A introduction to Data Mining and its applications with Point Clouds

Matthias Weil
School of Computation, Information and Technology, Informatics
Techincal University Munich
Munich, Germany
matthias.weil@tum.de

***Abstract*— In this paper, we focus on the necessary steps in Data Mining and how this process is applied with Point Clouds in Geospatial Data. We will shine light on several steps of a Data Mining Pipeline, including Data Cleaning, Feature Enginnering followed by model selection and important aspects that have to be considered when training the selected model.**

***Index terms*—Data Mining, Geospatial Data, PointClouds, 3D, Feature Extraction, Classification, Co-Registration**

I. Introduction

Data Mining and extracting data from Point Clouds are necessary skills when working with Geospatial Data sets. Many methods and how they can be practically applied are explained.

II. Data Mining

A. Data Cleaning

It is not uncommon for the data obtained to be of poor quality and difficult to work with. This may include the presence of missing values, outliers, or spelling mistakes, which are particularly prevalent in user-generated data. To enhance the comparability of the data, various approaches can be employed. If the data follows a normal distribution and statistical models that assume such a distribution are to be used, standardisation is the most common method. Otherwise, normalisation is the most frequently used technique to improve data comparability. Standardisation is not susceptible to outliers, and thus the removal of these as a preliminary stage can be omitted.

B. Feature Extraction

Machine learning algorithms require the given data to be in certain numerical formats. Additionally this allows the machine learning algorithm to receive a more effective set of inputs, which increases accuracy and lowers the computational need for machine learning algorithms.

1) *Over- and Underfitting:*

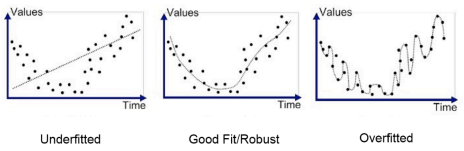

Figure 1: A common problem when working with a machine learning algorithm to create a model that discoveres patterns in unknown data. On the left is Underfitting, when the model is too simple to capture the complexity of the data. In the middle is the desired outcome of a data model. On the right is the example of Overfitting which occurs when a machine learning algorithms detects noise as data. Figure from [1].

C. Feature Selection

Feature selection is a critical part in data mining due to numerous reasons. As only relevant features are being considered, the dimensionality of the data set is reduced and is more efficient, accurate and less prone to overfitting. Furthermore less features lead to a better understanding of the learning result. The following figure explains the Hughes phenomenon, where a higher amount of features deteriorates the model's performance. Additional features may introduce noise, which lead to a decrease in classification accuracy [2].

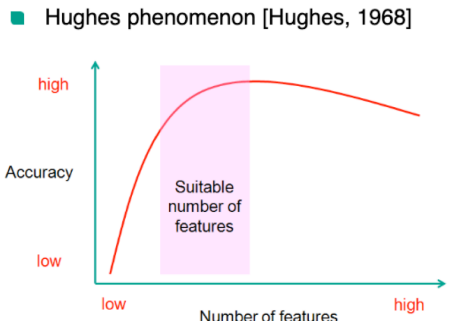

Figure 2: Hughes phenomenon, a increase of number of features decreases the accuracy of the classification model. Figure from [3].

There are three notable categories of feature selection. Filter-based methods that include statistical tests to assess the correlation between various features. These filters also remove features with low variance, since they are deemed to have little information. Because they are classifier-independent, simple

and efficient they are often used. Alternativley Wrapper-based methods or Embedded methods can be used[4].

D. Model Selection

Two overarching themes emerge when selecting a model: supervised and unsupervised learning models. Supervised machine learning includes classification which is heavily used in image recognition. Example of classification models that categorize data into labels, are Decision Trees, Support Vector Machines and k-Nearest Neighbors. Supervised learning also includes regression models which describes a functions that models the correlation of a independent variable and a target variable. The correlation of systolic blood pressure and amount of coffee a person drinks is a example of a linear regression.

In unsupervised classification the algorithm tries to assign each data point to a cluster. The number of clusters to be extracted depends on the algorithm. In the case of k-means, the user must input the desired number of clusters. Alternatively, in the case of DBSCAN, the user must definde the distance between two points to form a new cluster. These algorithms work without training labels, so the user has to understand the output clusters.

Deep learning as the machine learning algorithm called representation learning harvest the benefits, that the algorithm learns the features from the data itself. The algorithm is then capable of representing the data in a way that eases classification or regression [5].

E. Model Training

Models are prone to be overfitted when too many features are selected, sufficient data not being available to train the model or if the data is noisy. This noise may easily be detected as a pattern by the model. In order to detect overfitting the data set is split into a training and a validation set. After training, the model is presented the validation set and the perfomance is tracked. If they diverge, then it can be concluded that the model is overfitting. To tackle underfitting additonal features can be selected, the complexity of the model can be raised or a increase of duration of training the model.

F. Model Evaluation

To evaluate the model there are various methodlogies that may be applied. For regression Mean Absolute Error or Mean Squared Error are often applied. To evaluate a classification model one can calculate the Accuracy which is the proprotion of correctly classified instances. Furthermore, the Precision, that measures the quality of positive predictions, Recall, which details the number of correct predictions are being made, and the F1-score, which is the Harmonic mean of precision and recall, illustrate various metrics to evaluate a classifier.

