Chapter 10
Data Analytics and Machine Learning

Paula Järvinen, Pekka Siltanen, and Amit Kirschenbaum

Abstract In this chapter we give an introduction to data analytics and machine learning technologies, as well as some examples of technologies used in the DataBio project. We start with a short intdroduction of basic concepts. We then describe how data analytics and machine learning markets have evolved. Next, we describe some basic technologies in the area. Finally, we describe how data analytics and machine learning were used in selected pilot cases of the DataBio project.

10.1 Introduction

The goal of data analytics is to examine large quantities of data with the purpose of drawing conclusions about the data. Several techniques can be employed, each using similar methods but having a slightly different focus. The methods include, e.g., statistics, data mining, and machine learning (Fig. 10.1).

Data mining is defined as "a science of extracting useful information from large data sets or databases" [1]. Machine learning is "programming computers to optimize a performance criterion using example data or past experience" [2]. Sometimes the division between machine learning and data mining is done based on data sets. Data mining is focused on analyzing large databases, whereas in machine learning the focus is on learning patterns from data. The roots of data analysis are in statistics. The development of computers and their ability to store and manage large amounts of data has made possible large-scale statistical computation and has launched the development of new methods that would be tedious to perform manually.

A recent area of data analysis is visual data mining. Information visualization, data mining, and user interaction have evolved as separate fields in the past, but since the turn of the 2000s have become increasingly integrated as visual data mining.

P. Järvinen · P. Siltanen (✉)
VTT Technical Research Centre of Finland Ltd., Espoo, Finland
e-mail: Pekka.Siltanen@vtt.fi

A. Kirschenbaum
Institute for Applied Informatics (InfAI), University of Leipzig, Goerdelerring 9, 04109 Leipzig, Germany

C. Södergård et al. (eds.), *Big Data in Bioeconomy*,
https://doi.org/10.1007/978-3-030-71069-9_10

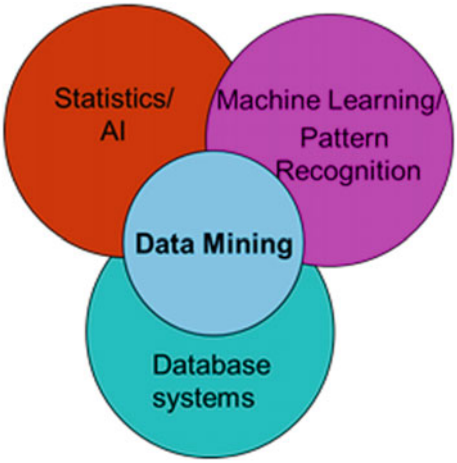

Fig. 10.1 Data analysis techniques [1]

The idea of visual data mining first emerged in 1999 when Wong [3] argued that rather than using visual data exploration and analytical mining algorithms as separate tools, a stronger data mining strategy would be to couple the visualizations and analytical processes into one data mining tool. Many data mining techniques involve mathematical steps that require user intervention, and visualization could support these processes. Visual data mining is not just about using visualization to exploiting data, it is an analytical mining process in which visualizations play a major role [4].

Artificial intelligence (AI) can be defined as "a system's ability to correctly interpret external data, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation [5]."

Machine learning has been used since the 1950s by researchers in order to analyze and extract information from data. It has only been during the last decade with the rise of the generalized usage of the graphics processing units (GPUs) that enabled the true development of neural networks and in particular what is nowadays referred to as deep learning [6]. This newly found computational power gave rise to methods that are capable of solving complex, real-world problems. The capacity of modern computers not only allows for computationally intensive methods, but also facilitates the analysis of huge amounts of data, the so-called big data, in a scale that was previously intractable. In contrast to previous methods, deep learning uses multiple layers of neural networks to build architectures capable of performing a specific task, such as classification, segmentation, detection, prediction, and generation of data.

Deep learning is capable of discovering correlations in the data without the need of handcrafted features. The lack of heuristics together with the abundance of computational resources makes deep learning methods ideally suited for handling big data problems. Further to that, machine learning offers the possibility for lifelong learning where the system is capable of adapting to changing conditions. While machine learning is often portrayed as a replacement for human intelligence, it is only a tool for digitalizing human expertise into a computer model. This model is only as good as the information humans supplied it with.

10.2 Market

Data analysis has been studied intensively, and numerous algorithms exist. It has applications in different business, science, and social science domains. A wide range of tools and commercial applications is available, some of which are highly competitive in markets, such as customer relationship management (CRM). There are also several statistics programs and packages available, both for casual users and specialists (Excel, SAS, SPS, R).

Big data analysis solutions can be classified into two categories: "Data Discovery and Visualization" and "Advanced Analytics" [7]). Data discovery and visualization solutions integrate and transform big data sources using data mining algorithms to find insights into business use. Advanced analytics solutions are focused on building use case-specific predictive or descriptive solutions using advanced modeling techniques, such as deep learning or advanced statistical methods.

