UNIVERSITY OF TWENTE.

Introduction to Machine Learning

Mahdi KHODADADZADEH

October 2021

ITC FACULTY OF GEO-INFORMATION SCIENCE AND EARTH OBSERVATION

Question

- You want to identify outliers in a dataset with house sale prices in Dhaka.
- What exploratory method would you suggest using for this purpose?
- Explain what the method does, and how to interpret its output?

Question

- **You want to identify outliers in a dataset with house sale prices in Dhaka.**
- **What exploratory method would you suggest using for this purpose?**
- **Explain what the method does, and how to interpret its output?**

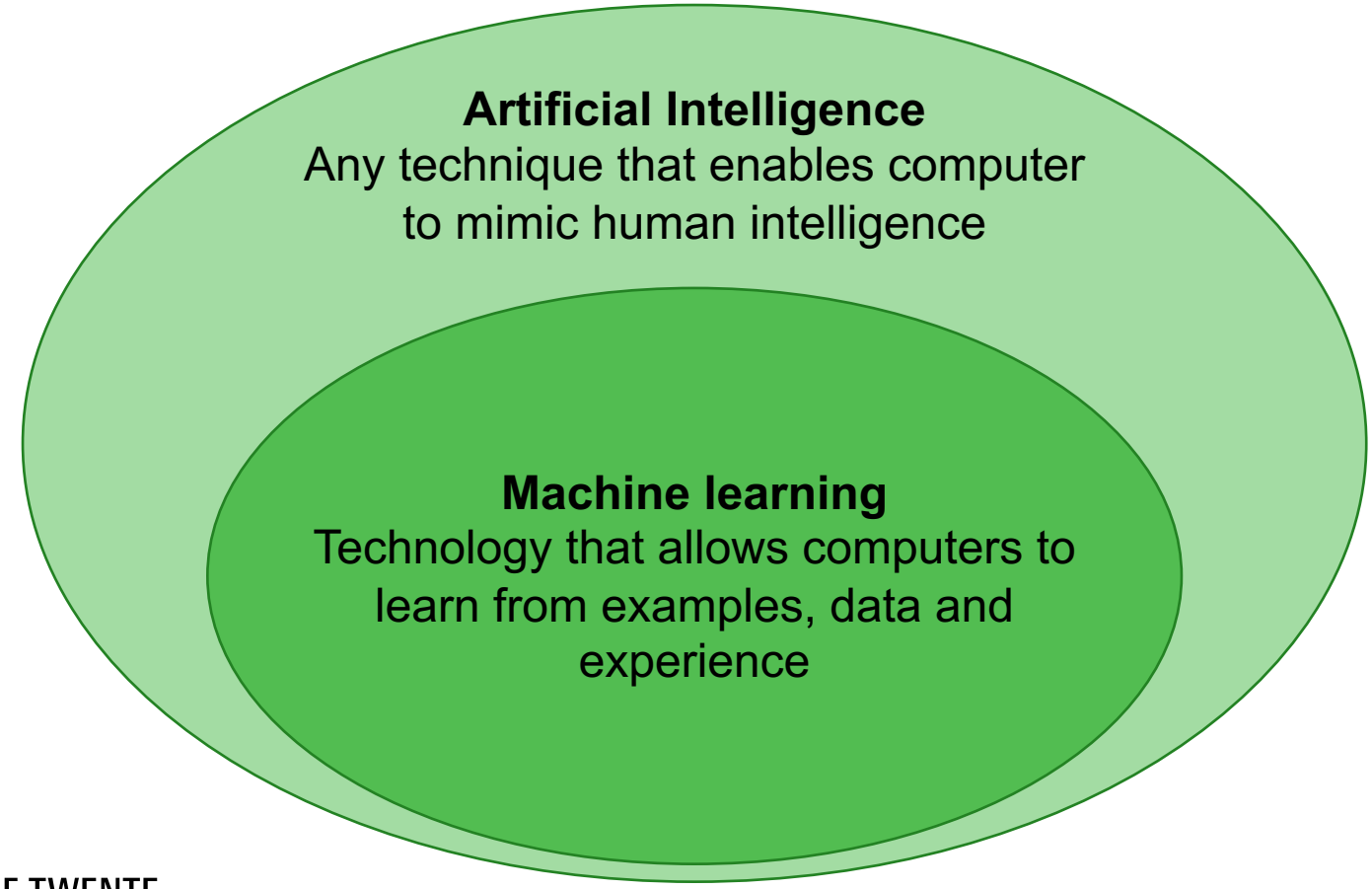
  - An outlier is an observation that is numerically distant from the rest of the data. A very straightforward method would be box map (or boxplot). It finds six bins categories to identify the lower and upper outliers. The definition of outliers is a function of a multiple of the inter-quartile range (IQR), the difference between the values for the 75 and 25 percentile. For example, outside 1.5 times the IQR above the upper quartile and below the lower quartile (Q1 - 1.5 * IQR or Q3 + 1.5 * IQR).

Outline

- **Machine Learning Basics**
  - Terminology and key concepts
- **Performance Evaluation**
  - Metrics for regression and classification
- **Variance and Bias**
- **Dimensionality Reduction**
  - Principal Component Analysis

Basics

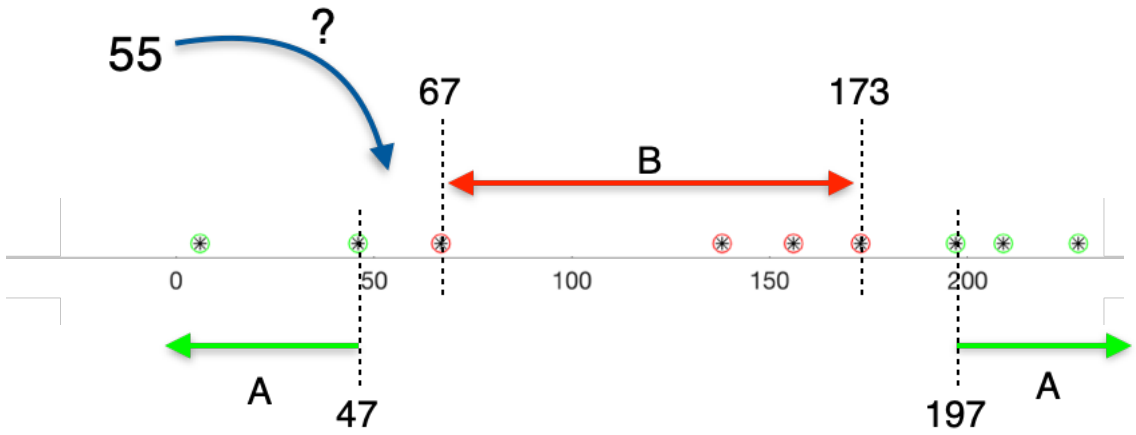
- What is Machine Learning?



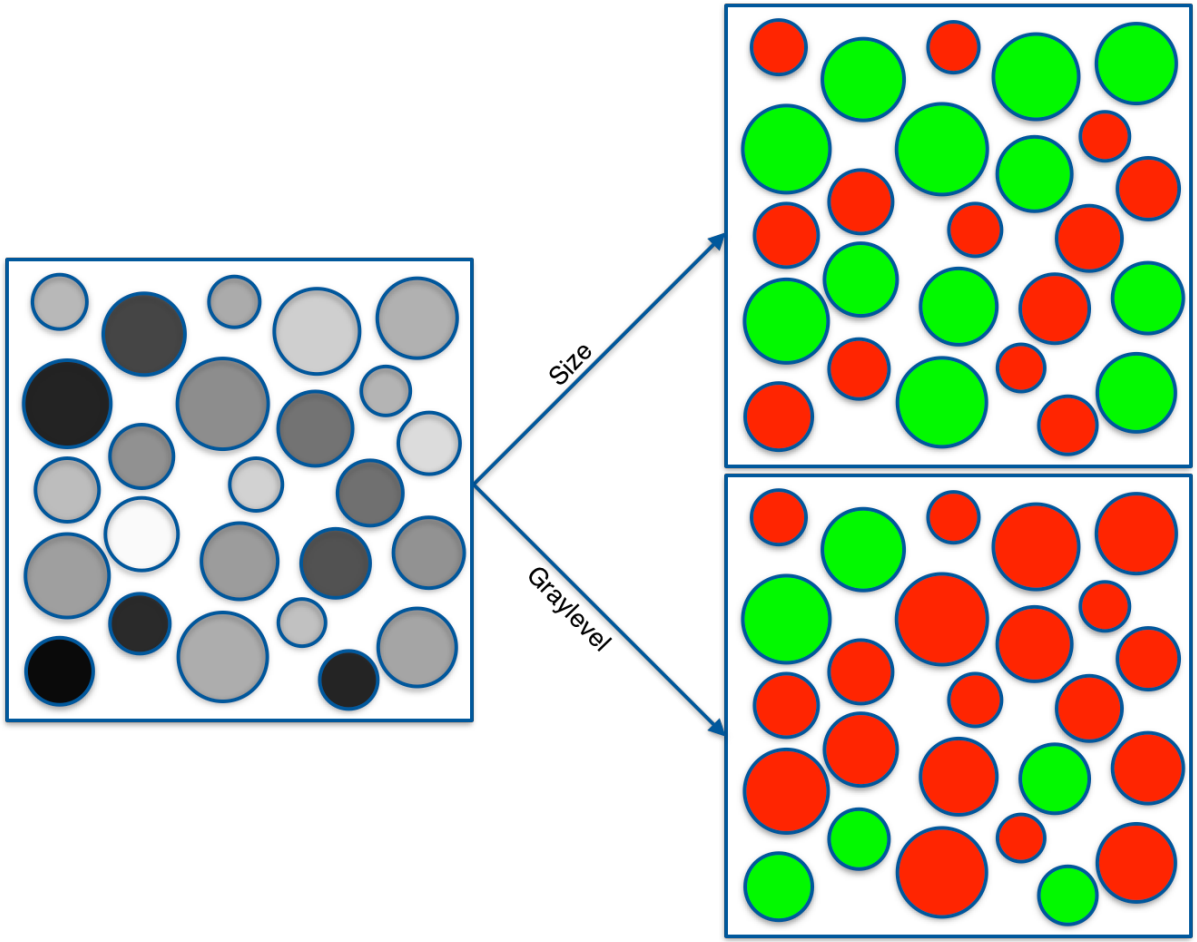
Basics

- Finding a rule that is consistent with the data
- Increasing the number of samples will improve the definition of the rule

