A ROADMAP TO A CLINICAL PREDICTION MODEL WITH COMPUTATIONAL INTELLIGENCE FOR PCOS

Dr.V.Krishnaveni

Associate Professor in Computer Science,
Kongu Arts and Science College (Autonomous),Erode, Tamilnadu – 638107

ABSTRACT

In recent times, the application of computational or machine intelligence in medical diagnosis is a new trend for large medical data applications. A Computer- Aided decision making system can assist a physician in diagnosing diseases of a patient through systematized intelligent data classification approaches. Data Mining techniques and algorithms such as clustering, classification, SVM and Naive Bayes algorithm have been used for early diagnosis of chronic diseases , to predict heart diseases and liver problems. They have also been used for diagnosis and prognosis of breast cancer and diabetic. Polycystic Ovarian Syndrome (PCOS) is a condition which leads to growth of ovarian cysts. It is a prevalent endocrine disorder in child-bearing women, which can lead to infertility , dangerous illness like Type 2 Diabetes-Gestational Diabetes, Cardiovascular and Cerebrovascular dysfunction and higher risk of mood and anxiety disorders. A recent study has revealed that about 18% of the women in India suffer from this syndrome. Even though PCOS has been identified as the most common endocrinal disease , the author has found that there are a very limited number of researches initiated towards building a prediction model for PCOS. Hence, In this paper , the author has outlined a prediction model for PCOS to assist the physicians for their final decision on their patients.

Keywords- Data mining, Classification, Polycystic Ovarian Syndrome

1. Introduction

In the era of personalized medicine, prediction of prevalent or incident diseases (diagnosis) or outcomes for future disease course (prognosis) became more important for patient management by health-care personnel. Clinical prediction models are used to investigate the relationship between future or unknown outcomes (endpoints) and baseline health states (starting point) among people with specific conditions. They generally combine multiple parameters to provide insight into the relative impacts of individual predictors in the model.

Clinical prediction models can inform patients and their physicians or other healthcare providers of the patient's probability of having or developing a certain disease and help them with associated decision-making (e.g., facilitating patient-doctor communication based on more objective information). Applying a model to a real world problem can help with detection or screening in undiagnosed high-risk subjects, which improves the ability to prevent developing diseases with early

interventions. Furthermore, in some instances, certain models can predict the possibility of having future disease or provide a prognosis for disease (e.g., complication or mortality). The Machine learning based models are a natural extension of classical statistical approaches but provide more effective methods to analyse very large datasets. In addition, the predictive capability of such models promises to be useful in developing decision support systems. That is, they can guide, diagnose, classify and personalize the treatment. The Physician roles will then likely change to more of a consultant than decision maker, who will advise, warn and help individual patients[21][22][26].

Polycystic ovarian syndrome is an endocrine system disorder with a collection of symptoms that are found as a result of a broad- spectrum hormonal disturbance. It is considered as one of the most common endocrine disorder in women at their reproductive age and a leading cause for infertility. PCOS commonly manifests during adolescence, and it's primarily characterized by ovulatory dysfunction and hyperandrogenism. Moreover, PCOS could lead to numerous complications that may heavily affect woman's health and the quality of life. Women with PCOS experience a diversity of symptoms/complications involving different systems, such as, gynecological disorders, (failure to ovulate, late menopause, endometrial cancer and infertility), metabolic (insulin resistance, diabetes type 2, dyslipidemia), cardiac (hypertension, and cardiovascular diseases), physical (central obesity, acne, hirsutism, hair loss and baldness), and psychological (depression, stress and anxiety). Prevalence of PCOS is highly variable ranging from 2.2% to 26% globally. In few Asian countries prevalence figures are ranging from 2% to 7.5% in China and 6.3% in Srilanka. There are few studies conducted in India. Studies done in South India and Maharashtra, prevalence of PCOS (by Rotterdam's criteria) were reported as 9.13% and 22.5% (10.7% by Androgen Excess Society criteria) respectively[1-3].