Frost and Sullivan estimate big data revenue at 2017 of $8.54 billion [7]. The revenue is expected to reach $40.65 billion in 2023. The market is expected to grow at a steady rate, as data discovery and visualization are expected to become more mainstream over this period and advanced analytics is expected to see more real-life use cases [7]. North America is expected to continue to be the largest market contributor, followed by Western Europe, having similar growth path.

Biggest user of data analytics techniques is business and finance, followed by governance and integrity (public sector), both over 15% of the market. In Frost and Sullivan estimations, bio-economy falls into the category of "Others," which in total covers 7.7% of the market.

According to the Zion Market Research [8], global machine learning market was valued at around USD 1.58 billion in 2017 and is expected to reach approximately USD 20.83 billion in 2024, growing at a compound annual growth rate (CAGR) of 44.06% between 2017 and 2024. Artificial intelligence experts have projected their idea that by 2050 all the intellectual tasks performed by the humans can be accomplished by the artificial intelligence technology. Some of the top applications of machine learning are financial services, virtual personal assistants, health care, government, marketing and sales, transportation, oil and gas, manufacturing, bioinformatics, computational anatomy, and more. The artificial intelligence (AI) market in agriculture is expected to register a CAGR of over 21.52%, during the forecast period of 2019–2024, offering services for the management of the crops yield, species breeding, disease detection.

Geographically, machine learning market is segmented into North America, Asia Pacific, Europe, Latin America, and Middle East and Africa. North America is predicted to govern the market in forecast period because of developed countries and their major focus on innovative technologies obtained from R&D sector. Asia-Pacific region is predicted to grow at the highest CAGR in forecast period due to increasing awareness regarding business productivity. In Asia, region vendors are offering competent machine learning proficiency due to which it is the highest potential region for the market. Moreover in Europe, the world-class research facilities, the emerging start-up culture, and the innovation and commercialization of machine intelligence technologies are stimulating the machine intelligence market. Among all regions, Europe has the largest share of intra-regional data flow. This, together with the machine learning technologies, is boosting the market in Europe.

10.3 Technology

10.3.1 Data Analysis Process

Data analysis is an iterative process starting with selecting the target data from the raw material and preprocessing and transforming it into a suitable form (Fig. 10.2). Data analysis uses several data types: database records, matrix data, documents, graphs, links, transaction data, transaction sequences, DNA sequence data, whole genome information, and spatiotemporal data. The quality of data may often cause problems. The data can contain noise, there may be missing values and duplicate data, and thus

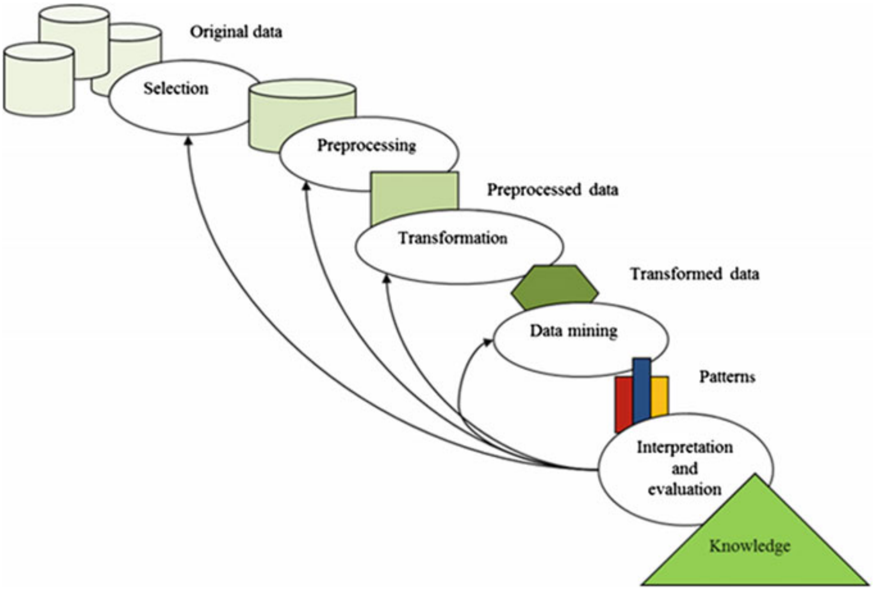

Fig. 10.2 Data mining process. Figure adapted from [1]

data cleaning phase is required before using the data. Other kinds of preprocessing may also be required, such as data aggregation, sampling, dimensionality reduction, subset selection, feature creation, and attribute transformation [1].

Next, the data is run through a data mining algorithm that creates patterns from the data. The user interprets and evaluates the results and starts a new iteration with possible modifications to the raw data, algorithm, and algorithm parameters.

10.3.2 Statistical Methods

Statistical methods are used for data exploration to gain a better understanding of the characteristics of data [1]. The central methods include, e.g., summary statistics, correlations, and visualizations. Summary statistics are numbers that summarize properties of the data. Amar et al. [9] have classified the statistical methods as

(1) computer-derived values: average, median, count, more complex values,
(2) finding extremum: finding data cases having the highest and lowest value of a defined attribute,
(3) determining range: finding a span of values of an attribute of data cases, and
(4) characterizing distributions: creating a distribution of a set of data cases with a quantitative attribute, e.g., to understand "normality." The visual methods utilize humans' ability to recognize patterns. Single variables are expressed in visual form, for instance as histograms and line charts.

