AUSTRIAN JOURNAL OF STATISTICS
Volume 28 (1999), Number 4, 195–201

Accuracy Assessment of Satellite Image Classification Depending on Training Sample

Georg Ruppert[1], Mushtaq Hussain[2,4], and Heimo Müller[3]
[1] Joanneum Research, Graz
[2] Eurostat, Luxembourg
[3] Technikum Joanneum, Graz

Abstract: The paper presents a method of predicting classification accuracy of remote sensing data by means of training set analysis. Various sampling plans were applied to satellite image and its complete ground truth to derive different training sets. The quality of these training sets was determined by quantifying the similarity of the training set distributions to the ones of the entire satellite image. Each training set was then used to learn a classifier. The paper shows how the accuracy of classifications that were carried out using these classifiers depends upon the quality of the corresponding training sets.

Zusammenfassung: Es wird eine Methode zur Voraussage der Klassifikationsgenauigkeit für Fernerkundungsdaten präsentiert, die durch Trainingsdatenanalyse funktioniert. Verschiedene Sampling Strategien wurden auf ein Satellitenbild und die dazugehörende Referenzdaten angewendet, um unterschiedliche Trainingsdatensätze zu erzeugen. Die Qualität dieser Trainingssätze wurde durch Beurteilung der Ähnlichkeit des jeweiligen Trainingsdatensatzes mit der Verteilung der gesamten Referenzdaten ermittelt. Jeder Trainingssatz wurde in weiterer Folge zum Training eines individuellen Klassifikators herangezogen. Es wird gezeigt, wie die erreichte Klassifikationsgenauigkeit für diese Klassifikatoren von der Qualität der Trainingsgebiete abhängt.

Keywords: Classification Accuracy, Remote Sensing, Ground Truth.

Acknowledgment: The authors want to express their gratitude to Prof. Reinhard Viertl from the Institute of Statistics, Probability Theory, and Actuarial Mathematics, Vienna University of Technology, for his suggestions and comments.

1 Introduction

Land use and land cover classification maps generated from remote sensing data are valuable management and planning tools. The classification of satellite images is a complex problem with many research areas involved. There exists a wide selection of publications ranging from threshold methods to highly complex classification methods like specially designed and adapted Neural Networks. Ground truth (training set and test set) plays an

[4]Disclaimer. While the paper was written the author was employed at Joanneum Research, Institute of Applied Statistics, Graz. It does not necessarily reflect the views of Eurostat, where the author now works.

important role in remote sensing classifications. Ground truth is mainly acquired from visual aerial photo interpretation of a part of the satellite image and terrestrial surveying (see Van der Wel and Jansen, 1994). Until recently, the idea of assessing the classification accuracy of remotely sensed data was treated more as an afterthought than as an integral part of any project. This paper focuses on the influence of the ground truth on the final classification results. One of the major problems for the producer of commercial land use classifications is the quality and representativity of the ground truth. If the information derived form the ground truth is not representative of the entire satellite image, problems with classification accuracy may arise. For this reason a method for assessing the classification accuracy depending on the ground truth quality is necessary. This method could be used to decide whether classification accuracy would be satisfying or additional ground truth had to be acquired in order to improve the accuracy to a predefined level. For this investigation a Landsat TM image of the Vienna forests consisting of 512 x 512 pixels (25mx25m large) representing an area of 163,84 km2 was used (see Bischof et al., 1992; Ruppert et al., 1997). The picture was taken with seven spectral channels. This paper classifies this area in four different classes: agricultural area, built-up land, forest, and water. The size of these four classes was quite different. Additionally, complete ground truth was established by examination of aerial photos of the whole area. Therefore, we are able to choose any arbitrary training set to learn the classifier while using the remaining pixels to estimate the classification accuracy.

