Remote Sens. **2014**, *6*, 964-983; doi:10.3390/rs6020964

OPEN ACCESS

remote sensing

ISSN 2072-4292

www.mdpi.com/journal/remotesensing

Article

Comparison of Classification Algorithms and Training Sample Sizes in Urban Land Classification with Landsat Thematic Mapper Imagery

Congcong Li [1], Jie Wang [2], Lei Wang [2], Luanyun Hu [3] and Peng Gong [2,3,4,5,*]

[1] State Key Laboratory of Remote Sensing Science, and College of Global Change and Earth System Science, Beijing Normal University, Beijing 100875, China; E-Mail: licc129@163.com

[2] State Key Laboratory of Remote Sensing Science, Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing 100101, China; E-Mails: wangjie@irsa.ac.cn (J.W.); wang@irsa.ac.cn (L.W.)

[3] Ministry of Education Key Laboratory for Earth System Modeling, Center for Earth System Science, Tsinghua University, Beijing 100084, China; E-Mail: hly90@qq.com

[4] Department of Environmental Science, Policy, and Management, University of California, Berkeley, CA 94720-3114, USA

[5] Joint Center for Global Change Studies, Beijing 100875, China

* Author to whom correspondence should be addressed; E-Mail: penggong@tsinghua.edu.cn; Tel.: +86-10-6279-5385; Fax: +86-10-6279-7284.

Received: 18 November 2013; in revised form: 10 January 2014 / Accepted: 13 January 2014 / Published: 24 January 2014

Abstract: Although a large number of new image classification algorithms have been developed, they are rarely tested with the same classification task. In this research, with the same Landsat Thematic Mapper (TM) data set and the same classification scheme over Guangzhou City, China, we tested two unsupervised and 13 supervised classification algorithms, including a number of machine learning algorithms that became popular in remote sensing during the past 20 years. Our analysis focused primarily on the spectral information provided by the TM data. We assessed all algorithms in a per-pixel classification decision experiment and all supervised algorithms in a segment-based experiment. We found that when sufficiently representative training samples were used, most algorithms performed reasonably well. Lack of training samples led to greater classification accuracy discrepancies than classification algorithms themselves. Some algorithms were more tolerable to insufficient (less representative) training samples than

others. Many algorithms improved the overall accuracy marginally with per-segment decision making.

Keywords: machine learning; maximum likelihood classification; logistic regression; support vector machine; tree classifiers; random forests

1. Introduction

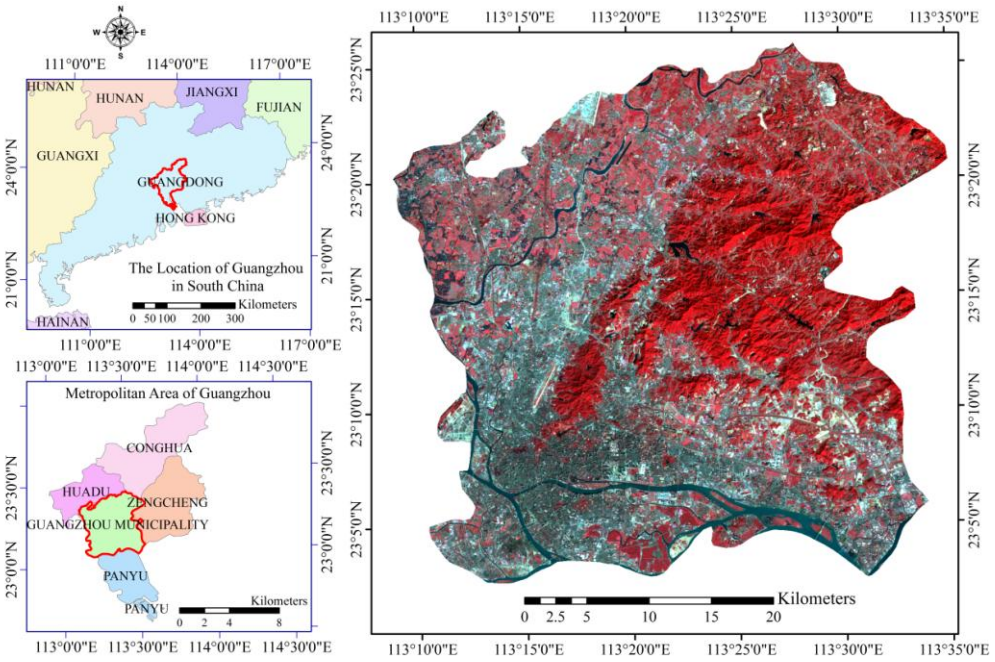
Since the launch of the first land observation satellite, Earth Resource Technology Satellite (later changed to Landsat-1) in 1972, substantial improvements have been made in sensor technologies. The spatial resolution has increased over 100 times, from the 80 m of the Landsat-1 to 0.41 m of the GeoEye-1 (Orbview-5) satellite. The spectral sampling frequency has increased nearly 100 times, from a few spectral bands to a few hundred spectral bands. Classification of land cover and land use types has been one of the most widely adopted applications of satellite data. Although a large number of algorithms have been developed and applied to map land cover from satellite imagery, and new algorithm proposers have reported improvements in accuracies of their mapping experiments (see [1] for a substantial review on classification algorithms), it is difficult to find a systematic comparison on the performance of newly proposed algorithms. This is particularly true for machine learning algorithms, as many of them have been introduced to the field of remote sensing for less than 10 years. Instead, classifier performance comparison has only been limited to the comparison of a new algorithm with a conventional classifier like the maximum likelihood classifier [2–4], or the comparison among a small number of two to three new algorithms [5]. Through meta-analysis of a large number of published literatures on land cover and land use mapping, Wilkinson [6] found that accuracy improvement of land cover and land use mapping by new algorithms are hardly observable. However, this kind of analysis compares classification accuracies in different literature reporting applications over different study areas and/or with different types of satellite data.

As the number of machine learning algorithms increases, it is beneficial for the user community of machine learning algorithms to gain a better knowledge on the performances of each algorithm. In the field of remote sensing image classification, a more comprehensive comparison of major machine learning algorithms is needed. This must be done with the same land cover and land use classification scheme and the same satellite image. It is generally believed that final image classification results are dependent of a number of factors: classification scheme, image data available, training sample selection, pre-processing of the data including feature selection and extraction, classification algorithm, post processing techniques, test sample collection, and validation methods [7]. The purpose of this research is to compare performances of 15 classification algorithms when applied to the same set of Landsat Thematic Mapper (TM) image acquired over Guangzhou City, China, while keeping the other factors the same. The urban area of Guangzhou has been selected for this purpose as it includes relatively complex land cover and land use patterns that are suitable for classification algorithm comparison. In addition to applying the algorithms on a pixel-by-pixel basis, we also tested the algorithms on a per-segment basis to compare the effect of including the object-based image analysis as a preprocessing step in the image classification process.

2. Study Site and Data

Our study site is located in the north of the Pearl River Delta (23°2'–23°25'N, 113°8'–113°35'E). As the capital city of Guangdong, Guangzhou is one of the fastest growing cities in China. It contains the core part of Guangzhou Municipality, and its rural-urban fringe (Figure 1). It can be divided into three regions: forest in the northeast, farmland in the northwest, and settlement in the south. As Guangzhou has been among the first group of cities that have undergone rapid development for over 20 years, it has been studied extensively for land use and land cover mapping and change detection (e.g., [8–10]).

Figure 1. The study area. The image displays the green, red and near infrared band of the TM data with blue, green, and red color guns.