Correlation is a basic statistical method of studying two variables. The prevailing method is the calculation of the Pearson correlation coefficient (r), where the correlation between two variables, x_i and y_i is calculated with the formula:

$$r = \sum_{i=0}^{n} \frac{(x_i - x)(y_i - y)}{nS_xS_y}$$

where n is the number of observation pairs, and S_x, S_y are the standard deviations, and x and y the means of the variables x_i and y_i. The correlation produces positive or negative values within the range -1 to 1. If the result is zero, there is no correlation between the variables. Values -1 and 1 indicate complete linear dependence between the variables, either negative or positive. Often the square of the correlation coefficient R^2 (also known as the coefficient of determination) is calculated. This value ranges from 0 to 1 and indicates how much one variable explains the variance of the other and is often expressed as a percentage. For instance, if R^2 is 0.32, 32% of the variance of a variable is explained by the other.

Correlations are visualized in the form of scatterplots. Exploration methods for higher dimensions use projections of data on a two-dimensional plane. These are called dimension reduction methods. They include principal component analysis

(PCA) and multidimensional scaling, as well as auto-encoders for neural networks. The result of PCA can be visualized as a two-dimensional plot.

10.3.3 Data mining

The goal of data mining is to extract useful information from large data sets [10]. Data mining can be categorized into different kind of tasks, corresponding the objectives of analysis: exploratory data analysis, descriptive modeling, predictive modeling, and discovering patterns and rules.

Exploratory data analysis (EDA) explores data without clear ideas of the findings. Visualization is effective EDA techniques, especially with relatively small and low-dimensional data sets. Bar charts, boxplots, histograms, and density plots are applicable with single variable data, scatterplots with two variable data. With multidimensional data, dimension reduction methods, such as principal component analysis, (PCA) are used. They produce informative low-dimensional projections of data that can be visualized in two-dimensional space.

The goal of descriptive methods is to describe the data. The methods include density estimation, clustering and segmentation, and models describing the relationships between variables. Clustering looks for groups of objects such that the objects in a group are similar (or related) to one another and different from (or unrelated to) the objects in other groups. The similarity of objects is defined based on similarity (or distance) measures. Euclidean distance can be used if attributes are continuous; otherwise, problem-specific measures are needed. Clustering has been an active research topic, and lots of algorithms are available. Algorithms include K-means clustering and its variants, hierarchical clustering, agglomerative clustering, and density-based clustering. Market segmentation is an application of clustering.

The purpose of predictive modeling is to build models that predict the value of one variable from the known values of other variables [10]. The predicted objects are predefined. Regression and classification are two much used predictive methods. Regression predicts a value of a continuous variable based on other variables using linear or nonlinear models [1]. Linear regression is easy to visualize, often shown as a line on a scatterplot diagram. The area is studied extensively and has its origins in statistics. It has various uses, both in commerce and science. Application examples include predicting sales based on advertising expenditure, stock markets, or wind as a function of temperature or humidity. Classification creates a model for a class attribute as a function of the values of other attributes. Unseen records are then assigned to the class. Models in both methods are developed with a learning data set, and the precision and accuracy of the models are evaluated with a test set. Several techniques have been developed including decision trees, Bayesian methods, rule-based classifiers, and neural networks. Classification is a much used method, and commercial applications are also available. Examples include classification of credit card transactions as legitimate or fraudulent, classification of e-mails as spam, or classification of news stories as finance, weather, entertainment, or sports [1].

Discovering patterns and rules involves finding combinations of items that occur frequently in databases. Sequential pattern discovery finds rules that predict strong sequential dependencies among different events. Association rule mining involves the prediction of occurrences of an item based on occurrences of other items. It produces dependency rules such as "buyers of milk and diapers are likely to buy beer." One special case of pattern discovery is anomaly detection. Anomalies are observations whose characteristics differ significantly from the normal profile. Methods of anomaly detection look for sets of data points that are considerably different from the remainder of the data. The methods build a profile of "normal" behavior and detect significant deviations from it. The profile can be patterns or summary statistics for the overall population. Types of anomaly detection schemes can be graphical-based, statistical-based, distance-based, or model-based. Credit card fraud detection, telecommunication fraud detection, network intrusion detection, and fault detection are examples of application areas [1].

