Data Mining and Pattern Recognition in Agriculture

Article *in* KI - Ku_nstliche Intelligenz · October 2013

DOI: 10.1007/s13218-013-0273-0

CITATIONS
39

READS
1,811

2 authors:

Christian Bauckhage
University of Bonn
453 PUBLICATIONS **7,647** CITATIONS

SEE PROFILE

Kristian Kersting
University of Bonn
394 PUBLICATIONS **6,676** CITATIONS

SEE PROFILE

Noname manuscript No.
(will be inserted by the editor)

Data Mining and Pattern Recognition in Agriculture

Christian Bauckhage · **Kristian Kersting**

Received: date / Accepted: date

Abstract Modern communication, sensing, and actuator technologies as well as methods from signal processing, pattern recognition, and data mining are increasingly applied in agriculture. Developments such as increased mobility, wireless networks, new environmental sensors, robots, and the computational cloud put the vision of a sustainable agriculture for anybody, anytime, and anywhere within reach. Yet, precision farming is a fundamentally new domain for computational intelligence and constitutes a truly interdisciplinary venture. Accordingly, researchers and experts of complementary skills have to cooperate in order to develop models and tools for data intensive discovery that allow for operation through users that are not necessarily trained computer scientists. We present approaches and applications that address these challenges and underline the potential of data mining and pattern recognition in agriculture.

1 Introduction

Facing a rapidly growing world population, answers to the daunting question of "How to feed a hungry world?" are in dire need. Challenges include climate change, water scarcity, labor shortage due to aging populations, as well as concerns as to animal welfare, food safety, and environmental impact. Addressing these issues, agriculture –arguably the oldest economic endeavor of humankind– is receiving a technological makeover and information technology makes its appearance in the fields.

Agricultural information is gathered and distributed by means of smartphones, portable computers, GPS devices, RFID tags, and other environmental sensors. Farming companies are working on automation technologies such as GPS steering to operate balers, combines, or harvesters [5]. Aiming at increased food safety, RFID technologies are used to track animals in livestock; for example, since 2010, European sheep farmers are required to tag their flocks and the European Commission has suggested to extend this to cattle. RFID technologies also provide new possibilities for harvest asset management. For instance, by adding RFID tags, bales can be associated with measured properties such as weight and moisture level [5].

Mobile communication networks and technologies which are now commonly deployed in many areas around the world have become a backbone of pervasive computing in agriculture. Researchers and practitioners apply them to gather and disseminate information as well as to market products or to do business [10,52]. As Farmers need to obtain and process financial, climatic, technical and regulatory information to manage their businesses, public and private institutions cater to their needs and provide corresponding data. For example, the U.S. Department of Agriculture, supplies information as to prices, market conditions, or newest production practices. Internet communities such as e-Agriculture allow users to exchange information, ideas, or procedures related to communication technologies in sustainable agriculture and rural development [5].

Agriculture is thus rapidly becoming a knowledge intensive industry. So far, however, much of the research and development in this regard has focused on sensing and networking rather than on computation. In this article, we survey recent work on computational intelligence in agriculture and precision farming and present two examples of our own work in this area. We point out specific research challenges and opportunities and thus hope to increase awareness of this new and exciting application domain.

C. Bauckhage[1,3] · K. Kersting[2,3]
[1]B-IT, University of Bonn, Germany
[2]IGG, University of Bonn, Germany
[3]Fraunhofer IAIS, Sankt Augustin, Germany
E-mail: firstname.lastname@iais.fraunhofer.de

2 Computational Intelligence in Agriculture

Looking at the scientific literature on precision farming, it appears that, most efforts so far were focused on the development and deployment of sensor technologies rather than on methods for data analysis tailored to agricultural measurements. In other words, up to now, contributions to computational intelligence in agriculture mainly applied off-the-shelf techniques available in software packages or libraries but did not develop specific frameworks or algorithms. Yet, efforts in this direction are noticeably increasing and in this section we survey recent work on data mining and pattern recognition in agriculture.

Computational intelligence in agriculture involves different areas of computer- and information science. Here, we focus on key areas such as knowledge and information management, geo-information systems, and signal processing.

Vernon et al.[51] highlight the importance of information systems for sustainable agriculture. While early work in this direction was focused on the design of (relational) databases, more recent approaches consider semantic web technologies for instance for pest control [31], farm management [49], or the integration of molecular and phenotypic information for breeding [7]. Others consider recommender systems and collaborative filtering to retrieve personalized agricultural information from the web [24] or the use of web mining, for instance, in localized climate prediction [11].

Geo-information processing plays a particular role in computational agriculture and precision farming. Research in this area considers mobile access to geographically aggregated crop information [25], region specific yield prediction [46], or environmental impact analysis [17]. It is clear that, in addition to information infrastructures, applications like these require advanced remote sensing or modern sensor networks. Distributed networks of temperature and moisture sensors are deployed in fields, orchards, and grazing land to monitor growth conditions or the state of pasture [10,52]. Space- or airborne solutions make use of technologies such as Thermal Emission and Reflection Radiometers or Advanced Synthetic Aperture Radar to track land degradation [8] or to measure and predict levels of soil moisture [29]. Other agricultural applications include plant growth monitoring [28] and automated map building [48].

A particularly interesting sensing modality consists in airborne hyper-spectral imaging which records spectra of several hundred wavelengths per pixel. With respect to plant monitoring this allows, for instance, for assessing changes of pigment compositions due to metabolic processes. This in turn allows for remotely measuring phenotypic reactions of plants due to biotic or abiotic stress [32].

Recently, hyper-spectral imaging is being increasingly used for near range plant monitoring in agricultural research. It enables basic research, for example regarding the molecular mechanisms of photosynthesis [39,40], but is also used in plant phenotyping, for instance as an approach towards understanding phenotypic expressions of drought stress [3, 23,44] (see also section 3 of this article).

Classical image analysis and computer vision techniques are being used in agriculture, too. Examples include automated inspection and sorting in agricultural production facilities [26,42], the detection of the activity of pests in greenhouses [6], or the recognition of plant diseases [47] (see also section 4 of this article).

Finally, artificial intelligence techniques are increasingly applied to address questions of *computational sustainability* [16]. Work in this area considers algorithmic approaches towards maximizing the utility of land [18], enabling sustainable water resource management [36], or learning of timber harvesting policies [13].

A common theme of the works in this short survey is that they require algorithms and architectures that can cope with massive amounts of data. Owing to the increased use of modern sensors, corresponding solutions have to cope with exploding amounts data recorded in dynamic and uncertain environments where there typically are many interacting components [16]. However, it appears that most work in this area so far did not involve specifically trained data scientists and that, from the point of view of computational intelligence, more efficient and accurate methods seem available. Yet, computer scientists entering the field must be aware that methods they bring have to benefit researchers and practitioners in agriculture. Practitioners "out in the fields" are in need of methods and tools that yield results they can work with, ideally run on mobile devices, i.e. under constrained computational resources, and in real time in order to help them in their daily work.

