IJCTA, 9(27), 2016, pp. 255-260 ISSN: 0974-5572
© International Science Press

A Comparison of the Perceptive Approaches for Preprocessing the Data Set for Predicting Fertility Success Rate

M. Durairaj* and Nandhakumar Ramasamy**

ABSTRACT

Medical diagnostics systems are evaluated by employing large information databases, but it endures many failures to extract data from Database. There is no sufficient tool available to discover the major relationship between the data. In such case, the core knowledge of healthcare data is extracted by applying the data mining methods. The extracted knowledge can be used for the perfect diagnosis and further treatment. Infertility is an emotional cause of fertility in all over the world over past years. Treatment for infertility includes set of procedures like IUI, IVF, ICSI, and GIFT to cure the disease. Predicting the success rate for the infertility treatment can be done by using the preprocessed data from the database. This paper explains the existing Preprocessingmethod and analyzes the accuracy of prediction rate after preprocessing. It is evident that the accuracy is increased up to 90% after preprocessing the raw data using the existing techniques. Hybriding different techniques together will provide a better result which is taken as the future direction of thiswork.

Keywords: Data Mining, Preprocessing, Accuracy, Naïve Bayes, Simple Logistic Regression, Prediction, In-Vitro Fertilization.

1. INTRODUCTION

Data Mining is examination of large volume of data to extract the hidden knowledge and unknown pattern relationship. Data mining methods are long process for the product development [10]. It involves from data collection to upgrade to a form of knowledge. This process consist many steps; those are Data cleaning, Integration, selection, mining, pattern Evaluation and Knowledge Discovery [7][8]. Knowledge discovery is a procedure to get high level knowledge from low level data. The below figure-I shows the knowledge discovery process.

Medical Data Mining is an area of challenge since the data involves in it are imprecise, inconsistent and the data is massive. The clinical decision are made by practioner knowledge and experience rather than on the rich data hidden in databases [4]. Some modes of errors are encountered such as expenses of medical cost which affect the emotional factor of patients. The medical histories of patients have a number of test results to diagnosis the patients. Some poor clinical decision may distraction consequences [23]. Hence the advantage of Data mining is taken for developing an intelligent diagnosis tool [6]. Accuracy of the prediction can be improved by using data classification algorithms. The researchers in the field of medical started identify and predicting the success rate with the aid of data mining methods.

This paper is organized as follows: Section 2 give description about Infertility and its importance on the society. Section 3 is about the data sets for experimentation. Section 4 describes the preprocessing methods and training of the data set using Simple Logistic Regression and Naïve Bayes classification function. Section 5 denotes proposed work and the Section 6 concludes the paper.

* Assistant Professor, *Email: durairaj.bdu@gmail.com*

** Research Scholar, School of Computer Science, Engineering & Applications, Bharathidasan University, Tiruchirappalli, *Email: nandhakumar.897@gmail.com*

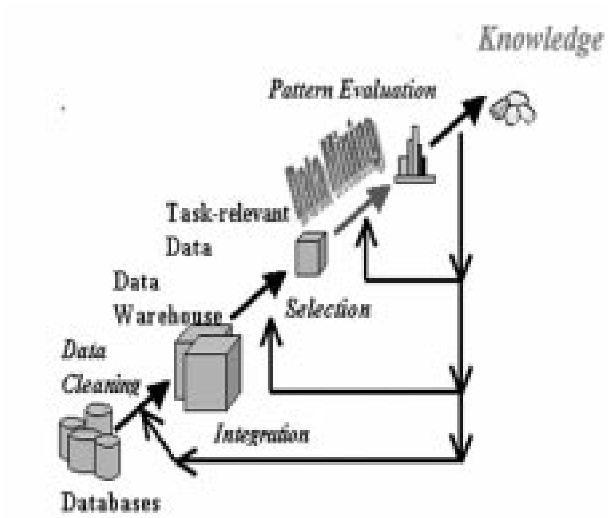

Figure 1: Data mining Knowledge discovery process

2. INFERTILITY DISEASE

The infertility is one of the world's most important problems in health care. It affects the patients mentally and physically [3]. Number of cycles and procedures are there to treat the person with infertility problem through artificial insemination methods.