10.3.4 Machine Learning

In machine learning, the idea is to learn things from data. The approach is to create mathematical models and adjust model parameters with the help of data until the model matches best the modeled phenomena. Machine learning utilizes theories from statistics combined with computer algorithms [2]. It has a strong overlap with data mining. Machine learning is focused on learning patterns from data whereas in data mining focus is on analyzing large databases. Machine learning methods can be divided into unsupervised and supervised learning. In unsupervised learning, there is only input data available, and the aim is to find patterns in data. In supervised learning, there is prior knowledge of the phenomena available in addition to the input data. Clustering belongs to unsupervised methods, whereas classification, regression, and bayesian methods are supervised. Another division is parametric and nonparametric methods. The parametric methods assume that the data is drawn from some probability distribution known before, and the model is created by estimating model parameters from data. Regression and classification methods are parametric methods. The nonparametric methods do not make such assumptions of the data but are based on finding similarities. They divide the input space into local regions, defined by a distance measure. Decision trees belong to nonparametric methods.

As in data mining, model validation is an important issue in machine learning. Input data is divided into learning part and validation part. The model is developed with the learning part and validated with validation part. Measures of the validity are model accuracy and precision.

Neural networks are a specific set of algorithms inspired by biological neural networks. The current deep neural networks (deep learning) work well in problems such as computer vision, speech recognition, and natural language processing. Currently, there are many available open-source frameworks, TensorFlow, PyTorch, Cafe, etc., that can be used for developing neural network models. These include highly optimized code that can be used for both training and using a model and thus greatly simplify the development process. Architectures for building models for specific tasks get published constantly in conferences and journals very often in an open manner. This has given rise to a variety of applications escaping the confines of academic research and reaching directly the market.

10.4 Experiences in DataBio

10.4.1 Data Analytics in Agriculture

10.4.1.1 Classification of Land Covering

This section describes the use of deep learning techniques for Earth observation data in the agriculture pilots in Part V of this book. The ongoing advancements in deep learning, and exemplary results obtained for different problems using spatiotemporal satellite images, have made deep neural networks quite popular for analyzing Earth observation data. The aim of the pilot was to design a pipeline based on deep neural networks to classify land cover using available satellite images from Sentinel-2A satellite. Initially, an investigation was done using only images and not taking advantage of the temporal nature of the signal. The results of this approach were not satisfying as the spatial information was not sufficient to differentiate crops with an adequate accuracy. For this reason, a new pipeline based on spatiotemporal data was designed. The new pipeline consisted of two steps: clean available training data, and then use this cleaned data for training crop classifiers. For the first step, instead of using traditional methods (based on data specific heuristics and handcrafted filters) to clean data, an RNN-based auto-encoder was trained to remove unreliable data. The encoder and decoder consisted of recurrent neural network (RNN) layers with long short-term memory (LSTM) cells [11]. The encoder learns the representations in latent space from the time series of pixels in crop parcels, while the decoder tries to reconstruct the time series. The representations are clustered in the latent space using K-means clustering. It is expected that most of the pixels will form one huge cluster while the outlier pixels will be away from this cluster. In this way, the parcels with clean pixels are selected and further used for training a pixel-level classifier network (inspired from [12] and [13]) for individual crops. Instead of training a neural network from scratch, the encoder part of the auto-encoder is used as initial layers of the classifier network. The pre-trained encoder network is appended with a dense layer and fine-tuned for the classification task. The classifier network produced a probability of being a particular crop for each input pixel. The details of the training for complete pipeline and obtained results can be found in [14].

The classifiers are trained for wheat, maize, and legumes for the data from regions in Greece, provided by NEUROPUBLIC for year 2016. Further, the classifiers are integrated in the DataBio online platform developed by Fraunhofer, where the probability of each pixel in the selected parcel belonging to a certain crop type can be obtained. This technology allows the user to identify the crop grown in a given area by using corresponding satellite imagery.

The presented pipeline shows the significance of data verification and provides an efficient way to create models by optimizing the efforts and the time of both engineers and experts. The data cleaning step done in an unsupervised manner increases the reliability of the data. An expert can further verify and refine these data groups by verifying only the boundary cases in the cleaning step. In this manner, the effort of the expert is optimized by focusing on targeted areas.

Additionally, the cleaning and classification done using time series of pixels (instead of parcels) are advantageous to us due to the following reasons:

- Lack of availability of fully labeled satellite images
- Due to the complexity of drawing parcel boundaries in low resolution satellite images, the pixel-level cleaning allows us to remove pixels corresponding to nearby road, lakes, etc., from the crop parcels.
- Instead of using image patches, the use of time series of individual pixel values for classification avoids influence of nearby pixels.
- The classification obtained at a pixel level enables sub-parcel level analysis which is very helpful in applications like damage assessment.

Although the presented pipeline performed well for the available data set, the results may not be as good for the following cases:

- The auto-encoder and the classifier both assume the variation of time series with in a crop type is low. In case of huge variation, we may need to subdivide the crop type for this approach to work.
- The training data corresponding to selected region in Greece creates the model depending on the temporal behavior of crops in that locality. This model may not work for the same crop having significantly different behavior in different regions across the world.
- The current model may not work well for data from other years as it may have a bias toward year 2016.

