AN IMPORTANCE OF PREPROCESSING THE MEDICAL DATA FOR EXTRACTING INFLUNTIAL FEATURE

Nandhakumar R[1] Ramaprabha T[2]

[1]PG and Research Department of Computer Science and Applications, Vivekanandha College of Arts and Sciences for Women (Autonomous), Elayampalayam, Tiruchengode
[2]Associate Professor, Department of Information Technology, Nehru Arts and Science College, Tirumalayampalayam. Coimbatore., Tamilnadu , India

Abstract *- In modern trend world, the every field have produced massive data and it may help to feature extraction. Based on the data the technical area of data mining will mining the data and produce a most successful prediction result for the upcoming problems and growth all other field. In the area of medical science produces a lot of massive data either organized or unorganized. As a technical data worker need to segregate the data and making an organized one. The pre-processing of medical data for feature reduction and extracting the influential parameter for prediction is the hectic task. In this paper, the preprocessing techniques applied in the In Vitro Fertilization (IVF) data to clean and find the influential data for classification and prediction. IVF data is nothing but the infertility treatment data which who move on Artificial Insemination. Based on the patients test results and survey of personal details that are combine into a formatted and it have 19 attributes. The preprocessing techniques like Genetic Algorithm (GA), Ant Colony Optimization (ACO) and Relative Reduct Algorithm (RR) are applied in the data for feature selection as well as removing the noisy data and redundant data from the data set. So from this study we can find which algorithm act effectively for reduces unwanted data and produce the highly influential data for classification.*

1.INTRODUCTION

The healthcare domain is recognized for its ontological complication and diversity of medical data standards and variable data quality. Making an effective and practically usable discovery in medical data is of ongoing importance in current decades due to the addition of privacy issues in patient data. Computer- aided knowledge discovery methods plays an important role in transformation and understanding of concepts related to health and illness. From the twentieth century, many countries have chosen e-health as a prioritized national program [1]. It provides standardized aggregation of patients' clinical information and health care services by providing instant access to this information for healthcare professionals and patients too. Unequivocal methods, tools and methodology are needed for the application of Data Mining in health care. Both structured and unstructured data is available for research as a result of progress in the computerization of data in the health care industry [2]. Even though, there are number of algorithms are available to classification related to specific domains in health care is still to be resolved. The health care industry is one of the most information intensive industries. Health care is an intensive research field and largest consumes of public health. Human medical data are the most rewarding and difficult of all biological data to mine and analyze [3] [4]. When the database is large, it is very difficult for the medical researchers, physicians and health care providers to use the stored data more effectively. The medical database usually contains data such as patient records, physician diagnosis and monitoring information to save lives. The medical decision support system was designed to reduce medical errors and costs, earlier disease detection and to achieve preventive medicine [6]. To improve the decision

making in medical problems, the data mining has been used in medial domain. Even biomedical researchers and health care managers find very difficult to detect the association between the risk and outcomes. So based on the high massive data, the process of data analysis is difficult to extract the knowledge. In this work, the infertility patient data samples collected for the pre-processing to find the highly influential parameters that make easy to find the knowledge of patient success of the treatment [8]. The 19 parameters of patient's data are collected from various places of Tamil Nadu and the data are organized into find the data cleaning like pre-processing [9] [10]. In this regard three major pre-processing techniques like Genetic Algorithm, Ant colony optimization, Relative Reduct Algorithm are used to preprocess the data set and find the highly suitable technique for these kind of infertility data.

2. DATA SET

The dataset with the list of attributes as illustrated in Table 1 is taken for experiments as suggested by the subject experts and medical practitioners [11].

Table 1: List of attributes chosen for experimentation

S. No	Test Parameters	Data type
1	Age (Female)	Numeric
2	Endometriosis	Yes/No
3	Ovulatory Factor	Yes/No
4	Hormonal Factor	Numeric
5	Cervical Factor	Numeric
6	Unexplained Factor	Yes/No
7	Semen Ejaculate Volume	Numeric
8	Liquefaction Time	Numeric
9	Sperm Concentration	Numeric
10	Sperm motility	Numeric
11	Sperm vitality	Numeric
12	Sperm morphology	Numeric
13	No. of oocytes retrieved	Numeric
14	No. of embryos transferred	Numeric
15	Male factor only	Yes/No
16	Severe male factor	Yes/No
17	Female factor only	Yes/No
18	Combined factor	Yes/No
19	IVF Treatment Results	Success/Failure

Collections of 300 patient's data are feed by using Mat lab environment to find the highly influential parameter and check with the classification technique for high accuracy and minimal error values.

3. EXPERIMENTS

In this work, the following tables shows, each preprocessing technique's performance and find which technique produce high accuracy compare with Naive Bayes (NB) and Multi- Layer Perception Network(MLPN) classifiers [12] [13]. The selected features of the data set are tested with the algorithms. The reduced features are listed in Table2. Also Chart 1 shows the number of features obtained by applying the selected and proposed algorithms.

