Data Mining for Decision Support in Multiple-Model System Identification

SANDRO SAITTA*
Ecole Pol. Féd. de Lausanne
Appl. Comp. and Mech. Lab
Station 18, 1015 Lausanne
SWITZERLAND

BENNY RAPHAEL
National University of Singapore
Department of Building
117566, Singapore
SINGAPORE

IAN F. C. SMITH
Ecole Pol. Féd. de Lausanne
Appl. Comp. and Mech. Lab
Station 18, 1015 Lausanne
SWITZERLAND

Abstract: Data mining techniques presented in the literature are usually used for prediction and they are tested on well known benchmark problems. System identification is a practical engineering problem and an abductive task which is affected by several kinds of modeling assumptions and measurement errors. Therefore, system identification is supported by multiple-model reasoning strategies. The objective of this work is to study the use of data mining techniques for system identification. One goal is to improve views of model-space topologies. The presence of clusters of models having the same characteristics, thereby defining model classes, is an example of useful topological information. Distance metrics add knowledge related to cluster dissimilarity. Engineers are thus better able to improve decision making for system identification.

Key–Words: knowledge extraction, data mining, clustering, system identification.

1 Introduction

Data Mining is useful in areas such as text categorization, speech processing, image recognition and gene classification among many others (see [8] for more areas). As stated in [3], the amount of data worldwide increases at twice the rate of Moore's Law. This is especially true in structural engineering, where the use of sensors has increased drastically in recent years. An overabundance of data can overwhelm engineers if data are not processed systematically.

Use of data mining in engineering is not new [6] [1]. Examples of applications include oil production prediction, traffic pattern recognition, composite joint behavior and joint damage assessment. However, all of these contributions use data mining as a predictive tool. There are engineering tasks in which it is more appropriate to use data mining as a descriptive tool, through obtaining a description of important characteristics of data. System identification is an example of this type of task. In system identification [5], the goal is to determine the state of a system including values of system parameters through comparisons of predicted behavior with measurements. Since many causes (models) might lead to the same consequences (sensor data), unique identification is rarely possible in most cases of complex structures. In previous work [10], a system identification methodology that generates several candidate models has been developed.

To help engineers in the system identification task, data mining techniques can be used to extract knowledge from candidate models. An example of useful information is relationships between parameters of these models. The number of classes of candidate models is also an important piece of information, since this indicates whether system identification is unique. For these assessments, techniques such as principal component analysis (PCA) [4] and k-means [12] are useful. K-means clustering has been successfully applied in domains such as relational databases [7] and gene expression data [15]. Even though clustering has been proposed for various applications by the data mining community, its application is not straightforward; there are many open research issues.

Hybrid data mining methods are proposed in the literature, for example, [9] and [14]. Most work combines data mining methods for better prediction. For example, [2] proposed a combination of PCA and k-means to improve prediction accuracy of DNA gene expression and Internet newsgroups. Visualization improvement is not the objective of this research. Hybrid data mining methods that aims to generate better descriptions of spaces of models have not been found in the literature.

While there are well accepted methods such as

*Corresponding author: Sandro Saitta, Grad. Research Assistant in Computer Science, IMAC, Ecole Polytechnique Federale de Lausanne, 1015 Lausanne, Switzerland.

cross-validation [13] for evaluating predictive models, quantitative methods are not available for evaluating the clustering of candidate models. The criterion for assessing the capability of clustering algorithms is subjective and dependent on the final goal of the knowledge discovery task.

In this paper, a combination of two data mining techniques is proposed to extract knowledge from models. PCA and k-means clustering are used to facilitate better understanding of the space of candidate models. An important objective is to obtain a view of model-space topologies. The main focus of this paper is on clustering model data, where three issues are addressed. Firstly, an evaluation of the quality of a set of clusters is performed specifically for the task of system identification. Secondly, the choice of the number of clusters is discussed. Finally, limitations of knowledge discovery are presented.

The paper is structured as follows. In Section 2, multiple-model system identification is presented. Section 3 explains how data mining techniques can be used to obtain useful knowledge for engineers. Section 4 contains the results of a case study and a discussion of limitations. The final section contains conclusions and a description of work in progress.

