Information-Theoretic Feature Selection for the Classification of Hysteresis Curves*

Vanessa Gómez-Verdejo[1,2], Michel Verleysen[2], and Jérôme Fleury[3]

[1] Department of Signal Theory and Communications, Universidad Carlos III de Madrid Avda. Universidad 30, 28911 Leganés (Madrid), Spain
[2] DICE - Machine Learning Group, Université catholique de Louvain, 3 Place du Levant, B-1348 Louvain-la-Neuve, Belgium
[3] Manufacture Française des Pneumatiques Michelin, Bât F32 - Site de Ladoux, 23 Place des Carmes-Deschaux - 63040 Clermont Ferrand

Abstract. This paper presents a methodology for functional data analysis. It consists in extracting a large number of features with maximal content of information and then selecting the appropriate ones through a Mutual Information criterion; next, this reduced set of features is used to build a classifier. The methodology is applied to an industrial problem: the classification of the dynamic properties of elastomeric material characterized by rigidity and hysteresis curves.

1 Introduction

Modeling data having specific structure properties is an important challenge in data analysis. Structures include trees, functions, multi-level data, graphs, and many others. Functional data, i.e. data that are intrinsically curves (despite the fact that they are usually known through a finite sampling) form an important class of data, as they are found in many industrial applications. In particular, functional data can consist in signals, spectra, hysteresis curves, etc.

In practice, working with functional data means to extract a sufficient number of appropriate characteristics (features) from the functions, and then analyzing the features in a quite traditional way. "Sufficient" and "appropriate" are however vague terms, that need to be defined more precisely in the context of a specific problem. This makes functional data analysis often application-driven in practice, relying on expert knowledge and lacking of general methodology. In addition, when expert knowledge is used to extract information from raw (functional) data, this may prevent the discovery of features that could increase the modeling accuracy but are not known by experts.

This paper presents a methodology for functional data analysis, which consists in extracting a large number of features, most of them revealing later to be useless for the analysis, and then selecting the appropriate ones through a principled variable selection procedure. To fit with nonlinear data models, the criterion used to select variables is the Mutual Information (MI) [1].

* The work of V. Gómez-Verdejo was supported by CM Pjt. S-0505/TIC/0223.

F. Sandoval et al. (Eds.): IWANN 2007, LNCS 4507, pp. 522–529, 2007.

The methodology is applied to the classification of industrial elastomeric material characterized by their dynamic properties (rigidity and hysteresis curves).

In the following of this paper, Section 2 describes the methodology to extract features from curves and selecting them for further modeling, and Section 3 applies it to an industrial application of rigidity and hysteresis curve classification.

2 Methodology

The general methodology to analyze functional data consists in four steps:

1. Feature creation. The idea here is to create a potentially large set of features, in order to avoid any risk to loose information about the functional data.
2. Feature selection. A reduced set of relevant features is selected among the initially created ones, in order to make easier the following stage design.
3. Classification. The reduced set of features is used to characterize the data and build a classifier.
4. Optional addition expert knowledge. If needed, features with high physical interpretation power can be used to replace automatically selected ones, with minimal loss of classification performances.

These steps will be described in the following; although, the first and third steps will be detailed in the application example in Section 3.

2.1 Feature Creation

Feature creation is by essence strongly linked to the application data. The principle is to create a set of features with maximal content of information. This means to take basic features such as rough measurements on the data (sampled curves), and all other features that *could* be relevant for the problem. Traditional possibilities include functional approaches such as the coefficients of splines or other aproximators of the curves, extraction of the numerical (usually first and second) derivatives, the location of extrema, the area under curve, etc. The idea here is to select a *sufficient* set, even if the price to pay is to increase the number of features. Variables that *could* (but do not necessarily *will*) play a role in the further analysis process should therefore be taken into account.

