Assessment of Breastfeeding practices in Ethiopia using different data mining techniques

Abebe Alemu[1]

Department of Computer Science
University of Gondar
Gondar,Ehtiopia
abecom12@yahoo.com

Yosef Berhanu[2]

Department of Computer Science
University of Gondar
Gondar,Ehtiopia
joshecomp@gmail.com

Dr. M. Mahalkshmi[3]

Department of Computer Science
University of Gondar
Gondar,Ehtiopia
magasree4312@gmail.com

Abstract

Breastfeeding is one of the critical issues in Ethiopia because researches show that 24.0% - 27.0% of infant death in Ethiopia is due to poor breastfeeding practices. UNICEF has been reported that a good promotion of breastfeeding practices is a most important strategic plan to reduce child mortality in developed and developing countries. Hence, it is important to identifying the determinate factors of poor breastfeeding practice, especially poor countries like Ethiopia. Poor Breastfeeding is a reasonable well-defined problem caused by many factors that are related to motherhood, environment, community and child. Therefore, it is very important to predict the determinate factors of poor breastfeeding practice in various communities in the country in order to come up with feasible intervention strategies to minimize the problem. This research intends to provide a survey of current techniques of knowledge discovery in large databases using data mining techniques which will be useful for medical practitioner to improve the breast feeding practices. The assessment was carried out with cross validation and percentage split of different data mining algorithms such as decision tree, Naive Bayes , Artificial Neural Network and Bagging.

Keywords: Datamining; J48; Bagging; Artificial Neural Network; Naive Bayes; Breast Feeding.

1. Introduction

The World Health Organization has described breastfeeding as an unequaled way of providing ideal food for the survival, healthy growth and development of infants and young children; it is also an integral part of the reproductive process with important implications for the health of mothers (Sterken, 1990).

Breast milk is the safest and most natural food for an infant. It provides an infant's complete nutritional needs up to four to six months of age. There is no need for other food or drink before this age. When the baby is fed on breast milk only, it is called exclusive breastfeeding. Exclusive breastfeeding provides the best nutrition and growth in infants, and continued growth with the introduction of solid foods at six months (Cattaneo and Buzzetti, 2001).

In addition to its nutritive value, breast milk also has a protective action against common infections (Grant, 1991). It contains many immunological factors, which protect infections of the gastrointestinal tract, allergies, certain metabolic and other diseases (Shah and Khanna, 1990). Babies, their mothers, their families, their community, their environment, even the economy of the country in which they live, all benefit from breastfeeding (Schubiger et al., 1997).Research shows that breastfeeding can save the lives of over 1,500,000 babies who die every year from diseases such as diarrhea and pneumonia. Breastfed babies have stronger immune systems and are healthier than bottle-fed babies (UNICEF, 2005).

Moreover, according to UNICEF, 1994 has advocated breastfeeding as one of the strategies for "Child Survival" and exclusive breastfeeding as a best protective way for infants against infection and malnutrition. Nowadays,

promotion of breastfeeding through Family Planning and MCH Programs is increasingly considered to be a public health policy priority especially in developing societies (Tin, 1995).

The Healthcare industry is among the most information intensive industries. Medical information, knowledge and data keep growing on a daily basis. It has been estimated that an acute care hospital may generate five terabytes of data a year (Huang et al., 1996). Medical informatics plays a very important role in the use of clinical data. In such discoveries application of data mining is important for the diagnosis of new diseases and the study of different patterns found when classification of data takes place (Fauci, 2008).

This paper aims to analyze several data mining techniques for classification of breast feeding factors. In this paper, we were considered the classification algorithms such as decision tree, Navie Bayes, Artificial Neural Network and bagging .The rest of the paper is organized as follows: Section 2 contains Data Description and Section 3 describes the Overview of the techniques employed. Detailed results are discussed in Section 4. Finally Conclusions and references are given.

