SEMANTIC ANALYSIS OF EO-1 HYPERION HYPERSPECTRAL DATA

Teodor Costăchioiu[1], Rodica Constantinescu[1] , Vasile Lăzărescu[1] and Mihai Datcu[1,2]

[1]Research Centre for Spatial Information, University Politehnica of Bucharest, Romania
[2]DLR - German Aerospace Center, Oberpfaffenhofen, Germany

ABSTRACT

In this paper we propose a new method for unsupervised data model discovery in hyperspectral data, by adopting the Latent Dirichlet Allocation (LDA) text modeling tool. The proposed method relies on defining a representation of hyperspectral data that allows LDA analysis, by defining a correspondence between hyperspectral profiles and visual words. As such, we can use Latent Dirichlet Allocation as a method for discovery of latent (intrinsic and natural) meaningful grouping of the spectral bands. Based on the model parameters we propose several use scenarios for hyperspectral data analysis, allowing us to identify semantically meaningful structures in the observed scene. Experiments using the proposed method were carried on an EO-1 Hyperion data set imaging an agricultural area in Romania, acquired in June 2009.

Index Terms— Latency analysis, Latent Dirichlet Allocation

1. 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.

Those disadvantages were overcomes 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 Minimum Noise Fraction 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 the third section 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.

2. 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 Probabilistic Latent Semantic Analysis (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. 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 to 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.

3. 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.

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.

To make possible the analysis of hyperspectral data using LDA a suitable bag-of-words representation of data has to be defined. While in language analysis the words are clearly defined, in hyperspectral domain we can define our own "visual words". Under the assumption that the spectral response for a pixel carries sufficient semantic information, in this paper we choose to consider the whole dataset as a corpus, the pixels as documents, and to define words from the spectral information. To achieve this we apply a linear quantization process to each spectral band, with a total number of 410 levels. Words are then defined as $\boldsymbol{w_{ij}} = \boldsymbol{B_i q_j}$ where $\boldsymbol{B_i}$ represents the i-th spectral band and $\boldsymbol{q_j}$ is the quantization level within the spectral band. Considering that we have 196 spectral bands, each quantized by 410 levels, in total we can obtain a dictionary of 80360 possible words. As from each spectral band only one word is extracted, all words will appear with a frequency of one in the associated document.

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. While LDA has the ability to generalize over unseen documents, this is limited only to the corpus from which the training set is extracted. Furthermore, as in this approach LDA learns

directly from the data, which may be contaminated by noise, the LDA model will be a joint model of data and noise, and its application for inference is restricted to only the image from which the model is trained.

As in the LDA model the number of topics (classes) is a user-given parameter, we have trained several LDA models for different number of topics and selected the model that returns the lowest value of perplexity as the model with the optimal number of classes. Once the LDA model is selected, we can perform inference over the whole data set to estimate the topic proportions inside each document.

4. EXPERIMENTS AND RESULTS

Experiments were performed on an EO-1 Hyperion data set acquired on 18 June 2009, imaging an agricultural area in Romania. Several land use/land cover types can be observed in the studied area: land use allocated to agriculture dominates, water bodies and patches of forests being also observed.

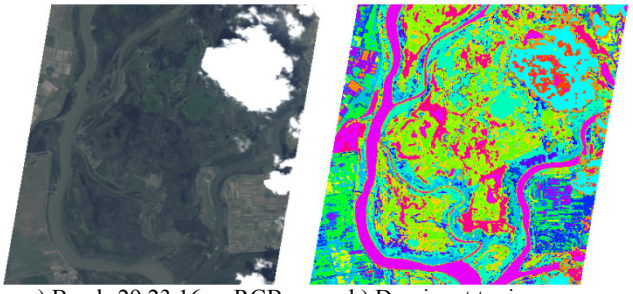
Following the proposed method, we have extracted visual words from the data. To reduce the complexity of the LDA model we have chosen to keep only the words that occur in the dataset. In our experiments by this approach we have obtained a dictionary of 22540 words.

A training set of 10% of the original data was selected, and we have trained several LDA models for different number of topics. As a result of LDA training we have obtained a data model characterized by the γ parameters of the topic over words distributions and by the topic mixture parameters β. This model is then used to perform inference over the whole dataset, obtaining the topic mixtures for each pixel.

Several use scenarios can be further defined based on the LDA model parameters and the estimated topic mixtures.

