Knowledge Mining with Genetic Programming Methods for Variable Selection in Flavor Design

Katya Vladislavleva
University of Antwerp
Belgium
katya@vanillamodeling.com

Kalyan Veeramachaneni
Massachusetts Institute of Technology
Cambridge, MA
kalyan@csail.mit.edu

Matt Burland
Givaudan Flavors Corp.
Cincinnati, OH
matt.burland@givaudan.com

Jason Parcon
Givaudan Flavors Corp.
Cincinnati, OH
jason.parcon@givaudan.com

Una-May O'Reilly
Massachusetts Institute of Technology
Cambridge, MA
unamay@csail.mit.edu

ABSTRACT

This paper presents a novel approach for knowledge mining from a sparse and repeated measures dataset. Genetic programming based symbolic regression is employed to generate multiple models that provide alternate explanations of the data. This set of models, called an *ensemble*, is generated for each of the repeated measures separately. These multiple ensembles are then utilized to generate information about, (a) which variables are important in each ensemble, (b) cluster the ensembles into different groups that have similar variables that drive their response variable, and (c) measure sensitivity of response with respect to the important variables. We apply our methodology to a sensory science dataset. The data contains hedonic evaluations (liking scores), assigned by a diverse set of human testers, for a small set of flavors composed from seven ingredients. Our approach: (1) identifies the important ingredients that drive the liking score of a panelist and (2) segments the panelists into groups that are driven by the same ingredient, and (3) enables flavor scientists to perform the sensitivity analysis of liking scores relative to changes in the levels of important ingredients.

Categories and Subject Descriptors

I.1.2 [**Computing Methodologies**]: Symbolic and Algebraic Manipulation—*algorithms*

General Terms

Algorithms, Design, Experimentation

GECCO'10, July 7–11, 2010, Portland, Oregon, USA.
Copyright 2010 ACM 978-1-4503-0072-8/10/07 ...$10.00.

Keywords

variable selection, ensemble modeling, sensory science, genetic programming, symbolic regression

1. INTRODUCTION

Variable selection is a process of identifying influential variables (attributes) that are discriminative and are necessary to describe a real or a simulated system and its performance characteristics. Understanding the relative importance of variables makes a design problem tractable by reducing the dimensionality of the original problem. It shortens the design time by facilitating insights and improves the generalization power of models. These factors usually drive the product costs down.

In this paper, we consider variable selection in datasets that are sparse and contain repeated measures because they present unique challenges for variable selection. Consider a set of explanatory variables $\overline{x} = \{x_1 \dots x_n\}$, a response variable y and an unknown function $\mathcal{F}$ that relates $\overline{x}$ to y. Sparsity implies that very few data samples that explain $\mathcal{F}$ are available relative to the number of the explanatory variables, i.e. n. The dataset contains repeated measures, if the same samples are passed to different measuring functions (or responses) that are denoted as $\mathcal{F}_s$ for $s = 1 \dots l$. If there is large variance for 1 sample's responses, this implies that there is no one single model for the entire dataset and one has to build a model for each $\mathcal{F}_s(\overline{x})$ measuring function.

In this paper, we adopt an ensemble based symbolic regression approach to provide multiple unbiased explanations of the input-output relationships in the data. There are several known advantages of symbolic regression over parametric regression. For example, symbolic regression can handle dependent and correlated variables and automatically discover various appropriate and diverse models. However, the *multiple model generating capability* of genetic programming (GP) is the strongest argument for using symbolic regression on sparse data sets. To our surprise it is often ignored (or taken for granted) and a GP with single-objective fitness driven selection, and a single best-of-the run final solution (see [?, ?, ?, ?] among others) is used.

In this paper, we exploit the *multiple model generating* capability of evolution. We employ a robust approach using ParetoGP which is symbolic regression via GP implemented with archiving (elite-based selection with elite preservation), two-objective selection and other defining features [?]. ParetoGP yields the aggregated final archive of multiple independent runs. We call this a model set, $\mathcal{M}$ and generate these model sets for each subset of data samples corresponding to a measuring function $\mathcal{F}_s(\overline{x})$. When repeated for all the measuring functions, symbolic regression creates rich sets of model ensembles.

We exploit this model set to propose two methods for calculating *variable importance*. Using the importance information, we further mine the data to conduct *sensitivity analysis* and identify similarity among measuring functions (or model sets).

We present our results empirically on a dataset from the area of sensory science provided by Givaudan Flavors Corporation, an international flavor and fragrance design company. In its data, flavors are mixtures of seven edible ingredients that enhance the perception of food products by impacting taste and smell pathways. The data, derived via design of experiments, contains 40 different flavors evaluated by 69 human panelists. Givaudan's urge to continually improve has driven its flavor scientists to seek new methods that will provide alternate answers regarding relevant ingredients within a flavor that drive liking. This has been our primary motivation for this work.

