A Survey on Data Mining Technologies for Decision Support System of Maternal Care Domain

Rutvij Mehta
Student, M.Tech
Dept. of Computer Engineering
CSPIT, Changa, Gujarat, India

Nikita Bhatt
Assistant Professor
Dept. of Computer Engineering
CSPIT, Changa, Gujarat, India

Amit Ganatra, PhD
Dean, Head of Department
Dept. of Computer Engineering
CSPIT, Changa, Gujarat, India

ABSTRACT

Data mining is becoming gradually popular and vital to healthcare organizations, finding useful patterns in complex data, transforming it into beneficial information for decision making. The latest statistics of WHO and UNICEF show that annually approximately 55,000 women die due to preventable pregnancy-related causes in India. Therefore, the current focus of health care researchers is to promote the use of e-health technology in developing countries. There have been many studies that apply data mining methods to recognize solutions for health care limitations in obstetrics and maternal care domain. Some of those studies included high risk pregnancy, prediction of preeclampsia, Identification of obstetric risk factors, discovering the risk factors of preterm birth, and predicting risk pregnancy in women performing voluntary interruption of pregnancy. This paper provides a survey and analysis of data mining methods that have been applied to maternal care domain.

Keywords

Maternal care, Data mining, Decision support system, High risk pregnancy, Classification mining techniques.

1. INTRODUCTION

Data and Computer science technologies are like two sides of a single coin. We can convert data into information and information into knowledge called Data Mining.

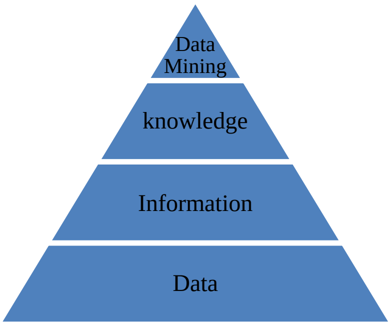

Figure 1: Mining from data

Every year around 55,000 women die in India due to pregnancy or childbirth-related complications [5][6][7]. The situation is so grave that UN has set a target of reducing Maternal Mortality Rate (MMR) by 75% till the year 2015-16 in its Millennium development goals (MDGs) [5][6][7]. Causes of maternal mortality are preventable. Causes can be easily removed through effective preventive countermeasures with use of

- E-health technology
- Remotely monitor patients in their homes by health professionals, and
- Using data mining techniques to raise alarms about high risk maternal patients.

Unfortunately very few systems available that uses data-mining techniques to determine the risk factors, which could lead to the death of a pregnant woman [1]. We should have a dataset for pregnancy related problems because the procedure of risk assessment for maternal patient is totally different. Pregnancy related problems show a particular pattern which otherwise might be measured as a normal. For example a blood Pressure of 130/90 is an alarm of high risk for pregnant women but for other women it might be a routine increase [1].

Current obstetric risk scoring systems do not make a precise prediction of the chances of an abnormal outcome and so cannot be used in formal decision analysis because of inaccuracy of data base: geographical variations and the “treatment paradox” [8]. In obstetrics and maternal care domain there have been many studies that apply data mining methods to recognize solutions for health care limitations. Some of those studies included high risk pregnancy, prediction of preeclampsia, Identification of obstetric risk factors, discovering the risk factors of preterm birth, and predicting risk pregnancy in women performing voluntary interruption of pregnancy [1] [2] [3] [13].

Therefore Data mining is becoming increasingly popular and essential to healthcare organizations, benefiting different health services, through many applications since identifying effective treatments and best practices until providing quality healthcare to the patients [3].By applying data mining methods and techniques, one can identified high risk pregnant women and provide consequence alert and suggestions for high risk maternal patients.

2. DATA MINING METHODS AND TECHNIQUES FOR MATERNAL CARE DOMAIN

Major Data Mining techniques are association rule mining, classification and cluster analysis [12].

2.1 Association Rule

Association mining discovers interesting correlation relationships between a large set of data items. It is more related when new rules are searched. Association rules are widely used in exploring the relationship between the symptom and syndrome type. Generated rules can be used for fast and better clinical decision-making [12].

Tom M. Mitchell presented approach [9] to generate rules similar to Clark's and Nisbett's CN2 [10]. Dataset contains 9,714 records of pregnant women. They wanted to improve ability to identify future high-risk pregnancies with early caesarean delivery. Each pregnant woman has features like age, diabetic or not, and first pregnancy or not. Rules learned automatically from data set. Rule predicts a 60% risk of early caesarean. There are total 215 features, they used three features. Training set accuracy is 63% and Test set accuracy is 60%. If training data is imperfect, although CN2 algorithm can work efficiently. Two algorithms are used in design like AQ and ID3. It creates a rule set like AQ but it is able to handle noisy data like ID3 [9][10]. Therefore they have mention Decision-tree learning algorithm C4.5 is also frequently used to formulate rules of this type.

