AUTOMATIC GENERATION OF DIMENSION REDUCTION PROCEDURES

S.V. Markevich, S.S. Chernova
Kharkevich Institute for Information Transmission Problems of Russian Academy of Sciences
Moscow, RUSSIA
e-mail: markevich@iitp.ru
Institute for System Analysis of Russian Academy of Sciences
Moscow, RUSSIA
e-mail: chernova@cpt-ran.ru

Abstract

The paper describes a technique used for constructing automatic procedure generators that create software modules implementing basic data analysis and processing procedures. The generators are data-specific and can be used as external libraries in the development of specialized engineering applications.

1 Introduction

Enhancing the operational reliability of an engineering object has always been a topical and challenging issue. Many baseline decisions regarding the components, structure, materials and other parameters are made at the conceptual design phases based on the analysis of various typical performances displayed by different configurations of the future object under diverse operating conditions.

Mathematical modeling and computational experiments dealing with analytical models of a future object and its environment have become widely adopted for the analysis and optimization of an engineering object's structure. The analysis of an object's performances leverages on the data about the reliability of its elements and similar objects, the properties of the materials used and any other available information. The computed performances can be added to the existing data bases and used in subsequent computational experiments.

Progress in the development of intelligent data analysis instruments calls for new applied task-specific engineering tools offering a synergy up-to-date data handling methods based on advances in general science and domain-specific techniques. The new tools should also be able to adapt to the specifics of a subject domain by selecting the most effective data processing algorithms and procedures for a specific set of data or a problem-oriented task.

Up-to-date cognitive technologies offer a host of data-based mathematical models that help reduce the design time and enable mass computations. The models are built upon the results of full-scale and/or computational experiments with various objects from the investigated class, with a minimum of knowledge from the subject field, i.e. process physics. In other words, the models are trained on a set of input and output data prototypes and are commonly termed "surrogate models" [1], [2].

The data-based models by nature are certain to work with the input data similar to the set of prototypes the model is based upon. To be applied to another set of prototypes, the model has to be either built anew or re-constructed by solving a statistical task in which the values of one model should be predicted based on the values of another model, with the input data being the same. This prompts an acute need for tools that would provide fully or semi-automatic generation of data-based models.

The paper describes a technique used for constructing automatic procedure generators that create software modules implementing basic data analysis and processing procedures. The generators are data-specific and can be used as external libraries in the development of specialized engineering applications.

2 Requirements to the procedure generator

Each data processing procedure is centered around the implementation(s) of one or several typical mathematical tasks that an expert in mathematical data analysis methods solves "manually" using standard multi-purpose publicly available software packages, such as MATLAB, MATHEMATICA, STATISTICA, etc. A task has to be solved anew each time the input data set is modified.

The choice of the solution technique is governed by the type of the targeted data structure. For example, the principal component analysis (PCA) is effective for linear data structures only. Some subject fields may already have their own specific dimension reduction procedures devised with regard for the 'physical' aspects of the data (parametric models [5]).

A mathematician familiar with various mathematical solution techniques and aware of the impact of a data structure on the quality of a procedure can manually select or construct an effective dimension reduction procedure by analyzing the data structure, combining various algorithms and making comparisons by conducting computational experiments.

Project engineering teams hardly ever have a mathematician on the staff and are unable or unwilling to hire an external expert, confidentiality of the data being one of the reasons.

To summarize, the procedure generator must meet a set of requirements.

Input data for the generator come in the form of a data set supplied by a domain expert and used as a learning data set when creating a new procedure.

A procedure is constructed using the known classes of functions and/or models that may include both universal models indifferent to the data specifics and domain-specific models taking into account the "physical" context of the data.

The classes of models should allow the data processing models to be used one after another. This means that all the models should conform to the same specification of inputs and outputs.

Each of the models can have its own set of parameters. The user should be able to specify a value for each parameter to make the generation process settings more accurate.

The generator's output is a software module that can be used as a component in the development of specialized engineering applications. The software module is implemented as a dynamic link library (DLL) that can be connected to the application and activated without prior assembly or compilation of the application.

3 An example of the automatic procedure generator implementation

The automatic procedure generator was developed and is successfully used for reducing the dimension of an object's digital description.

The generator implements a solution for the following mathematical task: use the given data set consisting of N n-dimensional vectors $\{X_1, X_2, \ldots, X_N\}$ to select dimension $m \leq n$ of the compressed vector and construct two procedures:

- a compression procedure that converts an n-dimensional vector X to a compressed m-dimensional vector $\lambda = (X)$;
- a reconstruction procedure that converts the compressed m-dimensional vector λ to the decompressed n-dimensional vector $X^* = R(\lambda)$.

The quality of this pair of procedures applied to the input vector X is predicated upon the reconstruction error measured as the distance $d(X, X^*) = |X - X^*|$ between the input vector X and the reconstructed vector $X^* = R(C(X))$ obtained by successive application of the compression and decompression procedures.

The quality of the dimension reduction procedure R applied to a specified data set is quantified by the root-mean-square decompression error.

The generator implements two independent solutions for the above task:

- minimizing dimension m of the compressed vector while staying at the specified root-mean-square error level ϵ;
- minimizing the root-mean-square reconstruction error for the specified dimension m of the compressed vector.

The generator core is composed of the following classes of algorithms:

- linear dimension reduction algorithms based on the artificial neural network (PCA) technology;
- non-linear dimension reduction algorithms based on the ANN technology;
- specific parametric models designed to provide analytical descriptions of the airfoil structure.

The automatically generated procedures is constructed as a combination of algorithms from all given classes. On generating process both the parameters of each algorithm and parameters of algorithm combination were selected to minimize the distance $d(X, X^*)$.

The automatically generated procedures were compared to standard ANN- and PCA-based dimension reduction procedures. The comparison results are summarized in the table below.

Dimension of compressed vector	ANN	PCA	Automatically generated procedure
$m = 2$	4,11E-03	4,32E-03	3,06E-03
$m = 4$	1,96E-03	2,09E-03	1,89E-03
$m = 6$	1,36E-03	1,19E-03	1,10E-03
$m = 8$	9,65E-04	7,74E-04	6,63E-04
$m = 10$	7,15E-04	4,70E-04	4,64E-04

Conclusion: the automatically generated procedures display uniformly lower mean reconstruction errors as compared to the standard ANN and PCA-based procedures.

References

[1] Kuleshov A. (2008). Cognitive technologies in adaptive models of complex objects. *Information technologies and computer systems.* Vol. **1**, pp. 18-29.

[2] Bernstein A., Kuleshov A. (2008). Mathematical Methods in Engineering Cognitive Technologies. *Survey in Applied and Industrial Mathematics, Ser. "Probability and Statistics"*. Vol. **15**, No. **3**, pp. 451-452.

[3] Bernstein A., Burnaev E., Dorofeev E. Sviridenko Y. Chernova S. (2008). Cascade dimension reduction procedures. *Proceedings of the 11th National Conference with International Participation on Artificial Intelligence, Dubna.* Vol. **1**, pp. 241-250.

[4] Bernstein A., Kuleshov A. (2008). Cognitive technologies in the problem of dimension reduction of geometrical object descriptions. *Information technologies and computer systems.* Vol. **2**, pp. 6-19.

[5] Ivanova E., Chernova S. (2009). Reduction of complex geometrical object dimension in the presence of particular parametric models. *Artificial Intelligence and Decision Making.* Vol. **3**, pp. 53-58.