Project proposal

Vegetation mapping and monitoring using hyperspectral data and advanced processing techniques.

ESR: Karoly Bakos
Supervisor: prof. Dr. Paolo Gamba
University of Pavia (UNIPV)

The researcher will be required to work on vegetation monitoring using hyperspectral airborne or space borne data in both mountainous and urban areas. The research will require the definition and the implementation of a full processing chain of hyperspectral datasets that is able to start with radiometrically calibrated data and ending with maps of vegetated species and tree status in various environments. Testing of known algorithms and development of innovative and advanced ones for each step of the defined chain will be required, with particular emphasis on geometric and atmospheric corrections and classification procedures based on known spectral of vegetated species. Joint spectral and spatial analyses will be also considered to implement the final mapping chain. Mandatory part of the Ph.D. will be also validation and accuracy assessment of the outputs of the processing chain, by means of both in situ spectral measurements and visual inspection.

Introduction

In this research the potential of hyperspectral datasets in vegetation research and monitoring is investigated with the aim of further increase the application possibilities of hyperspectral images mainly for producing generic vegetation maps. Present algorithms and methods are revised and new approaches are proposed both for data dimensionality reduction and for classification that are playing the key role in hyperspectral image interpretation as will be presented. For data dimensionality reduction a new data transformation methodology is suggested that for calculating the data transformation function takes into account the data cloud on a per class basis using a training dataset as opposed to traditional methods such as Principal Components Analysis (PCA) or Minimum Noise Fraction transformation (MNF) that are only taking into account the distribution of the dataset as a whole to derive the transformation functions. For data classification a new Decision Tree Classification (DTC) approach is proposed where the optimum DTC construction is propagated using a series of test classifications of a limited subset of the data.

With the improvements in dimensionality reduction it is not only desired to achieve a better data compression and therefore smaller data dimensionality but is also aimed that the transformed dataset will be better suited and prepared for data classification. By the DTC approach it is aimed that an ensemble of different classifiers and different input data sources on a per land cover class basis enables to achieve the highest overall mapping accuracy. Also the DTC approach can be easily extended when further classifiers or transformation algorithms are available for the users.

Hyperspectral datasets are more and more available and more and more used in certain applications of different kind. Both space borne and airborne instruments are available for data collection at a broad range of different waveband range and resolution. Presently many applications can be found in the technical literature using hyperspectral datasets. Most of them are dealing with detection of physical properties...
of different surfaces and static physical properties of objects on wrath’s surface and a few of them is addressing the problem of detecting vegetation properties. The latter applications based on the properties of hyperspectral data mainly focusing on the retrieval of different physical properties of the vegetative surface and optionally address the problem of linking those properties to biological entities of the different species. Opposed to multispectral remote sensing data hyperspectral images are not often used for detecting land cover and generate generic mapping products. This is because of the huge data dimensionality and the high amount of redundant information found in the data that makes the processing and interpretation procedure more difficult than in case of multispectral imagery. Therefore in many case instead of the traditional way of data classification a reverse approach is used and based on the retrieved physical properties of the surface the actual land covers desired be identified and in many case in hyperspectral data applications land cover maps originated from different source are used as external information.

Also in these but in any kind of processing of hyperspectral datasets data storage, processing and interpretation is generally an issue that is still to be solved in the future, In most application the data which already has huge dimensionality has to be processed in several steps each of which results further data that has to be stored too. Because there is a lack of framework for data storage, naming convention, documentation and meta data structure regarding to hyperspectral data it is particularly difficult to exchange interim results and intermediate data between different applications although using the same original dataset. Also it is difficult the reproduction of certain experiments as data complexity is significantly rising with each step of the data processing.

Aims and objectives

The broader perspective of this study is to make better use of hyperspectral image data and enable the usage of hyperspectral images for vegetation mapping. This is of importance as hyperspectral sensors are now more available but most of the cases there is still a need for multispectral data for generic mapping procedures. The ability of using hyperspectral data for generic mapping would also enable to use the same dataset in certain applications where now only the physical parameters are retrieved only to link those parameters against the actual land cover without the need for external land cover sources.

According to the brief introduction above the aim of the research project is to investigate the possibility of creating a comprehensive data processing methodology for land cover mapping using hyperspectral datasets and advanced processing techniques that is especially efficient for differentiates among different vegetation species.

According to the aim of the research to achieve it the identified objectives are as follows:

- Investigate the presently available feature selection and feature extraction techniques.
- Identify whether it is possible to further increase the positional of the data dimensionality reduction throughout novel algorithms.
- To investigate the data classification approaches available to interpret the data.
- Develop a composite methodology that is able to integrate the advantages of present classification algorithms while providing flexibility for end users to easily adopt it according to their datasets and available resources.
Background and Technical literature

A broad range of technical literature was reviewed in order to seek for solutions for each of the identified issues. Some of the problems are already addressed and results can be integrated into this research. Based on the technical literature research it became clear that although some of the problems that is subject of this study was already investigated but without the aim of a complete processing chain therefore the methodologies are specific and individual and the methodology to integrate them is still within the focus of present research. In the project the term processing chain is using the theoretical background of Gamba et al (2005) that consist of a two stage processing chain with different steps as illustrated in Figure 1.

![Figure 1. Generic concept of processing chain for hyperspectral data processing using Gamba et al (2005)](image)

During the first period of the project also an extensive test of different data processing chains were carried out with many of the presently available methodologies. The methodologies were assessed and overall well performing chains were identified. After the processing also an in depth analysis of each chain and the arising differences were assessed by comparing accuracy levels of each processing chain with each other on a per pixel per class bases. The visual representation of the applied processing chains can be seen below in Figure 2.

