TSML: Modeling Training Samples for Exchange among Image Classification Systems

Marinalva Dias Soares[1], **Luciano Vieira Dutra**[1], **Nandamudi L. Vijaykumar**[1]

[1]Instituto Nacional de Pesquias Espaciais (INPE)
Av. dos Astronautas, 1.758
Jd. da Granja – CEP: 12227–010
São José dos Campos, SP – Brazil

{marinalva,dutra}@dpi.inpe.br, vijay@lac.inpe.br

Abstract. *The variety of formats of Earth Science data has led to the data/application interoperability problem. The extraction of information from Earth observation data is, more commonly, made using classification techniques. Often users and researchers perform comparisons of classification results using different systems in search for better results. For this, it is necessary to use the same dataset, i.e, the same training samples and image. In doing so, they have difficulties in the interchange or reuse of training samples due to differences in their formats and structure, which leads them to repeat the task of extraction of samples in each system. This paper presents a novel contribution by proposing a standard model of storage of training samples, based on XML, to enable its interchange among different systems.*

1. Introduction

The characteristics of Earth Science data has led to the data/application interoperability problem. Interoperability is an issue to be tackled in several areas of computer science when dealing with different data formats and structures. When specialists in these areas have to deal with such data, it is essential to deal with such a variety of formats with a need to convert each of the formats to be understood by the reader.

These formats can range, for example, from simple free formats such as ASCII to complex formats such as GRid In Binary (GRIB), Hierarchical Data Format (HDF) or HDF for NASAs Earth Observing System (HDF-EOS). According [1], there are historical and practical reasons for having all these different standard data formats. Some agencies developed these formats for their own use and needs and others formats were developed by different communities within Earth Science for specific needs such as compactness and portability. However, as the Earth Science community became more connected and the science problems turned more complex, data sharing encouraged cross collaborations between different agencies and scientists. Users of Earth Science data now have to understand these different data formats in order to use them. In addition, the analysis applications used by the scientists must be modified every time a new data format is encountered, or the data must be translated into some other familiar format, which limits the utility of an application.

Currently, remote sensing data is widely used to provide information required at making-decision in several areas such as government, economy, environmental control, climate change, urban planning and others. Land cover/land use mapping is an important research activity for local to global scale studies, which is best conducted using remotely sensed satellite images [2]. The extraction of information from satellite images is, more commonly, made using classification techniques. Classification is a process of recognizing categories of objects and labeling entities (normally pixels) to be classified [3]. The classification process may therefore be considered as a form of pattern recognition, in which each pixel represents one point on the Earth's surface.

Procedures for supervised classification of images need to define samples. The selection of samples should be careful, because they must represent the classes, or patterns, one wants to find in the image and have an influence on the classification accuracy. The use of representative training data can help the classifier to produce more accurate and reliable results. Often users need to use classifiers of different systems (like Envi [4], Spring [5], Idrisi [6], MultiSpec [7], etc) in search for better classification results for analysis. For this, the same training samples should be used. But this task is made difficult because each system has its own definition to structure and storage of samples. Thus, to achieve interoperability in terms of data (especially in remote sensing) to perform classification of the same data using classifiers of different systems, many steps are necessary to transform the data from one format to another.

Even when systems are flexible enough to export samples in a certain format to another, there is still an issue of embedding some system specific information leading to an extra difficulty to be handled. Normally, the file structure to be exported changes a lot. The use of ASCII format has been seen as an interesting option for enabling the easy transfer of data among different systems, but usually these files have a specific syntax that difficults the development of generic readers. XML (eXtensible Markup Language) has a more flexible structure and syntax to facilitate the development of generic readers (or parsers). XML became an Internet Standard of W3C [9] as an important new technology of the universal format for structuring documents and data on the Web.

This paper presents a novel contribution of a standard format to structure and store training samples to enable the interchange among different image processing and pattern recognition systems. Section 2 describes some efforts to standardize earth observation data. Section 3 presents an overview of the proposed model and its contributions. Section 4 presents the conclusion.

