Product Quality Modelling Based on Incremental Support Vector Machine

J Wang¹, W Zhang², B Qin³ and W Shi⁴
Mechanical Engineering School, Inner Mongolia University of Science and Technology, Baotou, China

E-mail: ¹wjgkyc@imust.cn, ²manifold@imust.cn, ³nkdqb@163.com, ⁴yanstone8697@yahoo.com.cn

Abstract. Incremental Support vector machine (ISVM) is a new learning method developed in recent years based on the foundations of statistical learning theory. It is suitable for the problem of sequentially arriving field data and has been widely used for product quality prediction and production process optimization. However, the traditional ISVM learning does not consider the quality of the incremental data which may contain noise and redundant data; it will affect the learning speed and accuracy to a great extent. In order to improve SVM training speed and accuracy, a modified incremental support vector machine (MISVM) is proposed in this paper. Firstly, the margin vectors are extracted according to the Karush-Kuhn-Tucker (KKT) condition; then the distance from the margin vectors to the final decision hyperplane is calculated to evaluate the importance of margin vectors, where the margin vectors are removed while their distance exceed the specified value; finally, the original SVs and remaining margin vectors are used to update the SVM. The proposed MISVM can not only eliminate the unimportant samples such as noise samples, but also can preserve the important samples. The MISVM has been experimented on two public data and one field data of zinc coating weight in strip hot-dip galvanizing, and the results shows that the proposed method can improve the prediction accuracy and the training speed effectively. Furthermore, it can provide the necessary decision supports and analysis tools for auto control of product quality, and also can extend to other process industries, such as chemical process and manufacturing process.

1. Introduction

With the increase in data acquisition systems on production process, the collection of process data during production runs is becoming routine and plants are now becoming data rich but information poor [1,2]. Consequently, there is a real need to extract useful information and build empirical models between the past product quality and the process parameters for quality prediction. Moreover, the empirical models based on data are very important for obtaining high quality production, because they can enable plant operators to understand the status of the process and predict the quality before manufacturing.

Support vector machine is a new empirical model based on statistical learning theory and optimization methods; it aims to generate an optimal separating hyperplane by minimizing the generalization error. The decision hyperplane of SVM is determined by the most informative data instances, called support vectors (SVs). Actually, these SVs are just a subset of the entire training data which determine the generalization property of a SVM. In recent years, SVM has become one of the
most useful methods of solving the problems in machine learning with good generalization performance [3], and is widely applied in face recognition [4], fault diagnosis [5], etc.

However, with the increase in the size of the real-world data set, the training procedure of SVM always requires huge memory space and significant computation time. In order to solve this problem, many researches proposed incremental training methods to shorten the training time. The goal is to select a subset of training samples while preserving the performance as using all the training samples. Incremental learning for SVM was first introduced by Syed et al. [6], who presented an incremental learning procedure by partitioning the training samples into several subsets, called batch-SVM. At each incremental step, only SVs are added to the next incremental samples to retrain the SVM. In contrast to traditional SVM training method, the training samples are greatly reduced. But not all the incremental samples are useful for retraining the SVM, such as the redundant similar samples and noise samples. So Zeng et al. [7] proposed a new Incremental SVM learning method by KKT condition mutual inspection; here we called K-ISVM. The main concept is that two SVMs are trained by training samples and incremental samples respectively; and then the two SVMs are used to inspect the two subset samples to find out which samples violate the KKT condition; finally, the SVs of the two SVMs and the samples which violate the KKT condition are used to train a new SVM. It usually thinks that the samples, which violate the KKT condition, contain useful information that the original SVM has not learned before. But noise samples would also violate the KKT condition, and the operation of mutual inspection will take much time.

In this paper, we propose a modified incremental support vector machine (MISVM) for product quality prediction. Firstly, we train the SVM and collect the support vectors (SVs). Then the trained SVM is applied to examine the incremental samples, and the margin vectors which violate the KKT conditions are chosen. Thirdly, the distance between the margin vectors and the margin hyperplane is computed to evaluate the importance of each margin vectors, and the margin vectors whose distance is below the specified value are preserved, the others are eliminated. Finally, the original SVs and the remaining margin vectors are used to train a new SVM. The proposed MISVM can not only eliminate the unimportant samples such as noise samples, but also preserved the important samples. The superiority of MISVM is illustrated through three examples, two public data and one field data of zinc coating weight in strip hot-dip galvanizing.

The organization of this paper is as follows. The next section describes the theory of SVM for regression. In section 3, we describe the proposed the MISVM method. In section 4, the method of MISVM is experimentally evaluated. Finally, we conclude this paper in section 5.

