Chapter 12

TRUSTABLE SYMBOLIC REGRESSION MODELS: USING ENSEMBLES, INTERVAL ARITHMETIC AND PARETO FRONTS TO DEVELOP ROBUST AND TRUST-AWARE MODELS

Mark Kotanchek[1], Guido Smits[2] and Ekaterina Vladislavleva[3]
[1]*Evolved Analytics,LLC, Midland, MI, USA;* [2]*Dow Benelux B.V., Terneuzen, the Netherlands;*
[3]*Tilburg University, Tilburg, the Netherlands.*

Abstract Trust is a major issue with deploying empirical models in the real world since changes in the underlying system or use of the model in new regions of parameter space can produce (potentially dangerous) incorrect predictions. The trepidation involved with model usage can be mitigated by assembling ensembles of *diverse* models and using their consensus as a trust metric, since these models will be constrained to agree in the data region used for model development and also constrained to disagree outside that region. The problem is to define an appropriate model complexity (since the ensemble should consist of models of similar complexity), as well as to identify diverse models from the candidate model set.

In this chapter we discuss strategies for the development and selection of robust models and model ensembles and demonstrate those strategies against industrial data sets. An important benefit of this approach is that all available data may be used in the model development rather than a partition into training, test and validation subsets. The result is constituent models are more accurate without risk of over-fitting, the ensemble predictions are more accurate and the ensemble predictions have a meaningful trust metric.

Keywords: Symbolic regression, Pareto optimality, trust metrics, ensembles, confidence, robust solutions

1. Introduction

The problem with empirical models

Data-driven models are important in real-world applications since in many cases first-principle models either are not possible or practical, because of an absence of valid theory, complexity of input interactions, execution time requirements of a first-principles model or a lack of fundamental understanding of the underlying system. In these situations, a data-driven model is the only viable option to infer the current state of a critical variable, make predictions about future behavior, emulate the targeted system for optimization, or extract insight and understanding about driving variables and their influence.

Unfortunately, there is a problem: *Most empirical models are big piles of "trust me"*.

There is a fundamental limitation of data-driven models in that they are only 100% valid (assuming noise-free data) at the points at which there is data. Since there are an infinite number of models which will perfectly fit a finite data set, we typically impose a preference for simplicity (parsimony), impose a model form or use additional (test and/or validation) data sets to make sure that the model is valid at some other regions of parameter space and hope for the best in using the model.

Alas, these models are NOT valid if used outside the region of parameter space used for the model development ,and possibly not valid within that region, if the underlying model dynamics have changed, or if the model was over-fitted to the data. There is no easy way to detect that a developed model should not be trusted. Inappropriate use of an invalid model can be dangerous – either physically, financially, or both.

The symbolic regression-centric approachmodels

Conventional symbolic regression does not have an inherent advantage for developing trustable models. However, there are three pillars which do support that development:

- Pareto-aware symbolic regression algorithms to develop models of appropriate complexity,
- Interval arithmetic for identifying robust models, and
- Ensembles of diverse models to provide a trust metric.

Together these pillars support developing robust and trustable models – which is a unique and very significant capability for empirical models.

Pareto-aware symbolic regression. Pareto-aware symbolic regression algorithms (Kotanchek et al., 2006) which explicitly explore the trade-off between

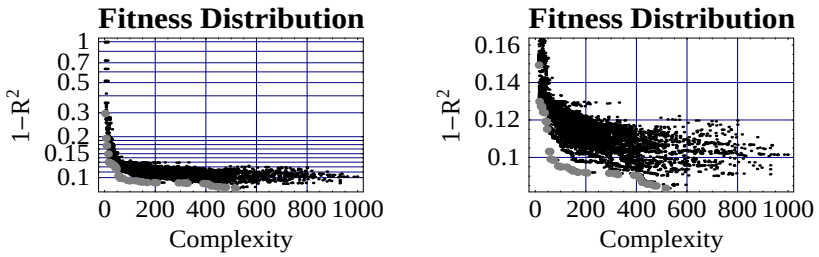

Figure 12-1. Here we show the Pareto front trading off accuracy (1-R^2) complexity for developed inferential sensor models. Note the knee of the curve indicates the point of diminishing returns and is the region from which we would likely prefer to retrieve our models.

model complexity and accuracy allow us the luxury of selecting models which provide the best balance between model accuracy and complexity – with the implicit assumption that overly complex models are at risk of being over-fitted and having pathologies. This behavior is illustrated in Figure 12-1, which shows the distribution of developed models against these two criteria. (The symbolic regression algorithm we use rewards models for being near the Pareto front – which is why the population of models are banded as shown.).

A benefit of being able to define a region of model fitness (accuracy vs. complexity) space which provides the best return on model complexity is that we can use ALL of the data in the model development – which enables quality model development even for systems with small or fat data sets. (This bold statement will be supported through the course of the rest of the chapter.)