A recent survey has revealed that approximately 46 per cent of Indians who are looking for medical assistance for giving birth to a child and are in the age group of 31 to 40 years are sterile [20]. This implies that they have been failed to give birth to a baby even after being at it for 2 years. The survey was conducted across 9 cities in the country – Delhi, Mumbai, Kochi , Bangalore, Ahmedabad, Hyderabad, Agra, Kolkata, and Chennai and took into account 2,562 patients. It was named "Helping Families" and had the endorsement of the Indian Society for Assisted Reproduction (ISAR) [11]. This survey was organized by a pharmaceutical company and involved doctors as well [22].

3. DATA SET

The data set used for the experimentation is collected from various Fertility clinics, Hospitals and Research centres in Tamil Nadu. This data set has 42 attributes. Among all the 42 attributes, 34 attributes is taken for the experiments based on the doctor's suggestion.

Table 1
Attributes used for this work

	Attributes used for this work			
Name	*Previous Surgery*	*Endometriosis*	*Liquefaction Time*	*Male Factor Only*
Unknown Factor	Pre-Existing Symptoms Of Depression	Tubal Infertility	Sperm Concentration	Severe Male Factor
Place	Fear And Negative Treatment Attitude	Ovulatory Factor	Sperm Motility	Female Factor Only
IvfTreatment	Psychological And Emotional Factors	Hormonal Factor	Sperm Vitality	Combined Factor
If Miscarriage	Difficulty In Tolerating Negative Emotions For Extended Time	Cervical Factor	Sperm Morphology	Unknown Factor
If Yes Miscarriage Caused	Uncertainty	Unexplained Factor	No.of Oocytes Retrieved	Place
Medical Disorders	Strain Of Repeated Treatment	Semen Ejaculate Volume	No.of Embryos Transferred	Ivf Treatment

The data set contains both numeric and nominal attributes.Among theselected 34 attributes, Numeric attributes are converted to nominal for better results.

4. DATA PREPROCESSING AND TRAINING THE DATA

4.1. Logistic Regression

Simple logistic regression has one nominal variable and one range variable, where the nominal variable is the dependent variable, and the range variable is the independent variable [2].

It separates simple logistic regression, with only one independent variable, from multiple logistic regressions, which has more than one independent variable [18]. Many people lump all logistic regression together, but it is useful to treat simple logistic regression separately, because it's simpler. Simple logistic regression is analogous to linear regression, rule out that the dependent variable is nominal, not a range [19]. One goal is to see in case the probability of getting a particular value of the nominal variable is associated with the range variable the other aim is to forecast the probability of getting a particular value of the nominal variable, given the range variable.

The logistic regression model is simply a non-linear transformation of the linear regression. The "logistic" distribution is an S-shaped distribution function which is similar to the standard-normal distribution (which results in a probity regression model) but easier to work with in most applications (the probabilities are easier to calculate) [17]. The logit distribution constrains the estimated probabilities to lie between 0 and 1.

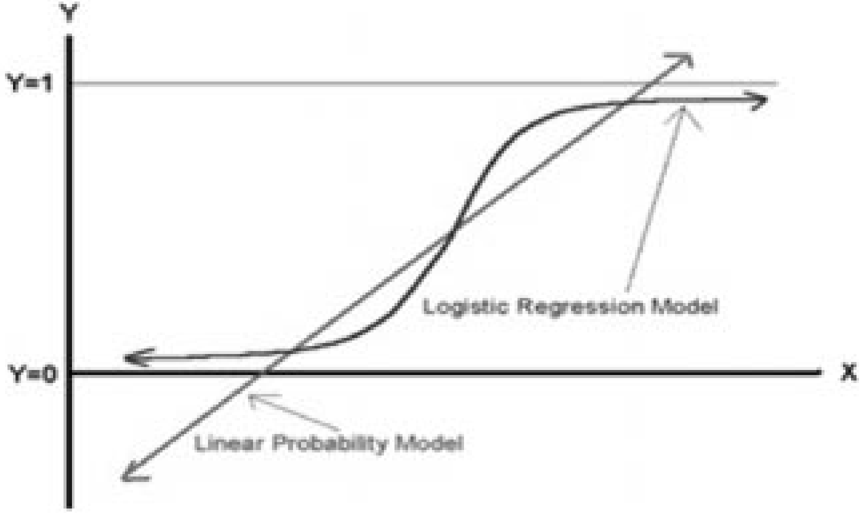

Figure 2: Comparison of probability and logistic Regression

For instance, the estimated probability is:

p = 1/[1 + exp(-a - BX)]