Lessons Learned

- Classification of crop using spatial data only does not have adequate performance and important information is in the temporal dimension.
- Some crops have similar varieties and can be covered with a common model. There are crops though whose varieties are very different, this approach would probably not succeed, and separate models for each sub-variety would be required.
- To develop models that can work for multiple years, implying different weather conditions throughout the season, would require further work on combining data measured on different dates each year. Similarly, multi-regional and global models would require much more data, as they would need to abstract the variation caused not only from local climates but also from a variety of different soils.

While this solution can benefit from further developments, it has the potential to form the baseline for methods targeting global scale satellite image analysis. The proposed approach for detection and classification of vegetation types operates on the sub-parcel level and is robust to noise in both the data and the labels.

10.4.1.2 Crop Detection and Monitoring

The free and open availability of Earth observation data is bringing land monitoring to a completely new level, offering a wide range of opportunities, particularly suited for agricultural purposes, from local to regional and global scale, in order to enhance the implementation of Common Agricultural Policy (CAP).

Terrasigna proposes an in-house developed fuzzy-based technique for crop detection and monitoring in Romania, based on combined free and open Sentinel-2 and Landsat-8 Earth observation data. The general methodology is based on the comparison between real crop behavior and the expected trends for each crop typology. It involves image processing, data mining, and machine learning techniques and is based on different categories of input data: Sentinel-2 and Landsat-8 SITS covering the time period of interest, farmers' declarations of intention with respect to crops types, as well as in situ/field data.

The machine learning technique used is an original one, developed taking into account the particularities of the CAP-monitoring process. The fuzzy approach allowed the use of all available scenes, provided they were not completely contaminated with clouds and shadows. The mixed time series, consisting of S2 and L8 scenes, are accompanied by relevance masks, which act as weights in the final fuzzy extraction process (i.e., drawing a firm conclusion using a series of vague and incomplete information). The strictly statistical character of the algorithm, which does not use phenology information or the intervention of a specialist with agronomic competences, makes the technique universal, being able to adapt to other regions and types of cultures, without difficulty.

The processing chain involves a series of well-defined steps:

- Image preprocessing (numerical enhancements for Sentinel-2 and Landsat-8 scenes, ingestion of external data, and clouds and shadows masking);
- Individual scene classification;
- Deriving crop probability maps at scene level;
- In the end, time series analysis allows the generation of overall crop probability maps and derived products.

The main goal of the approach within the DataBio project framework was to provide services in support to the National and Local Paying Agencies and the authorized collection offices for a more accurate and complete farm compliance evaluation—control of the farmers' declarations related to the obligation introduced by the current Common Agriculture Policy (CAP). The system produced three main types of results, all provided at a 10-meters spatial resolution as follows:

- Crop mask maps, which are pixel level maps, identifying some of the most important crop types;
- Parcel use maps, which are object-based maps, showing the most probable type of crop at plot level;
- Crop inadvertencies maps, which can be both pixel-based and object-based maps, revealing the areas for which the declared type of crop included in the LPIS appears to be different from the identified one. The pixel-based analysis states whether pixel values correspond or are different from typical spectral values of the declared crop types, whereas the object-based analysis reveals the plots for which the declared type of crop appears to be different from the one identified based on satellite imagery, based on a specific threshold.

Lessons Learned

The technology developed by Terrasigna is able to recognize a large number of crops families, of the order of tens. For Romania, it addressed the first most cultivated 32 crops families, which together cover more than 97% of the agricultural land. In 2018, the validation of results for a full agricultural season (full phonological cycle) against independent sources revealed promising results, with an accuracy higher than 95% for more than 10 crop types. The performance is quite uniform reported to parcels size and remains high even for parcels smaller than 1 ha. The highly automated proposed approach allows the performing of big data analytics to various crop indicators, being reliable, cost-, and time-saving. It leads to a more complete and efficient management of EU subsidies, strongly enhancing their procedure for combating non-compliant behaviors.

The most serious problems that had to be solved and that served as lessons were as follows:

- The use of data S2 and L8 together—which have a different format and resolution;
- Correction of the geographical positioning (georeferencing) automatically—which deeply affects the quality of the classification for small or narrow plots;
- Selecting the areas of interest from each image—which are not, as it might seem, the areas uncontaminated by clouds and shadows, but the areas where there is vegetal “activity”;
- The construction of an algorithm that takes into account the matrix of semantic confusion between cultures—which required finding the natural classes of cultures that can be followed simultaneously, without serious mutual confusion.

Geospatial services together with Copernicus data can provide a really powerful tool for monitoring agricultural dynamics. The end users, the National Paying Agencies, are able to benefit from the modern and effective near real-time service, based on the principles of sustainable agriculture and saving effort both in terms of costs and time. A continuous agricultural monitoring service based on the processing and analysis of Copernicus satellite imagery time series is not just a CAP compliance tool, but can also offer a great range of supplementary information for both public authorities and citizens.

The developed technique is replicable at any scale level and can be implemented for any other area of interest.