2. Using XML to Facilitate the Access to the Earth Observation Data

XML (eXtensible Markup Language) became an Internet Standard of W3C [9] as an important new technology of the universal format for structuring documents and data on the Web. XML allows communities to define their own elements (tags) and hierarchical structure to describe their data. With this flexibility, the user can create tags that express the meaning of the described data making the document semantically rich and more appropriate for interchange. The ideal situation is to take this semantic power to structure the data and to make explicit relationships among entities. And this can be achieved with tools based on ontologies. Thus, the use of XML as a standard for data exchange is unquestionable. In this respect, various efforts to standardization and exchange of data are based on XML, including the standards of metadata and ontologies.

Several organizations responsible for the establishment of standards are working on the development of geographic information standards, obtained or not by remote sensing. Organizations that are more involved in the standardization are Federal Geographic Data Committee (FGDC), which develops standards on geographic information for National Spatial Data Infrastructure (NSDI) [10], the Open GIS Consortium (OGC), which sets the industry specifications for interoperability of geo-processing software and services [11], and the International Organization for Standardization (ISO) Technical Committee 211 (TC 211), which sets the international standards on geographic information [12]. But there is no standard model or format for training samples or image structure for image processing and pattern recognition systems.

Many efforts in standardization and interoperability are based on XML. Some previous research (including in Brazil) based on XML that aims interoperability include: Earth Science Markup Language (ESML) [1], Geographic Markup Language (GML) [11], Remote Sensing Markup Language (RSML) [13] and the XML-based Brazilian format GeoBR [14, 15].

GML and GeoBR were developed with the aim of encoding geographic information in XML for storage and transport over the web. Being an open standard, GML has been widely used as a means for storage of spatial data from different GIS. GML provides a set of rules that allow the user to write her or his own schema to describe data in specific areas. However, some users see this as a disadvantage because every new GML Schema means a whole new XML to support in software.

But in GeoBR the user doesn't need to create her or his own schema. Moreover, GeoBR provides a dictionary of terms to provide semantic interoperability, while GML does not by itself solve the semantic integration of heterogeneous schemas among different geographic information systems.

RSML was proposed as an attempt to fill the gaps in formats more usual in remote sensing. Unlike GML and GeoBR, RSML presents a model of remote sensing metadata used by data producers and a model for raster remote sensing operators/algorithms. Therefore, no research was found reporting the use of RSML by others systems or applications.

ESML is not a data exchange format by itself. It is different from other works described for neither being a data format nor a data model, but a technology that allows applications to share information without the need to develop a data converter, even if they use different formats from each other. This is possible because ESML provides a standard method for describing the structure of a dataset in several common scientific data formats.

Using ESML and its API (Application Programming Interface), applications can understand and use a data file regardless of its format as long as the format has been fully described using ESML. Software developers can build data format independent of their scientific applications utilizing the ESML Library. ESML is extensible and new data format can be added in ESML Library.

The current ESML schema supports descriptions for unstructured data formats such as ASCII, Binary, GRid In Binary (GRIB), network Common Data Format (netCDF), Hierarchical Data Format - Earth Observing System (HDF-EOS) and WSR88D Level II. A separate ESML element is defined for each individual data format. Descriptions for additional data formats can be added to the schema. ESML schema is publicly available via the ESML web site http://esml.itsc.uah.edu.

Although the work listed above have been developed to provide interoperability among different systems that make use of such data, none of them has a common model for structure and storage of training samples so as to be reused by classifiers of different systems.

3. Description of the Solution Proposal to Interchange of Training Samples

The difficulties in exchanging training samples reported in Introduction led to the development of TSML (Training Samples Markup Language). TSML, still under development, is a standard format, based on XML, to structure and storage of training samples to enable the interchange among different image classification systems.

