Evolutionary Feature Selection for Classification: A Plug-in Hybrid Vehicle Adoption Application

Joseph S. Krupa
Civil & Envir. Engineering
University of Vermont
33 Colchester Ave.
Burlington, VT 05405
jkrupa@uvm.edu

Somdeb Chatterjee
Computer Science
University of Vermont
33 Colchester Ave.
Burlington, VT 05405
schatter@uvm.edu

Ethan Eldridge
Computer Science
University of Vermont
33 Colchester Ave.
Burlington, VT 05405
ejeldrid@uvm.edu

Donna M. Rizzo, Ph.D, P.E.
School of Engineering
University of Vermont
33 Colchester Ave.
Burlington, VT 05405
(802) 656-1495
drizzo@uvm.edu

Margaret J. Eppstein, Ph.D
Computer Science
University of Vermont
33 Colchester Ave.
Burlington, VT 05405
(802) 656-1918
Maggie.Eppstein@uvm.edu

ABSTRACT

We present a real-world application utilizing a Genetic Algorithm (GA) for exploratory multivariate association analysis of a large consumer survey designed to assess potential consumer adoption of Plug-in Hybrid Electric Vehicles (PHEVs). The GA utilizes an intersection/union crossover operator, in conjunction with high background mutation rates, to achieve rapid multivariate feature selection. We experimented with two alternative fitness measures based on classification results of a naïve Bayes quadratic discriminant analysis; one fitness function rewarded only for correct classifications, and the other penalized for the degree of misclassification using a quadratic penalty function. We achieved high classification accuracy for three different survey outcome questions (with 3-, 5-, and 7- outcome classes, respectively). The quadratic penalty function yielded better overall results, returning smaller feature sets and overall more accurate contingency tables of predicted classes. Our results help to identify what consumer attributes best predict their likelihood of purchasing a PHEV. These findings will be used to better inform an existing agent-based model of PHEV market penetration, with the ultimate aim of helping auto manufacturers and policy makers identify leverage points in the system that will encourage PHEV market adoption.

Categories and Subject Descriptors

I.2.8 [Artificial Intelligence]: Problem Solving, Control Methods, and Search – *Heuristic Methods*; I.6.5 [Simulation and Modeling]: Model Development – *Modeling Methodologies*; J.2 [Physical Sciences and Engineering]: *Mathematics and Statistics*; J.4 [Social and Behavioral Sciences]: *Economics*.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
GECCO'12, July 7–11, 2012, Philadelphia, Pennsylvania, USA.
Copyright 2012 ACM 978-1-4503-1177-9/12/07...$10.00..

General Terms

Algorithms, Management, Reliability, Performance, Design, Experimentation, Verification

Keywords

Genetic Algorithm, Self-Search, Statistics, Discriminant Analysis, ReliefF, Plug-in Hybrid Vehicles, Agent-based Model, Consumer Survey, Survey Analysis, Alternative Transportation

1. INTRODUCTION

In recent decades, there has been an exponential explosion in our ability to collect and store data for a variety of application domains [31]. However, the development of new data analysis tools to sufficiently mine these large data sets has lagged behind [3]. Consequently, much of the information embedded in large data sets remains undiscovered, resulting in a growing interest in automated tools for exploratory data analysis. Some newer approaches utilize crowd sourcing to help mine large volumes of data, such as that found in astronomy [21] and protein folding [17]. However, even these approaches require one to be quite specific about the type of information desired. Recently, Reshef et al., [27], proposed a promising new exploratory analysis methodology for discovering interesting non-parametric relationships between pairs of variables in large data sets. However, it is not computationally feasible to extend this approach to multivariate relationships.

When searching for unknown multivariate relationships in large data sets, the number of features that can be statistically associated with classified outcomes is ultimately limited by the sample size of the dataset and the dimensionality of the feature space [10]. Consequently, there has been significant work in the field of feature selection to identify optimal feature subsets that enable construction of a predictive classifier without simply over-fiting the data. Various statistical approaches have been applied to feature selection, including those based on entropy and *t*-statistics [22]. However, since the optimal size of a feature set depends on the problem, rather than the classifier [22], it may be difficult to determine the number of needed features.

The ReliefF data-mining algorithm [28] provides a way of estimating the relative strength of association between various features and class outcomes. Because ReliefF considers all features simultaneously, it can detect non-linear interactions between multiple features. For this reason, it has been applied with some success in wrapper-type iterative feature reduction approaches [19][29] designed to detect epistatic genetic interactions associated with disease [7][8][14][15][16][23][28]. However, the power of all wrapper-type ReliefF-based approaches is ultimately governed by the accuracy of the ReliefF weights during each iteration; and once features have been discarded (perhaps mistakenly), they are not recovered using these approaches. On the other hand, evolutionary algorithms are naturally attractive as feature selectors, because they allow the size of the evolving feature sets to shrink and grow as the space is explored and do not require pre-specification of the size of the optimal feature set. Consequently, many researchers have successfully applied genetic algorithms (GAs) [9][20][23][24][26][33] as feature selectors. In [4], a GA using a novel crossover method combining set intersection and set union was found to be very effective in identifying small non-linearly interacting sets of features in noisy data sets.

