2009 Ninth International Conference on Intelligent Systems Design and Applications

A Method to Point Out Anomalous Input-Output Patterns in a Database for Training Neuro-Fuzzy System with a Supervised Learning Rule

Valentina Colla, Nicola Matarese
Scuola Superiore Sant'Anna, SSSA
Pisa, Italy
Email: colla {n.matarese}@sssup.it

Leonardo M. Reyneri
Politecnico di Torino
Torino, Italy
Email: leonardo.reyneri@polito.it

***Abstract*—When designing a neural or fuzzy system, a careful preprocessing of the database is of utmost importance in order to produce a trustable system. In function approximation applications, when a functional relationship between input and output variables is supposed to exist, the presence of data where the similar set of input variables is associated to very different values of the output is not always beneficial for the final system to design. A method is presented which can be used to detect anomalous data, namely non-coherent associations between input and output patterns. This technique, by mean of a comparison between two distance matrix associated to the input and output patterns, is able to detect elements in a dataset, where similar values of input variables are associated to quite different output values. A numerical example and a more complex application in the pre-processing of data coming from an industrial database were presented.**

***Keywords*-Data preprocessing; Filtering technique; Distance evaluation;**

I. Introduction

A neuro-fuzzy system [1] is often designed to perform a sort of *functional association* between input and output patterns: not only in function approximation problems, but also in classification [2] and clustering operation [3], the neuro-fuzzy system is required to associate each input pattern to a desired output value or vector, independently on the meaning that is attribued to such output. In order to make this operation successful and meaningful, the designer must be sure that the same functional relationship is also somehow represented in the available data, i.e. one must be sure that similar values of the input variables correspond to quite similar output values.

In such case it is well known that a neuro-fuzzy system performs a sort of *averaging operation*, i.e. at the end of the training phase the estimated output will be more similar to the output value that is most often associated with a given input. This is a correct and robust behaviour, that, however, does not always lead to satisfactory results, especially when some *a priori* knowledge is available on the desired system behaviour. Thus it is mandatory to exploit such behaviour in the very preliminary phase of data preprocessing, by pointing out the patterns where similar input values are associated with very different outputs, and by eliminating the input-output couples which are considered less realistic (for instance with the support of the technical personnel working in the plant where the database comes from).

It must be underlined that in some cases the fact that in an experimental database similar input sets are associated with very different output patterns is normal, such as, for instance, when one deals with instable systems or when the target value depends on parameters or state variables which are not fed as input to the model. In such cases, the proposed method is not applicable, while it fits well to all the situations when the input-output relationship is "one to one" or "many to one."

The method proposed in this paper is based on *dissimilarity matrices*, which are utilized in more application fields, with the aim of recognizing interesting patterns among the collected data [4][5][6].

An alternative approach to the discussed problem could rely on the automatic estimation of reliability of the model output that has been proposed in [7], as this reliability is a direct consequence of the data quality, therefore the areas of the input space where the reliability is low are those where the above described "incoherent" input-output patterns most probably lie. However the method that is proposed here directly points out such set of data and is therefore simpler to exploit in an autonomous way by end-user, who are not often very experienced in the model development but have a deep knowledge of the field where the data are collected.

The paper is organized as follows: Sec. II describes the implementation of the proposed method and one alternative approach. Sec. III depicts an example of application of both proposed approach, that helps the reader to understand how the proposed techniques actually works, while Sec. IV describes a practical application to the preprocessing of an industrial database. Finally Sec. V provides some concluding remarks.

II. Proposed methods

Given a problem where N inputs are used to predict a single output value, let us consider a dataset that has been prepared to train, for instance, a feedforward neural network or any other kind of neuro-fuzzy system through

978-0-7695-3872-3/09 $26.00 © 2009 IEEE
DOI 10.1109/ISDA.2009.202

IEEE computer society

a supervised learning algorithm. The database contains P input-output couples $(\mathbf{X^P}, \mathbf{t^P})$, where $\mathbf{X^P}$ is a column vector with N entries and t^p is the scalar value to predict. By representing the whole database in the matrix formulation, the input matrix $\mathbf{X}$ has P columns and N rows, while the target values are stored in a column vector $\mathbf{T}$ with P entries.

The anomalous pattern we are looking for are those columns in $\mathbf{X}$ which are similar to each other, while the corresponding entries in $\mathbf{T}$ are very different. The final aim of the algorithm is just to point out these patterns each of one composed by two elements i.e. the incoherent rows of the matrix. More explicitly, if one of the two patterns that have been detected as mutually incoherent is present also in another (or more than one) couple of incoherent data, a group of pattern is formed where only one pattern is anomalous with respect to the other ones and this pattern is automatically eliminated from the database. Otherwise, i.e. if both the incoherent patterns are not present in any other anomalous couple, the advice of the expert of the field where the data have been originated is fundamental in order to take the decision of eliminating one, both or none of them. The criteria to take this decision depend on the application and are out of the scope of the present work.

