SiRCub – Brazilian Agricultural Crop Recognition System

Jordi Creus Tomàs[1]
Fabio Augusto Faria[1]
Júlio César Dalla Mora Esquerdo[2]
Alexandre Camargo Coutinho[2]
Claudia Bauzer Medeiros[1]

[1] Universidade Estadual de Campinas – UNICAMP/IC
Caixa Postal 6176 – 13083-852 – Campinas – SP, Brasil
{jordi, ffaria, cmbm}@ic.unicamp.br

[2] Empresa Brasileira de Pesquisa Agropecuária – Embrapa Informática Agropecuária
Caixa Postal 6041 – 13083-886 – Campinas – SP, Brasil
{julio.esquerdo, alex.coutinho}@embrapa.br

Abstract. This paper presents a novel approach to classify agricultural crops using NDVI time series. The novelty lies in i) extracting a set of features from the each and every NDVI curve, and ii) using them to train a crop classification model using a Support Vector Machine (SVM). Specifically, we use the TIMESAT program package to: 1) smooth the time series, 2) decompose them into *agricultural seasons* –a season is the period between sowing and harvesting–, and 3) extract the features for each season. The 11 crop features we extract include the start and end of season, its amplitude, and the curve gradients of the sprouting and senescence periods, among others. Once we have the collection of features, they are fed into an SVM system –we use the LIBSVM library–, together with a collection of annotations about the land use of the corresponding time series. These annotations represent the type of crop for a given location and agricultural season, and they are provided by the specialists of the Embrapa. As a result we obtain a classification model that allows for identifying different crop classes. Our methodology is generic and can be applied to a variety of regions and crop types. We have develop a system called SIRCUB (*Sistema de Reconhecimento de Culturas brasileiro*), which implements such methodology. Thus, we describe in this paper the architecture of the system and the crop model learning methodology.

Keywords: LULC, NDVI, Timesat, SVM, time series, séries temporais

1. Introduction

Crop monitoring in Brazil is a strategic question, since agribusiness represents approximately 25% of the Brazilian Gross Domestic Product (GDP). Timely information about cropping areas is important for the Government and producers, manly to support agricultural production forecasting systems. Such kind of crop identification and mapping is not an easy issue, due to the continental extension of the country, climate variations, different agricultural intensification processes, etc. In this sense, remote sensing and geoprocessing techniques have been extensively used to develop objective methodologies for crop mapping and monitoring on a regional and national basis.

In the last years, the analysis of time series of vegetation index images has been appointed by many research groups as the most appropriated to monitor and identify crops (JöNSSON; EKLUNDH, 2002; JAKUBAUSKAS; LEGATES; KASTENS, 2002; LU; WENG, 2007; WARDLOW; EGBERT; KASTENS, 2007; BROWN et al., 2013). These approaches explore the high spectral and temporal resolutions of orbital sensors like the MODIS[1], in order to evaluate the biomass

[1]The Moderate Resolution Imaging Spectroradiometer (MODIS) is on board the Terra and Aqua satellites.

condition of the crops throughout their phenological cycles. This temporal analysis does not process the images within the time series independently, instead it considers the evolution of each and every single pixel along the sequence of images, emphasizing in its spectral variation. Thereby, it is possible to monitor the dynamics of vegetation in specific temporal scales, including the characterization of phenological cycles, and apply this information to identify specific crop types However, some challenges still remain in this kind of approach, such as the spectral mixture due the low spatial resolution data, the similarity of different crop biomass curves, regional crop management practices.

This paper is organized as follows: Section 2 gives a background describing two tools – TIMESAT and LIBSVM– used in our system. Section 3 provides an overview of the related work. Next, Section 4 describes our classification technique and the architecture of our system. Finally, Section 5 concludes and discusses future work.

2. Background

Our approach for crop recognition relies on two tools: TIMESATand LIBSVM. The first is used to smooth the time series and extract the crop features from the series. Afterward, LIBSVMis used to create a classification model from features produced by TIMESAT.