Robust model development and identification. Millions of models will be explored in a typical symbolic regression effort. Even if we restrict our attention to the most attractive models, we could have tens of thousands of models that meet our nominal accuracy requirements, and are in the region at the knee of the Pareto front. From this abundance, we need to select models which are also robust – in the sense that they do not contain numerical pathologies which would cause operational problems – as well being accurate and not over-fitted.

Searching nonlinear models with multiple variables is a very computationally intensive exercise and does not have a guarantee of actually finding singularities. Maarten Keijzer (Keijzer, 2003) proposes an alternative based upon interval arithmetic which is very attractive due to its relative speed. This approach may be overly conservative, since some parameter combinations which cause singularities may not be achievable in the real-world due to variable coupling. Alternately, the existence of a pathology may actually be appropriate for some variable situations.

Rather than selecting models for robustness in post-processing, this criterion may be included during the evolutionary development. To some extent we can guide the development by either judicious choice of function building blocks or by eliminating ill-structured models as they are developed. Alternately, a nonlinearity metric can be included in the Pareto-based selection process. Since the efficiency of the multi-objective selection breaks down as the dimensionality of the objectives is increased, an attractive alternative is to use alternating fitness metrics wherein the parsimony aspect (e.g, model complexity, model nonlinearity, etc.) is switched each generation. Of course, the accuracy metric (1-R^2, scale-invariant noise power, norm, etc.) can also be toggled in this algorithmic variant. This approach has been shown (Vladislavleva and Smits, 2007) to improve both the efficiency of the model development, as well as the robustness of the models.

Diverse ensembles for accuracy plus trust. Diverse and independent models will be constrained to agree where there is data and, to a large extent, constrained to disagree away from those constraint points. Obviously, the degree of model divergence will depend upon how far the prediction point is away from the data points. Therefore, the consensus within an ensemble (i.e., a collection of diverse quality models) can implicitly detect extrapolation even in high-dimensional parameter spaces. This is a very unique capability as well as very valuable in real-world applications as a trust metric. Similarly, if the modeled system undergoes fundamental changes, the ensemble models will likely also diverge in their predictions which provides an early warning and awareness that would not otherwise be possible.

One of the keys to a good ensemble is that the models are of similar complexity as well as similar accuracy. The multi-objective Pareto front perspective, therefore, is very important in identifying models for possible inclusion.

Arguably the easiest way to identify diverse models is to look at the correlation of their error residuals and pick a diverse set based upon a joint lack of significant correlation. However, because the developed models share a common reference they will tend to be correlated; otherwise, they would not be the quality models that are wanted. Noisy data will increase that correlation since models will tend to navigate through the center of the fuzzy observed response surface. Because of these two factors, the definition of acceptable levels of correlation needs to be adjusted somewhat.

Although use of ensembles is relatively novel in symbolic regression and genetic programming, they have been a major factor in the industrial success of stacked analytic networks (Kordon et al., 2006) for fifteen years. Climate and meteorological prediction has also embraced ensembles to provide a confidence in weather event prediction (Hamill, 2002). Finally, the machine learning community is beginning to use ensembles in a classification framework (Wichard,

2006). Similar perspectives have also been used in stock portfolio definition (Korns, 2006).

Classic approaches to resolving this conundrum

We have a conundrum. On the one hand, we need a model and we must derive it from the available data and, on the other hand, we suspect that we cannot completely trust the developed models. As we shall see, there is a possible solution via a judicious selection of multiple symbolic regression models; however, let us first review how this problem is addressed by other modeling and machine learning approaches: linear statistics, neural networks and support vector machines.

General foundations for a trustable model. There are a number of data characteristics which make development of a trustable model easier:

- **Representative data** – the data represents the current (and future) response behavior;
- **Balanced data** – the data captures the dynamics and does not unduly represent any region of parameter space;
- **Significant inputs** – nuisance variables are not included in the data set;
- **Abundant data** – this enables coverage of the parameter space as well as definition of model validation data sets.

These ideals are often not achievable in real-world situations. Most systems of interest undergo changes with time either slowly (e.g., parts wear or the economy shifts) or in a step change (e.,g., a down-stream condenser is removed). We may also want the developed model to be transferrable to a similar but different system (e.g., to recommend treatment plans given a set of medical diagnostic tests on a new patient). If the system is in our control, we may be able to run a designed experiment to attempt to generate a balanced data set. However, that may not be an option if there are many variables and we do not know which are the truly significant ones, or if production or safety constraints limit the range or number of experiments which can be executed. If the system is not under our control, we can, obviously, not perform what-if experiments and we are constrained to the data stream that is captured.

Linear Statistics & Trustable Models. The classic linear statistics approach of a control chart is really only applicable to constant output and suffers from a latency problem in that new data must come in and be recognized as different from the expected behavior before action can be taken. If the charted

variable changes over time, this makes the recognition of a model failure much more difficult.