2.2 Feature Selection

The feature selection stage consists in choosing, among the created features, those that are most useful to solve the classification problem. This process has two goals. First, it reduces the data dimensionality, making easier the design of the classifier by reducing the *curse of dimensionality.* Second, having a low number of features makes the interpretation of the model easier: with an objective measure of the usefulness of features, practitioners may deduce information about the process, and drive their next measurement campaigns accordingly.

Feature selection procedures require to combine two key elements: a relevance criterion and a subset search procedure.

The relevance criterion scores a feature or a group of features according to its/their capacity for predicting the output. Among the possible criteria, the Mutual Information (MI) measure has been selected for the design of this system stage. MI measures the amount of information contained in X in order to predict Y; unlike correlation, MI measures *any* relation between X and Y, not only linear ones. Furthermore, the MI concept is directly applicable to groups of variables (i.e. X and/or Y vectors instead of scalars). Let us denote the marginal probability density function (pdf) of X and Y as $p_x(x)$ and $p_y(y)$ respectively, and the joint pdf of X and Y as $p_{x,y}(x,y)$. The MI is given by [2]:

$$I(X,Y) = \int\int p_{x,y}(x,y) \log \frac{p_{x,y}(x,y)}{p_x(x)p_y(y)}. \tag{1}$$

As pdf are not known in practice, the MI value has to be estimated. Traditional histograms and kernel-based estimators [3] could be used. Nevertheless, the search procedure below will use vectors X of increasing dimension. Histograms and kernels are not robust estimators when the dimension increases; a MI estimator based on nearest-neighbor search, provided by Kraskov et al. [4], and implemented in the MILCA toolbox [5], is preferred.

The other ingredient of the feature selection process is the search procedure that allows to find the most adequate subset of features (the ones that achieve the maximum MI value) without evaluating all the 2^N possible subsets among N variables, what it is unpractical from a computational time point of view. For this purpose a Forward-Backward algorithm is used, inspired from [6], but adapted here to the specific problem of hysteresis curve classification:

1. The first selected feature is the one, from the set of all the original features $\{X_1, \ldots, X_N\}$, that maximizes the MI with the output variable Y, i.e.,

$$X_1^{sel} = \text{argmax}_{X_j} \hat{I}(X_j, Y) \quad 1 \le j \le N \tag{2}$$

where $\hat{I}$ is the estimation of I and X_1^{sel} is the first selected feature.

2. The next components must be selected so that the MI between the output Y and the selected set of variables is maximized; in other words, if the algorithm is in t-th step and, for the time being, the features $\{X_1^{sel}, \ldots, X_{t-1}^{sel}\}$ have been selected, the next selected feature, X_t^{sel}, must be obtained as:

$$X_t^{sel} = \text{argmax}_{X_j} \hat{I}\left(\{X_1^{sel}, \ldots, X_{t-1}^{sel}, X_j\}, Y\right) \qquad 1 \le j \le N, \quad X_j \notin \{X_1^{sel}, \ldots, X_{t-1}^{sel}\}. \tag{3}$$

3. After adding X_t^{sel}, the backward procedure consists in checking if removing one by one any of the previous selected components, there is a MI increment. If the removal of several variables (one by one) leads to increasing the MI, the one (X_t^{rem}) that produces the largest increment is removed, i.e.,

$$X_t^{rem} = \text{argmax}_{X_j^{sel}} \hat{I}\left(\{X_1^{sel}, \ldots, X_{j-1}^{sel}, X_{j+1}^{sel}, \ldots, X_t^{sel}\}, Y\right) \quad 1 \le j \le t, \tag{4}$$

$$\text{if} \quad \hat{I}\left(\{X_1^{sel}, \ldots, X_{j-1}^{sel}, X_{j+1}^{sel}, \ldots, X_t^{sel}\}, Y\right) > \hat{I}\left(\{X_1^{sel}, \ldots, X_t^{sel}\}, Y\right)$$

4. Finally, the algorithm is stopped when the MI cannot be increased anymore, or, possibly, when a prefixed maximum number of features is reached.