2.1. TIMESAT

TIMESAT is a program package developed by Eklundh and Jönsson (JöNSSON; EKLUNDH, 2002, 2004) to process and analyze vegetation index time series, such as NDVI series.

In practice, given a time series ts it realizes the following tasks:

1. Seeks the smooth function f^s that best fits ts;
2. Decomposes the smoothened series f^s into *agricultural seasons*[2];
3. Extracts, for each season, a set of crop features (amount of eleven features) called *seasonality parameters*.

TIMESAT provides three different smoothing filters, of which the Savitzky-Golay filter is the most reliable since it is more sensitive to rapid changes and noise (TAN et al., 2011). In this software, the users may parameterize the time series processing with many distinct variables.

The eleven *seasonality parameters* (EKLUNDH; JöNSSON, 2011) extracted from each agricultural season are detailed next:

1. **Start of season** (beg): time in which the left edge of a fitted curve has increased to a user defined level (*e.g.*, 10% of the amplitude) measured from the left minimum level (a in Fig. 1).
2. **End of season** (end): time in which the right edge has decreased to a user defined level measured from the right minimum level (b in Fig. 1).
3. **Length of season** ($length$): period from the start to the end of the season (g).
4. **Base value** ($base$): the average of the left and right minimum values.
5. **Position of middle of season** (mid_x): computed as the mean value of the times for which, respectively, the left edge has increased to the 80% level (c in Fig. 1) and the right edge has decreased to the 80% level (d).
6. **Maximum of fitted data** (max): may occur at a different time compared with 5 (e).
7. **Amplitude** (amp): difference between the maximum value max and the base level $base$ (f).
8. **Left derivative** (l_der): calculated as the ratio of the difference between the left 20% and 80% levels and the corresponding time difference.

[2]An *agricultural season* corresponds to the period spanning a cycle of planting/harvesting of the crop.

9. **Right derivative** (r_der): calculated as the absolute value of the ratio of the difference between the right 20% and 80% levels and the corresponding time difference. It is thus given as a positive quantity.
10. **Large integral** (l_integ): integral of the function describing the season from the season start to the season end (i).
11. **Small integral** (s_integ): integral of the difference between the function describing the season and the base level from season start to season end (h).

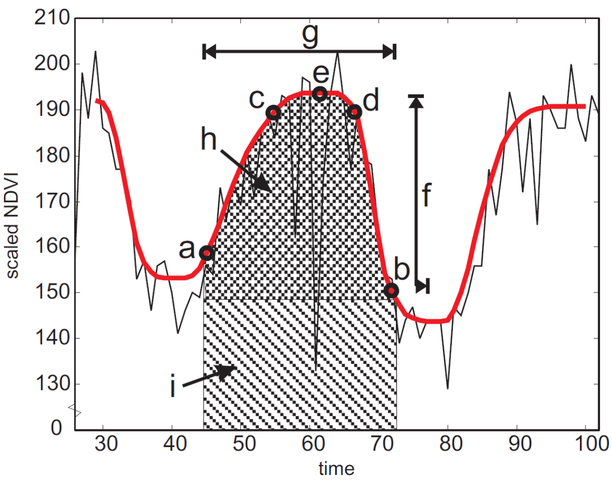

Figure 1: NDVI time series, fitted model function using a double logistic function (thick line) and seasonality parameters (a-i), extracted from (EKLUNDH; JöNSSON, 2011).

2.2. LIBSVM – Support Vector Machines

Support Vector Machine (SVM), proposed by Vapnik (BOSER; GUYON; VAPNIK, 1992), is one of the most popular machine learning techniques for classification in the literature. Initially, SVM was defined as a two-class classifier (a binary problem), but many subsequent works extended SVM to the multi-class classification problem (CRAMMER; SINGER, 2002; TSOCHANTARIDIS et al., 2004; JOACHIMS, 1999).

