Feature Selection for the Binary Classification of Severe-Sickness Experience in HIV-Exposed but Uninfected Infants Involving Multivariate Longitudinal Dataset

Anonymous Author(s)
Affiliation
Address
email

Abstract

This project was aimed at developing a procedure for inferring the combined effects of a set of input variables relating to a binary response (severe-sickness experience). Typically data analysis on similar multivariate longitudinal data considers the multiple responses separately when determining correlation to a class, for example inferring the difference in mean measurement of a variable for two classes. The devised procedure involved constructing a classification model with the available data. Features were first selected with linear SVM weights using different splits of inputs. Selected features were then recombined for further filtering. The final selected subset of features was able to predict the class of severe-sickness experience with 85% classification accuracy when using k-Nearest Neighbor algorithm.

1 Motivation

The group of HIV-exposed but uninfected infants (HEU) has been the center of HIV research studies in recent years. The interest lays particularly in the health of HEU compared with infants born to healthy parents (un-exposed, hereby referred to as UE). Several studies have observed higher morbidity rate [7], higher chance of contracting serious illness and infections for the HEU [12] [1]. These studies suggested evidence that HEU infants differ from UE infants in that they are more likely to become severely sick. Based on this evidence, further research efforts had been spent to uncover the cause for a weaker immune system in HEU, in order to understand the disease and to perform preventative measures to increase survival rate of HEU infants.

Research in this direction however, is restricted to observational studies since experimenting on the human immune system is either difficult or limited due to ethical concerns. However, observational studies are generally less powerful for making inferences, because the monitored responses may or may not be correlated with the outcome of interest, in this context, whether a child contracts severe sickness or does not.

Many efforts had been made to understand the immune system of HEU. While the majority of these studies kept track of a multitude of response variables, for example various anti-body responses [5], the thymic size and T-cell, B-cell levels [6], the type and frequency of diseases [9], data analysis were performed on individual responses independent of one another. But because longitudinal immunology studies take a long time and are costly, it is also important to examine the effect of immune parameters in conjunction in order to harvest the most information out of observational studies. The following sections of this paper will describe a procedure for identifying a subset of immune parameters most correlated to the outcome whether a child becomes severely sick when used in combination as inputs to a classification model.

2 Data Description

The data came from a pilot study conducted by the Kollmann Lab at the Child and Family Research Institute in Vancouver. Measurements of immune parameters were taken from two groups of infants, experimental (HEU) and control (UE) groups over a course of 2 years. The original data included 28 HEU, 30 UE, and 4 HIV subjects for a total of 62. The scope of this project is limited to 2 types of data: haematology measurements and cytokine levels measured via flow cytometry on monocytes, mPC, pDC cells, and the intercellular fluid. A total of 74 different measurements were taken for a single time period for each subject.

3 Goal and Challenges

The goal of this project was to reduce the set of 74 measured variables into a small subset most correlated with severe-illness experience, in other words, to perform feature selection. A binary classification model was constructed to model correlation to a binary target (0,1) with 1 indicating that subject experienced severe-illness. Apart from the small sample-size relative to the number of parameters measured, there were other challenges presented by this dataset. Following sections adress the challenges of missing data, and of modeling inter-correlated, multivariate longitudinal data.

3.1 Correlated Inputs

The data was transformed to orthogonal directions using principle component analysis (PCA), which addresses the problem of highly correlated inputs, as described in the book “The Elements of Statistical Learning” [3].

3.2 Missing Data

This data set contained a high percentage of missing data (see Figure 1). Case-wise deletion, mean imputation, MLP imputation, Maximum Likelihood with EM are amongst some of the ways to treat missing data [2]. Being the simplest solution, case-wise deletion had been proven to yield the poorest classification result [10] and introduce bias [4]. Due to the small number of subjects in this study, further reduction in available data was to be avoided. Performance of Expectation Maximization (EM) algorithm was very often amongst the top of the available choices [2] [10]. In this application correlation between inputs could not be ruled out, therefore missing entries were imputed via EM algorithm for PCA [11], by which the covariance structure of the data would be preserved. Imputation via PCA with EM allowed 34 more subjects (cases) to enter the analysis.