Early diagnosis of PCOS is important as it has been linked to an increased risk for developing several medical conditions including insulin resistance, type 2 diabetes, high cholesterol, high blood pressure and heart disease. PCOS is an emerging health problem during adolescence therefore promotion of healthy lifestyles and early interventions are required to prevent future morbidities.[4]

The research in Data mining has entered the age of 'Big Data'. The Medical Datasets now routinely involve thousands of heterogeneous attributes related to symptoms, medical history, clinical information, Blood test and hormone analysis of the patients. The analysis of these datasets is challenging, especially when the number of measurements exceeds the number of individuals, and may be further complicated by missing data for some subjects and variables that are highly correlated.

This paper suggests a roadmap for building a model for predicting those who are susceptible to the Poly-Cystic Ovary Syndrome (PCOS) by applying computing intelligence and data mining techniques.

2. The Framework of the model

The main components of the proposed predictive model for PCOS are (i) building of a sample dataset of clinical information of patients with and without PCOS from the Health Care Professionals, (ii) selection of the features relevant to the solution, (iii) building of a predictive model(using a Machine learning technique) by using the sample as a Training dataset,(iv) analysis of the quality of the model by using validation testing

2.1 Building a sample dataset with clinical information

The attributes related to PCOS have been observed and selected from various research papers and have been considered for building a dataset for the prediction model. The data set might contain data from five different categories. They are (a) Symptoms, (b) Medical History, (c) Clinical Signs and (d) Scan Findings and (e) Blood Test and Harmone Analysis[5,6,10,11,12,13]. The possible attributes of dataset have been listed as follows:

(a) Symptoms

Excess hair growth on the face, chest, stomach, thumbs, or toes, Baldness or thinning hair, Acne, Oily skin or dandruff , Patches of thickened dark brown or black skin, Obesity, Sleep apnea, Depression anxiety ,eating disorders, Pelvic pain, Menstrual problems

(b) Medical History

Diabetes , high blood pressure, heart disease ,endometrial cancer, Miscarriage or premature birth, Gestational diabetes or pregnancy-induced high blood pressure, Metabolic syndrome Type 2 diabetes or prediabetes, Cancer of the uterine lining (endometrial cancer)

(c) Clinical Signs

Amenorrhea ,oligomenorrhea , infertility , virilism, Anovulation

(d) Scan Findings

Presence of multiple cysts on the ovaries, enlarged ovaries

(e) Blood Test and Hormone analysis

Follicle-stimulating hormone (FSH) , Luteinizing hormone (LH) ,Testosterone , Estrogens a, sex hormone binding globulin (SHBG) , androstenedione , Human chorionic gonadotropin (hCG), Anti-Mullerian hormone (AMH):, Lipid profiles , Glucose test , Insulin resistance, DHEA, S/Testosterone, Lutenizing Hormone (LH) and Follicle Stimulating Hormone (FSH), Progesterone, Estradiol, Hemoglobin A1c, Cortisol, Fasting Insulin, Fasting Glucose, Prolactin, Thyroid Markers

2.2 Feature selection

Feature selection is an act of identifying a small subset of features to be employed for classification when a classification problem involving a large set of potential features. The data without feature selection may be redundant or noisy, and may degrade the accuracy rate of classification [36]. The main advantages of feature selection are as follows: (1) lowering computational cost and storage requirements, (2) minimizing the degradation of classification accuracy rate because of the finite

nature of training sample sets, (3) decreasing training and prediction time and, (4) facilitating understanding and visualization of data and (5) lowering the cost of clinical tests.

The optimal feature selection process has been considered as a NP-hard problem since the number of features is significantly large and only heuristics approaches are able to deal with them [39]. Therefore, the researchers have explored the use of heuristic algorithms for feature subset selection in which the features were either selected or eliminated sequentially to determine the final feature subset [7-9]. Heuristics such as mutual information, relevance, and relevance of each feature were also employed to find the optimal feature subset [14-18]. Several researchers have also explored the use of randomized population-based heuristic search techniques for feature subset selection such as Genetic algorithm, Particle Swarm Optimization, Ant Colony Optimization, Artificial Bee colony optimization and Harmony Search Optimization.