The Landsat Thematic Mapper (TM) image used here was acquired on 2 January 2009, in the dry season of this subtropical area. For classification on a single date image, there is no need to do atmospheric correction if the sky is clear, which is the case in this study [7,11]. Geometric correction was applied to the raw imagery by co-registering this image with a previously georeferenced TM image acquired in 2005. A total of 153 ground control points were selected from the image. A second order polynomial resulted in the root mean squared error of 0.44 pixels. The original image was radiometrically resampled with a cubic convolution algorithm (for classification purposes nearest neighbor or bilinear resampling would work as well). Due to its coarser resolution, we experimented with a 6-band set of the TM data by excluding the thermal band. In order to estimate the potential of satellite data of similar resolution but with only visible and near-infrared bands (e.g., the 32 m resolution multispectral camera on board the Disaster Monitoring Constellation satellites, the 30 m multispectral sensor on board China's Huanjing-1A satellite), we also experimented with a 4-band set of the TM data by further excluding the two middle infrared bands.

At the time of image acquisition, some fruit trees (such as Litchi) and several vegetables were in their blooming stage and some fruit trees (such as citrus) were in fruit-bearing stage. The elevation is

high in the northeast mountains and low in the southwest farmlands. Newly developed industrial areas are in the southeast.

3. Method

3.1. Classification System

The land cover and land use classification system was developed to reflect the major land types in this area with reference to Gong and Howarth [2,12], and Gong *et al.* [13] Their meanings are self-explanatory (Table 1). On this basis, totally there were 14 subclasses for training, which were divided according to the spectral characteristics [2]. For example, the Industrial/commercial was subdivided into 4 types due to the spectral differences by different roofing materials.

Table 1. Land use classification system.

Land-Use Types	Description
Water	Water bodies such as reservoirs, ponds and river
Residential area	Residential areas where driveways and roof tops dominate
Natural forest	Large area of trees
Orchard	Large area of fruit trees planted
Farmland	Fields where vegetables or crops grow
Industrial/commercial	Lands where roof tops of large buildings dominate
Cleared land/Land under construction	Lands where vegetation is denuded or where the construction is underway
Idle land	Lands where no vigorous vegetation grows

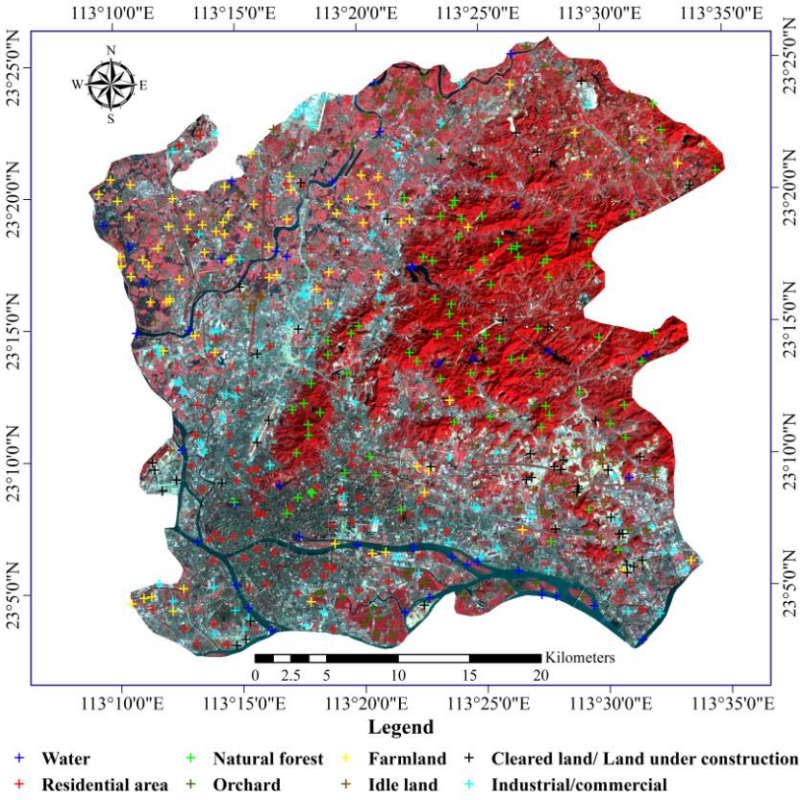
3.2. Training Samples

Training samples are primarily collected on a per-pixel basis to reduce redundancy and spatial-autocorrelation [7]. They were selected through image interpretation with intensive field visits over this area. Although more training samples are usually beneficial, as they tend to be more representative to the class population, a small number of training samples is obviously attractive for logistic reasons [14]. It is often recommended that a training sample size for each class should not be fewer than 10–30 times the number of bands [15–17]. This is usually okay for classifiers that require few parameters to be estimated like the maximum likelihood classifier when applied to a handful number of bands. With many classification algorithms, no previous study has reported an optimal number of training samples. To test the sensitivity of an algorithm to the size of training samples, we selected training samples uniformly from the images to make sure each subclass has 240 samples for later experiments. We sampled the training data to construct 12 sets of training samples with 20, 40, 60, 80, 100, 120, 140, 160, 180, 200, 220, and 240 pixels. For object-based method, we selected the segments contained the training pixels (240 pixels per subclass) as training objects.

3.3. Test Samples

We separately collected 500 pixels as test data, 138 of which were from field visits (done in April and December 2009, and June 2010), and the remaining were selected according to prior knowledge. The size of test sample for each land class was greater than 40 pixels (Figure 2). We used Kappa coefficient as the evaluation criterion [18].

Figure 2. The distribution of test samples.



3.4. Classification Process

We tested 15 algorithms [19–33] all from easily accessible sources [34–37]. These algorithms are selected because they are openly accessible and easy to use. As the number of algorithms is large, and they are clearly documented elsewhere, sources of references on the codes and documentation of the algorithms are provided in Table 2. Most algorithms require certain parameterizations. While the choice of optimal parameter set is desirable, it is extremely difficult to do so even with the original algorithm developer as the application conditions vary so widely from one environment to another and from one data type to another. However, it is generally safe to adopt the recommended range by the algorithm developers. In practice this is usually what has been done. Therefore, we designed experiments to cover a majority of parameter combinations for each algorithm (Table 2) while adopting the parameter ranges as recommended in the original sources of references. These algorithms were tested using both pixel-based and segment-based methods. For the two unsupervised classifiers, the Iterative Self-Organizing Data Analysis Technique (ISODATA) is a popular advanced clustering algorithm [19] while the Clustering based on Eigenspace Transformation (CBEST) is an efficient k-means algorithm [20].

The clusters obtained were grouped into informational classes by the same analyst who did the selection of training and test samples. It is assumed that the analyst is most familiar with the study area given sufficient field visits and consulting with local experts.

For segment extraction, we used BerkeleyImageSeg (http://berkenviro.com/berkeleyimgseg/) to perform image segmentations and then classified the segments by each algorithm. For the segmentation, the threshold is the most important parameter, which determines the size of the objects [38]. Here, four threshold values {5, 10, 15, 20} were examined. The shape parameter and compactness parameter were set to 0.2 and 0.7, respectively. The statistical spectral properties of the segments were then used in the segment-based classification. The features [39] are listed in Table 3. There are a total of 24 features. The parameters used for this method are selected according to the empirical values from the pixel-based classification, and taking the number of features used into consideration.

Table 2. Algorithm parameter set up and source of codes.