Yu Chen et al. had proposed SmartRule [11] designed for mining tabular data like spread sheets. They had mined pregnancy data to generate association rules. SmartRule can generate Maximum Frequent Item sets directly from tabular data. Their tool has the advantage over traditional analytical methods that possibly important data sources can be examined for signals of importance for clinical or public health practice without having to wait for a proper hypothesis to come by. The accuracy of data mining technique was compared by regression analysis with known published results.

2.2 Classification & Clustering

Classification is the process of discovering a set of models to predict the class of objects whose class label is unknown. The Resulting model is based on the analysis of a set of training data. Clustering is the process of analysing and grouping the data into different clusters or classes such that objects within the same cluster will be having more similarity to each other, but will be having different properties in other clusters [12].

M. Jamal Afridi et al. [1] have developed an intelligent health tool – Obstetrics and Gynaecology (OG) OG-Miner – that presents an innovative combination of data mining techniques for classification of high risk pregnant women. 1200 patients' record contains most important risk factors of maternal mortality like obstructed labour, hypertension, haemorrhage, septicaemia. They developed a hybrid classifier with the use of Naive Bayes in combination with IBk. They mentioned IBK classifier may fail to classify test instance that belongs to the confusion region. So system should not completely depend on instance based learning instead of other classification method must use for decision making process. From their analytical study they combined Naive Bayes with IBK. They performed Voting at the meta-level that uses the average probabilities as a combination rule. Accuracy of tool is nearly 98% on the collected dataset. Authors have mention with the use of this kind of intelligent tool high risk maternal patients are accurately identified. So high risk maternal patients are referred to the experts and they would get quality care. Authors want to implement rule learning classifier in future so that medical expert can edit diagnosis rules.

Sónia Pereira et al. presented approach [2] of Data Mining classification models to classify type of delivery by identification of Obstetric Risk Factors. Authors have collected data from the information systems and technologies used in the perinatal and maternity care unit of Centro Hospital of Oporto. Authors had explored four data mining algorithms Decision Trees, Generalized Linear Models, Support Vector Machine, and Naïve Bayes. Holdout sampling and Cross Validation were applied with 30% of the data and all data for testing respectively. Decision Tree without oversampling achieved best accuracy and specificity, 83.91% and 80.05%. GML technique achieved best sensitivity 91.11%. Authors have mentioned that instead of oversampling, Cross Validation as the sampling method improved result and other data mining techniques can be applied like Clustering which can allow creating groups linking precise pregnancy issues.

Hsiang-Yang Chen et al. had proposed approach [3] to discover the risk factors of preterm birth using data mining with neural network and decision tree C5.0.Dataset contains 910 records of maternal and paternal attributes. In proposed approach neural network was used first to explore the top 15 main risk factors of preterm birth with coefficients nearly 0.0300. Then decision tree was used to formulate 17 rules. 80% accuracy achieved with 10 rules. Authors have mentioned that their study will help medical staff and health workers to detect high risk pregnant women.

L.M. Taft et al. presented approach [4] that performance of classification algorithms can be improve by use of synthetic minority class oversampling techniques. Authors identified adverse drug events in women admitted for labor and delivery based on patient risk factors. SMOTE generate new instances from the existing cases without duplicating the original data. Authors had used naïve Bayes and decision tree because of their simplicity and graphical representation. Performance of the Naïve Bayes and the decision tree improved i.e. true positive rate of 0.32 in the raw dataset increased to 0.67.

Hsiang-Yang Chen et al. had proposed approach [14] to discover the risk factors of parenting stress using data mining with decision tree C5.0. Database collected from National Taiwan University. Authors have mentioned that DT is showing the classification route of risk factors better than the regression model. Regression analysis is also not identifying unknown important factors.

3. EMPIRICAL STUDY

There are many merits and demerits of each and every data mining algorithms. Performances of algorithms are depended on datasets characteristic. Comparative Study of different data Mining Algorithms can be used to develop classification model. Table 1 shows Comparative Study of different Data Mining Algorithms.

Here primary data is collected from company VRSSPL. Database has 970 records with 11 attributes like id, age, height(cm),weight(kg),hemoglobin(gm),gravida,pih,previous_caesarian,abortion>=3,previous_instrumental,High_Risk. First five attributes are continuous and others are categorical.

