Unsupervised classification of EO-1 Hyperion hyperspectral data using Latent Dirichlet Allocation

Costachioiu Teodor[1], Bashar Alzenk[1], Rodica Constantinescu[1] and Mihai Datcu[2]
[1]Applied Electronics and Information Engineering Dept., University Politehnica Bucharest, Romania
teodor.costachioiu@ceospacetech.pub.ro
[2] DLR - German Aerospace Center, Oberpfaffenhofen, Germany

Abstract — **In this paper, we perform land cover classification using hyperspectral data acquired by the EO-1 Hyperion spaceborne platform using Latent Dirichlet Allocation text modeling tool, experiments being carried on a Hyperion data scene acquired on 13 May 2011, covering an agricultural area located east of Bucharest, Romania.**

I. Introduction

The use of remotely sensed data plays an important role in global monitoring and assessment, with multiple applications such as urban monitoring, agriculture, soil mapping or disaster management. Airborne hyperspectral imaging sensors have been used by more than a decade [1], providing an almost continuous representation of the spectral response, with the tradeoff of reduced spatial resolution. The high cost of operating airborne hyperspectral sensors and their limited range have been restrictive factors in using hyperspectral data on a wide range of applications. This disadvantages were overcome in late 2000, with the launch of Hyperion sensor onboard the EO-1 remote sensing platform [2], since then EO-1 being the only source of spaceborne hyperspectral data. The EO-1 satellite is placed on a sun-synchronous orbit, at an altitude of 705 km. It is capable of imaging a 7,65 km swath, with a spatial resolution of 30m. The Hyperion sensor covers images a spectral range of 0.4 - 2.5 µm in 220 spectral bands, with a spectral resolution of 10nm. Research has proven that using EO-1 hyperspectral data offers significant advantages over the use of multispectral data in applications such as detecting crop diseases [3], discrimination between crop types [4] as well as for land use/land cover classification [5]. Among the current methods for EO-1 Hyperion data analysis we can mention the use of MNF transformation [6], the use of various spectral indices [7], the classification of the principal components performed using Niche Hierarchical Artificial Immune System [8] or the use of pixel unmixing methods [9].

In this paper we present a new approach towards EO-1 Hyperion data analysis by adopting the Latent Dirichlet Allocation (LDA) text modeling tool to discover interesting patterns in the hyperspectral data. The paper is structured as follows: a short introduction of Latent Dirichlet Allocation for text modeling in the second section. In section III we present the methodology related to finding a suitable representation of EO-1 data in order to allow analysis using LDA. Finally, section IV presents the experimental setup and the results we have obtained by the proposed method.

II. Latent Dirichlet Allocation

Considering the inherent complexity and high dimensionality of hyperspectral data, computerized analysis of this type of data has to overcome the semantic gap, defined as the problem of mapping from the low-level features to the high-level semantic concepts expected by the user [10]. Significant advances towards bridging the semantic gap have been made in text analysis domain. The Latent Semantic Analysis [11] used singular value decomposition to mimic the way people associate words. A probabilistic approach towards semantic modeling was proposed in PLSA, the distribution of topics (classes) over words being modeled as multinomial distribution [12]. PLSA also introduced the idea of a generative model at document level.

In 2003 Blei et.al. have introduced the Latent Dirichet Allocation model [13], in which topic mixtures are sampled from a Dirichlet distribution, thus completing the PLSA model by introducing a generative model at corpus level. As such, LDA has the ability to generalize over unseen documents from the same corpus.

In text analysis, the following definitions are necessary prior to describe the LDA generative process:

- the *corpus* is a collection of M documents and is denoted by $D = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_M\}$

- a *document* is a sequence of N words, denoted by $\boldsymbol{w} = \{w_1, w_2, \dots, w_N\}$, where w_n is the n-th word in the document.

- *words* are the basic units of discrete data, defined as being items from a vocabulary indexed by $\{1, \dots, V\}$. A word w_n can be modeled as a V-length, with the property that $w^n = 1$, and $w^u = 0$, for any $u \neq n$.

With the above definitions, for each document $\boldsymbol{w}$ in the corpus D the generative process of LDA can be described as follows:

1. Choose $N \sim Poisson(\xi)$
2. Choose $\theta \sim Dirichlet\ (\alpha)$
3. For each of the words w_n
   a. Choose a topic $z_n \sim Multinomial\ (\theta)$
   b. Choose a word w_n from $p(w_n | z_n, \beta)$, a multinomial probability conditioned by the topic z_n

The Poisson distribution was chosen to model document length based on empirical observations and is not critical for the generative process. Other document length distributions can be chosen as needed.

