Expert Systems with Applications 42 (2015) 4687–4696

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Modeling of steelmaking process with effective machine learning techniques

Dipak Laha [a], Ye Ren [b], P.N. Suganthan [b,*]

[a] Department of Mechanical Engineering, Jadavpur University, Kolkata, India
[b] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore

ARTICLE INFO

Article history:
Available online 30 January 2015

Keywords:
Steelmaking process
Modeling
Prediction
Random forests
Support vector regression
Artificial neural networks
Dynamic evolving neural-fuzzy inference system
5-Fold cross validation

ABSTRACT

Monitoring and control of the output yield of steel in a steelmaking shop plays a critical role in steel industry. The yield of steel determines how much percentage of hot metal, scrap, and iron ore are being converted into steel ingots. It represents the operational efficiency of the steelmaking shop and is considered as an important performance measure for producing a specific quantity of steel. Due to complexity of the steelmaking process and nonlinear relationship between the process parameters, modeling the input–output process parameters and accurately predicting the output yield in the steelmaking shop is very difficult and has been a major research issue. Statistical models and artificial neural networks (ANN) have been extensively studied by researchers and practitioners to model a variety of complex processes. In the present study, we consider random forests (RF), ANN, dynamic evolving neuro-fuzzy inference system (DENFIS) and support vector regression (SVR) as competitive learning tools to verify the suitability of applications of these approaches and investigate their comparative predictive ability. In the present investigation, 0.00001 of MSE is set as a goal of learning during modeling. Based on real-life data, the computational results depict that the training and testing MSE values of SVR and DENFIS are close to 0.00001 indicating that they have higher prediction ability than ANN and RF. Also, mean absolute percentage prediction errors of the proposed models confirm that the predicted yield based on each method is in good agreement with the testing datasets. Overall, SVR performs best and DENFIS the next best followed by ANN and RF methods respectively. The results suggest that the prediction precision given by SVR can meet the requirement for the actual production of steel.

1. Introduction

Monitoring and control of the output yield of steel of a steelmaking shop plays a critical role in steel industry. The yield of steel determines how much percentage of raw material hot metal, scrap, and the iron ore are being converted into steel ingots. It represents the operational efficiency of the steelmaking shop and is considered as an important performance measure for producing a specific quantity of steel. Due to the complexity of the heat and mass transfer along with a large number of chemical reactions involved in the steelmaking process, the nature of relationship between the input and output parameters in the process is complex and highly nonlinear and also the process data for these parameters collected from the steelmaking shop are quite noisy. Therefore, modeling the input–output process parameters and accurately predicting the output yield of steel in the steelmaking shop is very difficult and has been a major research issue.

For the purpose of modeling and predicting the output of a process, a considerable group of authors have used different methodologies such as multiple linear regression model, response surface methodology (RSM), and artificial intelligence approaches in a wide variety of complex processes such as ironmaking process (Tang, Zhuang, & Jiang, 2009), metal cutting process (Mukherjee & Routroy, 2012; Tsai, Li, & Chen, 2008; Pontes, de Paiva, Balestrassi, Ferreira, & da Silva, 2012), molding process (Erzurumlu & Oktem, 2007) and medicinal chemistry (Jalali-Heravi, Asadollahi-Baboli, & Shahbazikhah, 2008) to name a few. These studies reveal that artificial neural networks (ANN) have been shown to perform nonlinear data well and has better predictive quality as compared to multiple linear regression model and RSM due to its good learning ability while they suffer from a number of weaknesses such as unexplainable nature of relationships among the input and output parameters of the process, requiring much higher computational times and often tendency of

* Corresponding author.
E-mail addresses: dipaklaha_jume@yahoo.com (D. Laha), re0003ye@ntu.edu.sg (Y. Ren), epnsugan@ntu.edu.sg (P.N. Suganthan).

http://dx.doi.org/10.1016/j.eswa.2015.01.030
0957-4174/

overfitting leading to poor performance. Apart from modeling using ANN, recently some additional machine learning methods have emerged as alternative competitive learning tools such as random forest (RF), adaptive neuro-fuzzy inference system (ANFIS), dynamic evolving neural-fuzzy inference system (DENFIS) and support vector machine (SVM). A considerable number of studies have focused on developing prediction based modeling using RF, ANFIS, DENFIS and SVM in a variety of complex processes (e.g., Aich & Banerjee, 2014; Azadeh, Saberi, Gitiforouz, & Saberi, 2009; Kara, Boyacioglu, & Baykan, 2011; Kavousi-Fard, Samet, & Marzbani, 2014; Lariviere & Van den Poel, 2005; Oliveira, Oehler, San-Miguel-Ayanz, Camia, & Pereira, 2012; Sanchez Lasheras, Garcia Nieto, de Cos Juez, & Vilan Vilan, 2014; Zhou, Wu, Chan, & Tontiwachwuthikul, 2011).