After data is collected, performance should be analysed by considering the various factors like accuracy, time, mean absolute error etc. The experiment is carried out in Weka 3.7.13 tool. For evaluating performance of the classifier, accuracy and Mean absolute error are considered. Table 2 shows performance of dataset on different classifiers.

Table 1 Comparative Study

Data Mining Algorithms	**Comparative Study**
Decision Trees	• The graphical representation of DT is easily understood unlike black box algorithms, Neural Networks and Vector Machines [4]. • DT is showing the classification route of risk factors better than the regression model [14]. • Handling both continuous and discrete attributes [15]. • Handling training data with missing attribute values [15]. • Memory issues with large databases [16]. • Output attribute must be single and categorical [17]. • Complex tree for numeric datasets [17].
Naïve Bayes	• Simplicity of the NB model is easily understood unlike Neural Networks and Vector Machines[4] • Less computational time for training [16]. • Handle large datasets [16]. • Assumption for class conditional independence [12]. • Rules cannot be generated [18].
Artificial Neural Networks	• Handle noisy data [16] [18]. • Classify patterns for untrained data [16] [18]. • Appropriate for continuous valued inputs and outputs [16]. • Parallelization techniques can be used to speed up the computational process [16]. • Black box nature [17]. • Computational Complexity is high [17]. • Training time is high [16] [17].
K-Nearest Neighbour	• IBK classifier may fail to classify test instance that belongs to the confusion region [1] • Large storage requires [16]. • Retraining is not requiring if the new training pattern added to the existing training set [15]. • Easy to implement classification technique [16]. • Handle noisy training data and effective for large database [18].
Support Vector Machine	• Can be used in statically learning [15]. • Handle both linear and nonlinear data [15]. • Computational Complexity is high [16]. • Training time is high [16]. • Black box nature [18].
Association Rule	• CN2 algorithm work even when the training data is imperfect [9][10] • CN2 creates a rule set like AQ[9][10] • CN2 is able to handle noisy data like ID3 [9][10].

Table 2 Performance of different Classifiers on Accuracy and Mean absolute error.

Algorithm	Accuracy	Mean absolute error
Decision Trees	**97.01%**	**0.0565**
Naïve Bayes	87.84%	0.2286
K-Nearest Neighbour	84.54%	0.1554
Artificial Neural Networks	93.40%	0.0822
Support Vector Machine	76.39%	0.2361

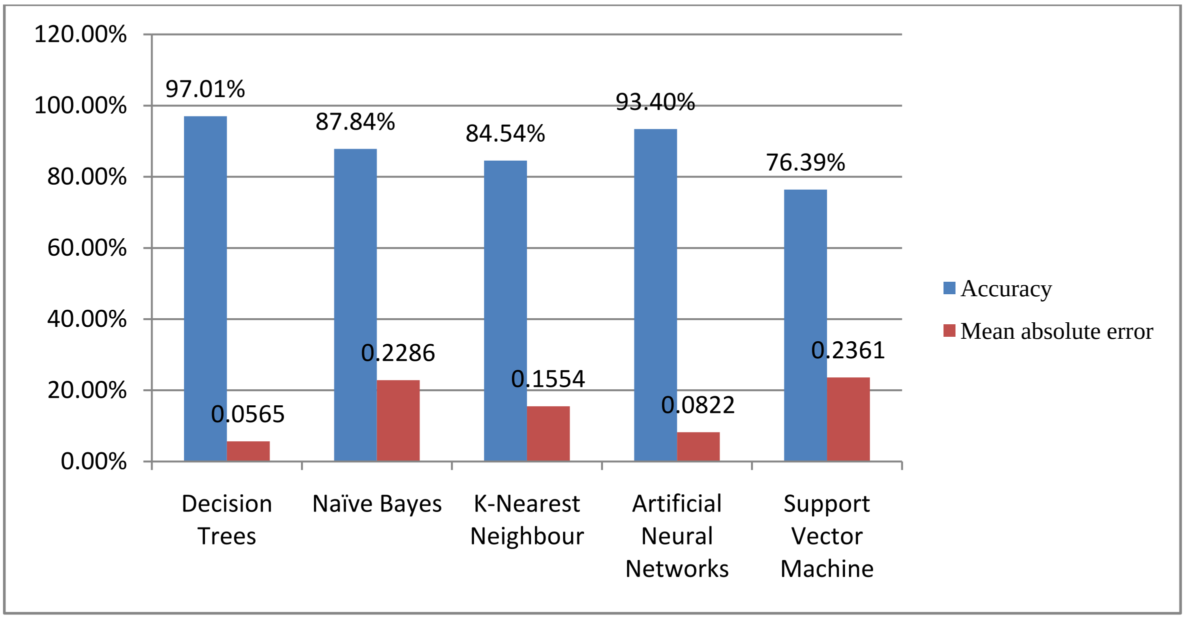

Figure 2 Performance of dataset on different classifier according to column chart

For collected dataset, Decision Tree gives better result. Support Vector Machine gives poor result. All other classifier gives an average result.