The rest of the paper is organized as follows. Section 2 presents the salient features of our sensory evaluation dataset and the challenges in modeling sparse and repeated measures data. Section 3 provides an overview of our approach. Section 4 presents the ParetoGP technique used to generate the ensemble of models. Section 5 presents our knowledge mining approach to derive variable importance from the ensembles. Section 6 and 7 present sensitivity analysis and clustering approach based on the variable importance derived in section 5. Section 8 presents the results on the empirical study we performed in the area of sensory science. Section 9 concludes our study.

2. DATA FEATURES, CHALLENGES

Our challenging dataset has been presented to us by Givaudan Flavors Corporation. Each flavor is a mixture of seven ingredients by concentration levels (unnormalized and unscaled), called *keys* that are denoted as $k_1, \ldots, k_7$. The maximum concentration levels for $k_1, \ldots, k_7$ are $(130, 80, 50, 20, 20, 20, 200)$ respectively. A total of 40 flavors are experimentally designed by combining keys at three levels each, corresponding to their *zero*, *mean*, and the *maximum* concentration. Care has been taken such that no two flavors have more than three similar keys. Notice that the number of combinations are very low compared to the total number of combinations that are possible even when only 3 levels are used for each key, which is $3^7 = 2187$. In reality, these levels can vary in fine-grained discrete intervals in between 0 and the maximum range.

An important feature of this data besides sparsity is multiple responses per sample. Each of the 40 flavors is rated by 69 panelists from panel $\mathcal{P} = \{P^{(1)}, \ldots, P^{(69)}\}$. They create 40×69 ratings, which we will call *liking scores*. The panelists rate the flavors on an integer hedonic scale from 1 (*'dislike extremely'*) to 9 (*'like extremely'*) via a neutral 5 (*'neither like nor dislike'*). This data presents us the variety of challenges envisioned for sparse and repeated datasets. First due to the fact that same samples are presented to different panelists, we have the different response values for the same inputs. In Figure 1 we demonstrate the variation in the raw liking scores per flavor. Note, that the variation in the liking is wide for all flavors and covers the entire range of liking score, i.e., 1-9. In other words, the differences in the liking preferences of the panel are too high to ignore, and averaging them per flavor will heavily reduce the information content in the data.

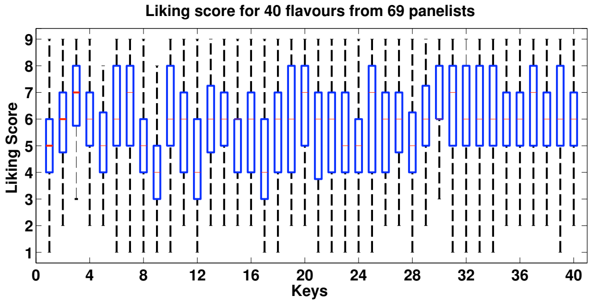


Figure 1: Variation in the liking scores assigned by all panelists to a given flavor over all 40 flavors. Box boundaries correspond to the interquartile range of 69 liking scores per flavor.

The goal of the paper, in the sensory science context, is to select variables that drive liking scores of panelists, and to understand the direction of the driving, i.e. the analysis of the changes in the liking scores caused by changes in the concentration of the keys (sensitivity analysis).

The conventional approach in flavor science is to explain the dependence between the key levels and the liking scores of the entire panel by an empirical model. This model is constructed to approximate the average assigned liking score per flavor, and is usually a low-order polynomial obtained by linear regression. Variable importance information is obtained from the analysis of model parameters. Variable sensitivity is studied based on predictions of the model.

We argue that one needs to build a model per panelist, extract as much information about the panelist from this model and then combine this information when necessary. However, due to the sparsity of data, it is hard to build models that are reliable (have good predicting capabilities on unobserved points) and robust (less error prone on observed points), i.e. models of high accuracy and no overfitting. This is our challenge in this paper and we approach this problem systematically using ensemble based symbolic regression.

3. OUR APPROACH

Our approach to solve the problem is presented in the Figure 2. In the figure we show three distinct steps. (a) First ParetoGP generates multiple models for a single panelist and these multiple models, that form an archive, are used to derive variable importance vectors. (b) Correlation analysis is performed on variable importance vectors for multiple measuring functions and the functions are clus-

