

A Genetic Automatic Ground-Truth Validation Method for Multispectral Remote Sensing Images

Noureddine Ghoggali and Farid Melgani

Dept. of Information Engineering and Computer Science, Univ. of Trento,
Via Sommarive, 14, I-38050 Trento, Italy
E-mail: {ghoggali, melgani}@dit.unitn.it

PREFERRED TOPIC: T3 – CLASSIFICATION AND DATA MINING TECHNIQUES

PRESENTATION: ORAL

ABSTRACT

The goal of an inductive learning algorithm is to build discriminant functions from part of the ground-truth samples (training set) so that the generalization capability of the resulting classifier is maximized on previously unseen samples. The quantification of the generalization capability is typically performed on another part of the ground-truth samples, termed as test set. Most of the works on automatic classification have focused efforts on improving the accuracy (generalization capability) of the classification process by acting mainly on three levers which are: 1) data representation; 2) discriminant function model; and 3) criterion on the basis of which the discriminant functions are optimized [1]. These works are however based on an essential assumption that is the ground-truth samples are of unquestionable quality. In this work, we will put this assumption under light and show that the accuracy of a classification process (whatever the kind of classifier used) critically depends on the quality of the adopted ground-truth.

The two well-known ground-truth collection approaches are: 1) on-situ observation approach; and 2) photo-interpretation approach [2]. Each of them has its own advantages and drawbacks but both are subject to errors in the labeling process. In the first approach, this may occur because of georeferencing problems while in the second one spectral mismatching errors by the human analyst are the main source of problems. Since the presence of noise in a learning (training and test) set has a direct negative impact on the classification process, the development of automatic techniques for assisting and/or validating the ground-truth collection procedure is in our opinion crucial.

In the literature, very scarce attention has been paid for coping with this issue, which is mainly faced through two different strategies. The first one consists in admitting anyway the presence of noise in the data but designing a sophisticated classifier which is less likely to be influenced by this presence [3]. The second strategy is based on the removal of “suspect” samples from the learning set. In [4], the suspect samples are identified and filtered (removed) from the learning set by means of an ensemble of three classifiers (i.e., C4.5, k-NN and linear classifiers). In particular, a sample is expected to be mislabeled if it is misclassified by the ensemble of classifiers.

In this paper, we propose an alternative approach that aims at interacting with the ground-truth expert by providing him/her with a binary information of the kind

“validated”/“invalidated” for each learning sample. For each invalidated sample, the expert may confirm or not the invalidation and thus correct or maintain the adopted labeling before creating the final learning set that will be exploited in the classification process. Our ground-truth validation approach is based on viewing the mislabeled sample detection issue as an optimization problem where it is looked for the best subset of learning samples in terms of class statistical separability. This problem is formulated within a genetic optimization framework [5]-[6]. Each chromosome represents a candidate solution for validating/invalidating the available learning samples. Accordingly, the chromosome is configured as a binary string. The genetic optimization process is guided by the joint optimization of two different criteria which are the maximization of the between-class statistical distance and the minimization of the number of invalidated samples. The former is expressed in terms of the Jeffries-Matusita distance measure [1]-[2]. The latter allows to get at convergence a Pareto front from which the ground-truth expert can select the best solution according to his/her prior confidence on the reliability of the ground-truth.

Experiments were conducted on both simulated data sets and real remote sensing images. The obtained results show that the proposed automatic validation method succeeds in detecting the mislabeled samples with a very high accuracy, even when up to 30 % of the learning samples are actually affected by noise. Moreover, we show how the removal of the detected mislabeled samples impacts very positively on the accuracy of different classifiers, namely the support vector machine (SVM), the k-nearest neighbor and the radial basis function neural network classifiers.

Keywords: class separability, genetic algorithms, ground-truth, mislabeled samples.

REFERENCES

- [1] R. O. Duda, P.E. Hart, and D.G. Stork, *Pattern classification*, New York: Wiley, 2nd edition, 2001.
- [2] Richards J. A. and Jia X., *Remote Sensing Digital Image Analysis: An Introduction*, Springer-Verlag, Berlin, 1999.
- [3] Y. Lia , L.F.A.Wesselsa,b. Dick de Riddera, M.J.T. Reindersa, “Classification in the presence of class noise using a probabilistic kernel Fisher method”, *Pattern Recognition*, vol. 40, pp. 3349–3357, 2007.
- [4] C.E. Brodley and M.A. Friedl, “Identifying mislabeled training data”, *Journal of Artificial Intelligence Research*, vol. 11, pp. 131-167, 1999.
- [5] Y. Bazi and F. Melgani, “Toward an Optimal SVM Classification System for Hyperspectral Remote Sensing Images”, *IEEE Trans. Geosci. and Remote Sens.*, vol. 44, pp. 3374-3385, 2006.
- [6] N. Ghoggali and F. Melgani, “Genetic SVM Approach to Semi-Supervised Multitemporal Classification”, *IEEE Geosci. and Remote Sens. Letters*, vol. 4, in press, 2008.