Viharos, Zs.J.; Monostori, L.; Vincze, T.; Training and application of artificial neural networks with incomplete data, *Lecture Notes of Artificial Intelligence,* LNAI 2358*, The Fifteenth International Conference on Industrial & Engineering Application of Artificial Intelligence & Expert Systems, Springer Computer Science, Springer-Verlag Heidelberg;* Cairns, Australia, 17-20 June, 2002, pp. 649-659.

Training and application of artificial neural networks with incomplete data

Zs. J. Viharos[1], L. Monostori[1], T. Vincze[2]

[1]Computer and Automation Research Institute, Hungarian Academy of Sciences, POB 63, H-1518, Budapest, Hungary
viharos@sztaki.hu
[1]Eötvös Loránd University, Egyetem Tér 1-3., H-1056, Budapest, Hungary

Abstract. The paper describes a novel approach for learning and applying artificial neural network (ANN) models based on incomplete data. A basic novelty in this approach is not to replace the missing part of incomplete data but to train and apply ANN-based models in a way that they should be able to handle such situations. The root of the idea is inherited form the authors' earlier research for finding an appropriate input-output configuration of ANN models [21]. The introduced concept shows that it is worth purposely impairing the data used for learning to prepare the ANN model for handling incomplete data efficiently. The applicability of the proposed solution is demonstrated by the results of experimental runs with both artificial and real data. New experiments refer to the modelling and monitoring of cutting processes. Keywords: Neural Networks, Machine Learning, Applications to Manufacturing

1. Introduction

Reliable process models are extremely important in different fields of computer integrated manufacturing [10]. Difficulties in modelling manufacturing processes are manifold: the great number of different machining operations, multidimensional, nonlinear, stochastic nature of machining, partially understood relations between parameters, lack of reliable data, missing parts of data sets, etc. A number of reasons back the required models: the design, optimisation, control and simulation of processes and the design of equipment [20]. Based on the knowledge applied, fundamental, heuristic and empirical models can be distinguished.

Model-based solutions are efficient techniques to make difficult problems more tractable. It can contribute to elaborating new algorithms, supporting decision makers, decreasing the risk in investments, and running the systems exposed to changes and disturbances more efficiently.

Learning denotes changes in the system adaptive in the sense that learning techniques enable the system to do the same or similar task more effectively next time [11]. Artificial neural networks (ANNs) are general, multivariable, nonlinear estimators. This soft computing technique can offer viable solutions especially for problems where abilities for real-time functioning, uncertainty handling, sensor integration, and learning are essential features [11]. Successful applications in manufacturing were reported on in the literature [12][17][23].

The paper illustrates a method and a related tool able to handle and apply incomplete data arising in the field of manufacturing. After describing some real-life cases and methods for handling data sets with missing parts, the developed algorithm will be detailed. Test results based on artificial data and an application in the field of production explains the behaviour and the advantages of this novel concept. Further research issues are also enumerated.

2. Missing data and their handling in different application fields

The problem of missing data arises in several fields of real-life applications. This section flashes some examples from several fields, together with the applied methods for handling missing data.

Valentini *et al.* used function-based interpolation to determine a value for missing data in their study for carbon dioxide exchange of a beech forest in Central Italy [19].

Analysing the atmosphere, Francis *et al.* faced incomplete data. They used ANNs for estimating a parameter one hour in advance but in 6.6% of the database the desired data were missing. Their solution is special from the point of view of applying the model used for the estimation and to exploit the specialities of time series to handle and to complete the missing data [3].

Different approaches were tested to replace missing data by Pesonen *et al.* while building up a neural network to study medical data received form University Hospital of Tampere and Savonlinna Central Hospital in Finland. About 20% of the patients' data were missing, mainly the parameter of leucocyte count. They compared four methods to replace the missing data: substituting means, random values, data based on the nearest neighbour and a neural network based substitution [16].

The markets of a commercial Bank in Latin America were the field of examination where data with missing parts were arising. The researchers applied clustering methods to build different market segments and they extended the incomplete part of the individual data vectors with the most probable values of the given segment [6].

An algorithm called Expectation Maximisation was elaborated by Ghahramani and Jordan to find appropriate values for the places of missing data. Their solution was interesting in the point of view that all of the dependencies among given and missing parameters and their distribution were taken into account in the replacement procedure [5].

