

Finding Rules for Food Products with FormRules

Background

Creating a high-protein nutritional cereal-based food poses a challenge for food scientists, since the cereals wheat and rice are generally low in protein. Soybeans, on the other hand, are high in protein, but adding them to a cereal-based food will improve the amino acid balance. Adding soybeans, though, often has an adverse effect especially to taste.

Formulation Data

A study reported by Ruguo Huo, in *Food Product Design: A Computer-Aided Statistical Approach* (1999) addresses the issue of creating a high protein cereal-based product by an extrusion process, to create an instantly soluble nutritional food with good sensory properties. The ingredients were water, soybean, parboiled rice and wheat – these were required to sum to the same weight, and were expressed as percentages in the data.

Process conditions were also included, in particular the temperature and the screw rotation speed appropriate to the extrusion process. 48 different experiments were carried out, following a simplex centroid mixture design for the 4 ingredients (which gave 12 experiments) together with 4 factorial trials (high and low values for the 2 process parameters).

The properties measured were the energy value (in kJ/100g), the protein content, protein quality, and a 'sensory factor' which was a composite of scores for appearance and colour, texture and viscosity, and odour and taste.

Models for Food Products

This example is a relevant one for data mining, since most of the conclusions – except perhaps for the sensory factor – will be obvious to most knowledgeable consumers. So, it is interesting to see whether **FormRules**, which has no prior knowledge of the system, can 'learn' accurate rules directly from the data.

The 48 data records were imported into **FormRules**, and 'mined' using neurofuzzy logic, first with the Structural Risk Minimization model assessment criterion with

C1 parameter of 1. This choice is appropriate since there are a reasonable number of data points in the experimental design. This choice gave very good models for the protein content and energy value, but a relatively poor model for the sensory factor (R^2 value of 0.53). This is perhaps not surprising, since the data for sensory factors, which were obtained from a panel of assessors, is likely to be more variable than the 'hard' numbers of energy value and protein content.

Consequently, we adopted a 'mixed' procedure, in which $C1=1$ was used for the energy value and protein content outputs, but the value of $C1$ for the sensory score was varied. $C1=0.896$ (the value calculated by the **FormRules** AutoScale function) gave a model with $R^2 = 0.71$. When we decreased $C1$ to 0.8 (which can be expected to give more complex models and a better fit to the data) R^2 did not improve – suggesting that there is considerable inherent scatter in the data and that better models will not be obtained.

Rules for Food Products

With $C1=1$, the relatively poor model suggested that only soybean was important for determining the sensory factor. The rules obtained were

IF Soybean is LOW THEN Sensory score is HIGH (0.70)
IF Soybean is MID THEN Sensory score is HIGH (0.77)
IF Soybean is HIGH THEN Sensory score is LOW (0.88)

where the values in parentheses are 'confidence levels'. This rule is in line with the expectations of experienced food formulators.

For the 'mixed $C1$ ' model, the sensory score depended on three inputs, as shown in the Figure below. Soybean, as expected remains important. However, the amount of rice, and the temperature, have a secondary effect.

The rules are:

IF Soybean is LOW THEN Sensory score is HIGH (0.86)
IF Soybean is MID THEN Sensory score is HIGH (1.00)
IF Soybean is HIGH THEN Sensory score is LOW (1.00)
and

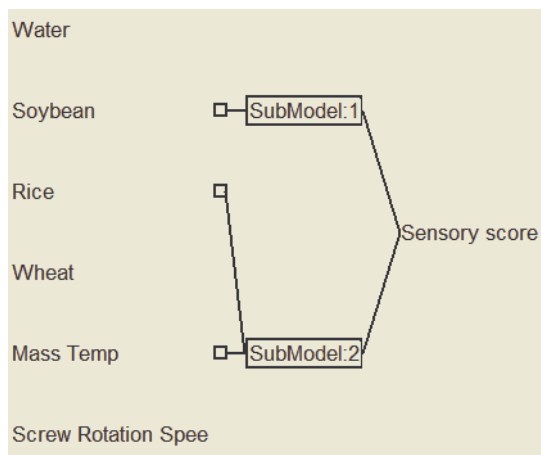
IF Mass Temp is LOW AND Rice is LOW THEN Sensory score is HIGH (0.63)

IF Mass Temp is LOW AND Rice is HIGH THEN Sensory score is HIGH (0.52)

IF Mass Temp is HIGH AND Rice is LOW THEN Sensory score is LOW (0.87)

IF Mass Temp is HIGH AND Rice is HIGH THEN Sensory score is HIGH (0.79)

The sensory score is therefore reasonably high as long as the extrusion mass temperature is low – the mid-range 'confidence levels' indicate that the sensory scores are mid-range rather than high or low. When the mass temperature is high, then the amount of rice is crucial in determining the sensory score – adding more rice improves the sensory score.



It is interesting to note that these conclusions are similar to those of the statistical study of Hu, who concluded that increasing rice and wheat would improve the sensory score, and that rice contributes more than wheat to the sensory score. However, it has been considerably easier to extract this information with **FormRules** than with the response surfaces from statistics.

Protein content depends on the amount of water and soybean (where **FormRules** identifies a weak interaction between the two inputs) and on the amount of wheat. The rules for the amount of wheat are not overly useful, and are

IF Wheat is LOW THEN Protein Content is HIGH (0.77)
IF Wheat is HIGH THEN Protein Content is HIGH (0.94)

This indicates that increasing the amount of wheat will increase the protein content. However, soybean remains the dominant contributor to the protein content, with the rules from the relevant submodel being

IF Soybean is LOW AND Water is LOW THEN Protein Content is LOW (1.00)
IF Soybean is LOW AND Water is MID THEN Protein Content is LOW (1.00)

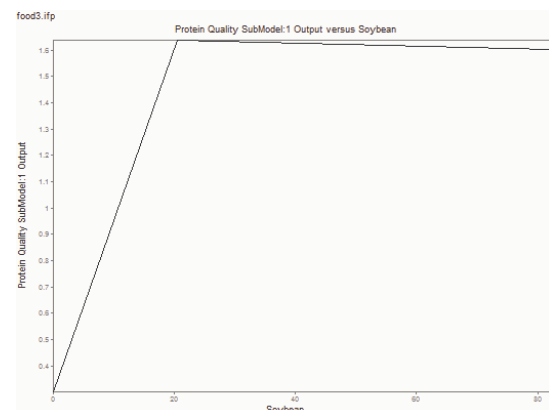
IF Soybean is LOW AND Water is HIGH THEN Protein Content is LOW (1.00)

IF Soybean is HIGH AND Water is LOW THEN Protein Content is HIGH (1.00)
IF Soybean is HIGH AND Water is MID THEN Protein Content is HIGH (1.00)
IF Soybean is HIGH AND Water is HIGH THEN Protein Content is HIGH (1.00)

These can be summarised simply by saying that if soybean is high, then protein content is high – and vice versa.

Energy value depends on all the ingredients, but not on the processing conditions – as expected. The rules are quite complex for this property, with several interactions between them. These interactions are perhaps the result of the experimental design, in which the amounts of the ingredients had to sum to 100%.

Protein quality depends only on the soybean amount, and is high as long as the soybean amount is greater than about 20%. The submodel plot is



Conclusions

The rules generated from **FormRules** are in line with those expected by experienced food formulators.

The analysis of the data was straightforward, with rules presented in a clear way that was easier to understand than the graphs produced from the statistical methods.

Small changes to the default C1 parameters were useful in improving the models generated by **FormRules**.

© 2005 **Intelligensys Ltd**
All rights reserved.