Finally, like many other text modeling tools, LDA assumes the *bag-of-words* model, in which the order of the words and the grammar are ignored, only the word count having importance in training a LDA model.

According top the LDA model the likelihood of a document can be expressed as:

$$p(\boldsymbol{w}|\alpha,\beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n} p(z_n|\theta)p(w_n|z_n,\beta)\right)d\theta$$

The above equation is intractable. Only an approximate solution can be found, using inference algorithms such as variational expectation maximization inference or Gibbs Sampling [14]. In this paper we used the first method proposed by Blei [15].

In the LDA model the dimensionality k of the Dirichlet distribution, and thus the number of topics, is considered known, and is a user-given parameter. Estimation of the number of topics can be performed using the measure of perplexity, defined as:

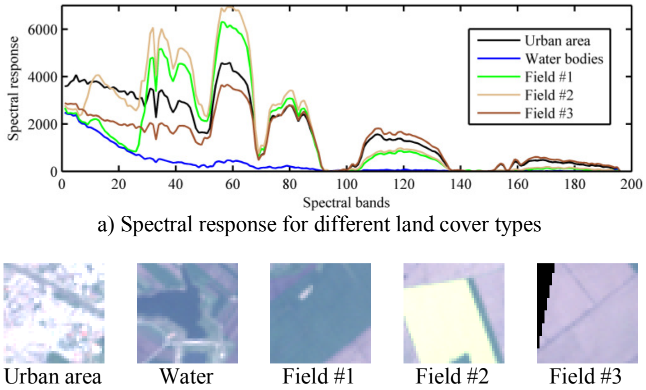
$$perplexity(D_{test}) = exp\left(-\frac{\sum_{d=1}^{M}\log p(\boldsymbol{w}_d)}{\sum_{d=1}^{M}N_d}\right)$$

where $D_{test} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_M\}$ is a test corpus consisting of M documents, each document having N_d words.

The use of LDA is not limited to text analysis. In fact, almost any collection of discrete data can be analyzed using LDA. To apply LDA outside text domain the user must define a synthetic language, finding equivalents for words, documents and corpus, while respecting the bag-of-words assumption. In remote sensing analysis this is often achieved by extracting a set of features from data, such as mean and variance from patches of Quickbird data [16], of spectral signatures and fuzzy templates extracted from patches of Landsat and Quickbird data [17], or the analysis of satellite image time series by extracting change descriptors between pairs of images [18], followed by a vector quantization process, usually performed using k-means.

III. LDA FOR HYPERSPECTRAL DATA ANALYSIS

Hyperspectral data captured by EO-1 Hyperion sensor is characterized by a medium spatial resolution, a pixel corresponding to an area of 30m x 30m. The hyperspectral sensor records 224 bands, on which only 204 channels are usable, due to insufficient signal at the extremes of the spectral range, respectively bands 1-9 and 225-224. A further overlapping of the VNIR bands 54-57 and SWIR bands 75-78 is observed, and used mainly for calibration between VNIR and SWIR detectors [19]. In this paper we have chosen to ignore the spectral response in bands 75-78, in order to avoid unnecessary redundancy and to reduce the complexity of the trained LDA model. In total, a number of 196 bands were selected for analysis.



a) Spectral response for different land cover types

b) Patches of image corresponding to the above spectral profiles

Fig.1 Spectral profiles recorded at pixel level

Considering the characteristics of the EO-1 Hyperion sensor, we aim to preserve the spatial resolution while fully exploiting the rich information recorded in the spectral domain. As such, the proposed method is applied at pixel level, the semantic meaning being extracted from the spectral information.

Examples of pixel-level spectral profiles for urban areas, water bodies and several types of land use in agriculture are displayed in figure 1. We can notice that urban areas are characterized by a high magnitude of the spectral response in bands 1-7. As infrared wavelengths are absorbed by water, we can observe a low response of water bodies in the infrared range. Land occupied by agriculture can be characterized by a high magnitude of the spectral response in bands 35-75, and by a moderate response in bands 110-130. Barren lands have a spectral profile characterized by a moderate response in bands 1-50, and a spectral profile similar to the urban areas in the rest of the spectral range.

