Achievements in Mechanical and Materials Engineering

COMMENT Worldwide Congress on Materials and Manufacturing Engineering and Technology
16th - 19th May 2005 Gliwice-Wisła, Poland

COMMITTEE OF MATERIALS SCIENCE OF THE POLISH ACADEMY OF SCIENCES, KATOWICE, POLAND
INSTITUTE OF ENGINEERING MATERIALS AND BIOMATERIALS OF THE SILESIAN UNIVERSITY OF TECHNOLOGY, GLIWICE, POLAND
ASSOCIATION OF THE ALUMNI OF THE SILESIAN UNIVERSITY OF TECHNOLOGY, MATERIALS ENGINEERING CIRCLE, GLIWICE, POLAND

13th INTERNATIONAL SCIENTIFIC CONFERENCE ON ACHIEVEMENTS IN MECHANICAL AND MATERIALS ENGINEERING

Modeling of manufacturing processes by learning systems: the naïve Bayesian classifier versus artificial neural networks

M. Perzyk, R. Biernacki and A. Kochański

Institute of Materials Processing, Warsaw University of Technology
Narbutta 85, 02-954 Warszawa, Poland, M.Perzyk@wip.pw.edu.pl

Abstract: Modeling capabilities of two types of learning systems are compared: the naïve Bayesian classifier (NBC) and artificial neural networks (ANNs), based on their prediction errors and relative importance factors of input signals. Simulated and real industrial data were used. It was found that NBC can be an effective and, in some applications, a better tool than ANNs.

Keywords: Modeling; Manufacturing process; Learning systems; Naïve Bayesian classifier; Artificial neural networks.

1. INTRODUCTION

Designing and controlling manufacturing processes can be aided by the use of mathematical tools. In practice, we often have to do with processes of the “black box” type processes, whose physical nature is either unknown or very complex. Modeling processes of this kind consists in establishing the relationship between the input and the output (i.e. result) signals on the basis of a certain number of observed cases (regression problem). Recently, it has been artificial neural networks (ANNs) that have been commonly used for this purpose. Their applications include, for example:

- predicting properties of products or materials on the basis of the parameters of the technological process involved;
- predicting equipment failures on the basis of selected signals;
- identifying the causes that lead to the appearance of manufacturing defects in products;
- designing based on the specific data which was collected in the industry and generalized by ANNs.

Despite of great successes, ANNs have several shortcomings, such as complex and time-consuming training process and ambiguity of the results [1,3,5]. This lack of prediction uniqueness does not appear in modeling with the help of other mathematical tools, such as polynomials or the naïve Bayesian classifier (NBC) [2] – a simple learning system making use of the probability calculus. The authors of the paper are not aware of any applications of NBC to the modeling of industrial processes.

The present research project is concerned with an analysis of the NBC’s applicability to the problems of modeling manufacturing processes, an analysis residing in the comparison of

prediction errors of the NBC and those of ANNs, as well as in the comparison of the relative importance factors of input signals for the data modeled with the help of both systems. Basically, ANNs are used by interrogating them, i.e. by the simple calculation of output values for assumed values of inputs. However, to facilitate performing the tasks such as the identification of the causes of defects in products, or critical factors for a failure of equipment, an analysis of the relative importance of the input signals seems to be most useful. Similarly, finding production parameters, which are most significant for obtaining particular physical or economical effects, can help in the optimization of manufacturing process.

2. RESEARCH METHODOLOGY

The data sets used in the present work differ in the character of the dependence between input and output signals. They include both industrial and simulated data. The sets of the latter type, in which the input-output dependencies are known, make possible a better assessment of the learning systems' prediction correctness. These simulated data sets were created on the basis of assumed simple equations of the type Y = f(X1, X2, ...) in the following way: all the input values were originally set at random between 0 and 1, then the output was calculated and, finally, ±20% noise of the Gaussian type distribution was added to the generated input values. In the present work, the following data sets were used:

Industrial data 1 – ductile cast iron strength as the function of its chemical composition.