2. Data Description and Preparation

The source of this research is the Standard type of DHS survey data from Ethiopia Demographic and Health Survey 2011 (EDHS). The 2011 Ethiopia Demographic and Health Survey (2011 EDHS) is part of the worldwide MEASURE DHS project which is funded by the United States Agency for International Development (USAID). The survey was implemented by the Ethiopian Central Statistical Agency (CSA).

The principal objective of the 2011 Ethiopia Demographic and Health Survey (EDHS) is to provide current and reliable data on children's nutritional status([Ethiopia-CSA] and International, 2012).This research is one of the nutrition topic research. Nutrition is the hottest research areas in developed and developing country especially breast feeding is a widespread hottest problem domain research in the world (Lawrence and Michael, 1994).

The original survey is kept in SPSS format with 928 attributes. The selections of the dataset are performed with the help of domain experts and DHS handbook manual (DHS, 2013).

In the dataset of Breastfeeding factors there are 11,360 number of instances and there are 11 attributes that are described as follows:

Table1: Selected attributes for Breast feeding classification

Name	Description
Region	Amhara,AddisAbaba,Harari,Somali,Benishangul-Gumuz,Oromiya,SNNP,Tigray,Affar,Gambela,Dire Dawa
Resident	Rural Urban
Educational level	Primary Level, No Education, Secondary Level
Watching Television	Yes ,No
Number of Childs	One or Two Three or Four More than Five
Wealth	Poor,Middle,Rich
Delivery Place	Home,Public Sector,Private Sector,Others
Child Alive	Yes or No
Duration of Breatfeeding	Still breastfeeding, Ever breastfed, not currently breastfeeding, Never breastfed
Size of the Child	Average, Smaller than average, Larger than average
Breastfeeding	Yes ,No

3. Overview of the techniques Employed

3.1. J48 Decision Tree:-

J48 is a tree like structure, where each node represents the attributes in the dataset.J48 handles both continuous and categorical attributes to build a decision tree. The selection of the root node depends on the attribute with maximum information gain value. The internal node of the tree denotes the test on the root node attribute. The leaf node holds the class label. In order to improve the accuracy,J48 uses pre pruning or post pruning algorithm to remove unnecessary branches.

3.2.Navie Bayes:-

The Naïve Bayes is a simple probabilistic classifier based on Bayes theorem with strong independence assumptions. Navie Bayes is a supervised learning method as well as statistical method for classification.It can be used to solve diagonistic and predictive problems. A naive Bayes classifier assumes that the presence or absence of a particular feature is unrelated to the presence or absence of any other feature, given the class variable. This algorithm uses Bayes formula ,which calculates the probability of instances d being in class C_j.

$$P(C_j|d)=\frac{P(d|C_j)*P(C_j)}{P(d)}$$

To simplify the task ,Navie Bayes classifier assumes attributes have independent distributions and thereby estimate

$$P(d|C_j)=P(d_1|C_j)*P(d_2|C_j)*\ldots\ldots P(d_n|C_j)$$

Where $P(d_i|C_j)$ represents the probability of class Cj generating the observed value for feature i.

3.3. Artificial Neural Network:-

ANN was invented by psychologist Frank Rosenblatt in 1958.It was intended to model how the human brain processed visual data and learned to recognize objects.ANNs could be useful tool for pattern matching and learning capabilities. An ANN has many different processing elements (neurons) which are operated by creating connections between them. Each connection is associated with some weights. Each processing element takes many input signals and produces one output signal based on an internal weighting system. The neurons are interconnected and organized into different layers. The input layer receives input and output layer produces the final output. There are one or more hidden layers between the input and output layers. Depending on the problem it must solve, there are methods for training ANN. One is Self Organizing ANN which is used discover pattern and relationship from large amount of data .Another one is Back Propagation ANN which is trained by human being to perform specific task. The later one is mainly used for cognitive research and problem solving applications.