To make possible the analysis of hyperspectral data using LDA a suitable bag-of-words representation of data has to be defined. Under the assumption that the spectral response for a pixel carries sufficient semantic information, in this paper we choose to consider pixels as documents, and the names of spectral bands as words. As such, the spectral information recorded for a pixel is equated with the histogram of word occurrence in the text domain, the band names being treated as words, and the magnitude of the spectral response in a given spectral band being considered as the frequency of the word occurrence in the document. By analogy with the text domain, the whole image is treated as a corpus.

Once the analogy between text analysis and hyperspectral data is defined, the next step is to discover a Latent Dirichlet Allocation model of the data. To achieve this we take advantage of the LDA ability to generalize and we select only a part of the data for training. As in the LDA model the number of topics (classes) is a user-given parameter, we train LDA models for different number of topics and we select the model that returns the lowest value of perplexity as the model with the optimal number of classes. Once the LDA model is trained, we can perform inference over the whole data set to estimate the topic proportions inside each document.

IV. Experiments and Results

Experiments were performed on an EO-1 Hyperion data set acquired on 13 May 2011, imaging an area located east of Bucharest, Romania. Several land use/land cover types can be observed in the studied area: land use allocated to agriculture dominates, several classes of agricultural use being observed. The area also features urban areas, water bodies and forests.

Following the proposed method, we have selected a training set of 10% of the original data, for which we have trained several LDA models for different number of topics. We found that the lowest value of perplexity is obtained for a number of 50 classes. As a result of LDA training a data model characterized by the β parameters of the topic over words distributions and by the topic mixture parameters γ. Based on these parameters we can infer the mixture of topics at document level.

Several use scenarios can be formulated based on the results of the LDA analysis.

Most commonly the goal is to classify the hyperspectral data into a number of S classes, the pixel being assigned to the class that maximizes the likelihood $S = argmax_S\, p(\boldsymbol{w}|\alpha,\beta)$. Based on a LDA model with 50 topics we have obtained the classification in figure 2.

In the above classification several interesting spectral classes can be observed. In agriculture for example, for land areas that in the visible range look similar, LDA found different classes, enabling us to observe patterns that otherwise would be difficult to distinguish. A detailed view of such area is presented in figure 3, in which LDA has identified two different classes in the same land area.

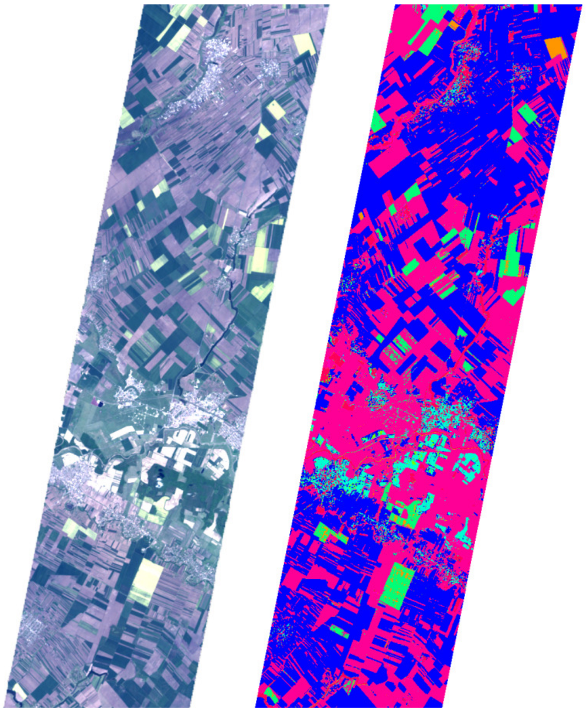


Fig.2 False color image of the studied area (bands 29, 23, 16 as RGB) and classification obtained by LDA

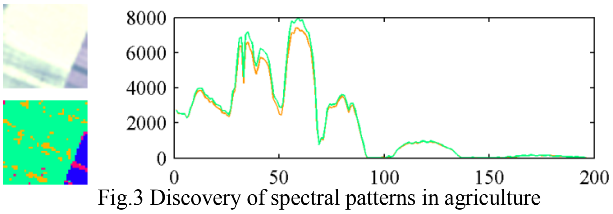


Fig.3 Discovery of spectral patterns in agriculture

In figure 3 we show false color RGB image of bands 29, 23 and 16, the classification result and the spectral response in bands 1-196 for the two selected classes, highlighting the differences discovered by LDA.

Starting from the above results, another use scenario of LDA is to perform the analysis of the distribution of topics (classes) over the words or in our case the spectral bands. Using the β parameters we can extract and plot the topic distributions corresponding to the classes of interest, as shown in figure 4.