III. Geospatial Data Mining

In the following we showcase a typical workflow for classification that involves first having a point cloud that can be exapnded by co-registration. Then neighborhood selection, followed by feature extraction and feature selection. Lastly we choose our classification model.

A. Co-Registration

When working with point cloud data there may be two acquisitions from the same structure, therefore it is benefical to apply Co-registration to combine the multiple point clouds of interest. Typical algorithms include iterative closest point alogrithm or feature based matching algorithms.

Figure 3: Co-registration of two point clouds, showing the Arc de triomphe from different angles. Here a feature based matching algorithm, along the corners of the building, might be more computationaly efficient. Figure from [6].

As can be seen in Figure 3, with two point clouds of the Arc de triomphe in paris, the co-registrated point cloud has more information about the subject of interest. Often co-registration is conducted to detect changes in the area, for example after natural disasters or after longer periods of time to show change in the scanned area.

B. Segmentation

Multiple homogenous regions or clusters with similar properties. There are different approaches to do this. Firstly euclidean clustering, then k-nearest neighbors and lastly spatial neighborhoods.

C. Neighborhood Selection

When working with point cloud data calculations based on neighborhoods are among the most important ones. They are necessary for any filtering, smoothing or interpolation step and for information extraction.

Spatial neighborhoods are defined either by a certain distance to other points or to a certain amount of fixed neighbors [3]. If its a certain distance then its considered a cylindircal neighborhood since it runs along the x and y axis. Otherwise if its 3D then they are called spherical neighborhoods because their height is also relevant. A point cloud neighborhood can also be defined by the fixed amount of neighbors. Fixed neighbors are also commonly called k-nearest neighbors and the size can be highly variable depending on the point density at that area.

Various forms of information can be supplied by a spatial neighborhood. A easily obtainable one is the point density. With the following formula the amount of points contained inside of a 1 m radius can be calculated.

$$D = \frac{k+1}{\frac{4}{3}\pi r^3 kNN}$$

With this information outliers can be removed easily through statistical outlier removal provided by the Point Cloud Library.

D. Feaure Extraction

After neighborhoods are defined, various features can be extracted. 3D-features can be calculated by using eigenvalues or looking at the geometric properties. Spatial neighborhoods allow the derivation of local surface roughness.

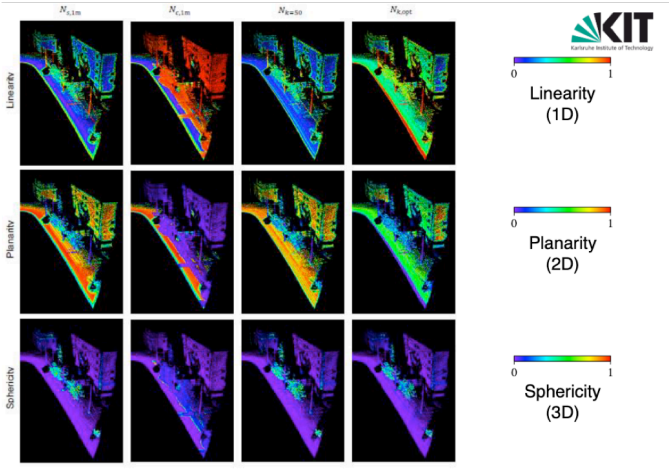

Figure 4: Features that can be easily derived from eigenwerte. Figure from [7].

Ground removal is a important procedure when working with point cloud data. Ground points are always static, so it is not cost-effective to process them, using up time and computing power. Notable algortihms for this purpose are RANSAC (Random Sample Consensus), Ground Plane Fitting and Patchwork++.

E. Feature Selection

F. Classification

The extracted features can then be used as input for classifieres that have been trained with representable data. Different classifiers can be used like, Random Forest, Support Vector Machines, Nearest Neighbor classifiers, Decision Tree, naïve bayesian classifier or Linear Discriminant Analysis.

References

[1] "over and Underfitting." [Online]. Available: https://medium.com/greyatom/what-is-underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6803a989c76

[2] "On the mean accuracy of statistical pattern recog- nizers.," 1968.

[3] M. Weinmann, B. Jutzi, S. Hinz, and C. Mallet, "Semantic point cloud interpretation based on optimal neighborhoods, relevant features and efficient classifiers," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 105, pp. 286–304, 2015, doi: https://doi.org/10.1016/j.isprsjprs.2015.01.016.

[4] J. B. Blomley R. and M. Weinmann, "Classification of air- borne laser scanning data using geometric multi-scale features and different neighbourhood types.," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2016, [Online]. Available: https://isprs-annals.copernicus.org/articles/III-3/169/2016/isprs-annals-III-3-169-2016.pdf

[5] [Online]. Available: https://3dgeo-heidelberg.github.io/etrainee/module3/05_pointcloud

[6] "point-cloud-registration." [Online]. Available: https://www.thinkautonomous.ai/blog/point-cloud-registration/

[7] M. Weinmann, B. Jutzi, C. Mallet, and M. Weinmann, "GEOMETRIC FEATURES AND THEIR RELEVANCE FOR 3D POINT CLOUD CLASSIFICATION," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pp. 157–164, 2017, doi: 10.5194/isprs-annals-IV-1-W1-157-2017.