Algorithm	Abbreviation	Parameter Type	Parameter Set	Source of Codes
ISODATA	ISODATA	Number of Clusters Maximum Iterations	100, 150 100	ENVI
CBEST	CBEST	Number of Clusters Maximum Iterations	100, 150 100	CBEST
Maximum-likelihood classification	MLC	Mean and covariance matrix	Estimated from training samples	openCV
K-nearest neighbor	KNN	K weight	1,3,5,7,9,11 No weighting, 1/distance	Weka
Logistic regression	LR	Ridge estimator	0, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}	Weka
C4.5	C4.5	MinNumObj Confidence	1, 2, 3, 4, 5, 6, 7, 8, 9, 10 0.05f, 0.1f, 0.2f, 0.3f, 0.4f, 0.5f	Weka
Classification and Regression Tree	CART	MinNumObj	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	Weka
Quick, Unbiased, Efficient, and Statistical Tree algorithm	QUEST	Split types MinNumObj	univariate, linear 1, 2, 3, 4, 5, 6, 7, 8, 9, 10	QUEST
Random Forests	RF	numFeature numTrees	For 6-bands: 1,2,3,4,5,6 For 4-bands: 1,2,3,4 20, 40, 60, 80, 100, 120, 140, 160, 180, 200	Weka
Support Vector Machine	SVM	kernelType Cost gamma	radial basis function 10, 20, 30, 40, 50, 60, 70, 80, 90, 100 2^{-2}, 2^{-1}, 1, 2^{1}, 2^{2}, 2^{3}, 2^{4}, 2^{5}, 2^{6}, 2^{7}	Libsvm

Table 2. *Cont.*

Algorithm	Abbreviation	Parameter Type	Parameter Set	Source of Codes
Radial Basis Function Network	RBFN	MaxIteration NumCluster minStdDev	500, 1,000, 3,000, 5,000, 7,000 2, 3, 4, 5, 6, 7, 8, 9 0, 0.01, 0.05, 0.1	Weka
Logistic model tree	LMT	minNumInstances weightTrimBeta splitOnResiduals	5, 10, 15, 20, 25, 30 0, 0.01, 0.05, 0.1 False (C4.5 splitting criterion), True (LMT criterion)	Weka
Bagging C4.5	B_C4.5	bagSizePercent numIterations classifier	20, 40, 60, 80, 100 10, 50, 100, 150, 200 C4.5	Weka
AdaBoost C4.5	AB_C4.5	weightThreshold numIterations classifier	40, 60, 80, 100 10, 20, 30, 40, 50, 60, 70 C4.5	Weka
Stochastic gradient boosting	SGB	n.trees shrinkage bag.fraction	500, 1000 0.05, 0.1 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9	R

Table 3. The features used in the objected-based classification method.

Feature
Maximum value of the segments for each spectral band (6 bands)
Mean or average values of the segments for each spectral band (6 bands)
Minimum values of the segments for each spectral band (6 bands)
The standard deviations of the pixels in the segments for each spectral band (6 bands)

3.5. Active Learning

Active learning is an algorithm for selecting effective training samples. This kind of algorithms adds unlabeled samples as training samples from the sample pool through human-machine interaction [40]. In this research, we used a margin-sampling algorithm [41,42], which takes advantage of SVM. It selected candidate samples lying within the margin of the model, and these samples are most conducive to the improvement of the classifier's performance. At the beginning, we randomly selected 20 samples for each class, and added 10 samples from the training set using margin sampling at a time. The best parameters of SVM are selected using simple grid search.

4. Results

4.1. Pixel-Based Classification

Table 4 shows the best pixel-based classification accuracies of the algorithms. For the two unsupervised algorithms, they could produce as good results as some of the supervised algorithms when we cluster 150 spectral clusters. This is usually a very large number of clusters for an image analyst. Thus, we did not experiment for more clusters. Most supervised algorithms produce satisfactory results when the training samples are sufficient (more than 200 samples per class).

However, MLC only requires 60 pixels to reach its highest accuracy. This indicates the high level of robustness and capability of generalization.

Table 4. Best classification accuracy for each algorithm using pixel-based approach.

Algorithm	Parameter Choice	Accuracy
ISODATA	6-Band, Number of Clusters = 150	0.864
	4-Band, Number of Clusters = 150	0.841
CBEST	6-Band, Number of Clusters = 150	0.850
	4-Band, Number of Clusters = 150	0.846
MLC	6-Band, 60 training samples	0.892
	4-Band, 60 training samples	**0.873**
KNN	6-Band, 240 training samples, K = 3, weight = 1/distance	0.855
	4-Band, 240 training samples, K = 3, weight = 1/distance	0.823
LR	6-Band, 220 training samples, Ridge estimator = 10^{-8}	**0.899**
	4-Band, 200 training samples, Ridge estimator = 10^{-8}	0.862
C4.5	6-Band, 220 training samples, MinNumObj = 7, confidence = 0.2f	0.866
	4-Band, 240 training samples, MinNumObj = 3, confidence = 0.5f	0.841
CART	6-Band, 240 training samples, MinNumObj = 1	0.857
	4-Band, 220 training samples, MinNumObj = 2	*0.818*
QUEST	6-Band, 100 training samples, split type = linear, MinNumObj = 8/9/10	0.875
	4-Band, 180 training samples, split type = linear, MinNumObj = 1-10	0.843
RF	6-Band, 240 training samples, numFeatures = 1, numTrees = 20	0.873
	4-Band, 200 training samples, numFeatures = 1, numTrees = 60	0.848
SVM	6-Band, 240 training samples, kernelType = radial basis function, C = 80, gamma = 2^3	0.885
	4-Band, 240 training samples, kernelType = radial basis function, C = 50, gamma = 2^4	0.855
RBFN	6-Band, 220 training samples, minStdDev = 0.01, NumCluster = 9, MaxIts = 1,000	0.887
	4-Band, 240 training samples, minStdDev = 0.01, NumCluster = 8, MaxIts = 3,000/5,000	0.859
LMT	6-Band, 160 training samples, weightTrimBeta = 0, splitOnResiduals = C4.5 splitting criterion, minNumInstances ≥ 5	0.885
	4-Band, 200 training samples, weightTrimBeta = 0.1, splitOnResiduals = LMT splitting criterion, minNumInstances ≥ 5	0.862
B_C4.5	6-Band, 240 training samples, bagSizePercent = 80, numIterations = 100 C4.5parameter (MinNumObj = 2, confidence = 0.3f)	0.862
	4-Band, 240 training samples, bagSizePercent = 60, numIterations = 10 C4.5parameter (MinNumObj = 1, confidence = 0.2f)	0.836
AB_C4.5	6-Band, 220 training samples, weightThreshold = 40, numIterations ≥ 10, C4.5parameter (MinNumObj = 7, confidence = 0.2f)	0.866
	4-Band, 240 training samples, weightThreshold = 40, numIterations ≥ 10, C4.5parameter (MinNumObj = 5, confidence = 0.3f)	0.841
SGB	6-Band, 240 training samples, bag.fraction = 0.2, shrinkage = 0.1, n.tree = 1,000	0.869
	4-Band, 220 training samples, bag.fraction = 0.2, shrinkage = 0.1, n.tree = 1,000	0.852

A small value of K (K = 3) for KNN is the better choice in this study, and the distance-based weighting improves the KNN results. For the simple classification tree algorithms (CART, C4.5, and QUEST), minNumObj means minimum number of samples at a leaf node, which determines when to

stop tree growing. All the three simple tree algorithms achieve high accuracies when this value is less than 10. In other words, they all grow big trees and then prune them. However, the LMT needs a large minNumInstances to build the tree. For RF, numFeatures means the number of features to be randomly selected at each node and numTrees means number of trees generated. Usually, the suggested value of numFeatures is $\sqrt{N}$, where N is the number of features [43]. However, in this research, we find a value smaller than $\sqrt{N}$ is more suitable. For SVM, we used radial basis function (RBF) kernel, the space affected by each support vector is reduced as the kernel parameter gamma increases. A slightly large gamma (2^3, 2^4) is the best choice for this research, which means more support vectors are used to divide the feature space. MinStdDev in RBFN is the minimum number of standard deviations for the clusters, controlling the width of Gaussian kernel function as gamma in SVM. numCluster is the number of clusters, determining the data centers of the hidden nodes. In this research, we found the numCluster equal to or slightly greater than the number of classes is a better choice. BagSizePercent in Bagging controls the percentage of training samples randomly sampled from the training sets with replacement. The results show that 60%–80% of the training set achieved better results. It is similar to weightThreshold in Adaboost, but the latter one resamples the training set according to the weight of the last iteration. It achieves good classification results using only 10 iterations. For SGB, bag.fraction controls the fraction of training set randomly selected without replacement. The best value of the sampling fraction is 0.2. This reduces the correlations between models at each iteration. The best shrinkage value, which is the learning rate is 0.1.