With this functional form:

- if you let a + BX = 0, then p = .50
- as a + BX gets really big, p approaches 1
- as a + BX gets really small, p approaches 0.

4.2. Bayesian Classification

The Bayesian Classification denoted a supervised learning method as well as a statistical method for classification [1]. Assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the results. It can solve diagnostic and predictive problems. This Classification is named Thomas Bayes (1702-1761), who derived the Bayes Theorem. Bayesian classification provides practical learning algorithms and forgoing knowledge and observed data can be combined. Bayesian Classification provides a useful perspective for understanding and appraising many learning algorithms. It calculates explicit probabilities for hypothesis and it is robust to noise in input data.

5. PROPOSED WORK

The proposed methodology for preprocessing is as follows:

- In the initial stage, the data set is preprocessed by using the Numeric to Nominal method.
- Replacing the missing value by using Replace Missing Value Techniques.
- After Data cleaning, the data is moved for training for accuracy.
- The next move is the extracting the redundant feature for predicting the knowledge.
- This knowledge is to be effected by using existing Data Mining techniques.
- The data is validated to find optimal reduction feature for prediction.

Data preprocessing is one of the major role in data mining. In the initial phase of data mining is training the data, during the discovery of knowledge it will be difficult with unwanted, irrelevant data or many noisy and unreliable data. In a medical data have many irrelevant data to disturbing the knowledge [2]. So preprocessing is an important step before training the data. Here totally 150 instances are trained before preprocessing. In this work, the two existing techniques are used for preprocessing the data and the results obtained from each of the techniques are compared.

The Table2 compares the activity performance of Simple Logistic and Naïve Bayes classifier before the preprocessing. The RMSE and correlation coefficient are evaluated and values are low. Correlation coefficient is measures the probabilistically correlation between actual and predicted value. If correlation is 1, it is perfect statistical and there is no correlation if it is 0.The Correlation value of the Logistic Regression is nearer to 0. Hence, training a network with NaiveBayes will yield better results than Logistic Regression [23].

The Infertility disease data set has both numeric and nominal data sets. The primary step of preprocessing is converting the numeric attribute to nominal attribute. The NumerictoNominal conversion is used for renovating the attributes as Nominal. The Logistic Regression and NaiveBayes networks are trained after preprocessing. The results acquired are also having promising results.

The Data is again subjected to Classification by using both NaiveBayes and Logistic Regression. The Feature Selection accuracy is calculated in both the cases. It is witnessed that NaiveBayes outplays Logistic Regression in accuracy. The Kappa Statistic value is higher in NaiveBayes. If the Kappa Statistic producing is 0.7 or greater than 0.7, then it is said to better statistic correlation. The correlation is found to be effective in the case of high Kappa value. Table 2 compares both the network for performance after converting numeric attributes to nominal.

Table 2
Comparing the Performance of Logistic Regression and NaiveBayes before preprocessing

Parameters	*Logistic Regression*	*Naïve Bayes*
Time taken	0.70	0.02
Correlation co-efficient	0.8764	0.3729
Mean Absolute Error	0.3975	0.9017
Relative Mean Squared Error	0.6025	1.1385
Relative Absolute Error (%)	39.1571	88.8746
Root Relative Absolute Error (%)	49.1125	92.8265

Table 3
Comparing the Performance of Logistic Regression and Naïve Bayes after converting attributes to Nominal

