93.9799%
Incorrectly Classified Instances	18	6.0201%
Kappa statistic	0.9396	
Mean absolute error	0.0004	
Root mean squared error	0.0142	
Relative absolute error	6.0403%	
Root relative squared error	24.577%	
Coverage of cases (0.95 level)	100%	
Mean rel. region size (0.95 level)	0.5559%	
Total Number of Instances	299	

H. J48 Classification Result for Result Dataset from NFRSE

Correctly Classified Instances	281	93.9799%
Incorrectly Classified Instances	18	6.0201%
Kappa statistic	0.9396	
Mean absolute error	0.0004	
Root mean squared error	0.0142	
Relative absolute error	6.0403%	
Root relative squared error	24.577%	
Coverage of cases (0.95 level)	100%	
Mean rel. region size (0.95 level)	0.5559%	
Total Number of Instances	299	

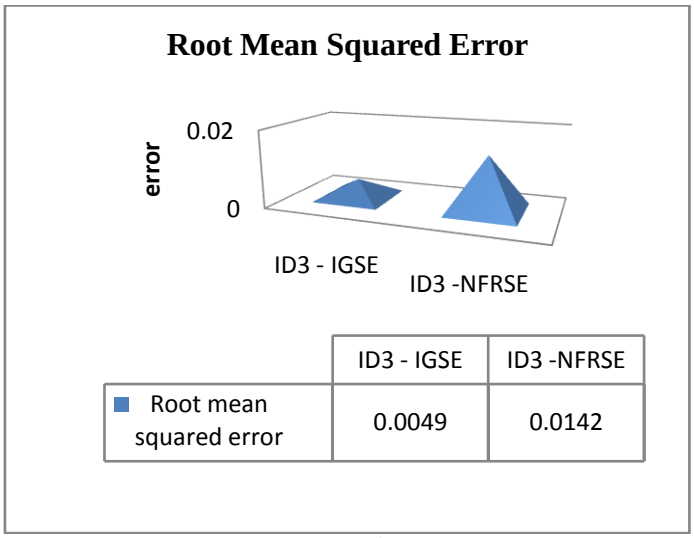

Figure 3: Graphical Representation of Root mean Squared Error

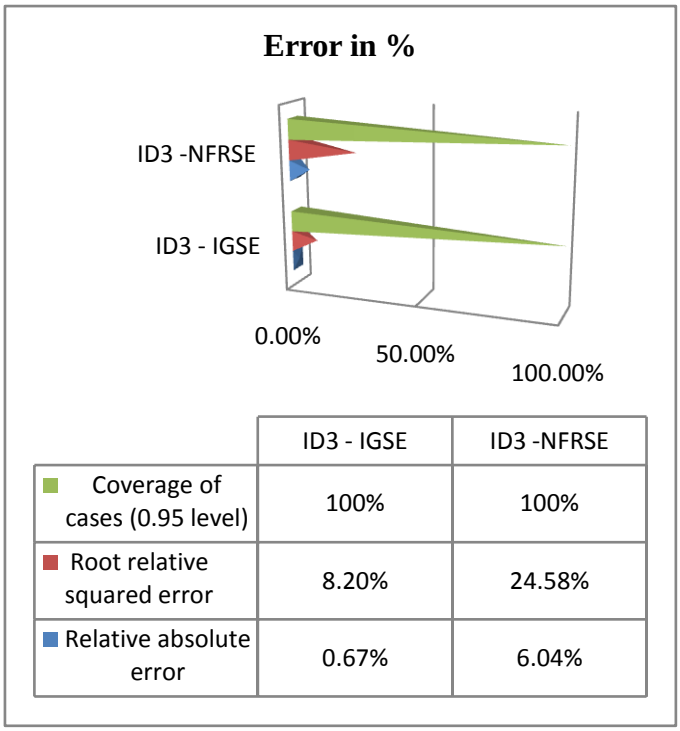

Figure 4: Graphical Representation of error in %

VIII. RESULT AND ANALYSIS

After analyzing the above experimental results of proposed method gives the less number of attributes when it is compared with other feature selection technique like Information Gain and it is useful in the decision making process of PCOS diagnosis of the patients using some important features in mean time with accuracy. In addition to accuracy we have to concentrate on the error rate. Here J48 decision tree classification produces the approximate error rate than another technique of decision tree ID3 for the result set obtained by the above feature selection techniques. By analysing the graphical representation, the root mean squared error of NFRSE ID3 gives less error rate than the ID3-IGSE. From all the above results, we can conclude that Neural Fuzzy Rough Set Evaluation gives better result than the other methods.

