Learning to Predict Dental Caries for Preschool Children

Group members: Fangzhou Guo SUNetID: fangzhou; Huaiyang Zhong SUNetID:hzhong34;
Yuchen Li SUNetID: yuchenli

1. Introduction

Dental caries, or tooth decay/cavity, is a dental disease caused by bacterial infection. Of people from different age groups, preschooler children requires more attention since caries has become the most common chronic childhood diseases. More importantly, a skewed distribution of the diseases has been observed in Europe, US and Singapore among the children or preschoolers, which indicate a small portion of the population endures a big portion of caries incidences. Therefore, there is still the need to improve on the current caries control to identify the high-risk individuals and prevent resurgence in children in developed countries like Singapore. Our project will study on the data such as questionnaire responses, oral examination and biological tests of certain preschoolers from Singapore and use suitable learning methods to predict whether they have carries now and whether they are likely to get dental caries in the future.

The final target of this project is to give a reasonable prediction on preschoolers' dental caries rate based on the chosen statistical learning model and to reduce the feature space in order to improve the utility of the models in the clinic context.

2. Data Description

In the data file provided by Prof Hsu, there are in total 1783 children entries (rows) and 64 features (columns). The features is summarized in the following Table 1:

Type of Variables	Detailed Type	Number
Input Variables	Questionnaire	57
	Clinical Examination	2
	Biological Test	5
Outcome Variables	Baseline Caries Experience	5
	Caries Increment	3

Table 1: Summary of the input and outcome variables

The questionnaire includes information from demographic background, socioeconomic status, oral health history, fluoride applications, utilization of dental care services, systemic diseases and regular medication, and parent's attitude towards caries. Clinical examination and biologic tests were administered to collect data on several clinical and biologic risk factors such as plaque amount, baseline caries existence, saliva flow rate, saliva buffering capacity, level of lactobacilli (LB), level of mutans (MS), and plaque PH.

In general, the questionnaire requires least amount of monetary input. The biologic tests incur the most cost and prolonged period of data collection. The oral examination lies in between.

The outcome variables are specified in the following Table 2

Disease Outcome	**Code and Definition of Dependent Variable**	**Type of Dependent Variable**	**Variable name in .sav file**
I: Baseline Caries	(0) dmft=0 (1) dmft>0	Dichotomous	dv1.dmft2.2
II: Caries Increment	(0) Δdmft=0 (1) Δdmft>0	Dichotomous	dv6.new.tooth

Table 2: Outcome Variables Description

Some attributes of the collected data are missing because of the missing fields in the collected questionnaires. In final project, we simply remove the entries with empty data.

3. Model

We construct the learning models based on 3 types of data space: Prediction (Full-Blown), Prediction (Screening), and Community Screening. Prediction (Screening) and Prediction (Full-Blown) are utilized to predict the incremental caries existence, i.e. number of new caries forming in next year. The difference between prediction screening and prediction full-blown lies in the use of expensive and long-waiting biological tests data. Community screening is used to predict baseline caries existence by employing the data only in questionnaire. It serves to identify high-risk children in a community setting since it is costly and time consuming to perform the oral examination directly.

The summary of the data space is presented in Table 3:

	Model Constructed		
	Prediction Model		**Community Screening**
	Screening	**Full-Blown**	
Outcome Variable: Incremental dmft? Baseline dmft?	Yes	Yes	Yes
Type of input variables	Questionnaire Oral Examination	Questionnaire Oral Examination Biological Test	Questionnaire

Table 3: Summary of Model Constructed

In addition to the initial model construction, we develop models with reduced feature space via feature selection. It will not only have influence on the model performance, but also improve the model utility in clinical trials.

4. Method

We build up 5 methods for the outcome variables (Baseline year caries existence and Incremental caries existence) for each of the above model. The five methods are SVM (Linear Kernel), SVM(Gaussian Kernel), SVM (polynomial Kernel), Logistic Regression and LDA. We construct these models because they are most frequently used methods for categorical prediction. We compare and contrast Linear (e.g. SVM (Linear Kernel)) vs Gaussian (e.g. SVM (rbf Kernel)). Based on the comparison, we get insight into the data and explore how to select different models based on the property of the data. We use Matlab to implement the methods. data is split out into training and testing set (70% for training and 30% for testing). For feature selection, we use wrapper method with aid of forward search.