3.4.Bagging:-

Bagging is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression . Although it is usually applied to decision tree methods, it can be used with any type of method.

4. Results and Discussion

Experiments were conducted in WEKA with 10 fold cross validation and percentage split. In 10 fold cross validation, the entire data set is used training set and then apply generated rules to the same dataset for testing. In the later case, we use percentage split 80%of the data used for training and the remaining 20% of the dataset are used for testing .According to our study, the cross validation has been proved to be statistically good in evaluating the performance of the classifier for large dataset. Quality of the classifier is evaluated with the help of accuracy, time taken to build a classifier and also ROC.

We have trained the classifiers to classify the breast feeding dataset as either yes(Breastfeed) or no(not Breastfeed).

Accuracy can be calculated with the help of confusion matix. The number of correctly classified instances is sum of diagonal values of the confusion matrix; all others are incorrectly classified instances. Accuracy can be calculated as follows:

		Predicted class	
		Positive (YES)	**Negative (NO)**
Actual class	**Positive (YES)**	A	B
	Negative (NO)	C	D

Fig 1. Confusion Matrix

Accuracy=TP+TN/Total Number of instances

Accuracy= A+D / (A+B+C+D)

Time shows the time complexity of each algorithm.ROC is defined as the number of testing data that can be classified to the total numbers of testing input data. Following table shows the comparative study of different algorithms for cross validation and percentage split.

Table 2: Comparison of different classification algorithms using 10 fold cross validation

Algorithm	Correctly Classified instances	Incorrectly classified instances	Time Taken(in sec's)	Accuracy	ROC Area
Naïve Bayes	10311	1074	0.07	90.56%	0.959
J48	10950	435	0.88	96.41%	0.989
ANN	10568	817	113.35	92.82%	0.934
Bagging	10945	440	6.18	96.13%	0.992

Table 3: Comparison of different classification algorithms using percentage split

Algorithm	Correctly Classified instances	Correctly Classified instances	Time Taken(in sec's)	Accuracy	ROC Area
Naïve Bayes	8194	914	0.03	89.96%	0.958
J48	8402	706	0.49	92.24%	0.948
ANN	8350	758	113.36	91.67%	0.948
Bagging	8427	681	6.48	92.52%	0.97

From the above tables ,we can see the highest accuracy is 96.41% provided by J48 and Bagging provides 96.13 in cross validation. In the case of percentage split, the highest accuracy is around 92% provided by bagging and J48 respectively.

Kappa statistic, mean absolute error and root mean squared error will be in numeric value only. We also show the relative absolute error and root relative squared error in percentage for references and evaluation. The results are shown in the figures 2 and 3.