Table 2: List of Attributes Reduced by Preprocessing Techniques

Original Attributes	GA	ACO	RR
Age	✓	✓	
Endometriosis	✓	✓	✓
Ovulatory Factor	✓	✓	✓
Hormonal Factor	✓	✓	✓
Cervical Factor	✓	✓	✓
Unexplained Factor	✓	✓	✓
Semen Ejaculate Volume	✓	✓	✓
Liquefaction Time	✓		
Sperm Concentration	✓	✓	✓
Sperm motility	✓	✓	✓
Sperm vitality		✓	✓
Sperm morphology	✓	✓	✓
No. of oocytes retrieved			
No. of embryos transferred			
Male factor only	✓	✓	✓
Severe male factor			
Female factor only	✓	✓	✓
Combined factor			
IVF Treatment	✓	✓	✓

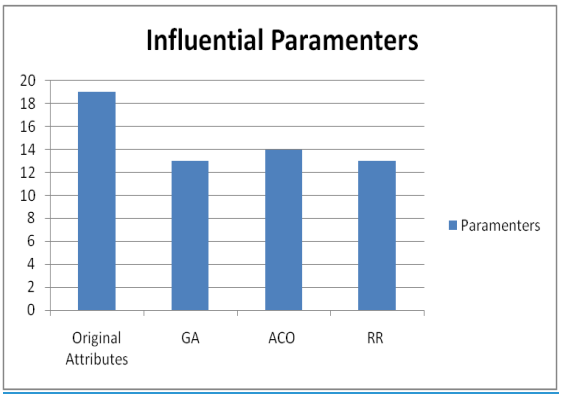

Chart 1: Attributes Reduced by Preprocessing Techniques

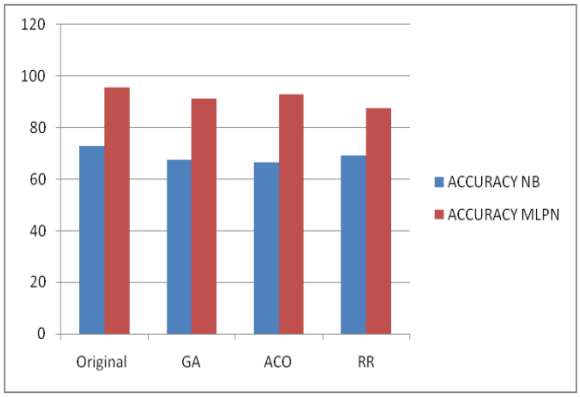

Chart 2: Comparison of accuracy of different algorithms with different classifiers

After reducing the number of attributes, the reduced data set is subjected to the classification. The existing classifiers like Naïve Bayes (NB) and Multi-Layer Perceptron Network (MLPN) are taken for comparative study [14]. The metrics like Accuracy, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE), Root Relative Absolute Error (RRAE), True Positive Rate (TPR), False Positive Rate (FPR), Precision, Recall, F-Measure, and Receiver Operating Characteristic (ROC) Area are calculated to measure the performance of the classifier [15]. These selected influential parameter set are subjected to different classifier and calculated accuracy. The results obtained are tabulated in Table 3 and illustrated in Chart 2.

Table 3: Comparison of accuracy of different algorithms with different classifiers

Data set with Attributes	Classification Algorithms	
	NB	MLPN
Original / Entire Attributes	72.81	95.61
Attributes obtained using GA	67.54	91.22
Attributes obtained using ACO	66.67	92.98
Attributes obtained using RR	69.30	87.72

Error Values like MAE, RMSE, RAE, RRAE for different Existing and proposed algorithms with different classifiers are measured. These measured error values are table 4 shows the comparison of error values with respect to attributes obtained by using preprocessing techniques.

Table 4 Comparison of Error Values of different algorithms with different classifiers

Dataset	Classifiers	Error Values			
		RRAE	RAE	RMSE	MAE
Original / Entire Attributes	NB	93.68	62.39	0.38	0.21
	MLPN	56.89	16.76	0.16	0.05
Attributes obtained using GA	NB	93.28	70.32	0.38	0.24
	MLPN	50.55	25.40	0.21	0.08
Attributes obtained using ACO	NB	96.07	69.24	0.39	0.23
	MLPN	43.97	23.00	0.18	0.077
Attributes obtained using RR	NB	92.03	67.14	0.37	0.22
	MLPN	34.11	36.07	0.23	0.12

Dataset	Classifiers	Error Values			
		RRAE	RAE	RMSE	MAE
Original / Entire Attributes	NB	93.68	62.39	0.38	0.21
	MLPN	56.89	16.76	0.16	0.05
Attributes obtained using GA	NB	93.28	70.32	0.38	0.24
	MLPN	50.55	25.40	0.21	0.08
Attributes obtained using ACO	NB	96.07	69.24	0.39	0.23
	MLPN	43.97	23.00	0.18	0.077
Attributes obtained using RR	NB	92.03	67.14	0.37	0.22
	MLPN	34.11	36.07	0.23	0.12

4. COMPARISION RESULT

Based on the values get from the Preprocessing and Classification, the data are processed in multiple manner. The each techniques specfically done their significatent menthod to data pre process and classification. Based on the classification accuracy and minimul error value the ACO and RR Algorithms are did high valued accuray and low level of error rate in the place of MLPN Classification. So these techniques can be enrich and refine for further enrichment.