From Table 4 we can see that the best classification accuracy for the 6-band case is achieved by logistic regression, followed closely by the maximum likelihood classifier, neural network, support vector machine, and logistic model tree algorithms. Opposite to this, the CBEST and KNN produced the lowest accuracies. The range of Kappa coefficient from the lowest to the highest is 0.049. For the 4-band case, in general, there is a 0.02 to 0.04 difference in Kappa for each algorithm, confirming the fact that with fewer spectral bands there is indeed accuracy loss. However, in this experiment, the accuracy drop is quite small implying that the inclusion of the two middle infrared bands of the TM would not add a lot of power in separability to the classification of our classes. The maximum likelihood classifier produced the highest accuracy of 0.873 for the 4-band case, only 0.026 inferior to the highest accuracy with the 6-band case. The accuracy range for the 4-band case is between 0.818 and 0.873.

4.2. Objected-Oriented Classification

Table 5 shows the best classification accuracies using objected-oriented method. The results of this kind of classification are largely depending on the segmentation [44]. The classification accuracies are the highest when the segmentation scale is set to 5 (the smallest). The best performer is SGB with an accuracy improvement of 0.025 over the best pixel-based classification results. This is followed closely by RF. The accuracies decrease with the increase of the threshold. A higher threshold produces larger objects. For the TM image, which is 30 m in resolution, fragmentation is relatively high in this urban area. High threshold brings more mixed information in the segments under this classification system. As small segments are relatively homogeneous, the classifiers utilizing statistical properties of the segments rather than individual pixel values improved the results.

Comparing Table 5 with Table 4, we can see all results are improved based on objected-oriented approach using spectral features only. Among them, SGB produced the best results, followed by RF, C4.5, LMT, LR, and MLC. From another perspective, these algorithms could deal with high-dimensional data.

Table 5. Best classification accuracy for each algorithm using objected-oriented approach.

	OB_5
MLC	0.898
KNN	0.871
LR	0.901
C4.5	0.912
CART	0.864
QUEST	0.882
RF	0.917
SVM	0.891
RBFN	0.894
LMT	0.908
B_C4.5	0.896
AB_C4.5	0.891
SGB	**0.924**

5. Discussions

5.1. Most Common Errors among the Classifiers

Figure 3 shows the test pixels of different classes that have been misclassified at least once. The clusters of repeatedly misclassified pixels are mainly found in the urban areas. The red circled area is the center of Guangzhou, where many residential, forest, commercial buildings, and construction land are mixed. The blue circled area is the new urban district, where bare land and industrial park are mixed. Most of the algorithms perform poorly in these complex areas given the fact that it is easy to have a wider range of spectral characteristics than it would normally have in natural environment within the same class. In addition, in the green circled area, there are black greenhouses and iron and steel enterprise. They are misclassified as residential area or water. From Figure 4 we can see residential area and water; natural forest and orchard; industrial/commercial and cleared land/land under construction, residential, water; farmland and idle land are more easily mixed in the feature space.

Tables 6 and 7 show that residential area and water, and residential area and industrial/commercial are well distinguished by ISODATA, CBEST, MLC, and LR. RBFN is good at distinguishing forest from orchard, while SVM is good at classifying industrial/commercial and cleared land/land under construction. LR, SVM, and LMT can better distinguish farmland and idle land.

Figure 3. The distribution of misclassified test samples.

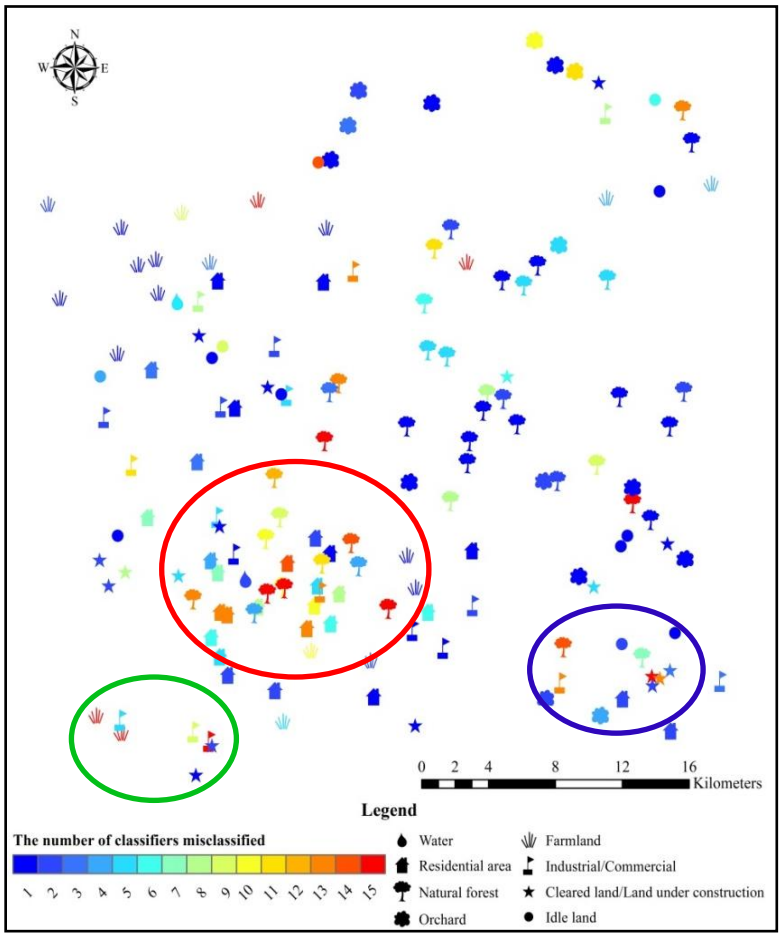


Figure 4. The distribution of the test samples in the feature space (principal component analysis (PCA) is used to reduce the dimension of data space).

