Peer review statement

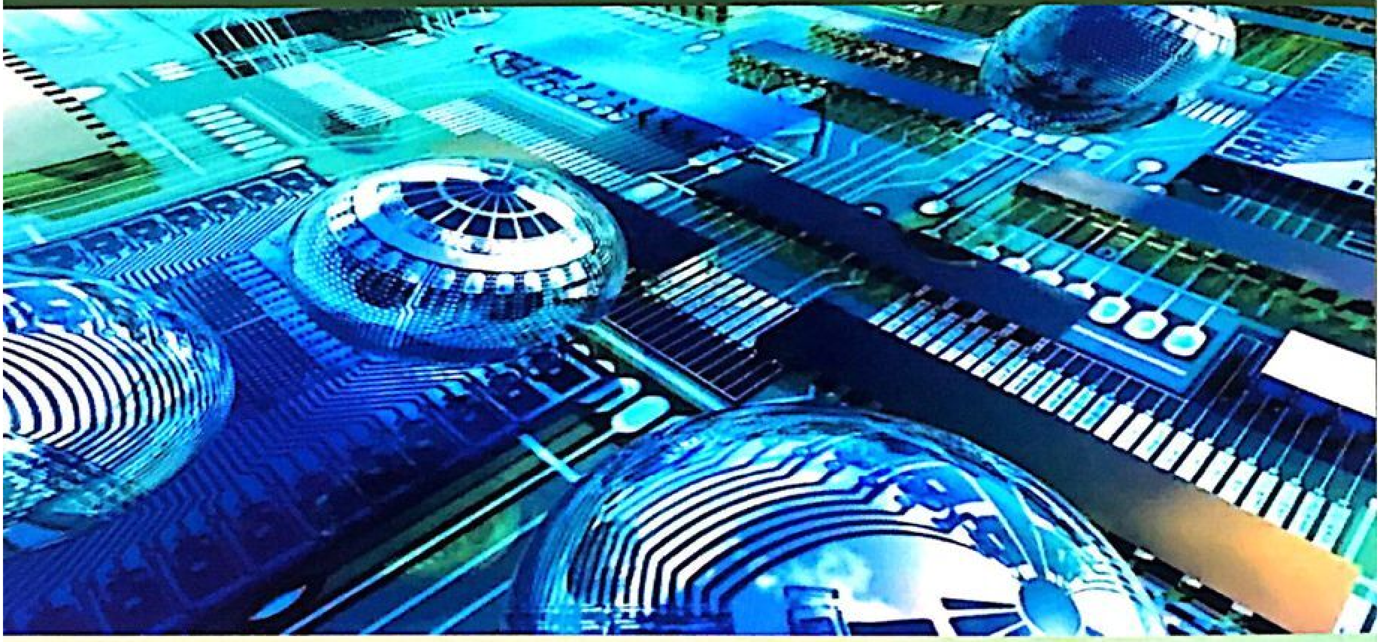
All papers published in this volume of *IOP Conference Series: Materials Science and Engineering* have been peer reviewed through processes administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a proceedings journal published by IOP Publishing.

