Predicting Preterm Birth Based on Maternal and Fetal Data

Yavar Naddaf, Mojdeh Jalali Heravi and Amit Satsangi

1. Introduction

Preterm birth is "a birth which takes place after at least 20, but less than 37, completed weeks of gestation. This includes both live births, and stillbirths" [15]. Preterm birth may cause problems such as perinatal mortality, serious neonatal morbidity and moderate to severe childhood disability. Between 6-10% of all births in Western countries are preterm and preterm deaths are the cause for more than two-third of all perinatal deaths [9]. While the recent advances in neonatal medicine has greatly increase the chance of survival of infants born after 20 weeks of gestation, these infants still frequently suffer from lifelong handicaps, and their care can exceed a million dollars during the first year of life [5 as cited in 6]. As a first step for preventing preterm birth, decision support tools are needed to help doctors predict preterm birth [6].

While there are large datasets containing historic data for preterm birth, finding an accurate way to predict preterm risk is still a challenging problem. As we will discuss in section 5, there are currently no classification or scoring models that can predict preterm birth with satisfactory accuracy. In this work, we apply a number of classification techniques (*e.g.*, Naïve Bayes, Decision trees, SVM, logistic regression, and associative classifier) on a dataset of historic maternal and newborn records to predict preterm birth. We also perform contrast rule mining in order to select attributes that discriminate the most between normal and preterm cases. However, our results show no major improvements over the existing prediction models. We discuss a possible explanation for the poor performance of the classification methods in section 6.

2. Dataset Description and Data Preparation

The original dataset was collected by Northern and Central Alberta Perinatal Outreach Program between 1992 and 2003. The dataset contains maternal and newborn data for 243948 cases, including 21193 preterm cases. There are 244 attributes, containing "maternal demographic information, medial history such as preexisting chronic illness, lifestyle information such as smoking and alcohol use, past reproductive history including previous [preterm] or [small for gestational age] delivery, and history with the current pregnancy such as presence of hypertension or toxemia" [10]. A large portion of the attributes is collected during or after delivery, and thus cannot be used for predicting preterm birth. Appendix I contains a list of the 46 attributes that were collected before delivery. As it is shown in appendix I, "Group B Step", "Maternal Hepatitis B", and "Steroid During Pregnancy" attributes contain a high ratio of missing values. These attributes were not used in any of the classification tasks. Furthermore, there are 2107 records with missing class labels, which are also omitted from the dataset.

3. Prediction Methods

This section provides an overview of different classification methods used for predicting preterm birth. Section 3.1 describes an associative classifier using the Eclat algorithm. Section 3.2 describes a number of other poplar classifiers that we also applied on the dataset. Section 3.3 provides the experimental results.

3.1 Associative classifier

Before running an associative classification, the dataset should be converted into a transactional database. Each transaction is in fact a record of a patient. Each item is either the presence or absence of a categorical or binary attribute or a specific interval of a numerical attribute[1]. The resulting transactional database contains 241841 transactions with an average of 41 items per transaction.

[1] Another transactional dataset was generated by ignoring the absence of attributes. However, this dataset was later discarded, since associative classifiers based on this dataset performed very poorly.

Due to the large number of transactions as well as the high average number of items per transaction, the number of frequent itemsets will be enormous. Algorithms that perform a depth first search to create the frequent itemsets are best suited to transaction sets with high average number of items per transaction. Eclat, developed by Zaki et al [14], is one such algorithm and was used in our current experiences. As part of the algorithm, the "low ranked specialized rules" are pruned to reduce the large number of generated rules.

Program runtime is another prohibitive problem with such a large dataset. Generating the rules for all the 241841 transactions takes over a month. Due to limited time to finish the project, only 1% of the transactions (*i.e.* about 2000 transactions) are used for associative rule mining.

Using Eclat implemented by Borgelt [2], an associative classifier is applied to classify the records into "preterm" and "normal" classes. Eclat generates a set of rules for each class based on the given minimum support and confidence. For a new patient record, all the rules that match this record are selected from both classes. The group of rules that has the highest average confidence will be the label for that new record.