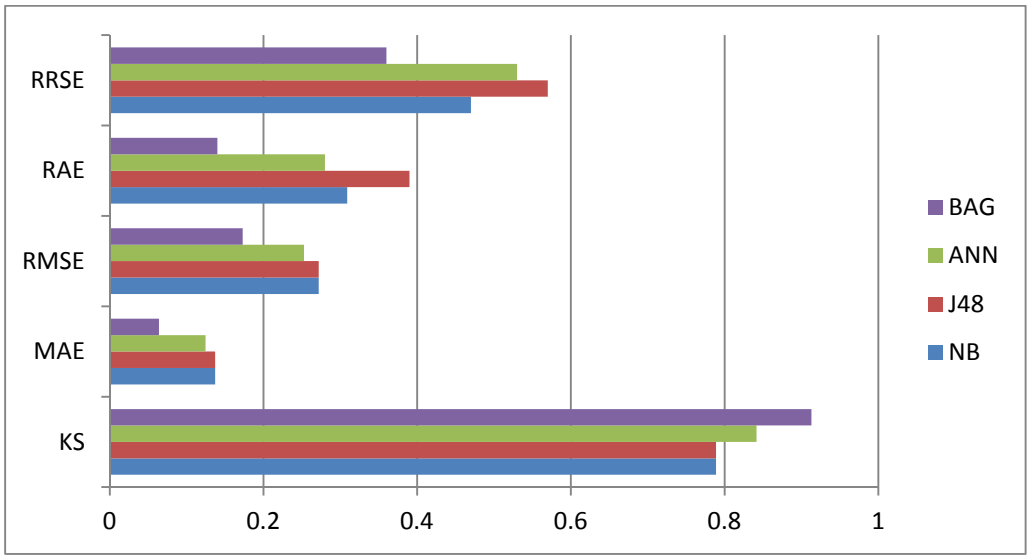

Fig 2. Classifiers with 10 folds cross validation

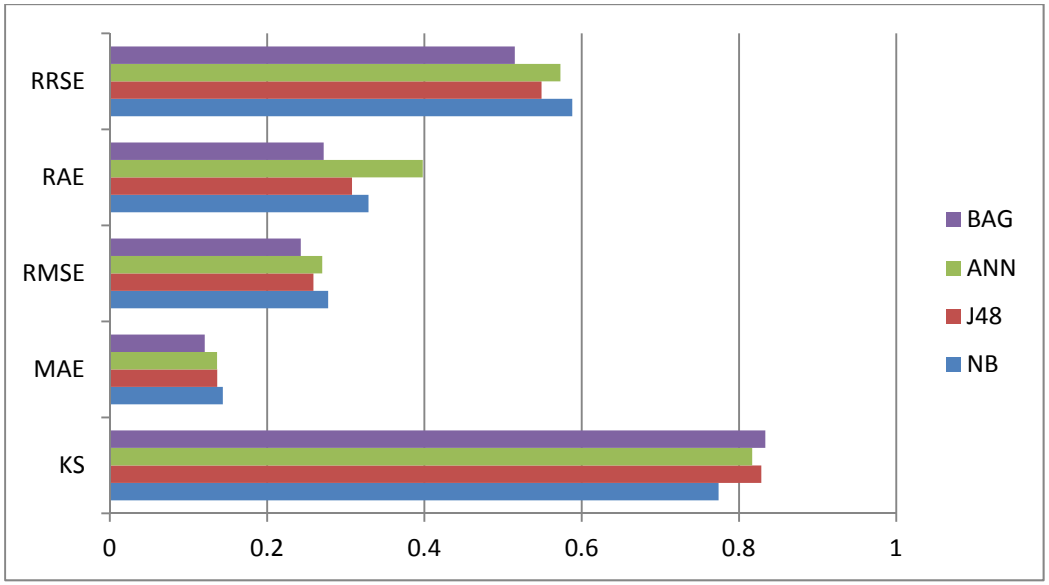

Fig 3. Classifiers with percentage split

Kappa Statistic is a measure of how closely the instances classified by the machine learning classifier. The average Kappa value from the selected algorithm is 0.80 shown in the figures. Based on kappa statistic criteria, the accuracy of this classification is substantial. From figures we can observe that the differences of errors from the selected algorithms.

5. Conclusion

Breastfeeding is a global issue for communities and governments that affect both women and their infants on many levels. Hence, the study of breastfeeding practice is one of the important mechanism for addressing the global problem. The study was conducted using classification techniques namely decision tree, naive bayes, artificial neural network and bagging. Experiment was conducted using two options cross validation and percentage split. Moreover, the finding of this research indicates that delivery place, maternal (mother) educational status, resident place, child weight and watching television are determinate factors of child breastfeeding practice. The results from this study can contribute towards in encouraging and support the decision for healthcare organization and health practitioner. From the analysis, it is concluded that J48 decision tree provides high accuracy (96.41%) for large data set compared to other algorithms. Bagging also provides the same accuracy, but it takes lot of time (6.48 sec's) to build the model.