Another limitation is an implicit assumption that the variables used in the model be uncorrelated. To some extent, this can be detected by examining the variance inflation factors (VIF); however, that tends to be a labor-intensive operation. The easiest way to avoid this issue is to use a principle components analysis to identify orthogonal variables as inputs. The downside of this approach is that the resulting models may lose interpretability and the construction of the input variables implicitly assumes a linear coupling of the input variables.

In each iteration of linear model building, we assume an *a priori* a model structure and look for the coefficients which best fit that model to the data. If the assumed model structure matches that of the targeted system, then we have a high quality model. If not, then we must revise the model structure and try again. Although this iterative process is numerically efficient, it can be quite inefficient from a human time standpoint. As such, we are generally restricted to searching low-order polynomials both from an assumption that simple models will be more robust as well as the human effort required to interactivity explore and refine candidate model structures; unfortunately, a mismatch between the true and mathematically convenient models can lead to accuracy and robustness problems[1].

Neural Networks & Trustable Models. Our goal in data modeling is to model the underlying system and not any noise which might be also present in the available data. Since the amount of noise in the data is not known *a priori*, the available data is partitioned into training, test and (possibly) validation subsets. The traditional neural network (NN) approach assumes a model structure and does a (nonlinear) search for the coefficients which best fit the model to the training data with the search being stopped when the performance against the test set degrades since, presumably, at that point the neural network is starting to model the noise rather than the system fundamentals. As an additional check, a validation set may be used to discriminate between developed NNs for the final selection. (Since the coefficient search is a very nonlinear optimization problem, the model coefficients – and, hence, model response behavior – which change with each randomly initialized training run even if the model structure is the same.)

Parsimony also is a factor for NN selection in that we prefer simple model structures. Towards that end, we generally include a direct connection between the input and output nodes as well as via the hidden layers. This produces an

[1]There is a synergy between linear statistics-based modeling and symbolic regression via genetic programming via the discovery of linearizing transforms (i.e., identifying metavariables – variable combinations or transforms) which allow the implicit mathematical assumptions to be satisfied, (Castillo et al., 2004)

underlying linear model which means that the NN has an easier time discovering the global response surface as well as providing a linear response behavior when the model is asked to extrapolate outside the training realm. Because of the preference for parsimony, NN development algorithms will suppress low-significance connections or iteratively evolve network structures in an attempt to achieve robustness.

As with the linear models, identifying that the model should not be trusted relies upon *post facto* assessment of model performance against new data which implies a corresponding lag on detection and response to system changes or new operating regimes. Additionally, if data is scarce, partitioning the data into viable training and test sets becomes an issue. Data balancing also becomes a serious issue in defining the various data subsets.

Support Vector Machines & Trustable Models. A support vector machines (SVM) or, more specifically in this case, support vector regression (SVR) identifies the key data points (vectors) in a data set and builds a response model by placing kernels (e.g., polynomials or radial basis functions) at those points. The predicted response is, therefore, the cumulative contribution of the kernels located at each of the support vectors. Judicious kernel selection can lead to models which are both accurate and extrapolate reasonably well.

The problem still remains that system changes or new operating regimes can only be detected *post facto*. SVR also suffers from the curse-of-dimensionality in that inclusion of spurious inputs can cause robustness problems.

Summary on non-symbolic regression approaches. All of these modeling techniques can produce excellent models for static situations where the operating region of the parameter space is well covered by the data used in the model development, the operating region does not change over time and the targeted system does not change over time. They all will tend to struggle if correlated inputs or spurious inputs are used in the model development. None will provide an assessment of deployed model quality except via after-the-fact analysis.

2. Ensemble Definition and Evaluation

In this section we walk through the process of model development and the definition and use of ensembles built from those models.

Developing diverse models

If the models used to define an ensemble are minor variations on a theme, they can all happily agree as they guide the user to march off a cliff. Hence model diversity is a critical ingredient in a successful ensemble. We have a

number of mechanisms we can use to assist the symbolic regression processing to produce diverse models:

- **Independent runs** – due to the founders effect as well as randomness, each evolutionary exercise will explore a different trajectory in the model search. The more searches we execute, the more likely we are to develop diverse models.
- **Different rescale ranges** – although GP can handle model building with data in the natural data ranges, we can often improve the efficiency of the model building process (sometimes very significantly) by rescaling input variables to a common range. The choice of range will affect both the ease of quality model discovery as well as structure of the developed models. Thus, model building with different rescale ranges helps to produce diverse model structures.
- **Use different subsets** – if data is abundant, we can use different data subsets in model development. This is related to the ordinal optimization approach to developing robust models wherein different subsets are used for each generation within a single evolution (Kotanchek et al., 2006), (Smits and Vladislavleva, 2006), except that here we are using a different subset for each evolution but maintaining that subset throughout the processing.
- **Use different functional building blocks** – obviously the choice of functional building blocks will influence the structure of the developed models. Hence, using a variety of function sets will produce a variety of model structures.
- **Change supplied variables** – Pareto-aware symbolic regression is quite powerful in that it will automatically select the most important variables – even of the supplied variables are correlated. However, for the purposes of an ensemble, we may wish to include intentionally myopic models which feature sub-optimal variables since they can contribute to overall robustness as well as to the quality and accuracy of the trust metric (Korns, 2006).

