Automatic selection of training samples for multispectral image classification

Cazes, T.B., Feitosa, R.Q., Mota, G.L.A.

Abstract. The present work presents and evaluates a method to automatically select samples to the supervised multispectral classification. In order to reach out this aim, the method considers simultaneously a pair of multitemporal images of the same area, but acquired in different dates. The main assumption of this method is the existence of a linear dependence between the spectral appearances of one class in both dates. In this work, a linear regression model is employed to find out the linear models. In the experiments, the automatically selected training set was compared to the an optimal training set in terms of amount of classification error. The results …..

1. Introduction

By and large, the interpretation of low resolution remote sensing data encompasses a sequence of manual and automatic steps. Considering both his knowledge background and additional data, in the manual steps, a photo interpreter calibrates/trains the automatic algorithms and solves the inconsistencies of the outcomes.

The interpretation process taken into consideration in this paper can be subdivided in following steps [1] [2]: 1) image segmentation; 2) selection of training set; 3) supervised classification of the image; 4) post-editing of the supervised classification result. The step 1 outlines the segments, objects, present in the input image. In the step 2, a photo interpreter selects some of the segments produced in the previous step in order to train a supervised classifier which, in the step 3, classifies the segments present in the input image. In the step 4, the photo interpreter solves the inconsistencies produced by the supervised classification. As a consequence, in order to automate this process, it is necessary to emulate digitally the steps 2) and 4).

The present paper aims at improving the degree of automation of the previously described process. More specifically, this work proposes a methodology to automate the selection of the training set (step 2). The proposed approach considers as inputs two images of the same area taken in different time instances – I_{t-1}, previous time image, and I_t, current time image – and the thematic map of the area in the instant $t-1$, TM_{t-1}. The automatically selected samples corresponds to segments whose classifications were not considered changed between t and $t-1$.

The remaining of this paper is organized as follows. The next section presents the automatic selection of training set. The section 3 presents and analyzes the experiments. Finally, the section 4 presents the conclusions.

The automatic selection of training set

The spectral appearances of the classes of relevance are affected by the conditions during the image capture, among which atmospheric factors, problem with sensor calibration and land humidity level. For that reason, it is usually employed a supervised classification algorithm, trained with the spectral levels in the current image – step 3 of the interpretation process.

The interpretation process considered in this work employed an objected oriented approach. The objects to be classified are segments, with homogeneous spectral response, outlined by the step 1 of the interpretation process. This allows that the interpretation process considers attributes as shape, texture, and space relation between objects [7][8][9]. The herein employed image segmentation procedure outlines the regions with homogeneous spectral response in both input images, I_{t-1} and I_t. To perform this task, it was implemented a variation of the watersheds algorithm presented in [6].

Comentário: Acertar as referências.

All segments outlined by the step 1 of the interpretation process are candidates to integrate the training set. The **fig. 1** summarizes the procedure of automatic selection of training set.

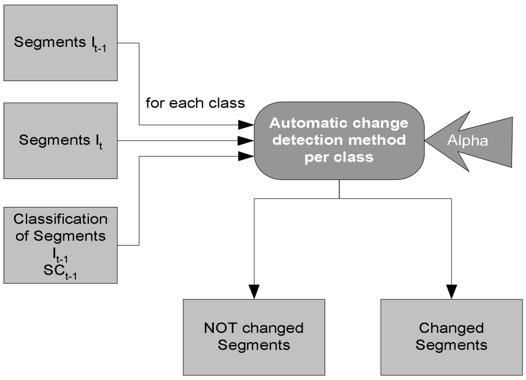

Fig. 1. The overview of the automatic selection of training set process.

The proposed method assumes some truths. First, the differences in the land cover between *t* and *t-1* are moderate. Second, natural events and image acquisition conditions are differently related with each class. But in each class they affect segments in same way. Finally, the images are assumed to be registered.

The automatic change detection method employed in this work presupposes that the spectral responses of the segments covered by one determined class in *t* and *t-1* are linearly dependent. Thus, if the spectral responses of one single segment in *t* and *t-1* fit to the linear model, it is considered as stable. On the other hand, if the spectral responses does not fit to the linear model it is considered changed. The stable segments compose the training set.