Industrial data 2 – the input quantities here are the production process parameters which are related to the green sand mould (12 parameters). The output quantity is the appearance of the gas porosity defect in steel castings. In this case, the original output was of a discrete type (category "1" – no defect, category "2" – defect).

Simulated data 1 – the set was obtained in accordance with the following equation: Y = X1 + 2•X2 + 3•X3 + 4•X4 + 5•X5.

Simulated data 2 – the set was obtained in accordance with the following equation: Y = (6•X1)3 + (10-3•X2)3 + ... + (1•X7)3 + ... (1•X12)3. The output values of this set have a continuous character.

Simulated data 3 – the output values Y of this set have a binary character, 0 or 1, depending on whether or not the sum above goes beyond a certain boundary value. The values of the input data in this case were identical as the values in the simulated data 2 set.

Simulated data 4 – the training set is of the same type as the *simulated data 3*. However, this time the number of records is much smaller.

The *simulated data 2, 3* and *4* were designed in a way making possible a comparative analysis of selected results from these sets with the results obtained for the *industrial data 2*.

The networks had MLP type architecture with one hidden layer having the number of neurons equal to the number of network's inputs. To improve learning results, the training method called 'simulated annealing' has been applied, combined with conventional back-propagation method [1].

The relative importance factor of input signals can be defined in several ways. In general, it should reflect the maximum extent of change of the system output due to possible changes of the inputs. The definition assumed in the present work is based on the maximum differences between the output values which can be obtained by changing the considered input (within the training interval), while the remaining inputs are kept at randomly assumed values. This is repeated a number of times and the average of the maximum differences is calculated for that input. The values thus obtained for each input are then normalized between

0 and 1 so that the most significant input gets the rank equal 1 and the least significant – the rank equal 0.

For NBC training, continuous variables were translated into attribute values and categories through the inclusion of a particular variable into an appropriate range of its original value, with a particular number assigned to the range. For each of the above sets, trials were made in setting the number of ranges (categories) for the input and output values. The assumed numbers of categories were those resulting in the lowest number of errors for the verifying data (or the training data in the absence of the verifying data). It should be noted that verifying sets are not necessary to train the NBC. They can be used only for assessment of the system's prediction ability in the case of new data.

The relative importance factor for NBC is calculated on the basis of the elements of trained system. An input Xi influences the category of output Y through the conditional probabilities of the form:

$$\Pr = \left(X_i = c_j^{(x)} \middle| Y = c_k^{(y)}\right) \tag{1}$$

where $c_j^{(x)}$ is an input category, $c_k^{(y)}$ - the output category.

The measure of an input's importance is defined as a maximum possible difference of the probability values given by (1) for that input. These maximum differences are calculated for each category of output $c_k^{(y)}$ and averaged over all possible output categories. The relative importance factors are obtained by normalizing those averages in the 0 – 1 interval.

3. RESULTS

From the six training data sets used in the present work the results obtained for two of them will be presented in this section. The average errors of output value prediction results obtained from the NBC and the ANNs were identified, taking into consideration their proportion in the characteristic value ranges. Figure 1 represents the distributions of prediction errors for the simulated data in accordance with the following equation: $Y = (6 \bullet X1)^3 + (10\text{-}3 \bullet X2)^3 + ... + (1 \bullet X7)^3 + ... (1 \bullet X12)^3$. The simulated output values of this set have a continuous character. The predictions of both systems modeling the complex relations have a high error ratio. It should be emphasized, however, that the ANN coped with its task slightly better, since the proportion of small error is larger in the case of this system.

One can observe large scatters of results for ANNs training sessions. Although the X1 signal has the highest average importance value, the next signals (X6 and X11) are not very much different. Taking into consideration the huge scatters, one cannot affirm that the neural network found a signal with such a predominant importance. On the other hand, the results of the classification made by the NBC are unequivocal in character. It is worth emphasizing that the ambiguity of the ANNs predictions, which shows up – among other things – in high scatters of relative importance factor values, seems to be unavoidable, despite all the efforts aimed at its reduction [4].

