CORRELATION OF ARTIFICIAL NEURAL NETWORK CLASSIFICATION AND NFRS ATTRIBUTE FILTERING ALGORITHM FOR PCOS DATA

K. Meena[1], M. Manimekalai[2], S. Rethinavalli[3]

[1]*Former Vice-chancellor, Bharathidasan University, Tricy, Tamilnadu, India*
[2]*Director and Head, Department of Computer Applications, Shrimati Indira Gandhi College, Trichy, TN, India*
[3]*Assistant Professor, Department of Computer Applications, Shrimati Indira Gandhi College, Trichy, TN, India*

Abstract

Mostly 5 to 15% of the women in the stage of reproduction face the disease called Polycystic Ovarian Syndrome (PCOS) which is the multifaceted, heterogeneous and complex. The long term consequences diseases like endometrial hyperplasia, type 2 diabetes mellitus and coronary disease are caused by the polycystic ovaries, chronic anovulation and hyperandrogenism are characterized with the resistance of insulin and the hypertension, abdominal obesity and dyslipidemia and hyperinsulinemia are called as Metabolic syndrome (frequent metabolic traits) The above cause the common disease called Anovulatory infertility. Computer based information along with advanced Data mining techniques are used for appropriate results. Classification is a classic data mining task, with roots in machine learning. Naïve Bayesian, Artificial Neural Network, Decision Tree, Support Vector Machines are the classification tasks in the data mining. Feature selection methods involve generation of the subset, evaluation of each subset, criteria for stopping the search and validation procedures. The characteristics of the search method used are important with respect to the time efficiency of the feature selection methods. PCA (Principle Component Analysis), Information gain Subset Evaluation, Fuzzy rough set evaluation, Correlation based Feature Selection (CFS) are some of the feature selection techniques, greedy first search, ranker etc are the search algorithms that are used in the feature selection. In this paper, a new algorithm which is based on Fuzzy neural subset evaluation and artificial neural network is proposed which reduces the task of classification and feature selection separately. This algorithm combines the neural fuzzy rough subset evaluation and artificial neural network together for the better performance than doing the tasks separately.

Keywords: *ANN, SVM, PCA, CFS*

----------***----------

1. INTRODUCTION

The polycystic ovary syndrome (PCOS) in women is pretended by the endocrinological which is the most common issue. The following changes like hirsuitism acne, oligo or amenorrhoea, anovulation, morphological change and showing the increased levels of serum androgen and these are demonstrated with the PCOS women on the evidence of the ovary on the ultrasonography. Diagnostically, the above is the current practice and it is the agreed criteria in Rotterdam 2003 [1]. About 50% of the general population, the patients with PCOS are obese which are the higher prevalence. Due to the condition of the metabolic element, the result of the long-term morbidity by the insulin resistance.

The small cysts and the clusters of pearl-sized which induces the PCOS frequently because it is referred as the polycystic when there are many number of cysts in the ovaries. The fluid-filled form of the immature eggs are contained by the cystic. The symptoms of the PCOS are contributed due to the large number of male hormones productions by the Androgen [2]. The programming of utero fetal is brings out by the PCOS phenotype which is the plays the major role in the adult age of women and it might be the cause for the PCOS. Due to the interaction of the genetic factors with the obesity, the metabolic characteristic and result of the menstrual disturbances are the last expressions of the phenotype of PCOS. [3] (factors of environment) which leads the women for developing of PCOS and it is genetically inclined.

Nowadays, data mining is the exploration of large datasets to extort hidden and formerly unknown patterns, relationships and knowledge that are complicated to detect with conventional statistical methods. In the emerging field of healthcare data mining plays a major role to extract the details for the deeper understanding of the medical data in the providing of prognosis [4]. Due to the development of modern technology, data mining applications in healthcare consist about the analysis of health care centres for enhancement of health policy-making and prevention of hospital errors, early detection, prevention of diseases and preventable hospital deaths, more value for money and cost savings, and detection of fraudulent insurance claims.