10.4.1.3 Farm Weather Insurance Assessment

Trying to identify the parameters (weather or soil related) with the dominant impact on the crop yield such as normalized difference vegetation index (NDVI) measurements, the following approach is considered. For the first phase of this analysis k-prototypes, clustering algorithm was applied for the profile building of the parcels. Using satellite, meteorological measurements and soil characteristics are aggregated on the level of one or two months considering a full growing season. The k-prototypes algorithm is based on the k-means paradigm but removes the numeric data limitation while preserving its efficiency [15]. After this phase of analysis, each one of the parcel linear regression models [16] is trained considering only the data that belongs to this cluster. In that way after the clustering procedure, we can use historical data of a parcel in order to identify in which cluster it belongs and make predictions for the NDVI values of an upcoming period using the corresponding linear regression model.

Lessons Learned

The main challenge of this approach is that the clustering analysis cannot work with missing values, so each one of the parcels is required measurements for the same months, otherwise the parcel must be excluded from the analysis. Another challenge is the sparsity of satellite data due to weather issues (e.g., cloudy days) making it difficult to create a "complete" or usable by the machine learning algorithms data set in terms of meteorological, satellite, and soil information for the same dates. In order to deal with that issue, interpolation and aggregation of the data were applied.

10.4.1.4 Crop Disease Detection Using Satellite Images

Automated crop monitoring is an essential aspect of smart agriculture, as it allows to improve yield estimation while reducing costs and environmental imprint. We conducted a study to forecast diseases in sorghum using remote sensing via satellite imagery as a proxy for crop health. Our method uses images from Sentinel-2 satellites, which regularly provide multispectral images for land monitoring. Images of a sorghum field with infected parts taken under different weather conditions, as captured by Sentinel-2 satellites, served as our training data [17]. We use the observation that there is a strong correlation between the physiological status of a plant and its chlorophyll content, i.e., diseases have a negative influence on the chlorophyll level [18] and derive NDVI from the recorded satellite images. NDVI is an indicator for vegetation vitality which measures the difference between near-infrared light that vegetation strongly reflects, and red, which vegetation absorbs. Healthy plants, that is, with a higher level of chlorophyll, reflect more near-infrared and green light compared to other wavelengths and absorb more red light.

Lessons Learned

Accurate data on disease outbreaks in the agricultural sector is usually not publicly available, e.g., due to data protection. This posed us the challenge of a small training set, which may lead to overfitting, a general problem in training machine learning methods. To overcome this, we perform data augmentation, i.e., artificially expand the training data set, to improve the ability of the learned model to generalize. Data augmentation is performed by small changes in data, in this case—image manipulation. Such operators include rotations, reflections, random excerpts, image zooming, or combinations thereof [19].

Mask region-based convolutional network (R-CNN) [20] is then used to train a model that determines which areas are infected. Mask R-CNN is a convolutional neural network that performs instance segmentation, i.e., identifies outlines of objects on a pixel level. In our case, the segmentation would be according to the NDVI values. This method showed great potential for the task at hand, and the model achieved mean average precision very close to 1.

10.4.2 Data Analytics in Fishery

10.4.2.1 Reducing Energy Consumption of Vessels in the Fishery Domain

This study aims at reducing the ecological and economical costs of fishery vessels, by optimizing their route and speed and thus decreasing fuel oil consumption. This process requires analysis of many observations collected over time. We collected thousands of observations per day from two boats for three years, where each observation involves dozens of features, e.g., speed and angle of wind, engine load, speed of the vessel, and, of course, and fuel consumption. The first step was creating predictive models for consumption of fuel oil per nautical mile. To this end, we compared two modeling techniques: the extreme gradient boosting framework XGBoost [21] and polynomial regression [22] and opted for the latter as it provides better results. We then explore two use cases: One considers calculating an optimal route (TSP) connecting several points; and the other, selecting a single sailing destination which optimizes the energy consumption. In both use cases, varying travel distances and weather conditions were taken into account. The locations are assumed to be known and in GPS format. The weather conditions are extracted in near real time from the Sentinel-3 mission API [23].

To determine the optimal speed and corresponding fuel oil consumption, we employ a gradient descent algorithm for each possible route segment. The algorithm uses wind data, speed exploration values, as well as some control variables to estimate internal machinery values, which in turn are employed to estimate the consumption, and the minimal value gets selected. By applying this optimization method, a reduction of about 3% of the fuel oil consumption was obtained.

Lesson Learned

A general lesson learned in this study is the importance of data preparation to control input data quality. An observation considering the reliability of different sensors which varied across the ships that lead to many outliers, which negatively impacted the model accuracy, and the creation of a unified model. In addition, the timelines of wind data provided by the Sentinel mission posed limits on the methods, as they were provided once or twice a day, depending on the region of interest, and accurate forecast of wind for a period of over six hours turns out to be better by classic meteorological methods than by statistical approach.

10.4.2.2 Analyzing Historical Measurement Data

Different data analysis methods were used, e.g., in a fishery pilot (see PartVII of this book) where measurements from several fishing ship motors were analyzed using VTT OpenVA application. The main goal was to analyze ship fuel consumption, but since we imported all the measurements available, the system can be used to analyze other variables as well. VTT OpenVA is an advanced analytics solution that was tailored to create an application where a data scientist can select different measurements and get different visualizations based on the user selection. In a DataBio fishery pilot case, there are 115 measurements from the motors four different ships that were analyzed. Measurements were stored in a 10s interval from a four-year period. More three billion measurement values were stored in a standard relational database (PostgreSQL, https://www.postgresql.org/).