1.1 Real-World Application: PHEV Market Penetration Survey Analysis

This paper is specifically motivated by a desire to detect novel multivariate associations using real-world consumer survey data, with the aim of discovering which demographic or attitudinal characteristics might best predict whether consumers are likely to become early adopters of plug-in hybrid vehicles (PHEVs).

PHEVs offer consumers many potential advantages over conventional vehicles, including the potential to reduce greenhouse gas emissions [5], while offering driving ranges that are not limited by battery capacity (since these vehicles can also utilize gasoline). However, it is not clear what combinations of governmental policies and manufacturer marketing strategies will be most cost-effective in promoting successful market penetration of this new vehicle technology. In prior work [6], we developed an agent-based model of PHEV consumer adoption to identify nonlinear interactions among potential leverage points and to inform policy-makers and manufacturers of these interactions. Unfortunately, we discovered a paucity of data regarding correlations and interactions between consumer demographics and attitudes that are necessary for initializing inputs in our agent-based model and designing accurate consumer choice rules. To address this issue, in a joint project between the University of Vermont and Sandia National Laboratories, we conducted an extensive online survey of 1000 American adults using the Amazon Mechanical Turk (AMT) crowd-sourcing platform [1][25]. The survey questions include basic demographics, general consumer purchasing behaviors, and more specific questions designed to extract opinions on the environment and comfort in adopting PHEV technology (see Table 1 for a sample of survey features) and to better inform our agent-based PHEV model.

The specific goal of this work is to apply a GA-based approach to explore the search space of over 10^{22} possible feature combinations, to identify which combination of PHEV survey questions ("features") most accurately predict outcomes of interest related to consumers' stated willingness to purchase PHEVs. In this preliminary study, we identify relationships for predicting three types of categorical outcomes, with 3, 5, and 7 classes, respectively. A more complete analysis of the PHEV survey data using this approach will eventually be used to better inform the agent decision-making rules and model described in [6], with the ultimate aim of helping vehicle manufacturers and policy makers prioritize investments toward potential leverage points in the system.

Table 1. Feature descriptions for a selected subset of features in PHEV survey.

Feat. #	Description	# Response Classes
1	Age	7
2	Gender	2
12	My purchasing habits are influence by what others buy.	7*
14	I often identify with others by making similar purchases.	7*
15	To learn more about a brand/product, I often ask my friends.	7*
17	I like to know what brands/products make good impressions on others.	7*
18	I often gather info from friends/family about a product before I buy.	7*
40	Manufacturer ads would influence a future vehicle purchase.	5*
43	Observations of what others drove in my area would influence a future vehicle purchase.	5*
48	How important do you feel it is for the USA to reduce transportation energy consumption?	5*
52	How have concerns of foreign oil dependence influenced your opinion in regards to feature 48?	5*
54	How have discussions with family/friends influenced your opinion in regards to feature 48?	5*
55	How would car company ads influence your opinion in regards to feature 48?	5*
57	How would battery lifetime affect your comfort in purchasing new PHEV tech?	5*
59	How would the potential inconvenience of recharging affect your comfort in purchasing new PHEV tech?	5*
63	How would realizing your GHG emissions could decrease significantly increase your comfort in purchasing new PHEV tech?	5*
64	How would realizing your fuel costs could decrease significantly increase your comfort in purchasing new PHEV tech?	5*
69	How would having PHEV battery swap stations, so as to avoid unexpected costs due to battery failure, increase your comfort in purchasing new PHEV tech?	5*

*Indicates a Likert response scale

2. METHODS: TEST OUTCOMES

2.1 PHEV Survey Data Set

A total of 1,000 PHEV surveys were collected. After a rigorous quality assurance analysis that rejected surveys with incomplete or inconsistent responses, 942 surveys remained for analysis. Likert-type ordinal responses (e.g., strongly agree, agree, neither agree nor disagree, etc.) were treated as equally-spaced integers. Responses that specified ranges of values were given real values equal to the mean of the selected answer (e.g., an income range of $75,000 to $99,999 was assigned a value of $87,500). Ranges of all survey questions were subsequently normalized to the range [0,1]. Of the over 100 questions in the survey, 74 questions were selected as potential features for this study. The median Pearson correlation between all pairs of these features was only 0.04, with a maximum pairwise correlation of 0.78, so no features were excluded as being redundant.

For this preliminary analysis, we selected 3 questions as test outcomes that we felt would be informative regarding the consumer's potential to purchase a PHEV. The analysis aim was to find subsets of the 74 features that could be used to predict the response classes of each of these 3 outcome questions. The first question had 3 possible response classes (designated as test C-3); the second question had 5 possible response classes (C-5); and the third had 7 possible response classes (C-7). These selected questions, and their possible responses, are described in Section 3. A preliminary analysis showed that none of the 74 individual features was strongly correlated with any of the 3 outcomes (maximum pairwise $R^2 \leq 0.13$), making a multivariate approach essential.

Because the class outcomes for all three problems were highly imbalanced, the data set was divided into a training set of 300 respondents and a prediction set of 642. These 300 respondents were selected to ensure the same outcome distributions as the entire survey. All reported resulting classification errors refer to assessment on the prediction set.