4. SUMMARY

This paper has presented major research accomplishments and techniques that immerged in the field of maternal care domain. By using DM techniques the chance of high risk maternal patients can be predicted which is helpful for timely detection and providing quality care. This paper has delivered the summary and analysis of data mining techniques used for maternal care domain.

In future scope, on the basis of this kind of empirical study combination of two or more algorithms can be used for classification model to overcome demerits of data mining algorithms. Therefore performance of classification model can be improved for specific datasets.

5. REFERENCES

[1] M. Jamal Afridi and Muddassar Farooq, OG-Miner: an Intelligent Health Tool For Achieving Millennium Development Goals (MDGs) in m-Health Environments, IEEE, Proceedings of the 44th Hawaii International Conference on System Sciences, pages 1-10, 2011

[2] Sónia Pereira, Filipe Portela, Manuel Filipe Santos, José Machado, António Abelha, Predicting Type of Delivery by Identification of Obstetric Risk Factors through Data Mining, Elsevier, Procedia Computer Science Volume 64 , 601 – 609, 2015

[3] Hsiang-Yang Chen, Chao-Hua Chuang , Yao-Jung Yang, Tung-Pi Wu , Exploring the risk factors of preterm birth using data mining, Elsevier, Expert Systems with Applications, Volume 38, Issue 5, May 2011, Pages 5384–5387, 2011

[4] L.M. Taft ,R.S. Evans , C.R.Shyua, M.J., Countering imbalanced datasets to improve adverse drug event predictive models in labor and delivery, Elsevier, Journal of Biomedical Informatics, Volume 42, Issue 2, April 2009, Pages 356–364, 2009

[5] U. Fund, “State of the World’s Children 2009. Maternal and Newborn Health,” New York: UNICEF, vol. 60, 2008.

[6] http://www.who.int/gho/maternal_health/en/

[7] http://unicef.in/Whatwedo/1/Maternal-Health

[8] Problem and pitfalls of risk assessment in antenatal care, An International journal of Obstetrics and Gynecology

[9] Tom M. Mitchell Machine learning algorithms enable discovery of important “regularities” in large data sets, communications of the ACM November 1999/vol. 42, no. 11

[10] Clark, P. and Niblett, R. The CN2 induction algorithm. Mach. Learn. 3, 4 (Mar. 1989), 261–284.

[11] Yu Chen, Lars Henning Pedersen, Wesley W. Chu,, Jorn Olsen: Drug Exposure Side Effects from Mining Pregnancy Data, ACM, Volume 9 Issue 1, Pages22–29,June-2007.

[12] Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, 2nd ed. ISBN 1-55860-901-6,March,2006.

[13] L. Kenny, W. Dunn, D. Ellis, J. Myers, P. Baker, and D. Kell, “Novel biomarkers for preeclampsia detected using metabolomics and machine learning,” Metabolomics, vol. 1, no. 3, pp. 227–234, 2005.

[14] Hsiang-Yang Chen, Ting-Wei Houa, Chao-Hua Chuang, TBPS Research Group, Applying data mining to explore the risk factors of parenting stress, Elsevier, Expert Systems with Applications 37 598–601,2010

[15] Dr. Meenu Dave, Priyanka Dadhich, International Journal of Information, Communication and Computing Technology, Vol 1, JUL 2013

[16] Hetal Bhavsar, Amit Ganatra, A Comparative Study of Training Algorithms for Supervised Machine Learning, International Journal of Soft Computing and Engineering, Volume-2, Issue-4, September 2012.

[17] Ajayi Adebowale, Idowu S.A, Anyaehie Amarachi A ,Comparative Study of Selected Data Mining Algorithms Used For Intrusion Detection, International Journal of Soft Computing and Engineering, Volume-3, Issue-3, July 2013

[18] A. S. Galathiya , A. P. Ganatra , and C. K. Bhensdadia , Improved Decision Tree Induction Algorithm with Feature Selection, Cross Validation, Model Complexity and Reduced Error Pruning, IJCSIT, Vol. 3 (2) ,3427-3431,2012.