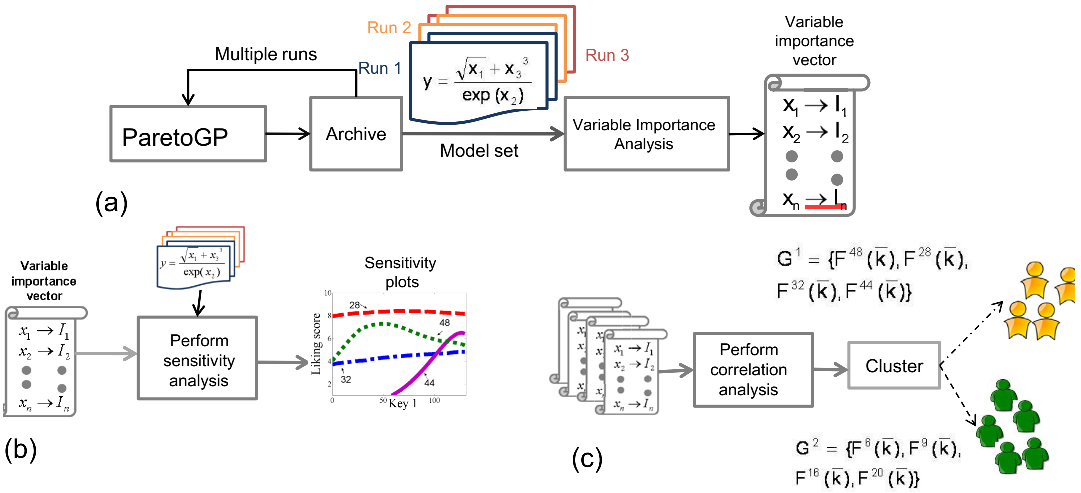

Run 2
Run 3
Multiple runs
Run 1
$y = \frac{\sqrt{x_1} + x_3^3}{\exp(x_2)}$
ParetoGP
Archive
Model set
Variable Importance Analysis
Variable importance vector
$x_1 \to I_1$
$x_2 \to I_2$
$x_n \to I_n$
(a)
Variable importance vector
Perform sensitivity analysis
Sensitivity plots
Liking score
Key 1
28
48
32
44
(b)
Perform correlation analysis
Cluster
$G^1 = \{F^{48}(\bar{k}), F^{28}(\bar{k}), F^{32}(\bar{k}), F^{44}(\bar{k})\}$
$G^2 = \{F^{6}(\bar{k}), F^{9}(\bar{k}), F^{16}(\bar{k}), F^{20}(\bar{k})\}$
(c)

Figure 2: Overview of our knowledge mining approach using ensemble based symbolic regression. (a) An archive of models is generated via multiple runs of ParetoGP. The archive is then analyzed for variable importances. (b) The variable importance information along with the archive of models is used for performing sensitivity analysis. (c) The variable importance for multiple measuring functions is used to perform correlation analysis and cluster them.

tered into groups that have similar influential variables.(c) Finally, the variable importance vectors are then used to analyze the sensitivity of the response variable with respect to most influential variables.

4. PARETOGP SYMBOLIC REGRESSION

Learning from sensory data is a perfect example of an application where **the** model does not exist. To gain prediction robustness on this sparse data, we use ensemble-based Pareto GP. The ensemble, also known as model set, $\mathcal{M}$ contains diverse but high-quality models, which are constrained to approximate *all* training data samples well (high-quality) and are also constrained to diverge in predictions on unobserved data samples (diverse). When a sufficient number of models are generated, all of them can be used to determine both the prediction (by unification of their predictions) and the disagreement at an arbitrary point of the original variable space. ParetoGP used here is a tree-based GP. An experiment consists of multiple independent runs called replicates. In a single run the algorithm performs the following operations:

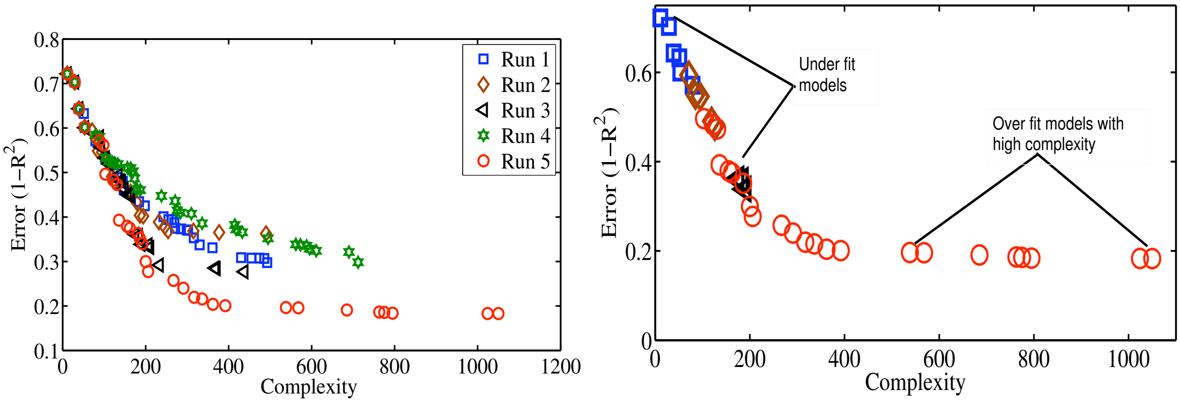
1. **Initialize models**: The following primitives are used for tree-based individuals of the maximal arity of four: $\{+, -, *, /, inverse,\ power(x, const), square,\ ln, exp\}$. The list of variables, which in our case are seven keys and real constants from $[-5, 5]$ are used as terminals. We rescaled our inputs variables to the range $\{0 \ldots 2\}$.