228	A
67	B
138	B
209	A
156	B
46	A
197	A
6	A
173	B

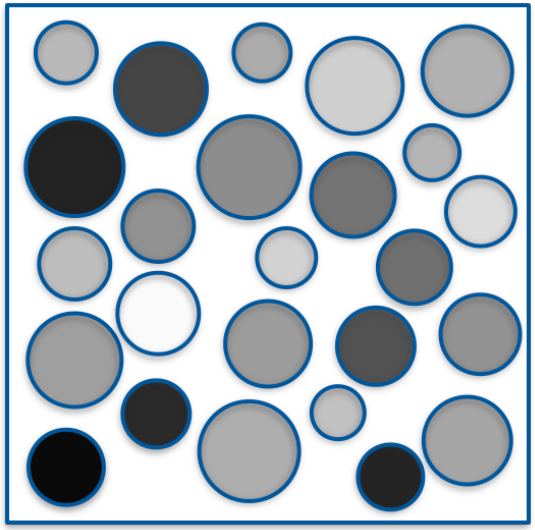


Basics



Basics

- In machine learning, a feature is a measurable property of a phenomenon being observed.

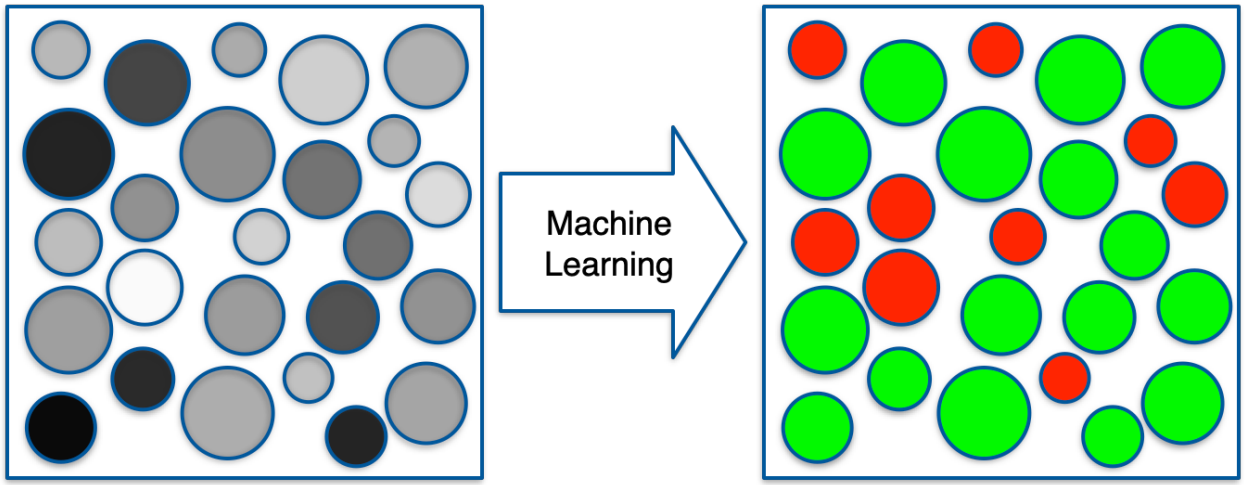


Size	Graylevel
0.17	0.23
0.92	0.86
0.79	0.69
0.38	0.20
0.21	0.37
0.83	0.31
0.58	0
0.29	0.82
1	0.38
0.08	0.27
0.88	0.14
0.63	0.48
0.13	0.13
0.67	0.32
0.96	0.26
0.46	1
0	0.19
0.25	0.85
0.50	0.62
0.71	0.29
0.75	0.25
0.04	0.24
0.33	0.10
0.42	0.49
0.54	0.35

Features

Feature Vector

Basics



- Design an algorithm that ingest data and learn a model of the data
- We do not pre-define and hard-code all the rules
- Modern Machine Learning algorithms are heavily data-driven

Basics

- **When should we use Machine Learning**?
  - Problems where there are no human experts, so data cannot be labelled or categorized
  - Problems where there are human experts, but it is very hard (or impossible) to define rules
  - Problems where there are human experts and the rules can be defined, but where it is not cost effective to implement

  - When we want to
    - ✓ Detect patterns/structures/themes/trends etc. in the data
    - ✓ Make predictions about future data and make decisions

Basics

- **Machine Learning Applications:**
  - Remote sensing data analysis
  - Image processing and computer vision, for face recognition, motion detection, and object detection
  - Computational finance, for credit scoring and algorithmic trading
  - Computational biology, for tumor detection, drug discovery, and DNA sequencing
  - Energy production, for price and load forecasting
  - Automotive, aerospace, and manufacturing, for predictive maintenance
  - Natural language processing

Basics

Open Access Review

Machine Learning in Agriculture: A Review

by Konstantinos G. Liakos [1], Patrizia Busato [2], Dimitrios Moshou [1,3], Simon Pearson [4] and Dionysis Bochtis [1,*]

[1] Institute for Bio-Economy and Agri-Technology (IBO), Centre of Research and Technology—Hellas (CERTH), 6th km Charilaou-Thermi Rd, GR 57001 Thessaloniki, Greece

[2] Department of Agriculture, Forestry and Food Sciences (DISAFA), Faculty of Agriculture, University of Turin, Largo Braccini 2, 10095 Grugliasco, Italy

[3] Agricultural Engineering Laboratory, Faculty of Agriculture, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece

[4] Lincoln Institute for Agri-food Technology (LIAT), University of Lincoln, Brayford Way, Brayford Pool, Lincoln LN6 7TS, UK

* Author to whom correspondence should be addressed.

Sensors **2018**, *18*(8), 2674; https://doi.org/10.3390/s18082674

Received: 27 June 2018 / Revised: 31 July 2018 / Accepted: 7 August 2018 / Published: 14 August 2018

(This article belongs to the Special Issue Sensors in Agriculture 2018)

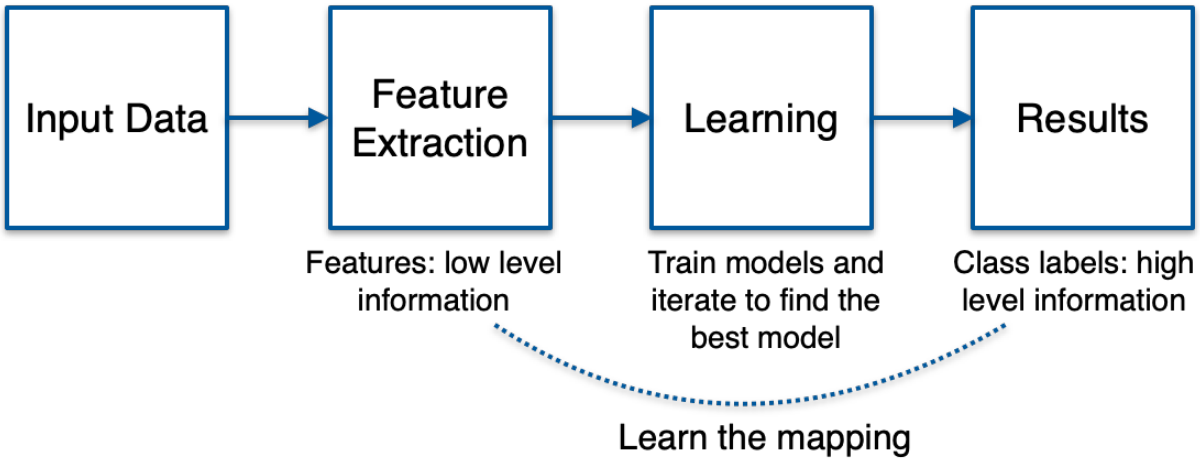
View Full-Text | Download PDF | Browse Figures | Citation Export