2.3 Building a predictive model

There are two important processes in building a predictive model. They are,

(1) choosing the right criteria to solve the problem and

(2) selecting the appropriate prediction method from the various Data mining techniques.

2.3.1 Different criteria for diagnosis of PCOS

There are different criteria available for diagnosing the PCOS. They are listed as follows:

Rotterdam criteria (2004) include at least two of the following three conditions: Clinical and/or biochemical hyperandrogenism (androgen excess), oligo-ovulation or anovulation (irregular ovulation or absence of ovulation), and polycystic ovaries (12 or more follicles in at least 1 ovary).

(Exclusion of all other disorders that can result in menstrual irregularity and hyperandrogenism, including adrenal or ovarian tumors, thyroid,dysfunction, congenital adrenal hyperplasia, hyperprolactinemia, acromegaly, and Cushing syndrome)

Androgen Excess Society (AES), 2006 considered PCOS as a primarily disorder of androgen excess or hyperandrogenism and defined PCOS by the presence of hyperandrogenism (clinical and/or biochemical), ovarian dysfunction and/or polycystic ovaries, and the exclusion of related disorders.

According to American Association of Clinical Endocrinologists (AACE) and the Androgen Excess and PCOS Society (AES) 2015, diagnosis of PCOS is based on presence of at least two of the following three conditions: chronic anovulation (Cycle length >35 days), hyperandrogenism (clinical or biological) and polycystic ovaries.

Tests for evaluation of PCOS:

Follicle stimulating hormone (FSH)

Lutenizing hormone

Testosterone

Estrogens

Sex hormone binding globulin (SHBG)

Androstenedione

Human chorionic gonadotropin (HCG)

Anti-Mullerian hormone

Tests to rule out other conditions with similar signs and symptoms:

Thyroid stimulating hormone (TSH) – to rule out thyroid dysfunction.

Cortisol– to rule out Cushing syndrome .

Serum Prolactin – to rule out elevated prolactin (hyperprolactinemia)

17-hydroxyprogesterone – to rule out the most common form of congenital hyperplasia.

Serum free insulin like growth factor-1 (IGF-1) – to rule out excess growth hormone

Dehydroepiandrosterone Sulfate (DHEAS) – to rule out an adrenal tumor

Other blood tests to check a woman's health and detect any complications:

Lipid profile - to help determine risk of developing Cardiovascular disease ; risk is associated with a low High density lipoproteins (HDL), high low density lipoproteins (LDL), high total cholesterol and/or elevated triglycerides (Dyslipidemia).

Glucose or Hemoglobin A1c (HbA1c) – can be used to detect diabetes; elevated in diabetes.

Insulin– often elevated in insulin resistance.

Imaging tests:

The following imaging studies may be used in the evaluation of PCOS:

Ovarian ultrasonography, preferably using transvaginal approach.

2.3.1 Selection of the appropriate data mining method

Classification techniques in healthcare: The objective of the classification is to assign a class to find previously unseen records as accurately as possible. If there is a collection of records (called as training set) and each record contains a set of attributes, then one of the attributes is class. The motive is to find a classification model for class attributes, where a test set is used to determine the accuracy of the model.[31]

The given data set is divided into training and test sets. The training set used to build the model and test set is used to validate it. Classification process consists of training set that are analyzed by a classification algorithms and the classifier or learner model is represented in the form of classification rules. Test data are used in the classification rules to estimate the accuracy. The learner model is represented in the form of classification rules, decision trees or mathematical formulae. Decision trees can be used to classify new cases. They can construct explicit symbolic rules that generalize the training cases. New cases can then be classified by comparing them to the reference cases.

There are many classification techniques and algorithms existing to build a predictive model, out of which some have been listed below: -[27]

1. Rule Induction 2. Support Vector Machine (SVM) Algorithm 3. ID3 Algorithm

4. C4.5 Algorithm 5. Naïve Bayes Algorithm 6. Artificial Neural Networks (ANN) Algorithm 7. K nearest neighbour algorithm

A brief discussion on the classification techniques most commonly used in Health Care Systems are given below.