2.1 Feature Selection with a GA

We employed the genetic algorithm that comes with the ®Matlab R2011a optimization toolbox to perform feature selection. The GA utilized binary chromosomes comprising of 74 bits, each representing whether a feature was to be selected for use as input to the fitness classifier. Twenty trials were conducted for each outcome, and for each of two fitness functions (described below). We used tournament selection (tournament size of 4), with mutually exclusive crossover 80% of the time and bit-flip mutation the remaining 20% of the time, and elitism of 2 individuals.

We implemented an intersection/union type of crossover, which previous work showed to be effective in feature selection applications [4]. Specifically, to cross two parents, we applied a bitwise AND operator 95% of the time (which returns the intersection of the feature sets of the two parents) and a bitwise OR operator 5% of the time (to help restore lost features) (see Figure 1). Because the crossover operator is fairly aggressive in reducing feature sets, we employed a relatively high mutation rate of 0.2, to maintain diversity in the gene pool.

A preliminary population sizing study was conducted using population sizes $\in$ {100, 500, 1000, 5000, and 10,000} individuals. Results were strongest and most consistent for the largest population size, so we used a population of 10,000 for all subsequent runs reported here.

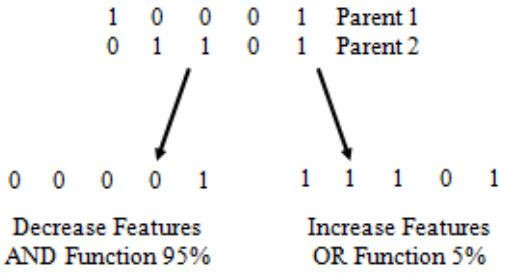

Figure 1. Sample illustration of intersect/union crossover function. Logical 'and' reduces feature sets for 95% of crossovers, and logical 'or' increases features sets for 5% of crossovers to preserve diversity in population.

Fitness was assessed based on classification accuracy of the prediction set, using a parametric discriminant analysis (DA) using the ®Matlab "classify" function in the statistics toolbox. Prior experimentation achieved the most robust classification results with the "diagquadratic" version of the DA, which fits multivariate normal densities, using a diagonal quadratic covariance matrix.

The predicted classes were converted into contingency matrices (P), as illustrated in Table 2 for a hypothetical example of a 5-class outcome for a 100 respondent survey. For any element P_{ij} the i^{th} row represents the observed class value reported by survey respondents, and the j^{th} column is the class predicted by the DA. The interclass distance $d = ||j-i||$ indicates the degree of misclassification represented by a given element. The elements on the main diagonal of the contingency table have an interclass distance of $d = 0$ and indicate respondents that the DA classified correctly (shown in large red font in the hypothetical example in Table 2). Those cells with interclass distance $d = 1$ (the diagonals just above and below the main diagonal, and marked in blue in Table 2) indicate the predicted class was only 1 level away from the observed class. We use the term *extreme* misclassification when a respondent is placed in the two opposite corners of the contingency table (e.g., see Table 2, elements $P_{1,5}$ and $P_{5,1}$, with $d = 4$). Adjacent to Table 2, we display the percent of respondents predicted at $d = 0$, and the percent of respondents found within the 3 center diagonals; where $d \leq 1$, fitness was computed from the values in P in two different ways, as follows.

Table 2. Sample Contingency Table (*P*) with sample calculated scores (see equation (1) and (2), and percentages within Interclass Distances (d) of 0 and 1.

		PREDICTED					
OBSERVED							
	6	7	4	3	0	Flat Score:	0.48
	2	**10**	8	0	0	Quad Score:	2.38
	1	3	**12**	3	1		
	1	1	3	**12**	3	Correct (d=0):	52%
	1	1	0	6	**12**	Correct (d≤1):	87%

The flat penalty fitness function calculates the score of each contingency table retuned by the DA based on the percent of respondents classified exactly (i.e., when $d = 0$). All other values

of d are penalized equally, as illustrated in Figure 2. Since the GA treats optimization as a minimization problem, fitness is computed as the proportion of respondents *not* classified on the main diagonal (i.e., where $d \geq 1$):

$$F(P) = 1 - \frac{\sum Diag(P)}{n}, \qquad (1)$$

where P is the contingency table returned by the DA, n is the 642 surveys in the prediction set, and $\sum Diag(P)$ is the summation of all respondents in the prediction set that were classified correctly. For this scheme, the best possible flat penalty score is 0.

Alternatively, we employed a quadratic penalty fitness function by differentially penalizing misclassifications according to their interclass distance (d) using a quadratic penalty (see Figure 2). For each cell of the contingency tables, its d value is squared and multiplied by the number of respondents predicted to be in that cell. This is summed over all cells and divided by the number of respondents predicted to have $d = 0$, shown below:

$$F(P) = \frac{(\sum_{i=1}^{r}\sum_{j=1}^{c} P_{ij}\|j-i\|^2)}{(\sum Diag(P))}, \quad (2)$$

where P is the contingency table and $d = \|j\text{-}i\|$. The best possible quadratic score is also 0. This type of penalty is appropriate for survey questions with more than 2 response classes where an ordinal relationship exists between classes. For example, misclassifying "strongly agree" as "agree" is better than misclassifying it as "strongly disagree."

