Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.12785/amis/072L11

Data De-noising Based on PCA-KNN Algorithm in Billet Surface Temperature Measurement

Huiyan Jiang[1], *Fengzhen Tang*[1], *Lingbo Zou*[1] *and Yenwei Chen*[2]

[1]Software College, Northeastern University, Shenyang, 110819, China
[2]College of Information Science and Engineering, Ritsumeikan University, Kusatsu-shi, Japan

Received: 5 Oct. 2012, Revised: 25 Dec. 2012, Accepted: 28 Dec. 2012
Published online: 1 Jun. 2013

Abstract: This paper first analyzes the soft-sensor technology and temperature measurement technology used in recent soft measurements, focuses on the accuracy and efficiency requirements of soft measurement in slab surface temperature detection. In this paper, we collects images noise processes of continuous casting slabs, and uses component-based OTSU segmentation method to extract the slab area and then implements Hough transform for edge correction; then measures the selected regions of interest in pixels and extracts color features using PCA for feature reduction; we extracts the improved data set with KNN algorithm for noise reduction, and the removal of contradictory data; the final regression models are used in prediction.

Keywords: Temperature soft measurement, Image processing, Regression model, Knowledge acquisition

1 Introduction

Since the 1990s, with the support vector machine theory put forward, more and more support vector machine soft measurement techniques were used [1]. Least squares support vector machine combined with a mixture of expert system with soft sensor modeling method were proposed and applied to the liquid steel temperature prediction [2]. The support vector machines have the advantages of simple topology, sound foundation, globally unique optimal solution, and could obtain the maximum useful information from a narrow distribution of the samples, which has strong data processing ability and learning generalization capacity. Therefore, SVM based soft temperature measurement will have a broad practical value.

Billet temperature's real-time computing model was put forward in 2007 [3]. We can use the "billet solidification heat transfer mathematical model" [4,5] coupled with online correction to obtain more accurately measures of the temperature effect. Fuzzy model [6] of slab surface temperature measurements was introduced. The fuzzy model uses gradient descent method to modify the model, which build a good model not only reflects the actual system, but also has a good practicability. RBF network [7] based soft-sensor model of billet temperature can achieve the accurate and timely forecasts, reduce fuel and reach the purpose of billet surface oxidation. Kang Ping [8] uses neural network software measurement methods based on the measured experimental data, to establish billet heating temperature in the soft sensor model, which can make a more accurate prediction to the temperature of the billet.

However, sampled data sets inevitably contain noise, including random errors, negligence statute of error and data error. Random error is the mean difference of the same large number of repeated measurements. As long as the casting process and the thermal imager run normally, random measurement error has a very small influence to the soft measure model. Gross error is usually due to system failure, instrument failure or human error. Statute of error data is cause by the lack of consideration of certain factors or loss data compression caused by the loss of information. The latter two gross errors will seriously affect the temperature of this soft measure model.

2 K-NN Regression

The Nearest Neighbor algorithm establishes a classification method which has no special assumptions in the form of the function; the only assumption is that the

* Corresponding author e-mail: hyjiang@mail.neu.edu.cn

function is a smooth function. That is to say that k-NN algorithm is a non-parameter estimation method, because it does not involve the parameters' estimation in an equation which is formed by the assumed function.

In the process of finding the nearest k neighbors, it is necessary to define a kind of measuring function to measure the distance between the samples. This function gives two samples of the size of the distance between the scalar judged by far and near. The scalar distance of two samples will be calculated by the function. In the way, we can judge which sample is closer to the new sample. The distance measurement function must satisfy the followed four properties. For arbitrary variables a, b, and c:

(1) Non-negativity $D(a,a) \geqslant 0$
(2) Reflexivity $D(a,b) = 0$ only when $a = b$
(3) Symmetry $D(a,b) = D(b,a)$
(4) Triangle inequality $D(a,b) + D(b,c) \geqslant D(a,c)$

The common metric distance function is the Euclidean distance. The Euclidean distance of the two attribute samples is defined as:

$$D_E(X,Y) = \sqrt{\sum_{i=1}^{l} (x_i - y_i)^2} \tag{1}$$

The basic steps of k-NN algorithm are:

1. Choose the training samples and record the category sign of the samples.
2. Calculate the distance between the new sample and the all the training samples (absolute distance, the Euclidean distance or Minkowski distance), find the k nearest samples of the new sample.
3. The category which most samples belong to will be taken as the category of the new sample. If there are two patterns of the same proportion, the pattern which includes the nearest sample will be chosen as the category that the new sample belongs to.
4. According to the largest share of the nearest samples, the category which the sample belongs to will be identified.

3 Data De-noising Model

Considering that the quality of this data is very important to the predictive model, we use k-Nearest Neighbor algorithm to remove noisy data. The intermediate target data item, to find his k-Nearest Neighbor, should use the function value of k-nearest neighbors to estimate target function value data item. If the estimated objective function value and the value of the data item itself vary widely, it indicates that the data function value of items are with conflicting data attributes, so the data item is a noise data, which should be removed.

