Genetic Algorithms and Classification Trees in Feature Discovery: Diabetes and the NHANES database

Alejandro Heredia-Langner[1], Kristin H. Jarman[1], Brett G. Amidan[1], and Joel G. Pounds[1]
[1]Pacific Northwest National Laboratory, 902 Battelle Boulevard, PO Box 999 Richland, Washington 99352 USA

***Abstract*— This paper presents a feature selection methodology that can be applied to datasets containing a mixture of continuous and categorical variables. Using a Genetic Algorithm (GA), this method explores a dataset and selects a small set of features relevant for the prediction of a binary (1/0) response. Binary classification trees and an objective function based on conditional probabilities are used to measure the fitness of a given subset of features. The method is applied to health data to find factors useful for the prediction of diabetes. Results show that our algorithm is capable of narrowing down the set of predictors to around 8 factors that can be validated using reputable medical and public health resources. Key Words: Genetic Algorithms, Decision Trees, NHANES, Diabetes.**

1. INTRODUCTION

The National Health and Nutrition Examination Survey (NHANES) database (www.cdc.gov/nchs/nhanes.htm) contains a complete health survey for a sample of the U.S. Population. Included in the survey are nutritional, demographic, and socioeconomic data as well as results of medical examinations and laboratory analyses for the study participants. The survey, which began gathering data in the 1960s, contains information from around 5000 adults and children per year and results are presented in a biannual format, which means that each two-year dataset contains information from about 10,000 respondents.

The NHANES dataset is a rich environment in which a supervised learning algorithm can be applied. The dataset contains hundreds of features in a variety of formats. There are continuous features (age, body-mass index, cholesterol level), ordered categorical features (for example, annual household income value is expressed as integers where, in general, a higher number represents a higher level of income) and unordered categorical features (pregnancy, for example). Data gathered from the NHANES survey has been used in the past to inform and monitor the effects of public policy decisions [1], [2] and by researchers to help test relationships between lifestyle or nutrition levels and medical conditions or illness [3].

Although the vast majority of the NHANES related research uses the data to focus on testing hypotheses involving a small number of previously selected predictors, there has been recent development in applying data mining and pattern recognition algorithms [4], [5] to the information gathered from the NHANES survey. The work in [4] used 2005-2006 laboratory and questionnaire data for 10348 participants and constructed classification trees for an attribute of interest (the respondent has high blood pressure, for example) using the rest of the variables as potential predictors. The work in [4] used total accuracy of prediction as the objective function and discarded decision trees that result in too low (below 80%) or too high (above 95%) predictive accuracy. The work in [4] aimed at discovering predictive relationships among variables in the dataset that may shed new light on the association between health conditions and lifestyle choices, but its broad application produced a myriad of results that may be difficult to sift through and validate. The work in [5] used data for a subset of 4979 respondents. They develop a clustering approach to find associations between conditions of interest (high blood pressure and high cholesterol, for example). Their aim was to explore the data for new and interesting disease associations that could then be substantiated (or disproved) by searching literature in the appropriate field. The work in [5] was only reported for people with known illnesses or conditions, excluding respondents without the diseases, and their results show only disease associations, not associations between diseases and other factors.

The NHANES database contains hundreds of features, some of which may be useful for classification purposes. A key challenge is to find a small set of features whose combined use is optimal in some sense for a classification task, while at the same time avoiding the computationally impractical task of testing every possible combination of factors. The approach presented in this paper uses a combination of decision trees and a Genetic Algorithm (GA) to optimize the selection of a set of predictors that are useful to describe a condition of interest.

2. IMPLEMENTATION AND RESULTS

Genetic Algorithms (GA) are a heuristic optimization technique with mechanisms inspired by the process of evolution and natural selection. A solution to a problem is represented as a string of characters and a number of these solutions are generated, typically at random, to form the initial parent population. New, or offspring, solutions are created by recombining information from selected parents, and the best performing offspring individuals are then selected to form the new parent population. To avoid premature convergence, some solutions in the new parent population are subjected to a mutation mechanism, which may alter their contents. A GA works by repeatedly applying the mechanisms of recombination, selection and mutation to an initial population of solutions until some measure of convergence has been reached [6]. Genetic algorithms are well suited to explore

large and complex problem spaces and are not deterred by noisy, constrained, or discontinuous objective functions. On the other hand, a GA cannot guarantee that an optimal solution will be found.