A. Finding similarities

For the proposed method, the steps required by the algorithm are the follows:

1) Static normalization of the entries of $\mathbf{X}$ and $\mathbf{T}$ with respect to their own maximum values. Two tranformed matrices π and τ are obtained, where
$$\begin{aligned} \pi_i^p &= \frac{x_i^p}{max_i\{x_i^p\}} \\ \tau_i^p &= \frac{t_i^p}{max_i\{t_i^p\}} \end{aligned} \tag{1}$$
2) Computation of the *euclidean distance* between each pair of columns of π and between the entries of τ, by obtaining two symmetrical square *distance matrices* related to the input and output patterns, named, respectively:
$$\begin{aligned} \mathbf{D_{I_{p,q}}} &= \sum_i (\pi_i^p - \pi_i^q)^2 \\ \mathbf{D_{O_{p,q}}} &= \sum_i (\tau_i^p - \tau_i^q)^2 \end{aligned} \tag{2}$$
whose entries located on the main diagonal are null. Due to the simmetry, the following step 3 can be performed by only considering the matrix upper triangle. The obtained matrices can also be referred as *dissimilarity matrices*.
3) A final comparison of the corresponding elements of the two normalised dissimilarity matrices, by considering that anomalies are pointed out when a "small" entry of $\mathbf{D_I}$ (namely, below a threshold λ_I) corresponds to a "large" entry of $\mathbf{D_O}$ (namely, above a threshold λ_O).

The definition of "small" and "large" is given through a comparison with some threshold values that should be independent on the number of patterns, and possibly on the particular dataset.

It is clear that the smaller λ_I and the larger λ_O, the smaller will be the fraction of patters identified as anomalous, therefore the optimal value shall be found both with some objective technique (as discussed further) and empirically, by running the method with different values of the threshold until a reasonable amount of patters is identified.

B. Finding the threshold

Firstly define the *proportionality factors* k_I and k_O: the value of the threshold λ_I on the entries of $\mathbf{D_I}$ is determined by considering a value of proportionality factor k_I as 0.15. On the other hand, the value of the threshold λ_O on the entries of $\mathbf{D_O}$ is determined by considering the fraction k_O of the overall range $[t^p_{min}, t^p_{max}]$ of the target variable, that can be considered a sensible difference between two output values. A reasonable value for k_O can lie in the range $[0.25, 0.5]$, but it can also be fixed depending on the application by exploiting the experience of the technicians that work in the practical context from which the considered database is obtained.

There are now several ways to determine reasonable values for the two thresholds λ_I and λ_O:

- *fixed values*, that is, two values k_I and k_O defined by the user according to its knowledge of the problem. This is the simplest, although most critical, method
$$\begin{aligned} \lambda_I &= k_I \\ \lambda_O &= k_O \end{aligned} \tag{3}$$
- *averaged*, that is, having them proportional to the average of the entries in the corresponding matrix:
$$\begin{aligned} \lambda_I &= k_I \frac{\sum_{p,q} \mathbf{D_{I_{p,q}}}}{P^2} \\ \lambda_O &= k_O \frac{\sum_{p,q} \mathbf{D_{O_{p,q}}}}{P^2} \end{aligned} \tag{4}$$
by properly choosing the proportionality factors as, for instance $k_I = 0.15$ and $k_O = 0.33$. The advantage of this method is that the threshold depends on the average distance between the patterns, therefore it depends on how spreaded are the patterns in the input (respectively, output) space, therefore within a set of very similar patterns (that is, with a low average distance), an anomalous pattern need not be far away from each other, while in a set of different patterns (that is, with a larger average distance), an anomalous pattern should be more separated from the others;

- *normalization*, that is, having them proportional to the largest entry in the corresponding matrix:

$$\lambda_I = k_I \max_{p,q} \mathbf{D_{I_{p,q}}}$$
$$\lambda_O = k_O \max_{p,q} \mathbf{D_{O_{p,q}}} \quad (5)$$

  The major drawback of this method is that the threshold depends on the distance between the two most different patterns, which has no relationship with the problem of finding the anomalous patterns;
- *relative values*, that is, two values defined by the user according to the largest possible entries of $\mathbf{D_I}$ and $\mathbf{D_O}$:

$$\lambda_I = k_I \sqrt{N}$$
$$\lambda_O = k_O \sqrt{M} \quad (6)$$

  where N and M are the dimensions of the input and output vectors, respectively, while $\sqrt{N}$ and $\sqrt{M}$ are the largest possible Euclidean distances between two normalized input and output vectors. This has the advantage, with respect to the previous one, to be indpendents of the two most different patterns.