2. **Perform multi objective evaluation**: The models are evaluated under two objectives. The first one, model error, is defined as $1 - R^2$, where R is a correlation coefficient between scaled predicted and scaled observed response. The second objective, model complexity, is defined as the sum of all subtrees of the tree-based genome of the GP individual. The goal is to minimize both error and complexity.

3. **Archive the best models and update**: An archive of individuals is created separately from the population and an elite-preservation strategy is employed. At generation $t + 1$, the archive, which is the elite set of best individuals discovered so far, gets updated. Its size is limited to $ArchiveSize$ by selecting the least-dominated individuals from the union of $Archive(t)$ and $Population(t + 1)$ in the objective space of model error and model complexity.

4. **Vary the models**: During each iteration, a new population is created using archive mutations and crossovers. In crossovers, parents are either both sampled from the archive, or one parent sampled from the archive, and one from the population (in both cases using Pareto tournament selection). This archive-based selection preserves genotypic diversity of individuals. The new individual is generated by using a sub-tree crossover with rate 0.9, and sub-tree mutation with rate 0.1. Every 10 generations, the population gets re-initialized to provide diversity and avoid inbreeding.

Other parameters for the ParetoGP are given in Table 1. A run is executed for a time interval and using all the observations because using complexity as a second objective and collecting multiple solutions in accuracy-complexity tradeoff space eliminates any requirement for an arbitrary maximum generation or cross-validation that would make the training data even more sparse. Some evolved models will "over-fit" but they can rationally be pruned post-hoc when the model set is finalized to be used for prediction. The time interval we chose is equivalent to 280 generations. Interval arithmetic is used to prune individuals with numerical inconsistencies. Linear scaling is used to enhance the effectiveness of evolution. At the end of an experiment, the models in the archives of each run are aggregated into an archive. The non-dominated solutions in this archive form the super Pareto front. This is illustrated in Figure 3.



(a) (b)

Figure 3: An exemplar ParetoGP simulation on the sparse data (a) Results from multiple runs of ParetoGP. Pareto fronts from each run show the trade-offs between model error $(1 - R^2)$ and model complexity. (b) A super Pareto front is generated by aggregating the Pareto fronts from multiple runs. The super Pareto front has 37 models.

Table 1: ParetoGP experimental parameters

Parameters	Comments
# replicates	5 unless stated otherwise
# generations	310
population size	1000
archive size	100
fitness	$1 - R^2$
complexity	expressional complexity
crossover rate	0.9
subtree mutation rate	0.1
population tournament	5
archive tournament	5

5. VARIABLE SELECTION

Most non-evolutionary modeling methods are vulnerable to producing solutions that contain insignificant inputs. This results in a fast deterioration of prediction performance of final solutions as more irrelevant variables in the data are considered in the model.

A conventional approach to identify the true dimensionality of the problem is to perform a principal component analysis or a factor analysis. The former reduces the problem dimensionality to a smaller number of meta-variables which are linear combinations of the original variables. The latter extracts the latent dimensionality of the problem by determining the number of factors that contain the same information as the matrix of mutual correlations of data variables. The potential problem of these approaches (in analysis of non-linear systems) is that they only take into account mutual correlations between variables, and hence miss the relevance of non-linear combinations of inputs to the response. As well, they do not select important variables from the original set, but create new variables in the new reduced set. They forfeit multicollinearity (which is most often present in real measurements). Finally, they are sensitive to outliers.

One of the unique capabilities of genetic programming is its built-in power to select significant variables and gradually omit the variables that are not relevant while evolving models. Variable selection based on genetic programming has been exploited in various applications where the significant inputs are generally unknown (for examples see [**?**, **?**, **?**, **?**, **?**, **?**, **?**]).

In this paper we consider two methods of variable presence analysis for the multiple models generated using ParetoGP. We also consider how relative variable importance can be calculated. Note that these methods could also be used on the population of solutions at the end of a standard GP run.

The methods generate a variable importance vector, V:

Definition 1. *A variable importance vector is a vector* $V = \{\mathcal{I}_1, \mathcal{I}_2, \ldots \mathcal{I}_n\}$ *of the importance of all explanatory variables* $\{x_1, x_2, \ldots x_n\}$*, in percents, arranged in the same order as the explanatory variables. The importances are relative if* $\sum_{k=1}^{n} I_k = 100$.