Rule induction:[23] is the process of extracting useful 'ifthen' rules from data based on statistical significance. A Rule based system constructs a set of if-then-rules. Knowledge represents has the form

IF conditions THEN conclusion

In the health care system it can be applied as follows:

(Symptoms) (Previous--- history) ------- > (Cause—of--- disease)

Decision tree:

It is a knowledge representation structure consisting of nodes and branches organized in the form of a tree such that, every internal non-leaf node is labeled with values of the attributes. The branches coming out from an internal node are labeled with values of the attributes in that node. Every node is labeled with a class (a value of the goal attribute).

Decision tree models are well suited for developing healthcare systems. They are inexpensive to construct, easy to interpret, easy to integrate with database system and they can produce comparatively better accuracy than other methods. There are many Decision tree algorithms such as HUNTS, CART, ID3, C4.5, SLIQ, and SPRINT.

Artificial neural networks (ANN)

Artificial neural networks (ANN) provide a powerful tool to help doctors analyze, model and make sense of complex clinical data across a broad range of medical applications. In medicine, ANNs have been used to analyze blood and urine samples, track glucose levels in diabetics, determine ion levels in body fluids and detect pathological conditions. A neural network has been successfully applied to various areas of medicine, such as diagnostic aides, medicine, biochemical analysis, image analysis and drug development [30].

2.4 Analysing the quality of the model by using validation testing

There are many Statistical Measures existing for Model Evaluation. They are, Sensitivity and specificity (Discrimination (ROC/AUC)), Predictive values: positive, negative(Likelihood ratio: [positive, negative]), Accuracy: Youden index, Brier score (Number needed to treat or screen), Calibration: Calibration plot, Hosmer-Lemeshow test(Model determination: R2) Statistical significance: P value (Magnitude of association, e.g., β coefficient, odds ratio) Model quality: AIC/BIC (Net reclassification index and integrated discrimination improvement) Net benefit (Cost-effectiveness).

Where

ROC - receiver operating characteristic;

AUC-area under the curve;

AIC- Akaike information criterion;

BIC- Bayesian information criterion.

Further, it is necessary to do separate internal and external validation on the Datasets to finalize the research findings . Internal validation can be done using a random sub sample or different years from

the development dataset or by conducting bootstrap resampling. This approach can particularly assess the stability of selected predictors, as well as prediction quality. Subsequently, external validation should be performed on an independent dataset from that which was previously used to develop the model. For example, datasets can be obtained from populations from other hospitals or centers or a more recently collected population . This process is often considered to be a more powerful test for prediction models than internal validation because it evaluates transportability, generalizability and true replication, rather than reproducibility.

3. Conclusion

For patient-centered perspectives, clinical prediction models are useful for several purposes: to screen high-risk individuals for asymptomatic disease, to predict future events of disease or death, and to assist medical decision-making. Data mining, a field that can uncover patterns from large repositories, has numerous applications such as building predictive models which can be extremely beneficial in the field of healthcare. In this paper, the author has proposed a road map to a prediction model for PCOS which is the most common endocrine disorder in women of reproductive age. The Road map is a four step plan and if followed strictly will guarantee an efficient prediction model for PCOS.

Acknowledgement

This work is a part of a Minor Research Project funded by University Grants Commission, New Delhi, India

References

[1]. Sunanda B, Nayak S. A study to assess the knowledge regarding PCOS (polycystic ovarian syndrome) among nursing students at NUINS. NUJHS. 2016;6(3).

[2]. Broder-Fingert S, Shah B, Kessler M, Pawelczak M, David R. Evaluation of adolescents for polycystic ovary syndrome in an urban population. J Clin Res Pediatr Endocrinol. 2009;1(4):188-93.