An interesting solution can be found in the paper of Keeler *et al.* for handling missing data. A data pre-processing module extends the missing part of the input vector before conveying it to the ANN model and serves various parameters about this extension, too. The so-called decision processor receives these together with the model output and uses both of them to decide about the necessary changes in the system [7].

The problem of missing data is found in several fields of manufacturing, as pointed out by Gardner and Bicker in the case of semiconductor wafer manufacturing, as well [4].

3. A novel approach to handle missing data by neural networks

Some methods were listed in the previous paragraph to solve the problem of missing data. These methods try to complete the missing part of the data vectors in several different ways. Instead of completing data vectors another aspect could be to generate neural networks, which can handle incomplete data directly. This was one of the main ideas during the research reported on here.

3.1. Earlier results behind the research

The authors have presented some research results to improve the applications of neural networks in the industrial field [12][13][14][15][21][22]. The introduction of a concept basically different from the assignment-oriented application of neural networks was an important milestone of the research. The assignment orientation of this modelling technique means that the model building phase was totally dependent on the given engineering assignment as explained in the next paragraph.

The expert facing a task recognises that a model is needed to solve it, since some parameters have to be determined based on other ones. An ANN can be used as a model if some requirements are satisfied, e.g., a learning data set is given. One of the most typical applications of this modelling technique is if the expert is not aware of the dependencies among parameters. In many cases neither the dependencies nor the type or form of possible dependencies are known. This is one of the reasons why the general estimation and learning capabilites of ANNs give a viable solution for the expert. Several attempts can be found in the literature to solve various engineering assignments by using neural network based models [1][2][8][9]. The authors referred to applied the same concept to select the input and output parameters of the applied neural network, namely, parameters are known in the application phase were selected as inputs and the model has to estimate the output parameters which are unknown in the assignment. Consequently, the predefined assignment determines the input-output configuration of the model.

The estimation accuracy of the applied models plays a very important role in the engineering field, too. A special requirement for modelling is to satisfy the accuracy demands acceptable in a certain situation. The requirement cannot be always met in the assignment dependent input-output configuration, i.e., it is not obvious that this setting realises the most appropriate mapping between the variables, consequently, a method finding the appropriate input-output configuration of the applied neural network model is needed. A method like this was presented by the authors in [21]. The developed algorithm determines the input-output configuration of the neural network model satisfying the accuracy requirements and finding the maximal number of parameters that can be estimated within the prescribed accuracy resulting that the model has good properties from modelling point of view. This model building process runs totally independent from the assignment, which was a new, corn idea behind. That is why the models can be considered as the general model of the system in question.

As to the application phase, a method was developed, by which various engineering tasks can be solved, using the same general model [13]. Applying the general model the method estimates the unknown parameters of the given assignment, based on the known ones. The estimation is totally independent from the input or output positions of the known or unknown parameters on the general model. This new concept ensures that several assignments (e.g. in planning, optimisation or control) can be solved by using the same general model of the system in question.

3.2. The basic idea in the introduced concept

The previous paragraph summarised in short a method to find the appropriate input-output configuration of an ANN model of a system. The method starts with the collection of a set of parameters expected to describe the analysed system. The aim is to decide about all of these parameters whether they should act as inputs or outputs and to give also the multilayer perceptron model. The method is a modified backpropagation learning algorithm. The first phase of the backpropagation (BP) algorithm is the so named forward step where the ANN estimates its output parameters (output vector) based on the input parameters (input vector) and the network structure. During the learning phase, after the forward stage the BP algorithm calculates the partial derivatives of the network error on the network weights and modifies them accordingly. The so named batch learning stores the changes in weights and summarizes them through one batch, so the weights are changed only after all of the learning data have been presented. The procedure can be illustrated in short, as follows:

```
repeat
{
        for all of the input-output vector pairs
        {
                forward (based on the input data vector),
                calculate the derivatives of the network weights,
                calculate the corresponding changes of the weights and sum them up
        }
        change the weights with their corresponding sum
} until a special criterion, e.g. the value of estimation error is higher than required.
```

Batch learning was also the basis of the convergence accelerator algorithm called SuperSab [18] applied in the algorithm described here.