Neuro-fuzzy modeling approaches such as ANFIS, DENFIS, which are combination of ANN and fuzzy system are preferred in order to explicate relationship between the parameters identified in ANN and based on fuzzy inference systems, membership function and fuzzy rules. There have been several studies applying ANFIS, DENFIS and ANN, demonstrating the superiority of ANFIS and DENFIS over ANN with respect to accuracy levels and effective interpretation of the relationship among the parameters of the process (Azadeh et al., 2009; Talei, Chua, & Quek, 2010; Zhou et al., 2011). As an alternative learning strategy and pattern recognition method, RF has recently attracted growing interest of researchers. It is a nonparametric approach based on combination of Classification and Regression Trees (CART). This method has been applied in various processes such as in forest management (Oliveira et al., 2012), customer relationship management (Lariviere & Van den Poel, 2005) involving complex interaction between input and output variables. These studies have investigated the potential applications of RF and multiple linear regression demonstrated that RF has a higher predictive performance and the generalization ability than multiple linear regression. However, it has a drawback of behaving as a black box like ANN, i.e., the relationship between input and output variables are not explainable because in this case the individual trees cannot be examined separately. Recently, SVM has been also successfully employed for its good predictability ability in complex processes. It is a global optimization model for classification and regression (SVR) has grown a great research interest to various research and industrial process applications since the development of SVM (Cortes & Vapnik, 1995). SVM performs superior prediction mainly due to the use of structural risk minimization principle instead of empirical risk minimization principle as is used in ANN modeling, resulting in improved generalization ability than ANN Tang et al., 2009; Yuvaraj, Murty, Iyer, Sekar, & Samui, 2013).

Although there have been several studies on the applications of RF, ANN, DENFIS and SVM for modeling the input–output parameters of different complex processes, modeling using these techniques especially on prediction of output yield of the steelmaking process was not reported in the literature. However, there are some literature on ANN modeling of a steelmaking process to predict output parameters, like temperature of the liquid metal and the volume of necessary oxygen blow (Falkus, Pietrzkiewicz, Pietrzyk, & Kusiak, 2003), metallurgical length, shell thickness at the end of the mould and billet surface temperature (Gresovnik, Kodelja, Vertnik, & Sarler, 2012), percentage of phosphorus in the final composition of steel (Monteiro & Sant'Anna, 2012; Shukla & Deo, 2007). A comprehensive description of modern steelmaking processes along with physical and mathematical modeling and solution methodologies based on AI-based techniques, especially ANN, GA, CFD, and FLUENT software is provided by Mazumdar and Evans (2009). Therefore, this research is motivated to verify the suitability of potential applications of RF, ANN, DENFIS and SVM to investigate the comparative performances of these approaches for modeling the input–output process parameters, especially on minimizing the error of prediction of the output yield in the steelmaking process.

Inspired by this motivation, this paper first provides the basic principles of RF, ANN, DENFIS and SVM algorithms to model and predict the steelmaking process parameters and then, verify the predictability of the output yield using these techniques based on real-life data set. The present study is carried out using the data for the input and output parameters of an open hearth process in steelmaking industry. The process data was collected from a steelmaking shop, located in Eastern India. The production data was collected on a monthly basis due to non-availability of daily-basis data.

The remaining of the paper is organized as follows. Section 2 gives a brief description of literature review of machine learning techniques. Section 3 presents the input and output variables of the steelmaking process under investigation. Section 4 provides the details of the methodologies of RF, ANN, DENFIS and SVR. The experimental settings and discussion of the results are reported in Section 5. In Section 6, we draw the conclusions, research limitation and future scope of work.

2. Previous work

Recently, there have been growing number of studies on prediction based modeling applied to a variety of complex processes that are frequently encountered in many industries and real-life situations. Prediction modeling of these complex processes due to their complex characteristics and practical importance have long received attention of researchers and practitioners. In the following section, we focus on the review of the previous studies on modeling complex processes using four effective machine learning techniques: random forests, artificial neural networks, dynamic evolving neuro-fuzzy inference system, and support vector machines.

2.1. Random forests (RF)

Since the introduction of random forests proposed by Breiman (2001), it has been increasingly gaining popularity due to its consistent, robust performance results and easy to implement. It is a nonparametric technique based on ensemble of CART. It has been successfully applied in a wide range of fields such as data classification (Liu, Wang, Wang, & Li, 2013), customer relationship management (Lariviere & Van den Poel, 2005), chemical mixtures (Cooper et al., 2014), medical data classification (Seera & Lim, 2014), customer churn prediction in banks (Xie, Li, Ngai, & Ying, 2009), forest management and (Oliveira et al., 2012). Liu et al. (2013) compared RF, SVM and ANN for electronic tongue data classification in order to recognize orange beverage and Chinese vinegar and have demonstrated that RF may be a promising method among them. Lariviere and Van den Poel (2005) employed random forests and regression forests in customer relationship management for predicting customer retention and profitability as measures of customer outcome and demonstrated that both these techniques provide better fit for the estimation and validation sample than ordinary linear regression and logistic regression models. Cooper et al. (2014) used RF to accurately predict the concentration of different hydrocarbons in water based on the sensor array responses and showed that the accuracy of this technique was not unduly affected by increasing mixture complexity. Seera and Lim (2014) proposed a hybrid algorithm combining RF, fuzzy min–max neural network and CART for solving medical data classification problem. RF is exploited to build diverse CART models so that the best tree can be selected for classification and rule extraction purposes. Oliveira et al. (2012) fist identified the main structural variables that seem to influence fire occurrence in