Research on SVM Remote Sensing Image Classification Based on Parallelization

Qiuhong Yang¹*

¹PLA Army Armored Force Academy Bengbu Campus,Bengbu Anhui,233000

*Corresponding author e-mail: yangqiuhong20201@aafe.cn

Abstract. With the development of technology, the feature dimension and data volume of remote sensing image classification have increased rapidly. However, when remote sensing image classification based on support vector machine (SVM) is used for large-scale data calculation, there are significant limitations in training time. This paper focuses on the parallel processing method of support vector machine (SVM). Based on the popular hybrid parallel support vector machine, a hybrid parallel support vector machine based on sample cross combination is proposed, and carried on the simulation experiment analysis in the stand-alone environment.

Keywords: Image Classification, Support Vector Machine, Parallel Support Vector Machine, Sample Crossover

1. Introduction
Remote sensing image classification is a key technology widely used in many fields such as environmental monitoring and portrait recognition, and its algorithm accuracy and speed directly affect the realization effect. Support vector machine (SVM) has been proved to be an effective method to solve the problems of small samples, non-linearity, high dimensionality, etc. It has been applied to remote sensing image classification in recent years. This method can automatically find high-quality classification support vectors, maximize the interval between classifications, and achieve the best balance between model complexity and learning ability. Compared with traditional statistical classification methods, it has better generalization ability. However, the remote sensing image classification based on support vector machine (SVM) also has its limitations. This article will conduct research on this to explore methods to improve the accuracy and speed of remote sensing image classification.

2. Support vector machine (SVM)
Support vector machine (SVM) is a class of generalized linear classifiers based on the principle of structural risk minimization and binary classification of data by supervised learning. If the feature space is linearly separable, construct the optimal classification hyperplane with low VC dimension in the high-dimensional feature space as the decision surface to maximize the separation between the two types of data; otherwise, map the input vector to the high-dimensional through the kernel function in the feature space, and construct the optimal classification hyperplane in this high-latitude feature space.
The optimal classification hyperplane is shown in Figure 2-1.

![Figure 2-1. SVM optimal classification hyperplane](image)

When calculating large-scale data based on SVM, the problem is divided into several sub-problems, and then the serial processing method or parallel processing method is used for calculation. In comparison between the two, the serial processing method is to process several sub-problems in turn, the training time is longer and the number of iterations is large, so the efficiency is lower; the parallel processing method is to process several sub-problems in parallel first, and then for integration, it is more efficient. With the development of technology, the feature dimension and data volume of remote sensing image classification have increased sharply, and the requirements for accuracy have also increased. In summary, this article chooses to conduct further research on parallel support vector machines (PSVM).

3. Parallel support vector machine (PSVM)

3.1. Principle of parallel support vector machine (PSVM)
There are two main principles of parallel support vector machine (PSVM): one is based on the improvement of support vector machine (SVM), which can be resolved and parallelized through algorithm improvements; the other is based on several support vectors Machine (SVM) realizes the parallel processing of the training data set. Obviously, the latter can not only improve the classification speed, but also improve the classification accuracy, and can be flexibly adjusted. Therefore, the current parallel support vector machine (PSVM) mainly adopts the second design principle: first classify the training samples, then use the support vector machine (SVM) to separately train each type of sample, and then use the local support vector obtained from the training set is combined and retrained until the classification accuracy reaches the predetermined requirement.

3.2. Parallel training mode
At present, the most commonly used parallel support vector machine design modes include Cascade-PSVM, Grouped-PSVM, Feedback-PSVM, and Hybrid Parallel Support Vector Machine (Hybrid-PSVM).

From the actual application effect: Cascade-PSVM has a faster training speed, has good parallelism in the training process, and has a slightly lower training accuracy; grouped parallel support vector machine (Grouped-PSVM) ) Can significantly reduce training time, and has the best parallelism compared to other modes, but there is a problem of lower training accuracy when there are more training samples; feedback-type parallel support vector machine (Feedback-PSVM) training The time is longer, the parallelism of the training process is lower, but the training accuracy is higher and more stable. It can be seen that the above three parallel support vector machine design modes have some advantages in practical applications, but the effect is still not ideal. Hybrid Parallel Support Vector Machine (Hybrid-PSVM) combines the above modes to ensure that the training speed and training accuracy can reach the ideal level. At present, a combination of cascading and grouping is the most common.