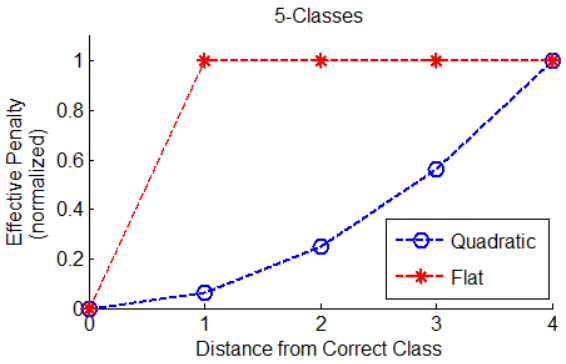

Figure 2. Effective penalties (normalized so the two types of penalties are shown on the same scale) for a 5 class problem, showing how misclassifications are weighted as a function of distance from the main diagonal of the contingency table.

3. GA RESULTS

In each of 20 trials on problem C-3, the same two features were identified as being those most strongly associated with survey respondents' stated willingness as to whether they (1) would not, (2) might, or (3) would seriously consider purchasing a PHEV for their next vehicle, while recognizing that these first generation PHEVs will only be available in compact car models. These features (features 59 and 69; see Table 3) were associated with perceived risks associated with the new PHEV battery technology (see Table 1). Using these two features alone, the DA was able to accurately classify 55% of respondents into their correct categories, and only misclassified 5% into the opposite extreme category (see Figure 3a).

Similarly, in each of 20 trials on problem C-5, the GA consistently identified two features (features 43 and 52, see Table 3) that best predict the degree to which survey respondents considered themselves to be early adopters of new technology in general. One of these features (43) relates to the degree to which social influences affect consumer vehicle purchasing decisions, and the other (52) deals with the degree to which concerns about foreign oil influence the respondents' opinions about the need to reduce transportation energy consumption (see Table 1). Although the DA was only able to accurately classify 34% of respondents into the correct (one of 5) ordinal response classes, classification accuracy increased to 78% when allowing for up to 1-class classification error ($d \leq 1$) (see Figure 3b). Since the 5 response categories were (1) strongly agree, (2) agree, (3) neutral, (4) disagree, (5) strongly disagree, being able to predict these responses to within one class value is quite useful.

Table 3. List of feature sets returned by GA for all 3 questions; survey questions corresponding to these features are shown in Table 1.

Problem	Features Selected in Best Individual
C-3	59, 69
C-5	43, 52
C-7	1,2,12,14,15,17,18,40, 48,54,55,57,59,63,64

We retrospectively used the DA to classify the C-3 and C-5 outcomes using exhaustive search for all possible 2-feature, 3-feature, and 4-feature combinations of the 74 possible input features. Of these 1,218,151 possibilities, the feature pairs selected by the GA (see Table 3) proved to have the best fitness according to both fitness functions (see Table 4). Thus, although

Table 4. GA statistics for all three cases and population size of 10,000 for 20 trials. Reduced Fitness and Number of Features refer to analysis done in Figure 5.

Flat Penalty					
Prob	**Avg. Fit.**	**Best Fit.**	**Avg. # Features**	**Reduced Fit.**	**# Feats. ≥ 50%**
C-3	0.452 ± 0.000	0.452	2.00 ± 0	0.452	2
C-5	0.665 ± 0.000	0.665	2.00 ± 0	0.665	2
C-7	0.649 ± 0.003	0.643	31.5 ± 9.5	0.639	29
Quadratic Penalty					
Prob	**Avg. Fit.**	**Best Fit.**	**Avg. # Features**	**Reduced Fit.**	**# Feats. ≥ 50%**
C-3	1.11 ± 0.00	1.11	2.00 ± 0	1.11	2
C-5	4.78 ± 0.00	4.78	2.00 ± 0	4.78	2
C-7	16.3 ± 0.52	14.8	13.4 ± 4	17.9	7