2. Support vector machine for regression

SVM is proposed to solve the classification problems at first. With the $\varepsilon$-insensitive loss function developing, it is extended to nonlinear system regression estimation problem, which shows good generalization performance [8].

Define the $\varepsilon$-insensitive loss function as below:

$$
|y - f(x)| = \begin{cases} 
0, & \text{if } |y - f(x)| \leq \varepsilon \\
|y - f(x)| - \varepsilon, & \text{if } |y - f(x)| > \varepsilon
\end{cases}
$$

(1)

where $\varepsilon$ is width of the insensitivity tube.

Only the points outside the $\varepsilon$ region contribute to the cost insofar, as the deviations are penalized in a linear fashion.

Let $T = \{ (x_i, y_i) | x_i \in \mathbb{R}^n, y_i \in \mathbb{R} \}$ be the training set for a regression problem, where $x_i$ the sample is point in an $n$-dimensional space, and $y_i$ is the target point. By defining a kernel matrix $K_{ij} = K(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j) \rangle$, the SVM can solve a nonlinear regression problem as a linear
one in a high dimensional feature space. We define a nonlinear decision function \( f(x) \) in feature space such that:

\[
f(x) = w \cdot \varphi(x) + b
\]  

(2)

The standard regression problem of a nonlinear SVM is as follows:

\[
\begin{align*}
\min & \quad \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{l} (\xi_i + \xi_i^*) \\
\text{s.t.} & \quad y_i - w \cdot \varphi(x_i) - b \leq \varepsilon + \xi_i \\
& \quad w \cdot \varphi(x_i) + b - y_i \leq \varepsilon + \xi_i^* \\
& \quad \xi_i, \xi_i^* \geq 0, i = 1, 2, \cdots, l 
\end{align*}
\]  

(3)

Where \( C > 0 \) is the penalty parameter, it determines the trade-off between the flatness of \( f \) and the amount up to which deviations larger than \( \varepsilon \) are tolerated. \( \xi_i, \xi_i^* \) are the slack variables which introduce to cope with otherwise infeasible constraints.

The dual is:

\[
\begin{align*}
\min & \quad \frac{1}{2} \sum_{i,j=1}^{l} (a_i - a_i^*)(a_j - a_j^*) K(x_i, x_j) \\
& \quad + \varepsilon \sum_{i=1}^{l} (a_i + a_i^*) - \sum_{i=1}^{l} y_i (a_i - a_i^*) \\
\text{s.t.} & \quad \sum_{i=1}^{l} (a_i - a_i^*) = 0 \\
& \quad a_i, a_i^* \in [0, C]
\end{align*}
\]  

(4)

After training, the decision function of the SVM can be written as:

\[
f(x) = \sum_{x_i \in SV} (a_i - a_i^*) K(x_i, x) + b
\]  

(5)

Where \( a_i, a_i^* \) are the Lagrange multipliers. If \( a_i \neq 0 \), or \( a_i^* \neq 0 \), their corresponding samples are called Support Vectors. It is often thought that SVs provide a sufficient representation of the samples for the given regression task. So solution is sparseness, this has guiding significance to incremental learning algorithm constructing [9].

3. **Modified incremental support vector machine for regression**

A traditional incremental learning algorithm is described as below [10]:

Presupposition: there exist initial training set \( A \) and incremental set \( B \), where \( A \cap B = \emptyset \). SVMA is the model of SVM which is trained by the training set \( A \), and SVsA are the corresponding support vectors.

Goal: finding out the margin support vectors in set \( B \) which violate the KKT condition of SVMA. The KKT conditions of the support vector machine for regression are as follows:

\[
\begin{align*}
& a_i = 0, \quad a_i^* = 0 \quad \Rightarrow |y_i - f(x_i)| < \varepsilon \\
& 0 < a_i, a_i^* < C \quad \Rightarrow |y_i - f(x_i)| = \varepsilon \\
& a_i = c, \quad a_i^* = c \quad \Rightarrow |y_i - f(x_i)| > \varepsilon
\end{align*}
\]  

(6)
The figure 1 shows the samples distribution of SVM, where the circular points are the initial data and the pentagon points are the incremental data, the thick solid line is the decision hyperplane, and the thin solid line is the margin hyperplane.

It is well to know that the vectors outside the $\varepsilon$ region will violate the KKT condition [11]. So, there few vectors of a trained SVM will exceed the $\varepsilon$ region. According to this character, we can think that, if one of the incremental samples violates the KKT condition, it may contain new information that the original SVM have not learned before. But if the sample violates the KKT condition greatly, it may be a noise sample.