3.3 Longitudinal Multivariate Data Modeling

Classifying longitudinal multivariate data is no trivial task by means of rigorous statistical methods. Part of the difficulty is a result of the lack of literature on classification model for longitudinal multivariate data. Verbeke suggested for this particular purpose a heterogeneity model (mixture model) based on linear mixed effects model [13]. These linear mixed effects models involve an estimated variance-covariance structure. Estimation could be achieved through EM [13], or first order Taylor-Approximation. [8]. In this project, linear models were constructed with PCA adjusted inputs for the purpose of bypassing the need of the covariance matrix.

4 Experimental setup

In preliminary data processing, missing values were first imputed via EM with PCA. For each available case, the slopes and means of measurements were the input features chosen to represent the linear model of longitudinal profile. The correlation between every possible pair of variables was

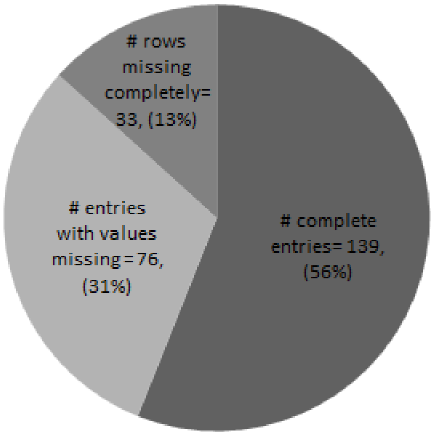

Figure 1: Figure showing the high percentage of missing data in the original data set. Each subject was measured at 4 time periods producing 4 entries per subject for a total of 248 entries.

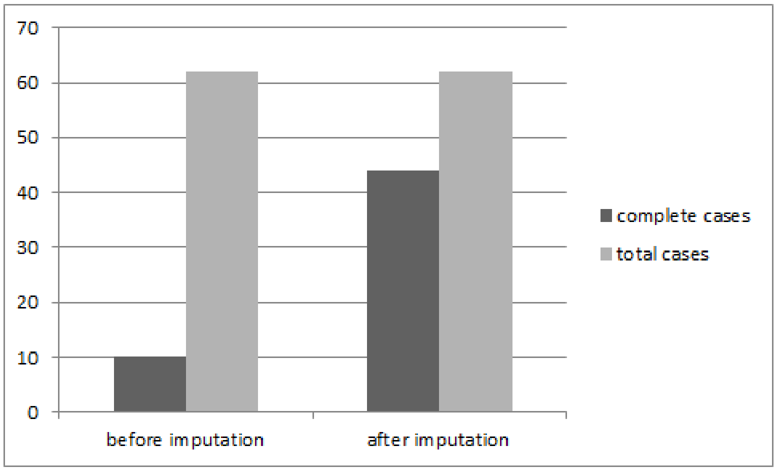

Figure 2: Significant improvement in the number of subjects with complete records for all four time periods (complete cases) was observed after data imputation. Data imputation was not performed on rows that were missing completely.

also calculated using the Pearson product-moment correlation coefficient. Table 2 summarizes the number of input pairs above certain correlation levels. Note that there were 2,701 possible pairings out of 74 input variables.

SVM, Nave Bayes, Random Forest, K-Nearest Neighbors, and Logistic Regression algorithms were compared and used as guidelines for feature selection. The algorithms were implemented with default parameters using the data-mining package Orange. Feature-selection was automated through the preprocess module in Orange based on linear SVM weights. Performance indicators (classification accuracy, area under the ROC curve, Brier score) of the various algorithms were evaluated through 50 fold cross validation.