Other than independent evolutions, the above mechanisms may or may not be important for developing independent candidate models. The relative importance of each diversity introduction mechanism as well as best-practices for their use is an open research area.

Selecting candidate models

Assuming that a large number of (hopefully diverse) models have been developed, we are now faced with the problem of reducing this abundance to

a manageable number as a precursor to the actual ensemble development. A sequence which we have found to be effective is:

1. Identify a rectangular region in fitness space which includes the knee of the Pareto front and select the models which lie within that region,
2. Select the robust models from that subset using an interval arithmetic test over the nominal data ranges (or an extension of that range, if extrapolation performance is important)
3. Select a manageable size subset from the robust models based upon their adjacency to the Pareto front. We will typically select between 500 and 3,000 models.

In Figure 2.0 we show the results of such a model down-selection processing. Although algorithmically simple and efficient, this step is quite important since the quality of the final ensemble will depend upon the quality of the candidate models. Unfortunately, definition of the region from which to draw the models as well as the number of final models to be included in the candidate set varies from problem to problem and requires engineering judgement as to what is appropriate.

Defining ensembles

Rather than selecting THE model from the candidate model set, our goal is to select a diverse ensemble of models. Diversity could be defined a number of different ways:

- Model **structure**,
- Constituent **variables** embedded within the model,
- Lack of prediction error **correlation**, etc...

Unfortunately, although intuitively pleasing, the first two of the above are difficult to quantify. Thus, our approach is to use a lack of residual correlation as the criteria for the ensemble definition with the implicit expectation that models which satisfy that criteria will also satisfy the others.

Algorithms to identify uncorrelated model sets. Using the relative lack of error correlation is an N^2 scaling problem which rapidly becomes intractable as the number of models increases. We generally use a divide-and-conquer approach wherein we randomly partition the models into manageable size subsets ($\sim$100 models) apply a selection algorithm:

1. Build the covariance matrix from the error residuals of the supplied models.

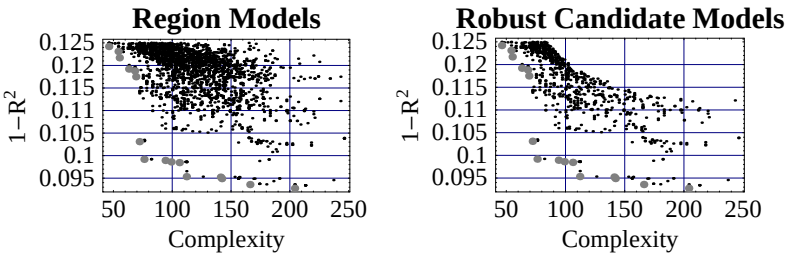

Figure 12-2. Here we show the result of focusing on a region of the fitness landscape for the inferential sensor. Note that by selecting models based upon their adjacency to the Pareto front, we can suppress the inclusion of models with a relatively high complexity and relatively low accuracy. In this case, the number of models went from 2,587 in the rectangular region to 1,576 robust models in the region to 817 in the subset used for ensemble definition. The robustness test selected models which did not have any interval-arithmetic-based pathologies on a ±50% expansion of the nominal training parameter ranges.

2. Select the most uncorrelated model pair that is less than a specified correlation threshold.
3. Select the most uncorrelated model relative to the previously selected models which meets the specified correlation threshold.
4. Repeat step 3 until no models remain which satisfy the correlation threshold.
5. If no models meet the independence threshold, return the most typical model based upon an eigenvalue analysis.
6. Merge the models returned from the subsets and repeat the algorithm to return the final set of uncorrelated models.

As inferred previously, the definition of uncorrelated needs to be adjusted relative to the standard linear statistics threshold of a 30% or less correlation. The default value we use for a threshold is 80% or less correlation; however, this threshold must generally be adjusted based upon the size of the data set as well as the amount of noise in the data set.

An alternate algorithm that is also very efficient is to:

1. Select a reference model,
2. Calculate the correlations of all models with respect to that reference and choose the model with the absolute lowest correlation with the reference model,
3. Re-calculate the correlations of all of the models relative to this new model,

4. Choose the model which has the lowest sum of absolute correlations relative to the selected models,

5. Repeat step 3 and 4 until the desired number of models have been retrieved or there are no models left which satisfy the significance threshold.