In this work, a GA is applied to a subset of the NHANES data to find a set of features that best predicts the presence of diabetes. At any given iteration, a solution to the feature selection problem consists of a vector with binary (1/0) entries, where a '1' indicates that the corresponding feature is present and may be used by a decision tree. The initial dataset includes 45 features (including demographic information, cholesterol data and body measures from 9965 respondents in the 1999-2000 NHANES database), and the initial population size is 35 solutions, each solution generated randomly. The recombination mechanism produces 175 offspring solutions (five times the size of the parent population). Each offspring solution is formed using two randomly selected individuals from the parent population. The new solution is created by joining alternating portions of each parent. Individuals in the offspring population are evaluated using a 90/10 train/test strategy, where the partitions are created anew in every generation to maintain, roughly, the proportions of (1/0) present in the overall dataset, and the best performing 35 individuals are selected for further processing. In the next step, up to 20% of the 35 individuals selected are mutated by having their contents randomly altered (it is possible that any given entry in a mutated solution remains unchanged). The GA was implemented in MatLab [7], using the CLASSREGTREE tree function.

A somewhat related approach was implemented in [8] who employed a GA for feature and instance selection using a support vector machine (SVM) and k-NN (nearest neighbor) classifiers on several datasets. The work presented in [8] focused solely on accuracy of prediction and their results suggest that in a dataset with many potential features, it is possible to greatly reduce the number of features without, in most cases, affecting classification performance. The work in [8] does not measure feature importance -- it is implicitly assumed that all features selected by their algorithm are equally important-- and their interest lies mostly in comparing accuracy of classification when feature and instance selection are used individually or together. The application of a GA for feature selection is also presented in [9].

During a run of the GA for the present work, the GA procedures are executed repeatedly, keeping track of the number of times each feature is present in every new parent population and the best objective function value in every new generation. The variable used for classification is the diabetes 1999-2000 set where the response was re-coded as follows: respondents with diagnosed diabetes or borderline diabetes are coded as '1', respondents without diabetes are coded as '0' and individuals that responded 'Don't know' or refused to answer are coded as 'N/A' (not available). The diabetes dataset is highly unbalanced, with around 5% of the respondents affected by the illness. Aside from cholesterol levels and body measures (such as waist circumference, height, weight), other demographic predictors (or features) include information about age, gender, ethnicity, education and income level, military veteran status, and others. Feature data were pre-processed to re-code values not useful for classification. This particular dataset was chosen to develop and test the optimization approach described in this document because findings on the relationship between diabetes and factors like age, ethnicity, socioeconomic and cholesterol data can be supported by numerous reputable sources (http://diabetes.niddk.nih.gov/dm/pubs/causes/#causes, and [10]).

The objective function developed for the GA produces conditional probability estimates of having diabetes or borderline diabetes for a given set of predictors and predictor levels. The simplest way of computing conditional probability estimates using binary classification trees involves computing the relative frequency of one of the classes at the leaves (terminal nodes) for a set of training data. Using relative frequencies as conditional probabilities is known to produce very poor estimates [11], because terminal nodes may have high purity (a single class assigned to it) but a very small number of observations. This is particularly true in highly unbalanced datasets like the one used for this work. Better probability estimates can be obtained by smoothing [12], curtailment [13], or averaging [14] probability estimates, or by applying a combination of these techniques. In this work, several different probability estimates were tested. These estimates were obtained using Laplace estimators, m-estimators and values from a Naïve Bayesian classifier, either alone or in combination. Probability estimates were used as inputs for an objective function in the form of the average negative cross entropy (NCE, [15]).

The task of developing and testing the algorithm was carried out including redundant predictors in the initial set of features. For example, NHANES contains several features related to family income. This approach was used to observe the behavior of the GA under different formulations of the objective function, as the goal is to develop an approach that can be applied to datasets that have received only a minimum of pre-processing. The objective function combined Laplace and binned Naïve Bayesian probability estimates [13], and aimed to minimize the average NCE of the test sets.

In total, the candidate set of predictors contained 45 variables, including demographic variables (age, ethnicity, gender, family income and others), results from blood analyses (total cholesterol, HDL cholesterol, C-reactive protein, Helicobacter pylori, fibrinogen, bone alkaline phosphate, N-telopeptides), and body measures (weight and BMI, waist circumference, arm circumference and others). Twenty generations of the GA produce the results shown in Figure 1.