The characteristic selection has been an energetic and productive in the field of research area through pattern recognition, machine learning, statistics and data mining communities. The main intention of attribute selection is to

choose a subset of input variables by eradicating features, which are irrelevant or of non prognostic information. Feature selection [5] has proven in both theory and practice to be valuable in ornamental learning efficiency, escalating analytical accuracy and reducing complexity of well-read results. Feature selection in administered learning has a chief objective in finding a feature subset that fabricates higher classification accuracy. The number of feature N increases because the expansions of the domain dimensionality. Among that finding an optimal feature subset is intractable and exertions associated feature selections have been demonstrated to be NP-hard. At this point, it is crucial to depict the traditional feature selection process, which consists of four basic steps, namely, validation of the subset, stopping criterion, evaluation of subset and subset generation. Subset generation is a investigation process that generates the candidate feature subsets for assessment based on a certain search strategy. Depends on the certain assessment, the comparison with the best prior one and each candidate subset is evaluated. If the new subset revolves to be better, it reinstates best one. Whenever the stopping condition is fulfilled until the process is repeated. There are large number of features that can exceed the number of data themselves often exemplifies the data used in ML [6]. This kind of problem is known as "the curse of dimensionality" generates a challenge for a mixture of ML applications for decision support. This can amplify the risk of taking into account correlated or redundant attributes which can lead to lower classification accuracy. As a result, the process of eliminating irrelevant features is a crucial phase for designing the decision support systems with high accuracy.

In this technical world, data mining is the only consistent source accessible to unravel the intricacy of congregated data. Meanwhile, the two categories of data mining tasks can be generally categorized such as descriptive and predictive. Descriptive mining tasks illustrate the common attributes of the data in the database. Predictive mining tasks execute implication on the present data in order to formulate the predictions. Data available for mining is raw data. The data collects from different source, therefore the format may be different. Moreover, it may consist of noisy data, irrelevant attributes, missing data etc. Discretization – Once the data mining algorithm cannot cope with continuous attributes, discretization [7] needs to be employed. However, this step consists of transforming a continuous attribute into an unconditional attribute, taking only a small number of isolated values. Frequently, Discretization often improves the comprehensibility of the discovered knowledge. Attribute Selection – not all attributes are relevant and so for selecting a subset of attributes relevant for mining, among all original attributes, attribute selection [8] is mandatory.

2. CLASSIFICATION TECHNIQUES

The most commonly used data mining technique is the classification that occupies a set of pre-classified patterns to develop a model that can categorize the population of records at large. The learning and classification is involved by the process called the data classification. By the classification algorithm, the training data are analyzed in the learning [8] [9]. The approximation of the classification rules, the test data are used in the classification. To the new data tuples, the rules can be applied when the accuracy is acceptable. To verify the set of parameters which is needed for the proper discrimination, the pre-classified examples are used in the classifier-training algorithm. The model which is called as a classifier, only after these parameters encoded by the algorithm. Artificial neural network, Naive Bayesian classification algorithm classification techniques are used in this paper.

3. FEATURE SELECTION TECHNQIUES

Before In general, Feature subset selection is a pre-processing step used in machine learning [10]. It is valuable in reducing dimensionality and eradicates irrelevant data therefore it increases the learning accuracy. It refers to the problem of identifying those features that are useful in predicting class. Features can be discrete, continuous or nominal. On the whole, features are described in three types.

1) Relevant, 2) Irrelevant, 3) Redundant. Feature selection methods wrapper and embedded models. Moreover, Filter model rely on analyzing the general qualities of data and evaluating features and will not involve any learning algorithm, where as wrapper model uses après determined learning algorithm and use learning algorithms performance on the provided features in the evaluation step to identify relevant feature. The Embedded models integrate the feature selection as a part of the model training process.

The collection of datas from medical sources is highly voluminous in nature. The various significant factors distress the success of data mining on medical data. If the data is irrelevant, redundant then knowledge discovery during training phase is more difficult. Figure 2 shows flow of FSS.

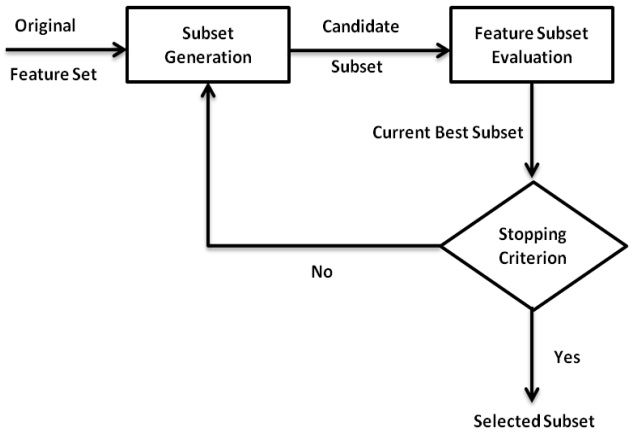

Fig -1: Feature Subset Selection

4. ARTIFICIAL NEURAL NETWORK

Moreover, the realistic provisional terms in neural networks are non-linear statistical data modelling tools. For discovering the patterns or modelling the complex relationships between inputs and outputs the neural network can be used. The process of collecting information from datasets is the data warehousing firms also known as data

mining by using neural network tool [11]. The more informed decisions are made by users that helping data of the cross-fertilization and there is distinction between these data warehouse and ordinary databases and there is an authentic manipulation.