From the perspective of farming professionals, problems are natural and real phenomena that may be addressed using scientific methods and advanced computing. To them purely theoretic concepts or mathematical abstractions are of little use. The world's food producers are highly technology-oriented people but with a purpose. Even if they may not be adequately trained in information technology, they actually do not need to be. They know their business and if a new technology does not fit into their work flows they will either ignore it or wait until it meets their needs.

In the next two sections, we present examples from our own work on data mining and pattern recognition in agricultural scenarios. We shall use these examples to underline the above challenges and to illustrate practical solutions for large-scale data and real time processing. We expect that, with the availability of more computational power combined with sensing and networking technologies, new forms of farming may emerge. In this sense, the following examples are meant to point out the potential of computational intelligence in modern precision farming.

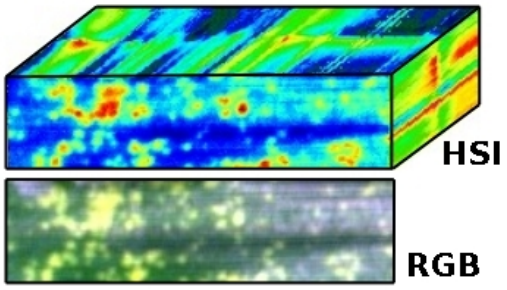


Fig. 1 While conventional RGB images record only three color values (red, green, and blue) per pixel, each pixel of a hyper-spectral image records how a whole spectrum of visible or invisible light waves is reflected from a scene (for technical details, see section 3.3).

3 Specific Example I: Drought Stress Recognition From Hyper-Spectral Image Analysis

In this section we present results from an ongoing project on recognizing and predicting levels of drought stress in plants based on the analysis of hyper-spectral images[1].

There are estimates that drought in conjunction with other abiotic stresses causes a depreciation of crop yields of up to 70% [9,37]. Because of global warming this trend is expected to increase, so that an improved understanding of how plants adapt to drought is called for to be able to breed more resistant varieties. Yet, mechanisms of stress resistance are characterized by complex interactions between the genotype and the environment which lead to different phenotypic expressions [35]. Progress has been made towards the genetic basis of drought related traits [27,30] and modern data analysis has lead to molecular insights into drought tolerance [1,20,38]. However, as genetic and biochemical research are time consuming and only moderately successful in predicting the performance of new lines in the field, there are increased efforts on phenomic approaches.

Hyper-spectral imaging provides an auspicious approach to plant phenotyping [39,40]. In contrast to conventional cameras, which record only 3 wavelengths per pixel, hyper-spectral cameras record a spectrum of several hundred wavelengths ranging from approximately 300nm to 2500nm (see Fig. 1). These spectra contain information as to changes of the pigment composition of leafs which are the result of metabolic processes involved in plant responses to stress. Supervised classification of hyper-spectral signatures can thus be used to predict biotic stress before symptoms become visible to the human eye [43,45].

However, scale poses a significant challenge in hyper-spectral image analysis, since the amount of phenotyping data easily grows into TeraBytes if several plants are monitored over time. For instance, each individual hyperspectral recording considered below consists of a total of about 2 (resp. 5.8) Billion matrix entries. Manually labeling such data as well as running established supervised classification algorithms therefore quickly becomes infeasible.

Addressing these challenges, we developed a novel prediction approach, called Dirichlet-aggregation regression (DAR). It does not require labeled data or statistical learning, does not make any assumption on the generating distribution of observed signatures, and employs a recent linear time, data-driven matrix factorization approach to represent hyper-spectral signatures by means of convex combinations of only few extreme data samples. Practical results show that DAR can predict the level of drought stress of plants well and before it becomes visible to the human eye.

3.1 Dirichlet Aggregation of Phenotypes

First, we briefly recall fundamentals of matrix factorization and how they lead to parametric probability distributions over phenotypes and to a formal notion of drought level.

Scientists working on plant phenotyping regularly need to find meaningful patterns in massive, high dimensional and temporally diverse observations. In our project, hyperspectral data of resolution $640 \times 640 \times 69$ were taken of 10 (resp. 12) plants l at 7 (resp. 20) days t. Each record can thus be viewed as a data matrix $\mathbf{X}^{t,l} \in \mathbb{R}^{m \times n}$ with $m = 640 \times 640$ and $n = 69$. Horizontally stacking the data matrices recorded in all experiment then results in a single matrix $\mathbf{X}$ with about 2 (resp. 5.8) Billion entries. Matrix factorization is commonly used to analyze such data. That is, the matrix is approximated as

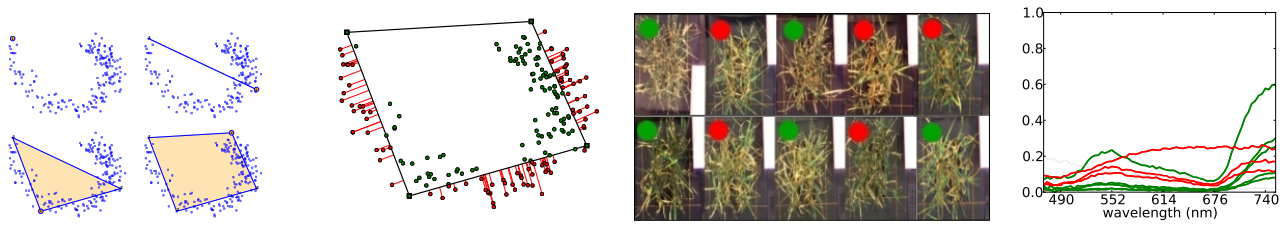
$$\mathbf{X} \approx \mathbf{WH} \tag{1}$$

where the matrix of basis elements $\mathbf{W} \in \mathbb{R}^{m \times k}$, the coefficient matrix $\mathbf{H} \in \mathbb{R}^{k \times n}$, and $k \ll \min\{m,n\}$. This allows for mapping high dimensional data $\mathbf{X}$ to a lower dimensional representation $\mathbf{H}$ and can thus mitigate effects due to noise, uncover latent relations, or facilitate further processing and ultimately help finding patterns in data.

Practitioners in biology usually resort to the Singular Value Decomposition or Non-negative Matrix Factorization. For large matrices both methods are time-consuming and also require deep mathematical knowledge in order to make sense of the results. Both approaches thus hardly characterize intricate data distributions in a unified parametric and interpretable form.

An alternative are interpretable factorization methods where the basis vectors in $\mathbf{W}$ are k columns selected from $\mathbf{X}$ so that they maximize the volume $Vol(\mathbf{W}^{m \times k}) = |\det \mathbf{W}|$ and the matrix of coefficient vectors $\mathbf{H}$ is restricted to convexity, i.e., $h_{ij} \geq 0$ and $\sum_i h_{ij} = 1$. Consequently, the basis

[1] Note that, for the time being, our methodology is tailored towards basic agricultural research carried out in the lab; yet, as costs of hyperspectral sensors are decreasing, corresponding technologies may soon be applicable in outdoor scenarios.