IMMAEE 2018

SEPTEMBER 15-16, 2018 - MELBOURNE, AUSTRALIA

2018 3RD INTERNATIONAL CONFERENCE ON INSULATING MATERIALS, MATERIAL APPLICATION AND ELECTRICAL ENGINEERING

Volume 452, Year 2018



IOP
Institute of Physics

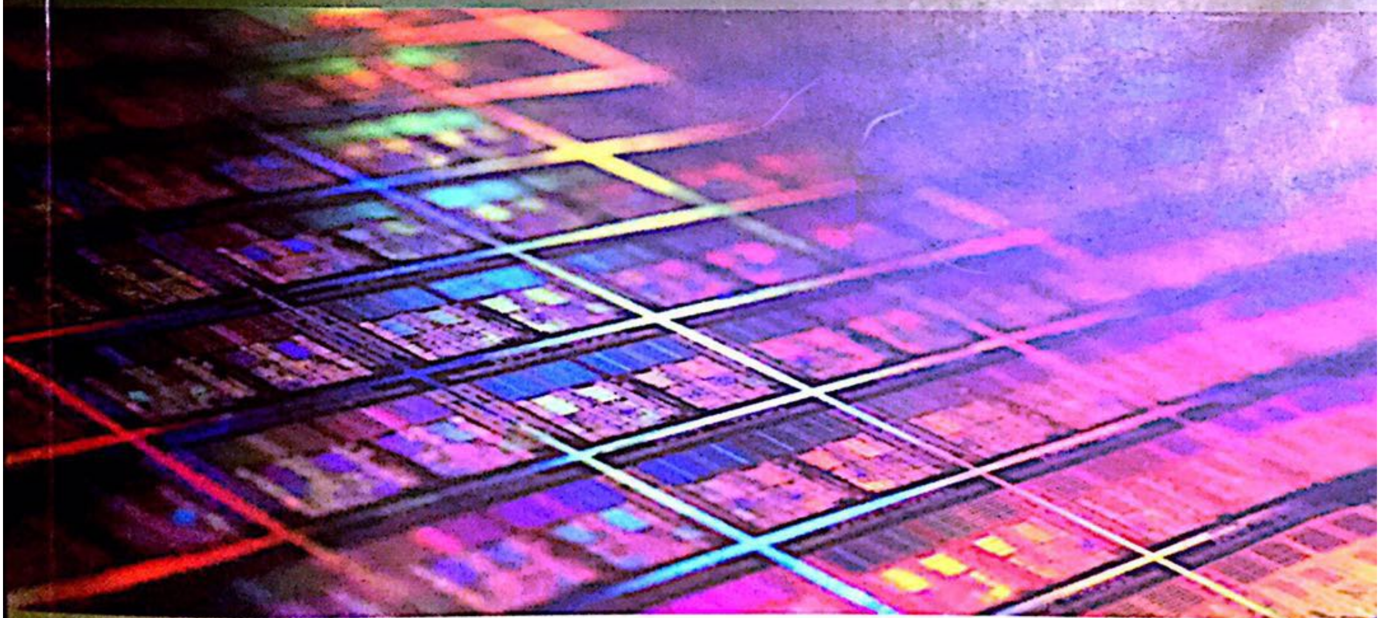
IOP Conference Series
Materials Science and Engineering
ISSN: 1757-8981

IOP Conference Series

Materials Science and Engineering

ISSN: 1757-8981

Volume 452, Year 2018



IOP Institute of Physics

Study on Handwritten Digit Recognition using Support vector machine

Xiaoning Zhou[1, *], Jie Li[2, a], Changjun Yang [3,4, b] and Jianming Hao[1, c]

[1]Nanhang Jincheng College, Nanjing, China
[2]College of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Nanjing, China
[3]Key Laboratory of Radiometric Calibration and Validation for Environmental Satellites, China Meteorological Administration (LRCVES/CMA) , Beijing 100081, China
[4]National Satellite Meteorological Centre, Beijing 100081, China

*Corresponding author e-mail: zxn0215@163.com, [a]jetly@126.com, [b]yangcj@cma.gov.cn, [b]alan8409@126.com

Abstract. A machine learning model of handwritten digit recognition based on SVC is established in this paper. Then the influence of sample number, kernel function parameters, penalty coefficients and other parameters on the prediction model is analysed. This results show that training samples have a significant impact on the model. There is an acceptable training number. Different kernel functions have a different effect on the accuracy of the model. The radial basis function is the best in recognition model. The recognition rate increases continuously with C, while the recognition rate increases first with gamma increases, and when gamma increases to a certain value, precision begins to decline.

1. Introduction

Image is one of the important means for information to express, store and transmit. Along with the rapid development of the computer technology and digital image processing technology, digital image classification and recognition techniques are required urgently. Image recognition, widely applied to handwriting recognition, face recognition, vehicle license plate recognition and so on, is the use of AI technology to enable computers to recognize information in images. For instance, the methods to recognize handwritten date, account digit and other numeric information are the key technique of image recognition.

Support vector machine (SVM) is a learning method based on statistical learning theories, which proposed by Vapnik. Basing on the principle of structural risk minimization, SVM can improve the generalization ability of the learning machine as much as possible. Even the decision rules obtained from limited training samples can still get small errors for independent test datasets. In recent years, SVM has been widely used in pattern recognition, regression analysis and feature extraction. Vapnik found that different kernel functions had little effect on SVM performance. The key factors affecting SVM performance are kernel function parameters and penalty coefficient. Therefore, the study of kernel function parameters and penalty coefficients is an important field to improve the performance of machine learning. A machine learning model of handwritten numeric information based on SVC is established in this paper. Then the influence of training number, kernel function and penalty coefficients parameters on the prediction model is analysed, providing reference for improving efficiency and accuracy of machine learning model.

2. Handwritten digits recognition based on SVM

2.1 The theory of SVC

A hyper-plane or set of hyper-planes in a high or infinite dimensional space, which can be used for classification, regression or other tasks, is constructed by a support vector machine. In general, since the larger margin the lower generalization error of the classifier, the hyper-plane, which has the largest distance to the nearest training data points of any class, achieve a good separation.