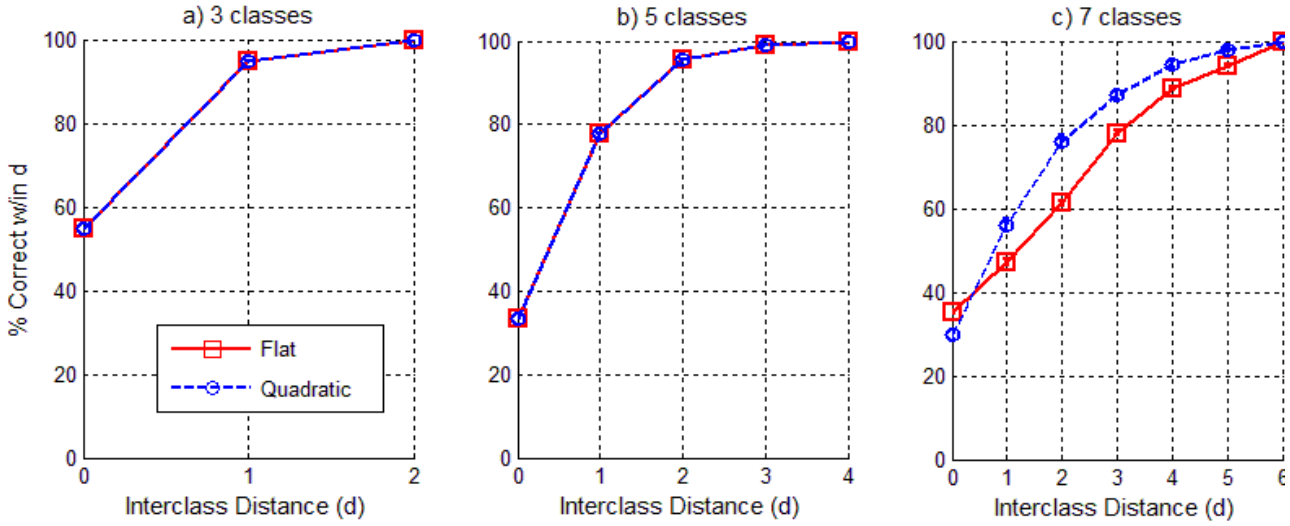

Figure 3. Percentages of predicted classifications within various interclass distances for the best individuals found using both the flat and quadratic penalty fitness functions applied to problem a) C-3, b) C-5, and c) C-7.

it is not computationally feasible to exhaustively explore all possible larger multivariate feature subsets, these results provide strong evidence that our evolutionary approach is effective.

The 7-class problem sought to determine what level of market penetration PHEVs would have to achieve for the respondent to feel comfortable in considering this new technology over a comparable vehicle with a more conventional fuel type, assuming that both vehicles had similar features and were within the purchaser's budget. The seven responses ranged from (1) the respondent would consider being an early adopter of a PHEV, even if they saw no other PHEVs in the current fleet, to only if they saw that (2) 1%-10%, (3) 11%-25%, (4) 26%-50%, (5) 51%-75%, (6) 76%-100% of vehicles on the road were PHEVs would they consider purchasing a PHEV, to finally (7) the respondent would never consider purchasing a PHEV, regardless of how many PHEVs they observed on the road. The GA results on this more difficult classification problem were more variable, between both fitness functions and across the 20 different runs with the same fitness function (see Table 4). Out of the 74 possible features, the flat penalty fitness function selected an average of 31.5 features, with a standard deviation of 9.5 features, whereas the quadratic penalty function found feature subsets with only an average of 13.4 features, and a standard deviation of 4 features. Despite these variations in feature subset size, the actual classification accuracies of the 20 trials were quite consistent for both fitness functions, as indicated by the errorbar plots shown in Figure 3c, where the error bars representing ± one standard deviation in classification errors are so small they are not visible behind the markers.

For C-7, the flat penalty fitness function was able to correctly classify a slightly larger percentage of respondents into their exact class than the quadratic fitness function (35% vs. 30% classification accuracy, see Figure 3c). However, there were more than twice as many features selected using the flat penalty fitness relative to the quadratic penalty fitness (see Table 4). By taking into account the degree of misclassification, the quadratic penalty approach excludes excess features that marginally increase classification accuracy only along the main diagonal. This is supported by the observation that the smaller feature sets obtained from the quadratic penalty fitness yielded better overall contingency tables. That is, if one allowed for slight misclassification in adjacent categories, the quadratic fitness function gave consistently better results (for example, accuracy within 2 class levels was 76% for the quadratic penalty fitness, but only 61% with the flat penalty fitness, see Figure 3c).

Because the number of selected features was so variable across the 20 trials on C-7 (especially when using the flat penalty fitness), yet the fitnesses of these different feature sets were so consistent, we suspected that there were excess features in the selected feature subsets that the DA was simply not assigning much weight to, and so were not enhancing the classification accuracy much if at all. In order to test this hypotheses, we created smaller feature sets for C-7 that included only those features that appeared in a consensus of at least 50%, 60%, 70% 80%, and 90% of the feature sets returned by the 20 trials of the GA. In general, the fitnesses were relatively insensitive to this reduction in set size for the flat penalty sets. For example, the size of the 50% consensus feature set from the flat penalty fitness function was reduced to 29 features with an actual increase in the classification accuracy (see Table 4, columns 5 and 6, flat penalty). With a stricter 90% consensus requirement the feature set was reduced to only 8 features, but there was only a minor decrease in the classification accuracy (see Figure 4a), indicating that there were many excess features in the non-reduced sets returned by the GA using the flat penalty fitness. Furthermore, since the classification accuracy was only noticeably degraded for $d = 0$ by this set reduction process (see Figure 4a) this supports our hypothesis that the flat penalty fitness was including extra features to maximize the number of respondents that are perfectly classified.