(a) SiVM (b) data reconstruction (c) RGB images of plants (d) extreme spectra

Fig. 2 Fast plant phenotyping using SiVM. From left to right: (a) didactic example illustrating how SiVM determines four extreme points from a set of 2D data; (b) didactic example of how any sample point can be expressed as a convex combination of selected extremes; while points inside of the convex hull of selected basis elements can be reconstructed exactly, points on the outside are approximated by their projection onto its closest facet; (c) actual examples of RGB images of plants on the fourth measurement day; these images were recorded from a top view camera in the controlled environment of a laboratory; corresponding hyper-spectral images were recorded from the same point of view and under the same conditions but are not shown here as they are difficult to visualize; (d) actual examples of different extreme high-dimensional spectra determined within the hyper-spectral recordings; each of these spectra corresponds to a hyper-spectral pixel and shows the fraction of light reflected at different wavelength; the automatically determined extreme spectra belong to images of "dry" (red) and "healthy" (green) leafs; it is noticeable that dry and healthy plants are not necessarily distinguishable from looking at the RGB images in (c).

vectors have a biological meaning since they correspond to observed signatures and can thus be easily be interpreted by an expert. However, the maximum-volume criterion is provably NP-hard [12].

An approximation, called Simplex Volume Maximization (SiVM), was recently introduced by Thurau et al. [50] and empirically proven successful. The method iteratively selects basis vectors among the observed data such that they are as far apart as possible. This is achieved by maximizing the value of the Cayley-Menger determinant computed from pairwise distances between data points selected so far and the next point under consideration. Figure 2(a) and (b) illustrate the SiVM procedure for comuting matrices $\mathbf{W}$ and $\mathbf{H}$ by means of a simple, two dimensional example; Fig. 2(d) shows extreme, high-dimensional data points, i.e. hyper-spectral signatures, determined from a collection of hyper-spectral images of plants; for details, we refer to [22].

From a geometric point of view, the columns $\mathbf{h}_1,\ldots,\mathbf{h}_n$ of $\mathbf{H}$ determined after running the SiVM procedure are data points that reside in a simplex spanned by the extreme elements in $\mathbf{W}$. Accordingly, there are natural parametric distributions to characterize the density of the $\mathbf{h}_i$. The best known one is the Dirichlet

$$\mathscr{D}(\mathbf{h}_i|\alpha) = B(\alpha)\prod_{j=1}^{c} h_{ij}^{\alpha_j-1} \tag{2}$$

where $\alpha = (\alpha_1,\alpha_2,\ldots,\alpha_c)$. The normalization constant $B(\alpha) = \Gamma(S(\alpha))/\prod_{j=1}^{c}\Gamma(\alpha_j)$ where $\Gamma(\cdot)$ is the gamma function and $S(\alpha) = \sum_{j=1}^{c}\alpha_j$. This distributional view on hyper-spectral data provides an intuitive measure of drought stress: the expected probability of observing a healthy spot, which we call the "drought stress level" of a plant.

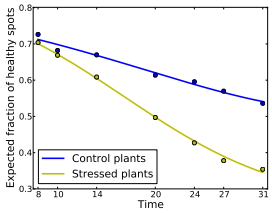
Given α, we note that the marginal distribution of the j-th reconstruction dimension follows a Beta distribution $\mathscr{D}(\alpha_j, S(\alpha)-\alpha_j)$ and the expected value of the j-th reconstruction dimension is $\mu_j = \alpha_j/S(\alpha)$. Thus, each α_j controls "aggregation" of mass of reconstructions near the corresponding column c_j which explains the term *Dirichlet aggregation*. Now assume that each dimension was labeled either "background", "healthy", or "dry". Averaging the expected values of "healthy" or "dry" dimensions and treating them as parameters of a Beta distribution yields the drought stress level of a plant.

Given the results in [22], a signature is is said to show "background" if the corresponding hyper-spectral pixel does not show a leaf. This is easy to verify since extracted prototypic or extreme signatures correspond to actually observed hyper-spectral pixels. If a signature is not background, it is "healthy" if it has a low reflectance (< 0.1) for 470nm – 540nm (chlorophyll a+b). In other words, the hyper-spectral pixel shows a green part or spot of a leaf. If this is not the case and the ratio of maximal reflectances observed at the wavelengths 470nm – 540nm (i.e., they are or turn brown) and at the wavelengths 700nm and above is > 0.5 (they start to overheat). In other words, the hyper-spectral pixel shows a dry part or spot of a leaf (see Fig. 2(d)).

3.2 Bayesian Dirichlet-Aggregation Regression

In order to track and to predict drought levels over time, we apply an approach proposed in [22]. We first select extreme columns from the overall data matrix $\mathbf{X}$. This captures global dependencies as we represent the complete data by means of convex combinations extreme data points selected across all time steps. Then, on the simplex spanned by the extreme points, we estimate Dirichlet distributions specified by $\alpha^{t,l}$ over all reconstructions per day t and plant l. This captures local dependencies. Finally, we compute the drought levels using the $\alpha^{t,l}$.

Using an inference mechanism introduced in [23], we extend this detection approach towards prediction. In sum-



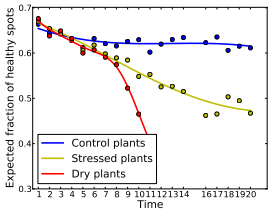
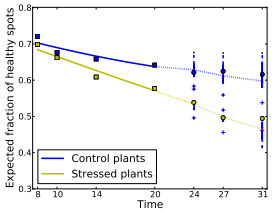


Fig. 3 Dirichlet-aggregation regression of drought levels over several days in 2010 (left) and in 2011 (right) using all hyper-spectral images available. Colors indicate controlled/stressed plants. While the x-axis indicates measurement days, the y-axis indicates the fraction of pixels in the analyzed hyper-spectral images that show healthy parts or spots of a plant predicted from our DAR model. Note that experiments in agricultural research cannot seamlessly be repeated at any time but have to adhere to seasonal growth cycles of plants. Accordingly, data not recorded in an experiment may not be available until a year later. In this example, in the experiments in 2010, plants were watered or stressed but not deliberately dried out. In the experiments in 2011, a third set of data was recorded from dry plants (cf. the experimental procedure in section 3.3).



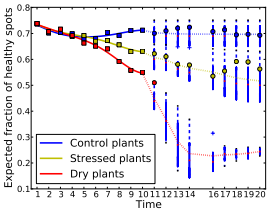


Fig. 4 Bayesian drought level predictions (over time indicated in days) for 2010 (left) and 2011 (right). While the x-axis indicates measurement days, the y-axis indicates the fraction of pixels in the analyzed hyper-spectral images that show healthy parts or spots of a plant. In both experiments, the drought levels of the second half of measurement days were predicted based on a DAR model (including the extraction of extreme spectra) obtained from the data gathered in the first half of measurement days. Colors indicate controlled/stressed plants. Again, for the measurements carried out in 2010, a control group of dry plants was not available.

mary, Dirichlet-aggregation regression (DAR) iterates the following steps until convergence:

1. Optimize the logarithm of the complete likelihood w.r.t. the hidden Dirichlet aggregations η_c^t for each plant.
2. Optimize the log-likelihood of all plants w.r.t. the hyper-parameters ϑ of a common Gaussian process prior.