[3]. Sanchez N. A life course perspective on polycystic ovary syndrome. Int J Womens Health. 2014;6:11522. 5

[4]. Sunita J. Ramanand, Balasaheb B. Ghongane, Jaiprakash B. Ramanand, Milind H. Patwardhan,2 Ravi R. Ghanghas,and Suyog S. Jain "Clinical characteristics of polycystic ovary syndrome in Indian women"Indian J Endocrinol Metab. 2013 Jan-Feb; 17(1): 138-145

[5]. Suhail A.R. Doia, Mona Al-Zaidb, Philip A. Towersc, Christopher J. Scottc,Kamal A.S. Al-Shoumera, “Steroidogenic alterations and adrenal androgen excess in PCOS”, ELSEVIER Science direct, steroids 71 (2006) 751–759

[6]. Antoni J. Duleba, “Medical management of metabolic dysfunction in PCOS”, ELSEVIER Steroids, Steroids 77 (2012) 306-311

[7]. P. Narendra and K. Fukunaga, “A branch and bound algorithm for feature subset selection”, IEEE Transactions on Computers, Vol. 26, 1977, pp. 917-922.

[8]. I. Foroutan and J. Sklansky, “Feature selection for automatic classification of non gaussian data”, IEEE Transactions on Systems, Man and Cybernetics, Vol. 17, 1987, pp. 187-198.

[9]. D. Koller and M. Sahami, “Hierarchically classifying documents using very few words”, Proceedings of International Conference on Machine Learning, 1997, pp. 535-539.

[10]. Robert A. Wild, “Dyslipidemia in PCOS”, ELSEVIER Steroids, Steroids 77 (2012) 295-299

[11]. Frank González, “Inflammation in Polycystic Ovary Syndrome: Under Ipinning of insulin resistance and ovarian dysfunction”, ELSEVIER Steroids, Steroids 77 (2012) 300-305

[12]. Anuja Dokras, “Mood and anxiety disorders in women with PCOS”, ELSEVIER Steroids, Steroids 77 (2012) 338-341

[13]. Tristan S.E. Hardy, Robert J. Norman, “Diagnosis of adolescent polycystic ovary syndrome”, ELSEVIER Steroids, Steroids 77 (2012) 751-754

[14]. H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information: criteria of maxdependency, max-relevance, and min-redundancy”, IEEE Transactionson Pattern Analysis and Machine Intelligence, Vol. 27, 2005, pp. 1226-1238.

[15]. Y. Sun and J. Li, “Iterative RELIEF for feature weighting: algorithms, theories, and applications”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 29, 2007, pp. 1035-1051.

[16]. G. Claeskens, C. Croux, and J. Kerckhoven, “An information criterion for variable selection in support vector machines”, Journal of Machine Learning Research, Vol. 9, 2008, pp. 541-558.

[17]. Z. Zhao and H. Liu, “Searching for interacting features”, Proceedings of the 2nd International Joint Conference on Artificial Intelligence, 2007, pp. 1156-1161.

[18]. R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, “Diagnosis of multiple cancer types by shrunken centroids of gene expression”, Proceedings of the National Academy of Sciences of the United States of America, Vol. 99, 2002, pp. 6567-6572.

[19]. Roy Homburg, “Pregnancy complications in PCOS”, ELSEVIER, Best Practice & Research Clinical Endocrinology & Metabolism, Vol. 20, No. 2, pp. 281–292, 2006

[20]. Ghada El-Kannishya, Shaheer Kamala, Amany Mousaa, Omayma Saleha,Adel El Badrawyb, Reham El farahatyc, Tarek Shokeird, “Endothelial function in young women with polycystic ovary syndrome (PCOS): Implications of body mass index (BMI) and insulin resistance”, ELSEVIER, Obesity Research & Clinical Practice (2010) 4, e49—e56

[21]. Divya Tomar and Sonali Agarwal, “A survey on Data Mining approaches for Healthcare” International Journal of Bio-Science and Bio-Technology Vol.5, No.5 (2013), pp. 241-266.

[22]. V. Manikantan and S. Latha, “Predicting the Analysis of Heart Disease Symptoms Using Medicinal Data Mining Methods” , International Journal on Advanced Computer Theory and Engineering (IJACTE), Volume-2, Issue-2, 2013 2319 – 2526.