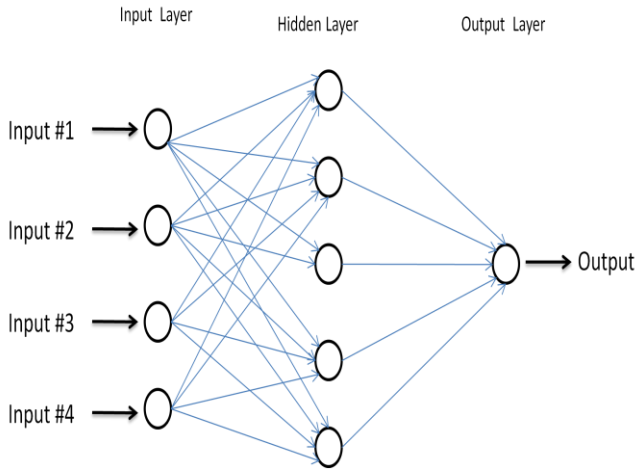

Fig -2: Example Artificial Neural Network

Among the algorithms the most popular neural network algorithms are Hopfield, Multilayer perception, counter propagation networks, radial basis function and self organizing maps etc. In which, the feed forward neural network was the first and simplest type of artificial neural network consists of 3 units input layer, hidden layer and output layer. There are no cycles or loops in this network. A neural network [12] has to be configured to fabricate the required set of outputs. Basically there are three learning conditions for neural network. 1) Supervised Learning, 2) Unsupervised Learning, 3) Reinforcement learning the perception is the basic unit of an artificial neural network used for classification where patterns are linearly separable. The basic model of neuron used in perception is the Mcculluchpitts model. The learning for artificial neural networks are as follows:

Step 1: Ret D= {{Xi,Yi}/i=1, 2, 3----n}} be the set of training example.
Step 2: Initialize the weight vector with random value,W (o).
Step 3: Repeat.
Step 4:For each training sample (Xi, Yi) ∈ D.
Step 5: Compute the predicted output Yi^ (k)
Step 6: For each weight we does.
Step 7: Update the weight we (k+1) = Wj(k) +(y i -yi^ (k))xij.
Step 8: End for.
Step 9: End for.
Step 10: Until stopping criteria is met.

5. NEURAL FUZZY ROUGH SET EVALUATION

This particular technique has been implemented by us in previous research paper [13] and in addition we have implemented the further scope of the research in this paper. 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 [14]. The knowledge of value of the conditional attribute Dj completely predicts the value of the decision feature E and it is indicated by the value of 1 and Dj and E values which are independent, then the attribute value Dj is irrelevant and it is indicated by the value zero [15]. 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 Dk = getFirstElement (SB).
Step 5: Then go to begin stage
Step 6: for each feature DK in SN
Step 7: If (σ DK (SB) <σ (SB)
Step 8: SN →Dk ; new old { } SB = SB ∪ Dk
Step 9: SB = max{I (SBnew),I (SBold)}
Step 10: Dk = getNextElement(SB) ;
Step 11: End until (Dk == null)
Step 12: Return SB;
End;

6. PROPOSED ALGORITHM

We proposed a new hybrid feature selection method by combing 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.

Input: T(K1; K2; :::; KM;L) // a training data set δ // a predefined threshold
output: Tbest {Abest(highest IG)} // an optimal subset
Step 1: begin
Step 2: for j= 1 to M do begin
Step 3: calculate TUj;l for Kj;
Step 4: if (Tuj,l ≥ δ)
Step 5: append Kj to T'list;
Step 6: end;
Step 7: order T'list in descending TUj,l value;
Step 8: Kp = getFirstElement(T'list);
Step 9: do begin
Step 10: Kq = getNextElement(T'list , Kp);