References

[1] Azevedo, A. I. R. L. 2008. "KDD, SEMMA and CRISP-DM: a parallel overview."
[2] Baylis, P. 1999. "Better health care with data mining." SPSS White Paper, UK.
[3] Belay, D. T. 2004. "FMOH and UNICEF join forces to promote safe breastfeeding" [Online]. [Accessed may 28 2013].
[4] Bener, A., Hoffmann, G., Afify, Z., Rasul, K. & Tewfik, I. 2008. "Does prolonged breastfeeding reduce the risk for childhood leukemia and lymphomas?" Minerva pediatrica, 60, 155-161.
[5] Bernt, K. & Walker, W. 1999. "Human milk as a carrier of biochemical messages. "Acta Paediatrica, 88, 27-41.
[6] Berson, A., Smith, S. & Thearling, K. 2000."Building data mining applications for CRM," McGraw-Hill New York.
[7] Cattaneo, A. & Buzzetti, R. 2001. "Quality improvement report: Effect on rates of breast feeding of training for the Baby Friendly Hospital Initiative." BMJ: British Medical Journal, 323, 1358.
[8] Chezem, J., Friesen, C. & Boettcher, J. 2003."Breastfeeding knowledge, breastfeeding confidence, and infant feeding plans: effects on actual feeding practices." Journal of Obstetric, Gynecologic, & Neonatal Nursing, 32, 40-47.
[9] Chung, M., Raman, G., Chew, P., Magula, N., Trikalinos, T. & LAU, J. 2007. "Breastfeeding and maternal and infant health outcomes in developed countries. "Evid Technol Asses (Full Rep), 153, 1-186.
[10] Cios, K., Swiniarski, R., Pedrycz, W. & Kurgan, L. "The Knowledge Discovery Process." Data Mining, 2007. Springer, 9-24.
[11] Colet, E. 2013. " Clustering and Classification: Data Mining Approaches" [Online]. Virtual Gold Inc. Available: http://www.tgc.com/dsstar/00/0704/101861.html [Accessed 7/17/2013 2013].
[12] Duffy, L. C., Faden, H., Wasielewski, R., Wolf, J. & Krystofik, D. 1997. "Exclusive breastfeeding protects against bacterial colonization and day care exposure to otitis media." Pediatrics, 100, e7-e7.
[13] Fauci, A. S. 2008."Harrison's Principles Of Internal Medicine," MCgraw-Hill Medical New York.
[14] Ferri, C., Flach, P. & Hernández-Orallo, J. "Learning decision trees using the area under the ROC curve." machine learning-international workshop then conference-, 2002. 139-146.
[15] Gershenson, C. 2003. "Artificial neural networks for beginners." Proceedings of the International Symposium on Engineering under Uncertainty: Safety Assessment and Management (ISEUSAM-2012), Springer.
[16] GRANT, J. P. 1991." The State of the World's Children." Oxford University Press.
[17] Han, J., Kamber, M. & Pei, J. 2006. Data mining: concepts and techniques, Morgan kaufmann.
[18] He, H., Jin, H., Chen, J., Mcaullay, D., Li, J. & Fallon, T. Analysis of breast feeding data using data mining methods. Proceedings of the fifth Australasian conference on Data mining and analystics-Volume 61, 2006. Australian Computer Society, Inc., 47-52.