A flag is used in the developed algorithm to indicate if a neuron of the ANN model is protected or not. If a neuron is in a protected state, it means neither it takes part in the forward calculation nor the derivative is calculated related to it, consequently, zero modification ratio is stored for all of the weights in the ANN structure that are directly connected to this neuron (both input and output weights). So the weights are also set into a protected state. Certainly this protected state influences the calculation of derivatives of other unprotected parts of the ANN structure. This technique combined with the sequential forward selection search method was used to find the appropriate input-output configuration of the ANN model [21].

The search method starts with a network having the same number of input and output neurons which is equal to the number of parameters used to describe the analysed system. During the search method if a neuron is unprotected on the input side, the corresponding output is protected and vica versa. This behaviour is shown in Figure 1.

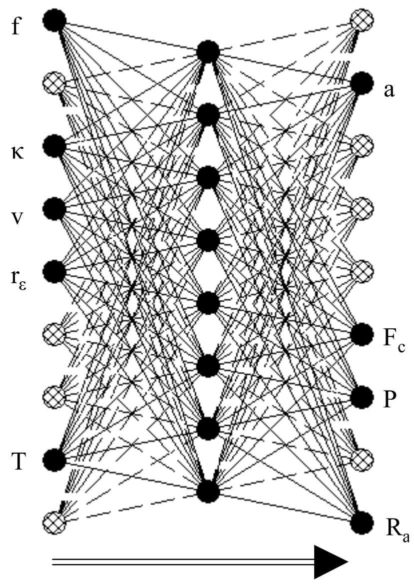

Figure 1. Protected and unprotected states of different neurons and corresponding weights. If a graphical object is not continuous it means that this part of the ANN structure is protected and, consequently, not used

During the input-output search the protection state of a neuron is changed only after a batch is over, namely, the protection of a neuron could not be changed until all the learning data have been processed by the ANN.

Considering a data vector set with missing data, it can be stated that generally, different parts of the data vectors can be missing, indicating that a learning method is needed, taking the incompleteness of individual data vectors into consideration. This requirement and the above method for changing the protection of a neuron lead to the idea to adapt the protection of a neuron according to the missing part of certain data vectors.

3.3. Description of the algorithm for handling incomplete data

The algorithm is based on the main idea of turning the neurons corresponding to the missing part of certain data vectors into protected state and leaves the other neurons in unprotected state. Further information is needed to realise this, i.e., to describe which part of the input- and output vector is missing. A flag called validity is used for indicating whether a date in the data vector is valid or not. It shows that a so named (binary) validity vector is attached to all of the input and output vectors to describe the validity state of the data incorporated in these vectors.

The protection of the input and output layer of the ANN structure changes according to the validity vector of the data vector, in question, namely if a data is invalid in the input or output vector, the corresponding neuron is set to protected, otherwise the neuron will be unprotected. It ensures that the protections of the input and output neurons change by all the learning data vector pairs. It is necessary to explain that the changes in weights are the same as in batch learning, namely, the changes remain summed during one batch of learning. If a learning vector indicates that a weight is set to be protected in the ANN structure, the corresponding value of sum for change the weight remains the same as that of the previous learning vector. At the beginning of one batch all the weights will be changed according to the calculated sum. The algorithm can be presented as a small cycle:

repeat
{
  for all the input-output vector pairs
  {
    set the protection of the network neurons according to the validity vector of the given data vector
    forward (based on the input data vector),
    calculate the derivatives of the network weights,
    calculate the corresponding changes of the weights and sum them up
    set all the neurons unprotected
  }
  change the weights with their corresponding sum
} until a special criterion, e.g. the value of estimation error is higher than required.

This modification shows that the introduced method gives a solution for handling the missing data in the learning and application phases of the ANN model, too. The test results of this new learning and application method will be presented in the next paragraphs.

4. Experimental tests

Some experimental tests were performed to evaluate the behaviour of the developed learning algorithm. The following situations were tested:

- based on artificial data
  - handling incomplete output data,
  - handling incomplete input data,
    - if the data are fully independent,
    - if the data are redundant,
- based on real measurements in the field of engineering: modelling a cutting process.

In the tests the behaviour of the introduced algorithm with varying ratio of missing data was studied. As it was outlined in the previous paragraph, the so named validity vector shows which part of the input and output data vector is missing. The test stages were started always with data vectors without any missing values, indicated by validity vectors. Different ratios of missing data were simulated through various ratios of zero values in the validity vectors. The network estimation errors were used for measuring the capability of a tested method. During the test runs the difference in the estimation errors originating form the various ratios of incomplete data was eliminated.