SVM, including C-SVC, v-SVC, one-class SVM, e-SVR, v-SVR and so on, is a set of supervised learning methods used for classification, regression and border detection. SVCs are classes capable of performing multi-class classification on a dataset.

Given training vectors $x_i \in R^n, i=1,\ldots,l$, in two classes, and a vector $y \in R^l$, and $y \in \{1,-1\}$, SVC solves the following primal problem:

$$\begin{aligned} &\min_{w,b,\xi} \frac{1}{2} w^T w + C\sum_{i=1}^{l} \xi_i \\ &subject\ to \quad y_i(w^T \cdot \phi(x_i)+b) \geq 1-\xi_i \\ &\xi_i \geq 0, i=1,\ldots,l \end{aligned} \tag{2-1}$$

Its dual is

$$\begin{aligned} &\min_{\alpha} \frac{1}{2} \alpha^T Q\alpha - e^T \alpha \\ &subject\ to \quad y^T \alpha = 0 \\ &0 \leq \alpha \leq C, i=1,\ldots,l \end{aligned} \tag{2-2}$$

Where *e* is the vector of all ones, $C > 0$ is the upper bound, *Q* is *n* by *n* positive semi-definite matrix, $Q_{ij} = y_i y_j K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ is the kernel. Here training vectors are implicitly mapped into a higher (maybe infinite) dimensional space by the function ϕ.

The decision function is:

$$\operatorname{sgn}\left(\sum_{i=1}^{l} y_i \alpha_i K(x_i, x) + b \right) \tag{2-3}$$

2.2 Kernel function

To solve specific problems, it is very important to choose the appropriate kernel function. There are four types of kernel functions:

1) Linear kernel function

$$k(x,z) = x \cdot z \tag{2-4}$$

2) Polynomial kernel function

$$k(x,z) = (x^T z + c)^d \tag{2-5}$$

3) RBF (Radial Basic Function)

$$k(x,z) = \exp(-\gamma |x - z|^2) \tag{2-6}$$

4) Sigmoid kernel function

$$k(x,z) = \frac{1}{1 + \exp(vx^T - c)} \tag{2-7}$$

There is no unified model for Kernel function selection. Kernel function only can be determined through experimental comparison, according to the characteristics of specific samples, for better parameters and more generalization performance. In some cases, the performance of kernel function is satisfied, while others are quite bad. For all chosen kernel functions, better accuracy can be achieved as long as the appropriate parameters are setted. Usually, it is appropriate to choose radial basis function (RBF). At present, the cross validation method is used to select the appropriate kernel function. The best kernel function is that with the least inductive error.

3. Recognition process and experiment design

3.1 Recognition process

The main steps of handwritten numeric recognition are as follows:

Firstly, handwritten digits are analyzed, and a series of pre-processing is carried out to extract a set of feature vectors. Then, those training feature vectors are sent to the training I/O of SVM to train the parameters and support vectors. After that, the feature vectors of the handwritten digits, which needs to be recognized, are extracted and sent to the prediction I/O of SVM. At last, the prediction results are outputted and the recognition rate is calculated (fig1).

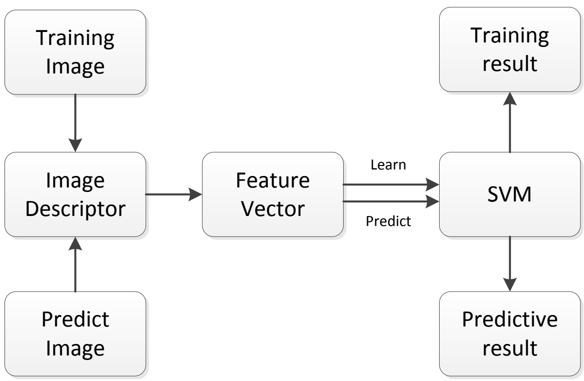


Figure 1. The flow chart of recognition using Support Vector Machine

3.2 Experiment design

3.2.1 Dataset Information

Collecting handwritten digits dataset is a very time-consuming task. The dataset is a copy of the test set of the UCI ML hand-written digits datasets. The digits dataset consists of 1797 images, in which each image is an 8*8 pixel one representing a handwritten digit. Those images in dataset are divided into 10 classes, and each class refers to a digit, is a number in the range of 0 to 9 (fig 2). The test set was used for writer-independent testing and is the actual quality measure.