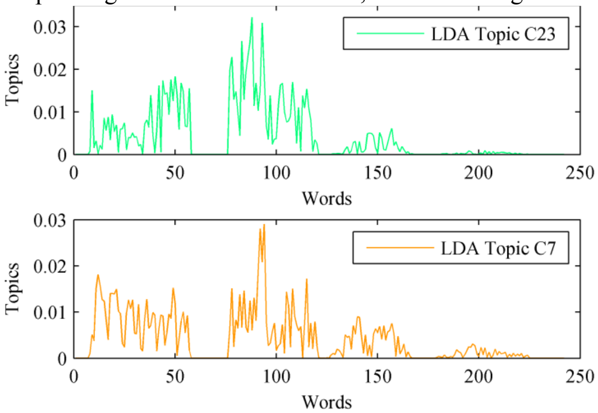


Fig.4 Distribution of topics over words for the LDA classes in fig. 3

Considering that at 30m resolution many pixels cover a range of land use/land cover types, another use scenario of LDA is to identify areas of an image in which a specific land cover class is present, by analyzing the topic proportions within documents. Such particular evolutions can be observed by analyzing LDA class C27, as in figure 5, where yellow indicates a higher presence of this class, while red and brown tones indicate a reduced contribution of this class in the topic mixture. Examples of topic mixtures for the different pixels are given in figure 6.

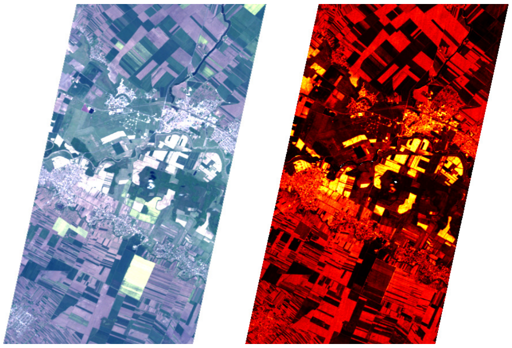
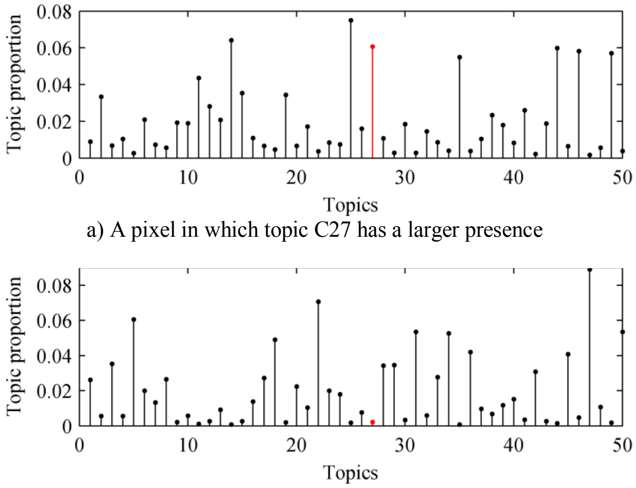


Fig.5 Proportion of class C27 within topic mixtures; yellow tones indicate a higher presence of this topic, dark tones indicate a reduced presence of this topic.



a) A pixel in which topic C27 has a larger presence

b) A pixel in which topic C27 has a reduced presence

Fig.6 Topic mixtures observed at document (pixel) level.

In many situations band combinations are used to highlight interesting features. By examining the distribution of topics over words we can easily select the best band combinations. For the topic #27, the five words that bring the most contribution correspond to spectral bands 71, 72, 43, 41, and 61. Based on this result we have selected the bands 72, 41 and 61 to create a false color image, the result being shown in fig.7

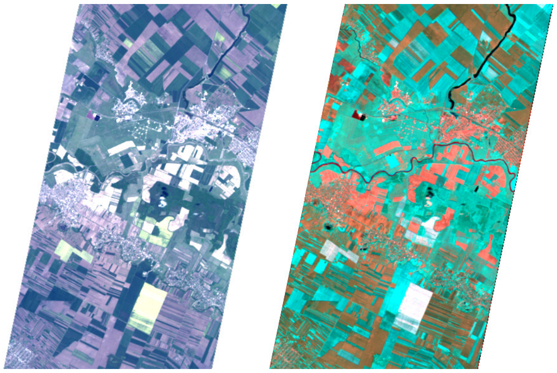


Fig.7 Words having the most importance to a topic can be used to select the best band combinations for generation of false color images

CONCLUSIONS

In this paper we have proposed a novel method for analysis of hyperspectral remotely sensed data by adopting Latent Dirichlet Allocation text modeling tool. The method is applied in the spectral domain, allowing us to preserve the spatial information of the original dataset. Several use scenarios are then formulated based on the analysis of the LDA model.