Selecting the ensemble components. The outlined methods to select the most diverse models from the candidate model set will generally NOT select from the models on the Pareto front. To some extent, we should expect this since the Pareto front models are, by definition, the optimal performing models and, as a result should be taking the most central path through the fuzzy response surface. Although diversity is important to provide a viable trust metric through the model consensus, not including the Pareto front models is not very emotionally satisfying since they are optimal and quite a bit of effort went into their development.

Since our ensemble will have two goals — consensus metric and prediction — we can make the argument that we should include models from the Pareto front. Towards that end, a strategy that seems to work well in practice is to build the operational ensemble from three components:

- Diverse models selected from the overall candidate set;
- Diverse models selected from the Pareto front of the candidate set;
- The "most typical" model from the Pareto front models.

Intuitively, this strategy should overload in the middle of the (fuzzy) response surface. In our ensemble evaluation scheme, the central models will dominate the prediction while the outer models will dominant the consensus metric. Figure 2.0 illustrates the distribution of the selected models for the inferential sensor.

Note that an ensemble is simply a container for the selected models each of which will be evaluated against parameter sets. Converting these disparate model predictions into an overall ensemble prediction and trust measure is the topic of the next section.

Ensemble evaluation

Ensemble evaluation consists of evaluating all of the constituent models embedded within the model and producing two numbers: a prediction and a consensus metric. The assumption is that the consensus is an indication of the trustworthiness of the prediction.

We generally target to have between 10 to 50 models in the final ensemble with the number of embedded models determining what is viable from a prediction and consensus evaluation standpoint. As a general rule for consensus,

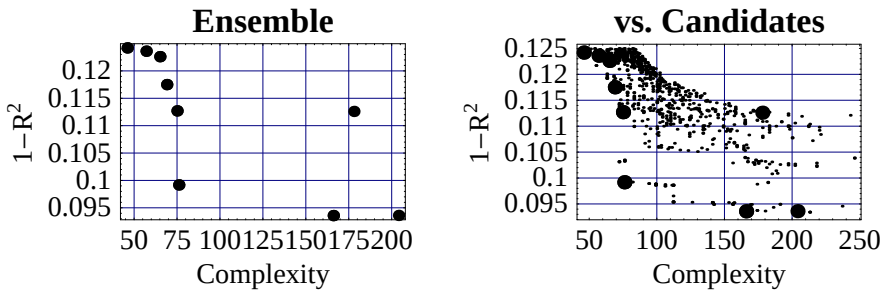

Figure 12-3. Here we look at the distribution of ensemble models on the fitness landscape relative to the overall candidate set. The correlation threshold for ensemble selection from the candidate model set was increased to 90% due to the relatively noisy data set.

we prefer robust statistical measures (mean, MAD, quantiles, etc.) as opposed to conventional statistical metrics (mean, standard deviation, etc.) since the conventional statistics are vulnerable to distortion from outliers. The conventional measures have the advantage that they are faster to evaluate since the constituent model outputs do not need to be sorted; however, that is generally not a significant issue from an operational viewpoint. Note that a common reference frame is important so conventional and robust metrics should not be generally mingled.

An attraction of the mean relative to the median as a ensemble prediction method is that the mean will be a smoother function since the median will undergo step changes as the dominant (median) model changes. A compromise which we use is to provide prediction smoothness as well as robustness is the "median average" — i.e., the average of the predictions from the 3–5 models surrounding the median.

We currently use the median average (averaging at least three models and more if a large ensemble is being used) with the spread (maximum - minimum prediction) used as the trust metric for small ensembles and either the spread, 10-90% quantile range or the standard deviation for large ensembles. However, the choice of ensemble prediction and consensus functions is an open research topic.

3. Ensemble Application

Our contention is that, "The consensus metric from a properly constructed ensemble will give an immediate warning that the ensemble prediction is suspect." This is a significant improvement for real-time systems over conventional methods since the delay until the prediction errors accumulate to a noticeable level is avoided along with the associated risks. Having a consensus metric

is also useful for off-line analysis since such information can guide additional data collection as well as give human insight.

Off-line data analysis

One of the major benefits of symbolic regression — the human insight derived from examining the structure of the evolved expressions — is lost when many disparate models are gathered into an ensemble. That said, it may be that examining the selected models may be more insightful than examining the Pareto front models because of the implicit diversity of structure.

Of course, the response surface of an ensemble can be explored in a similar fashion as an individual model; however, the real benefit of an ensemble could be the exploration of the consensus surface since the divergence of the ensemble models indicates locations in parameter space where data is missing. This could be useful in applications such as combinatorial chemistry which use iterative model development. In essence, this would support an adaptive design-of-experiments approach.

On-line prediction

On-line data analysis is where the use of ensembles has its greatest benefits since the consensus metric acts as a trust metric to warn that the model predictions may not be accurate as well as an indicator that new or refined models may need to be developed. Whether the deviation in the prediction of ensemble models is due to operating in new regions in parameter space or fundamental changes in the targeted system must be determined by the user — however, he or she has been warned!