4. Improved hybrid parallel support vector machine
According to the foregoing, this article improves on the Hybrid-PSVM (Hybrid-PSVM), and proposes
a hybrid parallel support vector machine based on sample cross combination, in order to improve the speed and accuracy of training at the same time, ensure a high level of parallelism.

Hybrid parallel support vector machines are mainly used to randomly cross merge the original sample subsets or partial support vector sets, so that when the original training set is decomposed, the probability of generating classification information loss is reduced and the training accuracy is improved. But in a parallel environment, usually the larger the number of sub-problems, the more layers of combined iteration and the longer the training time. Therefore, in order to improve the efficiency of training, it is necessary to reduce the scale of the cross sub-problems.

In the hybrid parallel support vector machine based on sample cross combination proposed in this paper, the original sample set SP is first divided into n sub-sample sets (n is an even number and n>0); then each sub-sample set is performed according to the degree of separation Sorting, among which [n/2] sub-sample sets are regarded as easy-to-separate sample sets S, and the remaining sub-sample sets are regarded as hard-to-separate sample sets P; the cross-combination of difficult-to-separate sample sets P is used to reduce the scale of cross sub-problems. In this research, the sample number is 8 as an example to explain and analyze, build a hybrid parallel support vector machine model, the specific structure is shown in Figure 4-1.

![Figure 4-1. Schematic diagram of the structure of the improved hybrid parallel support vector machine](image)

It can be seen from Figure 4-1 that in the first layer, 16 sub-questions are obtained by cross-combining the easy-to-divide sample set and the hard-to-divide sample set. Train the 16 sub-problems of the first layer to get 16 support vectors. Randomly divide the 16 support vectors into 24 sub-sample sets for training, and get the support vectors of the second layer; do the training of other layers in the same way, then combine the deepest support vector set and generate the global support vector set sv_final, and generate the classifier model Final_Model.

5. Experiment and analysis
In order to verify the performance of the improved hybrid parallel support vector machine proposed in
this paper, simulation experiments are carried out in a stand-alone environment, and compared with several other parallel support vector machines.

5.1. Simulation experiment
The stand-alone environment of the simulation test: 2.5GHZ dual-core CPU; 4G memory; NVIDIA RTX2060 graphics card. The experimental platform uses Windows 10 operating system, Matlab parallel toolbox and Libsvm application development kit. The remote sensing image data selected in this experiment is 206*1062 pixels, as shown in Figure 5-1.

![Figure 5-1. Experimental image](image)

The training samples selected in the experiment are shown in Table 5-1.

| Type | C_1  | C_2  | C_3  | C_4  | C_5  | C_6  | C_7  | C_8  | C_9  | C_{10} | C_{11} | C_{12} |
|------|------|------|------|------|------|------|------|------|------|--------|--------|--------|
| Number of training samples | 6012 | 4846 | 8036 | 5364 | 7638 | 5412 | 6326 | 6354 | 5846 | 9678   | 7532   | 5438   |
| Number of inspection samples | 2654 | 2388 | 3484 | 2550 | 2466 | 3052 | 2544 | 2533 | 2486 | 5688   | 4876   | 4922   |

5.2. Experimental results and analysis
After normalizing and formatting the training sample and test sample data, the improved hybrid parallel support vector machine based on sample cross combination proposed in this paper is used for training. The training results under different sample sizes are shown in Table 5-2.
Table 5-2. Training results of hybrid parallel support vector machine based on sample cross combination

| Number of initial training subsets | Number of cross-training subsets | Number of iterations | Number of SVS | Training time (s) | Overall classification accuracy (%) |
|-----------------------------------|----------------------------------|----------------------|---------------|-------------------|-------------------------------------|
| 1                                 | 1                                | 1                    | 28262         | 1423.87           | 85.15                               |
| 2                                 | 4                                | 2                    | 28481         | 841.56            | 84.53                               |
| 4                                 | 8                                | 3                    | 28346         | 568.75            | 85.16                               |
| 6                                 | 12                               | 3                    | 28353         | 235.31            | 85.16                               |
| 8                                 | 16                               | 3                    | 28319         | 137.32            | 85.25                               |
| 10                                | 20                               | 3                    | 28232         | 89.77             | 85.20                               |
| 12                                | 24                               | 3                    | 28335         | 60.48             | 85.22                               |
| 14                                | 28                               | 3                    | 28339         | 48.32             | 85.21                               |
| 16                                | 32                               | 3                    | 28345         | 36.15             | 85.20                               |
| 18                                | 36                               | 3                    | 28341         | 38.77             | 85.21                               |
| 20                                | 40                               | 3                    | 28510         | 43.62             | 84.18                               |
| 22                                | 44                               | 3                    | 28698         | 48.34             | 83.61                               |
| 24                                | 48                               | 3                    | 28698         | 56.21             | 82.15                               |
| 26                                | 52                               | 3                    | 29113         | 63.95             | 79.20                               |