![Figure 2. Graphical representation of the approach applied to while designing the different processing chains. Paths with highlighted arrows are identified to be more suitable than others. (FS-TD=Feature selection with transformed divergence; MNF=Minimum nois fraction rotation; PCA=Principal components rotation;SAM=Spectral angle mapper; NN=Neural Net)](image)

In the project more datasets and different classification chains were combined and the best processing chain was aimed to be identified. A classification and data processing system was set up that could carry out the task and after proceeding the results were analysed. The most relevant accuracy levels are shown in Table 1.

University of Pavia – Department of Electronics, via Ferrata 1. Pavia 27100 - ITALY
Table 1. Sample of overall accuracy levels obtained using different processing chain configurations.

<table>
<thead>
<tr>
<th>Input data source</th>
<th>Dimensionality reduction</th>
<th>Classification</th>
<th>Post processing</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNF</td>
<td>15 band</td>
<td>NN</td>
<td>+</td>
<td>87.96%</td>
</tr>
<tr>
<td>S</td>
<td>FS TD</td>
<td>ML</td>
<td>+</td>
<td>87.92%</td>
</tr>
<tr>
<td>S</td>
<td>FS TD</td>
<td>SAM</td>
<td>-</td>
<td>57.77%</td>
</tr>
<tr>
<td>S</td>
<td>FDAFE</td>
<td>ML</td>
<td>-</td>
<td>72.08%</td>
</tr>
<tr>
<td><strong>Multi stage DTC approach</strong></td>
<td></td>
<td></td>
<td></td>
<td>84.37%</td>
</tr>
</tbody>
</table>

**Proposed methodology**

**Dimensionality reduction**

The methodology proposed for dimensionality reduction is a feature extraction methodology that is instead of taking the whole data cloud to calculate transformation function for data transformation uses the training dataset too. For example PCA uses the maximal Euclidean distance found in the hyper body to formulate the transformation equation and tries to use the maximum data variation that is found in the dataset as a higher level structural basis. MNF transformation instead tries to limit the data range to a critical mass that is likely to be real data and defines the new data axes using hyper body for the calculations. The fact that in most studies MNF was found to be superior against PCA transformation when the data was classified it is expected that for data classification the maximum data variance is not playing the primary role.

In this study subset of the data that calibrated is aimed to be used and transformation functions are aimed to be calculated using this subset. The distribution of this training data within the feature space enables to extract more relevant components of the data that can improve the accuracy of the classification procedure.

In this research the possibility of using transformations where the data is projected based on the distribution of the training data of single classes is aimed to be used where at minimum as many components are generated as many classes can be found in the training data each of which will be a component in which a class is more obviously separable than the others and two additional components that will maximise the variance among the marginal values and the classes of the training data.

The former aim is addressed by calculating a component that minimises the variation of a class taking the line of the highest Euclidean distance within the data cloud of corresponding to a single class for each one of the training classes. With this it is possible to minimise the variance of the particular class within the component. The latter aim is addressed by calculating the additional component using the extreme data values from all of the training data and maximise the valuation of them in the component and secondly the mass centre of the training classes are used and new component is calculated that maximises the variance of the mass centres of the training classes.

As the above methodologies are not exist the algorithms has to be developed, implemented and their performance has to be tested. Also the computational costs have to be measured whether the methodology is feasible.
Data classification

For data classification a novel methodology is suggested that enables flexibility and high overall accuracy of data classification. Based on the complexity of hyperspectral data and the difficulties mentioned in technical literature regarding to hyperspectral data classification a multi stage Decision Tree structure is suggested for data classification. As the biggest issue arising when constructing a decision tree classification system is to identify the optimal construction of the decision tree (optimisation) the methodology proposed is mainly focusing on this task. In this research an optimisation procedure is introduced that is based on the predicted overall accuracy level of the decision tree that is propagated from a limited number of test classifications of the available data sources with the available classification methodologies. This feature enables the methodology to be able to adopt any new input source and classification algorithms when they become available for the end-users.

The methodology is consists of a test classification phase and an optimisation phase. First the available input data and the derived information (PCA, MNF, selected features ...etc) are classified using the available algorithms and confusion matrices are generated. These test classifications are single stage classifications where all input is classified into all of the available classes in one step.

After this a methodology is introduced that is able to convert the created confusion matrices and able to simulate classification result of a decision tree classification system composed using the available input data sources, classification methodologies varying over the decision tree nodes. For each node one input data source and one classification method can be selected with optional number of classes furthermore classes can be aggregated for further classification nodes. This means that with the methodology the overall accuracy any random decision tree structure can be propagated from the confusion matrices of the single stage test classifications.

Based on pure logic and that the Decision tree is a hierarchical structure and the design process follows a top down approach it is possible to keep the overall accuracy above a certain level by selecting for each node a classification that is above that certain level or it is also possible to select for each node the best possible classification therefore optimise the decision tree structure using the maximum possible accuracy level as design criteria. The concept of Decision Tree possible configurations is illustrated in Figure 3.

Figure 3.Possible DTC configurations of a simple 4 class classification problem. The implemented code should automatically select the best possibility based on propagated overall accuracy level of the DTC.

Because of the complexity of the calculations and the relatively high number of test confusion matrices (based on the number of input, classification methods, and classes) an algorithm is to implemented that is able to use standardised tabular format confusion matrices and generate the optimal design of the decision tree. This
will serve a list of decision nodes coupled with the classes to be detected and the classification method for the node that can be used to construct the decision tree in a commercial image processing software or in a manual fashion,

During the research it is particularly important to investigate the limitations of the methodology and compare it to standard classification methodologies. The computational costs are also to be measured as it is relevant for end users.