Comparisons of Machine-Learning Methods to Predict Diagnostic Values for Leprosy Infection Based on Protein Microarray Data

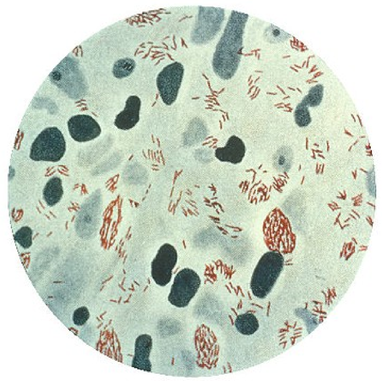
Yue Li[1], Rolando Pajon[3], Mik Bickis[2] and Anthony Kusalik[1]

[1]*Department of Computer Science and* [2]*Department of Mathematics and Statistics, University of Saskatchewan and* [3]*Department of Microbiology and Infectious Diseases, University of Calgary*

I. Introduction

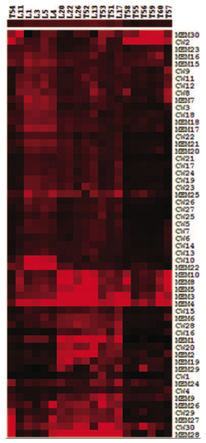
Skin lesion of tuberculoid (left) and lepromatous (right) leprosy infection as the basis of clinical diagnosis

M. Leprae colonies as the basis of bacterial index (BI) measurement



ELISA measurement of antigen PGL-I for leprosy diagnosis

M. Leprae protein microarray measuring humoral immunity pattern

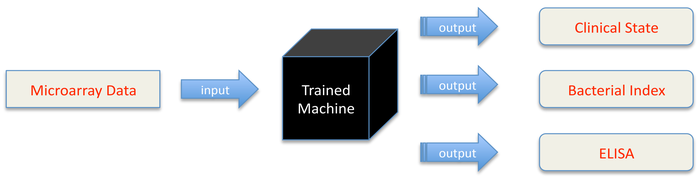


II. Motivation:

Protein microarrays provide a new opportunity for developing a more efficient diagnostic approach for infection diseases than the other three.

III. Hypothesis:

Using an appropriate machine-learning approach, humoral response measurements from protein microarrays can be used to reliably predict other diagnostic values including clinical diagnostic state, and BI or ELISA to a particular microbial antigen.



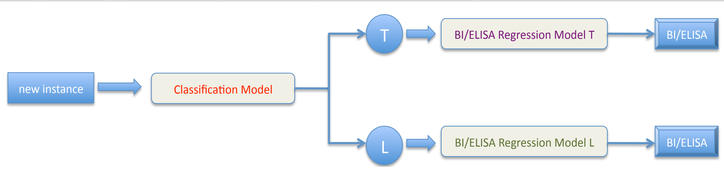
IV. Material

- Microarray data (data source: Groathouse *et al.*, 2006)
  - 56 native protein probes
  - 20 patient sera: 10 lepromatous (L) sera; 10 tuberculoid (T) sera
- ELISA to antigen PGL-I measurements for the 20 sera
- Bacterial index of the 18 sera (with 2 missing values)

V. Methods

General workflow of evaluation of each prediction model using leave-one-out cross validation. In the table, column 'S' stands for leprosy state, 'BI' for bacterial index and E for ELISA to PGL-I. 1 to 56 are the 56 native proteins.

Proposed approach: Regression by Stratification to predict BI/ELISA.