C. An alternative method

An alternative method to the one proposed in the previous sections is based on finding a "metric" which identifies the entries of the dissimilarity matrices which are lowest in $\mathbf{D_I}$ and largest in $\mathbf{D_O}$ at the same time, without thresholding the values.

This can be computed by computing an additional matrix $\mathbf{M}$ whose elements

$$m_{p,q} = \exp(-k \cdot \mathbf{D_{I_{p,q}}}^2) \cdot \mathbf{D_{O_{p,q}}}$$

are larger if the corresponding element of $\mathbf{D_I}$ are smaller and the corresponding element of $\mathbf{D_O}$ are larger. The largest entries in $\mathbf{M}$ will therefore correspond to potential anomalous patterns. In the above formula, k is a constant which can be used to optimize the performance of the proposed method.

The advantage of this method is that there is no threshold involved and the user can choose as many patterns as he likes, by properly choosing an appropriate number of largest entries, without the need of find the thresholds. In addition, this second method shows the considerable advantage that it does not require any time consuming comparison between two matices, but only the search for the most significant entries of the matrix M.

III. A simple example

Given a problem where 4 inputs are used to predict a single output value. Let us consider a dataset organised in rows, where the columns of X matrix correspond to the inputs and the column vector T contains the associated output value:

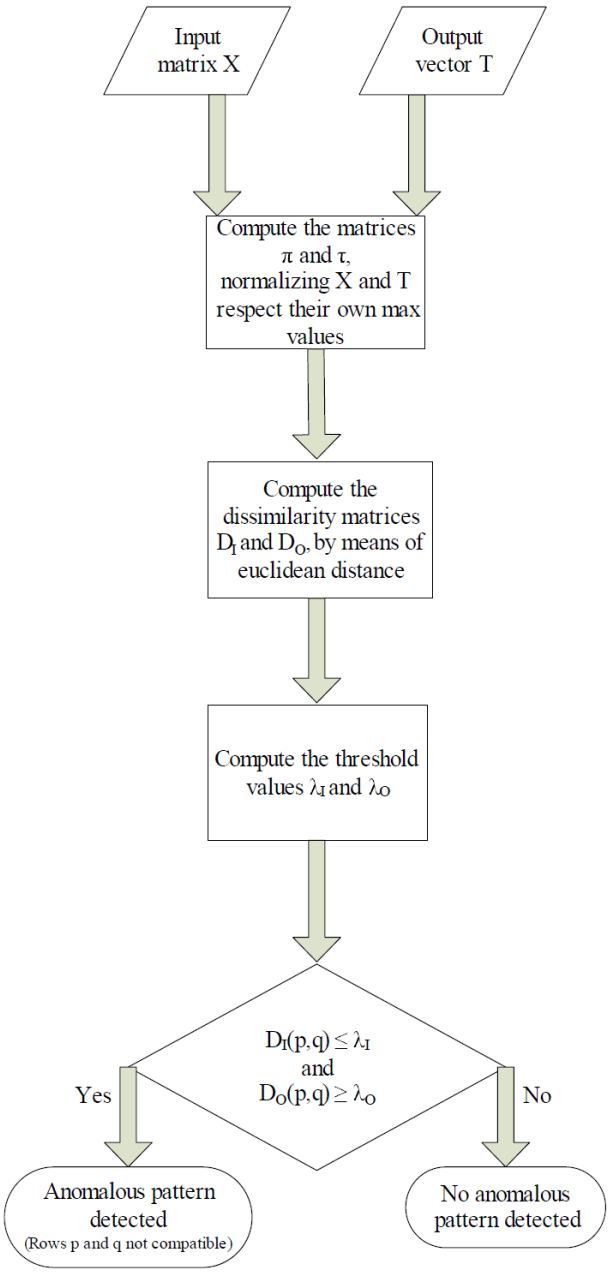

Figure 1: Flow chart of the proposed method

$$\mathbf{X} = \begin{bmatrix} 5 & 8 & 10 & 20 \\ 12 & 3 & 7 & 8 \\ 20 & 22 & 30 & 50 \\ 6 & 7 & 11 & 18 \\ 5 & 8 & 10 & 20 \end{bmatrix}$$