2 Multiple-Model System Identification

Traditionally, system identification is treated as an optimization problem in which the difference between model predictions and measurements is minimized. Values of model parameters for which model responses best match measured data are determined by this approach. However, this approach is not reliable because different types of modeling and measurement errors compensate each other such that the global minimum may be far away from the correct state of the system. Therefore, instead of optimizing one model, a set of candidate models is identified in our approach. These candidate models lie below a threshold which is computed using an estimate of the upper bound of errors due to modeling assumptions as well as measurements. Each model has a unique value for each model parameter.

An indication of the reliability of system identification is obtained through an examination of the characteristics of the population of candidate models. If model parameter values show wide variation, it means that either parameter values might have compensated for the effects of incorrect modeling assumptions or that the measurement system is inadequate. On the other hand, if solutions are located in a narrow, well-defined region of the search space, parameter estimates are likely to be accurate. Several distinct regions containing candidate models indicate the presence of multiple local minima in the objective function. These have been observed in our experiments.

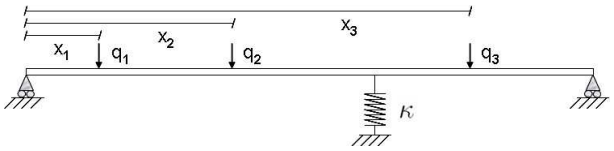

Figure 1: Beam used for the case study. Seven parameters are used.

Our approach is illustrated using a case study of a two-span beam (Figure 1). The structure is two meters long and its middle support is a spring. Using the methodology described in [11], several models are generated. There are seven parameters. They consist of three loads (position x_i on the beam and magnitude q_i) as well as the stiffness κ of the central spring. According to the error threshold discussed in [11], 500 models are identified. These models are used in this paper and their parameter values are considered as input points for data mining techniques.

Data mining techniques [12, 13] are applied to the data set containing model parameters in order to obtain useful knowledge for system identification tasks. In general, the objective is to determine the accuracy of diagnoses.

3 Mining Model Data

3.1 PCA

PCA is a linear method for dimensionality reduction [4, 12]. Ultimately, PCA finds a set of principal components (PC) that are sorted such that the first components explain most of the variability of the data. In the machine learning community, PCA is usually used as a preprocessing technique, for example before a supervised algorithm. Since the aim of this study is not predictive, PCA is used for knowledge extraction.

The main objective is to extract linear relationships among more than two parameters. It is also possible to obtain a general idea of relative importance of parameters. This is reflected in the coefficients of principal components. The coefficients of some parameters may always be zero. This means that these parameters have no importance in explaining the variability of the data. Therefore, they denote reliable parameters for system identification. Finally, clusters are found while visualizing the data using the first principal components. Even though PCA is not meant for clustering, it can be used to improve the clustering

Clustering procedure
1. Transform the data using PCA.
2. Choose the number *K* of clusters (Table 2).
3. **Loop** *i* from 1 to *N*
4. Run k-means with K clusters.
5. Calculate score function (SF).
6. **End**
7. Select clustering *i* with maximum SF

Table 1: Pseudo-code algorithm combining PCA and k-means to separate models into classes.

process as explained in the following Section.

3.2 Clustering

3.2.1 Feature Space Clustering (FSC)

K-means [12] is a widely applied clustering algorithm. Although it is simple to understand and implement, it is effective only if applied and interpreted correctly. The k-means algorithm divides the data into *K* clusters according to a given distance measure. Although the Euclidean distance is usually chosen, other metrics may be more appropriate.

The proposed methodology - combining PCA and k-means - is described next. First, the PCA procedure is applied to the models. Using the principal components the complete set of model predictions is mapped into the new feature space. Then, the k-means algorithm is applied to the data in the feature space. The final objective is to see if it is possible to separate models into clusters and to present results to the engineer in an understandable way. Table 1 presents the pseudo-code of the methodology used.

In addition to the limitations mentioned in [12], this methodology has two drawbacks. Firstly, the number of clusters has to be specified by the user a-priori. Strategies for estimating the number of clusters have been proposed in [12, 13]. One of these method is chosen here and adapted to the system identification context. Secondly, as stated above, the *K* initial centroids are chosen randomly. Therefore, running *P* times will result in *P* different clustering of the same data. A strategy for avoiding such a problem is described next.

3.2.2 Evaluation and Significance of FSC

The number of clusters of models is useful information for engineers performing system identification. When the methodology defined in [11] outputs M possible models, it does not mean that there are M different models of the structure. These M models might only differ slightly in a few values of parameters while representing the same model. In other situations, models might have important differences representing distinct classes which are referred to as clusters.