Figure 3 shows drought levels estimated by DAR averaged over groups of plants in two data sets considered in our project. As one can see, DAR nicely smoothes SiVM's "hard" drought level (shown as dots).

Having a Bayesian regression model at hand, we iteratively obtain predictions by making repeated one-step ahead predictions, up to the desired horizon. For the one-step ahead prediction at time t^*, we apply standard Gaussian process regression [41]. That is, the expectation of the hidden Dirichlet aggregation η_c^{t*} of a plant is computed as

$$\hat{\eta}_c^{t*} = \mathbf{K}_* \mathbf{K}^{-1} \hat{\boldsymbol{\eta}}_c^{1:\tau} \tag{3}$$

where $\mathbf{K}_*$ denotes the covariances between the new time t_* and the known ones $t = 1, \ldots, \tau$ computed with the kernel function using the estimated hyper-parameters $\hat{\vartheta}$. Given $\hat{\eta}_c^{t*}$, we compute the drought stress level as described in the previous section.

For the multiple-step ahead prediction task we follow the method proposed in [15]. That is, we predict the next time step using the estimate of the output of the current prediction as well as previous outputs (up to some lag U) as input, until the prediction k steps ahead is made. Thus, the prediction k steps ahead is a random vector with mean formed by the predicted means of the lagged outputs.

To summarize, based on hyper-spectral images, drought stress levels of plants are predicted as follows: (1) Using SiVM, we compute few extreme signatures, say 50, and label them accordingly. (2) On the simplex spanned by these extremes, we estimate the latent Dirichlet aggregation values per plant and time step using DAR. (3) Using the Gaussian process over the latent Dirichlet aggregation values, we compute the drought levels of each plant and time step using the labels of extreme spectra, i.e. "background", "healthy", and "dry". (4) Finally, we predict drought levels multiple steps ahead in time using the above Gaussian process approach.

3.3 Rainout Shelter Experiments

With the experiments reported next, we intended to investigate whether DAR can predict drought stress presymptomatically from hyper-spectral images and if DAR smoothing can lead to improved detection of drought stress patterns compared to the ones using SiVM only. SiVM and DAR were implemented in Python. Following [22], we ran SiVM using a variant of the Kolmogorov-Smirnov distance where we treated the reflectance signatures as empirical distributions. Once again, we point out that the methods presented in this section were developed for data analysis in basic agricultural research. That is to say, the data we analyze below was provided by biologists working on phenotyping of barley.

Data sets: We considered two sets of hyper-spectral images. Both data sets were recorded under semi-natural conditions in in rain-out shelters at the experimental station of the University of Bonn. For the controlled water stress, three barley summer cultivars Scarlett, Wiebke, and Barke were chosen. The seeds were sown in 11.5 liter pots filled with 17.5 kg of substrate Terrasoil. In 2010 (first data set) the genotype Scarlett was used in *two* treatments (well-watered

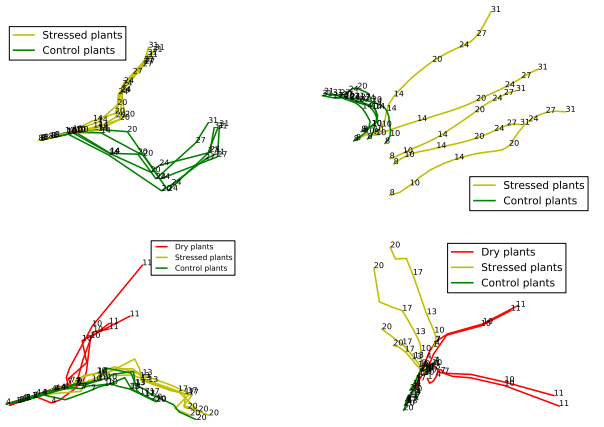


Fig. 5 Improved drought stress detection using DAR. From left to right: (1) Dirichlet traces for 2010 (two groups of measurements) without and (2) with DAR smoothing. (3) Dirichlet traces for 2011 (three groups of measurements) without and (4) with DAR smoothing. Colors indicate controlled/stressed plants; numbers denote the measurement days.

and with reduced water) with 6 pots per treatment. In 2011 (second data set) the genotypes Wiebke and Barke were used in pot experiments arranged in a randomized complete block design with *three* treatments (well-watered and two drought stressed) with 4 pots per genotype and treatment. The drought stress was induced either by reducing the total amount of water or by completely withholding water. In both cases, the stress was started at developmental stage BBCH31. By reducing the irrigation, the water potential of the substrate remained at the same level as in the well-watered pots for the first seven days but decreased rapidly in the following 10 days down to 40% of the control. For the measurements, the plants were transferred to the laboratory and illumination was provided by 6 halogen lamps fixed at a distance of 1.6 meters from the support where the pots where placed to record hyper-spectral pictures. These were obtained using the Surface Optics Corp. SOC-700 which records images of 640 pixels x 640 pixels with a spectral resolution of approximately 4 nm with up to 120 equally distributed bands in the range between 400 and 900 nm. In 2010, images were taken at 10 time-points, twice per week starting from day four of water-stress. This provided 70 data cubes of resolution $640 \times 640 \times 69$. We transformed each cube into a dense pixel by spectrum matrix. Stacking them horizontally resulted in a dense data matrix with about 2 Billion entries. In 2011 images were taken every consecutive day starting at the second day of watering reduction. Images were taken at 11 time-points for the non-irrigated plants and at 20 time-points for plants with reduced water amount. Applying the same procedure as for the data from 2010 resulted in a matrix of about 5.8 Billion entries.

Analysis Setup: We split the data from 2011 (resp. 2010) into a first half, denoted 2011.A (resp. 2010.A) and a second half, denoted as 2011.B (resp. 2010.B). Then, we extracted 50 extreme signatures from 2011.A (resp. 2010.A) and determined a DAR regression model on 2011.A (resp. 2010.A). We labeled the extreme signatures as "healthy", "dry", and "background", computed drought levels for 2011.A (resp. 2010.A) based on the DAR model, and used them to predict the drought levels for 2011.B (resp. 2010.B). We also considered the complete 2010 (resp. 2011) data to determine a corresponding DAR model and computed Euclidean embeddings as described in [22] using the smoothed αs. Finally, we note that SiVM can be parallelized so that plant phenotyping from the given data required only about 30 minutes. Estimating DAR models and making predictions happened within minutes.

Experimental Results: Figure 4 summarizes prediction results. In both cases, the notches of the boxes of the drought levels predicted for the different groups of plants at the last measurement day do not overlap. This suggests a statistically significant difference between the predicted medians. Moreover, the predictions match the "SiVM-only" values well. We conclude that DAR can fully distinguish all groups of plants. Thus, as in the fully observed case, we can group the extreme spectra based on their probability in the different groups at the last measurement days. The extreme signatures found were essentially identical to the ones found in [22]. We conclude that the automatically determined extreme signatures indeed conform to physiological plant knowledge. Moreover, since the symptoms do not become visible to the human eye at this time, it appears that data intensive technologies such as hyper-spectral imaging in deed provide benefits for research on plant phenotyping.