Given a linearly separable problem, the SVM technique tries to construct an optimal hyperplane or set of hyperplanes that maximizes the margin between two classes (*e.g.*, classes $+1$ and -1). This margin can be seen as the minimum distance of one point of one class to the other class. The instances placed on borders between the classes are called *support vectors*. When it is impossible to find a linear separator for the classes, the instances are mapped to other spaces through a non-linear mapping (kernel trick) (CRISTIANINI; SHAWE-TAYLOR, 2000), whose hyperplanes might be able to separate both constructed classes.

The most popular implementation of the SVM technique available in the literature is the LIBSVM library (CHANG; LIN, 2011). However, there are other implementations, such as WEKA (HALL et al., 2009), Shogun (SONNENBURG et al., 2010), or R (IHAKA; GENTLEMAN, 1993).

3. Related Work

Land Use/Land Cover (LULC) mapping has been widely studied during last 15 years. We can find a number of papers proposing different techniques, and with different purposes. We have classified the related work according to two criteria: the technique used, and the nature of their target classes. We have found 4 different types of techniques: i) time-frequency analysis using Fourier and wavelet transforms (SAKAMOTO et al., 2005; GALFORD et al., 2008; MARTÍNEZ; GILABERT, 2009), ii) harmonic analysis of time series (ZHANG et al., 2008; JAKUBAUSKAS; LEGATES; KASTENS, 2002; MINGWEI et al., 2008), iii) decision tree classifiers (WARDLOW; EGBERT, 2008; ZHANG et al., 2008), and iv) machine learning techniques –like SVM– (XUE; DU; FENG, 2014; ALCANTARA et al., 2012; PUERTAS; BRENNING; MEZA, 2013; BAO; CHI; BENEDIKTSSON, 2013).

On the other hand, if we consider the granularity of their target classes, we can classify the papers into two types: i) those that try to identify general land cover classes, *e.g.*, forest, grassland, shrubland, cropland, urban, paddy field, water (ZHANG et al., 2008; PUERTAS; BRENNING; MEZA, 2013; MARTÍNEZ; GILABERT, 2009; ALCANTARA et al., 2012; XUE; DU; FENG, 2014; BAO; CHI; BENEDIKTSSON, 2013); and ii) those that deal with specific crop types, *e.g.*, alfalfa, corn, sorghum, cotton, soybean, bean (WARDLOW; EGBERT; KASTENS, 2007; WARDLOW; EGBERT, 2008; JAKUBAUSKAS; LEGATES; KASTENS, 2002; MINGWEI et al., 2008). We summarize in Table 1 this twofold classification of the related work. For each paper, we provide the authors, year, and the location of the data they experimented with.

		Technique			
		Fourier/wavelet transforms	Harmonic analysis	Decision tree	Machine learning
Target classes	Land cover	Martínez e Gilabert (2009), Spain	Zhang et al. (2008), China	Zhang et al. (2008), China	Puertas, Brenning e Meza (2013), Chile; Alcantara et al. (2012), Europe; Xue, Du e Feng (2014), China; Bao, Chi e Benediktsson (2013), Botswana
	Crop types	Sakamoto et al. (2005); Japan; Galford et al. (2008), Brazil	Jakubauskas, Legates e Kastens (2002), Kansas; Mingwei et al. (2008), China	Wardlow, Egbert e Kastens (2007), Wardlow e Egbert (2008), Kansas	

Table 1: Related work classification.

4. Our Approach: the SIRCUB System

SIRCUB stands for *Sistema de Reconhecimento de Culturas brasileiro* (*i.e.*, Brazilian Agricultural Crop Recognition System), and it is a system we have developed at the Institute of Computing of the University of Campinas. SIRCUB allows for classifying different types of agricultural crops using NDVI time series, and it relies on SVM (Support Vector Machine). The methodology SIRCUB implements is generic and can be applied to a variety of regions and crops. We have developed a prototype of SIRCUB that implements almost all functionalities of the system.