From the results shown in Table 5-2, it can be seen that during the training process, as the number of training subsets increases, the training time shows a significant downward trend. After the number of training subsets exceeds 16, the training time shows a slow upward trend. Through analysis, it is believed that the reason for this situation is that the experiment was carried out in a stand-alone environment, and the number of concurrency is limited by memory, CPU, etc. When the number of training subsets is large, some parallel threads are actually performed in a serial manner, which leads to an increase in training time.

After normalizing and formatting the training sample and test sample data, under the condition that the initial training subset number is 8, the parallel support vector machines of different modes are used for training. The training results are shown in Table 5-3.

Table 5-3. Training results of different modes of parallel support vector machines

| Training method   | Classifier | Initial training subset number | Layers | Number of SVS | Training time (s) | Overall classification accuracy (%) |
|-------------------|------------|--------------------------------|--------|---------------|-------------------|-------------------------------------|
| Cascade-PSVM      | SVM        | 8                              | 4      | 28825         | 70683             | 84.82                               |
| EC-PSVM           | SVM        | 8                              | 4      | 28419         | 1405.77           | 85.21                               |
| JC-PSVM           | SVM        | 8                              | 4      | 28398         | 1295.57           | 85.25                               |
| Grouped-PSVM      | SVM        | 8                              | 1      | 31983         | 122.23            | 84.50                               |
| Hybrid-PSVM       | SVM        | 8                              | 2      | 29286         | 132.55            | 85.06                               |
| Improve           | SVM        | 8                              | 16     | 28319         | 137.32            | 85.25                               |

It can be seen from the experimental results shown in Table 5-3 that when the number of initial training subsets is 8, the training time of Cascade-PSVM is slightly longer, and the classification accuracy is at a medium level; the training speed of Grouped-PSVM is the fastest, but the
classification accuracy is also the lowest; although EC-PSVM and JC-PSVM have higher classification accuracy, the training time is very long; the improved Hybrid-PSVM in this article maintains both a faster training speed and a higher classification accuracy.

6. Conclusions
In summary, the experimental analysis results prove that the improved Hybrid-PSVM hybrid parallel support vector machine proposed in this paper has good performance in terms of training time and overall classification accuracy, and has relatively good parallelism under the premise of ensuring classification accuracy, it can significantly improve the calculation efficiency of large-scale training samples, so it has certain application value in the field of remote sensing image classification.

References
[1] He Zhouyu, Feng Xupeng, Liu Lijun, Huang Qingsong. Deep Strong Correlation Hash Learning Method for Large-Scale Image Retrieval[J]. Computer Research and Development, 2020, 57(11): 2375-2388.
[2] Liu Qing. Research on Image Retrieval Methods Based on Deep Learning [D]. Qingdao University of Science and Technology, 2020.
[3] Wang Ming. Research on Sugarcane Field Recognition Method Based on Support Vector Machine [D]. Guangxi Normal University, 2020.
[4] Li Chaohua, Tian Jingjing, Jiang Leyi. A Two-Layer System of Criminal Investigation Image Retrieval Method Based on Sorting Synthesis [J]. Computer and Digital Engineering, 2019, 47(12): 3164-3168.
[5] Wang Longhui. Research on Intelligent Analysis and Modeling Method of Control Valve Big Data Based on Parallel Computing [D]. Shandong University, 2019.
[6] Zhang Hongsheng, Ding Yonghong. Research on Parallel Stacked Support Vector Machine Based on Mixed Sample Training[J]. Journal of Jinling Institute of Technology, 2019, 35(03): 8-11.
[7] Wang Yan, Cao Jianfang, Li Yanfei. Large-scale Scene Image Classification Algorithm Combined with Hybrid Optimization[J]. Computer Technology and Development, 2019, 29(09): 86-91.
[8] Li Zhen, Kang Jie, Liu Zhaobang, Lu Qianqi, Xie Jing. A Skin Disease Image Retrieval Algorithm Based on Multi-Feature Combination and SVM Correlation Feedback[J]. Computer Knowledge and Technology, 2019,15(05):178-181.