The proposed MISVM are using the distance from the margin vectors which violate the KKT condition to the margin hyperplane to evaluate the importance of each margin vectors, and the margin vectors whose distance is below the specified bias are preserved, the others are eliminated. The detailed process is as below:

1. Use training set A to train the SVM which called SVMA, SVsA are the corresponding support vectors.
2. Apply the SVMA to inspect the incremental Set B, and then find out the margin vectors which violate the KKT condition in Set B.
3. Let a small positive value $D$ as the distance bias, where $D > \varepsilon > 0$.
4. Computer the distance
   \[ d(x_i) = |y_i - f(x_i)|, \text{where } f(x_i) = \sum_{k \in SV} (a_k - a_k^*) K(x_i, x_k) + b \]  
   (7)
5. Find out the margin vectors in set B which satisfy $\varepsilon \leq d(x_i) \leq D$, here we call SVsB;
6. Treat the SVsA and SVsB as new training sample to train a new SVM, end.

4. Experiments and comparisons
In this section, we present three experiments for testing our algorithm. The experimental data of two public data of Boston Housing and Computer Hardware and one field data of zinc coating weight in strip hot-dip galvanizing.

4.1. Experiments of two public dataset
In this section, we present two experiments for testing our rule extraction method. The experimental data of Boston Housing and Computer Hardware come from the University of California at Irvine [12]. Subsequently, we make comparisons with SVM, Batch-SVM and MISVM.
Boston Housing data contains 13 attributes, where there are 12 continuous attributes and 1 discrete attribute. The goal is to predict the houses average price of Boston. There are 506 samples in the dataset, where 400 samples are regarded as training set and the others as testing set. In the training set, we split the data into 4 subsets averagely, one subset is as the initial training data, and the other 3 subsets are as the incremental data. The Computer Hardware dataset contains six continuous attributes. The goal is to predict the CPU’s relative performance based on the other computer characteristics. There are 209 samples in the dataset, where 180 samples are regarded as training set and the others as testing set. In the training set, we choose 60 samples as the initial training samples randomly, and the other 120 samples are divided into 3 subsets averagely as the incremental samples.

In order to evaluate the experiment results and make comparisons with other methods, the mean square error (MSE) is applied as follows:

\[
\text{MSE} = \frac{1}{k} \sum_{i=1}^{k} (y_i - y_i^*)^2
\]  

Where \(y_i^*\) is the predicted value corresponding to the input sample \(x_i\).

The distance bias \(D\) is a very important value, so in this paper, we should make experiments to choose the optimized value. The performance of distance bias \(D\) is depicted in figure 2 and figure 3.

**Figure 2.** The performance of the distance bias \(D\) in Housing dataset.

**Figure 3.** The performance of the distance bias \(D\) in Computer Hardware dataset.
The two figures show that the best distance bias are $D = 0.24$ and $D = 0.08$ respective in Housing dataset and Computer dataset. In order to demonstrate the superiority of the proposed MISVM, we also make comparisons with SVM, Batch-SVM, and MISVM in prediction accuracy and training time (Units: second). The results of the comparisons are depicted in table 1, table 2, table 3 and table 4.

Two experimental results show that the incremental learning methods all have faster training speed than the traditional SVM; especially the proposed MISVM has not only the highest accuracy but also the fastest training speed. In these two experiments, we found that the prediction accuracy and the distance bias $D$ have high correlation. Take Housing dataset for example, the prediction accuracy increase with increased distance bias $D$ first, but after $D = 0.24$, the prediction accuracy decrease with increased distance bias $D$. The phenomenon demonstrates that the small value $D$ means little useful incremental samples, it will cause few information has been learned in SVM training; the big value $D$ means many useful incremental sample, it will cause over learning. So the best bias $D$ is the focal point of the proposed MISVM.

