Outlier quantification for multibeam data

Tobias Ziolkowski[1,2] · Agnes Koschmider[2] · Peer Kröger[3] · Colin Devey[1]

Abstract

This paper discusses the challenges of applying a data analytics pipeline for a large volume of data as can be found in natural and life sciences. To address this challenge, we attempt to elaborate an approach for an improved detection of outliers. We discuss an approach for outlier quantification for bathymetric data. As a use case, we selected ocean science (multibeam) data to calculate the outlierness for each data point. The benefit of outlier quantification is a more accurate estimation of which outliers should be removed or further analyzed. To shed light on the subject, this paper is structured as follows: first, a summary of related works on outlier detection is provided. The usefulness for a structured approach of outlier quantification is then discussed using multibeam data. This is followed by a presentation of the challenges for a suitable solution, and the paper concludes with a summary.

Introduction

Data analytics techniques such as data mining and machine learning can give valuable insights into the data. They allow rules that describe specific patterns within the data to be identified or can reveal hidden knowledge. Based on the analysis results, informed decisions can be made.

The most time-consuming step in the analytics pipeline from processing raw data to discovering knowledge is data pre-processing. This step includes activities for data integration, data enhancement, data transformation, data reduction, data discretization and data cleaning. The reason for the time-consuming nature of this activity is usually the quality of the data (i.e. missing or incomplete entries). Some approaches to improving quality can be found in the litera-ture [18]. These approaches are usually based on detecting and filtering of outliers. In statistics, outliers are defined as "*high measurements where the value is some standard deviation above the average*" [5]. In data engineering, outliers, commonly referred to as "*anomalies*", refer to "*something that is out of range*". This can, on the one hand, point to insignificant data or, on the other hand, to interesting and useful information about the underlying system. Hence, distinguishing the essence of outliers in terms of undesired or unwanted behavior versus surprisingly correct and informative data is of particular interest for the quality of data analysis.

The purpose of our work is to develop an outlier quantification framework making the analysis results explainable. As a use case, we selected ocean science (multibeam) data to calculate the outlierness for each data point. The benefit of outlier quantification is a more accurate estimation of which outliers should be removed or further analyzed. Fig. 1 shows, on the left, the convential process of outlier detection. The data is pre-processed and outlier techniques are interweaved in this step, resulting in analysis results such as clusters or patterns. The right-hand side of Fig. 1 shows a new approach to outlier detection. Outlier information is propagated through each step of the process from raw data to the analysis results in terms of meta-data annotations.

Although plenty of approaches exist that classify, filter and remove outliers, the number of approaches for explainable outlier quantification is limited. To shed light on the subject, this paper is structured as follows: the next section summarizes related works on outlier detection. The useful-

✉ Tobias Ziolkowski
  tziolkowski@geomar.de, tzi@informatik.uni-kiel.de

  Agnes Koschmider
  ak@informatik.uni-kiel.de

  Peer Kröger
  pkr@informatik.uni-kiel.de

  Colin Devey
  cdevey@geomar.de

1   Helmholtz Centre for Ocean Research Kiel, Kiel, Germany

2   Group Process Analytics, Computer Science Department, Kiel University, Kiel, Germany

3   Group Information Systems and Data Mining, Computer Science Department, Kiel University, Kiel, Germany

Fig. 1 *Left-hand side*: the process from raw data to clustering without outlier quantification. *Right-hand side*: the process with outlier quantification. Outliers are continuously annotated within the analytics pipeline
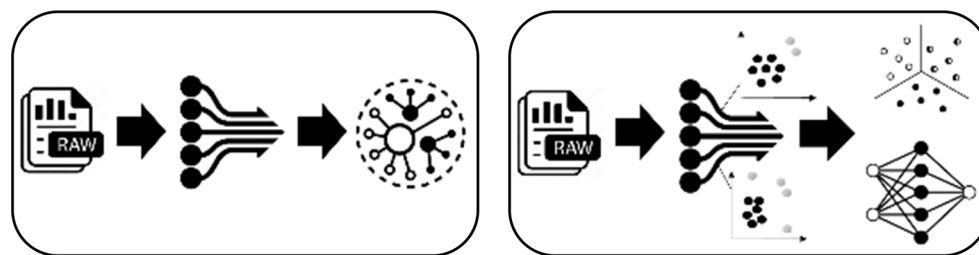
ness for a structured approach of outlier quantification is discussed with a use-case scenario using multibeam data. Finally, the last section sketches challenges for a suitable solution and concludes the paper with a summary.