5.1 Presence-weighted variable importance

This method analyzes variable presence rates in a subset of models $\mathcal{M}$ from the ensemble archive and considers variables relevant if they have a high presence rate. The *aggregated importance* of the variable $x_i, i = 1 \ldots, d$ computed on the basis of best models $\tilde{\mathcal{M}} = \{M_j\}, j = 1, \ldots, m$ is

$$\mathcal{I}_i^{(PW)}(x_i, \tilde{\mathcal{M}}) = \sum_{j=1}^{m} \frac{\delta(x_i, M_j)}{m}, \qquad (1)$$

where $\delta(x_i, M_j)$ is zero if x_i is not present in model M_j, and one otherwise. This aggregated variable importance provides a robust estimation of relevance *if* $\tilde{\mathcal{M}}$ is hand selected for high-quality (i.e., fitness and complexity) from an experiment-archive derived from many independent runs.

The second variable importance metric resolves the problem of hand selecting $\mathcal{M}$ by eliminating the need for it.

5.2 Fitness-weighted variable importance

Fitness-weighted variable importance is calculated using all models (in the archive or in both archive and population) (see [?]). It first uniformly distributes the fitness of each model over all variables present in it, thus assigning a variable a score per each model it is present in. Then, it sums up the scores over all models, $\mathcal{M} = \{M_j, j = 1, \ldots, m\}$.

$$\mathcal{I}_i^{(FW)}(x_i, \mathcal{M}) = \sum_{j=1}^{m} \frac{fitness(M_j)}{\sum_{i=1}^{d} \delta(k_i, M_j)} \delta(k_i, M_j), \qquad (2)$$

Since the fitness of a model is uniformly distributed over all its variables, this creates an explicit bias towards variables occurring in lower dimensional solutions. Thus, the overall aggregated scores of irrelevant variables (only present in over-fitting solutions) is much smaller than the overall score of relevant variables.

We use normalized fitness-weighted variable importances defined as

$$\mathcal{I}_i^{(NFW)}(x_i, \mathcal{M}) = \frac{\mathcal{I}_i^{(FW)}(x_i, \mathcal{M})}{\sum_i \mathcal{I}_i^{(FW)}(k_i, \mathcal{M})} \cdot 100\%. \qquad (3)$$

6. VARIABLE SENSITIVITY ANALYSIS

The variable importance vector enables a means of sensitivity analysis which supports efficient exploration of the design space to observe the response variable under selected conditions of the explanatory variables. Consider an explanatory variable set consisting of n variables where each variable can be explored in r discrete step sizes. The total number of design exploration samples is n^r which is generally intractable.

To alleviate this, the variable importance vector can be used. The distribution of the percentages in the variable importance vector informs the choice of downsizing the sampling. The effects of q influential variables, where $q << n$ can be explored while the *non-influential* $n - q$ variables are clamped to a finite set of combinations, $c << (n - q)^r$. The q *influential* variables can be exhaustively sampled over q^g. For each sample, the predicted response of the predictive model ensemble is calculated using a median-average method [?]. The predictive model ensemble is derived by boxing the ensemble-archive. See [?] for more details. These predictions for the q most relevant variables can subsequently be visualized to observe the measuring function's sensitivity under the clamped conditions. The values of q, c and g are selected based on the needs of the modeling application. It is sensible to also reduce g as the importance ranking of an influential variable decreases. This supports coarser grained sampling in dimensions where variable importance is less and higher grained sampling where it is higher.

7. MODEL CLUSTERING

For datasets consisting of repeated measures, i.e. when the same input variables are passed through different measuring functions (e.g., different people), we form a model ensemble for each measuring function and then extract variable importance vectors for each of them. We are next interested in how similar one measuring function is to another. This is equivalent to identifying people driven by the same key and, in sensory evaluation, is called segmentation. Segmentation enables design strategies for multiple, similar people and is highly useful.

We use the variable importance vector as the basis of similarity. Consider a model set denoted by $\mathcal{M}_s$ for a measuring function $\mathcal{F}_s$ and the corresponding variable importance vector as V_s. We compute pairwise correlation between each pair of vectors and construct a correlation matrix defined by C. The entry C_{ij} in this matrix is the Pearson correlation coefficient between variable importance vectors of model set, i and j given by

$$C_{ij} = \frac{\sum_{k=1}^{n} (V_i^k - \bar{V_i})(\bar{V_j}^k - \bar{V_j})}{(n-1) s_{V_i} s_{V_j}}, \qquad (4)$$

where s_{V_i}, s_{V_j} are sample standard deviations for V_i and V_j. We then apply a threshold θ to the matrix entries to determine clusters of model sets. A cluster is identified when all the pair-wise correlations exceed θ.

8. KNOWLEDGE MINING WITH GP

We now apply our methods to Givaudan's data per its description in Section 2. We proceed by running an experiment over each panelist's data and constructing variable importance vectors for each. We next correlate these variable importance vectors to cluster the panelists into groups. Finally, we conduct sensitivity analysis on group's model sets which identifies panelist segments. Panelists in a segment are generally influenced in the same way by the variables (keys) and these variables generate hedonic responses that move in the same (direct or inverse) direction.