4.1. Handling incomplete output data

The first phase was to test the case where the missing data were in the output vectors. Four types of possible methods for handling missing data were compared:

1. Excluding the data vectors containing missing data, from the data set.
2. Writing a fix value into the place of missing data. Usually, data are normalised [12] before the learning stage of ANN, which transforms all the learning vector components into a predefined interval. Because this interval runs usually from 0 to 1, the value for replacing missing data was set to 0.5.
3. Writing random value into the place of missing data. The normalisation is the reason to set this value from the 0 to 1 interval.
4. Using the introduced new algorithm for handling the data with missing data.

A thousand data vectors were generated for learning and testing purposes. Two, randomly set values (x and y) gave the input vectors. Output parameters were also two dimensional, with the calculated values of x and y^2, respectively. The network estimation errors were compared in the learning and testing phases, as well. The ratio of missing output data differed in the tests. Results are shown in Figure 2.

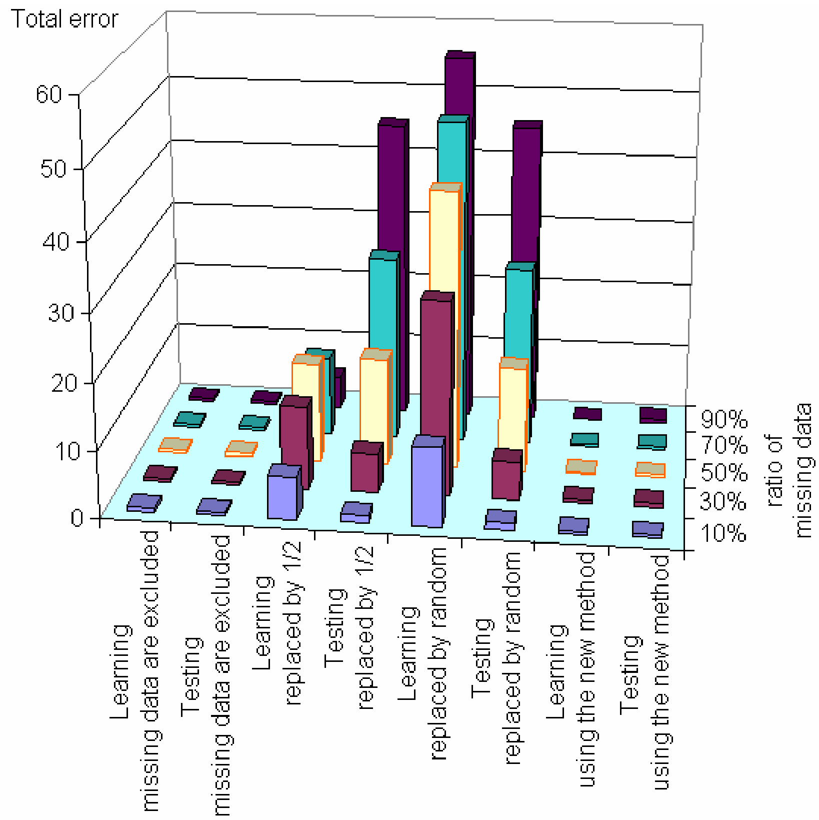

Figure 2. Test results of handling missing output data by different methods

The results show that the introduced method can learn dependencies among parameters, in the same way as in the case where no data were missing. It shows also that it is better to use the introduced method than to replace missing data by fix or random values. The results indicate that the method is able to handle missing output data.

4.2. Handling incomplete input data containing independent variables

A data set having three-dimensional input and output data vectors was used in this stage. Input data vectors consisted of randomly selected independent values of x, y and z, while output vectors incorporated calculated components of x*y, x*y*z and z. A hundred data with different missing ratio were used for learning. The analysis of the learning stage shows that the errors are mainly originated from the vectors where parts of inputs were missing. This experience leads to the question whether the data vectors without a missing part were enough to learn the dependencies among parameters and whether it could be better to eliminate the data vectors form the learning data set. Comparing the estimation errors between the ANNs trained with a data set having incomplete data vectors and trained with a data set where these data vectors were eliminated. This test was repeated for different ratios of complete data vectors. The results on the test data set are shown in Figure 3.