Measures	*Naïve Bayes*	*Logistic Regression*
Prediction Accuracy	95%	63%
Kappa statistic	0.9225	0.3
Mean Absolute Error	0.0321	0.1433
Root Mean Squared Error	0.127	0.3569
Relative Absolute Error (%)	12.4102	55.3535
Root Relative Squared Error (%)	35.3664	99.3946
True Positive Rate	0.950	0.637
False Positive Rate	0.039	0.305
Precision	0.950	0.648
Recall	0.950	0.637
F-measure	0.950	0.576
ROC	0.989	0.893

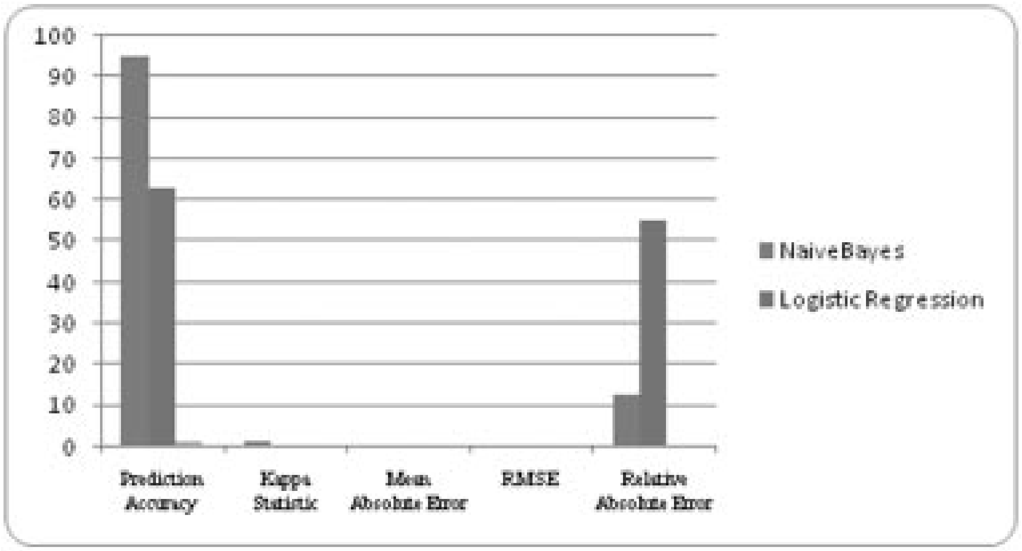

Figure 3: Naïve Bayes Performance over Logistic Regression

From the Table 2, it is inferred that both Logistic Regression and NaiveBayes perform better. But NaiveBayes outperforms Logistical Regression in exactness and also in Relative Absolute Error. It is also observed that the RAE value is better after preprocessing the data set.

6. CONCLUSION

Infertility disease prediction is a major challenge in health care industry. Selecting less number of attributes without influencing the accuracy of diagnosis is a challenging task in Data Mining. Removing and correcting all the noisy data and extracting the data from the medical data would help medical practitioners in many ways. Apart from the removal of unwanted data, extracting the feature is a important way for prediction of Fertility Success Rate. It is observed from the analysis that the preprocessing of data yields promising results. The preprocessing of data enhances the prediction accuracy and diagnosing and it was nearly 95%. The NaiveBayes network prediction is has high accuracy with low error rates when compared with Logistic. Extracting relevant features and proper training of the network will result in highly promising diagnosis. The future direction in this research is the extraction of relevant Redunt feature with hybrid techniques which would further improve the prediction accuracy.

REFERENCES

[1] Liangxiao. J, Harry. Z, Zhihua.C and Jiang.S, "One Dependency Augmented Naïve Bayes", *ADMA*, 186-194, 2005.