2 Process of Satellite Image Classifications

The user of a classification map usually provides the producer with a list of classes whose distributions or proportions are of interest. The producer then locates a number of pixels for which the corresponding class is known and uses these as training data (called ground truth) in order to establish the discriminating criteria whereby the remaining pixels (whose class is not known precisely) may be allocated to a class. Classification methods mainly used in the remote sensing community are maximum likelihood, nearest neighbor, and increasingly neural networks and computer learning methods like C4.5 (see Quinlan, 1993). C4.5 is a machine learning strategy which generates decision trees from training sets. A decision tree is a binary tree whose leaf nodes indicate the classes and non-leaf nodes represent the decisions. A pixel is classified by the decision tree by walking through its nodes until a leaf is encountered. At each decision node the outcome of the test determines the sub-tree where the path is continued. Thus C4.5 automatically generates binary decision trees. Feature names and thresholds used for classification are printed in the node of the decision tree, which provides an insight view into the decision process of the classifier not common to most other classifiers. This decision tree can easily be re-expressed by production rules in order to include them into existing remote sensing packages. The part of the ground truth which was not used for training purposes will afterwards be used for the assessment of the classification accuracy. Not all pixels can be classified correctly, because of fuzzy class boundaries, incorrectly assigned pixels or class distributions not representing reality. The classification accuracy therefore refers to the correspondence between the class label assigned to a pixel and the ‘true’ class known through ground

truth (see Congalton, 1991). If the ground truth does not represent all classes adequately the classification result and the corresponding accuracy may not be predicative.

3 Quality Assessment of Ground Truth

3.1 Sampling Plan

At the beginning, the entire data set has been clustered into 256 classes using the k-means algorithm. The corresponding membership matrix contains the mean values of the satellite channels as well as the pixel frequencies for each individual class. Next, six different training sets consisting of 3000 pixels were selected out of the entire image using the following techniques:

S1: simple random sample of 3000 pixels,

S2: systematic sample of 3000 pixels,

S3: random sample of 750 pixels for each class,

S4: random sample where the number of pixels in each class was proportional to the original size,,

S5: systematic sample of 750 pixels for each class,

S6: systematic sample where number of pixels in each class was proportional to the original size.

In the next step all the pixels of the six training sets S1 to S6 were allocated to the 256 classes using the previously calculated membership matrix for the entire population giving six distributions. For each training set the distribution was then compared with the distribution of the entire satellite image using an index based on chi-square distribution.

3.2 Goodness of Fit Test

The chi-square distribution is widely applied for goodness of fit test. Here a single array of categories of sample frequencies or proportions is tested against a pre-specified set which comprises the null hypothesis (see Edwards, 1972, pp. 53–55). The chi-square test on frequencies is quite general in its applicability to problems in both manipulative experiments and survey analysis. When used for frequency comparisons, the chi-square test is a non-parametric test, since it compares entire distributions rather than parameters of distributions. Thus, other than the need to avoid very small hypothetical frequencies, the test is relatively free of constraining assumptions. Cohen (1988, p. 216) defines an effect size index (w) which is a 'pure' number and increases with the degree of discrepancy between the distribution specified by the alternative hypothesis and that which represents the null hypothesis. This relative 'pureness' is achieved by working with relative frequencies, i.e,

proportions. The index w measures the discrepancy between the paired proportions over the cells in the following way:

$$w = \sqrt{\sum_{i=1}^{m} \frac{(P_{1i} - P_{0i})}{P_{0i}}}, \tag{1}$$

where

$P_{0i} =$ the proportion in cell i posited by the null hypothesis,

$P_{1i} =$ the proportion in cell i posited by the alternative hypothesis, and

$m =$ the number of cells.