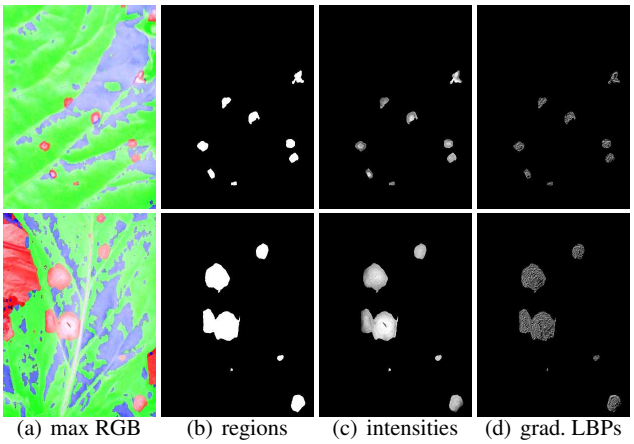
Figure 5 summarizes detection results. The Euclidean embeddings based on the DAR smoothing appear biologically plausible. Healthy plants are mapped closely together in a small region. Stressed plants cluster together only in an early period; later, they diverge due to the dispersion of senescence. Also, differences in senescence development between the different groups of plants are pronounced. Thus, we can again conclude that methods developed for data intensive discovery can benefit basic agricultural research.

4 Specific Example II: Leaf Spot Classification Using Smartphones

In this section, we present initial results from a project on using smartphones in farming. Smartphones have become widespread consumer products and prevalence and ease of use of their sensors suggest to explore their use in agriculture. Opportunities arise, for instance, from the availability

(a) Cercospora beticola (b) Phoma beta (c) Cercospora + Phoma

Fig. 6 Cell phone camera images of beet leafs showing leaf spots caused by Cercospora beticola and/or Phoma beta.



(a) max RGB (b) regions (c) intensities (d) grad. LBPs

Fig. 7 Results from three image preprocessing steps ((a) to (c)) and a feature extraction technique (d). The upper row shows results for Cercospora beticola, the lower row for Phoma beta. Panel (d) depicts local binary patterns of gradient magnitudes computed from (c).

of digital cameras in environments where they were not naturally present only a few years ago. Challenges are due to the fact that advanced signal processing on smartphones has to cope with constraints such as low battery life, restricted computational power, or limited bandwidth.

We aim at developing a system where farmers take pictures of plants they suspect to be infected by a disease. Pictures or information extracted therefrom are then sent to a central server and analysis results are to be reported back to the farmer while still in the field. The overall goal is not so much to assist farmers in classifying plant diseases but to provide an infrastructure that, as a nationwide system, allows for tracking how diseases spatially progress over the course of a farming season so as to facilitate the coordination of countermeasures.

Given the weak connection strengths out in the fields or the increased fees for high volume data transfer, farmers can hardly be expected to transmit pictures of sufficient resolution. If, on the other hand, image features were extracted by an app running on the cell phone, transmission times and costs would decrease. Yet, elaborate processing must be avoided since it might overburden the phone's battery or take too long to compute.

Addressing these issues, we tailor image preprocessing, feature extraction, and classification to the recognition of fungal pathogens that infect beet plants [2]. Beet is a widely cultivated crop but fungi frequently reduce yields. An early recognition of disease onsets assisted by our system may limit the amount of fungicides needed for pest control and can thus reduce costs and environmental burden. In particular, we attempt to automatically recognize symptoms of infection due to two common pests:

Cercospora beticola is a fungal pathogen that infects beet plants. Once infected, a plant shows leaf spots that are round blemishes with a definite edge between infected and healthy leaf tissue; while the border of a spot is typically darker and of brownish-reddish hue, spot interiors appear bright and greyish (see Fig. 6).

Phoma beta is a soil fungus that causes heart rot and blight of beets. Infected plants show rather large leaf spots of irregular shape and less pronounced edges; borders of a spot are chlorotic and of yellowish or light brown hue, spot interiors are characterized by slightly darker brownish hues; when several spots are present, they can grow together and form even larger blotches (see Fig. 6).

4.1 Preprocessing

Images that are to be analyzed in our scenario are recorded under uncontrollable conditions. Whenever farmers take snapshots of leafs in the field, scene illumination, camera angle, focal length, and distance to object are essentially random variables (s. Fig. 6). To alleviate diminishing effects due to some of these variables, the default setting of our system is to consider rather high resolution images. Given an RGB image I of 7.2 megapixels in JPG format, the following preprocessing steps proved to yield useful intermediate representations for later analysis:

1.) down-sample the input image I to an image D of 612×816 pixels; this reduced detrimental effect effects to to individual noisy pixels and facilitates subsequent computations.

2.) apply a max RGB filter to D to pronounce reddish/brownish image regions which hint at fungal leaf spots; that is for each (r,g,b) triplet in image D, we compute (r',g',b') to obtain an image D' where

$$r' = \begin{cases} 255, & \text{if } r > g \text{ and } r > b \\ r, & \text{otherwise} \end{cases}$$

and correspondingly for g' and b'; examples of images D' are show in Fig. 7(a).

3.) compute a binary image B from the RGB filtered image such that reddish pixels become foreground pixels and greenish/blueish pixels become background pixels.

4.) compute a region image R from B; this subsumes median filtering of B, connected component analysis, hole filling, filtering of regions adjacent to the image borders, as well as filtering of regions where the width/height ratio of their bounding boxes is is outside of $[\frac{1}{2}, 2]$; the latter is motivated from biological expertise: since leaf spots tend to be compact, elongated regions may be discarded from analysis; examples of images R are shown in Fig. 7(b).

5.) convert image D' to a grey scale image G; in G, suppress pixels that are background pixels in R and apply unsharp masking using

$$G(x,y) \leftarrow G(x,y) + \gamma\Big(G(x,y) - \big(\mathscr{N}_\sigma * G\big)(x,y)\Big) \tag{4}$$

where $*$ denotes convolution, $\mathscr{N}_\sigma$ is an isotropic Gaussian filter kernel of variance $\sigma = 3$ and the scaling constant is set to $\gamma = 1$; the information in G is analyzed subsequently; examples of images G are shown in Fig. 7(c).

4.2 Feature Extraction and Classification

In order to classify image regions resulting from the above preprocessing steps, we resort to texture features. Simple shape descriptors such as form factors [19], bounding box signatures [4], or entropies of curves [14] were tested but found to be lacking w.r.t. discriminative power. Again, as efficiency is key, texture features for analysis should be easy to compute. We therefore considered the use of histogram features. Elaborate methods for histogram analysis have been discussed in the previous section; for the application considered here, we found simpler approaches to achieve very good results. Especially local entropy features [19] led to high classification accuracy; in particular, we investigated the following methods:

a) for each region r_i in image G, compute a pixel intensity histogram h_i of n bins and use the entropy

$$e_i = -\sum_{k=1}^{n} h_{ik} \log(h_{ik}) \tag{5}$$

as a descriptor of r_i.

b) for each region r_i in G, apply contrast enhancement by means of histogram equalization, compute a histogram h_i from the contrast enhanced intensity values and use its entropy e_i as a descriptor of region r_i.

c) for each pixel (x,y) of region r_i in image G, compute local gradients using the Sobel operator $(-1,0,1)$, bin their magnitudes into a histogram h_i, and use its entropy e_i as a descriptor of region r_i.