$$\mathbf{T} = \begin{bmatrix} 70 \\ 50 \\ 150 \\ 75 \\ 10 \end{bmatrix}$$

The normalized input and output matrices, that will be indicated as π and τ, respectively, are:

$$\pi = \begin{bmatrix} 0.25 & 0.3636 & 0.3333 & 0.4 \\ 0.6 & 0.1364 & 0.2333 & 0.16 \\ 1 & 1 & 1 & 1 \\ 0.3 & 0.3182 & 0.3667 & 1 \\ 0.25 & 0.3636 & 0.3333 & 0.4 \end{bmatrix}$$

$$\tau = \begin{bmatrix} 0.467 \\ 0.333 \\ 1 \\ 0.5 \\ 0.0667 \end{bmatrix}$$

The corresponding input and output dissimilarity matrices $\mathbf{D_I}$ and $\mathbf{D_O}$ are:

$$\mathbf{D_I} = \begin{bmatrix} 0 & 0.492 & 1.331 & 0.085 & 0 \\ & 0 & 1.48 & 0.425 & 0.492 \\ & & 0 & 1.329 & 1.331 \\ & & & 0 & 0.085 \\ & & & & 0 \end{bmatrix}$$

$$\mathbf{D_O} = \begin{bmatrix} 0 & 1.133 & 0.533 & 0.033 & 0.4 \\ & 0 & 0.667 & 1.167 & 0.267 \\ & & 0 & 0.5 & 0.933 \\ & & & 0 & 0.433 \\ & & & & 0 \end{bmatrix}$$

A. Effectiveness of proposed solution

After the computation of the two dissimilarity matrices $\mathbf{D_I}$ and $\mathbf{D_O}$, a necessary step was to define the thereshold values in order to identify the anomalous patterns. During the phase of threshold computing, the values adopted for the proportionality factors k_I and k_O are respectively 0.15 and 0.33, which means that a small entry in $\mathbf{D_I}$ is in contrast with a large entry in $\mathbf{D_O}$, when the difference of the output value is bigger than $1/3$ of the whole range of the output space.

In Tab.I the threshold values obtained with all the metodologie proposed in Sec.II-B were depicted.

By comparing the two matrices $\mathbf{D_I}$ and $\mathbf{D_O}$, considering all the threshold values depicted in Tab.I, it is evident that applaying any of the proposed threshold values, the null entry $(1, 5)$ of $\mathbf{D_I}$ corresponds to high entries in $\mathbf{D_O}$, by confirming the fact that pattern No. 1 is not "compatible" with pattern No. 5. Analogously the entries $(4, 5)$ of $\mathbf{D_I}$, if compared with very high corresponding entries of $\mathbf{D_O}$, show that pattern No. 5 is also incompatible with pattern No. 4. Considering that the fifth row emerges in both the detected registration (i.e. both pairs $(1, 5)$ and $(4, 5)$), it must be considered as a wrong line in the dataset that came from a wrong registration during the process of industrial data collection.

Table I: Threshold values

Threshold type	$\lambda_\mathbf{I}$	$\lambda_\mathbf{O}$
Fixed values	0.15	0.33
Averaged	0.085	0.1074
Normalization	0.2224	0.3080
Relative values	0.30	0.33

B. Effectivenes of alternative method

After the application of the metric proposed in par. II-C on two dissimilarity matrices $\mathbf{D_I}$ and $\mathbf{D_O}$, with a value of k parameter as 10, the results symmetric matrix $\mathbf{M}$ is depicted below.

$$\mathbf{M} = \begin{bmatrix} 0 & 0.011 & 0 & 0.031 & 0.400 \\ & 0 & 0 & 0.027 & 0.024 \\ & & 0 & 0 & 0 \\ & & & 0 & 0.403 \\ & & & & 0 \end{bmatrix}$$

The matrix $\mathbf{M}$ contains two elements that have highest values than the other ones; in detail the pair $(1, 5)$ and $(4, 5)$. These two pairs indicate that the rows 1 and 5 are not compatible between them, such as the couple of rows 4 and 5 of the pattern composed by input matrix X and output column vector T.

Also in this alternative approach, like that the method proposed in Sec.III-A, the obtained results confirm that the fifth row must be considered as incoherent registration, which must be deleted. The same results in both the porposed method confirming the effectiveness of both proposed approaches.