In other words the index w is the square root of the non-centrality parameter λ, divided by the total sample size. The null hypothesis for goodness of fit tests is simply:

$$H_0 = P_{01}, P_{02}, ..., P_{0m} \mid \sum_{i=1}^{m} P_{oi} = 1 \tag{2}$$

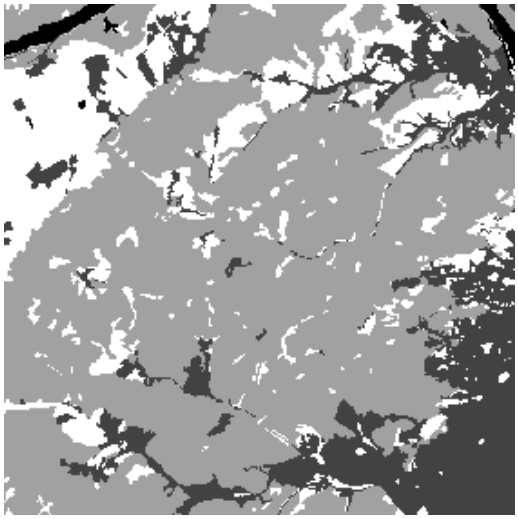
i.e., a specified distribution of proportions in m cells, summing to unity. A population of independent observations is posited as falling into m mutually exclusive and exhaustive classes with a specified proportion in each. The source of the null hypothesis in our case is the membership matrix based on the entire image, i.e., the proportion of the pixels allocated to the 256 classes. The alternative hypothesis is expressed by the proportions based on the individual samples S1 to S6. w in Formula 1 therefore shows the similarity of the distribution derived from samples with the distribution based on the entire population. It is clear that presuming identical distributions, the numerator of each cell's contribution is zero, hence $w = 0$. In general, the maximum value of w is infinity.

3.3 Goodness of Fit vs. Classification Accuracy

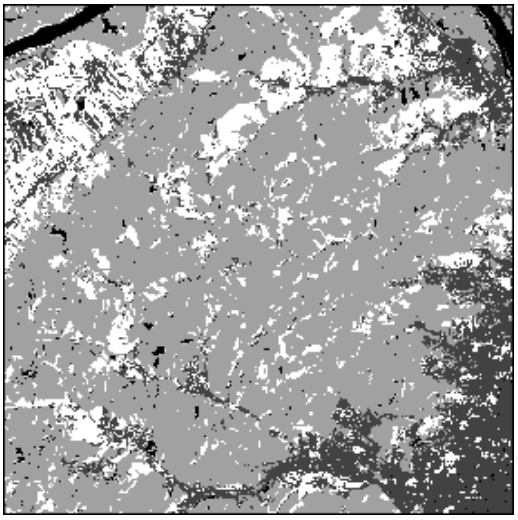
Each training set is used to generate a decision tree by C4.5. Consequently, the entire satellite image is classified separately by these decision trees giving images consisting of the four classes described above. Figure 1(a) shows the original Landsat TM image of the investigated area illustrated by the channels 3,2,1 as RGB. The complete ground truth, derived from aerial photo interpretation, is shown by Figure 1(b). The four classes are agricultural area shown in white, forest in light grey, built-up land in dark grey, and water surfaces in black. Figures 1(c) and 1(d) illustrate the classification of the worst and best result based on samples S3 and S6, respectively. As the entire ground truth was known it was possible to calculate the accuracies of the classification for each of the six decision trees giving the percentage of correctly classified pixels. Table 1 shows the calculated values of goodness of fit index w as well as the classification accuracy for the samples.

The Pearson correlation between the index w and the classification accuracy is -0.94, implying that the sample distributions with a greater departure from the population distribution lead to a lower classification accuracy.

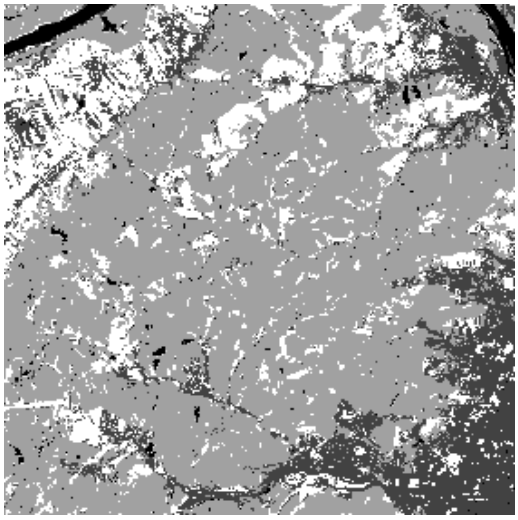
(a) Original Landsat TM image
RGB: channels 3, 2, 1.