In this paper, we define the weighted methods and properties from the weighted k-NN algorithm for improvement. We introduce W to be the weighted matrix of property rights of individual properties, and introduce $A = (\alpha_1, ...\alpha_k)$ as the weight vector that weighted the contribution of the k neighbors. Distance measuring function is defined as follows:

$$D(X_i, X_j) = \sqrt{\sum_{r=1}^{l} w_r (x_{ir} - x_{jr})^2} \tag{2}$$

Wtakes the diagonal matrix:

$$W = \begin{bmatrix} w_1 & 0 & \ldots & 0 \\ 0 & w_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & w_l \end{bmatrix} \tag{3}$$

The objective function of the new sample is calculated as follows:

$$\hat{f}(x) = \frac{\sum_{i=1}^{k} \alpha_i f(X_i)}{\sum_{i=1}^{k} \alpha_i} \tag{4}$$

In the formula, $\alpha_i = \frac{1}{D(x,X_i)^2}, i = 1,2,...k$.

The process of the algorithm is described as follows:

1. Select the training data set that is without noise.
2. Determine the attribute weight matrix W, clip threshold θ and the number of the nearest neighbor samples k. In this paper, we use the square of the coefficient $(e_1)^2$ of the training data set $Y_i^1 = e_1 Y_i^T$ which is the first principal component, as the value of property rights. Then use the maximum error value in the prediction using k-neighbor rules of training data set as the threshold θ;
3. Take any sample from the data set which needs to be clipped. According to the above-defined distance metric function and function value estimation algorithm rules of k-NN, we try to find nearest neighbors from the training data set, and estimate the value of the function $\hat{f}(x)$. Calculate the estimation error $error(x) = |\hat{f}(x) - f(x)|$, if $error(x)$ is greater than θ, the given threshold, then remove the sample. Otherwise, reserve the sample.
4. Repeat a and traverse the entire data set until all samples in D are considered, get the sets of data D' which gets rid of the contradictions of data.

4 Experiments and Results

4.1 Data De-noising Experiment

First select 100 samples of non-noise. The method will choose 10 pixels in the billet region of each image, which

is from the 10 images in the experiments, then carefully record and extract features.

Secondly determine the weight. The approach to determine the weight is to analysis the principal component of these 100 samples, the attribute weights matrix is the diagonal elements $w_1,...,w_{13}$. Parameter K=3, the query sample uses some of the training samples, getting the maximum prediction error of 28.179 and the data clipping threshold is set to 29. The dataset size needed to be clipped is 600. Experimental platform is Visual Studio 2008, operating environment is laptop of 2.53GHz with 2.00GB RAM. Running the algorithm consumes eight minutes, and 495 samples are left, excluding 105 samples, which includs100 samples extracted from the background region and 5 samples extracted from slab region.

4.2 Parameter Optimization

Typically, using a grid search method to select the parameters of SVR is time-consuming and blind. In this paper, particle swarm optimization algorithm is adopted for the SVR parameter selection, and experiments show that the method has higher prediction accuracy and fast convergence speed.

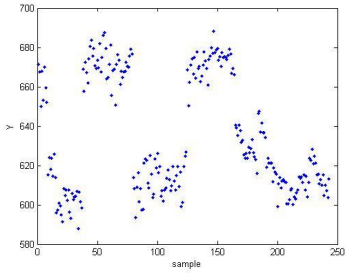
The training samples and test samples using in the experiments are after feature selection by $k = 3$. That is to use the first three principal components as feature vectors. In PSO algorithm, the population size m is set to 20, the learning factors $c_1 = c_2 = 2$. The algorithm's termination condition is the maximum number of iterations for the 100 samples or fitness value reaches 0.9. Search space of C is [1, 215]. Search space using in the experiments of γ is $[2^{-5}, 2^3]$. Search space of is $[2^{-5}, 2^{-1}]$. The results are in Table 1.

Table 1 Comparison of parameters selection between PSO and grid search

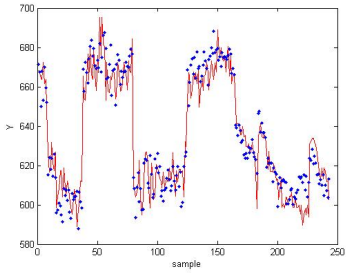
Method	C	γ	ε	MSE	Time(s)
PSO	128	0.031	0.5	7.683	513.243
Exhaustive search	8	2	0.25	13.288	904.34

4.3 Regression Model Experiment

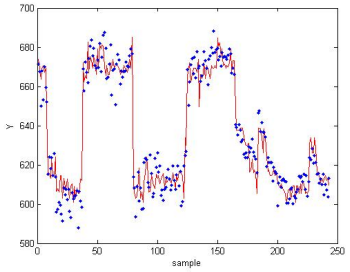
The training samples and testing samples used by our experiments are selected through feature $k = 3$, and training sample size is 352, the test sample size is 243 and the feature dimension is 3. Multiple linear regression models, $\varepsilon - SVR$ model and LS-SVR model were used, where model parameters are optimized by PSO algorithm. Distribution of data point samples for test is shown in Figure 1 (a), test results of regression using multiple linear regression model is drawn in Figure 1 (b) , the test results of regression $\varepsilon - SVR$ model is shown in Figure 1 (c), the test results of the LS-SVR regression model is described in Figure 1 (d) .