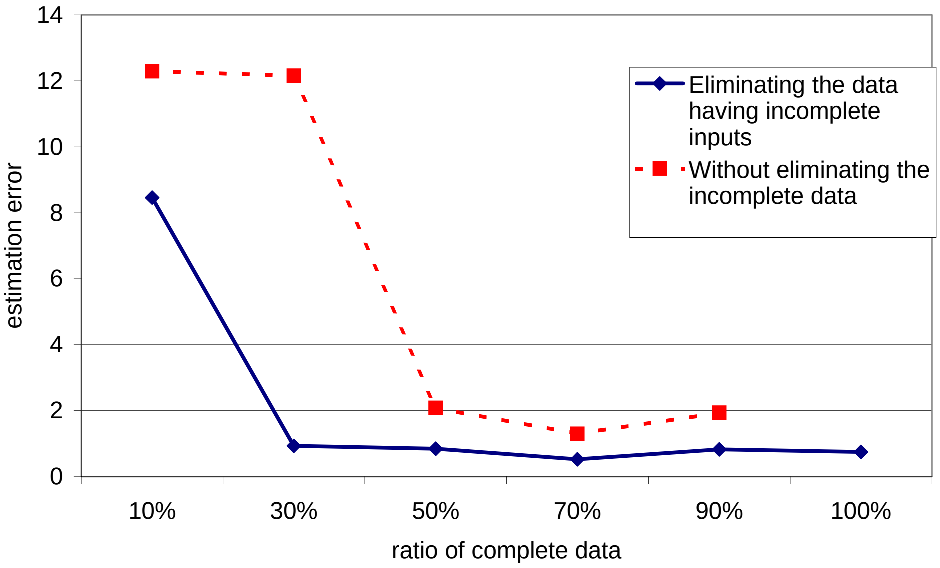

Figure 3. The4 figure shows that it is worth eliminating the training data vectors having incomplete, fully independent input data.

4.3. Handling incomplete but redundant input data

The previous paragraph outlined the benefits of eliminating data vectors from the learning data set if they have incomplete input vectors and the parameters are fully independent. This section analyses the introduced learning algorithm with redundant and incomplete input data. A simple task with two input parameters and one output parameter was selected. The input data are of the same value ensuring a high-level of redundancy (x, x). The output parameter is x^2. A thousand data with a third part of complete vectors, a third part were the first input, furthermore, a third part where the second input parameter was missing, were used for testing the models built up with different ratio of incomplete input vectors. Results related to estimation capabilities are shown in Figure 4.

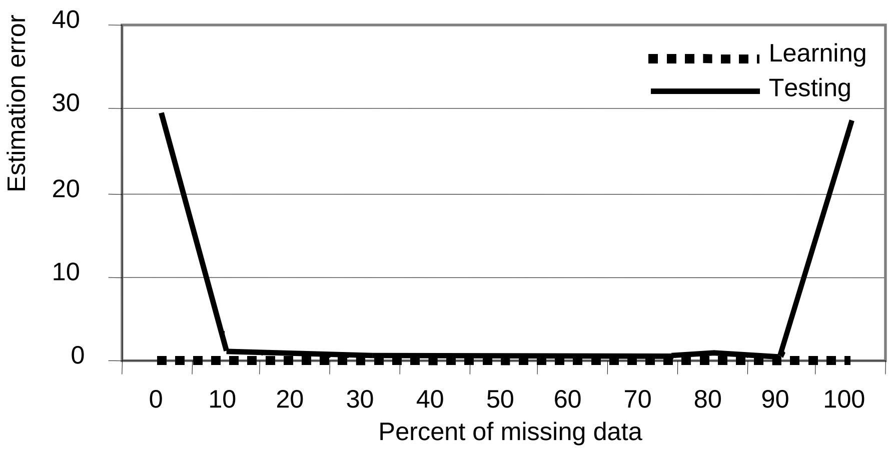

Figure 4. Estimation errors measured on the testing data set

Figure 4 shows an exciting result, namely, *if the ANN model is built up only with complete data or only with incomplete data, the estimation error is significantly higher than using a model built up on data set having complete and incomplete input data, as well.*

This fact gives the idea that if the input parameters are redundant it is worth impairing the data purposely to prepare the ANN model for handling incomplete input data successfully. This impairment means either setting some certain data as they would be missing, or extending a data set with the same but in some ratio impaired data.