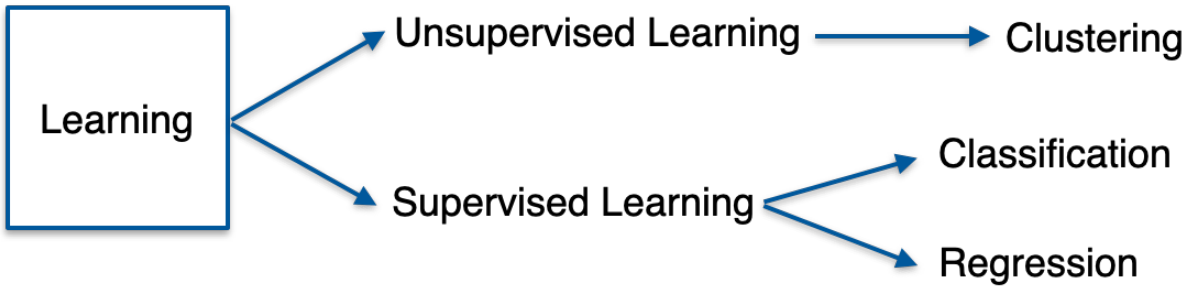
Abstract

Machine learning has emerged with big data technologies and high-performance computing to create new opportunities for data intensive science in the multi-disciplinary agri-technologies domain. In this paper, we present a comprehensive review of research dedicated to applications of machine learning in agricultural production systems. The works analyzed were categorized in (a) crop management, including applications on yield prediction, disease detection, weed detection crop quality, and species recognition; (b) livestock management, including applications on animal welfare and livestock production; (c) water management; and (d) soil management. The filtering and classification of the presented articles demonstrate how agriculture will benefit from machine learning technologies. By applying machine learning to sensor data, farm management systems are evolving into real time artificial intelligence enabled programs that provide rich recommendations and insights for farmer decision support and action.
View Full-Text

Keywords: crop management; water management; soil management; livestock management; artificial intelligence; planning; precision agriculture

Basics

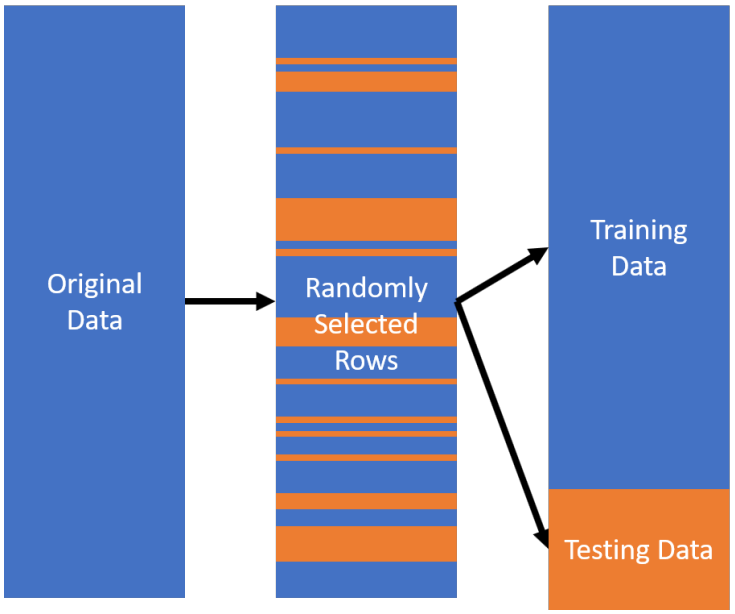

Input Data
Feature Extraction
Learning
Results
Features: low level information
Train models and iterate to find the best model
Class labels: high level information
Learn the mapping


Learning
Unsupervised Learning
Clustering
Supervised Learning
Classification
Regression

Supervised Learning

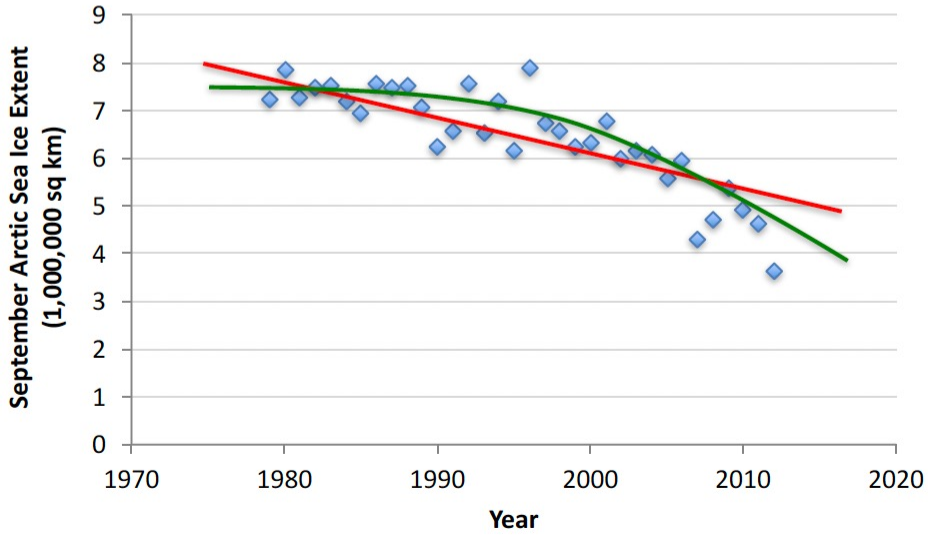
- Learning using labeled data
- Given (x1, y1), (x2, y2), ..., (xn, yn)
- Learn a function f(x) to predict y given x

sklearn.model_selection.train_test_split



Supervised Learning: Regression

- Learning using labeled data
- Given (x1, y1), (x2, y2), ..., (xn, yn)
- Learn a function f(x) to predict y given x
  - y is real-valued == regression



Supervised Learning: Regression

International Journal of Health Geographics

Home About Articles Collections Submission Guidelines

Research | Open Access | Published: 14 November 2017

Modelling and mapping tick dynamics using volunteered observations

Irene Garcia-Martí, Raúl Zurita-Milla, Arnold J. H. van Vliet & Willem Takken

International Journal of Health Geographics **16**, Article number: 41 (2017) | Cite this article

4067 Accesses | **11** Citations | **1** Altmetric | Metrics

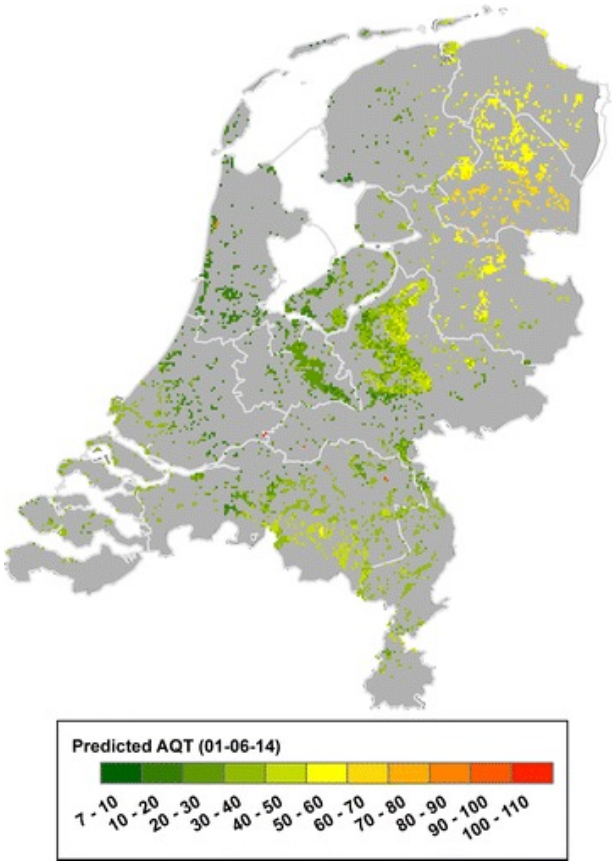
Abstract

Background

Tick populations and tick-borne infections have steadily increased since the mid-1990s posing an ever-increasing risk to public health. Yet, modelling tick dynamics remains challenging because of the lack of data and knowledge on this complex phenomenon. Here we present an approach to model and map tick dynamics using volunteered data. This approach is illustrated with 9 years of data collected by a group of trained volunteers who sampled active questing ticks (AQT) on a monthly basis and for 15 locations in the Netherlands. We aimed at finding the main environmental drivers of AQT at multiple time-scales, and to devise daily AQT maps at the national level for 2014.

Method

Tick dynamics is a complex ecological problem driven by biotic (e.g. pathogens, wildlife, humans) and abiotic (e.g. weather, landscape) factors. We enriched the volunteered AQT collection with six types of weather variables (aggregated at 11 temporal scales), three types of satellite-derived vegetation indices, land cover, and mast years. Then, we applied a feature engineering process to derive a set of 101 features to characterize the conditions that yielded a particular count of AQT on a date and location. To devise models predicting the AQT, we use a time-aware Random Forest regression method, which is suitable to find non-linear relationships in complex ecological problems, and provides an estimation of the most important features to predict the AQT.



Supervised Learning: Regression

Open Access Article

Drill-Core Mineral Abundance Estimation Using Hyperspectral and High-Resolution Mineralogical Data

by Laura Tuşa *, Mahdi Khodadadzadeh, Cecilia Contreras, Kasra Rafiezadeh Shahi, Margret Fuchs, Richard Gloaguen and Jens Gutzmer

Helmholtz Institute Freiberg for Resource Technology, Helmholtz-Zentrum Dresden-Rossendorf, Chemnitzer Straße 40, 09599 Freiberg, Germany

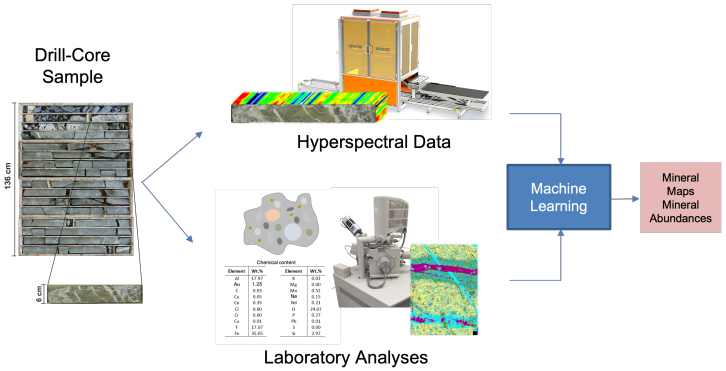
* Author to whom correspondence should be addressed.