In addition, we explored the merits of local binary patterns (LBPs) for our scenario. LBPs as introduced in [34] compress local texture information and are frequently reported to enable accurate classification. Accordingly, we also considered the following methods:

d) for each pixel (x,y) of region r_i in image G, compute an LBP descriptor from its 8-neighborhood; then, compute a histogram h_i of LBP values and use its entropy e_i as a descriptor; note that instead of applying the original approach in [34], we follow a suggestion in [21] and obtain LBPs from thresholding at the local median intensity rather that at the intensity of the neighborhood's center pixel.

e) for each region r_i in G, compute LBPs from contrast enhanced intensities, bin them into a histogram and use its entropy as a descriptor.

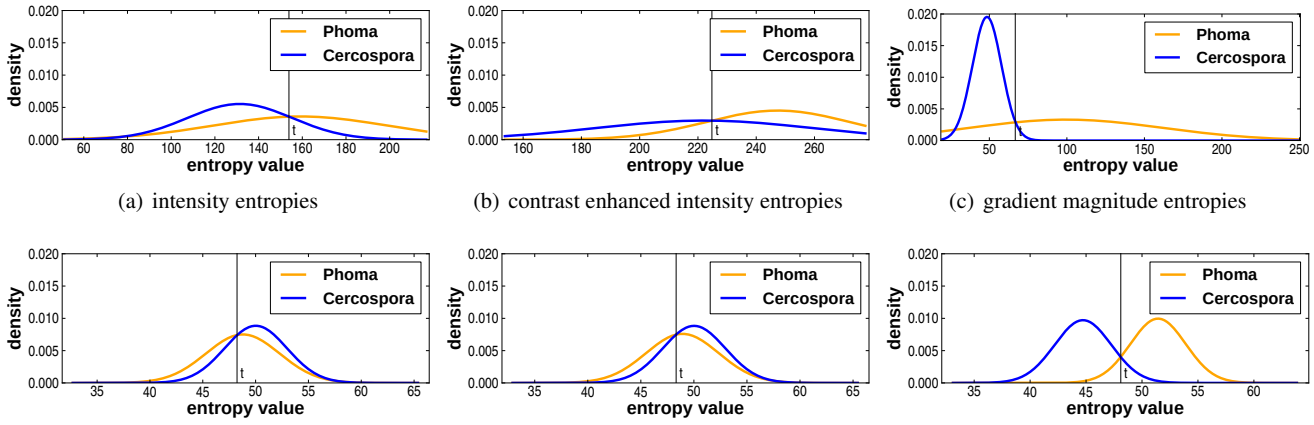
f) for each region r_i in G, compute LBPs from the region's gradient magnitudes, bin them into a histogram and use its entropy as a descriptor.

Figure 8 shows class specific entropy densities for the above features where the number n of histogram bins was set to 64. The densities were obtained from fitting Normal distributions to entropy values computed from a set of training images showing leafs infected by Cercospora beticola and Phoma beta. Entropies resulting from from histograms of local binary patterns of gradient magnitudes are particularly well separated (see Fig. 8(f)). In fact, a closer inspection of the corresponding data revealed that the seemingly high within class variances are due to very few outliers only and that the vast majority of data are located closely to their respective sample means. These results indicate that, in our scenario, leaf spot classification may be accomplished using naïve Bayes classifiers where an observation is assigned to that class for which the corresponding density is highest [33].

4.3 Experimental Procedure and Results

Next, we present first results obtained using the above approach to fungal leaf spot classification. We experimented with a set of 1000 smartphone (Motorola Razr, Samsung Galaxy SIII, Sony Ericsson Xperia) images that were taken in natural environments under uncontrolled conditions. Each of the images in our experiments showed leafs of plants infected by Cercospora beticola, Phoma beta, or both.

50% of the available images were randomly selected for training the remaining 50% for testing. The number of leaf spots in these images varied from less than ten to more than a hundred. In each image (training or testing) at least five but not more than ten spots were manually labeled by an expert as either caused by Cercospora or Phoma. 37% of the leaf spots used for training were caused by Cercospora beticola, the remaining 63% were due to Phoma beta. In testing, 14% of the images showed only Cercospora leaf spots, 9% of the images showed only Phoma leaf spots, and 77% of the images showed both types of leaf spots.



(a) intensity entropies (b) contrast enhanced intensity entropies (c) gradient magnitude entropies

(d) intensity LBP entropies (e) contrast enhanced intensity LBP entropies (f) gradient magnitude LBP entropies

Fig. 8 Class specific densities of six different entropy measures. The statistics were obtained from fitting normal distributions to entropy values computed from a training set of images of plants infected by Cercospora beticola and Phoma beta. The classification threshold t indicates where densities intersect. Entropy values resulting from histograms of local binary patterns (LBPs) of gradient magnitudes (see panel (f)) appear well separated. In fact, the within class variances in this case are due to only a few outliers; most entropy values in (f) cluster close to their means; the Bhattacharyya distance $0 \leq d_B \leq 1$ between the densities in (f) amounts to $d_B = 0.89$ indicating that the two classes are indeed well separable.

Since cell phone implementations of the proposed processing cascade are not yet available, experiments were run on a desktop computer (Intel core i3 processor 540, 4GB RAM) running Ubuntu Linux 10.04 using C implementations of our methods.

Table 1 presents computation times measured for different stages of the above image processing cascade averaged over 100 runs with the test images. These figures underline the efficiency of the proposed approach: preprocessing and extraction and classification of features run in fractions of a second. Given earlier experience with implementations of image processing algorithms on Android cellphones, we expect these runtimes to increase by a factor of 2 to 5 for software running on mobile devices.

Figure 9 compares recognition accuracies obtained from training and testing with manually labeled data. To produce the curves in this figure, the six different entropy-based texture descriptors discussed in section 4.2 were computed from histograms where the number n of histogram bins was varied from 4 to 128. Overall, the results corroborate what is apparent in Fig. 8. For histograms of more than $n = 16$ bins, entropies of gradient magnitude LBPs outperform the other descriptors considered in our experiments. Accuracies obtained from this feature exceed 97% and suggest that the proposed processing cascade can reliably distinguish leaf spots caused by Cercospora beticola or Phoma beta.

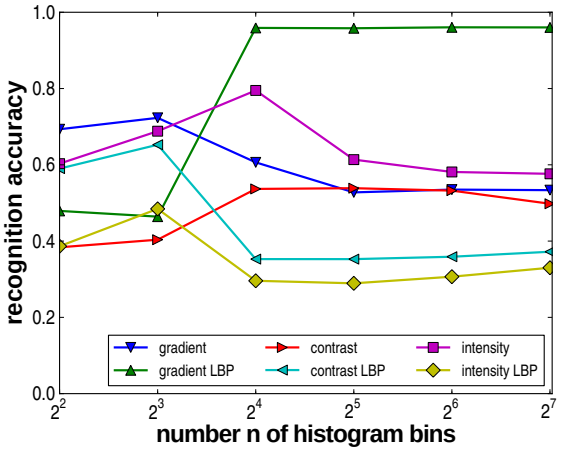


Fig. 9 Recognition accuracies obtained from experiments on leaf spot classification using different texture features. Naïve Bayes-classifiers based on LBPs of gradient magnitudes reach accuracies of 97% if the entropy measures are computed from histograms of more than 16 bins.