[2] M.J. Ruperez, J.D. Martin-Guerrero, C. Monserrat, M. Alcaniz, "Artificial Neural Networks for predicting dorsal pressure on the foot surface while walking" *Expert Systems With Applications*, 5349-5357, 2011

[3] M. Durairajand P. Thamilselvaln, "Applications of Artificial Neural Network for IVF Data Analysis and Prediction"*Journal of Engineering, Computers & Applied Sciences,* **2(9)**, 11-15, 2013

[4] M. Durairaj, K. Meena, S. Selvaraju, "Applying a data mining approach of rough sets on spermatological data analysis as predictors of in-vitro fertility of bull semen" *International Journal of Computer, Mathematical Sciences and Applications,* **2**, 189-199, 2008

[5] M. Durairaj and V.Ranjan, " Data mining Application in Healthcare Sector: A Study" *International Journal of Scientific & Technology Research*, **2(10)**, 29-35, 2013

[6] M.Durairaj, K.Meena, "Intelligent Classification Using Rough Sets and Neural Networks" *The Icfai journal of Information Technology*,75-85, 2007

[7] Guoqiang Peter Zhang "Neural Networks for Classification: A Survey" *IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications And Reviews,* 451 -462, 2000

[8] K.Srinivas, G.RaghavendraRao, A.Govardhan, "Analysis of Attribute Association in Heart Disease Using Data Mining Techniques" *International Journal of Engineering Research and Applications (IJERA),*680- 1683, 2012

[9] S.J.Kaufmann, J.L.Eastaugh, S.Snowden, S.W.Smye and V.Sharma, "The application of Neural Networks in predicting the outcome of in-vitro fertilization "*Human Reproduction,***12(7),** 1454-1457, 1997

[10] Kay Elder, & Brian Dale., "In- Vitro Fertilization",*United Kingdom at the University Press*,**2,** Cambridge, 2000

[11] M. Durairaj and R. Nandhakumar, "Data Mining Application on IVF Data for the Selection of Influential Parameters on Fertility" *International Journal Engineering and Advanced Technology,* **2(6),** 262 - 266, 2013

[12] NorsaliniSalim., "Medical diagnosis using Neural Networks", *Faculty of Information Technology, University Utara Malaysia*, Sintok, Kedah, 2004.

[13] LixiangShen, Francis, E.H. Tay,LiangshengQu and YudiShen, "Fault Diagnosis using Rough Sets Theory" computer industry. Muller H., Freytag J.," Problems, Methods, and Challenges in Comprehensive Data Cleansing" *Humboldt-Universitatzu Berlin*, Germany. 61-72, 2000

[14] Edwards, R. G., "The Bumpy Road to Human In-Vitro Fertilization. Nature Medicine" **7**:1091-1094, 2001

[15] B.S. Ahn, S.S.Cho and C.Y.kim, "The integrated methodology of rough set theory and artificial neural network for business failure prediction" *Expert System with applications/Elsevier/locate*, 65-74, 2000

[16] Dr. Durairaj. M, Sivagowry.S, " A Pragmatic Approach of Preprocessing the Data Set for Heart Disease Prediction", *International Journal of Innovative Research in Computer and Communication Engineering,* **2(11)**, November 2014.

[17] ShantakumarB.Patil, "Intelligent and Effective Heart Attack Prediction System using Data Mining and Artifical Neural Network", *European Journal of Scientific Research*, **31(4),** 642-656, 2009.

[18] Shanthakumar B. Patil, "Extraction of Significant patterns from Heart Disease Ware Houses for Heart Attack Prediction", *IJCSNS*,**9(2)**, 228-235, 2009.

[19] M. Durairaj and R. Nandhakumar(2014) "An Integrated Methodology of Artificial NeuralNetwork and Rough Set Theory for Analyzing IVF Data" *International Conference on Intelligent Computing Applications*, 978-1-4799-3966-4/14 © 2014 IEEE DOI 10.1109/ICICA..35, 2014

[20] Sudha.A, Gayathri.p and Jaishankar. N "Utilization of Data Mining Approaches for prediction of life Threatening Disease Survivability", *IJAC*, **l4(17),**876-885, March 2012.

[21] BhagyashreeAmbulkar and VaishaliBorkar "Data Mining in Cloud Computing",*MPGINMC, Recent Trends in Computing,* **ISSN 0975-8887**, 23-26, June 2012.

[22] Cengizcolak.M ,Cemizcolak and Hasan Kocatruk "Predictingcoronary artery disease using different artificial neural network models", *CAD and Artificial neural network*, 249-254, 2008.