4.1. System Architecture

Figure 2 depicts SIRCUB's architecture and data flow. The system can be divided into two phases: feature extraction and crop model learning.

The *feature extraction* phase is composed of the following modules: time series preprocessing, smooth function finder and feature extractor. In this phase, we have as input a collection of NDVI time series and a settings file. The *Time series preprocessing* module is in charge of removing signal noise –mainly cloud noise– from the NDVI time series. Next, the filtered time series and the settings file are fed into the TIMESAT modules. TIMESAT subsystem is responsible for smoothing the NDVI time series, decomposing them into agricultural seasons, and extracting the crop features from each season. The user can configure the time series processing through TIMESAT's *settings file*. Some of the parameters the user can define include: the smoothing filter, the seasonal parameter, the number of envelope iterations, or the window size of the Savitzky-Golay filter. At the end of this phase we obtain a collection of crop descriptors.

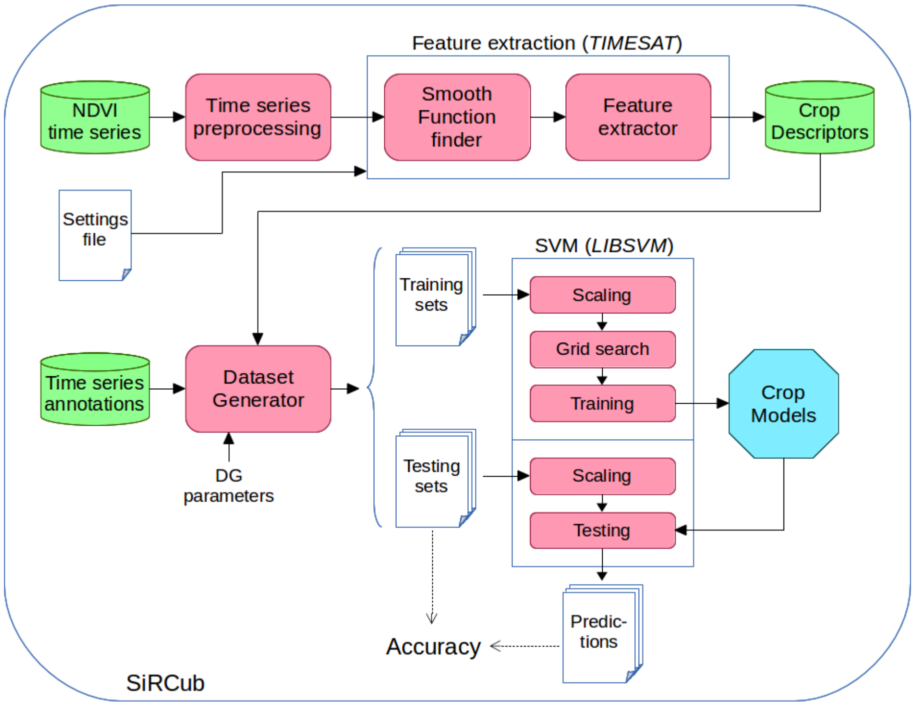

Figure 2: SIRCUB architecture and dataflow.

In the second phase, *crop model learning*, we can find the dataset generator and different modules of the LIBSVM package: scaling, grid search, training, and testing. The *Dataset Generator* is an important module, it is in charge of creating the instance datasets that feed the learning machine. Namely, it creates pairs of training and testing sets. It has as input the crop descriptors produced by the previous phase, and the time series annotations provided by the specialists. The user can define several parameters concerning the dataset generation. For instance, he/she can define the method used to generate the datasets (random resampling or n-fold cross-validation), or whether he/she wants balanced or unbalanced datasets. Once we have the pairs of training and testing sets, we execute the different SVM routines in order to create the agricultural crop models, and finally we test their accuracy.