[23]. Mahmood A. Rashid, Md Tamjidul Hoque and Abdul Sattar, “Association Rules Mining Based Clinical Observations”, pp.1-5.

[24]. Anoop Jain, Aruna Bajpai and Manish Kumar Rohila, “Efficient Clustering Technique for Information Retrieval in Data Mining”, International Journal of Emerging Technology and Advanced Engineering, pp.12-20.

[25]. K. Rajeswari, Mahadev Shindalkar, Nikhil Thorawade and Pranay Bhandari, “DSS Using Apriori Algorithm, Genetic Algorithm And Fuzzy Logic‖, Journal of Engineering Research and Applications (IJERA)”, Vol. 3, Issue 4, Jul-Aug 2013, pp.132-136.

[26]. R.Saravana Kumar and G.Tholkappia Arasu, "Rough Set Theory And Fuzzy Logic Based Warehousing Of Heterogeneous Clinical Databases", pp.1-22.

[27]. Tapas Ranjan Baitharu and Subhendu Kumar Pani, "A Survey on Application of Machine Learning Algorithms on Data Mining", International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-3, Issue-7, December 2013, pp. 17-20.

[28]. George D. Magoulas and Andriana Prentza, "Machine Learning In Medical Applications", pp.1-7.

[29]. Markus Brameier and Wolfgang Banzhaf, "A Comparison of Linear Genetic Programming and Neural Networks in Medical Data Mining", IEEE Transactions On Evolutionary Computation, Vol. 5, No. 1, February 2001, 17-26.

[30]. Dr. Yashpal Singh and Alok Singh Chauhan, "Neural Networks In Data Mining", Journal of Theoretical and Applied Information Technology, pp. 37-42.

[31]. Vanitha.L and Venmathi.A, "Classification of Medical Images Using Support Vector Machine" 2011 International Conference on Information and Network Technology IPCSIT vol.4 (2011), pp. 63-67.

[32]. D.Lavanya and Dr. K.Usha Rani, "Performance Evaluation of Decision Tree Classifiers on Medical Datasets", International Journal of Computer Applications (0975 – 8887) Volume 26– No.4, July 2011, pp. 1-4.

[33]. Prof. Dr. P. K. Srimani and Manjula Sanjay Koti, "Outlier Mining In Medical Databases By Using Statistical Methods", International Journal of Engineering Science and Technology (IJEST), Vol. 4 No.01 January 2012, pp. 239-246.

[34]. Sona Baby and Ariya T.K, "A survey paper of data mining in medical diagnosis", International Journal of Research in Computer and Communication Technology (IJRCCT) ,98-101 2014

[35]. S. Kiruthika Devi, S. Krishnapriya and Dristipona Kalita," Prediction of Heart Disease using Data Mining Techniques", Indian Journal of Science and Technology, Vol 9(39), DOI: 10.17485/ijst/2016/v9i39/102078, October 2016

[36]. Shih-wei Lin, Shih-Chieh Chen, " PSOLDA: A Particle Swarm Optimization approach for enhancing classification accuracy rate of linear discriminant analysis", Applied Soft Computing (2009) 1008-1015

[37]. SHIV SHAKTI," A Review on Data Mining Techniques Used in Healthcare Industry", International Journal in Multidisciplinary and Academic Research (SSIJMAR) Vol. 3, No. 1, February- March -2014 (ISSN 2278 – 5973)

[38]. Manaswini Pradhan," Analysis of Data Mining Techniques for Building Health Care Information System", International Journal of Engineering Technology, Management and Applied Sciences, January 2016 , Volume 4, Issue 1, ISSN 2349-4476

[39]. Mao, K. Z. " Feature subset selection for support vector machines through discriminative function pruning analysis", IEEE Transactions on Systems, Man, and Cybernetics, 34(1), 60–67,2004.

[40]. Abdullah A. Aljumah, Mohammed Gulam and Mohammad Khubeb Siddiqui," Application of data mining: Diabetics health care in young and old patients", Journel of king Saud University-Computer and Information Sciences,May 2012