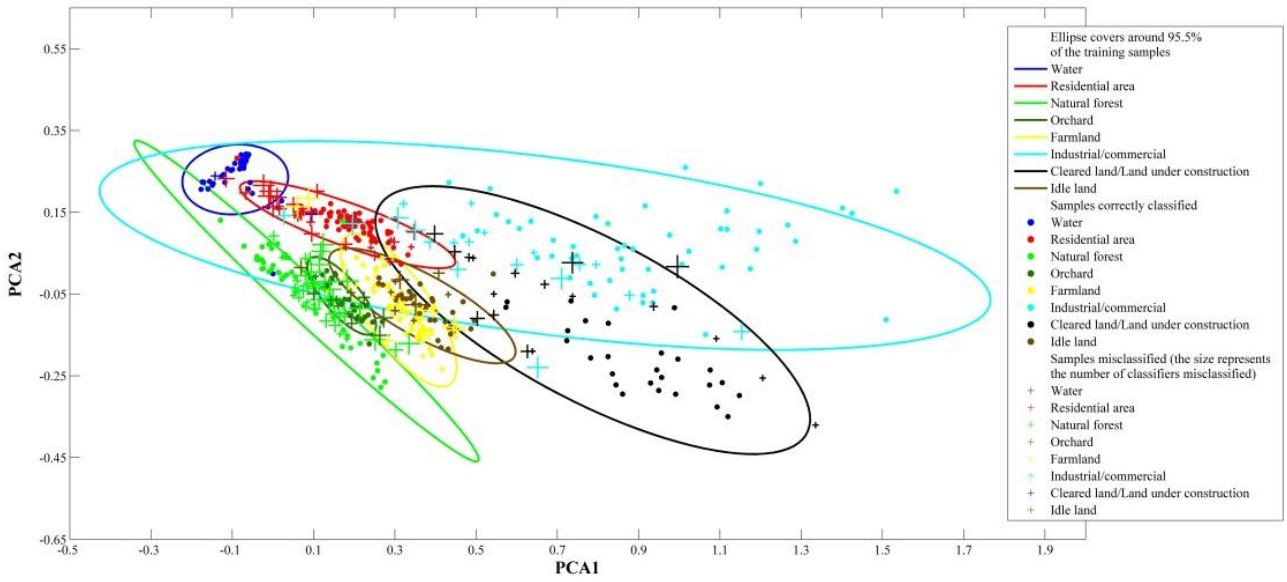


Table 6. The number of misclassified pixels by different algorithms.

Class	Water	Residential Area	Natural Forest	Orchard	Farmland	Industrial/ Commercial	Cleared Land/ Land under Construction	Idle Land
ISODATA	1	4	16	7	10	6	14	1
CBEST	0	7	13	8	12	7	7	11
MLC	1	5	20	2	6	8	3	2
LR	0	7	20	0	5	7	4	1
KNN	0	11	30	2	6	8	5	2
C4.5	2	10	15	4	8	11	5	3
CART	0	11	27	2	8	8	3	3
QUEST	0	14	18	4	5	7	4	2
RF	0	14	17	4	8	7	3	2
SVM	0	13	22	1	5	6	2	1
RBFN	0	10	12	3	7	8	3	6
LMT	0	8	21	1	4	9	5	2
B_C4.5	1	11	26	2	8	7	3	2
AB_C4.5	2	10	15	4	8	11	5	3
SGB	0	12	21	2	8	9	4	1

Table 7. The confusion matrix of the best result.

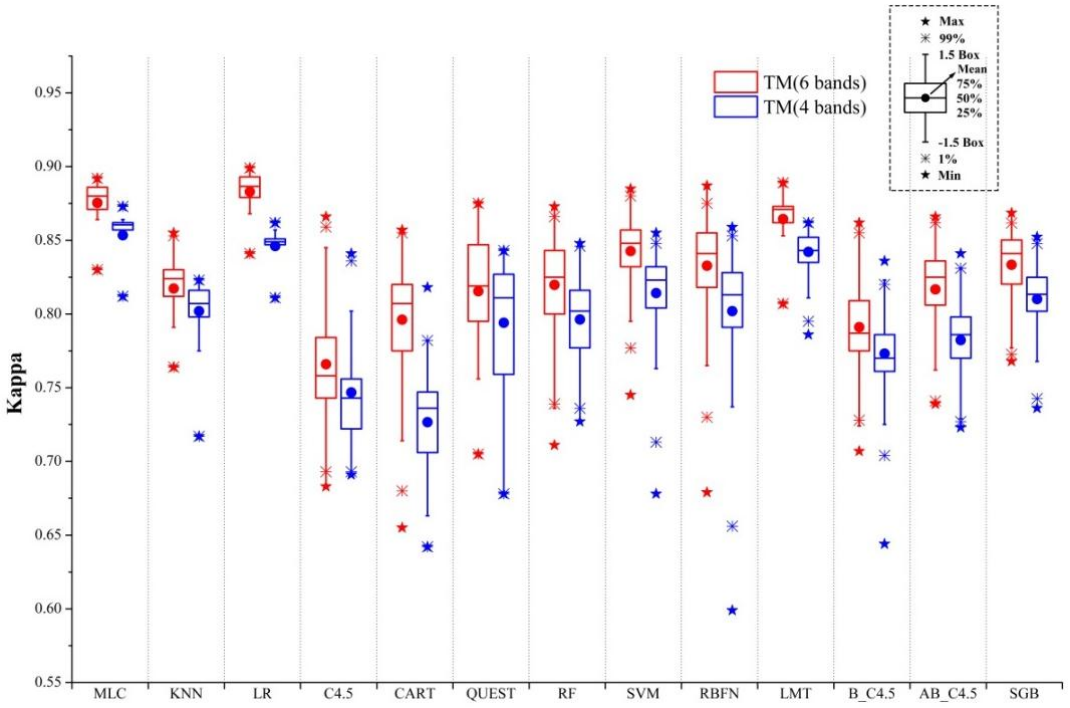
		Ground Truth (pixels)								
	Class	Water	Residential Area	Natural Forest	Orchard	Farmland	Industrial/ Commercial	Cleared Land/ Land under Construction	Idle land	Total
Classification results	Water	41	1	0	0	0	0	0	0	42
	Residential area	0	84	0	0	2	2	0	0	88
	Natural forest	0	0	63	0	1	0	0	0	64
	Orchard	0	0	18	48	0	0	0	0	66
	Farmland	0	0	2	0	72	0	0	1	75
	Industrial/ commercial	0	6	0	0	0	64	4	0	74
	Cleared land/ Land under construction	0	0	0	0	0	4	40	0	44
	Idle land	0	0	0	0	2	1	0	44	47
	Total	41	91	83	48	77	71	44	45	500

5.2. The Comparison of the Algorithms Using Pixel-Based Method

Figure 5 summarizes classification accuracies from all the parameter combinations listed in Table 2 and with different-sized training sets. The two unsupervised classifiers are not included as they were only tested with two parameter settings. We can see that all algorithms tested in this research could achieve high accuracies with sufficient training samples and proper parameters. MLC and Logistic regression have superior performances to other algorithms as their accuracy range is narrow and they

can be easily set to produce a high accuracy. Another traditional algorithm—K-nearest neighbor does not get as high accuracies as these two. From the ranges of the boxes, MLC, LR, and LMT are the most stable algorithms among all the algorithms.

Figure 5. Comparison of the pixel-based supervised classification.



For the tree classifiers, the box ranges are large. They are sensitive to the selection of parameters and training samples. C4.5 and CART tested in this research both select only one feature to split on the nodes, while QUEST uses a linear combination of features to split classes. The latter divides the feature space more reasonably and flexibly when the spectral distribution is complex. RF as an advanced tree algorithm uses Bagging algorithm to generate different training sample sets, and ensembles the different trees created by these training sets. Shown in the figure, RF is superior to Bagging (C4.5) and other simple trees. Compared with Bagging, RF splits the node using features randomly selected. This can reduce the correlation between the trees, and then improve the stability of the classification results. SVM and RBFN show similar performance in our experiments, but the parameters of these two classifiers are difficult to set. In general, users could not get the most out of the two algorithms because of the difficulties in parameter setting.