5. Application of the developed tools for a the cutting process

The above paragraphs explained the test results of the introduced algorithm. These tests were based on different artificial data sets, which ensured that dependencies among modelled parameters are known. The application described in this section is based on real machining measurements, which leads to the fact that the exact dependencies among the system parameters are unknown and the analysed data set is also originated from a real process. The turning process was analysed in this experiment. The same material was cut at different speed (v), depth of cut (a) and feed per revolution (f). During the experiments the tool wear was measured and only the tools of which the flank wear (VB) had not got over a certain level, were applied, i.e. they were considered as sharp ones. 119 different experiments were performed and during the process the three components of the cutting force, the power and the cutting temperature, and after the process, the tool wear and roughness (Ra) were measured. Features [15] calculated from the measured data were used to solve a process-monitoring task, indicating that the a, f, v parameters and average and variance value of each force components and of the resultant force can be selected as input parameters. The temperature (T), average power (P) and the specific energy appropriation (u) give the output vectors (Figure 5).

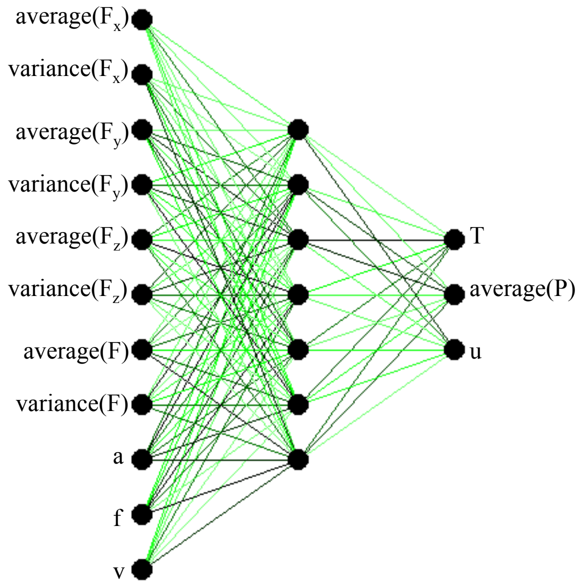

Figure 5. The ANN model for the monitoring task, in question

A hundred data were randomly selected for learning and the remaining 19 data were used for a testing process. A special ratio of missing data was used for building up the ANN model and a data set with another ratio of missing data for testing its estimation capability. The missing data were among the input and output data, as well. The estimation errors are shown in Figure 6.

Figure 6 shows several interesting outcomes:

- The best situation is to have missing data neither in the learning nor in the testing data set. No situation resulted in lower estimation errors.
- The estimation capabilities of ANN models trained on a missing ratio of 20%, 40%, 60% are very similar in their estimation capabilities.
- The estimation error of ANN models trained with no missing data is always higher than the errors of models trained with different ratio of missing data if the testing data could be incomplete. This experience conforms the idea from the previous paragraph, namely, it is worth impairing the data purposely to prepare the ANN model for handling incomplete data efficiently.
- A hundred data were used for learning and 19 for testing, which explains that Figure 6. cannot give advice in the cases of 80% incompleteness of learning or testing data sets.

Further interesting questions are generated by the figure:

- Is there an ideal ratio for incompleteness of learning data? Does it depend on the missing ratio of testing/application data?
- What is the allowed highest level of possible incompleteness?

These and similar questions are addressed in the research work of future.