IV. Application to an industrial database

An industrial database has been analysed, which refers to metal industry. In particular, in the considered application a system needs to be designed, which is capable to predict a particular property of a final product as a function of many variables referring to both the chemical composition of the raw material and the manufacturing process. However, several problems can occurr in the registration of some input variables as well as in the final measurement of the property to predict, mainly due to errors in raw material assignment or to sensor failures. The technical personnel working on the plant in most of the cases is capable to recognize the incoherent patterns and sometimes also to find out the error sources and possible causes, but, due to the high rithm of the production and to the fact that the database is collected in an automatic way, such analysis cannot be performed on

Table II: Results on industrial database

Number of patterns	1000
Incoherent patterns detected	15
Couples detected	10
Groups detected	5

line. On the other hand, due to the considerable number of data that are rapidely collected in the database, also the off-line analysis cannot be manually performed. In order to select only reliable data for the design of the system to develop, as well as to point out eventual systematic errors in the data registration, that might affect both the production control and the final system evaluation, a software needed to be developed which provides support to the technical personnel by pointing out incoherent input-output patterns. The input variables have been selected with the support of plant engineers and the procedure for such selection is out of the scope of the present paper, as well as the theoretical fundations for assessing the repeteability of the process, i.e. for justifying the assumption that a functional relationship holds between the selected input variables and the target property to predict.

In order to test the system performance, the procedure has been validated on a validation dataset, that was not used for the system development and contains data referring to products which were subjected to a particular quality control procedure. In this particular case, the reliability of the input-output patterns has been carefully checked and we know exactly which patterns are actually inconsistent with the rest of the database.

The database contains about 1000 patterns, and 15 sets of two or more incoherent input-ouput patterns are present. Such sets are not always constituted by couples of observations, in this case the $2/3$ of the total number are couples, while the remain $1/3$ are a group of incoherent patterns; in these groups there are several observations and in one of them an output value sensibly different from the other ones is present. The results obtained with both the proposed methods are summarized in Tab.II. Both the presented methods showed to be efficient in pointing out all the 15 incoherent patterns.

It must be underlined that in the proposed application, no further filtering stage was applied to the dataset as this was the specification for the systems. Obviously, some criteria for data selection could also be elaborated but they are not the object of the present investigation. Noticeably, when the group of incoherent observation is numerous and only one or a few of them show a target value very different from the other ones, the incoherent observations could also be inferred in an automatic way, while when only two observations are pointed out, the opinion of the expert is fundamental in order to chose the correct pattern. In case of lack of this information, both the observations would need to be discarded.

V. Conclusion

In this paper a novel metodology and an alternative approach that are able to detect anomalous data registrations (non-coherent association) between input and output patterns were presented. Considering data, that came from real industrial line, the proposed techniques are able to find in the industrial dataset, anomalous associations between input and output patterns, that in most case are possible consequence of wrong data registration. In the case of neural networks models, the goodness and robustness of the prediction is strictly related to the quality of the data that have been used for the training: if such data are strongly affected by noise and other kind of errors, obviously the model performance will be poor. Frequently in real-world applications the data quality is a function of the same input variables that are fed to the model itself. The step of identifying and removing the anomalous data registrations is a necessary requisite in the phase of the preparation of data for function approximation problems, classification and clustering, because these kind of operations associate input pattern to a desired output value. Therefore it is important to detect and delete the wrong registration, in the preprocessing phase, in order to make previously the exposed operation successful and meaningful.

References

[1] Jyh-Shing Roger Jang, and Cheuen-Tsai, "Neuro-Fuzzy Modeling and Control", Proceedings of the IEEE, IEEE, 1995, Vol.83 No.8

[2] Chuen-Tsai Sun, and Jyh-Shing Jang, "A Neuro-Fuzzy Classifier and Its Applications", Second IEEE International Conference on Fuzzy Systems, IEEE, 1993, Page(s):94 - 98 Vol.1

[3] R. Poluzzi, A. Savi, D. Vago, and G. Martina, "Neuro-fuzzy clustering techniques for complex acoustic scenarios", Neural Comput. & Applic, Springer, 2003, 12: 160165

[4] E. Pekalska, and R. P. W. Duin, "The Dissimilarity Representation for Pattern Recognition: Foundations And Applications", World Scientific Publishing Co. Pte. Ltd, 2005

[5] Y. Zhou, S. Yan, and T. S. Huang, "Detecting anomaly in videos from trajectory similarity analisys", Multimedia and Expo, 2007 IEEE International Conference on, 2007, Page(s): 1087-1090

[6] V. Cheng, Chun-Hung Li, J. T. Kwok, and Chi-Kwong Lic, "Dissimilarity learning for nominal data", Pattern Recognition Society, Elsevier, 2004, Page(s): 1471-1477

[7] L. M. Reyneri, V. Colla, M. Sgarbi, and M. Vannucci, "Self-Estimation of Data and Approximation Reliability Through Neural Networks", IWANN09, Springer, 2009