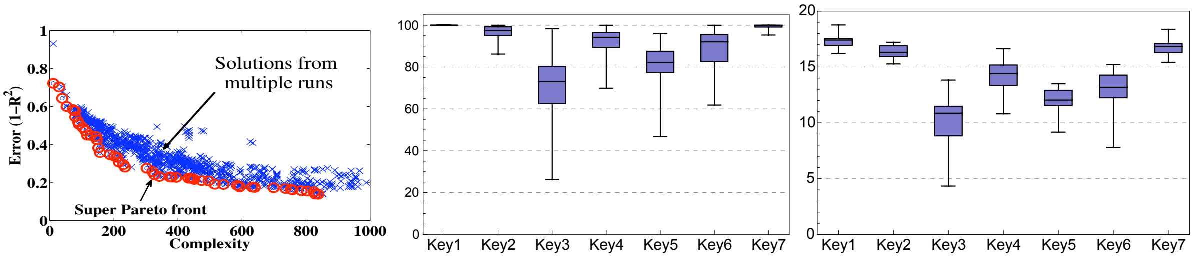
8.1 Modeling using ParetoGP

To confirm our intuition that modeling the panel as an aggregate will disrespect important inter-panelist differences and will be inaccurate, we preliminarily run a ParetoGP experiment to generate models for the entire panel altogether. With 69 panelists in the panel and 40 flavors for each panelist, there are 40 × 69 data points. We denote this model set (or ensemble-archive) as $\mathcal{M}_F$. The best GP models we can evolve with this data are poor. Five independent runs of 1000 generations each only minimize errors to $1 - R^2 = 0.91$ (with optimal error value 0.0 and the worst value 1.0). Trying other parameter settings and increasing the length of the GP runs does not improve the model set quality.

To respect the differences among different panel members, we separately model each panelist with 40 data points. The results of these 69 experiments on individual panelists are significantly better than on the entire panel both with respect to the area under Pareto fronts in the objective space and the lowest error. We denote the model set for panelist s as $\mathcal{M}_s$. We evaluate the normalized area under the Pareto curve metric for all the 69 panelists and analyze these 69 values to see if our results are consistent. The mean of the normalized area under curve is 20.257% (with the ideal best value of 0%) with a standard deviation of 6.81%. The distribution has a positive skew (0.35) indicating that most of the numbers are on the left of the mean. The area under the curve for the ensemble $\mathcal{M}_F$ generated for all the panelists together is 91.5% (with the worst theoretical value of 100% if no models are generated at all).

8.2 Variable Selection

Next we compute the normalized fitness-based variable importances for both $\mathcal{M}_F$ and each of the 69 panelist's $\mathcal{M}_s$.



(a) Pareto GP Solutions

(b) Presence-weighted Importance

(c) Normalized Fitness-weighted Importance

Figure 4: Models and variable importances for panelist $P^{(3)}$ (a) Archive of models generated by one trial of ParetoGP runs. (b) Variable importances over 50 trials based on the presence frequency in the archived models with error not exceeding 0.4. (c) Normalized fitness weighted importances of all archive solutions (box plots illustrate variation over 50 independent experiments).

As an example, we present a thorough analysis of variable selection for panelist $\mathcal{P}^{(3)}$.

For $\mathcal{M}_F$ ensemble constructed using full data, the variable importance vector is

$$V(\bar{k}) = (48, 30, 0, 1, 5, 0, 16)^T \qquad (5)$$

Its values imply that the most relevant ingredients explaining the overall liking scores of the panel are k_1, k_2 and k_7, while ingredients $k_3 - k_6$ are of minor importance.

Figure 4 illustrates modeling results for $\mathcal{P}^{(3)}$. Plot 4(a) shows the ensemble-archive with the ensemble-Pareto front colored red. For evaluation purposes we repeat 50 variable selection experiments and examine the variance in their variable importances for both methods (presence and fitness weighted) in plots 4(b) and 4(c). Plot 4(b) shows $\mathcal{I}^{(PW)}$ values in percentages. Plot 4(c) shows normalized $\mathcal{I}^{(FW)}$ values. It appears from plot 4(c) that using fitness weighting provides better discrimination between variables than presence weighting. A two-sided Wilcoxon-Mann-Whitney rank tests confirms that the differences among the fitness-weighted importances for all keys are statistically significant with p-values ranging between 10^{-16} and 10^{-4} for 95% confidence. The results also indicated that $I^{(FW)}$ metric from only five independent ParetoGP runs generates reliable and reproducible rankings of the input variables.

Plots 4(b) and (c) also indicate that the liking preferences of panelist $P^{(3)}$ are driven by all keys in the flavors because all interquartile ranges of importances are above 50% for presence-weighted and above 5% in the normalized fitness-weighted importance. The differences among the relative fitness-weighted importances are small, but their ranking is clear. The ranking is (in decreasing order): k_1, k_7, k_2, k_4, k_6, k_5, k_3. These variable importances, for one panelist, are markedly different from those derived for the panel overall.