For comparison purposes, both PCA treated data and original data were used. The slopes and means were also considered in isolation to see if either set was more important for predicting severe-illness experience. Performance indicator values from all algorithms were averaged, to produce a single measure representing the classification ability of the 3 splits (all features, slopes only and means only) of inputs. Upon running the classification algorithms with the 3 splits of inputs, features

Table 1: Amount of correlation within input variables represented by the number of input pairs above certain correlation levels

Correlation above	# Input pairs
0.5	60 pairs
0.6	44 pairs
0.7	17 pairs
0.8	4 pairs
0.9	2 pairs

Table 2: Classification accuracy averaged across all machine learning algorithms. Results shown for several splits of input variables and two possible treatments (PCA and feature selection) for each split. Feature selection using linear SVM weights provided improved accuracy for all splits of inputs

	with PCA		Original Data	
	Feature Selection	No Feature Selection	Feature Selection	No Feature Selection
All Inputs	0.653	0.548	0.632	0.575
Slopes Only	0.665	0.558	0.636	0.580
Means Only	0.631	0.593	0.626	0.573
With 2nd Stage Feature Selection:			0.686	N/A

auto-selected by the data-mining package were manually filtered and combined to form a new s of input features. Classification was again repeated for this new subset of manually selected features with further feature selection via linear SVM weights. The performance of all machine learning algorithms were evaluated against varying number of features (from 1 to 22) selected from this same subset.

5 Results and Discussion

Table 1 provides a summary of the amount of correlation between input data. Table 2, Figure 3 and Figure 4 summarizes classification performance of the various machine learning algorithms. As seen in Table 2, PCA adjusted inputs showed only a slight improvement compared to the original inputs. Further experimentation would be required to determine the statistical significance of this slight difference. This was most likely due the majority of inputs being weakly correlated. Generally, Pearsons product-moment correlation coefficient must be greater than 0.7 to be considered strongly correlated. Only 17 pairs of inputs demonstrated correlation above this threshold out a total of 2,701 pairs (less than 1%).

Most classifying algorithms investigated in this project showed the trend of performing better initially as features are increased, but again decreased in performance when too many features were added. See Figure 3. The K-Nearest-Neighbor algorithm in particular was most affected by the addition of unimportant features. Nave Bayes and Random Forest seemed less affected by addition of inputs, but further investigation would be required to infer this behavior in general. Note that other performance indicators, namely area under the ROC curve and Brier score, showed a similar pattern. See Figure 4.

A big improvement in performance was observed for using a subset of inputs as filtered by linear SVM weights. See Table 2. Even better classification results, as high as 85% accuracy when using K-Nearest-Neighbor, were obtained when selected features from the different input splits were further filtered. See Figure 3. While exhaustive search for the best subset would be unrealistic, multi-stage filtering and recombination seemed to produce good results even when only two filter stages were used as with this project. However, the number of subjects in this study (n = 62) was quite small. Further experimentation with large sample data would be required to provide stronger evidence. Future work would also include investigating filtering inputs from random splits of various sizes at each stage, and experimenting with the number of stages required to find a best result.