5 Conclusion

Agriculture, the oldest economic venture in the history of humankind, is currently undergoing yet another technological revolution. Sparked by issues pertaining to sustainability, climate change, and growing populations, solutions for precision farming are increasingly sought for and deployed in agricultural research and practice.

We surveyed recent work on computational intelligence in precision farming. From the point of view of pattern recognition and data mining, the major challenges in agricultural applications appear to be the following:

1. The widespread deployment and ease of use of modern, (mobile) sensor technologies leads to exploding amounts of data. This poses problems of BIG DATA and high-troughput computation. Algorithms and frameworks for data management and analysis need to be developed that can easily cope with TeraBytes of data.

Table 1 Average computation times for processing color images of 2448×3264 pixels.

		computation and classification of entropy feature based on a histogram of 128 bins of					
processing task	preprocessing	intensities	int. LBPs	enhanced int.	enh. int. LBPs	grad. magnitudes	grad. mag. LBPs
processing time	58 ms	5 ms	13 ms	15 ms	24 ms	18 ms	26 ms

2. Since agriculture is a truly interdisciplinary venture whose practitioners are not necessarily trained statisticians or data scientists, techniques for data analysis need to deliver interpretable and understandable results.
3. Mobile computing for applications "out in the fields" has to cope with resource constraints such as restricted battery life, low computational power, or limited bandwidths for data transfer. Algorithms intended for mobile signal processing and analysis need to address these constraints.

To illustrate these challenges, we presented two examples from our own work on computational intelligence in agriculture. First, we considered the problem of drought stress recognition in research on plant phenotyping from hyper-spectral imaging and second an application of mobile sensing for vision-based plant disease recognition. We presented algorithmic solutions that cope with TeraBytes of sensor recording and deliver useful, i.e. biologically plausible, and interpretable results. In particular, our approach was based on a distributional view of hyper-spectral signatures which we used for Baysian prediction of the development of drought stress levels. Prediction models of this kind have great potential as they provide better insights into early stress reactions and to identify the most relevant moment when biologists or farmers have to gather samples for invasive, molecular examinations.

We also presented a cascade of simple image processing and analysis steps of low computational costs that allows for reliably distinguishing different fungal leaf spots in natural, unconstrained images of leaves of beet plants. In this context, speed and efficiency are is pivotal because the proposed approach is supposed to be integrated into an online system that allows farmers in the field to take pictures of plants they suspect to be infected and have them analyzed in real time.

In conclusion these examples show that methods from the broad field of artificial intelligence, in particular from data mining and pattern recognition, can contribute to solving problems due to water shortage or pests. Together with the related contributions in our survey, it thus appears that developments such as mobile, wireless and positioning networks, new environmental sensors, and novel computational intelligence methods do have the potential of contributing to the vision of a sustainable agriculture for the 21st century.

Acknowledgements Parts of the work reported here were conducted within the project *SmartDDS* which is funded by the Bundesanstalt für Landwirschaft und Ernährung. Kristian Kersting was supported by the Fraunhofer ATTRACT fellowship "Statistical Relational Activity Mining". The authors gratefully acknowledge this support.

References

1. Abdeen, A., Schnell, J., Miki, B.: Transcriptome analysis reveals absence of unintended effects in drought-tolerant transgenic plants overexpressing the transcription factor abf3. BMC Genomics **11**(69) (2010)
2. Agrios, G.: Plant Pathology, 4th edn. Academic Press (1997)
3. Ballvora, A., Römer, C., Wahabzada, M., Rascher, U., Thurau, C., Bauckhage, C., Kersting, K., Plümer, L., Leon, J.: Deep phenotyping of early plant response to abiotic stress using non-invasive approaches in barley. In: G. Zhang, C. Li, X. Liu (eds.) Advance in Barley Sciences, chap. 26, pp. 301–316. Springer (2013)
4. Bauckhage, C.: Tree-based signatures for shape classification. In: Proc. ICIP (2006)
5. Bauckhage, C., Kersting, K., Schmidt, A.: Agriculture's technological makeover. IEEE Pervasive Computing **11**(2), 4–7 (2012)
6. Bechar, I., Moisan, S., M.Thonnat, Bremond, F.: On-line video recognition and counting of harmful insects. In: Proc. ICPR (2010)
7. Bergamaschi, S., Sala, A.: Creating and querying an integrated ontology for molecular and phenotypic cereals data. In: M. Sicilia, M. Lytras (eds.) Metadata and Semantics, pp. 445–445. Springer (2009)
8. Blanco, P., Metternicht, G., Del Valle, H.: Improving the discrimination of vegetation and landform patterns in sandy rangelands: a synergistic approach. Int. J. of Remote Sensing **30**(10), 2579–2605 (2009)
9. Boyer, J.: Plant productivity and environment. Science **218**, 443–448 (1982)
10. Burrell, J., Brooke, T., Beckwith, R.: Vineyard computing: Sensor networks in agricultural production. IEEE Pervasive Computing **3**(1), 38–45 (2004)
11. Chakraborty, S., Subramanian, L.: Location specific summarization of climatic and agricultural trends. In: Proc. WWW (2011)
12. Civril, A., Magdon-Ismail, M.: On selecting a maximum volume sub-matrix of a matrix and related problems. Theoretical Computer Science **410**(47–49), 4801–4811 (2009)
13. Crowley, M., Poole, D.: Policy gradient planning for environmental decision making with existing simulators. In: Proc. AAAI (2011)
14. Ebrahim, Y., Ahmed, M., Chau, S., Abdelsalam, W.: An efficient shape representation and description technique. In: Proc. ICIP (2007)
15. Girard, A., Rasmussen, C., Quinonero Candela, J., Murray-Smith, R.: Gaussian process priors with uncertain inputs – application to multiple-step ahead time series forecasting. In: Proc. NIPS (2002)
16. Gnomes, C.: Computational sustainability: Computational methods for a sustainable environment, economy, and society. The Bridge **39**(4), 5–13 (2009)
17. Gocht, A., Roder, N.: Salvage the treasure of geographic information in farm census data. In: Proc. Int. Congress European Association of Agricultural Economists (2011)