Related work

Existing outlier detection methods differ in the way they model and find the outliers and, thus, in the assumptions they rely on, implicitly or explicitly. In statistics, outlier detection is usually addressed by modelling the generating mechanism(s) of the normal data instances using a single or a mixture of multivariate Gaussian distribution(s) and measuring the Mahalanobis distance to the mean(s) of this (these) distribution(s). Barnett and Lewis [1] discuss numerous tests for different distributions in their classical textbook. As a rule of thumb, objects that have a distance of more than $3 \cdot \sigma$ to the mean of a given distribution (σ denotes the standard deviation of this distribution) are considered as outliers to the corresponding distribution. However, we are not aware of any approach that continuously tracks the outlier scores and updates the values within the analytics pipeline. Problems of these classical approaches are obviously the required assumption of a specific distribution in order to apply a specific test.

According to the data distribution, there are tests for univariate as well as multivariate data distributions, but all tests assume a single, known data distribution to determine an outlier. A classical approach is to fit a Gaussian distribution to a data set, or, equivalently, to use the Mahalanobis distance as a measure of outlierness. Sometimes, the data are assumed to consist of k Gaussian distributions and the means and standard deviations are computed data driven. However, mean and standard deviation are rather sensitive to outliers and the potential outliers are still considered for the computational step.

Related to the outlier detection techniques, many different approaches exist that have less statistically oriented but more spatially oriented ways of modelling outliers, particularly using distances between data objects. These models consider the number of nearby objects, the distances to nearby objects and/or the density around objects as an indication of the "outlierness" of an object [2, 10, 12, 13,

15]. However, all these approaches rely implicitly on the assumption that a globally fixed set of features (usually all available attributes) are equally relevant for the outlier detection process.

Outlier detection addresses the problem of discovering patterns in data that do not replicate the expected behavior. Although many approaches for outlier detection using supervised machine learning [6, 8] or signal processing based methods [9, 11] exist, the risk to unintendedly eliminate necessary signals if the sound data is unknown is present and a holistic approach is missing that combines different techniques, data distributions and tests and aim to provide a quantification.

The related work analysis attempts to identify apparent trends towards outlier detection, different techniques to find outliers and filter them. The next section discusses a suitable use case for outlier quantification. Particularly, we discuss multibeam (bathymetric) data for seafloor classification.

Use-case scenario

Pre-processing of bathymetric data is a time consuming task. Due to new technologies for data acquisition, in which a fan-shaped bundle of acoustic beams ("multibeams") is repeatedly transmitted (each transmission being called a "ping") from the ship perpendicular to the direction of travel (see first image in Fig. 2), a huge amount of data is collected. Not only the amount of data increases, but the data is noisy and contains many outliers.

Although the amount of data continues to grow, data processing steps, like outlier detection and filtering, are carried out manually by domain experts. This task is repetitive and subjective, so there is a need to ensure objectivity and a cleaning procedure which ensures traceability for outlier detection. In order to meet these goals, artificial neural networks (ANN), especially supervised machine learning (ML) methods, can be applied to reduce processing time and ensure objectivity and traceability. Figure 2 shows the pipeline for outlier quantification in multibeam data.