Remote Sens. **2020**, *12*(7), 1218; https://doi.org/10.3390/rs12071218

Received: 16 March 2020 / Revised: 8 April 2020 / Accepted: 9 April 2020 / Published: 9 April 2020

(This article belongs to the Special Issue Multispectral and Hyperspectral Remote Sensing Data for Mineral Exploration and Environmental Monitoring of Mined Areas)

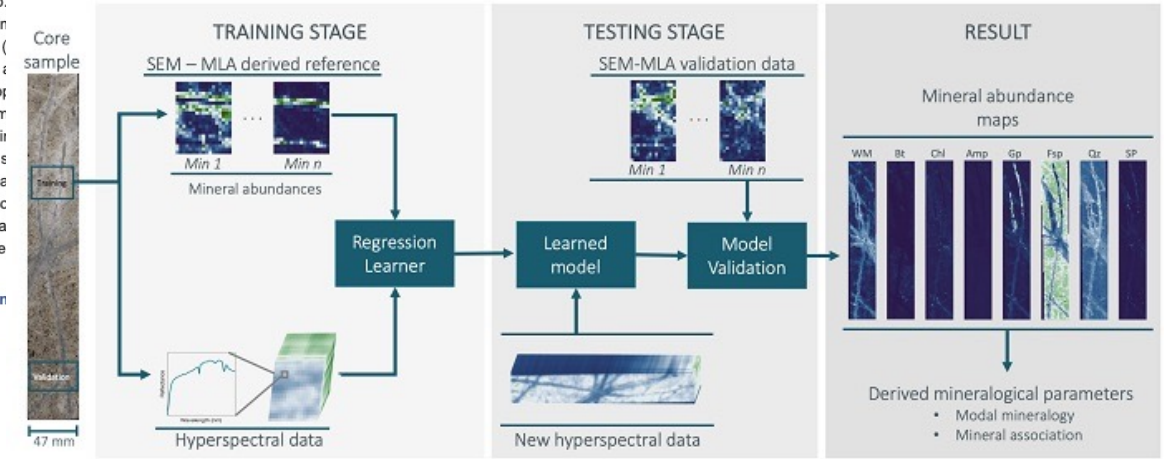
Download PDF | Browse Figures | Citation Export



Abstract

Due to the extensive drilling performed every year in exploration campaigns for the discovery and evaluation of ore deposits, drill-core mapping is becoming an essential step. While valuable mineralogical information is extracted during core logging by on-site geologists, the process is tir individual background. Hyperspectral short-wave infrared (complement traditional logging techniques and to provide a characterization. Additionally, Scanning Electron Microscop Analyser (SEM-MLA) provide exhaustive high-resolution m areas of the drill-cores. We propose to use machine learni the quantitative SEM-MLA mineralogical data to drill-core s core samples are obtained. Our upscaling approach increa physical-based data acquisition (hyperspectral imaging) cc procedure is tested on 5 drill-core samples with varying tra and neural network regression models. The obtained mine mineralogical parameters such as mineral association.

Keywords: hyperspectral imaging; drill-core; SWIR; min learning



Supervised Learning: Regression

ELSEVIER

Science of The Total Environment
Volume 636, 15 September 2018, Pages 52-60

A machine learning method to estimate $PM_{2.5}$ concentrations across China with remote sensing, meteorological and land use information

Gongbo Chen [a], Shanshan Li [a], Luke D. Knibbs [b], N.A.S. Hamm [c], Wei Cao [d], Tiantian Li [e], Jianping Guo [f], Hongyan Ren [d], Michael J. Abramson [a], Yuming Guo [a]

Show more

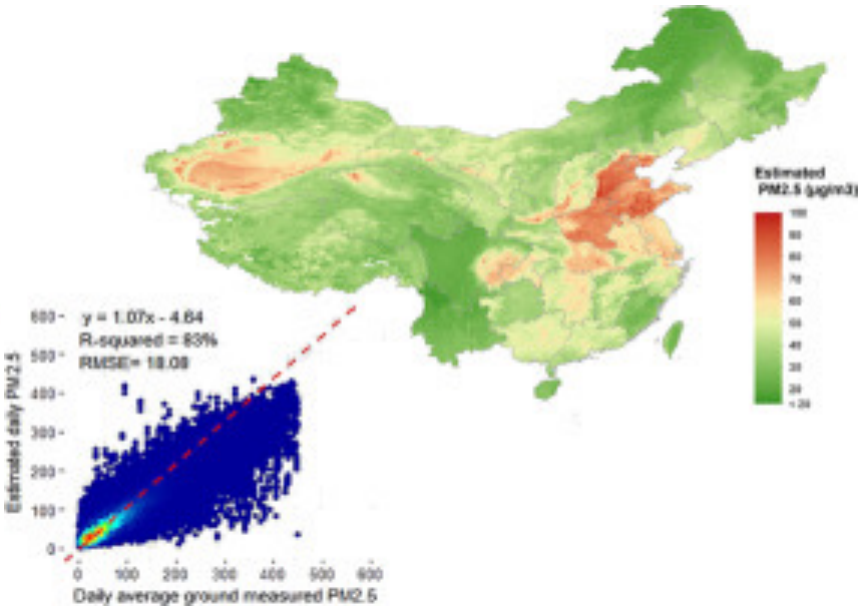
+ Add to Mendeley Share Cite

https://doi.org/10.1016/j.scitotenv.2018.04.251

Get rights and content

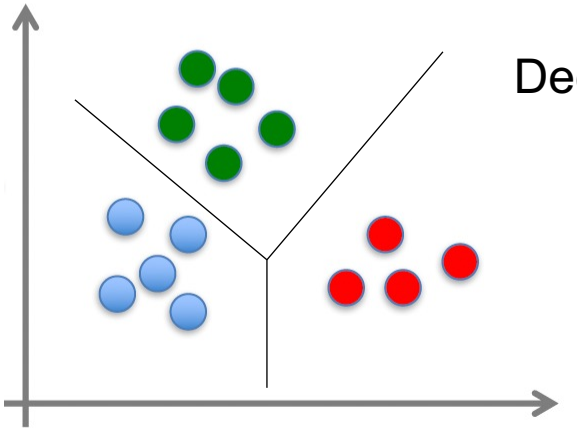
Highlights

- Historical exposure to $PM_{2.5}$ across China during 2005–2016 was estimated using AOD.
- The random forests model explained 83% of variability of ground measured $PM_{2.5}$.
- The machine learning method showed higher predictive ability than previous studies.



Supervised Learning: Classification

- Learning using labeled data
- Given (x1, y1), (x2, y2), ..., (xn, yn)
- Learn a function f(x) to predict y given x
  - y is categorical == classification



Decision Boundaries

Example

ELSEVIER

Remote Sensing of Environment
Volume 204, January 2018, Pages 509-523

Sentinel-2 cropland mapping using pixel-based and object-based time-weighted dynamic time warping analysis

Mariana Belgiu [a], Ovidiu Csillik [b, c]

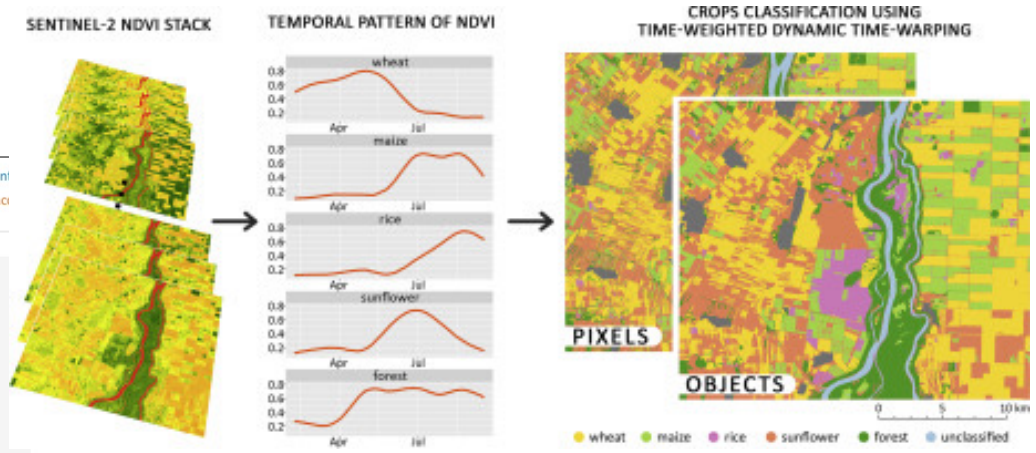
Show more

+ Add to Mendeley Share Cite

https://doi.org/10.1016/j.rse.2017.10.005

Highlights

- Time-weighted dynamic time warping (TWDTW) was evaluated for croplands mapping
- Object-based TWDTW performed better than pixel-based TWDTW
- TWDTW is robust in areas where inputs for training samples are limited
- Automatic time-series Sentinel-2 data segmentation produced satisfactory results
- Sentinel-2 proved to be a valuable satellite data source for croplands mapping



Example

IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, VOL. 8, NO. 6, JUNE 2015 2971