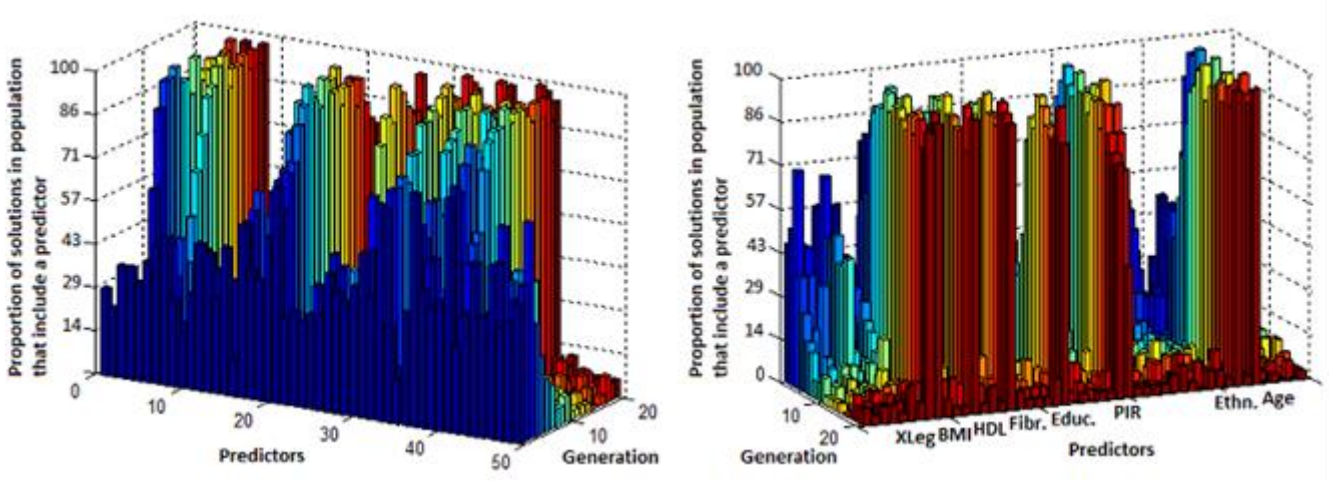

Figure 1. Proportion of predictors present in the population of solutions (z-axis) as a function of generation number (y-axis) with the first generation in the forefront (left panel) and the last generation in the forefront (right panel). The predictors that appear in a large proportion of the final population are Age, Ethnicity, Poverty income ratio (PIR), Education level, fibrinogen level, HDL cholesterol level, Body mass index (BMI), Upper leg length (XLeg).

Figure 1 shows the evolution in the proportion of predictors present in the parent population of the GA starting with the first generation (created at random) until the 20^{th} generation. The first generation, shown in the left panel of Figure 1, contains all predictors in roughly the same proportion. As the run progresses, some predictors were effectively eliminated, while a few others tended to be present in nearly all the individuals in the population of solutions. Results shown in Figure 1 were obtained after evaluating 35×5×20 = 3500 solutions, not necessarily distinct. The problem space consists of around 3×10^{13} possible solutions (a predictor is or is not present and there are 45 predictors available). This means that convergence has been achieved after exploring less than10^{-8} % of the problem space, in other words, a small fraction of all the solutions. Scientific support for the validity of the predictors selected by the GA can be found in [16],
[10], and
http://diabetes.niddk.nih.gov/dm/pubs/causes/#causes.

In addition to being efficient, the GA appears to be robust as well. The set of final features, shown in Figure 1 appears to be largely independent of the starting population. In particular, the GA was run several times with different, randomly selected starting populations and these runs produced essentially the same final population. The only notable differences were the substitution of a related variable for one of those listed above, for example waist circumference instead of body mass index (BMI). On the other hand, the final population doesn't necessarily contain only critical variables, those predictors that are important for the success of a classification tree. In fact, due to the way a GA processes information, it is possible for features to appear in the final population without having any role in the classification tree, simply because they happen to be chosen jointly with other, more critical predictors.

In the interest of finding the smallest, most critical feature set, it is useful to measure the relative importance of each predictor present in a given solution. In general, feature importance in decision trees is estimated by determining if the splitting variable improves the purity of the node (http://www.mathworks.com/ help/stats/classregtree.varimportance.html). Unfortunately, in some of our trials, calculating variable importance in this way produces results contrary to available information (ethnicity often appearing as having no importance as a factor affecting the incidence of diabetes, for example). A different approach to defining variable importance in decision trees is presented in [17]. In his approach, [17] proposes counting the predictor variables that direct an individual observation from the root to the leaf of the tree and apportioning importance accordingly. In this paper the approach proposed in [17] was implemented, using individuals in the test set to determine variable importance. Variable importance for a GA solution produced the results shown in Figure 2.