The model here described is designed to facilitate the exchange of training samples among different systems so that business representatives from each of these systems can quickly write routines for direct conversion to their internal format. This is easily achieved due to the ease of writing a parser to read XML and facilities provided by the structure of a well formed XML document. Being based on XML, TSML is readable, independent of the internal representation of each machine and flexible, allowing addition of new elements according to the needs of each system.

The conceptual model of TSML, which includes the main data object that can be part of a process of image classification, is shown in Figure 1.

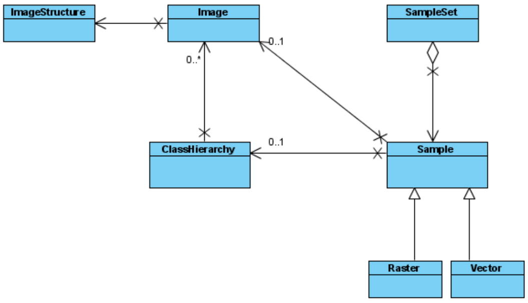


Figure 1. Conceptual Model of Samples

The model shown in Figure 1 shows that the training sample can be in raster and vector formats, can have a hierarchy of related classes (classes are the patterns that the samples represent the image, or priori knowledge about the area covered by the image, for example, forest, agriculture, pasture, water, etc). The class hierarchy should be independent of the training samples and also the image, so that can be exported/imported independent those data. The model also shows that the image has a associated structure that can be exported/imported among the systems. This structure consists, generally, in the segmentation of the image, which represents the regions of image.

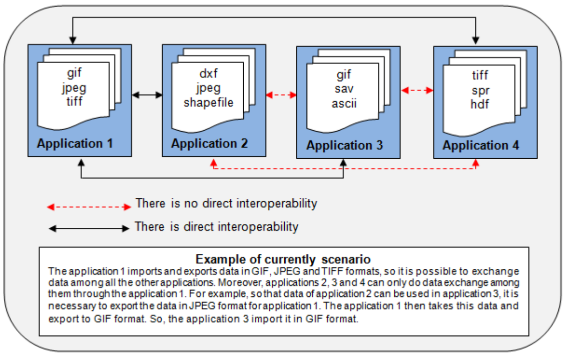
The XML Schema corresponding to the model should contain tags describing all data fields required for supervised classification of raster and vector data. The model is structured in XML so that it can easily be stored in a relational database. Each of the XML tags that store content may represent a field of a table in the database. This extends the generalization of the model to be used also for data mining algorithms.

The XML Schema of the model describes the elements and roles of compostition of one sample and it differs in the following aspects from other ASCII-based formats, such as Envi and Spring. For example:

- As TSML is based on XML, has a structured way for easy storage and reading of data related to training samples and image structure.
- All data related to the sample, for both raster and vector format, are contained in a single file that prevents the user having to work with multiple files depending on the type of information of the sample he or she wants to use.
- Each element may be an independent entity, so that each system exports or imports only the data necessary to perform the classification. For example, if the samples are georeferenced, the elements corresponding to the X and Y coordinates are not necessary to locate the sample in the image.
- The model can accomodate samples in raster or vector format, or both.
- The model can accommodate data from the training as the average vector, covariance matrix etc.

With this model, a single parser is able to retrieve samples in their raster or vector format and, especially, easily read only the relevant information for each system. If at any other time the system needs to recover other types of information in the file, no changes in the parser are necessary. In other words, write the parser once and use it in different ways.

The use of data from an application by other applications, actually, occurs as shown in the Figure 2.



Example of currently scenario

The application 1 imports and exports data in GIF, JPEG and TIFF formats, so it is possible to exchange data among all the other applications. Moreover, applications 2, 3 and 4 can only do data exchange among them through the application 1. For example, so that data of application 2 can be used in application 3, it is necessary to export the data in JPEG format for application 1. The application 1 then takes this data and export to GIF format. So, the application 3 import it in GIF format.

Figure 2. Scenario of data interchange among remote sensing system

The Figure 3 shows the interoperability among different applications through TSML.