In contrast, we see a clear decrease in performance for each level of feature subset reduction of the sets returned using the quadratic penalty function. For example, even at the relatively weak consensus requirement of 50%, the consensus feature set has been reduced to 7 features but shows a 21% degradation in fitness (see Table 4, columns 5 and 6, quadratic penalty), and

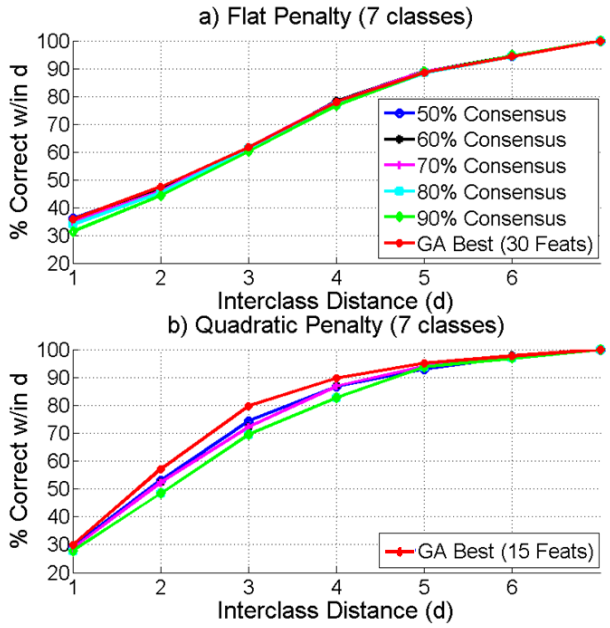

Figure 4: Percentages of those classified within various interclass distances (d). Percentages are of feature subset based on number of times a feature appeared out of all 20 trials for a) flat penalty fitness, and b) quadratic penalty fitness.

overall classification accuracy continues to decrease as the consensus threshold is raised to 90% (see Figure 4b). This provides evidence that the 15 features in the best individual returned by the GA using the quadratic penalty fitness function were probably all important contributors to the overall classification accuracy.

In Table 3, we report the 15-feature subset returned by the GA that had the best quadratic penalty performance. These features included age and sex; several features related to social and advertising influences on purchasing habits, and a desire to reduce transportation energy consumption with commensurate environmental and fuel savings benefits (selected feature numbers in Table 3 correspond to the survey questions in Table 1). Although it is not feasible to validate the optimality of this 15-feature subset through exhaustive search, we did investigate it in two ways. First, for comparison, we used an iterative ReliefF-based feature selection mechanism on the same problem (using the ®Matlab ReliefF function in the statistics toolbox). Specifically, we discarded the single feature that had the lowest ReliefF weight, assessed the fitness of the resulting feature subset with the quadratic penalty fitness function, re-evaluated the remaining feature set with ReliefF, and repeated the process. The resulting fitness measures of these nested feature subsets are shown in Figure 5, where it can be seen that the ReliefF approach never found any feature subset with fitness that was as good as that of the best 15-feature subset found by the GA. In fact, the fitness of the 15-feature subset found by the GA was over twice as good as that identified by ReliefF (see Figure 5). Second, we tried further pruning of the best GA-identified 15-feature subset with the ReliefF method; in all cases, the fitness was made worse by removing any of the selected 15-features, although minimally so for the first few removals (see Figure 5).

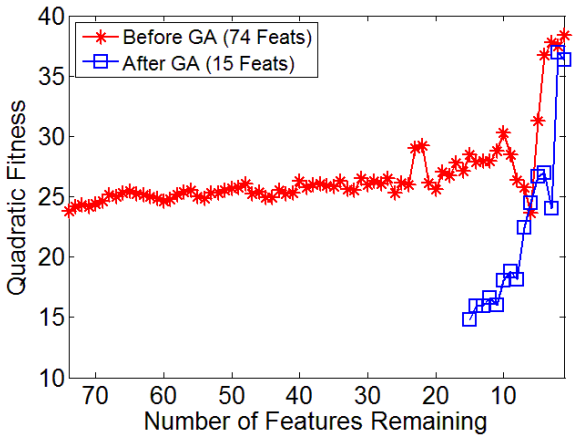

Figure 5. Feature reduction using ®Matlab's ReliefF algorithm for the 7 class problem, applied to the entire 74-feature set and to the 15-feature set found in the best individual returned by the GA.

4. DISCUSSION

In this paper, we applied a GA to preliminary exploratory analysis of a real-world data set comprising a consumer survey related to potential market penetration of PHEVs. Specifically, the GA was used to select subsets of survey questions that were then provided as input features to a quadratic discriminant analysis function for classifying respondents into categories for three different questions (with 3-, 5-, and 7- outcome classes, respectively) concerning attitudes likely to impact future purchase of the new PHEV technology. Misclassifications were penalized in two different ways. In one case, we penalized all misclassifications equally, regardless of how far the predicted class was from the reported class. In the other case, misclassifications were penalized by an amount that increased quadratically with the distance between the predicted class and the reported class.

On the 3-class and 5-class problems, the GA performed very consistently (see Tables 3 and 4, see Figure 3a-b), regardless of which fitness penalty was applied. In both of these problems, the GA selected 2-feature subsets that were subsequently verified to be optimal (at least to within all subsets up to size 4).