2.3 Classification

This work does not focus on the classification stage itself. Once a set of features has been created and selected, the problem becomes a traditional classification problem on vector data. Therefore any nonlinear classification tool [7], such as a Multi-Layer Perceptron (MLP), a Radial-Basis Function Network (RBFN), a Support Vector Machine (SVM), etc. could be used.

2.4 Incorporating Physical Knowledge About Variables

An important aspect of the above procedure is that knowledge about the variables can easily be incorporated between the feature selection and classification stages. In particular, one might be tempted to replace some of the selected variables by other ones having more physical interpretation, if the price to pay in terms of classification performance decrease is low.

More specifically, the feature selection results in a single set of variables. Nevertheless, nothing prevents the user to test other sets, i.e. to evaluate their information content by using the same estimator of mutual information as above. Of course, all combinations of variables cannot be tested, otherwise the benefit of the greedy feature selection would be lost. However, one can for example measure the mutual information between all pairs of single variables, and replace one by one non-interpretable variables by interpretable ones, choosing the latter as to maximize the mutual information with the deleted variable (they are more or less equivalent) and in the meantime minimize the loss of mutual information between the new set and the output (the replacement does not change a lot the information content of the set). The choice is intentionally left to the expert user, who is able to judge if the price to pay (in terms of performance decrease) is acceptable to gain interpretability.

It is insisted on the fact that use of the mutual information criterion makes this last procedure possible and measurable in an objective way, if needed by the application, without making it mandatory.

3 Rigidity and Hysteresis Curves Classification

In this section, the methodology described in the previous section is used to solve a specific, industrial problem of rigidity and hysteresis curves classification.

3.1 Original Data

In order to know the validity or conformity of a material, each sample of material undergoes a deformation process. First an external force is applied over the material; secondly, this force is removed which makes the material come back to

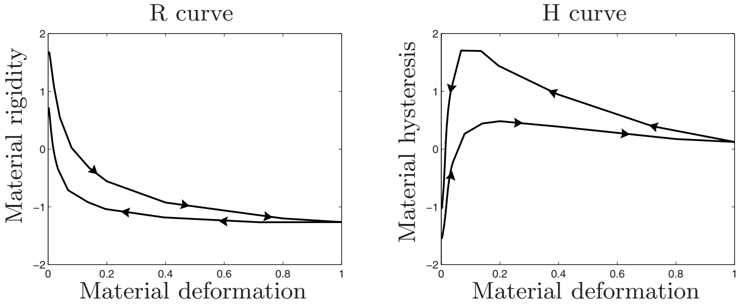

Fig. 1. Example of the normalized R and H curves

its original state. During this process, both the rigidity and the hysteresis of the material are measured for a number (here 23) of deformation values, resulting in two curves, called R and H curves respectively in the following. These two curves will be used to evaluate the validity of the material.

Figure 1 shows the typical shapes of these curves. Both the x- (deformation) and y-axes (rigidity/hysteresis) have been normalized, both for confidentiality reasons, and because this normalization does not influence the further processing.

For each material, a data vector with 47 components is measured:

- Component 1: temperature of the experiment.
- Components 2-24: values of the R curve in the material deformation positions (12 in the forward curve and 12 in the backward one, with the point corresponding to the largest deformation in common).
- Components 25-47: values of the H curve in the same deformation positions.

Learning samples belong to three classes called "conform", "non-conform" and "unknown" by experts; the last class is used when the experts do not know or disagree on the validity of the material and its measure. The classes are respectively labeled 1, -1 and 0. Despite the fact that the problem has three classes, this is one of the rare examples where more than two classes may be labeled with a single numerical value. The ordering introduced here is indeed legitimate, as the "unknown" class is clearly between the two other ones.

The training set provided by the industrial experts consists of 633 data (each with the 47 above components); 483, 112 and 83 are respectively from the "conform", "unknown" and "non-conform" classes. Besides, a test set with 168 data (respectively 119, 22 and 27) is provided.