As be seen in Figure 3, the applications access the format TSML through a parser that reads XML. The work in development will support the semantic equivalence through a dictionary of ontologies (or terms). The definition of equivalence is made through a function, which input data are the elements (terms) of TSML file and terms of the application. This dictionary can be updated by applications, which can also choose to export this dictionary with the file TSML for other applications. With XML, no conversion between formats is required, which promotes interoperability and reduces the effort required in developing a converter of unknown or complex formats.

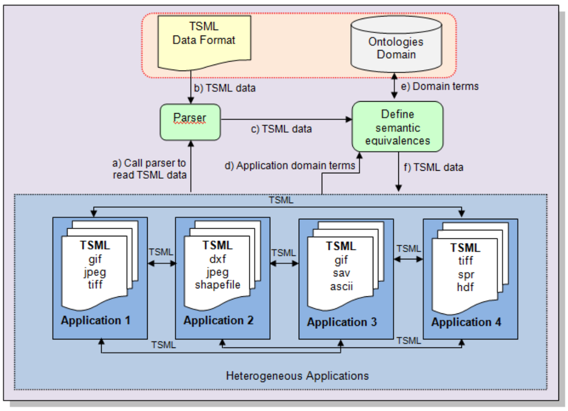
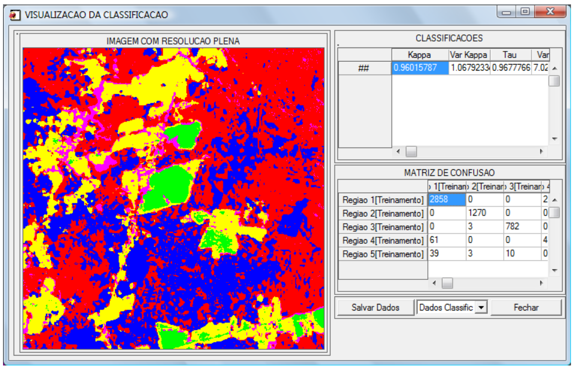


Figure 3. Interoperability among different applications through TSML

As XML is well structured, the development of parsers is easier. Since XML is independent of platform and operating system, the adoption of TSML is feasible and promising, because interoperability will occur among an unlimited number of applications. With this format the user does not have to define her or his own tags. The format already contains the necessary tags to identify any information of the sample, but the schema can be extended to accomodate any other specific information.

3.1. Case Study

To validate the exchange of samples TSML format, was a case study with the Envi and SACI. SACI (Sistema de Análise e Classificação de Imagens) is an internal system implemented at INPE (coded in IDL [4]) to meet to the specific needs of a group of researchers of the Image Processing Department. The original format for samples is the ".sav" binary format. For testing purposes, was performed a classification of Landsat-TM image, bands 3, 4 and 5 in SACI using the Maximum Likelihood classifier. The training samples were extracted in SACI and stored in the TSML format. The figures 4 and 5 show these steps. After, the samples was imported directly in Envi without any additional need for conversion to another intermediary format.



CLASSIFICACOES

	Kappa	Var Kappa	Tau	Var
##	0.96015787	1.0679233	0.9677766	7.02

MATRIZ DE CONFUSAO

	1[Treinan	2[Treinan	3[Treinan	4
Regiao 1[Treinamento]	2858	0	0	2
Regiao 2[Treinamento]	0	1270	0	0
Regiao 3[Treinamento]	0	3	782	0
Regiao 4[Treinamento]	61	0	0	4
Regiao 5[Treinamento]	39	3	10	0

Figure 4. Classification in SACI system with maximum likelihood algorithm

With the importation of samples in TSML format in Envi was possible to classify the same image, using the same samples, same classifier and compare the results. The classification results showed that there was no data distortion in import process.