2.1 Automatic change detection method

This method is based in the linear regression model. Basically, the regression model calculates a linear model associating the spectral appearances in *t* and *t-1* of the segments which were covered by one class in *t-1*. The stable segments, segments that fit to the linear model are selected for the training set. On the other hand, the outliers – segments that does not fit to linear model – probably not belong to same class in t and t-1, so, are considered changed and do not take part in the training set.

The problem is modeling the relationship between the intensity of each segment in two multispectral image pair for all *r* bands. Each class is assumed to follow its own regression model, given by

$$y_{ij} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_r x_{ir} + \varepsilon \qquad \textbf{(1)}$$

According to linear regression theory [5] the least square estimate of β is given by

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{x}^T \mathbf{x}\right)^{-1} \mathbf{x}^T \mathbf{y} = \mathbf{H}\,\mathbf{y} \qquad \textbf{(2)}$$

Evidence whether the model fits well to a segment or not (with significance level of α) can be drawn by computing the (1–α) confidence interval for the mean of each error, given by:

$$c_i = r_i \pm t_{\left(1-\frac{\alpha}{2}, n-r-1\right)} \hat{\sigma}_{(i)} \sqrt{1-h_i} \qquad \textbf{(3)}$$

where r_i is the raw residual for the segment, $t(1\text{-}\alpha/2, n\text{-}r\text{-}1)$ is the inverse Student´s t cumulative distribution function with *n-r-1* degrees of freedom at (1-α/2) , $\hat{\sigma}_{(i)}$ is the estimated standard deviation of the error and h_i is the *i*th diagonal element of the matrix **H.** If the confidence interval given by equation (3) does not include 0 (zero), the corresponding segment is likely to be an outlier.

3. Experiments

This section describes and analyses the experiments performed to evaluate the proposed approach.

3.1 Reference Data

The images employed in the experiments are situated in the Taquari Watershed, more exactly, in the County of Alcinópolis that belongs to the State of South Mato Grosso, Brazil. Such images – composites of the bands 5, 4 and 3 of the sensor ETM in the channels R, G and B respectively – take part in the scene 224-073 of the satellite LANDSAT. The images were acquired on August 5, 1999; August 7, 2000; and August 10, 2001.

The reference classification was produced by visual interpretation by a photo interpreter expert in vegetal cover classification. In this procedure, the photo interpreter had been aided by: the previously mentioned images, the drainage map, the digital elevation model and the frames of a videography performed in October 2001.
The **Table 1**, in order to summarize the reference classification, presents the number of segments of each class, the changed segments and the percentage of changed segments.

Comentário: Thiago, trocar a tabela conforme combinado.

Class	1999	2000	2001	Changed segments			Changed Segments %		
				99->00	00->01	99->01	99->00	00->01	99->01
C1	97	85	61	25	36	50	25.77%	42.35%	81.97%
C2	62	62	57	0	5	5	0.00%	8.06%	8.77%
C3	496	510	538	11	7	9	2.22%	1.37%	1.67%
C4	28	28	28	0	0	0	0.00%	0.00%	0.00%
C5	134	132	133	2	0	2	1.49%	0.00%	1.50%
C6	6	6	6	0	0	0	0.00%	0.00%	0.00%
Total	823	823	823	38	48	66	4.62%	5.83%	8.02%

Table 1. For each class, the number of segments, the changed segments and the percentage of changed segments in the reference classification.

3.2 Experiments description

The **Fig. 2** describes the experimental procedure.