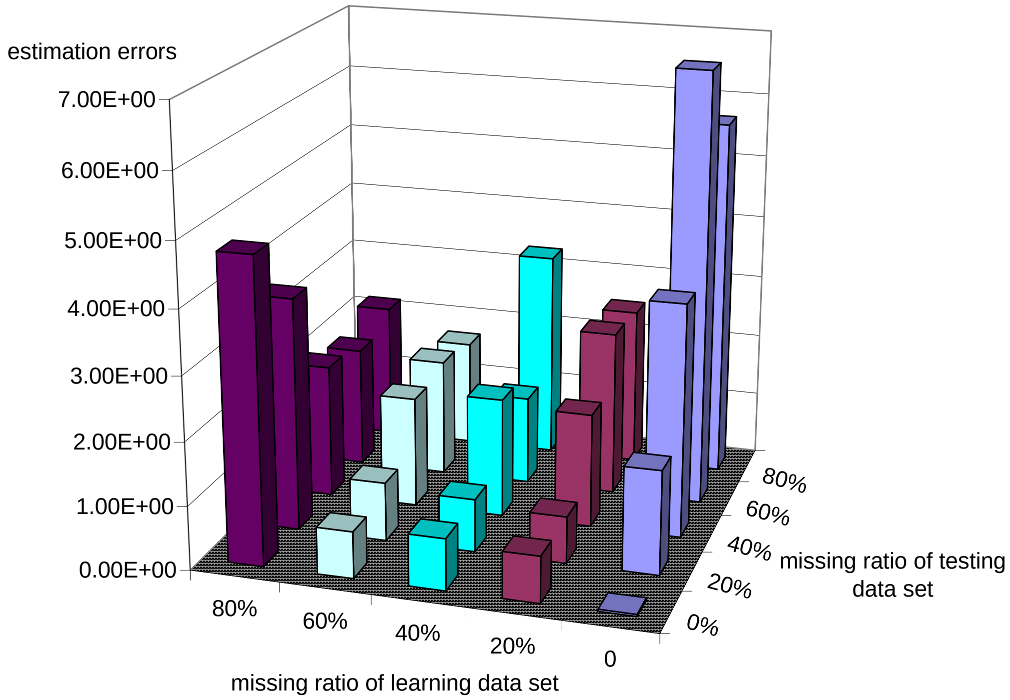

Figure 6. Estimation errors of different ANN models for monitoring a cutting process trained on a data set having a special ratio of missing data and tested on another data set with another ratio of missing data

6. Conclusions

A new approach is introduced in the paper for handling incomplete data in the learning and application phases of ANN based modelling. The basic ideas behind the research were not to extend the missing data with special values but to develop a learning algorithm for the applied neural network model to be able to handle data in a situation like this. The algorithm adapts the network structure to the individual data with missing components, it selects the so-called protection state of network neurons according to the missing/existing parts of data vectors. The developed algorithm was compared with three data-extending-methods and resulted in a model with superior estimation capabilities. The algorithm was tested through artificial data and it was found that it was completely able to handle missing output data. Tests showed that it is worth eliminating learning vectors with incomplete input vectors if input parameters are totally independent. Interesting result arose in the case of incomplete input data if input parameters are redundant: *it is worth impairing the data purposely to prepare the ANN model for handling incomplete data efficiently.* Tests for monitoring a turning process were performed and reinforced the previous conclusions in a real application field, as well.

7. Acknowledgements

The research was partially supported by Bolyai János Research Fellowship of Dr. Viharos Zsolt János and by the “Digital enterprises, production networks” project in the frame of National Research and Development Programme of the Ministry of Education (proj. No. 2/040/2001). A part of the work was covered by the *National Research Foundation,* Hungary, Grant No. T034632.

8. Literature

[1] Choi, G. H.; Lee, K. D.; Chang, N.; Optimisation of process parameters of injection modeling with neural network application in a process simulation environment, Annals of the CIRP, Vol. 43., No. 1., 1994, pp. 449-452

[2] Dini, G.; A neural approach to the automated selection of tools in turning, Proc. of 2nd AITEM Conf, Padova, Sept. 18-20, 1995, pp. 1-10

[3] Francis, N. M.; Brown, A. G.; Cannon, P. S.; Broomhead, D. S.; Nonlinear prediction of the hourly FOF2 time series in conjunctions with the interpolation of missing data points; Phys. Chem. Earth., Vol. 25, No. 4., 2000, pp. 261-265

[4] Gardner, R.; Bicker, J.; Using machine learning to solve tough manufacturing problems, International Journal of Industrial Engineering-Theory Applications and Practice, Vol. 7., No. 4., 2000, pp. 359-364

[5] Ghahramani, Z.; Jordan, M. I.; Learning form incomplete data; Report of the Massachusetts Institute of Technology, Artificial Intelligence Laboratory and Center for Biological and Computational Learning Department of Brain and Cognitive Sciences, C.B.C.L. Paper No. 108, A.I. Memo No. 1509, 1994

[6] Kamakura, WA.; Lenartowicz, T.; Ratchford, BT.; Prouctivity assesment of multiple retail outlets, Journal of retailing, Vol. 72, No. 4, 1996, pp. 333-356