Fusion of Hyperspectral and LiDAR Remote Sensing Data Using Multiple Feature Learning

Mahdi Khodadadzadeh, *Student Member, IEEE*, Jun Li, *Member, IEEE*, Saurabh Prasad, *Senior Member, IEEE*, and Antonio Plaza, *Fellow, IEEE*

***Abstract*—Hyperspectral image classification has been an active topic of research. In recent years, it has been found that light detection and ranging (LiDAR) data provide a source of complementary information that can greatly assist in the classification of hyperspectral data, in particular when it is difficult to separate complex classes. This is because, in addition to the spatial and the spectral information provided by hyperspectral data, LiDAR can provide very valuable information about the height of the surveyed area that can help with the discrimination of classes and their separability. In the past, several efforts have been investigated for fusion of hyperspectral and LiDAR data, with some efforts driven by the morphological information that can be derived from both data sources. However, a main challenge for the learning approaches is how to exploit the information coming from multiple features. Specifically, it has been found that simple concatenation or stacking of features such as morphological attribute profiles (APs) may contain redundant information. In addition, a significant increase in the number of features may lead to very high-dimensional input features. This is in contrast with the limited number of training samples often available in remote-sensing applications, which may lead to the Hughes effect. In this work, we develop a new efficient strategy for fusion and classification of hyperspectral and LiDAR data. Our approach has been designed to integrate multiple types of features extracted from these data. An important characteristic of the presented approach is that it does not require any regularization parameters, so that different types of features can be efficiently exploited and integrated in a collaborative and flexible way. Our experimental results, conducted using a hyperspectral image and a LiDAR-derived digital surface model (DSM) collected over the University of Houston campus and the neighboring urban area, indicate that the proposed framework for multiple feature learning provides state-of-the-art classification results.**

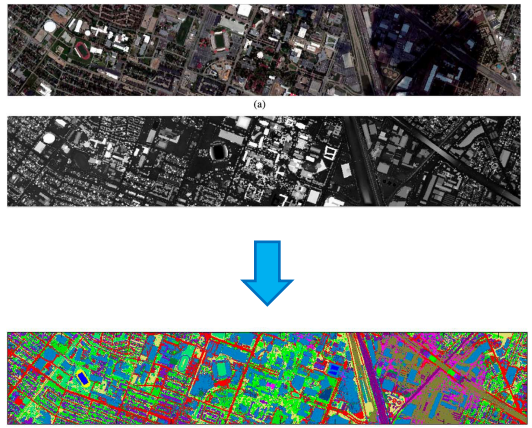
***Index Terms*—Digital surface model (DSM), hyperspectral, light detection and ranging (LiDAR), multiple feature learning.**

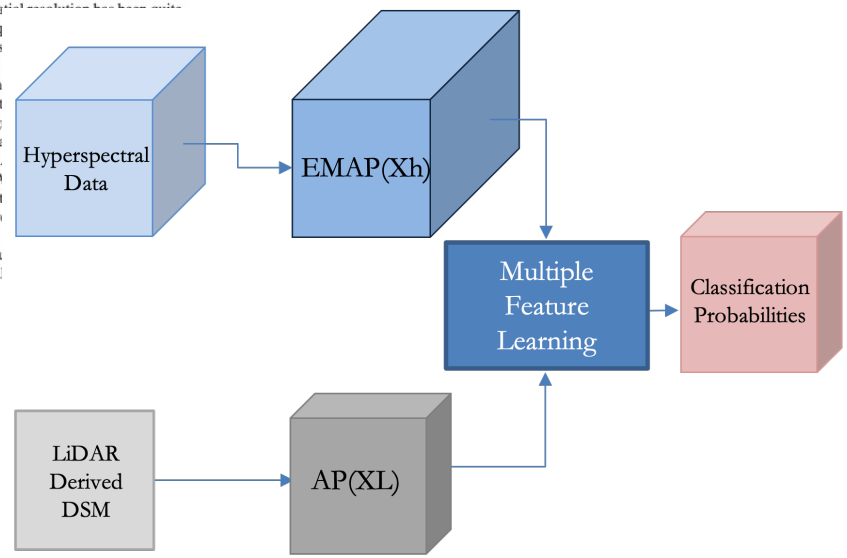
the Earth surface, using hundreds of (narrow) spectral bands typically covering the visible and near infra-red domains [1]. In hyperspectral imaging, also termed imaging spectroscopy [2], the sensor acquires a spectral vector with hundreds or thousands of elements from every pixel in a given scene. The result is the so-called hyperspectral image or hyperspectral data *cube*. It should be noted that hyperspectral images are spectrally smooth and spatially piece-wise smooth; this means that the values in neighboring locations and wavelengths are often highly correlated [3].

Hyperspectral image classification has been a very active area of research in recent years [4]. Given a set of observations (i.e., pixel vectors in a hyperspectral image), the goal of classification is to assign a unique label to each pixel vector so that it is well defined by a given class. The wider availability of hyperspectral data with high spa

important for classification techniq
the spatial resolution of the hypers
separate complex classes such as
ronments [4]. This aspect, togeth
or nonlinear) mixing happening at
cates the classification process sig
data coming from other sources ca
refine the results of classification.
light detection and ranging (LiDA
vide information about the height
LiDAR has been shown to be a v
classification purposes [7].

In the literature, many techniqu
fusion of hyperspectral and LiDA





Example

An Automatic Method for Subglacial Lake Detection in Ice Sheet Radar Sounder Data

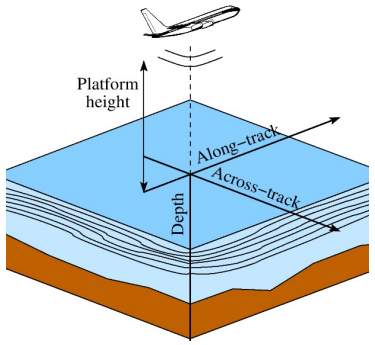
Ana-Maria Ilisei, *Member, IEEE*, Mahdi Khodadadzadeh, Adamo Ferro, and Lorenzo Bruzzone, *Fellow, IEEE*

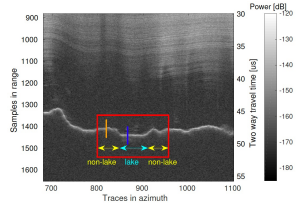
***Abstract*—During the past decades, radar sounder (RS) instruments have been effectively used to detect subglacial lakes (SLs). SLs appear as flat, smooth, and bright reflectors in RS radargrams. The visual interpretation has been the main approach to SL detection in radargrams. However, this approach is subjective and inappropriate for processing large amounts of radargrams. While the analysis of RS data for understanding the subglacial hydrology has recently received increased attention, the literature on the development of automatic methods specifically designed for SL detection is still limited. In order to fill this gap, in this paper, we propose a novel automatic technique for SL detection. The technique is made up of two steps: 1) feature extraction and 2) automatic detection. In the first step, we define and extract three families of features for discriminating between the lake and nonlake radar reflections. The features model locally the basal topography, the shape of the basal reflected waveforms, and the statistical properties of the basal signal. In the second step, we provide the features as input to a support vector machine classifier to perform the automatic SL detection. The proposed technique has been applied to radargrams acquired over two large regions in East Antarctica and Siple Coast. The obtained results, which are validated both quantitatively and qualitatively, confirm the robustness of the features and their capabilities to effectively characterize SLs. Moreover, they prove the potentiality of the method to process large amounts of radargrams and update the current SL inventory.**

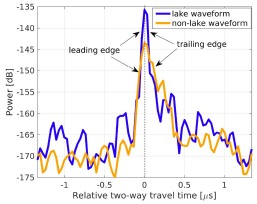
***Index Terms*—Automatic detection, ice sheet, radar sounder (RS), remote sensing, subglacial lakes (SLs).**

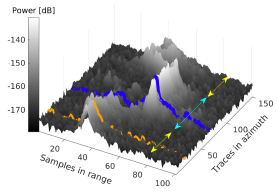
The fourth and most recent SL inventory in Antarctica, which is dated 2012 [16], reports 379 lakes, of which ≈30% have been detected by analyzing ice surface elevation changes in altimeter data [7]. Such lakes are also called active to highlight their observed dynamic behavior as a total or partial periodical discharge of their water [14]. The remaining ≈70% of inventoried Antarctic SLs have been detected in data acquired by airborne radar sounder (RS) instruments [15]. Recent analyses of RS data have also evidenced the presence of two SLs in Greenland [9] and a hypersaline SL complex in the Canadian Arctic [17]. Hereafter, we will focus our attention on the detection of SLs in RS data.