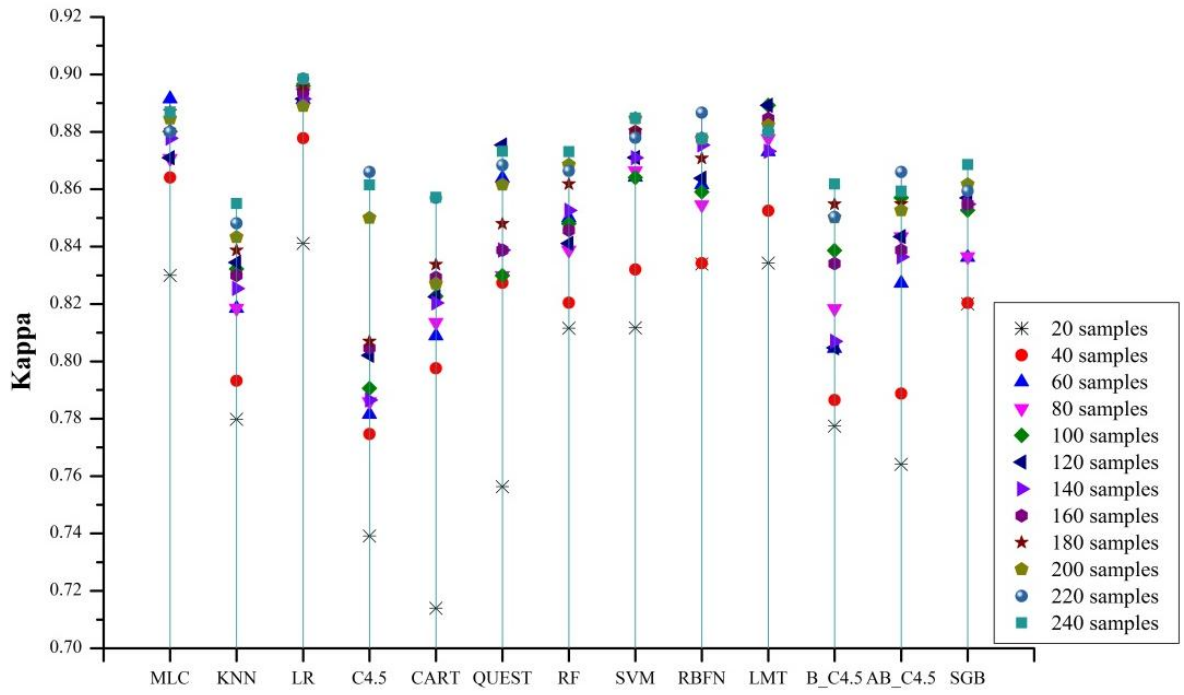
The Adaboost shows better results than Bagging. It focuses on the wrongly classified samples in the previous iteration rather than randomly selected samples. Bagging and Adaboost classifiers are both built on different training sample sets. Their maximum accuracies are not higher than that of C4.5 indicating a larger variability of single tree classifiers like C4.5 but better stability with ensemble classifiers through Bagging and Boosting. SGB is another boosting algorithm, and it outperformed Adaboost. It fits an additive function to minimize the residuals at each iteration. It relies on the small data set randomly selected while Adaboost relies on the incorrectly classified samples. LMT is built on different classifiers. The algorithm is a tree classifier and builds logistic regression models at its leaves. It takes advantage of the decision tree, building logistic regression at a small and relatively pure space.

The training sets have been divided into smaller subclasses according to the spectral characteristics. Therefore, LMT does not fully show its advantage.

5.3. The Impact of Different Training Set Sizes

Figure 6 shows the impact of training sample sizes on different classifiers. When the number of training samples is very small (e.g., 20, 40 samples), no algorithm performs well. The algorithms most affected by training sample size are the classification tree algorithms except for RF and Adaboost. They need sufficient samples to build the trees. MLC, LR, SVM, and LMT are the least affected algorithms. They could produce relative high accuracies using a smaller sized training set, and achieve stable results when there are 60 or more samples per class. All the algorithms except for MLC are improved in varying degrees by adding training samples. Generally speaking, MLC, LR, SVM, RBFN, and LMT could produce good results with small sized sample sets.

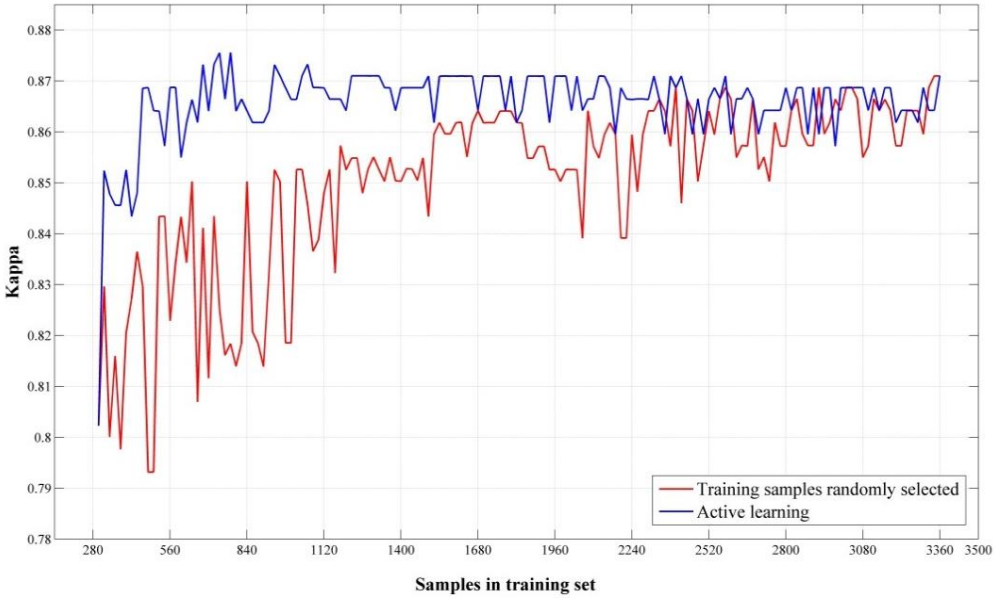
Figure 6. Pixel-based supervised classification with different sizes of training sets.



Active learning algorithm is used to test the representativeness of the training samples. Figure 7 shows that when the total number of training samples increases to 560–840 (40–60 samples for each subclass), about 1/6–1/4 of the entire training set, the classification results are satisfactory. The results are relatively stable when the training sample size further increases. In other words, the whole training set only contains 1/6–1/4 useful information. On the contrary, we randomly added the training samples without any rules, and then the results increased slowly and became stable when the training samples were representative enough (at about 2,800 samples, 200 samples for each subclass). That is why most of the algorithms achieve their highest accuracies when there are more than 200 samples per class.

Under such circumstances, using active learning algorithm to select training samples is an efficient way to achieving the optimal results before a large amount of trial and error tests. Active learning should be applied for representative sample selection to feed subsequent classification algorithms.

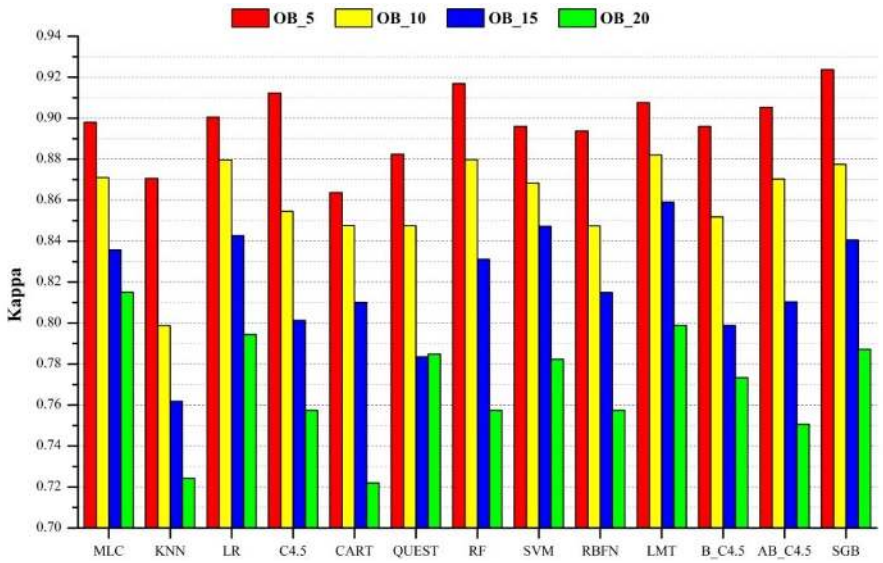
Figure 7. Active learning result and results based on training samples randomly selected.