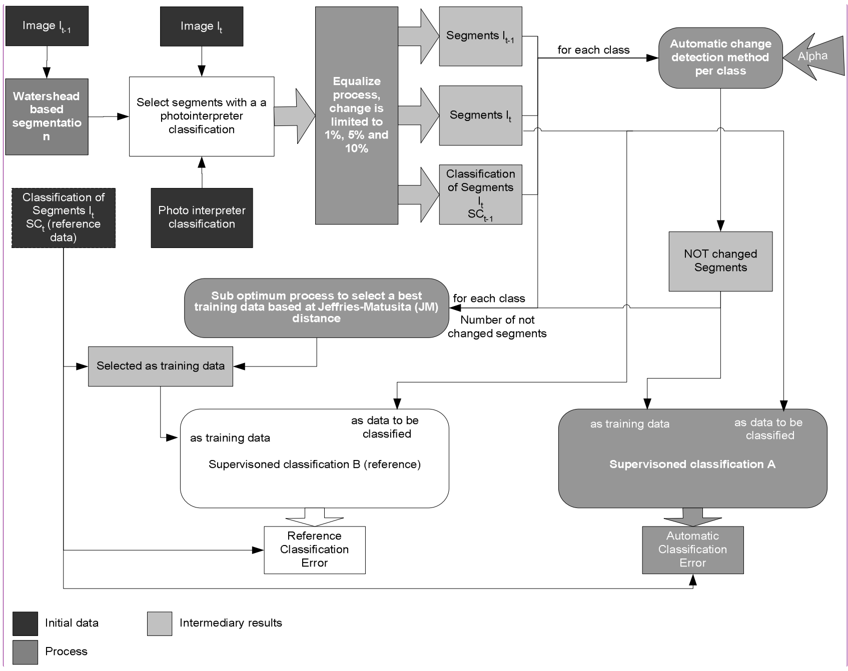

Fig. 2. Overview of the experimental procedure.

Comentário: Thiago, eliminar a segmentação e trocar o equalize process por equalization process. Trocar not changed by stable.

Analyzing the reference data summarized by **Table 1**, it can be observed that, for one single interval, distinct classes have considerably different percentage of changed segments. In order to guarantee that the same percentage of change for each class be delivered to the automatic change detection method, it is employed an equalization process which limit the amount of change to 1 %, 5 % and 10 %. Thus, it is possible to perform experiments under controlled conditions.

The set of segments delivered to the automatic change detection method contains a controlled amount of change. The linear regression model selects to the training set the stable segments inside the set of segments a controlled amount of change

To the evaluation of the automatically selected training set it is a sub optimally selected training set. The sub optimal training set contains the same number of segments than automatically selected training set.

3.2.1 Equalization procedure

This equalization procedure selects randomly part of the stable or changed segments in order to limit the total amount of change around an desired value. When the original number of changed segments for one determined class is zero, the equalized training set is composed by all stable segments available.

Class	change limited at 5%			change limited at 10%		
	Stable	Changed	%	Stable	Changed	%
C1	72	5	6.5%	72	10	12.2%
C2	62	0	0.0%	62	0	0.0%
C3	220	11	4.8%	110	11	9.1%
C4	28	0	0.0%	28	0	0.0%
C5	40	2	4.8%	20	2	9.1%
C6	6	0	0.0%	6	0	0.0%
Total	428	18	4.2%	298	23	7.7%

Table 2. – Stable, changed and percentage of changed in the equalized data set considering the reference data for 1999-2000.

3.2.2 Linear regression confidence interval values

A previous work of Mota [5] employed 15 %, equivalent to 85% of confidante interval, for the parameter alpha. In these experiments are evaluated the results for alpha equal 10% 30% an 50%, equivalent to confidante intervals of 90%, 70% and 50% respectively.

3.2.3 Sub-optimum training data selection

An exhaustive search of n segments in m total segments, are impossibly, however a sub-optimum method appear to be a good solution. In this work is selected a method called “*Sequential Backward Selection(SBS)”* in this method first all elements are selected then one by one elements is removed until we reached a required number of elements. A chose is made in this process, is that what element remove in each step. An SBS algorithm use a function to determine with element is better then another to be removed. This function is particular with problem, in this case we use a Jeffries-Matusita distance, called

JM, and this function is a form to calculate a distance between two distributions.

Function JM between distribution w_i e w_j is defined as:

$$JM_{ij} = \sqrt{2\left(1 - e^{-\phi}\right)} \quad \textbf{(4)}$$

and,

$$\phi = \frac{1}{8}\left(\mu_i - \mu_j\right)^T \left(\frac{C_i + C_j}{2}\right)^{-1}\left(\mu_i - \mu_j\right) + \frac{1}{2}\ln\left[\frac{\frac{1}{2}\left|C_i + C_j\right|}{\sqrt{\left|C_i\right| + \left|C_j\right|}}\right] \quad \textbf{(5)}$$

Where i and j are two distributions, Ci and Cj are covariance matrix of i and j respectively and μi e μj area means of each distributions.

JM function have an upper bound of 1.41(√2) and a lower bound of zero. When this function returns value near upper bound is then distributions are close, meantime when this value is near to lower bounds distributions are assumed totally separated.