Computing with each panelist $P^{(s)}$'s model set, M_s, we aggregate the 69 vectors of normalized fitness-weighted importances into a table of 69 rows, with their elements reflecting the relative importance of seven keys for predicting the liking scores of each panelist. Figure 5 shows large variation between the individual variable importances. The median provides a very 'coarse grained' look at the panel with less loss of information about panelist differences than when all panelists are modeled altogether. It is:

$$V_{(1,\dots 69,Median)}(\bar{k}) = (18, 15, 13, 17, 12, 19, 15)^T \qquad (6)$$

In contrast to the variable importance vector of Equation 5, it reveals that all variables are relevant for predicting the response. This contradiction highlights again the inappropriateness of compiling the data of 69 panelists all together.

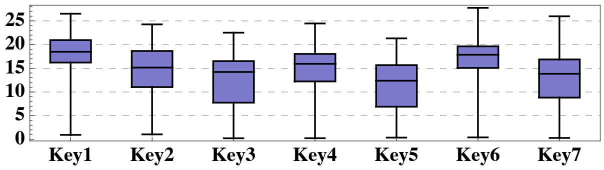


Figure 5: Normalized fitness-weighted variable importances for all panelists (each box-plot consists of 69 importance values).

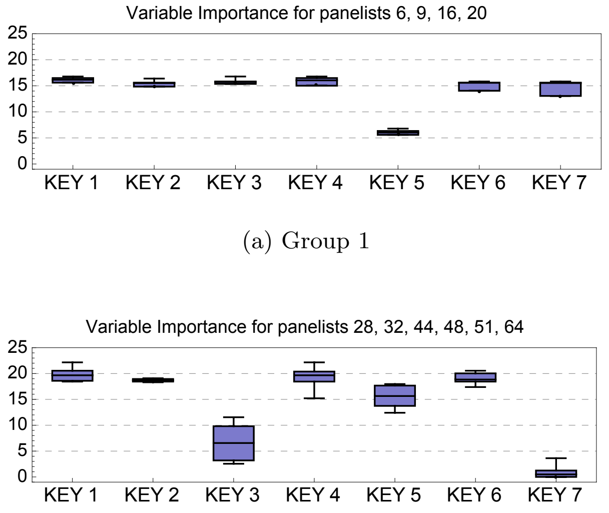
8.3 Segmentation of panelists

The combined fitness-weighted importance profiles depicted in Figure 5 clearly indicate that importance vectors are different, and therefore the models of individual panelist's likings contain different sets of inputs. We now work to cluster similar panelists (or model sets) into groups on the basis of variable importance. This will answer which panelists' liking is driven by the same keys. After this, we will segment a group using variable sensitivity analysis to identify panelists driven in the same direction by the same key.

8.3.1 Clustering to Form Groups

To find which panelists' liking is driven by the same keys, we compute pair-wise correlations between importance vectors V_i and V_j. By selecting pairs with correlations exceeding a threshold $\theta = 0.90$, we identify groups of panelists with similar variable importance vectors[1]. There are five groups of panelists with high pair-wise correlations of importances between all members in a group. Fitness-weighted variable

[1] The choice of the θ threshold is highly influential in the subsequent conclusions. We discuss this in Section 8.4.



(a) Group 1

(b) Group 2

Figure 6: Examples of clustering panelists into groups with similar variable importance information. The box plots show the variance in the variable importance within the group for each key. Compare with Figure 5.

importances of two groups are illustrated in Figure 6. Notice the high consistency in variable importance vectors per group, and the clear differences among the two groups. All the variables except k_5 are required to predict the individual liking scores of Group 1 consisting of panelists $6, 9, 16$ and 20. All variables except k_3 are required for Group 2 consisting of panelists $28, 32, 44, 48, 51$ and 64.

8.3.2 *Group Sensitivity Analysis for Segmentation*

We now perform sensitivity analysis to understand whether a key has direct or inverse relation with the liking score. By doing this we can identify panelists, for whom both the ith key is the most important variable, but one might hate it as its concentration increases and the other might like it. We use Group 1 and Group 2 as our exemplars. In Figure 7 we plot the individual ensemble predictions (median of predictions of ensemble members) of all panelists in the group for varying volume levels of k_1, while all other levels are clamped to their maximum. The step size in varying k_1 is domain related. The spread in ensemble predictions in Figure 7 justifies once again the differences in the panelists. In Group 1 one segment of panelists $\{6, 16\}$ have monotonically increasing predicted liking scores for increasing levels of k_1 , another segment, $\{9, 20\}$ shows decreasing liking scores as k_1 is varied in the interval $[0, 130]$. In Group 2 the liking score increases for all the panelists as k_1 is increased.