| Table 1. | Training Time of Different Methods in Housing Dataset. |
|----------|--------------------------------------------------------|
| Item     | SVM | Batch-SVM | K-ISVM | MISVM |
| Initial training 100 | 6.5 | 6.5 | 6.5 | 6.5 |
| 1st incremental 100 | 73.5 | 9.3 | 8.8 | 4.5 |
| 2nd incremental 100 | 440.0 | 16.4 | 14.7 | 7.6 |
| 3rd incremental 100 | 1233.6 | 21.4 | 15.5 | 9.9 |

| Table 2. | Prediction Accuracy of Different Methods in house dataset. |
|----------|----------------------------------------------------------|
| Item     | SVM | Batch-SVM | K-ISVM | MISVM |
| Initial training 100 | 0.0104 | 0.0104 | 0.0104 | 0.0104 |
| 1st incremental 100 | 0.0104 | 0.0211 | 0.0099 | 0.0075 |
| 2nd incremental 100 | 0.0117 | 0.0129 | 0.0115 | 0.0073 |
| 3rd incremental 100 | 0.0112 | 0.0172 | 0.0079 | 0.0071 |

| Table 3. | Training Time of Different Methods in Housing Dataset. |
|----------|--------------------------------------------------------|
| Item     | SVM | Batch-SVM | K-ISVM | MISVM |
| Initial training 100 | 0.6 | 0.6 | 0.6 | 0.6 |
| 1st incremental 100 | 1.4 | 1.0 | 0.9 | 0.8 |
| 2nd incremental 100 | 3.0 | 1.5 | 1.4 | 1.2 |
| 3rd incremental 100 | 7.5 | 2.2 | 1.8 | 1.6 |

| Table 4. | Prediction Accuracy of Different Methods in Computer Hardware Dataset. |
|----------|----------------------------------------------------------|
| Item     | SVM | Batch-SVM | K-ISVM | MISVM |
| Initial training 100 | 0.0047 | 0.0047 | 0.0047 | 0.0047 |
| 1st incremental 100 | 0.0057 | 0.0029 | 0.0039 | 0.0027 |
| 2nd incremental 100 | 0.0070 | 0.0049 | 0.0038 | 0.0020 |
| 3rd incremental 100 | 0.0088 | 0.0032 | 0.0030 | 0.0020 |
4.2. Experiment of Zinc Coating Weights

The zinc coating weights are the important quality of the strip hot-dip galvanizing. The air-knife is the key equipment to control the zinc coating weights. The prediction model of the zinc coating weights through the parameters of the air-knife has important significance to establish the reasonable set value, save raw materials and improve the production efficiency. The air knife is shown as figure 4.

A pair of air wiping jets is used to wipe excess zinc from the surface of the strip while the strip passes through a molten zinc pot. The technique of the air wiping has crucial influence on zinc coating weight, so we model the process of zinc coating production via the technology of the air wiping. The height of the air knife $H$ and the angle of the air knife $\theta$ are changeless, so they have no effects to the weight of zinc layer. According to industrial process, jet nozzle pressure $P$, nozzle to strip distance $d$, strip velocity $V$ and strip thickness $B$ are the most significant parameters. We model the production process of cold rolling strip hot-dip galvanizing based on the real industrial data which are obtained from a production line of an iron and steel corporation [13].

![Figure 4. Sketch of the air knife.](image)

The data set has four continuous attributes: $P$, $d$, $V$ and $B$. The goal is to predict the weight of the zinc layer according to those four technique parameters. There are 4 subsets samples in the data set corresponding to different production period respectively. Each subset contain 500 samples, where 400 samples are treated as training samples and 100 samples are treated as testing samples. The statistical analysis result of the original data is depicted in table 5.

| Variable | Subset 1 | Subset 2 | Subset 3 | Subset 4 |
|----------|----------|----------|----------|----------|
| $P$      | MAX: 25.5 | MIN: 8.8 | Mean: 16.4 | MAX: 33.3 | MIN: 9.6 | Mean: 19.8 | MAX: 44.3 | MIN: 7.5 | Mean: 25.8 | MAX: 46.9 | MIN: 10.5 | Mean: 31.6 |
| $d$      | MAX: 32.2 | MIN: 11.0 | Mean: 18.3 | MAX: 32.2 | MIN: 10.9 | Mean: 16.1 | MAX: 30.2 | MIN: 10.0 | Mean: 13.6 | MAX: 32.7 | MIN: 9.9 | Mean: 12.8 |
| $V$      | MAX: 72.0 | MIN: 40.9 | Mean: 63.2 | MAX: 120.0 | MIN: 37.8 | Mean: 87.0 | MAX: 150.0 | MIN: 60.0 | Mean: 12.1 | MAX: 150.0 | MIN: 43.9 | Mean: 134.4 |
| $B$      | MAX: 2.5 | MIN: 1.73 | Mean: 1.94 | MAX: 1.55 | MIN: 1.09 | Mean: 1.29 | MAX: 1.00 | MIN: 0.65 | Mean: 0.86 | MAX: 0.64 | MIN: 0.35 | Mean: 0.50 |
| Zinc weight | 146.2 | 48 | 88.1 | 149.0 | 40.0 | 90.7 | 145.0 | 47.0 | 88.0 | 144.0 | 42.0 | 79.8 |