Step 11: if (Kq <> NULL)
Step 12: do begin
Step 13: K'q = Kq;
Step 14: if (TUp,q ¸ TUq,l)
Step 15: remove Kq from T'list ;
Step 16: Kq = getNextElement(T'list , K 'q);
Step 17: else Kq = getNextElement(T'list , Kq);
Step 18: end until (Kq == NULL);
Step 19: Kp = getNextElement(T'list , Kp);
Step 20 end until (Kp == NULL);
Step 21 Tbest = T'list ;
Step 22 Tbest={X1,X2,..XN}
Step 23 for j=1 to N begin
Step 24 for k=j+1 to N begin
Step 25 P[Lm/(Xj,Xk)]=P[(Xj,Xk)/Lm]*P(Lm)
Step 26 P[L/(Xj,Xk)]= P[L1/(Xj,Xk)]+P[L2/(Xj,Xk)]+… ....+P[Ln/(Xj,Xk)]
Step 27 If(P[L/Xj,Xk)]>Ω)
Step 28 {
Step 29 if((P[L/Xj]>(P[L/Xk])
Step 30 Remove Xk from Tbest
Step 31 Tbest=IG(X)
Step 32 Else
Step 33 Remove Xj from Tbest
Step 34 Tbest=IG(X)
Step 35
Step 36 }
Step 37 end;

7. DATASET

The following dataset attributes is used for predicting the PCOS (Polycystic Ovarian Syndrome) disease for the women at earlier stage. The dataset is taken from the [16]. The dataset contains 26 attributes and 303 instances. The following table is the PCOS dataset.

Table 1: Attributes of PCOS dataset

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

8. EXPERIMENTAL RESULT

The experiment is done using Orange Data mining tool. The proposed algorithm is fed into the user defined coding block for the execution

Table 2: Number of Original and Reduced Attributes

Attribute Filtering Method	Total Number of Attributes	Number of Reduced Attributes
Neural Fuzzy Rough Set	26	5
ANN+NFRS	26	2

Table 3: Reduced Attributes by Neural Fuzzy Rough Set and Neural network

Attribute Filtering Method	Number of attributes	Number of filtered Attributes
ANN+NFRS	**26**	**2**

Table 4: Classifier Accuracy with full dataset

S.No	Classifiers	Correctly Classified Instances	Incorrectly Classified Instances	Accuracy of the classification
1	SVM	232	71	76.45%
2	ANN	254	49	83.70%
3	Classification Tree	228	75	75.25%
4	Naïve Bayes	251	52	82.75%

Table 5: Number of filtered attributes by each attribute filtering method

Attribute Filtering Method	Number of attributes filtered
Correlation based Feature Selection (CFS)	3 (3,5,4)
Information Gain	9 (3,5,2,21,4,15,19,11,6)
Gain ratio	10 (3,5,21,4,15,19,11,30,29,26)
Neural Fuzzy Rough Set (NFRS)	5 (2,3,11,30,4)
Principle Component Analysis (PCA)	11 (3,5,21,4,15,19,2,11,30,23,29)

Table 6: Accuracy of classifiers with filtered attributes

Attribute Filtering Method	Acc. of SVM (in %)	Acc. of ANN (in %)	Acc. Classification Tree (in %)	Acc. of Naïve Bayes (in %)	Aver-age (in %)
CFS	80.25	83.30	76.75	82.10	80.60
Informa-tion gain	72.65	75.50	73.30	74.25	73.92
Gain Ratio	71.45	74.85	72.60	73.35	73.10
NFRS	80.85	85.25	79.50	84.45	82.52
PCA	69.65	72.35	70.75	71.55	71.07

Table 7: Hybrid Feature Selection

Attribute Filtering Method	Filtered Attributes
NFRS Table 7: Hybrid Feature Selection +ANN	2 (2,3)

Table 8: Accuracy of classification after NFRS+ANN

Classification Algorithm	Accuracy (in %)
Support Vector Machine	76.58
Artificial Neural Network	83.83
Classification Tree	75.23
Naïve Bayes	82.85

Table 9: Comparison of accuracy of classifiers with NFRS+ANN algorithm as feature selector

Classification Method	Correctly Classified Instance	Accuracy (in %)
Support Vector Machine	232	76.45%
Artificial Neural Network	254	83.70%
Classification Tree	228	75.25%
Naïve Bayes	251	82.75%