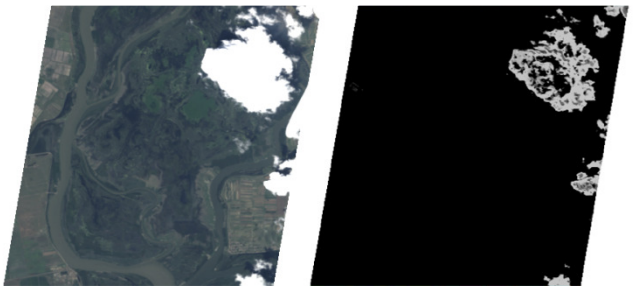
The first possible scenario is to use LDA to label each pixel considering a maximum likelihood of the topic mixtures, the result of labeling for a number of 20 topics being presented in figure 1.

A second scenario of analysis focuses on individual topics, providing support for highlighting semantically meaningful structures.



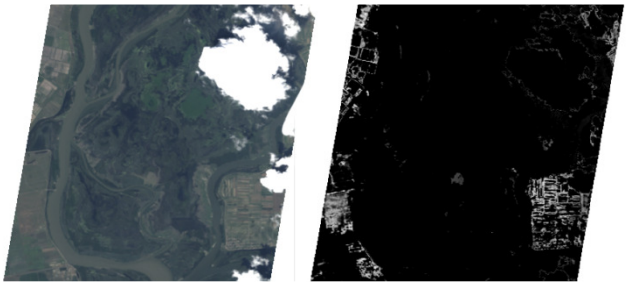
a) Bands 29,23,16 as RGB b) Dominant topics

Fig.1 Topics extracted using the LDA model



a) Bands 29,23,16 as RGB b) Cloud presence

Fig.2 Topics presence for cloud cover

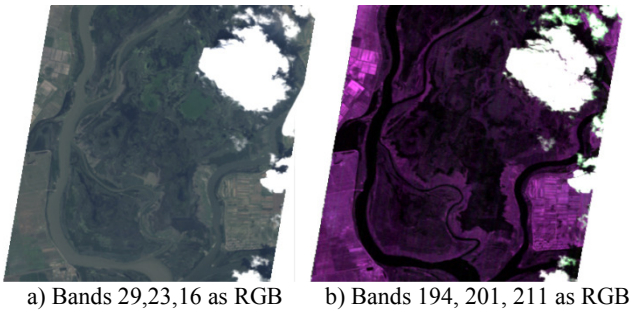


a) Bands 29,23,16 as RGB b) Agriculture

Fig.3 Topics presence for agriculture

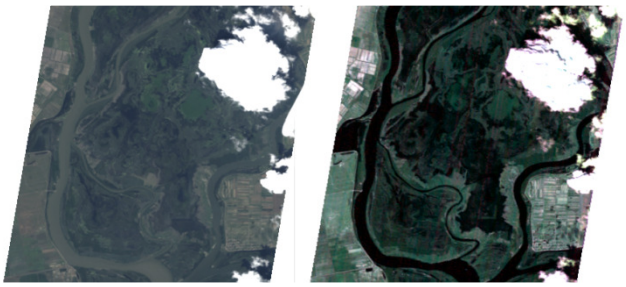
From the topic mixtures extracted by LDA we can extract only the topics of interest, and we can display the proportions of these topics within the pixel area. In figure 2 we show the results of the analysis a topic that can be identified as cloud cover, while in figure 3 we show the topic proportions for a topic corresponding to agriculture.

The last scenario we propose is based on the analysis of the conditional probabilities of the words and topics. This analysis allows us to select the words having the most contribution to topic. For example, considering the semantic topic corresponding to cloud cover in fig.2, the most important words are w_{21754}, w_{21675} and w_{22038}. Knowing these words allows us to identify the spectral bands and the range of the spectral bands corresponding to the words (in our case the spectral bands B_{203}, $B_{201}$and B_{211}. For the topic of agriculture in fig. 3 we found as the most significant the words w_{21379}, w_{19160} and w_{21917}, corresponding to the spectral bands B_{192}, B_{141} and B_{208}.



a) Bands 29,23,16 as RGB b) Bands 194, 201, 211 as RGB

Fig.4 Band combinations for highlighting cloud cover



a) Bands 29,23,16 as RGB b) Bands 194,141, 208 as RGB

Fig.5 Band combinations for highlighting a topic from agriculture

5. CONLCUSIONS

In this paper we have proposed a novel method for analysis of unsupervised 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.

Unlike data classification, the proposed method allows the discovery of latent (intrinsic and natural) meaningful grouping of the spectral bands. Based on such grouping we then formulate use scenarios that aim to identify semantically meaningful structures in the observed scene. Furthermore, the derived topics and the associated vial visual words can provide support for visualization of the hyperspectral data, making evident the meaningful structures in the observed scene.

Furthermore, through analysis of the relations between the topics and the extracted visual words, the topic discovery provides a powerful tool to support the creation of libraries of groups of hyperspectral signatures with semantic meaning.

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

[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.