5. Results and Discussion

5.1 Initial Models' Results

We gathered the results of the constructed models in the measurement of prediction error in Table 4-6.

Method	Training error	Test error
LDA	9.62%	14.15%
SVM (polynomial kernel)	0%	28.29%
SVM (rbf gaussian kernel)	0.42%	41.46%
SVM (linear kernel)	5.86%	15.61%
Logistic Regression	5.50%	15.61%

Table 4: Prediction Full-Blown Performance

Method	Training error	Test error
LDA	17.93%	23.81%
SVM (polynomial kernel)	3.79%	26.93%
SVM (rbf gaussian kernel)	3.79%	36.39%
SVM (linear kernel)	17.64%	25.85%
Logistic Regression	12.64%	22.79%

Table 5: Prediction Screening Performance

Method	Training error	Test error
LDA	24.20%	30.27%
SVM (polynomial kernel)	3.79%	35.37%
SVM (rbf gaussian kernel)	3.79%	36.05%
SVM (linear kernel)	24.64%	30.27%
Logistic Regression	17.12%	28.57%

Table 6: Community Screening Performance

Overall, prediction (full-blown) has best accuracy with prediction (screening) coming after. Community screening model has worst performance in the end. Such results showcase that high cost features like oral examination (clinical exam as in the table) and biologic tests play more dominant role in determining the outcomes. Moreover, we find that linear learning models such as Logistic Regression and Linear Discriminant Analysis have a relatively better performance than non-linear models. Our conjecture is that the data set tends to be easily over fitted by using a high-dimensional classification method (SVM using Gaussian kernel). The best classification comes from LDA, which has 14.15% classification error.

We included the receiver operating characteristics (ROC) graph for the above 3 data spaces in Figure 1-3

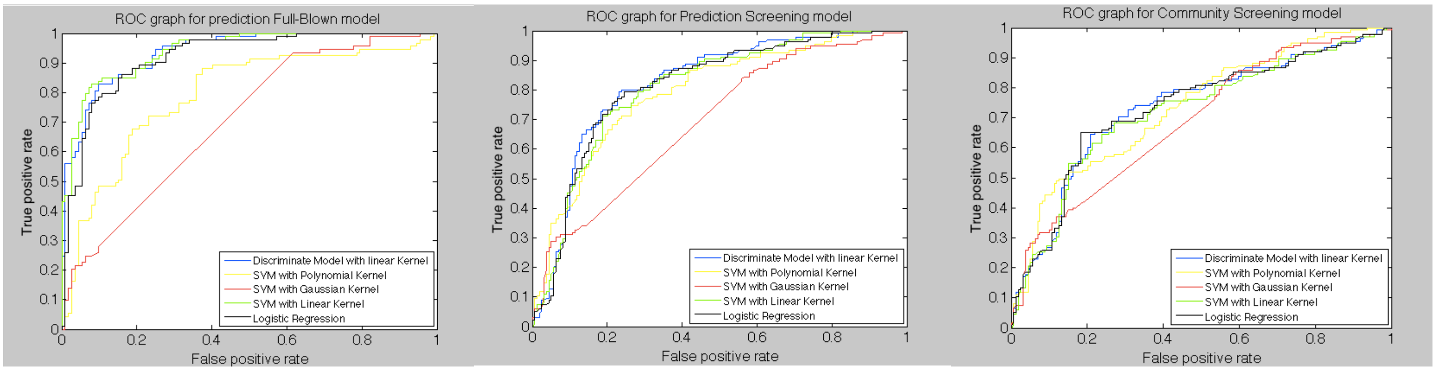

Figure 1-3: ROC Curve for each of the learning algorithms for each model

From the ROC curve, we can see that it is a tight fight among SVM(linear kennel), LDA and Logistics Regression in Prediction(Full-Blown) and Prediction(Screening), and they are well above the SVM(polynomial) and SVM(rbf Gaussian). However, as we delete the features from oral examination and biologic tests in the community-screening model, the advantage diminishes. The reason behind is that inappropriate feature space will distort the power of the good learning algorithms since the oral examination and biologic tests links more directly to the outcome of caries.