4. Example: An Inferential Sensor

Inferential sensors

Often in industry we want to monitor a variable which is hard to measure directly. Sometimes also called a "soft sensor", an inferential sensor uses measurements which can be reliably collected on-line to infer the state of the targeted response. One of the keys is to develop a function which will map from the easily observable parameters into the hard-to-observe response. The response may require special equipment or off-line lab testing which means that the training and test data is relatively precious. Of course, there are direct analogues of the inferential sensor to other domains such as financial modeling.

In this example, we have intentionally designed a test data set which spans a larger region of parameter space than the training set used to develop the symbolic regression models. This allows us to demonstrate the ability of a

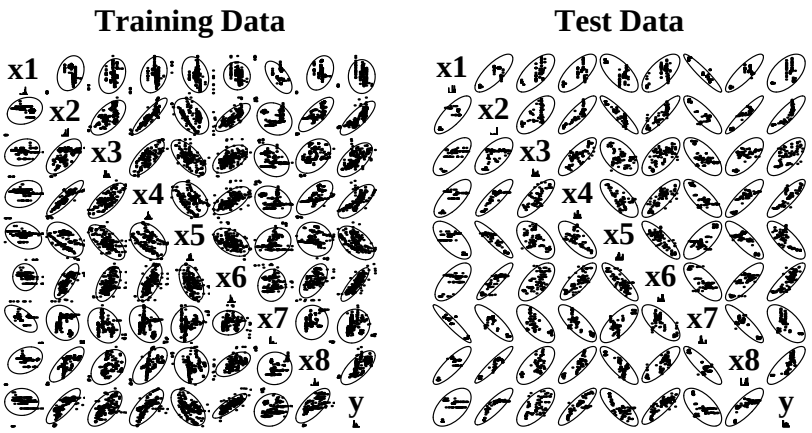

Figure 12-4. The test data set used spans a region of parameter space approximately ±20% larger than that of the training set used for model development. Note that all of the inputs are correlated with the output .

properly designed ensemble to identify when it is unsure of the validity its predictions.

The data

The industrial data in this example is relatively small: 8 input variables and a response. The data was intentionally designed such that the test set (107 data records) covered a larger region of parameter space than the training set (251 records) used in the model development. As illustrated in Figure 4.0, all of the input variables are correlated with the output (lbs/day); this correlation is even stronger with the test data than with the training data.[2]

The models

Although only nine models were selected for the final ensemble in this case, we will not show them in the interests of brevity. Note that four inputs were dominant in the sense that they were present in every model. We should also note that only two models only contained the dominant variables. The other models featured between five and eight inputs.

[2]The example modeling, analysis and figure generation were done using Evolved Analytics' DataModeler (*DataModeler*, 2007) add-on package for Mathematica.

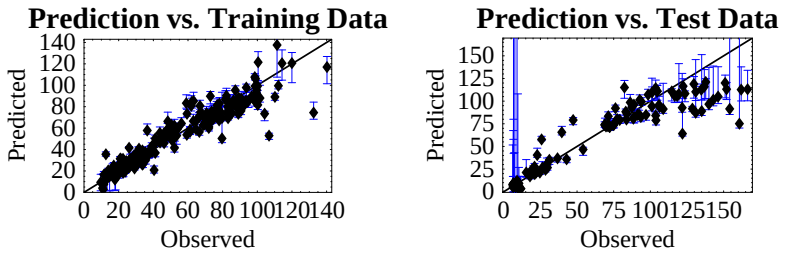

Figure 12-5. Here we show the ensemble performance against the training and test sets. The prediction here is the median model prediction and the consensus metric is the extrema points. The test set covers about 20% greater data range than used for the model development Notice that extrapolation is detected from the model divergences

Ensemble prediction performance

Figure 4.0 shows the model prediction performance (predicted vs. actual) for the training data set. The median model value is shown as is the extrema of the predictions. Note that the model does a relatively good job of prediction — with some problems for the few points at the high end of the range. However, these points do show a prediction divergence which is intuitively comforting.

Figure 4.0 also shows the model performance against the test data set — which was designed to test the ensemble prediction and detection of extrapolation abilities when encountering new regions of parameter space. There are three key points we must make about this graph:

- Extrapolation was clearly detected and flagged as shown by the consensus metric band.

- At the low end, the predictions were actually quite good; however, the extrapolation was detected.

- At the high end, the predictions are incorrect and are flagged as being untrustworthy. However, the ensemble predictions do show a graceful degradation since the predictions are also not wildly incorrect at any point.

The graceful degradation of the models is a result of choosing models from the appropriate region of fitness space (i.e., near the knee of the Pareto front) and further filtering the models by testing for robustness using interval arithmetic. From this foundation of robust and accurate models, we built the ensemble emphasizing diversity so that a valid trust metric would be generated.