4.2. Methodology

Our methodology can be described as follows:

1. Given a collection of NDVI time series TS, for each time series $ts \in TS$ we look for the smooth function f^s that best fits ts.
2. We decompose the smoothened time series f^s into agricultural seasons S, and for each season $s \in S$ we extract a set of 11 features that define the crop descriptor d for that season. These features include the beginning and end dates of the season, the curve gradients of the sprouting and senescence periods, the total area within the curve function, etc. (see the complete list in Section 2.1).

3. We learn the patterns of the crop curves using LIBSVM. The crop descriptors $D = \{d\}$ we found in previous step, and the annotations of land use A are then fed into the SVM. These annotations are the ground truth data provided by specialists from the Embrapa, who did an in the field gathering.
4. Finally, we obtain an agricultural crop model m that allows for classifying different types of crops.

5. Conclusion

This paper presents a new approach to deal with agricultural crop classification, as well as the architecture of SIRCUB, a system implementing the aforesaid approach. Our method is founded on: i) the use of a set of features to describe the crops on the NDVI time series, and ii) using such features to train a classification model for agricultural crops. The crop features are extracted using TIMESAT software, and they include the start and end of season, the sprouting and senescence gradients, the amplitude, etc. Furthermore, the crop models are created through use of a SVM technique, in particular we have used LIBSVM library. Our methodology is generic and can be applied to different regions and different types of crops.

At present, we are designing the scenarios to experimentally evaluate the effectiveness of our approach. Embrapa has provided a collection of more than 3000 annotated time series, which correspond to different farms located in the State of Mato Grosso, and span a period of 4 years (from 2009 to 2013). Each time series has two annotations per year, which were directly requested to the farmers. Thus, we have 8 annotations per time series, totalizing about 25,000 annotations.

In the future, we plan to extend our methodology by combining other methods. Some of the methods we will consider include prediction techniques like Conditional Random Fields (CRF) or hidden Markov models. Yet another enhancement might be complementary to the crop features, for example, use of data history of the land use during the learning process.

6. Acknowledgements

Work partially financed by FAPESP 2012/25169-5 fellowship, FAPESP/Cepid in Computational Engineering and Sciences (2013/08293-7), the Microsoft Research FAPESP Virtual Institute (NavScales project), CNPq (MuZOO Project), FAPESP-PRONEX (eScience project), INCT in Web Science, and individual grants from Capes and CNPq. We acknowledge as well Professor Omar Boucelma for his pertinent comments.

References

ALCANTARA, C. et al. Mapping abandoned agriculture with multi-temporal {MODIS} satellite data. *Remote Sensing of Environment*, v. 124, n. 0, p. 334 – 347, 2012.

BAO, J.; CHI, M.; BENEDIKTSSON, J. Spectral derivative features for classification of hyperspectral remote sensing images: Experimental evaluation. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, v. 6, n. 2, p. 594–601, April 2013.

BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. A training algorithm for optimal margin classifiers. In: *Workshop on Computational Learning Theory*. [S.l.: s.n.], 1992. (COLT '92), p. 144–152. ISBN 0-89791-497-X.

BROWN, J. C. et al. Classifying multiyear agricultural land use data from mato grosso using time-series {MODIS} vegetation index data. *Remote Sensing of Environment*, v. 130, n. 0, p. 39 – 50, 2013.

CHANG, C.; LIN, C. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, ACM, New York, NY, USA, v. 2, n. 3, p. 27:1–27:27, maio 2011.

CRAMMER, K.; SINGER, Y. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal Machine Learning Research*, v. 2, p. 265–292, 2002.

CRISTIANINI, N.; SHAWE-TAYLOR, J. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. [S.l.]: Cambridge University Press, 2000.

EKLUNDH, L.; JöNSSON, P. *Timesat 3.1 Software Manual*. Sweden, 2011.

GALFORD, G. L. et al. Wavelet analysis of {MODIS} time series to detect expansion and intensification of row-crop agriculture in brazil. *Remote Sensing of Environment*, v. 112, n. 2, p. 576 – 587, 2008. Soil Moisture Experiments 2004 (SMEX04) Special Issue.