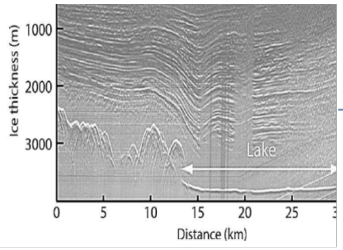
RSs are nadir-pointing instruments specifically designed for imaging the ice-sheet cross section. They emit low-frequency electromagnetic waves and measure the power reflected by subsurface mechanical and thermal discontinuities, from the surface to the basal interface below the RS platform along a predefined path. These measurements are recorded in 2-D matrices called radargrams. So far, for monitoring the Earth polar regions, glaciers, and ice caps, RSs have been operated in dedicated airborne campaigns, thus providing a large amount of radargrams. This offered the possibility to study the basal conditions and identify SLs over wide areas. In the future, the amount of RS data is expected to drastically increase,



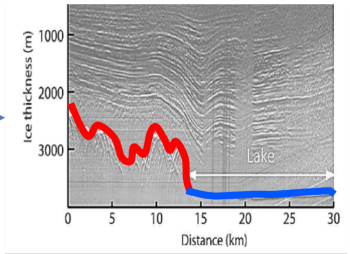






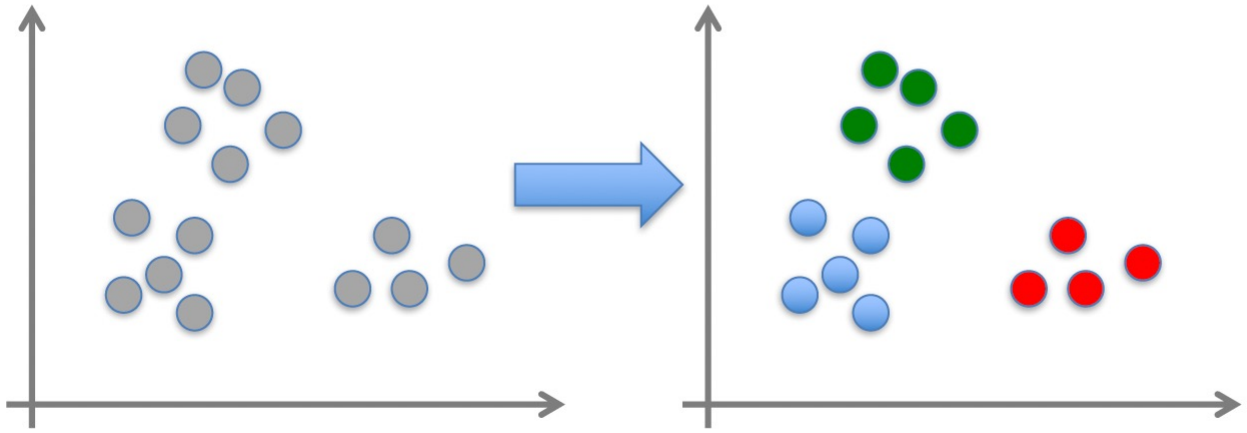


Automatic Basal Interface Detection → Feature Extraction → Classification



Unsupervised Learning - clustering

- Learning using unlabeled data
- Given x1,x2,...,xn (without labels)
- Output hidden structure behind the x's



Example

Review Article

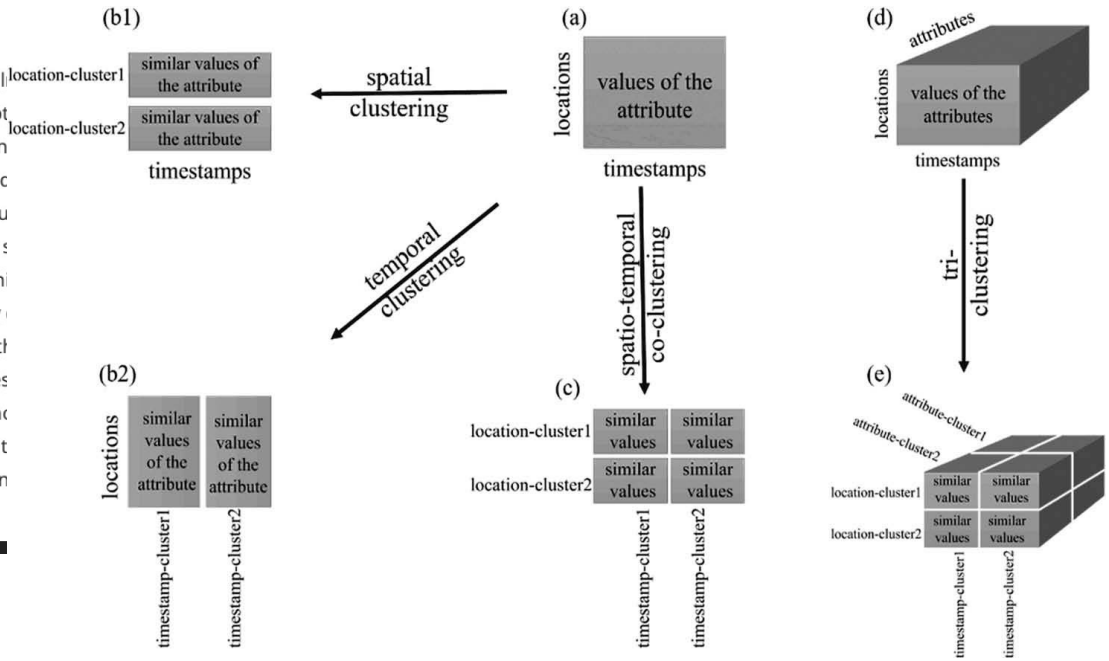
An overview of clustering methods for geo-referenced time series: from one-way clustering to co- and tri-clustering

Xiaojing Wu, **Changxiu Cheng**, Raul Zurita-Milla & Changqing Song

Pages 1822-1848 | Received 05 Jun 2019, Accepted 04 Feb 2020, Published online: 16 Feb 2020

ABSTRACT

Even though many studies have shown the useful
patterns, until now there is no systematic descript
(GTS) classified as one-way clustering, co-clusterin
suitable clustering method for a given dataset anc
overview of existing clustering methods for GTS, u
different methods to provide suggestions for the s
define a taxonomy of clustering-related geographi
using representative algorithms and a case study
more powerful in exploring complex patterns at th
way clustering and co-clustering methods yield les
However, the selection of the most suitable metho
computational complexity, and the availability of t
include novel clustering methods, thereby enablin
patterns.



Unsupervised Learning - clustering

IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, VOL. 47, NO. 8, AUGUST 2009 2973

Spectral–Spatial Classification of Hyperspectral Imagery Based on Partitional Clustering Techniques

Yuliya Tarabalka, *Student Member, IEEE*, Jón Atli Benediktsson, *Fellow, IEEE*, and Jocelyn Chanussot, *Senior Member, IEEE*

***Abstract*—A new spectral–spatial classification scheme for hyperspectral images is proposed. The method combines the results of a pixel wise support vector machine classification and the segmentation map obtained by partitional clustering using majority voting. The ISODATA algorithm and Gaussian mixture resolving techniques are used for image clustering. Experimental results are presented for two hyperspectral airborne images. The developed classification scheme improves the classification accuracies and provides classification maps with more homogeneous regions, when compared to pixel wise classification. The proposed method performs particularly well for classification of images with large spatial structures and when different classes have dissimilar spectral responses and a comparable number of pixels.**

***Index Terms*—Clustering, hyperspectral images, majority vote, segmentation, spectral–spatial classification.**

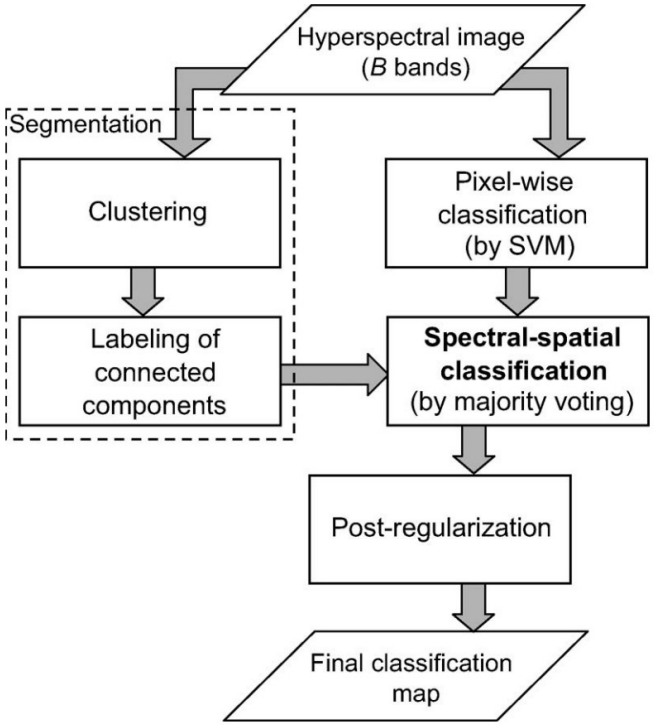
I. Introduction

THE ACCURATE classification of remote sensing images is an important task for many practical applications, such as precision agriculture, monitoring and management of the environment, and security and defense issues. The advent and growing availability of hyperspectral imagery, which records hundreds of spectral bands, has opened new possibilities in image analysis and classification. Examples of hyperspectral imaging systems are Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) [1], HYDICE [2], ARCHER [3], HyMap [4], and Hyperion [5]. They cover a range of 126–512 spectral channels, with the spatial resolution of 3–30 m per pixel. Thus, every pixel in a hyperspectral image contains values that correspond to the detailed spectrum of reflected light [6]. This rich spectral information in every spatial location increases the capability to distinguish different physical materials and objects, leading to the potential of a more accurate image classification.

An extensive literature is available on the classification of hyperspectral images where a wide range of pixel-level processing techniques is proposed, i.e., techniques that assign each pixel to one of the classes based on its spectral values. Maximum-likelihood or Bayesian estimation methods [7], decision trees [8], [9], neural networks [10]–[12], genetic algorithms [13], and kernel-based techniques [14], [15] have been investigated for this purpose. In particular, support vector machines (SVMs) have shown a good performance for classifying high-dimensional data when a limited number of training samples are available [14], [16], [17].