The shape of the consensus surface

The response surface of the ensemble is the ensemble prediction as a function of the input parameters. Similarly, the consensus surface is the model disagreement behavior (defined by the consensus measure) as a function of the input parameters. In Figure 5.0 we look at the shape of the example inferential sensor consensus surface under a ±20% extrapolation. Since detection of extrapolation in high-dimensional spaces is generally quite difficult, this is a very important attribute. We can search the consensus surface within the nominal operating range to identify points of model disagreement which would, therefore, be points where we might make extra effort to collect data to refine and improve the quality of the models.

5. Summary

The key message

The essence of this chapter is fairly simple:

- **All the data should be used** in the model development. To do otherwise is to intentionally produce myopic models.
- Models should be developed from a **multi-objective** viewpoint and chosen *post-facto* from the appropriate region of model fitness space.
- Models from this region should be tested for robustness via **interval arithmetic** to eliminate the risk of inappropriate pathologies.
- Ensembles of **diverse models** should be defined from this pool of accurate, simple and robust models.
- The **consensus of models** within a deployed ensemble should be monitored to detect operation in new regions of parameter space as well fundamental changes in the underlying system.

Following these recommendations does not obviate the need for monitoring the quality of the predictions that would be *de rigor* for a conventional machine learning model since it is possible that a system change could be undetected. However, the ability to deploy a model and have immediate assessment of the quality of model predictions is a huge improvement over conventional empirical modeling technologies.

Summary & Conclusions

In this chapter we discussed the development of ensembles of diverse robust models and their operational advantages due to the trust metric provided by the

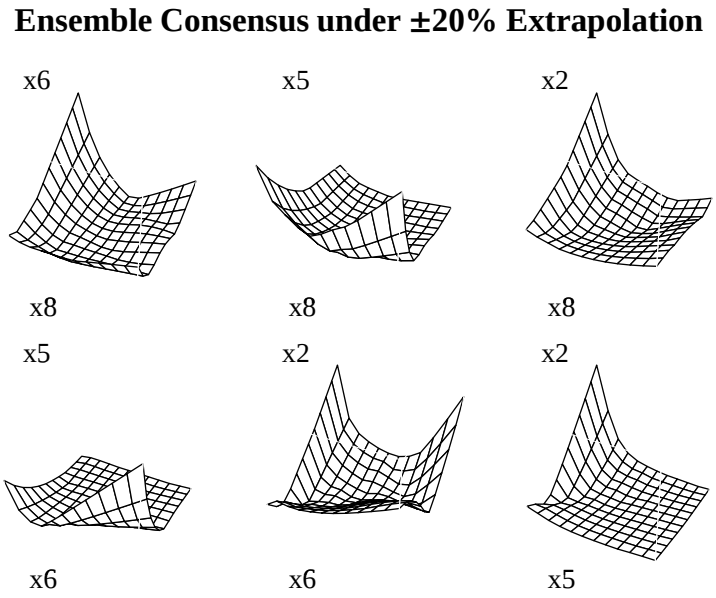

Figure 12-6. Here we look at the consensus surface plot of all pairwise combinations of the four variables which appear in every model within the ensemble (non-varying inputs are held constant at their mean values). As expected, the models diverge when extrapolating; this is an important capability since it is generally difficult to detect extrapolation in a high-dimensional parameter space.

consensus of those models. The associated industrial example demonstrated that an ensemble could detect that it was encountering new regions of parameter space and flag that fact via a consensus metric. From a practical standpoint, this is very significant. The trepidation of deploying a conventional empirical model is based upon four potential problems:

- The model can not handle **data outside the training space** and, as a result, could have wildly erroneous predictions;
- The **underlying system could change** and the model would happily predict assuming that the training data (system) was still appropriate;
- The chosen model was **over-fitted to the training data** and, as a result, could have wildly erroneous results even if new data was within previous operating conditions and the underlying system had not changed;
- Problems in model quality could only be detected after-the-fact with a further **detection delay** imposed to collect sufficient new data to declare the deployed model to be invalid.

The traditional way to mitigate the overfitting risk is to partition the data into multiple subsets, train against one of the subsets and select based upon the model performance against the multiple subsets. To a large extent using the Pareto front in symbolic regression and selecting models from the knee of the Pareto front naturally guards against either over-fitting or under-fitting. Using interval arithmetic to eliminate models with potential pathologies further mitigates the empirical modeling risk.

Conventionally, detecting that the model was no longer valid either due to the underlying system changing or due to input data outside the training range essentially involved diligent oversight on the part of the user to track the model output against reality and to make a judgement call as to whether errors were normal or if there was a structural problem with the model-reality match. The problem in practice is that the human oversight required is either not done or not done in a timely fashion; in any event, the problem with model prediction quality could only be discovered after the fact which implies an unavoidable time delay in response. A trusted incorrect model can be costly or dangerous. Using ensembles of diverse models mitigates this risk since immediate assessment of prediction quality is provided via the trust metric derived from the diversity of predictions from the models embedded in the ensemble.