18. Golovin, D., Krause, A., Gardner, B., Converse, S., Morey, S.: Dynamic resource allocation in conservation planning. In: Proc. AAAI (2011)
19. Gonzales, R., Woods, R.: Digital Image Processing, 3rd edn. Pearson Prentice Hall (2008)
20. Guo, P., Baum, M., Grando, S., Ceccarelli, S., Bai, G., Li, R., von Korff, M., Varshney, R., Graner, A., Valkoun, J.: Differentially expressed genes between drought-tolerant and drought-sensitive barley genotypes in response to drought stress during the reproductive stage. J. Experimental Botany **60**(12), 3531–3544 (2010)
21. Hafiane, A., Seetharaman, G., Palaniappan, K., Zavidovique, B.: Rotationally invariant hashing of median binary patterns for texture classification. In: Proc. ICIAR (2008)
22. Kersting, K., Wahabzada, M., Roemer, C., Thurau, C., Ballvora, A., Rascher, U., Leon, J., Bauckhage, C., Pluemer, L.: Simplex distributions for embedding data matrices over time. In: Proc. SDM (2012)
23. Kersting, K., Xu, Z., Wahabzada, M., Bauckhage, C., Thurau, C., Römer, C., Ballvora, A., Rascher, U., Leon, J., Plümer, L.: Presymptomatic prediction of plant drought stress using dirichlet-aggregation regression on hyperspectral images. In: Proc. AAAI (2012)
24. Kui, F., Juan, W., Weiqiong, B.: Research of optimized agricultural information collaborative filtering recommendation systems. In: Proc. ICICIS (2011)
25. Kumar, V., Dave, V., Bhadauriya, R., Chaudhary, S.: Krishimantra: Agricultural recommendation system. In: Proc. ACM Symp. on Computing for Development (2013)
26. Laykin, S., Alchanatis, V., Edan, Y.: On-line multi-sateg sorting algorithm for agriculture products. Pattern Recognition **45**(7), 2843–2853 (2012)
27. Lebreton, C., Lazic-Jancic, V., Steed, A., Pekic, S., Quarrie, S.: Identification of qtl for drought responses in maize and their use in testing causal relationships between traits. J. Experimental Botanic **46**(7), 853–865 (1995)
28. Lin, H., Cheng, J., Pei, Z., Zhang, S., Hu, Z.: Monitoring sugarcane growth using envisat asar data. IEEE Trans. Geoscience and Remote Sensing **47**(8), 2572–899 (2009)
29. Loew, A., Ludwig, R., Mauser, W.: Derivation of surface soil moisture from envisat asar wide swath and image mode data in agricultural areas. IEEE Trans. Geoscience and Remote Sensing **44**(4), 889–899 (2006)
30. McKay, J., Richards, J., Sen, S., Mitchell-Olds, T., Boles, S., Stahl, E., Wayne, T., Juenger, T.: Genetics of drought adaptation in arabidopsis thaliana ii. qtl analysis of a new mapping population, kas-1 x tsu-1. Evolution **62**(12), 3014–3026 (2008)
31. Medjahed, B., Gosky, W.: A notification infrastructure for semantic agricultural web services. In: M. Sicilia, M. Lytras (eds.) Metadata and Semantics, pp. 455–462. Springer (2009)
32. Mewes, T., Franke, J., Menz, G.: Data reduction of hyperspectral remote sensing data for crop stress detection using different band selection methods. In: Proc. IEEE Int. Geoscience and Remote Sensing Symp. (2009)
33. Mitchell, T.: Machine Learning. McGraw-Hill (1997)
34. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. PAMI **24**(7), 971–987 (2002)
35. Passioura, J.: Environmental biology and crop improvement. Functional Plant Biology **29**, 537–554 (2002)
36. Petrik, M., Zilberstein, S.: Linear dynamic programs for resource management. In: Proc. AAAI (2011)
37. Pinnisi, E.: The blue revolution, drop by drop, gene by gene. Science **320**(5873), 171–173 (2008)
38. Rabbani, M., Maruyama, K., Abe, H., Khan, M., Katsura, K., Ito, Y., Yoshiwara, K., Seki, M., Shinozaki, K., Yamaguchi-Shinozaki, K.: Monitoring expression profiles of rice genes under cold, drought, and high-salinity stresses and abscisic acid application using cdna microarray and rna gel-blot analyses. Plant Physiology **133**(4), 1755–1767 (2010)
39. Rascher, U., Nichol, C., Small, C., Hendricks, L.: Monitoring spatio-temporal dynamics of photosynthesis with a portable hyperspectral imaging system. Photogrammetric Engineering and Remote Sensing **73**(1), 45–56 (2007)
40. Rascher, U., Pieruschka, R.: Spatio-temporal variations of photosynthesis: The potential of optical remote sensing to better understand and scale light use efficiency and stresses of plant ecosystems. Precision Agriculture **9**(6), 355–366 (2008)
41. Rasmussen, C., Williams, C.: Gaussian Processes for Machine Learning. The MIT Press (2006)
42. Rocha, A., Hauagge, D., Wainer, J., Goldenstein, S.: Automatic fruit and vegetable classification from images. Computers and Electronics in Agriculture **70**(1), 96–104 (2010)
43. Römer, C., Bürling, K., Rumpf, T., Hunsche, M., Noga, G., Plümer, L..: Robust fitting of fluorescence sprectra for presymptomatic wheat leaf rust detection with support vector machines. Computers and Electronics in Agriculture **74**(1), 180–188 (2010)
44. Römer, C., Wahabzada, M., Ballvora, A., Pinto, F., Rossini, M., Panigada, C., Behmann, J., Leon, J., Thurau, C., Bauckhage, C., Kersting, K., Rascher, U., Plümer, L.: Early drought stress detection in cereals: Simplex volume maximization for hyperspectral image analysis. Functional Plant Biology **39**(11), 878–890 (2012)
45. Rumpf, T., Mahlein, A.K., Steiner, U., Oerke, E.C., Plümer, L.: Early detection and classification of plant diseases with support vector machines based on hyperspectral reflectance. Computers and Electronics in Agriculture **74**(1), 91–99 (2010)
46. Ruß, G., Brenning, A.: Data mining in precision agriculture: Management of spatial information. In: Proc. IPMU (2010)
47. Sankaran, S., Mishra, A., Ehsani, R., Davis, C.: A review of advanced techniques for detecting plant diseases. Computers and Electronics in Agriculture **72**(1), 1–13 (2010)
48. Satalino, G., Mattia, F., Le Toan, T., Rinaldi, M.: Wheat crop mapping by using asar ap data. IEEE Trans. Geoscience and Remote Sensing **47**(2), 527–530 (2009)
49. Schmitz, M., Martini, D., Kunisch, M., Mosinger, H.J.: agroxml: Enabling standardized, platform-independent internet data exchange in farm management information systems. In: M. Sicilia, M. Lytras (eds.) Metadata and Semantics, pp. 463–467. Springer (2009)
50. Thurau, C., Kersting, K., Wahabzada, M., Bauckhage, C.: Descriptive matrix factorization for sustainability: Adopting the principle of opposites. Data Mining and Knowledge Discovery **24**(2), 325–354 (2012)
51. Vernon, R. (ed.): Knowing Where You're Going: Information Systems for Agricultural Research Management. International Service for Agricultural Research (ISNAR) (2001)
52. Wark, T., Corke, P., Klingbeil, L., Guo, Y., Crossman, C., Valencia, P., Swain, D., Bishop-Hurley, G.: Transforming agriculture through pervasive wireless sensor networks. IEEE Pervasive Computing **6**(2), 50–57 (2007)