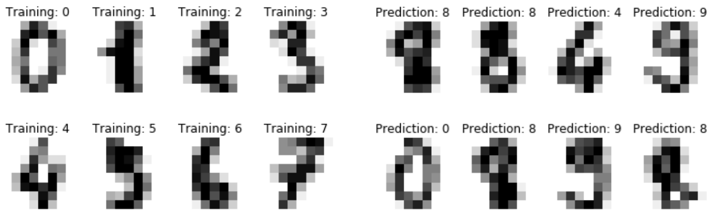


Figure 2. The training data examples (left) and forecast data examples (right)

3.2.2 Machine learning library

At present, there are many machine learning libraries in Python, of which scikit-learn is the most famous, simple and efficient tools for data mining and data analysis, it can be accessible to everybody, and reusable in various contexts. In scikit-learn, an estimator for classification is a Python object that implements the methods fit(x, y) and predict(T), and some forecast result is show in fig 2 .

We are given samples of each of the 10 possible classes (the digits 0 through 9) on which we fit an estimator to be able to predict the classes to which unseen samples belong. The cache_size is set to 500(MB). When training SVM with the kernel function of RBF, C and gamma are considered.

4. Experiment analysis

4.1 Influence of training sample numbers

Training samples have a very significant impact on the model. When the number of training samples increases, the running time increases dramatically. The running time of the model also increases with the number of predict, but the influence is not significant as the training samples (fig 3).

As the number of training samples increases, the accuracy of the model will increase significantly. However, when the predicted model reach a certain precision, increasing the number of training samples cannot contribute to the accuracy of the model. When the prediction number are 300, 500 and 800 with the training samples increase from 100 to 600, and the accuracy increases from less than 90% to more than 94%. But the accuracy of the model changes little with training number increasing up to 600(fig 3). So, it is no need to collect unlimited number of samples for training, there is an acceptable training number.

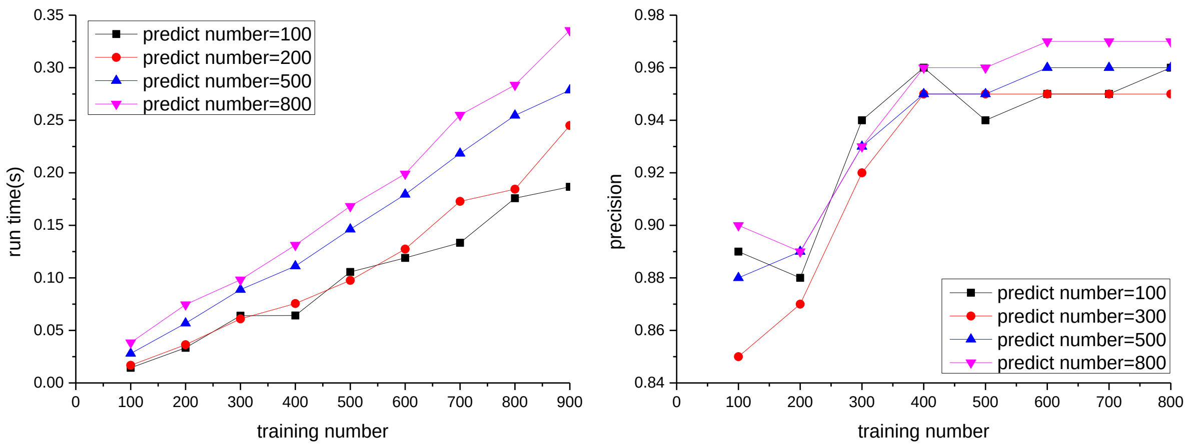


Figure 3. The relation of training number between run time (left) and precision (right)

4.2 Influence of kernel function

Different kernel functions have a different effect on the accuracy of the model. The experimental results show that the performance of RBF is the best. The recognition precision of handwritten digits can reach 93% - 100%, and the average recognition precision is 97% (fig 4). The recognition precision of Linear and Polynomial can also reach 83%-100%, and the average recognition precision is 93% and 94% respectively. However, the generalization performance of Sigmoid kernel function is weak, and the recognition precision vary from 48% to 95%, the precision of digits 2, 3 and 8 is much worse (no more than 55%), and the average precision is only 70%.

RBF has the strongest learning ability, linear and polynomial functions have the stronger learning ability, while sigmoid has the weakest learning ability.