REFERENCES

[1] P. K. Varshney , M. K. Arora, "Advanced Image Processing Techniques for Remotely Sensed Hyperspectral Data", ISBN 978-3642060014, Springer, 2010

[2] J. Pearlman, S. Carman, C. Segal, P. Jarecke, and P. Barry, "Overview of the Hyperion imaging spectrometer for the NASA EO-1 mission," in Proc. IGARSS, Sydney, Australia, 2001

[3] A.Apan, A Held, S. Phinn, J. Markley, "Detecting sugarcane'orange rust' disease using EO- I I-Iyperion hyperspectral imagery",International Journal of Remote Sensing, 25, 2, pp. 489--498, 2004.

[4] J S.Galvao, A R.Formaggio, D. ATisot, "Discrimination of sugarcane varieties in Southeastern Brazil with EO-l Hyperion data" , Remote Sensing of Environment, 94, pp523-534, 2005.

[5] J Bing Xu and Peng Gong, "Land use/Land cover classification with multispectral and hyperspectral EO-I data", Photogrammetric Engineering & Remote Sensing, 73, 8, pp. 955-965, 2007

[6] Ntouros, K.D.; Gitas, I.Z.; Silleos, G.N.; , "Mapping agricultural crops with EO-1 Hyperion data," Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, 2009. WHISPERS '09. First Workshop on , vol., no., pp.1-4, 26-28 Aug. 2009

[7] Datt, B.; McVicar, T.R.; Van Niel, T.G.; Jupp, D.L.B.; Pearlman, J.S.; , "Preprocessing EO-1 Hyperion hyperspectral data to support the application of agricultural indexes," Geoscience and Remote Sensing, IEEE Transactions on , vol.41, no.6, pp. 1246- 1259, June 2003

[8] Senthilnath, J.; Omkar, S. N.; Mani, V.; Karnwal, N.; P. B., S.; , "Crop Stage Classification of Hyperspectral Data Using Unsupervised Techniques," Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of , vol.PP, no.99, pp.1-7, 2012

[9] Soo Chin Liew; Chew Wai Chang; Kim Hwa Lim; , "Hyperspectral land cover classification of EO-1 Hyperion data by principal component analysis and pixel unmixing," Geoscience and Remote Sensing Symposium, 2002. IGARSS '02. 2002 IEEE International , vol.6, no., pp. 3111- 3113 vol.6, 2002

[10] Smeulders AWM, Worring M, Santini S, Gupta A, Jain R. "Content-Based Image Retrieval at the End of the Early Years". IEEE Trans Pattern Anal Mach Intell 2000;22(12):1349-80

[11] Landauer, T. K., & Dumais, S. T. (1997). "A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge". Psychological Review, 104, 211-240.

[12] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," J. Mach. Learn. Res., vol. 42, no. 1/2, pp. 177–196, Jan. 2001.

[13] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J.Mach. Learn. Res.*, vol. 3, no. 5, pp. 993–1022, Mar. 2003.

[14] T. L. Griffiths and M. Steyvers, "A probabilistic approach to semantic representation," in Proc. 24th Annu. Conf. Cognit. Sci. Soc., 2002, pp. 381–386

[15] Latent Dirichlet Allocation in C, http://www.cs.princeton.edu/~blei/lda-c/

[16] M. Lienou, H. Maitre, and M. Datcu, "Semantic annotation of satellite images using latent dirichlet allocation," Geosci. Remote Sens. Lett., vol. 7, no. 1, pp. 28–32.

[17] Bratasanu, D.; Nedelcu, I.; Datcu, M.; , "Bridging the Semantic Gap for Satellite Image Annotation and Automatic Mapping Applications," Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of , vol.4, no.1, pp.193-204, March 2011

[18] Văduva, C.; Costăchioiu, T.; Pătraşcu, C.; Gavăt, I.; Lăzărescu, V.; Datcu, M.; , "A Latent Analysis of Earth Surface Dynamic Evolution Using Change Map Time Series," Geoscience and Remote Sensing, IEEE Transactions on , vol.PP, no.99, pp.1-14, in press

[19] P. Barry, "EO-1/Hyperion science data user's guide, Level 1_B," TRW Space, Defense & Information Systems, Redondo Beach, CA, Rep. HYP.TO.01.077, 2001.