K-fold cross validation is too time consuming in the current setup. Instead, the classifier is built with a training set that contains 1% of the data and it is evaluated using 3 different test sets, each containing 1% of the data. A minimum confidence of 85% was chosen after trying the algorithm with a number of different minimum confidence values. The minimum support was set to be 1%.

3.2 Other popular classifiers

A number of popular classification algorithms are applied to the dataset for predicting preterm birth, and their prediction performance is compared with the associative classifier. The applied classifiers are Logistic Regression, Naive Bayes, C4.5 decision trees, Support Vector Machines, and Neural Networks. We used the implementation and the default parameters provided in Weka, an open source data-mining package [13]. For Naive Bayes and C4.5 Decision Trees, the numerical attributes are recoded into a set of categorical attributes. Three-fold cross validation is used to evaluate the performance of each classifier.

3.3 Prediction Performance

Table 1 contains various performance measures for each classification method. The best value for each performance measure is highlighted in bold.

Algorithm	True-Positive Rate	False-Postive Rate	Precision	Recall	F-Measure	AUC (by Weka)	AUC (Haneley and McNeil method)
Naïve Bayes	**0.281**	0.036	0.564	**0.281**	**0.375**	0.716	**0.6224**
Logistic Regression	0.207	0.014	0.713	0.207	0.321	**0.724**	0.5968
SVM with Linear kernel	0.155	**0.008**	**0.757**	0.155	0.257	0.573	0.5731
C4.5 Decision Tree	0.197	0.013	0.708	0.197	0.308	0.666	0.5918
Neural Network	0.228	0.02	0.657	0.228	0.338	0.711	0.6041
Associative Classifier	0.218	0.029	0.419	0.218	0.286	N/A	0.5944

Table 1: Predictive performance of various classification methods

AUC is the area under the Receiver Operating Characteristic (ROC) curve, and is "equivalent to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance" [4]. AUC is popular in medical data analysis, because it provides a representation of the trade-off between True Positive and False

Positive rates [12]. Furthermore, Huang and Ling claim that AUC is a superior method for comparing learning algorithms, and should replace accuracy as a performance measure in the future [8]. There are a number of methods for estimating the area under the ROC curve. In table 1, we present both the AUC estimation based on a method introduced by Haneley and McNeil [7], and the AUC as calculated by the Weka package.

It is evident from table 1 that all classification methods perform very poorly in predicting preterm birth based on the dataset. Within the tried methods, Naïve Bayes seems to outperform the rest of the algorithms.

4. Feature selection using contrast set mining

Contrast sets are conjunctions of attribute-value pairs that occur significantly more frequently in one group than the others. The main motivation behind using contrast-sets is to improve the prediction performance of the classification methods. The idea of using contrast sets to improve the accuracy of a classifier is first proposed by Dong and Li [3].

Two separate methodologies are tried for generating contrast-sets. The first method, called STUCCO, is proposed by Bay and Pazzani [1]. Bay and Pazzani create set-enumeration trees, and apply chi-square analysis to check for the independence between a parent node and its child node. We used STUCCO with minor modifications for generating contrast-sets. As mentioned earlier, the dataset used is enormously large, both in terms of the number of transactions as well as the average number of items per transaction. It was found that the set-enumeration tree that STUCCO constructs in the memory exceeded the size of the available buffer because of the enormous size of the data, and hence the program would terminate as soon as the buffer size was exceeded. Several approaches such as profiling and tracing were tried. However, STUCCO continued to terminate at data-sizes less than 0.5% of the original size of the data. As each simulation ran for a couple of days before crashing, this approach was given up.