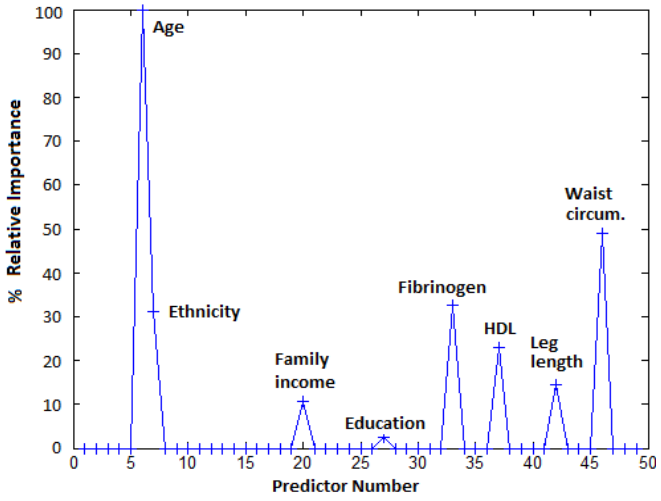

Figure 2. Relative importance of the predictors present in a GA solution. The y-axis represents the proportion of individuals in the testing set that are directed down the tree using the variables shown inside the box. HDL refers to HDL cholesterol. According to this analysis, age has, by far, the largest impact.

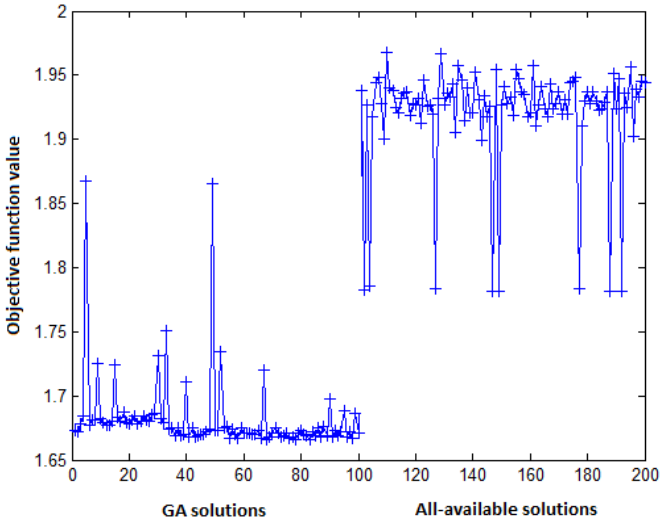

Figure 3. Average performance of GA-generated solutions (first 100 values shown) and "all-available" solutions using the same training/test sets. The objective function is the loss function described in the text. The GA solutions, in general, perform better than those for which all predictors are potentially available.

Combining the information from Figures 1 and 2 provides a much clearer picture of the importance of the predictors selected. Because Age was the variable at the root of the decision tree, it affected 100% of the individuals tested and is therefore the most important feature. At the other end, Education level impacted a relatively small percentage of the individuals in the testing set. Despite these differences, we are interested in all of these variables because the tree may identify relatively small portions of the space where an otherwise noncritical variable plays a big role in determining the presence or absence of the disease.

It is useful to compare the feature set obtained using this methodology to a situation where all the predictors are available for the construction of a decision tree. Figure 3 compares values of the average loss function applied to observations in the test sets used as the objective function of the 35 GA solutions in Figure 1 to results of 100 classification trees constructed from the complete set of predictors. In both instances, the same training/testing datasets, randomly created, have been used for every pair of GA/"all-available" solutions and the evaluation is over all folds, so that results can be directly compared.

Figure 3 shows that, on average, the GA solutions are better than the "all-available" solutions. This conclusion is likely a consequence of the greedy nature of the splitting algorithm when confronted with a large number of predictors, especially if some of the predictors are categorical. In addition to reducing the average loss function as shown in Figure 3, the GA also produces much more parsimonious solutions, with many fewer variables than the "all-available" case. This result is important because the "all-available" solutions generally produce many non-zero importance values, resulting in a confusing picture and making it difficult to discriminate between more and less important predictors.