The two features most strongly associated with whether or not a consumer would consider purchasing a PHEV for their next vehicle had to do with the respondent's comfort level concerning the new PHEV battery technology. This is consistent with the findings of other recent surveys indicating that consumer uncertainty about issues such as battery life, replacement costs, and recharging time will present some of the major obstacles to PHEV market penetration [30][34].

Somewhat more surprising, we found that the two features most predictive of the degree to which survey respondents considered themselves to be early adopters of new technology concerned the degree to which social influences affected consumer vehicle purchasing decisions, and the degree to which concerns about foreign oil influence the respondents' opinions about the need to reduce transportation energy consumption. These results were unexpected, since the question was phrased generically about adoption of new technology of *any* kind, not just related to the new PHEV technology, and the significance of these findings requires further investigation.

In our recently published agent-based model for studying potential PHEV market penetration [6], we implement a social threshold effect designed to model different levels of consumer comfort in adopting new technologies, motivated by the classic works of Granovetter (1978) and Watts (2002) [11][32]. Each agent assesses the proportion of PHEVs owned by other agents in its local spatial neighborhood and/or within its social network. If this proportion does not exceed the agent's personal threshold, then, the agent will not even consider purchasing a PHEV, regardless of any other financial or environmental costs and benefits. Individual agent thresholds are heterogeneous in our simulated populations, reflecting the varying levels of discomfort among people regarding adoption of the new PHEV technology [2]. However, we found that there was little published information to help us estimate reasonable distributions for this threshold, to use in initializing the model. The 7-class survey question (C-7) was thus designed to obtain that information. It is interesting that the 15 features found to be most predictive of these classes included basic demographics (age and sex), social and advertising influences on purchasing habits, and a desire to reduce transportation energy consumption. We plan to use this knowledge to better inform realistic distributions and correlation between these agent attributes, all of which are already part of our agent based model of vehicle consumers [6].

Although this preliminary analysis of the PHEV survey data is interesting and informative, much work remains to be done. For example, we plan to see how sensitive the feature selection is to a non-parametric counter-propagation neural network (CPNN) classifier [12]. Although it is computationally much more costly, the CPNN relaxes parametric assumptions required by the DA classifier and has proven useful in other similar applications [4][13]. We also plan to apply the method to a much wider range of potential outcome questions that are present in our PHEV survey data set. Features selected by the described method will be studied using other approaches. For example, we plan to apply multivariate regression and symbolic regression (using genetic programming [18]) to try to estimate the functional relationships between the features and predicted outcomes. These results will be used, in part, to further inform agent choice rules in our agent-based model.

5. ACKNOWLEDGMENTS

This work was funded in part by the United States Department of Transportation through the University of Vermont Transportation Research Center, a workforce development sub-award from Sandia National Laboratories supported by the U.S. Dept. of Energy through Inter-Entity Work Order M610000767, and the Vermont Advanced Computing Center which is supported by the NASA award NNX 06AC88G. We thank Kiran Lakkaraju, Christine E. Warrender, Brad Lanute, and Diann Gaalema for their help in developing and implementing the PHEV survey.

6. REFERENCES

[1] AMT: <https://www.mturk.com/mturk/welcome>

[2] Curtin, R., Shargo, Y., and Mikkelsen, J., Plug-in Hybrid Electric Vehicles. Reuters/University of Michigan, Surveys of Consumers.

[3] Deelman, E., and Gil, Y. 2006. Final Report of the NSF Workflows in Distributed Environments: Experiences and Challenges. *National Science Foundation*, (May 2006).

[4] DeHaas, D., Craig, J., Ricket, C., Haake, P., Stor, K., Eppstein, M.J. "Feature Selection and Classification in Noisy Epistatic Problems using a Hybrid Evolutionary Approach", extended abstract, *Genetic and Evolutionary Computation Conference* (GECCO), 2007.

[5] Duvall, M.C.C., Knipping, M., and Alexander, E. Environmental assessment of plug-in hybrid electric vehicles. *Nationwide greenhouse gas emissions* 1, 1015325 (2007).

[6] Eppstein, M.J., Grover, D.K., Marshall, J.S., and Rizzo, D.M., 2011. An agent-based model to study market penetration of plug-in hybrid electric vehicles. *Energy Policy* 39 (2011), 3789-3802.

[7] Eppstein, M.J. and Haake, P., "Very Large Scale ReliefF for Genome-Wide Association Analysis", *IEEE Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, (2008), 112-119.

[8] Eppstein, M.J., Payne, J.L., White, B.C., and Moore, J.H. 2007. Genomic mining for complex disease traits with 'Random Chemistry', *Genetic Programming and Evolvable Machines (special issue on Medical Applications),* 8, (2007),395-411. (DOI 10.1007/s10710-007-9039-5)

[9] Ferri, F., Pudil, P., Hatef, M., and Kittler, J. 1994. Comparative study of techniques for large-scale feature selection. *In: Pattern Recognition in Practice IV, Multiple Paradigms, Comparative Studies and Hybrid Systems* 16, (1994), 403-413.

[10] Gauderman, W.J. 2002. Sample size requirements for association studies of gene-gene interaction. *Am. J. Epidem*., 155, 5 (2002), 478-484.

