Journal of Materials Processing Technology 164–165 (2005) 1430–1435

Journal of Materials Processing Technology

www.elsevier.com/locate/jmatprotec

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

Marcin Perzyk*, Robert Biernacki, Andrzej Kochański

Institute of Materials Processing, Warsaw University of Technology, Narbutta 85, 02-954 Warszawa, Poland

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, which physical nature is either unknown or very complex. Modeling processes of this kind consist 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. The same problem may be solved by networks with different architectures and topology of connections. Moreover, for a particular network, each learning brings about different values synapse weights related to reaching local minima of network errors [2,5,6].

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) [3] – 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.

* Corresponding author. Tel.: +48 22 8429347; fax: +48 22 8499797.
E-mail address: m.perzyk@wip.pw.edu.pl (M. Perzyk).

0924-0136/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.jmatprotec.2005.02.043

2. Learning systems making use of the naïve Bayesian classifier

Those systems which make use of probabilistic methods are typically based on the Bayesian theorem and equation [1]. They include the Bayesian classifiers (the optimal classifier, without any practical importance, and the NBC) as well as the Bayesian networks [3].

The Bayesian theorem is applicable in the case of events with known outcomes whose occurrence depends on factors preceding them. We calculate the probability that a given factor is the cause of a particular event (the effect). The use of the NBC requires calculating of probabilities on the basis of an appropriately prepared training data set, which consists of examples described by means of attributes and the target concept. The use of the NBC consists in the following:

- Establishing the so-called attributes (categories) for input quantities and the categories of the output quantity (the target concept). It should be noted that the NBC requires the use of quantities which are nominal or serial, but not continuous in character. The "value" of an attribute or of the target concept is defined through marking its membership in a particular category, which in turn may be identified verbally. For instance, something may be "very small", "small" "medium", "big" and "very big" (five categories) or "good" and "bad" (two categories).
- Creating the training data set which consists of records including the values of attributes of all input quantities, as well as the corresponding value of the output quantity.
- On the basis of the training data set the probabilities of particular categories of the output quantity are estimated, as well as the probabilities of particular values of attributes (categories) of all input quantities for particular output categories. This phase is referred to as the training of the classifier.
- On the basis of the estimated probabilities obtained in the training phase one can calculate the probabilities of the occurrences of each category of the output quantity for any given case (any input vector). The solution given by the NBC is that value of the output quantity category, which scored the highest probability value. If the output categories were created by assigning particular ranges of continuous output values to the numbers standing for the categories, it is possible to change the number of the calculated category for example into a mean original value from such a range.

3. Research methodology

3.1. Data sets

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 $Y = f(X_1, X_2, \ldots)$ 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 [4]. In addition to the training subset containing 700 records, the verifying subset was also created, which contained 90 records.
- *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). The training data set contained 172 records and there was no verifying set. The data in question were presented and analyzed in detail in the work of two of the present authors [5].
- *Simulated data 1*. The set was obtained in accordance with the following equation: $Y = X_1 + 2X_2 + 3X_3 + 4X_4 + 5X_5$. The training set contained 1000 records, while the verifying one – 200 records.
- *Simulated data 2*. The set was obtained in accordance with the following equation: $Y = (6X_1)^3 + (10^{-3}X_2)^3 + \cdots + (1X_7)^3 + \cdots + (1X_{12})^3$. The output values of this set have a continuous character. The training set contains 1000 records, while the verifying one – 200 records.
- *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. The boundary value was chosen in such a way that the proportion of the $Y = 1$ value in the set was about 15%. The number of records in the set is as in the *simulated data* 3.
- *Simulated data 4*. The training set is of the same type as the *simulated data* 3. However, this time the number of records is 172 and there is no verifying data set.

The *simulated data* 2–4 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.

3.2. Artificial neural networks

The networks had the MLP-type architecture with one hidden layer having the number of neurons equal to the number of the network's inputs. To improve the learning results, the training method called 'simulated annealing' has been applied, combined with the conventional back-propagation method. The simulated annealing method was used for a better selection of the initial values of the synapses while the