As a suitable use case for outlier quantification multibeam bathymetry raw data of RV MARIA S. MERIAN during cruise MSM88 [19] with records in the Atlantic that took place between 2019-12-19 and 2020-01-14 could be

Fig. 2 Pipeline for outlier quantification in multibeam data with an artificial neural network (*ANN*). *CSV* comma-separated values

multibeam data → data selection → feature calculation → create csv → ANN training → ANN → outlier detection → analyze detected outlier
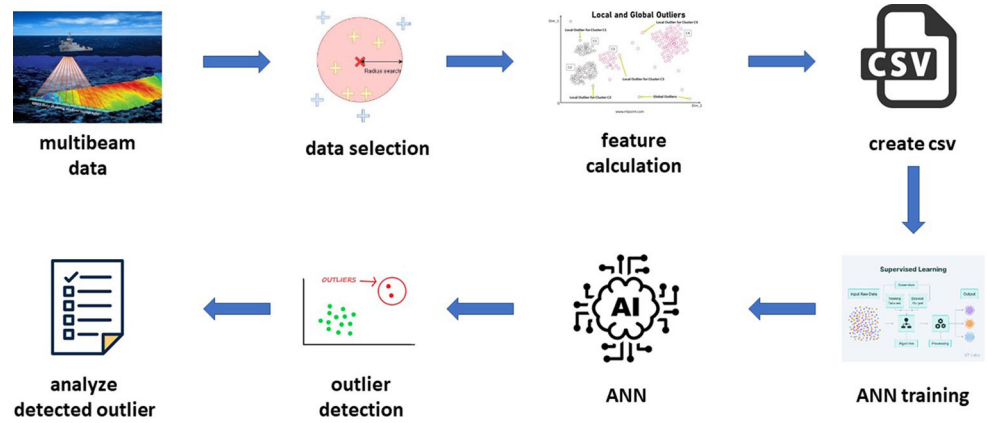

Fig. 3 Route of the MSM88/2 expedition between Cape Verde and Barbados. RV MARIA S. MERIAN cruise is located between *red anchor points*

used. The data were collected using the Kongsberg EM 122 system and cover an area of 153,121 square kilometres.

Next, the following analytics pipeline can be applied for this data set. *Multibeam data* is saved as .all files. The depth amplitude goes from 5244 to 5840 m. Figure 3 shows the location of the survey in the Atlantic. Firstly, multibeam data is transformed into a generally readable comma-separated values (csv) format containing latitude, longitude and depth values. Additionally, the backscattering strength (BS) is calculated and added to the csv file. BS data is a measure of intensity of the acoustic return and is used to detect and quantify the bottom echoes, so several seabed types like coral reefs, seagrass, salt or mud can be taken into account.

A prerequisite for supervised learning is the need for labelled data. So, for outlier detection a domain expert manually labelled all outliers in the collected data set. Each sounding thus receives an additional attribute and a flag is saved. The data set is 59.5 GB in size, so that the usual data processing steps cause high computational costs and the runtime for processing the data is very high. This challenge is described further in the next section.

A moving window data pattern can be applied to the data for *data selection*. Moving window algorithms are data-centric, because the moving window changes position iteratively while being centered on one sounding. The red cross in Fig. 2 for data selection is the centre of the mov-

ing window and the yellow plus signs are the points being selected to calculate the local neighbourhood. Only one parameter, which is the search radius around the sounding, is needed. Although the method is time consuming, the local neighbourhood calculation is representative and is suitable for detecting the local neighbourhood for each sounding. Local neighbourhoods are saved in an additional file.