[11] Granovetter, M. 1987. Threshold models of collective behavior. *American Journal of Sociology* 83, 6 (May 1987), 1420-1443.

[12] Hecht-Nielsen, R. 1987. Counterpropagation Networks. *Applied Optics* 26, 23 (1987) 4979-4983.

[13] Manukyan, N., Eppstein, M.J., Horbar, J.D., Leahy, K.A., Kenny, M.J., Mukherjee, S., and Rizzo, D.M. "Evolutionary Mining for Multivariate Associations in Large Time-Varying Data sets: A Healthcare Network Application", extended abstract accepted, *Genetic and Evolutionary Computation Conference* (GECCO), 2012.

[14] McKinney, B.A., Reif, D.M., White, B.C., Crowe, Jr., J.E., and Moore, J.H. 2007. Evaporative cooling feature selection for genotypic data involving interactions, *Bioinfomatics* 23, 16 (June 2007), 2113-2120.

[15] Moore, J.H, and White, B.C. 2007. Tuning ReliefF for genome-wide genetic analysis. *Lecture Notes in Computer Science* 4447, (2007), 166-175.

[16] Moore, J.H., and White, B.C. Genome-wide genetic analysis using genetic programming: The critical need for expert knowledge. In *Genetic Programming Theory and Practice IV*, (2006), 11-28.

[17] Khatib, F., Cooper, S., Tyka, M.D., Xu, K., Makedon, I., Popović, Z., Baker, D., and Players, F. 2011. Algorthm discover by protein folding game players. *Proceedings of the National Academy of Sciences of the United States of America* 108, 47 (Nov. 2011), 18949-18953.

[18] Koza, John R. 1994. Genetic Programming II: Automatic Discovery of Reusable Programs. *The MIT Press*, Cambridge, MA.

[19] Kononenka, I., Šimec, E., and Robnik-Šikonja, M. 1997. Overcoming the Myopia of Inductive Learning Algorithms with ReliefF. *Applied Intelligence* 7, 1 (Jan. 1997), 39-55.

[20] Kudo, M., and Sklansky, J. Comparison of Algorithms that Select Features for Pattern Classifiers. *Pattern Recognition,* 33, (2000), 25-41.

[21] Lang, D., and Hogg, D.W. 2011. Searching for comets on the World Wide Web: The orbit of 17PHolmes from the behavior of photographers. Cornell University. arXiv:1103.6038v1 [astro-ph.IM]

[22] Liu, H., Li, J., and Wong, L. 2002. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Informatics* 13, (2002), 51-60.

[23] Oh, I.-S., Lee, J.-S., and Moon, B.-R.Hybrid Genetic Algorithms for Feature Selection, *IEEE Trans. Patt. Anal. And Mach. Intel*. 26, (2004), 1424-1437.

[24] Pernkopf, F., and O'Leary, P. Feature Selection for Classification Using Genetic Algorithms with a Novel Encoding, *Proceedings of the 9th International Conference on Computer Analysis of Images and Patterns*, (2001), 161-168.

[25] Pontin, J. 2007. Artifical Inteligence, With Help From Humans. *The New York Times*. (Mar. 2007).

[26] Raymer, M.L., Punch, W.F., Goodman, E.D., Kuhn, L.A., and Jain, A.K. Dimensionality reduction using genetic algorithms. *IEEE Trans. Evol. Comp.* 4, (2000), 164—171.

[27] Reshef, D.N., Reshef, Y.A., Finucane, H.K., Grossman, S.R., McVean, G., Trunbaugh, P.J., Lander, E.S., Mitzenmacher, M., and Sabeti, P.C. 2011. Detecting Novel Associations in Large Data Sets. *Science* 334, (Dec. 2011), 1518-1524.

[28] Robnik-Sikonja, M., and Kononenko, I. Theoretical and empirical analysis of ReliefF and RReliefF. *Mach. Learning*, 53, (2003) 23-69.

[29] Robnik-Šikonja, M., and Kononenko, I. 1997. An Adaptation of ReliefF for Attribute Estimation in Regression. *In Proceedings of 14th International Conference on Machine Learning* , (Nashville, Tennessee, 1997).

[30] Sovacool, B.K., and Hirsh, R.F. 2009. Beyond batteries: an examination of the benefits and barriers to plug-in hybrid electric vehicles (PHEVs) and a vehicle-to-grid (V2G) transition. *Energy Policy* 37, 3 (Mar. 2009), 1095-1103.

[31] Szalay, A., and Gray, J. 2006. 2020 Computing: Science in an exponential world. *Nature* 440 , 7083 (Mar. 2006), 413-414. doi:10.1038/440413a.

[32] Watts, D. 2002. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences* 99, 9 (Apr. 2002), 5766-5771.

[33] Yang, J. and Honavar, V. Feature subset selection using a genetic algorithm. *IEEE Intelligent Systems* 13, (1998), 44-49.

[34] Zypryme Research and Consulting. 2010. The Electric Vehicle Study. <http://www.zpryme.com/SmartGridInsights/The_Electric_Vehicle_Study_Zpryme_Smart_Grid_Insights_Airbiquity_Sponsor_December_2010.pdf>