Eclat is considered as an alternative method for generating the association rules. The problem with Eclat, as mentioned in section 3.1, is the running time. However we are able to get the rules for 1% of original dataset. Once the association rules are generated, contrast sets are found using an approach proposed by Satsangi and Zaiane [11]. The original program is modified in order to accommodate for the different format as well as a different logic that was used for rule pruning. To generate the contrast sets, we look at the support difference of itemsets in the "normal" and "preterm" classes. Itemsets with a support difference larger than a threshold are selected as part of contrast sets. The attributes extracted from the generated contrast sets are selected, and the classification methods mentioned in section 3.1 and 3.2 are repeated. As illustrated in Figure 1 and table 2, the selected features do not improve the prediction performance of any of the classifiers. Indeed, the prediction performance is decreased among all classification methods.

Algorithm	True-Positive Rate	False-Postive Rate	Precision	Recall	F-Measure	AUC (by Weka)	AUC (Haneley and McNeil method)
Naïve Bayes	0.181	0. 014	0. 678	0.181	0.286	0.676	0.5835
Logistic Regression	0.171	0.012	0.698	0.171	0.275	0.676	0.5795
SVM with Linear k ernel	0.133	**0.009**	0.711	0.133	0.224	0.562	0.5618
C4.5 Decision Tree	0.164	0.011	**0.712**	0.164	0.266	0.629	0.5763
Neural Network	**0.195**	0.02	0.617	**0.195**	**0.296**	**0.677**	**0.5873**
Associative Classifier	0.137	0.013	0.513	0.137	0.216	N/A	0.5621

Table 2: Predictive performance of various classification methods using the selected attributes

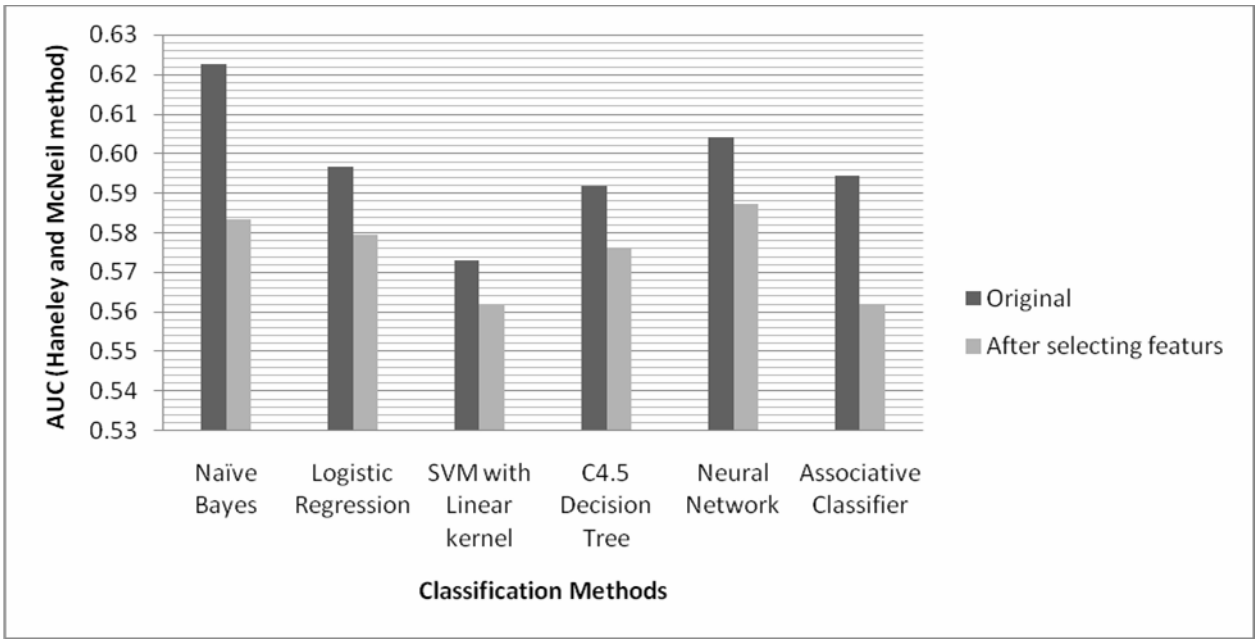

Figure 1: Comparizon of AUC (Haneley and McNeil method) for classifying using the original dataset and classifying using only rhe selected features