Diverse model ensembles enable some very profound practical advantages for real-world model deployment:

- We can now **use ALL available data in the model development**. If data is precious, we are essentially removing the fogged glasses we placed on

the modeling process to avoid the risk of over-fitting. Traditionalists are going to be VERY disturbed at this; however, there is no real need for the test/training/validation/etc. data partitioning!

- We can **let the ensemble warn us when the model output should not be trusted**. The ensemble can detect that the ensemble is extrapolating into new regions of parameter space and warn us immediately. Timely awareness allows the human to make appropriate judgements as to whether processing parameters need to be adjusted or a new model is required to reflect fundamental changes and avoid costly surprises.

Although we believe that the use of symbolic regression ensembles represents a profound opportunity in real-world model deployment, it should not be construed that we advocate the abdication of human responsibility for ensuring that the models are accurate and applicable. Any empirical model should be monitored and calibrated, as appropriate. Some symbolic regression models that we have developed are still in active use in the chemical process industry close to a decade after the original development so it is possible to develop robust and effective models.

Issues and future efforts

The god-father of ensembles for symbolic regression are the stacked analytic networks (Kordon et al., 2003), which have been deployed in industrial application for over fifteen years as of this writing. Lessons learned from that technology is that ensemble models should be of similar complexity but diverse. Unfortunately, those two characteristics are harder to define in a symbolic regression context. Some of the open issues we are currently addressing are:

- The nature of empirical modeling is that model predictions will be highly correlated – otherwise, they would not be quality models. Hence, we need to define an appropriate correlation threshold to declare model independence. This threshold will change depending upon the system complexity, number of data points and noise levels.

- Consensus (or, rather, lack of consensus) is the key trust metric of an ensemble. We currently use the spread or a summary statistic (e.g., standard deviation) depending upon the number of models in the ensemble. Is there something better?

References

Castillo, Flor, Kordon, Arthur, Sweeney, Jeff, and Zirk, Wayne (2004). Using genetic programming in industrial statistical model building. In O'Reilly,

Una-May, Yu, Tina, Riolo, Rick L., and Worzel, Bill, editors, *Genetic Programming Theory and Practice II*, chapter 3, pages 31–48. Springer, Ann Arbor.

Hamill, Thomas (2002). An overview of ensemble forecasting and data assimilation. In *Preprints of the 14th conference on Numerical Weather Prediction*, Ft.Lauderdale, USA. American Meteorological Society.

Keijzer, Maarten (2003). Improving symbolic regression with interval arithmetic and linear scaling. In Ryan, Conor, Soule, Terence, Keijzer, Maarten, Tsang, Edward, Poli, Riccardo, and Costa, Ernesto, editors, *Genetic Programming, Proceedings of EuroGP'2003*, volume 2610 of *LNCS*, pages 70–82, Essex. Springer-Verlag.

Kordon, Arthur, Smits, Guido, Kalos, Alex, and Jordaan, Elsa (2003). Robust soft sensor development using genetic programming. In Leardi, R., editor, *Nature-Inspired Methods in Chemometrics: Genetic Algorithms and Artificial Neural Networks*. Elsevier, Amsterdam.

Kordon, Arthur, Smits, Guido, and Kotanchek, Mark (2006). Industrial evolutionary computing. In *GECCO 2006: Tutorials of the 8th annual conference on Genetic and evolutionary computation*, Seattle, Washington, USA. ACM Press.

Korns, Michael F. (2006). Large-scale, time-constrained symbolic regression. In Riolo, Rick L., Soule, Terence, and Worzel, Bill, editors, *Genetic Programming Theory and Practice IV*, volume 5 of *Genetic and Evolutionary Computation*, chapter 16, pages –. Springer, Ann Arbor.

Kotanchek, Mark, Smits, Guido, and Vladislavleva, Ekaterina (2006). Pursuing the pareto paradigm tournaments, algorithm variations & ordinal optimization. In Riolo, Rick L., Soule, Terence, and Worzel, Bill, editors, *Genetic Programming Theory and Practice IV*, volume 5 of *Genetic and Evolutionary Computation*, chapter 3, pages –. Springer, Ann Arbor.

Smits, Guido and Vladislavleva, Ekaterina (2006). Ordinal pareto genetic programming. In *Proceedings of the 2006 IEEE Congress on Evolutionary Computation*, Vancouver. IEEE Press.

DataModeler (2007). Add-on analysis package for *Mathematica*.

Vladislavleva, Ekaterina and Smits, Guido (2007). Order of non-linearity as a complexity measure for models generated by symbolic regression via genetic programming. In *review at IEEE Trans. on Evolutionary Computation (sumbitted)*.

Wichard, Joerg (2006). Model selection in an ensemble framework. In *Proceedings of the IEEE World Congress on Computational Intelligence WCCI 2006, Vancouver, Canada*.