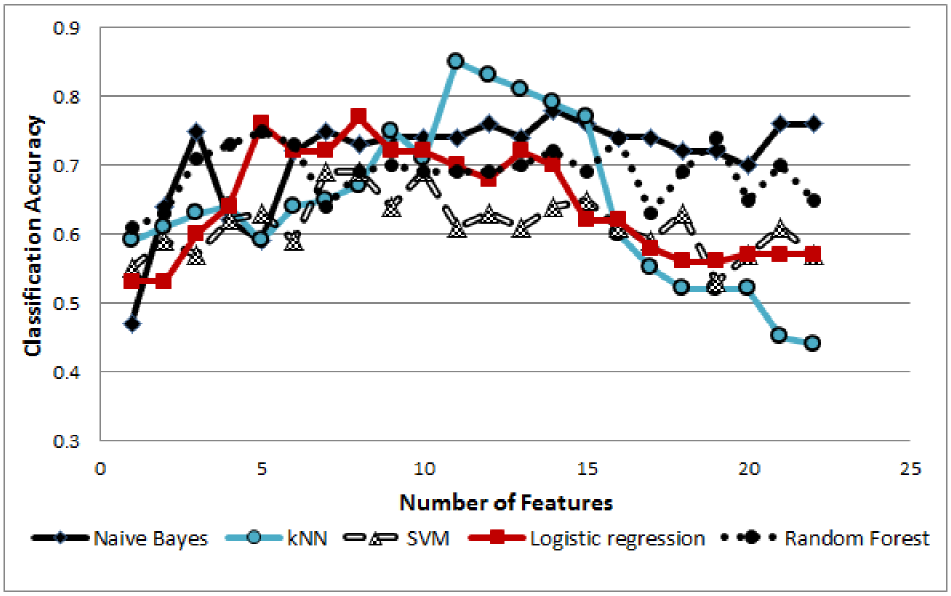

Figure 3: Classification accuracy of various machine learning algorithms using the manually selected feature subset. As the number of features increased initially, all algorithms showed improvement in performance. Note the drastic decrease for K-Nearest-Neighbor, from 0.85 to 0.44, between 11 to 22 features.

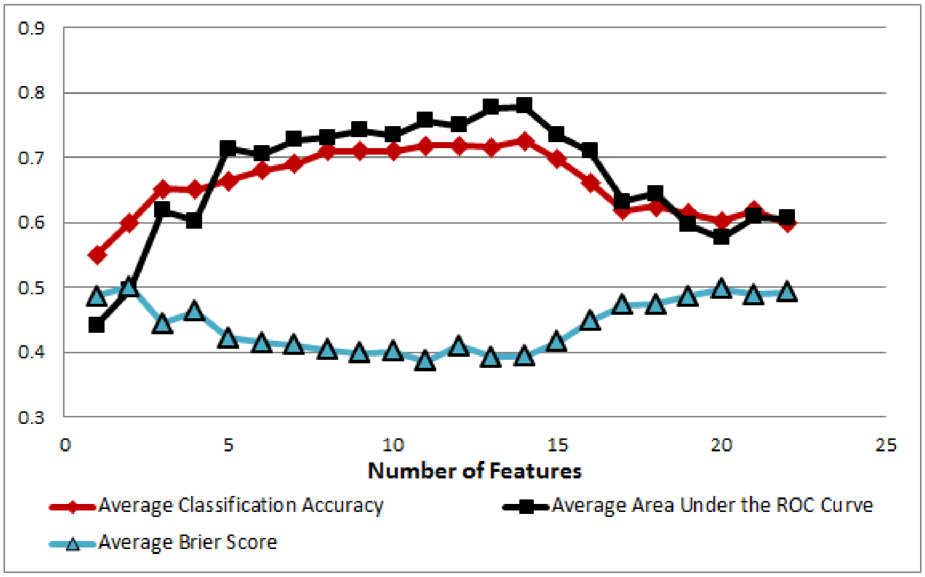

Figure 4: Average performance of all machine learning algorithms used in this project. Performance was measured in terms of classification accuracy (CA), area under the ROC curve (AUC), and Brier score. Note that high values CA and AUC indicate better performance, while low Brier score indicates better performance. The performance of algorithms increased initially with addition of features, but later decreased when too many features were in the active set.

6 Conclusion

Feature selection was defined as the ultimate goal for applying machine learning methods to this particular observational study as another way to extract information from multivariate longitudinal data. This study successfully extracted a subset of features, when considered in conjunction, provided high classification rate of target binary variable, namely severe-sickness experience. Classification without any feature selection was only slightly better than random. With feature selection, the classification accuracy increased above 60%. The procedure of 2-stage filtering used in this project showed as high as 85% accuracy with the K-Nearest-Neighbor algorithm. Other similar studies in biology often also produce data with very high dimensions. Performing an exhaustive search for the best subset would be very time consuming if not unrealistic. As an extension of the procedure followed in this project, further research would focus on developing a search procedure with multi-stage feature selection with random recombination of features as a favored alternative for obtaining the best feature subset compared with an exhaustive search method.