An analysis application using the data was implemented. Users of the application can select measurements from different ships on the selected time period, and the VTT OpenVA proposes available analysis methods, based on the measurement type. In the pilot, 18 different analysis results were shown as visualizations, and 55 single value performance indicators were calculated.

Lessons Learned

VTT OpenVA is designed to be an interactive application, but DataBio experiences show that when the amount of measurements becomes large—billions of measurements—it is hard to achieve real-time interaction, because database query response times grow to at least several seconds. This could be mitigated by more powerful database servers and specialized commercial databases, but the goal of the pilot was to use standard servers and databases, allowing easy transfer of the system to the server maintained by the system users.

To make queries faster, VTT OpenVA automatically distributes measurements into a large number of database tables instead of making queries to one huge table. Naturally, the way users use this kind of big data analysis tool, e.g., in the fishery pilot the normal time period of the data that is in practice analyzed is about three weeks, which is the average time that a fishing ship is out from harbor. Even though querying the data takes some time, actual time taken to analyze and visualize these relative small data sets is quite short.

10.4.2.3 Oceanic Tuna Fisheries Immediate Operational Choices

Exus analytics framework was integrated in the pipeline of pilot fishery to predict main engine performance and faults in advance. For the prediction of the main engine performance, a neural network was used to perform multivariate regression in order to estimate a regression model for multiple variables taking as input a considerably lower number of values. The main benefits of the neural networks are their ability to capture complex relationships between the inputs and their requirement of high number of data. The choice of this machine learning algorithm was also based on related work for fault diagnosis in engines used in vessels [24, 25].

Based on historical vessel data sets, a preprocessing of two stages is applied. First only data that corresponds to the steady state of the engine is considered. After extracting the steady-state engine data, the min-max normalization is applied for all features.

Various architectures for the number of hidden layers and units have been tested, and for the best model selection, the data set has been split into training and validation sets. The model that performed the lowest validation error is selected as the best one.

For the prediction of engine faults, the predicted variables (based on historical data) are compared to the actual values of the vessel measurements. When the variance of these differences is higher than a threshold, it is considered as an engine fault.

Lessons Learned

In this pilot, we have partial lack of knowledge about when actual faults happened or when variance on the values is due to the wear and tear. For that, only the steady data is considered for fitting the best line, so trends can be identified and when actual measurements appear different behavior from the normal trends give an early warning, even though the setting of thresholds for the identification of abnormal behavior of the vessel is challenging due to the variation of the historical data sets (e.g., different periods/years might report different statistical measurements) (Fig. 10.3).

10.4.2.4 Real-Time Data Classification for Automatic Fish Detection

The main goal of the experiment was to deploy an effective classification approach, relying only on acoustic data, that can form the basis of a real-time fish detection tool.

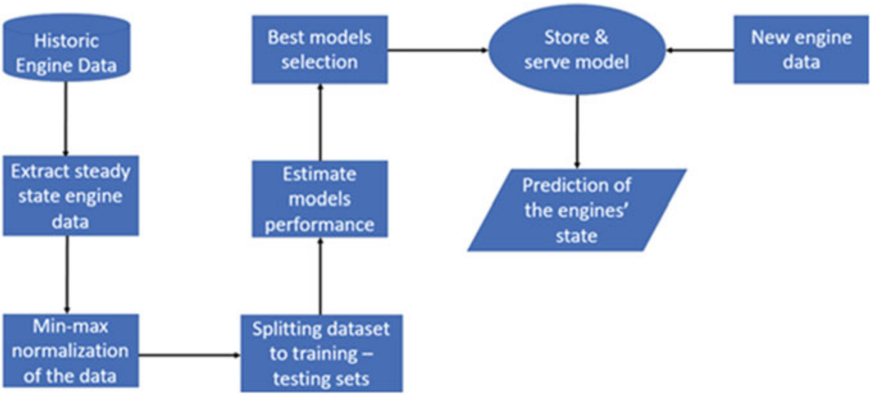
Historic Engine Data
Extract steady state engine data
Min-max normalization of the data
Splitting dataset to training – testing sets
Estimate models performance
Best models selection
Store & serve model
New engine data
Prediction of the engines' state

Fig. 10.3 Workflow of the Oceanic tuna fisheries immediate operational choices—pilot

For the study, echosounder sample output was appropriately preprocessed in order to produce the *mean volume backscattering strength* (MVBS) values for five frequencies: 18, 38, 70, 120, and 200 kHz. The problem with echosounder data is that the data set is quite unbalanced with respect to the presence of fish or not. In the samples that we used, about 5% of the measurements correspond to fish presence, while 95% measurements not. As a result, a random classifier can appear falsely effective.