When predictive performances are evaluated, the classification error rate is usually used. If the aim is to make predictions on unseen data sets, the most common way to judge the results is through cross-validation [13]. In this work, the evaluation process is different since the goal is not prediction. Results are evaluated in two ways. Firstly, a criterion is used to evaluate the performance of the clustering procedure. Secondly, from a decision support point of view, the performance is evaluated by users.

The main theme in this Section is to develop a metric in order to evaluate results obtained by the proposed approach. Without a metric, the way clusters are seen and evaluated is subjective. Furthermore, it is not possible to know the real number of cluster in the data since the task is unsupervised learning and this means that the answer - the number of clusters - is unknown. In this paper, the results obtained by the clustering technique are evaluated using a score function (*SF*). The score function combines two aspects: the compactness of clusters and the distance between clusters. The first notion is referred to as within class distance (wcd) whereas the second is the between class distance (bcd). In this research the wcd and the bcd are defined respectively in Equation 1 and 2. From a clustering viewpoint, the objectives are to minimize the first aspect and to maximize the second, i.e. to maximize the SF of Equation 3.

$$wcd = \frac{\sum_{i=1}^{K} \left(\sum_{x \in C_i} dist(c_i, x) \right)}{\sum_{i=1}^{K} size(C_i)} \tag{1}$$

$$bcd = \frac{\sum_{i=1}^{K} dist(c_i, c_{tot}) \cdot size(C_i)}{\sum_{i=1}^{K} size(C_i) \cdot K} \tag{2}$$

$$SF = \frac{bcd}{wcd} \tag{3}$$

where K is the number of clusters, C_i the cluster i, c_i its centroid and c_{tot} the centroid of all the points. The functions $dist$ and $size$ define respectively the Euclidean distance between two points (each point is a model which is represented by parameter values) and the number of points in a cluster. From a system identification point of view, bcd values indicate how different the K situations are. Values of wcd give overviews of sizes of groups of models.

Controlling Randomness
1. **Loop** *i* from 1 to *N*
2. **Loop** *j* from 1 to *P*
3. Run k-means with *j* clusters.
4. Calculate score function (SF).
5. **End**
6. **End**
7. Select P corresponding to maximum SF.

Table 2: Procedure to limit the effect of the random choice of the starting centroids when determining the number of clusters.

Clusters	**bcd**	**wcd**	**SF**
2	0.69	1.91	0.36
3	0.57	1.54	0.37
4	0.48	1.24	**0.39**
5	0.41	1.09	0.37
6	0.35	1.03	0.34
7	0.30	0.98	0.31
8	0.27	0.93	0.29

Table 3: Comparison of values for between class distance (bcd), within class distance (wcd) and score function (SF) for various numbers of clusters. The randomness is controlled by $N = 100$.

It is important that an engineering meaning in terms of model-based diagnosis can be given to these two distances. They are both related to the space of models for the task of system identification using multiple models. The wcd represents the spread of models within one cluster. Since it gives information on the size of the cluster, a high wcd means that models inside the class are widely spread and that the cluster may not reflect physical similarity. The bcd is an estimate of the mean distance between the centers of all clusters and therefore, it provides information related to the spread of clusters. For example, a high bcd value means that classes are far from each other and that the system identification is not reliable.

As explained in Section 3.2.1, the number of clusters (in system identification, the number of classes of models) for a data set is unknown. The procedure to determine the best number of clusters is to run the procedure for P different numbers of clusters. The criterion used to check if the number of cluster is appropriate is the score function of Equation 3. The higher the value of the SF, the more suitable the number of clusters. In Section 3.2.1, one of the weakness of the procedure highlighted was the random choice of the K first centroids. One solution is to run the algorithm *N* times and to select the maximum value for the score function. Therefore, randomness is controlled by *N*. The pseudo-code of the mentioned procedure is given in Table 2.

To conclude this section, the score function defined above serves two purposes. First, it gives an idea of the performance of the clustering procedure. Second, it allows choice of a realistic value for the number of clusters. However, this number must be interpreted with care as explained in Section 4. Reducing the random effect of the procedure is done through several runs of the algorithm to compute the score function value. Finally, the number of clusters could be fixed by the expert and therefore this may be considered to be domain knowledge.

4 Results and Limitations

The case study used to illustrate the proposed data mining approaches is explained in Section 2. In this study, PCA is used to discover independant parameters and to understand to what extend are model parameters linearly related. Applying PCA on the models, a set of PC are obtained. The focus is on studying these PC instead of the models in the feature space.