5. CONCLUSION AND FUTURE ENHANCEMENT

In the area of medical data mining, the data cleaning is the major process to find the influential paramenter. In this regard, applying multiple level of techniques that are used to finetune the data for effective feature reduction. In this paper the ACO and RR algoritms are perform well based on the error level. In the Classification based on the pre processed data, MLPN have the good perfromance in the minimum iteration. In future, the hybrid of multiple preprocessing techniques will be give more effective tuning of the data and find the highly influential parameters that make a perfect impact of feature reduction, analysis and prediction.

REFERENCES

[1] Milewski, R., Milewska, A. J., Czerniecki, J., Leśniewska, M., &Wołczyński, S. (2013). Analysis of the demographic profile of patients treated for infertility using assisted reproductive techniques in 2005–2010. GinekologiaPolska. 84(7), 609–614.

[2] S.J Kaufmann, J.L.Eastaugh, S. Snowden, S.W.Smye, V.Sharma, "*The application of neural network in predicting the outcome of in-vitrofertilization*", HumanReproduction vol.12 no.7 pp. 1454-1457, 1997.

[3] Asli Uyar, AyseBener, H.NadirCiray, Mustafa Bahceci, "*Handling the Imbalance Problem of IVF Implantation Prediction*", IAENG International Journal of Computer Science, May 2010.

[4] Asli Uyar, AyseBener, H.NadirCiray, Mustafa Bahceci, "*ROC Based Evaluation and Comparison of Classifiers for IVF Implantation Prediction*", Institute of Computer Sciences, Social-Informatics andTelecommunication Engineering, LNICST 27, pp. 108-111, 2010.

[5] M. Durairaj, R. Nandhakumar, "*Integrating Ant Colony Optimization and Relative Reduct Algorithm for predicting the Success Rate for In-Vitro Fertilization*", IJETAE, Volume 8, Issue 2, Aug 2018.

[6] M. Durairaj, K. Meena, "*A Hybrid Prediction System Using Rough Setsand Artificial Neural Networks*", International Journal of InnovativeTechnology and Creative Engineering, vol. 1, no. 7, July 2011.

[7] M. Durairaj, R. Nandhakumar, "A Comparison of the Perceptive Approaches for Preprocessing the Data Set for Predicting Fertility Success Rate", IJCTA, Volume 9, Issue 27, 2016.

[8] David Gil, Jose Luis Girela, Joaquin De Juan, M. Jose Gomez-Torres, Magnus Johnsson, "*Predicting seminal quality with artificial intelligence methods*", Elsevier, Expert Systems withApplications,2012

[9] M. Durairaj, R. Nandhakumar, "*Data Mining Application on IVF Data For The Selection of Influential Parameters on Fertility*", IJEAT, Volume 2, Issue 6, Aug 2013.

[10] M. Durairaj, P. Thamilselvan, "*Applications of Artificial NeuralNetwork for IVF Data Analysis and Prediction*", Journal of Engineering, Computers and Applied Sciences, Volume 2, No 9, September 2013.

[11] M. Durairaj, R. Nandhakumar, "Feature Reduction by Improvised Hybrid Algorithm for Predicting the IVF Success Rate", IJARCS, Volume 8, Issue 1 Feb 2017.

[12] Claudio Manna, Loris Nanni, Alessandra Lumini, SebastianaPappalardo, "*ArtificialIntelligence Techniques for Embryo and OocyteClassification*", Elsevier, Reproductive BioMedicine 2013.

[13] M. Durairaj, R. Nandhakumar, "*An Integrated Methodology of Artificial Neural Network and Rough Set Theory for Analyzing IVF Data*", IEEE International Conference on Intelligent ComputingApplication, 2014

[14] M. Durairaj, Nandhakumar Ramasamy, "*Intelligent Prediction Methods and Techniques Using Disease Diagnosis in Medical Database: A Review*", International Journal of Control theory and Applications, Volume 8, issue 5, 2015.

[15] M. Durairaj, Nandhakumar Ramasamy, "*Intelligent Prediction Methods and Techniques Using Disease Diagnosis in Medical Database: A Review*", IJCTA, Volume 8, issue 5, 2015.