3.2 Creation of New Features

Since the R and H curves allow experts to know the validity or not of each material, this step consists in creating a new set of features, trying to extract the maximum information from the curves. Previous knowledge on what exact information to extract from the curves is not available. Therefore, according to the methodology described in Section 2, a set of 191 potentially interesting features, which are described in the Table 1, is created .

Table 1. Description of the features

Feature description	Feature number	
	R curve	H curve
temperature of the experiment	1	
original values	2-24	97-119
area under the curve	25	120
numerical first derivatives	26-47	121-142
widths of the curve	48-58	143-153
coefficients of 5^{th} degree polynomial	59-70	154-165
coefficients of linear approximation	71-74	166-169
coefficients of quadratic approximation	75-80	170-175
maximum and minimum points	81-88	176-183
statistical information (moments)	89-96	184-191

3.3 Feature Selection

The 191 features are entered into the selection procedure described in the previous section. However, because of the high number of features in this application, a hierarchical approach is used to reduce the computation time. It consists in applying the selection procedure independently over each curve and, besides, over each group of features; the search algorithm thus is divided in three substages, working over a moderate number of features and accelerating the feature selection process. The process and the selected features are shown in Table 2. As a result of this process only five features are selected, all of them from the H curve. The remaining (uninteresting) ones are discarded for the next step.

3.4 Classification

This stage consists in the design of a classifier that will allow to estimate to which class each data belongs; for that purpose, a Multi Layer Perceptron (MLP) with one hidden layer and a hyperbolic tangent as activation function for both hidden and output layer has been employed. The number of hidden neurons has been fixed to 23 by a Cross Validation process and, for the output layer, only one neuron has been used; in this way, an output value in the range $[-1, 1]$ will indicate, with the help of two thresholds, the estimated class for each data.

To prevent the MLP from overfitting, early stopping over a random partition of the learning set (80% for training and 20% for validation) has been used. After training, the MLP provides for each new sample an output value in the $[-1, 1]$ range. Two thresholds η_+ and η_- should therefore be fixed, in order to discriminate respectively between the "conform" and "unknown" classes, and between the "non-conform" and "unknown" ones. Mid-range thresholds (-0.5 and 0.5) would not be adequate for three reasons: first because they do not necessarily lead to a minimum number of misclassifications in general, second because the costs of misclassifications could be unequal between classes (in this applicataion for example, the risk of classifying non-conform materials in the conform class

Table 2. Different stages of the feature selection process

	Initial subsets	Substage 1 (over subsets)	Substage 2 (over curves)	Substage 3 Final selection
Experiment conditions: temperature $\{1\}$				$\{134, 146, 150, 170, 189\}$
R Curve	Orig. feat.: $\{2-24\}$	$\{3, 5, 6, 10, 13, 19\}$	$\{91, 92, 95\}$	
	Area: $\{25\}$	$\{25\}$		
	Derivatives: $\{26-47\}$	$\{32, 33, 37, 38, 42\}$		
	Width: $\{48-58\}$	$\{48, 50, 51, 55, 57, 58\}$		
	Poly. coef.: $\{59-80\}$	$\{63, 72, 74, 75, 79\}$		
	Max.-Min.: $\{81-88\}$	$\{81, 83, 87, 88\}$		
	Statistics: $\{89-96\}$	$\{91, 92, 95\}$		
H Curve	Orig. feat.: $\{97-119\}$	$\{97, 105, 109, 117\}$	$\{134, 146, 150, 170, 189\}$	
	Area: $\{120\}$	$\{120\}$		
	Derivatives: $\{121-142\}$	$\{123, 129, 134, 136, 141\}$		
	Width: $\{143-153\}$	$\{145, 146, 148, 149, 150\}$		
	Poly. coef.: $\{154-175\}$	$\{163, 165, 168, 170, 172\}$		
	Max.-Min.: $\{176-183\}$	$\{176, 177, 178, 179, 183\}$		
	Statistics: $\{184-191\}$	$\{184, 185, 188, 189\}$		