(b) Ground truth



(c) Classification based on sample S3



(d) Classification based on sample S6

Figure 1: Satellite image and results

3.4 Future Activities

These first results seem promising, but are only based on six samples derived from one single data set. It is necessary to repeat the experiments with additional samples as well as additional data sets in order to validate the results. In doing so it is important to take samples of different sizes and qualities into consideration. As a side effect of this intensive validation process a rather large set of tuples of index w and classification accuracy will be built. Successful validation with a large number of samples presumed this set of tuples will be used to learn a **prediction function**. This prediction function can then be applied to new problems providing information about possible classification accuracy before even learning a classifier. One application in remote sensing would be to assess the possible accuracy of user provided ground truth to show eventual restrictions about accuracy to

Table 1: Goodness of fit and accuracy

Sample	w	accuracy
S1	0.35	85.0
S2	0.37	85.7
S3	2.08	82.9
S4	0.32	85.7
S5	2.15	83.3
S6	0.33	86.4

the user. If there is no such possibility, it will be difficult to explain to the user where the problems come from.

In order to reach the above goal we collected all satellite classification data relevant to our problem. However, data is rare and satellite image processing, machine learning, and calculation of statistical indicators had to be conducted using different software packages in this study. This turned out to be inefficient and time intensive. For an efficient continuation of the study we intend to implement a software tool integrating the steps described above. Data collected in the meantime will enable us to continue here.

4 Conclusion

Both from the user's and from the producer's point of view the question of the relationship between ground truth quality and classification accuracy is of immense importance. The method described in this paper addresses this problem. Once the results are validated as mentioned in the previous section it will be possible to find out the relationship between ground truth quality and classification accuracy and thereby to predict the classification accuracy for a given ground truth.

References

H. Bischof, W. Schneider, and A.J. Pinz. Multispectral classification of landsat-images using neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 30(3): 482–490, 1992.

J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Assoc. Publishers, New Jersy, 1st edition, 1988.

R.S. Congalton. A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing Environment*, 37:35–46, 1991.

A.L. Edwards. *Experimental Design in Psychological Research*. Holt, Rinehart & Winston, New York, 1st edition, 1972.

R.J. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, California, 1st edition, 1993.

G. Ruppert, M. Schardt, G. Balzuweit, and M. Hussain. A hybrid classifier for remote sensing applications. *International Journal of Neural Systems*, 8(1):63–68, 1997.

F. Van der Wel and L. Jansen. Accuracy assessment of satellite derived land-cover data: A review. *Photogrammetric Engineering & Remote Sensing.*, 60(4):419–426, 1994.

Authors' addresses:

DI. Georg Ruppert
Institute of Digital Image Processing
Joanneum Research
Wastiangasse 6
A-8010 Graz
Austria

Tel.: +43 316 876 1755
Fax: +43 316 876 1720
Email: georg.ruppert@joanneum.ac.at
http://www.joanneum.ac.at

Dr. Mushtaq Hussain
Eurostat
Rue Alcide de Gasperi
L-2920 Luxembourg

Tel.: +35 2 4301 35811
Fax: +35 2 4301 35989
Email: mushtaq.hussain@eurostat.cec.be

Dr. Heimo Müller
Informations Design
Technikum Joanneum
Alte Poststraße 152
A-8020 Graz
Austria

Tel.: +43 316 876 8610
Fax: +43 316 876 8601
Email: heimo.mueller@fh-joanneum.at
http://www.fh-joanneum.at