5. Related work

Goodwin and Maher have previously attempted to predict preterm birth using a very rich dataset provided by Duke University's Medical Center TMR perinatal data repository [6]. The dataset contains 63167 medical records collected between 1988 and 1997. Each patient record includes between 4000 to 5000 clinical and demographic variables. A team of TMR experts plus an ACOG[2] board certified obstetrician and an AWHONN[3] certified perinatal nurse performed the data cleaning and dealing with data redundancy issues, which took about a full year to complete. After cleaning the data and data reduction, 32 demographic variables plus 393 clinical variables are chosen for preterm birth prediction. Five prediction models are applied on the data: neural networks, logistic regression, CART decision trees, and two custom software packages PVRuleMiner and FactMiner, which the authors provide very little details about. Each classification method is applied once on demographic variables only, and once on all variables. The area under ROC curve is used to evaluate the classifiers. The achieved prediction performances are comparable to the results we get in our experiments. For instance, in Goodwin and Maher experiences, Neural Networks achieve an AUC of 0.64 on the 35 demographic variables only, and an AUC of 0.66 when the rest of the 393 clinical variables are added. The custom software package, FactMiner, achieves the best prediction performance; with an AUC of 0.725 on demographic variables only, and an AUC of 0.757 on all variables. An interesting conclusion of the authors is that "the 32 demographic variables produce results that are nearly as good as models with hundreds of additional variables".

Goodwin and Maher also provide a summary of various accepted risk assessment scoring tools, and report that the ability of these tools to positively predict preterm birth is within a very poor accuracy of 17% to 38%. The authors quote Dr Creesy, one of the pioneers of preterm risk scoring tools, as acknowledging that the preterm risk scoring tools "have not worked".

[2] American College of Obstetrics and Gynecology

[3] Association of Women's Health, Obstetrics, and Neonatal Nurses

6. Conclusion

In this project we applied a number of classification methods on a dataset of historic maternal and newborn records to predict preterm birth. The prediction performance of all algorithms is very poor, and even doing feature selection by contrast sets does not improve the results.

As discussed in section 5, previous attempts on predicting preterm birth using a larger and more feature-rich dataset have resulted in similarly poor prediction performance. This suggests that predicting preterm birth is a very challenging problem. Also, given that Goodwin and Moore find the demographic variables to be the most discriminative, it may be the case that the set of available clinical attributes do not cover the space required for predicting preterm birth, and as such cannot be predictive using any classification algorithm.

References

[1] Bay S.D. and Pazzani M.J., Detecting group differences: Mining contrast sets. Data Mining and Knowledge Discovery, 5(3): 213-246, 2001.

[2] Borgelt C.. Efficient implementations of apriori and éclat, Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations (FIMI'03), Florida, USA, 19. November 2003.

[3] Dong G. and Li.J., Efficient mining of emerging patterns: discovering trends and differences. Conference on Knowledge Discovery in Data Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, 43 - 52, 1999.

[4] Fawcett T, An introduction to ROC analysis. Patt Recog Letters 27: 861–874, 2006.

[5] Feldman W.E. and Wood B., The economic impact of high risk pregnancies. *Journal of Health Care Finance, 24* 64-71, 1997.

[6] Goodwin, L. and Maher, S. Data mining for preterm birth prediction. In *Proceedings of the 2000 ACM Symposium on Applied Computing - Volume 1* (Como, Italy). J. Carroll, E. Damiani, H. Haddad, and D. Oppenheim, Eds. SAC '00. ACM, New York, NY, 46-51, 2000.

[7] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*.,143:29-36, 1982.

[8] Huang, J. and Ling, C., Using AUC and Accuracy in Evaluating Learning Algorithms. *IEEE Trans. on Data and Knowledge Engineering*, 17(3)(3):299-310, 2005.

[9] Lumley J., Defining the problem: the epidemiology of preterm birth , BJOG: An International Journal of Obstetrics and Gynaecology 110 (s20) , 3–7, 2003.