To tackle this problem, the acoustic data set was resampled before being fed to the classifiers of the study. The comparison was made based on the *kappa coefficient*, which is more reliable in cases of unbalanced data sets. The methods tested were Naïve Bayes, K-nearest neighbors (K-NN), and SVM, both with linear and radial kernels. PCA was also examined as a preprocessing method. All classification approaches were tested on MVBS values for different combinations of the five frequencies measured.

Lessons Learned on the DataBio Use

From the process and the analyses carried out within DataBio and with respect to the specific pilot, the main conclusion and lesson learned are that many different classification algorithms should be tested, in order to identify the most efficient ones for the specific data set types. Because of the nature of the acoustic data sets, it was really challenging to identify the proper training subsets for the machine learning algorithms. This resulted in the need for a number of iterations with the pilot owner (SINTEF) in order to ensure that the algorithms are accurate enough.

References

1. Tan, P. N., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining*, First edition, Addison Wesley.

2. Alpaydin, E. (2020). *Introduction to machine learning*. MIT press.
3. Wong, P. C. (1999). Guest editor's introduction: Visual data mining. *IEEE Computer Graphics and Applications, 19*(5), 20–21.
4. Ferreira de Oliveira, M. C., & Levkowitz, H. (2003). From visual data exploration to visual data mining: A survey. *Visualization and Computer Graphics, IEEE Transactions, 9*(3), 378–394.
5. Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons, 62*(1), 15–25.
6. Arel, I., Rose, D. C., & Karnowski, T. P. (2010). Deep machine learning-a new frontier in artificial intelligence research. *IEEE Computational Intelligence Magazine, 5*(4), 13–18.
7. Frost & Sullivan. (2013). Global big data analytics market, forecast to 2023. RESEARCH CODE: K2AF-01-00-00-00, Frost & Sullivan.
8. "Machine Learning Market by Service (Professional Services, and Managed Services), for BFSI, Healthcare and Life Science, Retail, Telecommunication, Government and Defense, Manufacturing, Energy and Utilities, Others: Global Industry Perspective, Comprehensive Analysis, and Forecast, 2017–2024" (2019), Zion Market Research.
9. Amar, R., Eagan, J. & Stasko, J. (2005). Low-level components of analytic activity in information visualization. In J. T. Stasko & M. O. Ward (eds) *IEEE Symposium of Information Visualization (INFOVIS) 2005*, *IEEE Computer Society*, 23–25 Oct., p. 111.
10. Hand, D. J., Mannila, H., & Smyth, P. (2001) *Principles of data mining*, First edition, MIT press.
11. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.
12. Russwurm, M., & Koerner, M. (2017). Multi-temporal land cover classification with long short-term memory neural networks. *ISPRS—International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 42,* 551–558.
13. Mou, L., Ghamisi, P., & Zhu, X. X. (2017). Deep recurrent neural networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing, 55,* 3639–3655.
14. Purwar, P., Rogotis, S., Chatzipapadopoulus, F., Kastanis, I. (2019). "A reliable approach for pixel-level classification of land usage from spatio-temporal images". In *2019 6th swiss conference on data science (SDS)* (pp. 93–94).
15. Huang, Z. (1997). Clustering large data sets with mixed numeric and categorical values. In *Proceedings of the 1st pacific-asia conference on knowledge discovery and data mining, (PAKDD),* pp 21–34.
16. Draper, N., Smith, H. (1981). *Applied regression analysis.* Wiley.
17. Habyarimana, E., Piccard, I., Zinke-Wehlmann, C., De Franceschi, P., Catellani, M., Dall'Agata, M. (2019). Early within-season yield prediction and disease detection using sentinel satellite imageries and machine learning technologies in biomass sorghum. *Lecture Notes in Computer Science*, *11771*, 227–234. https://doi.org/10.1007/978-3-030-29852-4_19.
18. George, A. F. H., Houghton, J. D, & Brown, S. B. (1987). Tansley review no. 11. the degradation of chlorophyll—a biological enigma. *The New Phytologist*, *107*(2), 255–302.
19. Mikołajczyk, A., Michał, G. (2018). Data augmentation for improving deep learning in image classification problem. In *2018 international interdisciplinary Ph.D. workshop (IIPhDW)* (pp. 117–122). IEEE.
20. He, K., Gkioxari, G., Dollár, P., Girshick, R. (2017). Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision* (pp. 2961–2969).
21. Xgboost. https://xgboost.readthedocs.io/en/latest/get_started.html. Accessed: 2019.
22. Hastie, T., Gareth, J., Witten, D., Tibshirani, R. (2014). An introduction to statistical learning.
23. Sentinel-3 api. https://coda.eumetsat.int/#/home. Accessed: 2019.
24. Antory, D., et al. (2005). Fault diagnosis in internal combustion engines using non-linear multivariate statistics. *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering, 219*(4), 243–258.
25. Basurko, O. C., & Uriondo, Z. (2015). Condition-based maintenance for medium speed diesel engines used in vessels in operation. *Applied Thermal Engineering, 80,* 404–412.