[7] Keeler E.; David, J.; Hartman, H.; Jon, E.; Ferguson, J.; Bruce, R.; Method for operating a neural network with missing and/or incomplete data, United States Patent, No. 5 819 006

[8] Knapp, G.M.; Wang, Hsu-Pin; Acquiring, storing and utilizing process planning knowledge using neural networks, J. of Intelligent Manufacturing, Vol. 3, 1992, pp. 333-344

[9] Liao, T.W.; Chen, L.J.; A neural network approach for grinding processes: modeling and optimization, Int. J. Mach. Tools Manufact., , Vol. 34, No. 7, 1994, pp. 919-937

[10] Merchant, M.E.; An interpretive look at 20th century research on modelling of machining, Inaugural Address, Proc. of the CIRP International Workshop on Modelling of Machining Operations, Atlanta, Georgia, USA, May 19, 1998, pp. 27-31

[11] Monostori, L., Márkus, A., Van Brussel, H., Westkämper, E.; Machine learning approaches to manufacturing, Annals of the CIRP, Vol. 45., No. 2., 1996, pp. 675-712

[12] Monostori, L.; A step towards intelligent manufacturing: Modelling and monitoring of manufacturing processes through artificial neural networks, Annals of the CIRP, Vol. 42., No. 1., 1993, pp. 485-488

[13] Monostori, L.; Viharos Zs. J.; Markos, S.; Satisfying various requirements in different levels and stages of machining using one general ANN-based process model; Journal of Materials Processing Technology, Elsevier Science SA, Lausanne, 107: (1-3). NOV 22, 2000, pp. 228-235

[14] Monostori, L.; Viharos, Zs.J.; Hybrid, AI- and simulation-supported optimisation of process chains and production plants, Annals of the CIRP, Vol. 50., No. 1., 2001, pp. 353-356

[15] Monostori, László; Egresits, Csaba; Hornyák, József; Viharos, Zsolt János.; Soft computing and hybrid AI approaches to intelligent manufacturing. Book, Lecture Notes in Artificial Intelligence, 1416, Proceedings. 11th International Conference on Industrial & Engineering Applications of Artificial Intelligence & Expert Systems, Castellon, Spain, jun 1-4., 1998, pp. 763-774

[16] Pesonen, E.; Eskelinen, M.; Juhola, M.; Treatment of missing data values in a neural network based decision support system for acute abdominal pain, Artificial Intelligence in Medicine, Vol. 13, 1998, pp. 139-146

[17] Rangwala, S.S., Dornfeld, D.A.; Learning and optimisation of machining operations using computing abilities of neural networks, IEEE Trans. on SMC, 19/2, March/April, 1989, pp. 299-314

[18] Tollenare, T.; SuperSAB: fast adaptive backpropagation with good scaling properties, Neural Networks Vol. 3., 1990, pp. 561-573

[19] Valentini, R.; DeAngelis, P.; Matteucci, G.; Monaco, R.; Dore, S.; Mugnozza, GES.; Seasonal net carbon dioxide exchange of a beech forest with the atmosphere, Global Change Biology, Vol. 2., No. 3., 1996, pp. 199-207

[20] Van Luttervelt, C.A. Childs, T.H.C., Jawahir, I.S., Klocke, F., Venuvinod, P.K., 1998, Present situation and future trends in modelling of machining operations, Annals of the CIRP, 47/2:587-626

[21] Viharos, Zs. J.; Monostori, L; Automatic input-output configuration of ANN-based process models and its application in machining, Book: Lecture Notes of Artificial Intelligence - Multiple Approaches to Intelligent Systems, Conference, Cairo, Egypt, May 31-June 3, Springer Computer Science Book, Springer-Verlag Heidelberg, 1999, pp. 659-668

[22] Viharos, Zs.J.; Monostori, L.; Optimisation of process chains and production plants using hybrid, AI- and simulation based general process models; The Fourteenth International Conference on Industrial & Engineering Applications of Artificial Intelligence & Expert Systems, Book: Lecture Notes of Artificial Intelligence, Springer Computer Science Book, Springer-Verlag Heidelberg; Budapest, Hungary, 4-7 June, 2001, pp. 827-835

[23] Warnecke, G.; Kluge, R.; Control of tolerances in turning by predictive control with neural networks, Journal of Intelligent Manufacturing, 9/4, 1998, 281-287