should be strictly limited), and third because classes are not equally populated. The choice of the thresholds is thus left to the application experts, guided by the ROC curves calculated over the training data. Concretely, the experts have selected the values $\eta_+ = 0.8$ and $\eta_- = -0.1$ which result in the following TPR (True Positive Rate) and FPR (False Positive Rate): TPR = 71.10% and FPR = 5.66% for the “conform” class, and TPR = 98.42% and FPR = 53.47% for the “non-conform” one. The close to zero value (-0.1) for the η_- threshold may seem strange at first sight. However, it is justified by the fact that the “non-conform” class is underpopulated with respect to the two other ones, making the learning of the MLP giving less weight (in a sum-of-squares error criterion) to this class. Giving an increased weight to this class (therefore avoiding a too high FPR) is then possible by moving the threshold towards the selected value.

3.5 Experimental Results

In this section the effectiveness of the proposed methodology is analyzed comparing the proposal system with a system designed by the experts; concretely, the experts have selected a set of variables considered more relevant for the problem resolution and, next, they have trained and tested a classifier with the same data sets than it has been used for designing the proposal system.

In the Table 3 the TPR and FPR that both the proposed system and the experts’ system achieve in each class are presented. Comparing these rates, it can be observed that the proposed system presents a clear advantage with respect to the experts’ system in all the cases. In particular, the TPR is increased by 3.35%, 7.33% and 1.71% and the FPR is reduced by 0.2%, 3.88% and 0.38%, respectively for the “conform”, “unknown” and ” no conform” classes.

Table 3. Results achieved by the proposed system and the experts' system

			Proposal		Reference	
"conform"	TPR		84.86%		81.51%	
	FPR	"unknown"	7.1%	7.8%	8%	8%
		"no-conform"	0.7%		0%	
"unknown"	TPR		72.55%		65.22%	
	FPR	"conform"	11.74%	18.04%	15.07%	21.92%
		"no-conform"	6.30%		6.85%	
"no-conform"	TPR		64.67%		62.96%	
	FPR	"conform"	0.62%	2.44%	0%	2.82%
		"unknown"	1.82%		2.82%	

The application of the proposed general methodology to the classification of the dynamic properties of elastomeric material has provided a reduced set of features with enough problem knowledge to be able to obtain results even better than those achieved by a system designed "ad-hoc" for the problem resolution.

4 Conclusions

This paper presents a methodology to extract knowledge from a functional classification problem. Blind feature creation is combined to automatic extraction to build an efficient set of features. Prior expertise can be incorporated in the procedure, though the latter is able to identify features not known in advance. The methodology is applied to an industrial classification problem of dynamic properties of elastomeric material, and proves to improve both the quality and the interpretation of the results.

References

1. Battiti, R.: Using the mutual information for selecting features in supervised neural net learning. IEEE Trans. Neural Networks 5, 537–550 (1994)
2. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley, N. Y (1991)
3. Scott, D.W.: Multivariable Density Estimation: Theory, Practice and Visualization. Wiley, N. Y (1992)
4. Kraskov, A., Stögbauer, H., Grassberger, P.: Estimating mutual information. Phys. Rev. E 69, 66138 (2004)
5. Astakhov, S., Grassberger, P., Kraskov, A., Stögbauer, H.: MILCA software (Mutual Information Least-dependant Component Analysis) Avalible at http://www.klab.caltech.edu/kraskov/MILCA/
6. Rossi, F., Lendasse, A., François, D., Wertz, V., Verleysen, M.: Mutual information for the selection of relevant variables spectrometric nonlinear modelling. Chemometrics and Intelligent Laboratory Systems 80, 215–226 (2006)
7. Haykin, S.: Neural Networks: A Comprehensive Foundation, 2nd edn. Prentice-Hall, Englewood Cliffs (1999)