5.4. Algorithm Performances with Low-Quality Training Samples

The quality of training samples is reduced with the increase of thresholds in segmentation as segment statistics are becoming increasingly contaminated by information from potentially other classes. We could use this to assess algorithm performance. An algorithm showing good results on different thresholds performs well with low-quality training sets. Figure 8 shows that MLC, LR, RF, SVM, LMT, and SGB algorithms could deal with the contaminated information better than others. They could build more robust classifiers from weak training sets. In comparison, Adaboost algorithm shows good results only when the segmentation scale is smaller than 15. KNN results decrease sharply when the segmentation scale is greater than 10. As we have described hereinbefore, MLC, LR, SVM, RBFN, and LMT could perform well using a small training set. However, with deteriorating training samples, MLC, LR, SVM and LMT could still perform well.

Figure 8. Objected-based supervised classification results.



6. Summary

In this study, we compared 15 classification algorithms in the classification of the same Landsat TM image acquired over Guangzhou City, China, using the same classification scheme. The algorithms were tested on a pixel-based and segment-based classification. In the pixel-based decision making, the algorithms were tested with two band sets: a 4-band set including only the visible and near infrared TM bands and a 6-band set with all TM bands excluding the thermal and panchromatic bands. All supervised classifiers were tested with 12 sets of different sized training samples. In the segment-based decision making, the algorithms were tested with different segment sizes determined by different scale factors. All tests were evaluated by the same set of test samples with the total overall accuracy measured by the Kappa coefficient. The results can be summarized in the following:

(1) The 4-band set of TM data by excluding the two middle infrared bands resulted in Kappa accuracies in the range between 0.818 and 0.873. The inclusion of the two middle infrared bands in the 6-band case increased this range to 0.850 and 0.899. This indicates the potential loss of overall accuracies in urban and rural urban fringe environments with the lack of middle infrared bands could be within 3%–5%.

(2) Unsupervised algorithms could produce as good classification results as some of the supervised ones when a sufficient number of clusters are produced and clusters can be identified by an image analyst who is familiar with the study area. The accuracy of the unsupervised algorithms produced better than 0.841 Kappa accuracies for the eight land cover and land use classes.

(3) Most supervised algorithms could produce high classification accuracies if the parameters are properly set and training samples are sufficiently representative. In this condition, MLC, LR, and LMT algorithms are more proper for users. These algorithms can be easily used with relatively more stable performances.

(4) Insufficient (less representative) training caused large accuracy drops (0.06–0.15) in all supervised algorithms. Among all the algorithms tested, MLC, LR, SVM, and LMT are the least affected by the size of training sets. When using a small-sized training set, MLC, LR, SVM, RBFN, and LMT performed well.

(5) In segment-based classification experiments, most algorithms performed better when the segment size was the smallest (with a scale factor of 5). At the scale of 5, SGB outperformed all other algorithms by producing the highest Kappa values of 0.924 and this is followed by RF. All algorithms are less sensitive to the large increase of data dimensionality. MLC, LR, RF, SVM, LMT, and SGB algorithms are the best choices to do the classification. They could produce relatively good accuracy at different scales.

With the increasing number of new algorithms emerging rapidly, there is a need to assess their performance and sensitivities to various kinds of environments. This need is best addressed by developing standard image sets with adequate classification scheme and sufficiently representative training and testing samples. This research represents one of such attempts. However, more datasets containing high quality training and test samples should be established for different types of remotely sensed data sets over typical environments in the world to support more objective assessment of new algorithms. Only when more comprehensive test data sets covering major environmental types of the

world can we make more appropriate selection of algorithms for a particular application of remote sensing classification. Another important aspect that has not been assessed in this research is feature extraction and use of non-spectral features whose effectiveness has been demonstrated in the literature [45–51]. Furthermore, use of multisource data including optical, thermal and microwave data in urban land classification should be systematically evaluated [52,53]. Lastly, more analysis of the representativeness of training samples should be done in developing algorithm test image sample sets [54]. These will be evaluated in a future research.

Acknowledgments

This work was partially supported by the National High Technology Research and Development Program of China (2009AA12200101, 2013AA122804), the National Natural Science Foundation of China (41001275), the Dragon Program and a research grant from Tsinghua University (2012Z02287). Field work has been done with assistance by Zheng Lin and Xia Li at Sun Yat-Sen University.

Author Contributions

All authors contributed extensively to the work presented in this paper.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Lu, D.S.; Weng, Q.H. A survey of image classification methods and techniques for improving classification performance. *Int. J. Remote Sens.* **2007**, *28*, 823–870.
2. Gong, P.; Howarth, P.J. Land-use classification of SPOT HRV data using a cover-frequency method. *Int. J. Remote Sens.* **1992**, *13*, 1459–1471.
3. Pu, R.L.; Landry, S.; Yu, Q. Object-based detailed land-cover classification with high spatial resolution IKONOS imagery. *Int. J. Remote Sens.* **2011**, *32*, 3285–3308.
4. Shackelford, A.K.; Davis, C.H. A hierarchical fuzzy classification approach for high resolution multispectral data over urban areas. *IEEE Trans. Geosci. Remote Sens.* **2003**, *41*, 1920–1932.
5. Tzeng, Y.C.; Fan, K.T.; Chen, K.S. An adaptive thresholding multiple classifiers system for remote sensing image classification. *Photogramm. Eng. Remote Sens.* **2009**, *75*, 679–687.
6. Wilkinson, G.G. Results and implications of a study of fifteen years of satellite image classification experiments. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 433–440.
7. Gong, P.; Howarth, P.J. An assessment of some factors influencing multispectral land-cover classification. *Photogramm. Eng. Remote Sens.* **1990**, *56*, 597–603.
8. Fan F.L.; Wang, Y.P.; Qiu, M.H.; Wang, Z.S. Evaluating the temporal and spatial urban expansion patterns of Guangzhou from 1979 to 2003 by remote sensing and GIS methods. *Int. J. Geogr. Inf. Sci.* **2009**, *23*, 1371–1388.
9. Fan, F.L.; Weng, Q.H.; Wang, Y.P. Land use and land cover change in Guangzhou, China, from 1998 to 2003, based on Landsat TM/ETM+ imagery. *Sensors* **2007**, *7*, 1223–1342.