References

[1] C. Epalza, T. Goetghebuer, M. Hainaut, F. Prayez, P. Barlow, A. Dediste, A. Marchant, and J. Levy. High Incidence of Invasive Group B Streptococcal Infections in HIV-Exposed Uninfected Infants. *Pediatrics*, 126(3):e631–e638, 2010.

[2] P. García-Laencina, J. Sancho-Gómez, and A. Figueiras-Vidal. Pattern classification with missing data: a review. *Neural Computing & Applications*, 19:263–282, 2010.

[3] T. Hastie, R. Tibshirani, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer series in statistics. Springer, 2001.

[4] J. Honaker and G. King. What to do About Missing Values in Time Series Cross-Section Data. *American Journal of Political Science*, 54:561–581, 2010.

[5] C. E. Jones, S. Naidoo, C. De Beer, M. Esser, B. Kampmann, and A. C. Hesseling. Maternal HIV Infection and Antibody Responses Against Vaccine-Preventable Diseases in Uninfected Infants. *JAMA: The Journal of the American Medical Association*, 305(6):576–584, 2011.

[6] L. Kolte, V. Rosenfeldt, L. Vang, D. Jeppesen, I. Karlsson, L. P. Ryder, K. Skogstrand, and S. Dam Nielsen. Reduced Thymic Size but No Evidence of Impaired Thymic Function in Uninfected Children Born to Human Immunodeficiency Virus-infected Mothers. *The Pediatric Infectious Disease Journal*, 30(4), 2011.

[7] A. Koyanagi, J. H. Humphrey, R. Ntozini, K. Nathoo, L. H. Moulton, P. Iliff, K. Mutasa, A. Ruff, and B. Ward. Morbidity Among Human Immunodeficiency Virus-exposed But Uninfected, Human Immunodeficiency Virus-infected, and Human Immunodeficiency Virus-unexposed Infants in Zimbabwe Before Availability of Highly Active Antiretroviral Therapy. *The Pediatric Infectious Disease Journal*, 30(1), 2011.

[8] G. Marshall, R. De la Cruz-Mesía, A. E. Barón, J. H. Rutledge, and G. O. Zerbe. Non-linear random effects model for multivariate responses with missing data. *Statistics in Medicine*, 25(16):2817–2830, 2006.

[9] M. M. Mussi-Pinhata, L. Freimanis, A. Y. Yamamoto, J. Korelitz, J. A. Pinto, M. L. S. Cruz, M. H. Losso, and J. S. Read. Infectious Disease Morbidity Among Young HIV-1Exposed But Uninfected Infants in Latin American and Caribbean Countries: The National Institute of Child Health and Human Development International Site Development Initiative Perinatal Study. *Pediatrics*, 119(3):e694–e704.

[10] M. B. Richman, T. B. Trafalis, and I. Adrianto. Missing Data Imputation Through Machine Learning Algorithms. In S. E. Haupt, A. Pasini, and C. Marzban, editors, *Artificial Intelligence Methods in the Environmental Sciences*, pages 153–169. Springer Netherlands, 2009.

[11] S. Roweis. Em algorithms for pca and spca. In *in Advances in Neural Information Processing Systems*, volume 10, pages 626–632, 1998.

[12] A. L. Slogrove, M. F. Cotton, and M. M. Esser. Severe Infections in HIV-Exposed Uninfected Infants: Clinical Evidence of Immunodeficiency. *Journal of Tropical Pediatrics*, 56(2):75–81, 2010.

[13] G. Verbeke and G. Molenberghs. *Linear mixed models for longitudinal data*. Springer series in statistics. Springer, 2000.