To improve classification results, the contextual information should be considered for incorporation into the classifiers. Spectral–spatial classification aims at assigning each image pixel to one class using a feature vector based on the following: 1) its own spectral value (the spectral information) and 2) information extracted from its neighborhood (referred to as the spatial information in the following). One of the approaches of spectral–spatial classification consists in including the information from the closest neighborhood to classify each pixel. These fixed-window-based methods that use morphological filtering [15], morphological leveling [18], [19], or Markov random fields [20] have shown improvements in classification accuracies compared to the pixel wise methods, when applied to hyperspectral images. However, the use of these methods raises the problem of scale selection, particularly when small or complex structures are present in the image.

Another approach to include spatial information in classification consists in performing image segmentation. *Segmentation* can be defined as an exhaustive partitioning of the input image into regions, each of which is considered to be homogeneous with respect to some criterion of interest (homogeneity criterion, e.g., intensity or texture) [21]. These regions form a segmentation map that can be used as spatial structures for a spectral–spatial classification.



Unsupervised Learning - clustering

Open Access Article

Hierarchical Sparse Subspace Clustering (HESSC): An Automatic Approach for Hyperspectral Image Analysis

by Kasra Rafiezadeh Shahi [1,*], Mahdi Khodadadzadeh [1], Laura Tusa [1], Pedram Ghamisi [1], Raimon Tolosana-Delgado [2] and Richard Gloaguen [1]

[1] Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Helmholtz Institute Freiberg for Resource Technology (HIF), Division of "Exploration Technology", Chemnitzer Str. 40, 09599 Freiberg, Germany

[2] HZDR-HIF, Division of "Modelling and Valuation", Chemnitzer Str. 40, 09599 Freiberg, Germany

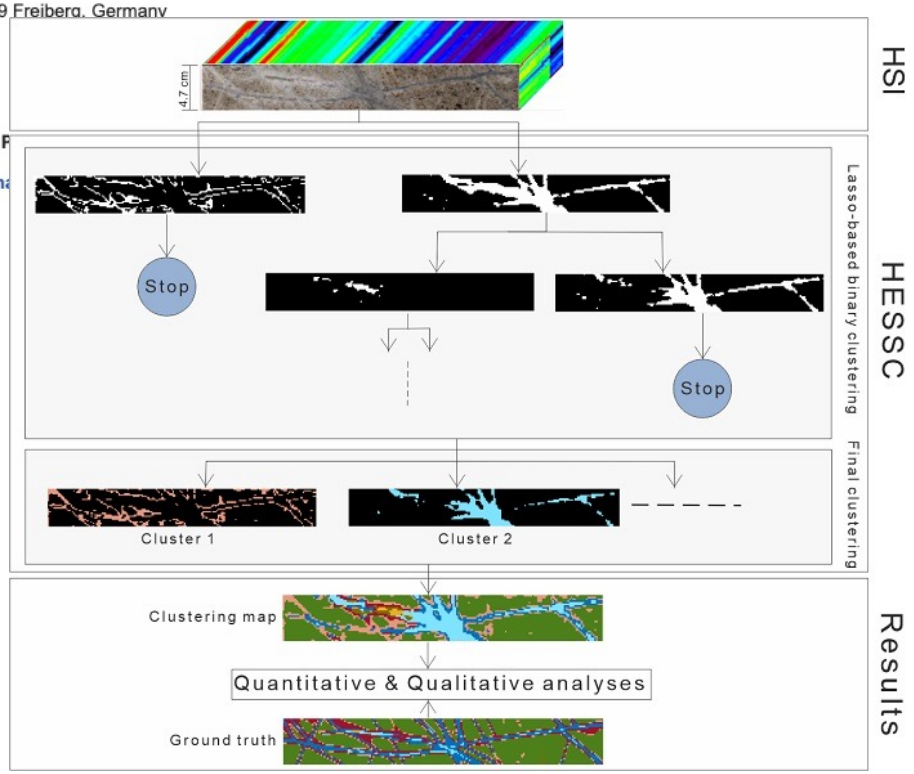
* Author to whom correspondence should be addressed.

Remote Sens. **2020**, *12*(15), 2421; https://doi.org/10.3390/rs12152421

Received: 20 May 2020 / Revised: 21 July 2020 / Accepted: 26 July 2020 / P

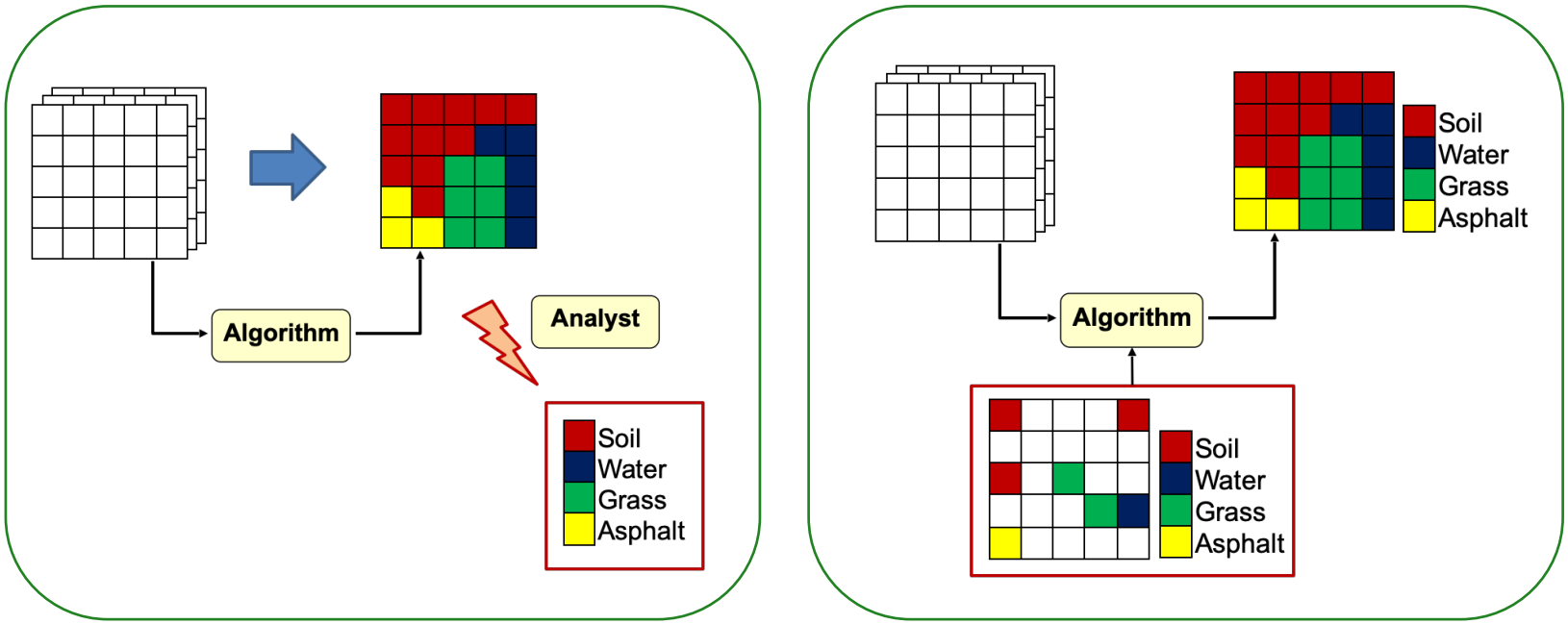
(This article belongs to the Special Issue Hyperspectral and Multispectral Ima

Download PDF | Browse Figures | Citation Export



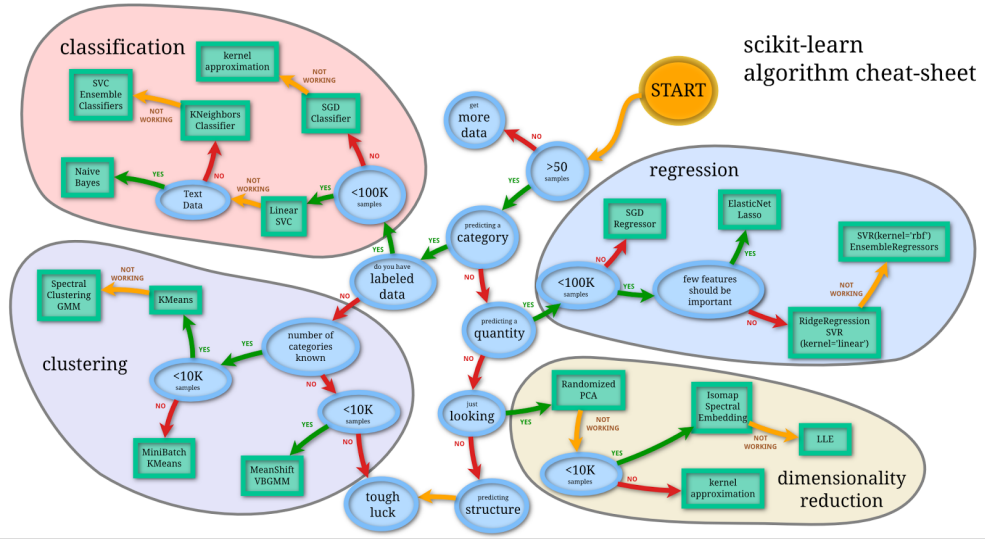
Question

- **supervised or unsupervised learning?**



Basics

- There are several supervised and unsupervised machine learning algorithms, and each takes a different approach to learning
- There is no best method! Finding the right algorithm is partly just trial and error
- First intuitive understanding of the problems, then identifying methods