In order to train ML algorithms to automatically detect outliers in multibeam data, a proper description of
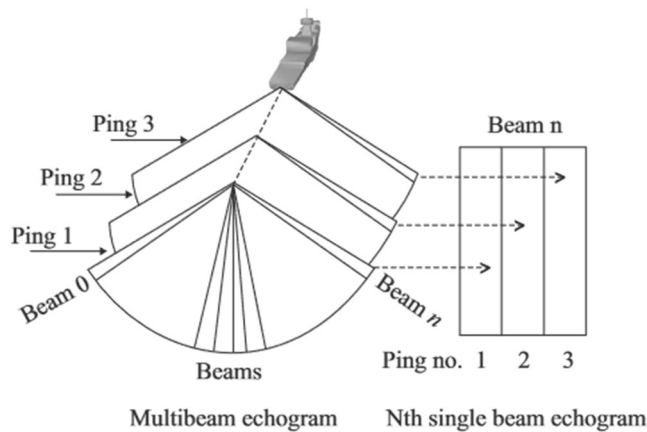

Fig. 4 Definition of a beam and a ping presented in ping/beam view [16]

Fig. 5 Georeferential (spatial) view of multibeam data [7]

the soundings is needed. We use the local neighbourhood, to calculate *features* for each sounding. These features are used for ANN training, so a trained network is generated, which is able to detect and flag outliers in multibeam data.

Depending on the attributes which should be calculated for each local neighbourhood, the raw view, spatial view or sequential view is suitable. Multibeam data can be handled with a dual representation; a ping/beam view as time-series where data is stored in a matrix (see Fig. 4) or as an absolute georeferential view where each sounding is represented as a triplet containing latitude, longitude and depth values (see Fig. 5). Raw features are based on the raw data set collected on the MSM88 expedition like the BS or the depth values. Spatial features include attributes like a local outlier factor or the standard deviation for each local neighbourhood. The sequential view is suitable for bad ping detection.

All calculated attributes are added to the *csv file* to gain metadata and a proper description of each sounding that can be utilized to train ML algorithms. These data and their associated description are used by ML algorithms for training, so in this use case these data are the basis to decide whether a sounding is an outlier or not.

To evaluate this approach, MB-System can be applied to the dataset to automatically detect outliers with the implemented outlier detection methods. MB-System is an open source software package for the processing and display of bathymetry and backscatter imagery data derived from multibeam, interferometry and sidescan sonars. MB-System detects outliers with simple interpolation methods or by adoption of alternate values. Finally, all detected outliers by MB-System can be contrasted to the outliers detected with the presented supervised ML approach to verify the accuracy.

Conclusion and research challenges

This paper discusses the challenges of applying a data analytics pipeline for a large volume of data as can be found in natural and life sciences. To address this challenge, we attempt to elaborate an approach for the improved detection of outliers. We discuss an approach for outlier quantification for bathymetric data. The approach presented in this paper contributes to the concept of cross domain fusion (CDF) as follows. The data-driven pipeline presented in this paper aims to replace or complement the predominately used model-driven approach in the domain of seafloor classification. We are convinced that a data-driven approach can give more insights than traditional approaches do. For this, however, several challenges must be addressed in order to provide a solution.

Challenge 1 Disciplines like natural and life-cycles have a large volume of data. This calls first for techniques to efficiently pre-process the data. We found that conventional pre-processing must be fine-tuned and adjusted to run algorithms for data integration and transformation. Even then it is difficult to calculate and summarize all features in one data set needed for training to enable the ANN to detect outliers. Moreover, the resulting csv file will be very large,

so that the training of the ANN, depending on the method used, is also challenging. For example, linear regression to predict a binary target is simple to implement, but there is a risk of underfitting.

Challenge 2 The number of approaches to accurately recognize objects is limited. While these techniques have been deeply studied for shallow water for instance [4, 14, 17], they fail for deep sea. Seafloor classification tasks should satisfy the precondition that the area covered by several consecutive pings belongs to the same seafloor type [3]. This precondition is easily met in shallow water, but it is difficult to ensure in the deep sea because, due to the fan-shaped nature of the beam bundle, the width of seafloor insonified by one ping is proportional to depth, and so consecutive pings cover a much larger area. This shows how challenging object recognition is in large data sources with certain properties like depth.

Challenge 3 Due to the complex pre-processing of the data, there is a great range of uncertainty in the analysis result. The analysis result should be interpreted more as a fuzzy value with a certain range. In addition, ANN methods like gradient boosting are very fast and powerful, but the results are not easily interpretable. Once again, this hampers transparency and explainability of the analysis result.