IX. CONCLUSION

The decision tree classification techniques like ID3 and J48 and the feature selection technique of information gain evaluation technique are also overviewed in this paper and the feature selection technique called as Neural Fuzzy Rough Subset Evaluation which is proposed in this paper. As a result of the above analysis, I finally concluded that for selecting the attributes, the neural fuzzy rough subset evaluation technique gives the better result in the purpose of decision making whereas to reduce the error rate, the J48 classification method are also used for the above purpose.

REFERENCES

[1] T. M. Barber, M. I. McCarthy, J. A. H. Wass and S. Franks, "Obesity and polycystic ovary syndrome", Clinical Endocrinology (2006) 65, 137–145.
[2] "Polycystic Ovary Syndrome", The American College of Obstetricians and Gynecologistics.
[3] Frederick R. Jelovsek, "Which Oral Contraceptive Pill is Best for Me?", pp.1-4.
[4] Marja Ojaniemi and Michel Pugeat, "An adolescent with polycystic ovary syndrome", European Journal of Endocrinology (2006) 155 S149–S152.
[5] Seddigheh Esmaeilzadeh, Mouloud Agajani Delavar, Zahra Basirat, Hamid Shafi, "Physical activity and body mass index among women who have experienced infertility", Fatemezahra Infertility and Reproductive Health Research Center, Department of Obstetrics and Gynecology, Babol University of Medical Sciences, Babol, Iran, pp. 499-505.
[6] Daniela Jakubowicz, Julio Wainstein and Roy Homburg, "The Link between Polycystic Ovarian Syndrome and Type 2 Diabetes: Preventive and Therapeutic Approach in Israel", IMAJ, VOL 14 July 2012, pp.442-447.
[7] Lyndal Harborne, Richard Fleming, Helen Lyall, Naveed Sattar and Jane Norman,"Metformin or Antiandrogen in the Treatment of Hirsutism in Polycystic Ovary Syndrome", Journal of Clinical Endocrinol Metab, September 2003, 88(9):4116–4123.
[8] Dimitrios Panidisa, Konstantinos Tziomalosb,Efstathios Papadakisa, Ilias Katsikisa, "Infertility Treatment in Polycystic Ovary Syndrome: Lifestyle Interventions, Medications and Surgery", Front Horm Res. Basel, Karger, 2013, vol 40, pp 128–141.
[9] I.H. Witten, E. Frank and M.A. Hall, "Data mining practical machine learning tools and techniques", Morgan Kaufmann publisher, Burlington 2011.
[10] J. Han and M. Kamber, "Data mining concepts and techniques", Morgan Kaufmann, San Francisco 2006.
[11] B. Pfahringer, "Supervised and unsupervised discretization of continuous features", Proc. 12th Int. Conf. Machine Learning, 1995, pp. 456-463.
[12] Guyon, N. Matic and V. Vapnik, "Discovering informative patterns and data cleaning", In: Fayyad UM, Piatetsky-Shapiro G, Smyth P and Uthurusamy R. (ed) Advances in knowledge discovery and data mining, AAAI/MIT Press, California, 1996, pp. 181- 203.
[13] W. Daelemans, V. Hoste, F.D. Meulder and B. Naudts, "Combined Optimization of Feature Selection and Algorithm Parameter Interaction in Machine Learning of Language", Proceedings of the 14th European Conference on Machine Learning (ECML-2003), Lecture Notes in Computer Science 2837, Springer-Verlag, Cavtat-Dubrovnik, Croatia, 2003, pp. 84-95.
[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, Vol. 4, No. 2, 2000.
[15] Yongheng Zhao and Yanxia Zhang "Comparison of decision tree methods for finding active objects", pp.no 2-8.
[16] Jay Gholap "Performance Tuning of J48 Algorithm for Prediction of Soil Fertility", pp.no 1-4.
[17] Pradipta Maji and Sankar K. Pal, "Feature Selection Using f-Information Measures in Fuzzy Approximation Spaces" IEEE Transactions On Knowledge And Data Engineering, Vol. 22, No. 6, June 2010.
[18] Lei Yu, Huan Liu, "Efficient Feature Selection via Analysis of Relevance and Redundancy" Journal of Machine Learning Research 5 (2004) 1205–1224.
[19] PCOS Dataset Link: ftp://ftp.ncbi.nlm.nih.gov/geo/datasets/GDS4nnn/GDS4987/