HALL, M. et al. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, ACM, New York, NY, USA, v. 11, n. 1, p. 10–18, 2009. ISSN 1931-0145.

IHAKA, R.; GENTLEMAN, R. 1993. Homepage: `http://cran.r-project.org/`.

JAKUBAUSKAS, M. E.; LEGATES, D. R.; KASTENS, J. H. Crop identification using harmonic analysis of time-series AVHRR NDVI data. *Computers and Electronics in Agriculture*, v. 37, n. 1–3, p. 127 – 139, 2002.

JOACHIMS, T. Advances in kernel methods. In: SCHöLKOPF, B.; BURGES, C. J. C.; SMOLA, A. J. (Ed.). [S.l.: s.n.], 1999. cap. Making Large-scale Support Vector Machine Learning Practical, p. 169–184.

JöNSSON, P.; EKLUNDH, L. Seasonality extraction by function fitting to time-series of satellite sensor data. *Geoscience and Remote Sensing, IEEE Transactions on*, v. 40, n. 8, p. 1824–1832, Aug 2002.

JöNSSON, P.; EKLUNDH, L. TIMESAT—a program for analyzing time-series of satellite sensor data. *Computers & Geosciences*, v. 30, n. 8, p. 833 – 845, 2004.

LU, D.; WENG, Q. A survey of image classification methods and techniques for improving classification performance. *Int. J. Remote Sens.*, Taylor & Francis, Inc., Bristol, PA, USA, v. 28, n. 5, p. 823–870, jan. 2007.

MARTíNEZ, B.; GILABERT, M. A. Vegetation dynamics from {NDVI} time series analysis using the wavelet transform. *Remote Sensing of Environment*, v. 113, n. 9, p. 1823 – 1842, 2009.

MINGWEI, Z. et al. Crop discrimination in northern china with double cropping systems using fourier analysis of time-series {MODIS} data. *International Journal of Applied Earth Observation and Geoinformation*, v. 10, n. 4, p. 476 – 485, 2008.

PUERTAS, O. L.; BRENNING, A.; MEZA, F. J. Balancing misclassification errors of land cover classification maps using support vector machines and landsat imagery in the maipo river basin (central chile, 1975–2010). *Remote Sensing of Environment*, v. 137, n. 0, p. 112 – 123, 2013.

SAKAMOTO, T. et al. A crop phenology detection method using time-series {MODIS} data. *Remote Sensing of Environment*, v. 96, n. 3–4, p. 366 – 374, 2005.

SONNENBURG, S. et al. The shogun machine learning toolbox. *J. Mach. Learn. Res.*, v. 11, p. 1799–1802, ago. 2010.

TAN, B. et al. An enhanced timesat algorithm for estimating vegetation phenology metrics from modis data. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, v. 4, n. 2, p. 361–371, June 2011.

TSOCHANTARIDIS, I. et al. Support vector machine learning for interdependent and structured output spaces. In: *Proceedings of the Twenty-first International Conference on Machine Learning*. [S.l.: s.n.], 2004. (ICML '04), p. 104–.

WARDLOW, B. D.; EGBERT, S. L. Large-area crop mapping using time-series MODIS 250 m NDVI data: An assessment for the U.S. Central Great Plains. *Remote Sensing of Environment*, v. 112, n. 3, p. 1096 – 1116, 2008.

WARDLOW, B. D.; EGBERT, S. L.; KASTENS, J. H. Analysis of time-series {MODIS} 250 m vegetation index data for crop classification in the u.s. central great plains. *Remote Sensing of Environment*, v. 108, n. 3, p. 290 – 310, 2007.

XUE, Z.; DU, P.; FENG, L. Phenology-driven land cover classification and trend analysis based on long-term remote sensing image series. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, v. 7, n. 4, p. 1142–1156, April 2014.

ZHANG, X. et al. Land cover classification of the North China Plain using MODIS_EVI time series. *ISPRS Journal of Photogrammetry and Remote Sensing*, v. 63, n. 4, p. 476 – 484, 2008.