3. CONCLUSIONS

Using a dataset from the NHANES database, an optimization methodology that employs binary classification trees, genetic algorithms and a probability-based loss function has been employed to build decision trees with a small number of features, effectively and efficiently pruning a large number of variables down to a small number of highly important predictors. The predictors for diabetes found (age, ethnicity, income, education level, HDL cholesterol level, fibrinogen level, and two body measures) can be validated through reputable sources in the medical and public health fields. As implemented, the methodology allows for a more complete understanding of a complex variable space, including (1) the elimination of uninformative or redundant features, (2) the discovery of the most important predictors, (3) the level at which a given predictor is useful for discrimination, and (4) the relative importance of the predictors found. This approach allows to efficiently mine a database, identifying a small but important set of predictors for diabetes without having to elicit

input from subject-matter experts or start from a well-defined hypothesis. In its current form, the decision trees produced by the GA can be examined to indicate combinations of features and feature levels that make a difference between subpopulations with high and low probabilities of diabetes. These findings may be useful in furthering understanding of factors that can be changed, cholesterol levels and body mass index, for example, to improve probabilistic health outcomes when other risk factors, such as age and ethnicity, are present.

In future work currently underway, the methodology shown in this paper is being applied to other health-related responses, using larger sets of predictors from NHANES, with the objective of discovering identifying features for conditions that are not well understood and developing probabilistic predictions when a given set of predictor levels are present.

ACKNOWLEDGMENT

The research described in this paper is part of the Signatures Discovery Initiative at Pacific Northwest National Laboratory. It was conducted under the Laboratory Directed Research and Development Program at PNNL, a multiprogram national laboratory operated by Battelle for the U.S. Department of Energy.

REFERENCES

[1] Annest, J.L., Pirkle , J.L., Makuc ,D., Neese, J.W., Bayse, D.D., Kovar, M.G, "Chronological trend in blood lead levels between 1976 and 1980," in New England Journal of Medicine, 308, 1373-1377, 1983.

[2] Yetley, E.A., Johnson, C.L., "Folate and vitamin B-12 biomarkers in NHANES: history of their measurement and use," The American Journal of Clinical Nutrition, May 18, 2011, 1S-10S, 2011.

[3] Li, C., Ford, E.S., Zhao, G., Croft, J.B., Balluz, L.S., Mokdad, A.H., "Prevalence of self-reported clinically diagnosed sleep apnea according to obesity status in men and women," National Health and Nutrition Examination Survey, 2005-2006. Preventive Medicine, 2010, 51(1), 18-23, 2010.

[4] Lee, J.W., Lin, Y.H., Smith, M., "Dependency mining on the 2005-06 National Health and Nutrition Examination Survey," Data. Presented at the American Medical Informatics Association 2008, Washington DC, 2008.

[5] Xing, Z., Pei, J., "Exploring disease association from the NHANES data: Data mining, pattern summarization, and visual analytics," International Journal of Data Warehousing and Mining, 6(3), 11-27, 2010.

[6] Goldberg, D.E., "Genetic Algorithms in Search, Optimization & Machine Learning," Addison Wesley, MA, 1989.

[7] MatLab. Version 7.11.0.584 (R2010b). The Mathworks Inc., Natick, MA, 2010.

[8] Tsai, C-F., Eberle, W., Chu, C-Y., "Genetic algorithms in feature and instance selection," Knowledge-Based Systems 39 (2013), 240-247, 2013.

[9] Leardi, R., Boggia, R. and Terrile, M. "Genetic algorithms as a strategy for feature selection". Journal of Chemometrics, Vol. 6, Issue 5, 267-281, 1992.

[10] Boyle, J.P., Honeycutt, A.A., Venkat Narayan, K.M., Hoerger, T.J., Geiss, L.S., Chen, H., Thompson, T.J., "Projection of diabetes burden through 2050," Diabetes Care, Vol. 24, Number 11, 1936-1940, November 2011.

[11] Provost, F., Domingos, P., "Tree Induction for Probability-based Ranking," Journal of Machine Learning, 52, 3, 199-215, 2003.

[12] Chawla, N.V., Cieslak, D.A., "Evaluating Probability Estimates from Decision Trees," American Association for Artificial Intelligence, 2006.

[13] Zadrozny, B., Elkan, C., "Learning and making decisions when costs and probabilities are both unknown," Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 204-213, 2001.

[14] Tumer, K., Ghosh, J., "Theoretical Foundations of Linear and Order Statistics Combiners for Neural Pattern Classifiers," Technical Report 95-02-98, The Computer and Vision Research Center, University of Texas, Austin, TX, 1995.

[15] Quiñonero-Candela, J., Rasmussen, C.E., Sinz, F., Bousquet, O., Schölkopf, B., "Evaluating Predictive Uncertainty Challenge," MLCW 2005, LNAI 3944, 1-27, 2006.

[16] Kafle, D.R., Shrestha, P., "Study of fibrinogen in patients with diabetes mellitus," Nepal Medical College Journal, 12(1), 34-37, 2010.

[17] Neville, P.G., "Decision Trees for Predictive Modeling," The SAS Institute, 1999.