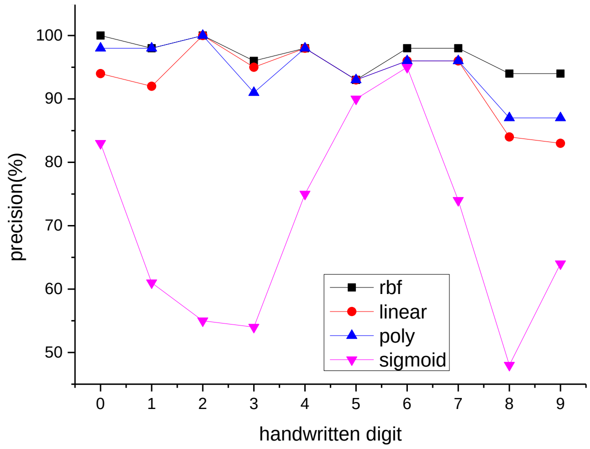


Figure 4. The precisions between different kernel functions

4.3 Influence of parameters

The setting of initial parameters is very important, and the parameters directly affect the generalization ability of SVM. It is necessary to select suitable parameters to train SVC models through tests.

When training an SVM with the RBF kernel, C and gamma is considered. The recognition rate increases continuously with C increases. When C increases to a certain extent, the recognition rate changes little, the maximum accuracy is 0.97. The recognition rate increases first with gamma increases, and when gamma increases to a certain value (about 0.001-0.002), precision begins to decline (fig. 5).

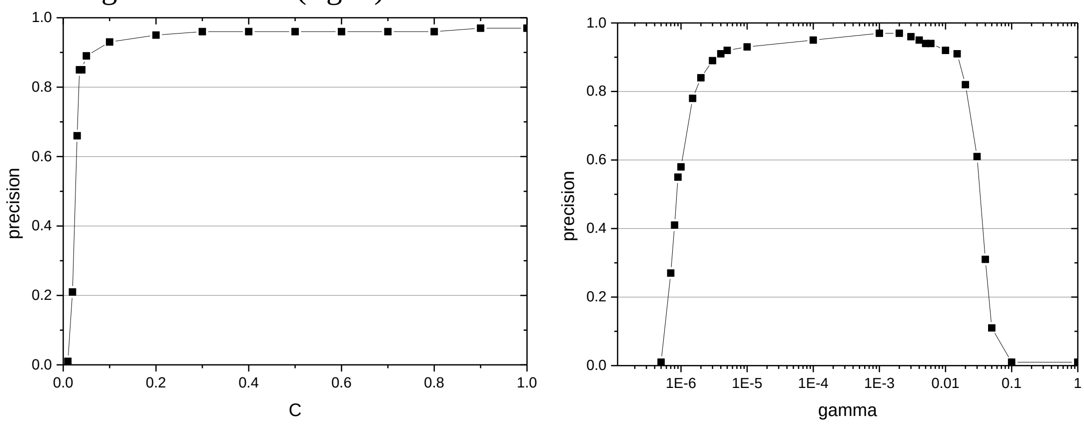


Figure 5. The relationship between C (left),gamma(right) and recognition precision

5. Conclusion

(1)Training samples have a significant impact on the model. As the number of training samples increases, the accuracy of the model will increase significantly. When the predicted model reach a best precision, increasing the number of training samples cannot improve the accuracy of the model. There is an acceptable training number.

(2)Different kernel functions have a different effect on the accuracy of the model. The experimental results show that the performance of RBF is the best. RBF has the strongest learning ability, linear and polynomial have the stronger learning ability, while sigmoid has the weakest learning ability.

(3) The recognition rate increases continuously with C. When C increases to a certain extent, the recognition rate changes little, the maximum accuracy is 0.97. The recognition rate increases first with gamma increases, and when gamma increases to a certain value, precision begins to decline.

This work was financially supported by xxx fund.

References

[1] Vladimir N. Vapnik An Overview of Statistical Learning Theory. IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 10, NO. 5, SEPTEMBER 1999

[2] Vladimir N.Vapnik.The Nature of Statistical Learning Theory, Springer, N Y, (1995)

[3] Stitson M O et al .Theory of support vector machines, Technical Report CSD-T R-96-17, Depar tment of Computer Science, Roy al Holloway University of London ,(1996)

[4] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[5] C. Kaynak. Methods of Combining Multiple Classifiers and Their Applications to Handwritten Digit Recognition, MSc Thesis, Institute of Graduate Studies in Science and Engineering, Bogazici University (1995).

[6] Guyon I, Boser B E, Vapnik V. Automatic Capacity Tuning of Very Large VC-dimension Classifiers[J]. Advances in Neural Information Processing Systems, 1992, 5:147--155.