References

1. Barnett V, Lewis T (1994) Outliers in statistical data, 3rd edn. John Wiley&Sons. ISBN 978-0-471-93094-5
2. Breunig MM, Kriegel HP, Ng RT, Sander J (2000) Lof: Identifying density-based local outliers. Proc. ACM Int. Conf. On Management of Data (SIGMOD)
3. Fonseca L, Calder B (2007) Clustering acoustic backscatter in the angular response space. Proceedings of the US Hydrographic Conference, Norfolk
4. Fonseca L, Brown C, Calder B, Mayer L, Rzhanov Y (2009) Angular range analysis of acoustic themes from Stanton Banks Ireland: a link between visual interpretation and multibeam echosounder angular signatures. Appl Acoust 70:1298–1304
5. Freedman D (2005) Statistical models : theory and practice. Cambridge University Press, Cambridge
6. Gu X, Akoglu L, Rinaldo A (2019) Statistical analysis of nearest neighbor methods for anomaly detection. In: Proceedings of the 33rd conference on neural information processing systems (NIPS2019) Vancouver, 8–14 December 2019, pp 10921–10931
7. Hey R, Martinez F, Höskuldsson A, Eason ED, Sleeper J (2016) Multibeam investigation of the active North Atlantic plate boundary reorganization tip. Earth Planet Sci Lett. https://doi.org/10.1016/j.epsl.2015.12.019
8. Hsu J, Wang Y, Lin K, Chen M, Hsu JH (2020) Wind turbine fault diagnosis and predictive maintenance through statistical process control and machine learning. IEEE Access 8:23427–23439
9. Ikeda S, Toyama K (2000) Independent component analysis for noisy data—MEG data analysis. Neural Netw 13:1063–1074
10. Jin W, Tung A, Han J (2001) Mining top-n local outliers in large databases. Proc. ACM Int Conf on Knowledge Discovery and Data Mining (KDD)
11. Koganeyama M (2003) An effective evaluation function for ICA to separate train noise from telluric current data. In: Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003 Nara, 1–4.April 2003, pp 837–842
12. Kriegel H-P, Kröger P, Schubert E, Zimek A (2009) LoOP: local outlier probabilities. Proc. Int Conf. On Information and Knowledge Management (CIKM)
13. Kriegel HP, Kröger P, Schubert E, Zimek A (2011) Interpreting and unifying outlier scores. Proc. SIAM Int. Conf. On Data Mining (SDM)
14. Lamarche G, Lurton X, Verdier A, Augustin J (2011) Quantitative characterisation of seafloor substrate and bedforms using advanced processing of multibeam backscatter. Application to Cook Strait, New Zealand. Cont Shelf Res 31:93–S109
15. Liu FT, Ting KM, Zhou ZH (2012) Isolation-based anomaly detection. ACM TKDD 6(1):1–39. https://doi.org/10.1145/2133360.2133363
16. Myounghee K (2011) Analysis of the ME70 multibeam echosounder data in echoview—current capabilityand future directions. J Mar Sci Technol. https://doi.org/10.51400/2709-6998.2197
17. Preston JM (2009) Automated acoustic seabed classification of multibeam images of banks. S Appl Acoust 70(10):1277–1287
18. Suriadi S, Andrews R, ter Hofstede AHM, Wynn MT (2017) Event log imperfection patterns for process mining: Towards a systematic approach to cleaning event logs. Inf Syst. https://doi.org/10.1016/j.is.2016.07.011
19. Wölfl A-C, Devey CW (2020) Multibeam bathymetry raw data (Kongsberg EM 122 entire dataset) of RV MARIA S. MERIAN during cruise MSM88/2. GEOMAR—Helmholtz Centre for Ocean Research, Kiel https://doi.org/10.1594/PANGAEA.918716