[19] Huang, H., Tsai, W., Bhattacharya, S., Chen, X., Wang, Y. & Sun, J. Business rule extraction from legacy code. Computer Software and Applications Conference, 1996. COMPSAC'96., Proceedings of 20th International, 1996. IEEE, 162-167.
[20] Kantardzic, M. 2011. Data mining: concepts, models, methods, and algorithms, John Wiley & Sons.
[21] Kim, Y., Choi, J.-Y., Lee, K.-M., Park, S. K., Ahn, S.-H., Noh, D.-Y., Hong, Y.-C., Kang, D. & Yoo, K.-Y. 2007. "Dose-dependent protective effect of breast-feeding against breast cancer among ever-lactated women in Korea."European journal of cancer prevention, 16, 124-129.
[22] Kyi, A. K. 2000." Factors affecting breastfeeding in the Philippines: an analysis of 1998 NDHS data." Mahidol University.
[23] Mcvea, K. L., Turner, P. D. & Peppler, D. K. 2000."The role of breastfeeding in sudden infant death syndrome." Journal of Human Lactation, 16, 13-20.
[24] Milley, A. 2000. "Healthcare and Data Mining Using data for clinical, customer service and financial results", Health Management Technology, 21, 44-45.
[25] Olson, D. L., Delen, D., Olson, D. L. & Delen, D. 2008. "Data Mining Process.", Advanced Data Mining Techniques, 9-35.
[26] Pandey, M. 2009." Applications of Data Mining in Healthcare System." Journal of Healthcare Information Management.
[27] Phyu, T. N. "Survey of classification techniques in data mining.", Proceedings of the International MultiConference of Engineers and Computer Scientists, 2009. 18-20.
[28] Pineda, R. G. 2006."Breastfeeding practices in the neonatal intensive care unit before and after an intervention plan". University Of Florida.
[29] Rogers, G. & Joyner, E. 2005. "Mining Your Data for Healthcare Quality Improvement "Journal of Healthcare Information Management—Vol 19(2): 65.
[30] Schubiger, G., Schwarz, U. & Tonz, O. 1997."UNICEF/WHO baby-friendly hospital initiative: does the use of bottles and pacifiers in the neonatal nursery prevent successful breastfeeding?" European journal of pediatrics, 156, 874-877.
[31] Shah, I. H. & Khanna, J. 1990. "Breast-feeding, infant health and child survival in the Asia-Pacific context." Asia Pacific Population Journal (ESCAP), 5, 4, 25-44.
[32] Stephens, S. & Tamayo, P. 2003."Supervised and unsupervised data mining techniques for the life sciences", Curr. Drug Discov, 34, 36.
[33] Sterken, E. 1990."Role Of The World Health Organization In The Promotion Of Breast-Feeding", Canadian Family Physician, 36, 1546.
[34] Thompson, J. 2005."Breastfeeding: benefits and implications. Part two. Community practitioner", the journal of the Community Practitioners'& Health Visitors' Association, 78, 218.
[35] Thompson, R. E., Kildea, S. V., Barclay, L. M. & Kruske, S. 2011."An account of significant events influencing Australian breastfeeding practice over the last 40 years." Women and Birth, 24, 97-104.
[36] UNICEF 2005. "Protection, promotion and support of breast-feeding in Europe: current situation." Public Health Nutr, 8, 39-46.
[37] UNICEF 2006. "Progress for Children: a report card on nutrition number 4." New York, Ref Type Report.
[38] UNICEF 2008. "Progress for children–a report card on nutrition," New York, Ref Type: Report.
[39] Vikas Chaurasia & Saurabh Pal "Data Mining Approach to detect Heart Dieses" ,IJACSIT, Vol. 2, No. 4, 2013, Page: 56-66, ISSN: 2296-1739
[40] Willinger, M., James, L. S. & Catz, C. 1991."Defining the sudden infant death syndrome (SIDS): deliberations of an expert panel convened by the National Institute of Child Health and Human Development." Fetal & Pediatric Pathology, 11, 677-684.
[41] Witten, I. H. & Frank, E. 2005. "Data Mining: Practical machine learning tools and techniques", Morgan Kaufmann.
[42] Witten, I. H., Frank, E. & Hall, M. A. 2011."Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques,", Elsevier,2005
[43] Ethiopia-Csa], C. S. A. & International, I. 2012. "Ethiopia Demographic And Health Survey 2011."
[44] Shewayenesh, G. 2007. Assesment Of Breastfeeding Practice In Yeka Sub-City Addis Ababa, Ethiopia. Addis Ababa University.