4. SUMMARY AND CONCLUSIONS

The comparative analysis, performed on the basis of results obtained for all six data sets applied in the present work, leads to the conclusion that the prediction errors of a NBC may

be lower or higher that the errors of an ANNs. For two identical simulated data sets with binary output, which differed only with respect to size, the NBC produced a smaller error (percentage of false categories) than an ANNs.

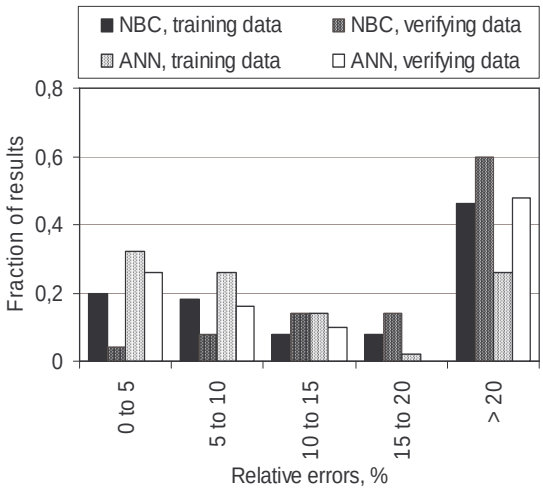

Figure 1. The comparison of output value prediction errors obtained from the NBC and the ANNs (minimum errors from 10 training sessions) for the simulated data set

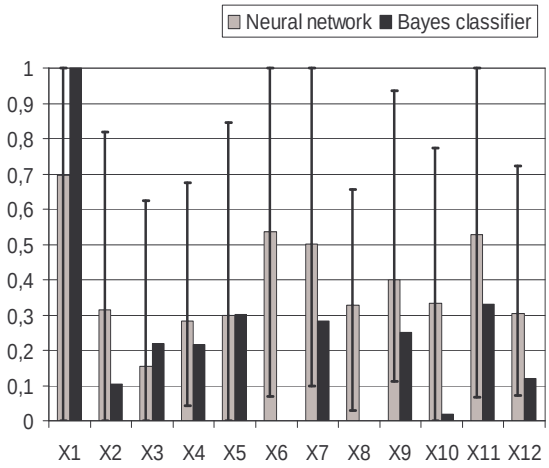

Figure 2. Relative importance factor values for the input signals for the simulated data set (the scatters of values obtained from 10 training sessions of ANNs are also shown)

The relative importance factor of input quantities for the NBC in its present form leads to a low degree of precision of results in the case of signals with lower importance. For an output of a binary type the NBC identified a clearly conspicuous signal and hence, its result was better that that obtained from ANNs.

Apparently the main advantage of the NBC in comparison with ANNs is its uniqueness, as well as the simplicity of its use. It may also be assumed that the NBC is less demanding as far as the size of the training set is concerned (for instance, the training does not require the use of a verifying set).

It seems that the NBC can make a learning system which is useful in industrial applications and which, in some cases, may be better than ANNs. However, further research is necessary, especially on extending the possibilities of interpretation of the results of both systems.

REFERENCES

1. T. Masters, Practical Neural Network Recipes in C++, Academic Press, Inc., 1993.
2. T.M. Mitchell, Machine Learning, McGraw-Hill, 1997.
3. M. Perzyk, A. Kochański, Detection of causes of casting defects assisted by artificial neural networks, Journal of Engineering Manufacture, Proceedings of the Institution of Mechanical Engineers, Part B. Vol. 217, (2003) p. 1279-1284.
4. M.A. Yescas, H.K.D.H. Bhadeshia, D.J. MacKay, Estimation of the amount of retained austenite in austempered ductile irons using neural networks, Materials Science and Engineering A311 (2001), p. 162-173.