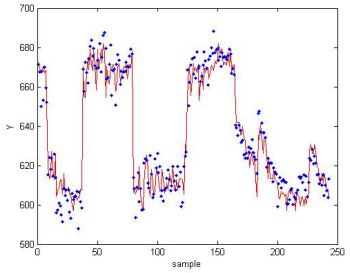
(a) Data points of samples



(b) multiple linear regressions



(c) $\varepsilon - SVR$



(d) LS-SVR

Fig. 1 Distribution of data point samples and several test results using different methods.

Comparison of the test results for three models is shown in Table 2.

Table 2 Comparison of parameters selection between PSO and grid search

Method	C	γ	ε	ME	Time(s)
MLR	–	–	–	28.000	0.003
$\varepsilon - SVR$	13	0.5	7.683	20.930	0.437
LS-SVR	9.484	0.13	–	20.940	0.074

As is shown in Table 2, multiple linear regression models has the worst testing accuracy, however, the time spent on training is the least, because its training process is simple matrix operations. $\varepsilon - SVR$ model has the best test accuracy, but also have the most expensive training time, as the method need to solve a constrained quadratic problem, and the number of constraints is equal to sample size. LS-SVR has a slightly worse test accuracy than the $\varepsilon - SVR$, but the training time is far less than $\varepsilon - SVR$. So the conclusion is to select the different regression models according to different needs.

Acknowledgement

This research is supported by the National Natural Science Foundation of China (No. 50834009, No. 60973071 and No. 61272176).

References

[1] B.S Hong, L.T Fan and R.S John, Monitoring the process occurring of epoxy/graphite fiber composites with a recurrent neural network as a soft sensor, Engineering Application of Artificial Intelligence, 11, 293-306(1998).
[2] L.C Zhou and H.Z Yang, Soft Sensor Modeling Based on KPCA and Support Vector Machines, Computer Simulation,25(10), 94-97(2008).
[3] R.L Liu, S.J Mu, H.Y Su and J Chu, Modeling soft sensor based on support vector machine and particle swarm optimization algorithms, 23(6), 895-906(2006).
[4] X.H Hao, Z. Wen, Y.M An, G.F She and N.F Xue, On Line Real-Time Computing Model and Application of Continuous Casting Slab Charging Temperature., Journal of System Simulation,19(14), 3324-3326(2007).
[5] X.H Hao, Z. Wen and Y.M An, The Comparative Studies between Heat Transfer Mathematical Model and BP Networks Prediction Model during Solidification of Continuous Casting Slab, 4, 10-14(2006).
[6] C. Bozzanca, S. Licitra, and L.Fortuna, Neural Networks for Benzene Percentage Monitoring in Distillation Columns. Proc. of the Third Int, 12, 391-395(1999).
[7] L. Jiang, D.H Wang and L.H Zhu, Study on Soft Sensor Prediction Model for Slab Temperature Based on RBF Neural Network, Hot Working Technology, 38(21), 93-96(2009).
[8] C. Douglas, N.Ngan, Locating Facial Region of a Head-and-Shoulders Color Image, Automatic Face and Gesture Recognition, Proc. Third Int'l Conf, 124-129(1998).

Huiyan Jiang is professor of computer science at Northeastern University, China. Director of Research Laboratory of Multimedia Medical Information Processing Technology, Member of China Society of Image and Graphics. Her main research fields include image analysis, computer-aided diagnosis(CAD), 3D visualization and pattern recognition. She has published research articles in reputed international journals of computer and engineering sciences. She is editor of AMT(Advances in Multimedia Technology).

Fengzhen Tang received her Master of Engineering degree in 2011 in Northeastern University, China, a member of Pattern Recognition group in MMIT, and a PhD candidate in Birmingham University, UK. She is now focusing on machine learning and data mining, large-scale optimization algorithms.

Lingbo Zou received his Bachelor of Engineering degree in 2010 in Northeastern University, China, a member of Pattern Recognition group in MMIT. He is now focusing on machine learning and data mining, large-scale optimization algorithms, tumor detection and image analysis.

Yenwei Chen received a D.E. degree in 1990, both from Osaka University, Osaka, Japan. He is currently a professor with the college of Information Science and Engineering, Ritsumeikan University, Kyoto, Japan. His research interests include intelligent signal and image processing, radiological imaging and soft computing.