VI. Results

A. Binary Classification of Leprosy Sera Sample into T or L form

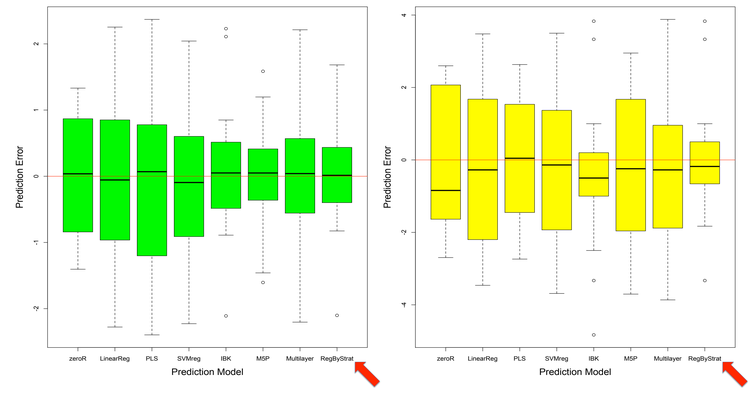
Accuracy of eighteen classification models in predicting leprosy states

Model	zeroR	DMNB	VP	Bagging	KM	J48	JRip	LWL	LTree
Accruracy %	0	35	50	60	65	65	70	75	75
Model	DTable	NB	BNet	LRegr	SVM	NNet	RF	AdaBoost	NN
Accruracy %	75	75	80	80	80	80	80	85	85

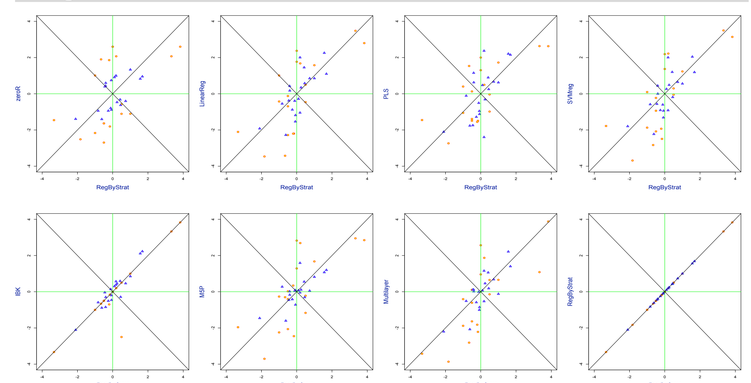
Classifiers full name list: zeroR: predict the mode of nominal class; DMNB: Discriminative Multinomial Naive Bayes; Bagging: meta-method using fast decision tree learner; KM: classification via simple K-means cluster; J48: C4.5 decision tree classifier; JRip: Repeated Incremental Pruning to produce error reduction; LWL: Locally Weighted Learning; LTree: Logistic Boost Decision Tree; DTable: Decision Table majority classifier; NB: Naive Bayes; BNet: Bayes Net; LRegr: Logistic Regression; SVM: Support Vector Machine; NNet: Neural Network; RF: Random Forest with 10 decision trees ; AdaBoost: Adaptive Boosting using Decision Stump as classifier; NN: Nearest Neighbor using normalized Euclidean distance.

B. PGL-I ELISA and BI Predictions

Boxplots of prediction errors of ELISA (left) and BI (right) for eight classification methods for predicting continuous class.



ELISA and BI prediction errors pairwise comparison of Regression By Stratification model against other seven methods. Yellow circle represents BI, purple triangle ELISA measurements.



Regression model list: zeroR: predict the mean of continuous class; LinearReg: Multivariate Linear Regression; PLS: Partial Least Squared Regression; SVMreg: Support Vector Machine for Regression; IBK: K-Nearest Neighbor using Linear Nearest Neighbor Search; M5P: M5 Model Tree Improved; Multilayer: Multilayer Perceptron; RegByStrat: Regression By Stratification method uses IBK+M5P for predicting ELISA and IBK+IBK for BI.

VII. Conclusion and Future Work

Serological reactivity patterns can be exploited for predicting other diagnostic values via appropriate machine-learning methods. Potential improvements of the regression by stratification method are: 1) use a mixed training set for each regression model, 2) incorporate a preceding feature selection stage, and 3) optimize parameters of the models used in the method. Many other approaches are under close investigation.

VIII. Acknowledgment

All the machine learning methods are used via: WEKA The University of Waikato