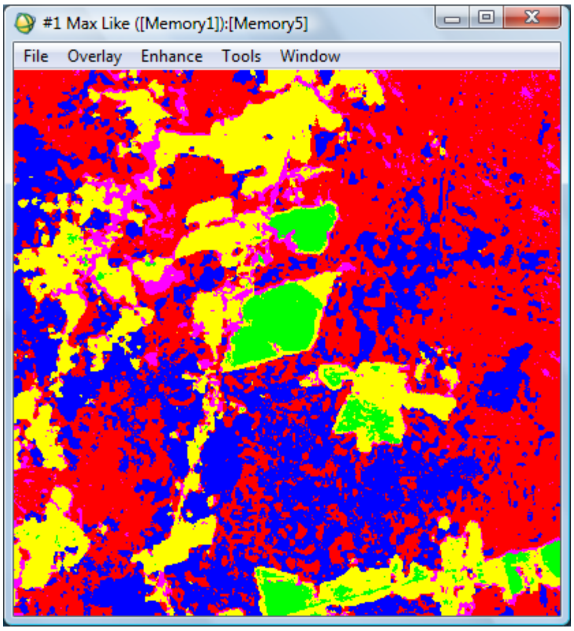


Figure 5. Classification in Envi system with maximum likelihood algorithm

4. Conclusion

This paper has showed the proposal of a standard model of training samples for exchange among different systems. Being based on XML, this model allows the exchange of samples with minimal effort to import the data without their loss or alteration. The implementation of the model for samples in vector format and structure of images is still under development, but tests for validation of the import samples in raster format were made in an important system for the analysis of images, the Envi. The tests showed the feasibility of the model through the direct import of data into XML in Envi for classification. The results of the classification were similar to those of the original, the SACI. These results confirm the advantages of a model based on XML.

References

[1] RAMACHANDRAN, R.; GRAVES, S.; CONOVER, H.; MOE, K. Earth science markup language (esml): a solution for scientific data-application interoperability problem. Computer and Geociences, v. 30, p. 117-124, 2004.

[2] KAVZOGLU, T. Increasing the accuracy of neural network classification using refined training data. Environmental Modelling and Software 24 (2009), pp. 850-858. www.elsevier.com/locate/envsoft

[3] MATHER, P. M. Computing Processing of Remotely-Sensed Images. John Wiley. 2004.

[4] ITT Visual Information Solutions. Envi Homepage. http://www.ittvis.com/ProductServices/ENVI.aspx

[5] CAMARA, G.; SOUZA, R. C. M.; FREITAS, U. M.; GARRIDO, J. Spring: Integrating remote sensing and gis by object-oriented data modelling. Computer and Graphics, v. 20, p. 395-403, 1996.

[6] CLARK LABS, Clark University. Idrisi Homepage. http://www.idrisi.com

[7] BIEHL, L., LANDGREBE D. MultiSpec - A Tool for Multispectral-Hyperspectral Image Data Analysis. 13th Pecora Symposium, Sioux Falls, SD, August 20-22, 1996. http://cobweb.ecn.purdue.edu/~biehl/MultiSpec/

[8] WITTEN, I. H., FRANK, E. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, June 2005. ISBN 0-12-088407-0.

[9] W3C. XML - eXtensible Markup Language. http://www.w3.org/XML/

[10] FGDC Homepage. http://www.ofgdc.gov/

[11] OGC Homepage. http://www.opengeospatial.org/

[12] ISO/TC 211 Geographic information/Geomatics. http://www.isotc211.org

[13] MOHANTY, K. K. Rsml: A proposed xml based format and processing language for remote sensing data. Indian Cartographer, p. 330-335, 2002.

[14] JUNIOR, P. L.; CAMARA, G.; PAIVA, J. A.; MONTEIRO, A. M. V. Intercambio de dados geograficos: Modelos, formatos e conversores. In: Anais do I WorCap - Workshop de Computação Aplicada. Sao Jose dos Campos, Brasil: INPE, p. 36-38, 2001.

[15] JUNIOR, P. de O. L. GEOBR: Intercambio Sintatico e Semantico de Dados Espaciais. INPE, 2002. Dissertacao de Mestrado. http://mtc-m16.sid.inpe.br/col/sid.inpe.br/jeferson/2004/09.20.09.49/doc/publicacao.pdf