In other words for each step in SBS algorithm is removed an item that's remaining items has a minor function JM.

3.2.4 Classification

In this work, both sub optimal and automatically selected training sets are employed to train supervised maximum probability classifiers. The performance of these classifier is checked in terms of the percentage of error – considering the reference data – while employed classify the equalized set.

3.3. Results and analysis

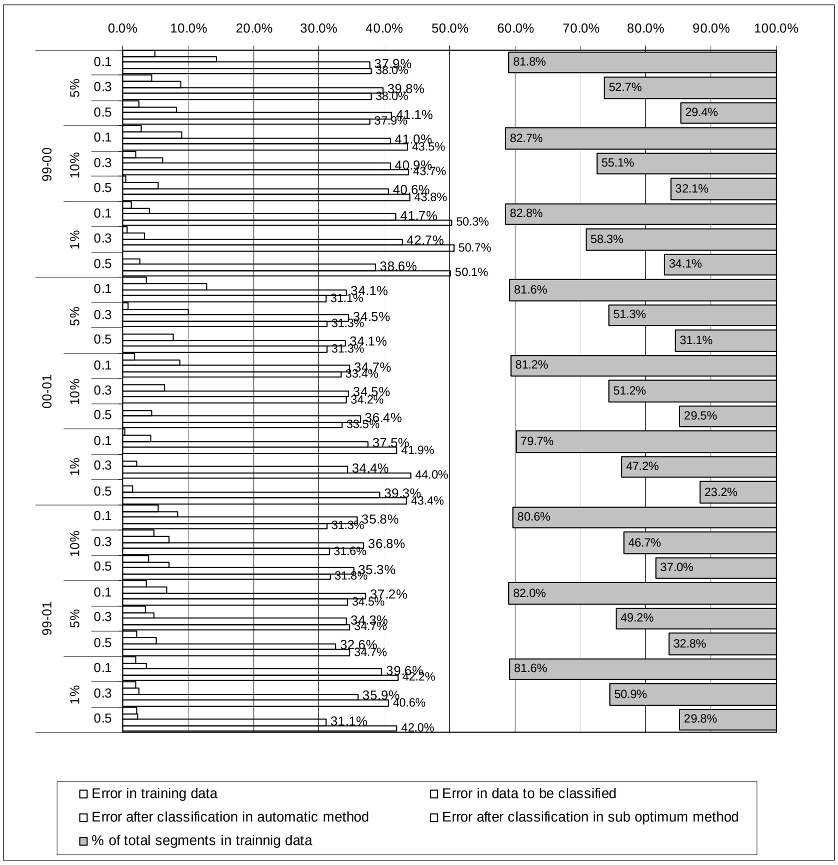

Fig. 3. Proposed method performance

4. Conclusion and future works

5 . References

[1] Mather, P.M. (1999). Computer Processing of Remotely-Sensed Images. An Introduction. Ed.Wiley, Second edition.

[2] Richards, J.A.,JIA X., Remote Sensing Digital Image Analysis – An Introduction, 3rd Ed. Springer Verlag,1999.

[3] Feitosa, R.Q., Meirelles, M.S.P., Blois, P.A., Using Linear Regression for Automation of Supervised Classification in Multitemporal Images.

[4] CHAVEZ, P. S. Jr, Image –Based Atmospheric Corrections – Revisited and Improved, PE&RS, 1996, pp. 1025-1036.

[5] MOTA. G.

[6] Gonzales. R.C.,Woods, R.E., Digital Image Processing Reading, MA: Addison-Wesley, 1992.

[7] Andrade, A., Botelho, M. F., Centeno, J., 2003. Classificação de imagens de alta resolução integrando variáveis espectrais e forma utilizando redes neurais artificiais. In: *XI Seminário Brasileiro de Sensoriamento Remoto*, Belo Horizonte, Brazil, pp. 265-272.

[8] Darvish, A., K. Leukert, W. Reinhardt, 2003. Image Segmentation for the Purpose of Object-Based Classification, In: *International Geoscience and Remote Sensing Symposium*, Toulouse, France.

[9] Yan, G., 2003. Pixel based and object oriented image analysis for coal fire research. Master Thesis, ITC, Netherlands.