9. RESULT AND ANALYSIS

The above experiment is done using PCOS disease dataset which contains 26 attributes and 303 instances. From the table 1, the technique NFRS (Neural Fuzzy Rough Set) evaluation gives 5 filtered attributes out of 26 attributes, whereas, when NFRS is merged with artificial neural network (ANN), it generates the reduced number of attributes 2 than the NFRS. And table 2 also gives the result as same as the above. Table 3 gives the classification accuracy with full dataset. From the table 3, Artificial Neural Network (ANN) gives the more accuracy result than the SVM (Support Vector Machine), Classification Tree and Naïve Bayes algorithm. Table 4 gives the result of the various attribute filtering methods, from that result, we can conclude that the Neural Fuzzy Rough Set (NFRS) generates the least number of filtered attributes than the other techniques like CFS, PCA, Information gain and gain ratio. Ratio The table 5 gives result for the accuracy of classifiers with filtered attributes. From the table 5, we can conclude that the feature selection technique NFRS and classification technique of ANN gives the more accuracy classification result than the others. Table 6 gives the result of the hybrid feature selection method. In the table 7, after the merging of NFRS+ANN, the result of the classification accuracy of ANN is more than the others. And also from table 8, after the NFRS+ANN, the classified accuracy of the classification for ANN is higher than the other techniques.

10. CONCLUSION

In this paper, from the classification techniques like Artificial Neural Network (ANN), Support Vector Machine (SVM), classification tree and Naïve Bayes algorithm, the ANN classification techniques gives the more classification result than the others. The attribute filtering Neural Fuzzy Rough Set (NFRS) generates the less number of attributes than the Correlation based Feature Selection (CFS), Information gain method, Gain ratio method and Principle Component Analysis (PCA). In this paper, we proposed a new method using NFRS and ANN. This proposed method gives the more accuracy in the classification result as well as attribute filtering. From this paper, we can conclude that instead of doing the classification and feature selection separately we can do together for the best result for predicting the PCOS disease among the women using PCOS data set.

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.no:1-4.
[4]. Hian Chye Koh and Gerald Tan, “Data Mining Applications in Healthcare”, Journal of Healthcare Information Management — Vol. 19, No. 2, pp.no: 64-72.\
[5]. S.Saravanakumar, S.Rinesh, “Effective Heart Disease Prediction using Frequent Feature Selection Method”, International Journal of Innovative Research in Computer and Communication Engineering, Vol.2, Special Issue 1, March 2014, pp.no: 2767-2774.
[6]. Jyoti Soni, Uzma Ansari, Dipesh Sharma and Sunita Soni, “Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers”, International Journal on Computer Science and Engineering (IJCSE), Vol. 3 No. 6 June 2011, pp.no:2385-2392.
[7]. Aieman Quadir Siddique, Md. Saddam Hossain, “Predicting Heart-disease from Medical Data by Applying Naïve Bayes and Apriori Algorithm”, International Journal of Scientific & Engineering Research, Volume 4, Issue 10, October-2013, pp.no: 224-231.
[8]. Vikas Chaurasia, Saurabh Pal, “Early Prediction of Heart Diseases Using Data Mining Techniques”, Carib.j.SciTech,2013,Vol.1, pp.no:208-217.
[9]. Shamsher Bahadur Patel, Pramod Kumar Yadav, Dr. D. P.Shukla, “Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques”, IOSR Journal of Agriculture and Veterinary Science (IOSR-JAVS), Volume 4, Issue 2 (Jul. - Aug. 2013), PP 61-64.

[10]. M. Anbarasi, E. Anupriya, N.CH.S.N.Iyengar, "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm", International Journal of Engineering Science and Technology, Vol. 2(10), 2010, 5370-5376.
[11]. D Ratnam, P HimaBindu, V.Mallik Sai, S.P.Rama Devi, P.Raghavendra Rao, "Computer-Based Clinical Decision Support System for Prediction of Heart Diseases Using Naïve Bayes Algorithm", International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, pp.no: 2384-2388.
[12]. Selvakumar.P, DR.Rajagopalan.S.P, "A Survey On Neural Network Models For Heart Disease Prediction", Journal of Theoretical and Applied Information Technology, 20th September 2014. Vol. 67 No.2, pp.no:485-497.
[13]. Dr. K. Meena, Dr. M. Manimekalai, 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.
[14]. 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.
[15]. Lei Yu, Huan Liu, "Efficient Feature Selection via Analysis of Relevance and Redundancy" Journal of Machine Learning Research 5 (2004) 1205–1224.
[16]. PCOS Dataset Source - ftp://ftp.ncbi.nlm.nih.gov/geo/datasets/GDS4nnn/GDS4987/