The main conclusion here concerns the relationships between parameters. Parameters are related since a few PC explains nearly all the variability in the data. However, PCA is able to discover only linear relationships. The only conclusions that can be made when using PCA are related to the presence of linearity in the data.

Results from the feature space clustering procedure of Section 3.2 are now presented. The first step is to run the procedure described in Table 2. Output is shown in Table 3. According to this table, four clusters are most appropriate for this data since this corresponds to the maximum value for the score function. This result can also be seen in Figure 2, representing the value of the SF evolving with the number of clusters. In this case, there is only a global maximum. However, in certain cases, local maxima can appear. The engineer is thus able to choose the most reliable number of clusters among the local maxima.

Once the number of clusters is fixed, the procedure outlined in Table 1 is followed. To judge the improvement of the methodology with respect to the standard k-means algorithm, the two techniques are compared. Figure 3 shows the improvement from a visualization point of view. The top part of Figure 3 corresponds to standard k-means. The bottom part is the result of the methodology described in this paper. It is evident that our methodology is better able to present results visually to engineers.

This methodology has a number of limitations.

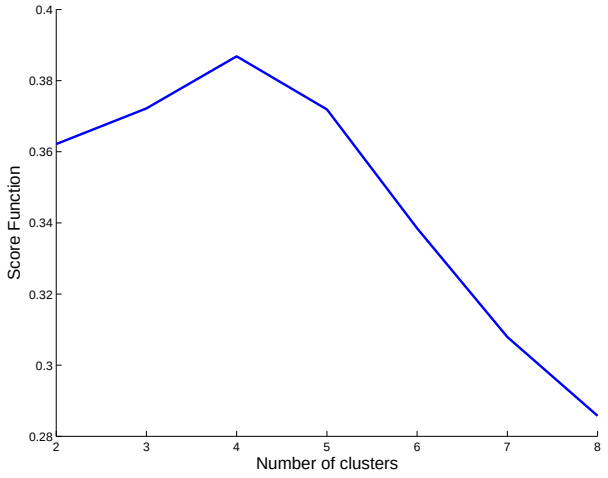

Figure 2: Evolution of the SF with regards to the number of clusters.

Firstly, results of data mining have to be interpreted carefully. The user thus has an important role in ensuring that the methodology is successful. Secondly, even if the methodology is well applied, results are not necessarily the most appropriate. For example, data might be noisy (poor sensor characteristics), or may have missing values (low sensor quality) or may be missing useful information (bad sensor configuration) and this may preclude obtaining useful results.

An example of challenges associated with applying data mining to system identification is given below. Assume that, after applying data mining methodology, three clusters of models are obtained. The methodology alone is not able to *interpret* these clusters. Suppose that two clusters group similar information. Although the clustering algorithm has generated three clusters, only the user is able to identify that there are only two clusters that have physical meaning. Therefore, data mining is only able to suggest possible additional knowledge. The process of acquiring knowledge that is of practical use for decision support is left for the engineer.

5 Conclusions

A combination of data mining techniques has been proposed for system identification tasks. Relationships between model parameters and clusters of models are examples of useful information for engineers. In order to evaluate clustering, a score function has been developed which is adapted to the specific task of system identification. Main conclusions are listed below:

- Standard data mining techniques such as PCA have to be adapted to obtain meaningful results in the area of system identification