$a$ unit: Kpa.  
$b$ unit: mm.  
$c$ unit: m·min$^{-1}$.  
$d$ unit: g·m$^{-2}$

We apply the 400 samples of subset 1 to train the first SVM model, and then to predict the rest 100 testing samples. In order to show the advantage of Incremental SVM, the first SVM is also tested by the others subset testing samples. Then we apply the proposed MISVM method to make incremental
training with the training samples of subset 2, subset 3, and subset 4 respectively, the prediction results are depicted in table 6.

According to table 6, we can find that the prediction accuracy of the first SVM is high in subset 1, and the accuracy of the other subsets is comparatively poor. After incremental training, the prediction accuracy is improved significantly. For example, the prediction accuracy is improved from 0.0303 to 0.0083 in subset 2. The prediction accuracy is getting better after incremental training; the effect is depicted in subset 4. So the three incremental training demonstrate that the proposed MISVM is suitable and effective for product quality modelling in process industry.

Table 6. The detailed results of zinc weight.

| Item               | Subset 1 | Subset 2 | Subset 3 | Subset 4 |
|--------------------|----------|----------|----------|----------|
| Initial training   | 0.0073   | 0.0303   | 0.0401   | 0.0325   |
| 1st incremental    | -        | -        | 0.0083   | 0.0132   |
| 2nd incremental    | -        | -        | 0.0055   | 0.0132   |
| 3rd incremental    | -        | -        | -        | 0.0119   |

5. Conclusions
In this paper, to reduce memory cost by deleting unnecessary data, we have presented a novel method of modified incremental support vector machine called MISVM. The distance from the margin vector to the decision hyperplane is used to evaluate the importance of each margin vectors; the advantages are that the samples which contain useful information can be learned to update the original SVM and the redundant samples are deleted. In order to show the effectiveness of the proposed MISVM, we make experiments on to public dataset, and the results show that training data can be deleted greatly while keeping the generalization ability comparable to Batch-SVM and K-ISVM, the prediction accuracy and the training speed is improved. Finally, we apply the MISVM method to build product quality model for the zinc coating weight in strip hot-dip galvanizing process, and obtain the good performance. Furthermore, it can provide the necessary decision supports and analysis tools for auto control of product quality, and also can extend to other process industries, such as chemical process and manufacturing process.

Acknowledgments
This research is supported in part by the National Natural Science Foundation of China (No. 51005014), Natural Science of Foundation Inner Mongolia (No.2011ZD08), Inner Mongolia High School Research Project (No.JZY11150), Baotou Science and Technology Bureau intellectual Property Development Project (2011Z1006), Intramural Foundation of Inner Mongolia University of Science and Technology (No.2009NC032 and No.2010NC032).

References
[1] Chiang L H, Russell E L, and Braatz R D 2011 Fault detection and diagnosis in industrial systems (London: Springer-Verlag)
[2] Jia F, Martin E B and Morris A J 2000 International Journal of Systems Science 31 1473
[3] Duan H, Shao X, Hou W, He G and Zeng Q 2009 Pattern Recognition Letters 30 p 1284
[4] Osuna E, Freund R and Girosit F 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Massachusetts: IEEE) pp 33–62
[5] Yuan S and Chu F 2007 Journal of Barton shock, 26 29
[6] L Shen 2008 Application of particle swarm optimization and support vector machine in simulating and predicting stream water Quality (Zhejiang, Zhejiang Normal University)
[7] Syed N, Liu H and Sung K IEEE International Conference on Data Mining (San Jose, California, 29 November - 02 December) Morgan Kaufmann publishers pp. 352-356.
[8] Zeng G W and Ma J 2002 Journal of XiAm University 41 687
[9] Du S and Wu T 2003 Journal of system Simulation 15 1580
[10] Deng N 2004 *A new method of data mining: support vector machine* (Beijing: Science Press)

[11] Li K and Huang H 2003 *Journal of Northern Jiaotong University* **27** 34

[12] Zhou W, Zhang L and Jiao L 2001 *Acta electronica Sinica* **27** 590

[13] http://archive.ics.uci.edu/ml/index.html. UCI–MLR: University of California Irvine Machine Learning Repository, California Irvine

[14] ZHU L 2006 *Steel Materials Heat Coating Zinc* (Beijing: Beijing Chemical Industry Press)