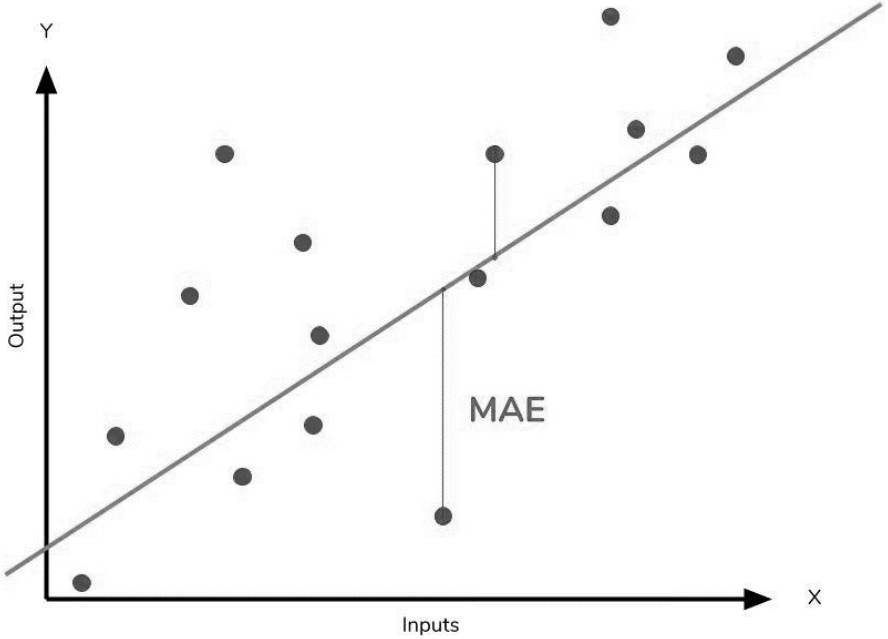


Performance Evaluation

- **Metrics for regression**
  - Mean absolute error
  - Mean squared error
  - Root MSE (RMSE)
  - R-squared
- **Variance & Bias**

- **Metrics for classification**
  - Accuracy score
  - Precision & Recall
  - Confusion matrix
  - F1-score

Performance Evaluation 1: metrics for regression



$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

Performance Evaluation 2: metrics for regression

Mean absolute error OR **mean squared error?**

- MSE is always bigger than MAE
- Each residual in MAE contributes **proportionally** to the total error
- Each residual in MSE contributes **quadratically** to the total error
- How you want your model to treat **outliers**, or extreme values, in your data?
  - *Is it a noise or a real phenomenon?*

$$MSE = \frac{1}{n}\Sigma\left(y - \widehat{y}\right)^2$$

Performance Evaluation 3: metrics for regression (c)

- **Root Mean Squared Error (RMSE)**
  - Squared root of MSE
  - Error metric has **similar units** → interpretation easier
  - Like MSE, RMSE is affected by outliers
  - Is a measure of how large your residuals are spread out

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

Performance Evaluation 4: metrics for regression

- **R-Squared** defines the degree to which the variance in the dependent variable (target or response) can be explained by the independent variable (features or predictors)

  - MSE, MAE, RMSE are not intuitive in interpretation
  - R-squared ranges from 0 to 1
  - Tells how good a model is **performing**

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

Performance Evaluation 5: metrics for classification

- A **confusion matrix** is an N X N matrix, where N is the number of classes being predicted.

	Predicted NO	Predicted YES
Actual NO	TN	FP
Actual YES	FN	TP

- **True Negatives (TN):** We predicted “NO”, and they are actually “NO”.
- **False Positives (FP):** We predicted “YES”, but they actually have “NO” labels.
- **False Negatives (FN):** We predicted “NO”, but they actually do have “YES” labels.
- **True Positives (TP):** We predicted “YES”, and they are actually “YES”.

Performance Evaluation 5: metrics for classification

- A **confusion matrix** is an N X N matrix, where N is the number of classes being predicted.

	Predicted NO	Predicted YES	
Actual NO	50	10	60
Actual YES	5	100	105

Accuracy: Overall, how often is the classifier correct?
(TP+TN)/total = (100+50)/165 = 0.91

Misclassification Rate: Overall, how often is it wrong?
(FP+FN)/total = (10+5)/165 = 0.09

Performance Evaluation 5: metrics for classification

- A **confusion matrix** is an N X N matrix, where N is the number of classes being predicted.

	Predicted NO	Predicted YES	
Actual NO	50	10	60
Actual YES	5	100	105

True Positive Rate (Sensitivity or Recall): When it's actually yes, how often does it predict yes? **TP/Actual Yes** = 100/105 = 0.95

True Negative Rate (Specificity): When it's actually no, how often does it predict no? **TN/Actual NO** = 50/60 = 0.83

Performance Evaluation 5: metrics for classification

- A **confusion matrix** is an N X N matrix, where N is the number of classes being predicted.

	Predicted NO	Predicted YES	
Actual NO	50	10	60
Actual YES	5	100	105

False Positive Rate: When it's actually no, how often does it predict yes?
FP/Actual NO = 10/60 = 0.17

Precision: When it predicts yes, how often is it correct?
TP/predicted yes = 100/110 = 0.91

Performance Evaluation 6: metrics for classification

- A **confusion matrix** is an N X N matrix, where N is the number of classes being predicted.

	Predicted NO	Predicted YES	
Actual NO	0	60	60
Actual YES	0	105	105

True Positive Rate (Sensitivity or Recall): 105/105 = 1
Precision: 105/165 = 0.64

$$F_1\text{–score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

Performance Evaluation 7: metrics for classification

- **Confusion matrix** for multi-class classification

	Predicted class 1		…	Predicted class n
Actual class 1	C_{11}	C_{12}	…	C_{1n}
Actual class 2	C_{21}			
	…			
Actual class n	C_{n1}			C_{nn}

Overall Accuracy:
$(C_{11} + C_{22} + \ldots + C_{nn})$/total

Class "i" Accuracy:
(C_{ii})/Actual class "i"

Average (Class) Accuracy:
Average of all the class accuracies

Performance Evaluation 7: metrics for classification

- **Confusion matrix** for multi-class classification

	Predicted class 1		…	Predicted class n
Actual class 1	C_{11}	C_{12}	…	C_{1n}
Actual class 2	C_{21}			
…	…			
Actual class n	C_{n1}			C_{nn}

Kappa

$$K = \frac{c \times s - \sum_k^K p_k \times t_k}{s^2 - \sum_k^K p_k \times t_k}$$

$c = \sum_k^K C_{kk}$ the total number of elements correctly predicted

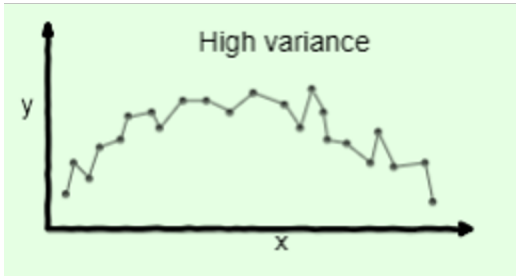
$s = \sum_i^K \sum_j^K C_{ij}$ the total number of elements

$p_k = \sum_i^K C_{ki}$ the number of times that class k was predicted (column total)

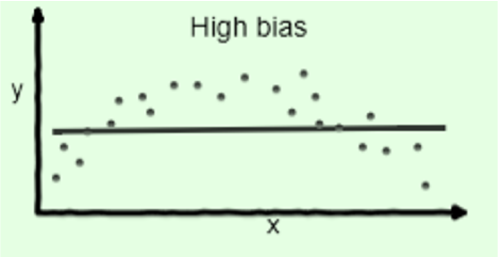
$t_k = \sum_i^K C_{ik}$ the number of times that class k truly occurs (row total)

Grandini, Margherita, Enrico Bagli, and Giorgio Visani. "Metrics for multi-class classification: an overview." arXiv preprint arXiv:2008.05756 (2020).

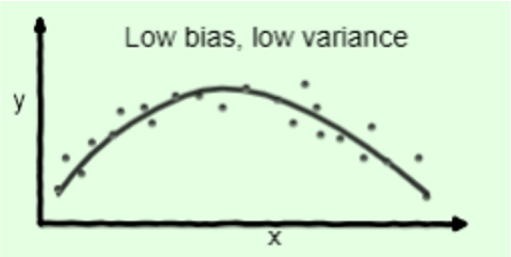
Variance & Bias



overfitting → Models are somewhat accurate but inconsistent on averages. A small change in the data can cause a large error



underfitting → Models are consistent but inaccurate on average



Good balance → Trade off between variance and bias

Variance & Bias

Is there a way to find when we have a high bias or a high variance?

- *High Bias*
  - High training error
  - Validation error/ test error ≈ training error
- *High Variance*
  - → Low training error
  - → High validation error/ test error

How do we fix high bias or high variance in the data set?

- *High Bias*
  - → More input features
  - → More complex models
- *High Variance*
  - → More training data
  - → Reduce input features

Basics

- ***Question 3:*** *A researcher possesses datasets to estimate the visits of Van Gogh Museum in Amsterdam. She has the number of visits over the last 200 days. Also, she has a volunteered geographical information dataset to be used as a proxy on the number of people at each location and each date. These are geolocated twitter messages at a buffer distance of 1 km around the museum. There are historical twitter data of the last 200 days and the researcher continues streaming data on a daily basis.*
  - Is this a regression or classification task?
  - What is the size of the training dataset?
  - What is the independent variable?