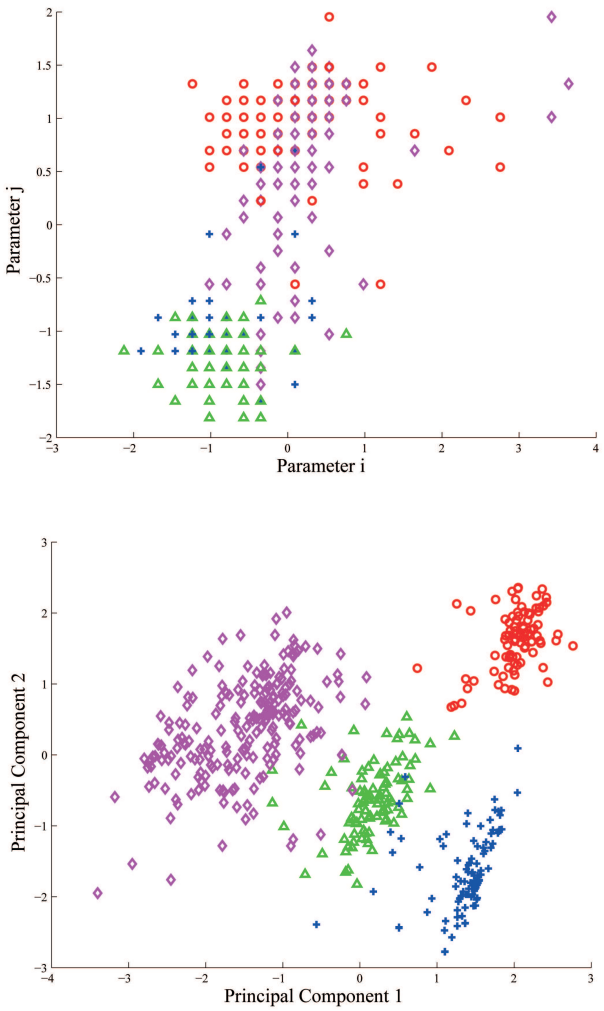

Figure 3: Comparison between standard k-means (top) and the proposed algorithm (bottom). In the first case, two parameters are chosen to display the clusters. In the second, the two first principal components are used.

- Combining data mining techniques, such as PCA and k-means, helps improve visualization of data

- Evaluation of results obtained through clustering is difficult. The score function that has been developed in this work helps in the evaluation

- Application of data mining to complex tasks such as system identification requires considerable expertise

Future work involves the use of other data mining techniques to discover hidden relationships among model parameters. Strategies for models containing a

varying number of parameters are also under development. Finally, a general framework integrating data mining techniques in the overall system identification process will be developed.

Acknowledgements: This research is funded by the Swiss National Science Foundation through grant no 200020-109257. The authors recognize Dr. Fleuret for fruitful discussions on data mining techniques and Dr. Kripakaran for his help concerning the case study.

References:

[1] C. Alonso, J.J. Rodriguez, and B. Pulido. Enhancing consistency based diagnosis with machine learning techniques. *Lecture Notes in Computer Science*, 3040:312–321, 2004.

[2] C. Ding and X. He. K-means clustering via principal component analysis. In *Proceedings of the* 21^{st} *International Conference on Machine Learning*, ACM International Conference Proceeding Series, page 29. ACM Press, 2004.

[3] U. Fayyad and R. Uthurusamy. Evolving data mining solutions for insights. *Communications of the ACM*, 45(8):28–31, 2002.

[4] I.T. Jolliffe. *Principal Component Analysis*. Springer, 2002.

[5] L. Ljung. *System Identification - Theory For the User*. Prentice Hall, 1999.

[6] H.G. Melhem and Y. Cheng. Prediction of remaining service life of bridge decks using machine learning. *Journal of Computing in Civil Engineering*, 17(1):1–9, 2003.

[7] C. Ordonez. Integrating k-means clustering with a relational dbms using sql. *IEEE Transactions on Knowledge and Data Engineering*, 18(2):188–201, 2006.

[8] S.K. Pal and P. Mitra. *Pattern Recognition Algorithms for Data Mining*. CRC Press, 2004.

[9] X. Pan, X. Ye, and S. Zhang. A hybrid method for robust car plate character recognition. *Engineering Applications of Artificial Intelligence*, 18(8):963–972, 2005.

[10] Y. Robert-Nicoud, B. Raphael, and Ian Flemming Campbell Smith. System identification through model composition and stochastic search. *Journal of Computing in Civil Engineering*, 19(3):239–247, 2005.

[11] S. Saitta, B. Raphael, and Ian Flemming Campbell Smith. Data mining techniques for improving the reliability of system identification. *Advanced Engineering Informatics*, 19(4):289–298, 2005.

[12] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison Wesley, 2006.

[13] A. Webb. *Statistical Pattern Recognition*. Wiley, 2002.

[14] L.J. Xu, Y. Yan, S. Cornwell, and G. Riley. Online fuel tracking by combining principal component analysis and neural network techniques. *IEEE Transactions on Instrumentation and Measurement*, 54(4):1640–1645, 2005.

[15] K.R. Zalik. Biclustering of gene expression data. In *Proceedings of the 5th WSEAS Int. Conf. on Simulation, Modelling and Optimization*, pages 225–230, 2005.