[10] Newburn-Cook C. V., White D., Svenson L. W., Demianczuk N. N. and Bott N., Where and to What Extent is Prevention of Low Birth Weight Possible?, Western Journal of Nursing Research, Vol. 24, No. 8, 887-904, 2002.

[11] Satsangi A., Zaiane O. R., Contrasting the Contrast Sets: An Alternative Approach, Eleventh International Database Engineering and Applications Symposium (IDEAS 2007), Banff, Canada, September 6-8, 2007.

[12] Sebag M., Lucas N., and Az´e J.. ROC-based Evolutionary Learning: Application to Medical Data Mining. In Proc. of EA 2003, pages 384–396, 2003.

[13] Witten I. H. and Frank E., "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[14] Zaki, M. J.; Parthasarathy, S.; and Li, W. 1997. A localized algorithm for parallel association mining. In 9th ACM Symp. Parallel Algorithms and Architectures.

[15] Zegers-Hochschild F., Nygren K.-G., Adamson G.D., de Mouzon J., Lancaster P., Mansour R. , Sullivan E. on behalf of The International Committee Monitoring Assisted Reproductive Technologies, The ICMART glossary on ART terminology, Human Reproduction 2006 21(8):1968-1970

Appendix I: Attributes collected before delivery

	Attribute definition	Attribute type	% of missing values
1	Pregnancy type. Singleton/multiple gestation	Numeric	0%
2	The woman's age in year at delivery time	Numeric	1.24%
3	Total number of pregnancies including the current pregnancy	Numeric	0.02%
4	Total number of babies born, excluding the current pregnancy	Numeric	2.22%
5	Total number of pregnancy losses	Numeric	2.22
6	Mother's weight	Numeric	2.22%
7	Mothers' height less than 152 cm	Binary	2.21%
8	Diabetes controlled by diet	Binary	2.22%
9	Diabetes documented retinopathy	Binary	2.22%
10	Insulin dependent diabetes	Binary	2.22%
11	Heart disease - asymptomatic	Binary	2.23%
12	Heart disease - symptomatic	Binary	2.23%
13	Hypertension 140/90 or greater	Binary	2.23%
14	Anti hypertensive drug used	Binary	2.22%
15	Chronic renal disease	Binary	2.23%
16	Other medical disorders	Binary	2.23%
17	Past neonatal death	Binary	2.23%
18	Past stillbirth	Binary	2.23%
19	Past abortion	Binary	2.23%
20	Past preterm	Binary	2.23%
21	Previous cesarean section	Binary	2.16%
22	Previous small for gestational age	Binary	2.23%
23	Previous large for gestational age	Binary	2.24%
24	previous RH isoimmunization - unaffected infant	Binary	2.24%
25	previous RH isoimmunization - affected infant	Binary	2.24%
26	previous major congenital anomaly Downs, Heart, CNS defect etc	Binary	2.24%

27	current large for gestational age	Binary	2.24%
28	current small for gestational age	Binary	2.24%
29	current polyhydramnios or oligohydramnios	Binary	2.24%
30	current malpresentation	Binary	2.24%
31	bleeding < 20 weeks gestation	Binary	2.23%
32	bleeding >= 20 weeks gestation	Binary	2.23%
33	pregnancy induced hypertention	Binary	2.24%
34	proteinuria >= 1+	Binary	2.24%
35	gestational diabetes	Binary	2.24%
36	blood antibodies	Binary	2.24%
37	Anemia	Binary	2.24%
38	poor weight gain	Binary	2.24%
39	smoker anytime during pregnancy	Binary	2.22%
40	major fetal anomaly	Binary	22.42%
41	acute medical disorder	Binary	22.43%
42	alcohol >= 3 drinks on any one occasion during pregnancy	Binary	22.43%
43	alcohol >= 1 drink per day throughout pregnancy	Binary	22.43%
44	group B strep	Binary	65.75%
45	maternal hepatitis B	Binary	95.99%
46	Sex	M or F	0.51%
47	steroids were given during pregnancy	Binary	63.13%