10. Seto, K.; Woodcock, C.; Song, C.H.; Huang, X.; Lu, J.; Kaufmann, R.K. Monitoring land-use change in the Pearl River Delta using Landsat TM. *Int. J. Remote Sens.* **2002**, *23*, 1985–2004.
11. Song, C.; Woodcock, C.E.; Seto, K.C.; Lenney, M.P; Macomber, S.A. Classification and change detection using Landsat TM data: When and how to correct atmospheric effects? *Remote Sens. Environ.* **2001**, *75*, 230–244.
12. Gong, P.; Marceau, D.; Howarth, P.J. A comparison of spatial feature extraction algorithms for land-use mapping with SPOT HRV data. *Remote Sens. Environ.* **1992**, *40*, 137–151.
13. Gong, P.; Wang, J.; Yu, L.; Zhao, Y.; Zhao, Y.; Liang, L.; Niu, Z.; Huang, X.; Fu, H.; Liu, S.; *et al.* Finer resolution observation and monitoring of global land cover: First mapping results with landsat TM and ETM+ data. *Int. J. Remote Sens.* **2013**, *34*, 2607–2654.
14. Foody, G.M.; Mathur, A.; Sanchez-Hernandez, C.; Boyd, D.S. Training set size requirements for the classification of a specific class. *Remote Sens. Environ.* **2006**, *104*, 1–14.
15. Mather, P.M. *Computer Processing of Remotely-Sensed Images*, 3rd ed.; John Wiley & Sons, Ltd.: Chichester, UK, 2004.
16. Piper, J. Variability and bias in experimentally measured classifier error rates. *Pattern Recognit. Lett.* **1992**, *13*, 685–692.
17. Van Niel, T.; McVicar, T.; Datt, B. On the relationship between training sample size and data dimensionality: Monte Carlo analysis of broadband multi-temporal classification. *Remote Sens. Environ.* **2005**, *98*, 468–480.
18. Congalton, R.G.; Green, K. *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*; CRC Press: Boca Raton, FL, USA, 1999; p. 137.
19. Ball, G.H.; Hall, D.J. *Isodata, a Novel Method of Data Analysis and Pattern Classification*; Stanford Research Institute: Menlo Park, CA, USA, 1965.
20. Chen, Y.; Gong, P. Clustering based on eigenspace transformation—CBEST for efficient classification. *ISPRS J. Photogramm. Remote Sens.* **2013**, *83*, 64–80.
21. Aha, D.W.; Kibler, D.; Albert, M.K. Instance-based learning algorithms. *Mach. Learn.* **1991**, *6*, 37–66.
22. Le Cessie, S.; van Houwelingen, J.C. Ridge estimators in logistic regression. *Appl. Stat.* **1992**, *41*, 191–201.
23. Quinlan, R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann Publishers: San Mateo, CA, USA, 1993.
24. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; Wadsworth Press: Monterey, CA, USA, 1984.
25. Loh, W.Y.; Shih, Y.S. Split selection methods for classification trees. *Stat. Sin.* **1997**, *7*, 815–840.
26. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.
27. Chang, C.C.; Lin, C.J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1–27.
28. Bishop, C.M. *Neural Networks for Pattern Recognition*; Oxford University Press: New York, NY, USA, 1995.
29. Niels, L.; Mark, H.; Eibe, F. Logistic model trees. *Mach. Learn.* **2005**, *95*, 161–205.

30. Sumner, M.; Frank, E.; Hall, M.A. Speeding up Logistic Model Tree Induction. In Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases, Porto, Portugal, 3–7 October 2005; pp. 675–683.
31. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140.
32. Freund, Y.; Robert, E. Schapire: Experiments with a New Boosting Algorithm. In Proceedings of the 13th International Conference on Machine Learning, Bari, Italy, 3–6 July 1996.
33. Friedman, J.H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **2002**, *38*, 367–378.
34. Bradski, G. The OpenCV library. *Dr. Dobb's J. Softw. Tools* **2000**, *25*, 120–125.
35. Witten, I.H.; Frank, E. *Data Mining, Practical Machine Learning Tools and Techniques*, 2nd ed.; Morgan Kaufmann Publishers: San Francisco, CA, USA, 2005; p. 525.
36. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. The WEKA data mining software: An update. *SIGKDD Explor.* **2009**, *11*, 10–18.
37. Ridgeway, G. Generalized Boosted Models: A Guide to the GBM Package. Available online: https://r-forge.r-roject.org/scm/viewvc.php/*checkout*/pkg/inst/doc/gbm.pdf?revision=17&root=gbm&pathrev=18 (accessed on 21 September 2009).
38. Clinton, N.; Holt, A.; Scarborough, J.; Yan, L.; Gong, P. Accuracy assessment measures for object based image segmentation goodness. *Photogramm. Eng. Remote Sens.* **2010**, *76*, 289–299.
39. Clinton, N. BerkeleyImageSeg User's Guide. CiteSeerX. 2010. Available online: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.173.1324&rep=rep1&type=pdf (accessed on 16 January 2014).
40. Tuia, D.; Volpi, M.; Copa, L.; Kanevski, M.; Munoz-Mari, J. A survey of active learning algorithms for supervised remote sensing image classification. *IEEE J. Sel. Top. Signal Process.* **2011**, *5*, 606–616.
41. Campbell, C.; Cristianini, N.; Smola, A.J. Query Learning with Large Margin Classifiers. In the Proceedings of the 17th International Conference on Machine Learning, Stanford, CA, USA, 29 June–2 July 2000.
42. Schohn, G.; Cohn, D. Less is More: Active Learning with Support Vector Machines. In Proceedings of the 17th International Conference on Machine Learning, San Francisco, CA, USA, 29 June–2 July 2000.
43. Strobl, C.; Malley, J.; Tutz, G. Supplement to "An Introduction to Recursive Partitioning: Rational, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests". Available online: http://supp.apa.org/psycarticles/supplemental/met_14_4_323/met_14_4_323_supp.html (accessed on 11 November 2010).
44. Wang, L.; Sousa, W.; Gong, P. Integration of object-based and pixel-based classification for mangrove mapping with IKONOS imagery. *Int. J. Remote Sens.* **2004**, *25*, 5655–5668.
45. Gong, P.; Howarth, P.J. The use of structural information for improving land-cover classification accuracies at the rural-urban fringe. *Photogramm. Eng. Remote Sens.* **1990**, *56*, 67–73.
46. Gong, P.; Howarth, P.J. Frequency-based contextual classification and grey-level vector reduction for land-use identification. *Photogramm. Eng. Remote Sens.* **1992**, *58*, 423–437.
47. Xu, B.; Gong, P.; Seto, E.; Spear, R. Comparison of different gray-level reduction schemes for a revised texture spectrum method for land-use classification using IKONOS imagery. *Photogramm. Eng. Remote Sens.* **2003**, *6*, 529–536.

48. Liu D.; Kelly, M.; Gong, P. A spatial-temporal approach to monitoring forest disease spread using multi-temporal high spatial resolution imagery. *Remote Sens. Environ*. **2006**, *101*, 167–180.
49. Cai, S.S.; Liu, D.S. A comparison of object-based and contextual pixel-based classifications using high and medium spatial resolution images. *Remote Sens. Lett.* **2013**, *4*, 998–1007.
50. Wolf, N. Object features for pixel-based classification of urban areas comparing different machine learning algorithms. *Photogramm. Fernerkund.* **2013**, *3*, 149–161.
51. Luo, L.; Mountrakis, G. Converting local spectral and spatial information from a priori classifiers into contextual knowledge for impervious surface classification. *ISPRS J. Photogramm. Remote Sens.* **2011**, *66*, 579–587.
52. Weng, Q.H. Remote sensing of impervious surfaces in the urban areas: Requirements, methods, and trends. *Remote Sens. Environ.* **2012**, *117*, 34–49.
53. Gamba, P. Human settlements: A global challenge for EO data processing and interpretation. *IEEE Proc.* **2013**, *101*, 570–581.
54. Mountrakis, G.; Xi, B. Assessing reference dataset representativeness through confidence metrics based on information density. *ISPRS J. Photogramm. Remote Sens.* **2011**, *78*, 129–147.