8.4 Discussion

There are many choices in this knowledge mining process: e.g. what data to aggregate and thresholds such as θ. They should, in general, be made by an expert on the system being modeled. A choice could depend on exogenous goals like market targeting. For example, Givaudan could decide to

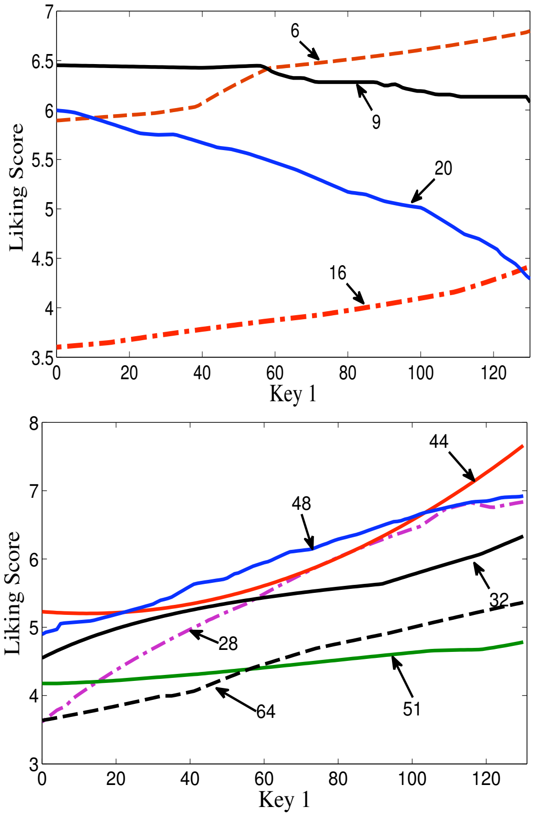


Figure 7: Comparison of sensitivity of liking of panelists within Group 1 (top) and Group 2 (bottom) based upon varying the levels of key k_1 while k_3, $k_5 - k_7$ clamped to their maximum, and k_2 and k_4 clamped to zero.

use average ratings of the panel. This would allow them to design flavors that maximize the liking scores according to this information. In this example, strong inter-panelist differences contra-indicate this approach. We observe that all variables are important for modeling the liking score of the entire panel and that there exist fundamental differences in the driving variables among individual panelists. This implies that an approach of designing flavors for the entire panel is likely to generate designs that will be suited to a broad population with a lesser degree of liking. Alternatively, if it is affordable to segment the panel into multiple segments and design flavors that satisfy these smaller segments, each resulting design would have higher likability inside a segment but less suitability across the broad population.

The advantage of our approach is that all analysis decisions are postponed until the moment when the decision trade-offs become clear to the domain expert. To understand the trade-offs, the domain experts have access to efficient sensitivity analysis methods which will allow them to finally identify the directions in which the liking scores of panelists are steered by important keys.

9. SUMMARY AND FUTURE WORK

In summary, we have presented an approach to variable selection and sensitivity analysis using genetic programming model ensembles. While the variable selection via genetic programming is by far not new, we believe the presented study to be of interest to the GP community for the following reasons:

1. It is effective on sparse data such as that from the sensory evaluation domain and the application area of flavor and fragrance design. We address the sparseness of the data by creating ensemble archives of Pareto genetic programming experiments that furnish the model sets for variable important analysis that can be driven by presence or fitness weighting. The standard GP-approach for variable selection, which analyses variable presence in a successful solution, does not work in this context because the single GP solutions are very inadequate and their variable importance statistics are not reliable.

2. It is effective on data, such as sensory evaluation data, where the variation in response values for the same input values (i.e. repeated measures) is extremely high. Our approach consists in modeling the repeated measurement functions (i.e. panelists) independently. This avoids the problem of disrespecting relevant variances in the responses per sample. Our approach retrieves reliable variable importance information from developed models, and then combines these variable importances for the entire panel to obtain robustness.

3. We demonstrate a new means of knowledge mining with GP methods by conducting data analysis in the space of variable importances. In this new space, we observe the evidence that the response and explanatory variable relationship differs among measurement functions (e.g. panelists) and exploit rather than inaccurately average the differences. The information of variable importance facilitates model similarity clustering on this basis and efficient sensitivity analysis.

We are enthusiastic about the results, primarily because they confirm that genetic programming symbolic regression methodology has evolved into a mature field capable of routinely solving real-world problems. In this case study, genetic programming allowed us to decompose a seemingly unsolvable problem (few samples with multiple responses of high variation) into a sequence of solvable problems generating insights at each step.

The most exciting feature of the study is its efficiency - the complete analysis when automated takes a night (or a lot less if multiple cores are available). This, in combination with flavor optimization (see [**?**]) opens up opportunities for new on-line protocols of flavor design, generating new insights in days instead of months. Additionally, panel segmentation, derived on the basis of liking being influenced by the same ingredients in the same direction, will allow a clearer understanding of the hedonic responses to a product suite. When affordable, it may enable the development of products for particular segments leading to higher consumer satisfaction.

10. ACKNOWLEDGMENTS

We thank the anonymous reviewers for very useful comments on the paper and thank Mark Kotanchek and Guido Smits for their technical support.