5.2 Feature Selection Results and Reduced Models' Results

To fully explore the performance of the statistical models and reduce the workload of clinicians by shrinking the feature space, we perform feature selection using LDA for Prediction Full-Blown, Prediction Screening and community screening. We selected LDA since it performs almost as good as logistic regression and it's time efficient. The selected features for each model are listed in the following:

For the Prediction Full-Blown model, we selected: Singapore Nationality, Advise of diet and caries, LB level, MS bacteria level, average PH of saliva and baseline caries existence. For Prediction Screening, we only select main reasons of tooth decay – heatiness and baseline year caries existence. For the Community Screening model, we select race of Indian, children taken care by child care service and parents awareness of children's caries.

The performance after the feature selection is listed in Table 7:

Models	Training Error	Test Errors
Prediction (Full-Blown)	0.1025	0.1561
Prediction (Screening)	0.2201	0.2415
Community Screening	0.2872	0.3537

Table 7: Reduced Model Performance

The ROC comparison of data spaces before and after feature selection is shown in Figure 4 – 6.

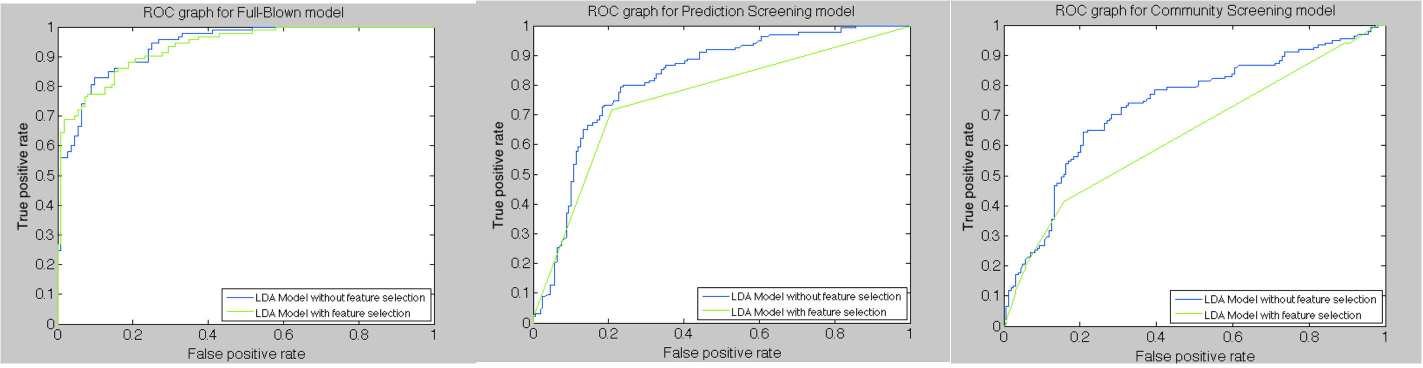

Figure 4-6: ROC Curve for before and after feature selection

We observe that the performance of reduced model is slightly inferior for high true positive region in Prediction Full Blown model, even though the performance is similar. However, for other two cases, the performance is indeed inferior in the other two models. The problem can be related to the feature selection method we employed. Since we only use wrapper method with forward search, it will stop early if the algorithm finds an addition feature which does not improve on its performance. However, there might exist a set of features (does not include current selected features) that has better performance but is neglected because of the non-exhaustive search. Therefore, the performances are inferior to the full models, even though in theory, the feature selection would improve the overall performance when the subsets of features were exhaustively searched.

6. Conclusion:

In this project, various statistics models with employment of different machine learning algorithms are developed and compared with each other. In addition, we figure out different subset of features for each model, which contribute to strike a balance between model performance and clinical utility. Nevertheless, there is much more opportunities in future to further develop this Caries Risk Assessment Model. A summary of further development is followed:

- Extending our model to different outcome variables, i.e. number of caries, position of caries surfaces, etc.
- Optimizing the feature selection process, i.e. use backward/genetic search and incorporate a penalty cost for number of features selected to avoid the increasing number of features selected in the backward/genetic search.
- Determination of optimal sample size to reduce the amount of effort in collecting ample amout of data which are very costly in research settings.

Bibliography:

Gao X.L, Hsu C.Y-S., Xu L., Hwarng H.B., Goh T., Koh D.,(2010) “Building Caries Risk Assessment Models for Children” J DENT RES 2010; 89; 637 originally published online Apr 16, 2010; DOI: 10.1177/0022034510364489

Selwitz, RH (2013). "Dental caries". PubMed. Lancet. Retrieved 10 December 2013